ARTIFICIAL NEURAL NETWORK TECHNIQUES TO INVESTIGATE POTENTIAL INTERACTIONS BETWEEN BIOMARKERS

CHRISTOPHE LEMETRE

A thesis submitted in partial fulfilment of the requirements of Nottingham Trent University for the degree of Doctor of Philosophy

September 2010

To Mathilde and my parents...

Copyright statement

Acknowledgements

First of all I would like to deeply thank my supervisor Prof. Graham Ball as well as Dr Lee Lancashire without whom none of this work could have seen the day. Their tremendous help from the ideas and concepts up to this very thesis, and their endless support and motivation have been so much appreciated that there would be hardly a way to thank them.

A great thank should go as well to Prof. Robert Rees for his valuable support, advice and motivation.

A great thank you goes to my parents. They never let me down, even in the moments of doubt and lack of self-confidence. Their eternal and unconditional support have been and will always be a deep source of inspiration and a constant source of perseverance, without which all of this would not have been possible.

I would like to thank Mathilde for her unvaluable motivation, her trust in me was incredibly helpful and helped to get the self-confidence needed. I hope one day, I will have an opportunity to give her back the same. She's the greatest fiance any man could ever imagine.

Finally I would like to thank all my friends from Nottingham: Maria, David, Cristina, Juan, Souvik, Hans and all the others. The time spent altogether helped to leave the most difficult moments behind, and forget about things too serious.

Abstract

High-throughput technologies in biomedical sciences, including gene microarrays, supposed to revolutionise the post-genomic era, have barely met the great expectations they inspired to the biomedical community at first.

Current efforts are still focused toward improving the technology, its reproducibility and accuracy. In the meantime, computational techniques for the analysis of the data from these technologies have achieved great progresses and show encouraging results.

New approaches have been developed to extract relevant information out from these results. However, important work needs to be further conducted in order to extract even more meaningful and relevant information.

These techniques offer great possibilities to explore the overall dynamic held within a living organism. The potential information contained in their output can reveal important leads at deciphering the interconnection, interaction or regulation influences that can exist between several molecules.

In front of an increasing interest of the scientific community toward the exploration of these dynamics, some groups have started to develop solutions based on different technologies to extract these information related to interactions. Here we present an Artificial Neural Network-based methodology for the study of interactions in gene transcriptomic data. This will be applied and validated in a breast cancer context.

This manuscript will discuss the methodological optimisation to identify biomarkers of interest from high-throughput transcriptomic technologies; and it will show how the algorithms were brought forward to identify the potential relationship that may exist between the markers identified. It will illustrate and highlight the robustness of the methods by discussing some examples of application in different breast cancer studies.

The present thesis will show that despite the great difficulty to obtain gold validation to prove the robustness of the approach; it has been possible to identify some relevant features able to highlight the promises held by this preliminary development of the method. The results obtained by trying to

identify the correlated component within an artificial dataset suggest some interesting ability of the approach. Additionally, when applied to the van't Veer dataset (van't Veer et al., 2002), the list of selected transcrpits held two different isoforms for two different genes, and the method identified the strong correlation between the 2 forms. Finally, the results involving the transcripts for DTL, TK1 and CDC45L have been shown to overlap with the result of a similar work from Gevaert *et al.* (2006) on the van't Veer dataset using a different method involving a Bayesian network with Markov blanket.

Ultimately, this thesis will try to discuss the advantages or limitations as well as the potential application and future hopes around the methods introduced.

Abbreviations

AI	Artificial Intelligence
ANN	Artificial Neur Network
ASCII	American Standard Code for Information Interchange
AUC	Area Under the Curve
BP	Back Propagation
BRCA1	breast cancer, 1
BRCA2	breast cancer, 2
CA9	Carbonic Anhydrase IX
CDCA7	Cell Division Cycle Associated 7
CEA	Carcinoembryonic Antigen
DNA	Deoxyribonucleic Acid
DTL	Denticleless homolog (Drosophila)
EGFR	epidermal growth factor receptor
ER	Oestrogen Receptor
HER2*neu*	Human Epidermal growth factor Receptor 2
HIF-1α	Hypoxia Induced Factor 1 α
IPA	Ingenuity Pathway Analysis

KEGG	Kyoto Encyclopedia of Genes and Genomes
KNN	K-Nearest Neighbours
LDA	Linear Discriminant Analysis
MA	Microarray
MCCV	Monte Carlo Cross Validation
miRNA	MicroRNA
MLP	MultiLayer Perceptron
MS	Mass Spectrometry
MSE	Mean Squared Error
NCBI	National Center for Biotechnology Information
NIH	National Institutes of Health
NPI	Nottingham Prognostic Index
ODEs	Ordinary Differential Equations
p53 or tp53	tumour protein 53
PARP	poly(ADP-Ribose) polymerase
PCA	Principal Component Analysis
PECI	Peroxisomal Δ3,Δ2-enoyl-CoA isomerase
PGK1	Phosphoglycerate Kinase 1
PR	Progesterone Receptor
qRT-PCR	Quantitative Real Time Polymerase Chain Reaction
RAMP	Retinoic Acid-regulated nuclear Matrix-associated Protein

RNA	Ribonucleic Acid
ROC	Receiver Operating Characteristic
RPROP	Resilient Backpropagation
SOMs	Self-organising maps
SVMs	Support Vector Machines
TLU	Threshold Logic Unit
TNM	Tumour, Node, Metastasis
TSPYL5	Testis-Specific Y-encoded-like protein 5
WHO	World Health Organization

Contents

Chapter 1

Introduction

1.1 The challenge of biomarker identification

The latest development in genomics and proteomics research have begun to open the door for the future generation of medicine. One potential to lead to *personalised medicine* where the hope would be to provide early, fully customised and adapted therapy for a particular individual to a particular condition or disease. In the first place, this involves an early detection/diagnosis of the disease, but yet it is just one part of the challenge. Another is to clearly identify the characteristics of the conditions. Several diseases are known to present high heterogeneity (*e.g.* breast cancer is known to have several types and even several subtypes (Kapp et al., 2006; Perou et al., 2000)). Not only does this heterogeneity exists for the disease, but combined with the natural biological variation from one individual to another, it results in very different phenotypes and types of behaviour. As a consequence different responses occur from one patient to another for a similar disease under a similar treatment. Therefore, early and accurate phenotyping and identification of the sub-types of the disease can reveal crucial information relevant for subsequent treatment and prognosis, and ultimately patient outcome.

One of the first requirements for patient-tailored treatment is to obtain early and adapted diagnostics for diseases with high incidence in the population such as cancer. It is well established that an early diagnosis

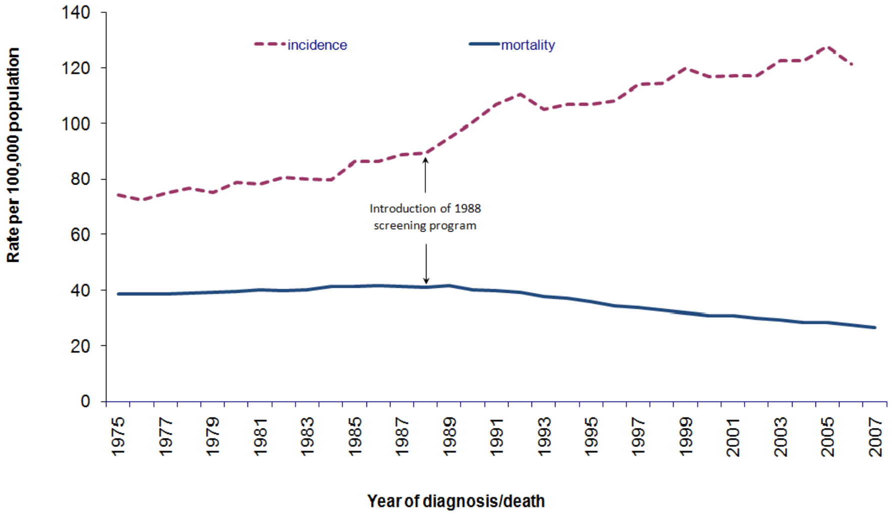


Figure 1.1: Evolution of incidence and mortality rate of female breast cancer in Great Britain from 1975 to 2009 (source: Cancer Research UK)

will greatly increase the prognostic outcome for a patient, by managing the disease and providing appropriate treatment in the early development of the disease, and subsequently preventing it to reach a critical development (*e.g.* metastatic stage in cancer). This has been clearly proven over the years with the improvements of screening techniques such as mammography for breast cancer. Figure 1.1 highlights this fact as it shows a clear drop of mortality rate around 1989 in England, whilst incidence is slightly increased. The evolution of incidence is mainly explained by the worlds first national breast screening programme set up in England in 1988. However, the drop in mortality rate observed can also be attributed to the early successful treatments with Tamoxifen on ER positive patients in the late 80's. These screening techniques have proven their efficiency by diagnosing early and asymptomatic breast carcinomas and subsequently allowed to provide treatments earlier in the disease evolution to a larger cohort of women, consequently leading to a drop in the mortality rate. Nevertheless, they only rely on visual and microscopic observation through pathological methods, and then can only pick tumours that have already evolved to a later stage or grade. As a result, the following step for earlier detection of malignancies needs to be carried out at the molecular scale. This is the whole interest of biomarkers.

Although there is no real consensus among the scientific community or health organisations to precisely define a biomarker, two of them can be noticed:

- A characteristic that can be measured and evaluated as an indicator of normal biological processes, pathological processes or pharmacologic responses to therapeutic interventions National Institutes of Health (NIH) Biomarkers Definitions Working Group, 1998).
- Any substance, structure or process that can be measured in the body or its products and can influence or predict the incidence of outcome of disease. (World Health Organization (WHO) International Programme on Chemical Safety)

However, for these two definitions, a key expression in common is : "*can be measured*". A biomarker is a characteristic or a molecule that can be measured and/or quantified with any instrument or method. Therefore, if the biomarker is to be involved in a potential future assay, or diagnosis tool, the accessibility to the marker needs to be relatively easy (from blood stream or urine). Finding biomarkers using invasive tissue based approaches can not be easily transferable to routine clinical aspect. Consequently, an ideal biomarker would be any peptide, protein or generally a molecule easily accessible via body fluids (such as blood or urine); and that would only be detected in affected patients and not in healthy individuals (Chatterjee and Zetter, 2005; Lane and Crawford, 1979; Linzer and Levine, 1979), hence with a sensitivity of 100%. However, a biomarker is only as good as its ability to accurately identify true positive cases and to distinguish as well true negative ones (*i.e.* not only the sensitivity is important but the specificity is key too) (Diamandis, 2004b).

One of the earliest markers to be used as a biomarker for cancer was Carcinoembryonic Antigen (CEA) described in 1965 by Dr. Joseph Gold for the prediction of colon cancer (Gold and Freedman, 1965). He found that CEA, normally present in foetal tissues, was also present in blood stream of patients with colon cancer (Chatterjee and Zetter, 2005). This particular antigen had been found later to be an indicator of breast cancer prognosis outcome and treatment response (Krieger et al., 1984; Park et al., 2008) as well as CA 15-3 (Chatterjee and Zetter, 2005; Ebeling et al., 2002). It is later, in 1979, that one of the best known cancer marker, tumour protein 53 (p53), is almost simultaneously identified by different groups of Arnold Levine, David Lane, and Lloyd Old (Deleo et al., 1979). This protein will be reported to have key role in Deoxyribonucleic Acid[1] (DNA) damage repair and tumour suppression which will open new direction in breast cancer research (Levine et al., 2004). Still in the focus of breast cancer markers, it is only a decade later that the most important step is accomplished in the identification of 2 major genes for breast cancer management, when in 1994 Mark Skolnick *et al.* discovered the breast cancer, 1 (BRCA1) gene (Miki et al., 1994) and Mike Stratton *et al.* identified the breast cancer, 2 (BRCA2) gene (Miki et al., 1994). BRCA genes are believed to be responsible for about 80% of inherited cancers, and women with BRCA1 mutations have about 85% risk of developing breast cancer for lifetime (Antoniou et al., 2003; King et al., 2003).

We can cite as well that a marker can not only be a molecular entity, but can be an index calculated from a combination of factors describing a sample for several types of outcome. It is the case of the Nottingham Prognostic Index (NPI) which allows the stratification of a population by calculating an index value, this value gives insight about the 5-year survival chance for an individual after surgery removal of breast carcinoma (Haybittle et al., 1982). The NPI is based on the tumour grade score (G) according to the modified Bloom-Richardson grading, tumour size(S), and number of positive lymph nodes score (L) (1 for no node, 2 for 1-3 nodes, 3 for more than 3 nodes). Multivariate analysis found these 3 parameters as important to be involved in the NPI, and the expression is :

$$NPI = G + L + (0.2 \times S) \tag{1.1}$$

[1] DNA is the long-term storage nucleic acid holding the genetic information of a living organism contained within the nucleus of its cell(s) for eukaryotic organisms.

Biomarkers have three main application domains:

- Diagnostics
- Prognostics
- Prediction of therapeutic outcome

Nowadays, the early diagnosis of diseases and their reccurence remain the most frequent use of biomarkers. However, future hopes are to improve the other 2 areas of application. With new biomarkers, accurate tests and predictive models could provide clinicians and patients with accurate prognosis and prediction of response to treatments (Chatterjee and Zetter, 2005). A more precise phenotyping of tumour characteristics would give an insight to which sort of therapeutic strategy had a chance of success. Thus, maximising the therapeutic and consequently prognostic outcome for the patient.

However, considering all this hope, the path leading toward personalised medicine still is a long and difficult one, and the biomedical community is still facing important challenges to achieve this objective (Ginsburg and McCarthy, 2001). The current lack in accurate and reliable biomarkers for critical diseases (*e.g.* cancers) remains a crucial challenge for the biomedical and the bioinformatics community, and the identification of new biomarkers still requires a tremendous amount of work (Rifai et al., 2006). So far, only classical methods, mainly based on the judgment of the pathologists (involving an obvious element of subjectivity) are being employed to characterise the clinical state of a patient regarding a disease, an infection, or the grade of a disease. The discovery of new biomarkers could then allow for a quicker, more reliable and efficient method of diagnosis for these pathologies. Furthermore, such biomarkers could provide new potential targets for future therapies.

This task is made difficult mainly by the high heterogeneity of some diseases, furthermore, diseases are almost always multifactorial (Barabási and Oltvai, 2004). Therefore, it can be worthless focusing on a single biomarker, and it should be recomended to consider a panel of biomarkers, a fingerprint which results in providing more reliable and accurate models.

In this thesis we show that it is possible, not only to extract relevant information and accurate biomarkers from high-throughput technologies such as microarray technology, but literally distill out from the data even more information regarding the whole dynamic in place within the sample at the moment of extraction, collect deeper information and assess how the relevant biomarkers studied may be co-expressed, and therefore interact somehow. This is possible as we know that the system studied are highly complex and dynamic, and therefore hold this information.

1.2 Breast cancer

Cancer is a generic term to designate diseases affecting the regulation of the cells' cycle, growth and migration. It has been found they involve important changes in the genetic information contained in the cells; and it has been shown that these genetic mutations can lead to the apparition of oncogenes (which enhance cell proliferation), and the silencing of cell proliferation regulation genes or tumor suppressor genes (Hanahan and Weinberg, 2000). When the highly regulated behaviour of the cell division is corrupted, they proliferate in an anarchic way, they expand without dying, and invasion can eventually lead to the death of the individual.

A large focal point of this research has been to cancer, and in particular breast malignancies. The reason for this is mainly due to the fact that 120 women out of 1,000 in the UK population have a lifetime risk to develop a breast cancer (source: breast cancer UK), as a consequence a large amount of information as well as samples have been collected and stored during years for the purpose of prospective study. These databases can then provide valuable information and datasets to further develop, test and validate approaches and models developed by scientific research. Besides, this cancer is known to be highly heterogeneous, several types and subtypes exist, with different outcomes, treatments and prognosis for each of them (Sørlie, 2009). That makes this cancer a real challenge for the biomedical community to find appropriate diagnosis tools, biomarkers to characterise the type of breast cancer. Characterising this type early in the development of the disease would help finding which treatment should give the best outcome (and which would not) and therefore apply directly a therapeutic solution, tailored to the patients' tumour phenotypes, that have optimum chances of maximising the patients' benefit.

Breast cancer has a high incidence in western countries, every year 45,000 women are diagnosed with breast cancer in the United Kingdom, 12,000 will die from it. It is the second cause of death from cancer for women in the UK. And one woman out of nine will be diagnosed with breast cancer in her lifetime (source: cancer research UK).

One of the main difficulties encountered by researchers in breast cancer is the high heterogeneity of the disease. This cancer can exist in many different types and sub-types. First of all, the cell type of the origin of the cancer (ductal or basal) makes it very different. Thereafter, several subtypes can exist regarding the status of the cells for oestrogen or progesterone hormones, or Her2 protein (Bertucci and Birnbaum, 2008; Sørlie et al., 2001, 2003). Greater understanding of underlying molecular events can result in a more accurate choice of therapeutic strategies thus favouring better results and outcomes for the patients at the end of the curing process.

1.3 The interest in Systems Biology

It is becoming clear that the era of reductionism (*i.e.* the removal or filtering out of important variables) is potentially limited with respect to genomics and proteomics information derived from histological systems in biology and biomedical sciences (van Regenmortel, 2004). Until lately, reductionism was still the current paradigm in biology and biomedical sciences. However, it appears clearly now that such a way of researching starts to show some limitations (Gierasch and Gershenson, 2009). Therefore, the scientific community is beginning to consider a more comprehensive approach toward life science : *Systems biology* (Ahn et al., 2006). According to Ahn and coworkers, system biology is the study of biological processes from a wider point of view, considering the overall dynamics of the parts to form an entity.

So far research groups have mainly focused their work toward biomarker identification, however it clearly appears that identifying single markers is highly reductionist and has reached its limits in terms of predictability, and more particularly in terms of understanding the biological relevance of a particular marker. Therefore, it was sought to investigate further to, firstly identify not only a single marker, but a combination of markers able to accurately discriminate together the population between two or multiple classes; and secondly identify potential interaction that may exist between the markers contained in a dataset. Indeed, by looking only at potential single biomarkers, we try to model a real-world system that is non-static, but completely dynamic and highly complex, as most of these markers are interacting with each other, as stated by Barabási & Oltvai (Barabási and Oltvai, 2004) :

> *It is increasingly clear that a discrete biological function can only rarely be attributed to an individual molecule.*

While the identification of a panel of genes or proteins is an important work, investigation of interactions within a panel of contextually relevant genes associated with a given clinical question has the potential to elucidate biological function and identify functions and interactions that associate with the disease, thus studying *"interactome"*[2] of a given disease. The study of the interactome offers the further potential for identification of new disease specific pathways that may be used to identify therapeutic targets.

When one considers the study of the interactome an issue of complexity arises. If one were, for example, to consider a whole genome, the number of interactions would be massive. For example, in an Affymetrix array one would find in the order of 600 million potential interactions. Clearly,

[2] The term interactome has been firstly described in 1999 by sanchez and coworkers as *"the complete repertoire of interactions potentially encoded by the genome of an organism"* (Sanchez et al., 1999).

interpretation of this is impossible due to complexity. In order to study the gene interactome in a reasonable fashion it is necessary to first define the context of the question. In an earlier study, our group developed a method of screening genes in the context of a given question (Lancashire et al., 2008). This reduces the parameter set to a manageable size while addressing the context of the problem allowing one to analyse the interactome.

In addition, it is now believed that the paradigm governing the idea of transcriptional regulation as a simple on-off switch is no longer acceptable, and scientists should rather consider it as a more complex system, similar to a complex logic circuit, where several inputs may have contradictory influence toward the expression of a gene (Duggan et al., 1999; McAdams and Shapiro, 1995).

This thesis will describe a novel approach simultaneously considering all of the genes in a set identified to be predictive of metastasis in breast cancer explored previously by the van't Veer study (van't Veer et al., 2002). This is achieved by the development and comparison of a series of Artificial Neur Network (ANN) models that each of which use all available inputs to predict a single omitted input. This is repeated for all single inputs within the set. By examination of the weights and predictive performance of each model, a matrix of interactions can be determined. The advantage of this approach over others is that the multi-factorial consideration of each input allows determination, for a given pair of parameters, the magnitude of interaction, whether it is inhibitory or stimulatory and whether the interaction is bi or unidirectional. Once the process is achieved, filtering out the non-significant interactions with a certain decision threshold will ultimately result in an interaction map giving insight on what is relevant.

1.4 Types of data used and analysed

Metabolic changes occurring due to a disease should have impact on the presence (qualitative) and abundance (quantitative) of particular molecules that can be monitored within samples such as serum or urine (Claudino et al., 2007). Nowadays, the development of the new *"omics"* technologies enables researchers to massively monitor the comprehensive profiling of several types of molecules (*e.g.* DNA, Ribonucleic Acid (RNA)[3], MicroRNA (miRNA)[4], and subsequently compare different samples to identify any significant qualitative or quantitative alteration occurring between a cohort of control samples and a cohort of samples carrying a particular disease; or between two stages of a same disease.

These technologies require multi-disciplinary teams to firstly produce and then decipher the vast amount of highly complex data produced at the end of a Microarray (MA) experiment in order to extract the maximum relevant knowledge (Duggan et al., 1999).

The research described in the present manuscript focused on several types of data to analyse, genomic MA data in particular. Although, some other types of data have been analysed too, such as Mass Spectrometry (MS), flow cytometry or environmental data, they won't be further discussed here, for concision and consistency purposes.

1.4.1 Microarrays for genomic profiling

Statistical analysis of the genomic profile of an individual can lead to the discovery of a pattern of genes, or fingerprint, associated with a disease or phenotype of interest. Therefore, assessing the gene expression profile for an individual could not only be exploited as a valuable diagnostic or phenotyping tool, but could enhance our knowledge of the particular dynamic

[3] RNA is a nucleic acid-based molecule existing outside of the nucleus of the cells. Several types exist, inluding rRNA, tRNA, miRNA, etc... The most abundant and important is messenger RNA (mRNA) that plays a role of intermediaite between DNA molecules and protein or peptide products.

[4] miRNAs are small RNAs (15-20 pair base) believed to play a regulatory role in transcription of DNA.

within the system of a given disease. MAs are a powerful and accurate new technology that potentially allow the detection of hundreds of thousands of genes contained in a biological sample within a single experiment (Lancashire et al., 2009a). As a consequence, MAs have the perspective to draw the whole map of the human genome and massively discover genes of interest related to a particular phenotypes, ultimately leading to important clinical developments for patient care and management (Lander, 1999).

MA techniques were directly derived from Southern blotting techniques developed in 1975 (Southern, 1975). Edwin Southern's method relied on the hybridisation of DNA fragments, previously migrated by electrophoresis, using probes to detect the migration of the DNA sequence of interest. The evolutions from Southern's technique combined to the development of cDNA libraries (Schena et al., 1995) triggered the development of DNA MAs. These evolutions consisted firstly in the use of non-porous surface, such as glass, which allowed their second evolution: miniaturisation (Lander, 1999). As a consequence, spotting cDNA probes onto a solid surface in microscopic wells, the MAs are able to simultaneously measure the expression level of an important number of genes. The first real use of genomic (DNA) MA was reported in 1995 by Schena and coworkers (Schena et al., 1995). They described a method to measure the differential gene expression of 45 genes in *Arabidopsis thaliana*, by spotting gene-specific cDNA probes on a solid glass surface and presenting to them fluorescent labeled RNAs from the organism.

Principle

The detection of presence and abundance of labeled nucleic acids mainly relies on the principle of DNA/RNA hybridisation on the array, where two strands of matching DNA/RNA pairs bind to each other. They consist of a collection of very short sequences of nucleic acids chemically bound to a solid surface. Each of these sequences (the probes) are complementary to a sequence of a gene, therefore each probe represents a particular gene, or a specific segment of a gene. As a result, if a particular gene is present in

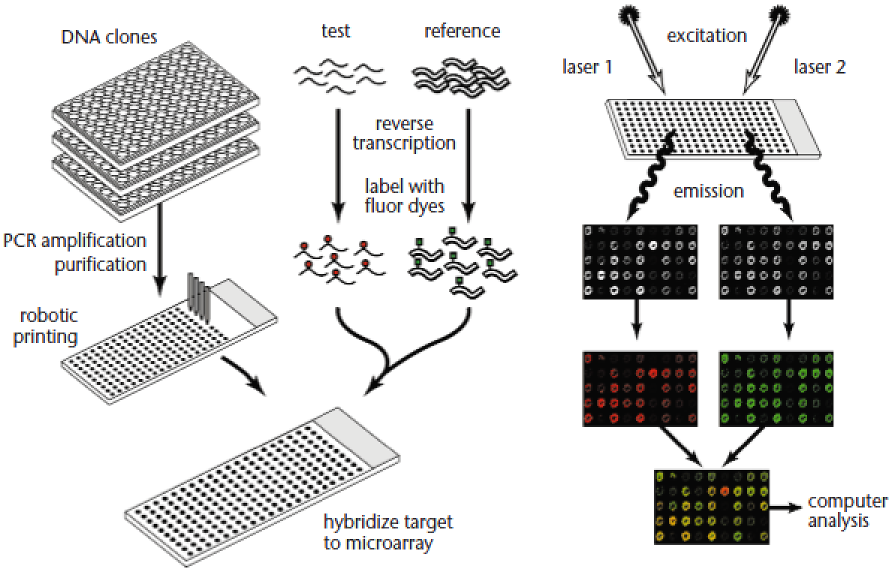


Figure 1.2: The different steps for a gene microarray experiment (Duggan et al., 1999)

the sample, the probe of the corresponding gene will appear as present by fluorescence.

Methods

The process of MAs can be split into 4 main steps (Stekel, 2003), summarised in Figure 1.2.

- Sample preparation and labelling
- Hybridisation
- Washing
- Image acquisition

The early stage of a MA experiment consists of the extraction of the molecules of interest (RNA, cDNA, miRNA...) from the samples. The molecules are then labeled as desired. An example of a labelling method is fluorescent dye binding to the oligonucleotide with Cytosine 3 (excited by green laser) and Cytosine 5 (excited by red laser). Commonly, two samples are used differently labeled with the cytosines (*e.g.* healthy and disease subjects) (Stekel, 2003).

Once labeling is complete, the molecules are presented to the array, where matching strands can bind to each other. Therefore, when labeled DNA/RNA fragments from the sample match DNA probes, heteroduplexes are formed by complementary base pairing (hybridisation) as described by Watson and Crick (Watson and Crick, 1953), which shows the presence of the gene targeted by the specific probe, and highlights this presence later by measuring the level of fluorescence.

Eluting is a critical step, as only labeled molecules that bound to the probes are needed, it is essential that non-bound molecules are washed out.

Finally, an image of the fluorescence emitted by the dyes still bound to the probes is captured. The intensity of the fluorescence is then numerically translated, which is supposed to be proportional to the binding intensity and therefore to the level of expression of the gene corresponding to the probe.

Limitations

However, one of the main limitations of this method is situated in the advantage itself. The massive production of data provide a dramatic highly dimensional dataset, which can barely be processed by conventional statistical methods. For example, with Affymetrix GeneChip 1.0 ST MAs (designed to target all known and predicted exons in human, mouse and rat genomes), each individual case studied contains information for approximately 1.2 million exon clusters corresponding to over 1.4 million probesets (Lancashire et al., 2009a). In addition, this issue, combined with the inherent important biological variability commonly found in biological samples, induces an important difficulty to identify key features within a dataset. This clearly stresses the need for feature selection/extraction strategy prior any deeper analysis.

More importantly other issues arose from a reproducibility aspect. For a robust and non-questionable identification of important markers, the data analysed need to be reproducible within samples, between sample runs and even across different instruments (Diamandis, 2004a). Technical solution employed so far use replicates and filtering with averaging of samples to assess the good reproducibility of the experiments. Lack of good reproducibility projects suspicion onto any identified predictive signature, subsequently making validation and then clinical application difficult (Matharoo-Ball et al., 2007c).

With this question comes as well the question raised by normalisation methods. The question of normalisation is key and several methods have been described so far. The lack of gold standards toward this normalisation issue reduces the chances to obtain highly reproducible and inter-experimental studies, particulary when applying normalisation methods to purposes it was not meant to (Lim et al., 2007).

Therefore, innovative analysis methods need to be applied to extract meaningful information from the array.

1.4.2 Simulated dataset

Despite the obvious advantages of using well-characterised simulated datasets for the testing of new analysis tools; clear precautions need to be taken in order to address appropriately any analysis concerning this type of dataset. Indeed, the recent editorial from Rocke et al. in Bioinformatics (Rocke et al., 2009) commented that our lack in the knowledge of actual biological correlation between sample replicates, combined with the unknown dependency between a biological state of a cell and transcript expression, and the unknown changes from one state to another, makes artificial data only valuable for development stage and unworthy for purely comparing different methods.

1.5 Nature of the problems

Although expression MAs have brought high hopes and expectations, they have brought with them tremendous challenges too. They have been proven to suffer from different limitations as previously discussed.

However innovative computational analysis solutions have been developed and have been proven to be effective and successful at identifying markers of interest regarding particular questions. But few research has been carried out to deeply explore the information hold by the complexity of expression arrays which can provide a broader view of interaction or regulatory pathways and networks (Bar-Joseph et al., 2003).

1.6 Statistical methods for biomarker identification

The real interest in MA experiments is in the analysis of multiple expression experiments in order to identify similar patterns of expression (Quackenbush, 2001). Although, multiplying the number of expression arrays ultimately leads to an increasing amount of data correlating with an increasing complexity; thus requiring robust methods to extract the relevant information from them. Computer based methods are then needed to decipher this complexity, and extract as much meaningful information as possible. Several methods exist for this purpose, some of them (i) are derived from the field of classical statistics, such as clustering techniques or regression analysis. Other methods (ii) come from the more advanced field of machine learning approaches ranging from Support Vector Machines (SVMs) to Bayesian Networks, and including ANNs.

We will briefly introduce here the most common methods employed to handle genomic MA data, previously discussed, before focusing on ANNs in the following chapter, the approach chosen for this Ph.D. project.

1.6.1 Standard statistical approaches

Classical statistics used to be gold-standard methods for the analysis and classification of genomic MA data in the early years of MAs. This section presents some of these methods.

Clustering methods

Historically, clustering (and particularly hierarchical clustering) was the first approach adopted by scientists in order to analyse results from MA technologies. These approaches group the samples (or the genes in the transposed fashion) together in clusters on the basis of their distance in the expression space (Eisen et al., 1998). Several algorithms for the measure of this distance exist, and generate the diversity of clustering techniques.

One of the most commonly applied clustering technique for genomic MA experimental data is ***hierarchical clustering***. The method introduces a measure of distance between the samples (or the genes in the way round), and groups them in clusters according to this distance, repeating this iteratively until all the samples have been used. The distance basically represents how close the objects to analyse are to one another. The most common distance measure is euclidian distance, but others may be used such as the Manhattan distance or Maximum Norm.

Finally, the samples are organised and grouped in a phylogenetic tree-like structure, allowing a rapid visual interpretation of the results (*i.e.* 2 samples with similar gene expression patterns have a small distance, and fall into the same cluster). As a result, genes with similar biological function may fall into the same cluster or different clusters with short distance. Different methods exist to determine the hierarchy of the clusters, amongst them the most famous are Single-linkage, Complete-linkage, Average-linkage or Weighted pair-group average. This method has been widely applied in many famous studies in breast cancer (Perou et al., 2000; van't Veer et al., 2002) mainly because it is a simple method with a quick and easy visualisation of the results (Quackenbush, 2001).

Contrary to hierarchical clustering, ***k-means clustering*** assigns the objects (either samples or genes) within clusters without looking at the relationship that might exist between them (Quackenbush, 2001). It partitions the data space in a predetermined number of clusters (k) and randomly assigns the objects to one of the clusters. An average expression vector is calculated for each cluster from the expression vector of the objects they contain, and is used to calculate the distance between the clusters. The objects are then iteratively moved to other clusters, each time recalculating the new distance obtained. Objects are moved until optimal distance is reached (*i.e.* the distances, either inter or intra-cluster, are less consistent).

Alternative implementation of the algorithm allows assignment of a seed object (sample or gene) for each cluster, using it in a semi-supervised fashion (with *a priori* knowledge). On the hypothesis of $k = 2$ groups (*e.g.* controls *versus* disease) with the extreme expression patterns assigned to each cluster, it is possible to assess in which class individual samples would fall, and how closely.

Self-organising maps (SOMs), or Kohonen maps, named after the statistician Teuvo Kohonen who developed this method (Kohonen, 1989), are a form of cross-combination of clustering techniques with the field of ANNs, and can be seen as an unsupervised neural network (Törönen et al., 1999). Similarly to k-means clustering, this approach creates groups and assigns the genes to each group based on their similarity to their reference vector (Quackenbush, 2001). The difference with k-means clustering stands in the organisation of the groups and in the calculation of the reference vector for each of them. Here the groups are represented by a grid of nodes defined by the user and geometrically arranged in a k-dimensional space (Tamayo et al., 1999), commonly a two-dimensional space for easy visualisation (*c.f.* Figure 1.3). Firstly, the groups are randomly set with their respective vectors. Then the algorithm iteratively selects a gene at random in the data space and calculates the distance (with respect to the distance metric system chosen) with the closest group. Then the groups are moved toward this particular gene inversely proportional to their distance with it by adjusting their reference vector (*i.e.* the closest group has its reference vector the most adjusted, and the farthest, the least). This adjustment is given by the following expression, with the position of the group N at iteration i being $f_i(N)$, and at the following iteration, the data point P is selected with its nearest group N_P :

$$f_{i+1}(N) = f_i(N) + \tau(d(N, N_p), i)(P - f_i(N)) \tag{1.2}$$

The learning rate τ gives the rate that decreases with the distance of the group N with N_P. This process is repeated over a number of iterations i (20,000 - 50,000).

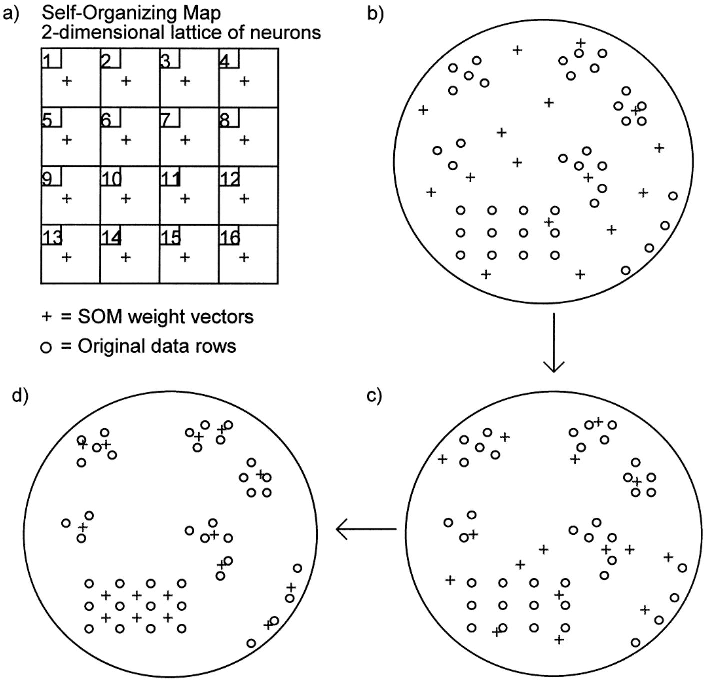


Figure 1.3: Principle of Self-Organizing Map. a) Example of a SOM consisting of a 4 × 4 matrix of neurons. b) The weight vectors (+) of neurons are first initialised with random profiles. c) Intermediate configuration during the learning process where weight vectors are moving towards the data profiles (O). d) Finally, weight vectors come to represent groups of data profiles. (Törönen et al., 1999)

Mangiameli and co-workers compared the robustness of SOMs with hierarchical clustering (Mangiameli et al., 1996) and demonstrated that SOMs were performing better for the analysis of what they called *"messy data"*.

The choice of geometric configuration for SOMs, or k-means clustering, is driven by the user, thus introducing an element of subjectivity. Additionally, this application of cluster analysis to class prediction have been criticised (Simon et al., 2003). Therefore, it is suggested to rely on another source of information, such as Principal Component Analysis (PCA), to enlighten this choice of number of clusters that best represents the data space.

Principal Component Analysis

PCA is a method of choice for dimensionality reduction in datasets (Lancashire, 2006), a particular issue when studying gene MA data, as the number of genes studied P far exceeds the number of samples n ($P >> n$). It is an exploratory multivariate statistical technique used to simplify complex data space (Raychaudhuri et al., 2000) by translating the data space into a new space defined by the *principal components*, relating to the variance (shape) of the data. These principal components are orthogonally arranged linear combinations of the original variables identified by the method and that can explain most of the variation observed within the dataset, which can be due to redundancy of information in the data space (Raychaudhuri et al., 2000). As a result, it is difficult or even impossible to link the principal components to the actual variables of the original dataset, making it difficult to identify the important genes in the system studied for the microarray experiment (Lancashire, 2006). This limitation stresses that PCA is only a data reduction tool and highlights the need of supplementary analysis tool.

All these techniques are usually qualified as unsupervised methods since they do not require any subset of data for which outcome is known in order to construct a model based on the known examples, although this is not entirely true since some clustering techniques are supervised (as mentioned earlier) (Eisen et al., 1998).

Classical unsupervised methods have been proven to suffer from several issues. Cluster techniques, and particularly hierarchical clustering, have been mis-employed for class comparison and class prediction and have been proven to be non-effective for that particular matter (Simon et al., 2003;

Tamayo et al., 1999) as well as PCA (Tan et al., 2005). The choice amongst the different clustering techniques, combined with the different distance measurement methods introduce an element of subjectivity and result in the production of different outcomes of the analysis (Quackenbush, 2001). But their main pitfall stands in their inability to cope with high-dimensionality. In high-dimensional space, it would be likely that for a pair of adjacent objects falling into the same cluster there would exist dimensions in which those points fall far apart from each other. Thus, techniques using distance functions with all input features may lack effectiveness (Domeniconi et al., 2004). However, some groups apply these techniques as a simple preliminary step for feature selection, prior to deeper analysis of the dataset, using machine learning approaches (Liu et al., 2005b).

1.6.2 Machine Learning and supervised approaches for classification and prediction problem

These methods are mostly used for their predictive ability. They can classify individual samples to groups they belong to, thanks to their expression profile and genomic expression patterns. This is an important step for allowing the stratification of patients and be able to accurately predict, for instance, their tumour phenotypes in the case of cancers.

K-Nearest Neighbours

K-Nearest Neighbours (KNN) is one of the most simple approaches for general non-parametric sample classification (Vadrevu and Murty, 2010). This method considers the comparison of expression profiles with an unknown class to other ones for which the class is known. The class for the unknown sample is then determined to be the same than the class for the sample with the most similar profile and for which class is known. This method follows three steps for the classification of a new sample with an unknown class:

- Assessment of the gene expression profile for the sample studied.
- Identification of the nearest known samples thanks to a distance measurement (usually Euclidian distance).
- The class of the sample is determined by the class of the nearest samples.

This is achieved by considering two important factors :

- k : the number of nearest samples to look at.
- l : the acceptable margin for classification to be successful.

As a result, the class of the sample to be classified is determined only if l of the k nearest samples are of the same class; if not, the sample class remains undetermined. Then, if $k = 3$ and $l = 3$, the algorithm only considers the 3 nearest neighbours (k) and the class of the sample to be classified is determined only if the 3 of them (l) are of the same class (Stekel, 2003).

KNN algorithms have been shown to have good classification performances for a wide range of real-world data (Xiong and Chen, 2006). However, they can suffer from major disadvantages, mainly due to poor efficiency with high-dimensionality and noise in the dataset (Vadrevu and Murty, 2010).

KNN have been also applied for class prediction problem with gene MA data in breast cancer studies. Miller and coworkers compared the performances of several supervised training methods (including KNN, Linear Discriminant Analysis (LDA) and SVM) for the prediction of p53 (mutant versus wild type) in breast cancer survival on different datasets (Miller et al., 2005). They observed that, despite very similar prediction accuracy (between 84.9% and 85.7%), KNN performed with slightly higher specificity[5] (95.3%) than SVM (94.3%) and LDA (84.6%). However, LDA was found to have higher sensitivity[6] (79%) compared to KNN and SVM (53%).

[5] Specificity is a statistical measure of the ability for a test to classify the proportion of negative cases correctly identified by the test as such.

[6] Sensitivity is a statistical measure of the ability for a test to classify the proportion of positive cases correctly identified by the test as such.

Linear discriminant analysis

LDA attempts to separate the data into two sub-groups by calculating the optimal linear line that best split the population. Calculation of this discriminating line is conducted by taking into account sample variation within similar classes, and minimising it between classes. As a result, any additional sample has its class determined by the side of the discriminating line it falls.

LDA can outperform other linear classification methods as LDA tries to consider the variation within the sample population. Nevertheless, LDA still suffers from its linear characteristic, and often fails to accurately classify non-linear problems which is mostly the case in biomedical sciences (Stekel, 2003). This is the reason why non-linear classifiers are recommended.

Support Vector Machines

SVMs are a form of supervised learning approach for non-linear modeling(Vapnik and Lerner, 1963). They are an evolution of LDA in the sense that they work by separating the data into 2 sub-groups with a straight line or an hyperplane.

In the common example of a two-class classification problem, SVMs firstly attempt to find a linear *"‘maximal margin hyperplane"’* able to accurately discriminate the classes (Dreiseitl et al., 2001). If no such linear hyperplane can be found, usually due to the inherent non-linearity of the dataset, the data are mapped into a high-dimensional feature space using *kernel functions* in which the two classes can now be separated by a *hyperplane* which corresponds to a non-linear classifier (Furey et al., 2000). The class of an unknown sample is then determined by the side of the hyperplane it falls.

SVMs are one of the most popular classification tool in the biomedical community, and have a strong competitive interest with ANNs. However, no evidence has clearly demonstrated that SVMs perform better than ANNs for medical prospect (Lancashire et al., 2009a).

1.7 Alternative interaction inference algorithm

Considering the tremendous promise expression arrays hold for a comprehensive understanding of the molecular biology in living systems, some groups have recently started to focus their research into the investigation and exploration of the potential gene interactions and relationships assessable with expression array data. Different approaches have been described, and some of them are discussed in this section.

Bar-Joseph and coworkers presented a paper describing a technique using DNA-binding data to infer potential gene regulatory networks (Bar-Joseph et al., 2003), arguing that it is necessary not to consider only expression data but a combination with other sources in order to overcome the variability from post-transcriptional modifications.

Hart et al. in 2006 tried to implement an ANN-based gene regulatory network inference method, which they applied to yeast cell cycle (Hart et al., 2006). Contrary to the method presented here, the group trained the ANN only to predict a set of cell cycle expression classes with a pool of 204 expression regulators in order to predict which of the regulators is predictive of a class.

Another ANN-based application for gene-gene interaction modelling has been described by Günther and coworkers (Günther et al., 2009), in which they conclude that ANN-based methods are more reliable than logistic regression at predicting gene-gene interactions. However, the study only aimed at comparing the methods with simulation datasets for two-loci disease scenario.

Gevaert and coworkers developed a Bayesian network method with a Markov blanket algorithm to investigate the connection between the genes involved in the van't Veer dataset (Gevaert et al., 2006). This is an interesting application and will be useful to confirm some of the findings obtained by the approach presented in this study.

Several other groups made the choice of Bayesian networks for these algorithms. It is the case of Jansen et al. who described an approach using genomic data to predict protein-protein interactions (Jansen et al., 2003), as well as Hartemink et al. and Needham et al. (Hartemink et al., 2002; Needham et al., 2009).

Another effective method preferred by mathematician groups is based on Ordinary Differential Equations (ODEs). Christley and coworkers published in 2009 a paper based with this methodology for gene expression data in which they incorporated *a priori* knowledge on gene interaction or regulation, such as CHIP-seq data, in order to decrease the modelling error (Christley et al., 2009).

Recently, Jupiter and colleagues presented a web-based tool named StarNet2 which aims to identify gene regulatory networks from MA co-expression data (Jupiter et al., 2009). However, this method only interrogates MA data for correlation in a pairwise fashion.

Other methods such as likelihood approach (Liu et al., 2005a), random walks (Chipman and Singh, 2009) or synergy networks (Watkinson et al., 2008) can be cited.

These methods for gene regulatory networks inference have been reviewed by Schlitt Brazma (Schlitt and Brazma, 2007), and by Lee Tzou (Lee and Tzou, 2009).

Beside these methods, a number of groups focused their studies on the use of functionality to define interactions, particularly for protein-protein interaction studies (Schwikowski et al., 2000; Shoemaker and Panchenko, 2007; Spirin and Mirny, 2003).

Also, literature-based tools seeking to compile every reported link between biomolecules have seen the day as well. Some well-known examples are:

- Kyoto Encyclopedia of Genes and Genomes (KEGG) PATHWAY database (Kanehisa and Goto, 2000).

- Ariadne Pathway Studio (Nikitin et al., 2003).

- Ingenuity Pathway Analysis (IPA) (Ingenuity ® Systems Inc., www.ingenuity.com).
- GeneGo (GeneGo Inc., http://www.genego.com/).

The main disadvantage of many of these methodologies is that they provide only limited information regarding the nature of the interactions. In many cases only the presence of an interaction is considered. Furthermore, when an interaction is considered, it is only in the context of the target, and not considered as it is interacting with the whole gene marker pool. In the approach discussed here, the overall panel of genes considered is questioned iteratively for each gene as target, output to be predicted, so a global interaction map can be extrapolated.

Bar-Joseph *et al.* argued that only expression data for gene regulatory network inference should be avoided due to the variability induced by post-transcriptional modifications, and a combination of data sources should be preferred (Bar-Joseph et al., 2003). In the case of this research, the post-transcriptional modifications are overcomed by modelling simultaneously tens to hundreds of cases. And a particular care has been taken with the integration of different sources of data, in order to avoid inconsistency and biases from cross-platform technologies.

1.8 Aims of the study

The biomedical technologies discussed previously provide a dramatic amount of data to researchers who don't always know how to handle such quantity and complexity of information (Lancashire et al., 2009a). Therefore, the questions are: what do they want out of these experiments, what are the answers sought, and how can we maximise the gain from these productive yet expensive technologies?

It has been demonstrated earlier in this manuscript that conventional statistical approaches suffer from great limitations for the purpose of class assignment and prediction. Hence, the analysis of the data from these biomedical technologies, enhanced the field of computational intelligence in biomedical sciences.

Several machine learning methdods have seen the day since then and amongst them are ANNs (discussed in the following chapter). A number of biomedical applications of ANN-based studies have been reported to be succesful (Drew and Monson, 2000). Comparison with other machine learning methods have been conducted, and showed that ANNs perform similarly or even better to other solutions like SVMs (Lancashire et al., 2009a).

However, these applications mainly aimed to identify signatures predictive of particular questions, and even fewer intended to bring a step further the analysis of the important information held within the experimental dataset.

Therefore, it was sought within this research to :

- Develop new approaches for the analysis of highly-dimensional data for biomarker identification using ANN techniques.

- Optimise the approaches for speed and accuracy in order to overcome limitations of the data. These methods should allow quicker and reliable results for the identification of biomarkers. In addition, developing standalone solutions with the ability to be portable under alternative systems, will allow smoother and faster application of the methods.

- Develop new approaches for identification of interaction networks, particularly in the field of genomic technologies.

- Optimise these approaches for speed and accuracy in order to obtain the most reliable results for the exploration of interactions between markers.

- Application of all the approaches developed in breast cancer case studies, to potentially identify new predictive signatures for outcomes, and new patterns of interaction or relationships between the identified elements. Ultimately, identifying key metabolic changes between two categories of populations could lead to breakthrough discoveries of potential new treatment strategies.

Chapter 2

Theories behind Artificial Neural Networks and Stepwise parameterisation

The following chapter will introduce the concepts and parameterisation of ANNs. It will present a brief history and the basic principles around ANNs and their architecture. The chapter will introduce as well the Stepwise parameterisation of ANNs with its principles and algorithm. The chapter will finally briefly present an application of the original implementation of the Stepwise approach in a breast cancer study.

2.1 Artificial Neural Networks

2.1.1 Introduction

ANNs are commonly described as computational modelling systems. They are a form of machine learning algorithms that can *learn* from patterns (Khan et al., 2001; Lancashire et al., 2005b) in order to predict answers to particular questions in complex data (*i.e.* non-linear, highly-dimensional and noisy data). The models produced by ANNs have been shown to be able to predict well for unseen data, meaning that ANNs have a reliable generalisation ability (Lancashire et al., 2009a). Therefore they represent

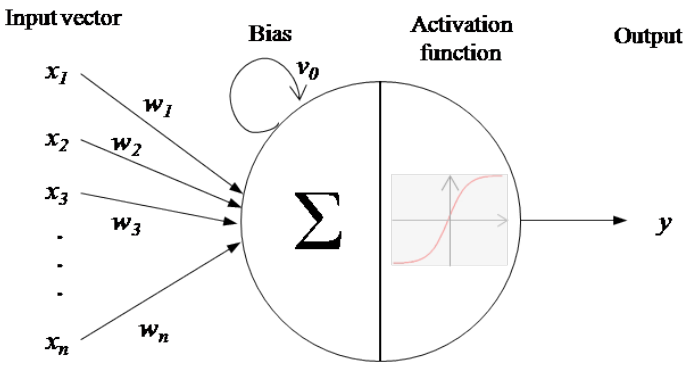


Figure 2.1: Schematic representation of the Artifical Neuron described by McCulloch and Pitts in 1943 (Rojas, 1996)

one of the most robust and reliable methods to handle complex data generated by the biomedical technologies discussed earlier (Drew and Monson, 2000).

ANNs are constituted by single elements, nodes named *neurons* in reference to the biological inspiration (discussed later). Figure 2.1 represents an example of a single artificial neuron. A neuron j receives n inputs i with their respective value x (the input vector) that has been scaled (usually between 0 and 1 or between -1 and +1). Each of them is modified by the weight w of the link between the input i and the neuron j. The neuron calculates then the sum of the weighted values for all the x inputs (Bishop, 1995; Lancashire et al., 2009a).

$$v_j = \sum_{i=1}^{n} (w_{ji} x_i) \tag{2.1}$$

The sum is then added to the bias value v_{j0} of the neuron j. The final value is then passed through a transfer (or activation) function ϕ in order to obtain a final non-linear output y_j for neuron j.

$$y_j = \phi(v_j + v_{j0}) \tag{2.2}$$

The transfer function ϕ is commonly a sigmoid function of the form:

$$\phi(v) = \frac{1}{1 + e^{-v}} \tag{2.3}$$

or an hyperbolic tangent function:

$$\phi(v) = \frac{e^{v} - e^{-v}}{e^{v} + e^{-v}} \tag{2.4}$$

So that we obtain:

$$y_j = \phi \left(v_{j0} + \sum_{i=1}^{n} (w_{ji} x_i) \right) \tag{2.5}$$

The output of a particular neuron can eventually become the input of a subsequent neuron, and information is transmitted and, if appropriate, altered until finally, this output is propagated up to the ultimate neuron of the network and provide an overall output of the global ANN. This overall output can be the predicted class of the current case for a two-group classification model often coded as 0 and 1 or as -1 and $+1$, or a discrete value for a regression model (Lancashire et al., 2009a).

2.1.2 Historical Background

It is believed that the early stages of theoretical neurophysiology triggered in parallel the preliminary works of neuromathematics and neurocomputing between 1890 and 1949 (Basheer and Hajmeer, 2000; Nelson and Illingworth, 1991). The preliminary descriptions of biological neurons initiated the early research on the artificial neuron. The neuroscientist Warren S. McCulloch and the logician Walter Pitts (Hecht-Nielsen, 1988) firstly described the concept of this artificial neuron (also known as the Threshold

Logic Unit (TLU), or the "McCulloch and Pitts neuron") in 1943 as a mathematical function derived from simulating the basic functions of biological neurons (McCulloch and Pitts, 1943).

This work described a neuron and the input vector this one receives (*c.f.* Figure 2.1). Mimicking the basic mechanic of the biological neuron, the artificial neuron stated by McCulloch and Pitts is a binary element, *i.e.* existing under two clearly distinct states y : activated or inhibited (positive or negative, 1 or 0), with a threshold θ defining the state of the neuron. The neuron receives signals from one or several (n) other input(s) (x_i with $i = 1...n$) (the input vector), and emits the signal toward others with its output. Each input is weighted by a fixed value w (-1 or $+1$), which alters the influence of the input. Therefore, signals from the input received can either be excitatory or inhibitory and have the same importance. But if any signal received is inhibitory (weighted as -1), it will not be further transmitted; in other words, all the signals received need to be excitatory in order to activate the neuron, so behaving as a "all-or-none" process: the equivalent of the logic threshold function. The weighted sum of the input vector is then calculated and the result compared to the threshold θ, resulting in the following rule :

$$y = \begin{cases} 1, & \text{if} \quad \left(\sum_{i=1}^{n} w_i x_i \right) \geq \theta \\ 0, & \text{if} \quad \sum_{i=1}^{n} (w_i x_i) < \theta \end{cases} \tag{2.6}$$

Later in 1958, the psychologist Frank Rosenblatt formulated the concept of the *Perceptron* (Rosenblatt, 1958) in an attempt to describe the perception (which gave the origin of the name for the Perceptron) especially in retina (Kanal, 2004). The Perceptron stood as an evolution of the recently born artificial neuron of McCulloch and Pitt integrated in the first neural network. The concept of the Perceptron mainly relies on two things : the weighted connections and its learning ability (Rojas, 1996). Initially, weight connections are stochastically determined, the learning process is achieved by the alteration of the nodes or the weights of the interconnections between the nodes of the network. For this purpose, a set of examples is used to learn, or *train* the network. By adapting the weights (w) of the connections between the nodes, the Perceptron has the ability to identify inputs or connections which provide stronger response toward the goal to achieve (*e.g.* character recognition problems in the context of Rosenblatt's work (Hecht-Nielsen, 1988). Similarly than for a biological neural network, the learning process involves a closer connection between the neurons involved (*i.e.* a higher weight) for the particular function required according to the *Cell assembly theory* formulated by the psychologist Donald Olding Hebb earlier in 1949 (Hebb, 1949):

> *Let us assume that the persistence or repetition of a reverberatory activity (or "trace") tends to induce lasting cellular changes that add to its stability. When an axon of cell A is near enough to excite a cell B and repeatedly or persistently takes part in firing it, some growth process or metabolic change takes place in one or both cells such that A's efficiency, as one of the cells firing B, is increased.*

This theory initiated the whole field of learning rules in artificial neural networks, brought forward with Rosenblatt's Perceptron. However, efficient at character recognition, this Perceptron was only able to model linear systems. Nevertheless, a growing interest in neurocomputing started at that moment.

Contemporary to Rosenblatt, the field of Artificial Intelligence (AI) was not gathering as much interest (*i.e.* funding) as the Perceptron due to this fashion. This led to the publication of a book by the cognitive scientist Marvin Minsky and the mathematician Seymour Papert, pioneers in AI, (Minsky and Papert, 1969) which aimed to stress the limitations of the Perceptron, especially their inefficiency at modelling non-linear systems. Most of these limitations were already known by the community (Basheer and

Hajmeer, 2000; Hecht-Nielsen, 1988), but they highlighted their importance (Lancashire, 2006). This publication had a major impact and stopped the works on the field of ANNs, with many researchers switching interest from ANNs toward AI (Basheer and Hajmeer, 2000).

Interest in ANNs only reappeared in the middle of the 1980's, initially with the work of John Joseph Hopfield in 1982 (Hopfield, 1982) and 1984 (Hopfield, 1984) who introduced the Hopfield network. But it is admitted that 1986 was the year of the regain of interest with a major breakthrough for ANNs when Rumelhart, Hinton and Williams published their work on the Back-Propagation algorithm (Rumelhart et al., 1986) initially described by Paul Werbos in 1974 (Basheer and Hajmeer, 2000; Werbos, 1994), and which is an adaptation of Widrow and Hoff's work on the Delta rule (Widrow and Hoff, 1960).

Nowadays, ANNs are still widely employed in a broad range of application domain, but as well on the theoretical aspect of neurocomputing. The large interest in the use of ANNs can be easily assessed by checking the number of daily publications or conferences organised around it.

2.1.3 Biological Neural Network homology

We have seen previously that early research in neuroscience and neurophysiology triggered and inspired the first scientists in the field of ANNs. We will discuss here the extent to which the analogy between the biological and the artificial world enabled this inspiration, bringing further the field of neurocomputing.

McCulloch and Pitts in 1943 initiated the work on ANNs with artificial neuron inspired by the biological neuron (McCulloch and Pitts, 1943). The word "neuron" comes from the greek *neuron* which means *nerve*, it is the building element of the nervous tissue composing the most part of the nervous system and almost exclusively located in the brain. The first description of a neuron is attributed to Santiago Ramón y Cajal in the late

1890's who was granted with the Nobel prize in physiology and medicine in 1906 (Jain et al., 1996) for this work. The nervous tissue is composed of billions of these neurons which have various types and length according to their function and location within the body (Basheer and Hajmeer, 2000). Neurons are composed of three main parts: the core of the cell (the body) is known as the *soma*; several extensions named *dendrites*; and another longer extension, the *axon* as shown in Figure 2.2. As for a regular cell, the soma contains a nucleus that holds the genetic information in DNA, and cytoplasm with the usual organelles.

However, the neuron is a unique and highly specific cell type, not only because of its morphology (the axon of the same neuron can be as long as one meter), but as well because of its physiology. The neuron is able to receive an electric signal on its dendrites, and if appropriate, it will transmit this signal towards other neurons or cells via its soma and consequently its axon. The transmission of the signal occurs at the extremities of the axon, which can split in several strands. At these extremities are the *synapses* which make the contact with the dendrites of the following neuron. The real transmission from one neuron to another is made possible by these synapses; when the signal arrives to the synapse, chemicals named *neurotransmitters* are released and travel through the synaptic cleft until they reach the following cell. Once the neurotransmitters reach the cell's membrane of the receiving neuron, they are captured by receptors which will eventually trigger and propagate an electric signal through this neuron depending if the threshold of transmission for the neuron is met and depending on the type of synapse (either inhibitory or excitatory) (Basheer and Hajmeer, 2000; Jain et al., 1996). The process of transmission is orientated with a *presynaptic* neuron and a *postsynaptic* neuron. The transmission is made possible thanks to a local depolarisation.

Therefore, ANNs are inspired by the way in which the biological neural network learns and processes information. Their concept simulates the behaviour of the biological neural network. Indeed, as discussed previously, learning involves minor adjustments to the synaptic connections between neurons as described by Donald Olding Hebb (Hebb, 1949), biological neural networks are believed to learn by following this rule. The more a synapse is used to transmit a signal toward a receiving neuron, the more reinforced this synapse will be; in the conversely, the less a synapse is used, the weaker it becomes, it could even eventually disappear. Similarly, the learning process with ANNs is based on the interconnections between the processing elements, the nodes that constitute the network topology, and learning occurs by adjusting the intensity of the interconnection weights through a number of iterations by presenting examples to the network (Basheer and Hajmeer, 2000). Figure 2.2 represents both biological and artificial neurons in a comparative manner.

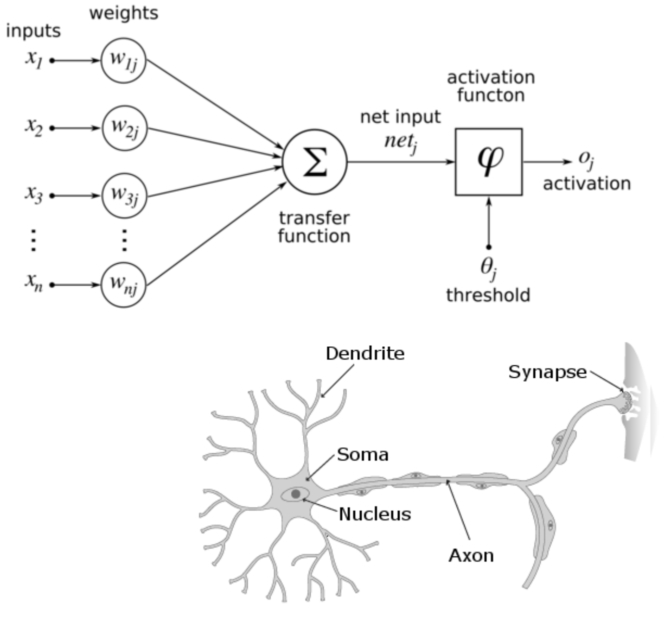


Figure 2.2: Comparison of schematic artificial and biological neuron.

2.1.4 Architectures

Although, this work involved only MultiLayer Perceptron (MLP) architecture, ANNs exist in several other form of architectures which will be briefly discussed in the following section. It has been estimated that around

50 different types of ANNs coexist (Pham, 1994). The architecture refers to how the nodes of the network are organized, mainly the layers of nodes. The choice of a particular architecture by the user will be guided mainly by the type of problem to be solved (*e.g.* some ANNs are appropriate for data modelling but not for perceptual problems (Basheer and Hajmeer, 2000)). For the purpose of this manuscript the focus will be on ANNs in their most commonly used form, the MLP, but other ANN based approaches exist; for example radial basis function networks and recurrent neural networks.

Perceptron

The Perceptron as described by Rosenblatt (Rosenblatt, 1958) is the most simple architecture of ANNs. It consists of a single artificial neuron that processes information from n inputs and provide a response thanks to an appropriate learning step. As Minsky and Papert pointed out in 1969 (Minsky and Papert, 1969), the major limitation of Perceptron is its inability to model non-linear problems. This limitation triggered the work on the evolution of the Perceptron: The MLP.

MLP

Improvement of the Perceptron has been brought by Rumelhart. The Perceptron can stand as a particular case of a MLP; indeed the Perceptron is a one-hidden node single-layer MLP. In the MLP, ANNs are organised into several layers, with each layer having a number of respective neurons, or processing elements, that constitute that layer (Figure 2.3). Simply put, the majority of ANNs have a similar topology consisting of an input layer, one or more hidden layers, and an output layer. The number of hidden layers and the number of neurons in each layer is dependent on the complexity of the problem, i.e. the number of input neurons. The input layer interacts with the external environment to receive the input data as a vector of predictor variables, each represented as a node. This information is passed through to

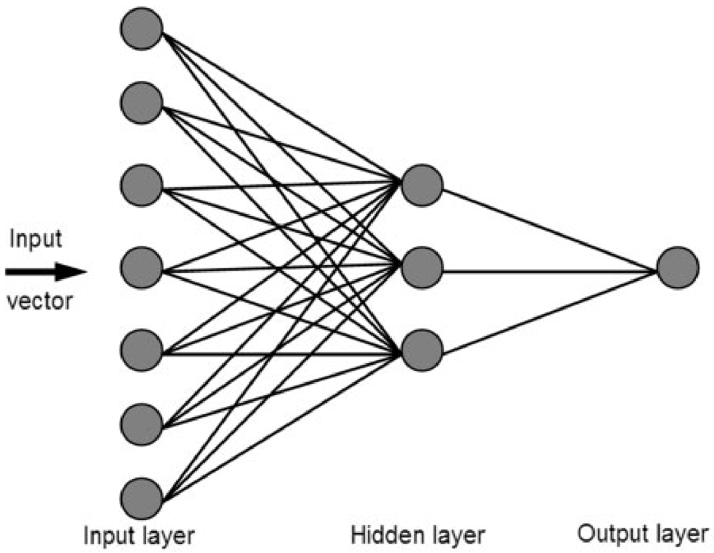


Figure 2.3: Schematic representation of a classical single-hidden layer MLP with m hidden nodes.

the first hidden layer, and multiplied (thus modified) by a set of associated weights. These products are summed and fed through a non-linear transfer function (*e.g.* sigmoid, hyperbolic tangent) which scales and then produces an output, similar to the axon of the neuron.

Other architecture

Several other architectures of ANNs are worth mentioning, but won't be discussed further in this thesis. They include Radial Basis Function or Recurrent Neural Networks.

2.1.5 Learning rules

Learning rules are algorithms that define how the weights linking the inputs to a neuron and two neurons with each other should be iteratively adjusted during the epochs of the training process in order to get optimal solution. The ultimate aim of the learning process is to find the optimal set of weights that will provide a trained ANN able to give solution that is the closest to the expected solution, in other words with the smallest error (Rumelhart et al., 1986). ANNs must then be trained to efficiently compute the gradient in order to be capable of accurately modelling a set of cases and predicting their output. There are two major learning paradigms; supervised and unsupervised.

Supervised learning algorithms

Supervised learning involves providing the network with a set of cases that have values for the inputs as well as the known desired outputs. The output of the network is then compared with the true output to calculate error by assessing the network performance as learning progresses. The interconnecting weights are initially randomised (*e.g.* [-1, 1]) so that predictions after completion of the first training cycle are essentially random. One of the most popular forms of supervised learning, *Back Propagation (BP)*, compares the error between the true output and the predicted output and then feed this error back through the layers of the network. The weights are adjusted so that after completion of the next training cycle (or epoch) the error decreases according to:

$$\omega_{ki}(\tau) = \eta\delta_k x_i \tag{2.7}$$

Each weight update ω_{ki} at the current (τ^{th}) cycle is updated in proportion to the input value to which the weight is applied x_i, the error in the output of the unit δ_k and a constant known as the learning rate η (Bishop, 1995). The weight change of a neuron is proportional to the influence an input had on the error during training and the learning rate is a constant which controls the size of these weight changes. The larger the learning rate, the faster learning will proceed; however too large a value may lead to oscillation or non-convergence of the model (Basheer and Hajmeer, 2000). Each time a pattern is presented to the network, the weights leading to an output node are modified slightly during learning in the direction required to result in a smaller error the next time the same pattern is presented, until a target error is reached or no improvement of the error is observed. A momentum term α, may be applied to help prevent the network becoming trapped in local minima, or being stuck along flat regions in error space. This occurs with a slight alteration to the weight update rule by making the weight update on the τ^{th} iteration depend on the update that occurred during the $(\tau - 1)^{th}$ iteration:

$$\omega_{ki}(\tau) = \eta\delta_k x_i + \alpha\omega_{ki}(\tau - 1) \tag{2.8}$$

This helps to speed up the time it takes for the network to reach convergence by gradually increasing the step size of the search in regions where the gradient is not changing. As with the learning rate, effectively choosing values for these constants depends on the particular problem of interest and experimentation is important here to find optimal values. In our own experiences, a learning rate of 0.1 combined with a momentum of 0.5 has proved successful (Lancashire et al., 2005a, 2008). The target error that needs to be minimized is often determined as the total sum-of-squares based on the difference between the output and target vector as follows:

$$\epsilon = \frac{1}{2}\sum_{j=1}^{n}\left(d_j - y_j\right)^2 \tag{2.9}$$

where n is the number of cases, d_j is the target network output for case j and y_j is the network predicted output for case j. Alternative error functions also exist, such as the Mean Squared Error (MSE), or the maximum conditional likelihood fitting, but will not be dwelt upon here. This learning process is an extension of the generalised delta rule, and is commonly known as BP (Rumelhart and McClelland, 1986; Rumelhart et al., 1986; Werbos, 1994).

It is crucially important that the data used in training the network should be reasonably large in order to contain all the information necessary to be able to recognise which of the predictor variables are important amongst the vast amounts of noise and individual variation that is expected to cloud important information in complex -omics datasets. If the network outputs fail to show good discrimination on an independent test dataset, over-fitting may have occurred and training must be continued or repeated. Over-fitting can occur when the number of parameters in a model exceeds the number of cases. It is in essence a memorisation of the training data (and any associated random noise) (Ransohoff, 2004; Simon et al., 2003). In order for the network to be trained to a satisfactory level which maintains generalisation for new data, it is vital to employ an appropriate regularisation technique (discussed in a subsequent section). Once learning is complete the weights are stored and can be used to predict future cases in separate test datasets. Other learning algorithms have also been proposed. These include (but are not limited to) QuickProp (Fahlman, 1988), Resilient Backpropagation (RPROP) (Riedmiller and Braun, 1993) and the Levenberg - Marquardt algorithm (Levenber, 1944; Marquardt, 1963).

Unsupervised learning algorithm

Unsupervised learning occurs when the network attempts to map the inputs to outputs without any external assistance. Therefore the network itself governs how it groups the cases based upon the input data. This is sometimes referred to as self organisation, and Kohonens self organising maps (discussed previously) (Kohonen, 1989) are the most popular form of neural network-based unsupervised learning. Other forms of unsupervised learning include PCA, independent components analysis, hebbian learning and autoassociators.

2.1.6 Regularisation

One of the main issues to solve with supervised learning is *over-training* or *over-fitting*. Such a phenomena occurs if the network achieves too well the tasks required for the training set, the examples used for learning. Consequently, the trained ANN can only model data from the training set, and poorly generalise for unseen cases (Lancashire et al., 2009a). Commonly the main purpose of modelling is to simulate a real world system and therefore a model is judged on its ability to generalise to new data. For ANNs, the risk of low generalisation is mainly attributed to over-training of the model, leading to over-fitting and subsequently poor predictive performance during independent validation. Due to the fact that even a linear model would over-fit in high dimensions, ANNs must be appropriately regularised during training in order to achieve sufficiently high predictive performances. In order to address this, regularisation techniques need to be applied during training. Several options for regularisation exist and methods can be chosen according to the type of data or generalisation performance that is required. We will briefly discuss now some of the most common forms.

Weight decay

One of the simplest regularisation methods to implement is weight decay. In weight decay, the error function includes a penalty term: for example, the sum of squared weights and biases multiplied by a decay constant that controls how much the penalty term should affect the resulting error function:

$$\lambda \sum_{i=1}^{n} w_i^2 \tag{2.10}$$

Since over-fitted models are more likely to contain unusually large weights, this approach aims to penalise such large weights, in order to keep weight values smaller than they naturally otherwise would converge at, thus keeping the activation of the neurons in the linear range (Bishop, 1995).

Resampling and early stopping

A common solution to address this issue is to stop the training process once it has reached a satisfactory error level at modelling. According to Ntzani and Ioannidis (Ntzani and Ioannidis, 2003), independent validation is only conducted in about 10% of MA studies published. Given the fact that these complex datasets are likely to be non-linear in nature, one may not have prior information regarding the intricacies of the data. As such it is vital to estimate the performance of these models on new data in order to be confident that over-fitting has been avoided. It was stated earlier that the BP algorithm should stop training once the network has achieved an acceptable level, however, the question remains as to what is considered to be an acceptable level, and what can be done to ensure that the model will be capable of generalising to additional future cases. If training is terminated solely on the basis of a set number of iterations the model is at risk of over-fitting. The most universal approach to address this problem is resampling. Typically in ANN-resampling approaches, the data is split into different subsets, where a percentage of the total sample set is used to train and optimize the ANN (the training set) and (sample size permitting) the remaining are partitioned for validation during training (the validation set) and external testing after the modelling is complete (the independent test set). This validation approach will be referred to as Monte Carlo Cross Validation (MCCV).

A common, yet efficient, regularisation technique known as early stopping mechanism monitors the network error with respect to a validation or test dataset. This process signals to stop training either when a predetermined number of iterations have completed, or when the prediction accuracy of the model begins to worsen for the validation or test dataset, a sign of over-fitting. The weights resulting in the minimum validation or test error are then selected. Once the network has completed the learning process, it is further validated using the test data split, to give an unbiased estimation of the networks likely performance on future unseen cases. Examples of this approach can be found in (Lisboa and Taktak, 2006) and (Basheer and Hajmeer, 2000).

Bayesian regularisation

The Bayesian regularisation approach involves modifying the target function (such as the sum of squared errors) in order to improve the models generalisation ability (Kelemen and Liang, 2003). The Bayesian regularisation aims to smooth the cost function by adding to it a regularisation parameter based on the sum of squared weights. To reduce bias, the weights and variables of the network are assumed to follow a Gaussian distribution and are assigned prior probabilities, optimised according to the Bayesian framework of Mackay (MacKay, 1992). Network training then attempts to find the trade-off between minimising the model complexity and model error, as such minimising both the bias and variance (Kelemen and Liang, 2003). Methods such as automatic relevance determination will identify and remove unnecessary parameters from the model since the Bayesian approach provides an estimate for the entire distribution of model parameters rather than a single optimal set of weights. Model comparison is based on highest evidence, rather than cross validation, and as such Bayesian regularisation maximises the data available as it does not require a validation set since all the training data can be used for model fitting. A review of Bayesian methods for supervised neural networks can be found in (MacKay, 1995), and an example of its application in a MA study can be found in (Kelemen and Liang, 2003).

2.2 Stepwise approach principle

ANNs have shown remarkable abilities for data-mining and pattern recognition as previously demonstrated (Bishop, 1995). However, data-mining in the field of biomedical sciences remains highly challenging especially for the analysis of genomic data. Genomic technologies (*e.g.* MAs) have triggered a new era in biomedical research by allowing the simultaneous assessment of the level of expression of many genes within a single experiment. This era comes with its own challenges, mostly regarding the analysis of the data.

One of the main limitations for the application of ANNs in genomic technologies is that, despite their ability to cope with high dimensionality within the data, they can still suffer from what is known as the *curse of dimensionality*, firstly described by Bellman (Bellman, 1961) as *the exponential growth of the input space as a function of dimensionality.* In other words, the importance of a particular feature of genes can be hidden amongst the vast amount of the other gene expression vectors. It occurs mainly when the number of variables P is far higher than the number of cases n (P » n), leading to a data space with noise and irrelevant inputs, consequently leading to poorer performances of the model to unseen data, hence poorer generalisation (Bishop, 1995). Recommended methods to overcome such an issue usually involves the application of pre-processing procedures or feature extraction algorithms (Bishop, 1995) as used for dimensionality reduction. However, given the disadvantages and advantages of the approach, the idea of independently testing each gene for the particular question, and subsequently test multiple combinations with the best performing single genes, assessing and comparing their predictive performance, arise and gave the opportunity to increase the gain of reliable information out from the data.

2.2.1 Stepwise algorithm

This approach consists basically in a wrapping method of a classical ANN that enables identification of patterns within genomic data in an additive manner by finding the best individual variables performing the highest to classify the dataset regarding the question studied and build the model by iteratively adding subsequent variables to further improve the classification. Initially, the approach uses each gene from the microarray experiment as an individual input in the ANN, thus creating n individual models (n being the number of genes studied in the experiment). Subsequently, the n models are compared and sorted according to their predictive performances for unseen cases (cases used in the Test subset from the MCCV). Thus, we can rank each gene according to their modelling ability for the question investigated on blind data. Then, the best input (*i.e.* gene) is selected and further tested in 2-input models with the remaining genes, therefore creating $(n - 1)$ models. The best 2-input model is selected similarly, and is further improved by testing its performances combined to the remaining $(n - 2)$ genes. The process is repeated until no improvement of predictive performance of the model is observed or until it has reached an optimal level of performance (Lancashire et al., 2008). The algorithm consists as follows (Lancashire et al., 2005c):

1. Each of the variables is used as a single input in a one-input model, creating the n single models.

2. Each model is then trained over 50 events of MCCV[1], meaning that all the samples were randomly reshuffled to ensure that all are considered blind for a number of models, in order to improve the ability of the network to generalise well for unseen cases.

3. The predictions and MSE across the 50 sub-models for test subset are monitored and recorded for each single-input model, and these inputs are then ranked based on their MSE.

4. The input within the model predicting the best (*i.e.* with the lowest error) is then selected for the second step.

[1] 50 MCCV folds were found to be the number for which the models started to reach consistency (Lancashire, 2006).

5. At the following step, the input that performed the best in the previous one is used as the basis for two-input models.
6. The remaining inputs $(n-1)$ are then sequentially added to create $(n-1)$ two-input models.
7. 50 sub-models are then trained for each of these two-input models, and their performance is monitored as explained earlier.
8. The performances allow us to rank the best two-input model, and select the combination of two inputs for the third step.
9. The process is repeated until no improvement in network performance is observed, or if any early-termination condition is met.

A more detailed version of the algorithm is shown in Figure 2.4, and a diagram in Figure 2.5.

This approach has been successfully applied in many different studies conducting to key discoveries of reliable and validated biomarkers mainly in cancer and especially breast cancer (Lancashire et al., 2008, 2009b). Some examples of these applications will be discussed further in the present manuscript.

2.2.2 ANN architecture for Stepwise parametrisation

Originally, the Stepwise approach consisted of a Visual Basic coded program named "Stepwise ANN modelling" interfacing the Neural Networks package of the software ©Statistica (StatSoft, Inc. (2004). STATISTICA (data analysis software system), version 7. www.statsoft.com.). Despite the numerous disadvantages of choosing to interface an external software, this solution allowed the developer to initially assess the validity and robustness of the approach without the heavy task to develop a complete ANN software. Therefore, it was possible to define the optimal settings and procedures to obtain the most efficient and reliable approach for the analysis

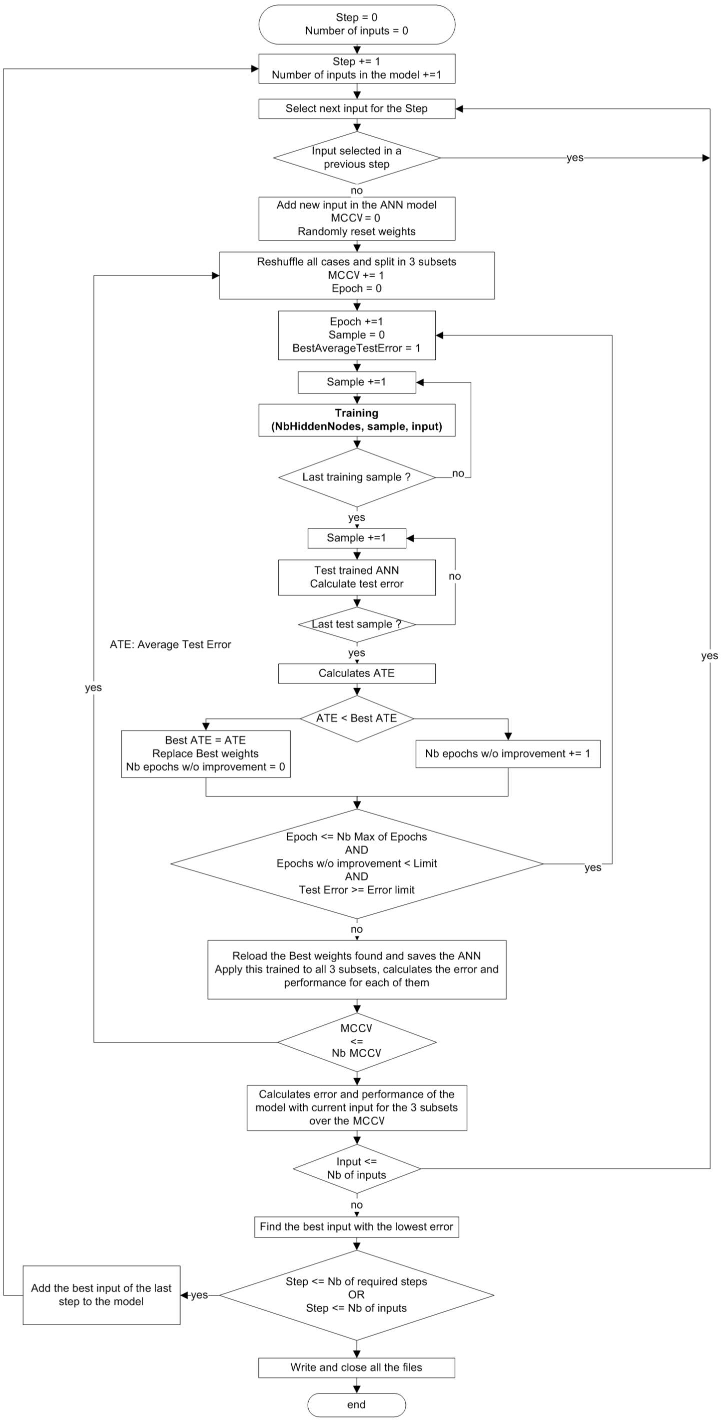

Step = 0
Number of inputs = 0
Step += 1
Number of inputs in the model +=1
Select next input for the Step
Input selected in a previous step
yes
no
Add new input in the ANN model
MCCV = 0
Randomly reset weights
Reshuffle all cases and split in 3 subsets
MCCV += 1
Epoch = 0
Epoch +=1
Sample = 0
BestAverageTestError = 1
Sample +=1
Training
(NbHiddenNodes, sample, input)
Last training sample ?
no
yes
Sample +=1
Test trained ANN
Calculate test error
no
Last test sample ?
yes
ATE: Average Test Error
Calculates ATE
yes
yes
ATE < Best ATE
Best ATE = ATE
Replace Best weights
Nb epochs w/o improvement = 0
Nb epochs w/o improvement += 1
Epoch <= Nb Max of Epochs
AND
Epochs w/o improvement < Limit
AND
Test Error >= Error limit
yes
no
Reload the Best weights found and saves the ANN
Apply this trained to all 3 subsets, calculates the error and performance for each of them
MCCV
<=
Nb MCCV
Calculates error and performance of the model with current input for the 3 subsets over the MCCV
Input <=
Nb of inputs
no
Find the best input with the lowest error
Add the best input of the last step to the model
yes
Step <= Nb of required steps
OR
Step <= Nb of inputs
Write and close all the files
end

Figure 2.4: Schematic representation of the Stepwise algorithm

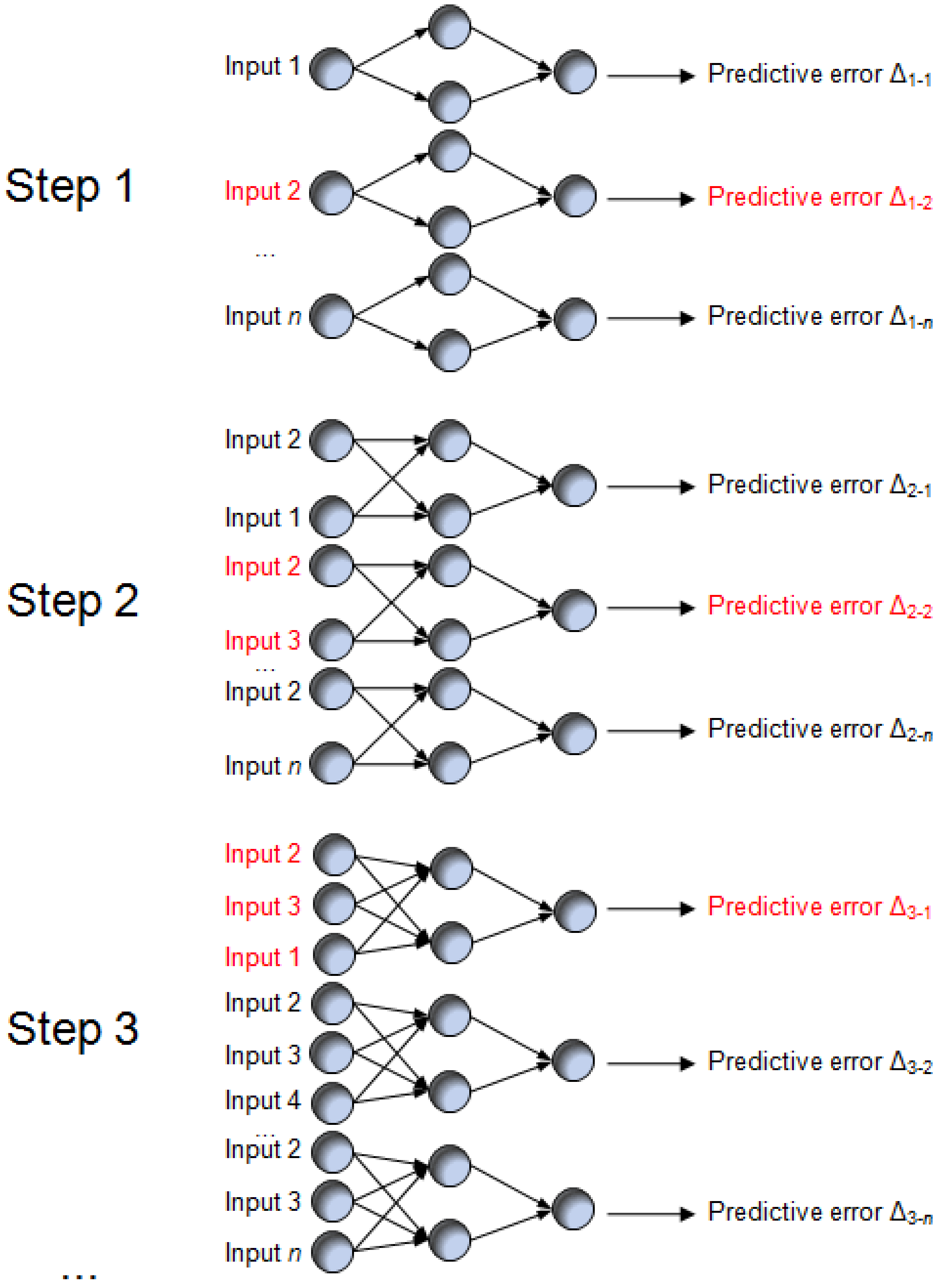

Step 1
Input 1
Input 2
...
Input n
Predictive error $\Delta_{1\text{-}1}$
Predictive error $\Delta_{1\text{-}2}$
Predictive error $\Delta_{1\text{-}n}$
Step 2
Input 2
Input 1
Input 2
Input 3
...
Input 2
Input n
Predictive error $\Delta_{2\text{-}1}$
Predictive error $\Delta_{2\text{-}2}$
Predictive error $\Delta_{2\text{-}n}$
Step 3
Input 2
Input 3
Input 1
Input 2
Input 3
Input 4
...
Input 2
Input 3
Input n
Predictive error $\Delta_{3\text{-}1}$
Predictive error $\Delta_{3\text{-}2}$
Predictive error $\Delta_{3\text{-}n}$
...

Figure 2.5: Diagram depicting the principle of the Stepwise algorithm (in red are the models at each step found to have the best modelling performance, i.e. the lowest modelling error, hence selected for the subsequent steps).

of complex genomic or proteomic data (Lancashire, 2006; Lancashire et al., 2008).

The Stepwise approach is based on a single hidden layer MLP with two hidden nodes using a BP learning algorithm. An early termination strategy is employed in order to avoid over-fitting of the data on the training subset with a maximum number of epochs of 3000, a window of 1000 epochs (*i.e.* training stops when no improvement of the model is observed after 1000 epochs) and a MSE of 0.01. Learning rate is set at 0.1 and momentum at 0.5 for the BP algorithm. Initial weights of the network are randomly set within a range between -1 and +1. A MCCV strategy is applied as well, in order to ensure the good generalisation ability of the model obtained, randomly splitting respectively 60%, 20% and 20% of the cases for training, validation and independent test purposes. The MCCV is repeated 50 times resampling all the cases in order to have them randomly affected differently to the different subsets. The same settings were used for all the biomarker discovery studies conducted during this Ph.D. project. However, some further studies were carried out to assess how the methodology behaves.

2.3 Example of application of the original Stepwise approach to a gene array in a breast cancer study

2.3.1 Introduction

The study to be discussed here focused on the exploitation of a dataset previously described and used in a study by van't Veer and colleagues (van't Veer et al., 2002). This group firstly identified a set of 70 genomic markers to predict distant metastases in breast cancer. Transposing such a signature into clinical routine then becomes challenging; the development of an assay based on 70 gene transcripts would be really complicated, however it has been recently transposed into clinical aspect with the Mammaprint assay from Agendia (Slodkowska and Ross, 2009). Therefore, it could be interesting to reduce the number of markers to investigate, with a similar or even better predictive ability. Using the ANN stepwise approach, it has been possible to reduce the number of markers down to 9 able to predict the outcome of distant metastasis for breast cancer patients with an accuracy of about 98%. The results of that study showed that the expression of Carbonic Anhydrase IX (CA9) was able, as a single predictor, to predict the metastatic events with an accuracy of 70%. The identified signatures were subsequently tested against a second validation cohort of 295 cases (van de Vijver et al., 2002). Here again, CA9 expression on its own was able to predict metastatic outcome with an accuracy of 63%. Finally, investigation of the protein expression of CA9 was conducted by immunohistochemistry which was carried out on an independant set of samples. This work contributed to (Lancashire et al., 2009b).

2.3.2 Materials and methods

Dataset and ANN analysis

The data were extracted from a microarray experiment in breast cancer (van't Veer et al., 2002). They were directly imported as a Microsoft Excel format from *http://www.rii.com/publications/2002/vantveer.html*. It consisted in a pool of 78 patients, with mixed ER status and no lymph node positive cases, and with 24,481 values for each corresponding to the Log10 expression ratio of the genes. This particular dataset was chosen as it is a well known and characterised dataset. However, this can be considered as an historical dataset and therefore involves some pitfalls. Indeed, the quality of the data relies on the quality of the technology back in that time. Even more, the quality of the samples has to be replaced in the context, in 2002 the technologies for detection and diagnosis were not the same than nowadays, as a result, the samples collected may present higher aggressiveness since the tumours may have been detected at a later stage in the development of the tumour.

The stepwise approach described earlier was applied in order to identify the genes of importance regarding the question of distant metastasis. The input vector was represented as analogue value with the log10 ratio of the expression of each gene investigated. The output for the individuals who presented a distant metastasis within 5 years was coded as 1, and 0 if no evidence of metastasis was found within the five year period. A classical architecture with MLP was applied in the ANN developed, with two hidden nodes in the hidden layer, and a sigmoidal transfer function. Learning rate was set at 0.1 and momentum at 0.5. Here again, overfitting of the ANN on the training dataset was avoided using a MCCV approach that created 50 submodels, each with different random split of the cases in 3 different subsets: 60% of the cases in the training set, 20% in the validation set, and 20% in the independent test set. Early stopping strategy interrupted model development after 1,000 epochs without improvements, or when minimal error of 0.01 was obtained.

Patient selection and tissue microarray

As mentioned in (El-Rehim et al., 2005), tissue microarray blocks containing 555 breast tumour samples from patients involved in the Nottingham Tenovus Primary Breast Carcinoma Series were used for the validation. The tissue samples were coming from tumour periphery avoiding necrotic tissue. All the cases involved in the study were well characterised with complete follow-up data with protein expression, pathology and clinical information, including information on tumour reccurence, metastasis and survival. Patients characterised with ER positive phenotype had received adjuvant endocrine therapy whilst individuals with poorer Nottingham Prognostic Index received chemotherapy. For CA9, the protein expression was further validated on a separate cohort of 245 patients for which again, comprehensive follow-up data was available, including histopathological information. These patients were managed at the Royal Marsden Hospital, where they received therapeutic surgery followed by similar therapy as previously described. For all tumour samples collected, grade and size were determined by pathologists according to a modified Bloom-Richardson scoring system (Bloom and Richardson, 1957) and the Tumour, Node, Metastasis (TNM) staging criteria.

CA9 immunohistochemistry

The immunohistochemically stained tissue microarrays were scored by pathologists blindly to the patients' clinical and pathological information. Staining was conducted in neoplastic cells of tumours and in stromal fibroblasts. A previous study suggesting the involvement of CA9 in prognostic outcome was the reason for the staining in stromal fibroblasts of this particular protein (Colpaert et al., 2003). Positive or negative staining in the samples were recorded respectively as 0 or 1. No scoring was conducted for altered tissues or tissue showing no invasive carcinoma.

2.3.3 Results

Model development

The analysis of the dataset by the stepwise approach identified a gene signature of 9 genes able to predict with 98% sensitivity and 94% specificity for prognostic prediction, with an Area Under the Curve (AUC) of 0.971 with Receiver Operating Characteristic (ROC) curve analysis. A summary of the results at each step of the model development is shown in Table 2.1.

Following this development stage, the identified 9-gene signature has been tested upon a different set of data consisting of 19 previously unused samples (therefore completely blind to the model developed). This set included 7 metastasis-free patients and 12 metastatic ones. The 9-gene model succeded to correctly discriminate the 19 samples showing the relevance of

Step	Input added	Gene name	Description	Cumulative accuracy	Error	Response
1	NM_001216	CA9	Carbonic anhydrase IX	70%	0.44	Pos
2	Contig52778_RC	EST		80%	0.38	Weak
3	Contig35076_RC	EST		83%	0.38	Neg
4	Contig40557_RC	FLJ13409	EST	87%	0.35	Pos
5	AB032973	LCHN	LCHN protein	80%	0.4	Pos
6	AB004064	TMEFF2	Transmembrane protein with EGF-like and 2 follistatin-like domains 2	95%	0.23	Pos
7	NM_006101	HEC/KNTC2	Kinetochore associated 2	95%	0.22	Weak
8	AF161451	HSPC333	HSPC337	96%	0.17	Neg
9	Contig33475	EST		98%	0.15	Weak

Table 2.1: The 9-gene signature identified at each step of the model development by the stepwise approach

the signature.

Response curve analysis for the 9 genes was also conducted to further question the influence of each gene to the phenotype. This showed that 7 out of the 9 genes have a strong discriminatory response; in addition, 4 of them (CA9 being one of them) showed a positive relationship between increased gene expression and higher probability of developing metastasis (see Table 2.1). Conversely, 3 genes showed the reverse relationship, meaning that a higher expression level of the gene resulted in a lower metastatic risk. The 2 genes having weak response can still play a role in regulating or modulating the other ones, which could explain why they appear in the signature obtained by additive approach.

Validation of the signature

The application of the 9-gene signature managed to correctly classify the population of 19 samples, thus showing the good generalisation ability of the identified model, where the 70-gene signature identified in the original paper misclassified 2 out of the 19 samples. A subsequent validation stage was accomplished using the NKI295 dataset (van de Vijver et al., 2002) including a cohort of 295 patients to the 9-gene model identified. The Kaplan-Meier analysis of the two groups obtained after the discrimination of the population by the signature showed significant difference for both overall survival ($P<0.001$)[2] and metastasis-free survival ($P<0.001$). The signature obtained was able to discriminate the groups defined by the original 70-gene signature from the van't Veer study into prognostically distinct groups ($P<0.001$). On the other hand, in a Cox regression model, using age, nodal status, Oestrogen Receptor (ER) status, tumour size, type of therapy and van't Veer's 70-gene signature, the 9-gene signature was a predictor of overall survival ($P = 0.012$, Hazard ratio = 1.89) and metastasis-free survival ($P = 0.003$, Hazard ratio = 1.92). In addition, CA9 gene expression showed a significant ($P<0.001$) positive association with tumours of basal-like phenotype and an inverse association with luminal type cancers ($P<0.001$). Further investigation followed to assess the feasibility of an immunohistochemically-based assay transferable to routine clinical practice. For more details about this and about the immunohistochemical validation of CA9 expression, the reader is referred to the paper (Lancashire et al., 2009b) in Appendix I.

2.4 Discussion

The focus of this study was to develop an optimal yet minimal gene signature able to accurately predict the outcome of distant metastasis in breast cancer patients using our regular ANN method coupled with the stepwise approach with a previously published dataset (van't Veer et al., 2002). The original aim of this work was to at least be able to obtain a similar degree of accuracy with the van't veer study. The stepwise approach, specifically designed to identify reliable biomarkers within complex datasets, was able to identify a set of 9 genes with the ability to accurately predict (98% sensitivity) metastatic outcome for breast cancer patients. Thus meeting our goals to improve the original identified set of 70 genes predicting with an accuracy of 83%, with a very smaller number of genes in the signature.

[2] P is an indicator of the statistical significance of the difference of the mean between two variables over a population.

The main gene identified was CA9, and was able, by itself to predict 70% of metastatic cases from the first cohort, and 63% from the NKI295 cohort (van de Vijver et al., 2002). The group of Carbonic anhydrases, in which belongs CA9, is induced by the Hypoxia Induced Factor 1 α (HIF-1α). They play a role in protecting cancer cell from death by neutralising acidic conditions from hypoxia-induced glycolysis. In addition, it seems that CA9 has a role in degradation of the extracellular matrix and induction of growth factor, which could enhance cancer migration and invasion (Kallergi et al., 2009). More details can be found in the published article (Lancashire et al., 2009b).

2.5 Summary

This chapter focused on the ANN principles, theories and architectures. Additionally, it showed the extension of the architectures, and how it can be optimised for the purpose sought here: the identification of potential new biomarkers from high-throughput data in biomedical sciences. In order to highlight this, it finally discussed a successful application in a study aiming to identify new, reliable and more importantly validated biomarkers in breast cancer for the prediction metastatic events. The results obtained in this example clearly showed the robustness of the approach presented here. However, clear limitations occurred and the approach requested some optimisation for more reliable and smoother applications.

Chapter 3

Optimisation of the Stepwise approach

This chapter will focus on the optimisation of the techniques discussed in the previous chapter for the purpose of the identification of biomarkers. We will detail how the optimisation was carried out, and how the improvements were assessed. This optimisation step helped to improve the selection step of important genes necessary for the interaction algorithm as it will be further discuss in the following chapter. Finally, another example of application of the newly developed stepwise approach will be discussed.

3.1 Stepwise approach optimisation

Although the program presented previously has shown great advantages, this solution quickly showed some important limitations, mainly due to the constraints of using an external software, not only because of the cost issue for the licence, but as well as the maintenance or upgrading side (*e.g.* the version 8 of ©Statistica was no more supporting cross-validation), and unexpected crashes after about a day of process.

3.1.1 Visual Basic software development

As a result, it was decided to develop an integrative and standalone new application in Visual Basic able to build ANN MLPs, and able to integrate several types of analysis. This software was developed in Visual Basic in the first place, in order to obtain a fast stage of development and allow a rapid validation and application stage, despite the poor speed of execution inherent to Visual Basic code. The software used a classical single hidden layer ANN MLP core, around which were wrapped several layers of algorithms in order to use the Neural Network core in different ways. The further algorithms incorporated into the new system, additionally to the Stepwise, at that stage were :

- **Sensitivity analysis** is a well-known type of analysis that tests a dataset by iteratively removing single elements to monitor how the removal affects the predictive performance of the model. This performance obtained is assessed and a ratio is given to the removed input.
- **Cascade analysis** iteratively removes correctly classified cases by the ANN with all the inputs successively tested.

These ANN-based approaches have been successfully implemented in the software and were working correctly. However, only the Stepwise approach has been intensively applied to studies, and only this particular one will be further discussed and described.

The software has been designed so that it can integrate any type of algorithm to wrap around the ANN kernel of the software, therefore, any idea can be quickly implemented as another layer around the core to use the Neural Networks.

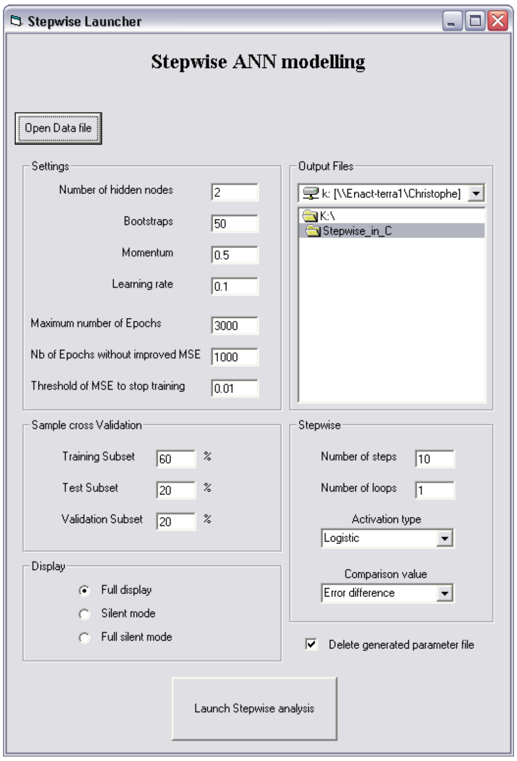
3.1.2 Application in C

Despite the great quality of the approach developed and presented in the previous chapter, several improvements were needed in front of the limitations encountered during application stage. One major issue being the poor speed of execution, which is the case for the solution presented in the previous section. Therefore, it was considered a redevelopment of the solution in a different and much faster language : C. Not only does this language allow a faster speed of execution; but as well, its portability makes it possible to apply the program in different environments such as Linux or Unix. These environments are well known to have better speed of execution, and more reliable memory management, a key issue in our particular situation for massive data analysis.

The development stage of the application in C, was straight forward. It revealed itself as a simple translation from Visual Basic to C, with no major difficulty beside the usual memory management that can arise with such code. Indeed, a careful management of the memory allocated needs to be carried out as it usually is the pitfall of programming in C.

For an easy setting of the parameters needed by the Stepwise approach, the strategy of an external parameter file was chosen. It is indeed easier to alter the particular features of the parameters in a separate file read by the main program, rather than altering them within the code and recompiling the whole program everytime needed.

Finally, in order to obtain a more user-friendly program, and make it available for biologists, a simple interface has been created in Visual Basic, so that it is possible for scientists unaware of command lines to utilise the software. Figure 3.1 present a snapshot of this interface. It allows the user to quickly set the appropriate parameters, although they are set by default to our regular settings. The user can specify the path where the file to analyse is stored, and the folder where the results will be saved. When this is done, the user can hit the "‘Run"’ button, the interface creates the parameter file and launches the Stepwise program in C. Finally, this one starts to read the parameter file, and can get started with the analysis itself. In addition, this interface has been designed so that it was easy to implement some security features (such as licence key).



Number of hidden nodes: 2
Bootstraps: 50
Momentum: 0.5
Learning rate: 0.1
Maximum number of Epochs: 3000
Nb of Epochs without improved MSE: 1000
Threshold of MSE to stop training: 0.01

k: [\\Enact-terra1\Christophe]
K:\
Stepwise_in_C

Training Subset: 60 %
Test Subset: 20 %
Validation Subset: 20 %

Number of steps: 10
Number of loops: 1
Activation type: Logistic
Comparison value: Error difference

● Full display
○ Silent mode
○ Full silent mode

☑ Delete generated parameter file

Figure 3.1: Screenshot of the Stepwise interface in Visual Basic

3.1.3 Monitoring improvements of optimised Stepwise

For the purpose of assessing the improvements brought by the optimised software, a quick monitoring has been carried out. It mainly consisted of assessing how the speed of processing improved with the newly developed program, and if any improvements of modelling error could be observed.

The monitoring has shown some great improvements in terms of speed and predictive error. Indeed, the time of process has been divided by up to 5 times, and the predictive error on the validation subset has been divided by a factor 2 on average (from 0.379 to 0.184). This assessment was conducted and determined by analysing the publicly available dataset from West et al. for 10 independent repeats (West et al., 2001). This particular dataset had been chosen as it is a well known, and it is been studied and used intensively by the group. The signature used for the assessment purpose consisted of recreating the predictive error ranking of the 7 markers previously identified as the best predictors for lymph node status in breast cancer for 49 patients by prior analysis (Lancashire et al., 2008):

- AFFX-CreX-3_st
- M83221_at
- S79862_s_at
- U39817_at
- U63139_at
- M83652_s_at
- U30894_at

The predictive error on the validation subset (Figure 3.3) and the time of execution to train the models (ranking the 7 genes by predictive error) were then monitored over 50, 75 and 100 bootstraps (Figure 3.2). In short, the monitoring consisted in the assessment of predictive error and speed of execution for the stepwise to find the same rank of genes in the steps than previously found and reported (Lancashire et al., 2008).

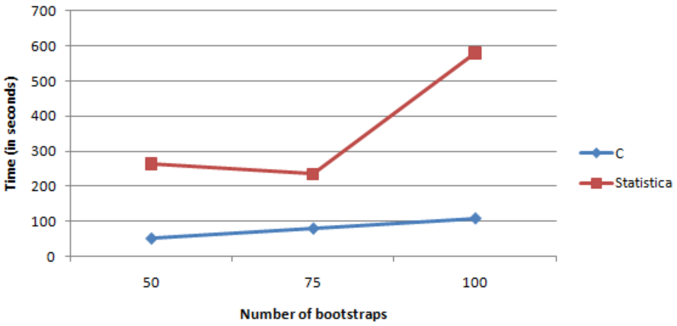


Figure 3.2: Time of execution for Statistica and the stepwise approach coded in C

On average it is now possible to run a 10-step stepwise analysis in the same time as a single-step run with the previous software when interfacing Statistica. We are then able to provide a quicker answer to biologists problems with a model and biomarkers of interests in a much shorter time period: instead of speaking in week(s), we can now speak in day(s). In addition, the average error, through different numbers of bootstraps have been compared between the stepwise in C and with Statistica. For the validation subset (Figure 3.3), it clearly appears that a better and more accurate modelling process is achieved with the Stepwise coded in C. This can be explained by the choice of a sigmoid transfer function for the program in C rather than a linear transfer function used by Statistica for the output node. In addition, since this monitoring has been achieved on the validation subset, we can argue that this improvement is not due to any sort of over-fitting reason, as this subset is completely unseen during the training process of the ANN.

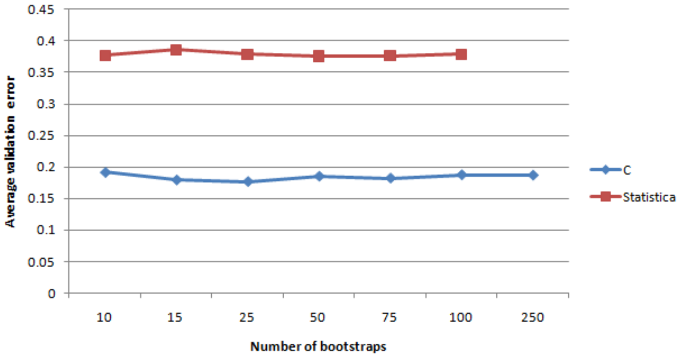


Figure 3.3: Comparison of the average predictive error on validation subset for different number of cross validation folds (number of bootstraps).

3.1.4 MCCV repeats assessment

The survey has been conducted as well to asses the ideal number of repeats for MCCV to consider for an appropriate balance between a speed of processing and a reliable modelling.

The results (Figures 3.2 and 3.3) showed that no real improvement could be really observed after 50 folds and this number remains ideal to keep an optimal balance between the time of execution and the development of a robust model, with accurate predictive ability and good generalisation. These results confirmed the results obtained by Dr. Lancashire previously (Lancashire, 2006).

Looping system

Several other improvements have been implemented. These include a looping system, which simply repeats an analysis completely independently to previous ones. This helps to run a Stepwise analysis over a number of repeats without having to be in front of the screen each time it needs to be re-launched for the user.

3.1.5 Output files

A clearer output filing system and summary file using American Standard Code for Information Interchange (ASCII) file format has been implemented as well in order to clarify the output of the analysis. An overall summary is now written at the end of the analysis to show to the final user the overall results, for each step and each loop. Even details on each bootstrap are recorded, in the way that if anything went wrong during the training process, it is possible to have an insight into the development of the model and track down where can be situated a problem that appeared.

3.2 Application of the optimised approach : identification of miRNA signatures to predict ER, PR and HER2/*neu* status for breast tumours

The study presented here refers to the discovery of significant validated diagnostic markers for breast malignancies. There is indeed great interest and efforts around discovery of new biomarkers for this particular form of cancer. The high heterogeneity of phenotypes in breast carcinomas still makes it one of the most challenging diseases in the clinical management of the patients, especially from a diagnosis and therapeutic point of view. Such heterogeneity tends to harden the work of clinicians who need accurate phenotyping of the tumours in order to provide the most precise and efficient treatments and therefore to maximise the benefit from the therapy for the patients. Highly divergent phenotypes result from a wide range of different genotypes. Therefore, identifying the underlying genotypes (*i.e.* genomic profiling) that can govern the mechanism of the subtypes in malignancies will lead to breakthrough discoveries ultimately leading to new clinical management of patients.

3.2.1 Introduction

miRNAs have seen a growing interest lately, as shown by the increasing number of publications during these years (Figure 3.4). They are now recognised as highly abundant regulatory molecules and they can offer another potential therapeutic strategy for malignancies, or a new prospect for diagnostic tools for the characterisation and subtyping of carcinomas. Importantly, miRNAs have been recently shown to be dysregulated in breast carcinomas, and some miRNAs play a role in tumorogenesis, metastasis and invasion (Lowery et al., 2008). Therefore comes the interest to identify miRNAs signatures of importance and monitor the correlation between these signatures and the clinicopathological classification of the tumours and the patients.

In the present study, it was sought to assess potential roles of several miRNAs expressed in breast tumour tissue related to the ER, and status of the patients. By identifying those miRNAs, we could then build a model thanks to ANN modelling, in order to predict patients' status according to the level of expression of the identified miRNA signatures. Those identified signatures could then provide new insight in breast carcinogenesis and breast cancer subtyping subsequently leading to potential new diagnosis tools and treatments.

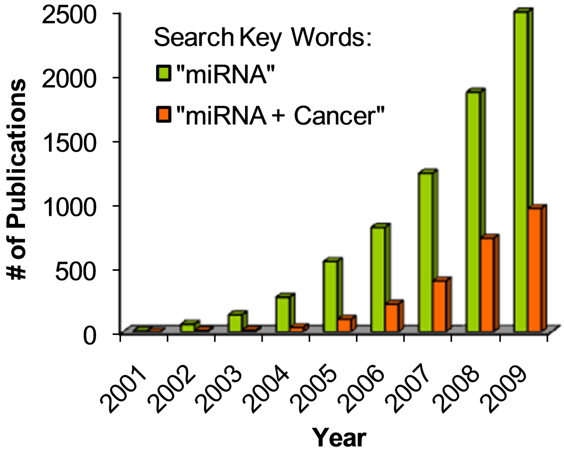


Figure 3.4: Increasing number of publications concerning total miRNAs (in green) and miRNAs in cancer (in red) between 2001 and 2009 (source: www.mirnatherapeutics.com).

3.2.2 Materials and methods

The samples were collected by surgically extracting breast tumour tissue from patients during primary curative resection at Galway University hospital, Galway, Ireland. The initial pool of patients consisted in 29 early-stage invasive ductal breast cancer tumours (Table 3.1) used for microarray miRNA expression experiment. A second cohort of 95 breast tumour samples, and 17 tumour-associates normal breast tissue specimen used as control references (Table 3.2) were used for validation purpose with . Immunohistochemistry determined the ER, PR and HER2/*neu* status for each patient involved.

After extraction, the RNA samples devoted to microarray were Cy-dye labelled and hybridised on LNA miChip microarray platform (Exiqon version 7, 453 miRNA sequences). Slightly modified RNA extraction protocol was employed for the samples devoted to validation with qRT-PCR. Prior ANN analysis, preprocessing methods were applied to capture intensity level of expression for miRNAs. The resulting data from the microarray were submitted to the genomic data repository platform Gene Expression Omnibus of the (http://www.ncbi.nlm.nih.gov/geo/) under the accession reference: GSE15885.

Patient	Age (years)	Tumour size (mm)	Lymph node status	Grade	Stage	ER	PR	HER2	Subtype
1	49	23	Negative	1	2A	P	P	N	Luminal A
2	52	30	Negative	3	2A	N	N	P	Her2 overexpressing
3	57	45	Negative	3	2A	N	N	P	Her2 overexpressing
4	51	21	Negative	3	2A	P	P	N	Luminal A
5	68	15	Negative	3	1	P	N	N	Luminal A
6	42	22	Negative	3	2A	N	N	N	Triple negative
7	54	26	Negative	3	2A	N	P	N	Luminal A
8	35	22	Negative	3	2A	P	P	N	Luminal A
9	50	16	Negative	3	1	N	N	N	Triple negative
10	49	25	Negative	2	2A	N	N	N	Triple negative
11	59	20	Negative	3	1	N	P	N	Luminal A
12	58	22	Negative	3	2A	P	P	N	Luminal A
13	58	18	Negative	1	1	N	P	P	Her2 overexpressing
14	66	22	Negative	3	2A	P	P	N	Luminal A
15	56	17	Negative	1	1	N	N	N	Triple negative
16	48	30	Negative	3	2A	N	N	P	Her2 overexpressing
17	60	26	Negative	3	2A	P	P	N	Luminal A
18	56	29	Negative	2	2A	P	P	N	Luminal A
19	50	3	Negative	2	1	P	P	N	Luminal A
20	40	7	Negative	1	1	P	P	N	Luminal A
21	40	6	Negative	2	1	P	P	N	Luminal A
22	58	35	Negative	2	2A	P	P	N	Luminal A
23	64	34	Negative	3	2A	P	P	N	Luminal A
24	66	26	Negative	1	2A	P	P	N	Luminal A
25	84	16	Negative	2	1	N	P	N	Luminal A
26	57	7	Negative	3	1	N	P	N	Luminal A
27	68	35	Negative	3	2A	P	P	N	Luminal A
28	40	20	Negative	2	1	P	P	P	Luminal B
29	49	35	Negative	3	2A	N	N	N	Triple negative

Table 3.1: Clinical and pahtological information for the original 29 patient cohort for microarray experiment.

Breast cancer clinicopathological characteristic	Number of patients (n = 95)
Median (interquartile range) tumour size (mm)	23.5 (17.75 to 35.0)
Histologic subtype	
Invasive ductal	80
Invasive lobular	13
Colloid/mucinous	1
Tubular	1
Tumour-associated normal	17
Intrinsic subtype	
Luminal A (ER/PR+, HER2/neu-)	47
Luminal B (ER/PR+, HER2/neu+)	21
Her2 overexpressing (ER-, PR-, HER2/neu+)	11
Triple-negative (ER-, PR-, HER2/neu-)	11
Missing data	5
Grade	
1	14
2	26
3	53
Missing data	2
Nodal status	
Node-negative	50
N1	17
N2	17
N3	11
Oestrogen receptor status	
Positive	62
Negative	32
Missing data	1
Progesterone receptor status	
Positive	58
Negative	33
Missing data	4
Her2/neu status	
Positive	32
Negative	59
Missing data	4
UICC stage	
Stage 1	23
Stage 2a	29
Stage 2b	8
Stage 3a	14
Stage 3b	4
Stage 3c	8
Stage 4	9

Table 3.2: Clinical and pathological information for the the second cohort with 95 patients involved for the independent validation using PCR.

For this study, the ANN stepwise approach has been applied, using a three-layer feedforward MLP with a BP algorithm and sigmoidal transfer function as described in chapter 2. The data were initially normalised between 0 and 1 for each variable casewise. The individual miRNA intensity values were presented to the ANN as input, and classes to predict (ER, PR and HER2/*neu*) were presented to the output layer (coded as 0 for negative and 1 for positive samples). The cross validation with an early termination strategy was applied to the approach in order to avoid overtraining of the model.

qRT-PCR quantification was performed to further validate the findings obtained with the microarray experiment coupled with the ANN stepwise approach. The relative quantity of miRNA expression was obtained with the comparative cycle threshold method, or $\Delta\Delta$Ct method. Analysis of these results was performed using the Kolmogorv-Smirnov normality test.

3.2.3 Results

The analysis of the microarray by the ANN stepwise method allowed us to identify fingerprints of miRNAs that can accurately predict status of patients associated to ER, PR, and HER2/*neu* for breast tumour samples. The predicted performances of the models developed are obtained by recreating a complete ANN just using the identified miRNA signatures on the same population of data. The lack of a second blind microarray dataset, that would be best for validation purpose, was tackled by applying a MCCV strategy for this model interrogation procedure on the original dataset. Further validation of the signatures found were performed using qRT-PCR on the second cohort of samples.

The most advanced model of miRNA signatures predicting for ER status consisted in a pool of 6 miRNA transcripts (miR-342, miR-299, miR-217, miR-190, miR-135b, miR-218). This model, when interrogated on the same population of data and trained over 50 MCCV repeats, was able to discriminate ER positive and negative phenotypes with a median performance of

miRNAs fingerprints identified by stepwise analysis at each step for each status studied					
Step	miRNA	mRNA targets	Mean squared error	Median performance (%)	Response
ER					
1	*miR-342*	-	0.132	83.3%	(+)
2	*miR-299-3p*	-	0.087	100%	(-)
3	*miR-217*	-	0.07	100%	(+)
4	*miR-190*	-	0.06	100%	(-)
5	*miR-135b*	-	0.057	100%	(-)
6	*miR-218*	LAMB3	0.047	100%	(+)
PR					
1	*miR-520g*	-	0.186	83.3%	(-)
2	*miR-377*	-	0.129	83.3%	(+)
3	*miR-527-518a*	-	0.086	100%	(-)
4	*miR-520f-520c*	-	0.07	100%	(+)
HER2/*neu*					
1	*miR-520d*	-	0.109	100%	(+)
2	*miR-181c*	Tcl1	0.086	100%	(-)
3	*miR-302c*	Cyclin D_1	0.062	100%	(*)
4	*miR-376b*	-	0.05	100%	(+)
5	*miR-30e-3p*	Ubc9	0.047	100%	(*)

Table 3.3: Fingerprints of miRNAs used in the expression signature at each step of model development for ER status, PR status and HER2/*neu* status. A (+) in response signifies that increased miRNA expression leads to increased probability of receptor positive status; on the opposite a (-), indicates that increased miRNA expression leads to increased probability of receptor negative status; (*) means weak response, possibly interacting to modify the response of other miRNAs.

100%.

The second model developed, to predict PR status, was found to be most accurate with a set of 4 miRNAs (miR-520g, miR-377, miR-527-518a, miR-520f-520c). This model was able to discriminate the patients regarding their status for PR receptors with a median performance of 100%.

Finally, the third model developed was looking at the optimal miRNA signature that could accurately predict the HER2/*neu* status of the patients. The identified pattern of miRNA by the trained ANNs was composed by :miR-520d, miR-181c, miR-302c, miR-376b, miR-30e. This signature as well was able to identify with 100% median predictive performance of the status, over 50 submodels run. These results are sumarised in Table 3.3.

The major advantage of an additive approach like the Stepwise algo-

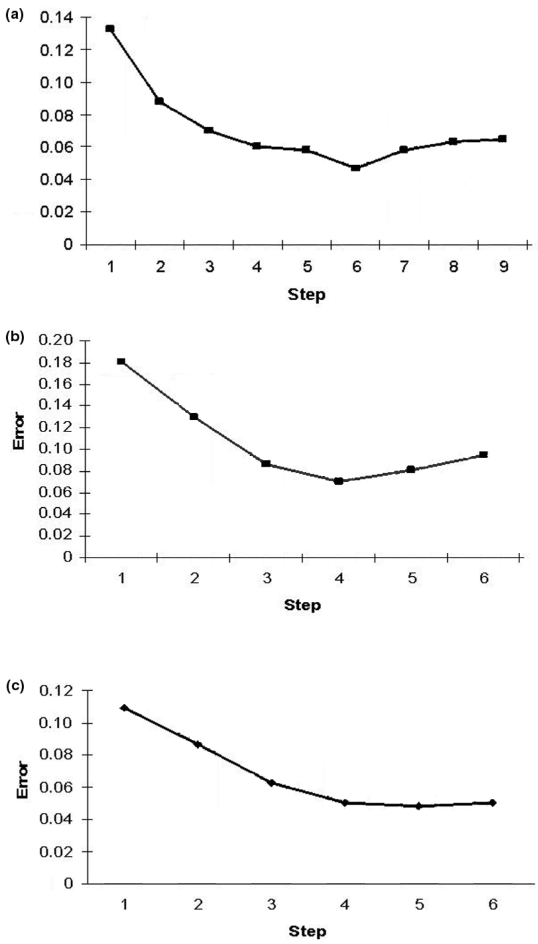


Figure 3.5: Evolution of the predictive errors at each step of development for the 3 models (a: for estrogen receptor status; b: for progesterone receptor status; c: HER2/*neu* status).

rithm presented here, consists in its potential to identify and further build and improve a molecular fingerprint. Figure 3.5 clearly shows how the predictive error drops at each step until the developed model reaches an optimal performance by iteratively adding new markers to the model. After optimal performance is reached, no further improvement is observed.

Model interrogation

The identified miRNA signatures found to have a role for each studied phenotypes were investigated for responses and population discrimination ability. In order to identify how the outcome studied is behaving regarding the level of expression for each of the miRNA included in the model, an

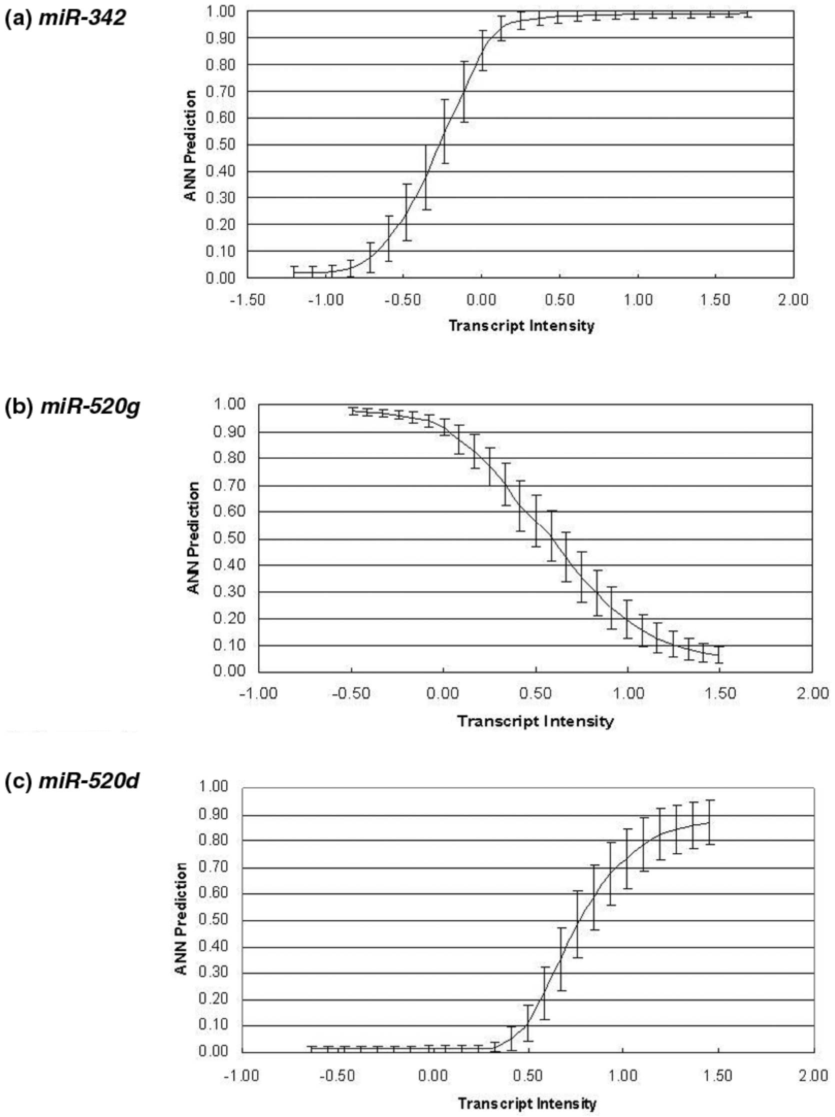


Figure 3.6: Response curves of the miRNAs identified to be the best single predictor for ER status (a); PR status (b); and HER2/*neu* status (c) (respectively miR-342, miR-520g, miR-520d).

analysis was performed using each of the identified miRNAs. The analysis consisted of presenting the values of the studied miRNAs from the signature as an input vector to the ANN that will predict the status of each individual, repeated for each of the 3 models (see Table 3.3). Figure 3.6 only displays the results of the response analysis for the best single miRNA for each of the 3 models. As a result, this figure showed that miR-342 has a positive influence to ER status, meaning that there are higher chances that an individual will present a positive status to ER when the intensity of this miRNA is found higher. Similarly, for HER2/*neu* status, miR-520d gives a positive response. Conversely, miR-520g is found to have a negative response, indicating that when its expression decreases, there are more chances for the PR status to be negative.

Further interrogation of the developed models with the identified miRNA fingerprints consisted of re-stratifying the population by predicting their status according to the output obtained with the respective models. This way of interrogating a particular model allows us to position any new individual within the population structure and visualise how this new patient fits into the overall population.

The 3 models were able to accurately classify the population accordingly to the status of the individuals as shown in Figure 3.7. The model for ER status, composed by 6 miRNA signatures, was able to perfectly distinguish positive from negative status individuals, with a very clear threshold, showing a great ability to accurately discriminate both populations. Moreover, this model is able to consistently predict the ER status as nearly no variation is observed in the 50 different submodels developed. This is shown by the error bars in the first population chart of Figure 3.7, which are obtained by the confidence intervals positive and negative for each individual over the average of the 50 submodels.

The third population chart shows again excellent discrimination ability for the model of miRNA signature at predicting HER2/*neu* status of the individuals. But this result needs to be considered carefully as only 5 out 27 patients present a positive phenotype. Therefore, although the model

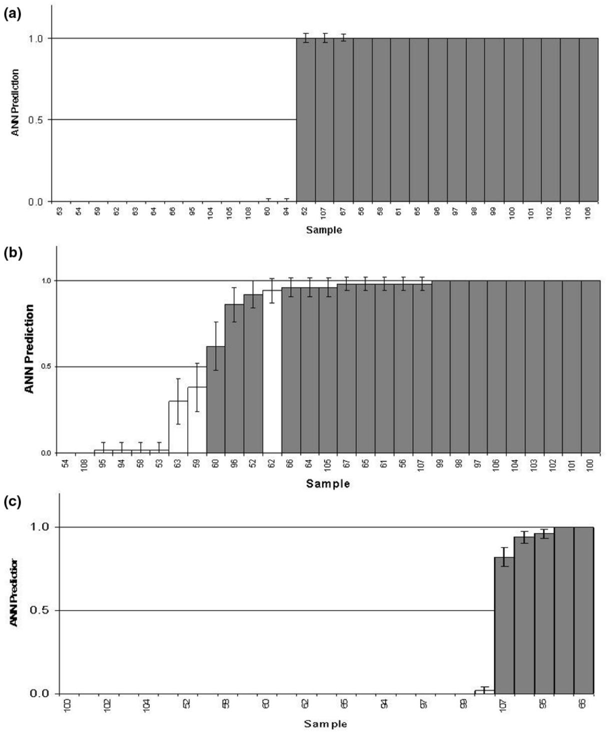


Figure 3.7: Population distibution of the model interrogation for the 3 models developed: the ER model (a); the PR model (b) and the HER2/*neu* model (c).

presents good reproducibility with low error bars, since the population is not balanced and only few individuals have a positive HER2/*neu* status, it is difficult to conclude about the absolute robustness of the model.

Validation

Validation of the identified signatures was mainly conducted with qRT-PCR by testing a subset of these identified miRNAs in the samples of the second cohort. The results can be seen in Figure 3.8.

miR-342, identified as best single predictor for ER status, was investigated by qRT-PCR, in order to identify its importance regarding the ER status of the patients. The analysis revealed that no significant difference in expression of miR-342 was observed between tumour and tumour-associated normal tissue (P value=0.6 from paired t test). However, the analysis showed that a significantly (P=0.04 from independent t test) higher expression of this miRNA was observed in ER positive tumour samples (n=62) compared to ER-negative (n=32). These results show that the expression of the miRNA is not specific to the tumour tissue, but it is significantly higher in ER-positive samples. In addition, it confirms the positive association between this miRNA and the ER positive status (*i.e.* higher likelihood of positive ER status with higher miR-342 expression level) obtained with the response curve. This is validating the results obtained by *in silico* method. Interestingly, miR-342 expression was also found higher for HER2/*neu* positive tumours (n=59) compared to negative tumours (n=32) (P=0.001 from independent t test), where the ANN analysis for HER2/*neu* status identified miR-342 at 62nd position out of 352 miRNAs studied.

Regarding PR status, miR-520g was found as best predictor using the ANN approach. The expression of this miRNA was investigated as well for validation by qRT-PCR in order to confirm the results. Similarly to the previous one, the expression of this miRNA was found to have no significant (P=0.228 from paired t test) difference between tumour and tumour-associated normal breast tissue; however, significant (P=0.032 from independent t test) increase of this expression could be observed in PR-negative tumour samples (n=33) compared to PR-positive ones (n=58), confirming here again the results obtained with the response curve analysis.

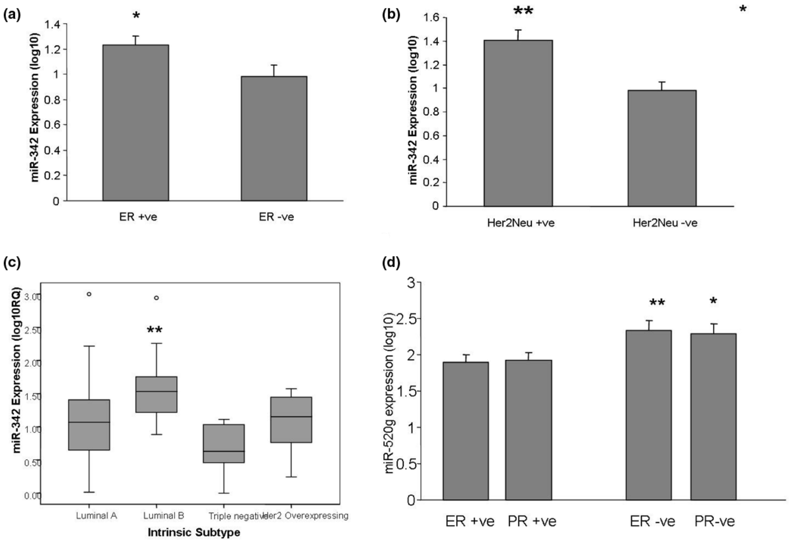


Figure 3.8: qRT-PCR Results for miR-342 and miR-520g for ER and PR status

3.2.4 Discussion and conclusion

Molecular identification and subtyping of phenotypes in cancer is showing an increasing interest for biomedical research. With this new focus, breast cancer is even more seen as a highly heterogeneous disease each subtype having its own clinical, pathological and molecular specificity. Therefore, the identification of the subtypes of the disease is critical to gain an optimised treatment for the individual and subsequently maximised prognostic outcome. Subclasses of breast carcinomas involve the status of ER, PR and HER2/*neu* (Lowery et al., 2009):

- **luminal-A subtype:** ER+ / HER2/*neu*-
- **luminal-B subtype:** ER+ / HER2/*neu*+
- **basal-like subtype:** ER- / PR- / HER2/*neu*-
- **HER2/*neu*-overexpressing subtype:** ER- / HER2/*neu*+

Not only are these receptors key factors for tumour subclassification, they also play a key role in the therapeutic strategy that will be employed. Specific drugs to treat breast cancer have an effect on one of these receptors. As an example, ER signalling can be totally inhibited with pure anti-oestrogen drugs such as fulvestrant. Moreover, the status of these receptors will have an impact on the response of the patient to a particular treatment, for example, HER2/*neu*-positive tumours will be less responsive to endocrine-based treatment (DeLaurentiis et al., 2005). The most challenging tumour subtype is the basal-like (or triple negative) as it usually presents the most aggressive behaviour that comes generally with a poorer prognostic outcome. Despite promising research of treatment for this subtype with platinum agents, -targeted agents or inhibitors, no effective specific treatment has been found for this particular subgroup. Therefore, it is clear that ER, PR and HER2/*neu* receptors are key indicators for breast cancer management although their regulation is very poorly understood so far. This is the reason why we decided to investigate potential miRNAs associated with them.

The microarray profiling have been shown to be remarkably useful for the assesment of the gene or miRNA expression of individuals. Despite the robustness of this method, it is still barely applied to routine clinical practice, mainly due to the lack of reliable computational method of the data. Classical statistical approaches have been proven problematic for such analysis (Lancashire et al., 2009a), mainly because of the highly dimensional and complex data that emerge from a microarray experiment. Moreover, several artifacts can still affect the outcome of such method (noise, experimental variability, chip and samples variability...), subsequently leading to poor performance of the models developed. A study by Michiels and colleagues (Michiels et al., 2005) demonstrated that the outcome of a particular microarray experiment is highly dependent on the computational analysis method employed and particularly dependent on the selection of the training set of patients. This is the reason why ANNs were chosen for this study. Their ability to cope with noisy, complex and highly dimensional data, coupled with a robust Cross validation procedure, to overcome the problem raised by Michiels et al., make ANNs a good choice for the data analysis. They have already been successfully applied in a number of contexts where markers of biological relevance have been identified, including polycystic ovarian syndrome (Matharoo-Ball et al., 2007b), melanoma (Matharoo-Ball et al., 2007c), prostate cancer (Matharoo-Ball et al., 2007a) and breast cancer (Lancashire et al., 2008).

The study here focused on the identification of miRNA transcript signatures predictive of ER, PR and HER2/neu status with microarray data analysed by ANN-based stepwise approach. The breast tumour samples selected for the study were early stages (stages 1 and 2a) without nodal invasion. The stepwise approach succeeded to identify optimal miRNA signatures able to accurately predict receptor statuses. Although the model performed perfectly for the dataset, it is required to perform further validation on a larger dataset, using alternative methods like PCR.

The miRNA signatures identified for ER status (miR-342, miR-299, miR-217, miR-190, miR-135b, miR-218), for PR status (miR-520g, miR-377, miR-527-518a, miR-520f-520c) and for HER2/*neu* status (miR-520d, miR-181c, miR-302c, miR-376b, miR-30e) include miRNAs that have previously been identified as dysregulated in different cancers, including breast cancer (Blenkiron et al., 2007; Mattie et al., 2006) and involved in the regulation of cell functions such as growth, apoptosis, migration and invasion (Cheng et al., 2005; Grady et al., 2008; Huang et al., 2008). This confirmed that the signatures identified are relevant in breast cancer context. In addition, confirmation by qRT-PCR of the results from microarray experiment gave a validation of the findings. A set of 8 miRNAs was validated in the same sample by stem-loop qRT-PCR, and a significant positive correlation in sample-to-sample expression was found between the two techniques (qRT-PCR and microarray). Moreover, qRT-PCR confirmed that the expression of the top ranking miRNAs in ER and PR (miR-342 and miR-520g) was related to the phenotypic characteristic of the patients in an independent set of 95 tumour samples (see Figure 3.8).

miR-520g was found important in the PR signature as well as ranked in the step 1 of the analysis for ER status. These results were validated by the results from qRT-PCR. It appears that it is the first report of implication of this miRNA dysregulation in breast cancer. However, it seems that this miR-520g is computationally related to a number of genes involved in breast cancer, such as ABCG2 (BCRDP) (Griffiths-Jones et al., 2008). ABCG2 expression in cancer cells has been shown to confer a drug-resistant phenotype and correlates with response to anthracyclines in breast cancer (Huang et al., 2008).

The validation qRT-PCR results, as well as the literature review of the miRNAs identified suggested then that these miRNAs are relevant and can be significant in the context of the phenotypes of breast cancer studied here, thus showing the robustness of the approach employed.

3.3 Summary

The Stepwise approach, despite the robustness of its original version, suffered several issues and limitations. As a result, different aspects of the method needed improvements, optimisation and tests in order to satisfy the expected robustness and accuracy at predicting models and fingerprints. Following the optimisation stage, the robustness of the redeveloped solution was illustrated by a successful study published in a peer-reviewed journal which allowed the identification of a new predictive miRNA signature for ER, PR and HER2/*neu* statuses in breast malignancies (Lowery et al., 2009) (see Appendix II).

Chapter 4

Interaction algorithm development

This chapter will now discuss the main focus of this Ph.D. research: the development and optimisation of a software solution aiming to infer networks of interaction/relationship that could exist between biomarkers using an ANN based approach. It will introduce the main idea and basic principle, and will further develop the optimisation and assessment to select the optimal algorithms and methods to extract the most accurate results.

4.1 Introduction

In the previous chapter were discussed some examples of applications of the stepwise approach to identify signatures of discrete variables (regardless of their nature) able to accurately predict a particular outcome in a disease. Although this approach has proven its robustness and reliability, it is still simplifying a complex problem to answer a particular question. In addition, although it is possible to identify a set of markers that seems to cooperatively have an influence on the context investigated, it is impossible to answer how these markers may be connected to each other, and it is impossible to explain why they seem to be involved in the same sort of interaction, complex or pathway. As discussed earlier, understanding the system of a disease, its dynamic, requires to use a more comprehensive approach and consider the global system rather than just some key features. In other words, we might have been able to identify several spots; so now, how is it possible to join the spots one to another and try to explain how they might behave between each other? We have previously seen that several groups have been trying to answer these questions using different types of approaches with different technologies. We are going to describe here a method based on ANN technology able to infer the potential co-expression, and therefore relationship, between several markers in a multivariate fashion.

4.2 Theory and development of the method

4.2.1 Theory

It has been well established and discussed that correlating and co-expressed molecules are highly likely to be involved in some sort of molecular complex or pathway of any nature, including regulatory patterns (Barabási and Oltvai, 2004; Quackenbush, 2001). Based on this statement, the primary hypothesis of the development of the algorithm is :

> *If any of the markers (e.g. genes) contained in an expression array of individuals have some influence on the expression of other markers (either positive or negative), we might be able to observe and monitor significantly correlating expression profiles between these interacting markers through the population of individuals. In other words, the influence that one input has upon the prediction of any other given input is proportional to the relationship between the two.*

Therefore, there might be other ways (in addition to investigating single markers for prediction), to investigate in depth a dataset in order to extract such information. Moreover, with the huge quantity of data contained within a single expression array (in excess of 1 million exons on a single chip), it might even be possible, with the expressions of all the interacting markers, to recreate the whole interaction map of the markers contained within the specific dataset in a particular context. We show here the development of an algorithm based on ANNs able to indicate the potential network of interactions between markers present in the dataset.

Such an algorithm could then stress the central role of key features within the particular context of a given disease, providing then not only a greater understanding of the system biology of the disease, but also new potential therapeutic targets for future treatments.

Several groups have recently started to focus some research toward this new scope as discussed in the introduction. Also, other alternatives so far have been to compile all the existing literature about systems biology and interacting molecules with several databases and softwares. This is the case with Ingenuity, which is used in this project to validate findings.

4.2.2 Algorithm

The main idea of the approach here focuses on the iterative calculation of the influence that multiple variables may have upon a single one. Indeed, if it has been possible to use all the variables to explain one particular categorical outcome (such as a survival or a status in a disease), it should be possible to use one of these variables as an outcome and all the remaining to explain the level of expression of the first one, and then be able to find out the influence everyone has on the expression level of this particular variable. Once this process for the first variable is complete, the algorithm iteratively repeats the procedure for each of the remaining variables. Briefly, it consists of looking at how the level of expression of every single input can be predicted with all the others of the original dataset, and subsequently repeat it iteratively for all of them. This algorithm is summarised in Figure 4.1, and explained by the diagram in Figure 4.2.

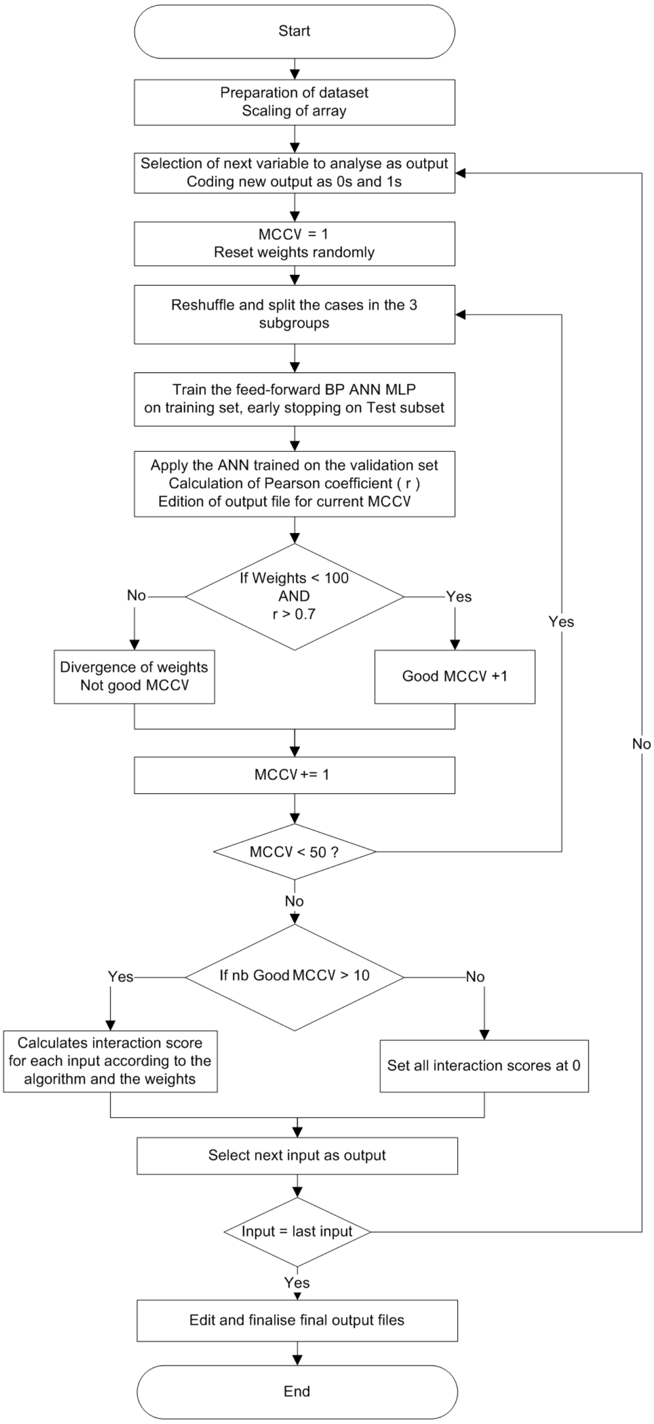
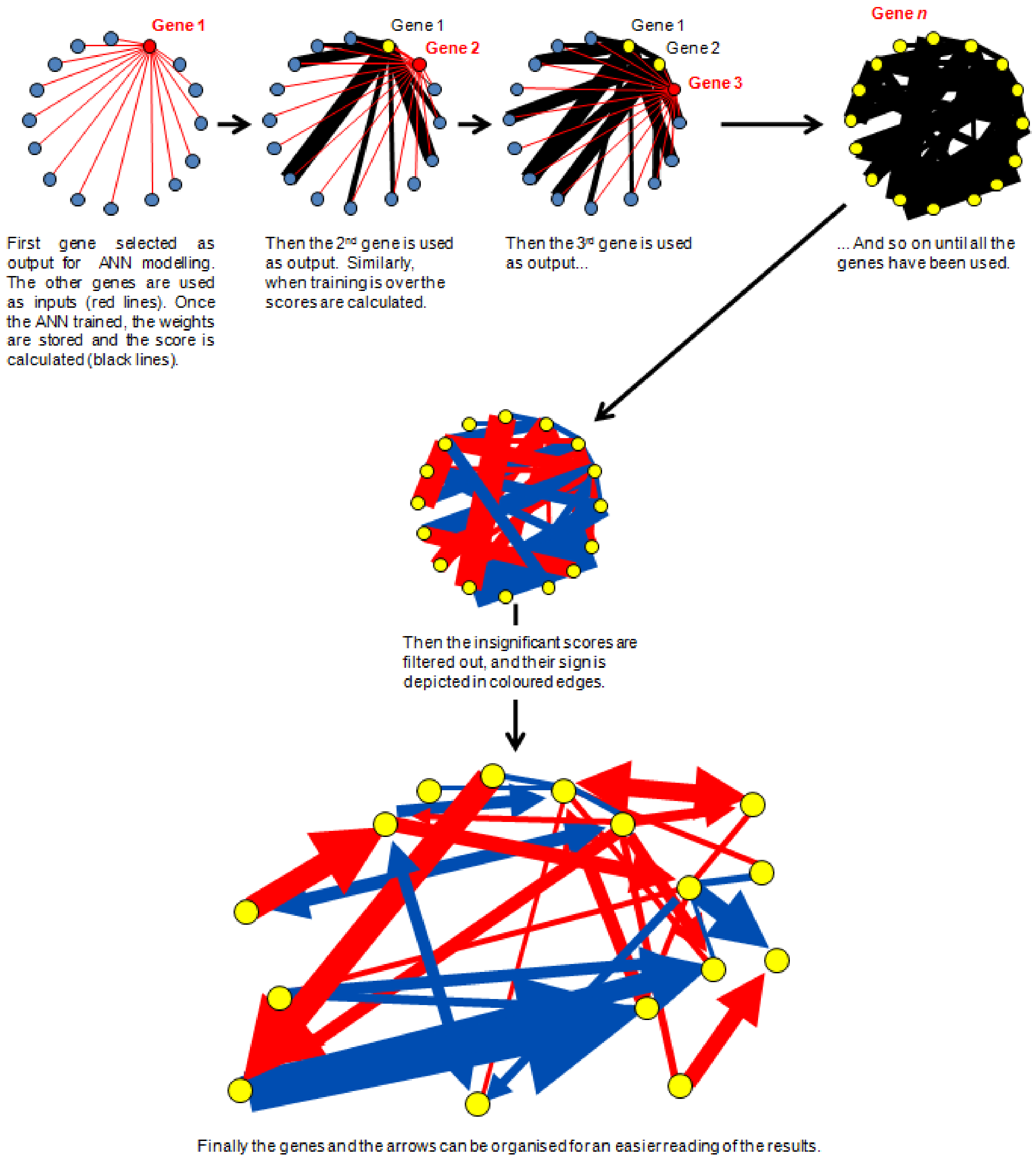

Start
Preparation of dataset
Scaling of array
Selection of next variable to analyse as output
Coding new output as 0s and 1s
MCCV = 1
Reset weights randomly
Reshuffle and split the cases in the 3 subgroups
Train the feed-forward BP ANN MLP
on training set, early stopping on Test subset
Apply the ANN trained on the validation set
Calculation of Pearson coefficient (r)
Edition of output file for current MCCV
If Weights < 100
AND
r > 0.7
No
Yes
Yes
Divergence of weights
Not good MCCV
Good MCCV +1
No
MCCV += 1
MCCV < 50 ?
No
If nb Good MCCV > 10
Yes
No
Calculates interaction score
for each input according to the
algorithm and the weights
Set all interaction scores at 0
Select next input as output
Input = last input
Yes
Edit and finalise final output files
End

Figure 4.1: Schematic representation of the simplified algorithm for interactions



First gene selected as output for ANN modelling. The other genes are used as inputs (red lines). Once the ANN trained, the weights are stored and the score is calculated (black lines).

Then the 2nd gene is used as output. Similarly, when training is over the scores are calculated.

Then the 3rd gene is used as output...

... And so on until all the genes have been used.

Then the insignificant scores are filtered out, and their sign is depicted in coloured edges.

Finally the genes and the arrows can be organised for an easier reading of the results.

Figure 4.2: Diagram of the principle of the algorithm for interactions.

At each step (*i.e.* for each variable tested), looking at the weights of the trained ANN (with the lowest predictive error) that links each input to the output gives an insight to establish a correlation between the input vector and the predicted output, and it is then possible to extract the interaction value, or **score**, with these weights. The variable considered as output is firstly converted into a boolean value in order to apply a categorical prediction of the value by the ANN. Therefore, the ANN is developed to predict if the input vector can explain an over or under-expression of the targeted variable.

An MCCV strategy is applied as well over 50 repeats with complete random re-sampling of the cases independently for each variable studied. For each of those repeats, a correlation analysis is conducted comparing the series of the actual expected values of the targeted variable (output), with its predicted values by the trained ANN, for all the cases of the test subset. From this, the Pearson correlation coefficient r (Rodgers and Nicewander, 1988) is calculated and gives a level of confidence for each of the 50 models tested at predicting the output. If the Pearson coefficient is below a threshold of 0.7 for more than 10 bootstraps out of 50, the model is no longer considered as able to accurately predict the level of expression of the variable tested, and the interaction scores targeting this particular variable will be set at 0. This method allows us to apply a sort of pre-filter in order to remove the most insignificant interactions.

To illustrate this, let's consider 10 genes identified to be of interest regarding a particular question (*e.g.* cancer against control cases). The first gene with its expression level is used as output, and all the 9 remaining genes are used as input vector. The ANN is trained with the training subset of the population, and subsequently tested against the test subset to assess how the model with the 9 genes performed to predict the level of expression of the first one. This is repeated over 50 folds, each time randomly reshuffling the cases over the subsets. If the 50 models trained have performed efficiently enough, the weights are derived according to a scoring algorithm (see page 107) to link each gene of the input vector to the output gene with a particular score (either positive or negative). In contrast, if the models did not perform satisfyingly enough, all the scores are set to 0. The method then carries on, using the second gene as output and gene 1 and 3 to 10 as inputs, and recreates and trains 50 new ANN models. The same process is repeated to obtain the scores linking the 9 genes to the second one, and so on until the 10 genes have been used as the output of the ANN.

As a result, a Microsoft Excel spreadsheet is finally produced as output containing the scores between every single input with all the others successively, with a direction of interaction given by the positivity or negativity of the scores obtained since the weights within the trained ANN can either be positive or negative. Moreover, there is an insight of bi-directivity information since a variable is tested as output and input.

Once the information regarding all the interactions is available, it is then possible to construct the network of interactions and display it in a visual way, using in-house solutions or external software such as Cytoscape (Shannon et al., 2003).

4.3 Workflow of the method

4.3.1 Workflow

The standard workflow applied for the studies conducted using this method is summarised in the Figure 4.3. It consisted of a preliminary selection of the important variables by applying a Stepwise analysis. The first step of this analysis provided a rank order of the most important variables contained in the dataset regarding a particular question (*e.g.* cancer versus normal, or responder versus non-responder to therapy). The top 100 variables, once ranked by predictive error from the lowest to the highest, are further selected. Indeed, despite the ability of the ANNs to cope with highly dimensional datasets, we need to apply here a feature selection of the most important variables for two reasons. Firstly, the dimension of the result is nearly quadratic from the initial number of variables, for n variables selected for interactions, each are tested against the remaining $n-1$, therefore the final number of interactions is $n(n-1)$ or $n^2 - n$. Therefore, reducing the number of investigated variables to the most relevant ones, will help to reduce the time of processing. Secondly, since the main core of the scoring system of the algorithm is based on the weights within the ANN, having too many inputs would literally hide the importance of a particular input amongst all the others.

4.3.2 Pre-processing step

An initial stage of the method consists in rescaling the whole array of data in order to normalise the variables. They are independently rescaled between 0 and 1 across all the cases. This allows application of a consistent and normalised set of data to the algorithm.

4.3.3 Application of the algorithm

Subsequently, the actual interaction process is carried out using the algorithm developed and discussed above as seen in Figure 4.1.

4.3.4 Filtering the interaction map

From this process an array of interaction scores between variables can be extracted as a table in a Microsoft Excel spreadsheet format. Since every single interaction between any pair of variables contained in the dataset have been investigated, it appears that some (most) of the relationships are not actual or of enough significance. Moreover, such an algorithm produces a huge amount of data, as for n inputs, there are $n(n-1)$, or $n^2 - n$, potential interactions. Therefore, it is necessary to filter out the non-significant

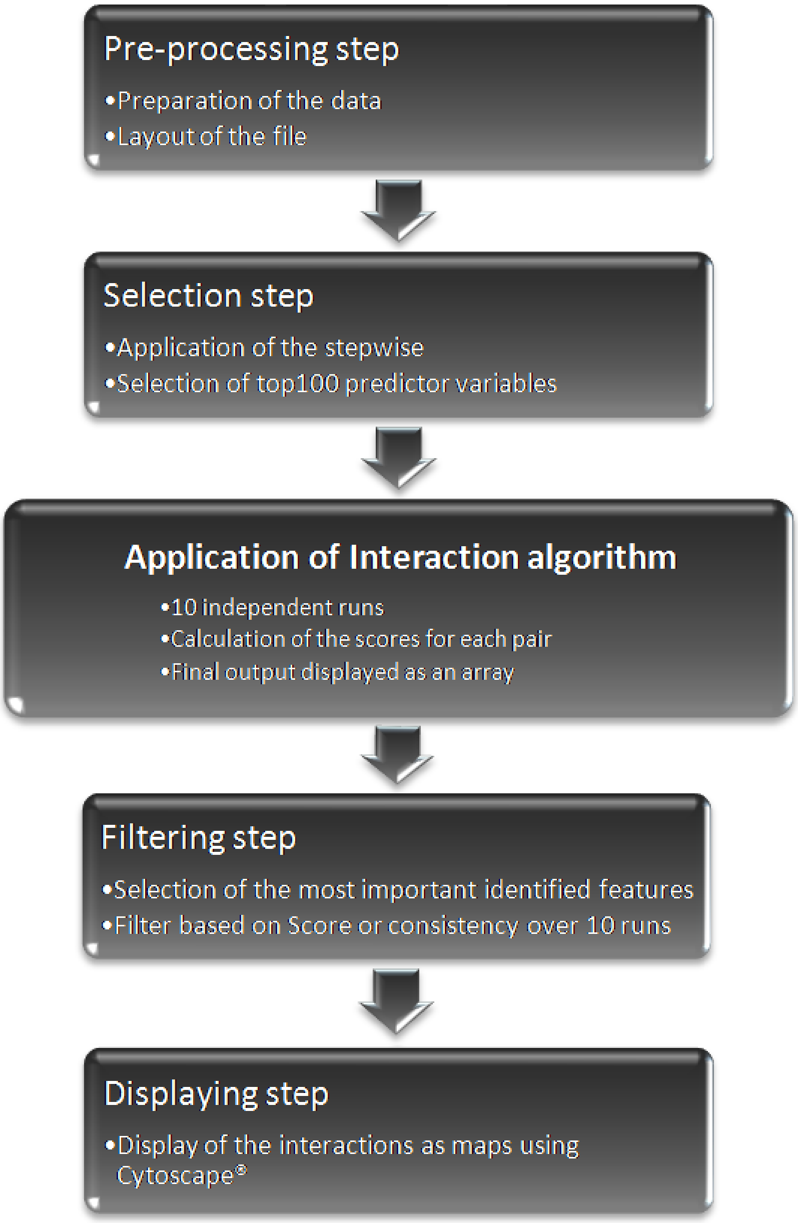

Pre-processing step
•Preparation of the data
•Layout of the file
Selection step
•Application of the stepwise
•Selection of top100 predictor variables
Application of Interaction algorithm
•10 independent runs
•Calculation of the scores for each pair
•Final output displayed as an array
Filtering step
•Selection of the most important identified features
•Filter based on Score or consistency over 10 runs
Displaying step
•Display of the interactions as maps using Cytoscape®

Figure 4.3: Workflow of the interaction algorithm application

interactions, and only keep the most interesting of them. Originally it has been decided to consider as significant the interactions with the highest absolute values of interaction score (either positive or negative). The sign of the interaction score is preserved so that it is possible to determine the direction for the mapped interaction indicating whether it is inhibitory or stimulatory. Some other ways of filtering have been considered, such as taking the average interaction values for 10 successive interaction models, and keeping only the interactions with the lowest standard error across the 10 models, or only those interactions at the tails of the distribution. Some of this filtering strategies are discussed in the following section discussing the optimisation of the approach.

4.3.5 Visualisation of Interaction maps

The concept of the interaction mapping in the present case uses the network theory as described by Barabási and Oltva (Barabási and Oltvai, 2004), where a single marker is symbolised by a node, and any relationship between two markers is represented with a directed edge, and/or an arrow setting one of the nodes as a source, and the other one as a target. In order to build up and visualise such interaction networks, the open source software Cytoscape (Shannon et al., 2003) has been used.

Although the whole concept relied on the basis of interaction map, alternative displaying method could be considered. One to be mentioned is the heatmap. Rather than presenting the results as a list of each possible pair of interactions, the results are presented as a matrix with each variable present in both column and rows. The interaction between each pair of variables is represented by a spot where the variables cross in the matrix, and the intensity of interaction score is coded by a gradual scale of blue colours for negative scores and red colours for positive scores. Scores set at 0 by the algorithm are coded in white, and the spot at the cross between the same variable is coded in black. An example of a heatmap for interaction scores is presented in Figure 4.4.

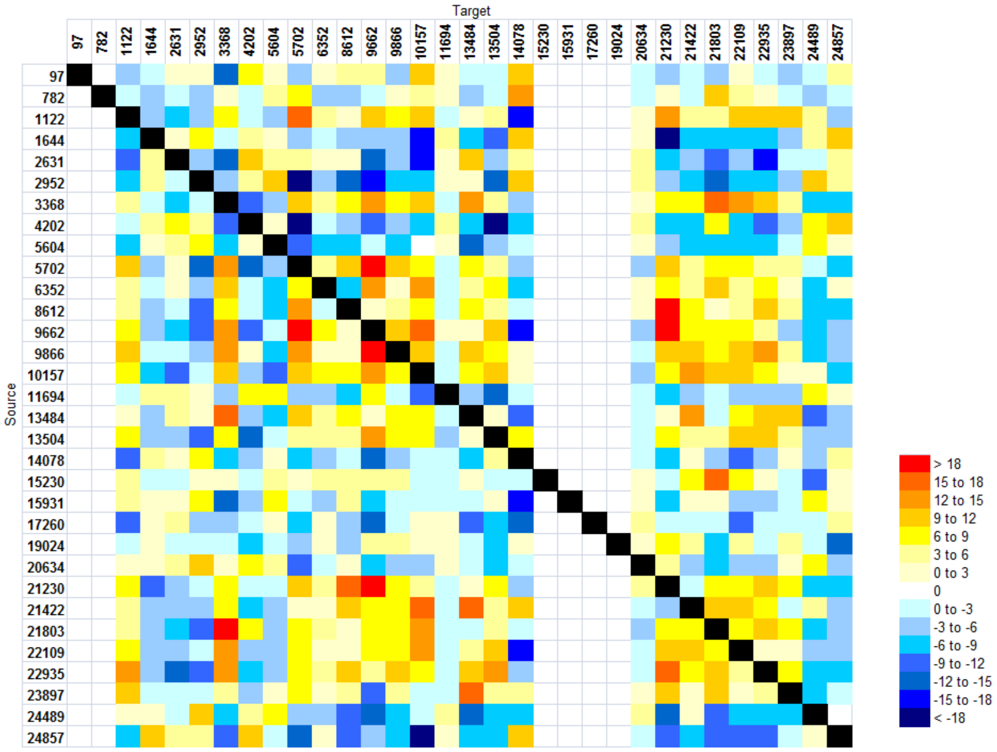


Figure 4.4: Example of a heatmap for interaction scores applied to a simulated dataset.

From this stage of development of the approach, the algorithm has been applied to the van't Veer dataset and results are described in the following chapter.

4.4 Optimisation of algorithm and method

4.4.1 Application on simulated dataset

A simulated dataset has been created and used for the purpose of further improving the algorithm. The dataset has been blindly generated using a custom R script[1] and consisted of a population of 100 cases with 25,000 variables containing 32 highly correlating variables (*c.f.* Correlation matrix in Table 4.2). The list of the 32 variables can be seen in Table 4.1. The other 24,968 variables were randomly generated by the R script.

	Position in dataset	Position in ranking at first step of Stepwise
1	12565	**328**
2	5209	79
3	16983	**9679**
4	12864	**4996**
5	14078	27
6	20634	25
7	97	30
8	15023	35
9	5604	22
10	11694	23
11	2631	24
12	24489	21
13	1644	11
14	4202	15
15	24857	9
16	2952	12
17	21230	13
18	13484	10
19	21422	16
20	8612	19
21	21803	18
22	6352	17
23	1122	20
24	23897	14
25	9866	3
26	9662	1
27	22109	4
28	13504	8
29	5702	6
30	22935	7
31	3368	5
32	10157	2

Table 4.1: List of the 32 highly-correlating variables of the simulated dataset. Positions in ranking presented in bold show the variables not present in the top 100 in step 1 of the stepwise analysis.

Results of the stepwise analysis

Compliant to the workflow of the method, a preliminary step for feature selection involved the application of the stepwise algorithm. A single

[1] R is a programming language and a mathematical environment specialised in the statistical analysis of data. It is popular amongst the bioinformatics community.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	1	0.01	0.09	0.17	0.14	0.03	0.30	0.06	0.19	0.16	0.16	0.18	0.12	0.20	0.20	0.21
2	0.01	1	-0.02	-0.06	0.11	0.20	0.14	0.16	0.18	0.21	0.25	0.19	0.18	0.22	0.32	0.31
3	0.09	-0.02	1	-0.03	0.01	0.13	-0.06	0.01	0.15	0.09	0.13	0.05	0.13	0.09	0.14	0.11
4	0.17	-0.06	-0.03	1	0.22	-0.14	0.18	0.05	0.12	0.05	0.01	-0.02	0.17	0.14	0.19	0.19
5	0.14	0.11	0.01	0.22	1	0.14	0.32	0.23	0.29	0.24	0.29	0.24	0.38	0.31	0.41	0.41
6	0.03	0.20	0.13	-0.14	0.14	1	0.11	0.27	0.40	0.31	0.39	0.42	0.37	0.47	0.34	0.50
7	0.30	0.14	-0.06	0.18	0.32	0.11	1	-0.08	0.27	0.32	0.28	0.28	0.31	0.42	0.41	0.37
8	0.06	0.16	0.01	0.05	0.23	0.27	-0.08	1	0.28	0.24	0.22	0.19	0.28	0.32	0.39	0.31
9	0.19	0.18	0.15	0.12	0.29	0.40	0.27	0.28	1	0.52	0.51	0.59	0.54	0.56	0.57	0.67
10	0.16	0.21	0.09	0.05	0.24	0.31	0.32	0.24	0.52	1	0.47	0.55	0.58	0.62	0.56	0.57
11	0.16	0.25	0.13	0.01	0.29	0.39	0.28	0.22	0.51	0.47	1	0.44	0.56	0.63	0.60	0.50
12	0.18	0.19	0.05	-0.02	0.24	0.42	0.28	0.19	0.59	0.55	0.44	1	0.59	0.63	0.61	0.69
13	0.12	0.18	0.13	0.17	0.38	0.37	0.31	0.28	0.54	0.58	0.56	0.59	1	0.69	0.75	0.69
14	0.20	0.22	0.09	0.14	0.31	0.47	0.42	0.32	0.56	0.62	0.63	0.63	0.69	1	0.76	0.70
15	0.20	0.32	0.14	0.19	0.41	0.34	0.41	0.39	0.57	0.56	0.60	0.61	0.75	0.76	1	0.72
16	0.21	0.31	0.11	0.19	0.41	0.50	0.37	0.31	0.67	0.57	0.50	0.69	0.69	0.70	0.72	1
17	-0.20	-0.27	-0.15	-0.04	-0.38	-0.33	-0.38	-0.19	-0.60	-0.60	-0.60	-0.67	-0.76	-0.71	-0.76	-0.71
18	-0.33	-0.21	-0.20	-0.19	-0.38	-0.39	-0.38	-0.36	-0.64	-0.57	-0.47	-0.66	-0.67	-0.69	-0.72	-0.67
19	-0.27	-0.29	-0.10	-0.15	-0.30	-0.45	-0.33	-0.38	-0.61	-0.59	-0.59	-0.57	-0.68	-0.72	-0.73	-0.70
20	-0.22	-0.37	-0.07	-0.13	-0.33	-0.34	-0.29	-0.32	-0.59	-0.57	-0.49	-0.62	-0.64	-0.65	-0.70	-0.71
21	-0.21	-0.24	-0.01	-0.07	-0.36	-0.45	-0.38	-0.25	-0.61	-0.53	-0.61	-0.65	-0.66	-0.61	-0.70	-0.73
22	-0.23	-0.15	0.02	-0.12	-0.41	-0.37	-0.31	-0.29	-0.67	-0.57	-0.50	-0.64	-0.62	-0.65	-0.63	-0.71
23	-0.29	-0.17	-0.11	-0.20	-0.47	-0.35	-0.40	-0.28	-0.60	-0.55	-0.63	-0.55	-0.69	-0.67	-0.72	-0.70
24	-0.16	-0.22	-0.07	-0.13	-0.33	-0.36	-0.44	-0.23	-0.56	-0.60	-0.53	-0.67	-0.66	-0.70	-0.71	-0.69
25	-0.24	-0.29	-0.13	-0.08	-0.38	-0.42	-0.43	-0.29	-0.68	-0.56	-0.62	-0.69	-0.71	-0.71	-0.77	-0.77
26	-0.31	-0.31	-0.18	-0.12	-0.44	-0.47	-0.37	-0.34	-0.63	-0.55	-0.65	-0.69	-0.73	-0.78	-0.77	-0.80
27	-0.17	-0.27	-0.17	-0.09	-0.45	-0.45	-0.35	-0.33	-0.65	-0.59	-0.63	-0.67	-0.73	-0.76	-0.79	-0.77
28	-0.18	-0.26	-0.10	-0.10	-0.34	-0.46	-0.40	-0.34	-0.63	-0.67	-0.61	-0.67	-0.74	-0.81	-0.76	-0.79
29	-0.22	-0.31	-0.09	-0.16	-0.41	-0.48	-0.43	-0.36	-0.66	-0.60	-0.58	-0.68	-0.72	-0.80	-0.80	-0.82
30	-0.20	-0.34	-0.09	-0.16	-0.40	-0.46	-0.40	-0.33	-0.66	-0.63	-0.71	-0.70	-0.75	-0.79	-0.80	-0.79
31	-0.26	-0.23	-0.13	-0.16	-0.42	-0.43	-0.48	-0.31	-0.67	-0.62	-0.66	-0.70	-0.71	-0.79	-0.81	-0.78
32	-0.29	-0.33	-0.07	-0.15	-0.38	-0.47	-0.32	-0.40	-0.73	-0.64	-0.68	-0.67	-0.77	-0.77	-0.81	-0.79

	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
1	-0.20	-0.33	-0.27	-0.22	-0.21	-0.23	-0.29	-0.16	-0.24	-0.31	-0.17	-0.18	-0.22	-0.20	-0.26	-0.29
2	-0.27	-0.21	-0.29	-0.37	-0.24	-0.15	-0.17	-0.22	-0.29	-0.31	-0.27	-0.26	-0.31	-0.34	-0.23	-0.33
3	-0.15	-0.20	-0.10	-0.07	-0.01	0.02	-0.11	-0.07	-0.13	-0.18	-0.17	-0.10	-0.09	-0.09	-0.13	-0.07
4	-0.04	-0.19	-0.15	-0.13	-0.07	-0.12	-0.20	-0.13	-0.08	-0.12	-0.09	-0.10	-0.16	-0.16	-0.16	-0.15
5	-0.38	-0.38	-0.30	-0.33	-0.36	-0.41	-0.47	-0.33	-0.38	-0.44	-0.45	-0.34	-0.41	-0.40	-0.42	-0.38
6	-0.33	-0.39	-0.45	-0.34	-0.45	-0.37	-0.35	-0.36	-0.42	-0.47	-0.45	-0.46	-0.48	-0.46	-0.43	-0.47
7	-0.38	-0.38	-0.33	-0.29	-0.38	-0.31	-0.40	-0.44	-0.43	-0.37	-0.35	-0.40	-0.43	-0.40	-0.48	-0.32
8	-0.19	-0.36	-0.38	-0.32	-0.25	-0.29	-0.28	-0.23	-0.29	-0.34	-0.33	-0.34	-0.36	-0.33	-0.31	-0.40
9	-0.60	-0.64	-0.61	-0.59	-0.61	-0.67	-0.60	-0.56	-0.68	-0.63	-0.65	-0.63	-0.66	-0.66	-0.67	-0.73
10	-0.60	-0.57	-0.59	-0.57	-0.53	-0.57	-0.55	-0.60	-0.56	-0.55	-0.59	-0.67	-0.60	-0.63	-0.62	-0.64
11	-0.60	-0.47	-0.59	-0.49	-0.61	-0.50	-0.63	-0.53	-0.62	-0.65	-0.63	-0.61	-0.58	-0.71	-0.66	-0.68
12	-0.67	-0.66	-0.57	-0.62	-0.65	-0.64	-0.55	-0.67	-0.69	-0.69	-0.67	-0.67	-0.68	-0.70	-0.70	-0.67
13	-0.76	-0.67	-0.68	-0.64	-0.66	-0.62	-0.69	-0.66	-0.71	-0.73	-0.73	-0.74	-0.72	-0.75	-0.71	-0.77
14	-0.71	-0.69	-0.72	-0.65	-0.61	-0.65	-0.67	-0.70	-0.71	-0.78	-0.76	-0.81	-0.80	-0.79	-0.79	-0.77
15	-0.76	-0.72	-0.73	-0.70	-0.70	-0.63	-0.72	-0.71	-0.77	-0.77	-0.79	-0.76	-0.80	-0.80	-0.81	-0.81
16	-0.71	-0.67	-0.70	-0.71	-0.73	-0.71	-0.70	-0.69	-0.77	-0.80	-0.77	-0.79	-0.82	-0.79	-0.78	-0.79
17	1	0.67	0.64	0.72	0.68	0.67	0.74	0.73	0.77	0.80	0.76	0.75	0.78	0.80	0.78	0.78
18	0.67	1	0.74	0.67	0.63	0.66	0.66	0.75	0.76	0.71	0.74	0.71	0.76	0.77	0.78	0.76
19	0.64	0.74	1	0.68	0.70	0.65	0.68	0.64	0.77	0.73	0.77	0.72	0.73	0.77	0.76	0.81
20	0.72	0.67	0.68	1	0.61	0.54	0.64	0.64	0.67	0.68	0.68	0.68	0.74	0.74	0.72	0.72
21	0.68	0.63	0.70	0.61	1	0.69	0.68	0.69	0.74	0.73	0.74	0.71	0.74	0.76	0.80	0.77
22	0.67	0.66	0.65	0.54	0.69	1	0.67	0.65	0.69	0.75	0.72	0.71	0.73	0.74	0.76	0.78
23	0.74	0.66	0.68	0.64	0.68	0.67	1	0.74	0.78	0.77	0.77	0.73	0.79	0.80	0.76	0.77
24	0.73	0.75	0.64	0.64	0.69	0.65	0.74	1	0.73	0.67	0.71	0.72	0.76	0.76	0.76	0.73
25	0.77	0.76	0.77	0.67	0.74	0.69	0.78	0.73	1	0.85	0.82	0.78	0.84	0.85	0.84	0.83
26	0.80	0.71	0.73	0.68	0.73	0.75	0.77	0.67	0.85	1	0.82	0.81	0.85	0.82	0.85	0.85
27	0.76	0.74	0.77	0.68	0.74	0.72	0.77	0.71	0.82	0.82	1	0.81	0.82	0.81	0.86	0.85
28	0.75	0.71	0.72	0.68	0.71	0.71	0.73	0.72	0.78	0.81	0.81	1	0.80	0.84	0.81	0.82
29	0.78	0.76	0.73	0.74	0.74	0.73	0.79	0.76	0.84	0.85	0.82	0.80	1	0.83	0.86	0.84
30	0.80	0.77	0.77	0.74	0.76	0.74	0.80	0.76	0.85	0.82	0.81	0.84	0.83	1	0.86	0.85
31	0.78	0.78	0.76	0.72	0.80	0.76	0.76	0.76	0.84	0.85	0.86	0.81	0.86	0.86	1	0.86
32	0.78	0.76	0.81	0.72	0.77	0.78	0.77	0.73	0.83	0.85	0.85	0.82	0.84	0.85	0.86	1

Table 4.2: Correlation matrix of the 32 correlating variables of the simulated dataset

step has been carried out in order to extract the rank order for the 25,000 variables regarding their predictive ability to discriminate the 2 sub-groups created in the simulated dataset. The selection of the top 100 variables from this ranking gave the dataset to further submit to the interaction algorithm (Table 4.3). The results showed that the Stepwise analysis successfully identified, among the top 100 variables, 29 out of the 32 correlating variables (hence 90.6%).

Rank order	Input name	Median Test Performance	Average Test Error	Correlating variable
1	9662	100.0%	0.052	**Yes**
2	10157	100.0%	0.062	**Yes**
3	9866	100.0%	0.065	**Yes**
4	22109	100.0%	0.066	**Yes**
5	3368	100.0%	0.071	**Yes**
6	5702	100.0%	0.072	**Yes**
7	22935	100.0%	0.073	**Yes**
8	13504	95.0%	0.087	**Yes**
9	24857	95.0%	0.090	**Yes**
10	13484	95.0%	0.091	**Yes**
11	1644	95.0%	0.095	**Yes**
12	2952	95.0%	0.095	**Yes**
13	21230	95.0%	0.100	**Yes**
14	23897	90.0%	0.103	**Yes**
15	4202	90.0%	0.104	**Yes**
16	21422	90.0%	0.107	**Yes**
17	6352	90.0%	0.115	**Yes**
18	21803	85.0%	0.119	**Yes**
19	8612	85.0%	0.127	**Yes**
20	1122	90.0%	0.130	**Yes**
21	24489	85.0%	0.139	**Yes**
22	5604	85.0%	0.142	**Yes**
23	11694	85.0%	0.146	**Yes**
24	2631	80.0%	0.158	**Yes**
25	20634	70.0%	0.208	**Yes**
26	17260	70.0%	0.210	No
27	14078	70.0%	0.219	**Yes**
28	19253	65.0%	0.225	No
29	782	65.0%	0.228	No
30	97	65.0%	0.230	**Yes**
31	15931	65.0%	0.230	No
32	11172	60.0%	0.231	No
33	10009	60.0%	0.231	No
34	3375	60.0%	0.233	No
35	15023	60.0%	0.233	**Yes**
36	15230	55.0%	0.233	No
37	4123	60.0%	0.233	No
38	19024	65.0%	0.233	No
39	356	65.0%	0.234	No
40	3675	60.0%	0.235	No
41	22571	60.0%	0.235	No
42	11469	60.0%	0.236	No
43	23974	60.0%	0.236	No
44	12494	60.0%	0.236	No
45	9061	55.0%	0.236	No
46	9261	60.0%	0.236	No
47	2612	62.5%	0.236	No
48	3220	55.0%	0.236	No
49	22878	55.0%	0.237	No
50	18812	65.0%	0.237	No
51	19440	60.0%	0.237	No
52	13098	60.0%	0.237	No
53	9640	60.0%	0.237	No
54	14904	60.0%	0.237	No
55	20764	60.0%	0.237	No
56	18111	55.0%	0.238	No
57	6212	60.0%	0.238	No
58	14360	57.5%	0.238	No
59	14362	60.0%	0.238	No
60	10155	60.0%	0.238	No
61	16789	60.0%	0.238	No
62	9278	60.0%	0.238	No
63	16564	60.0%	0.238	No
64	20584	60.0%	0.238	No
65	13915	60.0%	0.238	No
66	22846	60.0%	0.238	No
67	21921	65.0%	0.238	No
68	22731	60.0%	0.238	No
69	10230	62.5%	0.239	No
70	12805	60.0%	0.239	No
71	20729	60.0%	0.239	No
72	17574	55.0%	0.239	No
73	23203	60.0%	0.239	No
74	16235	60.0%	0.239	No
75	13462	57.5%	0.239	No
76	506	60.0%	0.239	No
77	10319	60.0%	0.239	No
78	24778	60.0%	0.240	No
79	5209	60.0%	0.240	**Yes**
80	7854	55.0%	0.240	No
81	17223	60.0%	0.240	No
82	21231	60.0%	0.240	No
83	14961	60.0%	0.240	No
84	9306	55.0%	0.240	No
85	8220	55.0%	0.240	No
86	13184	60.0%	0.240	No
87	14629	60.0%	0.240	No
88	16162	60.0%	0.240	No
89	15915	60.0%	0.240	No
90	18367	55.0%	0.240	No
91	22253	57.5%	0.241	No
92	6065	55.0%	0.241	No
93	15337	60.0%	0.241	No
94	9504	60.0%	0.241	No
95	11135	55.0%	0.241	No
96	18573	55.0%	0.241	No
97	5651	60.0%	0.241	No
98	24981	55.0%	0.241	No
99	15064	55.0%	0.242	No
100	10331	55.0%	0.242	No

Table 4.3: Results for the first step of the stepwise analysis

4.4.2 Optimisation of the algorithm

Using the artificial dataset and the selection of the top 100 best predictive variables; an experiment has been designed to further investigate potential optimised algorithm for the scoring system (*i.e.* the core of the algorithm).

Experimental design for the test of different algorithms

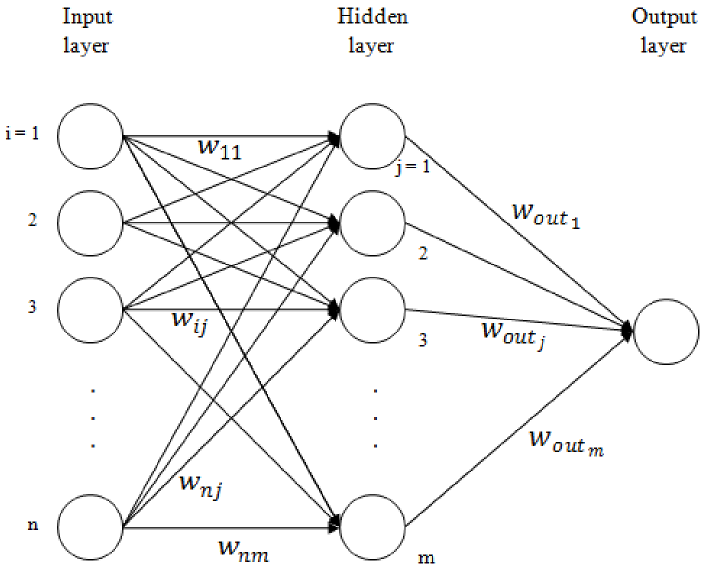


Figure 4.5: Representation of the MLP and its nomenclature used for the test of the 14 algorithms

Based on a classical MLP (*c.f.* Figure 4.5), 14 different algorithms (*c.f.* Table 4.4) were implemented and tested. All the weights connecting a single input (i) to the final output will be termed here as a *"path"*. For a single input i, the algorithms for the calculation of the score were defined as shown in Table 4.4.

For each of the 14 algorithms tested, the behaviour of the number of hidden nodes have been monitored too, applying each algorithm with alternatively 2, 5 and 10 hidden nodes. In addition, the number of selected variables from the stepwise analysis were tested: given the fact that 32 variables (see Tables 4.1 and 4.3) were supposed to be highly correlated

Algorithm	Description	Expression
Algorithm 1	Multiplied weights for sparse network of first path	$(w_{11}.w_{out_1})$
Algorithm 2	Multiplied weights for sparse network of last path	$(w_{nm}.w_{out_m})$
Algorithm 3	Multiplied weights for sparse network of one random path	$(w_{ix}.w_{out_x})$ with $x = rand(i)$
Algorithm 4	Multiplied weights for sparse network of the Max path	$Max(w_{ij}.w_{out_j})$
Algorithm 5	Sum of all the paths with multiplication of the weights	$\sum_{j=1}^{m} \left(w_{ij}.w_{out_j}\right)$
Algorithm 6	Product of all the paths with multiplication of the weights	$\prod_{j=1}^{m} \left(w_{ij}.w_{out_j}\right)$
Algorithm 7	Average value for all paths of multiplied weights	$\frac{\sum_{j=1}^{m} \left(w_{ij}.w_{out_j}\right)}{m}$
Algorithm 8	Added weights for sparse network of first path	$(w_{11} + w_{out_1})$
Algorithm 9	Added weights for sparse network of last path	$(w_{nm} + w_{out_m})$
Algorithm 10	Added weights for sparse network of one random path	$(w_{ix} + w_{out_x})$ with $x = rand(i)$
Algorithm 11	Added weights for sparse network of the Max path	$Max(w_{ij} + w_{out_j})$
Algorithm 12	Sum of all the paths with addition of the weights	$\sum_{j=1}^{m} \left(w_{ij} + w_{out_j}\right)$
Algorithm 13	Product of all the paths with addition of the weights	$\prod_{j=1}^{m} \left(w_{ij} + w_{out_j}\right)$
Algorithm 14	Average value for all paths of added weights	$\frac{\sum_{j=1}^{m} \left(w_{ij}+w_{out_j}\right)}{m}$

Table 4.4: Table of the 14 different algorithms tested

(see Table 4.2), the top 32 variables, as well as the top 100 (compliantly to our standard approach) were applied for the tests. Finally, each of the algorithms was tested over 10 independent runs to be able to monitor the consistency of the results.

For the purpose of optimisation, several results were recorded and analysed subsequently to the tests for each of the algorithm considered. These results mainly consisted of:

- The analysis of the Pearson correlation coefficient obtained by the correlation analysis between the predicted scores for each pair of the 32 variables actually correlated with their actual correlation value from the correlation matrix. This is justified as the two series compared (predicted interaction scores against actual correlation for the pairs) are linear. This coefficient was calculated for each of the algorithms and for each run, and averaged over the 10 runs.

- The signs obtained were compared in order to identify if any mismatch of signs existed between the scores obtained from the algorithms and the sign of the actual correlation value.
- A monitoring of the false positive rate has been carried out. It was obtained by taking the list of the top 10, top 50 and top 100 for positive and negative pairs of variables in terms of interaction score, and assessing out of each of these selections the number of pairs of variables where both of them were from the original pool of 32 variables correlating.
- Finally, as all the algorithms have been tested successively with 2, 5 and 10 hidden nodes in the hidden layer, a comparison of the performances for each parameterisation has been conducted.

Results

Assesment of performance with correlation coefficients

The results for the correlation values are summarised and reported in Table 4.5. It has been easy to observe that the algorithms 8 to 14 worked almost randomly to accurately predict interaction scores compared to the actual correlation values; this is clearly shown in Figures 4.6 and 4.7. This can be explained by the fact that these 7 algorithms consider only a part of the problem by just adding the weights in the paths. In addition, no significant result was obtained with the 6^{th} algorithm after the first test (not reported). The runs were consistently giving no scores as they were always diverging quickly and giving too high values. Therefore the tests on this algorithm were not further conducted.

	2 Hidden Nodes				5 Hidden Nodes				10 Hidden Nodes			
	top 32		top 100		top 32		top 100		top 32		top 100	
	Pearson's coefficient	Correct Signs	Pearson's coefficient	Correct Signs	Pearson's coefficient	Correct Signs	Pearson's coefficient	Correct Signs	Pearson's coefficient	Correct Signs	Pearson's coefficient	Correct Signs
Algorithm 1	0.788	88.16%	0.857	92.12%	0.653	80.31%	0.866	89.41%	0.617	78.17%	0.842	89.53%
Algorithm 2	0.787	88.73%	0.872	91.13%	0.619	81.31%	0.862	92.24%	0.599	77.46%	0.723	88.18%
Algorithm 3	0.797	89.59%	0.869	91.26%	0.648	81.03%	0.838	90.27%	0.590	77.75%	0.856	88.67%
Algorithm 4	0.808	91.44%	0.874	91.87%	0.675	82.31%	0.884	92.61%	0.638	80.46%	0.861	90.15%
Algorithm 5	0.805	89.16%	**0.865**	**91.26%**	0.653	80.31%	0.871	90.89%	0.607	79.32%	0.866	88.55%
Algorithm 6	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Algorithm 7	0.787	88.87%	0.865	91.50%	0.651	79.60%	0.868	89.78%	0.611	77.89%	0.823	88.67%
Algorithm 8	-0.569	33.10%	-0.287	42.24%	-0.011	53.07%	-0.022	54.43%	0.000	51.93%	-0.097	54.19%
Algorithm 9	-0.599	33.38%	-0.325	40.02%	0.056	54.21%	-0.174	53.45%	-0.051	52.07%	-0.083	52.22%
Algorithm 10	-0.534	33.38%	-0.243	44.21%	0.056	54.21%	-0.236	47.41%	-0.125	51.21%	-0.046	56.40%
Algorithm 11	-0.543	33.81%	-0.243	41.50%	-0.163	50.93%	-0.175	48.52%	-0.055	51.78%	-0.126	51.60%
Algorithm 12	-0.612	32.10%	-0.340	42.24%	-0.041	53.21%	-0.163	50.25%	-0.090	52.21%	-0.185	53.20%
Algorithm 13	-0.679	18.83%	-0.379	40.27%	0.009	51.21%	0.045	49.14%	-0.064	43.94%	0.033	51.35%
Algorithm 14	-0.623	31.95%	-0.280	41.38%	-0.094	52.07%	-0.013	50.62%	-0.066	53.07%	-0.219	52.34%

Table 4.5: Results for the test of the 14 algorithms with the Pearson's coefficient and percentage of correct signs.

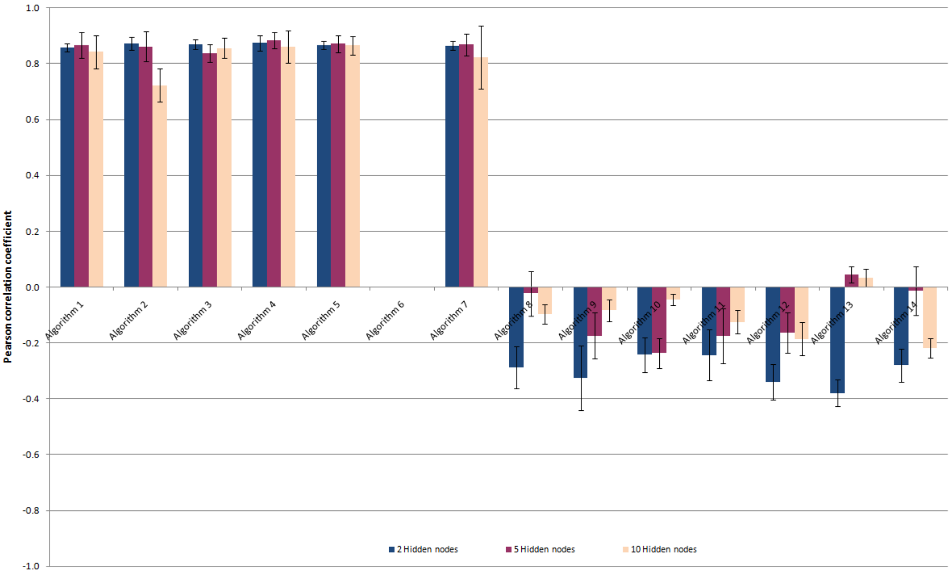


Figure 4.6: Results of Pearson correlation coefficient for each algorithm with 2, 5 and 10 hidden nodes for the selection of the top 100 variables out from stepwise analysis

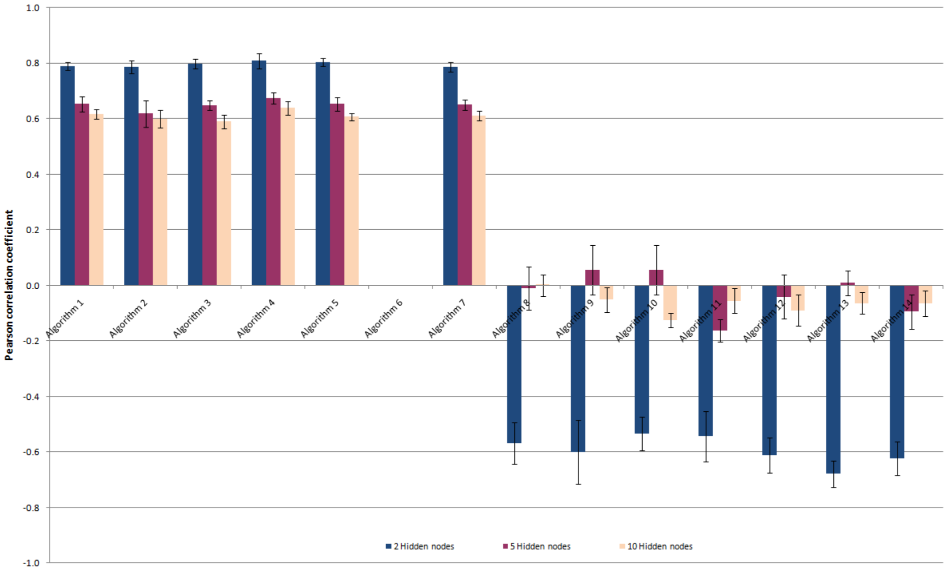


Figure 4.7: Results of Pearson correlation coefficient for each algorithm with 2, 5 and 10 hidden nodes for the selection of the top 32 variables out from stepwise analysis

Assessment of the signs

Similarly, attention has been brought to the signs obtained for the predicted interaction scores by the algorithm for each pair of variables, compared to the sign for the actual correlation coefficient for same pair. It has been carried out by taking out from the whole interaction results the pairs of variables that are both from the 32 highly correlating set. With these pairs, the number of the matching signs (between actual correlation value and predicted interaction score) was recorded in order to give a percentage of correct signs. The results are presented in Table 4.5.

Assessment of the true positive rate

The other major parameter on which it was crucial to focus on was the true positive rate. In other words : How many actually highly correlating pairs of variables can we pick within the pair of interactions predicted by the algorithm?

The results, presented in Table 4.6, show some encouraging results for the selection of 32 variables. However, this is not really relevant, there is no surprise that about 90% of the actual correlations are picked since the top 32 selection contained 27 (84.38%) of the correlating variables. It is then more relevant to focus on the top 100 variables selected. And the results for the selection of 100 variables showed poor rate for true positives with about 35% to 50%. In addition, we can see that adding more hidden nodes to the MLP doesn't improve the interaction results in terms of true positive rate. All these results suggest that the true positive rate is not really satisfactory. However, this is not exactly due to the algorithm itself or scoring system, but more a question of filtering. Therefore, improvements could be brought by adapting the filtering method.

			Algo 1	Algo 2	Algo 3	Algo 4	Algo 5	Algo 6	Algo 7	Algo 8	Algo 9	Algo 10	Algo 11	Algo 12	Algo 13	Algo 14
2 Hidden Nodes	top 32	top10 +	90%	100%	100%	90%	100%	n/a	100%	100%	100%	100%	100%	100%	100%	100%
		top50 +	90%	88%	88%	86%	88%	n/a	88%	100%	100%	100%	100%	100%	94%	100%
		top100 +	87%	87%	86%	89%	88%	n/a	88%	99%	100%	99%	100%	100%	91%	100%
		top10 -	90%	90%	90%	100%	100%	n/a	90%	90%	100%	100%	100%	100%	100%	100%
		top50 -	96%	96%	98%	96%	96%	n/a	96%	90%	94%	92%	94%	90%	90%	90%
		top100 -	97%	96%	96%	95%	98%	n/a	96%	89%	92%	87%	90%	89%	87%	87%
	top 100	top10 +	20%	40%	20%	20%	10%	n/a	30%	100%	80%	100%	90%	90%	30%	100%
		top50 +	36%	38%	32%	34%	34%	n/a	36%	94%	90%	90%	76%	94%	16%	94%
		top100 +	33%	35%	29%	35%	33%	n/a	40%	93%	86%	88%	72%	93%	12%	92%
		top10 -	30%	10%	40%	60%	40%	n/a	60%	60%	30%	30%	70%	80%	0%	70%
		top50 -	46%	26%	40%	60%	42%	n/a	50%	54%	46%	34%	48%	50%	16%	60%
		top100 -	44%	29%	44%	56%	50%	n/a	46%	42%	50%	39%	42%	44%	12%	51%
5 Hidden nodes	top 32	top10 +	70%	60%	50%	60%	40%	n/a	70%	40%	90%	90%	60%	40%	80%	50%
		top50 +	80%	86%	78%	70%	74%	n/a	80%	74%	82%	82%	78%	74%	78%	78%
		top100 +	83%	83%	83%	81%	79%	n/a	83%	79%	78%	78%	83%	79%	85%	85%
		top10 -	90%	90%	90%	80%	80%	n/a	80%	60%	80%	80%	30%	80%	100%	40%
		top50 -	94%	92%	88%	92%	88%	n/a	94%	78%	74%	74%	60%	84%	88%	70%
		top100 -	92%	95%	85%	89%	88%	n/a	92%	81%	74%	74%	37%	83%	85%	80%
	top 100	top10 +	0%	0%	0%	0%	0%	n/a	0%	0%	0%	10%	0%	20%	0%	30%
		top50 +	0%	0%	0%	0%	0%	n/a	0%	0%	6%	6%	0%	16%	2%	30%
		top100 +	0%	0%	0%	0%	0%	n/a	0%	0%	10%	8%	0%	17%	7%	28%
		top10 -	0%	0%	0%	0%	0%	n/a	0%	0%	0%	0%	0%	30%	10%	0%
		top50 -	0%	0%	0%	0%	0%	n/a	0%	0%	2%	0%	0%	12%	20%	0%
		top100 -	0%	0%	4%	2%	0%	n/a	0%	1%	2%	0%	0%	9%	21%	0%
10 Hidden nodes	top 32	top10 +	40%	50%	50%	60%	60%	n/a	60%	70%	70%	70%	70%	40%	90%	40%
		top50 +	68%	76%	78%	82%	74%	n/a	78%	80%	88%	78%	76%	74%	88%	76%
		top100 +	75%	79%	77%	84%	72%	n/a	81%	82%	90%	78%	83%	80%	86%	78%
		top10 -	80%	80%	70%	90%	80%	n/a	90%	60%	80%	80%	80%	70%	90%	80%
		top50 -	90%	86%	86%	90%	92%	n/a	92%	70%	60%	72%	62%	82%	88%	76%
		top100 -	90%	83%	82%	90%	88%	n/a	92%	72%	69%	72%	31%	56%	88%	50%
	top 100	top10 +	0%	10%	0%	0%	0%	n/a	0%	0%	0%	0%	0%	0%	0%	0%
		top50 +	0%	4%	0%	0%	0%	n/a	0%	0%	0%	0%	0%	0%	4%	0%
		top100 +	0%	4%	0%	1%	0%	n/a	2%	0%	0%	1%	0%	0%	6%	1%
		top10 -	0%	0%	0%	0%	0%	n/a	0%	0%	0%	0%	0%	0%	10%	0%
		top50 -	0%	4%	0%	0%	0%	n/a	0%	0%	0%	0%	0%	10%	2%	0%
		top100 -	0%	2%	0%	1%	0%	n/a	1%	0%	2%	1%	2%	7%	8%	0%

Table 4.6: Results for the true positive rate for the 14 algorithms.

Number of hidden nodes

It was then interesting to assess as well if altering the number of hidden nodes had any effect on the results. For that purpose, the tests were conducted as described previously and iteratively and independently repeated with 2, 5 and 10 hidden nodes in the single hidden layer of the MLP.

The results showed that the number of hidden nodes does not make much of difference in terms of correlation results or in terms of signs (see Table 4.5). Moreover, it appears that 2 hidden nodes are doing better for the selection of the top 32 variables than for the top 100 from the stepwise analysis. However, this can be easily explained by the inherent bias involved in the top 100 variables. Indeed, the Tables 4.1 and 4.3 show that 27 out of 32 (84.38%) are from the actual pool of 32 variables correlating originally in the dataset; whereas it automatically drops with the selection of the top 100 since 29 out of the 100 (29%) are from this pool of variables. Therefore, it appears normal that the importance of the number of hidden nodes has less influence for the selection of the top 100 variables.

In addition, the analysis of the results concerning the time of process indicated that it is taking about 2 and then 3 times more time to run the process on average for all the 14 algorithms, for 5 and 10 hidden nodes respectively compared to running the algorithm with 2 hidden nodes (shown in Figure 4.8). Since no actual improvement could be observed with 5 and 10 hidden nodes in terms of Pearson coefficient study and signs, and considering the processing time, it seems to be sufficient to use a two hidden node-hidden layer for the interaction analysis purpose.

4.4.3 Optimisation of the filtering method

After completing the tests of the algorithms, the screening of the true positive rate indicated that the filtering method, based on the absolute value of the interaction scores (*c.f.* page 98), was inefficient (*i.e.* gave low true

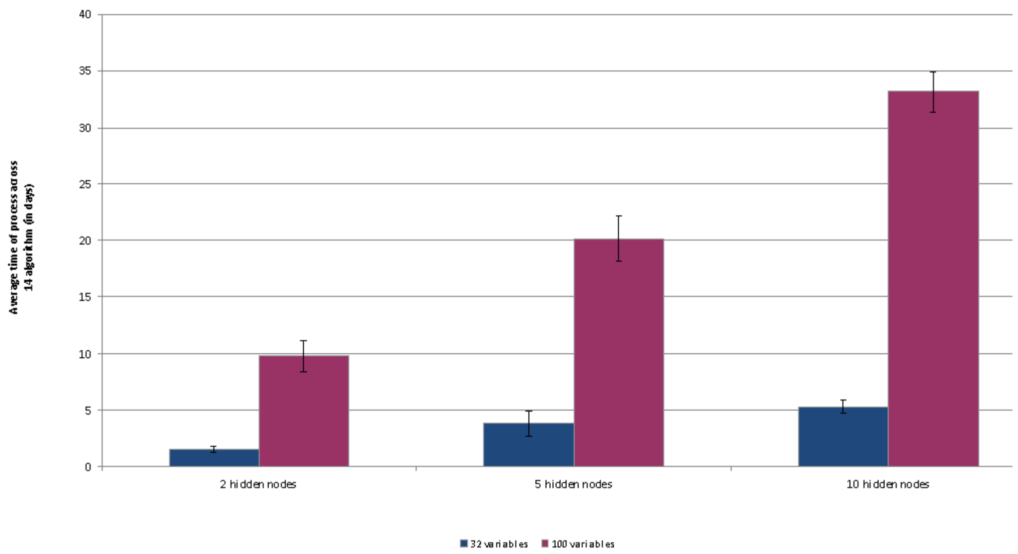


Figure 4.8: Average time spent by each algorithm for 10 runs, with 2, 5 and 10 hidden nodes (on Intel Core2 duo E840 machines at 3.00 GHz with 2Gb memory running under Microsoft Windows XP SP3 system).

positive rate values for top 100+/- across all algorithms - *c.f.* results in Table 4.6), and needed to be further optimised. Therefore, since 10 runs of each test had been carried out, the consistency of the scores obtained were used for filtering rather than using the only scores themselves. This consisted of taking the Coefficient of Variation C_v, a relative measure of the dispersion for a particular distribution (Hendricks and Robey, 1936), where the Standard Deviation σ of the interaction scores found for each pair of variables is divided by the mean value of the score μ for that particular pair, in order to avoid its influence upon the deviation. Thus, the expression for the new filtering value is :

$$C_v = \frac{\sigma}{\mu} \tag{4.1}$$

From the first results of the correlation coefficient, only the 5 first algorithms were further tested with this new filter method (results shown in Table 4.7), and only the selection of the top 100 variables from the stepwise was applied. From these results, it clearly appears that the new filtering method brings a much more reliable strategy at deciphering the significant interactions, especially shown by Figures 4.9 for positive scores and 4.10 for negative scores. This is particularly true when the number of hidden nodes is increased: with the new filtering, the true positive rates with 5 or 10 hidden nodes can reach the rates obtained with 2 hidden nodes.

			Algo 1	Algo 2	Algo 3	Algo 4	Algo 5
2 Hidden nodes	top 32	top10 +	90% (-)	90% (-10)	90% (-10)	90% (-)	90% (-10)
		top50 +	92% (+2)	92% (+4)	88% (-)	92% (+6)	88% (-)
		top100 +	92% (+5)	91% (+4)	88% (+2)	92% (+3)	89% (+1)
		top10 -	70% (-20)	90% (-)	90% (-)	100% (-)	100% (-)
		top50 -	88% (-8)	90% (-6)	90% (-8)	92% (-4)	96% (-)
		top100 -	90% (-7)	91% (-5)	90% (-6)	92% (-3)	96% (-2)
	top 100	top10 +	50% (+30)	60% (+20)	20% (-)	10% (-10)	90% (+80)
		top50 +	56% (+20)	68% (+30)	56% (+24)	44% (+10)	74% (+40)
		top100 +	58% (+25)	64% (+29)	52% (+23)	53% (+18)	69% (+36)
		top10 -	90% (+60)	30% (+20)	70% (+30)	50% (-10)	60% (+20)
		top50 -	64% (+18)	50% (+24)	64% (+24)	68% (+8)	66% (+24)
		top100 -	59% (+15)	49% (+20)	59% (+15)	69% (+13)	63% (+13)
5 Hidden nodes	top 32	top10 +	90% (+20)	90% (+30)	100% (+50)	90% (+30)	70% (+30)
		top50 +	86% (+6)	88% (+2)	90% (+12)	92% (+22)	86% (+12)
		top100 +	87% (+4)	89% (+6)	86% (+3)	89% (+8)	86% (+7)
		top10 -	80% (-10)	100% (+10)	80% (-10)	100% (+20)	100% (+20)
		top50 -	92% (-2)	92% (-)	86% (-2)	96% (+4)	98% (+10)
		top100 -	93% (+1)	87% (-8)	92% (+7)	92% (+3)	94% (+6)
	top 100	top10 +	50% (+50)	70% (+70)	50% (+50)	80% (+80)	80% (+80)
		top50 +	62% (+62)	60% (+60)	60% (+60)	56% (+56)	58% (+58)
		top100 +	54% (+54)	54% (+54)	59% (+59)	52% (+52)	61% (+61)
		top10 -	100% (+100)	70% (+70)	90% (+90)	80% (+80)	80% (+80)
		top50 -	74% (+74)	68% (+68)	88% (+88)	62% (+62)	70% (+70)
		top100 -	63% (+63)	61% (+61)	72% (+68)	60% (+58)	62% (+62)
10 Hidden nodes	top 32	top10 +	70% (+30)	90% (+40)	90% (+40)	90% (+30)	100% (+40)
		top50 +	86% (+18)	84% (+8)	88% (+10)	92% (+10)	94% (+20)
		top100 +	86% (+11)	84% (+5)	89% (+12)	89% (+5)	86% (+14)
		top10 -	100% (+20)	90% (+10)	100% (+30)	100% (+10)	100% (+20)
		top50 -	92% (+2)	90% (+4)	96% (+10)	96% (+6)	88% (-4)
		top100 -	92% (+2)	89% (+6)	91% (+9)	93% (+3)	88% (-)
	top 100	top10 +	60% (+60)	70% (+60)	60% (+60)	80% (+80)	100% (+100)
		top50 +	56% (+56)	62% (+58)	62% (+62)	72% (+72)	76% (+76)
		top100 +	56% (+56)	56% (+52)	56% (+56)	72% (+71)	70% (+70)
		top10 -	90% (+90)	100% (+100)	100% (+100)	90% (+90)	90% (+90)
		top50 -	78% (+78)	80% (+76)	82% (+82)	94% (+94)	94% (+94)
		top100 -	76% (+76)	73% (+71)	70% (+70)	90% (+89)	87% (+87)

Table 4.7: Results for the test of the 5 first algorithms for true positive rate using the coefficient of variation for filtering. In brackets are shown the positive or negative evolution compared to the previous filtering method.

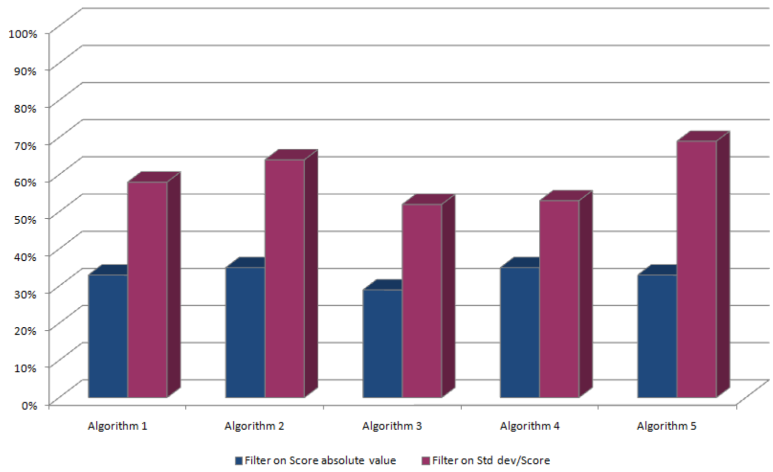


Figure 4.9: Comparison of the improvements of true positive rate for the selection of the 100 pairs of variables with the highest positive interaction score for algorithms 1 to 5.

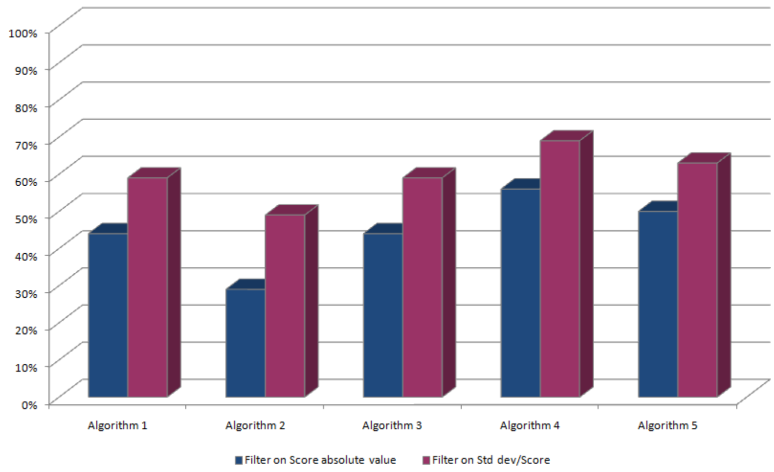


Figure 4.10: Comparison of the improvements of true positive rate for the selection of the 100 pairs of variables with the highest negative interaction score for algorithms 1 to 5.

4.5 Discussion

The increasing number of genomic array studies has allowed the potential to assess the relative expression intensity of gene products in the particular context of a disease system, hence allowing the evaluation of the potential regulatory relationships between genes. Since ANN techniques have shown great abilities at predicting categorical outcomes in disease (*c.f.* chapter 1 and 2), it was considered to develop a novel approach wrapping an ANN core with a different shell.

This chapter described the development of an original algorithm seeking to predict and infer interaction networks, or regulatory networks in the case of genomic data. The chapter discussed the principles and hypothesis on which the overall process of development was founded; and it discussed the methodologies employed for the optimisation of the approach. The algorithm of the approach itself and the workflow of its application has been as well introduced. As a result, following the theoretical stage of the method developed, it is possible to discuss its application for real-world data in the following chapter. The first development of the approach was applied to the van't Veer dataset discussed in the following chapter (chapter 5). Subsequently, it has been assessed and further improved using a simulated dataset, and reapplied to the same dataset, presented in the next chapter as well.

The optimisation consisted of testing 14 different algorithms (*c.f.* Table 4.4) in order to improve the scoring system. Another aspect tested and improved was the filtering step. Therefore, this chapter described how the simulated dataset was used to test the different algorithms. Each of these 14 algorithms were tested at trying to identify the most correlated pairs of variables within the dataset. Several aspects were taken into account to decide which of the algorithm was performing the best at predicting the actual correlated pairs: the Pearson's coefficient depicting the correlation obtained between the scores predicted and the actual correlation values for the correlated pairs of variables; the percentage of signs of correlation predicted; as well as the true positive rate (*i.e.* the percentage of actually correlated pairs in the highest scores of predicted pairs by the algorithms).

The first observation showed that the 5 first algorithms were predicting accurately the correlation values of the pairs actually correlated, with a range of very close Pearson values from 0.857 (for algorithm 1) to 0.874 (algorithm 4) for 2 hidden nodes and the top 100 selection of variables from the stepwise analysis (used for feature selection). The other algorithms (6 to 14) were giving poor results and were not further tested. The tests showed as well that the number of hidden nodes was not affecting the predictive ability of the approach as similar range of Pearson values were observed (0.838 to 0.884 for 5 hidden nodes; and 0.723 to 0.866 with 10 hidden nodes). As a results, it was decided to focus the work with 2 hidden nodes as the performance was not affected, whilst it was dramatically increasing the time of computing (by a factor 2 and then 3 for respectively 5 and 10 hidden nodes - *c.f.* Figure 4.8).

Another great advantage of an ANN based algorithm is that the scoring system can interpret the resulting weights in a positive or negative manner. Therefore, not only it is possible to attribute a weighted score for each pair of variables, but as well a sign; hence giving a clue of whether the interaction is positive or negative. This particular aspect was assessed as well during optimisation by comparing the sign of the predicted interaction score with the sign of the actual correlation value in the original simulated dataset. Here again the algorithms 1 to 5 provided accurate close results with a range of correct signs from 91.13% (algorithm 2) to 92.12% (algorithm 1) for top 100 feature selection and 2 hidden nodes. Similar observation was made as previously regarding the number of hidden nodes.

Another point explored during the optimisation process was the true positive rate. For that purpose a different approach toward filtering process has been explored. Rather than filtering the relevant interacting pairs on their absolute values of scores, as it was the case at the early stage of the development, it has been decided to filter them on their consistency over multiple runs. As a result, the filtering consisted of calculating the coefficient of variation for each interacting pair, and only keep the pairs with the lowest absolute value (for both positive and negative scores). The coefficient of variation gives an insight about the consistency of the scores (as it uses the standard deviation) but without the influence of the score. The results shown in Table 4.7 and in Figures 4.9 and 4.10 show an improvement in terms of true positive rate with this new filter value. Indeed, the results show, with 2 hidden nodes and for the 100 features selected, an improvement of 18 (algorithm 4 with 53% true positive rate) to 36 (algorithm 5 with 69% true positive rate) points in the 100 highest positive interaction scores. Similarly, for the 100 highest negative interaction scores, algorithm 4 presented a 69% true positive rate (improving by 13 points the true positive rate of the previous filtering method) and algorithm 5 presented 63% true positive rate (showing the same improvement).

The simulated dataset allowed to drastically improve the approach, in terms of accuracy of prediction for the scores but as well to improve the filtering step by employing a more relevant approach in order to maximise the true positive rate. Nevertheless, the use of such an artificial dataset is not free of pitfalls, as it has been pointed in (Rocke et al., 2009). Using a simulated dataset reduces the complexity and does not take into account the highly interconnected mechanisms taking place in genuine data from gene expression arrays. It only simplified the problem in a pairwise fashion. However, improving and assessing the performance of an approach on a simplified example gives an insight about the abilities of the method on less complex data, allowing to further build a robust approach once it is able to accurately model such data. Once the best performing algorithm identified, applying a real-world data will then allow a second level of validation with the biological relevance of the findings.

4.6 Conclusion

Considering all these results, the scoring algorithm number 5 has been selected, as follows :

$$\sum_{j=1}^{m} \left(w_{ij} . w_{out_j} \right) \tag{4.2}$$

This algorithm presented one of the best performances on the simulated dataset for true positive rate (*i.e.* 63% and 69% respectively for top 100 negative and positive interactions), but also good results in the correlation study of the predicted scores against the actual correlation value of similar pairs (0.865 for Pearson's correlation coefficient). Combined with a more adapted filter of interaction scores, based on the coefficient of variation, they were implemented in the software and the workflow of the method. Subsequently, the dataset from van't Veer has been re-applied in order to get more accurate and up-to-date results, and can be seen in the following chapter.

Chapter 5

Application of the interaction algorithm

Within this chapter will be discussed applications of the interaction algorithm presented in chapter 4. These applications involved the use of the dataset from the van't Veer study (van't Veer et al., 2002). The chapter will report the applications of the approach before and after its optimisation, it will review some of the validations carried out on the results. In particular, section 5.1 will discuss the results from using the originally developed algorithm; subsequently, section 5.2 will discuss the results obtained from using the same dataset

5.1 Application of the original algorithm to the van't Veer dataset

The first application of the originally developed algorithm concerned the dataset published by van't Veer (van't Veer et al., 2002). This particular study from the Dutch group focused on the metastatic risk for patients with breast cancer. This dataset contained 78 samples from breast cancer patients with different metastatic outcome for the prognostic prediction. Out of these 78 patients, 34 of them developed distant metastases within 5 years (which constituted the poor prognostic group) and 44 remained metastasis-free during at least 5 years (the good prognostic group). cRNA from tumour samples were extracted and two hybridisation per sample had been conducted on MA chips screening for 24,481 genes. This dataset had been previously analysed for biomarkers related to metastatic risk, as shown in chapter 2 of the present manuscript (Lancashire et al., 2009a). It was then interesting to further investigate and try to understand how the identified genes could be related to each other, in any sort of association or interaction.

5.1.1 Pre-processing step using the stepwise algorithm

The first step of the method consisted of screening of the rank order for the genes found to be associated with metastatic risk in breast cancer in terms of predictive ability for each of them. For this purpose, the stepwise approach had been applied using a classical MLP with one hidden layer containing 2 hidden nodes. An extensive cross validation strategy had been used, randomly re-splitting the cases over 50 times in the 3 different subsets for training, validation and independent test purpose with respectively 60%, 20% and 20% of the cases in each. Learning rate and momentum were respectively set at 0.1 and 0.5. Training was performed with a BP algorithm for 3,000 epochs, and an early stopping threshold was set at 1,000 epochs if no improvement of the model was observed when applied on the validation subset. The Stepwise approach was applied here for just the first step in order to identify the most relevant genes for this particular set regarding the question of distant metastasis outcome, and give a ranking of those genes. The interaction algorithm could then provide with an insight of the potential correlation and/or interaction that may exist between those genes.

Figure 5.1 shows the distribution of the genes ranked by their predictive error for unseen data after the stepwise analysis. From the shape of the graph, it is clear that the very first genes have a higher ability to predict (with the lowest error) metastatic risks for the patients, and therefore are most relevant for the system being investigated by interaction analysis.

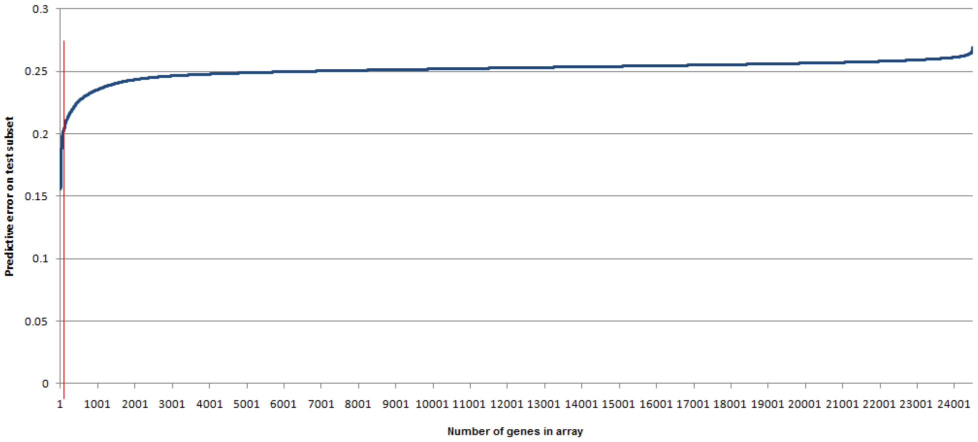


Figure 5.1: Distribution of the genes for their predictive error for unseen data. Red line indicates the position of the 100^{th} gene.

According to what is seen in Figure 5.1, and compliant to the workflow, the first 100 genes were selected for further interaction analysis. To assess the interactions within the genes associated with prognostic outcome we selected the top 100 predictors from the screening and submitted them to the interaction analysis. These genes had a range of predictive ability for unseen data from 73.86% for the first gene to 60.79% for the 100^{th}, and are presented in Table 5.1.

Importantly, we can observe from the results that CA9, previously discussed in this manuscript, was found again as the best predicting gene for metastatic risk. Although this is not so surprising since the very same method was used for the same dataset, it is still a validation of the developed method showing a good reproducibility and consistency.

Rank order	Input number	GenBank accession nb.	Test Perf.	Test Error	Gene name	Reference	Rank order	Input number	GenBank accession nb.	Test Perf.	Test Error	Gene name	Reference
1	2025	NM_001216	71.87%	0.438	CA9	(Span et al., 2003)	51	2207	AA425275	65.32%	0.471	COMMD6	
2	118	AI742029	72.79%	0.444	TBC1D10A		52	2242	NM_002266	61.71%	0.471	KPNA2	(Dankof et al., 2007)
3	1514	NM_016598	67.58%	0.448	ZDHHC3		53	1228	NM_006763	65.71%	0.472	BTG2	(Boiko et al., 2006)
4	1336	AA830802	69.76%	0.448	AGTPBP1		54	677	H17914	65.46%	0.472	ZNF420	(Tian et al., 2009)
5	275	AI458063	71.08%	0.451	ZBTB45		55	674	AJ011306	65.12%	0.472	EIF2B4	
6	1185	AI347425	73.86%	0.455	MOG		56	985	NM_005744	65.39%	0.472	ARIH1	
7	692	NM_006117	71.37%	0.455	PECI		57	1819	AW162411	66.38%	0.472	TMEM218	
8	1409	AA534406	71.59%	0.456	GALNT5		58	45	NM_001667	65.32%	0.473	ARL2	(Beghin et al., 2008)
9	642	NM_020678	69.93%	0.456	LRTM1	(Egland et al., 2006)	59	1066	D86957	64.39%	0.473	SEPT8	
10	1113	NM_016017	63.93%	0.456	UCHL5	(Wicks et al., 2005)	60	974	NM_006461	62.37%	0.473	SPAG5	
11	1443	NM_016448	68.37%	0.459	DTL	(Ueki et al., 2008)	61	1827	AI183575	66.41%	0.473	DNAJC5G	
12	1460	AA973313	67.81%	0.460	TSPYL5	(Kim et al., 2010)	62	1774	NM_018313	65.45%	0.473	PBRM1	(Horikawaa and Barrett, 2002)
13	170	AI885466	59.75%	0.461	C13orf37		63	450	AF052162	65.75%	0.474	LPCAT1	
14	1053	NM_006544	66.28%	0.461	EXOC5		64	1143	NM_006625	62.23%	0.474	SFRS13A	
15	909	AL080059	71.22%	0.461	TSPYL5	(Kim et al., 2010)	65	1640	NM_018131	63.36%	0.474	CEP55	(Inoda et al., 2009)
16	1081	NM_006581	67.00%	0.463	FUT9	(Bogoevska et al., 2006)	66	950	NM_006407	63.71%	0.474	ARL6IP5	
17	505	NM_003766	67.15%	0.463	BECN1	(Won et al., 2010)	67	718	NM_004702	62.39%	0.474	CCNE2	(Payton et al., 2002)
18	1855	AA524093	68.96%	0.464	FBXO41		68	25	AL035297	62.24%	0.474	SFT2D2	
19	497	NM_003748	70.22%	0.464	ALDH4A1	(Yoon et al., 2004)	69	1918	BE739817_RC	62.88%	0.474	IFNAR1	
20	706	NM_013438	66.63%	0.465	UBQLN1		70	1474	AL137615	67.02%	0.475	MKNK2	(Chrestensen et al., 2007)
21	322	NM_003504	62.92%	0.465	CDC45L	(Takahashi et al., 2008)	71	493	NM_005196	63.07%	0.475	CENPF	(O'Brien et al., 2007)
22	1684	AF161451	70.23%	0.465	NSMCE1		72	1351	AI992158	66.61%	0.475	CDCA7	
23	1283	AI14038	68.71%	0.465	FOXL2		73	1642	NM_018136	62.71%	0.475	ASPM	
24	1724	NM_018265	66.16%	0.465	C1orf106		74	2034	NM_000507	62.77%	0.475	FBP1	
25	283	NM_003450	68.85%	0.466	ZNF174		75	1510	AA759198	65.90%	0.476	-	
26	936	NM_014400	67.08%	0.466	LYPD3	(Fletcher et al., 2003)	76	1912	AF073519	67.17%	0.476	SERF1A	
27	143	NM_003258	63.88%	0.466	TK1	(O'Neill et al., 1992)	77	1446	NM_016458	67.19%	0.476	C8orf30A	
28	86	NM_003163	65.96%	0.466	STX1B		78	1809	NM_018391	63.93%	0.476	FLJ11328	
29	2075	NM_000599	69.04%	0.466	IGFBP5	(Ahn et al., 2010)	79	258	R73468	66.54%	0.476	SLC25A16	
30	1078	AI918032	68.08%	0.466	RUNDC1		80	1812	NM_000286	66.77%	0.476	PEX12	
31	736	AI479831	64.95%	0.466	MTHFR		81	188	AF016903	63.30%	0.477	AGRN	
32	1153	NM_014675	66.35%	0.467	CROCC		82	2115	NM_002066	64.66%	0.477	GML	
33	568	NM_004583	68.06%	0.467	RAB5C		83	248	AW137071	62.42%	0.477	SMARCE1	(García-Pedrero et al., 2006)
34	1167	NM_006681	69.19%	0.467	NMU		84	1384	NM_014968	66.17%	0.477	KIAA1104	
35	1675	AI583960	68.92%	0.467	DHX58		85	1142	NM_006623	61.65%	0.477	PHGDH	
36	663	NM_003977	69.55%	0.467	AIP		86	1082	NM_006582	64.84%	0.477	GMEB1	
37	2152	AI738508	61.64%	0.467	ECT2		87	366	AL133447	65.93%	0.477	EDC3	
38	278	NM_004163	66.46%	0.467	RAB27B	(Hendrix et al., 2010)	88	1811	NM_000284	63.88%	0.477	PDHA1	
39	1961	NM_000419	67.28%	0.468	ITGA2B		89	578	AF131828	66.46%	0.477	C9orf25	
40	32	D42044	66.11%	0.468	KIAA0090		90	1226	AW162331	66.45%	0.477	ALKBH7	
41	1236	AI554061	64.09%	0.468	QSOX2		91	98	AI393558	66.88%	0.478	CLSTN2	
42	980	NM_014489	71.12%	0.468	PGAP2		92	1791	NM_018354	62.22%	0.478	C20orf46	
43	138	NM_003239	66.63%	0.469	TGFB3	(Laverty et al., 2009)	93	307	AI912791	67.26%	0.478	FBXO16	
44	1096	AI653719	64.77%	0.470	TTLL7		94	616	NM_013360	63.36%	0.478	ZNF222	
45	738	AF257175	69.66%	0.470	PECI		95	1942	AI082587	66.92%	0.478	IL17RA	
46	1442	NM_016444	64.37%	0.470	ZNF226		96	1465	AA834945	65.52%	0.478	LIN9	(Pandis et al., 1995)
47	572	NM_003862	63.12%	0.470	FGF18		97	1186	AI741080	67.14%	0.478	B3GALNT2	
48	1832	AL122101	62.45%	0.470	TEF		98	428	NM_002916	60.79%	0.478	RFC4	
49	102	NM_001756	68.45%	0.471	SERPINA6		99	742	AI828164	65.25%	0.478	FAM69B	
50	1363	NM_016359	62.41%	0.471	NUSAP1	(Wadia et al., 2010)	100	1814	NM_000291	62.70%	0.478	PGK1	(Zhang et al., 2005)

Table 5.1: Rank order of the first 100 genes identified by the first step of the stepwise sorted by test error with gene names when available. The references indicate previous evidence of involvement of the gene in breast cancer in the literature.

5.1.2 Results

Global population

Following this initial selection step, the global population (*i.e.* both poor and good prognostic groups combined) had been firstly applied and ran through the interaction algorithm. The dataset was built using the selected 100 genes for all the 78 cases. Once the format of the dataset was appropriately defined, it had been applied to the interaction algorithm. After running 10 runs of the original interaction algorithm, the resulting matrix has been created in a Microsoft Excel format with each output as target and each input for each output as source for the interaction pair, hence with $N = n(n-1)$ (N=9900 with n=100) interaction pairs. For each of these pairs, the interaction score was averaged over the 10 runs, and its consistency monitored with standard deviation and confidence interval. These gene-paired interactions, were displayed using Cytoscape (Figure 5.2). From this interaction map, it clearly appears that no relevant information can be elucidated due to its high dimensionality and complexity, and as a consequence, stresses the need for a supplementary filtering step.

Figures 5.3 and 5.4 display the distribution of the ranked interaction values. Clearly, it is possible to observe that the majority of the interaction scores have very low absolute values, and only few scores have higher absolute value, hence following a normal distribution with very small variance (*cf.* Figure 5.3 and Figure 5.4). Overall the values range from -10.9 to +17.2. These may be of biological significance upon further investigation.

During the process of the interaction analysis, the overall predictive performance of the models used to define the interaction map had been constantly monitored by determining the Pearson's correlation coefficient between actual and predicted level of expression of the targeted gene used as output. Therefore, it was possible to constantly assess the degree of error of the model, and get a level of confidence for the interactions identified. In the present application, we have been able to get an averaged r^2 value of 0.665.

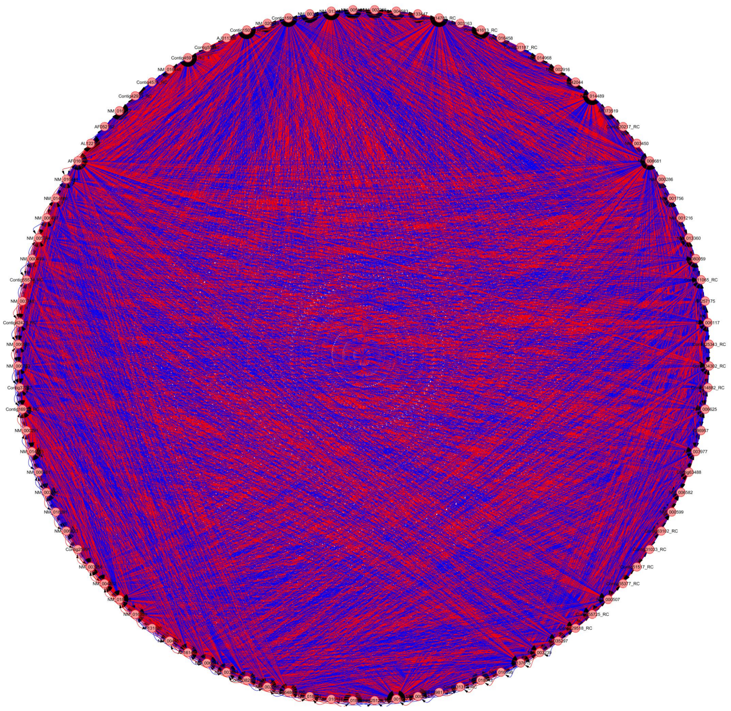


Figure 5.2: Complete display of the whole interactome obtained from the interaction algorithm for the top 100 genes before any filter applied.

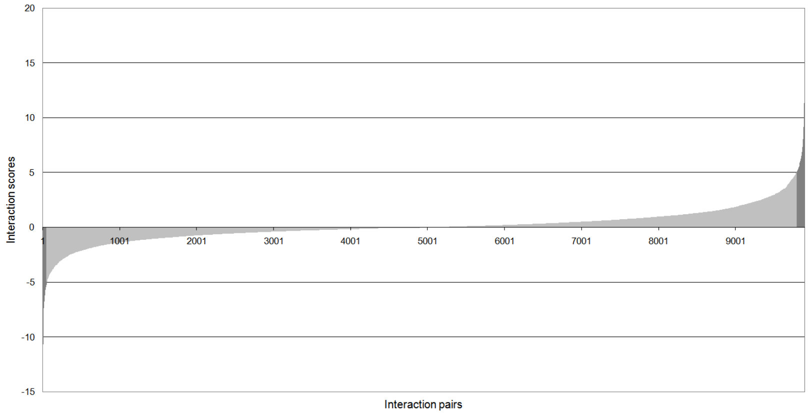


Figure 5.3: Distribution of the interactions sorted by interaction scores. The interactions in dark grey at the extremities represent the strongest interactions at level 5 (*i.e.* with scores above 5 and below -5).

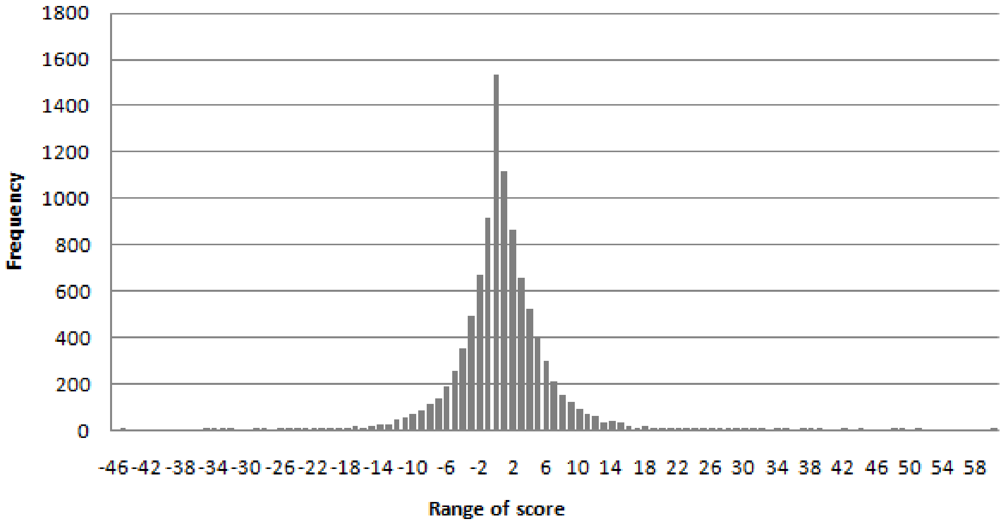


Figure 5.4: Distribution of frequencies of interaction scores.

Considering the difficulty of interpretation of the results emerging from the interaction analysis, mainly due to high dimension of the matrix, it was decided to apply successive filters in order to remove the scores with the lowest absolute value (Figures 5.5 and 5.6). This decision was based on the hypothesis that low interaction scores might reveal less likely interactions between the genes. When a filter removing links of absolute interaction score between -5 and +5 was applied (level 5 filtering), the number of links drastically decreased from 10,000 to 163 (1.63% of the original interactome). This greatly simplified the interaction map, facilitating interpretation and understanding of the key features within the global interaction map (obtained with Cytoscape) as shown in Figure 5.7 where red edges indicate a positive interaction (supposedly stimulation), and blue a negative one (supposedly inhibition).

From this map, we could then identify the highest positive interaction involving NM_016448 (Retinoic Acid-regulated nuclear Matrix-associated Protein (RAMP) - also known as Denticleless homolog (Drosophila) (DTL)) which has already been described to have a role in cancer cell proliferation (Crowe et al., 2003), and Contig42933_RC, which corresponds to the Genbank accession number R73468, referring to an unknown gene sequence which has already been strongly associated with metastasis in breast cancer

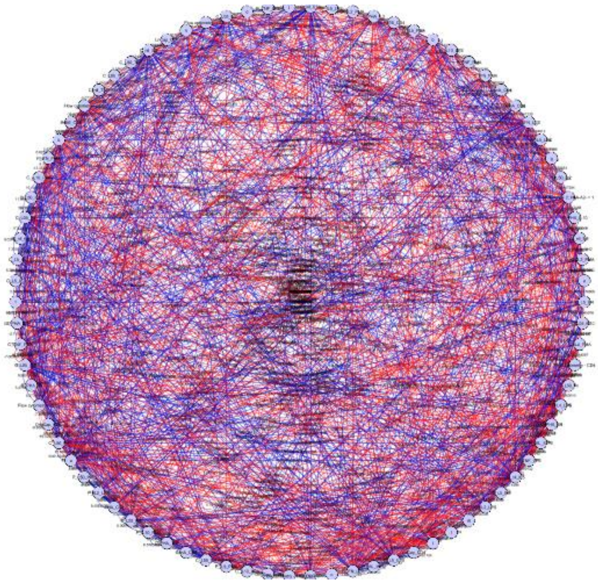


Figure 5.5: Display of the interaction map reduced to level 2 (*i.e.* with scores above 2 and below -2).

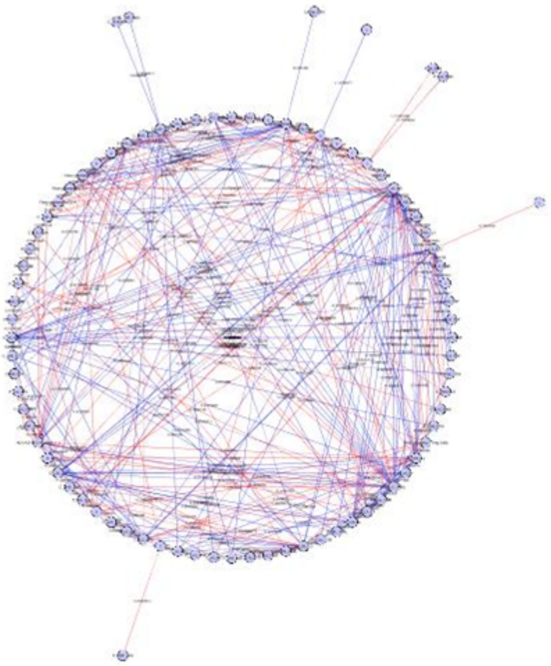


Figure 5.6: Display of the interaction map reduced to level 4 (*i.e.* with scores above 4 and below -4).

(Takahashi et al., 2004). Conversely in the negative way, we identified the strongest interaction between NM_016448 again with Contig55725_RC, which corresponds to the Cell Division Cycle Associated 7 (CDCA7) which has been shown to be over expressed in human tumours (Osthus et al.,

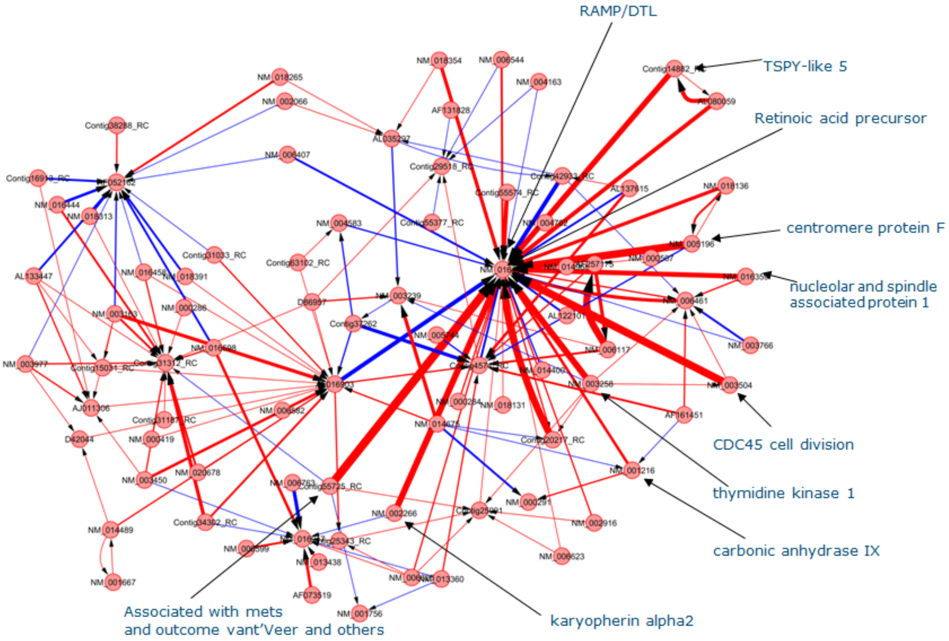


Figure 5.7: Display of the interaction map reduced to level 5 (*i.e.* with scores above 5 and below -5).

2005). Moreover, we can clearly identify some key nodes involved in dense regions of interactions, here again NM_016448 (DTL) appears as a key feature within the whole pathway of interaction, interacting strongly (within the interactions over 5) with 35 other genes, where most of these interactions (33) have NM_016448 as the target.

According to these preliminary observations, it appears that DTL has a key role in the whole interactome in this breast cancer study. Indeed, the literature has demonstrated its implication in breast cancer (Crowe et al., 2003) with the ability to decrease the expansion of cancer cells. Moreover, the mRNA sequence for CA9, previously identified as the best predictor for metastasis in that case, belongs to one of the features interacting with the Retinoic acid-regulated nuclear matrix-associated protein (NM_016448). In addition, the results show an interaction between CA9 and the mRNA for Phosphoglycerate Kinase 1 (PGK1), interactions between both have already been described in literature in relation to cancer (Winter et al., 2007).

However, all these results are only coming from a very early step of the development process for this new method, and clearly needs to be validated. As a first validation step, it was decided to analyse the literature involving the 100 genes. This work was facilitated by the database from Ingenuity (Ingenuity® Systems, www.ingenuity.com), which compiles all the published work on identified interactions between any sort of molecule or drug.

5.1.3 Validation using Ingenuity Pathway Analysis

The submission of the 100 selected input values in the Ingenuity Pathway Analysis (IPA) (Ingenuity® Systems, www.ingenuity.com) database gave 62 matches (*c.f.* Table 5.2). Once these matching genes were isolated, they were submitted to the pathway database of IPA® and a map of interactions had been designed (*c.f.* Figure 5.8) using the Path Explorer tool with the whole set of 62 genes as Set A and Set B and exploring any direction.

This analysis tool employed in IPA® in order to identify the potential relations existing between the 62 molecules introduced several others as intermediates so that not only direct relations could be added to the pathway, but indirect ones as well, with one degree of freedom (*i.e.* meaning that only one intermediate candidate was added between 2 molecules). Therefore the analysis introduced an extra set of 79 molecules (Table 5.3) to the 62 originally identified as matching between the top 100 selected with the Ingenuity databas entries (Table 5.2). As a consequence, the comparison of the maps obtained became a little less meaningful, since not all the 100 genes were present, but even more extra intermediate molecules were added. However, it is still possible to discuss the indirect relations identified by IPA® between two of the 62 molecules with their intermediates. They can still hold relevant information toward the validation of the findings from the interaction algorithm.

	Nb.	Accession nb.	Symbol	Name/Description
1	2025	NM_001216	CA9	Carbonic Anhydrase IX
2	1514	NM_016598	ZDHHC3	Zinc finger, DHHC-type containing 3
3	642	NM_020678	LRTM1	Leucine-rich repeats and transmembrane domains 1
4	1443	NM_016448	DTL	Denticleless homolog (drosophilia)
5	1053	NM_006544	EXOC5	Exocyst complex component 5
6	909	AL080059	TSPYL5	TSPY-like 5
7	1081	NM_006581	FUT9	Fucosyltransferase 9 (α-(1,3)-fucosyltransferase)
8	505	NM_003766	BECN1	Beclin 1 (autophagy related)
9	497	NM_003748	ALDH4A1	Aldehyde dehydrogenase 4 family, member A1
10	706	NM_013438	UBQLN1	Ubiquilin 1
11	322	NM_003504	CDC45L	CDC45 cell division cycle 45-like (S. cerevisiae)
12	1684	AF161451	NSMCE1	Non-SMC element 1 homolog (S. cerevisiae)
13	1724	NM_018265	C1orf106	Chromosome 1 open reading frame 106
14	283	NM_003450	ZNF174	Zinc finger protein 174
15	936	NM_014400	LYPD3	LY6/PLAUR domain containing 3
16	143	NM_003258	TK1	Thymidine kinase 1, soluble
17	2075	NM_000599	IGFBP5	Insulin-like growth factor binding protein 5
18	1153	NM_014675	CROCC	Ciliary rootlet coiled-coil, rootletin
19	568	NM_004583	RAB5C	RAB5C, member RAS oncogene family
20	1167	NM_006681	NMU	Neuromedin U
21	663	NM_003977	AIP	Aryl hydrocarbon receptor interacting protein
22	278	NM_004163	RAB27B	RAB27B, member RAS oncogene family
23	1961	NM_000419	ITGA2B	Integrin, alpha 2b
24	32	D42044	KIAA0090	KIAA0090 protein
25	980	NM_014489	PGAP2	Post-GPI attachment to proteins 2
26	138	NM_003239	TGFB3	Transforming growth factor, beta 3
27	738	AF257175	PECI	Peroxisomal D3,D2-enoyl-CoA isomerase
28	572	NM_003862	FGF18	Fibroblast growth factor 18
29	1832	AL122101	TEF	Thyrotrophic embryonic factor
30	102	NM_001756	SERPINA6	Serpin peptidase inhibitor, clade A(α-1 antiproteinase, antitrypsin), member 6
31	1363	NM_016359	NUSAP1	Nucleolar and spindle associated protein 1
32	2242	NM_002266	KPNA2	Karyopherin α 2 (RAG cohort 1, importin alpha 1)
33	1228	NM_006763	BTG2	BTG family, member 2
34	674	AJ011306	EIF2B4	Eukaryotic translation initiation factor 2B, subunit 4 δ
35	985	NM_005744	ARIH1	Ariadne (Drosophila) homolog, ubiquitin-conjugating enzyme E2-binding protein, 1
36	45	NM_001667	ARL2	ADP-ribosylation factor-like 2
37	1066	D86957	SEPT8	Septin 8
38	974	NM_006461	SPAG5	Sperm associated antigen 5
39	1774	NM_018313	PBRM1	Polybromo 1
40	450	AF052162	LPCAT1	Lysophosphatidylcholine acyltransferase 1
41	1143	NM_006625	SFRS13A	Splicing factor, arginine/serine-rich 13A
42	1640	NM_018131	CEP55	Centrosomal protein, 55kDa
43	950	NM_006407	ARL6IP5	ADP-ribosylation-like factor 6 interacting protein 5
44	25	AL035297	SFT2D2	SFT2 domain containing 2
45	1918	BE739817_RC	IFNAR1	Interferon (α, β and ω) receptor 1
46	1474	AL137615	MKNK2	MAP kinase interacting serine/threonine kinase 2
47	1642	NM_018136	ASPM	asp (abnormal spindle) homolog (Drosophila), microcephaly associated
48	2034	NM_000507	FBP1	Fructose-1,6-bisphosphatase 1
49	1912	AF073519	SERF1A	Small EDRK-rich factor 1A (telomeric)
50	1446	NM_016458	C8orf30A	Chromosome 8 open reading frame 30A
51	1812	NM_000286	PEX12	Peroxisomal biogenesis factor 12
52	188	AF016903	AGRN	Agrin
53	2115	NM_002066	GML	Glycosylphosphatidylinositol (GPI) anchored molecule like protein
54	1142	NM_006623	PHGDH	Phosphoglycerate dehydrogenase
55	1082	NM_006582	GMEB1	Glucocorticoid modulatory element binding protein 1
56	366	AL133447	EDC3	Enhancer of mRNA decapping 3 homolog (S. cerevisiae)
57	1811	NM_000284	PDHA1	Pyruvate dehydrogenase (lipoamide) α 1
58	578	AF131828	C9orf25	Chromosome 9 open reading frame 25
59	1791	NM_018354	C20orf46	chromosome 20 open reading frame 46
60	616	NM_013360	ZNF222	Zinc finger protein 222
61	428	NM_002916	RFC4	Replication factor C (activator 1) 4, 37kDa
62	1814	NM_000291	PGK1	Phosphoglycerate kinase 1

Table 5.2: 62 matching genes found by Ingenuity Pathway Analysis database entries ranked by predictive ability in step 1 of stepwise analysis

	Symbol	Name/Description	Family
1	AHR	aryl hydrocarbon receptor	ligand-dependent nuclear receptor
2	ARF6	ADP-ribosylation factor 6	transporter
3	ARNT	aryl hydrocarbon receptor nuclear translocator	transcription regulator
4	ATF4	activating transcription factor 4 (tax-responsive enhancer element B67)	transcription regulator
5	ATN1	atrophin 1	transcription regulator
6	ATXN3	ataxin 3	other
7	BRCA1	breast cancer 1, early onset	transcription regulator
8	Ca	Calcium	chemical drug
9	CD47	CD47 molecule	other
10	CDC2	cell division cycle 2, G1 to S and G2 to M	kinase
11	CDKN1A	cyclin-dependent kinase inhibitor 1A (p21, Cip1)	other
12	CDKN2A	cyclin-dependent kinase inhibitor 2A (melanoma, p16, inhibits CDK4)	transcription regulator
13	CDT1	chromatin licensing and DNA replication factor 1	other
14	CREB1	cAMP responsive element binding protein 1	transcription regulator
15	CREBBP	CREB binding protein	transcription regulator
16	CREM	cAMP responsive element modulator	transcription regulator
17	CSNK1A1	casein kinase 1, alpha 1	kinase
18	CTNNB1	catenin (cadherin-associated protein), β 1, 88kDa	transcription regulator
19	DLG4	discs, large homolog 4 (Drosophila)	kinase
20	E2F1	E2F transcription factor 1	transcription regulator
21	E2F4	E2F transcription factor 4, p107/p130-binding	transcription regulator
22	EP300	E1A binding protein p300	transcription regulator
23	EPAS1	endothelial PAS domain protein 1	transcription regulator
24	EPB41	erythrocyte membrane protein band 4.1 (elliptocytosis 1, RH-linked)	other
25	EWSR1	Ewing sarcoma breakpoint region 1	other
26	FGFR3	fibroblast growth factor receptor 3	kinase
27	FN1	fibronectin 1	enzyme
28	FXR2	fragile X mental retardation, autosomal homolog 2	other
29	GATA1	GATA binding protein 1 (globin transcription factor 1)	transcription regulator
30	GFI1B	growth factor independent 1B transcription repressor	transcription regulator
31	GRB2	growth factor receptor-bound protein 2	other
32	HGS	hepatocyte growth factor-regulated tyrosine kinase substrate	other
33	HIF1A	hypoxia inducible factor 1, α subunit	transcription regulator
34	Histone h3		group
35	HNF1A	HNF1 homeobox A	transcription regulator
36	HNF4A	hepatocyte nuclear factor 4, alpha	transcription regulator
37	HSP90AB1	heat shock protein 90kDa alpha (cytosolic), class B member 1	other
38	Importin α		group
39	KCNMA1	potassium large conductance calcium-activated channel, subfamily M, α 1	ion channel
40	KPNB1	karyopherin (importin) beta 1	transporter
41	MAPK1	mitogen-activated protein kinase 1	kinase
42	MAPK13	mitogen-activated protein kinase 13	kinase
43	MAPK14	mitogen-activated protein kinase 14	kinase
44	MIR122	microRNA 122	microRNA
45	MIR17	microRNA 17	microRNA
46	MIRN324	microRNA	
47	MYC	v-myc myelocytomatosis viral oncogene homolog	transcription regulator
48	NFKB1	nuclear factor of kappa light polypeptide gene enhancer in B-cells 1	transcription regulator
49	PDE6D	phosphodiesterase 6D, cGMP-specific, rod, delta	enzyme
50	PI3K		complex
51	PIN1	peptidylprolyl cis/trans isomerase, NIMA-interacting 1	enzyme
52	PMS1	PMS1 postmeiotic segregation increased 1 (S. cerevisiae)	other
53	POLA2	polymerase (DNA directed), alpha 2 (70kD subunit)	enzyme
54	PPARA	peroxisome proliferator-activated receptor alpha	ligand-dependent nuclear receptor
55	PPARG	peroxisome proliferator-activated receptor gamma	ligand-dependent nuclear receptor
56	PRMT1	protein arginine methyltransferase 1	enzyme
57	RARB	retinoic acid receptor, beta	ligand-dependent nuclear receptor
58	RELA	v-rel reticuloendotheliosis viral oncogene homolog A (avian)	transcription regulator
59	SAP25	sin3A-binding protein, SAP25	other
60	SFN	stratifin	other
61	SLC2A4	solute carrier family 2 (facilitated glucose transporter), member 4	transporter
62	SMAD1	SMAD family member 1	transcription regulator
63	SMARCA4	SWI/SNF related, actin dependent regulator of chromatin, subfamily a, member 4	transcription regulator
64	STAT2	signal transducer and activator of transcription 2, 113kDa	transcription regulator
65	STAT4	signal transducer and activator of transcription 4	transcription regulator
66	TGFBR1	transforming growth factor, β receptor 1	kinase
67	THRB	thyroid hormone receptor, β	ligand-dependent nuclear receptor
68	TP53	tumor protein p53	transcription regulator
69	TSC22D1	TSC22 domain family, member 1	transcription regulator
70	UBC	ubiquitin C	other
71	UBQLN4	ubiquilin 4	other
72	UNC119	unc-119 homolog (C. elegans)	other
73	VTN	vitronectin	other
74	YWHAB	tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, β	other
75	YWHAG	tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, γ	other
76	YWHAQ	tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, θ	other
77	YWHAZ	tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, ζ	enzyme
78	ZFP36	zinc finger protein 36, C3H type, homolog (mouse)	transcription regulator
79	ZNF143	zinc finger protein 143	transcription regulator

Table 5.3: 79 additional molecules added by the IPA® Path Explorer analysis tool for direct and indirect interactions.

This can be seen as a valid method to confirm the results presented. However, it suffered several important limitations, one of which is inherent to the Ingenuity database. It compiles only published and established work on interacting molecules. Therefore, it is only reliable to a certain point. It can validate findings if an interaction identified by the method is present in the database. However, if a strong interaction identified by the algorithm is not present in the Ingenuity database, this does not necessary mean that there is no actual interaction taking place there; but would just mean that no group have previously demonstrated its existence in a laboratory.

Moreover, this database is subject to changes, and from one date to another, additional interactions may have been added, and even more, some others may have been removed.

In addition, only 62 out of 100 genes matched in the IPA® database, which dramatically reduced the possibility to identify matching interactions, and 16 of the 62 matches could not find any interacting molecule according to IPA®. As well, since no direct connection could be observed, IPA® added extra molecules as intermediates to connect them, which added some complexity to the interaction map, and limited the validation of the findings.

Although no clear overlapping could be observed between the results from the interaction algorithm and the results from IPA®, as for example, there is no evidence in any sort of connection between CA9 and DTL in IPA® which was found by the algorithm; it was possible to identify with IPA® an indirect connection between CA9 and PGK1 (NM_000291) via HIF1α where it was found by the interaction algorithm with an interaction score of 6.54. Indeed, PGK1 has been reported to be one of the several genes of the glycolytic pathway to be upregulated by HIF1α (Hu et al., 2003; Wang et al., 2005; Winter et al., 2007).

A complete study of the interaction links will not be conducted, but it is possible to describe some interactions detected that appeared as well in the IPA® analysis. Among some of the highest positive interaction scores identified by the algorithm is the link between DTL and CDC45L (NM_003504).

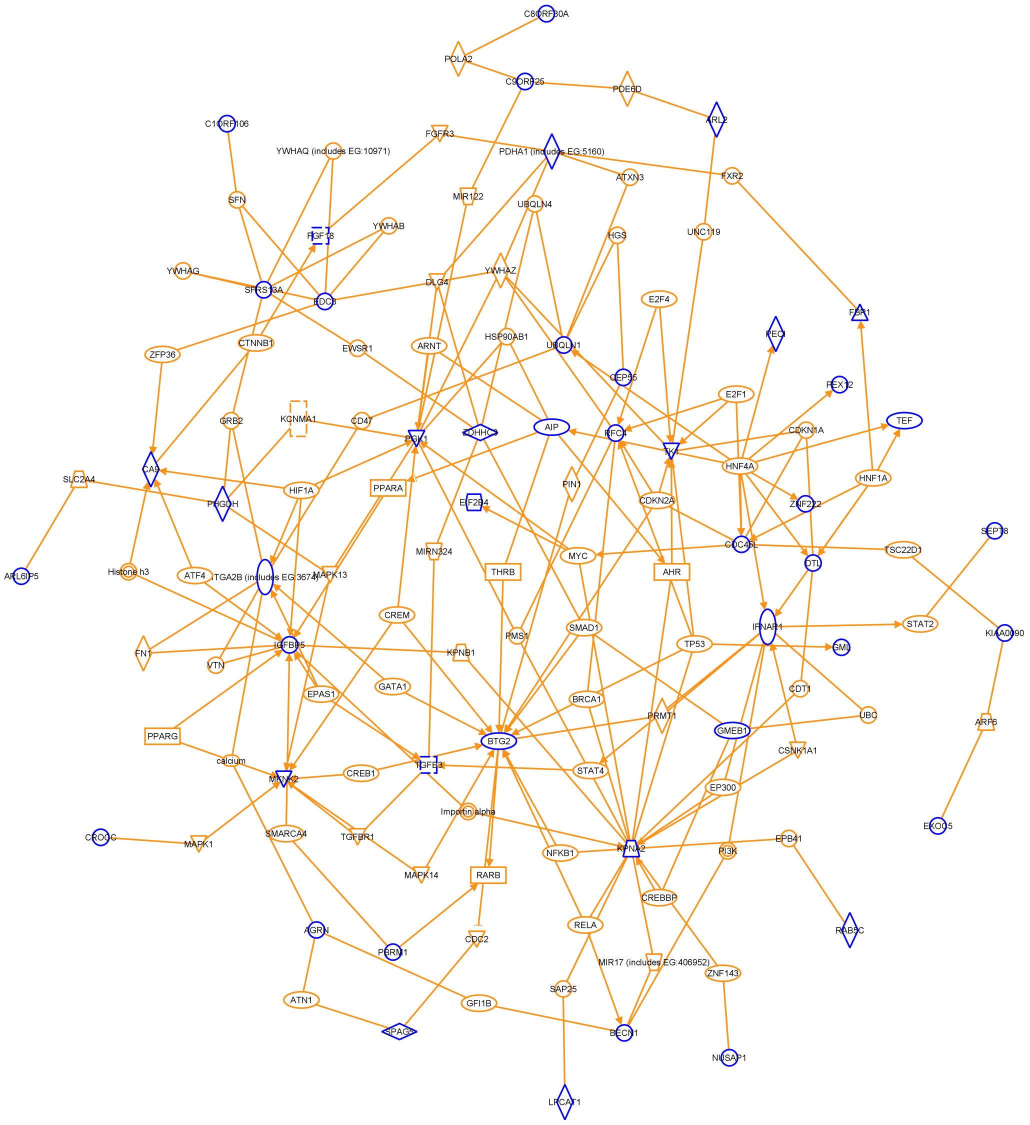


Figure 5.8: Results obtained by querying Ingenuity Pathway with 62 matching genes using the Path Explorer tool with the complete pool of 62 genes for Set A and Set B and any direction. In blue are the 62 genes from the top 100 identified genes, and in orange are the additional molecules added by the Path Explorer tool from Ingenuity.

IPA® found this link intermediated by CDKN1α, CDT1 or HNF-1α. HNF-1α protein is involved in gene regulation of DTL (or RAMP) and CDC45L (Odom et al., 2004). Therefore, it seems to make sense a strong association between DTL and CDC45L (with a score of 15.49) has been identified, as they would be following a similar coexpression pattern under the influence of HNF-1α. Other evidence showed that DTL, combined with other molecules, is likely to promote the degradation of CDT1 (Higa et al., 2006); and CDT1 enhances the expression of CDC45 (Ballabeni et al., 2009). Similarly, the interaction algorithm identified a relation that exists between DTL and KPNA2, confirmed by IPA® via CDT1 here again. CDT1 has just been described to have its degradation enhanced by a complex involving DTL (Higa et al., 2006), and according to the work of Sugimoto and co-workers, CDT1 protein is involved in the formation of a protein-protein complex with KPNA2 (Sugimoto et al., 2008).

Only one of the predicted negative interactions out of the 20 with highest absolute value interaction score matched with the results in IPA®. DTL again was found to have a strong negative influence upon TEF with an interaction score of -7.36. This was found to be the case by IPA® via the transcription factor HNF-1α.

Although several interactions matched with the database query, most of them were not (5 out of top 20 for positive scores, and only one out of top 20 for negative scores). Therefore, this approach was not entirely conclusive according to the validation of the algorithm. As a consequence, since no clear overlapping result could be observed for validation purpose, the algorithm has been optimised using an artificial dataset. Once optimised, the same dataset has been re-applied. The results are presented in the following section.

5.2 Optimisation of the method: new application of the van't Veer dataset

Following the optimisation of the algorithm, the van't Veer set has been reapplied in order to further improve and refine the interactome previously described. The whole set of data has been firstly analysed without consideration of the population characteristics and disease subclasses (*i.e.* all the cases where submitted to the analysis). Subsequently, the population was split in two groups considering the metastatic risk status of the individuals, and thus analysing separately the interactions that may occur in the good prognostic group and in the poor prognostic group. Finally, the resulting interactomes obtained from the algorithm are compared and analysed to further identify potential alteration of interacting pairs that may occur from one group to another.

5.2.1 Global population

As shown previously, the whole population of the dataset has been firstly applied to infer the interaction patterns for the selection of top 100 genes. The results are displayed as a map, here again using Cytoscape, in Figure 5.9.

Interestingly, we could firstly note that two of the probes belonging to the top 100 selection (with the Genbank numbers : NM_006117 and AF257175) are actually related to the same gene : Peroxisomal Δ3,Δ2-enoyl-CoA isomerase (PECI). Indeed, AF257175 refers to a cDNA sequence of the PECI gene, and NM_006117 refers to an alternative mRNA transcript sequence of the same gene. As a result, the algorithm was able to identify a very strong interaction score for the two pairs (see Table 5.4) : 47.389 between NM_006117 and AF257175 and 37.82 between AF257175 and NM_006117; which actually highlights a highly correlating expression pattern, thus confirming in this case the fact that both probes referred to the same gene product. And as a consequence, this artifact, resulting

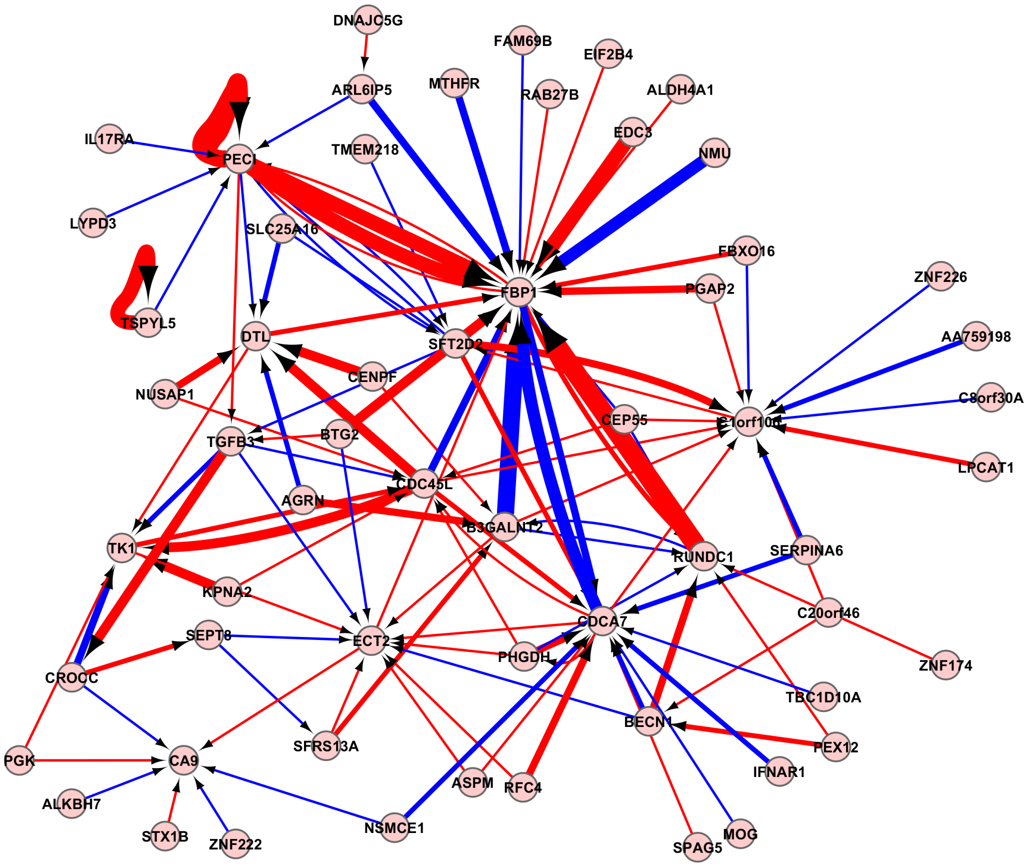


Figure 5.9: Display of resulting interaction map with algorithm 5, filter applied to all the interaction pairs with $C_v \leq 0.25$ and score > 10.

in those strong interaction scores, internally validates the robustness and particularly the reproducibility of the approach.

In addition, similar phenomena can be observed with Testis-Specific Y-encoded-like protein 5 (TSPYL5) and further validates the approach. This particular gene was represented indeed by two probes with Genbank numbers AA973313 (Contig14882_RC) and AL080059. These two isoforms of TSPYL5 were predicted interacting with scores of 59.36 for the interaction between AA973313 toward AL080059 and 37.60 between AL080059 toward AA973313.

In this application, it seems that DTL does not appear anymore to be so highly relevant in the present network, although it was the case when analysed by the previous version of the algorithm. We can observe as well sev-

	Source GenBank Nb.	Source Name	**Interaction Score**	Target Name	Target GenBank Nb.	Filter value
1	AI912791	FBXO16	18.622	FBP1	NM_000507	0.094
2	AI738508	ECT2	11.174	CA9	NM_001216	0.122
3	NM_016359	NUSAP1	9.262	ECT2	AI738508	0.144
4	NM_000291	PGK	14.572	CA9	NM_001216	0.147
5	AF052162	LPCAT1	16.649	C1orf106	NM_018265	0.154
6	NM_006582	GMEB1	7.095	STX1B	NM_003163	0.155
7	NM_018136	ASPM	10.182	ECT2	AI738508	0.156
8	NM_006117	**PECI**	**47.389**	**PECI**	AF257175	0.157
9	NM_000284	PDHA1	7.935	ECT2	AI738508	0.157
10	NM_006763	BTG2	26.544	FBP1	NM_000507	0.157
11	AJ011306	EIF2B4	8.968	PECI	AF257175	0.157
12	NM_016448	DTL	6.604	ECT2	AI738508	0.159
13	NM_004702	CCNE2	6.764	ECT2	AI738508	0.167
14	NM_006461	SPAG5	5.296	ECT2	AI738508	0.168
15	NM_005196	CENPF	9.157	ECT2	AI738508	0.174
16	NM_003163	STX1B	5.313	PGAP2	NM_014489	0.176
17	AI992158	CDCA7	13.370	ECT2	AI738508	0.176
18	NM_003239	TGFB3	9.224	PECI	NM_006117	0.177
19	NM_000599	IGFBP5	9.992	CA9	NM_001216	0.177
20	AF257175	**PECI**	**37.819**	**PECI**	NM_006117	0.178
21	AJ011306	EIF2B4	11.207	FBP1	NM_000507	0.178
22	NM_018136	ASPM	13.477	CDCA7	AI992158	0.181
23	NM_006625	SFRS13A	10.980	ECT2	AI738508	0.185
24	NM_013360	ZNF222	6.985	PECI	NM_006117	0.187
25	NM_018313	PBRM1	6.760	STX1B	NM_003163	0.187
26	AA759198	AA759198	6.125	STX1B	NM_003163	0.189
27	NM_014489	PGAP2	12.276	C1orf106	NM_018265	0.189
28	NM_018131	CEP55	8.176	ECT2	AI738508	0.189
29	NM_006461	SPAG5	10.491	CDCA7	AI992158	0.190
30	NM_002916	RFC4	22.253	CDCA7	AI992158	0.190
31	NM_020678	LRTM1	5.720	AA759198	AA759198	0.191
32	AI741080	B3GALNT2	11.466	C1orf106	NM_018265	0.192
33	NM_006623	PHGDH	10.555	ECT2	AI738508	0.192
34	NM_000507	FBP1	18.265	RUNDC1	AI918032	0.194
35	NM_003163	STX1B	11.068	CA9	NM_001216	0.194
36	AL133447	EDC3	7.143	STX1B	NM_003163	0.194
37	AI479831	MTHFR	5.264	AA759198	AA759198	0.195
38	NM_002916	RFC4	10.757	ECT2	AI738508	0.196
39	NM_016458	C8orf30A	5.005	AA759198	AA759198	0.196
40	NM_003504	CDC45L	18.138	CDCA7	AI992158	0.196
41	NM_020678	LRTM1	5.404	ZNF174	NM_003450	0.197
42	NM_003977	AIP	6.904	AA759198	AA759198	0.197
43	AL035297	SFT2D2	22.696	C1orf106	NM_018265	0.197
44	NM_014675	CROCC	17.745	SEPT8	D86957	0.197
45	NM_006625	SFRS13A	19.520	B3GALNT2	AI741080	0.197
46	AI992158	CDCA7	14.535	C1orf106	NM_018265	0.200
47	NM_016448	DTL	9.912	CDC45L	NM_003504	0.200
48	AF016903	AGRN	22.772	B3GALNT2	AI741080	0.201
49	NM_016448	DTL	10.548	TK1	NM_003258	0.201
50	NM_018354	C20orf46	5.211	TSPYL5	AA973313	0.203
51	NM_000507	FBP1	11.523	PECI	NM_006117	0.203
52	NM_000599	IGFBP5	9.104	PECI	NM_006117	0.204
53	NM_003258	TK1	19.079	CDC45L	NM_003504	0.204
54	AI992158	CDCA7	7.787	CA9	NM_001216	0.205
55	NM_018131	CEP55	10.652	CDC45L	NM_003504	0.205
56	AI738508	ECT2	5.438	PECI	AF257175	0.206
57	AI828164	FAM69B	5.020	ZNF174	NM_003450	0.206
58	AI554061	QSOX2	8.327	ECT2	AI738508	0.206
59	NM_006623	PHGDH	12.384	CDC45L	NM_003504	0.206
60	NM_000419	ITGA2B	5.379	STX1B	NM_003163	0.207
61	NM_000291	PGK	12.288	TK1	NM_003258	0.207
62	AW162411	TMEM218	7.385	RUNDC1	AI918032	0.207
63	NM_006117	PECI	11.011	TGFB3	NM_003239	0.207
64	AI918032	RUNDC1	50.522	FBP1	NM_000507	0.208
65	AF257175	PECI	8.123	TGFB3	NM_003239	0.208
66	NM_003977	AIP	6.520	STX1B	NM_003163	0.210
67	NM_016359	NUSAP1	10.471	CDC45L	NM_003504	0.212
68	AF257175	PECI	38.872	FBP1	NM_000507	0.215
69	NM_003504	CDC45L	7.694	ECT2	AI738508	0.216
70	AI183575	DNAJC5G	10.771	ARL6IP5	NM_006407	0.217

Table 5.4: Top 70 predicted positive interacting pairs by the algorithm filtered and sorted by Coefficient of variation of score and with scores > 5. In bold are the interacting pairs of the isoforms of PECI.

	Source GenBank Nb.	Source Name	**Interaction** **Score**	Target Name	Target GenBank Nb.	Filter value
1	NM_005196	CENPF	-8.712	PECI	AF257175	0.118
2	NM_003504	CDC45L	-21.411	FBP1	NM_000507	0.132
3	NM_013438	UBQLN1	-7.098	RUNDC1	AI918032	0.135
4	NM_018265	C1orf106	-5.061	ZNF226	NM_016444	0.138
5	NM_014400	LYPD3	-9.973	PECI	AF257175	0.147
6	R73468	SLC25A16	-8.574	ECT2	AI738508	0.155
7	NM_006681	NMU	-7.876	RUNDC1	AI918032	0.163
8	AI742029	TBC1D10A	-9.479	ECT2	AI738508	0.164
9	AI885466	C13orf37	-7.641	PECI	AF257175	0.166
10	NM_018131	CEP55	-13.571	FBP1	NM_000507	0.170
11	NM_003239	TGFB3	-10.991	ECT2	AI738508	0.173
12	LOC51630	UCHL5	-6.332	RUNDC1	AI918032	0.174
13	NM_002916	RFC4	-6.380	TSPYL5	AA973313	0.176
14	NM_000507	FBP1	-22.315	CDCA7	AI992158	0.177
15	AI918032	RUNDC1	-6.457	ECT2	AI738508	0.177
16	AL080059	TSPYL5	-12.444	PECI	AF257175	0.178
17	AI082587	IL17RA	-9.374	PECI	NM_006117	0.179
18	AI082587	IL17RA	-6.773	CA9	NM_001216	0.179
19	NM_000599	IGFBP5	-9.926	RUNDC1	AI918032	0.179
20	NM_016458	C8orf30A	-10.899	C1orf106	NM_018265	0.180
21	NM_006681	NMU	-32.387	FBP1	NM_000507	0.180
22	NM_014400	LYPD3	-10.005	PECI	NM_006117	0.182
23	NM_018131	CEP55	-12.094	RUNDC1	AI918032	0.183
24	NM_014489	PGAP2	-5.629	ECT2	AI738508	0.184
25	NM_006763	BTG2	-6.047	CA9	NM_001216	0.184
26	NM_006763	BTG2	-12.659	ECT2	AI738508	0.184
27	NM_014675	CROCC	-14.744	CA9	NM_001216	0.187
28	AL035297	SFT2D2	-9.395	PECI	NM_006117	0.187
29	AI479831	MTHFR	-6.546	PECI	AF257175	0.187
30	AI082587	IL17RA	-11.740	PECI	AF257175	0.188
31	D86957	SEPT8	-12.825	ECT2	AI738508	0.190
32	AI828164	FAM69B	-14.348	FBP1	NM_000507	0.190
33	AI918032	RUNDC1	-8.518	SFT2D2	AL035297	0.191
34	NM_005196	CENPF	-8.927	PECI	NM_006117	0.191
35	NM_006407	ARL6IP5	-7.200	CDC45L	NM_003504	0.192
36	NM_001756	SERPINA6	-16.452	CDCA7	AI992158	0.193
37	NM_003239	TGFB3	-8.677	SFRS13A	NM_006625	0.193
38	AF131828	C9orf25	-8.598	PECI	AF257175	0.194
39	AI912791	FBXO16	-13.402	C1orf106	NM_018265	0.194
40	AL035297	SFT2D2	-10.963	PECI	AF257175	0.195
41	R73468	SLC25A16	-12.790	SFT2D2	AL035297	0.195
42	AW137071	SMARCE1	-9.165	ECT2	AI738508	0.196
43	AI885466	C13orf37	-6.198	STX1B	NM_003163	0.196
44	NM_018136	ASPM	-6.246	PECI	AF257175	0.197
45	AL035297	SFT2D2	-21.792	FBP1	NM_000507	0.199
46	AI479831	MTHFR	-6.221	PECI	NM_006117	0.200
47	NM_001216	CA9	-6.284	PECI	NM_006117	0.200
48	AL137615	MKNK2	-6.335	ECT2	AI738508	0.202
49	NM_003239	TGFB3	-10.260	CDC45L	NM_003504	0.204
50	AI741080	B3GALNT2	-45.063	FBP1	NM_000507	0.205
51	AI347425	MOG	-11.578	CDCA7	AI992158	0.205
52	AI479831	MTHFR	-23.630	FBP1	NM_000507	0.206
53	AW137071	SMARCE1	-8.157	CA9	NM_001216	0.207
54	NM_014675	CROCC	-9.844	CDC45L	NM_003504	0.207
55	NM_003766	BECN1	-5.444	CDC45L	NM_003504	0.208
56	AI992158	CDCA7	-32.059	FBP1	NM_000507	0.208
57	AI082587	IL17RA	-8.471	ECT2	AI738508	0.208
58	NM_018136	ASPM	-7.071	RUNDC1	AI918032	0.208
59	FLJ11328	USP48	-5.541	SPAG5	NM_006461	0.209
60	AI741080	B3GALNT2	-12.538	RUNDC1	AI918032	0.211
61	AA534406	GALNT5	-5.786	TK1	NM_003258	0.213
62	AI992158	CDCA7	-7.135	FBXO16	AI912791	0.213
63	AA425275	COMMD6	-9.389	C1orf106	NM_018265	0.213
64	AL137615	MKNK2	-9.502	CENPF	NM_005196	0.213
65	D86957	SEPT8	-12.783	SFRS13A	NM_006625	0.214
66	AI885466	C13orf37	-8.309	PECI	NM_006117	0.216
67	NM_003766	BECN1	-10.807	ECT2	AI738508	0.216
68	NM_006117	PECI	-11.376	DTL	NM_016448	0.216
69	R73468	SLC25A16	-17.114	DTL	NM_016448	0.217
70	AW162411	TMEM218	-10.560	SFT2D2	AL035297	0.218

Table 5.5: Top 70 predicted negative interacting pairs by the algorithm filtered and sorted by Coefficient of variation of score, and scores > 5.

eral important hubs from the map (Figure 5.9), particularly around FBP1 (Fructose-1,6-bisphosphatase 1) that receives several very important connections (21 links with interaction scores ranging from 11.21 up to 50.52 for positive scores; and from -13.57 to -45.06). Furthermore, other hubs such as CDC45L, ECT2, TK1, CDCA7 and Corf106 can be noticed.

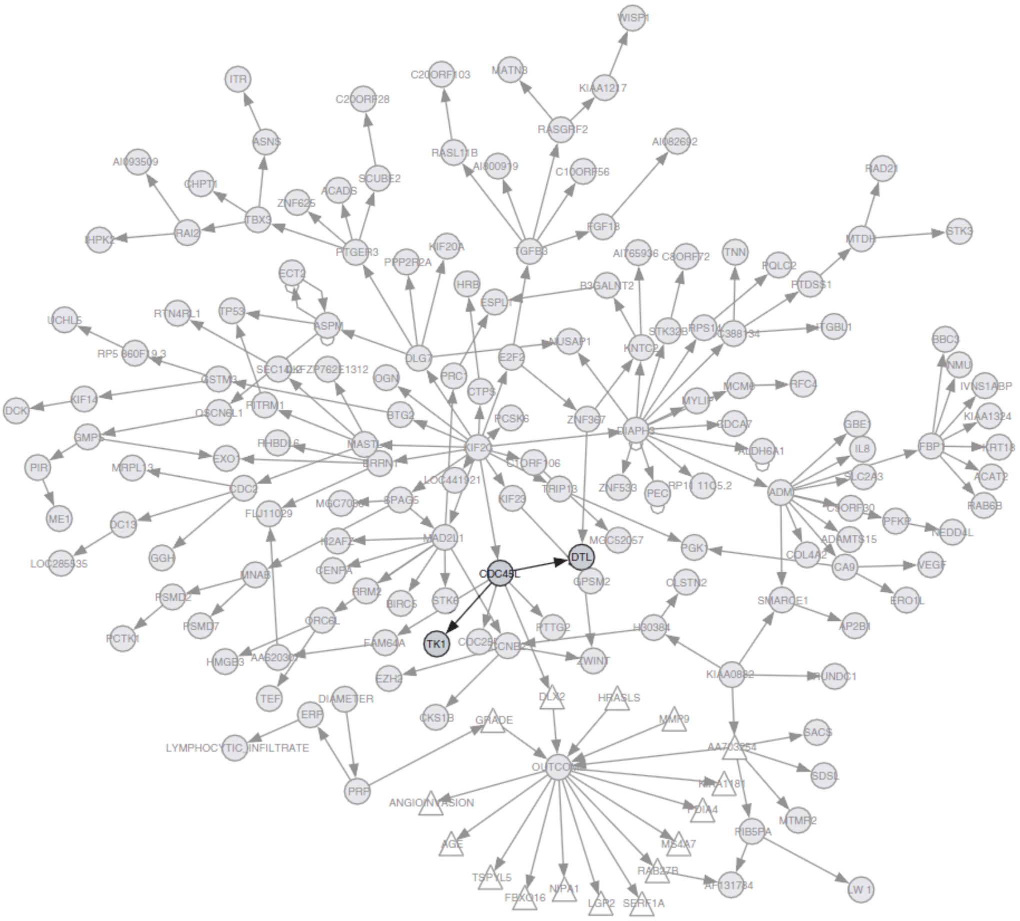


Figure 5.10: CDC45L, DTL and TK1 interactions (shown in bold) predicted by the solution proposed by Gevaert and coworkers (Gevaert et al., 2006)

It was possible to observe as well from the Table 5.4 that one of the highest positive scores, with one of the best filter value (ranked number 1) corresponded to the interaction between Contig42421_RC (for the F-box protein 16) and FBP1 (Alexe et al., 2006). In addition, within the top 70 interactions predicted with positive scores presented in Table 5.4, CA9 presented two scores higher than 10 with ECT2 (rank 2, score=11.17) and PGK (rank 4, score=14.57).

The hub observed with CDC45L, DTL and TK1 (Figure 5.11) has already been inferred with an alternative *in-silico* method developed by Gevaert and co-workers (Gevaert et al., 2006). The method proposed by Gevaert's group predicted interactions between CDC45L toward DTL and TK1 (see Figure 5.10). These particular interactions have been inferred by the present method too.

Comparison with IPA® results

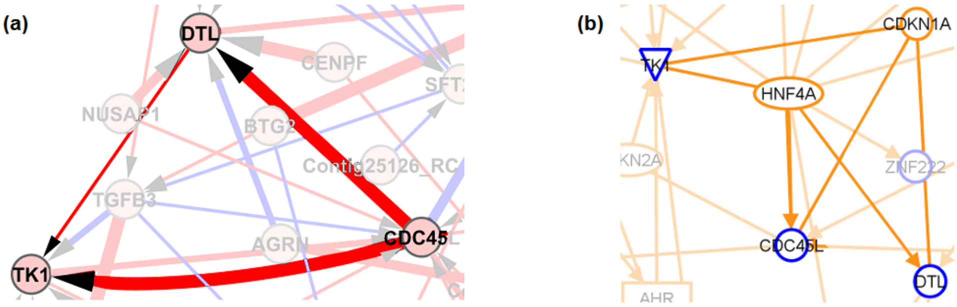


Figure 5.11: CDC45L, DTl and TK1 interactions predicted by the algorithm (a) and found by Ingenuity Pathway Analysis (b).

From the results presented earlier in this chapter (*c.f.* Figure 5.8), a rapid comparison of the interactions obtained with the refined algorithm and method has been conducted, similarly to what has been done previously.

An interesting connection can be observed with BTG2 expression influenced by FBP1. Although it is not shown by the map in Figure 5.9, because of a slightly higher filter value (0.32), there is a strong positive score predicted of 9.72, suggesting this influence. According to IPA® results, it has been reported that BTG2 expression is promoted by the expression of TP53 (a very well known gene involved in several cancers) (Amundson et al., 2005), which have been reported to bind with HSPA8, itself binding to FBP1 as well (Stelzl et al., 2005).

The network identified between CDC45L, DTL and TK1, which was confirmed by another inferring method (Gevaert et al., 2006) as mentioned earlier, is confirmed as well by IPA®, as seen in Figure 5.11. These 3 markers were found to be connected or related to CDKN1A. Their relationship with CDKN1A are:

- TK1 binds to CDKN1A (Huang and Chang, 2001).
- CDC45L binds to CDKN1A (Ramachandran et al., 2004).
- DTL expression promoted by CDKN1A (Wu et al., 2002).

Nevertheless, IPA® failed to show any direct relationship between the 3 markers. This can be explained by the fact that CDKN1A may be a key molecule intermediating between those 3 markers; and since it doesn't belong to the list of top 100 genes found by the preliminary analysis, it doesn't appear in the results for interaction. Consequently, the 3 genes are found by the algorithm to be directly connected event though no direct connection was found by IPA®.

However, from the interaction map predicted by the algorithm, FBP1 seems to appear as a key hub concentrating important vectors of influence towards it. The results with IPA® failed to confirm this important hub as it only identified a single indirect connection with PDHA1, intermediated by FXR2.

Nevertheless, despite some interesting indirect matches with the results obtained with IPA, it showed limited validation of the findings. The main reasons for this lack of overlapping results have been already discussed, in the previous section, and consists mainly of the fact that only very few interaction or relationships between molecules have been described and validated *in vitro* or *in vivo*.

5.2.2 Good prognostic group and poor prognostic group interaction analysis

The interest of the approach presented in this thesis does not only rely on the analysis of interactions that might exist between genes in the global population, but stands as well in the separate study of the interactions that could exist in the subpopulations, and ultimately compare the results to identify patterns of interactions significantly altered between the two subpopulations.

With the van't Veer study, the population could be separated regarding their status to recurrence. The population has been split into two subgroups with respect to the metastatic outcome of the individuals. The Good Prognostic Group received the individuals who showed no relapse until five years after treatment, and the Poor Prognostic group received the individuals who presented recurrence before five years.

Figure 5.12 presents the interaction map for the analysis of the good prognostic group. From this map, FBP1 does not seem to appear as a central hub here anymore. However, it seems that BCEN1, CCNE2, ZNF226 as well as ECT2 concentrate a number of interactions around them. We can see as well, that the group composed by DTL, CDC45L and TK1 is showing a similar pattern of inter connection than seen earlier in the global population analysis.

For the poor prognostic group, Figure 5.13 shows the resulting interaction map. We can observe here that some other hubs stood out from other markers, involving ARL2, FOXL2, Contig55725_RC (or CDCA7) or FBP1. However, the filer for this subpopulation did not keep the interactions involving DTL, TK1 and CDC45L, suggesting a lower consistency of these connections in the model.

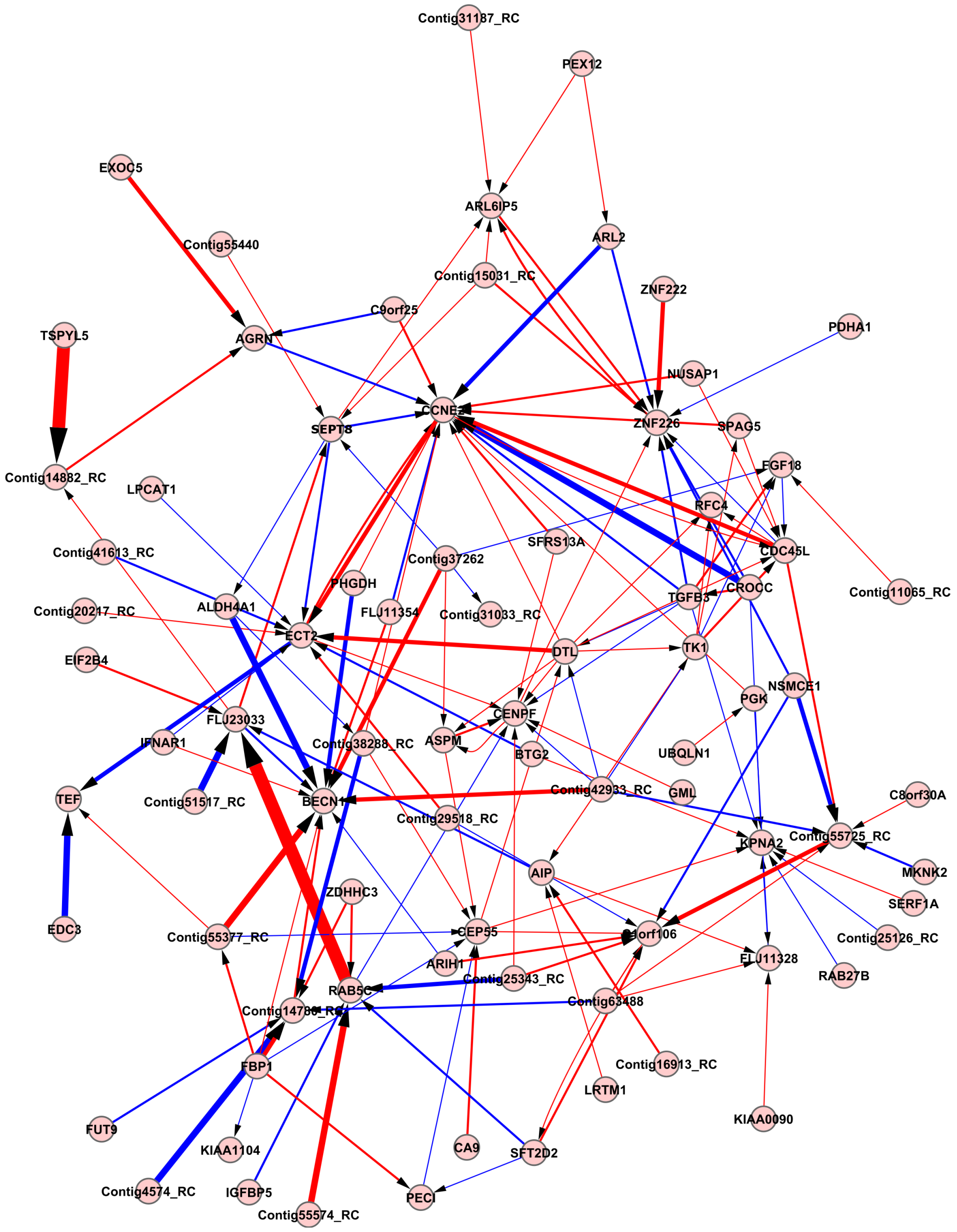


Figure 5.12: Display of resulting interaction map for good prognostic group, ranked and filtered by Coefficient of variation for values below 0.25. Red edges indicate positive interactions, and blue edges indicate negatives ones.

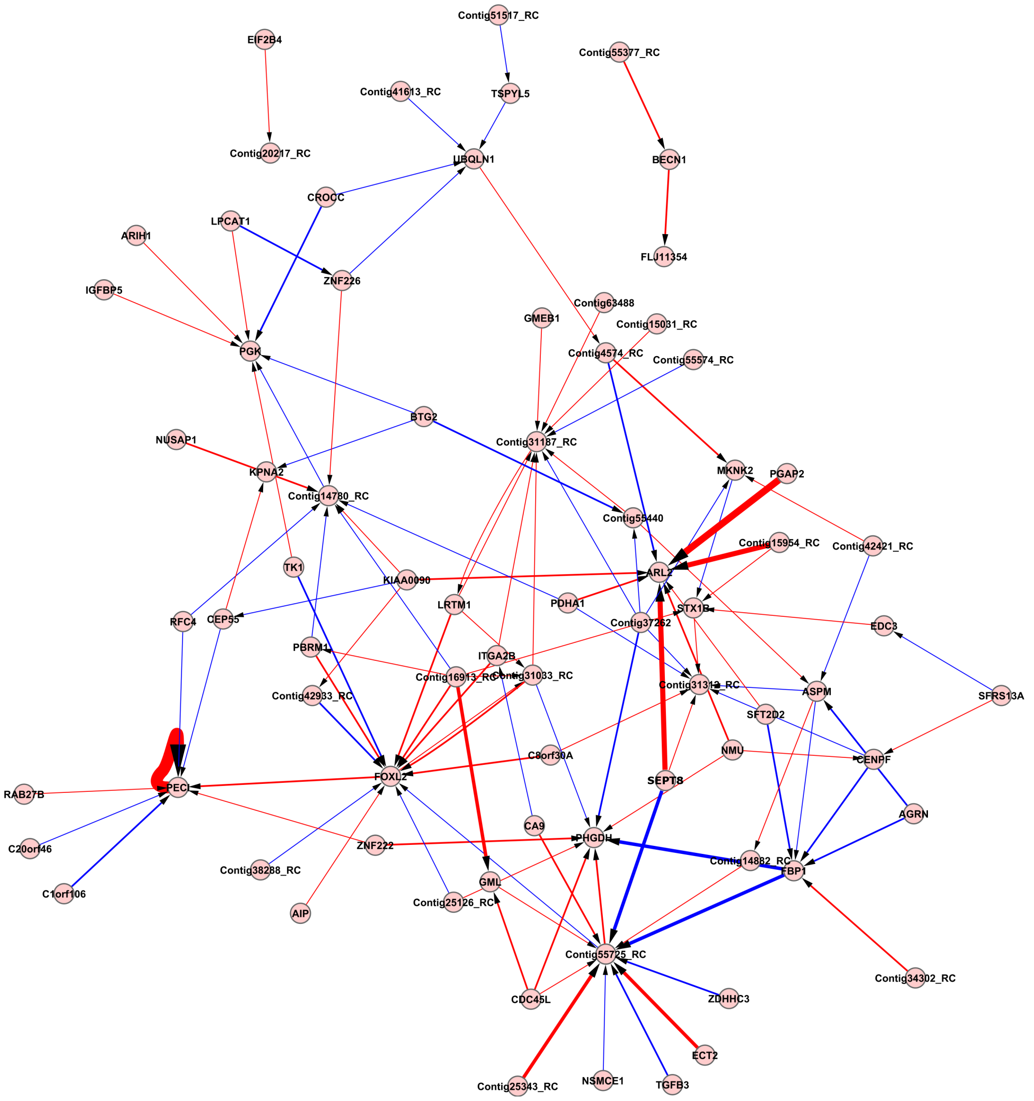


Figure 5.13: Display of resulting interaction map for poor prognostic group, ranked and filtered by Coefficient of variation for values below 0.25. Red edges indicate positive interactions, and blue edges indicate negatives ones.

5.2.3 Differential patterns

With the separate analysis of the two sub-groups from the original dataset, the most interesting aspect to investigate was to monitor the changes that may have occurred between those two groups. In other words: is there a possibility to identify strong patterns that constantly change from the good prognostic group to the poor prognostic one. The hope here would then be to identify new altered interactions or signalling paths that may have changed for the poor prognostic group and then may explain the difference of outcome for them, with the potential to lead to future new therapeutic strategies.

This idea has been applied to the two sub-populations from the van't Veer set. The original tables with scores for both groups were compared. The absolute value of interaction scores for each matching pair in both analysis were substracted (good prognostic group score - poor prognostic score), giving a new value for each interacting pair:

$$(|Score_{GPG}| - |Score_{PPG}|) \tag{5.1}$$

This value was therefore either positive (illustrating that higher interaction is found for the good prognostic group) or negative (for the opposite). Finally, displaying map for differential patterns with these new values using Cytoscape gave the result seen in Figure 5.14, when a filter only kept values of difference with an absolute value higher than 10. In this particular case, the coloured edges don't translate anymore an interaction, properly speaking, but highlights a difference of interaction. Therefore, the blue edges indicate a directed interaction that is significantly higher in the poor prognostic group, and the red edges indicate significantly higher interaction in the good prognostic group.

From the results, several significant alterations could be observed. It mainly involved TEF receiving a number of much higher influences in the good prognostic than in the Poor Prognostic Group; among them is an interaction from CA9, these alterations are amongst the most important for the interaction found higher in GPG. A similar pattern is observed for FLJ23033 (also known as TTL7 or Tubulin Tyrosine Ligase-like family, member 7). We can hypothesise from this that as higher interactions seem to take place toward TEF in good prognostic group, some alterations of these interactions is taking place for poorer prognostic individuals and therefore this could open a lead toward potential new therapeutic strategy. However, all of this is highly hypothetical, and needs deeper investigation to validate if the approach is correct.

As a consequence, because of the very nature of the method and its results, it is particularly challenging and difficult to validate with existing identifications of altered interactions in previous studies, and in the literature.

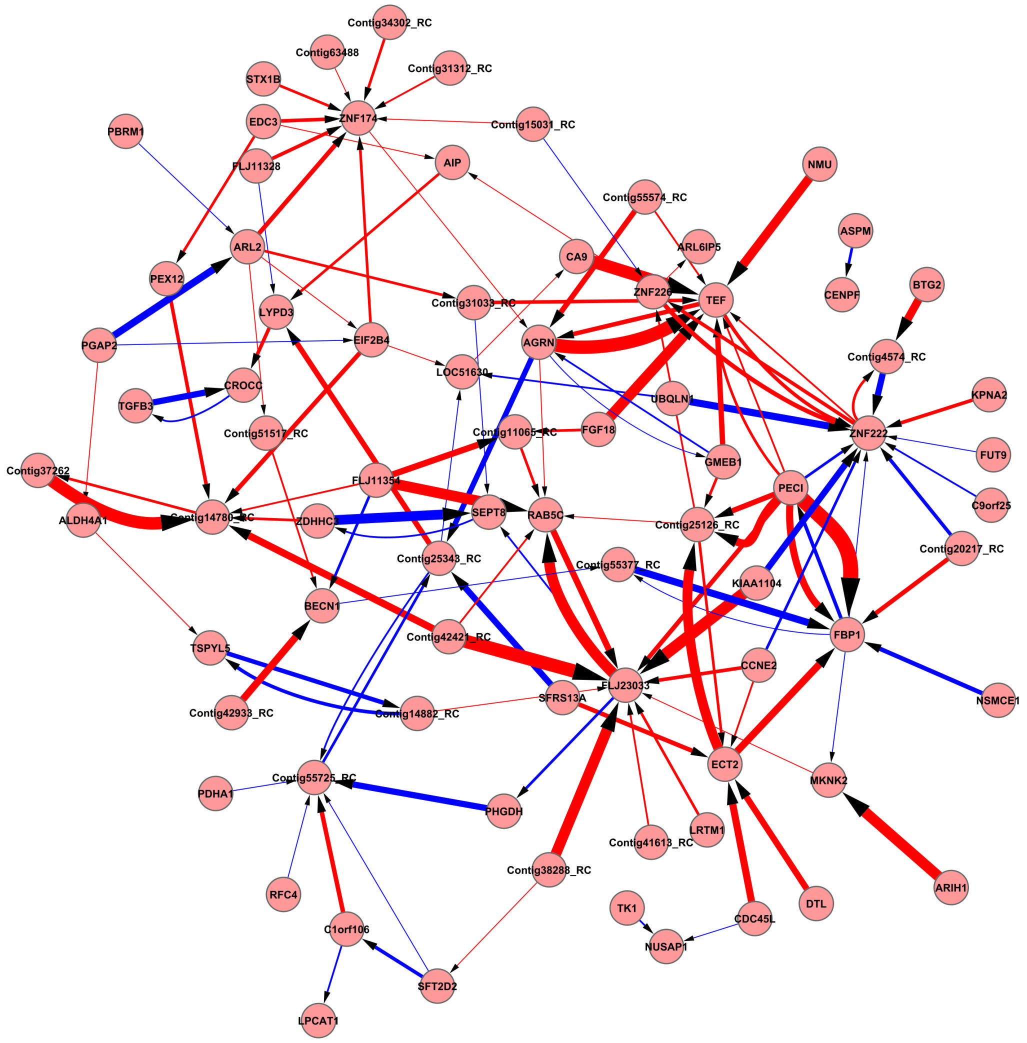


Figure 5.14: Display of resulting differential patterns of interactions between the good prognostic group and the poor prognostic group. Red edges represent interactions found to be higher in the good prognostic group, and blue edges represent interactions higher in the poor prognostic group.

5.3 Validation with the NKI295 cohort

Following the study of van't Veer and co-workers (van't Veer et al., 2002) and their 70-gene signature predictive of breast cancer metastatic risk, van de Vijver and co-workers provided one of the first major validation of this signature using the same protocol (van de Vijver et al., 2002) on a larger cohort of patients. The pool of samples consisted in 295 patients with breast cancer, showing different phenotypes of ER positivity, or treatment. This dataset has been chosen to validate the previous results obtained with the van't Veer set since they used the same microarray, and as a result, the same gene expression were monitored, and similar normalisation techniques were applied.

5.3.1 Methodology

A similar workflow (*c.f.* Figure 4.3) has been employed to proceed with the NKI295 dataset. The 100 genes previously identified were selected with their expression vector for each of the individuals. Prior interaction analysis, the list of the patients redundant with the van't Veer cohort were removed in order to guarantee a non-questionable comparison as all the patients are therefore completely unseen. The interaction study has been conducted only on the overall population here, as it was primarily sought to assess and validate the reproducibility of the method by comparing similar results with the global population.

5.3.2 Results

The interaction scores resulting from the analysis for the overall population in the NKI295 cohort are shown in Tables 5.6 and 5.7. These tables summarise the 50 most important interactions for both positive and negative scores, and for both NKI295 and van't Veer cohort. This allows the comparison of the results obtained by the algorithm. When a similar pair was found present in both cohort within the top50 interactions negative and positive, the pair was highlighted in bold in the table. Therefore, it is possible to observe that 12 of the top50 positive interactions and 8 out of the top 50 negative interactions were found overlapping for both cohorts. These results show that, although it is not perfect, this predictive interaction approach showed an encouraging reproducibility, and worth optimisation to improve further this reproducibility.

NKI295 cohort			van't Veer cohort		
Source name	Interaction scores	Target name	Source name	Interaction scores	Target name
TSPYL5	**64.499**	**TSPYL5**	**TSPYL5**	**59.361**	**TSPYL5**
PGK1	49.253	CA9	**RUNDC1**	**50.522**	**FBP1**
RAB5C	38.685	DHX58	B3GALNT2	48.759	AGRN
CA9	38.594	PGK1	**PECI**	**47.389**	**PECI**
TGFB3	36.274	01-sept	ALKBH7	43.707	FBP1
ECT2	35.119	IFNAR1	**ZDHHC3**	**41.56**	**SEPT8**
KPNA2	34.530	DHX58	**PECI**	**38.872**	**FBP1**
PECI	**32.729**	**PECI**	**PECI**	**38.257**	**FBP1**
ASPM	30.443	CENPF	IFNAR1	37.919	AGRN
RUNDC1	**29.526**	**FBP1**	**PECI**	**37.819**	**PECI**
ZNF222	27.253	ZNF226	TSPYL5	37.604	TSPYL5
ZNF226	26.949	ZNF222	EDC3	37.163	FBP1
C9orf25	**26.804**	**C20orf46**	ZBTB45	36.447	AGRN
DHX58	26.326	RAB5C	RUNDC1	34.979	BECN1
ALDH4A1	26.086	FBP1	STX1B	34.742	AGRN
TSPYL5	**25.827**	**TSPYL5**	**C9orf25**	**33.898**	**C20orf46**
PECI	25.325	ALKBH7	**EXOC5**	**33.364**	**IFNAR1**
BTG2	24.781	SEPT8	TGFB3	31.687	SEPT8
LRTM1	24.193	ALKBH7	**TSPYL5**	**31.265**	**IFNAR1**
RAB5C	**23.970**	**FBP1**	STX1B	31.053	SEPT8
SFRS13A	23.814	SERF1A	TGFB3	30.297	CROCC
C13orf37	23.675	ZNF420	FLJ11354	29.538	AGRN
DHX58	23.657	FBP1	SFT2D2	29.492	C20orf46
PBRM1	23.522	C20orf46	**NSMCE1**	**28.889**	**FBP1**
EXOC5	**23.372**	**IFNAR1**	RAB5C	28.308	FLJ11354
PECI	**23.254**	**PECI**	BECN1	28.236	SMARCE1
IFNAR1	22.968	ECT2	**RAB5C**	**28.018**	**FBP1**
CLSTN2	22.798	MTHFR	CDC45L	27.992	DTL
SFRS13A	22.741	IFNAR1	FLJ23033	27.860	SEPT8
C9orf25	22.454	FBXO41	CDCA7	27.735	DTL
QSOX2	22.435	CDCA7	CENPF	27.690	DTL
CEP55	21.963	CDCA7	KPNA2	27.081	TK1
ZDHHC3	**21.860**	**SEPT8**	FLJ11354	27.049	BECN1
FBP1	21.817	DHX58	BTG2	26.544	FBP1
PECI	**21.782**	**FBP1**	CDC45L	26.429	TK1
FGF18	21.667	TGFB3	CENPF	26.166	AGRN
ZNF420	21.327	IFNAR1	AGRN	25.309	IFNAR1
EXOC5	21.129	ZNF420	FOXL2	24.989	40429
ZNF226	21.098	ZNF420	PHGDH	24.835	CDCA7
TSPYL5	**20.972**	**IFNAR1**	BECN1	24.735	FLJ11354
CROCC	20.910	GMEB1	NUSAP1	24.701	SMARCE1
RAB27B	20.871	FBP1	IFNAR1	24.488	EXOC5
TGFB3	20.870	TEF	BTG2	24.305	SEPT8
PECI	**20.729**	**FBP1**	LIN9	24.301	DTL
BECN1	20.711	DHX58	FLJ11354	24.111	RAB5C
TEF	20.664	FBXO41	CROCC	23.967	TGFB3
CCNE2	20.538	DHX58	MKNK2	23.898	FBP1
SEPT8	20.501	TGFB3	FLJ11354	23.826	FBP1
SFRS13A	20.498	ECT2	C20orf46	23.756	FBP1
NSMCE1	**20.486**	**FBP1**	FLJ11328	23.458	IFNAR1

Table 5.6: Comparison of the 50 strongest positive scores obtained in both NKI295 and van't Veer interaction analysis. In bold are the interacting pair occurring in both lists.

NKI295 cohort			van't Veer cohort		
Source name	Interaction scores	Target name	Source name	Interaction scores	Target name
SLC25A16	-33.821	DHX58	B3GALNT2	-45.063	FBP1
PDHA1	-28.333	ALKBH7	ZNF222	-35.355	SERPINA6
CDCA7	**-27.870**	**FBP1**	PHGDH	-34.984	FBP1
C13orf37	-27.355	GMEB1	**CA9**	**-33.053**	**FBP1**
C13orf37	-26.527	SEPT8	SERF1A	-32.469	AGRN
ARL6IP5	-24.320	C9orf25	NMU	-32.387	FBP1
PGK1	-24.173	GMEB1	**CDCA7**	**-32.059**	**FBP1**
SFT2D2	**-23.238**	**FBP1**	C9orf25	-29.522	AGRN
SLC25A16	-23.229	C20orf46	PDHA1	-28.772	IFNAR1
AIP	-22.737	TSPYL5	CEP55	-26.596	AGRN
PEX12	-22.624	C20orf46	C1orf106	-26.507	FBP1
PECI	-22.558	DHX58	GALNT5	-26.213	FBP1
TEF	-21.943	C20orf46	C13orf37	-25.685	RAB5C
LYPD3	-21.816	FBXO41	PDHA1	-25.467	AGRN
QSOX2	-21.637	C13orf37	**CDCA7**	**-24.693**	**SERPINA6**
FBP1	**-21.463**	**CDCA7**	CDCA7	-24.360	IFNAR1
ZNF226	-21.408	PGK1	**ARL2**	**-24.160**	**SERPINA6**
ZNF420	-21.035	DHX58	SERPINA6	-24.041	SEPT8
ZNF226	-20.953	GMEB1	AA759198	-23.868	BECN1
C9orf25	-20.871	ALKBH7	MTHFR	-23.630	FBP1
EIF2B4	-20.743	IFNAR1	ECT2	-23.548	SEPT8
MKNK2	-20.361	TSPYL5	PGK	-23.305	CROCC
C1orf106	**-20.226**	**SERPINA6**	TSPYL5	-23.190	AGRN
ARL6IP5	-19.962	GMEB1	CA9	-23.120	NSMCE1
ALKBH7	-19.559	SEPT8	**BTG2**	**-23.066**	**C20orf46**
ARL2	-19.544	ZNF226	DTL	-22.550	AGRN
RAB27B	-19.203	GMEB1	CROCC	-22.527	TK1
GMEB1	-19.202	C13orf37	**C1orf106**	**-22.520**	**SERPINA6**
RAB27B	-19.046	C9orf25	**FBP1**	**-22.315**	**CDCA7**
SEPT8	-19.008	ALKBH7	EXOC5	-22.005	FBP1
CROCC	-18.833	C20orf46	B3GALNT2	-21.865	SEPT8
PGK1	-18.727	ZNF226	**SFT2D2**	**-21.792**	**FBP1**
BTG2	**-18.663**	**C20orf46**	ARL6IP5	-21.789	FBP1
CA9	**-18.496**	**FBP1**	CDC45L	-21.411	FBP1
SERF1A	-18.301	GMEB1	SEPT8	-21.238	FBP1
SEPT8	-18.239	C13orf37	RFC4	-21.181	AGRN
TSPYL5	-18.016	DHX58	ASPM	-20.898	AGRN
SFRS13A	-18.004	FBP1	SERPINA6	-20.822	C1orf106
COMMD6	-17.990	DHX58	IFNAR1	-20.762	CDCA7
C1orf106	-17.907	DHX58	AGRN	-20.670	DTL
PGK1	-17.865	SMARCE1	TGFB3	-20.617	SERPINA6
ZBTB45	-17.842	DHX58	CDCA7	-20.477	C20orf46
RAB27B	-17.696	SERPINA6	CDCA7	-20.414	SEPT8
ARL6IP5	-17.691	FBXO41	ARIH1	-20.329	SMARCE1
ARL2	**-17.531**	**SERPINA6**	EDC3	-20.303	IFNAR1
PECI	-17.422	C20orf46	ARL2	-20.121	SMARCE1
FBXO16	-17.413	GMEB1	ARL6IP5	-19.799	AGRN
QSOX2	-17.322	ALKBH7	NSMCE1	-19.703	CDCA7
SMARCE1	-17.258	DHX58	TK1	-19.493	CROCC
CDCA7	**-17.218**	**SERPINA6**	RAB5C	-18.789	C13orf37

Table 5.7: Comparison of the 50 strongest negative scores obtained in both NKI295 and van't Veer interaction analysis. In bold are the interacting pair occurring in both lists.

5.4 Discussion

The application of the first developed algorithm and then of its optimised version have been discussed in the present chapter. The challenge here was then to validate the approach at another level, with the results obtained, after the optimisation and validation obtained with the simulated dataset as discussed in chapter 4.

The first level of validation here tried to identify if any of the predicted interactions from real world data had been previously reported. For this reason, IPA® appeared as an extremely useful tool from its very nature: by compiling all the reported interactions in the literature from any study, IPA® allows to query the database and provides a direct insight into what has been described until now about any molecule in the literature. As a result, the query of the selected 100 genes gave 62 matching results, for which additional molecules were found to be intermediates in their interactions, adding an extra pool of 79 molecules.

Despite the challenge of finding existing and previously described features in literature that were predicted by the algorithm, it has been possible to identify some key components already described (CDC45L, TK1 and DTL) by different solutions or with different platforms with the van't Veer dataset (van't Veer et al., 2002).

The IPA® study presented evidence of interactions of these 3 components with CDKN1A, known to be inhibitor of DNA synthesis and to be involved in DNA damage repair due to transcriptional expression increase induced by the tumour suppressor p53 (Abbas et al., 2008), making sense in the context of cancer. It helped to identify that TK1 was described to bind with CDKN1A according to (Huang and Chang, 2001). This association would then perturb the DNA synthesis inhibitory function of CDKN1A, thus blocking its DNA repair activity. Similarly, in (Ramachandran et al., 2004) have been described interactions occurring between CDC45L and

CDKN1A (and with CDKN2A as well), found to be binding together. As well, CDKN1A was found connected to DTL by IPA® according to (Abbas et al., 2008). They showed an involvement of both features in ubiquitylation activity.

The results obtained with the global population showed as well an interesting hub concentrating several important interactions towards FBP1: 21 links with some of the most important positive and negative scores (from 11.21 to 50.52, and from -13.57 to -45.06). Interestingly, FBP1 transcript has been recently described as a regulator of *c-myc*, a well known proto-oncogene, in renal cancer (Weber et al., 2008), they described as well its frequent up-regulation in urothelial and prostate cancers, suggesting an important role in such tumour regulation. However, no evidence has been found of its implication in breast cancer until now. Nevertheless, this gives an important new confirmation of the relevance of what has been predicted by the approach, thus validating the findings.

The chapter described as well that, by selecting two different isoforms of two genes (PECI and TSPYL5) in the feature selection step, the approach has been able to pick up this artifact by predicting as the most highly interacting genes the 2 isoforms for both genes. Both isoforms of PECI were predicted with a score of 37.82 and 47.39 ranking respectively 5th and second in the global population with the van't Veer dataset. For TSPYL5, the two isoforms were found to be interacting with scores of 37.60 and 59.36. However, these two last interaction pairs, despite their very high scores, showed lower consistency (coefficients of variation of 0.24 and 0.35) and only ranked respectively 361st and 957th out of 9,900, explaining why these pairs did not appear in Table 5.4.

It is interesting to note as well that, the isoforms of the two genes were all predicted with positive scores, which one would expect considering they have the same pattern of expression. This suggests that the scoring system has reached an interesting level of accuracy in terms of signs of scores, at least for positive scores, but still requires further improvements.

To confirm the consistency of the findings, the comparison with the NKI295 cohort showed as well that amongst the highest interaction scores were found again the isoforms of both PECI and TSPYL5 (32.73 and 23.25 for PECI, and 64.5 and 25.83 for TSPYL5). These similar results between the two cohorts suggest an encouraging reproducibility of the method and the algorithm.

The comparison of the results for the population of the van't Veer study and the NKI295 study showed that amongst the 50 highest predicted interaction scores, 13 pairs (26%) were common in the positive scores, and 8 pairs (16%) in the negative scores. Although not perfect, these results proved an encouraging reproducibility of the algorithm and the method.

The analysis of the interactions occuring in the 2 subpopulations (good and poor prognostic groups) offered an interesting perspective for the analysis of the results. Splitting the populations in subgroups offers the ability to identify the potential changes occuring in the dynamic of the interactions between the groups. Identifying these patterns could then stress important dysregulation and provide new insight toward future therapeutic angles.

The study of the altered interaction in the van't Veer cohort showed some strong values as well, depicting really higher interactions taking place in the GPG, and others really higher in the PPG. The most altered interaction, increased in the GPG (red edges in Figure 5.14) is the one from PDHA1 toward TEF. As discussed, several important alterations could be observed between the two groups. These alterations involved TEF, TTL7, or ZNF222. No clear evidence of these altered interactions could be found in the literature and therefore validating such results is really difficult. However they offer the potential of new interesting leads toward novel therapeutic approaches or management in breast carcinomas.

The conjunction of all these aspects allowed to initiate the process of validation of the method after the first level of validation with the simulated dataset. These real world data showed interesting preliminary results by the method developed.

Chapter 6

Conclusion and discussion

6.1 Preamble

The recent advances achieved lately by important technological innovations has pushed forward the ability of the scientific community to enhance discovery, and, as a consequence, improve our knowledge of our environment. This is particularly true for the biomedical sciences. The latest technological achievements have incredibly enhanced the possibility for researchers to explore and discover the physiology or the physiopathology of some particular disease. It is now almost possible to assess the expression of the entire genome for an organism, even for a human individual (Ishkanian et al., 2004).

However, technological achievements may not always be met by the appropriate tools to properly analyse and interpret their output. The delay in time between new technologies and the emergence of the most appropriate tools to interpret their output can lead to misinterpretation, or even worse to wrong conclusions and false discovery. It is crucial that, besides the production of new technologies, a careful interpretation of the results is carried out.

Furthermore, although satisfactory analysis tools may see the day for the appropriate analysis of the output for a particular technology, they might offer a robust but yet superficial interpretation of the amount of data provided by such high-throughput technologies.

This motivated the reason why one of the aims of this study was to develop a new approach for deeper exploration of the potential information being held within an expression MA dataset.

It was sought, within this study, to develop, optimise and apply robust computational solutions to accurately identify potential markers specific to a question within a disease system, and subsequently identify the potential patterns of relationship between the markers of interest with an innovative ANN-based approach. The original contribution stood in the optimisation implemented in the existing approaches, and more importantly in the development of an original ANN based algorithm with the scope of interaction network inference.

6.2 Stepwise analysis and biomarker discovery

There has been a decade now since the appearance and rise of genomic MAs, which allowed an increasing screening over the years for gene expression level, assessed against particular phenotypes to identify key genes involved in different disease systems, or cell conditions. However, we have seen that despite this decade of research, relatively few biomarkers have emerged from this technology despite the great potential they carry for this purpose (Rifai et al., 2006).

This manuscript demonstrates that these technologies, and in particular their output, need to be appropriately addressed by robust and reliable methods in order to get the most accurate results; and we have shown the development and optimisation of one particular algorithm involving ANNs with the scope to investigate for potential biomarkers related to a question within a particular system like a disease (*e.g.* breast cancer). These methods have provided with interesting results, and have provided already with new potential biomarkers with the possibility to develop assays from them for future clinical routine application.

Among these newly identified biomarkers, this thesis reviewed the identification of CA9 as a marker of prediction for metastatic risk for breast cancer patients. The study involved the van't Veer breast cancer cohort (van't Veer et al., 2002) for which the originally developed Stepwise algorithm identified a smaller and more accurate fingerprint than the 70 gene model presented by van't Veer and coworkers and applied in the Mammaprint® test. The identification of CA9 was further validated by immunohistochemical staining from pathologists confirming the validity of CA9 as a biomarker for breast tumours, but confirming as well the robustness of the approach. More details can be found in the published paper (Lancashire et al., 2009b).

The further optimisation of the method has carried great improvements, mainly standing in a more reliable and accurate solution, but more importantly a faster one. The redevelopment using the C programming language allowed great efficiency to increase the speed of execution. It has been demonstrated that the number of 50 MCCV repeats show a good balance between time of execution with the modelling performance obtained.

It has been demonstrated as well that this solution has already given insight in some important applications, particularly in breast cancer; and some results were even validated by *in vitro* experimentation. This was illustrated by the example of applications presented previously where miRNAs could be related to be important in subclasses of breast cancer. The optimised Stepwise approach successfully identified sets of miRNAs predictive of ER, PR and HER2/*neu* status in breast cancer. Subsequent RQ-PCR investigations confirmed the presence of the miRNAs identified for each signature, confirming the validity of the model developed and the rosbustness of the approach.

As a result, the two studies discussed in these first chapters demonstrated the strength of the approaches developed, and the reliability of the models, signatures and markers identified.

6.3 Interaction algorithm

Within this study it was mainly sought to develop an algorithm for the exploration of interactions between genes within gene MA data. This algorithm uses a multifactorial ANN based approach to model the potential interactions. This approach, incorporating an ANN based interaction mapping, offers the potential to identify key components and interactions in a multifactorial fashion overcoming the limitations from previous studies with simple pairwise interaction inference methods, or based upon biomolecular functions; and facilitating the identification of unidirectional influences. Moreover, the reliability of ANNs will offer the benefit to incorporate multiplatform data for exploration, and therefore not only focus on a particular system, but investigate on a wider scale for cross-platform interactions (*e.g.* gene expression level in DNA with mRNA transcripts). However, for such cross-platform application, extreme care need to be taken as mentioned by Bitton et al. (Bitton et al., 2008). The inherent variability from one experiment to another would introduce a bias that would not allow valid conclusion to be drawn.

6.3.1 Optimisation

Chapter 3 demonstrated the development, and moreover the optimisation of the interaction algorithm and the method for its application. It showed great improvements after optimisation, in terms of predictive ability of correlated and interacting pairs of markers. In addition, it allowed identification of the most accurate scoring algorithm and optimal number of hidden nodes for the optimal balance between accuracy and time consumption. This led to an important leap towards a more effective algorithm that enabled more advanced validation step as shown by the results with the simulated dataset.

6.3.2 Validation

Previously, as discussed in the second chapter, Lancashire and coworkers (Lancashire et al., 2009a) identified a rank order of genes of biological relevance to metastasis from the van't Veer cohort (van't Veer et al., 2002). CA9, appeared as the strongest feature from the screening step. The interaction algorithm presented here showed CA9 to have a strong interaction with PCK1, both already described to interact and to be involved in breast cancer (Winter et al., 2007). Moreover, DTL, known to have a role in cancer cell proliferation (Crowe et al., 2003), has been found to be strongly influenced within the whole interactome. We have shown other features with biological significance, which helped to start building the validation for the method developed.

We have seen that, despite the difficulty to confirm the results with IPA, it was possible to identify potential relations between genes or gene products predicted by the algorithm with IPA via intermediate molecules. In addition it has been possible to overlap with some results of previous groups (Gevaert et al., 2006), particularly for the relation between CDC45L, TK1 and DTL again. Finally, we have shown, using a larger cohort of patients, that the algorithm was able to show encouraging results regarding reproducibility, with results and particularly interaction scores for gene pairs matching between the two cohorts.

The possible reasons why definite validation of the results could not be obtained are:

- There might be biases introduced by the technical experimentation for the MA, resulting in artifacts of the expression monitored from one gene to another.

- IPA only provides a database compiling what has been reported in the literature. Therefore, there is a chance that the interactions requested have not been published yet.

- Interactions could be identified only with intermediate molecules with IPA due to the low number of genes interrogated (100).

- Actual interactions might not be identified by the algorithm if they only take place under some particular conditions, or might be hidden among the others due to the heterogeneity of the population.

6.3.3 Advantages

The method described here brings some novelties to existing approaches discussed in the introduction. The ANN technology coupled with a robust cross-validation technique constitute an efficient solution and allow the complexity and bias induced by expression array technologies to be overcome (Hart et al., 2006). One of the main advantages compared to other proposed techniques stands in the consideration of the overall panel of genes rather than just single ones in a pairwise fashion or subgroup. However, the scoring system based on a weight analysis of the ANN does not yet permit a comprehensive analysis of a broad panel of genes, and this will require further work.

Additionally, the optimisation process coupled with the artificial dataset, allowed to assess and even to significantly improve one of the most important aspect of network inference modeling tools: true positive rate (Goldberg and Roth, 2003). It is important to clearly understand and be aware of the level of confidence it is possible to have in the results of the inferred network, and the optimisation work carried out and presented in this thesis gives the opportunity to have an idea of how genuine is the interaction map deciphered.

Furthermore this novel approach, contrary to most alternatives, has the potential to be applied to several types of data from different technologies like MS for exploring interactions between peptides or proteins.

6.3.4 Limitations

Nonetheless, despite encouraging primary validation, the approach may have few limitations. Firstly, one of the main limitations may come from the origin of the data itself. Indeed, we are well aware that the collection of experimental data involves a certain level of bias. However, as we can only investigate and explore for relative interactions, none of the interactions are absolute, but relative to all the others within the dataset. Such a limitation is not a constraint as far as the same procedure was employed for the whole dataset, but would have an increased influence once the approach is applied to a cross-platform dataset, investigating transcriptomics data for instance, and such an application will require extreme care in normalisation (Bitton et al., 2008).

Another important limitation has been highlighted by the validation with Ingenuity. Although the algorithm described here considers the markers in a multifactorial approach, it still suffers from the limited amount of genes that can be explored. Only very few connections could be confirmed using IPA, in an indirect way in addition. Since not all the gene products are questioned, there might be important features of connection missed by the algorithm. The restriction of taking 100 markers to investigate their potential relationships reaches the limit of validation, and it is highly unlikely that out of the 100 markers, they have been directly described to be interacting. However, we have been able to find, with the help of the literature and IPA that, even if some markers have not yet been described directly, indirect pathways may have been described, and show that there is a likelihood that what was predicted could actually be what is really taking place *in vivo*. Nevertheless, this highlights a major limitation, maybe not for the algorithm, but more for the method, and particularly the feature selection step. Reducing the system to only 100 markers can give a biased insight. The awareness of this limitation, is crucial, and it needs to be addressed.

Another limitation could be identified as well, coming from the complexity of the overall interactome obtained which appeared to be important too, and the relevance of the filtering strategy employed has been shown to be critical (*c.f.* chapter 3). It will be challenging to propose and address alternative filtering strategies and be able to only keep the most accurately predicted and significant interactions, meanwhile not to remove any interaction that can have relevance in the actual interactome. In chapter 3, it has been demonstrated that a relevant filtering method could increase the true positive rate (*i.e.* the number of interactions actually taking place out of the total number of interactions predicted by the algorithm) up to 70% in average (for 2 hidden nodes and top 100 selection).

But the most important limitation here stands in the modelling process. The expression level of the gene considered as output in the ANN is modelled in a categorical fashion. The other genes are then used to predict, by training the ANN, if the targeted gene is over or under expressed compared to its median value accross the overall population. Therefore a lot of information is missed as it is simplifying an analogic problem in a numeric way, reducing the precision that could be acquired by using the continuous data of the actual expression values.

Therefore, this tool, although not entirely validated, provides an insight to what can happen in interactions and offers trails to follow for future interaction studies on the bench.

6.3.5 Future work

Re-development of the method

One of the main limitations so far within this algorithm stood in the prediction of a categorical (binary) output, as discussed previously. This is one of the first and most critical improvements to carry out, by using a continuous output for the ANN. It will then be interesting to reassess the complete study with the artificial dataset, and compare the improvement of performance with the previous results.

Another interesting challenge to tackle will be the display strategy, we have seen that another consideration can be a heatmap, and it would be an interesting development of the program.

A crucial final point to work on will be the number of inputs integrated in the interaction analysis. Unless studying simple organisms, it is unlikely that a system can be reduced to a pool of 100 genes. Therefore, for accurate and meaningful results, the program needs improvements in terms of feature selection and coping with higher inputs for the scoring algorithm.

Further validation

Further validation is required in order to confirm the findings of the algorithm and subsequently validate the method itself. The next objective would be to validate it against existing concurrent methods such as bayesian network-based methods (Hartemink et al., 2002; Jansen et al., 2003; Needham et al., 2009), as it as been carried out with IPA and the Gevaert method (Gevaert et al., 2006) in this work. Although, since no gold standard method has been described and fully validated to date, due to the youth of this domain, developing a validation protocol would be challenging, and require a certain amount of time. Nevertheless, beside using alternative *in silico* based techniques, the ultimate validation of the findings would be *in vitro* or *in vivo* techniques, using for instance gene silencing techniques. The observation of the interaction between 2 molecules could be validated by silencing the expressing gene of one of them, and monitor the effect it might have on the other one. Although conducting such a project would become really time consuming and require a multidisciplinary project and team.

Range of applications

Cross platform applications (miRNAs and cDNA for miRNA targets), or even transcriptomics (mRNA, cDNA) can be considered as potential future applications. However, as discussed previously, multi-experiment datasets would hold biased informations, and in addition extreme care needs to be taken when considering post-transcriptional alterations (*e.g.* alternative splicing) that take place in a living organism.

We have seen that, despite the fact it may suffer from some disadvantages and limitations and needs to be further improved, it is clear, according to our preliminary results, that a novel powerful ANN-based tool for interaction discovery has been developed in its early stages. This has the potential to investigate biological systems further, and possibly discover new potential targets for therapies.

Bibliography

T. Abbas, U. Sivaprasad, K. Terai, V. Amador, M. Pagano, and A. Dutta. Pcna-dependent regulation of p21 ubiquitylation and degradation via the crl4cdt2 ubiquitin ligase complex. *Genes & Development*, 22:2496–2506, 2008.

A. C. Ahn, M. Tewari, C.-S. Poon, and R. S. Phillips. The limits of reductionism in medicine: Could systems biology offer an alternative? *PLOS Medicine*, 3(6):e208, 2006.

B. Y. Ahn, A. N. Elwi, B. Lee, D. L. Trinh, A. C. Klimowicz, A. Yau, J. A. Chan, A. Magliocco, and S.-W. Kim. Genetic screen identifies insulin-like growth factor binding protein 5 as a modulator of tamoxifen resistance in breast cancer. *Cancer Research*, 70(8):3013–3019, 2010.

G. Alexe, S. Alexe, D. E. Axelrod, T. O. Bonates, I. I. Lozina, M. Reiss, and P. L. Hammer. Breast cancer prognosis by combinatorial analysis of gene expression data. *Breast Cancer Research*, 8:R41, 2006.

S. A. Amundson, K. T. Do, L. Vinikoor, C. A. Koch-Paiz, M. L. Bittner, J. M. Trent, P. Meltzer, and A. J. F. Jr. Stress-specific signatures: expression profiling of p53 wild-type and -null human cells. *Oncogene*, 24: 4572–4579, 2005.

A. Antoniou, P. D. P. Pharoah, S. Narod, H. A. Risch, J. E. Eyfjord, J. L. Hopper, N. Loman, H. Olsson, O. Johannsson, . . Borg, B. Pasini, P. Radice, S. Manoukian, D. M. Eccles, N. Tang, E. Olah, H. Anton-Culver, E. Warner, J. Lubinski, J. Gronwald, B. Gorski, H. Tulinius, S. Thorlacius, H. Eerola, H. Nevanlinna, K. Syrjäkoski, O.-P. Kallioniemi, D. Thompson, C. Evans, J. Peto, F. Lalloo, D. G. Evans, and D. F. Easton. Average risks of breast and ovarian cancer associated with brca1 or brca2 mutations detected in case series unselected for family history: A combined analysis of 22 studies. *American Journal of Human Genetics*, 72(5):1117–1130, 2003.

A. Ballabeni, R. Zamponi, G. Caprara, M. Melixetian, S. Bossi, L. Masiero, and K. Helin. Human cdt1 associates with cdc7 and recruits cdc45 to chromatin during s phase. *The Journal of Biological Chemistry*, 284: 3028–3036, 2009.

Z. Bar-Joseph, G. K. Gerber, T. I. Lee, N. J. Rinaldi, J. Y. Yoo, F. Robert, D. B. Gordon, E. Fraenkel, T. S. Jaakkola, R. A. Young, and D. K. Gifford. Computational discovery of gene modules and regulatory networks. *Nature Biotechnology*, 21(11):1337–1342, 2003.

A.-L. Barabási and Z. N. Oltvai. Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, 5(2):101–113, 2004.

I. A. Basheer and M. Hajmeer. Artificial neural networks: fundamentals, computing, design, and application. *Journal of Microbiological Methods*, 43(1):3–31, 2000.

A. Beghin, E.-L. Matera, S. Brunet-Manquat, and C. Dumontet. Expression of arl2 is associated with p53 localization and chemosensitivity in a breast cancer cell line. *Cell Cycle*, 7(19):3074–3082, 2008.

R. E. Bellman. *Adaptive Control Processes: A Guided Tour.* Princeton University Press, 1961.

F. Bertucci and D. Birnbaum. Reasons for breast cancer heterogeneity. *Journal of Biology*, 7(2):6, 2008.

C. M. Bishop. *Neural Networks for Pattern Recognition.* Oxford University Press Inc, New York, USA, 1995.

D. A. Bitton, M. J. Okoniewski, Y. Connolly, and C. J. Miller. Exon level integration of proteomics and microarray data. *BMC Bioinformatics*, 9: 118–128, 2008.

C. Blenkiron, L. D. Goldstein, N. P. Thorne, I. Spiteri, S.-F. Chin, M. J. Dunning, N. L. Barbosa-Morais, A. E. Teschendorff, A. R. Green, I. O. Ellis, S. Tavaré, C. Caldas, and E. A. Miska. Microrna expression profiling of human breast cancer identifies new markers of tumor subtype. *Genome Biology*, 8(10):R214, 2007.

H. Bloom and W. Richardson. Histological grading and prognosis in breast cancer; a study of 1409 cases of which 359 have been followed for 15 years. *British Journal of Cancer*, 11(3):359–377, 1957.

V. Bogoevska, A. Horst, B. Klampe, L. Lucka, C. Wagener, and P. Nollau. Ceacam1, an adhesion molecule of human granulocytes, is fucosylated by fucosyltransferase ix and interacts with dc-sign of dendritic cells via lewis x residues. *Glycobiology*, 16(3):197–209, 2006.

A. D. Boiko, S. Porteous, O. V. Razorenova, V. I. Krivokrysenko, B. R. Williams, and A. V. Gudkov. A systematic search for downstream mediators of tumor suppressor function of p53 reveals a major role of btg2 in suppression of ras-induced transformation. *Genes & Development*, 20(2): 236–252, 2006.

S. K. Chatterjee and B. R. Zetter. Cancer biomarkers: knowing the present and predicting the future. *Future Oncology*, 1(1):37–50, 2005.

A. M. Cheng, M. W. Byrom, J. Shelton, and L. P. Ford. Antisense inhibition of human mirnas and indications for an involvement of mirna in cell growth and apoptosis. *Nucleic Acids Research*, 33:1290–1297, 2005.

K. C. Chipman and A. K. Singh. Predicting genetic interactions with random walks on biological networks. *BMC Bioinformatics*, 10:17, 2009.

C. A. Chrestensen, J. K. Shuman, A. Eschenroeder, M. Worthington, H. Gram, and T. W. Sturgill. Mnk1 and mnk2 regulation in her2-overexpressing breast cancer lines. *The Journal of Biological Chemistry*, 282(7):4243–4252, 2007.

S. Christley, Q. Nie, and X. Xie. Incorporating existing network information into gene network inference. *PLOS One*, 4(8):e6799, 2009.

W. M. Claudino, A. Quattrone, L. Biganzoli, M. Pestrin, I. Bertini, and A. D. Leo. Metabolomics: Available results, current research projects in breast cancer, and future applications. *Journal of Clinical Oncology*, 25 (19):2840–2846, 2007.

C. G. Colpaert, P. B. Vermeulen, S. B. Fox, A. L. Harris, L. Y. Dirix, and E. A. V. Marck. The presence of a fibrotic focus in invasive breast carcinoma correlates with the expression of carbonic anhydrase ix and is a marker of hypoxia and poor prognosis. *Breast Cancer Research and Treatment*, 81(2):137–147, 2003.

D. L. Crowe, R. Kim, and R. A. Chandraratna. Retinoic acid differentially regulates cancer cell proliferation via dose-dependent modulation of the mitogen-activated protein kinase pathway. *Molecular Cancer Research*, 1:532–540, 2003.

A. Dankof, F. R. Fritzsche, E. Dahl, S. Pahl, P. Wild, M. Dietel, A. Hartmann, and G. Kristiansen. Kpna2 protein expression in invasive breast carcinoma and matched peritumoral ductal carcinoma in situ. *Virchows Archive*, 451(5):877–881, 2007.

M. DeLaurentiis, G. Arpino, E. Massarelli, A. Ruggiero, C. Carlomagno, F. Ciardiello, G. Tortora, D. D'Agostino, F. Caputo, G. Cancello, E. Montagna, L. Malorni, L. Zinno, R. Lauria, A. R. Bianco, and S. D. Placido. A meta-analysis on the interaction between her-2 expression and response to endocrine treatment in advanced breast cancer. *Clinical Cancer Research*, 11(13):4741–4748, 2005.

A. B. Deleo, G. Jayt, E. Appellat, G. C. Duboist, L. W. Lawt, and L. J. Old. Detection of a transformation-related antigen in chemically induced sarcomas and other transformed cells of the mouse. *PNAS*, 76(5):2420–2424, 1979.

E. Diamandis. Mass spectrometry as a diagnostic and a cancer biomarker discovery tool: opportunities and potential limitations. *Molecular and Cellular Proteomics*, 3(4):367–378, 2004a.

E. P. Diamandis. How are we going to discover new cancer biomarkers? a proteomic approach for bladder cancer. *Clinical Chemistry*, 50:793–795, 2004b.

C. Domeniconi, D. Papadopoulos, D. Gunopulos, and S. Ma. Subspace clustering of high dimensional data. In *Proceedings of the SIAM International Conference on Data Mining*, 2004.

S. Dreiseitl, L. Ohno-Machado, H. Kittler, S. Vinterbo, H. Billhardt, and M. Binder. A comparison of machine learning methods for the diagnosis of pigmented skin lesions. *Journal of Biomedical Informatics*, 34(1):28–36, 2001.

P. J. Drew and J. R. Monson. Artificial neural networks. *Surgery*, 127(1): 3–11, 2000.

D. J. Duggan, M. Bittner, Y. Chen, P. Meltzer, and J. M. Trent. Expression profiling using cdna microarrays. *Nature Genetics*, 21:10–14, 1999.

F. Ebeling, P. Stieber, M. Untch, D. Nagel, G. Konecny, U. Schmitt, A. Fateh-Moghadam, and D. Seidel. Serum cea and ca 15-3 as prognostic factors in primary breast cancer. *British Journal of Cancer*, 86: 1217–1222, 2002.

K. A. Egland, X. F. Liu, S. Squires, S. Nagata, Y.-G. Man, T. K. Bera, M. Onda, J. J. Vincent, R. L. Strausberg, B. Lee, and I. Pastan. High expression of a cytokeratin-associated protein in many cancers. *PNAS*, 103(15):5929–5934, 2006.

M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *PNAS*, 95(25):14863–14868, 1998.

D. M. A. El-Rehim, G. Ball, S. E. Pinder, E. Rakha, C. Paish, J. F. Robertson, D. Macmillan, R. W. Blamey, and I. O. Ellis. High-throughput protein expression analysis using tissue microarray technology of a large well-characterised series identifies biologically distinct classes of breast cancer confirming recent cdna expression analyses. *International Journal of Cancer*, 116(3):340–350, 2005.

S. E. Fahlman. Faster-learning variations on back-propagation: An empirical study. In *1988 Connectionist Models Summer School*, 1988.

G. C. Fletcher, S. Patel, K. Tyson, P. J. Adam, M. Schenker, J. A. Loader, L. Daviet, P. Legrain, R. Parekh, A. L. Harris, and J. A. Terrett. hag-2 and hag-3, human homologues of genes involved in differentiation, are associated with oestrogen receptor-positive breast tumours and interact with metastasis gene c4.4a and dystroglycan. *British Journal of Cancer*, 88(4):579–585, 2003.

T. S. Furey, N. Cristianini, N. Duffy, D. W. Bednarski, M. Schummer, and D. Haussler. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, 16(10):906–914, 2000.

J. M. García-Pedrero, E. Kiskinis, M. G. Parker, and B. Belandia. The swi/snf chromatin remodeling subunit baf57 is a critical regulator of estrogen receptor function in breast cancer cells. *The Journal of Biological Chemistry*, 281(32):22656–22664, 2006.

O. Gevaert, F. D. Smet, D. Timmerman, Y. Moreau, and B. D. Moor. Predicting the prognosis of breast cancer by integrating clinical and microarray data with bayesian networks. *Bioinformatics*, 22(14):e184–e190, 2006.

L. M. Gierasch and A. Gershenson. Post-reductionist protein science, or putting humpty dumpty back together again. *Nature Chemical Biology*, 5(11):774–777, 2009.

G. S. Ginsburg and J. J. McCarthy. Personalized medicine: revolutionizing drug discovery and patient care. *Trends in Biotechnology*, 19(12):491–496, 2001.

F. Günther, N. Wawro, and K. Bammann. Neural networks for modeling gene-gene interactions in association studies. *BMC Genetics*, 10:87, 2009.

P. Gold and S. O. Freedman. Demonstration of tumor-specific antigens in human colonic carcinomata by immunological tolerance and absorption techniques. *Journal of Experimental Medicine*, 121(3):439–462, 1965.

D. S. Goldberg and F. P. Roth. Assessing experimentally derived interactions in a small world. *PNAS*, 100(8):4372–4376, 2003.

W. M. Grady, R. K. Parkin, P. S. Mitchell, J. H. Lee, Y.-H. Kim, K. Tsuchiya, M. K. Washington, C. Paraskeva, J. K. V. Willson, A. M. Kaz, E. M. Kroh, A. Allen, B. R. Fritz, S. D. Markowitz, and M. Tewari. Epigenetic silencing of the intronic microrna hsa-mir-342 and its host gene evl in colorectal cancer. *Oncogene*, 27(27):3880–3888, 2008.

S. Griffiths-Jones, H. K. Saini, S. van Dongen, and A. J. Enright. mirbase: tools for microrna genomics. *Nucleic Acids Research*, 36:D154–D158, 2008.

D. Hanahan and R. A. Weinberg. The hallmarks of cancer. *Cell*, 100(1): 57–70, 2000.

C. E. Hart, E. Mjolsness, and B. J. Wold. Connectivity in the yeast cell cycle transcription network: Inferences from neural networks. *PLoS Computational Biology*, 2(12):1592–1607, 2006.

A. J. Hartemink, D. K. Gifford, T. S. Jaakkola, and R. A. Young. Bayesian methods for elucidating genetic regulatory networks. *IEEE Intelligent Systems*, 17(2):37–43, 2002.

J. L. Haybittle, R. W. Blamey, C. W. Elston, J. Johnson, P. J. Doyle, F. C. Campbell, R. I. Nicholson, and K. Griffiths. A prognostic index in primary breast cancer. *British Journal of Cancer*, 45(3):361–366, 1982.

D. O. Hebb. *The organisation of behaviour*. Wiley, New York, USA, 1949.

R. Hecht-Nielsen. Applications of counterpropagation networks. *Neural Networks*, 1(2):131–139, 1988.

W. A. Hendricks and K. W. Robey. The sampling distribution of the coefficient of variation. *Annals of Mathematical Statistics*, 7(3):129–132, 1936.

A. Hendrix, D. Maynard, P. Pauwels, G. Braems, H. Denys, R. V. den Broecke, J. Lambert, S. V. Belle, V. Cocquyt, C. Gespach, M. Bracke, M. C. Seabra, W. A. Gahl, O. D. Wever, and W. Westbroek. Effect of the secretory small gtpase rab27b on breast cancer growth, invasion, and metastasis. *Journal of the National Cancer Institute*, 102(12):866–880, 2010.

L. A. Higa, D. Banks, M. Wu, R. Kobayashi, H. Sun, and H. Zhang. L2dtl/cdt2 interacts with the cul4/ddb1 complex and pcna and regulates cdt1 proteolysis in response to dna damage. *Cell Cycle*, 5(15):1675–1680, 2006.

J. J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. *PNAS*, 79(8):2554–2558, 1982.

J. J. Hopfield. Neurons with graded response have collective computational properties like those of two-state neurons. *PNAS*, 81(10):3088–3092, 1984.

I. Horikawaa and J. C. Barrett. cdna cloning of the human polybromo-1 gene on chromosome 3p21. *DNA Sequence*, 13(4):211–215, 2002.

C.-J. Hu, L.-Y. Wang, L. A. Chodosh, B. Keith, and M. C. Simon. Differential roles of hypoxia-inducible factor 1α (hif-1α) and hif-2α in hypoxic gene regulation. *Molecular and Cellular Biology*, 23(24):9361–9374, 2003.

D.-Y. Huang and Z.-F. Chang. Interaction of human thymidine kinase 1 with p21waf1. *The Biochemical Journal*, 356:829–834, 2001.

Q. Huang, K. Gumireddy, M. Schrier, C. le Sage, R. Nagel, S. Nair, D. A. Egan, A. Li, G. Huang, A. J. Klein-Szanto, P. A. Gimotty, D. Katsaros, G. Coukos, L. Zhang, E. Puré, and R. Agami. The micrornas mir-373 and mir-520c promote tumour invasion and metastasis. *Nature Cell Biology*, 10(2):202–210, 2008.

S. Inoda, Y. Hirohashi, T. Torigoe, M. Nakatsugawa, K. Kiriyama, E. Nakazawa, K. Harada, H. Takasu, Y. Tamura, K. Kamiguchi, H. Asanuma, T. Tsuruma, T. Terui, K. Ishitani, T. Ohmura, Q. Wang, M. I. Greene, T. Hasegawa, K. Hirata, and N. Sato. Cep55/c10orf3, a tumor antigen derived from a centrosome residing protein in breast carcinoma. *Journal of Immunotherapy*, 32(5):474–485, 2009.

A. S. Ishkanian, C. A. Malloff, S. K. Watson, R. J. deLeeuw, B. Chi, B. P. Coe, A. Snijders, D. G. Albertson, D. Pinkel, M. A. Marra, V. Ling, C. MacAulay, and W. L. Lam. A tiling resolution dna microarray with complete coverage of the human genome. *Nature Genetics*, 36:299–303, 2004.

A. K. Jain, J. Mao, and K. Mohiuddin. Artificial neural networks: a tutorial. *Computer IEEE*, 29(3):31–44, 1996.

R. Jansen, H. Yu, D. Greenbaum, Y. Kluger, N. J. Krogan, S. Chung, A. Emili, M. Snyder, J. F. Greenblatt, and M. Gerstein. A bayesian networks approach for predicting protein-protein interactions from genomic data. *Science*, 302(5644):449–453, 2003.

D. Jupiter, H. Chen, and V. VanBuren. Starnet 2: a web-based tool for accelerating discovery of gene regulatory networks using microarray co-expression data. *BMC Bioinformatics*, 10:332, 2009.

G. Kallergi, H. Markomanolaki, V. Giannoukaraki, M. A. Papadaki, A. Strati, E. S. Lianidou, V. Georgoulias, D. Mavroudis, and S. Agelaki. Hypoxia-inducible factor-1? and vascular endothelial growth factor expression in circulating tumor cells of breast cancer patients. *Breast Cancer Research*, 11:R84, 2009.

L. N. Kanal. *International Encyclopedia of the Social & Behavioral Sciences*, chapter Perceptrons, pages 11218–11221. Elsevier Ltd., 2004.

M. Kanehisa and S. Goto. Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1):27–30, 2000.

A. V. Kapp, S. S. Jeffrey, A. Langerød, A.-L. Børresen-Dale, W. Han, D.-Y. Noh, I. R. Bukholm, M. Nicolau, P. O. Brown, and R. Tibshirani. Discovery and validation of breast cancer subtypes. *BMC Genomics*, 7: 231, 2006.

A. Kelemen and Y. Liang. Bayesian regularized neural network for multiple gene expression pattern classification. In *Proceedings of the International Joint Conference on Neural Networks*, pages 654–659, 2003.

J. Khan, J. S. Wei, M. Ringnér, L. H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C. R. Antonescu, C. Peterson, and P. S. Meltzer. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural network. *Nature Medicine*, 7(6):673–679, 2001.

E. J. Kim, S. Y. Lee, T. R. Kim, S. I. Choi, E. W. Cho, K. C. Kim, and I. G. Kim. Tspyl5 is involved in cell growth and the resistance to radiation in a549 cells via the regulation of p21waf1/cip1 and pten/akt pathway. *Biochemical and Biophysical Research Communications*, 392(3):448–453, 2010.

M. King, J. Marks, J. Mandell, and N. Y. B. C. S. Group. Breast and ovarian cancer risks due to inherited mutations in brca1 and brca2. *Science*, 302 (5645):643–646, 2003.

T. Kohonen. *Self-organization and associative memory.* Springer-Verlag New York, Inc., New York, NY, USA, 1989.

G. Krieger, H.-E. Wander, M. Kneba, M. P. t, G. Bandlow, and G. A. Nagel. Metastatic breast cancer with constantly low cea blood levels. a subgroup with unfavorable prognosis? *Journal of cancer Research and Clinical Oncology*, 108(3):341–344, 1984.

L. J. Lancashire. *Multi-Layer Perceptron Artificial Neural Network Predictive Modelling of Genomic and Mass Spectrometry Data in Bioinformatics.* PhD thesis, Nottingham Trent University, 2006.

L. J. Lancashire, S. Mian, I. O. Ellis, R. C. Rees, and G. R. Ball. Current developments in the analysis of proteomic data: Artificial neural network data mining techniques for the identification of proteomic biomarkers related to breast cancer. *Current Proteomics*, 2(1):15–29, 2005a.

L. J. Lancashire, O. Schmid, H. Shah, and G. R. Ball. Classification of bacterial species from proteomic data using combinatorial approaches incorporating artificial neural networks, cluster analysis and principal components analysis. *Bioinformatics*, 21:2191–2199, 2005b.

L. J. Lancashire, S. Ugurel, C. Creaser, D. Schadendorf, R. Rees, and G. Ball. Utilizing artificial neural networks to elucidate serum biomarker patterns which discriminate between clinical stages in melanoma. In *Computational Intelligence in Bioinformatics and Computational Biology*, 2005c.

L. J. Lancashire, R. C. Rees, and G. R. Ball. Identification of gene transcript signatures predictive for estrogen receptor and lymph node status using a stepwise forward selection artificial neural network modelling approach. *Artificial Intelligence in Medecine*, 43(2):99–111, June 2008.

L. J. Lancashire, C. Lemetre, and G. R. Ball. An introduction to artificial neural networks in bioinformatics–application to complex microarray and mass spectrometry datasets in cancer studies. *Briefings in Bioinformatics*, 10(3):315–329, 2009a.

L. J. Lancashire, D. G. Powe, J. S. Reis-Filho, E. Rakha, C. Lemetre, B. Weigelt, T. M. Abdel-Fatah, A. R. Green, R. Mukta, R. Blamey, E. C. Paish, R. C. Rees, I. O. Ellis, and G. R. Ball. A validated gene expression profile for detecting clinical outcome in breast cancer using artificial neural networks. *Breast Cancer Research and Treatment*, 120 (1):83–93, 2009b.

E. S. Lander. Array of hope. *Nature*, 21(Supplement):3–4, 1999.

D. P. Lane and L. V. Crawford. T antigen is bound to a host protein in sy40-transformed cells. *Nature*, 278:261–263, 1979.

H. Laverty, L. Wakefield, N. Occleston, S. O'Kane, and M. Ferguson. Tgf-beta3 and cancer: a review. *Cytokine & Growth Factor Reviews*, 20(4): 305–317, 2009.

W.-P. Lee and W.-S. Tzou. Computational methods for discovering gene networks from expression data. *Briefings in Bioinformatics*, 10(4):408–423, 2009.

K. Levenber. A method for the solution of certain non-linear problems in least squares. *Quarterly Journal of Applied Mathmatics*, 2(2):164–168, 1944.

A. J. Levine, C. A. Finlay, and P. W. Hinds. P53 is a tumor suppressor gene. *Cell*, 116(2):S67–S70, 2004.

W. K. Lim, K. Wang, C. Lefebvre, and A. Califano. Comparative analysis of microarray normalization procedures: effects on reverse engineering gene networks. *Bioinformatics*, 23(13):i282–288, 2007.

D. I. H. Linzer and A. J. Levine. Characterization of a 54k dalton cellular sv40 tumor antigen present in sv40-transformed cells and uninfected embryonal carcinoma cells. *Cell*, 17(1):43–52, 1979.

P. J. Lisboa and A. F. Taktak. The use of artificial neural networks in decision support in cancer: A systematic review. *Neural Networks*, 19(4): 408–415, 2006.

Y. Liu, N. Liu, and H. Zhao. Inferring protein-protein interactions through high-throughput interaction data from diverse organisms. *Bioinformatics*, 21(15):3279–3285, 2005a.

Z. Liu, D. Chen, and H. Bensmail. Gene expression data classification with kernel principal component analysis. *Journal of Biomedicine and Biotechnology*, 2005(2):155–159, 2005b.

A. J. Lowery, N. Miller, R. E. McNeill, and M. J. Kerin. Micrornas as prognostic indicators and therapeutic targets: potential effect on breast cancer management. *Clinical Cancer Research*, 14(2):306–365, 2008.

A. J. Lowery, N. Miller, A. Devaney, R. E. McNeill, P. A. Davoren, C. Lemetre, V. Benes, S. Schmidt, J. Blake, G. Ball, and M. J. Kerin. Microrna signatures predict oestrogen receptor, progesterone receptor and her2/neu receptor status in breast cancer. *Breast Cancer Research*, 11:R27, 2009.

D. J. C. MacKay. A practical bayesian framework for backpropagation networks. *Neural Computation*, 4(3):448–472, 1992.

D. J. C. MacKay. Probable networks and plausible predictions? a review of practical bayesian methods for supervised neural networks. *Network: Computation in Neural Systems*, 6:469–505, 1995.

P. Mangiameli, S. K. Chen, and D. West. A comparison of som neural network and hierarchical clustering methods. *European Journal of Operational Research*, 93(2):402–417, 1996.

D. W. Marquardt. An algorithm for the least-squares estimation of nonlinear parameters. *Journal of the Society for Industrial and Applied Mathematics*, 11(2):431–441, 1963.

B. Matharoo-Ball, G. Ball, and R. Rees. Clinical proteomics: Discovery of cancer biomarkers using mass spectrometry and bioinformatics approaches - a prostate cancer perspective. *Vaccine*, 25(2):B110–B121, 2007a.

B. Matharoo-Ball, C. Hughes, L. Lancashire, D. Tooth, G. Ball, C. Creaser, M. Elgasim, R. Rees, R. Layfield, and W. Atiomo. Characterization of biomarkers in polycystic ovary syndrome (pcos) using multiple distinct proteomic platforms. *Journal of Proteome Research*, 6(8):3321–3328, 2007b.

B. Matharoo-Ball, L. Ratcliffe, L. Lancashire, S. Ugurel, A. K. Miles, D. J. Weston, R. Rees, D. Schadendorf, G. Ball, and C. S. Creaser. Diagnostic biomarkers differentiating metastatic melanoma patients from healthy controls identified by an integrated maldi-tof mass spectrometry/bioinformatic approach. *Proteomics*, 1(6):605–620, 2007c.

M. D. Mattie, C. C. Benz, J. Bowers, K. Sensinger, L. Wong, G. K. Scott, V. Fedele, D. Ginzinger, R. Getts, and C. Haqq. Optimized high-throughput microrna expression profiling provides novel biomarker assessment of clinical prostate and breast cancer biopsies. *Molecular Cancer*, 5:24, 2006.

H. McAdams and L. Shapiro. Circuit simulation of genetic networks. *Science*, 269:650–656, 1995.

W. S. McCulloch and W. Pitts. A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biology*, 5(4):115–133, 1943.

S. Michiels, S. Koscielny, and C. Hill. Prediction of cancer outcome with microarrays: a multiple random validation strategy. *The Lancet*, 365 (9458):488–492, 2005.

Y. Miki, J. Swensen, D. Shattuck-Eidens, P. A. Futreal, K. Harshman, S. Tavtigian, Q. Liu, C. Cochran, L. M. Bennett, W. Ding, R. Bell, J. Rosenthal, C. Hussey, T. Tran, M. McClure, C. Frye, T. Hattier, R. Phelps, A. Haugen-Strano, H. Katcher, K. Yakumo, Z. Gholami, D. Shaffer, S. Stone, S. Bayer, C. Wray, R. Bogden, P. Dayananth, J. Ward, P. Tonin, S. Narod, P. K. Bristow, F. H. Norris, L. Helvering, P. Morrison, P. Rosteck, M. Lai, J. C. Barrett, C. Lewis, S. Neuhausen, L. Cannon-Albright, D. Goldgar, R. Wiseman, A. Kamb, and M. H. Skolnick. A strong candidate for the breast and ovarian cancer susceptibility gene brca1. *Science*, 266(5182):66–71, 1994.

L. D. Miller, J. Smeds, J. George, V. B. Vega, L. Vergara, A. Ploner, Y. Pawitan, P. Hall, S. Klaar, E. T. Liu, and J. Bergh. An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. *PNAS*, 102(38):13550–13555, 2005.

M. Minsky and S. Papert. *Perceptrons: An Introduction to Computational Geometry.* MIT Press, Cambridge, Mass. (USA), 1969.

C. J. Needham, I. W. Manfield, A. J. Bulpitt, P. M. Gilmartin, and D. R. Westhead. From gene expression to gene regulatory networks in arabidopsis thaliana. *BMC Systems Biology*, 3:85, 2009.

M. M. Nelson and W. Illingworth. *A practical guide to neural nets.* Addison-Wesley Publishing, Reading, MA (USA), 1991.

A. Nikitin, S. Egorov, N. Daraselia, and I. Mazo. Pathway studio?the analysis and navigation of molecular networks. *Bioinformatics*, 19(16):2155–2157, 2003.

E. E. Ntzani and J. P. Ioannidis. Predictive ability of dna microarrays for cancer outcomes and correlates: an empirical assessment. *The Lancet*, 362(9394):1439–1444, 2003.

S. L. O'Brien, A. Fagan, E. J. Fox, R. C. Millikan, A. C. Culhane, D. J. Brennan, A. H. McCann, S. Hegarty, S. Moyna, M. J. Duffy, D. G. Higgins, K. Jirström, G. Landberg, and W. M. Gallagher. Cenp-f expression is associated with poor prognosis and chromosomal instability in patients with primary breast cancer. *International Journal of Cancer*, 120(7): 1434–1443, 2007.

D. T. Odom, N. Zizlsperger, D. B. Gordon, G. W. Bell, N. J. Rinaldi, H. L. Murray, T. L. Volkert, J. Schreiber, P. A. Rolfe, D. K. Gifford, ErnestFraenkel, G. I. Bell, and R. A. Young. Control of pancreas and liver gene expression by hnf transcription factors. *Science*, 303(5662): 1378–1381, 2004.

K. L. O'Neill, M. Hoper, and G. W. Odling-Smee. Can thymidine kinase levels in breast tumors predict disease recurrence? *Journal of the National Cancer Institute*, 84(23):1825–1828, 1992.

R. C. Osthus, B. Karim, J. E. Prescott, B. D. Smith, M. McDevitt, D. L. Huso, and C. V. Dang. The myc target gene jpo1/cdca7 is frequently overexpressed in human tumors and has limited transforming activity in vivo. *Cancer Research*, 65(13):5620–5627, 2005.

N. Pandis, M. R. Teixeira, A.-M. Gerdes, J. Limon, G. Bardi, J. A. Andersen, I. Idvall, N. Mandahl, F. Mitelman, and S. Heim. Chromosome abnormalities in bilateral breast carcinomas. cytogenetic evaluation of the clonal origin of multiple primary tumors. *Cancer*, 76(2):250–258, 1995.

B.-W. Park, J.-W. Oh, J.-H. Kim, S. Park, K.-S. Kim, J. Kim, and K. Lee. Preoperative ca 15-3 and cea serum levels as predictor for breast cancer outcomes. *Annals of Oncology*, 19(4):675–681, 2008.

M. Payton, S. Scully, G. Chung, and S. Coats. Deregulation of cyclin e2 expression and associated kinase activity in primary breast tumors. *Oncogene*, 21(55):8529–8534, 2002.

C. M. Perou, T. Sørlie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, C. A. Rees, J. R. Pollack, D. T. Ross, H. Johnsen, L. A. Akslen, Øystein Fluge, A. Pergamenschikov, C. Williams, S. X. Zhu, P. E. Lønning, A.-L. Børresen-Dale, P. O. Brown, and D. Botstein. Molecular portraits of human breast tumours. *Nature*, 406:747–752, 2000.

D. Pham. Neural networks in engineering. In G. R. et al., editor, *Application of Artificial Intelligence in Engineering IX*, 1994.

J. Quackenbush. Computational analysis of microarray data. *Nature Reviews Genetics*, 2(6):418–427, 2001.

N. Ramachandran, E. Hainsworth, B. Bhullar, S. Eisenstein, B. Rosen, A. Y. Lau, J. C. Walter, and J. LaBaer. Self-assembling protein microarrays. *Science*, 305:86–89, 2004.

D. F. Ransohoff. Rules of evidence for cancer molecular-marker discovery and validation. *Nature Reviews Cancer*, 4:309–314, 2004.

S. Raychaudhuri, J. M. Stuart, and R. B. Altman. Principal components analysis to summarize microarray experiments: application to sporulation time series. In *Pacific Symposium on Biocomputing*, pages 455–466, 2000.

M. Riedmiller and H. Braun. A direct adaptive method for faster backpropagation learning: The rprop algorithm. In *IEEE International Conference on Neural Networks*, pages 586–591, 1993.

N. Rifai, M. A. Gillette, and S. A. Carr. Protein biomarker discovery and validation: the long and uncertain path to clinical utility. *Nature biotechnology*, 24(8):971–983, 2006.

D. M. Rocke, T. Ideker, O. Troyanskaya, J. Quackenbush, and J. Dopazo. Papers on normalization, variable selection, classification or clustering of microarray data. *Bioinformatics*, 25(6):701–702, 2009.

J. L. Rodgers and W. A. Nicewander. Thirteen ways to look at the correlation coefficient. *The American Statistician*, 42(1):59–66, 1988.

R. Rojas. *Neural networks: a systematic introduction.* Springer, Berlin (Germany), 1996.

F. Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386–408, 1958.

D. E. Rumelhart and J. L. McClelland. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition - Volume 1: Foundations.* The MIT Press, 1986.

D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *Nature*, 323:533–536, 1986.

C. Sanchez, C. Lachaize, F. Janody, B. Bellon, L. Röder, J. Euzenat, F. Rechenmann, and B. Jacq. Grasping at molecular interactions and genetic networks in drosophila melanogaster using flynets, an internet database. *Nucleic Acids Research*, 27(1):89?94, 1999.

M. Schena, D. S. aand Ronald W. Davis, and P. O. Brown. Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science*, 270(5235):467–470, 1995.

T. Schlitt and A. Brazma. Current approaches to gene regulatory network modelling. *BMC Bioinformatics*, 8(Suppl 6):S9, 2007.

B. Schwikowski, P. Uetz, and S. Fields. A network of protein-protein interactions in yeast. *Nature Biotechnology*, 18:1257–1261, 2000.

P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13:2498–2504, 2003.

B. A. Shoemaker and A. R. Panchenko. Deciphering protein-protein interactions. part ii. computational methods to predict protein and domain interaction partners. *PLoS Computational biology*, 3(4):595–601, 2007.

R. Simon, M. D. Radmacher, K. Dobbin, and L. M. McShane. Pitfalls in the use of dna microarray data for diagnostic and prognostic classification. *Journal of the National Cancer Institute*, 95(1):14–18, 2003.

E. Slodkowska and J. S. Ross. Mammaprint 70-gene signature: another milestone in personalized medical care for breast cancer patients. *Expert Review of Molecular Diagnostics*, 9(5):417–422, 2009.

E. M. Southern. Detection of specific sequences among dna fragments separated by gel electrophoresis. *Journal of Molecular Biology*, 98(3):503–517, 1975.

P. N. Span, J. Bussink, P. Manders, L. V. A. M. Beex, and C. G. J. Sweep. Carbonic anhydrase-9 expression levels and prognosis in human breast cancer: association with treatment outcome. *British Journal of Cancer*, 89:271–276, 2003.

V. Spirin and L. A. Mirny. Protein complexes and functional modules in molecular networks. *PNAS*, 100(21):12123–12128, 2003.

T. Sørlie. Introducing molecular subtyping of breast cancer into the clinic? *Journal of Clinical Oncology*, 27(8):1153–1154, 2009.

T. Sørlie, C. M. Perou, R. Tibshirani, T. Aas, S. Geisler, H. Johnsen, T. Hastie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, T. Thorsen, H. Quist, J. C. Matese, P. O. Brown, D. Botstein, P. E. Lønning, and A.-L. Børresen-Dale. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *PNAS*, 98(19):10869–10874, 2001.

T. Sørlie, R. Tibshirani, J. Parker, T. Hastie, J. S. Marron, A. Nobel, S. Deng, H. Johnsen, R. Pesich, S. Geisler, J. Demeter, C. M. Perou, P. E. Lønning, P. O. Brown, A.-L. Børresen-Dale, and D. Botstein. Repeated observation of breast tumor subtypes in independent gene expression data sets. *PNAS*, 100(14):8418–8423, 2003.

D. Stekel. *Microarray bioinformatics.* Cambrigde University Press, 2003.

U. Stelzl, U. Worm, M. Lalowski, C. Haenig, F. H. Brembeck, H. Goehler, M. Stroedicke, M. Zenkner, A. Schoenherr, S. Koeppen, J. Timm, S. Mintzlaff, C. Abraham, N. Bock, S. Kietzmann, A. Goedde, E. Toksöz, A. Droege, S. Krobitsch, B. Korn, W. Birchmeier, H. Lehrach, , and E. E. Wanker. A human protein-protein interaction network: a resource for annotating the proteome. *Cell*, 122(6):957–68, 2005.

N. Sugimoto, I. Kitabayashi, S. Osano, Y. Tatsumi, T. Yugawa, M. Narisawa-Saito, A. Matsukage, T. Kiyono, and M. Fujita. Identification of novel human cdt1-binding proteins by a proteomics approach: Proteolytic regulation by apc/ccdh1. *Molecular Biology of the Cell*, 19 (3):1007–1021, 2008.

H. Takahashi, K. Masuda, T. Ando, T. Kobayashi, and H. Honda. Prognostic predictor with multiple fuzzy neural models using expression profiles from dna microarray for metastases of breast cancer. *Journal of Bioscience and Bioengineering*, 98(3):193–199, 2004.

S. Takahashi, T. Moriya, T. Ishida, H. Shibata, H. Sasano, N. Ohuchi, and C. Ishioka. Prediction of breast cancer prognosis by gene expression profile of tp53 status. *Cancer Science*, 99(2):324–332, 2008.

P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E. S. Lander, and T. R. Golub. Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. *PNAS*, 96(6):2907–2912, 1999.

Y. Tan, L. Shi, W. Tong, and C. Wang. Multi-class cancer classification by total principal component regression (tpcr) using microarray gene expression data. *Nucleic Acids Research*, 33(1):56–65, 2005.

C. Tian, G. Xing, P. Xie, K. Lu, J. Nie, J. Wang, L. Li, M. Gao, L. Zhang, and F. He. Krab-type zinc-finger protein apak specifically regulates p53-dependent apoptosis. *Nature Cell Biology*, 11:580–591, 2009.

P. Törönen, M. Kolehmainen, G. Wong, and E. Castrén. Analysis of gene expression data using self-organizing maps. *FEBS Letters*, 451(2):142–146, 1999.

T. Ueki, T. Nishidate, J. H. Park, M. L. Lin, A. Shimo, K. Hirata, Y. Nakamura, and T. Katagiri. Involvement of elevated expression of multiple cell-cycle regulator, dtl/ramp (denticleless/ra-regulated nuclear matrix associated protein), in the growth of breast cancer cells. *Oncogene*, 27 (43):5672–83, 2008.

S. H. R. Vadrevu and S. U. Murty. A novel tool for classification of epidemiological data of vector-borne diseases. *Journal of Global Infectious Diseases*, 2(1):35–38, 2010.

M. J. van de Vijver, Y. D. He, L. J. van't Veer, H. Dai, A. A. M. Hart, D. W. Voskuil, G. J. Schreiber, J. L. Peterse, C. Roberts, M. J. Marton, M. Parrish, D. Atsma, A. Witteveen, A. Glas, L. Delahaye, T. van der Velde, H. Bartelink, S. Rodenhuis, E. T. Rutgers, S. H. Friend, and R. Bernards. A gene-expression signature as a predictor of survival in breast cancer. *The New England Journal of Medicine*, 347(25):1999–2009, 2002.

M. H. van Regenmortel. Reductionism and complexity in molecular biology. *EMBO Reports*, 5(11):1016?1020, 2004.

L. J. van't Veer, H. Dai, M. J. van de Vijver, Y. D. He, A. A. M. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards, and S. H. Friend. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415:530–536, 2002.

V. Vapnik and A. Lerner. Pattern recognition using generalized portrait method. *Automation and Remote Control*, 24(6):774–780, 1963.

P. P. Wadia, M. Coram, R. J. Armstrong, M. Mindrinos, A. J. Butte, and D. B. Miklos. Antibodies specifically target aml antigen nusap1 after allogeneic bone marrow transplantation. *Blood*, 115(10):2077–2087, 2010.

V. Wang, D. A. Davis, M. Haque, L. E. Huang, and R. Yarchoan. Differential gene up-regulation by hypoxia-inducible factor-1? and hypoxia-inducible factor-2? in hek293t cells. *Cancer Research*, 65:3299–3306, 2005.

J. Watkinson, X. Wang, T. Zheng, and D. Anastassiou. Identification of gene interactions associated with disease from gene expression data using synergy networks. *BMC Systems Biology*, 2:10, 2008.

J. D. Watson and F. H. C. Crick. Genetical implications of the structure of deoxyribonucleic acid. *Nature*, 171(4361):964–967, 1953.

A. Weber, I. Kristiansen, M. Johannsen, B. Oelrich, K. Scholmann, S. Gunia, M. May, H.-A. Meyer, S. Behnke, H. Moch, and G. Kristiansen. The fuse binding proteins fbp1 and fbp3 are potential c-myc regulators in renal, but not in prostate and bladder cancer. *BMC Cancer*, 8(369), 2008.

P. J. Werbos. *The roots of backpropagation.* Wiley, New York (USA), 1994.

M. West, C. Blanchette, H. Dressman, E. Huang, S. Ishida, R. Spang, H. Zuzan, J. John A. Olson, J. R. Marks, and J. R. Nevins. Predicting the clinical status of human breast cancer by using gene expression profiles. *PNAS*, 98(20):11462–11467, 2001.

S. J. Wicks, K. Haros, M. Maillard, L. Song, R. E. Cohen, P. ten Dijke, and A. Chantry. The deubiquitinating enzyme uch37 interacts with smads and regulates tgf-beta signalling. *Oncogene*, 24(54):8080–8084, 2005.

B. Widrow and M. E. Hoff. Adaptive switching circuits. In *IRE WESCON Convention Record*, volume 4, pages 96–104, 1960.

S. C. Winter, F. M. Buffa, P. Silva, C. Miller, H. R. Valentine, H. Turley, K. A. Shah, G. J. Cox, R. J. Corbridge, J. J. Homer, B. Musgrove, N. Slevin, P. Sloan, P. Price, C. M. West, and A. L. Harris. Relation of a hypoxia metagene derived from head and neck cancer to prognosis of multiple cancers. *Cancer Research*, 67:3441–3449, 2007.

K. Y. Won, G. Y. Kim, Y. W. Kim, J. Y. Song, and S.-J. Lim. Clinicopathologic correlation of beclin-1 and bcl-2 expression in human breast cancer. *Human pathology*, 41(1):107–112, 2010.

Q. Wu, P. Kirschmeier, T. Hockenberry, T.-Y. Yang, D. L. Brassard, L. Wang, T. McClanahan, S. Black, G. Rizzi, M. L. Musco, A. Mirza, and S. Liu. Transcriptional regulation during p21waf1/cip1-induced apoptosis in human ovarian cancer cells. *The Journal of Biological Chemistry*, 277 (39):36329–36337, 2002.

H. Xiong and X.-W. Chen. Kernel-based distance metric learning for microarray data classification. *BMC Bioinformatics*, 7:299–309, 2006.

K.-A. Yoon, Y. Nakamura, and H. Arakawa. Identification of aldh4 as a p53-inducible gene and its protective role in cellular stresses. *Journal of Human Genetics*, 49(3):134–140, 2004.

D. Zhang, L. K. Tai, L. L. Wong, L.-L. Chiu, S. K. Sethi, and E. S. C. Koay. Proteomic study reveals that proteins involved in metabolic and detoxification pathways are highly expressed in her-2/neu-positive breast cancer. *Molecular & Cellular Proteomics*, 4:1681–1696, 2005.

List of Figures

List of Tables

Communications

Peer-reviewed publications

1. Vafadar-Isfahani, B., Laversin, S.A.S., Ahmad, M., Ball, G., Coveney, C., **Lemetre, C.**, Miles, A.K., van Schalkwyk, G., Rees, R., Matharoo-Ball, B.
   Serum biomarkers which correlate with failure to respond to immunotherapy and tumor progression in a murine colorectal cancer model.
   Proteomics Clinical Applications 2010, **4**(8-9):682-696.

2. Lancashire, L.J., Powe, D.G., Reis-Filho, J.S., Rakha, E., **Lemetre, C.**, Weigelt, B., Abdel-Fatah, T.M., Green, A.R., Mukta, R., Blamey, R., Paish, E.C., Rees, R.C., Ellis, I.O., Ball, G.R.
   A validated gene expression profile for detecting clinical outcome in breast cancer using artificial neural networks.
   Breast Cancer Research Treatment 2010, **120**(1):83-93.

3. **Lemetre, C.**, Lancashire, L.J., Rees, R.C., Ball, G.R.
   Artificial Neural Network Based Algorithm for Biomolecular Interactions Modeling.
   J. Cabestany et al. (Eds.): *IWANN 2009, Part I, LNCS* **5517**:877-885. Springer-Verlag Berlin Heidelberg.

4. Lowery, A.J., Miller, N., Devaney, A., McNeill, R.E., Davoren, P.A., **Lemetre, C.**, Benes, V., Schmidt, S., Blake, J., Ball, G., Kerin, M.J.
   MicroRNA signatures predict oestrogen receptor, progesterone receptor and HER2/neu receptor status in breast cancer.
   Breast Cancer Research 2009, **11**(3):R27.

5. Lancashire, L.J., **Lemetre, C.**, Ball, G.R.
   An introduction to Artificial Neural Networks in Bioinformatics-Application to complex mixroarray and mass spectrometry datasets in cancer studies.
   Briefings in Bioinformatics 2009, **10**(3):315-329.

Oral presentations and conferences

- Cancer Bioinformatics Workshop
  Cambridge, UK *September 2010*

- British Breast Cancer Research Conference
  Nottingham, UK *September 2010*

- Nottingham Trent University seminar (final year talk)
  Nottingham, UK *June 2010*

- IWANN'09 (International Work-Conference on Artificial Neural Networks)
  Salamanca, Spain *June 2009*

- The Nottingham Trent University Annual Research School Conference
  Nottingham, UK *May 2008*

- ENACT Framework 6 Programme
  Nottingham, UK *May 2008*

- 6th East Midlands Proteomics Workshop
  Nottingham, UK *November 2007*

- ENACT Framework 6 Programme
  Stockholm, Sweden *September 2007*

Posters

2010
Lemetre, C., Lancashire L.J., Ball, G.R.
ANN Algorithm for interaction network inference in genomic array studies.
Cancer Bioinformatics Workshop Cambridge,UK

2009
Matharoo-Ball, B., **Lemetre, C.**, Minthon, L., Coveney, C., Vafadar-Isfahani, B., Boocock, D., Miles, A., Kalsheker, N., Rees, R.C., Janciauskiene, S.M., Ball, G.R., Morgan K.
Mass Spectrometric Techniques for the Identification of Differentially Expressed Proteins in Alzheimers Disease.
Proteomics Europe Barcelona, Spain

2008

Lemetre, C., Lancashire, L.J., Ward, S.J., Adams, V.C., Rees, R.C., Ball G.R.
Artificial Neural Network Based Algorithm For Biomolecular Interactions Modelling.
BSPR meeting Cambridge, UK

Adams, V.C., Ball, G.R., **Lemetre, C.**, Rees, R.C., Ward, S.J.
An immune screening platform to enrich for cancer vaccine responsive HRPCa patients.
ASCO-EORTC Annual Meeting on Molecular Markers in Cancer Chicago, USA

Matharoo-Ball, B., **Lemetre, C.**, Adams, V., Inderberg Suso, E.M., Ward, S.J., Gaudernack, G., Ball, G.R., Rees R.C.
Prostate Cancer Biomarkers to predict response to Immunotherapy.
7th HUPO Annual World Congress Amsterdam, Netherlands

Lemetre, C., Lancashire, L.J., Ward, S.J., Adams, V.C., Rees, R.C., Ball G.R.
Artificial Neural Network Based Algorithm For Biomolecular Interactions Modelling.
NTU Annual Research School Conference Nottingham, UK

Appendix I

Breast Cancer Res Treat
DOI 10.1007/s10549-009-0378-1

PRECLINICAL STUDY

A validated gene expression profile for detecting clinical outcome in breast cancer using artificial neural networks

L. J. Lancashire · D. G. Powe · J. S. Reis-Filho · E. Rakha · C. Lemetre · B. Weigelt · T. M. Abdel-Fatah · A. R. Green · R. Mukta · R. Blamey · E. C. Paish · R. C. Rees · I. O. Ellis · G. R. Ball

Received: 11 February 2009 / Accepted: 13 March 2009

Abstract Gene expression microarrays allow for the high throughput analysis of huge numbers of gene transcripts and this technology has been widely applied to the molecular and biological classification of cancer patients and in predicting clinical outcome. A potential handicap of such data intensive molecular technologies is the translation to clinical application in routine practice. In using an artificial neural network bioinformatic approach, we have reduced a 70 gene signature to just 9 genes capable of accurately predicting distant metastases in the original dataset. Upon validation in a follow-up cohort, this signature was an independent predictor of metastases free and overall survival in the presence of the 70 gene signature and other factors. Interestingly, the ANN signature and CA9 expression also split the groups defined by the 70 gene signature into prognostically distinct groups. Subsequently, the presence of protein for the principal prognosticator gene was categorically assessed in breast cancer tissue of an experimental and independent validation patient cohort, using immunohistochemistry. Importantly our principal prognosticator, CA9, showed that it is capable of selecting an aggressive subgroup of patients who are known to have poor prognosis.

Keywords Breast cancer · Prognosis · Bioinformatics · Survival · Hypoxia · Biomarkers

Abbreviations

ANN	Artificial neural networks
BCSS	Breast cancer specific survival
CA9	Carbonic anhydrase IX
EGF	Epidermal growth factor
DFI	Disease-free interval
EST	Expressed sequence tag

L. J. Lancashire and D. G. Powe contributed equally.

Electronic supplementary material The online version of this article (doi:10.1007/s10549-009-0378-1) contains supplementary material, which is available to authorized users.

L. J. Lancashire
Clinical and Experimental Pharmacology, Paterson
Institute for Cancer Research, University of Manchester,
Manchester M20 4BX, UK
e-mail: llancashire@picr.man.ac.uk

L. J. Lancashire · C. Lemetre · R. C. Rees · G. R. Ball (✉)
John Van Geest Cancer Research Centre, Nottingham Trent
University, Clifton Campus, Clifton Lane, Nottingham
NG11 8NS, UK
e-mail: graham.balls@ntu.ac.uk

D. G. Powe · E. Rakha · T. M. Abdel-Fatah ·
A. R. Green · R. Mukta · E. C. Paish · I. O. Ellis (✉)
Department of Histopathology, Nottingham University Hospitals
Trust and University of Nottingham, Nottingham NG7 2UH, UK
e-mail: ian.ellis@nottingham.ac.uk

J. S. Reis-Filho
The Breakthrough Breast Cancer Research Centre, Institute of
Cancer Research, Chester Beatty Laboratories, 237 Fulham
Road, London SW3 6JB, UK

R. Blamey
Department of Surgery, Breast Institute, City Hospital
Nottingham, Nottingham NG5 1PB, UK

B. Weigelt
Signal Transduction Laboratory, London Research Institute,
Lincoln's Inn Fields Laboratories, 44 Lincoln's Inn Fields,
London WC2A 3PX, UK

Published online: 04 April 2009

HR	Hormonal receptors
HIF-1α	Hypoxia induced factor 1 alpha
ROC	Receiver operating characteristic
RMH	Royal marsden hospital
TMA	Tissue microarray
TNP	Triple negative phenotype
AUC	Area under the curve

Introduction

Breast cancer is a heterogeneous disease where the outcome and response to therapy is often uncertain due to the complex network of overlapping and interacting molecular pathways. New strategies are needed to maximise therapeutic outcomes while limiting unnecessary over-treatment, achievable through customised treatment regimens. Previous studies have shown the ability of microarrays [1] to successfully predict clinical outcome in a variety of malignancies [2–6]. In particular, the molecular classification of malignant breast tumours using high throughput technologies including expression arrays and immunohistochemistry screening on tissue microarrays (TMAs), has successfully identified a number of biologically relevant subgroups [5, 7–10], showing good association between group membership and prediction of clinical outcome, targeted treatment and sensitivity to therapeutics [11–13]. Although there has been little overlap between different studies, more recent meta-analyses have demonstrated that different signatures identify similar groups of patients who have tumours with high proliferation rates [14, 15]. However, these meta-analyses have also demonstrated that most signatures reported to date have a relatively poor discriminatory power in oestrogen receptor negative disease [15]. Determining an optimal subset of predictive markers from microarray data is daunting due to the number of potential biomarker combinations present in these complex datasets. As an example, the seminal gene expression array data of van’t Veer et al. [13] comprised in excess of 24,000 variables (gene transcripts) per sample. More recent generations of gene chip now contain in excess of one million variables, further highlighting the requirements for robust computational analysis methods and emphasising the difficulties in translating these results to routine clinical practice.

Given the obvious advantages of analysing high density microarrays offering large (or even complete) genome coverage, powerful approaches are required for determining prognostic gene subsets in breast cancer. One such approach utilises Artificial Neural Networks (ANNs) to assess the prognostic potential of each gene transcript individually in a univariate procedure, and then adding further genes in a sequential, multivariate manner to improve upon the classification accuracy [16]. ANNs are a form of artificial intelligence inspired by learning in human neuronal systems and have been shown to be capable of modelling complex systems with high predictive accuracies on blind data [3, 17–19]. ANN models are developed by iteratively changing a network of weights, in response to predictive error. Predictions are made by mathematically modifying weights generated from input values (e.g. gene transcript intensity), in turn producing a predicted output value (for example, predicted survival). Moreover, the importance of the individual inputs in generating these predictions may be determined to define optimal subsets of biomarkers within the system being analysed. In a previous study [16], we developed a novel iterative stepwise approach to ANN modelling. In this study, we have applied ANN to van’t Veer’s dataset [13] to determine a minimal set of biomarkers required for the prediction of metastasis in patients with breast cancer. We identified a panel comprising just nine genes predicting tumour metastasis with 98% accuracy. The principal prognostic indicator had a prediction accuracy of 70% when used independently in the model and was found to be the hypoxia-associated enzyme carbonic anhydrase IX (CA9). The prognostic gene panel was validated on a second gene expression dataset consisting of 295 cases [20], with CA9 expression displaying an accuracy of 63% in predicting the development of metastasis in a categorical yes/no fashion. This increased to 66% when the remaining genes in the signature were included and was shown to be an independent predictor of both overall survival and metastasis free survival in this second cohort. Consequently, we investigated the immunohistochemical protein expression of CA9 as a prognostic and predictive indicator in an independent patient TMA containing 552 unselected breast cancers, and in 390 full-face breast excision tumour blocks comprising an experimental and validation cohort of 160 and 230 patients, respectively.

Materials and methods

ANN model development to identify a prognostic gene signature for metastasis

The ANN modelling used a supervised learning approach, multi-layer perceptron architecture with a sigmoidal transfer function, where weights were updated by a back propagation algorithm [21]. Learning rate and momentum were set at 0.1 and 0.5 respectively. The ANN architecture utilised five hidden nodes in the hidden layer and randomised initial weights. The output node was coded as 0 if the patient showed no evidence of metastasis within 5 years, and 1 if metastasis was evident. Data were downloaded in Microsoft Excel format from http://www.rii.com/publications/2002/vantveer.html. This initial set consisted of 78 samples each with 24,481 corresponding variables specifying the Log_{10} expression ratio of each gene. Prior to ANN training, the data was randomly divided into three subsets; 60% for training, 20% for validation (to assess model performance during the training process) and 20% for testing (to independently test the model on data completely blind to the model). This Monte–Carlo cross validation procedure [22] avoids over-fitting of the data, and has been shown to outperform and to be more consistent than the commonly used leave-one-out cross validation [23, 24], which may be a poor candidate for estimating the prediction error [25].

The forward stepwise approach to biomarker identification using ANNs has been previously described in detail (for specific details the reader is referred to [16]). This method develops a predictive model containing a parsimonious gene expression signature accurately classifying the cases according to the development of metastasis. Receiver Operating Characteristic (ROC) curves were generated to provide statistics regarding the sensitivity, specificity and area under the curve (AUC) of the model.

Patient selection and TMA preparation

Six paraffin processed TMA blocks containing 555 consecutive primary operable invasive breast carcinomas from patients involved in the Nottingham Tenovus Primary Breast Carcinoma Series between 1986 and 1993, were used as detailed previously [10]. The TMA construction involved sampling donor tissue cores from the tumour periphery and avoiding regions of obvious necrosis. In addition, 160 full face paraffin blocks of breast cancer were selected for comparison because of observed heterogeneity of CA9 distribution using immunohistochemistry. All cases used in this study are well characterised and have data on tissue protein expression for tumour-relevant biomarkers, comprehensive pathology and long term clinical follow-up data [10] including information on local, regional and distant tumour recurrence, and survival outcome. Patients with ER positive tumours were treated with adjuvant endocrine therapy whereas patients with a moderate and poor Nottingham Prognostic Index received chemotherapy.

CA9 protein expression was further validated on a cohort of 245 patients diagnosed and managed at the Royal Marsden Hospital (RMH) between 1994 and 2000. Patients were selected on the basis of being eligible for therapeutic surgery, being followed up at the RMH, having representative histological blocks in the RMH pathology files, and receiving standard anthracycline-based adjuvant chemotherapy. All patients were primarily treated with therapeutic surgery followed by anthracycline-based chemotherapy. Adjuvant endocrine therapy was prescribed for patients with ER positive tumours (tamoxifen alone in 96.4% of the patients for the available follow-up period). Complete follow-up was available for 244 patients, ranging from 0.5 to 125 months (median = 67 months, mean = 67 months). Tumours were graded according to a modified Bloom–Richardson scoring system [26] and size was categorised according to the TNM staging criteria. The project was approved by the Ethics and R&D committees at NUH and RMH.

CA9 immunohistochemistry and morphometry

Four micron thick paraffin-processed TMA and full face sections were subjected to microwave antigen retrieval in citrate buffer (pH 6.0), and then immunohistochemically stained with an antibody against CA9 on a TechMate immunostainer (DakoCytomation, Cambridge, UK). The CA9 rabbit polyclonal antibody (Abcam 15086, Cambridge, UK) was used at an optimised working dilution of 1:2,500 with a labelled streptavidin biotin (LSAB) technique. Sections were counterstained in haematoxylin and mounted using DPX mounting medium. Negative control sections had non-immune serum substituted for the primary antibody and positive control sections comprising high-grade ovarian cancer with necrotic foci were included in each immunohistochemistry run.

The immunohistochemically stained TMA and full face sections were scored with observers blinded to the clinicopathological features of tumours and patients' outcome. Staining was assessed in the cell membrane of morphologically unequivocal neoplastic cells of tumours and in stromal fibroblasts. The presence of CA9 staining in stromal fibroblasts was recorded because it has previously been suggested to be of prognostic significance [27]. Presence of tumour membrane and fibroblast CA9 staining was recorded '1' for affirmative and '0' for negative. Damaged tissue cores and those that did not contain invasive carcinoma were excluded from scoring.

Univariate and multivariate statistics

The Chi square test was used for testing the association between CA9 protein expression and other biomarkers scored as categorical variables, to produce contingency tables (Version 15, SPSS Inc., IL, USA). Similarly, the presence or absence of tumour-associated membranous and normal stromal cell cytoplasmic CA9 staining was categorically scored as positive or negative, regardless of its extent or staining intensity. Kaplan–Meier survival plots were produced to estimate disease-free interval (DFI), breast cancer specific survival (BCSS) and the development time for metastasis formation. DFI was expressed as the number of months from diagnosis to the occurrence of invasive local recurrence, local LN relapse or distant relapse. Survival rates were compared using the log rank (Mantel–Cox) test. A P-value of less than 0.05 was deemed significant with 95% confidence intervals.

Results

Development of a signature to predict development of distant metastasis using ANNs

ANN analysis identified a gene expression signature consisting of nine genes which predicted patient prognosis with 98% sensitivity and 94% specificity, with an AUC of 0.971 when assessed by ROC curve analysis. The overall screening process assessed over eleven million individual models. A summary of performance for the models at each step is shown in Table 1 and Supplementary Fig. 1. To further validate the model, an additional set of 19 samples were downloaded from the same location as the first series and used as a second order validation set, as in the original manuscript [13]. This set consisted of 7 patients who remained metastasis free, and 12 who developed metastases within 5 years. The novel nine gene expression signature correctly diagnosed all 19 samples, further emphasising the models predictive power. The response curves for these genes were also analysed, with seven of the nine having strong discriminatory responses (Supplementary Fig. 2 shows the response curve for CA9. The association between increased expression and development of metastases is clearly seen).

As seen in Table 1, four of the nine genes showed a positive association between increased expression and the probability of developing distant metastases, as output by the model. Of those four genes, CA9 gave the highest accuracy (70%) for predicting metastases. On the contrary, three genes showed an inverse association between increased expression and the predicted likelihood of metastases. In addition, two genes showed a weak response in the predicted probability of developing metastases, possibly modulating the responses of other genes in an additive fashion.

Validation of ANN findings

Since the ANN gene signature was capable of predicting the development of metastases to a high degree, the expression of these genes were further explored and validated using the NKI295 dataset [20] which includes gene expression data for a 295 patient cohort. Using the ANN 9 gene signature to classify this series of cases into two groups showed a significantly reduced overall survival ($P < 0.001$) and metastasis free survival ($P < 0.001$) between groups in univariate Kaplan–Meier analysis (Supplementary Fig. 3). Interestingly, the ANN signature was also able to split the groups defined by the original 70 gene signature into prognostically distinct groups ($P < 0.001$). In a multivariate Cox regression model adjusted for age, nodal status, tumour size, ER status, therapy type (chemotherapy or hormonal) and van't Veer's 70 gene signature, the ANN signature was shown to be an independent predictor of metastasis free survival ($P = 0.003$, Hazard ratio = 1.92) and overall survival ($P = 0.012$, Hazard ratio = 1.89) in this larger cohort (Supplementary Table 1a, b). Furthermore, analysis of CA9 gene expression in the NKI295 dataset showed a significant positive association with tumours of a basal-like phenotype ($P < 0.001$) and an inverse association with luminal type cancers ($P < 0.001$). These findings led us to investigate if

Table 1 Summary of the nine genes used in the gene expression signature at each step of model development

Step	Input added	Gene name	Description	Cumulative accuracy (%)	Error	Response
1	NM_001216	CA9	Carbonic anhydrase IX	70	0.44	Positive
2	Contig52778_RC		EST	80	0.38	Weak
3	Contig35076_RC		EST	83	0.38	Negative
4	Contig40557_RC	FLJ13409	EST	87	0.35	Positive
5	AB032973	LCHN	LCHN protein	80	0.40	Positive
6	AB004064	TMEFF2	Transmembrane protein with EGF-like and two follistatin-like domains 2	95	0.23	Positive
7	NM_006101	HEC/KNTC2	Kinetochore associated 2	95	0.22	Weak
8	AF161451	HSPC333	HSPC337	96	0.17	Negative
9	Contig33475		EST's	98	0.15	Weak

Table details the identity of the input added at each step, the gene name (where known) and description. The model accuracy and error when applied to the independent validation data splits are also shown, together with the direction of response of the gene as it correlated with metastases our gene expression findings could be translated into a routine immunohistochemistry practice for the principal prognosticator CA9.

CA9 protein distribution in breast tumours within the Nottingham patient cohort

CA9 staining was heterogeneously distributed in the cell membrane of tumour tissue and in the cytoplasm of stromal fibroblast. CA9 staining of tumours was predominantly associated with necrotic glandular foci (Fig. 1a, b) but in contrast, positively stained fibroblasts did not always show close association with necrotic malignant tissue. In TMAs, 552 cores were readable but comparison with the full face sections showed lack of concordance (Supplementary Table 1). Membranous CA9 expression was under-represented in TMAs due to heterogeneity in CA9 localisation and because of avoidance of necrotic regions during TMA construction. Membranous CA9 expression was identified in 26/552 (4.7%) TMA cores compared with 29/160 (18.1%) full face cases. For these reasons, only data from the full face sections was used.

Associations between CA9 expression and other clinicopathological variables

Membranous CA9 staining was significantly increased in younger patients with high histological grade cancers ($P < 0.001$; Table 2). However, membranous CA9 expression showed no significant association with menopausal

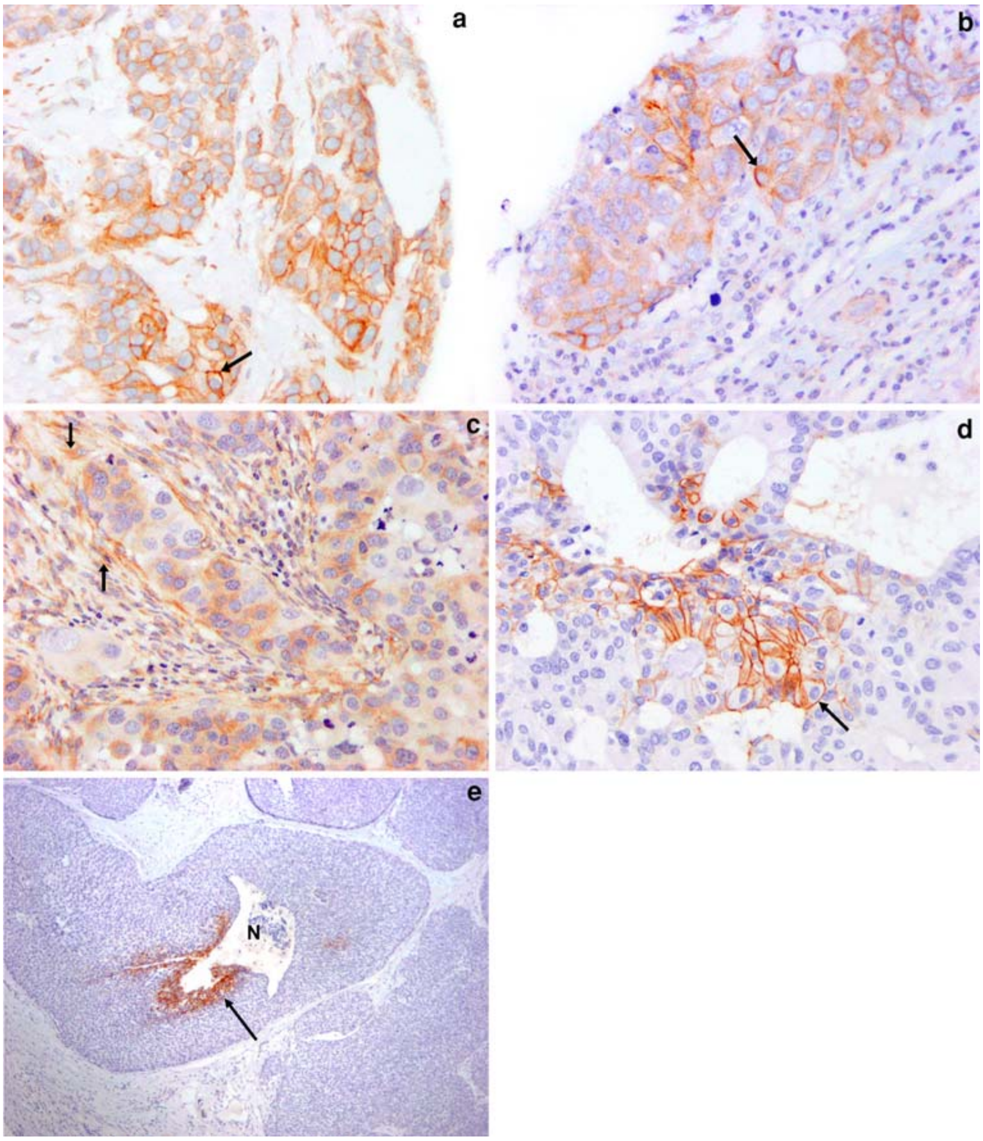


Fig. 1 CA IX immunostaining (*arrow*) was detected in breast tumour cell membrane (**a**), cytoplasm (**b**) and stroma (**c**) in TMA sections but its frequency was reduced due to its heterogeneous localisation. The latter is demonstrated in full face sections of ductal cancer (**d**), especially in cases showing glandular necrosis (*N*) (**e**) associated with hypoxia. Original magnification **a–d** 20×; **e** 4×

Table 2 Distribution of patients according to localisation of CA9 immunostaining by age, tumour grade and lymph node involvement

	Cytoplasmic staining	Membrane staining	Fibroblast staining
Age distribution			
Grade 1	51 (39–59)	44 (44)	55.6 (54–57)
Grade 2	54.5 (28–69)	53.5 (51–56)	49 (41–69)
Grade 3	47 (25–66)	48.8 (25–66)	51.1 (28–67)
CA9 distribution (%)	CAIX (+) CAIX (−)	CAIX (+) CAIX (−)	CAIX (+) CAIX (−)
Grade 1	8/89 (95.5) 20/71 (71.8)	1/29 (3.4) 26/131 (19.8)	3/26 (11.5) 24/134 (17.9)
Grade 2	26/89 (30.8) 31/71 (50)	3/29 (10.3) 54/131 (41.2)	7/26 (26.9) 49/134 (36.5)
Grade 3	55/89 (57.5) 20/71 (15)	25/29 (86.2) 49/131 (37.4)	16/26 (61.5) 55/134 (41)
Node involvement (%)			
Lymph node negative	85/89 (95.5) 51/71 (71.8)	28/29 (96.5) 106/131 (80.9)	25/26 (82) 105/134 (78.3)
Lymph node positive	4/89 (4.5) 20/71 (28.2)	1/29 (3.4) 23/131 (17.5)	1/26 (3.8) 23/134 (17.1)

status, tumour size, lymphovascular invasion ($P = 0.056$) or lymph node metastases ($P = 0.051$; Table 3a, b).

Membrane expressing tumours showed a strong negative association with the steroid hormonal receptors (HR) [ER, PgR and androgen receptor ($P < 0.001$ each)], and the luminal cytokeratin CK19 ($P = 0.015$). Importantly, tumours expressing membranous CA9 showed a triple negative phenotype (ER−, PgR−, HER2−) [28] and expressed basal-like markers [CK5/6 ($P = 0.001$), CK14 ($P = 0.02$), BRCA1 nuclear positivity ($P = 0.002$), p53 ($P = 0.001$) and P-cadherin ($P = 0.01$). No association with E-cadherin expression was seen (Table 3a, b).

Fibroblast expression

Stromal fibroblast CA9 staining was seen in 26 (16%) cases; 5 of them showed coexisting membranous and stromal cell CA9 expression (Table 2). Stromal expression showed no significant association with tumour size or menopausal status (Table 3c, d). CA9 expression showed a trend towards association with p53 ($P = 0.06$) and lymph node involvement ($P = 0.051$), but showed no significant associations with the other clinicopathological variables including HR, E-cadherin, HER2, CK56 or CK14 (Table 3c, d).

Survival analysis

No significant association between membranous CA9 immunohistochemical expression in cancer cells or stromal cells was observed with BCSS, DFI or local/regional recurrence.

CA9 protein expression in the validation patient cohort

CA9 protein expression was validated in a cohort of 245 patients, of which 230 could be evaluated for CA9 immunohistochemical expression. Membranous CA9 protein expression was present in 29 cases. Similar to the experimental patient group, the validation cohort showed a significant negative association with ER and PgR expression (Table 3a, b), and was significantly associated with triple negative basal-like tumours ($P < 0.001$). Similar to the Nottingham patient group, the validation group showed no significant association between CA9 expression in cancer cells or fibroblasts and other clinicopathological variables including tumour size, vascular invasion, or patients' outcome in terms of BCSS and DFI. However, CA9 staining in the validation group differed in showing a negative significant association with ER ($P = 0.033$), CK5/6 ($P = 0.01$), and CK14 ($P = 0.001$), and an absence of borderline association with lymph node involvement ($P = 0.268$; Table 3c, d).

Discussion

The aim of our study was to derive a minimal gene expression signature predictive of the outcome of breast cancer patients by applying an ANN approach to analyse a previously published dataset of breast cancer [13]. We hypothesised that this signature would be capable of predicting survival to at least the degree of accuracy obtained in the original study. Using an ANN approach developed specifically for the identification of optimal biomarker subsets in complex data, we found just nine genes were necessary to predict metastatic spread with sensitivity of 98%. This compares favourably with the computational approach used in the original manuscript [13] that resulted in the identification of a prognostic panel comprising 70 genes with a prediction accuracy of 83%. The principal prognostic indicator in our signature was identified as CA9, and this gene correctly predicted metastasis in 70% in the original cohort (van't Veer's) and in 63% of the validation

Table 3 Association between CA9 IHC protein expression with biological markers and clinical parameters assessed in full face sections of breast tumours, according to the cytoplasmic, membranous, or stromal staining pattern of localisation

Parameter	Experimental cohort				Validation cohort			
	Number of samples (%)		χ^2	*P* value	Number of samples (%)		χ^2	*P* value
	CA9(−)	CA9(+)			CA9(−)	CA9(+)		
a								
Tumour size								
Small	59 (88.1)	8 (11.9)	3.194	0.074	180 (86.5)	28 (13.5)	3.117	0.210
Large	70 (76.9)	21 (23.1)			12 (85.7)	2 (14.3)		
Menopausal status								
Premenopausal	41 (74.5)	14 (25.5)	4.567	0.102	–	–	–	–
Postmenopausal	72 (85.7)	12 (14.3)			–	–	–	–
ER								
Negative	24 (53.3)	21 (46.7)	27.196	<0.001	28 (66.6)	14 (33.3)	16.946	<0.001
Positive	87 (91.6)	8 (8.4)			161 (90.9)	16 (9.1)		
PgR								
Negative	38 (64.4)	21 (35.6)	13.746	<0.001	41 (74.5)	14 (25.5)	8.586	.003
Positive	73 (90.1)	8 (9.9)			148 (85)	16 (15)		
AR								
Negative	23 (56.1)	18 (43.9)	16.401	<0.001	–	–	–	–
Positive	79 (87.8)	11 (12.2)			–	–	–	–
P-cadherin								
Negative	53 (86.9)	8 (13.1)	4.110	0.043	–	–	–	–
Positive	56 (72.7)	21 (27.3)			–	–	–	–
E-cadherin								
Negative	47 (87.0)	7 (13.0)	2.984	0.084	60 (85.7)	10 (14.3)	0.055	0.973
Positive	66 (75.0)	22 (25.0)			111 (86.7)	17 (13.3)		
b								
c-erbb2								
Negative	96 (79.3)	25 (20.7)	0.283	0.413	161 (87.5)	23 (12.5)	2.003	0.157
Positive	21 (84.0)	4 (16.0)			25 (78.1)	7 (21.9)		
CK5/6								
Negative	111 (87.4)	16 (12.6)	27.806	<0.001	172 (89.5)	20 (10.5)	11.685	0.001
Positive	8 (38.1)	13 (61.9)			14 (63.6)	8 (36.4)		
CK14								
Negative	103 (83.1)	21 (16.9)	6.456	0.011	179 (89)	22 (11)	13.089	<0.001
Positive	11 (57.9)	8 (42.1)			12 (60)	8 (40)		
P53								
Negative	91 (87.5)	13 (12.5)	14.276	<0.001	130 (86.6)	20 (13.4)	0.207	0.649
Positive	23 (59.0)	16 (41.0)			48 (84.2)	9 (15.8)		
Vascular invasion								
Absent	87 (79.8)	22 (20.2)	1.353	0.508	61 (83.5)	12 (16.4)	0.762	0.408
Present	42 (87.5)	6 (12.5)			130 (87.8)	18 (12.2)		
Lymph node involvement								
Absent	106 (79.1)	28 (20.9)	3.801	0.051	68 (85)	12 (15)	0.481	0.531
Present	23 (95.8)	1 (4.2)			121 (88.3)	16 (11.7)		
Tumour recurrence								
Absent	112 (83.6)	22 (16.4)	2.208	0.137	–	–	–	–
Present	17 (70.8)	7 (29.2)			–	–	–	–
Overall survival	–	–	2.976	0.085	–	–	1.310	0.253
DFI	–	–	2.756	0.097	–	–	2.870	0.093

Table 3 continued

Parameter	Experimental cohort				Validation cohort			
	Number of samples (%)		χ^2	*P* value	Number of samples (%)		χ^2	*P* value
	CA9(−)	CA9(+)			CA9(−)	CA9(+)		
c								
Tumour size								
Small	58 (86.6)	9 (13.4)	0.827	0.363	170 (81.7)	38 (18.3)	3.703	0.157
Large	73 (81.1)	17 (18.9)			14 (100)	0 (0)		
Menopausal status								
Premenopausal	42 (76.4)	13 (23.6)	4.567	0.102	–	–	–	–
Postmenopausal	71 (85.5)	12 (14.5)			–	–	–	–
ER								
Negative	36 (80.0)	9 (20.0)	0.183	0.669	30 (71.4)	12 (28.6)	4.562	0.033
Positive	78 (114)	16 (17.0)			151 (85.3)	26 (14.7)		
PgR								
Negative	51 (86.4)	8 (13.6)	0.986	0.321	42 (23.2)	139 (76.8)	2.023	0.155
Positive	64 (80.0)	16 (20.0)			13 (34.2)	25 (65.8)		
AR								
Negative	34 (85.0)	6 (15.0)	0.288	0.592	–	–	–	–
Positive	73 (81.1)	17 (18.9)			–	–	–	–
P-cadherin								
Negative	54 (88.5)	7 (11.5)	1.258	0.262	–	–	–	–
Positive	62 (81.6)	14 (18.4)			–	–	–	–
E-cadherin								
Negative	23 (82.1)	5 (17.9)	0.029	0.865	163 (92.6)	13 (7.4)	1.692	0.429
Positive	96 (83.5)	19 (16.5)			106 (82.8)	22 (17.2)		
d								
c-erbb2								
Negative	9,100 (83.3)	20 (16.7)	0.161	0.688	151 (82)	33 (18)	0.100	0.751
Positive	20 (80.0)	5 (20.0)			27 (84.4)	5 (15.6)		
CK5/6								
Negative	104 (82.5)	22 (17.5)	0.129	0.72	164 (85.4)	28 (14.6)	6.692	0.010
Positive	18 (85.7)	3 (14.3)			14 (63.6)	8 (36.4)		
CK14								
Negative	101 (82.1)	22 (17.9)	1.932	0.165	173 (86)	28 (14)	16.621	<0.001
Positive	18 (94.7)	1 (5.3)			10 (50)	10 (50)		
P53								
Negative	90 (87.4)	13 (12.6)	3.533	0.060	129 (86)	21 (14)	5.570	0.025
Positive	29 (74.4)	10 (25.6)			41 (71.9)	16 (28.1)		
Vascular invasion								
Present	91 (83.5)	18 (16.5)	1.482	0.477	123 (83.1)	25 (16.9)	0.029	0.865
Absent	39 (83)	8 (17)			60 (82.2)	13 (17.8)		
Lymph node involvement								
Absent	106 (79.1)	28 (20.9)	3.801	0.051	63 (78.8)	17 (21.2)	1.226	0.268
Present	23 (95.8)	1 (4.2)			116 (88.5)	21 (11.5)		
Tumour recurrence								
Absent	113 (85.0)	20 (15.0)	1.460	0.227	–	–	–	–
Present	18 (75.0)	6 (25.0)			–	–	–	–
Overall survival	–	–	1.989	0.158	–	–	0.120	0.7280
DFI	–	–	1.431	0.232	–	–	0.700	0.4034

P values refer to χ^2 or log rank test for overall survival. Significance level = <0.05

cohort [20]. In this validation cohort, the ANN 9 gene signature was showed to be an independent predictor of both metastasis free and overall survival, and interestingly, was able to split the groups defined by the original 70 gene signature into prognostically distinct groups.

A further aim of our study was to investigate if our ANN-derived minimal gene panel for predicting poor prognosis in breast cancer could be successfully translated into routine practice. To test this, we studied the immunohistochemical localisation of the principle prognosticator CA9 in unselected breast cancer. Carbonic anhydrases are induced by hypoxia induced factor 1 alpha (HIF-1α) and assist cancer cells in avoiding death by neutralising acid pH conditions associated with hypoxia-induced glycolysis. Furthermore, it has been proposed that CA9 promotes tumour migration and invasion via its role in extracellular matrix degradation and through the induction of growth factors [29]. These important roles suggest that not only is CA9 a key candidate prognostic biomarker for determining clinical outcome, but because of its resistance to degradation, it could be a more robust marker of hypoxia than HIF-1α protein [30]. Previously, a number of studies have shown that over-expression of CA9 is functionally important in several tumour types including colorectal [31], cervical [32] and uterine [33] cancers, and sarcomas [34]. Although the contribution of CA9 as a prognostic marker in breast cancer has been obscured by conflicting reports, some authors [35] demonstrated that its expression is associated with tumours characterised by a basal-like phenotype and showing reduced patients' survival, emphasising the relationship between CA9 expression and poor prognosis.

In this study, we found membranous expression of CA9 is associated with tumours showing aggressive features including younger patients' age, high grade ductal cancers, basal-like phenotype (CK5/6+, CK14+; ER−, PgR−, HER2−) and BRCA1 positivity. Such patients showed a tendency towards reduced breast cancer specific survival and disease free interval even in the absence of lymph node involvement. It should be noted, however, that immunohistochemical expression of CA9 was not significantly associated with outcome of breast cancer patients.

Immunohistochemical assessment of CA9 was shown to be heterogeneously distributed and was frequently associated with regions showing necrotic foci. Donor tissue used in TMA construction specifically avoided necrotic regions resulting in under-representation of CA9 expression. For this reason, results of full face sections were considered in our study. Supporting our concern about the unsuitability of TMAs for studying CA9 expression, Brennan et al. [35] also identified a reduced frequency (11%) of membranous expression in TMAs when compared with larger samples of tumours.

In agreement with others [27, 36, 37], CA9 expression was identified in the cell membrane of tumour cells and in the cytoplasm of stromal fibroblast cells. The experimental and validation patient cohorts were concordant for membrane staining. In agreement with other studies [35, 38] our data provide further evidence that CA9 occurs in tumours with features of aggressive clinical behaviour, including loss of hormonal receptors, showing poor response to adjuvant endocrine therapy [38].Previously, it was reported that hypoxia can down-regulate ER expression via transcriptional nuclear factors and this might explain the observation seen in the current study [39]. In addition, hypoxia is reported to promote basal tumour-like features (ER−/HER2-negative, CK5-positive) due to up-regulation of SLUG gene expression [40]. Here, our data showed that 62% membrane CA9-expressing tumours significantly associate with the basal markers CK5/6 [41], and have a triple negative phenotype (TNP) [28], supporting the recent findings of Van den Eynden et al. [42]. More recently, it has been proposed that the use of five immunohistochemical markers (ER−, PgR−, HER2−, CK5/6+, EGFR+) can identify a basal subgroup with a worse prognosis (10 year BCSS, 62%) than that seen in TNP (10 year BCSS, 67%) [43]. We showed that 12/29 (41.3%) cases of membranous CA9 fall in the five marker subgroup and, similar to Nielsen et al. [43], we found no lymph node involvement despite their poor prognosis. In addition, 16/26 (61.5%) of the membrane CA9 group were positive for BRCA1 nuclear IHC positivity [44].

The biological significance of CA9 localization in fibroblasts is not readily understood but it has been proposed that it might be caused by the effect of HIF-1α induction factors in these cells due to reasons other than hypoxia [45]. Further work is required to explore the significance of fibroblast CA9 staining.

Other genes identified in our expression signature were more compatible with a tumour suppressor function, including *TMEFF2* and *HEC*. *TMEFF2* encodes for a transmembrane protein containing an epidermal growth factor (EGF)-like motif and two follistatin domains. Our data showed a negative correlation between *TMEFF2* expression and the development of distant metastases, supporting the study of Gery et al. [46] who showed that *TMEFF2* could suppress the growth of prostate cancer cells. More recently [47], it was proposed that *TMEFF2* suppression may contribute to the oncogenic properties of c-Myc, thereby promoting cell proliferation, differentiation, and apoptosis. *HEC* (also known as kinetochore-associated 2), was shown here to be associated with metastases with increased expression. Similar findings have been reported [48] where *HEC* was identified as part of an 11 gene signature predictive of disease recurrence and distant metastasis in prostate and breast cancer.

Furthermore, elevated *HEC* expression has been shown to be associated with poorer prognosis in non-small cell lung carcinomas [49], and therefore a potential target for treatment of cancers, highlighted further still by Gurzov and Izquierdo [50]. Four of the nine genes identified in our panel represent expressed sequence tags (EST's) and the associated gene is therefore of unknown function. However, given their predictive capability with regard to survival, further analysis is justified.

To conclude, using powerful ANN methodologies, we have identified a minimal gene signature that is predictive of outcome at least with a similar degree of accuracy to that obtained in van't Veer's study [13]. Interestingly, this gene signature was shown to have a similar accuracy in predicting the development of metastasis and to be an independent predictor of outcome (metastasis free and overall survival) in a larger validation series from the same group [21]. Moreover, using immunohistochemistry we confirmed its practical and translational application. In agreement with van't Veer et al. [13] we have shown that whilst single genes are capable of discriminating between different disease states, multiple genes in combination enhance the predictive power of these models. Our signature predicted the hypoxic marker CA9 as the principal indicator of poor clinical outcome and although assessment of CA9 protein expression showed no significant association with patients' outcome when compared with our prediction gene panelta, CA9 expression showed association with variables of poor prognosis and aggressive behaviour. In particular, CA9 is associated with basal-like and triple negative cancers. Further studies of all nine genes in combination using immunohistochemistry are warranted to assess the prognostic value of this signature in routine practice.

Acknowledgments This study was supported by a grant from the UK HEFCE and ENACT (The European Network for the identification and validation of antigens and biomarkers in cancer and their application in clinical tumour immunology) and the Breast Cancer Campaign (2005). The authors thank Dr. Kay Savage for production of the Royal Marsden tumour tissue sections. Thanks also to the John and Lucille Van Geest Foundation.

References

1. Brown PO, Botstein D (1999) Exploring the new world of the genome with DNA microarrays. Nat Genet 21((1)(Suppl)):33–37. doi:10.1038/4462
2. Bhattacharjee A, Richards WG, Staunton J et al (2001) Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. Proc Natl Acad Sci USA 98(24):13790–13795. doi:10.1073/pnas.191502998
3. Khan J, Wei JS, Ringner M et al (2001) Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat Med 7(6):673–679. doi:10.1038/89044
4. Rosenwald A, Wright G, Chan WC et al (2002) The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. N Engl J Med 346(25):1937–1947. doi:10.1056/NEJMoa012914
5. Sorlie T, Perou CM, Tibshirani R et al (2001) Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc Natl Acad Sci USA 98(19):10869–10874. doi:10.1073/pnas.191367098
6. West M, Blanchette C, Dressman H et al (2001) Predicting the clinical status of human breast cancer by using gene expression profiles. Proc Natl Acad Sci USA 98(20):11462–11467. doi:10.1073/pnas.201162998
7. Callagy G, Cattaneo E, Daigo Y et al (2003) Molecular classification of breast carcinomas using tissue microarrays. Diagn Mol Pathol 12(1):27–34. doi:10.1097/00019606-200303000-00004
8. Paik S, Shak S, Tang G et al (2004) A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. N Engl J Med 351(27):2817–2826. doi:10.1056/NEJMoa041588
9. Perou CM, Sorlie T, Eisen MB et al (2000) Molecular portraits of human breast tumours. Nature 406(6797):747–752. doi:10.1038/35021093
10. Abd El-Rehim DM, Ball G, Pinder SE et al (2005) High-throughput protein expression analysis using tissue microarray technology of a large well-characterised series identifies biologically distinct classes of breast cancer confirming recent cDNA expression analyses. Int J Cancer 116(3):340–350. doi:10.1002/ijc.21004
11. Masters JR, Lakhani SR (2000) How diagnosis with microarrays can help cancer patients. Nature 404(6781):921. doi:10.1038/35010139
12. Naderi A, Teschendorff AE, Barbosa-Morais NL et al (2007) A gene-expression signature to predict survival in breast cancer across independent data sets. Oncogene 26(10):1507–1516. doi:10.1038/sj.onc.1209920
13. van 't Veer LJ, Dai H, van de Vijver MJ et al (2002) Gene expression profiling predicts clinical outcome of breast cancer. Nature 415(6871):530–536. doi:10.1038/415530a
14. Fan C, Oh DS, Wessels L et al (2006) Concordance among gene-expression-based predictors for breast cancer. N Engl J Med 355(6):560–569. doi:10.1056/NEJMoa052933
15. Wirapati P, Sotiriou C, Kunkel S et al (2008) Meta-analysis of gene expression profiles in breast cancer: toward a unified understanding of breast cancer subtyping and prognosis signatures. Breast Cancer Res 10(4):R65. doi:10.1186/bcr2124
16. Lancashire LJ, Rees RC, Ball GR (2008) Identification of gene transcript signatures predictive for estrogen receptor and lymph node status using a stepwise forward selection artificial neural network modelling approach. Artif Intell Med 43(2):99–111. doi:10.1016/j.artmed.2008.03.001
17. Ball G, Mian S, Holding F et al (2002) An integrated approach utilizing artificial neural networks and SELDI mass spectrometry for the classification of human tumours and rapid identification of potential biomarkers. Bioinformatics 18(3):395–404. doi:10.1093/bioinformatics/18.3.395
18. Lancashire L, Schmid O, Shah H et al (2005) Classification of bacterial species from proteomic data using combinatorial approaches incorporating artificial neural networks, cluster analysis and principal components analysis. Bioinformatics 21(10):2191–2199. doi:10.1093/bioinformatics/bti368
19. Matharoo-Ball B, Ball G, Rees R (2007) Clinical proteomics: discovery of cancer biomarkers using mass spectrometry and bioinformatics approaches—a prostate cancer perspective. Vaccine 25(Suppl 2):B110–B121

20. van de Vijver MJ, He YD, van 't Veer LJ et al (2002) A gene-expression signature as a predictor of survival in breast cancer. N Engl J Med 347(25):1999–2009. doi:10.1056/NEJMoa021967
21. Rumelhart DE, Hinton GE, Williams RJ (1986) Learning representations by back-propagating errors. Nature 323:533–536
22. Picard RR, Cook RD (1984) Cross-validation of regression models. J Am Stat Assoc 79(387):575–583
23. Xu QS, Liang YZ, Du YP (2004) Monte Carlo cross-validation for selecting a model and estimating the prediction error in multivariate calibration. J Chemometr 18(2):112–120. doi:10.1002/cem.858
24. Shao J (1993) Linear model selection by cross-validation. J Am Stat Assoc 88(422):486–494. doi:10.2307/2290328
25. Efron B (1986) How biased is the apparent error rate of a prediction rule? J Am Stat Assoc 81(394):461–470. doi:10.2307/2289236
26. Bloom HJ, Richardson WW (1957) Histological grading and prognosis in breast cancer; a study of 1409 cases of which 359 have been followed for 15 years. Br J Cancer 11(3):359–377
27. Colpaert CG, Vermeulen PB, Fox SB et al (2004) The presence of a fibrotic focus in invasive breast carcinoma correlates with the expression of carbonic anhydrase IX and is a marker of hypoxia and poor prognosis. Breast Cancer Res Treat 81(2):137–147. doi:10.1023/A:1025702330207
28. Nielsen TO, Hsu FD, Jensen K et al (2004) Immunohistochemical and clinical characterization of the basal-like subtype of invasive breast carcinoma. Clin Cancer Res 10(16):5367–5374. doi:10.1158/1078-0432.CCR-04-0220
29. Raghunand N, He X, van Sluis R et al (1999) Enhancement of chemotherapy by manipulation of tumour pH. Br J Cancer 80(7):1005–1011. doi:10.1038/sj.bjc.6690455
30. van Berkel M, van der Groep P, Schvarts A et al (2002) HIF-1alpha/CAIX coexpression in invasive human breast cancer. Breast Cancer Res Treat 76(Suppl 1):S146
31. Saarnio J, Parkkila S, Parkkila AK et al (1998) Immunohistochemical study of colorectal tumors for expression of a novel transmembrane carbonic anhydrase, MN/CA IX, with potential value as a marker of cell proliferation. Am J Pathol 153(1):279–285
32. Liao SY, Brewer C, Zavada J et al (1994) Identification of the MN antigen as a diagnostic biomarker of cervical intraepithelial squamous and glandular neoplasia and cervical carcinomas. Am J Pathol 145(3):598–609
33. Hockel M, Schlenger K, Aral B et al (1996) Association between tumor hypoxia and malignant progression in advanced cancer of the uterine cervix. Cancer Res 56(19):4509–4515
34. Brizel DM, Scully SP, Harrelson JM et al (1996) Tumor oxygenation predicts for the likelihood of distant metastases in human soft tissue sarcoma. Cancer Res 56(5):941–943
35. Brennan DJ, Jirstrom K, Kronblad A et al (2006) CA IX is an independent prognostic marker in premenopausal breast cancer patients with one to three positive lymph nodes and a putative marker of radiation resistance. Clin Cancer Res 12(21):6421–6431. doi:10.1158/1078-0432.CCR-06-0480
36. Chia SK, Wykoff CC, Watson PH et al (2001) Prognostic significance of a novel hypoxia-regulated marker, carbonic anhydrase IX, in invasive breast carcinoma. J Clin Oncol 19(16):3660–3668
37. Tomes L, Emberley E, Niu Y et al (2003) Necrosis and hypoxia in invasive breast carcinoma. Breast Cancer Res Treat 81(1):61–69. doi:10.1023/A:1025476722493
38. Trastour C, Benizri E, Ettore F et al (2007) HIF-1alpha and CA IX staining in invasive breast carcinomas: prognosis and treatment outcome. Int J Cancer 120(7):1451–1458. doi:10.1002/ijc.22436
39. Kronblad A, Hedenfalk I, Nilsson E et al (2005) ERK1/2 inhibition increases antiestrogen treatment efficacy by interfering with hypoxia-induced downregulation of ERalpha: a combination therapy potentially targeting hypoxic and dormant tumor cells. Oncogene 24(45):6835–6841. doi:10.1038/sj.onc.1208830
40. Storci G, Sansone P, Trere D et al (2008) The basal-like breast carcinoma phenotype is regulated by SLUG gene expression. J Pathol 214(1):25–37. doi:10.1002/path.2254
41. Fadare O, Tavassoli FA (2007) The phenotypic spectrum of basal-like breast cancers: a critical appraisal. Adv Anat Pathol 14(5):358–373. doi:10.1097/PAP.0b013e31814b26fe
42. Van den Eynden GG, Smid M, Van Laere SJ et al (2008) Gene expression profiles associated with the presence of a fibrotic focus and the growth pattern in lymph node-negative breast cancer. Clin Cancer Res 14(10):2944–2952. doi:10.1158/1078-0432.CCR-07-4397
43. Cheang MC, Voduc D, Bajdik C et al (2008) Basal-like breast cancer defined by five biomarkers has superior prognostic value than triple-negative phenotype. Clin Cancer Res 14(5):1368–1376. doi:10.1158/1078-0432.CCR-07-1658
44. Turner NC, Reis-Filho JS (2006) Basal-like breast cancer and the BRCA1 phenotype. Oncogene 25(43):5846–5853. doi:10.1038/sj.onc.1209876
45. Vleugel MM, Greijer AE, Shvarts A et al (2005) Differential prognostic impact of hypoxia induced and diffuse HIF-1alpha expression in invasive breast cancer. J Clin Pathol 58(2):172–177. doi:10.1136/jcp.2004.019885
46. Gery S, Sawyers CL, Agus DB et al (2002) TMEFF2 is an androgen-regulated gene exhibiting antiproliferative effects in prostate cancer cells. Oncogene 21(31):4739–4746. doi:10.1038/sj.onc.1205142
47. Gery S, Koeffler HP (2003) Repression of the TMEFF2 promoter by c-Myc. J Mol Biol 328(5):977–983. doi:10.1016/S0022-2836(03)00404-2
48. Glinsky GV, Berezovska O, Glinskii AB (2005) Microarray analysis identifies a death-from-cancer signature predicting therapy failure in patients with multiple types of cancer. J Clin Invest 115(6):1503–1521. doi:10.1172/JCI23412
49. Hayama S, Daigo Y, Kato T et al (2006) Activation of CDCA1-KNTC2, members of centromere protein complex, involved in pulmonary carcinogenesis. Cancer Res 66(21):10339–10348. doi:10.1158/0008-5472.CAN-06-2137
50. Gurzov EN, Izquierdo M (2006) RNA interference against Hec1 inhibits tumor growth in vivo. Gene Ther 13(1):1–7. doi:10.1038/sj.gt.3302595

Appendix II

Research article

Open Access

MicroRNA signatures predict oestrogen receptor, progesterone receptor and HER2/*neu* receptor status in breast cancer

Aoife J Lowery[1], Nicola Miller[1], Amanda Devaney[1], Roisin E McNeill[1], Pamela A Davoren[1], Christophe Lemetre[2], Vladimir Benes[3], Sabine Schmidt[3], Jonathon Blake[3], Graham Ball[2] and Michael J Kerin[1]

[1]Department of Surgery, Clinical Science Institute, University Hospital/National University of Ireland Galway, Galway, Ireland
[2]John Van Geest Cancer Research Centre, School of Science & Technology, Nottingham Trent University, Clifton Campus, Clifton Lane, Nottingham NG11 8NS, UK
[3]European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany

Corresponding author: Nicola Miller, nicola.miller@nuigalway.ie

Received: 13 Nov 2008 Revisions requested: 19 Dec 2008 Revisions received: 25 Mar 2009 Accepted: 11 May 2009 Published: 11 May 2009

Breast Cancer Research 2009, **11**:R27 (doi:10.1186/bcr2257)
This article is online at: http://breast-cancer-research.com/content/11/3/R27

Abstract

Introduction Breast cancer is a heterogeneous disease encompassing a number of phenotypically diverse tumours. Expression levels of the oestrogen, progesterone and HER2/*neu* receptors which characterize clinically distinct breast tumours have been shown to change during disease progression and in response to systemic therapies. Mi(cro)RNAs play critical roles in diverse biological processes and are aberrantly expressed in several human neoplasms including breast cancer, where they function as regulators of tumour behaviour and progression. The aims of this study were to identify miRNA signatures that accurately predict the oestrogen receptor (ER), progesterone receptor (PR) and HER2/*neu* receptor status of breast cancer patients to provide insight into the regulation of breast cancer phenotypes and progression.

Methods Expression profiling of 453 miRNAs was performed in 29 early-stage breast cancer specimens. miRNA signatures associated with ER, PR and HER2/*neu* status were generated using artificial neural networks (ANN), and expression of specific miRNAs was validated using RQ-PCR.

Results Stepwise ANN analysis identified predictive miRNA signatures corresponding with oestrogen (*miR-342*, *miR-299*, *miR-217*, *miR-190*, *miR-135b*, *miR-218*), progesterone (*miR-520g*, *miR-377*, *miR-527-518a*, *miR-520f-520c*) and HER2/*neu* (*miR-520d*, *miR-181c*, *miR-302c*, *miR-376b*, *miR-30e*) receptor status. *MiR-342* and *miR-520g* expression was further analysed in 95 breast tumours. *MiR-342* expression was highest in ER and HER2/*neu*-positive luminal B tumours and lowest in triple-negative tumours. *MiR-520g* expression was elevated in ER and PR-negative tumours.

Conclusions This study demonstrates that ANN analysis reliably identifies biologically relevant miRNAs associated with specific breast cancer phenotypes. The association of specific miRNAs with ER, PR and HER2/*neu* status indicates a role for these miRNAs in disease classification of breast cancer. Decreased expression of *miR-342* in the therapeutically challenging triple-negative breast tumours, increased *miR-342* expression in the luminal B tumours, and downregulated *miR-520g* in ER and PR-positive tumours indicates that not only is dysregulated miRNA expression a marker for poorer prognosis breast cancer, but that it could also present an attractive target for therapeutic intervention.

Introduction

Breast cancer is a heterogeneous disease that encompasses a range of phenotypically distinct tumour types. Underlying this heterogeneity is a spectrum of molecular alterations and initiating events that manifest clinically through a diversity of disease presentations and outcomes. Novel therapeutic strategies are increasingly being investigated and implemented, but unpredictable response and the development of

ABCG2: ATP-binding cassette sub-family G member 2; ANN: artificial neural network; BCRP: breast cancer resistance protein; bp: base pairs; ΔΔCt: comparative cycle threshold; *E*: PCR amplification efficiencies; ER: oestrogen receptor; HER2/*neu*: v-erb-b2 erythroblastic leukaemia viral oncogene homolog 2 receptors; miRNA: microRNA; PR: progesterone receptor; RQ-PCR: real-time quantitative polymerase chain reaction; RT: reverse transcriptase; UTR: untranslated region.

resistance to adjuvant therapy remain major challenges in the clinical management of breast cancer patients.

The key to optimizing and targeting therapy lies in a more complete understanding of the complex molecular interactions that underlie breast cancer and contribute to its heterogeneous nature. Breast-cancer-related genes have been extensively investigated, largely through the development of high-throughput array-based gene expression profiling platforms. The substantial datasets that have ensued have enabled us to decipher in depth some of the molecular intricacies associated with breast cancer, and have expanded our knowledge of the genetic pathways associated with breast carcinogenesis, resulting in classification systems predictive of outcome [1,2].

Breast tumours can now be classified into major subtypes on the basis of gene expression – luminal, v-erb-b2 erythroblastic leukaemia viral oncogene homolog 2 receptors (HER2/*neu*) overexpressing and basal like – and further analysis has identified additional subtypes within the original subgroups [3]. The expression of specific genes such as the oestrogen receptors (ERs) and HER2/*neu* are indicative of outcome in breast cancer patients, and the clinically relevant subgroupings are based broadly on ER/progesterone receptor (PR)/HER2/*neu* status. The ability to classify breast cancers in this manner has obvious beneficial implications for the development of targeted therapies; multigene prognostic and predictive tests have been developed, have been commercialized and have become established as tools in breast cancer diagnostics [4], although as yet there is little knowledge regarding the precise regulation of these genes and receptors.

MicroRNAs (miRNAs) are short (~22 bp), single-stranded, noncoding RNAs that have recently been recognized as a highly abundant class of regulatory molecules. They are thought to regulate up to one-third of the human genome via sequence-specific regulation of post-transcriptional gene expression by targeting mRNAs for cleavage or translational repression [5]. miRNAs have recently been identified as key players in cellular processes including self-renewal, differentiation, growth and death [6], all of which are dysregulated in carcinogenesis. There is increasing evidence to suggest that miRNAs may be responsible for a large proportion of breast cancer heterogeneity. A number of miRNAs have been shown to be dysregulated in breast cancer [7-10], and specific miRNAs functioning as regulators of tumorigenicity, invasion and metastasis have been identified [11-14]. Furthermore, miRNA regulation of ER and HER2/*neu*, known to be of prognostic significance in breast cancer, has been demonstrated [15,16]. As each miRNA can target up to 200 mRNA sequences, and mRNAs can have multiple miRNA target sites [5], it is probable that further miRNA regulators of these genes remain to be determined.

Expression profiling of miRNA to classify breast tumours according to clinicopathological variables currently used to predict disease progression is of particular interest. Firstly, profiling highlights the potential to identify novel prognostic indicators, which may contribute to improved selection of patients for adjuvant therapy. This approach has already shown promise with genomic signatures [2], and miRNA profiles appear to have superior accuracy to mRNA profiling [17]. Furthermore, the identification of miRNAs with regulatory roles in clinically distinct breast tumour samples could identify novel targets for therapeutic manipulation.

Despite its apparent clinical application, microarray technology remains deficient with regard to its translation into routine clinical practice. There has been little overlap between the breast cancer gene sets, leading to questions regarding their biological significance and reproducibility [18]. Array technology is highly dependent on bioinformatics, mathematics and statistics to produce biologically relevant results. The generation of high-complexity microarray data has necessitated the development of novel data analysis methodologies that can cope with data of this nonlinear and highly dimensional nature. Current conventional methods such as hierarchical clustering have shown limitations for the modelling and analysis of high-dimensionality data [19].

Artificial neural networks (ANNs) are a form of artificial intelligence that can learn to predict, through modelling, answers to particular questions in complex data. The models produced by ANNs have been shown to have the ability to predict well for unseen data and have the ability to cope with complexity and nonlinearity within the dataset [20,21]; these features of ANNs means they have the potential to identify and model patterns in this type of data to address a particular question. ANNs are therefore able to determine patterns or features (for example, in genes or proteins) within a dataset that can discriminate between subgroups of a clinical population (for example, disease and control), or disease grades [22]. Indeed, this discrimination has been previously demonstrated in different tumour types [22,23]. These patterns can combine into a fingerprint that can accurately predict the subgroups.

Our aims in the present study were to identify miRNA signatures using ANNs that accurately predict the ER, PR and HER2/*neu* status of breast cancer patients, thus identifying potential biologically relevant miRNAs and providing further insight into breast cancer aetiology and regulation.

Materials and methods

Patients and samples

Breast tumour specimens were obtained from patients during primary curative resection at Galway University Hospital, Galway, Ireland. Matched tumour-associated normal breast tissue was also obtained from a subset of these patients where possible. Following excision, tissue samples were immediately snap-frozen in liquid nitrogen and stored at -80°C until RNA extraction. Prior written and informed consent was obtained from each patient and the study was approved by the ethics review board of Galway University Hospital. The initial cohort for microarray analysis consisted of 29 early-stage, invasive ductal carcinoma breast tumour specimens. A larger cohort of fresh-frozen breast tumour (n = 95) and tumour-associated normal breast tissue (n = 17) specimens was used for validation and further analysis of selected miRNAs. Clinical and pathological data relating to the clinical samples are presented in Tables 1 and 2.

The ER, PR and HER2/*neu* status of the patients was determined by immunohistochemistry on formalin-fixed, paraffin-embedded sections of clinical specimens as part of routine pathology to guide clinical decision-making regarding adjuvant therapy. Immunohistochemistry was performed using a rabbit monoclonal antihuman ER antibody (clone SP1; Dako, Cam-

Table 1

Clinical and pathological data for breast tumours analysed by microarray

Number	ID	Age (years)	Tsize (mm)	Lymph node status	Grade	UICC stage	ER	PR	HER2/*neu*	Subtype
1	52	49	23	Negative	1	2A	P	P	N	Luminal A
2	53	52	30	Negative	3	2A	N	N	P	Her2 overexpressing
3	54	57	45	Negative	3	2A	N	N	P	Her2 overexpressing
4	56	51	21	Negative	3	2A	P	P	N	Luminal A
5	58	68	15	Negative	3	1	P	N	N	Luminal A
6	59	42	22	Negative	3	2A	N	N	N	Triple negative
7	60	54	26	Negative	3	2A	N	P	N	Luminal A
8	61	35	22	Negative	3	2A	P	P	N	Luminal A
9	62	50	16	Negative	3	1	N	N	N	Triple negative
10	63	49	25	Negative	2	2A	N	N	N	Triple negative
11	64	59	20	Negative	3	1	N	P	N	Luminal A
12	65	58	22	Negative	3	2A	P	P	N	Luminal A
13	66	58	18	Negative	1	1	N	P	P	Her2 overexpressing
14	67	66	22	Negative	3	2A	P	P	N	Luminal A
15	94	56	17	Negative	1	1	N	N	N	Triple negative
16	95	48	30	Negative	3	2A	N	N	P	Her2 overexpressing
17	96	60	26	Negative	3	2A	P	P	N	Luminal A
18	97	56	29	Negative	2	2A	P	P	N	Luminal A
19	98	50	3	Negative	2	1	P	P	N	Luminal A
20	99	40	7	Negative	1	1	P	P	N	Luminal A
21	100	40	6	Negative	2	1	P	P	N	Luminal A
22	101	58	35	Negative	2	2A	P	P	N	Luminal A
23	102	64	34	Negative	3	2A	P	P	N	Luminal A
24	103	66	26	Negative	1	2A	P	P	N	Luminal A
25	104	84	16	Negative	2	1	N	P	N	Luminal A
26	105	57	7	Negative	3	1	N	P	N	Luminal A
27	106	68	35	Negative	3	2A	P	P	N	Luminal A
28	107	40	20	Negative	2	1	P	P	P	Luminal B
29	108	49	35	Negative	3	2A	N	N	N	Triple negative

ER, oestrogen receptor; HER2/*neu*, v-erb-b2 erythroblastic leukaemia viral oncogene homolog 2 receptors; ID, identification; N, negative confirmed; PR, progesterone receptor; P, positive confirmed; Tsize, Tumour size in mm; UICC, stage of breast tumour according to the international union against cancer staging criteria.

Table 2

Clinical and pathological data for breast tumours in the independent validation cohort

Breast cancer clinicopathological characteristic	Number of patients (n = 95)
Median (interquartile range) tumour size (mm)	23.5 (17.75 to 35.0)
Histologic subtype	
Invasive ductal	80
Invasive lobular	13
Colloid/mucinous	1
Tubular	1
Tumour-associated normal	17
Intrinsic subtype	
Luminal A (ER/PR$^+$, HER2/*neu*$^-$)	47
Luminal B (ER/PR$^+$, HER2/*neu*$^+$)	21
Her2 overexpressing (ER$^-$, PR$^-$, HER2/*neu*$^+$)	11
Triple-negative (ER$^-$, PR$^-$, HER2/*neu*$^-$)	11
Missing data	5
Grade	
1	14
2	26
3	53
Missing data	2
Nodal status	
Node-negative	50
N1	17
N2	17
N3	11
Oestrogen receptor status	
Positive	62
Negative	32
Missing data	1
Progesterone receptor status	
Positive	58
Negative	33
Missing data	4
Her2/*neu* status	
Positive	32
Negative	59
Missing data	4
UICC stage	
Stage 1	23
Stage 2a	29
Stage 2b	8
Stage 3a	14
Stage 3b	4
Stage 3c	8
Stage 4	9

ER, oestrogen receptor; HER2/*neu*, v-erb-b2 erythroblastic leukaemia viral oncogene homolog 2 receptors; PR, progesterone receptor; UICC, stage of breast tumour according to the international union against cancer staging criteria.

bridgeshire, UK) and a polyclonal rabbit antihuman PR antibody (Dako). The Allred scoring method was used for expression scoring of ER and PR based on proportion and intensity. In brief, the proportion score represented the estimated percentage of tumour cells staining positive (0 = 0%; 1 = 1%; 2 = 1 to 10%; 3 = 10 to 33%; 4 = 33 to 66%; 5 = > 67%), and the intensity of staining was scored as follows: 1 = weakly positive; 2 = moderately positive; 3 = strongly positive. The total score was derived from the following equation, a score of 0 being negative and a score of 2 to 8 being positive:

$$\text{Percentage of positive cells} + \text{intensity of staining} = \text{total score}$$

Membranous staining was scored for HER2/*neu* according to the HercepTest (Dako) as follows: 0 = negative; 1 = weak incomplete membranous staining of > 10% cells (negative); 2 = weak – moderate complete membranous staining of > 10% of cells (equivocal-fluorescence *in situ* hybridization was used to assess amplification in these cases); 3 = strong complete membranous staining of > 30% of cells (positive).

miRNA microarray

RNA extraction

Depending on whether samples were destined for microarray or RQ-PCR analysis, slightly modified RNA extraction methods were employed. For the microarray experiment, total RNA was required. Breast tumour tissue (50 to 100 mg) was homogenized using a bench-top homogenizer (Polytron® PT1600E; Kinematica AG, Littau-Luzem, Switzerland) in 1 ml QIAzol lysis reagent (Qiagen, Crawley, UK). Total RNA was isolated from homogenized breast tissue using the RNeasy® Tissue Mini Kit (Qiagen) according to the manufacturer's instructions. For RQ-PCR, miRNA was selectively isolated from approximately 100 mg tissue.

Large RNA fractions (> 200 nucleotides) and small RNA fractions (< 200 nucleotides) were isolated separately using the RNeasy Plus Mini Kit and RNeasy MinElute® Cleanup Kit (Qiagen) according to the supplementary protocol: purification of miRNA from animal cells. The concentration and purity of total RNA were assessed using a NanoDrop™ 1000 spectrophotometer (Nanodrop Technologies, Wilmington, DE, USA). RNA integrity was assessed using the RNA 6000 Nano LabChip Series II Assay with the 2100 Bioanalyzer System (Agilent Technologies, Palo Alto, CA, USA). Electropherograms and gel-like images were evaluated using the Agilent 2100 Expert software (version B.02.03; Agilent Technologies, Palo Alto, CA, USA), which generated the RNA integrity number to ensure that only RNA of good integrity was used in these experiments (RNA integrity number range, 7.6 to 9.5). The miRNA concentration and purity were also assessed by NanoDrop™ 1000 spectrophotometry. Small miRNA-enriched fractions were analysed using the Small RNA Assay on the Agilent 2100 Bioanalyzer.

RNA labelling and microarray hybridization

Total RNA was Cy-dye labelled and hybridized on miRNA microarray chips as previously described [24]. Briefly, 5.5 μg total RNA was 3' ligated to Cy dye-linked 2'-deoxyuridine-5'-triphosphate using T4 RNA ligase (catalogue number 2141; Ambion, Woodward, Austin, TX, USA), in the presence of RNase inhibitor (catalogue number 2682; Ambion, Woodward, Austin, TX, USA), ATP (Grade I, catalogue number A2383-1G; Sigma-Aldrich Corp. St. Louis, MO, USA), and polyethylene glycol 50% aqueous solution (PEG 6000, catalogue number 81304; Fluka, Sigma-Aldrich Corp, St. Louis, MO, USA). Following a 12-hour to 16-hour incubation, labelled RNA was washed in ethanol, and precipitated in sodium acetate (3 M) using linear acrylamide. Labelled RNA was hybridized to LNA™ miChip array platforms (Exiqon version 7, containing 453 miRNA sequences) over 16 hours at 54°C using a rotational hybridization chamber. Arrays were subsequently washed in varying stringency washes, rinsed, drained and scanned using a GenePix 4000AL laser scanner (Axon Instruments, Foster City, CA, USA).

Data processing

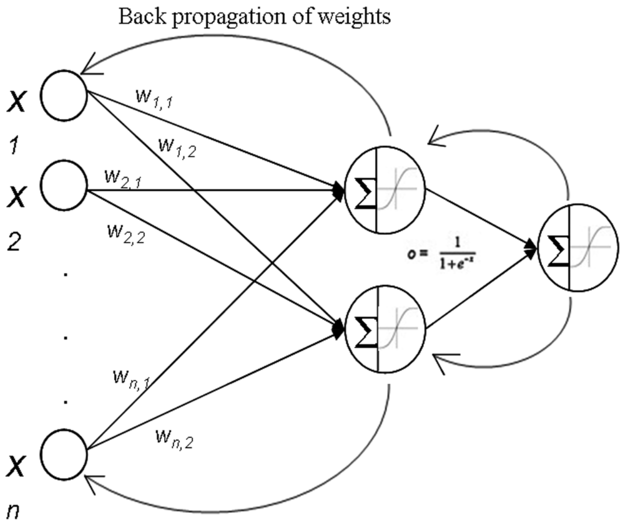
Images generated by the GenePix 4200AL scanner were imported to GenePix 6 microarray analysis software (Axon Instruments, Foster City, CA, USA). Artefact-associated spots were removed by both software-guided and visual-guided flags. Empty and control data were filtered out. Signal intensities were measured according to the local background subtraction method as a function of the median of foreground pixels minus the median of background pixels. The median spot intensities were then normalized to the median intensity per chip using custom R scripts. All microarray data were submitted to the Gene Expression Omnibus [GEO:GSE15885].

Artificial neural network algorithms and architecture

Within the present study, a three-layer multilayer perceptron modified with a feedforward back-propagation algorithm and a sigmoidal transfer function [25] was employed (Figure 1). The learning rate and momentum were respectively set at 0.1 and 0.5. Automatic pre-processing normalized the data between 0 and 1 for each variable. The intensity values for the miRNA for each individual were represented in the input layer, the hidden layer contained two hidden nodes, and the class (related to ER, PR or HER2/*neu*) was represented in the output layer coded as 0 for negative and 1 for positive.

A randomly selected subset of the cases devolved for training purpose is presented to the network to train it (training data) while it is constantly monitored with a randomly select subset of unseen cases (test data). These test data are used to stop the training process once the model has reached predetermined conditions such as an optimal error value preventing overtraining. Once training is stopped, the efficiency of the model is further assessed by presenting a third, randomly selected, blind subset to the model to determine performance

Figure 1



Multilayer perceptron with sigmoidal activation function. Weights are adjusted at the end of each epoch by the back-propagation algorithm.

for unseen cases not involved in the training process. This subset selection process was repeated up to 50 times for randomly selected subsets, a process known as Monte – Carlo cross-validation. The suite of 50 models produced was analysed and screened for model optimization purposes.

Model optimization

An additive stepwise approach was employed (as described previously [21]) to identify an optimal set of markers explaining variation in the population of each of the questions explored: ER, PR and HER2/*neu* status, for miRNA microarrays. In brief, the stepwise approach consists of taking each single variable as an input to the ANN, and training 50 submodels with Monte–Carlo cross-validation. Each single input model subset is then analysed and the median classification performance (based on the predictive error for the blind test set) is determined. The median performance for all single inputs is then analysed and the inputs ranked accordingly. The best predictor input (with the lowest error) is then selected and a second single variable added, creating a two-input model. This was repeated for all of the variables in the dataset, and the best pair was determined again based on the classification error. Further inputs are then added in stepwise fashion (generating three-input models, four-input models, and so on) until no further improvement is obtained and an optimal model with the best predictive performance is generated.

cDNA synthesis and RQ-PCR

RQ-PCR quantification of miRNA expression was performed using TaqMan MicroRNA® Assays (Applied Biosystems, Foster City, CA, USA) according to the manufacturer's protocol. Small RNA (5 ng) was reverse-transcribed using the MultiScribe™-based High-Capacity cDNA Archive kit (Applied Biosystems). RT-negative controls were included in each batch of reactions. PCR reactions were carried out in final volumes of 20 μl using an ABI Prism 7000 Sequence Detection System (Applied Biosystems). Briefly, reactions consisted of 1.33 μl cDNA, 1× TaqMan® Universal PCR Master Mix, 0.2 μM TaqMan® primer–probe mix (Applied Biosystems). Reactions were initiated with a 10-minute incubation at 95°C followed by 40 cycles of 95°C for 15 seconds and 60°C for 60 seconds. *miRNA-16* and *let-7a* were used as endogenous controls to standardize miRNA expression [26]. An interassay control derived from a breast cancer cell line (ZR-75-1) was included on each plate. All reactions were performed in triplicate. The threshold standard deviation for intra-assay and inter-assay replicates was 0.3. The percentage PCR amplification efficiencies (*E*) for each assay were calculated, using the slope of the semi-log regression plot of cycle threshold versus log input of cDNA (10-fold dilution series of five points), with the following equation:

$$E = \left(10 - 1/\text{slope} - 1\right) \times 100$$

A threshold of 10% above or below 100% efficiency was applied.

Relative quantification

The relative quantity of miRNA expression was calculated using the comparative cycle threshold (ΔΔCt) method [27]. The geometric mean of the cycle threshold value of the endogenous control genes was used to normalize the data, and the lowest expressed sample was used as a calibrator.

Statistical analysis of RQ-PCR miRNA expression data

The Kolmogorov–Smirnov normality test was applied; as the values of miRNA expression displayed a non-normal distribution, data were standardized by $\log_{10}$ transformation. Associations between miRNA expression and standard prognostic factors (patient age, tumour size, tumour grade, axillary nodal status, hormonal status and HER2/*neu* status) were examined using *t* tests, analysis of variance and Pearson correlations. The above tests were performed in SPSS® (version 14.0; SPSS Inc., Chicago, IL, USA). $P < 0.05$ was considered statistically significant.

Results

miRNA signatures predictive of ER, PR and HER2/*neu* status

Using the ANN to analyse miRNA array expression data, we identified distinct miRNA expression signatures predictive of ER, PR, and HER2/*neu* status in breast tumour samples. The ER signature consisted of six miRNA transcripts (*miR-342*, *miR-299*, *miR-217*, *miR-190*, *miR-135b*, *miR-218*), and discriminated cases correctly with a median accuracy of 100% when classifying between ER-positive and ER-negative phenotypes. Similarly, four miRNA transcripts (*miR-520g*, *miR-377*, *miR-527-518a*, *miR-520f-520c*) were identified that predicted tumour PR status with 100% accuracy, and HER2/*neu* status was predicted with 100% accuracy by a signature of five miRNAs (*miR-520d*, *miR-181c*, *miR-302c*, *miR-376b*, *miR-30e*) (Table 3).

These reported accuracies are from separate validation data splits where the samples were treated as blind data over 50 models with extensive Monte–Carlo cross-validation. At each step of the model, additional miRNA transcripts were selected; the addition of key miRNA transcripts improved the predictive capabilities of the signature. When there was no further improvement in performance with regards to predictive error, no additional miRNA transcripts were added as the signatures were now considered to contain the optimum miRNAs to most accurately model the data. Figure 2 shows the performance of the models at each step of the analyses, and it is evident that the selection and addition of key transcripts led to an overall improvement in the error associated with predictive capabilities of the model for blind data. After step 6, step 4 and step 5 for the ER, PR and HER2/*neu* data, respectively, no further steps were conducted as no significant improvement in performance with regards to predictive error could be achieved. At this point the models were considered to contain the miRNAs that most accurately predicted receptor status.

A detailed examination of the ranked model performance for the most predictive individual miRNA transcripts in step 1 of the analysis is presented in Table 4. There are a number of miRNA transcripts capable of classifying samples effectively, independently of the miRNA ranked highest in terms of predictive ability. All of these miRNAs are considered important in step 1 of the analysis; however, they are not independent of each other and may all explain the same variation in the data. These miRNAs are not subsequently identified as important in the following steps of the analysis, and as a result are not all present in the final signatures. The miRNA signatures that are included in the final model each explain additional variation in the patient data, and the combination of these transcripts contributes to the final predictive power of the model. Table 3 summarizes the performances of the network models at each step of the analysis; the transcripts in this table composed the final miRNA signatures for ER, PR and HER2/*neu* status, respectively.

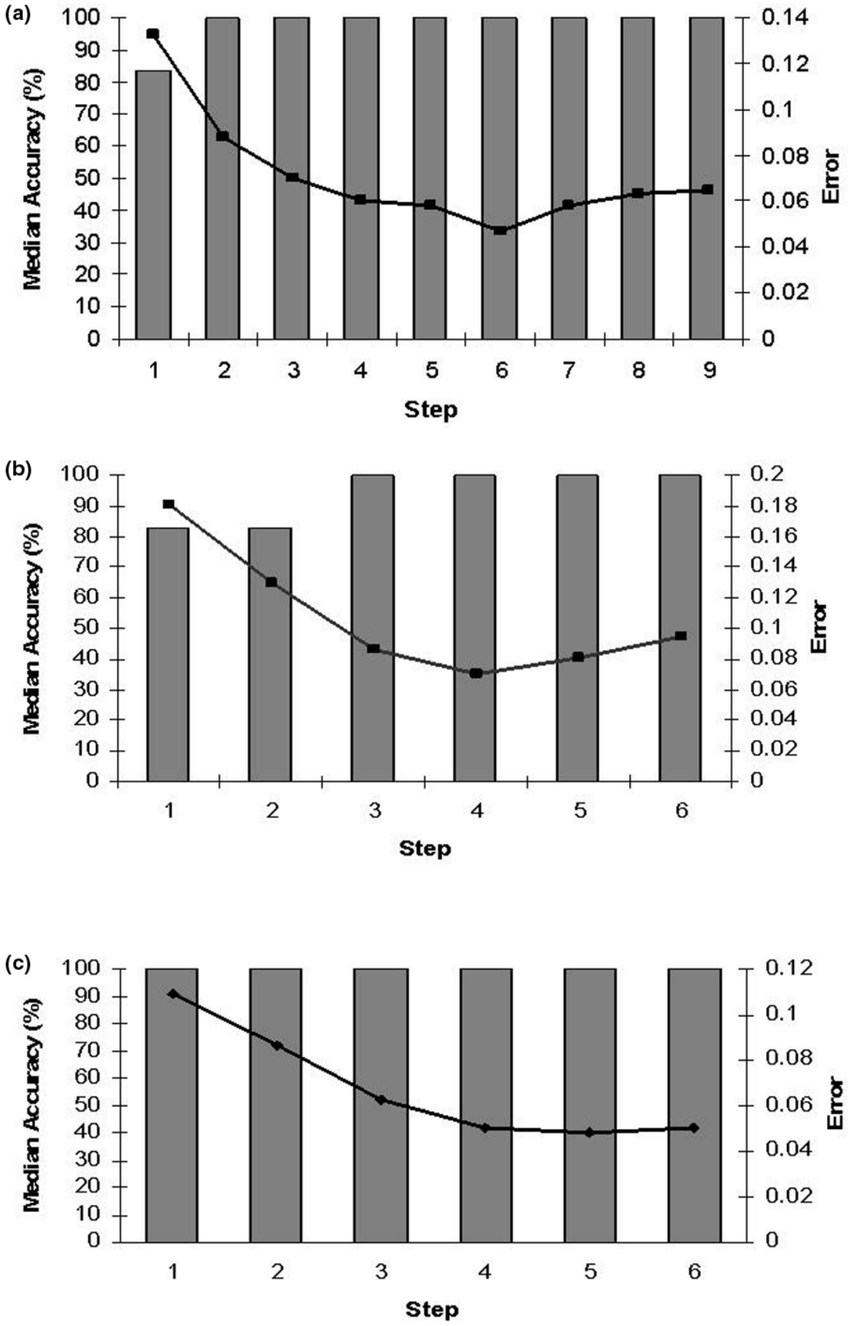
Table 3

Summary microRNAs used in the expression signature at each step of model development

Rank	miRNA	Chromosomal location	Validated mRNA targets	Mean squared error	Median accuracy (%)	Response[a]
ER status						
1	*miR-342*	14q32.2, intronic	-	0.132	83.3	(+)
2	*miR-299-3p*	14q32.31, intergenic	-	0.087	100	(-)
3	*miR-217*	2p16.1, intergenic	-	0.07	100	(+)
4	*miR-190*	15q22.2, intronic	-	0.06	100	(-)
5	*miR-135b*	1q32.1, intronic	-	0.057	100	(-)
6	*miR-218*	4p15.31, intronic	LAMB3	0.047	100	(+)
PR status						
1	*miR-520g*	19q13.42, intergenic	-	0.186	83.3	(-)
2	*miR-377*	14q32.31, intergenic	-	0.129	83.3	(+)
3	*miR-527-518a*	19q13.42, intergenic	-	0.086	100	(-)
4	*miR-520f-520c*	19q13.42, intergenic	-	0.07	100	(+)
HER2/*neu* status						
1	*miR-520d*	19q13.42, intergenic	-	0.109	100	(+)
2	*miR-181c*	19q13.12, intergenic	Tcl1	0.086	100	(-)
3	*miR-302c*	4q25, intronic	Cyclin D_1	0.062	100	(*)
4	*miR-376b*	14q32.31, intergenic	-	0.050	100	(+)
5	*miR-30e-3p*	1p34.2, intronic	Ubc9	0.047	100	(*)

Summary microRNAs (miRNAs) used in the expression signature at each step of model development for oestrogen receptor (ER) status, progesterone receptor (PR) status and v-erb-b2 erythroblastic leukaemia viral oncogene homolog 2 receptors (HER2/*neu*) status. [a](+), increased miRNA expression leads to increased probability of receptor positive status; (-), increased miRNA expression leads to increased probability of receptor negative status; (*), weak response, possibly interacting to modify the response of other miRNAs.

Figure 2



Performance of the models at each step of the analyses. Model performance with each input addition over the course of the analysis for **(a)** oestrogen receptor (ER) status – 6 optimal transcripts. After the addition of the six optimal microRNA transcripts, the accuracy of the model has reached 100% and there is no further improvement in the error. At this point the model is considered to contain the transcripts that most accurately model the data. Columns represent median model accuracy; lines represent mean squared error for the predictions at each step. **(b)** progesterone receptor (PR) status – four optimal transcripts, and **(c)** v-erb-b2 erythroblastic leukaemia viral oncogene homolog 2 receptor (HER2/*neu*) status – five optimal transcripts.

Table 4

Summary of step 1 of the stepwise analysis of the ER, PR, and HER2/*neu* signatures

Rank	MicroRNA	Chromosomal location	Validated mRNA targets	Mean squared error	Median accuracy (%)
ER status					
1	*miR-342*	14q32.2, intronic	-	0.132	84
2	*miR-520g*	19q13.42, intergenic	-	0.198	73
3	*miR-107*	10q23.31, intronic	-	0.200	73
4	*miR-149*	2q37.3, intronic	-	0.201	69
5	*miR-520g-h*	19q13.42, intergenic	-	0.203	73
6	*miR-155*	21q21.3, exon	AGTR1, AID, TP53INPI	0.208	70
7	*miR-30c*	1p34.2, intronic	-	0.210	67
8	*miR-382*	14q32.31, intergenic	-	0.211	67
PR status					
1	*miR-520g*	19q13.42, intergenic	-	0.180	83.3
2	*miR-520d*	19q13.42, intergenic	-	0.181	83.3
3	*let-7d*	9q22.32, intronic	SMC1A	0.185	67
4	*miR-328*	16q22.1, intronic	CD44, BCRP	0.189	83.3
5	*miR-373*	19q13.41intergenic	E-Cadherin, lats2 CSDC2, CD44, RAD23B	0.189	83.3
6	*miR-217*	2p16.1, intergenic		0.196	67
7	*miR-504*	Xq26.3, intronic		0.198	67
8	*miR-485-3p*	14q32.31, intergenic		0.201	83.3
HER2/*neu* status					
1	*miR-520d*	19q13.42, intergenic		0.109	87.5
2	*miR-30b*	8q24.22intergenic		0.111	83.3
3	*miR-217*	2p16.1, intergenic		0.114	83.3
4	*miR-363*	Xq26.2, intergenic		0.115	83.3
5	*miR-383*	8p22, intronic		0.115	83.3
6	*miR-377*	14q32.31, intergenic		0.120	87.5
7	*miR-130a*	11q12.1, intergenic	GAX, HOXA5	0.121	83.3
8	*miR-422a*	15q22.31, intergenic		0.122	83.3

ER, oestrogen receptor; PR, progesterone receptor; HER2/*neu*, v-erb-b2 erythroblastic leukaemia viral oncogene homolog 2 receptors.

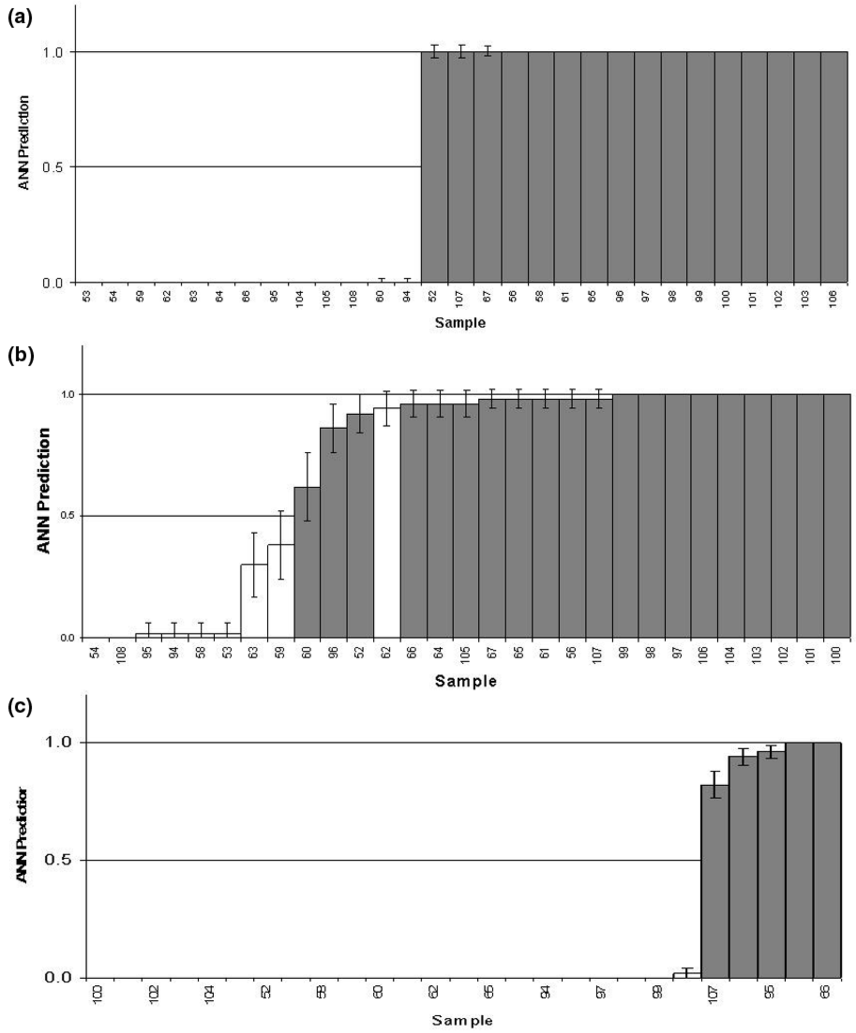
Sample population analysis

Figure 3 shows population structures for ER, PR and HER2/*neu* status. The transcript signature determined from the ANN model was used to position patients into population structures based upon the ANN predicted probability of the individual falling into a given receptor status class. By ranking the probabilities for individuals, the population structure is determined. The developed ANN model may be used to predict probability of receptor status and thus position new individuals within the population structure.

miRNA response curve analysis

To determine the effect of each individual miRNA on class prediction, the ANN model was presented with controlled input values representing discreet intervals across the range of the miRNA of interest (keeping all others at their mean value). The predicted probability in the output class of interest (that is, ER, PR and HER2/*neu* status) was determined under these input conditions and response curves were plotted. This enabled an understanding of how the miRNAs govern the tumour sample classification by assessing the strength of response. The response can be discriminatory (crosses the 0.5 class threshold) or co-factorial (does not cross the 0.5 class threshold). Such analysis identifies whether specific miRNA expression is

Figure 3



Population analysis for receptor status. Population analysis for **(a)** oestrogen receptor (ER) status. Using the transcript signature from the ANN model, it is possible to be able to place a patient with unknown ER status within this population structure, with 100% accuracy from an ANN prediction, **(b)** progesterone receptor (PR) status, and **(c)** HER2/*neu* status. White, receptor-negative patients; grey, receptor-positive patients. *y* axis, artificial neural network (ANN) prediction with 0 being a receptor-negative prediction and 1 a receptor-positive prediction. Error bars indicate a 95% confidence interval.

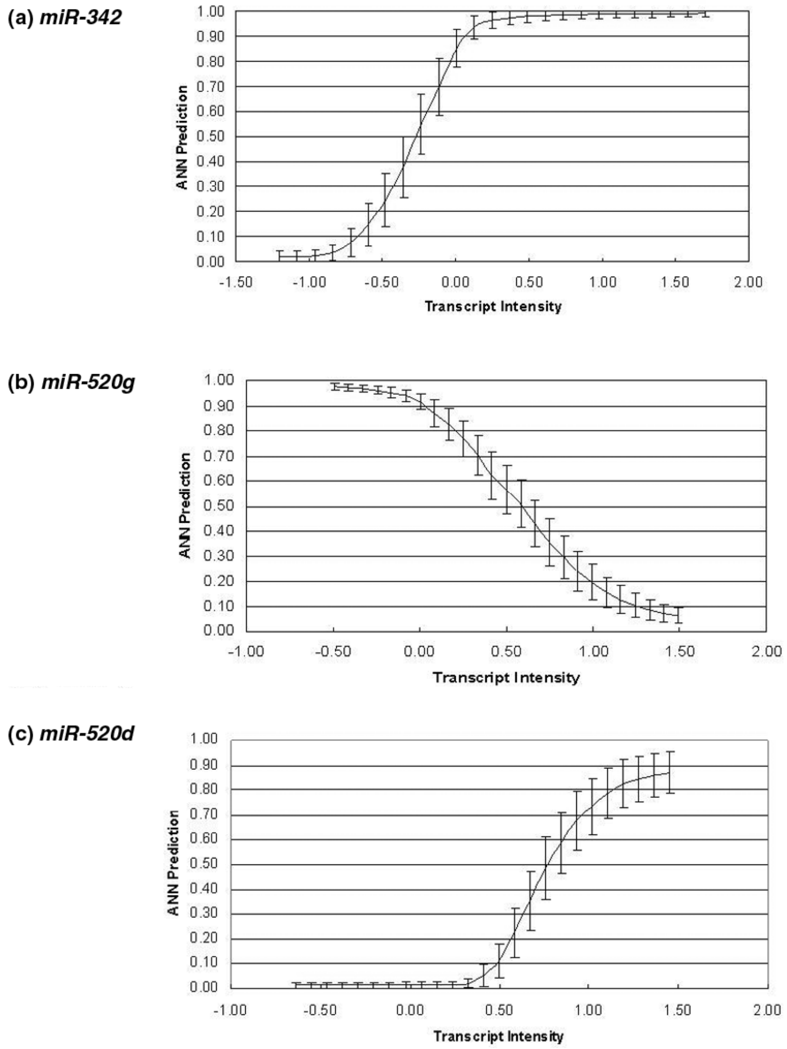
increased or decreased with respect to the receptor status, providing an indication of their possible biological role.

The analysis is performed using the trained ANN model and adjusting an input variable of interest to monitor the affect of this adjustment on the output variable. The output, with respect to the changing input value, is plotted to produce a response graph. The response graphs for *miR-342*, *miR-520g* and *miR-520d** in relation to ER, PR and HER2/*neu* status, respectively, are shown in Figure 4. Some miRNAs showed that with increased expression, the probability of receptor positivity increased; conversely, other miRNAs showed that with increased expression, the likelihood of the sample being classed as receptor-positive decreased. This highlights potential regulatory roles for these miRNAs through inhibition of the receptors themselves or of their co-regulators. Table 3 includes information on how the level of expression of each miRNA correlates with the receptor status.

Coordinated expression of miRNA clusters

The expression of miRNAs from the same chromosomal location was shown to be coordinated in our dataset. Figure 5 shows pairwise scatterplots for miRNAs transcribed from adjacent chromosomal regions. This highly correlated expression of adjacent miRNAs is in keeping with their processing from primary polycistronic transcripts.

Figure 4



Response curves for *miR-342*, *miR-520g* and *miR-520d*. Response curves for **(a)** *miR-342*, **(b)** *miR-520g* and **(c)** *miR-520d**. Figures show the intensity of each transcript plotted against the artificial neural network (ANN) prediction with respect to the sample being classified as either (a) oestrogen receptor (ER)-positive or ER-negative, (b) progesterone receptor (PR)-positive or PR-negative and (c) v-erb-b2 erythroblastic leukaemia viral oncogene homolog 2 receptor (HER2/*neu*)-positive or HER2/*neu*-negative. Error bars indicate 95% confidence intervals.

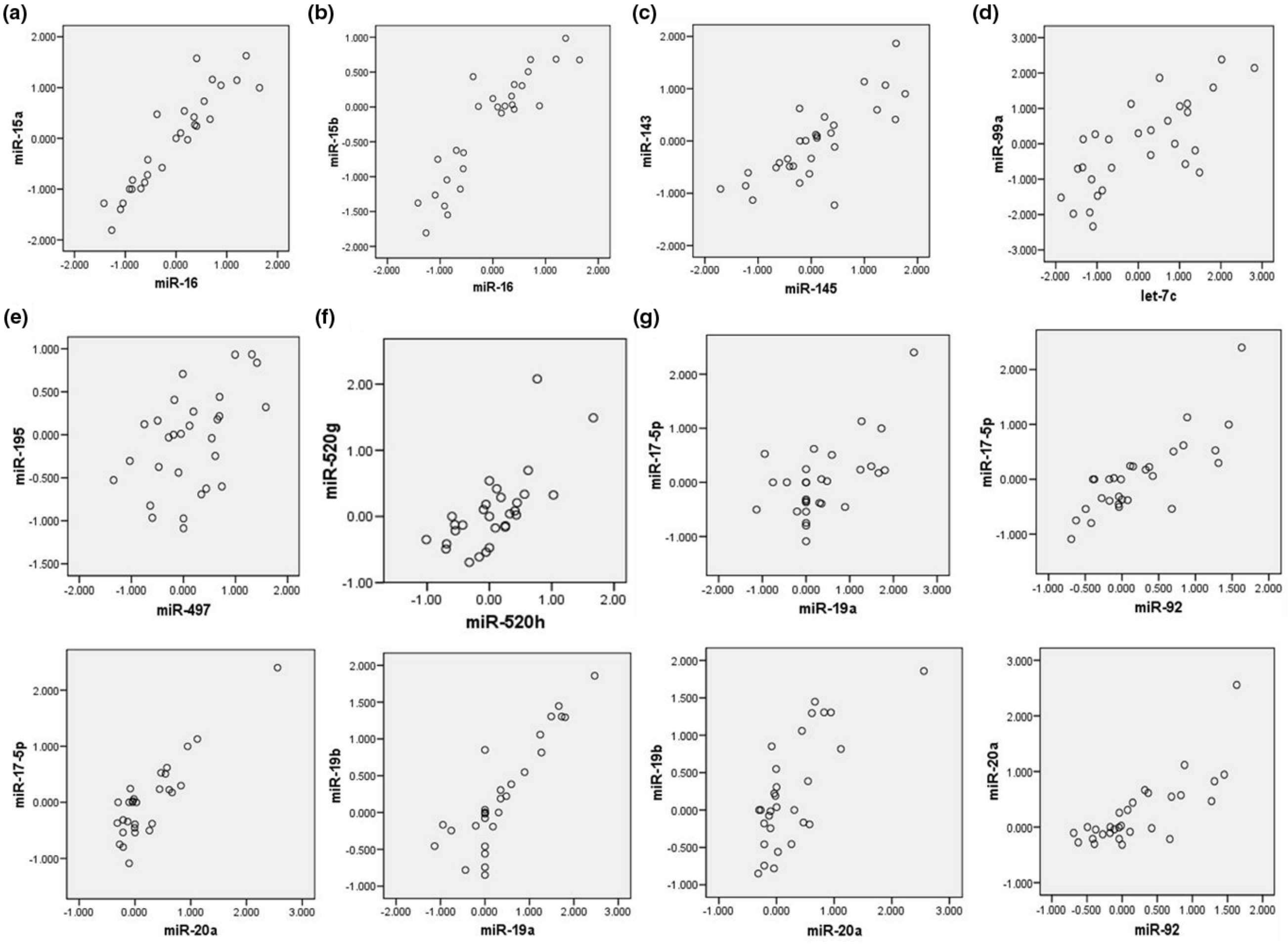
PCR validation

To confirm expression results obtained from the microarray analysis we carried out RQ-PCR on a subset of miRNAs. There was good correlation in sample-to-sample expression patterns between the two techniques (Figure 6).

Validation/interrogation of identified miRNAs

The first miRNA identified by the ANN model in relation to ER status was *miR-342*. The expression of *miR-342* was further analysed in a cohort of 95 breast tumours, 17 of which had matched tumour-associated normal tissue. RQ-PCR of mature *miR-342* in these samples showed no significant difference in expression between tumour and tumour-associated normal tissue (P = 0.6, paired t test). Within the tumour samples, the expression of *miR-342* was significantly higher in ER-positive tumours (n = 62) compared with ER-negative tumours (n = 32) (P = 0.04, independent t test), confirming the association with ER positivity identified in the ANN response curve analysis. *miR-342* expression was also higher in the HER2/*neu*-positive tumours (n = 59) versus the HER2/*neu*-negative tumours (n = 32) (P = 0.001, independent t test). The expression of *miR-342* was highest in the luminal B subtype of breast cancers and was lowest in the triple-negative/basal subtype (P = 0.001, analysis of variance; Figure 7). There was no association of *miR-342* with other clinicopathological parameters, including PR status, grade, stage or nodal status.

Figure 5



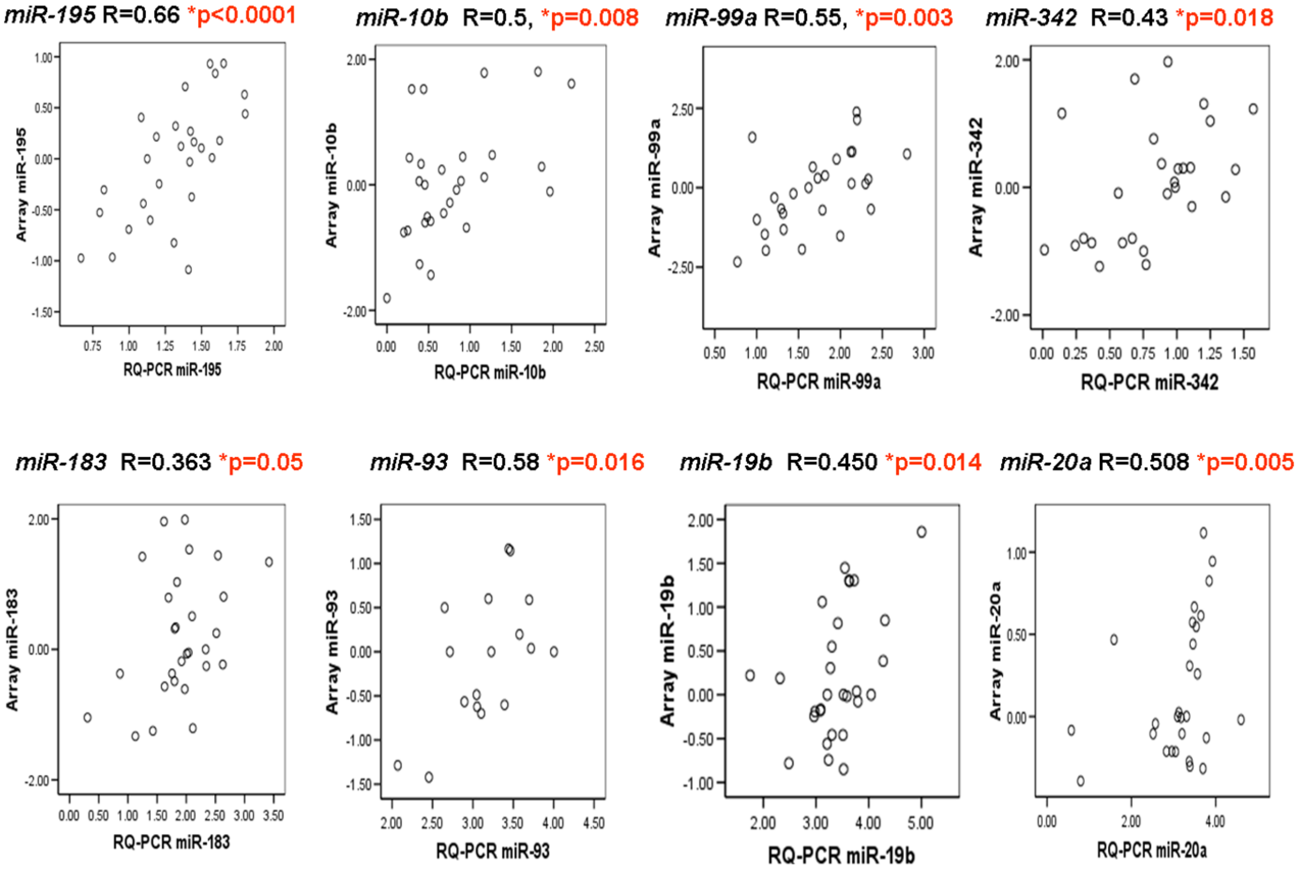
Coordinate expression of co-located microRNAs. Scatterplots of expression values for microRNAs located adjacently on the same chromosome. **(a)** *miR-16* and *miR-15a*; Ch13q14.3. **(b)** *miR-16* and *miR-15b*; Ch3q26.1. **(c)** *miR-143* and *miR-145*; Ch5q14. **(d)** *miR-99a* and *let-7c*; Ch21q16. **(e)** *miR-195* and *miR-497*; Ch17p13.1. **(f)** *miR-520g* and *miR-520h*; Ch19q13.42. **(g)** *miR-17-5p*, *miR-18a*, *miR-19a*, *miR-19b*, *miR-20a*, *miR-92*; Ch13q31.3.

miR-520g was the top-ranked miRNA in the PR status signature (Table 3) and the second-ranked miRNA predictive of ER in step 1 of the analysis (Table 4). The expression of *miR-520g* was also analysed using RQ-PCR. There was a significant positive correlation between *miR-520g* microarray expression and RQ-PCR ($R = 0.4$, $P = 0.029$, Pearson). In the cohort of 95 breast tumours with 17 matched tumour-associated normal breast tissue tissues there was no significant difference in *miR-520g* expression between tumour and tumour-associated normal breast tissue ($P = 0.228$, paired *t* test). Within the tumour samples, *miR-520g* expression was significantly higher in PR-negative breast tumours (n = 33) compared with PR-positive tumours (n = 58) ($P = 0.032$, independent *t* test). The *miR-520g* expression was also significantly higher in ER-negative tumours (n = 32) compared with ER-positive tumours (n = 62) ($P = 0.005$, independent *t* test). There was no significant association of *miR-520g* with other tumour characteristics, including HER2/*neu* status, tumour size, grade, stage or nodal status.

Discussion

In the wake of molecular profiling and the identification of intrinsic subtypes, breast cancer is now considered a heterogeneous group of disease entities with distinct clinical, pathological and molecular features. This biologic heterogeneity has implications for treatment; response to therapy can be predicted by subtyping tumours based on their expression profiles [2]. The molecular subclasses of breast cancer that are predictive of prognosis are based on their expression of specific genes including ER and HER2/*neu*: luminal-A subtype, ER^+/HER2/neu^-; luminal-B subtype, ER^+/HER2/neu^+; basal-like subtype, ER^-/PR^-/HER2/neu^-; HER2/*neu*-overexpressing subtype, ER^-/HER2/neu^+ [1]. The expression of these receptors alone has also been shown to have an effect on chemo-

Figure 6



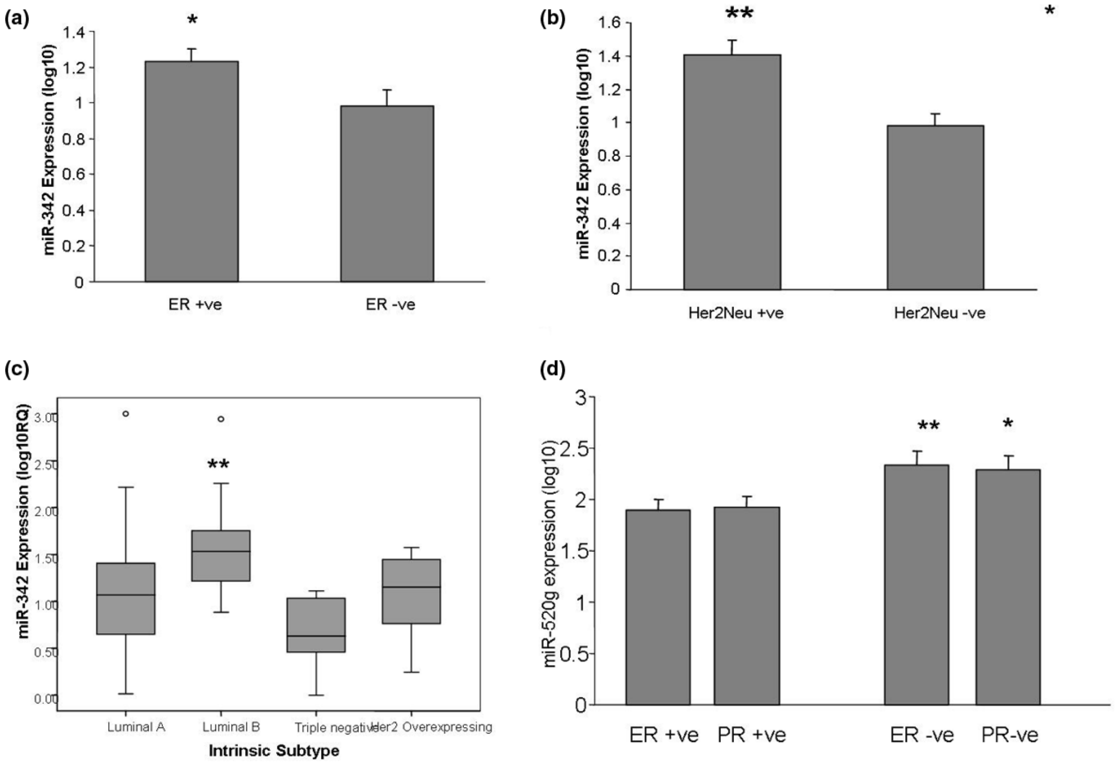
Correlation between microRNA expression on microarray and RQ-PCR. For a subset of microRNAs (miRNAs) and samples we performed RQ-PCR to independently assess miRNA expression. RQ-PCR data are normalized using *let-7a* and *miR-16*. There is generally good correlation between miRNA expression using the two techniques. probe-specific differences were observed, however. *R* value using Pearson correlation, $P < 0.05$ significant.

therapy sensitivity [28]. Furthermore, the only targeted therapies currently used in the management of breast cancer are directed at these receptors; ER-positive tumours are treated with endocrine therapy in the form of selective ER modulators, pure anti-oestrogens such as fulvestrant that completely inhibits ER signalling, or aromatase inhibitors that deplete extragonadal oestrogen synthesis. The monoclonal antibody trastuzumab has been developed to target the HER2/*neu*, while lapatinib inhibits HER2/*neu*-associated tyrosine kinase activity.

The specific combination of receptor status has a significant impact on the outcome of these targeted therapies; HER2/*neu*-positive breast cancer is less responsive to any type of endocrine treatment [29]; approximately one-half of HER2/*neu*-positive breast cancers are also ER-positive, and this breast cancer subgroup (luminal B) is thus more refractory to endocrine therapy – despite the ER-positive status. In addition, many patients with HER2/*neu*-positive breast cancers do not respond or eventually evade trastuzumab by both *de novo* and acquired mechanisms of therapeutic resistance. The subset of patients who are HER2/*neu*-negative and ER-negative (basal like/triple negative) are a particular therapeutic challenge as they typically exhibit aggressive clinical behaviour and poorer prognosis. Focused research has revealed promising strategies for treating this subtype of breast cancer, including platinum agents, epidermal growth factor receptor (EGFR)-targeted agents and poly(ADP-Ribose) polymerase (PARP) inhibitors; however, there is as yet no specific target for effective tailored therapy in this subgroup.

Clearly the hormone (ER and PR) and HER2/*neu* receptors are vitally important to the current classification and management of breast cancer; however, there is little knowledge regarding the precise regulation of these receptors. For this reason we sought to identify miRNAs associated with these receptors.

Figure 7



Expression of *miR-342* and *miR-520g* in breast tumours. RQ-PCR detection analysis shows that expression levels of *miR-342* are increased in: **(a)** oestrogen receptor (ER)-positive tumours compared with ER-negative tumours ($P = 0.04$), **(b)** v-erb-b2 erythroblastic leukaemia viral oncogene homolog 2 receptor (HER2/*neu*)-positive compared with HER2/*neu*-negative tumours ($P = 0.001$), and **(c)** luminal-B subtype of breast tumours ($P = 0.001$). **(d)** *miR-520g* expression is increased in ER-negative tumours compared with ER-positive tumours ($P = 0.005$) and in progesterone receptor (PR)-negative tumours compared with PR-positive tumours ($P = 0.032$). MicroRNA expression presented as $\log_{10}$ of the relative quantity. *$P < 0.05$, **$P < 0.005$.

Microarray profiling is a useful strategy for examining global gene and miRNA expression [17]. Messenger RNA profiling has been central to breast cancer subtyping. Adaptation of microarray-devised gene sets into routine clinical practice, however, has been hindered by the apparent lack of consensus between gene sets. One reason for this hindrance is that the classical computational analysis of such highly dimensional microarray data has proved problematic as it is not robust enough. The inherent noise (for example, experimental error, sample and chip variability) can significantly interfere with the development of accurate predictive models, and their performance is compromised by their modelling of extraneous portions of the dataspace. Michiels and colleagues questioned the robustness of the analysis of several microarray studies, and found that the molecular signatures were largely dependent on the selection of patients in training sets and that several of the largest studies addressing cancer prognosis failed to classify patients better than randomly [30].

ANNs were chosen as the bioinformatics tool for microarray data analysis for the present study due to their ability to cope with complex data and the potential for modelling data of high nonlinearity. For this reason, they have been widely applied to a range of domains including character/face recognition [31], stockmarket predictions [32], or survival prognosis for trauma victims [33]. ANN model development is achieved by a training process involving the adjustment of the weighted interconnections between nodes within the neural network over a defined number of epochs. This adjustment occurs by the iterative propagation of the predictive error back through the entire network with a learning algorithm (for example, the back-propagation algorithm used in the present study). ANNs have already been successfully applied in a number of contexts where markers of biological relevance have been identified, including polycystic ovarian syndrome [34], melanoma [22], prostate cancer [35] and breast cancer [36].

The miRNA expression profiles have shown superior accuracy to mRNA signatures at classifying tumours [17]. The novel application of ANNs to the analysis of miRNA array data should serve to enable breast tumours to be classified according to their miRNA expression profile, and should also focus attention upon a relatively small number of molecules that might warrant further biochemical/molecular characterization to assess their suitability as potential therapeutic targets.

In the present study, miRNA transcript signatures predictive of ER, PR and HER2/*neu* status were generated from microarray data using an ANN model (Tables 3 and 4). The breast tumours selected for the array experiment were relatively homogeneous in terms of other clinicopathological parameters, all being early stage (stages 1 and 2a) and free of nodal disease. In the first step of the analysis, miRNAs capable of classifying tumour samples according to receptor status with an accuracy of 67 to 87% were identified. Sequential selection and addition of miRNAs to the ANN successfully identified an optimum miRNA set based on predictive performance.

While the model shows high confidence for the dataset analysed (100% predictive accuracies), further validation is required on larger datasets and validation of the miRNA sets identified using alternative methods such as PCR.

Confirmation of the expression data from the microarray by RQ-PCR was used for validation in this dataset; the expression patterns of a subset of eight miRNAs was validated in the same sample set by stem-loop RQ-PCR, and there was significant positive correlation in sample-to-sample expression patterns between the two techniques (Figure 6, $P < 0.05$). Furthermore, the expression patterns and phenotypic associations of the top-ranking miRNAs *miR-342* and *miR-520g* were validated in an independent sample set of 95 tumours (Figure 7).

The miRNA signatures generated for ER status (*miR-342*, *miR-299*, *miR-217*, *miR-190*, *miR-135b*, *miR-218*), for PR status (*miR-520g*, *miR-377*, *miR-527-518a*, *miR-520f-520c*) and for HER2/*neu* status (*miR-520d*, *miR-181c*, *miR-302c*, *miR-376b*, *miR-30e*) include miRNAs that have previously been identified as dysregulated in breast cancer and other cancers [7,9,37-43] and involved in the regulation of cell functions such as growth, apoptosis, migration and invasion [38,42,43]. This finding suggests that the miRNAs thus identified are biologically relevant and their selection is not arbitrary or a result of the highly dimensional nature of the data.

Notably, two chromosomal locations account for a number of the dysregulated miRNAs in these predictive sets: Ch19q13 (*miR-520g*, *miR-520d*, *miR-527-528a*, *miR-520f-520c*, *miR-181c*) and Ch14q32 (*miR-342*, *miR-299*, *miR-377*, *miR-376b*). Allelic deletions on chromosome 14q32 are frequently observed in various tumours, including renal cell carcinoma [44], neuroblastoma [45], colorectal carcinoma [46], bladder cancer [47], ovarian carcinoma [48], meningioma [49] and breast carcinoma [50].

Approximately one-third of human miRNAs are organized in clusters, which may represent a single transcriptional unit and coordinated regulation – possibly leading to synergistic biological effects, as suggested by the inclusion of miRNAs from adjacent chromosomal locations in our signatures. This may contribute to our finding that while single miRNAs are capable of distinguishing between different breast tumours (step 1; Table 4), multiple miRNAs in combination significantly enhance the predictive power of these models (step 2; Table 3). Our finding of co-expression of other neighbouring miRNAs not included in the predictive signatures (Figure 5) is in concordance with previous studies [7,51] and is probably due to shared regulatory elements.

A primate-specific conserved miRNA family is located at Ch19q13.42 [52]. Two miRNAs from this location, *miR-373* and *miR-520c*, have previously been shown to stimulate cancer cell migration and invasion in both *in vitro* and *in vivo* models and to be expressed at increased levels in metastatic breast cancer [43]. The miRNAs from this family were associated with ER, PR and HER2/*neu* status in our analysis. Similar seedpairing in miRNA families indicates that they may function through the same pathways and share mRNA targets – such as CD44, identified as a target of *miR-373* and known to correlate with survival in breast cancer patients [53]. It is likely that this particular miRNA family has a significant regulatory role in breast cancer.

miR-520g was ranked as the top miRNA in the PR signature and also was identified in step 1 of the analysis as an ER-predictive miRNA. Both of these findings were validated using RQ-PCR in a larger, more heterogeneous cohort of 95 breast tumours (Figure 7d). To our knowledge this is the first report of *miR-520g* dysregulation in association with ER and PR status in breast cancer. Importantly, *miR-520g* is computationally predicted to target a number of breast-cancer-related genes including ABCG2 (BCRP) [54]. ABCG2/BCRP is an ATP-binding cassette transporter that is often associated with multidrug resistance due to its ability to remove substrates from a cell against a concentration gradient [55]. ABCG2 expression in cancer cells has been shown to confer a drug-resistant phenotype and correlates with response to anthracyclines in breast cancer [56]. The regulation of ABCG2/BCRP is controlled via oestrogen and progesterone response elements [57,58], and the steroid hormones have been shown to impact on ABCG2 expression [57,59,60].

Recent studies have shown that ABCG2 expression is also regulated by miRNAs including *miR-328* [61], leading to increased mitoxantrone sensitivity, and by miRNAs from the Ch19q13.42 cluster. Specifically, ABCG2 is downregulated by *miR-519c* in drug-sensitive cells via a binding site in the 3' UTR that is not present in their drug-resistant counterparts [62], and *miR-520h* targets ABCG2 in hematopoietic stem cells during their differentiation into progenitor cells [63]. *miR-*

520g shares sequence homology with *miR-520h*, and these miRNAs were coordinately expressed in our dataset (Figure 5); it is therefore probable that *miR-520g* may also be a regulator of ABCG2. This hypothesis warrants further investigation; identification of miRNA binding sites in the 3' UTR of genes such as ABCG2 that promote multidrug resistance could enable the delivery of specific miRNAs from this cluster to tumours in an attempt to repress ABCG2 and to increase sensitivity to existing therapeutic agents.

The ER-status predictor *miR-342*, identified as having the strongest response curve, was also chosen for further characterization. Expression of *miR-342* in the larger cohort of breast tumours (n = 95) using RQ-PCR confirmed the microarray findings of an association between *miR-342* and ER positivity. Furthermore, we report the first findings of an association between *miR-342* and HER2/*neu* positivity. Increasing evidence suggests that *miR-342* plays an important role in the carcinogenic process, particularly in the hormonally regulated breast cancer. *miR-342* is dysregulated in multiple myeloma [64] and has been shown to be epigenetically silenced by methylation in colorectal carcinoma [42]. *In vitro* studies have demonstrated that introduction of a *hsa-miR-342* mimic to colorectal cancer cells induces apoptosis, suggesting a potential tumour suppressor role for this miRNA [42].

Previous miRNA profiling studies in breast cancer have identified associations between *miR-342* and ER, intrinsic breast cancer subtype and tumour grade [7,9]. A recent study has shown downregulation of *miR-342* in tamoxifen-resistant breast cancer cells compared with tamoxifen-sensitive breast cancer cells, suggesting a potential role as a biomarker of drug sensitivity [65]. To our knowledge this is the largest number of primary breast tumours in which *miR-342* has been quantitated using RQ-PCR. Our findings of increased *miR-342* expression in both ER-positive and HER2/*neu*-positive tumours is of particular interest as the luminal B (ER^+/HER2/neu^+) and triple-negative tumours present particular therapeutic challenges. In the present study, *miR-342* has emerged as a potential candidate for regulation of ER/HER2/*neu* expression that warrants further functional investigation to elucidate its mRNA targets and its precise role in breast carcinogenesis.

Conclusions

Our novel use of ANN to analyse miRNA expression profiles has identified biologically relevant miRNAs capable of discriminating between tumours with differing hormone receptor status in breast cancer. This approach contributes to the understanding of miRNA expression profiling in breast cancer, and the selection of the most predictive signatures has identified specific individual miRNAs and families of miRNAs that are promising candidates for future functional studies. These miRNAs have a potential influence on the behaviour of breast cancer subtypes in addition to their role as potential biomarkers. Uncovering the miRNA layer of genetic regulation will be part of the optimal approach to targeted therapy in breast cancer; this involves improving our understanding of molecular targets such as ER, PR and HER2/*neu* in addition to identifying novel molecular pathways and targets in order to predict response and to identify pathways of primary and acquired resistance to therapy.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

AJL performed the experiments, was responsible for data analyses and drafted the manuscript. NM conceived of, designed and supervised experimental work and manuscript editing. AD and PAD contributed to RQ-PCR data. REM contributed to sample preparation and array experiments, and participated in preliminary data analysis. VB, SS, JB were responsible for conducting microarray hybridizations and preliminary data analysis at EMBL Heidelberg. GB and CL designed bioinformatics models for interrogation of the array dataset. MJK contributed throughout the experiment, critically reviewed the manuscript and participated clinically in sample provision. All authors read and approved the final manuscript.

Acknowledgements

The authors would like to acknowledge the National Breast Cancer Research Institute and the John and Lucile Van Geest foundation for their financial support. AJL is supported by a Clinical Scientist fellowship award from Molecular Medicine Ireland. The authors gratefully acknowledge Ms Emer Hennessy for continued technical assistance and for curation of the Department of Surgery, BioBank, NUIG. They also wish to thank Ms Catherine Curran for collation of clinical and histological data.

References

1. Sorlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, Hastie T, Eisen MB, Rijn M van de, Jeffrey SS, Thorsen T, Quist H, Matese JC, Brown PO, Botstein D, Eystein Lonning P, Borresen-Dale AL: **Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications.** *Proc Natl Acad Sci USA* 2001, **98:**10869-10874.
2. Potti A, Dressman HK, Bild A, Riedel RF, Chan G, Sayer R, Cragun J, Cottrill H, Kelley MJ, Petersen R, Harpole D, Marks J, Berchuck A, Ginsburg GS, Febbo P, Lancaster J, Nevins JR: **Genomic signatures to guide the use of chemotherapeutics.** *Nat Med* 2006, **12:**1294-1300.
3. Chin SF, Teschendorff AE, Marioni JC, Wang Y, Barbosa-Morais NL, Thorne NP, Costa JL, Pinder SE, Wiel MA van de, Green AR, Ellis IO, Porter PL, Tavaré S, Brenton JD, Ylstra B, Caldas C: **High-resolution aCGH and expression profiling identifies a novel genomic subtype of ER negative breast cancer.** *Genome Biol* 2007, **8:**R215.
4. Ross JS, Hatzis C, Symmans WF, Pusztai L, Hortobagyi GN: **Commercialized multigene predictors of clinical outcome for breast cancer.** *Oncologist* 2008, **13:**477-493.
5. Lewis BP, Burge CB, Bartel DP: **Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets.** *Cell* 2005, **120:**15-20.
6. Bartel DP: **MicroRNAs: genomics, biogenesis, mechanism and function.** *Cell* 2004, **116:**281-297.
7. Blenkiron C, Goldstein LD, Thorne NP, Spiteri I, Chin SF, Dunning MJ, Barbosa Morais NL, Teschendorff AE, Green AR, Ellis IO, Tavaré S, Caldas C, Miska EA: **MicroRNA expression profiling of human breast cancer identifies new markers of tumor subtype.** *Genome Biol* 2007, **8:**R214.

8. Volinia S, Calin GA, Liu CG, Ambs S, Cimmino A, Petrocca F, Visone R, Iorio M, Roldo C, Ferracin M, Prueitt RL, Yanaihara N, Lanza G, Scarpa A, Vecchione A, Negrini M, Harris CC, Croce CM: **A microRNA expression signature of human solid tumors defines cancer gene targets.** *Proc Natl Acad Sci USA* 2006, **103:**2257-2261.
9. Mattie MD, Benz CC, Bowers J, Sensinger K, Wong L, Scott GK, Fedele V, Ginzinger D, Getts R, Haqq C: **Optimized high-throughput microRNA expression profiling provides novel biomarker assessment of clinical prostate and breast cancer biopsies.** *Mol Cancer* 2006, **5:**24.
10. Iorio MV, Ferracin M, Liu CG, Veronese A, Spizzo R, Sabbioni S, Magri E, Pedriali M, Fabbri M, Campiglio M, Ménard S, Palazzo JP, Rosenberg A, Musiani P, Volinia S, Nenci I, Calin GA, Querzoli P, Negrini M, Croce CM: **MicroRNA gene expression deregulation in human breast cancer.** *Cancer Res* 2005, **65:**7065-7070.
11. Ma L, Teruya-Feldstein J, Weinberg RA: **Tumour invasion and metastasis initiated by microRNA-10b in breast cancer.** *Nature* 2007, **449:**682-688.
12. Tavazoie SF, Alarcón C, Oskarsson T, Padua D, Wang Q, Bos PD, Gerald WL, Massagué J: **Endogenous human microRNAs that suppress breast cancer metastasis.** *Nature* 2008, **451:**147-152.
13. Bhaumik D, Scott GK, Schokrpur S, Patil CK, Campisi J, Benz CC: **Expression of microRNA-146 suppresses NF-kappaB activity with reduction of metastatic potential in breast cancer cells.** *Oncogene* 2008, **27:**5643-5647.
14. Zhu S, Wu H, Wu F, Nie D, Sheng S, Mo YY: **MicroRNA-21 targets tumor suppressor genes in invasion and metastasis.** *Cell Res* 2008, **18:**350-359.
15. Adams BD, Furneaux H, White BA: **The micro-ribonucleic acid (miRNA) miR-206 targets the human estrogen receptor-alpha (ERα) and represses ERα messenger RNA and protein expression in breast cancer cell lines.** *Mol Endocrinol* 2007, **21:**1132-1147.
16. Scott GK, Goga A, Bhaumik D, Berger CE, Sullivan CS, Benz CC: **Coordinate suppression of ERBB2 and ERBB3 by enforced expression of micro-RNA miR-125a or miR-125b.** *J Biol Chem* 2007, **282:**1479-1486.
17. Lu J, Getz G, Miska EA, Alvarez-Saavedra E, Lamb J, Peck D, Sweet-Cordero A, Ebert BL, Mak RH, Ferrando AA, Downing JR, Jacks T, Horvitz HR, Golub TR: **MicroRNA expression profiles classify human cancers.** *Nature* 2005, **435:**834-838.
18. Massague J: **Sorting out breast-cancer gene signatures.** *N Engl J Med* 2007, **356:**294-297.
19. Simon R, Radmacher MD, Dobbin K, McShane LM: **Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification.** *J Natl Cancer Inst* 2003, **95:**14-18.
20. Khan J, Wei JS, Ringnér M, Saal LH, Ladanyi M, Westermann F, Berthold F, Schwab M, Antonescu CR, Peterson C, Meltzer PS: **Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks.** *Nat Med* 2001, **7:**673-679.
21. Lancashire L, Schmid O, Shah H, Ball G: **Classification of bacterial species from proteomic data using combinatorial approaches incorporating artificial neural networks, cluster analysis and principal components analysis.** *Bioinformatics* 2005, **21:**2191-2199.
22. Matharoo-Ball BR, Lancashire L, Ugurel S, Miles AK, Weston DJ, Rees R, Schadendorf D, Ball GR, Creaser CS: **Diagnostic biomarkers differentiating metastatic melanoma patients from healthy controls identified by an integrated MALDI-ToF mass spectrometry/bioinformatic approach.** *Proteomics* 2007, **1:**605-620.
23. Ball G, Mian S, Holding F, Allibone RO, Lowe J, Ali S, Li G, McCardle S, Ellis IO, Creaser C, Rees RC: **An integrated approach utilizing artificial neural networks and SELDI mass spectrometry for the classification of human tumours and rapid identification of potential biomarkers.** *Bioinformatics* 2002, **18:**395-404.
24. Castoldi M, Benes V, Hentze MW, Muckenthaler MU: **miChip: a microarray platform for expression profiling of microRNAs based on locked nucleic acid (LNA) oligonucleotide capture probes.** *Methods* 2007, **43:**146-152.
25. Bishop C: *Neural Networks for Pattern Recognition* Oxford: Oxford University Press; 1995.
26. Davoren PA, McNeill RE, Lowery AJ, Kerin MJ, Miller N: **Identification of suitable endogenous control genes for microRNA gene expression analysis in human breast cancer.** *BMC Mol Biol* 2008, **9:**76.
27. Livak KJ, Schmittgen TD: **Analysis of relative gene expression data using real time quantitative PCR and the 2(-$\Delta\Delta$C(T)) method.** *Methods* 2001, **25:**402-408.
28. Hayes DF, Thor AD, Dressler LG, Weaver D, Edgerton S, Cowan D, Broadwater G, Goldstein LJ, Martino S, Ingle JN, Henderson IC, Norton L, Winer EP, Hudis CA, Ellis MJ, Berry DA, Cancer and Leukemia Group B Investigators: **HER2 and response to paclitaxel in node-positive breast cancer.** *N Engl J Med* 2007, **357:**1496-1506.
29. De Laurentiis M, Arpino G, Massarelli E, Ruggiero A, Carlomagno C, Ciardiello F, Tortora G, D'Agostino D, Caputo F, Cancello G, Montagna E, Malorni L, Zinno L, Lauria R, Bianco AR, De Placido S: **A meta-analysis on the interaction between HER-2 expression and response to endocrine treatment in advanced breast cancer.** *Clin Cancer Res* 2005, **11:**4741-4748.
30. Michiels S, Koscielny S, Hill C: **Prediction of cancer outcome with microarrays: a multiple random validation strategy.** *Lancet* 2005, **365:**488-492.
31. Phung SL, Bouzerdoum A: **A pyramidal neural network for visual pattern recognition.** *IEEE Trans Neural Netw* 2007, **18:**329-343.
32. Mizuno HK, Yajima M, Komoda HN: **Application of neural network to technical analysis of stock market prediction.** *Studies Informatics Control* 1998, **7:**111-120.
33. DiRusso SM, Chahine AA, Sullivan T, Risucci D, Nealon P, Cuff S, Savino J, Slim M: **Development of a model for prediction of survival in pediatric trauma patients: comparison of artificial neural networks and logistic regression.** *J Pedriatr Surg* 2002, **37:**1098-1104.
34. Matharoo-Ball B, Hughes C, Lancashire L, Tooth D, Ball G, Creaser C, Elgasim M, Rees R, Layfield R, Atiomo W: **Characterization of biomarkers in polycystic ovary syndrome (PCOS) using multiple distinct proteomic platforms.** *J Proteome Res* 2007, **6:**3321-3328.
35. Matharoo-Ball B, Ball G, Rees R: **Clinical proteomics: discovery of cancer biomarkers using mass spectrometry and bioinformatics approaches – a prostate cancer perspective.** *Vaccine* 2007, **25(Suppl 2):**B110-B121.
36. Lancashire LJ, Rees RC, Ball GR: **Identification of gene transcript signatures predictive for estrogen receptor and lymph node status using a stepwise forward selection artificial neural network modelling approach.** *Artif Intell Med* 2008, **43:**99-111.
37. Szafranska AE, Davison TS, John J, Cannon T, Sipos B, Maghnouj A, Labourier E, Hahn SA: **MicroRNA expression alterations are linked to tumorigenesis and non-neoplastic processes in pancreatic ductal adenocarcinoma.** *Oncogene* 2007, **26:**4442-4452.
38. Cheng AM, Byrom MW, Shelton J, Ford LP: **Antisense inhibition of human miRNAs and indications for an involvement of miRNA in cell growth and apoptosis.** *Nucleic Acids Res* 2005, **33:**1290-1297.
39. Martinez I, Gardiner AS, Board KF, Monzon FA, Edwards RP, Khan SA: **Human papillomavirus type 16 reduces the expression of microRNA-218 in cervical carcinoma cells.** *Oncogene* 2008, **27:**2575-2582.
40. Schetter AJ, Leung SY, Sohn JJ, Zanetti KA, Bowman ED, Yanaihara N, Yuen ST, Chan TL, Kwong DL, Au GK, Liu CG, Calin GA, Croce CM, Harris CC: **MicroRNA expression profiles associated with prognosis and therapeutic outcome in colon adenocarcinoma.** *JAMA* 2008, **299:**425-436.
41. Dixon-McIver A, East P, Mein CA, Cazier JB, Molloy G, Chaplin T, Andrew Lister T, Young BD, Debernardi S: **Distinctive patterns of microRNA expression associated with karyotype in acute myeloid leukaemia.** *PLoS ONE* 2008, **3:**e2141.
42. Grady WM, Parkin RK, Mitchell PS, Lee JH, Kim YH, Tsuchiya KD, Washington MK, Paraskeva C, Willson JK, Kaz AM, Kroh EM, Allen A, Fritz BR, Markowitz SD, Tewari M: **Epigenetic silencing of the intronic microRNA hsa-miR-342 and its host gene EVL in colorectal cancer.** *Oncogene* 2008, **27:**3880-3888.
43. Huang Q, Gumireddy K, Schrier M, le Sage C, Nagel R, Nair S, Egan DA, Li A, Huang G, Klein-Szanto AJ, Gimotty PA, Katsaros D, Coukos G, Zhang L, Puré E, Agami R: **The microRNAs miR-373 and miR-520c promote tumour invasion and metastasis.** *Nat Cell Biol* 2008, **10:**202-210.

44. Kovacs G, Frisch S: **Clonal chromosome abnormalities in tumor cells from patients with sporadic renal cell carcinomas.** *Cancer Res* 1989, **49:**651-659.
45. Takayama H, Suzuki T, Mugishima H, Fujisawa T, Ookuni M, Schwab M, Gehring M, Nakamura Y, Sugimura T, Terada M: **Deletion mapping of chromosomes 14q and 1p in human neuroblastoma.** *Oncogene* 1992, **7:**1185-1189.
46. Bando T, Kato Y, Ihara Y, Yamagishi F, Tsukada K, Isobe M: **Loss of heterozygosity of 14q32 in colorectal carcinoma.** *Cancer Genet Cytogenet* 1999, **111:**161-165.
47. Chang WY, Cairns P, Schoenberg MP, Polascik TJ, Sidransky D: **Novel suppressor loci on chromosome 14q in primary bladder cancer.** *Cancer Res* 1995, **55:**3246-3249.
48. Bandera CA, Takahashi H, Behbakht K, Liu PC, LiVolsi VA, Benjamin I, Morgan MA, King SA, Rubin SC, Boyd J: **Deletion mapping of two potential chromosome 14 tumor suppressor gene loci in ovarian carcinoma.** *Cancer Res* 1997, **57:**513-515.
49. Tse JY, Ng HK, Lau KM, Lo KW, Poon WS, Huang DP: **Loss of heterozygosity of chromosome 14q in low- and high-grade meningiomas.** *Hum Pathol* 1997, **28:**779-785.
50. Kerangueven F, Noguchi T, Coulier F, Allione F, Wargniez V, Simony-Lafontaine J, Longy M, Jacquemier J, Sobol H, Eisinger F, Birnbaum D: **Genomewide search for loss of heterozygosity shows extensive genetic diversity of human breast carcinomas.** *Cancer Res* 1997, **57:**5469-5474.
51. Baskerville S, Bartel DP: **Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes.** *RNA* 2005, **11:**241-247.
52. Zhang R, Wang YQ, Su B: **Molecular evolution of a primate-specific microRNA family.** *Mol Biol Evol* 2008, **25:**1493-1502.
53. Diaz LK, Zhou X, Wright ET, Cristofanilli M, Smith T, Yang Y, Sneige N, Sahin A, Gilcrease MZ: **CD44 expression is associated with increased survival in node-negative invasive breast carcinoma.** *Clin Cancer Res* 2005, **11:**3309-3314.
54. Griffiths-Jones S, Saini HK, Van Dongen S, Enright AJ: **miRBase: tools for microRNA genomics.** *Nucleic Acids Research* 2008:D154-D158.
55. Robey RW, To KK, Polgar O, Dohse M, Fetsch P, Dean M, Bates SE: **ABCG2: a perspective.** *Adv Drug Deliv Rev* 2009, **61:**3-13.
56. Burger H, Foekens JA, Look MP, Meijer-van Gelder ME, Klijn JG, Wiemer EA, Stoter G, Nooter K: **RNA expression of breast cancer resistance protein, lung resistance-related protein, multidrug resistance-associated proteins 1 and 2, and multidrug resistance gene 1 in breast cancer: correlation with chemotherapeutic response.** *Clin Cancer Res* 2003, **9:**827-836.
57. Ee PL, Kamalakaran S, Tonetti D, He X, Ross DD, Beck WT: **Identification of a novel estrogen response element in the breast cancer resistance protein (ABCG2) gene.** *Cancer Res* 2004, **64:**1247-1251.
58. Wang H, Lee EW, Zhou L, Leung PC, Ross DD, Unadkat JD, Mao Q: **Progesterone receptor (PR) isoforms PRA and PRB differentially regulate expression of the breast cancer resistance protein in human placental choriocarcinoma BeWo cells.** *Mol Pharmacol* 2008, **73:**845-854.
59. Imai Y, Ishikawa E, Asada S, Sugimoto Y: **Estrogen-mediated post transcriptional down-regulation of breast cancer resistance protein/ABCG2.** *Cancer Res* 2005, **65:**596-604.
60. Wang H, Zhou L, Gupta A, Vethanayagam RR, Zhang Y, Unadkat JD, Mao Q: **Regulation of BCRP/ABCG2 expression by progesterone and 17β-estradiol in human placental BeWo cells.** *Am J Physiol Endocrinol Metab* 2006, **290:**E798-E807.
61. Pan YZ, Morris ME, Yu AM: **MicroRNA-328 negatively regulates the expression of breast cancer resistance protein (BCRP/ABCG2) in human cancer cells.** *Mol Pharmacol* 2009, **75:**1374-1379.
62. To KK, Zhan Z, Litman T, Bates SE: **Regulation of ABCG2 expression at the 3' untranslated region of its mRNA through modulation of transcript stability and protein translation by a putative microRNA in the S1 colon cancer cell line.** *Mol Cell Biol* 2008, **28:**5147-5161.
63. Liao R, Sun J, Zhang L, Lou G, Chen M, Zhou D, Chen Z, Zhang S: **MicroRNAs play a role in the development of human hematopoietic stem cells.** *J Cell Biochem* 2008, **104:**805-817.
64. Ronchetti D, Lionetti M, Mosca L, Agnelli L, Andronache A, Fabris S, Deliliers GL, Neri A: **An integrative genomic approach reveals coordinated expression of intronic miR-335, miR-342, and miR-561 with deregulated host genes in multiple myeloma.** *BMC Med Genomics* 2008, **1:**37.
65. Miller TE, Ghoshal K, Ramaswamy B, Roy S, Datta J, Shapiro CL, Jacob S, Majumder S: **MicroRNA-221/222 confers tamoxifen resistance in breast cancer by targeting p27(Kip1).** *J Biol Chem* 2008, **283:**29897-29903.

Appendix III

Artificial Neural Network Based Algorithm for Biomolecular Interactions Modeling

Christophe Lemetre[1], Lee J. Lancashire[2], Robert C. Rees[1], Graham R. Ball[1]

[1] The van Geest Cancer Research Center, Nottingham Trent University, School of Science and Technology, Clifton campus, Clifton lane, Nottingham NG11 8NS, United Kingdom.
[2] Clinical and Experimental Pharmacology, Paterson Institute for Cancer Research, University of Manchester, Manchester M20 4BX, United Kingdom.

Abstract. With the advent of new genomic platforms there is the potential for data mining of genomic profiles associated with specific subclasses of disease. Many groups have focused on the identification of genes associated with these subclasses. Fewer groups have taken this analysis a stage further to identify potential associations between biomolecules to determine hypothetical inferred biological interaction networks (*e.g.* gene regulatory networks) associated with a given condition (termed the interactome). Here we present an artificial neural network based approach using the back propagation algorithm to explore associations between genes in hypothetical inferred pathways, by iteratively predicting the level of expression of each gene with the others, with respect to the genes associated with metastatic risk in breast cancer based on the publicly available van't Veer data set [1]. We demonstrate that we can identify a subset of genes that is strongly associated with others within the metastatic system. Many of these interactions are strongly representative of likely biological interactions and the interacting genes are known to be associated with metastatic disease.

Keywords: artificial neural networks; breast cancer; metastasis; interactions; interactome.

1 Introduction

1.1 The interactome problem and its potential

Recently, with the advent of gene expression array platform technologies, a large number of groups have focused on the profiling of a range of diseases and conditions. While huge efforts have focused on the generation of data, fewer groups have addressed issues of appropriate analysis. The data generated by microarray platform technologies is non-linear and highly dimensional with significant redundancy. This necessitates analysis strategies that appropriately identify components of functional relevance. To date, many groups have analysed microarray data for the identification of gene signatures that associate with specific clinical questions. For example van't Veer et al. [1] and West et al. [2] focused on genes associated with metastases in breast cancer. However, not all of the potential information contained within these datasets has been deeply investigated and by limiting our efforts to single biomarkers we are attempting to model a real-world system that is dynamic, highly complex and correlated.

More recently groups have considered the use of microarray data for the investigation of gene regulations in terms of a more systems level view of the processes associated with disease. Barabási & Oltva stated that "*it is increasingly clear that a discrete biological function can only rarely be attributed to an individual molecule.*" [3], but on a highly and complex interaction of biomolecules. While the identification of gene signatures remains important, investigating an extension of this in describing how these gene sets change in expression, and subsequently change the expression of other genes of functional relevance with a given clinical question has the potential to elucidate novel disease specific pathways that may be used to identify potential therapeutic targets.

1.2 Interaction studies

Several techniques for the exploration of interaction networks in both proteins and genes have already been described. Examples include likelihood approaches [4] and Bayesian methods [5]. A number of groups focused their studies on the use of functionality to define interactions [6-8]. Schlitt & Brazma present a review of methodologies for gene regulatory networks [9]. However, the main disadvantage of many of these methodologies is that they provide only limited information regarding the nature of the interactions. In many cases only the presence of an interaction is considered. Furthermore, interactions are often only considered in the context of the target, and not within the whole gene marker pool.

To overcome current limitations we propose a novel approach based on back-propagation (BP) artificial neural networks (ANNs) that simultaneously considers all genes in a dataset. This will be introduced using genes associated with distant metastasis in breast cancer explored previously by the van't Veer study [1]. This is achieved by the development and comparison of a series of ANN models, each of which uses all available inputs (genes) to predict the expression of a single omitted input. This is repeated for all single inputs within the set. By examination of the weights and predictive performance of each model a matrix of interactions can be determined. The advantage of this approach over others is that the multi-factorial consideration of each input allows the magnitude of interaction to be determined for a given pair of parameters, whether it be inhibitory, stimulatory, bi or unidirectional. Once the initial screening is complete, non-significant interactions are removed using a decision threshold based upon the absolute values of the association between each input determined by the matrix of interactions. This study adopts an approach based on iterative prediction of each single input expression from all the others in a defined set. The results from each of these models are subsequently integrated into an interaction map based on the weights of each sub model. All sub models used to define the interaction map are extensively validated by Monte Carlo Cross Validation (MCCV). Previously, other groups have only investigated inferences by an analysis of large complicated ANN models associated with particular classes of expression

[10-11]. This study advances upon these methods by allowing the direct prediction of associations between expressions of genes. The problems analysed are broken into small components that are unlikely to suffer from dimensionality issues associate with model complexity. Further, through a process of MCCV, early stopping and optimisation on a validation dataset, the findings are more likely to reflect a real-world solution and not an overfitting of the dataset of interest.

1.3 Artificial Neural Networks

ANNs are from the field of artificial intelligence and can "learn" from patterns by example [12-13], by adapting the connectivity between the nodes of the network. ANNs have been widely used in biology and Lisboa et al. [14] reviews their use in cancer applications. A major advantage of ANNs is their ability to cope with noisy and non-linear data, such as that found in microarray studies. Learning in ANNs occurs by adapting the weights of the connections between nodes of subsequent layers.

1.4 The dataset

The dataset we used here consisted of the gene profiling with microarray from the seminal van't Veer breast cancer study [1]. This was, downloaded in excel format from http://www.rii.com/publications/2002/vantveer.html, and contained 78 samples, each with 24,481 inputs representing the expression ratio of each gene.

2 Materials and Methods

2.1 Data-preprocessing

The dataset was utilized as presented by the van't Veer data set [2]. All the data were normalized to get a coherent dataset. The level of expression of every single gene was rescaled between 0 and 1 through all the cases.

2.2 Pre-screening of the data to determine genes associated with metastatic risk

The ANN algorithm described in [13] and [15] was successfully applied to the dataset to screen microarray data for genes of relevance to a given biological question. Here, we have utilised this approach to select the most important genes that can accurately predict the metastatic risk class described in the van't Veer study [1]. Gene microarray intensities from the arrays were used as single inputs to the ANN model. The output node represented two classes; 0 represented no distant metastases within 5 years, 1 represented cases who developed distant metastases within 5 years. The performance for each single gene model was determined based on prediction accuracy on a test dataset using cross validation. This produced a rank order of genes based on their predictive performance for the metastatic risk class. The top 100 genes defined in this analysis were further used in our interaction algorithm.

2.3 The interaction algorithm

The main idea of the approach lies on the iterative calculation of the influence that multiple variables may have on a single one. Indeed, if we have been able to use all the variables to explain one particular outcome, we can now use one of these variables as an outcome and all the remaining to explain the level of expression of the first one, and therefore be able to find out the influence everyone has on this particular gene expression level. Once the first input is complete, we repeat the procedure for all the other variables.

The interaction algorithm presented in this study is based on a three-layer feed forward Perceptron with a BP algorithm and a sigmoid transfer function [16]. The hidden and output layers consisted of 2 and 1 nodes respectively. Training was performed through 3000 epochs, terminated by either a window of 1000 epochs without improvement of the Mean Squared Error (MSE) on the validation subset, or an MSE threshold of 10^{-2}. Momentum and learning rate were set respectively at 0.5 and 0.1 as previously shown [13]. Training was repeated 50 times for each model, with MCCV applied prior to training to randomly split the patients in three different subsets; here the training set comprised 60% of the cases, and validation and test sets contained 20% each.

To define an interaction map for a given set of genes, the first input gene in the dataset was defined as an output. All others were set as inputs and applied to the ANN architecture and algorithm described above. The weights of the trained ANN model were stored. This process was iteratively repeated for all inputs in the dataset, treating each one in turn as an output. The weights relating a given input to a given output were then analyzed based on the sum of weights leading from an input to the output, to determine the intensity of the relationship between a source (input) and target (output). Analysis of the weights across all of the potential associations provided a rank order of their strength.

2.4 Visualisation of Interaction maps

The concept of the interaction mapping uses the network theory as described previously [3], where a single marker is symbolised by a node and any relation between two markers is represented with a directed edge, and/or an arrow setting one of the node as a source, and the other one as a target. Cytoscape® [17] was used for interaction visualisation.

2.5 Filtering the interaction map to determine the key interactions for the metastatic risk system

A matrix of interactions between variables was generated and extracted. Every association between any pair of variables contained in the dataset was investigated. Most interactions were found to be of non-significance. As this algorithm produces a huge amount of results (*n* inputs giving *n(n-1)* associations), non-significant associations were filtered out so that only the highest absolute values remained. A filter value of 5 was used and the sign of the interaction was preserved so we could determine a direction for the mapped interaction indicating whether it was inhibitory or stimulatory.

3 Results

The approach was tested on the Van't Veer dataset [1]. In the first instance this consisted of screening the rank order of genes in terms of strength of predicting metastasis. The distribution of the genes ranked by their predictive error for test data was obtained (data not shown here), and clearly showed that the top genes have a higher ability to predict, and therefore are most relevant for the system being investigated by interaction analysis. The best predictor in the gene set was the mRNA sequence referred to as NM_001216 which codes for carbonic anhydrase IX (CA9), a protein involved in breast cancer prognosis and relapse [18].

To assess the interactions within the genes associated with prognostic outcome top 100 predictors were submitted to the interaction analysis. These genes had a range of predictive ability for test data between 72% for the first gene to 62.7% for the 100th. Analysis of the resulting weights from the interaction analysis produced a matrix of 9900 potential associations. It clearly appears that no relevant information can be elucidated from this interaction map due to its high dimensionality and complexity.

Figure 1 displays the distribution of the ranked interaction values. The majority of the interactions have low absolute values ranging from -10.9 to +17.2. The distribution of interaction indicates that there are very few strong links compared to the overall population of links. These warranted further investigation.

During the interaction analysis, the overall predictive performance of the models was assessed to define the interaction map by determining the correlation coefficient between actual and predicted levels. Therefore, the error values of the model were constantly assessed to produce a level of confidence for the interactions identified. In this dataset the average r^2 values was 0.665.

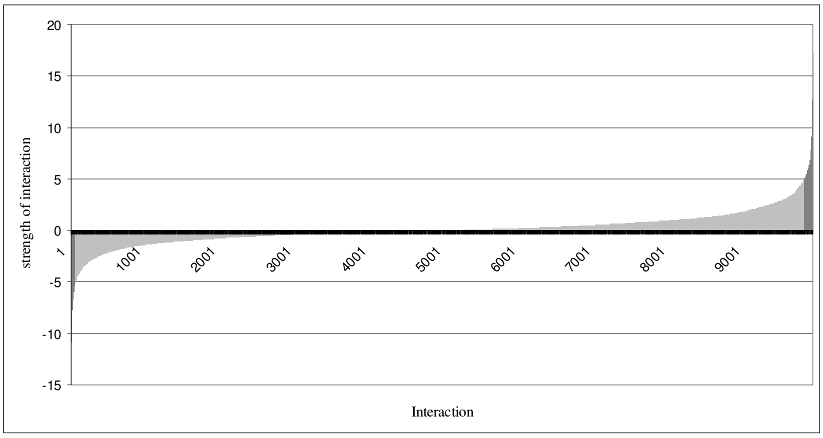


Fig. 1. Distribution of the interaction sorted by values. The interactions in dark grey at the extremities represent the strongest interactions at level 5.

A filter removing links between -5 and +5 was applied which resulted in the number of links decreasing from 10,000 to 163 (1.63% of the original interactome). This greatly simplified the interaction map, facilitating interpretation and understanding of the key features within the global interaction map as shown in Figure 2.

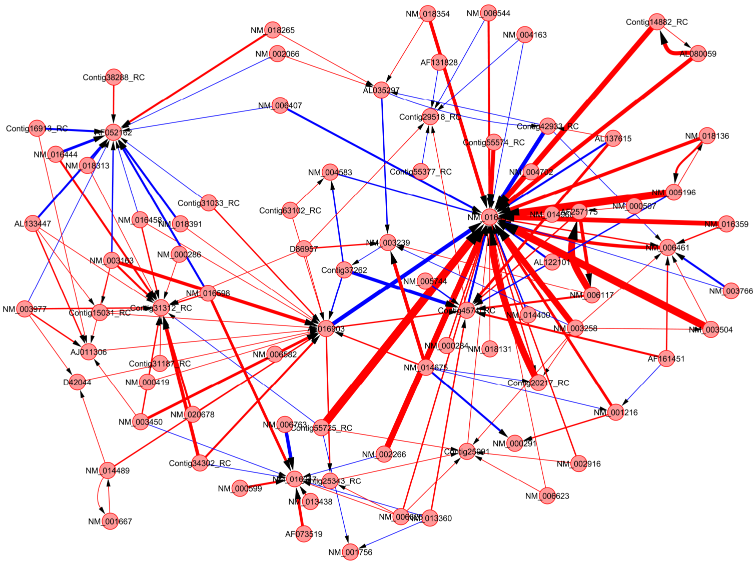


Fig. 2. Display of the interaction map reduced to level 5. Red is a positive interaction (stimulation). Blue is a negative interaction (inhibition).

From this map, we could identify the highest positive interaction involving NM_016448 (Retinoic acid-regulated nuclear matrix-associated protein) which is known to have a role in cancer cell proliferation [19], and Contig42933_RC, which corresponds to the genbank accession number R73468, referring to an unknown gene which has already been strongly associated with metastasis in breast cancer [20]. In the opposite negative direction, the strongest interaction was between NM_016448 and Contig55725_RC, the cell division cycle associated 7 (CDCA7). This has been shown to be over expressed in human tumours [21].

Moreover, we can clearly identify some key nodes involved in dense regions of interactions. Again, NM_016448 (Retinoic acid-regulated nuclear matrix-associated protein) appeared as a key feature within the whole interaction network, interacting strongly (within the strength of interactions over 5) with 35 other genes. Significantly, most of these interactions (33) have NM_016448 as the target.

According to these preliminary observations it appears that Retinoic acid-regulated nuclear matrix-associated protein has a key role in the whole interactome in this breast cancer study. This is in agreement with the literature which has demonstrated its implication in breast cancer [19] with the ability to decrease the expansion of cancer cells. Moreover, the mRNA sequence for CA9, previously identified in the screening step as the best predictor for metastasis in this case series, belongs to one of the features interacting with the Retinoic acid-regulated nuclear matrix-associated protein (NM_016448). In addition, we show an interaction between CA9 and the mRNA for phosphoglycerate kinase 1 (PGK1). Interestingly, interactions between both have already been described in relation to cancer [22].

4 Discussion

Within this study we have sought to develop an algorithm for the exploration of interactions within gene microarray data. This algorithm uses a multifactorial ANN based approach to model interactions.

4.1 Advances

This approach, incorporating an ANN based interaction mapping, offers the potential to identify key components and interactions in a multifactorial fashion. This may overcome limitations from previous studies using simple pairwise interactions, or those based upon biomolecular function, thus facilitating the identification of unidirectional influences. Moreover, the robustness of ANNs allows us to incorporate multiplatform data for exploration, and therefore not only focus on a particular system, but investigate on a wider scale for cross-platform interactions.

4.2 Validation

Earlier using the data supplied by West et al. [2], we have demonstrated that we can identify a rank order of key genes of biological relevance [15]. Similarly we have found a rank order of genes of biological relevance to metastasis here. CAIX, appeared as the strongest feature from the screening step, and was seen to have a strong interaction with PCK1, both already described to interact and to be involved in breast cancer [22]. Moreover, the retinoic acid-regulated nuclear matrix-associated protein, known to have a role in cancer cell proliferation [19], has been found to be strongly influenced within the whole interactome. We have shown other features with biological significance, confirming a level of confidence we can have for the model developed and the overall method.

However, despite these initial promising results, further validation both from the literature (and databases such as Ingenuity, or KEGG) and utilising gene silencing techniques needs to be conducted.

4.3 Limitations

The method is without its limitations. Firstly, the main limitation may come from the origin of the data itself. We are well aware that the collection of experimental data involves a certain level of bias. However, as we can only investigate and explore for relative interaction, none of the interaction is absolute, but relative to all the others within the dataset. This limitation may have an increased influence once the approach is applied to a cross-platform dataset.

Another important limitation comes from the complexity of the overall interactome obtained, and the relevance of the filtering strategy employed. It will be challenging to address a filtering strategy with an acceptable false discovery rate.

4.4 Future work

The filtering strategy will have a key role in the future development of the approach. New approaches are under consideration, such as monitoring the deviation that multiple models can have and therefore filtering out any interactions that are not consistent over multiple runs of the process.

Further validation of the method is fundamental. This will involve databases queries and literature reviewing. Under consideration is an automatised process to query the databases and literature. In addition, comparison with other existing and validated methods will be key, in order to prove the efficiency of this approach. Several alternative methods are under considerations, such as Bayesian Networks [23].

An additional approach that could be further investigated is using the interaction mapping process to investigate the differences between the interactomes of two different classes, for example the metastatic *versus* the non-metastatic breast cancer patients. This may elucidate further features of interest associated with the breast cancer interactome.

5 Conclusion

A novel ANN based approach to predict the interactions which may exist between the components in a given dataset coming from high-throughput technologies has been described. The main advantages of the approach being its multifactorial character, and its reliance upon ANNs to allow for the inclusion of highly dimensional and non-linear data. Furthermore, the approach allows us to employ cross-platform datasets, and therefore we can apply the method to a wider scale. Preliminary results show this to be a novel and powerful tool for interactions discovery, which will allow for the investigation of biological systems and potentially provide a greater understanding of the underlying processes with a view to novel targets discovery.

Acknowledgments. We gratefully acknowledge the support of the John and Lucille van Geest foundation, and of the European Union 6th framework program (Contract nb. LSHC-CT-2004-503306).

References

1. van 't Veer, L.J., Dai, H., van de Vijver, M.J., He, Y.D., Hart, A.A., Mao, M., Peterse, H.L., van der Kooy, K., Marton, M.J., Witteveen, A.T., Schreiber, G.J., Kerkhoven, R.M., Roberts, C., Linsley, P.S., Bernards, R., Friend, S.H.: Gene expression profiling predicts clinical outcome of breast cancer. Nature 415(6871),530--536 (2002)
2. West, M., Blanchette, C., Dressman, H., Huang, E., Ishida, S., Spang, R., Zuzan, H., Olson, J.A. Jr., Marks, J.R., Nevins, J.R.: Predicting the clinical status of human breast cancer by using gene expression profiles. PNAS 98(20), 11462--11467 (2001)
3. Barabási, A.L, Oltvai, Z.N.: Network biology: understanding the cell's; functional organization. Nat. Rev. Genet. 5(2), 101--113, (2004)
4. Liu, Y., Liu, N., Zhao, H.: Inferring protein–protein interactions through high-throughput interaction data from diverse organisms. Bioinformatics 21(15), 3279--3285, (2005)
5. Hartemink, A.J., Gifford, D.K., Jaakkola, T.S., Young, R.A.: Bayesian Methods for Elucidating Genetic Regulatory Networks. IEEE Intelligent Systems, 17(2), 37--43, (2002)
6. Spirin, V., Mirny, L.A.: Protein complexes and functional modules in molecular networks. PNAS 100(21), 12123--12128, (2003)
7. Shoemaker, B.A., Panchenko, A.R.: Deciphering Protein–Protein Interactions. Part II. Computational Methods to Predict Protein and Domain Interaction Partners. PLoS Comput Biol. 3(4), e43, (2007)
8. Schwikowski, B., Uetz, P., Fields S.: A network of protein–protein interactions in yeast. Nat. Biotechnol. 18(12), 1257--1261, (2000)
9. Schlitt, T., Brazma, A.: Current approaches to gene regulatory network modeling. BMC Bioinformatics 8, S9, (2007)
10. Hart, C.E., Mjolsness, E., Wold B.J.: Connectivity in the Yeast Cell Cycle Transcription Network: Inferences from Neural Networks. PLoS Comput Biol. 2(12), 1592—1607, (2006)
11. Rui Xu, Wunsch, D.C., Frank, R.L.: Inference of Genetic Regulatory Networks with Recurrent Neural Network Models Using Particle Swarm Optimization. Computational Biology and Bioinformatics 4(4), 681--692, (2007)
12. Khan, J., Wei J.S., Ringnér M., Saal L.H., Ladanyi M., Westermann F., Berthold F., Schwab M., Antonescu C.R., Peterson C., and Meltzer P.S.: Classification and diagnostic prediction

of cancers using gene expression profiling and artificial neural networks. Nat Med, 7(6), 673--679, (2001)

13.Lancashire, L., Schmid, O., Shah, H., and Ball, G.: Classification of bacterial species from proteomic data using combinatorial approaches incorporating artificial neural networks, cluster analysis and principal components analysis. Bioinformatics, 21(10), 2191--2199, (2005)

14.Lisboa, P.J., and Taktak, A.F.: The use of artificial neural networks in decision support in cancer: A systematic review. Neural Networks, 19(4), 408--415, (2006)

15.Lancashire, L., Rees R.C., and Ball, G.R., Identification of gene transcript signatures predictive for estrogen receptor and lymph node status using a stepwise forward selection artificial neural network modelling approach. Artif Intell Med, 43(2), 99--111, (2008)

16.Rumelhart, D.E., and McClelland, J.L.: Parallel Distribution Processing: Explorations in the Microstructure of Cognition, Vol. 1: Foundations. Cambridge, MA, USA MIT Press, (1986)

17.Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., Ideker, T.: Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res, 13(11), 2498--2504, (2003)

18.Span, P.N., Bussink, J., Manders, P., Beex, L.V., and Sweep, C.G.: Carbonic anhydrase-9 expression levels and prognosis in human breast cancer: association with treatment outcome. British Journal of Cancer ,89(2), 271--276, (2003)

19.Crowe D.L., Kim R., and Chandraratna, R.A.S.: Retinoic Acid Differentially Regulates Cancer Cell Proliferation via Dose-Dependent Modulation of the Mitogen-Activated Protein Kinase Pathway. Molecular Cancer Research 1, 532--540, (2003)

20.Takahashi, H., Masuda, K., Ando, T., Kobayashi, T., and Honda, H.: Prognostic Predictor with Multiple Fuzzy Neural Models Using Expression Profiles from DNA Microarray for Metastases of Breast Cancer. Journal of Bioscience and Bioengineering, 98(3), 193--199, (2004)

21.Osthus, R.C., Karim, B., Prescott, J.E., Smith, B.D., McDevitt, M., Huso, D.L., and Dang C.V.: The Myc Target Gene JPO1/CDCA7 Is Frequently Overexpressed in Human Tumors and Has Limited Transforming Activity In vivo. Cancer Research, 65, 5620--5627, (2005)

22.Winter, S.C., Buffa, F.M., Silva, P., Miller, C., Valentine, H.R., Turley, H., Shah, K.A., Cox, G.J., Corbridge, R.J., Homer, J.J., Musgrove, B., Slevin, N., Sloan, P., Price, P., West, C.M., Harris, A.L.: Relation of a hypoxia metagene derived from head and neck cancer to prognosis of multiple cancers. Cancer Research, 67(7), 3441-3449, (2007)

23.Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N.J., Chung, S., Emili, A., Snyder, M., Greenblatt, J.F., Gerstein, M.: A Bayesian Networks Approach for Predicting Protein-Protein Interactions from Genomic Data. Science, 302(5644), 449-453, (2003)

Appendix IV

BRIEFINGS IN BIOINFORMATICS. VOL 10. NO 3. 315–329
Advance Access publication March 23, 2009

doi:10.1093/bib/bbp012

An introduction to artificial neural networks in bioinformatics—application to complex microarray and mass spectrometry datasets in cancer studies

Lee J. Lancashire, Christophe Lemetre and Graham R. Ball

Submitted: 14th November 2008; Received (in revised form): 13th February 2009

Abstract

Applications of genomic and proteomic technologies have seen a major increase, resulting in an explosion in the amount of highly dimensional and complex data being generated. Subsequently this has increased the effort by the bioinformatics community to develop novel computational approaches that allow for meaningful information to be extracted. This information must be of biological relevance and thus correlate to disease phenotypes of interest. Artificial neural networks are a form of machine learning from the field of artificial intelligence with proven pattern recognition capabilities and have been utilized in many areas of bioinformatics. This is due to their ability to cope with highly dimensional complex datasets such as those developed by protein mass spectrometry and DNA microarray experiments. As such, neural networks have been applied to problems such as disease classification and identification of biomarkers. This review introduces and describes the concepts related to neural networks, the advantages and caveats to their use, examples of their applications in mass spectrometry and microarray research (with a particular focus on cancer studies), and illustrations from recent literature showing where neural networks have performed well in comparison to other machine learning methods. This should form the necessary background knowledge and information enabling researchers with an interest in these methodologies, but not necessarily from a machine learning background, to apply the concepts to their own datasets, thus maximizing the information gain from these complex biological systems.

Keywords: *artificial neural networks; bioinformatics; genomics; mass spectrometry; microarray; proteomics*

INTRODUCTION

The intention of this review is to provide researchers with an understanding of the potential benefits of using artificial neural network (ANN)-based approaches within a biomedical context. They may be applied for classification, predictive modelling and biomarker identification within data sets of high complexity. The focus within this review is on transcript or gene expression data generated from DNA microarray (MA) analysis, or peptide/protein level data generated by mass spectrometry (MS). In ‘Artificial neural networks’ section the concepts behind ANN learning will be introduced and described detailing their advantages and disadvantages. This will include details on how robust models are generated, tested and validated using suitable

The first two authors contributed equally to this work.

Corresponding author. Lee J. Lancashire, Clinical and Experimental Pharmacology, Paterson Institute for Cancer Research, University of Manchester, Manchester M20 4BX, UK. Tel: +44-16-1446-3172; Fax: +44-16-1446-3109; E-mail: llancashire@picr.man.ac.uk

Lee J. Lancashire is a postdoctoral researcher at the Paterson Institute for Cancer Research. He received a PhD in bioinformatics from the Nottingham Trent University in 2006. His main research areas are in machine learning and bioinformatics.

Christophe Lemetre has a degree in computer sciences and bioinformatics and is currently a PhD student under the supervision of Dr Graham Ball, working on machine learning algorithms in bioinformatics.

Graham R. Ball is a reader in bioinformatics at the Nottingham Trent University. He has 16 years experience in the development of biological applications for Artificial Neural Network algorithms.

cross validation approaches. In 'Regularization' section the reader will be made aware of techniques that must be applied during the modelling process in order to obtain reliable results, a principal consideration in highly dimensional datasets. In 'Experimental methods requiring robust bioinformatics' section the MS and MA technologies will be outlined. In 'Data complications in proteomics and genomics' section, issues of high dimensional input data and the importance of reproducibility will be examined. In 'Recent applications' section, examples of publications detailing how ANNs are currently being used in genomic MA and proteomic MS studies will be summarized. 'Comparison to other machine learning methods' section provides highlighted case studies where ANNs have performed favourably in comparison to other common statistical and machine learning methodologies. 'Future trends' section briefly outlines the advanced steps necessary once a validated ANN biomarker signature has been discovered. 'Conclusions' will sum up the review. Researchers with an interest in the potential benefits that ANN approaches may bring to their laboratories should then be able to apply them to their own datasets, maximizing the information to be gained from the analysis of complex biological systems.

Background

There are a number of steps required in order to identify and validate a biomarker so that it can be used in a clinical setting [1], and despite the increasing use of high-throughput technologies such as MS and gene MAs, there remains a lack of clinically useful biomarkers emerging for diseases such as cancer. There may be several reasons for this, such as the reported lack of reproducibility of these approaches [2–5], and the sheer mass of data being generated, which is often extremely noisy, and is becoming progressively complex. This is particularly true in the field of '-omics', where for example, in the recent Affymetrix GeneChip 1.0 ST MAs (designed to target all known and predicted exons in human, mouse and rat genomes), each individual case studied contains information for approximately 1.2 million exon clusters corresponding to over 1.4 million probesets. Thus teasing out the key components from these datasets requires the use of mathematical models running on hardware capable of efficient analyses. The discovery of new biomarkers could facilitate more reliable, efficient and less subjective methods to assist the human expert in the diagnosis of disease, as well as providing new potential targets for future therapies.

With this in mind, it is clear that the identification of new biomarkers still requires a concerted, multidisciplinary effort. This necessitates the requirement for specific computational tools for data-mining, and as such remains a major challenge in bioinformatics [6]. One such tool are ANNs [7], a form of machine learning from the field of artificial intelligence utilized in many areas of bioinformatics and medicine [8] due to their ability to cope with noisy, non-linear and highly dimensional datasets, in particularly when appropriate regularization strategies are employed and when combined with appropriate feature reduction methodologies or forward selection methods such as that proposed in [9]. Using ANNs, it is possible to analyse these sophisticated datasets in identifying novel gene or protein signatures (biomarkers or fingerprints) of biological systems in an endeavour to identify specific phenotypes for diagnosis of disease, establishing a patient's clinical outcome, or even predicting a patient's response to therapy.

ARTIFICIAL NEURAL NETWORKS

ANNs are inspired by the way in which the human brain learns and processes information, with the ability to handle complex (non-linear) features within data in order to generalise and predict well for future cases. Their concept simulates the behaviour of a biological neural network; in humans, learning involves minor adjustments to the synaptic connections between neurons, in ANNs, the learning process is based on the interconnections between the processing elements that constitute the network topology.

McCulloch and Pitts first described the concept of the artificial neuron in 1943 as a mathematical function derived from simulating the basic functions of biological neurons [10]. This manuscript will focus on ANNs in their most common form, the multilayer perceptron (MLP), but other ANN-based approaches exist; for example radial basis function networks and recurrent neural networks. In the MLP, ANNs are organized into several layers, with each layer having a number of respective neurons, or processing elements, that constitute that layer (Figure 1). Simply put, the majority of ANNs have a similar topology consisting of an input layer, one or more hidden layers and an output layer. The number

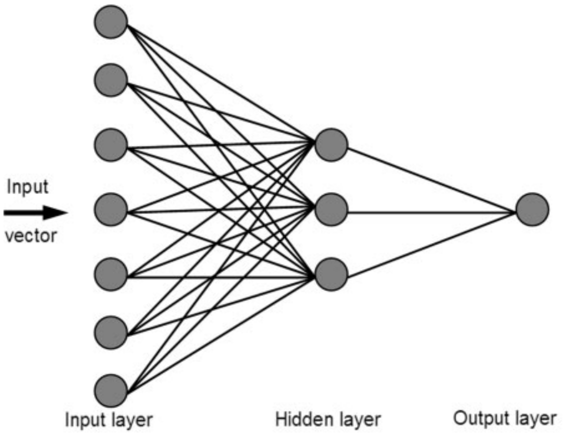


Figure 1: Architecture of a typical multi-layered perceptron artifical neural network.

of hidden layers and the number of neurons in each layer is dependent on the complexity of the problem, i.e. the number of input neurons. The input layer interacts with the external environment to receive the input data as a vector of predictor variables, each represented as a node. This information is passed through to the first hidden layer, and multiplied (thus modified) by a set of associated weights. These products are summed and fed through a non-linear transfer function (e.g. sigmoid, hyperbolic tangent) which scales and then produces an output, similar to the axon of the neuron. The calculation of the output for each neuron is then as follows:

$$v_k = \sum_{i=1}^{n} w_{ki} x_i$$

and

$$y_k = \Phi(v_k + v_{k_0})$$

where x_1, $x_2 \ldots x_{\mathrm{n}}$ are the input signals converging to neuron k. ω_{k1}, $\omega_{k2} \ldots \omega_{k\mathrm{n}}$ are the weights connecting neuron k. v_k is the net input. y_k is the output of the neuron where v_{k0} is a bias term and $\phi(.)$ is the activation function commonly of the form:

$$\Phi(v) = \frac{1}{1 + e^{-v}}$$

for the sigmoid activation function and:

$$\Phi(v) = \frac{e^{v} - e^{-v}}{e^{v} + e^{-v}}$$

for the hyperbolic tangent activation function.

Ultimately this modified information reaches the node(s) in the output layer, the result of which is the output of the entire ANN, for example the predicted class for a given case, or a continuous numerical output in a regression model. In a two group classification problem, the output in the training examples is usually represented as 0 and 1, or −1 and 1. The interconnecting weights are crucial to the system and also enable a variable strength to be given to each input variable included in the model, whether it is excitatory or inhibitory.

ANN learning

ANNs must be trained to efficiently compute the gradient as to be capable of accurately modelling a set of cases and predicting their output. There are two major learning paradigms; supervised and unsupervised. Supervised learning involves providing the network with a set of cases that have values for the inputs as well as the known desired outputs. The output of the network is then compared with the true output to calculate error by assessing the network performance as learning progresses. The interconnecting weights are initially randomized (e.g. [−1, 1]) so that predictions after completion of the first training cycle are essentially random. One of the most popular forms of supervised learning is to compare the error between the true output and the predicted output and then feed this error back through the layers of the network. The weights are adjusted so that after completion of the next training cycle (or epoch) the error decreases according to:

$$\omega_{ki}(\tau) = \eta \delta_k x_i$$

Each weight update ω_{ki} at the current (τ^{th}) cycle is updated in proportion to the input value to which the weight is applied x_i, the error in the output of the unit δ_k and constants known as the learning rate η [11]. The weight change of a neuron is proportional to the influence an input had on the error during training and the learning rate is a constant which controls the size of these weight changes. The larger the learning rate, the faster learning will proceed; however too large a value may lead to non-convergence of the model. Each time a pattern is presented to the network, the weights leading to an output node are modified slightly during learning in the direction required to result in a smaller error the next time the same pattern is presented, until a target error is reached or no improvement of the error is observed. The larger the learning rate, the larger the weight changes and the faster the learning will proceed. If the learning rate is too small, training will be slowed down, however, oscillation or non-convergence can occur if the learning rate is too large [12]. A momentum term, α, may be applied to help prevent the network becoming trapped in local minima, or being stuck along flat regions in error space. This occurs with a slight alteration to the weight update rule by making the weight update on the τ^{th} iteration depend on the update that occurred during the (τ -1^{th}) iteration:

$$\omega_{ki}(\tau) = \eta\delta_k x_i + \alpha\omega_{ki}(\tau - 1)$$

This helps to speed up the time it takes for the network to reach convergence by gradually increasing the step size of the search in regions where the gradient is not changing. As with the learning rate, effectively choosing values for these constants depends on the particular problem of interest and experimentation is important here to find optimal values. In our own experiences for MA and MS data, a learning rate of 0.1 combined with a momentum of 0.5 has proved successful [9, 13]. The target error that needs to be minimized is often determined as the total sum-of-squares based on the difference between the output and target vector as follows:

$$\varepsilon = \frac{1}{2}\sum_{j=1}^{n}(d_j - y_j)^2$$

where n is the number of cases, d_j is the target network output for case $_j$ and y_j is the network predicted output for case j. Alternative error functions also exist, such as the mean squared error, or the maximum conditional likelihood fitting, but will not be dwelt upon here. This learning process is an extension of the generalized delta rule, and is commonly known as back-propagation [14–16]. It is crucially important that the data used in training the network should be reasonably large in order to contain all the information necessary to be able to recognize which of the predictor variables are important amongst the vast amounts of noise and individual variation that is expected to cloud important information in complex '-omics' datasets. If the network outputs fail to show good discrimination on an independent test dataset, over-fitting may have occurred and training must be continued or repeated. Over-fitting can occur when the number of parameters in a model exceeds the number of cases. It is in essence a memorization of the training data (and any associated random noise) [17, 18]. In order for the network to be trained to a satisfactory level which maintains generalization for new data, it is vital to employ an appropriate regularization technique (discussed later in the review). Once learning is complete the weights are stored and can be used to predict future cases in separate test datasets. Other learning algorithms have also been proposed. These include (but are not limited to) QuickProp [19], RPROP [20] and the Levenberg–Marquardt algorithm [21, 22].

Unsupervised learning occurs when the network attempts to map the inputs to outputs without any external assistance. Therefore the network itself governs how it groups the cases based upon the input data. This is sometimes referred to as self organization, and Kohonen's self organizing maps [23] are the most popular form of neural network-based unsupervised learning. Other forms of unsupervised learning include principal components analysis, independent components analysis, hebbian learning and autoassociators. Although unsupervised learning algorithms are an active area of research, it is beyond the scope of this review to explain and review their application in detail and consequently this manuscript will focus on the use of supervised neural networks. For a more detailed discussion on unsupervised pattern recognition in high-throughput genomics and proteomics see [24].

Advantages and disadvantages of artificial neural networks

As ANNs are loosely based on the way a biological neuron is believed to organize and process information, they have many advantages in their ability to derive meaning from large complex datasets. First, they do not rely on data to be normally distributed, an assumption of classical parametric analysis methods. They are able to process data containing complex (non-linear) relationships and interactions that are often too difficult or complex to interpret by conventional linear methods. Another advantage is that they are fault tolerant, i.e. they have the ability of handling noisy or fuzzy information, whilst also being able to endure data which is incomplete or contains missing values. In addition to this (like other machine learning methods), they are capable of generalization, so they can interpret information which is different to that of the training data, thus representing a 'real-world' solution to a given problem by their ability to predict future cases or trends based on what they have previously seen. Thus, trained ANNs can be used as standalone executable systems in order to predict the class of an unknown case of interest, and therefore have the potential application in diagnosis. Finally, there are several techniques that can be used to extract knowledge from trained ANNs, and the importance of individual variables can be easily recovered using various methods such as the analysis of interconnecting network weights [25], sensitivity analysis and rule extraction [26]. This, from a biological perspective, is perhaps one of the most useful aspects of ANN modelling. Gevrey *et al.* [27] review this subject in more depth.

Like all approaches, ANNs also have their limitations. Training of ANNs can potentially be time consuming depending on the complexity of the data being modelled, and as the number of hidden layers required to capture the features of the data increases, so does the time taken for training to complete. As such, only one or two hidden layers are commonly used. Over-fitting may be a problem in ANNs, which is a memorization of the training cases causing the network to perform poorly on future cases. The one major barrier which researchers usually associate with ANNs is that it is not always apparent how they reach a solution, and because of this they have been referred to as 'black boxes' [28–31].

Further limitations originate from the data itself. Experimental data may suffer from high background variation that is difficult for computational algorithms to interpret. The challenges in terms of reproducibility of some technologies has also been investigated [18, 32–39], rendering validation with a separate cohort of samples virtually impossible. The old adage 'garbage in, garbage out' can be strongly applied to modelling with ANNs, and thus the quality of the model output is highly dependent upon the quality of the input data. If the input data is not representative of the 'real world' scenario, the model is compromised. To overcome these issues, several techniques for pre-processing the data have been proposed, and the reader is referred to [40–45] for more examples, and for a guide to considerations regarding study design see [2].

Implementing artificial neural networks

Implementing ANNs is usually performed with statistical computer software packages, or open source equivalents in R (http://www.r-project.org/index.html) and Weka (http://www.cs.waikato.ac.nz/ml/weka/). A comprehensive list of ANN software packages can be found at ftp://ftp.sas.com/pub/neural/FAQ6.html#questions.

REGULARIZATION

Commonly the main purpose of modelling is to simulate a real world system and therefore a model is judged on its ability to generalize to new data. In ANNs the risk of low generalization is mainly attributed to over-training of the model, leading to over-fitting and subsequently poor predictive performance during independent validation. Due to the fact that even a linear model would over-fit in high dimensions, ANNs must be appropriately regularized during training in order to achieve sufficiently high predictive performances. In order to address this, regularization techniques need to be applied during training. Several options for regularization exist and methods can be chosen according to the type of data or generalization performance that is required. This section will now briefly introduce some of the most common forms.

Weight decay

One of the simplest regularization methods to implement is weight decay. In weight decay, the error function includes a penalty term, for example the sum of squared weights and biases multiplied by a decay constant that controls how much the penalty term should affect the resulting error function. Since over-fitted models are more likely to contain unusually large weights, this approach aims at penalizing such large weights, in order to keep weight values smaller than they naturally otherwise would converge at, thus keeping the activation of the neurons in the linear range [7].

Resampling and early stopping

According to Ntzani and Ioannidis [46], independent validation is only conducted in about 10% of MA studies published. Given the fact that these complex datasets are likely to be non-linear in nature, one may not have prior information regarding the intricacies of the data. As such it is vital to estimate the performance of these models on new data in order to be confident that over-fitting has been avoided. It was stated earlier that the back-propagation algorithm should stop training once the network has achieved an acceptable level, however, the question remains as to what is considered to be an acceptable level, and what can be done to ensure that the model will be capable of generalizing to additional future cases. If training is terminated solely on the basis of a set number of iterations the model is at risk of over-fitting. The most universal approach to address this problem is resampling. Typically in ANN-resampling approaches, the data is split into different subsets, where a percentage of the total sample set is used to train and optimize the ANN (the training set) and (sample size permitting) the remaining are partitioned for validation during training (the validation set) and external testing after the modelling is complete (the test set). A crude regularization technique known as the early stopping mechanism monitors the network error with respect to a validation or test dataset. This process signals to stop training either when a predetermined number of iterations have completed, or when the prediction accuracy of the model begins to worsen for the validation or test dataset, a sign of over-fitting. The weights resulting in the minimum validation or test error are then selected. Once the network has completed the learning process, it is further validated using the test data split, to give an unbiased estimation of the networks likely performance on future unseen cases. Examples of this approach can be found in [8] and [12].

Bayesian regularization

The Bayesian regularization approach involves modifying the target function (such as the sum of squared errors) in order to improve the models generalization ability. The Bayesian regularization aims to smooth the cost function by adding to it a regularization parameter based on the sum of squared weights. To reduce bias, the weights and variables of the network are assumed to follow a Gaussian distribution and are assigned prior probabilities, optimized according to the Bayesian framework of Mackay [47]. Network training then attempts to find the trade-off between minimizing the model complexity and model error, as such minimizing both the bias and variance [48]. Methods such as automatic relevance determination will identify and remove unnecessary parameters from the model since the Bayesian approach provides an estimate for the entire distribution of model parameters rather than a single optimal set of weights. Model comparison is based on highest evidence, rather than cross validation, and as such Bayesian regularization maximizes the data available as it does not require a validation set since all the training data can be used for model fitting. A review of Bayesian methods for supervised neural networks can be found in [49], and an example of its application in a microarray study can be found in [48].

Cross validation

There are a number of cross validation approaches used to give an unbiased estimation of the error rate. Examples of these will now be discussed.

First, in Monte Carlo resampling, a training, validation and test set are randomly constituted, with a predetermined number of cases in each subset. All three sets may be randomized, or alternatively the test subset may be kept constant, with the training and validation sets drawn at random a number of times, to enable comparison between models for validation data [7].

Bootstrapping has been shown to be an effective strategy for estimating the error of predictive values in neural network models, and therefore is a reliable approach in determining generalization of the network [50]. In bootstrapping, rather than repeatedly analysing subsets of data (as in the Monte Carlo approach), subsamples of the data are analysed, where many 'pseudo-replicates' are created by resampling the original data. Here, cases are drawn at random from the data set, with equal probability, in order to replicate the process of sampling multiple datasets. The 0.632 bootstrap error estimator has been preferred in small sample microarray classification [51, 52].

k-fold validation is an effective approach when the number of samples is not efficient enough to split the data into three subsets. In a widely used version of this called leave one out cross validation [53, 54], *N* divisions are made (where *N* is the total number of cases in the dataset) and in each division the network is trained on all of the samples except one, which is set aside for test purposes. This process is repeated so that all of the samples are used once for testing. Tenfold validation is commonly used when the number of samples is relatively high (e.g. >100) whilst leave-one-out methods are useful when the training set is lower (e.g. <100) or when the number of features is higher than the number of examples. This multiple cross validation helps to minimize overlap of the test set compared to resampling. For an overview of assessing the accuracy of prediction algorithms for classification problems, the interested reader is directed to [55].

EXPERIMENTAL METHODS REQUIRING ROBUST BIOINFORMATICS

The advent of these high-throughput techniques has increased the potential for identification of new biomarkers massively. These methods facilitate the comprehensive profiling of samples representing disease states. The hurdle to overcome with these technologies is now the sheer complexity of the data generated. This complexity is necessary to represent coverage (or even partial coverage given current technological limitations) of the genome or proteome. MAs are one of the methods commonly used for the high throughput sample profiling at the transcript level, whilst MS is being used to detect changes at the protein level. These technologies are therefore complementary to one another in describing biological systems, and the basic principles will be briefly outlined.

Microarrays

A DNA MA consists of a solid surface, onto which DNA molecules have been chemically bonded. The purpose of MAs is to detect the presence and the abundance of labelled nucleic acids in a given biological sample, which will then hybridize to the DNA on the array, and become detectable via the label. The source of the labelled nucleic acids is the mRNA of the sample of interest, so therefore the purpose of a MA is to measure gene expression. As there may be thousands of different DNA molecules bonded to an array, it is possible to measure the expression of many thousands of genes simultaneously, leading to the potential for extremely high throughput analysis. There are two major types of MA technology used today; firstly cDNA and secondly oligonucleotide arrays, such as those marketed by Affymetrix. For a more detailed explanation of the technology, the reader is referred to [56], or more specifically [57] and [58] for cDNA and oligonucleotide MAs respectively.

Mass spectrometry

MS approaches, more specifically MALDI (matrix-assisted laser/desorption ionization) and a modification of this named SELDI (surface enhanced laser desorption/ionization) TOF (time of flight) MS are now being readily used to generate proteomic profiles of biological samples. Simply, a mass spectrometer consists of an ion source, a mass analyser to measure the mass/charge ratio (*m*/z) of the analytes which have been ionized (mass spectrometers do not measure mass directly, but rather the mass to charge ratio of ions formed), and finally a detector that records the number of ions at each *m/z* value. This generates a spectrum according to the time of flight of the ion, directly related to its mass, or a 'fingerprint' for the sample being analysed. For an overview of the method see [59]. These analyses have an inherent ability to generate profiles consisting of hundreds of thousands of points, with each point representing a protein mass, a peptide mass or a fragment of the above. This high dimensionality provides an obstacle and limits many analysis methods.

DATA COMPLICATIONS IN PROTEOMICS AND GENOMICS

Dimensionality and complexity

Biological '-omics' datasets are unusual in that there is a very large p (input variables) and relatively small n (cases). As the dimensionality of the input data space ($p \times n$) increases, it becomes exponentially more difficult to find a global optimum for the parameter space. This has been termed 'the curse of dimensionality' [60], and often leads to an input space with many irrelevant or noisy inputs, which coupled with the wide heterogeneity commonly found in biological samples, make it difficult to identify the truly important markers with predictive algorithms performing badly as a result of them modelling extraneous portions of the data space. Conventional statistical theory would indicate that for a valid representation of the population one should have a model where $n > p$, and some rules state that to have confidence in results there should be at least 10 events for each variable [61]. Clearly some form of dimensionality reduction/variable selection algorithm is required to satisfy this, because acquiring a data set containing hundreds of thousands of samples is not feasible. Ma and Huang [62] review the topic of feature selection in bioinformatics, and for a review to approaches for dimensionality reduction in biomarker studies the reader is referred to [63].

Reproducibility

Superimposed on the dimensionality issues are those of data quality. In order to identify biomarkers the data should be reproducible within samples, between sample runs and across multiple instruments (at least instruments of the same model) [64]. This can be optimized through the use of technical and experimental replicates, where filtering and averaging of samples are methods which are commonly used to assess reproducibility and increase the confidence in the profiles for comparison. Technical replicates provide information on the variability that occurs when performing a particular assay, whilst experimental (or biological) replicates give a measure of the natural sample to sample variation. Lack of reproducibility decreases the validity of markers and makes validation and ultimately clinical use difficult [65]. Low reproducibility within the data adds to the issues of dimensionality by making the relevant features within data sparser with respect to the overall noise. Low replication and poor data quality can lead to the introduction of features not representative of disease, but of sample run, sample collection, storage and preparation. This introduces random, noisy and unimportant features within the data, further increasing the problem of data analysis.

RECENT APPLICATIONS

This section will now highlight recent applications of ANN technologies in MAs and MS. Since the majority of studies involving the use of ANNs are in tumour diagnosis, the following will focus on the field of cancer. Table 1 summarizes the majority of studies using ANNs with these technologies since 2001, and a selection of these will now be discussed in more detail.

Table 1: Cancer studies using artificial neural networks to analyse microarray and mass spectrometry data since 2001

Platform	Cancer type	Number of cases	Number of classes	References
MA	Astrocytoma	65	2	[86]
MA	Astrocytoma	60	2	[69]
SELDI-TOF	Astrocytoma	12	2	[80]
MA	Breast	58	2	[68, 71]
MA	Breast	10	2	[87]
MA	Breast	49	2	[9]
MA	Breast	78	2	[69, 88]
MA	Breast	15	2	[69]
SELDI-TOF	Breast	40	2	[89]
SELDI-TOF	Breast	82	2	[90]
MA	Colorectal	62	2	[69, 91]
SELDI-TOF	Colorectal	147	2	[82, 92]
SELDI-TOF	Colorectal	93	2	[83]
MA	Oesophageal	28	2	[93]
MA	Leukaemia	72	2	[94]
MA	Leukaemia	64	2	[95]
MA	Leukaemia	38	2	[69]
MA	Leukaemia	57	3	[69]
MALDI-TOF	Liver	132	2	[84]
SELDI-TOF	Liver	106	2	[96]
SELDI-TOF	Liver	182	2	[97]
MA	Lung	32	2	[69]
MA	Lymphoma	40	2	[72]
MA	Lymphoma	220	2	[75]
MALDI-TOF	Melanoma	100	2	[65]
SELDI-TOF	Melanoma	205	2	[85]
MA	Myeloma	105	2	[77]
MA	Neuroblastoma	56	2	[76]
MA	Ovarian	54	2	[98]
MA	Prostate	102	2	[69]
SELDI-TOF	Renal	138	2	[81]
MA	SRBCT	88	4	[66–70]

SRBCT: small round blue cell tumours.

Genomics

The seminal paper by Khan *et al.* [66] was perhaps the first major application showing the potential advantages of using ANNs for these complex datasets. Here they used principal components analysis (PCA) followed by ANNs to classify 88 round blue-cells tumours into four diagnostic categories based on cDNA MA analysis of over 6000 genes. Due to the high accuracy of the models developed the authors eluded to the potential use of ANN-based methodologies 'as an adjunct to routine histological diagnosis'. This dataset was made available for the scientific community to download and has since formed the basis for several more studies using various ANN-based algorithms in the successful classification of these samples [67–70].

In [71], Gruvberger and colleagues used PCA for dimensionality reduction followed by ANN analysis to predict the oestrogen receptor (ER) status of 58 tumours from their gene expression profiles. Here they performed a series of classifications using different sets of 100 genes and showed the ANN performance to be good discriminators on this data. As a result of using ANNs, they hypothesized that the classification was not only controlled by a few differentially expressed genes, but a more complex expression pattern existed involving a larger number of genes.

In predicting long term survival of 40 patients with large B-cell lymphoma, O'Neill and Song [72] used the data generated by Alizadeh *et al.* [73] containing 12 078 transcripts representing expression levels for 4026 genes. This was the first time ANNs were shown to have the ability to perfectly classify (100% accuracy) this type of high dimensional data, and also provided a robust solution for reducing unknown noise and redundancies in datasets whilst maintaining correct classifications.

Using the data made accessible by Rosenwald *et al.* [74], Ando and co-workers [75] described the use of fuzzy neural networks as an approach to variable selection in the expression profiling of 220 diffuse large B-cell lymphoma patients in an effort to predict survival from 7384 genes. Here, using just four genes, ANNs were shown to predict outcome with a classification accuracy of 73%. The analysis in the original manuscript achieved a lower accuracy using more genes in a Cox model. Moreover, the authors showed that by increasing the number of genes in their model to 35 (many of which were clinically relevant to the prognosis of lymphoma), the accuracy increased to 91%. They were able to extract informative rules from their models, with a view to using these approaches in future approaches focused on personalized medicine.

Wei *et al.* [76] used cDNA MAs to analyse 56 tumour samples from patients with neuroblastoma. Total 37 920 data points for each of the samples remained to be analysed after the removal of poor quality data. Due to this complexity, the authors chose to utilize the power of ANNs in order to develop a predictor of survival. Using all of the data in a model, high accuracies were achieved (88%). What is more, they proposed an ANN base gene minimization strategy and identified a signature of 19 genes, some of which had previous affiliations as prognostic markers. This subset of 19 genes had the ability to correctly classify 98% of the patients and further partition the patients into subgroups according to survival status. They concluded that ANN-based approaches such as this would allow therapies to be tailored in a patient specific manner according to their gene expression profiles.

Using ANNs to analyse a 7129 gene expression dataset derived from 74 patients diagnosed with multiple myeloma and 31 normal bone marrow cases, Narayanan *et al.* [77] showed how genes that were consistently positive or negatively expressed could be identified from large datasets. They achieved this by using the interconnecting weights of the trained ANN model, and demonstrated how ANNs could be utilized as a powerful method for dimensionality reduction by identifying 39 genes with 100% generalization on unseen cases. Many of these genes had been previously linked to cancer. Furthermore, the authors described how symbolic knowledge can be extracted from these trained ANN models in order to create simple rules. For example, if gene *x* is present then myeloma, and if gene *y* is absent then normal. This made clear the potential for the use of ANNs in a clinical setting.

In one of our own studies [9], we presented a novel stepwise algorithm using ANNs so that optimal predictive gene signatures can be identified from highly complex, noisy and heterogeneous datasets. Using the dataset published by West *et al.* [78] we identified gene subsets highly predictive for ER status and lymph node status in 49 breast cancer cases analysed by MA containing 7129 gene transcript intensities per patient. As with other studies using ANNs, many of these genes had previously been associated with cancer. When the models were applied to a completely separate 88-patient cohort dataset made available by Huang *et al.* [79], accuracies of 88% and 83% were seen for predicting ER and lymph node status respectively. This manuscript also showed how ANNs could be used in the interrogation of predictive biomarkers to provide an insight into how the increased or decreased expression affects the class of interest, enabling rules for molecular classification to be derived.

Proteomics

One of the first major applications of ANNs for the analysis of MS data was in the classification of astrocytoma by Ball *et al* [80]. They showed the early promise of utilizing SELDI-TOF MS technology combined with intensive computer algorithms for protein expression screening in cancer patients. Here ANNs were used to screen ~100 000 data points generated by SELDI-TOF MS, and by scrutinizing the interconnecting network weights, the authors were able to assign a relative importance value to each ion in terms of its contribution to the classification. The top 50 ions were identified, which could be grouped into several sub-groups according to their mass. Furthermore, an additive approach was performed in order to find the optimal combination of ions in terms of predictive ability. This led to the identification of two ions that in combination were able to predict tumour grade with an accuracy of 94%.

Rogers *et al.* [81] also used SELDI-TOF MS in their study on urinary proteins in renal cancer. Here, ANNs were utilized in an effort to detect early onset of disease, and identify indicative biomarkers. Following pre-processing using peak identification, ANN models were built and trained using several types of controls (healthy controls and benign cases combined with healthy controls). Both peak presence/absence (categorical), as well as actual peak intensities (continuous) were used, with the latter shown to be more efficient. This highlighted the importance and extra information gain that is achieved using actual intensity data to capture the heterogeneity in biological systems rather than peak presence/absence. Moreover, this study utilizing ANNs achieved superior results to the urinary protein assays that were available at the time for bladder cancer.

With a current lack of reliable biomarkers for colorectal cancer, Chen *et al.* [82] proposed the use of proteomics combined with ANN analysis for the discovery of key proteins able to distinguish colorectal cancer patients from a healthy population. To achieve this, MS profiles were generated by SELDI-TOF MS for an age and gender matched cohort of 55 colorectal cancer cases and 92 healthy controls. Initially analysis by cluster analysis showed 54 peaks of interest, culminating in the identification of four candidate biomarkers significantly elevated in colorectal cancer patients. These four ions were then used in an ANN model to build a classifier and discriminate healthy controls from cancer. Here, this approach was shown to outperform discriminant analysis and achieve a sensitivity of 91% and specificity of 93%.

Similarly, Ward *et al.* [83] were also interested in data mining SELDI-TOF MS data for reliable biomarkers of colorectal cancer. They performed proteomic profiling on 62 colorectal cancer patients and 31 non-cancer controls. First, feature selection by *t*-test was conducted, with statistically significant differentially expressed peaks selected for ANN training. The final ANN model included seven peaks and was able to classify with high sensitivity (95%) and specificity (91%), and outperformed CEA (a marker of proven benefit in prognosis and benefit) in discriminating colorectal cancer.

In an effort to improve the prognosis of breast cancer patients through early diagnosis, Hu *et al.* [54] also used SELDI-TOF-MS to explore for reliable tumour markers in serum. They performed screening of the serum proteome in 49 breast cancer patients, 51 patients with benign breast diseases and 33 healthy controls. Total 253 mass peaks were identified using discriminant analysis in classifying between breast cancer and benign, and also between breast cancer and benign plus controls. Using a stepwise approach to assess the predictive ability for each peak, an ANN was able to narrow down the number of markers of interest to just four peaks. These were able to accurately predict the outcome of cancer with a sensitivity of 76% and specificity of 90% for the blind test set. This four-peak model did not result in a statistically significant reduction predictive performance compared to the 253 peak model, and therefore the four-peak model was shown to be more parsimonious in discriminating cancer patients from healthy controls.

Luk *et al.* [84] focused their work on hepatocellular carcinoma biomarkers, investigating differentially expressed proteins between tumour and adjacent healthy liver tissue. Here, proteomic profiling was performed using MALDI-TOF MS and 2D gel electrophoresis followed by analysis by ANNs and decision trees. Both techniques proved to be excellent discriminators of the two phenotypes, with ANNs superior in both training and validation data.

Mian *et al.* [85] were interested in profiling the serum proteome in the classification of early and late stage melanoma, and also predicting disease progression. Here, screening of the patients' proteome was performed with MALDI-TOF MS, showing an interesting signal with significantly higher intensity in 25% of the stage IV samples. ANN modelling in the lower mass range of the spectrum was shown to accurately classify between disease stages and also between progressors and non-progressors. Interestingly, when predicting disease progression, this ANN approach was shown to outperform S100-β, a widely utilised correlate of tumour burden in melanoma.

COMPARISON TO OTHER MACHINE LEARNING METHODS

There have been a number of studies comparing ANNs with other statistical and machine learning approaches to data analysis. Some of these will now be briefly reviewed, outlining how ANNs have performed compared to other statistical and machine learning methods when applied to biological data. This brief discussion will include but will not be focused singly on MA and MA methodology benchmarking studies, as few have been published.

Dreiseitl *et al.* [99] compared the ability of *K*NN, logistic regression, ANNs, decision trees and SVMs in classification of skin lesion data. The authors found logistic regression, ANNs and SVMs to give almost identical results, with *k*-nearest neighbours and decision trees performing the worst. Interestingly, even the worst of the five methods (decision trees) achieved sensitivity and specificity values comparable to human experts indicating these approaches may be of use to assist human decisions in the medical arena.

Sargent [100] carried out a review on 28 cases comparing ANNs with other statistical approaches when applied to medium and large data sets with more than 200 cases. ANNs outperformed regression in 36% of the studies, and was outperformed in 14% of the studies, with the results being similar in the remaining cases.

Pal *et al.* [67] used an ANN-based approach in categorizing subgroups of cancer from microarray data. They identified a smaller number of biomarkers when compared to other machine learning tool such as SVMs whilst performing equally well, suggesting ANNs found a more parsimonious solution. The study performed in [101] was a direct comparison of SVMs and ANNs in the detection of mammographic CAD. Overall, the authors found a similar performance between the two techniques, with ANNs slightly outperforming SVMs in detection and diagnosis in the test set of data.

Song and co-workers [102] compared various machine learning techniques to more classical statistical approaches in the prediction of outcome in two datasets. They used ANNs (single and multi-layered), logistic regression, least squares linear separation and support vector machines (SVMs) to determine the risk of death in a population of patients with cardiac problems. They found the multi layered ANN to be consistently better than the other approaches, suggesting that the ability of the ANN to model non-linear data was providing additional information regarding the datasets leading to higher predictive capabilities [103].

Eftekhar and colleagues [104] made a comparison between ANNs and logistic regression models to study patients with head injury trauma. It was reported that ANNs significantly outperformed the logistic models in discrimination and calibration (goodness of fit) in 77.8% of cases but under-performed in 68% of cases when comparing model accuracies.

In the study by Hu *et al.* [54] the authors compared their ANN model with other commonly used machine learning techniques such as SVMs and decision trees. They showed ANNs to be more reliable than the other methods in the discrimination of cancer patients from normal controls from mass spectrometry data.

Shen and Tan [105] used different coding strategies and feature selection methods in comparing SVMs to other machine learning methods on two cancer microarray datasets. Here, ANNs achieved similar results to SVMs and outperformed K-nearest neighbour and C4.5 decision tree approaches.

Another direct comparison between ANNs and SVMs was performed by Romero and Toppo on a variety of benchmark datasets [106]. Overall, ANNs obtained similar accuracies to SVMs and the two approaches remained competitive across the different datasets.

In 2008, Peterson and co-workers [69] performed a comparison of a large number of machine learning methods (including ANNs, SVMs, K-nearest neighbour, linear discriminant analysis and logistic regression) in the classification of DNA microarrays in cancer research. One of the main findings here was that at the greatest level of sample size ANNs out-performed all other methods resulting in the greatest area under the curve.

Judson *et al.* [107] performed a comparison of six machine learning approaches in complex simulated datasets. They showed that, particularly when using a large number of features, ANNs and SVMs were always the top performers, whereas recursive partitioning and regression trees and K-nearest neighbours were always the poorest.

In a study investigating heart rate variability before a Paroxysmal atrial fibrillation event using ANNs and SVMs [108], Chesnokov showed ANNs provided better results in terms of sensitivity, specificity and positive predictive value compared to SVM which became biased towards positive cases.

Muselli and co-workers [109] proposed an ANN-based method for gene selection microarray data. In both the artificial and the real gene expression data, they showed that SVMs exhibited poor performance compared to the ANN-based method.

FUTURE TRENDS

As with the development of a novel therapeutic agent, model systems representing novel biomarker expression signatures (be it gene or protein expression) must be validated carefully and extensively in a medical setting. Trained models of these biomarker signatures need to be incorporated into simple software solutions so that medical practitioners who are unsavvy in machine learning techniques can simply enter the biomarker profiles from their patients and receive an instant prediction with an acceptable degree of confidence. If it can be shown that the application of such models leads to an improvement in medical care towards the holy grail of cost effective 'personalized medicine', then these ANN software applications may be more widely acceptable and made more readily available to assist patient care in a larger number of hospitals and clinics.

CONCLUSIONS

Rapidly advancing technologies in genomics and proteomics have increased the complexity of data being generated, and with that the requirement for robust data mining approaches in order to analyse and extract panels of biomarkers from biological systems. This review introduces one such approach, artificial neural networks, as a robust tool able to digest these datasets and identify the key components (biomarkers), thus providing an increased understanding of the biological system being modelled whilst also pointing out potential therapeutic targets for focusing future research. Representative works in this field and comparisons with other popular statistical and machine learning techniques are highlighted to provide the interested reader with the sufficient background information required so that they can utilize the potential power of these approaches in the modelling of their own complex datasets.

Key Points

- Artificial neural networks are introduced; background theory, advantages and disadvantages are described.
- Modelling considerations when analysing high dimensional data are discussed, specifically the importance of regularization.
- Complex datasets generated by microarray and mass spectrometry experiments are outlined.
- Recent applications of artificial neural networks in analysing microarray and mass spectrometry data for predictive modelling and biomarker identification are reviewed.
- Example studies comparing artificial neural networks to other modelling approaches are highlighted.

Acknowledgements
The John and Lucille van Geest Foundation.

References

1. Rifai N, Gillette MA, Carr SA. Protein biomarker discovery and validation: the long and uncertain path to clinical utility. *Nat Biotechnol* 2006;**24**:971–83.
2. Simon R, Radmacher MD, Dobbin K. Design of studies using DNA microarrays. *Genet Epidemiol* 2002;**23**: 21–36.
3. Check E. Proteomics and cancer: running before we can walk? *Nature* 2004;**429**:496–7.
4. Dickie GL, Fleminger S, Altman DG, *et al*. Statistics notes. *Br Med J* 1994;**309**:539.
5. Garber K. Debate rages over proteomic patterns. *J Natl Cancer Inst* 2004;**96**:816–8.
6. Baldi P, Brunak S. *Bioinformatics: The Machine Learning Approach*, 2nd revised edn. Cambridge, MA: MIT Press, 2001.
7. Bishop C. *Neural Networks for Pattern Recognition*. Oxford: Oxford University Press, 1995.
8. Lisboa PJ, Taktak AF. The use of artificial neural networks in decision support in cancer: A systematic review. *Neural Netw* 2006;**19**(4):408–15.
9. Lancashire LJ, Rees RC, Ball GR. Identification of gene transcript signatures predictive for estrogen receptor and lymph node status using a stepwise forward selection artificial neural network modelling approach, *Artif Intell Med* 2008;**43**(2):99–111.
10. McCulloch WS, Pitts W. A logical calculus of the ideas immanent in nervous activity. 1943. *Bull Math Biol* 1990;**52**: 99–115; discussion 173–97.
11. Mitchell TM. *Machine Learning*. USA: McGraw-Hill Education, 1997.
12. Basheer IA, Hajmeer M. Artificial neural networks: fundamentals, computing, design, and application. *J Microbiol Methods* 2000;**43**:3–31.
13. Lancashire LJ, Mian S, Ellis IO, *et al*. Current developments in the analysis of proteomic data: artificial neural network data mining techniques for the identification of proteomic biomarkers related to breast cancer. *Current Proteomics* 2005; **2**:15–29.
14. Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. *Nature* 1986; **323**:533–6.
15. Rumelhart DE, McClelland JL. *Parallel Distribution Processing: Explorations in the Microstructure of Cognition*. Cambridge, MA: MIT Press, 1986.
16. Werbos P. *Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Science*. Cambridge, MA: Harvard University, 1974.
17. Ransohoff DF. Rules of evidence for cancer molecular-marker discovery and validation. *Nat Rev Cancer* 2004;**4**: 309–14.
18. Simon R, Radmacher MD, Dobbin K, *et al*. Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. *J Natl Cancer Inst* 2003;**95**:14–8.
19. Fahlman SE. Faster-learning variations on back-propagation: An empirical study. In Sejnowski TJ, Hinton GE, Touretzky DS (eds). *1988 Connectionist Models Summer School*. San Mateo, CA: Morgan Kaufmann, 1988.

20. Riedmiller M, Braun H. Direct adaptive method for faster backpropagation learning: the RPROP algorithm. In: The *IEEE International Conference on Neural Networks,* 1993. pp. 586–91.
21. Levenberg K. A method for the solution of certain non-linear problems in least squares. *Quart Appl Math* 1944;**2**: 164–8.
22. Marquardt DW. An algorithm for the least-squares estimation of nonlinear parameters. *SIAM J Appl Math* 1963;**11**: 431–41.
23. Kohonen T. *Self-Organization and Associative Memory*. Berlin: Springer, 1989.
24. Boutros PC, Okey AB. Unsupervised pattern recognition: an introduction to the whys and wherefores of clustering microarray data. *Brief Bioinform* 2005;**6**:331–43.
25. Olden JD, Joy MK, Death RG. An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. *Ecol Modell* 2004;**178**: 389–97.
26. Silva I, Cortez P, Santos MF, *et al*. Rating organ failure via adverse events using data mining in the intensive care unit. *Artif Intell Med* 2008;**43**:179–93.
27. Gevrey M, Dimopoulos I, Lek S. Review and comparison of methods to study the contribution of variables in artificial neural network models. *Ecol Modell* 2003;**160**:249–64.
28. Smith AE, Nugent CD, McClean SI. Evaluation of inherent performance of intelligent medical decision support systems: utilising neural networks as an example. *Artif Intell Med* 2003;**27**:1–27.
29. Tung WL, Quek C, Cheng P. GenSo-EWS: a novel neural-fuzzy based early warning system for predicting bank failures. *Neural Netw* 2004;**17**:567–87.
30. Wall R, Cunningham P, Walsh P, *et al*. Explaining the output of ensembles in medical decision support on a case by case basis. *Artif Intell Med* 2003;**28**:191–206.
31. Duh MS, Walker AM, Ayanian JZ. Epidemiologic interpretation of artificial neural networks. *Am J Epidemiol* 1998; **147**:1112–22.
32. Chiorino G, Mello Grand M, Scatolini M, *et al*. From single gene to integrative molecular concept MAPS: pitfalls and potentials of microarray technology. *J Biol Regul Homeost Agents* 2008;**22**:7–16.
33. Shi L, Jones WD, Jensen RV, *et al*. The balance of reproducibility, sensitivity, and specificity of lists of differentially expressed genes in microarray studies. *BMC Bioinformatics* 2008;**9**(Suppl. 9):S10.
34. Shi L, Perkins RG, Fang H, *et al*. Reproducible and reliable microarray results through quality control: good laboratory proficiency and appropriate data analysis practices are essential. *Curr Opin Biotechnol* 2008;**19**:10–8.
35. Kiehntopf M, Siegmund R, Deufel T. Use of SELDI-TOF mass spectrometry for identification of new biomarkers: potential and limitations. *Clin Chem Lab Med* 2007;**45**: 1435–49.
36. Callesen AK, Christensen R, Madsen JS, *et al*. Reproducibility of serum protein profiling by systematic assessment using solid-phase extraction and matrix-assisted laser desorption/ionization mass spectrometry. *Rapid Commun Mass Spectrom* 2008;**22**:291–300.
37. Callesen AK, Vach W, Jorgensen PE, *et al*. Reproducibility of mass spectrometry based protein profiles for diagnosis of breast cancer across clinical studies: a systematic review. *J Proteome Res* 2008;**7**:1395–402.
38. Schiffer E, Mischak H, Theodorescu D, *et al*. Challenges of using mass spectrometry as a bladder cancer biomarker discovery platform. *World J Urol* 2008;**26**:67–74.
39. Tiss A, Smith C, Camuzeaux S, *et al*. Serum peptide profiling using MALDI mass spectrometry: avoiding the pitfalls of coated magnetic beads using well-established ZipTip technology. *Proteomics* 2007; **7**(Suppl. 1):77–89.
40. Wang J. Computational biology of genome expression and regulation—a review of microarray bioinformatics. *J Environ Pathol Toxicol Oncol* 2008;**27**:157–79.
41. Barla A, Jurman G, Riccadonna S, *et al*. Machine learning methods for predictive proteomics. *Brief Bioinform* 2008;**9**: 119–28.
42. Grant GR, Manduchi E, Stoeckert CJ, Jr. Analysis and management of microarray gene expression data. *Curr Protoc Mol Biol* 2007; Chapter 19: Unit 19.6.
43. Phan JH, Quo CF, Wang MD. Functional genomics and proteomics in the clinical neurosciences: data mining and bioinformatics. *Prog Brain Res* 2006;**158**:83–108.
44. Wong JW, Cagney G, Cartwright HM. SpecAlign – processing and alignment of mass spectra datasets. *Bioinformatics* 2005;**21**:2088–90.
45. Wong JW, Durante C, Cartwright HM. Application of fast Fourier transform cross-correlation for the alignment of large chromatographic and spectral datasets. *Anal Chem* 2005;**77**:5655–61.
46. Ntzani EE, Ioannidis JP. Predictive ability of DNA microarrays for cancer outcomes and correlates: an empirical assessment. *Lancet* 2003;**362**:1439–44.
47. Mackay DJC. A practical Bayesian framework for back-propagation networks. *Neural Comput* 1992;**4**:448–72.
48. Kelemen A, Liang WL. Bayesian regularized neural network for multiple gene expression pattern classification. *Proc Int Joint Conf Neural Networks* 2003;**1–4**:654–9.
49. MacKay DJC. Probable networks and plausible predictions—a review of practical Bayesian methods for supervised neural networks. *Network: Comput Neural Syst* 1995;**6**: 469–505.
50. Tibshirani R. A comparison of some error estimates for neural network models. *Neural Comput* 1996;**8**:152–63.
51. Ambroise C, McLachlan GJ. Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc Natl Acad Sci USA* 2002;**99**:6562–6.
52. Braga-Neto UM, Dougherty ER. Is cross-validation valid for small-sample microarray classification? *Bioinformatics* 2004;**20**:374–80.
53. Braga-Neto U, Dougherty E. Exact performance of error estimators for discrete classifiers. *Pattern Recogn* 2005;**38**: 1799–814.
54. Hu Y, Zhang S, Yu J, *et al*. SELDI-TOF-MS: the proteomics and bioinformatics approaches in the diagnosis of breast cancer. *Breast* 2005;**14**:255.
55. Baldi P, Brunak S, Chauvin Y, *et al*. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics* 2000;**16**:412–24.
56. Brown PO, Botstein D. Exploring the new world of the genome with DNA microarrays. *Nat Genet* 1999;**21**:33–7.

57. Schena M, Shalon D, Davis RW, *et al.* Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 1995;**270**:467–70.

58. Lipshutz RJ, Fodor SP, Gingeras TR, *et al.* High density synthetic oligonucleotide arrays. *Nat Genet* 1999; **21**:20–4.

59. Petricoin EF, Ardekani AM, Hitt BA, *et al.* Use of proteomic patterns in serum to identify ovarian cancer. *Lancet* 2002;**359**:572–7.

60. Bellman RE. *Adaptive Control Processes. Princeton, NJ:* Princeton University Press, 1961.

61. Katz MH. Multivariable analysis: a primer for readers of medical research. *Ann Intern Med* 2003;**138**:644–50.

62. Ma S, Huang J. Penalized feature selection and classification in bioinformatics. *Brief Bioinform* 2008;**9**:392–403.

63. Hilario M, Kalousis A. Approaches to dimensionality reduction in proteomic biomarker studies. *Brief Bioinform* 2008;**9**:102–18.

64. Diamandis EP. Mass spectrometry as a diagnostic and a cancer biomarker discovery tool: opportunities and potential limitations. *Mol Cell Proteomics* 2004;**3**:367–78.

65. Matharoo-Ball B, Ratcliffe L, Lancashire L, *et al.* Diagnostic biomarkers differentiating metastatic melanoma patients from healthy controls identified by an integrated MALDI-TOF mass spectrometry/bioinformatics approach. *Proteomics Clin Appl* 2007;**1**:605–20.

66. Khan J, Wei JS, Ringner M, *et al.* Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat Med* 2001;**7**: 673–9.

67. Pal NR, Aguan K, Sharma A, *et al.* Discovering biomarkers from gene expression data for predicting cancer subgroups using neural networks and relational fuzzy clustering. *BMC Bioinformatics* 2007;**8**:5.

68. Peterson C, Ringner M. Analyzing tumor gene expression profiles. *Artif Intell Med* 2003;**28**:59–74.

69. Peterson LE, Coleman MA. Machine learning-based receiver operating characteristic (ROC) curves for crisp and fuzzy classification of DNA microarrays in cancer research. *Int J Approx Reason* 2008;**47**:17–36.

70. Xuan J, Wang Y, Dong Y, *et al.* Gene selection for multiclass prediction by weighted fisher criterion. *EURASIP J Bioinform Syst Biol* 2007;**2007**:Article No. 64628.

71. Gruvberger S, Ringner M, Chen Y, *et al.* Estrogen receptor status in breast cancer is associated with remarkably distinct gene expression patterns. *Cancer Res* 2001;**61**:5979–84.

72. O'Neill MC, Song L. Neural network analysis of lymphoma microarray data: prognosis and diagnosis near-perfect. *BMC Bioinformatics* 2003;**4**:13.

73. Alizadeh AA, Eisen MB, Davis RE, *et al.* Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* 2000;**403**:503–11.

74. Rosenwald A, Wright G, Chan WC, *et al.* The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *N Engl J Med* 2002;**346**: 1937–47.

75. Ando T, Suguro M, Kobayashi T, *et al.* Multiple fuzzy neural network system for outcome prediction and classification of 220 lymphoma patients on the basis of molecular profiling. *Cancer Sci* 2003;**94**:906–13.

76. Wei JS, Greer BT, Westermann F, *et al.* Prediction of clinical outcome using gene expression profiling and artificial neural networks for patients with neuroblastoma. *Cancer Res* 2004;**64**:6883–91.

77. Narayanan A, Keedwell EC, Gamalielsson J, *et al.* Single-layer artificial neural networks for gene expression analysis. *Neurocomputing* 2004;**61**:217–40.

78. West M, Blanchette C, Dressman H, *et al.* Predicting the clinical status of human breast cancer by using gene expression profiles. *Proc Natl Acad Sci USA* 2001;**98**: 11462–7.

79. Huang E, Cheng SH, Dressman H, *et al.* Gene expression predictors of breast cancer outcomes. *Lancet* 2003;**361**: 1590–6.

80. Ball G, Mian S, Holding F, *et al.* An integrated approach utilizing artificial neural networks and SELDI mass spectrometry for the classification of human tumours and rapid identification of potential biomarkers. *Bioinformatics* 2002;**18**: 395–404.

81. Rogers MA, Clarke P, Noble J, *et al.* Proteomic profiling of urinary proteins in renal cancer by surface enhanced laser desorption ionization and neural-network analysis: identification of key issues affecting potential clinical utility. *Cancer Res* 2003;**63**:6971–83.

82. Chen YD, Zheng S, Yu JK, *et al.* Artificial neural networks analysis of surface-enhanced laser desorption/ionization mass spectra of serum protein pattern distinguishes colorectal cancer from healthy population. *Clin Cancer Res* 2004; **10**:8380–5.

83. Ward DG, Suggett N, Cheng Y, *et al.* Identification of serum biomarkers for colon cancer by proteomic analysis. *Br J Cancer* 2006;**94**:1898–905.

84. Luk JM, Lam BY, Lee NP, *et al.* Artificial neural networks and decision tree model analysis of liver cancer proteomes. *Biochem Biophys Res Commun* 2007;**361**:68–73.

85. Mian S, Ugurel S, Parkinson E, *et al.* Serum proteomic fingerprinting discriminates between clinical stages and predicts disease progression in melanoma patients. *J Clin Oncol* 2005;**23**:5088–93.

86. Petalidis LP, Oulas A, Backlund M, *et al.* Improved grading and survival prediction of human astrocytic brain tumors by artificial neural network analysis of gene expression microarray data. *Mol Cancer Ther* 2008;**7**:1013–24.

87. Ellis M, Davis N, Coop A, *et al.* Development and validation of a method for using breast core needle biopsies for gene expression microarray analyses. *Clin Cancer Res* 2002;**8**:1155–66.

88. Blazodonakis ME, Zervakis M. Support vector machines and neural networks as marker selectors in cancer gene analysis. *Studies in Computational Intelligence* 2008;**109**: 237–58.

89. Mian S, Ball G, Hornbuckle J, *et al.* A prototype methodology combining surface-enhanced laser desorption/ionization protein chip technology and artificial neural network algorithms to predict the chemoresponsiveness of breast cancer cell lines exposed to Paclitaxel and Doxorubicin under in vitro conditions. *Proteomics* 2003;**3**: 1725–37.

90. Hu Y, Zhang SZ, Yu JK, *et al.* Diagnostic application of serum protein pattern and artificial neural network software in breast cancer. *Ai Zheng* 2005;**24**:67–71.

91. Kim K-J, Cho S-B. Prediction of colon cancer using an evolutionary neural network. *Neurocomputing* 2004;**61**: 361–79.

92. Chen YD, Zheng S, Yu JK, *et al*. Application of serum protein pattern model in diagnosis of colorectal cancer. *Zhonghua Zhong Liu Za Zhi* 2004;**26**:417–20.

93. Kan T, Shimada Y, Sato F, *et al*. Prediction of lymph node metastasis with use of artificial neural networks based on gene expression profiles in esophageal squamous cell carcinoma. *Ann Surg Oncol* 2004;**11**:1070–8.

94. Bicciato S, Pandin M, Didone G, *et al*. Pattern identification and classification in gene expression data using an autoassociative neural network model. *Biotechnol Bioeng* 2003;**81**:594–606.

95. Catchpoole D, Lail A, Dachuan G, *et al*. Gene expression profiles that segregate patients with childhood acute lymphoblastic leukaemia: An independent validation study identifies that endoglin associates with patient outcome. *Leukemia Res* 2007;**31**:1741–7.

96. Wang J-X, Zhang B, Yu J-K, *et al*. Using ANN and serum protein pattern models in liver cancer diagnosis. *Natl Med J China* 2005;**85**:189–92.

97. Ward DG, Cheng Y, N'Kontchou G, *et al*. Changes in the serum proteome associated with the development of hepatocellular carcinoma in hepatitis C-related cirrhosis. *Br J Cancer* 2006;**94**:287–92.

98. Tan TZ, Quek C, Ng GS, *et al*. Ovarian cancer diagnosis with complementary learning fuzzy neural network. *Artif Intell Med* 2008;**43**:207–222.

99. Dreiseitl S, Ohno-Machado L, Kittler H, *et al*. A comparison of machine learning methods for the diagnosis of pigmented skin lesions. *J Biomed Inform* 2001;**34**:28–36.

100. Sargent DJ. Comparison of artificial neural networks with other statistical approaches: results from medical data sets. *Cancer* 2001;**91**:1636–42.

101. Garcia-Orellana CJ, Gallardo-Caballero R, Macias-Macias M, *et al*. SVM and neural networks comparison in mammographic CAD. *Conf Proc IEEE Eng Med Biol Soc* 2007;**2007**:3204–7.

102. Song X, Mitnitski A, Cox J, *et al*. Comparison of machine learning techniques with classical statistical models in predicting health outcomes. *Medinfo* 2004;**11**: 736–40.

103. Chernushevich IV, Loboda AV, Thomson BA. An introduction to quadrupole-time-of-flight mass spectrometry. *J Mass Spectrom* 2001;**36**:849–65.

104. Eftekhar B, Mohammad K, Ardebili HE, *et al*. Comparison of artificial neural network and logistic regression models for prediction of mortality in head trauma based on initial clinical data. *BMC Med Inform Decis Mak* 2005;**5**:3.

105. Shen L, Tan EC. Reducing multiclass cancer classification to binary by output coding and SVM. *Comput Biol Chem* 2006;**30**:63–71.

106. Romero E, Toppo D. Comparing support vector machines and feedforward neural networks with similar hidden-layer weights. *IEEE Trans Neural Netw* 2007;**18**: 959–63.

107. Judson R, Elloumi F, Setzer RW, *et al*. A comparison of machine learning algorithms for chemical toxicity classification using a simulated multi-scale data model. *BMC Bioinformatics* 2008;**9**:241.

108. Chesnokov YV. Complexity and spectral analysis of the heart rate variability dynamics for distant prediction of paroxysmal atrial fibrillation with artificial intelligence methods. *Artif Intell Med* 2008;**43**:151–65.

109. Muselli M, Costacurta M, Ruffino F. Evaluating switching neural networks through artificial and real gene expression data. *Artif Intell Med* 2008; Sep 10. [Epub ahead of print]: doi: 10.1016/j.artmed.2008.08.002.