SYSTEMS BIOLOGY OF BREAST CANCER

GOPAL KRISHNA RAMADAS DHONDALAY

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

November 2013

Nottingham Trent University
School of Science and Technology
Clifton lane
Nottingham NG11 8NS.

Dedicated to my parents and family…

Copyright statement

Acknowledgements

Firstly, it is my privilege to thank ***Prof Graham R Ball*** for his very valuable guidance at each and every stage of this research. Without his extremely motivating support this research could not have happened. I also thank ***Dr Tony Allen*** for his expert advice and apt critics which have refined my work.

My thanks are due to ***Prof Robert Rees*** and ***Prof Graham Pockley*** for their encouragement and motivation throughout my entire research period.

I also thank my colleague ***Dr Dong L Tong*** for her friendly support and valuable discussions during our lunch together.

I would also like to thank ***Dr Desmond Powe*** for his support in clinical samples and biological validation.

I would like to thank entire staff and colleagues from J**ohn van Geest cancer Research Centre** for their help.

I would like to thank all my friends from the Nottingham Trent University and also from India who are directly or indirectly helped me throughout.

I also thank my better half ***Shwetha*** for her lovely support at every moment.

Finally, it's my pleasure to thank my parents and family for their ever encouraging blessings and unstinted support.

Contents

List of figures

List of tables

Abbreviations

ANN	Artificial neural network
AR	Androgen receptor
ATE	Average test error
AUC	Area under curve
BCSS	Breast cancer specific survival
BP	Back-propagation
BRCA1	Breast cancer 1
BRCA2	Breast cancer 2
CA12	Carbonic anhydrase 12
cDNA	complementary Deoxyribonucleic acid
ChIP	Chromatin Immuno Precipitation
CISH	Chromosomal In-Situ Hybridization
DMS	Distant metastasis-free survival
DNA	Deoxyribonucleic acid
ELISA	Enzyme Linked Immuno Sorbant assay
ER	Oestrogen receptor
ERE	Oestrogen Responsive Elements
FISH	Fluorescent In-Situ Hybridization
GEO	Gene Expression Omnibus
HER2/neu/ErbB2	Human epidermal growth factor 2
HIF1	Hypoxia inducible factor 1

IGFBP	Insulin like growth factor binding protein
IHC	Immunohistochemistry
K-NN	K-Nearest Neighbours
LDA	Linear discriminant analysis
MCCV	Monte Carlo Cross Validation
MLP	Multi-layer perceptron
mRNA	messenger Ribonucleic acid
MSE	Mean square error
NPI	Nottingham Prognostic Index
PCA	Principle Component Analysis
PCR	Polymerase Chain Reaction
PR	Progesterone receptor
RBF	Radial basis function
RFS	Relapse-free survival
RNA	Ribonucleic acid
RNN	Recurrent neural network
ROC	Receiver Operating Characteristic
RT	Room temperature
RT-PCR	Reverse Transcriptase-Polymerase Chain Reaction
SERMS	Selective estrogen receptor modulators
SOMs	Self-Organising Maps
SVM	Support vector machine
TMA	Tissue MicroArray

ABSTRACT

Breast cancer, with an alarming incidence rate throughout the globe, has attracted significant investigations to identify disease specific biomarkers. Among these, oestrogen receptor (ER) occupies a central role where overexpression is a prognostic indication for breast cancer. The cross-talk between the responsible contenders of ER-associated genes potentially play an important role in the disease aetiology. Investigation of such cross talk is the focus of this thesis.

The development of high throughput technologies such as expression microarrays has paved the way for investigating thousands of genes at a time. Microarrays with their high data volume, multivariate nature and non-linearity pose challenges for analysing using conventional statistical approaches. To combat these challenges, computational researchers have developed machine learning approaches such as Artificial Neural Networks (ANNs).

This thesis evaluates ANNs based methodologies and their application to the analysis of microarray data generated for breast cancer cases of differing oestrogen receptor status. Furthermore they are used for network inferencing to identify interactions between ER-associated markers and for the subsequent identification of putative pathway elements. The present thesis shows that it is possible to identify some ER-associated breast cancer relevant markers using ANNs. These have been subsequently validated on clinical breast tumour samples highlighting the promise of this approach.

This thesis will also demonstrate the novel application of ANNs in systems biology of ER, PR and Her2. Furthermore in this research, the integration of ER, PR and Her2 systems have been undertaken to represent a broader view of the breast cancer system.

Finally, this thesis will discuss the advantages, limitations, potential application and future potential applications of the methods evaluated.

Chapter 1

INTRODUCTION

Chapter abstract: Breast cancer is the most common cancer diagnosed in the west and the incidence rate is increasing throughout the globe. Oestrogens play a major role in the mammary gland tissue development and preparation of the uterus for fertilization. Oestrogens initiate signalling through oestrogen specific receptors, whose overexpression is a prognostic indication in breast cancer. To identify specific sub group of patients that would benefit from endocrine and hormonal therapy, researchers have attempted molecular sub-typing of breast cancer using microarrays. However, the use of microarrays, with the potential of investigating thousands of genes at the same time, pose challenges for conventional statistical approaches in terms of data volume, highly multivariate nature, complexity and non-linearity. To combat these challenges, computational researchers have developed machine learning approaches. These approaches have an ability to learn the data pattern and classify samples or predict the outcome. Such process have the potential to enormously benefit high throughput biological data analysis.

1.1. Breast cancer

Cancer is a disease characterised by an uncontrolled growth of cells, leading to benign (harmless) or metastatic (detrimental) tumours. Among the repertoire of cancers, breast cancer is the most commonly diagnosed. Due to the lack of effective therapeutic markers, the global burden of breast cancer is increasing with time, accounting for one in six of all cancer related deaths. Thus, makes it the second leading cause of death after lung cancer throughout the globe (Jemal et al., 2011). Although, breast cancer rates vary largely with race, ethnicity and geographic region, caucasian women have higher incidence compared to black African American women (Ma and Jemal, 2013). In the United Kingdom alone, one third of all newly diagnosed cases are of breast cancer, with the rate of diagnosis increasing from 1 to 400 per 100,000 cases each year from 1960 to 2010 (http://www.cancerresearchuk.org/).

1.2. Oestrogen Receptors

Oestrogens are steroidal hormones which function as signalling molecules. There are three basic oestrogen molecules: **Estrone** (E1; biologically inactive), **Estradiol** (E2 or 17β-estrone; most tumourigenic) *and* **Estriol** (E3; most beneficial). Both men and women have the same oestrogens which vary in concentrations. In women, they are produced and secreted by the ovaries, adrenal glands and breasts. Oestrogens mainly stimulate and maintain sexual

characteristics and reproduction. The breast and uterus, which play central roles in sexual reproduction, are two of the main targets of oestrogens. In addition, they also found to act on brain, bone, liver and heart tissue (Enmark and Gustafsson, 1999).

Oestrogens exert their function on various tissues by travelling through the blood stream. When an oestrogen molecule enters a cell and binds with their specific receptor, called an *Oestrogen Receptor* (ER) found in the nucleus, wherein the oestrogen-receptor complex is formed and a cascade of signalling events governing cellular proliferation, growth, and gene regulation ensues.

Oestrogens play a vital physiological role in mammary tissue development prior to the induction of lactation (Anderson, 2002), and uterus to facilitate adhesion of fertilized egg in adult females (Shao et al., 2012). Overexpression of ER leads to increased cellular stimulation resulting in a decreased repair of genetic material potentially leading to breast tumour development (Anderson, 2002). Contrastingly, females who have undergone oophorectomy before the age of 30, are known to be at lower risk of breast cancer development and circulating sex hormones in post-menopausal women are found to be predictive for hormone receptor positive breast cancers in 20 year follow-up study (Zhang et al., 2013).

Clinically, ER-positive (ER+) patients benefit from enhanced disease free survival rates with better response to hormonal therapy (tamoxifen), compared to women with aggressive ER-negative (ER-) breast tumours (Lemieux and Fuqua, 1996). About three-fourth of advanced ER+/Progesterone receptor (PR)+ breast

cancer patients responds to tamoxifen treatment, whereas only 10% of ER-/PR- patients will also respond similarly (Zwart et al., 2011).

Although ER expression is largely associated with a better outcome for the patients, this is somewhat paradoxical since low levels of ER in normal female breast depends on the menstrual cycle phase (Zhang et al., 2013). Even though ER expression is a good prognostic indication in breast cancer, some randomised control trials have confirmed that hormonal therapy is ineffective in invasive breast cancer that lacks ER expression (EBCTCG, 2005), hence ER status is also an excellent negative predictive factor.

Studies have also shown that ER expression alone is a relatively poor predictor for hormonal therapy response in tumours (EBCTCG, 2005). Incomplete knowledge of the ER signalling pathway suggests that an investigation of ER-associated genes may serve as better markers for disease progression and response to therapy. Thus, ER expression is now recognised as a prognostic marker in breast cancer, whether or not the usage of anti-oestrogen such as tamoxifen as adjuvant endocrine therapy for patients could be decided.

1.2.1. ERα and ERβ subunits:

The ER consists of two different forms, namely **alpha** (α) and ***beta*** (β) encoded by two different genes *ESR1* (NM_000125; Ch.6q24-q27) and *ESR2* (NM_00104275; Ch.14q21-q24) respectively. Both ERα and ERβ are also able to

form homodimer or heterodimers among themselves upon activation (Li et al., 2004). They have sequence and functional homology at their ligand binding and DNA binding domains, wherein ligand binding domain transactivates gene transcription in the absence of the ligand and the DNA binding domain binds with **Oestrogen Responsive Elements** (ERE) to initiate oestrogen-inducible gene transcription (Klinge, 2001). Both ERα and ERβ (ERs) are widely expressed in different tissue with notable differences in expression patterns (Couse et al., 1997). For example, ERβ is expressed much lower compared to ERα at mRNA level in breast tumours (Dotzlaw et al., 1997), and in healthy tissue (Shaw et al., 2002).

1.2.2. The ER pathway

In response to pituitary hormones, the oestrogens are synthesised in the ovary from cholesterol. Estradiol can be converted to 17β-estrone in an reversible reaction catalysed by the *Cytochrome P-450* enzyme, forming the by-products 4-hydroxy estrone and 16α-hydroxy estrone. Estradiol and estrone can both be converted to estriol, which is the active form of oestrogen. Being highly electronegative, 4-hydroxy estrone and 16α-hydroxy estrone (i) can actively bind with DNA bringing about conformational or transcriptional changes of specific genes or (ii) can bind with extra cellular oestrogen to facilitate aberrant expression of genes (Hayashi et al., 2003).

Oestrogens specifically binding to ERs present in the cell nucleus, produces conformational changes to form ER-complexes which later binds to promoter specific sites on the ERE DNA sequences and regulate the transcriptional activation of associated genes through binding with co-activators (Klinge, 2001). Oestrogen, by itself, serves as a transcriptional factor and upon binding with co-activators regulates the transcription of genes. Figure 1 is a schematic representation of oestrogen metabolism coupled with involvement of oestrogen in regulation of genes.

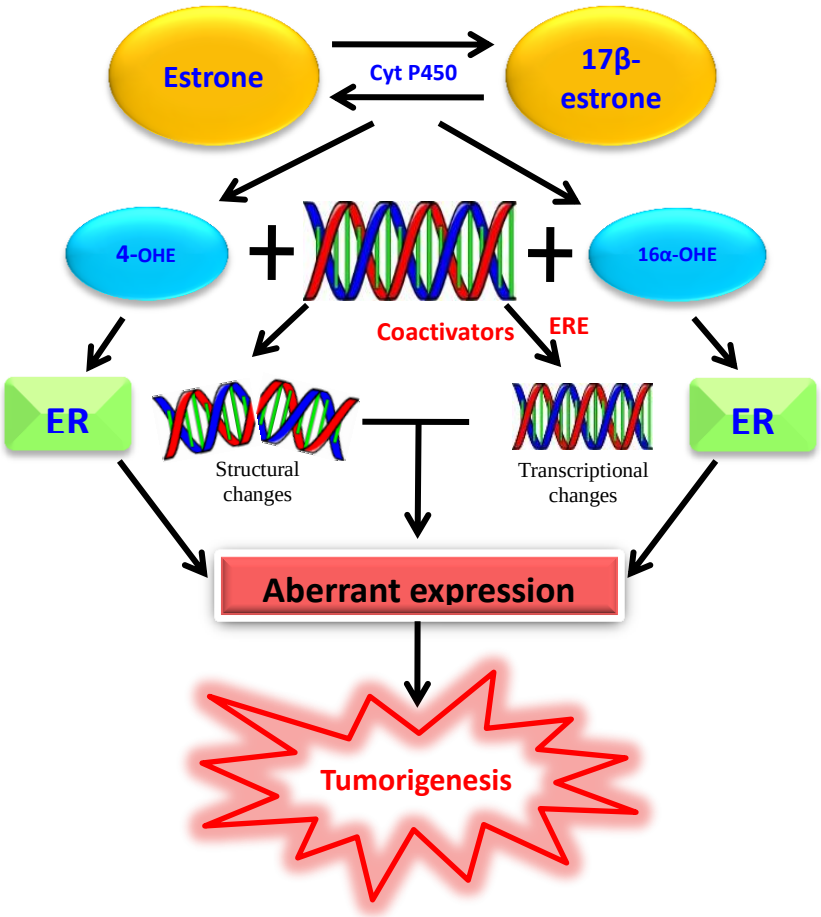


Figure 1: Schematic representation of ER mediated gene expression. Estrone metabolism and involment of extrinsic oestrogen in the aberrent gene expression. ER acting as trascriptional activator for oncogenes results in tumourigenisis.

ER interacts with growth factors to bring about up-regulation of their expression and/or others. ER up-regulates growth factor receptor tyrosine kinases such as human epidermal growth factor receptor 2 (*HER2/Neu/ErbB-2*) and insulin-like growth factor receptor-1 (*IGFR1*) (Schiff et al., 2004). Conversely, *IGFR1* can also modulate and enhance ER activity. This bi-directional cross-talk between ER and the growth factor receptor pathway plays a significant role in acquired and *de novo* resistance to endocrine therapy in breast cancers (Nicholson et al., 2005). ER also up-regulates signalling molecules such as membrane proteins, adaptor molecules and kinases such as proline, glutamic acid and leucine rich protein-1 (*PELP-1*) (Chakravarty et al., 2010).

Depending on the cell type, oestrogen exhibits different behaviours in tissues having beneficial effects. For example, in the liver oestrogens regulate the metabolism of high density lipoproteins and facilitates removal of fatty plaques from the inner walls of arteries and also regulates osteocyte matrix formation in bones (Lee et al., 2003). Overexpression of oestrogens is associated with cellular proliferation in glandular epithelium in the breast and endometrium in the uterus (Pozharisski, 2005). On the contrary, imbalanced expression of oestrogens in breast and uterus may lead to detrimental effects, causing cancer. Even though, over expression of ER is seen in the majority of breast tumours, it is somewhat of a paradox (Ricketts et al., 1991, Schiff et al., 2004).

1.2.3. ER in breast cancer

ER is overexpressed in almost 70% of breast cancers and plays an important role in the biology of breast cancer (Cordera and Jordan, 2006), ER directly stimulates mammary cell proliferation (Bocchinfuso and Korach, 1997) and division (Levin, 2009). This leads to increased genomic mutations through ER-induced reactive oxygen species production (Okoh et al., 2011). Furthermore, evidence of accumulated genotoxicity through ER metabolism has also been found (Dhillon and Dhillon, 1995). Both these processes are likely to disrupt the cell cycle, DNA repair mechanisms and alter programmed cell death – **Apoptosis** causing tumorigenesis (Bocchinfuso and Korach, 1997). Although ERα is highly associated with well differentiated tumours, ESR1 variants tend to have differences in risk of disease development (Deroo and Korach, 2006).

ER has long been considered as a "good" prognostic factor predicting favourable disease outcome and treatment in breast cancer (Samaan et al., 1981). Although the prognostic value of ER is still controversial, ER status is presently a valuable predictive factor for the success of endocrine therapy in breast cancer (Lopez-Tarruella and Schiff, 2007).

To combat the disease, **Selective Estrogen Receptor Modulators** (SERMS) like **tamoxifen** are used as ER antagonist in primary breast cancer patients (Fabian and Kimler, 2005). Furthermore, in post-menopausal women, **raloxifene** as SERMS (Cummings et al., 1999) and **anastrozole** as aromatase inhibitors are widely used (Mokbel, 2004). A number of clinical studies have

shown that patients with ER+ tumours enjoy improved disease-free survival rates and respond better to hormonal therapy compared to women with clinically aggressive ER- tumours (Rasmussen et al., 1989).

1.3. Classification of breast cancer

Breast cancer is classified into numerous subtypes according to various systems. The real-world drive for classification is to define each subtype of breast cancer in a manner which helps to select specific treatment approaches to achieve good outcome or better response to treatment. Classification of breast cancer is usually based on histological appearance of tissue in the tumour under the microscope, but recently emerging methods like nucleic acid based classification are also being developed and in trial.

On a histopathological basis, breast cancer consists of several types derived from the epithelial lining of mammary ducts or lobules. Among them, the invasive ductal carcinoma represents a dominant group, with invasive lobular carcinomas, invasive papillary and micropapillary carcinomas with a high frequency. Mixed type with lobular and ductal features and metaplastic carcinomas are lower in frequency (Webster et al., 2005).

Based on the differentiation of cells in the tissue, breast cancer can be divided into three grades as; low grade (well differentiated), intermediate grade

(moderately differentiated) and high grade (poorly differentiated) with a worsening of prognosis from low to high grade (Elston, 2005).

Tumour aggression, measured in terms of size of tumour (T), involvement of the lymph node (N) and metastatic activity (M), classifies breast cancer into five stages ranging from 0-4 representing good to worst prognosis respectively (Elston, 2005).

Hormone receptor expression classifies breast cancer mainly into ER and PR, along with *HER2/neu* amplification or over expressing types (Tanos et al., 2012).

Cancer poses heterogeneity in pathology, morphology, and also in response to treatment. Such clinical heterogeneity was simplified by dividing tumours into molecular classes by gene expression patterns of DNA microarrays for human acute leukaemia (Golub et al., 1999) later extended to breast cancer.

1.3.1. Molecular classification

The molecular taxonomy of breast cancer was initiated by Perou et. al. in the beginning of 21th century (Perou et al., 2000). Using cDNA microarrays of 65 breast carcinomas and an unsupervised hierarchical clustering method, they grouped genes on the basis of similarities and variations. They classified breast cancer cases into three main subtypes as **luminal**, **basal** and **HER2/neu** subtypes based on differential expression of genes. A balanced correlation between

oestrogen receptor α gene and protein expression in the tumours expressing transcription factors GATA3, Xbox binding protein 1 and hepatocyte nuclear factor 3α was found. Interestingly, the gene expression pattern revealed relationship between specific genes with specific tumour subtypes and also facilitated biological interpretation. Thus, they proved that the phenotypic diversity of breast cancer can be represented by gene expression variation.

Extending the Perou et. al. work, Sorlie et. al. classified 85 breast carcinomas into six subtypes as **luminal A, luminal B, luminal C** predominating ER positivity and **basal-like, ERBB2+** and **normal-like** group (Sorlie et al., 2001). Luminal A subtype was characterised by higher expression of ER alpha, ER-regulated *LIV-1*, *GATA3*, trefoil factor 3 (*TFF3*), X-box binding protein 1 and hepatocyte nuclear factor 3 alpha; luminal B and luminal C showed moderate expression of luminal specific genes; Basal-like subtype was characterised by higher expression of keratin 5/17, fatty acid binding protein 7 and lamin; whereas the *ERBB2+* subtype had overexpression of *ERBB2* and *GRB7*. This study was revised further by considering 115 breast carcinomas with 534 "intrinsic" gene signature along with two independent validation cohorts of samples from van't Veer and West datasets (van de Vijver et al., 2002, West et al., 2001). Cross comparison between the subtypes revealed that luminal B, basal-like and ERBB2+ was found to be negative for the luminal A genes previously known and had their own distinct set of genes. The luminal A/luminal B distinction was less clear compared to the *luminal/basal-like* distinction. This was also proved by *BRCA1* mutational patterns in the subtypes. It was observed that *BRCA1* mutation

was primarily associated with the basal-like subtype. When clinical outcome, such as time to develop distant metastasis was studied, it was observed that, luminal A subtype as best, basal-like and ERBB2+ as worst and luminal B has an intermediate response to therapy (Sorlie et al., 2003).

Simultaneously, Sotiriou et al made extensive studies to investigate the association of molecular subtypes of breast cancer with clinical outcomes. In one of their studies, clinical outcomes such as relapse-free survival (RFS) and breast cancer survival (BCS) were investigated with nodal status among 99 breast carcinomas (Sotiriou et al., 2003). It was observed that 16 genes were significantly associated with RFS. Based on the distinct differential gene expression profiles, basal-like showed two subgroups as Basal-like 1 and Basal-like 2, whereas the luminal subtypes showed three subgroups such as Luminal-like 1, Luminal-like 2 and Luminal-like 3. Luminal 1 had best outcome with 80% 10 years RFS and luminal 2 had worst outcome with 40% 10 years RFS. In another study, to investigate the association of tumour grade with the risk of recurrence of the disease and to compare the histological grade with gene expression grade, 64 ER+ carcinomas were considered using a 97 gene signature. Clinical outcomes such as RFS and distant metastasis-free survival (DMS) were investigated with a gene expression grade index. It was observed that RFS was more strongly associated with gene expression grade than the histological grade (Sotiriou and Desmedt, 2006, Sotiriou and Pusztai, 2009).

1.4. Breast cancer heterogeneity

Breast cancer exhibits diverse heterogeneity in pathological, molecular and clinical parameters. Calza et al. had classified tumour samples by considering clinical factors such as age at diagnosis, tumour size, tumour grade, lymph node status, receptor status, hormone replacement therapy and expression of p53 mutation (Calza et al., 2006). They observed breast cancer tumour classification to be closely related to ER status. ER positive protein expression was observed in luminal A, luminal B and normal-like tumours. Basal-like and ERBB2+ tumours expressed high ER negativity. Younger pre-menopausal women were found in the *basal* subgroup with tumours having a high grade (Elston grade) and being genomically unstable. Furthermore, the majority of basal tumours had a p53 mutation and most of the patients were present or former hormone therapy users. *ERBB2+* tumours were observed in elderly patients having a very high grade, a large tumour size, with *p53* mutation and almost 70% patients showing ER and PR positivity.

The luminal B group showed a similar complex pattern as the *ERBB2+* group, but with a less aggressive nature. The development of metastasis in the luminal B group with respect to tumour size and the receptor status was low, though a grade of III was observed. Luminal A tumours on the other hand were found in post-menopausal women with a smaller tumour size. These were genomically more stable, of lower grade and with wild type p53 expression. Very few patients in this group used hormone replacement therapy. Considering

prognosis it was observed that *ERBB2* expressing tumours had the same poor outcome as tumours expressing cytokeratin 5 and 17 of the basal type. Worst recurrence free survival was observed in *ERBB2+* tumours. Luminal B had a lower survival compared to the luminal A *and* normal-like group which had the best prognosis when treated with endocrine treatment. However, for the untreated patients the basal-like subgroup shared its survival pattern with luminal B and *ERBB2+* expressing tumours.

Another system was developed by Farmer et. al. (2005), classifying breast cancers by the presence of the androgen receptor (AR) and androgen signalling pathway in an negative ER cohort (Farmer et al., 2005). These were called as molecular apocrine containing tumours outside the *basal* group (Doane et al., 2006). This group is defined by *ERBB2* amplification along with genes like *AR, SPDEF, FOXA1, XBP1, CYB5, TFF3, NAT1, APOD, ALCAM* and *AR* presenting activation and response to the androgen signalling pathway.

Herschkowitz et. al. through their study in murine models identified a gene subset of claudin low expression as a biologically different entity with characteristic of stem cells (Herschkowitz, 2010). This was observed to belong to metaplastic tumours with altered PI3K/AKT pathway; hormone receptor negative and low levels of ER associated *GATA3* regulated genes and enriched with cell adhesion markers associated to stem cell function and epithelial mesenchymal transition (Hennessy et al., 2009).

Thus, from the above studies it was certain that transcriptional and genomic alterations differentiate tumour subtypes, suggesting that these aberrations could arise from transformed progenitor or stem cells possessing distinct biological properties (Sotiriou and Pusztai, 2009). It was observed that almost all cancers show *BRCA1* mutations, irrespective of them being sporadic or hereditary possessing a basal like triple negative (ER-, PR-, HER2-) phenotype associating with poor prognosis and response to treatment. Luminal A tumours are low grade, indolent, and respond to therapy, while luminal B, *ERBB2+* and ER-ve tumours show incomplete response to endocrine treatment. Aggressive *ERBB2+* tumours show sensitivity to an anti-HER2 monoclonal antibody, Trastuzumab, while similar basal-like tumours show sensitivity to chemotherapy.

The additional clinical value of the molecular classification and the tumour grade is limited by close association of the ER, PR and HER2 receptors. Since the molecular classification is able to provide molecular differences underlying the phenotypic expression of breast cancer which can enable revelation of new treatment targets, improvement in clinical trials is being witnessed. Detection of androgen receptor pathway in ER and PR negative tumours, DNA repair pathways in *BRCA1* and *BRCA2* mutation carriers and basal subtype by molecular studies also provide a new direction for future studies (Sotiriou and Pusztai, 2009).

1.5. Biomarker identification in breast cancer

Biomarkers can be the indicators of biological, pathological and pharmacological responses of a therapeutic intervention. They can provide prognostic or/and predictive information regarding the therapy, which can be characteristically measured and evaluated from body fluids like urine, plasma and blood or by invasive techniques with tumour tissue samples by immunohistochemistry (IHC) or custom gene expression arrays. The clinical usefulness of a biomarker depends on the relative and absolute difference between marker positive and negative states enabling to explore treatment options to treat a particular disease (Andre and Pusztai, 2006, Dowsett and Dunbier, 2008, Duffy, 2005, Oldenhuis et al., 2008).

1.5.1. Predictive and Prognostic biomarkers in breast cancer

Predictive markers are the set of markers which quantify the response of disease in terms of specific treatment's sensitivity or resistance (Duffy et al., 2011). Different breast cancer subtypes show varied responses towards a particular therapy, making it impossible to have a common treatment regimen for all cancer types. For any given cancer type, it is likely that only a small group of the patients will benefit from the systemic treatment and some are likely to suffer from the severe side effects of the adjuvant therapy (Shapiro and Recht, 2001).

Hence, predictive markers play a vital role in the determining the likelihood of patient response to a treatment in advance.

Candidates of predictive biomarkers in breast cancer are;

- Topoisomerase II alpha is considered as an important marker for response to anthracycline and other cytotoxic drug therapy (Dowsett and Dunbier, 2008).
- As marker for monitoring advance cancer therapy, Carcinoembryonic Antigen (CEA) and tissue peptide antigen (TPA) are helpful.

Prognostic biomarkers provide information regarding disease outcome irrespective of the therapy. Generally, ER and PR have both predictive and prognostic importance as markers of endocrine and chemotherapy treatment. Overexpression of *HER2* has established itself as a predictor of response to trastuzumab therapy or targeted antibody based therapy in response to Herceptin, whereas ER and HER2 as a pair are considered as positive and negative indicators of chemotherapy benefits respectively.

Candidates of prognostic predictive biomarkers in breast cancer are;

- As markers for prognosis assessment, traditional clinical factors like tumour size, tumour grade and lymph node status are important.
- Overexpression of non-histone protein Ki67 and proliferating cell nuclear antigen (PCNA), tyrosine kinase 1 (TK1) are important markers for proliferative activity with response to chemotherapy and have prognostic value in node negative tumours.

- Tumour growth factor (TGF) alpha and epidermal growth factor receptor (EGFR) as transcription factors serve as oncogenic markers.
- Overexpression of vascular epithelial growth factor (VEGF) as markers of angiogenesis.
- As markers for risk assessment in breast cancer, the germline mutations in *BRCA1* and *BRCA2* has been showed to have a strong association with tumorigenesis with hereditary and familial history.
- p53 as a cell cycle regulator whose accumulation represents cell signalling disorder.
- Overexpression of Bcl2 and imbalance of Bcl2/Bax ratio in cytoplasm as apoptotic activity initiators as apoptotic markers.
- Overexpression of oncogenic c-myc, loss or TGF-β II receptor and retinoic acid receptor alpha (*RARA*) as markers of cellular differentiation.
- Biological prognostic markers include uro-plasminogen (uPA) and its inhibitor PAI-1, E-cadherin as metastatic markers.
- A 62 gene signature developed by Yu et. al., called as Nottingham Prognostic Index (NPI) is a widely used clinicopathological staging system in breast cancer prognostication (Yu et al., 2004).

There is an ever increasing number of identified contender biomarkers involved in the overall molecular and cellular events that take place in the breast cancer genesis and prognosis. However, the known markers and their relationships in breast cancer development are still unclear. As a result, the use of traditional quantitative techniques of receptor binding assays such as Enzyme

Linked Immuno Sorbant Assay (ELISA) and IHC becomes non-economic and laborious if one need to investigate multiple markers in limited precious samples.

Recently, the usefulness of biomarkers accounted with breast cancer stem cells, addressed the need for the improvement in experimental designs and methods to detect biomarkers in clinical samples, work done by Golmen-Polar et. al. has validated their utility as predictors of the disease outcome, propensity for metastasis and response to treatment (Gokmen-Polar et al., 2011).

Developing genomic technologies have helped in the discovery of new potential markers and gene expression signatures which provide predictive and prognostic information of tumours (Dowsett and Dunbier, 2008). Simultaneous advances in the field of microarrays, micro fluidics and nanoparticles have opened a huge potential for the researchers to quantitatively estimate the expression of thousands of markers in minimal samples. One such high throughput advance is the field of gene expression estimation is based on transcriptomic microarrays.

1.6. Gene expression profiling

Gene expression profiling is a branch of molecular biology, wherein the quantitative expression of thousands of genes is carried out simultaneously, to create a comprehensive depiction of cellular process and function. The gene expression profiling, apart from its vast uses in different fields of biosciences, is extensively used in oncological studies.

1.6.1. Microarray

Microarrays comprise a grid of thousands of DNA spots (probes) attached to a solid substrate (usually a glass slide) in an orderly manner. Each spot contains millions of copies of identical DNA molecules (cDNA/genomic/oligonucleotides short sections) specific to given gene. The spots are printed on the slide by photolithographic synthesis (Lipshutz et al., 1999) or by piezoelectric (bubble jet) process (Okamoto et al., 2000).

Microarray technology, used in monitoring genome wide expression levels of genes in a given organism, has become an indispensable tool for biologists (Cooper, 2001).

1.6.2. Principles

Microarrays are commonly used for comparing expressions of a set of genes from cells maintained in a particular condition (tumour/diseased) to the same set of genes from a reference cell maintained in normal condition (normal). Figure 2 illustrates the general steps involved in a microarray experiment setup.

Firstly, the process of cDNA probe preparation involves total mRNA extraction from the samples (both tumour and normal) and Reverse Transcription-Polymerase Chain reaction (RT-PCR) for the production of cDNA clones. These cDNA clones are then labelled with fluorescent dyes, usually Cy3 (green) for normal sample and Cy5 (red) tumour sample, to distinguish sample conditions.

Later, these fluorescently tagged cDNA probes are hybridized on to microarrays by taking equal amounts of nucleic acid contents from both sample conditions. At this point, cDNA probes in the sample will hybridize to specific spots on microarray containing its complimentary sequence. Repeated washes are done to remove non-specific binding of probes. The amount of cDNA bound to a spot will be directly proportional to the initial number of RNA molecules present for that gene in both samples. Later, a suitable wavelength of laser is used to excite and detect the red and green fluorescent dyes. The amount of fluorescence emitted upon excitation corresponds to the amount of the bound nucleic acid (Murphy, 2002).

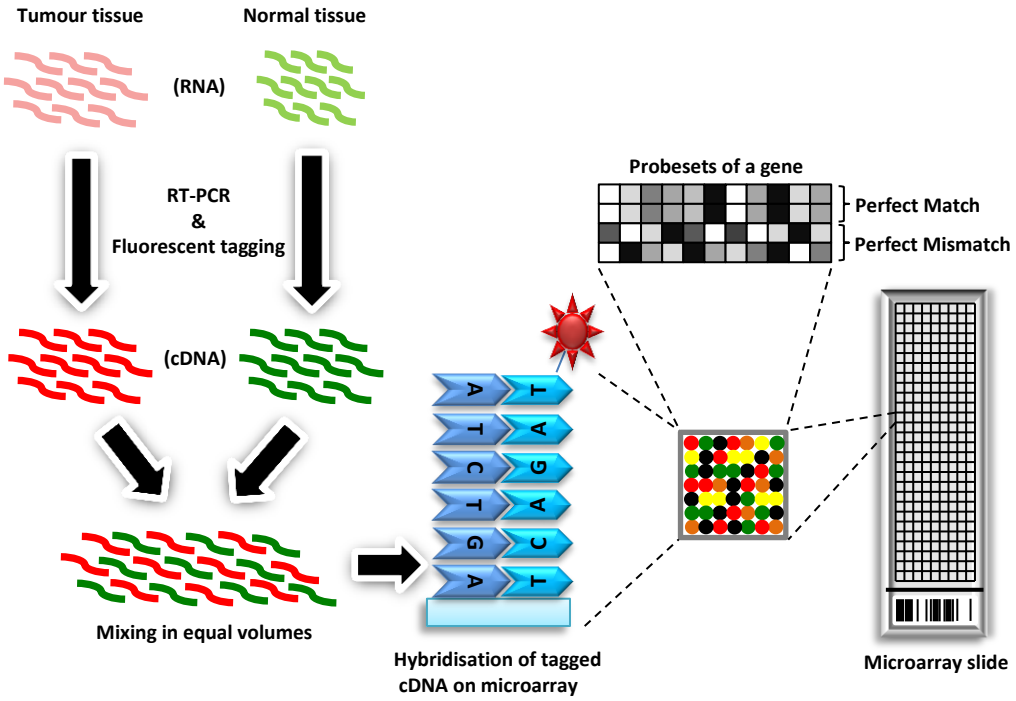


Figure 2: Schematic representation of Microarray technique. Representation of various steps involved in the microarray working and methodology. Microarray experimental design involves cDNA probe preparation and differential labelling with fluorescent dyes, simultaneous hybridization of both samples with microarray and scanning for the signals.

For instance, if cDNA of a particular gene from tumour was in greater abundance than that from normal sample, one would get the spot in red colour and conversely, the spot would be in green colour. If the gene was expressed in equal extent in both conditions of sample, the spot would be in yellow and if the gene was not expressed in both conditions, the spot would be black.

Lastly, the image analysis of microarray would incorporate high resolution image capture by using charge-coupled device camera of each spots which corresponds to a gene has an associated fluorescence value representing the relative expression level of that gene.

1.6.3. Applications of microarray in breast cancers

Use of cDNA expression microarrays in breast cancer is ever increasing. Significant usage started with the identification of distinct gene expression patterns defining breast cancer subtypes (Perou et al., 2000). Later they were successfully used to distinguish cancers associated with *BRCA1* and *BRCA2* mutations (Hedenfalk et al., 2002, Sorlie et al., 2001), to determine ER status (Gruvberger et al., 2001, Perou et al., 2000, van 't Veer et al., 2002), and to determine lymph-node status (Ahr et al., 2002, West et al., 2001). Numerous further milestones have been achieved.

1.6.4. Limitations

High throughput data structures from the biological experimental setups poses three limitations; Firstly, the data being highly non-linear. Simple linear methods have proveed to be inefficient in classifying samples compared to non-leniear methods (Lancashire et al., 2010). Secondly, the data points are highly multivariate comprising of multiple varaibles (genes) per sample. Thridly, the data possesses high dimensionality.

Although high-throughput technologies carry high expectations, they have a few disadvantages too. Their inherent problem is their massive data structure making simple correlation and linear regression statistical methods impossible to deal with the aspect of dimensionality of the data, hence, the "*curse of dimensionality*" as coined by Bellman (Widrow and Hoff, 1960).

Simultaneous investigation of tens of thousands of genes with a lack of standardised analytical steps for conversion of enormous amount of noisy data points into reliable and interpretable biological information can be troublesome (Simon, 2003).

Conventional statistical approaches, with simple linear functions and correlational methods, are unable to address the non-linearity and high dimension in data points. The aggregation of non-linearity, reproducibility and dimensionality of high-throughput biological data, therefore poses challenges for the bioinformaticians to develop algorithms to tackle the inherent problems (Lancashire et al., 2009). In addition, the combined biological variability

commonly found in biological samples, makes it difficult to identify key features within biological datasets. Hence, machine learning approaches have been exploited and extensively used to address the problems.

1.7. Machine learning for biomarker identification

Machine learning is a branch of artificial intelligence, a scientific discipline concerned with the design and development of algorithms that allow computers to learn characteristic patterns based on the empirical data. As defined by Tom Mitchell, "*a computer is said to learn from experience with respect to some class of tasks and performance measure, if its performance at tasks as measured by performance improved with experience*" (Mitchell, 1997).

Machine learning algorithms are classified as unsupervised learning, supervised learning, or reinforcement learning methods based on the desired output.

1.7.1. Unsupervised machine learning approaches

Unsupervised learning is a type of machine learning wherein the desirable output are not labelled. The machine is made to learn from the intrinsic characteristic patterns of input variables only. In other words, in the biological data instance, by learning the patterns in the variables, the algorithm classifies samples in to desirable classes.

1.7.1.1. Clustering methods

Clustering is an unsupervised machine learning approach used to classify data into groups with similar patterns that are characteristic to the features or variables. For instance in this clustering machine learning nomenclature we refer a gene or a sample as an **Object** and a set of objects having similar pattern as a **Cluster**. Clustering methods can be Hierarchical or Non-Hierarchical (Tefferi et al., 2002).

HIERARCHICAL CLUSTERING

Hierarchical clustering is a method of cluster analysis wherein the data patterns are clustered based on the specific relationships of the features of clusters resembling a phylogenetic tree. Hierarchical clustering can be of **Aggregative/Agglomerative** or **Divisive** method. Figure 3 illustrates the general flow of analysis.

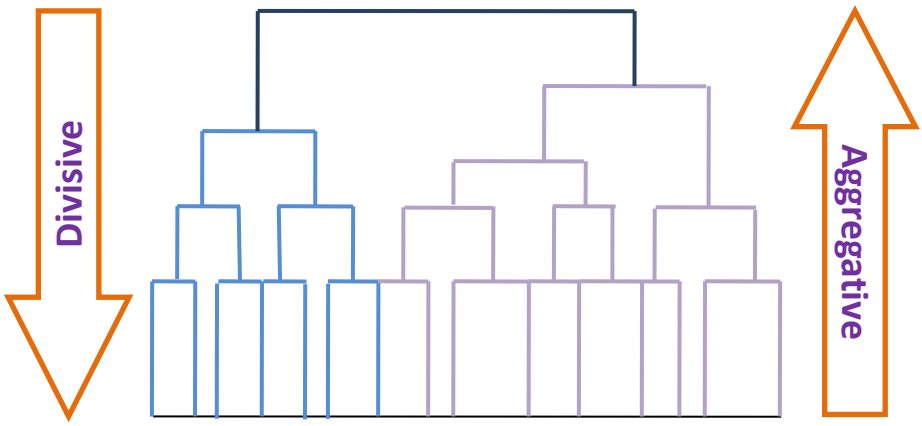


Figure 3: Schematic representation of Hierarchical Clustering method. Aggregative and Divisive Hierarchical clustering methods. Arrows indicates the direction of analysis process.

Aggregative/Agglomerative Hierarchical Clustering:

In Aggregative/Agglomerative hierarchical clustering method, clustering starts with an assumption that each object is a cluster. All the objects are subsequently merged into a single cluster based on the similarities of their pairwise distant measures, calculated in an iterative manner. This can then be represented in the form of a dendogram, where the distance from the branch point corresponds to the distance between the clusters (Tefferi et al., 2002).

Divisive Hierarchical Clustering:

Divisive hierarchical clustering, is contrary to aggregative/agglomerative clustering, starts with an assumption that all objects form a single cluster. Later, this is broken down into two or more clusters that have similar patterns. The divisive process is then repeated iteratively until all objects have been successfully separated (Chidananda and Ravi, 1995).

Hierarchical clustering methods in biological transcriptomic data mining context are extensively used to build genes of coexpression i.e., genes with related expression patterns, assuming that such groups contains functionally related proteins such as enzymes for a specific pathway (Stryhn and Christensen, 2013).

NON-HIERARCHICAL CLUSTERING

In non-hierarchical clustering methods, the data patterns are clustered without specifying the relationships of the features of the clusters. This method still requires a predetermination of the number of input clusters.

1.7.1.2. K-means clustering

The principle of K-mean clustering method arbitrarily groups objects into predetermined number of clusters (Initialization). The centroid-average expression profile of each cluster (Iteration) with respect to other clusters is then taken in an iterative manner for a fixed number of times (Convergence). The result is the convergence state where the composition of clusters remains unaltered by further iterations, usually 20,000-100,000. This form of non-hierarchical clustering method has become popular with its advantage of being scalable for large datasets (Khan and Ahmad, 2004). K-means clustering methods are used in various fields ranging from computer vision (Mignotte, 2008) to agriculture (Baskar et al., 2010).

1.7.1.3. Self-Organising Maps

Self-Organizing Maps (SOMs) are also called as Kohonen maps (Kohonen, 1990). The working of SOM starts from choosing the number and orientation of the clusters with respect to each data points i.e., during the initialization step, a grid of nodes is projected onto the expression space and each data point is assigned its closest node. Then, one data point is chosen at random and the assigned node is moved towards it, simultaneously moving all other data points depending on how much closer they are to the selected data point. The iteration

step is continued for fixed number of times or until the convergence is attained. SOMs are extensively used in data exploration (Stryhn and Christensen, 2013).

1.7.1.4. Principle component analysis

The principle of Principle Component Analysis (PCA) is to redistribute the data space of samples into a new data space (defined as Principle Component) depending of the variance among samples. These principal components are orthogonally arranged linear combinations of the original variables identified by the method and that can explain most of the variance, due to redundancy of information in the original data space (Raychaudhuri et al., 2000). PCA is usually considered as a tool for data reduction to tackle the issue of high dimensionality among biological data sets (Raychaudhuri et al., 2000).

1.7.2. Supervised machine learning approaches

Supervised learning is a type of machine learning wherein the desirable output(s) are labelled. The machine is made to learn from the intrinsic characteristic patters of the input variables together with a comparison of the actual current output and the desirable output. Here, some of the common supervised learning methods are mentioned.

1.7.2.1. K-Nearest Neighbours

K-Nearest Neighbours (K-NN) are popular for their non-parametric simple classification characteristic wherein an expression profile of known class sample is compared with an unknown. The processes involves the assessment of gene expression profiles of samples, identification of nearest neighbours depending on the applied distance measurement (usually Euclidian) and finally determining the class of the unknown sample by its nearest neighbour's class. (Vadrevu and Murty, 2010, Xiong and Chen, 2006). The limitations to this approach is the assumption that similar variables, being equally relevant, will have similar classifications becomes computationally complex as the number of variables increases (Cruz and Wishart, 2007). K-NN method with their varied applications, have been recently used in classification of microarray data (Parry et al., 2010).

1.7.2.2. Linear-discriminant analysis

Linear Discriminant Analysis (LDA) uses an approach of subgrouping the data by calculating the best splitting optimal linear line between populations. The whole process depends on the inter-sample and inter-group variability, giving it an added advantage to outperform when compared with other linear classification methods. However, LDA performs poorly on non-linear real-world biological data sets, with the major application being in the field of language analysis, image analysis and face recognition (Haykin and Gwynn, 2009).

1.7.2.3. Support Vector Machines

Support Vector Machines (SVMs) are a learning paradigm based on finding the equation for an optimal line which can maximally separate the classes, thus creating a hyper-plane from the data points which maximises the difference between them. In this high-dimensional space, an appropriate (linear/non-linear) separator, such as a kernel function can be applied to calculate the scalar product of the data points (Dreiseitl et al., 2001).

Among the supervised learning methods, K-NN has been shown to have good classification performances for a wide range of real-world data (Xiong and Chen, 2006). Miller and co-workers have compared the performances of K-NN, SVM and LDA supervised algorithms for the prediction of p53 status in breast cancer datasets (Miller et al., 2005). According to them, in spite of higher prediction accuracies (89.4% and 85.7%), K-NN performed with slightly higher specificity (95.3%) than that of SVM (94.3%) and LDA (84.0036%).

Apart from above mentioned basic supervised and unsupervised machine learning methods, the following methods are also used for specific applications in pattern recognition:-

Decision trees are simple predictive models which maps the conjunction of features (branches) to classification (leaves) and explicitly represents structured graph or flow-charts for decision and decision making (Mitchell, 1997). These are used in machine learning and data mining to predict the value of a target variable

based on several input variables. Decision trees are also called **Classification Trees** or **Regression Trees**.

Bayesian Networks are a probabilistic graphical model representation wherein the association between a set of variables (symptoms) and their conditional independencies (disease) is investigated. The classification is based on the posterior probability of an event (Jiang et al., 2011). Due to its probabilistic nature, Bayesian networks are also called **Belief Networks** or **Acyclic Networks**.

Genetic algorithms are a branch of artificial intelligence wherein a heuristic search is applied to mimic the process of natural evolution. A genetic algorithm is applied to optimise a population of solutions depending on the relation between genotype and reproductive success of an organism, inspired by genetic inheritance, crossover, mutation and natural selection (Muni et al., 2006). Collectively, genetic algorithms represent an evolutionary programming field called **Genetic Programming**.

1.7.3. Reinforcement machine learning approach

Reinforcement learning, unlike supervised learning, is a strategy of exploration by exploitation of current knowledge is applied and input/output relationship is unknown. The characteristic feature of reinforcement learning is the ability of online-learning i.e., the programme learns from the environment and

optimises itself to the inputs provided to enhance the performance of output (Nikolic and Fu, 1986).

Microarray data being highly complex with non-linear nature (reference) and high dimension

Furthermore, **artificial neural networks** are a method that has been exploited in the field of pattern recognition and biological data analysis, as detailed further in Chapter 2.

1.8. Databases and gene signatures using machine learning

The implication of microarrays in breast cancer diagnosis and prognosis has also facilitated the use of gene signatures corresponding to particular disease state. In a classical gene signature study in breast cancer, a 70 gene prognostic signature has been developed by using 78 lymph node negative breast cancer patients by van't Veer et al. (van 't Veer et al., 2002). The work has been further validated by van de Vijver et al. on a bigger cohort of 295 lymph node positive and negative samples treated with adjuvant therapy along with control non-treated breast cancer samples (van de Vijver et al., 2002). Later in 2005, another 76 gene signature for distant metastasis was applied using 115 samples for training and 171 samples for validating on lymph node negative breast cancer patients (Wang et al., 2005). The 76 gene signature was further validated by Foekens et al. on a multi centric cohort

of 180 lymph node negative patients who were untreated for adjuvant therapy (Foekens et al., 2006). Furthermore, the van't veer and Wang studies have been independently validated on 302 breast cancer patients from 5 different centres within the TRANSBIG project (Buyse et al., 2006, Desmedt et al., 2007). Table 1 summarizes some of the popular studies done by various groups to study different aspect of breast cancer using microarrays.

Simultaneously, the use of microarrays and high throughput technology like mass spectrometry has equally contributed for the advent and experimentations in their high dimensional data. Computational researchers and statisticians have used various methods and models for analysis of high-throughput data.

Table 1: Summary of the breast cancer microarray studies.

Study	Year	Class	Method	No. Of samples	Gene panel
West et al.	2001	Oestrogen receptor	Bayesian regression	49	40
Vijver et al.	2002	Distant metastasis	SC with clustering	248	70
vant Veer et al.	2002	Survival	SC with clustering	117	70
Minn et al.	2005	Distant metastasis	HC	62	48
WangY et al.	2005	Distant metastasis	MCR	286	76
Buyse et al.	2006	Survival	MCR	198	76
Chin et al.	2006	Transcriptional Aberrations	MCR	145	9
LiuQ et al.	2009	Pathological complete response	NBC, NMSC	278	34

Popular breast cancer microarray datasets from 2001 to 2009.
(Abbreviation: SC=Supervised clustering; HC=Hierarchical clustering; MCR=Multivariate Cox Regression; NBC=Naïve Bayes Classier; NMSC=Nearest mean scale classier).

1.9. Challenges for biomarker identification

In medicine, biomarkers are the measurable indicators of a particular disease state, progress of disease or the effect of treatment. Biomarkers could be a

biomolecules such as proteins or enzymes or even genes. The characteristic differential expression of a gene between the diseased state and normal state of samples could serve as a marker. Enormous efforts have been made from high-throughput transcriptomic data to identify biomarkers characteristic for specific diseases. As a result, it is well known that trying to find a single biomarker or a gene responsible for the complex disease such as breast cancer is inappropriate. A combination of genes or gene signatures in association may give more information about the behaviour of the disease. As the number of investigative genes increases, the needs for faster computational and statistical approaches arise. Hence, machine learning approach such as **Artificial Neural Networks** has been exploited and will be the focus of following chapter.

Chapter 2

ARTIFICIAL NEURAL NETWORKS

Chapter abstract: An Artificial Neural Network (ANN) is a computational machine learning approach inspired by physiological working of biological neuron. The basic perceptron model incorporates products of inputs and weights, which are then processed with a threshold function to mimic the output of samples in training. The perceptron model is advanced with the use of multi-layer perceptron (MLP) network. The training of the MLP model incorporates back propagation of error algorithm for the update of weights. Computational researchers have developed various variants of ANN. In this study, we have used an in-house developed, Stepwise-ANN model for classification of biological samples.

2.1 Introduction

Artificial Neural Networks (ANNs) are a branch of machine learning inspired by the working of a human brain. They are characterised by high complexity, non-linearity and with parallel information processing abilities. ANN

is also a computational machine designed to model or classify in a way in which human brain performs a particular task or function.

"A neural network is a massive parallel distributed processor made up of simple processing units that has a natural propensity for storing experimental knowledge and making it available for use"- (Haykin and Gwynn, 2009).

The basic structural units of human brain are the neurons which are inter-connected in a systematic manner for efficient transfer and processing of information. Neurons have the feature of **plasticity** - adaptation to surrounding environment (Churchland and Sejnowski, 1994). Likewise, artificial neurons form the basic units of computational neural networks and are able to adjust their internal weights in order to optimise the network outputs for a given training set.

2.2 Biological neural networks

The brain is the central part of nervous system comprising a massive neural network of approximately 10 million neurons in the cerebral cortex region in humans (Pascual-Leone et al., 2005). One end of the network consists of the **receptors** which receive external stimuli or information and pass it on to the neuronal network which continually receives and processes the information. Processed information then travels to the other end of the network, the **effectors** which exhibit appropriate response to the processed information. The whole

process of information transmission is a forward process with the scope for feedback between the contenders.

The interactions between the neurons are controlled by **Synapses** through chemical reactions which can induce an excitation or inhibition, but not both, of a subsequent neuron. At a synapse, a presynaptic input (electrical signal) releases a transmitter substance (chemical signal) which diffuses across the synaptic junction between neurons to produce a postsynaptic output (electrical signal). In a biological neuron, the *Axons* acts as transmission lines and **Dendrites** as receptors for the impulses. Together they constitute a neural microcircuit. With this neurobiological analogy as the source of inspiration, ANNs have emerged as an efficient machine learning tool.

2.3 Historical background

The history of ANNs starts from the preliminary efforts of neurophysiologists and neuro mathematicians to model the working of a biological neuron during the late 1900s. Neuroscientists Warren S. McCulloch and Walter Pitts first described a mathematical function with simulated bias function as an artificial neuron, called the **Threshold Logic Unit** in 1943 (McCulloch and Pitts, 1943). Later in 1958, the psychologist Frank Rosenblatt formulated the concept of the "Perceptron" in an attempt to describe perception (Rosenblatt, 1962). The perceptron (along with the artificial neuron) with its

ability of learning based on weighted connections between neurons, forms the basis for today's advanced neural network architectures (Rojas, 1996).

This theory of perceptron led to a formation of whole new field of learning methods and concepts in perception and classification. Simultaneous advances in Artificial Intelligence, and recognition of the limitations of perceptron methods, led to the diversification of interests of researchers from neural networks to artificial intelligence (Basheer and Hajmeer, 2000).

The interest in ANN among the research community reappeared when the Back-Propagation (BP) algorithm adopted from Delta rule of Widrow and Hoff, was published by Rumelhart, Hinton and Williams (Rumelhart et al., 1986, Widrow and Hoff, 1960). The BP algorithm offered an output error correction strategy wherein the successive correction of weights was based on the errors generated in classification during training. The learning algorithm produces accurate classification prediction. ANNs, inspired by mechanisms of information processing thorough a neuron, have been applied with success in various fields of sciences such as physical, chemical, astronomical, agricultural, biological etc.

2.4 Artificial neurons: structure and architecture

2.4.1 The perceptron

Perceptrons, coined by Rosenblatt, are the simplest form of the neural networks consisting of a single neuron with threshold activation function along

with adjustable synaptic weights and bias (Rosenblatt, 1962). These processing elements can be regarded as the basis functions of a generalised linear discriminant used for the classification of patterns said to be linearly separable. An algorithm is applied in a learning procedure to adjust the free parameters of the neural network. Rosenblatt, in **Perceptron convergence theory**, proved that if the patterns used to train the perceptron are drawn from two linearly separable classes, then the perceptron algorithm converges and positions the decision hyper plane between the two classes. A schematic representation of a single layer perceptron is represented in Figure 4.

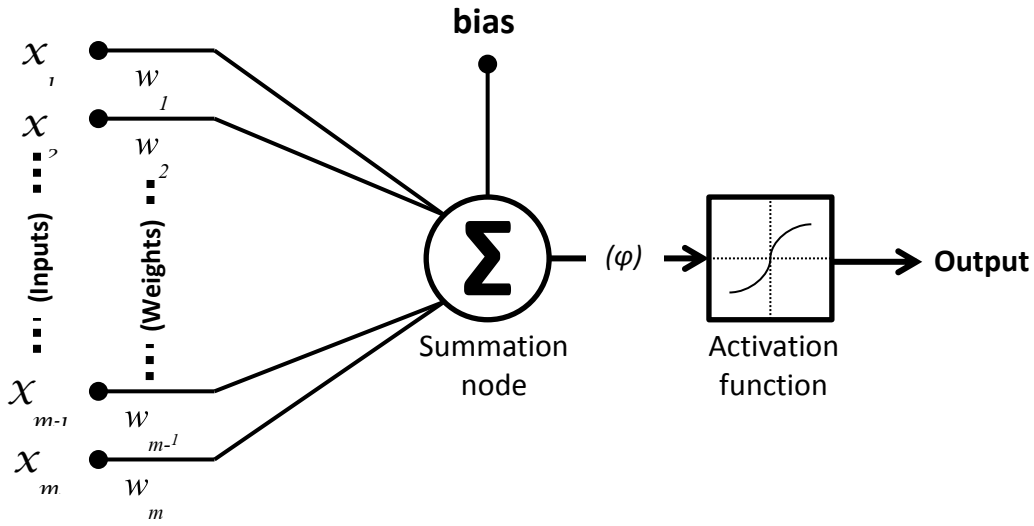


Figure 4: Schematic representation of single layer perceptron model. The number of input neurons is equal to the number of input variables, correspondingly equal number of weights. All the weights are passed on to the hidden neuron along with a bias. Activation function acts as a squashing function to scale-down the prediction.

If the inputs of the perceptron are x_1, x_2,…., x_m, external bias applied as b and the corresponding synaptic weights are w_1, w_2,…, w_m, then the applied hard limiter (φ) is given by;

$$\varphi = \sum_{i=1}^{m} w_i x_i + b \qquad \textbf{(Equation 2.41)}$$

Activation function: Activation function introduces non-linearity into the neural processing. Some commonly used activation functions are: Logistic sigmoid function and Hyperbolic tangent function.

A **Logistic sigmoidal function** is the most common activation function used to model the 'S-shape' behaviour of the data. The **Hyperbolic tangent function** is defined as the ratio between the hyperbolic sine and the cosine functions.

Sigmoid function: $$f(x) = \frac{1}{1 + e^{-x}} \qquad \textbf{(Equation 2.42)}$$

Tangent function: $$\tanh(x) = \frac{\sin h(x)}{\cos h(x)} = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \qquad \textbf{(Equation 2.43)}$$

where x is the data point and e is the *Euler's number* which is the base of the natural logarithm. These functions are widely used because their derivatives are easy to compute and can be expressed directly as a function of the net input (Shenouda, 2006).

Feedforward mechanism: The feedforward neural network algorithm was the first and perhaps the simplest type of artificial neural network algorithm developed (Widrow and Hoff, 1960). In this network, the information moves in only one direction, i.e., forward, from the input nodes to the output nodes without any cycles or loops in the network.

A major limitation of single layer perceptrons is that they are limited to the classification of linearly separable patterns. To overcome this practical limitation, multilayer perceptron models were developed.

2.4.2 Multilayer perceptron

Multilayer perceptron (MLP) neural networks comprise multiple perceptrons existing in multiple layers. The network exhibits a high degree of connectivity which is determined by the extent of the synaptic weights of the network (Haykin and Gwynn, 2009).

The basic multilayer perceptron representation can comprise of three or more layers namely; a layer consisting of input nodes as **Input layer**, output nodes as **Output layer** and a sandwiched **Hidden layer**. Each node of any layer is connected with the every nodes of the previous layer representing a fully connected architecture. In a multilayer perceptron, the hidden neurons in the hidden layer play a crucial role, acting as feature detectors. As the learning process continues, hidden neurons gradually map salient features that characterise

the training data by a process of nonlinear transformation of input data into feature space where pattern-classification may be more easily carried out. Indeed, it is this feature space which distinguishes multilayer perceptrons from Rosenblatt's single layer perceptrons. Figure 5 gives a schematic representation of multilayer perceptron model.

2.4.3 Back-propagation algorithm

In a MLP network, if an output neuron produces an incorrect classification, the derivatives of the error with respect to the weights can be evaluated and can be used to minimise the error function.

In order to minimise the error function of weight values in the back-propagation learning method, update of the weights at each and every weight vectors of the previous layer for every node has to be accomplished. This rule is famously known as the **Delta rule** or **Least Mean Square** method (Widrow and Hoff, 1960). The delta rule for a perceptron can be stated as

$$\Delta w_{ji} = \eta\left(t_j - y_j\right)x_i \qquad \textbf{(Equation 4)}$$

where Δw is the weight difference, η is the learning rate (described in detail in Section 2.5.2), t_j as the target vector and y_i as the input vector.

To evaluate the derivatives of the error function, Rumelhart et al developed an algorithm called the **Back-Propagation algorithm** (Rumelhart et al., 1986), illustrated in Figure 6.

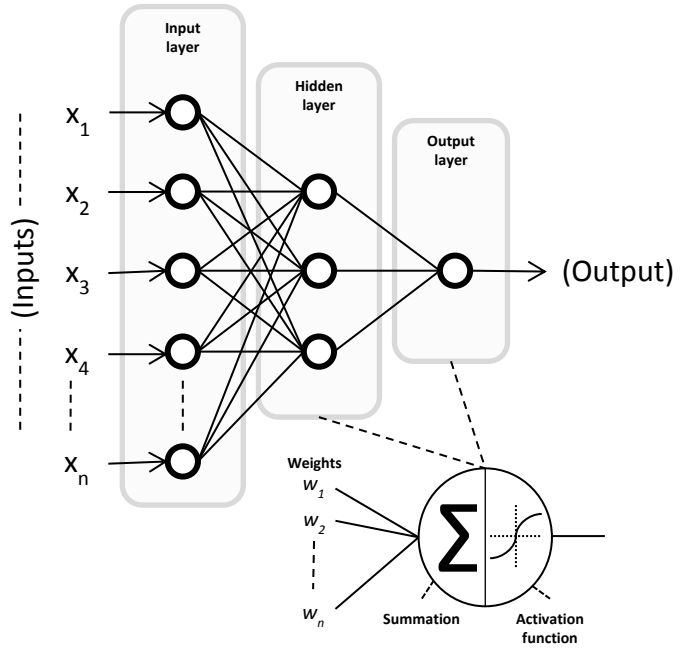


Figure 5: Multilayer Perceptron (MLP). Schematic representation of MLP with single layer of hidden layer. Each node at input layer corresponding to the total number of variables connected with weights. Computational mechanism at hidden and output layer is also depicted in the figure.

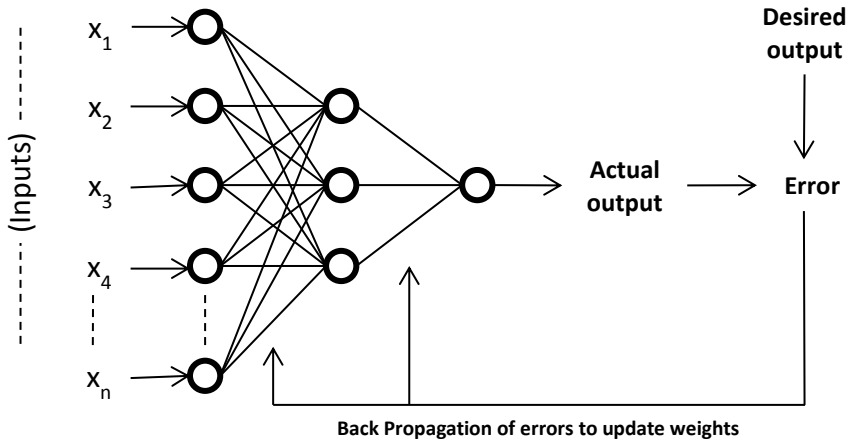


Figure 6: Illustration of Error back-propagation. A schematic representation of Error back-propagation algorithm. The actual output of the model is compared with the desired output and the error generated is back-propagated to alter the weights corresponding to the output.

2.4.4 MLP training

The training process is the most crucial process in a learning algorithm where the machine learns the pattern of the data through the inputs and generalises the parameters for a global predictive performance.

Consider a multilayer perceptron consisting of source nodes at the input layer, one or more nodes at the output layer and one or more hidden layers. Let $y_j(n)$ denotes the function signal produced at the output of the neuron j in the output layer by the stimulus $x(n)$ applied to the input layer. Correspondingly, the error produced at the output of neuron j is define by:-

$$e_j(n) = d_j(n) - y_j(n) \qquad \textbf{(Equation 5)}$$

where $d_j(n)$ is the j^{th} element of desired response vector and $y_j(n)$ is the j^{th} element of output vector.

The *instantaneous error* (E) of neuron j is defined as

$$E_j(n) = \frac{1}{2} e_j^2(n) \qquad \textbf{(Equation 6)}$$

The *Total instantaneous error* at all the neurons in the output layer (C) will then be

$$E(n) = \sum_{j \in c} E_j(n) = \frac{1}{2} \sum_{j \in c} e_j^2(n) \qquad \textbf{(Equation 2.47)}$$

Over a training set of N samples. The *averaged error* or the *empirical risk* can be defined as

$$E_{av}(N) = \frac{1}{N}\sum_{n=1}^{N} E(n) = \frac{1}{2N}\sum_{n=1}^{N}\sum_{j\in c} e_j^2(n) \qquad \textbf{(Equation 8)}$$

The goal of the learning process is to minimise the error (the difference between the desired output and supervised output). The same phenomena is exploited as the Back-propagation algorithm.

2.4.5 Gradient descent algorithm

Gradient descent, also called as **Steepest descent**, is one of the simplest network training algorithms. It is a first order optimisation of parameter algorithm for learning machines (Anderson, 1995). There are two versions of gradient descent; Batch and Sequential gradient descent.

In the batch version, the initial learning is carried out following randomisation of the weight vectors. Later on, the weight vectors are iteratively re-evaluated and updated so that the distance from the gradient is reduced at each step i.e., moving in the direction of negative gradient on an epoch-by-epoch basis. The advantages of this version include an accurate estimation of gradient vector, convergence of the gradient descent to the local minimum and parallelisation of the learning process.

In the sequential or on-line learning version, the error function gradient is evaluated and updated for just one pattern at a time i.e., the adjustment of the synaptic weights are done on an example-by-example basis.

2.4.5.1 Escaping local minima with momentum

In a MLP network, with more than one layer of adaptive weights, the non-linear functionality of the weights will correspond to the error function. Thus, there may exist more than one minima in the weight space. The one with the smallest value of the error function, defined as ***'Global minima'***, compared to the others, which are **'Local minima'**. The goal of modelling is to reach the global minima. Since negative gradient learning works according to the error surface topologies, there may be instances where learning is trapped in a local minima. In such situations, very small change in the synaptic weights increases the cost of the error function. So trapping the network to address this problem adding an appropriate momentum term to the weight vector adjustment can help the network escape local minima (Morris, 1993).

2.4.5.2 Generalisation and avoiding over-fitting

Generalisation is the ability of an algorithm to learn the patterns in the training data in order to successfully classify unseen data or the neural network's ability to produce a correct input-output mapping. **Overfitting** is the incorporation

of features that are unnecessary for learning patterns of data (Hawkins, 2004). To determine if a model has achieved generalisation, the dataset can be split into three subsets - a **Training set** for training the neural net in order to minimise the error, a **Test set** to optimise the stopping criteria of the training in order to prevent overfitting of the neural net, and finally a **Validation set** to determine the performance of trained patterns for completely unseen data. Model performance on this validation set determines the generalisation ability.

There are several ways of splitting the dataset. A well-known method is the Monte Carlo Cross Validation (MCCV) method wherein the dataset is randomly divided into three subsets as training, test and validation subsets respectively (Xu and Liang, 2001).

2.5 Optimisation of ANN parameters in predictive modelling

2.5.1 Randomisation of initial network weights

The speed of network convergence depends on the choice of initial weights. The weight update between the neurons depends on the weight specific derivative of the error function and also on the activation values of the neuron. Large weight can cause the network too quickly converge to a local minima, whereas small weights cause slow network learning (Fausett, 1994). To ensure that the initial signal of the hidden or output unit does not fall into the region where the

derivative of the sigmoid function is too small, the initial randomised weights of the network must not be of too large variance. If the derivative of the sigmoid function is small, the weight updates of the hidden or output neurons will be close to zero, making learning extremely slow. Thus, random initialisation of network weights results in appropriate convergence of the network (Rumelhart et al., 1986).

2.5.2 Learning rate and Momentum

The back-propagation algorithm facilitates an 'approximation' to the trajectory in weight space by the gradient descent method. If the learning rate (η) is too small, the change in the weights from one iteration to another becomes small, resulting in smoother trajectory of weights and improved learning, but at the cost of time consumption. On the contrary, if the learning rate is too large, large changes in the weight results in unstable or oscillatory network. Addition of a **momentum** (α) term to the delta rule of learning avoids instability of the network as shown below

$$\Delta w_{ji}(n) = \alpha \Delta w_{ji}(n-1) + \eta \delta_j(n) y_i(n) \qquad \textbf{(Equation 2.59)}$$

Where Δw_{ji} is weight difference, α is momentum constant, η is the learning rate, δ_j is the local gradient and y_i is the output signal.

The Momentum (α) parameter adds inertia to the movement through weight space, smoothing the oscillations of the network and leading to faster

convergence towards the global minimum. The optimum momentum range must lie between $0 \leq \alpha \leq 1$.

2.5.3 Hidden layer size

A hidden neuron will not have any specified desired response, but the back propagation of error signal needs to be determined in order to update the input layer weights. The optimal size of the hidden layer is very crucial. If too few hidden units are used, then the network will fail to achieve satisfactory performance, whereas too many hidden units will facilitate overfitting giving a poor performance on validation dataset.

In the literature, two approaches to the estimation of the proper size of the hidden layer are discussed. One is to start with a small network and iteratively increase the number of nodes in the hidden layer(s) until satisfactory learning is achieved, termed as constructive technique (Hernandez-Espinosa and Fernandez-Redondo, 2002). However, the problems are that these networks are also more likely to be trapped in a local minima (Bebis and Georgiopoulos, 1994), and a large number of networks must be trained to find the optimum network structure, which is time consuming. The second approach is to begin with a larger network and make it smaller by iteratively eliminating nodes in the hidden layer(s) or interconnections between nodes, termed as **Pruning** (Kavzoglu and Mather, 1998). Optimum brain damage, optimum brain surgery and skeletonization are the major pruning techniques in use.

As documented and shown by Lancashire et al., using two nodes in the hidden layer can successfully converge the variance of weights and give a higher performance on the validation subset (Lancashire et al., 2009), we have utilised the model architecture with two nodes in the hidden layer.

2.6 Variants of neural network architectures

- **Feed-forward Neural Network** – The feed-forward neural network was the first and arguably most simple type of artificial neural network devised. In this network the information moves in only one direction i.e., forward; from input nodes to the output nodes through hidden nodes (if any). There are no cycles or loops in the network (Bishop and Hinton, 1995).
- **Radial Basis Function Neural Network** – A Radial Basis Function (RBF) network has a built in distance criterion with respect to a centre. RBFs are powerful techniques for interpolation in multidimensional space. These neural networks have the advantage of not suffering from local minima in the same way as multilayer perceptrons, but have the disadvantage of requiring good coverage of the input space by radial basis functions (Aslan et al., 2008).
- **Kohonen Self-organizing Neural Network** – The self-organizing map is a form of unsupervised learning. A set of artificial neurons learn to map points in an input space to coordinates in an output space. The input space can have different dimensions and topology from the output space and the they will attempt to preserve these (Kohonen, 1990).

- **Learning Vector Quantization Neural Network** – Learning Vector Quantization can also be interpreted as a neural network architecture wherein the prototypical parameterisation of classes is represented with an appropriate distance measure. Hence, it is a form of distance-based classification (Burrascano, 1991).
- **Recurrent Neural Networks** – Recurrent neural networks (RNNs) are models with bi-directional data flow. Recurrent neural networks can be used as general sequence processors. Various types of Recurrent neural networks are fully recurrent network (Hopfield network and Boltzmann machine), Simple recurrent networks, Echo state network, long short term memory network, Bi-directional RNN, Hierarchical RNN, and Stochastic neural networks (Williams and Zipser, 1989).
- **Modular Neural Network** – Biological studies have shown that the human brain functions not as a single massive network, but as a collection of small networks. This realization gave birth to the concept of modular neural networks, in which several small networks cooperate or compete to solve problems (Happel and Murrae, 1994).

Other Special Types of Neural Networks

- **Holographic associative memory** – Holographic associative memory represents a family of analogues, correlation-based, associative, stimulus-response memories, where information is mapped onto the phase orientation of complex numbers operating (Khan, 1998).

- **Instantaneously Trained Neural Networks** – Instantaneously trained neural networks were inspired by the phenomenon of short-term learning that seems to occur instantaneously and have been extensively used in web based search algorithms (Kak, 1999).
- **Spiking Neural Networks** – Spiking neural networks are models which explicitly take into account the timing of inputs. The network input and output are usually represented as series of spikes (delta function or more complex shapes). These have an advantage of being able to process information in the time domain (signals that vary over time) (Jin et al., 2008b).
- **Dynamic Neural Networks** – Dynamic neural networks not only deal with nonlinear multivariate behaviour, but also include learning of time dependent behaviour such as various transient phenomena and delay effects (Sinha et al., 2000).
- **Cascade Correlation Neural Networks** – Cascade Correlation is a supervised learning architecture that begins with a minimal network and then automatically trains and adds new hidden units one by one, creating a multi-layer structure (Riley et al., 2010).
- **Neuro-Fuzzy Neural Networks** – A neuro-fuzzy network is a fuzzy inference system encoded within an artificial neural network. Depending on the inference system type, there are several layers that simulate the processes involved in a fuzzy inference engine i.e., fuzzification, aggregation and defuzzification. Embedding an fuzzy inference system in a

general structure of an ANN has the benefit of using available ANN training methods to find the parameters of a fuzzy system (Kasabov, 2001).

- **Compositional Pattern-producing Neural Networks** – Compositional pattern-producing networks are a variation of ANNs which differ in their set of activation functions and how they are applied. While typical ANNs often contain only sigmoid functions (and sometimes Gaussian functions), these can include both types of functions and many others (Jin et al., 2012).

2.7 Stepwise ANN approach

ANNs have proven their ability in the field of data-mining and high-throughput analysis of microarray data. However, one of the major limitation of ANNs in high throughput data analysis is the **Curse of Dimensionality** stated by Bellman as "*the exponential growth of the input space as a function of dimensionality*" (Wilcox, 1961). The importance of a particular feature of a gene can be hidden amongst the vast amount of other gene expression vectors. This occurs when the number of variables is far larger than the number of variables in the data. This additional noise from irrelevant inputs in the data space can hinder the performance of the model on unknown or validation data (Bishop and Hinton, 1995). Various pre-processing and data-reductionist approaches such as Principle Component Analysis (PCA) are widely used for tackling dimensionality aspect, but generalization and feature extraction aspects are still challenging.

A Stepwise ANN approach enables identification of patterns within the data space by finding the set of individual variables giving the best performance to classify the dataset. The learning in the model is attained by an iterative manner along with the subsequent addition of variables. Initially the approach uses each gene from the microarray experiment as an input for the model. A total of *n* individual models are created, each representing a gene of the dataset. Subsequently, all the models are compared and sorted according to their predictive performances for the unseen cases from test class of samples. These learned weights and model specification are then applied for the further (*n-1*) instances of input facilitating a stepwise-additive features contributing to the maximum predictive performance during each iteration. The whole process is repeated until no further improvement in predictive performance is seen in the model or until the model reaches an optimal level of performance. The overall step-by-step events happening during Stepwise ANN training can be summarized as below

- Each variable from the sample data set are assigned as the sole input node and used in the model.
- Each model is trained for 50 times with random reshuffling of samples with 60% of the dataset used for training, 20% for testing and the remaining 20% for validation using Monte Carlo Cross Validation (MCCV) strategy. This random reshuffling of samples produces a more generalised model with improved predictive ability of networks for unseen cases.

- From a separate test subset of samples, the predictions and Mean Square Errors (MSE) are recorded for each input model and ranked based of their MSE.
- The input within the model predicting the best is then selected for the second step.
- Subsequently, in the further steps, the input that performed the best in the previous run is removed from the input and remaining n-1 inputs are used as the basis for the next-input models, every time creating n-1 models.
- The process is repeated until no improvement in the network performance is seen, or until the early termination threshold is met with the test set of samples.

Figure 7 is a schematic representation of stepwise-ANN training. Briefly, The architecture of the stepwise ANN consisted of a single hidden layer MLP with two hidden nodes using a BP learning algorithm. The learning process is halted with an early stopping strategy to overcome over-fitting of the data by using a test subset of samples with maximum of 3000 epochs and a minimum window threshold of 1000 epochs if no further improvement in learning is seen. The BP algorithm was controlled with a learning rate of 0.1 and momentum of 0.5. The initial weights of the model are set between 1 to -1.

During this entire Ph.D. work, with some minor modifications, the architecture and the generalization settings of the stepwise ANN was maintained

as bespoke through an in-house developed and validated stepwise ANN software (Lancashire et al., 2009, Lancashire et al., 2008).

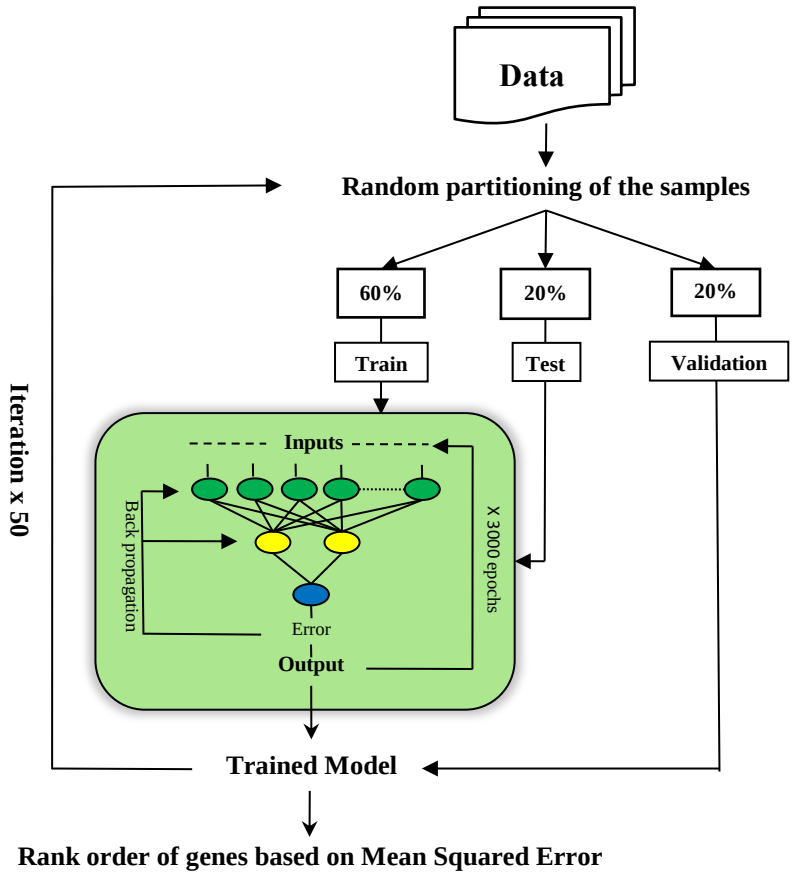


Figure 7: Stepwise-ANN model. Schematic representation of Stepwise-ANN model. The data analysis begin with a random stratification of samples into training, test and validation subsets. Feed forward ANN with back-propagation of errors was incorporated with Monte Carlo Cross validation strategy. The final ranking of best predictive gene for classification uses MSE.

2.8 Interaction algorithm

With the advent of gene expression profiling technologies, many researchers have analysed high throughput data for the identification of biomarkers and

signatures associated with specific disease or disease state. Others have also attempted to decipher the biological pathways. Some methods such as protein or gene interaction networks have already been described, among which Bayesian based approaches are common (Hartemink et al., 2002). Some have even tried to interpret the functionality of the interaction in protein complexes (Shoemaker and Panchenko, 2007, Spirin and Mirny, 2003). Karlebach and Shamir in their work have focused on gene regulatory network modelling (Karlebach and Shamir, 2008).

However, the potential information contained within them is not exploited fully. Recently, some researchers have used microarray data to investigate gene regulation with respect to particular gene and gene-gene regulation studies in the quest for more information. Often, the information concerning the interaction is limited and measured from the perspective of the target and not within the complete gene set. Moreover, Barabasi and Otva state that "*it is increasingly clear that a discrete biological function can rarely be attributed to an individual molecule*" (Barabasi and Oltvai, 2004). Thus, identification of a relevant gene signature for each specific disease is seen to be important, with the investigation of association between these genes being crucial.

The limitation of current interaction inferencing approaches are that they consider only one, or at best, very few counter-acting genes for the study. The interactions between the complete gene set with respect to each other can define how a given gene set change in expression subsequently changes the expression of other functionally relevant genes.

To overcome these limitations, Lemetre *et al*. has proposed a novel approach that considers the interactions of a whole gene set using an ANN trained with error back-propagation algorithm and Monte Carlo Cross Validation (MCCV) strategy (Lemetre et al., 2009).

2.8.1 Overview of interaction algorithm

This interaction algorithm measures the predictiveness of one gene with respect to all the others i.e., it predicts the expression of a single input gene by considering all the other genes as inputs, until all the inputs are predicted iteratively within the whole gene set. The weights connecting a given input to the output is then analysed by taking a sum of the weights leading to the output. This effectively represents intensity of the relation between a source (input) and the target (output). The predictive performance of the models and weights are monitored to represent the interaction vector. The input-output interaction vector determines the direction and the value of the weight representing the magnitude of the vector. Applying a threshold limit to all these interaction vectors allows a researcher to infer the information such as bi or unidirectional, stimulatory or inhibitory nature of the interaction. This study progresses upon the former methods by allowing direct prediction of association between expressions of genes.

2.8.2 Interaction model development

The interaction model is based on a bespoke ANN algorithm (Lemetre et al., 2009). Briefly, the interaction algorithm consists of a three-layer feed forward perceptron model trained with the error back-propagation algorithm. The activation function used in the hidden and output layer is of sigmoidal nature. A two and one node architecture in the hidden and output layers is maintained throughout. The learning was performed through 3000 epochs with a termination criteria of 1000 epochs without improvement of MSE on the validation subset or MSE threshold of 0.1. The momentum and learning rate were set at 0.5 and 0.1 respectively. Prior to the training, a MCCV strategy was applied for random segregation of samples into subsets as fraction of 60:20:20 for training, testing and validation respectively. The whole process of training was repeated 50 times independently. The weights generated with respect to input-output relation were recorded. The input to output connection represents the directionality of the association, the value of weights represents the intensity of the association between the source and target and the sign of the weight represents the nature of the association. By considering the weights with respect to input-output relation, a matrix of interactions can be generated.

2.8.3 Filtering

The matrix of interactions generated represents the association between any pair of variables contained in the dataset, producing huge amount of results

wherein *n* number of inputs gives *n(n-1)* number of associations. Hence, most of the associations may be non-significant. Non-significant associations can be removed by an arbitrary threshold to select only relevant and significant associations (Lemetre et al., 2009).

2.8.4 Visualisation of model

An interaction map of the matrix of interactions is then produced with the source and targets being symbolised by nodes, the association between them being represented by the edge, the intensity of association by the thickness of the edge and an arrow giving the directionality. For the visualisation of interaction maps, a freely available visualisation software, Cytoscape® version 2.7.3 was used (www.cytoscape.org).

2.9 Comparison of ANN with other methods

In recent years, considerable attention has been given to the development of sophisticated techniques for exploring relevant information from datasets. One such aspect which has been taken with great concern is classification. In medical applications, the goal is often to predict a patient's outcome based on available data. For studies with a binary endpoint (e.g., yes/no, alive/dead), regression based methods are used frequently. Daniel Sargent has reviewed 28 studies and

emphasised the usage of ANN along with other statistical methods such as Logistic and Cox proportional hazards regression methods for classification of medical data (Sargent, 2001). This review concludes that ANN outperforms regression-based methods.

ANNs have also been compared with logistic regression and SVMs in identification of circulatory biomarkers, wherein ANNs outperformed SVMs and logistic regression in classification and has confirmed the outperformance of ANN over other studied machine learning methods (Lancashire et al., 2010).

2.10 Advantages and disadvantages of ANN

Advantages of ANNs

- Ease to optimize, resulting in cost-effective and flexible in handling large data sets.
- They can process complex and non-linear data which are difficult by conventional linear methods.
- They can tolerate incomplete and fuzzy data.
- Based on their training tends, they can generalize and predict the future outcome of even a completely new test sample.
- With their reproducibility and accuracy for prediction, they can potentially support clinical decision making.

Disadvantages of ANNs

Like all other machine learning approaches, ANN also inherits limitations.

- The major limitation being the time consuming factor depending on the dimensionality and complexity of the data being modelled. As the total number of hidden layers required to capture the features of the data increases, so the time taken for training also increases.
- Over-fitting of the learned patterns by an ANN represents memorization through training process, and causes poor performance of network.
- Lastly, the apparent 'Black Box' situation exist wherein the path to reach a solution is unclear in an ANN, creating a barrier among researches for effective use of real-world data scenario.

2.11 ANNs in medicine and Biomarker discovery

The use of machine learning, and in particular the use of ANNs is increasing and occupying a central position as a predictive modelling tool in areas as diverse as computer-aided medical diagnosis, biological sequence analysis and prediction of peptide cleavage sites (Jagla and Schuchhardt, 2000). In an editorial letter, Gohari et al has supported the use of ANN in identification of prognostic factors in colorectal cancer patients (Gohari et al., 2011).

Lisboa et al. has extensively reviewed the ANN's contribution in clinical diagnosis, prognosis of disease and survival analysis in the domains of oncology, critical care and cardiovascular medicine, primarily focussing on randomised

controlled and clinical trials (Lisboa, 2002). The review emphasises that the ANN's role in evidence-based predictions is to reduce the information overload. The review also states that the use of ANNs in exploratory data analysis wherein complex hypothesis generation along with standard statistical methods is undertaken. A review on the usage of ANN in cancer clinical decision support can also be found (Lisboa and Taktak, 2006).

Agotonovic-Kustrin et al has reviewed ANN's application in pharmaceutical research and has recognised the potential application in pattern recognition, interpretation of analysed data, modelling drug response, drug and dosage prediction, formulation of product, modelling quantitative structure-property relationship, protein structure and function prediction. Analysis of pharmacokinetics-pharmacodynamics data is emphasised (Agatonovic-Kustrin and Beresford, 2000).

ANNs have been used in lung cancer cell identification through an automatic pathological diagnosis procedure named Neural Ensemble-based Detection method (Zhou et al., 2002). ANNs have also been used in diagnosis of breast cancer (Abbass, 2002), polycystic ovary syndrome (Matharoo-Ball et al., 2007), down syndrome (Coppede et al., 2010) and also for headache (Mendes et al., 2010).

ANNs are also used in identification of biomarker (Lancashire et al., 2005, Gonzalez et al., 2010), and survival analysis in breast cancer (Chi et al., 2007), using microarray and proteomic datasets (Lancashire et al., 2009). Lancashire has

also extended the ability of ANNs in identification of biomarkers in circulating blood (Lancashire et al., 2010).

ANNs have been found efficient in handling complex biological data from high-throughput techniques such as gene microarray (Huang et al., 2003, Lancashire et al., 2010) and mass spectrometry (Ball et al., 2002). ANNs have also been used in predictive non-linear modelling (Almeida, 2002). ANNs are not only used for the classification of samples, but also in inferring gene-gene interactions (Lemetre et al., 2009) and in discovering DNA regulatory elements (Firpi et al., 2010). Finally, ANNs in combination with genetic algorithms have also been used for feature selection (Tong and Schierz, 2011).

2.12 Systems biology approach using ANN

The advent of high throughput technologies has enabled a new and promising field of research called Systems Biology. Systems biology is a biology-based inter-disciplinary science of study that emphasises the intricate interactions of biological systems, using a more holistic perspective than the traditional reductionism, approach to biological and biomedical research (Kitano, 2002b).

"The reductionist approach has successfully identified most of the components and many of the interactions but, unfortunately, offers no convincing concepts or methods to understand how system properties emerge...the pluralism of causes and effects in biological networks is better addressed by observing,

through quantitative measures, multiple components simultaneously and by rigorous data integration with mathematical models" (Sauer et al., 2007).

Some have even argued that the systems biology can be seen as a conceptual approach to biological research that consciously combines 'reductionist' and 'integrationist' research, to understand the nature and maintenance of system as a whole (Kohl and Noble, 2009).

The challenges in systems biology is the complexity and dimensionality of data along with bias or noise induced by high throughput techniques. ANNs coupled with a robust cross-validation technique have been found to be efficient in addressing these challenges (Hart et al., 2006), and have motivated their application in systems biology of breast cancer.

Chapter 3

CLASSIFICATION OF OESTROGEN RECEPTOR STATUS IN A BREAST CANCER MICROARRAY DATASET USING ARTIFICIAL NEURAL NETWORKS

Chapter abstract: The classification of samples based on ER status (ER+/ER-) is vital for predicting the prognosis and endocrine treatment in breast cancer. Furthermore, the investigation of predictive gene signature for the ER class would benefit an understanding of the causal influence of ER-associated genes. Aiming to bespoke aspects, in this study we have utilised an in-house developed Stepwise-ANN predictive model to classify breast cancer microarray dataset samples based on ER status and to generate a predictive gene signature for the ER class. The stepwise-ANN model identified *ESR1* as the best predictive gene for the ER class with a consistent performance of 90.45% on the test subset and 89.45% on the validation subset of samples. The classification accuracy of the ANN for ER status was found to be 87.72% with 95.09% sensitivity and 77.19% specificity. The ANN showed higher accuracy of 90.97% when using the median-based gene expression for *ESR1* gene coding for ERα. The gene expression based classification also showed higher sensitivity and specificity. The ontological investigation of the top 100 predictive markers for ER status was dominated with catalytic activity in molecular process

group and metabolic process in biological function group. Finally, the Stepwise-ANN also showed consistent performance on external validation of breast cancer datasets.

3.1 Introduction

3.1.1 Need for ER associated gene markers

By quantitative Reverse Transcription-Polymerase Chain Reaction (RT-PCR), immunohistochemistry (IHC) and Fluorescent In-Situ Hybridisation (FISH) on breast cancer samples, multiple potential biomarkers and proliferation signatures with respect to a particular disease state or condition can be studied (Cheang et al., 2009). Through gene expression profiling on archival breast Tissue MicroArray (TMA) blocks of tumour tissue samples, it was observed that, the ontological distribution of previously studied genes can be evaluated to determine their role in breast cancer (Cimino et al., 2008). Meta-analysis of breast cancer microarray cDNA data for the expression of genes along with their transcriptional activators can facilitate tumour prediction in early stage breast cancer (Mehra et al., 2005). For improvement in the prognostic and therapeutic aspects of breast cancer, gene amplification by array-comparative genomic hybridisation has also been carried out using gene amplification techniques such as Chromosomal In-Situ Hybridisation (CISH) and FISH (Arnould et al., 2012). The conventional practice of identifying a single gene responsible for phenotype and cellular functionality lacks information

regarding the influence of associated genes, which can be overcome by investigating gene signatures using microarray technology (Simon, 2003). A gene signature representing disease state, such as patient diagnosis and classification, can help clinicians in treatment decisions appropriate to individual patients.

The advantage of using ANNs in this study was supported with their ability to handle non-linearity which is a characteristic of biological data. The successful implementation of ANN in learning the gene expression patterns in microarray data by Lancashire at el has provoked this study (Lancashire et al., 2009). ANNs with their capacity of sample classification and gene signature development depending on the predictability for the class could be used to shortlist contenders associated with the classification feature. Thus, the evaluation of information shared by the ER-associated genes on ER system could be carried out.

3.2 Aims and Objectives

- To successfully classify breast cancer samples based on the ER status derived from immunohistochemical method using an in-house developed Stepwise ANN model.
- To generate a gene signature for ER status in breast cancer microarray dataset.
- To investigate the efficiency of the predictive model for the ER status classification.
- To validate the predictive efficiency of ANN for ER classification on multiple microarray datasets.

3.3 Materials and Methods

3.3.1 E-GEOD-20194 dataset

A pre-normalized microarray breast cancer cDNA expression profile was downloaded from the freely available EMBL-EBI database library (http://www.ebi.ac.uk/arrayexpress). The cDNA microarray profile was found to be developed on an Affymetrix HG-U133A array from the E-GEOD-20194 study (Shi et al., 2010). The breast cancer microarray dataset contained 22283 probesets corresponding to genes across 278 primary breast cancer samples before neoadjuvant therapy administration. The samples were categorized into 2 classes based on IHC ER status as ER-positive (164 samples) class and the remaining as ER-negative (114 samples).

3.3.2 Other breast cancer datasets

Apart from E-GEOD-20194, other breast cancer cDNA microarray datasets developed on HG-U133A array were also selected from the *Gene Expression Omnibus* (GEO) library *ArrayExpress*.

E-GEOD-2034: This dataset was submitted to the GEO library as GEO2034 series. It consists of 180 lymph-node negative relapse free patients and 106 lymph-node negate patients that developed a distant metastasis. The dataset was designed to develop a gene signature to predict the distant metastasis in lymph-node-negative primary breast cancer patients (Wang et al., 2005). In their study, Wang et al. proposed a 76 gene signature predicted by univariate Cox's regression model

coupled with bootstraps and novel relapse score. The gene signature comprised 60 genes for ER-positive and 16 genes for ER-negative samples without any overlaps of genes was developed (Wang et al., 2005).

E-GEOD-7390: This study was designed to independently validate the E-GEOD-2034 study on a cohort from the TRANSBIG Consortium (Desmedt et al., 2007). Gene expression profiling of 198 fresh frozen samples of lymph node-negative breast cancer patients was carried out at Bordet Institute, Brussels, Belgium. This confirmed the performance of the 76-gene signature published earlier to be a good predictor for 5-years distance metastasis and overall survival in a so called 'good profile' group compared to a 10-years group.

E-TABM-158: This dataset was designed to explore the association of genomic copy number abnormalities in aggressive early-stage breast cancer patients using gene expression profiling along with clinical outcomes. A 66-gene signature was developed which had altered copy numbers in sub categories of breast cancer and was found to be a potential therapeutic target (Chin et al., 2006).

Table 2 is a summary of the breast cancer datasets selected for validation of the Stepwise-ANN findings.

3.3.3 Stepwise ANN

An in-house developed Stepwise-ANN algorithm (Lancashire et al., 2009), was used for the prediction of ER status in samples from the E-GEOD-20194 breast

cancer dataset. All 22283 probesets were used as input variables for training the ANN to predict ER status. The training was performed with a momentum of 0.5 and learning rate of 0.1. The complete stepwise ANN coupled with back propagation of error algorithm including randomised sample partitioning (60% for training, 20% for testing and remaining 20% for validation) with MCCV strategy was iterated for 10 independent iterations (or runs). The predictive performance and errors, for each input variable across all samples, were recorded for each independent loop (complete run of predictive model) and for each subset of samples under training, test and validation. The rank order of the genes under each iteration was based on their Mean Square Error (MSE) values. Minimum Average Test Error (ATE) was used for short listing the predicting genes for ER status.

Table 2: Summary of breast cancer datasets considered for validation.

GEO ID	Total no. of probes	Total no. of samples	Class	
			ER-positive	ER-negative
E-GEOD-2034	22283	286	209	77
E-TABM-158	22215	118	75	43
E-GEOD-7390	22283	198	134	64

Summary of breast cancer datasets considered for validating Stepwise-ANN findings. Datasets are selected from Gene Expression Omnibus (GEO) library.

3.4 Results and Discussions

3.4.1 Stepwise ANN consistency

The in-house Stepwise-ANN implementation for predicting IHC ER status of breast cancer samples from microarray dataset E-GEOD-20194 yielded consistent results across 10 independent iterations of learning. The top ranked probe with predictive performances and errors from all 10 individual iterations are summarized in Table 3. The mean training performance of the single gene was found to be 90.45%. The validation subset of samples showed a mean performance of 89.45% with error rate of 11.70% indicating robustness and consistent performance across ten independent iterations of analysis.

Table 3: Summary of Stepwise-ANN outcome.

Run	Probe ID	Gene Symbol	Training Performance	Test Performance	Validation Performance	Training Error	Test Error	Validation Error
1	205225_at	ESR1	0.9042	0.8929	0.8909	0.1149	0.1102	0.1179
2	205225_at	ESR1	0.8982	0.8929	0.8909	0.1175	0.1165	0.1187
3	205225_at	ESR1	0.9102	0.9107	0.9091	0.1174	0.1178	0.1181
4	205225_at	ESR1	0.9072	0.9107	0.8909	0.1065	0.1071	0.1120
5	205225_at	ESR1	0.9042	0.8929	0.8909	0.1122	0.1142	0.1160
6	205225_at	ESR1	0.9042	0.9286	0.8909	0.1173	0.1111	0.1250
7	205225_at	ESR1	0.9042	0.9018	0.8909	0.1068	0.1088	0.1083
8	205225_at	ESR1	0.9042	0.8929	0.8909	0.1085	0.1128	0.1184
9	205225_at	ESR1	0.9042	0.9110	0.9090	0.1130	0.1120	0.1150
10	205225_at	ESR1	0.9042	0.9107	0.8909	0.1156	0.1110	0.1208
		MEAN	**0.9045**	**0.9045**	**0.8945**	**0.1130**	**0.1122**	**0.1170**
		SD	**0.0029**	**0.0119**	**0.0076**	**0.0043**	**0.0033**	**0.0045**

Summarization of Stepwise-ANN classification for ER status. Top ranked gene from ten independent iterations was selected based on minimum averaged test errors recorded under each loop.

The cross examination of performances under all ten independent analysis was also found to be consistent with standard deviation of 0.76% and 0.45% in performance and error rates respectively in the validation subset of samples indicating reproducibility of the model performance.

The top ranked probe found to be highly associated with the IHC ER status prediction in the samples was 205225_at, coding for the gene *ESR1*, the ERα subunit of the receptor. *ESR1* was found to be highly associated with the ER class prediction in all the random sample subset segregations indicating consistent marker selection for class prediction and parameterisation capacity of the ANN model.

Furthermore, the performance of Stepwise-ANN was investigated with respect to errors generated in test and validation subsets of sample segregation under independent runs. Figure 8 illustrates the performance of Stepwise-ANN for classification of samples for ER status. The overall averaged error generated by Stepwise-ANN under 10 independent models in test subset was found to be of range 0.097 to 0.115 and in validation subset was found to be of range 0.113 to 0.125. The error generated by ranked 1 probeset (205225_at representing *ESR1* gene) under 5 independent repeats of Stepwise-ANN comprising 10 independent models each was also investigated. The test errors generated was found to be in range from 0.1048 to 0.1111 and validation errors was found to be in range from 0.11 to 0.12. The performance of Stepwise-ANN was investigated with respect to validation subset under 5 independent repeats comprising 10 independent models and was found to be in range from 0.8945 to 0.9027. Hence, the Stepwise-ANN performance was found to be consistence across independent models and independent repeats also.

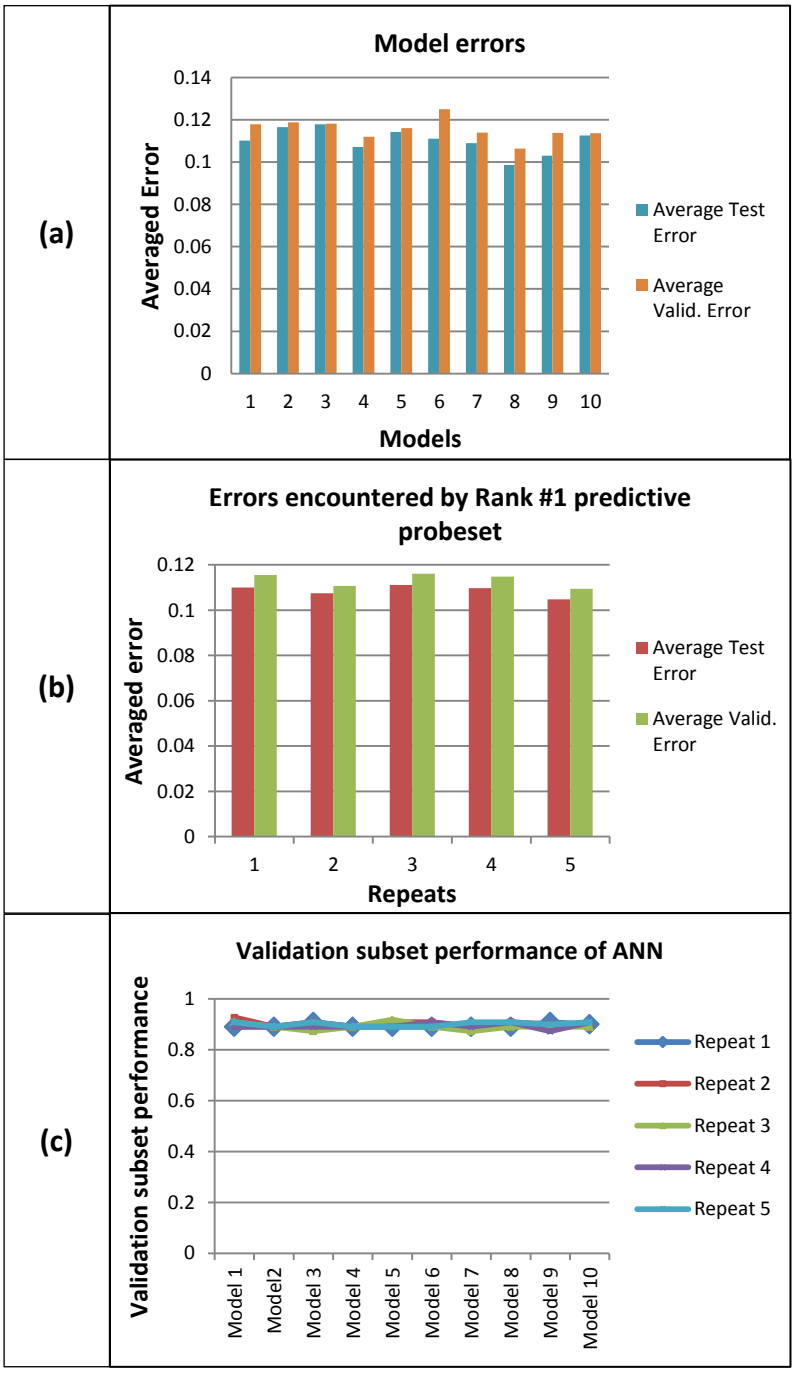


Figure 8: ANN model performance consistency. (a) Test and Validation subset errors generated by ANN under 10 independent models. **(b)** Error encountered by ranked 1 predictive probeset under Test and Validation subset at 5 independent repeats of ANN model. **(c)** Validation subset performance of 10 independent models and 5 independent repeats of ANN.

3.4.2 Classification accuracy of Stepwise-ANN

The ability of a predictive model to effectively classify the unseen data would largely depend on the nature of the learned pattern from the seen data used for the training. The stepwise ANN model used was found to efficiently classify all the 278 samples from the dataset used with an accuracy of 87.72% with a single gene model. By considering actual and predicted classes of samples, the class prediction accuracy was confirmed with 95.09% of sensitivity (true positive rate; ability to identify positive cases) and 77.19% specificity (true negative rate; ability to identify negative cases) respectively yielding to the overall prediction accuracy of 87.72%.

The ROC curve was found to be in the upper left quadrant of the graph representing an efficient model for class prediction. The Area Under Curve (AUC) or Receiver Operating Characteristic (ROC) area under stratified sample subsets was also found to be 93.02% for training (N=166), 91.70% for test (N=56) and 93.46% for validation (N=56) subset with random MCCV strategy implied. The prediction accuracy was found to be 90.45% in training, 90.45% in test and 89.45% in validation subset of the samples, confirming the efficiency of the predictive model and the classification ability of the Stepwise-ANN towards the unseen data. Table 4 shows the classification accuracy and ROC area for ER class prediction.

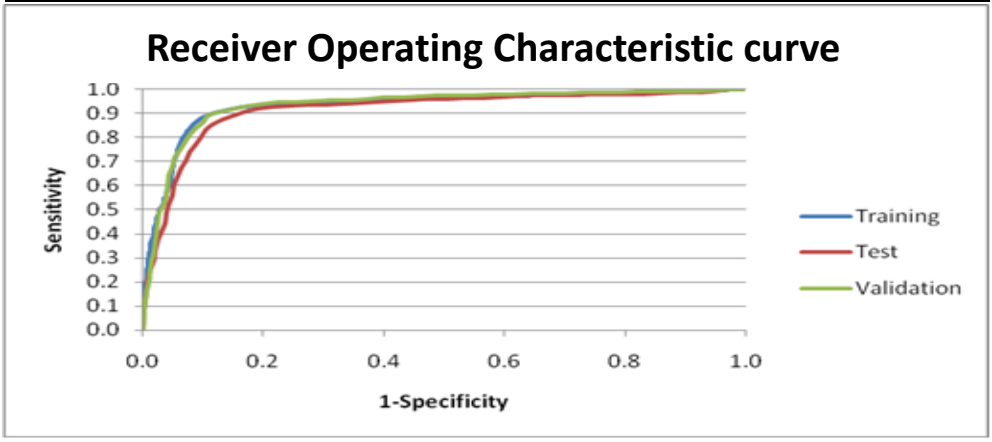
3.4.3 Top100 probe list for ER status from Stepwise-ANN

The Stepwise-ANN with consistent high performance in all the 10 independent predictive models generated a list of best predictive markers for the ER class from

the features and samples given for training. The MSE values from the testing subset of samples were used for the generation of rank order. The gene with the minimum error value in all the predictive models was given the highest rank and so on. Here, for the sake of simplicity, the 100 top genes were selected for further analysis and information mining to justify findings (Table 5). The top 100 genes were found to have high statistical significance having p-values in below $4.49X10^{-3}$. Hence, the top ranked 100 probes were considered for further analysis.

Table 4: Sensitivity, Specificity, Classification accuracy and ROC analysis.

True Positive	True Negative	False Positive	False Negative	Sensitivity	Specificity	Accuracy
155	88	26	9	95.09%	77.19%	87.72%



	Training samples	Test samples	Validation samples
Sample size	166	56	56
Prediction accuracy (%)	90.45	90.45	89.45
ROC area	**0.9302**	**0.9170**	**0.9346**

Tabulation of classification accuracies recorded by Stepwise-ANN for all samples selected during random stratification of samples into training, test and validation subsets. Sensitivity and specificity of the predictive model is also represented in terms of true and false positivity and negativity. Receiver operating Characteristic (ROC) curve analysis and ROC area under curve values represents models accuracy.

Table 5: Summary of top ranked 100 genes from stepwise ANN.

Rank Order	Probe set ID	Gene Symbol	p-value	Regulation
1	205225_at	ESR1	4.49E-05	Up
2	209603_at	GATA3	8.98E-05	Up
3	212956_at	KIAA0882	1.35E-04	Up
4	214164_x_at	FLJ20151	1.80E-04	Up
5	215867_x_at	CA12	2.24E-04	Up
6	209602_s_at	GATA3	2.69E-04	Up
7	214440_at	NAT1	3.14E-04	Up
8	204508_s_at	FLJ20151	3.59E-04	Up
9	209604_s_at	GATA3	4.04E-04	Up
10	203963_at	CA12	4.49E-04	Up
11	218195_at	FLJ12910	4.94E-04	Up
12	212960_at	KIAA0882	5.39E-04	Down
13	209173_at	AGR2	5.83E-04	Up
14	210735_s_at	CA12	6.28E-04	Up
15	203628_at	IGF1R	6.73E-04	Up
16	211712_s_at	ANXA9	7.18E-04	Up
17	218976_at	JDP1	7.63E-04	Up
18	212196_at	IL6ST	8.08E-04	Up
19	205009_at	TFF1	8.53E-04	Up
20	202089_s_at	LIV-1	8.98E-04	Up
21	205696_s_at	GFRA1	9.42E-04	Up
22	218211_s_at	MLPH	9.87E-04	Up
23	205862_at	GREB1	1.03E-03	Up
24	200670_at	XBP1	1.08E-03	Up
25	212496_s_at	KIAA0876	1.12E-03	Down
26	202088_at	LIV-1	1.17E-03	Down
27	209623_at	MCCC2	1.21E-03	Down
28	218807_at	VAV3	1.26E-03	Up
29	204667_at	FOXA1	1.30E-03	Up
30	215729_s_at	TONDU	1.35E-03	Up
31	212195_at	IL6ST	1.39E-03	Up
32	210652_s_at	C1orf34	1.44E-03	Down
33	217838_s_at	RNB6	1.48E-03	Down
34	205186_at	DNALI1	1.53E-03	Up
35	208682_s_at	MAGED2	1.57E-03	Up
36	209460_at	NPD009	1.62E-03	Up
37	221765_at	UGCG	1.66E-03	Down
38	212442_s_at	LASS6	1.71E-03	Down
39	211000_s_at	IL6ST	1.75E-03	Down
40	212492_s_at	KIAA0876	1.80E-03	Down
41	219051_x_at	MGC2601	1.84E-03	Down
42	206401_s_at	MAPT	1.89E-03	Up
43	209459_s_at	NPD009	1.93E-03	Down
44	212638_s_at	WWP1	1.98E-03	Up
45	201508_at	IGFBP4	2.02E-03	Down
46	215552_s_at	ESR1	2.06E-03	Up
47	210085_s_at	ANXA9	2.11E-03	Down
48	205471_s_at	DACH	2.15E-03	Up
49	201826_s_at	CGI-49	2.20E-03	Down
50	214552_s_at	RAB5EP	2.24E-03	Up

Rank Order	Probe set ID	Gene Symbol	p-value	Regulation
51	203749_s_at	RARA	2.29E-03	Up
52	219197_s_at	SCUBE2	2.33E-03	Down
53	204881_s_at	UGCG	2.38E-03	Up
54	204798_at	MYB	2.42E-03	Down
55	205355_at	ACADSB	2.47E-03	Down
56	218259_at	KIAA1243	2.51E-03	Down
57	206754_s_at	CYP2B6	2.56E-03	Up
58	219414_at	CLSTN2	2.60E-03	Down
59	203929_s_at	MAPT	2.65E-03	Down
60	203144_s_at	KIAA0040	2.69E-03	Down
61	204623_at	TFF3	2.74E-03	Down
62	216381_x_at	AKR7A3	2.78E-03	Down
63	208615_s_at	PTP4A2	2.83E-03	Down
64	205354_at	GAMT	2.87E-03	Down
65	200711_s_at	SKP1A	2.92E-03	Down
66	222125_s_at	PH-4	2.96E-03	Down
67	204863_s_at	IL6ST	3.01E-03	Down
68	213419_at	APBB2	3.05E-03	Down
69	203256_at	CDH3	3.10E-03	Up
70	213712_at	ELOVL2	3.14E-03	Down
71	205597_at	C6orf29	3.19E-03	Down
72	201983_s_at	EGFR	3.23E-03	Down
73	35666_at	SEMA3F	3.28E-03	Down
74	203928_x_at	MAPT	3.32E-03	Up
75	212148_at	PBX1	3.37E-03	Down
76	205472_s_at	DACH	3.41E-03	Up
77	219913_s_at	CRNKL1	3.46E-03	Up
78	201754_at	COX6C	3.50E-03	Down
79	205081_at	CRIP1	3.55E-03	Down
80	212446_s_at	LASS6	3.59E-03	Down
81	221710_x_at	FLJ10647	3.64E-03	Down
82	41660_at	CELSR1	3.68E-03	Down
83	204497_at	ADCY9	3.73E-03	Up
84	203627_at	IGF1R	3.77E-03	Down
85	200810_s_at	CIRBP	3.82E-03	Down
86	208873_s_at	DP1	3.86E-03	Up
87	212099_at	RHOB	3.90E-03	Down
88	212151_at	PBX1	3.95E-03	Down
89	208617_s_at	PTP4A2	3.99E-03	Down
90	214404_x_at	PDEF	4.04E-03	Down
91	218806_s_at	VAV3	4.08E-03	Down
92	212441_at	KIAA0232	4.13E-03	Down
93	204862_s_at	NME3	4.17E-03	Down
94	205066_s_at	ENPP1	4.22E-03	Up
95	220192_x_at	PDEF	4.26E-03	Down
96	212637_s_at	WWP1	4.31E-03	Down
97	209696_at	FBP1	4.35E-03	Up
98	40148_at	APBB2	4.40E-03	Down
99	209443_at	SERPINA5	4.44E-03	Down
100	206469_x_at	AKR7A3	4.49E-03	Down

Tabulation of top 100 ranked gene probes found associated with ER status class prediction through in-house developed Stepwise-ANN analysis. The average of test subset error for each gene under each independent 10 iterations of stepwise analysis was considered for ranking. The probes with minimum error is ranked highest in the table.

The probe 209603_at representing *GATA3* gene appeared at second rank position for ER classification. The same gene was found to have another two probes on the array as 209602_at and 209604_s_at appearing in rank positions 6 and 9

respectively. The *GATA3* genes codes for the transcriptional factor belonging to the GATA family. The gene is located at chromosome 10p15 in humans. They are responsible in the regulation of T cells (Marine and Winoto, 1991), inner ear (Milo et al., 2009), adrenal gland and brain with central nervous system in mouse (Zhao et al., 2008). *GATA3* is also found to be responsible for self-renewal of hematopoietic stem cells (Buza-Vidas et al., 2011). Wilson et al 2008 have confirmed the reciprocal association of *GATA3* gene with the ER α gene from their meta-analysis study using microarrays (Wilson and Giguere, 2008). *GATA3* was found to be a prognostic marker (Mehra et al., 2005), and associated with hormonal responsiveness in breast cancer (Fang et al., 2009).

Another important gene found to be associated with ER status in the breast cancer samples was *CA12*. *CA12* represents the Carbonic anhydrases gene which is a member of zinc metalloprotiens family that catalyses the reversible hydration reaction of carbon dioxide. The gene is located at 15q22.2 in humans. *CA12* is found up regulated in renal cells (Tureci et al., 1998). They have been found to have a role in oestrogen metabolism and are good prognostic factor for invasive breast carcinoma (Watson et al., 2003). The stepwise-ANN identified three probes as 215867_x_at, 203963_at and 210735_s_at at ranks 5^{th}, 10^{th} and 14^{th} respectively.

To address the issue of probability of occurrence of any gene by random chance, we considered ranking each probeset based on their ATE values and p-value based on rank position with respect to all 22283 probesets.

Furthermore, the heat map of the top 100 predictive genes (Figure 9) was done. The predictive genes considered were able to classify samples clearly. The expression pattern revealed genes having distinct regulation in samples. Some genes were also found to be having contrasting expression pattern compared to others potentially segregating samples into different sub groups of breast cancer.

3.4.4 Comparison of ANN result with original study

In the original study by Liu et al using E-GEOD-20194, different machine learning algorithms such as Recursive Feature Elimination of Supervised Vector Machines and Gradient and Sequential forward method of Leave-one-out supervised machine learning algorithms were compared for predicting the complete pathological response using Naïve bayes classier. Liu et al has mentioned that the lack of commonality between the markers was expected due to the difference in the classier algorithms used to signify the effect of the feature selection method on the dataset. In justification for the lack of inconsistency, the Liu et al study suspects that the pre-filtering methods used for the data generation was independent on the learning methods, thus representing the weakness of filtering method which ignores the involvement of features in the data. Moreover, a discrepancy among the top ranked gene from the different classier used was observed and accepted (Liu et al., 2009).

We compared our results with the original study using E-GEOD-20194 breast cancer data set. A total of 9 genes overlapped with the genes reported by Liu et al

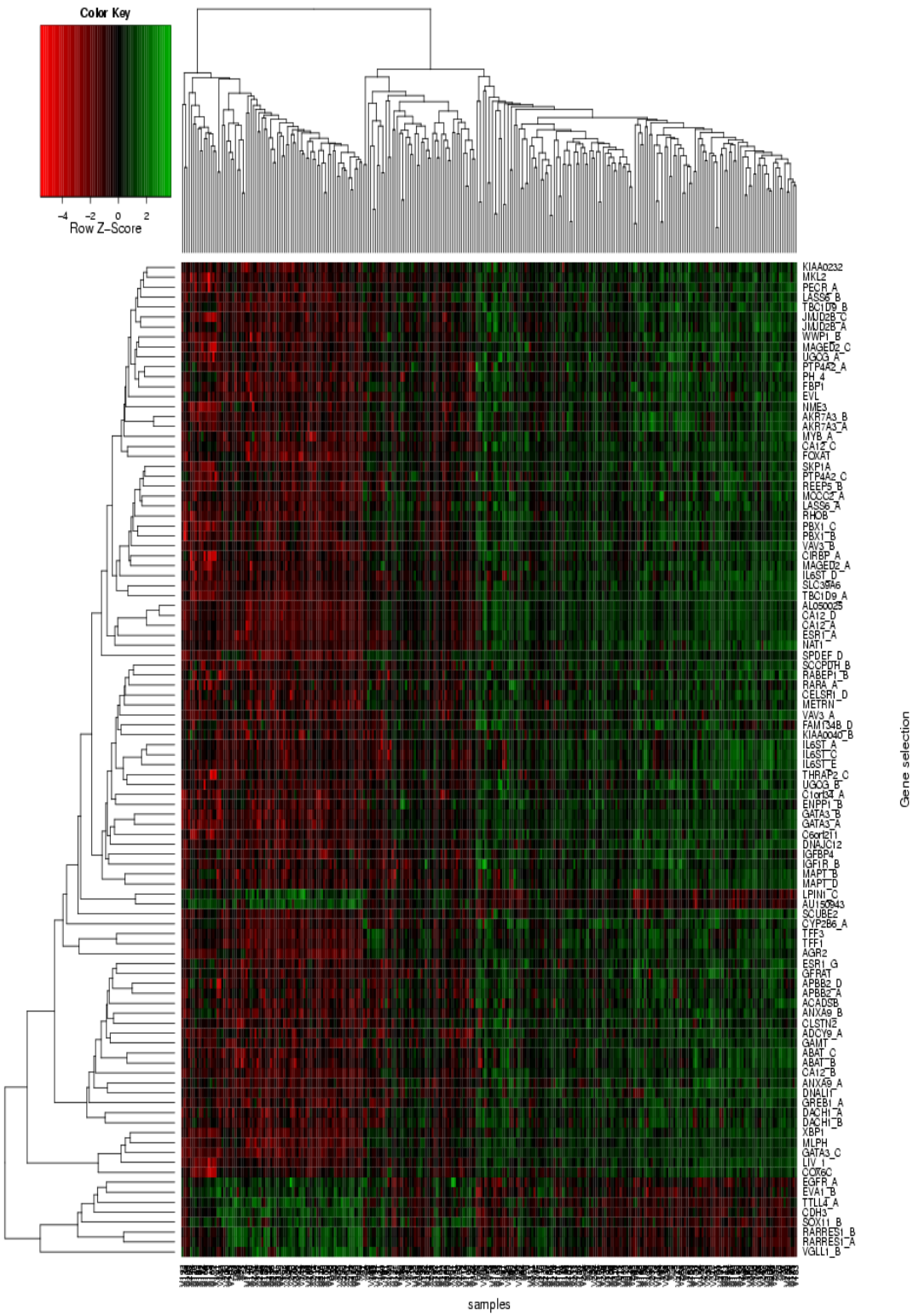


Figure 9: Heat map of the top 100 predictive genes. The heat map represents a dendogram of hierarchically clustered genes. The rows of heat map represents expression of genes (variants of the same genes are symbolized by alphabetical suffix) across corresponding samples in columns. Each cell is colorized based on the level of expression of that gene in that sample.

with our top 100 ER-associated genes from Stepwise-ANN method. These overlap genes are *EGFR* (201983_s_at), *ESR1* (205225_at), *IL6ST* (211000_s_at and 212195_at), *MAPT* (203929_s_at), *DACH* (205471_s_at), *GATA3* (209602_s_at), *GREB1* (205862_at), and *SERPINA5* (209443_at). Due to the unavailability of predictive performances of each genes from Liu et al., the rank order of genes could not be compared with our study.

3.4.5 ER-class prediction comparison between protein and gene expression based classification

The conventional protein expression based classification of samples through IHC method incorporates higher false positive rates due to inherent lower sensitivity and specificity of the technique (Gown, 2008). Upon classification of samples for ER status (IHC) provided by dataset using Stepwise-ANN method, a total of 34 samples were misclassified with 87.72% classification accuracy using a single high predictive gene at rank one. The sensitivity and specificity of the assay was found to be 95.09% and 77.19% respectively.

To improve the classification accuracy, the samples were classified based on the ESR1 gene expression values, since the *ESR1* gene coding ERα receptor was the best predictive gene for ER classification. The expression profile of the ESR1 gene across all 278 samples was considered for a median expression level based classification. Any sample's *ESR1* expression below the median value was considered to be ER-negative and those greater being ER-positive. The same

Stepwise-ANN model was applied to predict the *ESR1* gene expression based classification of the samples. The misclassification of samples was reduced to 25 with an improved classification accuracy of 90.97%. The sensitivity and specificity was also improved to 96.31% and 83.33% respectively. Promisingly, the false-positive samples were decreased from 26 to 5 which could be clinically beneficial to patients with endocrine and hormonal therapy. Figure 10 is the schematic representation of ER and *ESR1* expression distribution in samples. Table 6 is the ANN model performance for ER (protein) and *ESR1* (gene) expression based classification.

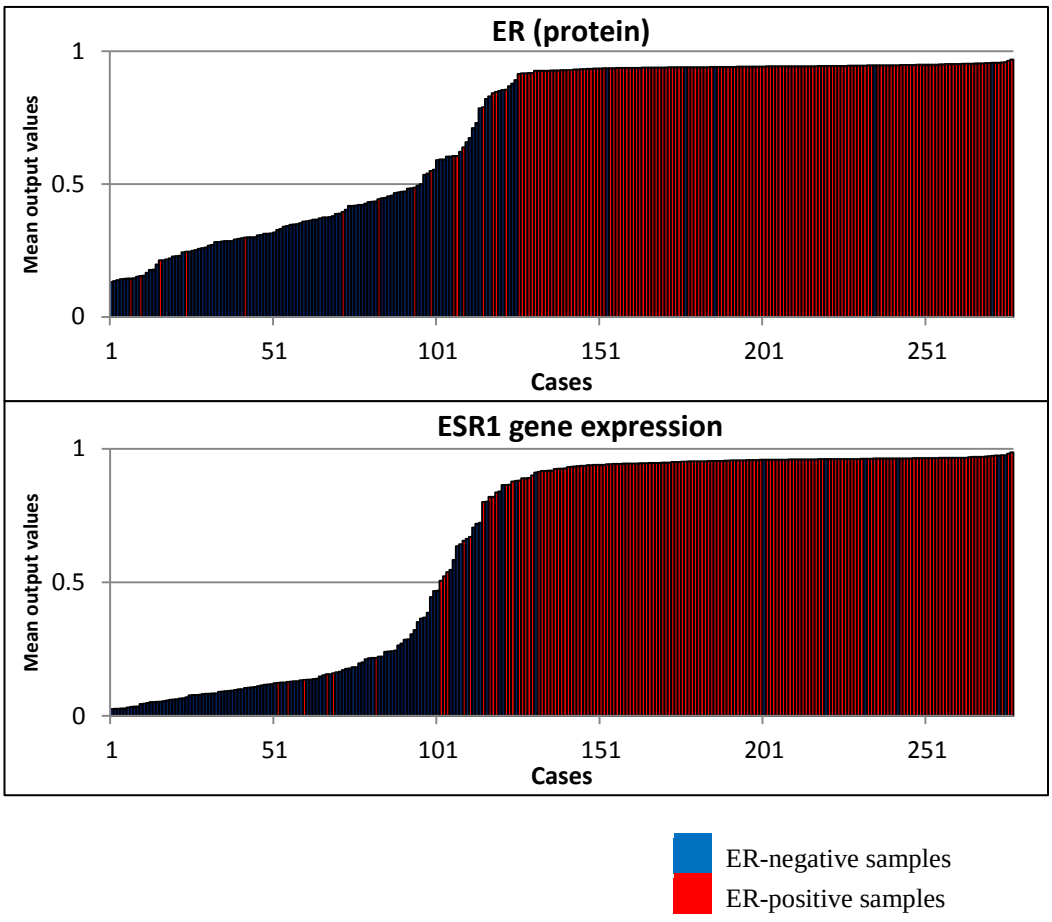


Figure 10: Population distribution of samples with IHC and gene expression based ER classification. Population distribution chart of breast cancer samples under protein by IHC and gene expression based ER classification. The prediction output values are normalised.

Table 6: Stepwise-ANN performance for ER (protein) and *ESR1* (gene) expression based classification.

	TP	TN	FP	FN	Sensitivity	Specificity	Accuracy
ER (protein)	155	88	26	9	95.09%	77.19%	87.72%
***ESR1* (gene)**	181	88	5	4	96.31%	83.33%	90.97%

The classification accuracy of the ANN model compared with respect to ER (protein) and ESR1 (gene) based classification measured in terms of sensitivity and specificity. TP = True Positive, FP = False Positive, TN = True Negative, FN = False Negative.

3.4.6 Investigation of probe variants

The HG U-133A array used in our study was found to contain multiple probes for particular genes. These were typically used for the internal calibration processes of the array among which some were house-keeping genes (for example, hum_alu_at representing human albumin gene). They may represent mutational variants of the wild type and different isoforms of the same genes, and some of the probes may correspond to a single protein or to particular polypeptides of the same protein. Furthermore, the difference in expression profile of these probe variants may also due to the degradation of mRNA in patient tumour samples.

3.4.6.1 Rank positions

The investigation of the top 100 predictive probesets for ER classification showed a total of 18 genes having multiple probesets. Among them, there were two

probesets representing the *ESR1* gene, 205225_at and 215552_s_at, which were found to be at rank 1 and 45 respectively. The stepwise-ANN also identified three probesets of *CD12* as 215867_x_at, 203963_at and 210735_s_at at ranks 5th, 10th and 14th respectively. The *GATA3* gene was found to have three different probesets 209603_at, 209602_s_at and 209604_s_at positioned at 2nd, 6th and 9th ranks. Table 7 is the tabulation of rank positions of repeated probes of the same gene found in the top 100 predictive probesets for ER classification using the Stepwise-ANN method. The appearance of multiple probes of the same gene at high rank order justifies that different probes of the same gene have similar ability to classify patterns for ER classes. It also shows that predictive models of ANN were efficient enough to recognize the similar expression pattern among the gene's probe variants.

3.4.6.2 Population distribution charts

Population distribution is the schematic representation of expression values of a genes or probe across samples to determine the expression profile patterns. The probe variants of the *ESR1* gene from the top 100 predictive probes for ER classification shows varied expression profiles across ER-positive and ER-negative samples i.e., the ranked number one, 205225_at probe shows a clear segregation of ER-positive and ER-negative samples, whereas the other probe, 215552_s_at at rank 46, classifies samples with a lower rate (Figure 11(a)). This suggests that the appearance of probes of the same gene at lower rank order contains higher classification error compared the higher ranked probes. On the other hand, different probes of the same gene were found to have a differential expression in ER-positive

Table 7: List of repeated probes for gene found in top 100 Stepwise-ANN.

Rank	Input ID	Accession number	Gene symbol	Gene name
68	213419_at	U62325	APBB2	amyloid beta (A4) precursor protein-binding, family B, member 2
98	40148_at	U62325	APBB2	
60	216381_x_at	AL035413	AKR7A3	Aldo-keto reductase family 7, member A3 (aflatoxin aldehyde reductase)
100	206469_x_at	NM_012067	AKR7A3	
17	211712_s_at	BC005830	ANXA9	Annexin A9
48	210085_s_at	AF230929	ANXA9	
10	203963_at	NM_001218	CA12	Carbonic anhydrase XII
13	210735_s_at	BC000278	CA12	
50	205471_s_at	AW772082	DACH1	dachshund homolog 1 (Drosophila)
69	205472_s_at	NM_004392	DACH1	
1	205225_at	NM_000125	ESR1	estrogen receptor 1
45	215552_s_at	AI073549	ESR1	
4	214164_x_at	BF752277	FLJ20151	
7	204508_s_at	BC001012	FLJ20151	
2	209603_at	AI796169	GATA3	GATA binding protein 3
5	209602_s_at	AI796169	GATA3	
8	209604_s_at	BC003070	GATA3	
37	211000_s_at	AB015706	IL6ST	interleukin 6 signal transducer
64	204863_s_at	BE856546	IL6ST	
3	212956_at	AI348094	KIAA0882 /TBC1D9	TBC1 domain family, member 9
12	212960_at	BE646554	KIAA0882 /TBC1D9	
19	202089_s_at	NM_012319	LIV-1 /SLC39A6	solute carrier family 39 (zinc transporter), member 6
29	202088_at	AI635449	LIV-1 /SLC39A6	
44	206401_s_at	J03778	MAPT	microtubule-associated protein tau
59	203929_s_at	AI056359	MAPT	
72	203928_x_at	AI870749	MAPT	
36	209460_at	AF237813	NPD009	
42	209459_s_at	AF237813	NPD009	
86	214404_x_at	AI435670	SPDEF	SAM pointed domain containing ets transcription factor
94	220192_x_at	NM_012391	SPDEF	
66	208615_s_at	BF795101	PTP4A2	protein tyrosine phosphatase type IVA, member 2
93	208617_s_at	AF208850	PTP4A2	
38	221765_at	AI378044	UGCG	UDP-glucose ceramide glucosyltransferase
55	204881_s_at	NM_003358	UGCG	
26	218807_at	NM_006113	VAV3	vav 3 guanine nucleotide exchange factor
88	218806_s_at	AF118887	VAV3	
43	212638_s_at	BF131791	WWP1	WW domain containing E3 ubiquitin protein ligase 1
96	212637_s_at	AU155187	WWP1	

The probe variants present in top 100 genes selected by Stepwise-ANN method. A total of 18 genes with multiple probe IDs were found to be associated with ER status in the array used. The ranks are sorted based on alphabetical order of gene symbols.

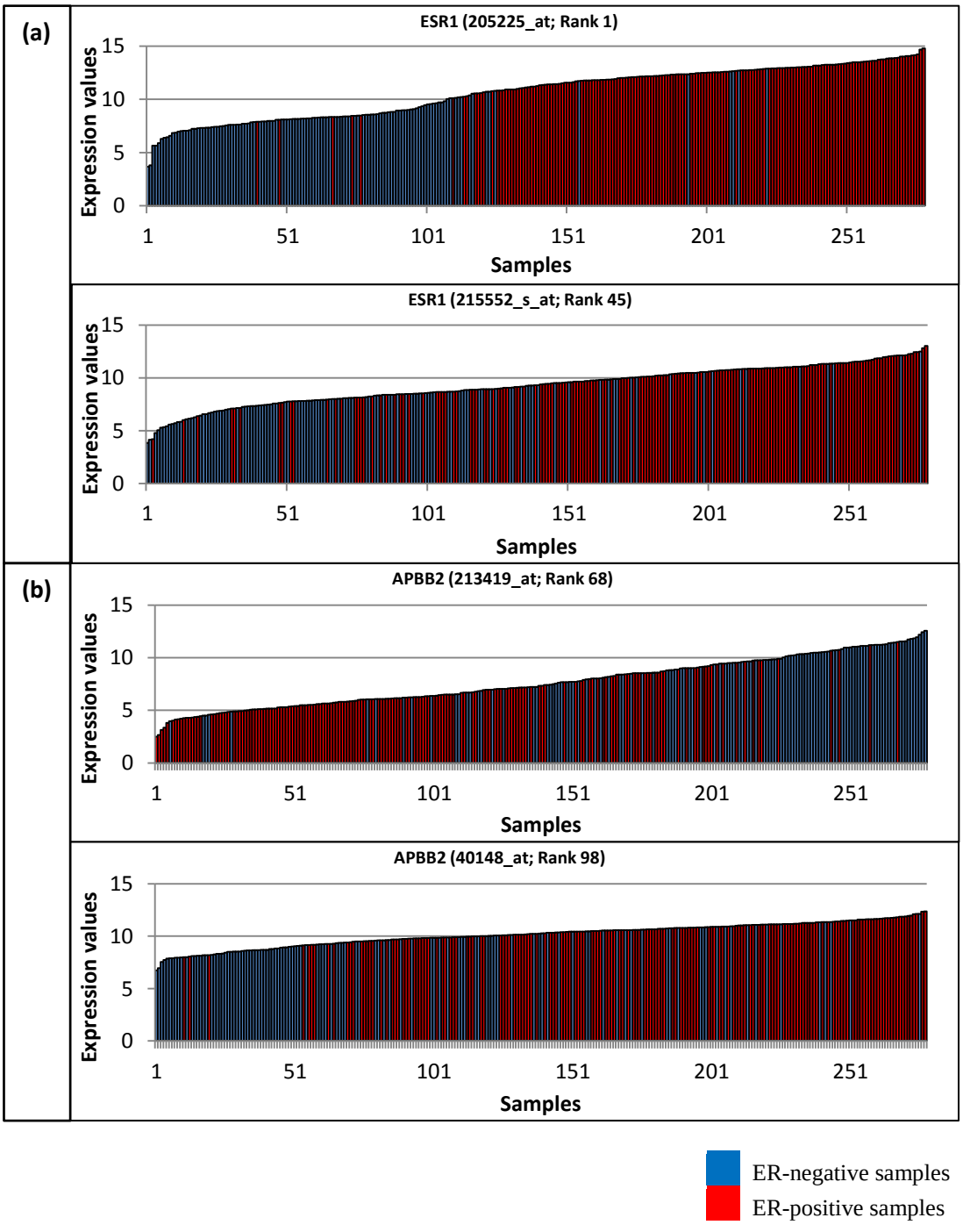


Figure 11: Population distribution of *ESR1* and *APBB2* probe variants. Schematic representation of population distribution of expression values of ESR1 probes across the total population of samples. **(a)** The probe variants of *ESR1* in ER-positive and ER-negative samples show a clear difference in the pattern of expression. **(b)** The probe variants of *APBB2* in ER-positive and ER-negative samples show a clear difference in the pattern of expression. Both the probes show an opposite expression profile among the samples.

and ER-negative samples. For example, the two probes of *APBB2* (Amyloid beta (A4) precursor protein-binding, family B) gene, 213419_at ranked 68th and 40148_at ranked at 98th positions, showed a clear contrasting expression pattern among the ER-positive and ER-negative samples. 213419_at had a higher expression profile in ER-positive samples, whereas, 40148_at was found to have an increased expression in ER-negative samples (Figure 11(b)).

3.4.7 Gene Ontology of the top 100 probes

An investigation of the involvement of ER-associated genes in disease formation and prognosis may provide information regarding the molecular or functional pathway of ER-associated genes. It may also identify biomarkers associated with ER and its functionality. To investigate the ontological information of the genes predictive for the ER class, the top 100 probes selected from the Stepwise-ANN model were submitted to an online open access gene-protein interaction database called *Panther* (Thomas et al., 2003). The advantage of using the Panther database for the mining of ontological information was that all these genes were found to be present and updated in the Panther database when accessed on 1st of November 2011.

3.4.7.1 Molecular function

The ontological information regarding the molecular functions associated with the ER-associated top 100 genes from Stepwise-ANN model were found to be

dominated by the catalytic activity (40.31%), DNA binding activity (27.69%), and transcriptional regulatory activity (13.31%). Other molecular functions shared by the ER-associated genes were enzyme regulatory function, and with receptor activity function. The molecular ontology is represented in Figure 12 and detailed information in Table 8.

3.4.7.2 Biological function

The biological functionality of the ER-associated genes was dominated by metabolic processes (19.76%), developmental process (17.37%), cellular process (11.38%) and cell communication (10.78%). Other functions such as cell cycle, immune response, system process, cell adhesion, cell component organisation, apoptosis, transport, stimulus response, reproduction, precursor and homeostatic functions were also found. The biological ontology is represented in Figure 13 and detailed information in Table 8.

3.4.7.3 Pathways

The top 100 genes found associated with the ER status by the Stepwise-ANN were also investigated for ontological pathways (Figure 14) and found to be involved in various pathways concerning with the cell cycle, immune responses in terms of B-cell and T-cell activation and regulation, nucleotide metabolism and synthesis, Vitamin-D metabolism and Want signalling pathways. Promisingly, the gene *IL6ST* was found to be involved in immune response and genes such as E-cad and P-cad were found to be highly involved in Catherin signalling pathway.

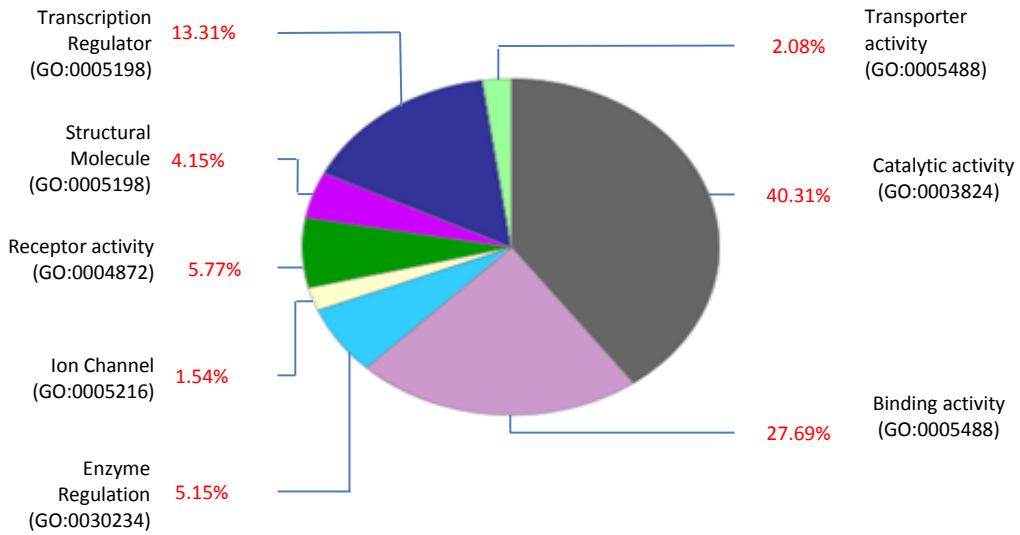


Figure 12: Molecular functional ontology information of ER-associated genes. Molecular functional ontology information of top 100 ER-associated genes from Stepwise-ANN model. The genes are grouped based on their molecular function along with the percentage of distribution.

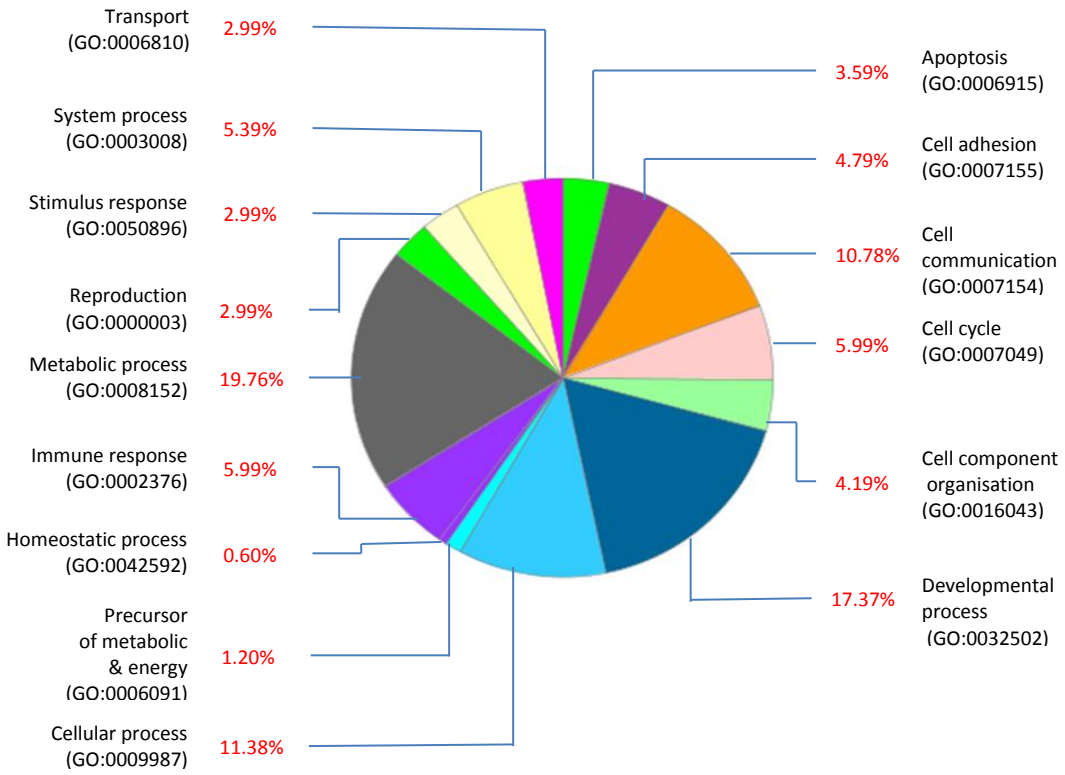


Figure 13: Biological functional ontology information of ER-associated genes. Biological functional ontology information of top 100 ER-associated genes from Stepwise-ANN model. The genes are grouped based on their biological function along with the percentage of distribution.

Table 8: Detailed ontological information of ER-associated genes from Stepwise-ANN model.

Ontology	Genes involved
MOLECULAR FUNCTIONS	
Binding	GATA3, TFF1, KIAA0876, RNB6, RARA, SEMA3F, DP1, SOX11, VAV3, IL6ST, CRNKL1, KIAA0882, ANXA9, MLPH, FOXA1, MAPT, TFF3, CIRBP
Catalytic	CYP2B6, GATA3, PTP4A2, AKR7A3, CRNKL1, MCCC2, KIAA0882, UGCG, CIRBP, CA12, ENPP1, FBP1, SKP1A, NME3, WWP1, ADCY9, AGR2, EGFR, ACADSB, CRIP1, NAT1
Enzyme regulation	VAV3, KIAA0882, MLPH, SERPINA5
Ion Channel	AKR7A3
Receptor Activity	RARA, IL6ST, DP1, CDH3, CELSR1, CDH3, EGFR, ESR1
Structural molecular Activity	RNB6, MAPT, DNALI1, CRIP1
Transcriptional Regulators	GATA3, KIAA0876, RARA, DP1, SOX11, CRNKL1, FOXA1
Transport Activity	DP1, NAT1
BIOLOGICAL FUNCTIONS	
Apoptosis	DP1, MAPT, MAGED2, MYB, EGFR, CRIP1
Cell cycle	DP1, JDP1, PTP4A2, FOXA1, CIRBP, MYB, SKP1A, PDEF, EGFR, ESR1
Cell Adhesion	SEMA3F, CDH3, CLSTN2, MAGED2, CELSR1, IGFBP4, CDH3, EGFR
Cell communication	RARA, SEMA3F, TFF1, IL6ST, VAV3, ANXA9, FOXA1, TFF3, CDH3, CLSTN2, CELSR1, IGFBP4, PDEF, CDH3, EGFR, ESR1, CRIP1, NAT1
Cell component organization	FOXA1, KIAA0882, MAPT, CDH3, CELSR1, CDH3, CRIP1
Cell process	GATA3, RARA, SEMA3F, CRNKL1, IL6ST, ANXA9, FOXA1, KIAA0882, CIRBP, MAPT, TFF3, CDH3, CELSR1, PDEF, WWP1, CDH3, EGFR, KIAA1243, CRIP1, DP1, JDP1, PTP4A2, RARA, RNB6, SEMA3F, TFF1, IL6ST, VAV3, ANXA9, FOXA1, KIAA0882, CIRBP, MAPT,TFF3, CDH3, CLSTN2, MAGED2, MYB, SKP1A, CELSR1, IGFBP4, PDEF, CDH3, DNALI1, EGFR, ESR1, CRIP1, NAT1
Homeostasis	AGR2
Immune response	JDP1, SEMA3F, TFF1, IL6ST, VAV3, TFF3, PDEF, EGFR, XBP1, CRIP1
Metabolic Process	CYP2B6, DP1, GATA3, JDP1, KIAA0876, PTP4A2, RARA, AKR7A3, CRNKL1, MCCC2, ANXA9, FOXA1, UGCG, CIRBP, DP1, CA12, ENPP1, FBP1, MYB, SERPINA5, SKP1A, NME3, PDEF, WWP1, ADCY9, AGR2, EGFR, XBP1, ESR1, ACADSB, CRIP1, NAT1
Precursor generation	CYP2B6, ACADSB
Reproduction	KIAA0876, CIRBP, MAGED2, EGFR, ESR1
Stimuli response	JDP1, IL6ST, VAV3, PDEF, CRIP1
System Process	VAV3, FOXA1, CIRBP, APBB2, CDH3, CDH3, ESR1, CRIP1, NAT1
Transport	ANXA9, KIAA0882, MLPH, DP1, NAT1
Unknown GO list	FLJ20151, AL050025, FLJ12910, H05812, AW242916, LIV-1, AL049265, C1orf34, TONDU, PD009, BG289001, MGC2601, CGI-49, AI424243, KIAA0040, PH-4, C6orf29, BF508639, AL049381, FLJ10647, BF967998, AI263909, AI658534, AI830698

The ontological information about the biological processes carried out on top 100 ER-associted genes. The genes were grouped according to their molecualr and biological functions.

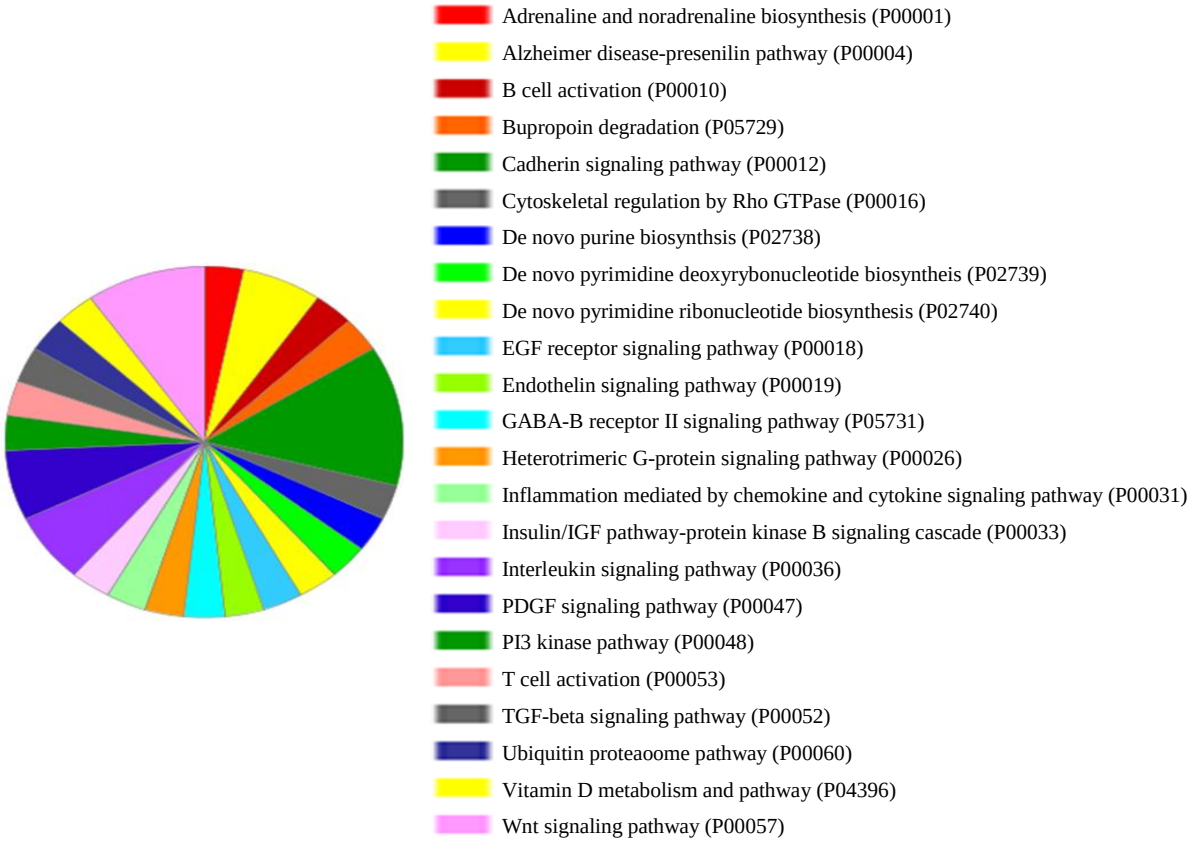


Figure 14: Pathway ontology information of ER-Associated genes. Pathway ontology of information of top 100 ER-associate genes from Stepwise-ANN model. The genes are grouped based on the pathway.

3.4.8 Cross validation with other breast cancer datasets

The findings from the Stepwise-ANN model of the E-GEOD-20194 dataset, were compared with three other breast cancer datasets namely E-GEOD-2034 (Wang et al., 2005), E-TABM-158 (Chin et al., 2006) and E-GEOD-7390 (Desmedt et al., 2007). To validate the findings, an investigation of predictive performance and rank position of predictive genes was compared between the breast cancer datasets.

To validate the predictive performance, the ER class prediction was performed using the Stepwise-ANN model by considering standardised settings (described earlier in Chapter 2. Section 2.7). The classification accuracy and ROC area were investigated for comparison. Among the datasets considered for cross validation, the classification accuracy of the rank one predictor for ER classification in E-GEOD-20194 dataset was found to be highest (90.97%) followed by E-TABM-158 (89.74%), E-GEOD-2034 (89.51%) and E-GEOD-7390 (87.30%). The investigation was extended to study the ROC area for validation subset of samples, and was found to be highest in E-GEOD-20194 (93.58%) followed by E-GEOD-2034 (92.11%), E-GEOD-7390 (88.08%) and E-TABM-158 (80.99%). The detailed findings are tabulated in Table 9 and Figure 15.

Table 9: Stepwise-ANN classification performance in breast cancer datasets.

Dataset	TP	TN	FP	FN	Sensitivity (%)	Specificity (%)	Accuracy (%)	ROC area (%)
E-GEOD-2034	190	66	11	19	90.90	85.71	89.51	92.11
E-TABM-158	73	32	11	1	98.64	74.41	89.74	80.99
E-GEOD-7390	124	48	15	10	92.53	76.19	87.30	88.08
E-GEOD-20194	157	95	19	6	96.31	83.33	90.97	93.58

The Stepwise-ANN classification performance in breast cancer datasets considered for cross validation. Sensitivity, specificity, classification accuracy and ROC area are represented in percentages. TP=True Positive, TN=True Negative, FP= False Positive and FN=False Negative.

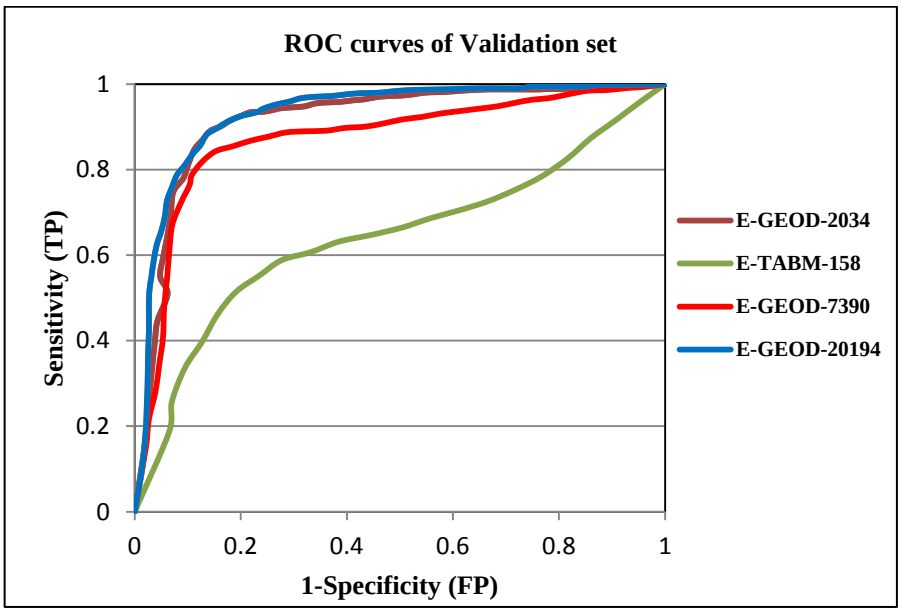


Figure 15: ROC Curve analysis of breast cancer datasets. Receiver Operative Characteristic (ROC) curve of validation subset of samples from breast cancer datasets. The true positivity (Sensitivity) was compared with False positivity (1=Specificity).

To investigate the consistency of the Stepwise-ANN in selecting efficient predictive markers across the breast cancer datasets, we considered the rank position of the predictive markers. The ranking of each probes was assigned with respect to the averaged error recorded in the test subset of samples. Then, the top 10 predictive probes were selected after cumulative ranking of each probe across the datasets used for cross validation. The rank position of the 10 highly predictive probes were checked across the datasets used. Table 10 is the tabulation of rank position of the 10 best predictive probes across breast cancer datasets.

Table 10: Rank position of top 10 probes across breast cancer datasets.

Gene Symbol	Probe ID	Rank order of genes in breast cancer datasets				p-value
		E-GEOD-2034	E-TABM-158	E-GEOD-7390	E-GEOD-20194	
ESR1	205225_at	4	1	1	1	1.63E-17
GATA3	209603_at	3	2	7	2	3.42E-16
GATA3	209604_s_at	1	4	9	8	1.17E-15
TBC1D9	212956_at	6	14	2	3	2.05E-15
CA12	214164_x_at	8	7	6	4	5.47E-15
CA12	215867_x_at	7	13	4	6	8.89E-15
GATA3	209602_s_at	2	10	22	5	8.95E-15
CA12	203963_at	17	5	19	10	6.57E-14
TBC1D9	212960_at	31	3	11	12	4.99E-14
DNAJC12	218976_at	34	12	16	16	4.25E-13

The rank position of top 10 probes across breast cancer datasets. The probes are selected based on the averaged predictive performances across datasets. p-value represents the statistical significance of the probe cumulated across all datasets studied.

The probe representing the *ESR1* gene (ERα) was found to be at the top rank in 3 out of 4 datasets studied. The rank position of probes did not differ much across the datasets since the top 10 probes were all under the rank order of 35. The rank position and p-value of the probe ranking also confirmed the consistency of the Stepwise-ANN findings across the datasets.

3.5 Conclusion

The Stepwise-ANN, an in-house developed algorithm, showed the ability to identify the most significant gene signature from the E-GEOD-20194 breast cancer microarray data. We applied random cross-validation on the data to obtain the unbiased top 100 genes from the dataset. We used a predictive ANN model coupled with back-propagation of errors algorithm to perform a thorough search on the entire gene population to ensure that the most significant gene subset was always selected by our model.

The Stepwise-ANN model was able to produce a consistent high performance with mean training, testing and validation subsets being 90.45%, 90.45% and 89.45% respectively in all the 10 independent iterations of the predictive models. Furthermore, the Stepwise-ANN model showed a high classification accuracy of 90.97%. The sensitivity and specificity analysis showed 96.31% and 83.33% respectively signifying the reliability of the findings. The consistency and reproducibility of the model was further demonstrated by the

listing of a single gene most associated with the ER-class at every independent iteration.

The gene *ESR1*, coding for ERα subunit was found to be directly associated with the ER-status in the breast cancer, validating the feature selection capability of the Stepwise-ANN. The *ESR1* gene which is highest ranked by the ANN shows that it is a strong predictor of ER status in breast cancer.

From this model, nine other contenders associated genes; *EGFR, ESR1, IL6ST* (X2), *MAPT, DACH, GATA3, GREB1, IL6ST* and *SERPINA5* were found to overlap with the original study of Liu et al (LiuQ 2009).

Interestingly, the Stepwise-ANN model was efficient enough to pull out the differences between the expression profiles of close variants of the same gene, some being similar in expression profile and some with contrasting expression profiles across the ER-positive and ER-negative samples.

The consistency of the Stepwise-ANN in selecting the most predictive marker for ER-class was investigated and proved in three different breast cancer datasets in terms of classification performance of the model and rank position of probes.

A part of this chapter is published in the International Conference on Machine Learning and Cybernetics (ICMLC), 2011 with an article entitled "ESTROGEN RECEPTOR STATUS PREDICTION FOR BREAST CANCER USING ARTIFICIAL NEURAL NETWORK" (Dhondalay et al., 2011) (*Annexure 1*). This peer reviewed paper of the bespoke methodology and findings

on the E-GEOD-20194 dataset, emphasised the implementation of the in-house developed Stepwise ANN method for ER classification on a larger cohort (>200) of samples.

Chapter 4

OESTROGEN RECEPTOR NETWORK INFERENCE WITH ARTIFICIAL NEURAL NETWORKS

(using pre-identified ER-associated markers by Stepwise-ANN approach)

Chapter abstract: The overall functionality of a cell at any given time is governed by the complex interactions among its molecules. A conventional approach using knocking out a molecule of interest to investigate the physiological and functional alterations in the biological behaviour of a disease is tiresome and time consuming. To circumvent this, computational network inferencing approaches have been found to be promising. In this study, we have attempted to investigate the interactions between ER-associated markers utilising an ANN based gene-gene network inferencing approach. The ANN-based network inference identified breast cancer associated markers such as *FOXA1, GATA3, CA12* and others with ER status specificity. Clinical validation of ER-associated marker, *VAV3* was found insignificant with clinicopathological markers, breast cancer specific markers and clinical outcome in patients. On the other hand, *DACH1* was found to be an independent prognostic marker for disease free survival and distance metastasis in clinical samples.

4.1 Introduction

One characteristic of a biological cell at a given time is the complex interaction of its various elements such as DNA, RNA, proteins and small molecules (Alberts et al., 2002). It is a major goal of functional genomics to explain the functional interactions between genomic and proteomic functional elements in a cellular pathway or network (Hieter and Boguski, 1997). The vital challenge in biomedicine is the need to establish the relationship between the physiological processes and the action of interacting contenders within a disease.

Traditionally the interaction of a single biological element is determined by knocking out the particular gene or molecule of interest and studying the physiological and functional changes arising that could be attributed to the knockout using a reductionist approach (Van Regenmortel, 2004). However, biology is more complex than this and biological networks consist of hundreds or thousands of potential interactions. To test all of these experimentally for even a single pathway would involve multiple knockout studies taking an enormous amount of time to achieve. To circumvent this, bioinformaticians have developed computational algorithms which predict and infer interactions and the relation between contender markers.

4.2 Network Inference

Network inference is a field of computational systems biology which predicts and infers biological network by employing algorithms wherein the

interactions among the contender of the system are logically connected to each other representing the behaviour of the system (D'Haeseleer et al., 1999). The simplest network in systems graph is a pairwise relation between **Nodes** (called vertices) through **Edges** (called links). The nodes of cellular system may be genes, mRNA, proteins and/or other molecules. Edges are comprised of links between a source node and a target node representing the relation between the nodes. Depending on the availability of information, edges can have positive or negative values (representing activation or inhibition), weights (quantifying confidence levels), strength, and/or reaction speeds.

The advent of high throughput post genomic technologies has enabled the cumulative study of thousands of biological molecules at a time. Network inference methods using high-throughput data depend on probing for patterns of partial correlation or conditional probabilities that indicate causal influence at any given time or across multiple times (Hayete et al., 2007).

There are generally five types of biological networks that are widely inferred by using computational algorithms. They are;

- ***Gene-gene interaction networks*** wherein the correlational association of expression between genes are investigated using experimental microarray data as inputs (Basso et al., 2005, Hartemink, 2005).
- ***Protein-protein interaction networks*** wherein the mere association of a protein with other proteins is exploited without consideration of the functional and

conformational changes brought about by the association (Vazquez et al., 2003).

- ***Transcriptional regulatory networks*** exploit the fact that DNA (gene) transcribes to RNA (mRNA) which is then translated to protein. Some studies have utilised mRNA expression levels from microarray experiments as inputs in the networks to predict the target gene (Gutierrez-Rios et al., 2003). For example, the chromatin immune precipitation (ChIP) method explores transcriptional factors governing other gene's expression (Wang et al., 2007).
- ***Signal transduction network*** wherein amino acid modifications such as phosphorylation, dephosphorylation, methylation or ubiquitylation (representing activation or inactivation status of a protein) are of prime importance (Steffen et al., 2002). The signalling networks are also complicated by the fact that post-transcriptional and post-translational modifications may alter the concentration of involved or resultant protein (Beck et al., 2006), requiring complex and refined statistical intervention.
- ***Metabolic interaction networks*** use measurement of the concentration of metabolites from experimental setups as inputs for the algorithms (Mahadevan and Palsson, 2005), wherein metabolites act as nodes and clubbing the information of enzyme kinetics and energy conversions has facilitated inference of complex interactions (Craciun et al., 2013).

Recently, a number of computational methods have been established for inferring gene regulatory networks using Bayesian networks (Beal et al., 2005, Husmeier, 2003), differential equations (Chen et al., 2005, de Jong, 2002) and

optimization techniques (Tsai and Wang, 2005), in several organisms such as *Saccharomyces cerevisiae* (Xu et al., 2004), and Drosophila (Sun et al., 2003). Moreover, efforts from bioinformatics has been directed to the development and use of various network inference algorithms focussing on the linearity of datasets (D'Haeseleer et al., 1999), with target-regulator pair methods to predict functional interactions (Arkin et al., 1997).

The cDNA microarray data with its added advantage of thousands of genes in single experimental setup serves as an ideal platform for investigating ER-associated genes. In the past, several investigators have employed various machine learning methods for prediction and have struggled to overcome the problem of non-linearity of biological microarray datasets. Gruvberger et al has utilised ANNs to predict ER-associated gene signature in breast cancer patients (Gruvberger et al., 2001). However, an investigation of contribution of each contenders for prediction and the association of genes with respect to each other and was not attempted, until Lemetre et al proposed a novel method for quantifying interactions between predictive genes using ANN (Lemetre et al., 2009).

4.3 Aims and Objectives

- To identify the importance of ER-associated genes in breast cancer disease, through an investigation of the contribution between ER-associated genes using ANN.

- To investigate the interactions in ER stratified (ER+ and ER-) samples in order to understand why ER dependent subgroup of breast cancer have varied response to treatment
- To visualise the constructed interaction network models.
- To corroborate the findings in published literature.
- To validate ER-associated markers in clinical breast tumour samples.

4.4 Methodology

4.4.1 Gene-Gene Interaction analysis

The pre-identified top 100 genes from the Stepwise-ANN method were applied to an ANN based gene-gene network inference algorithm developed by Lemetre et al. (Lemetre et al., 2009). This approach utilizes a feed forward back propagation algorithm to predict the expression level of a given probe from the other 99 probes in the set. The model is trained to optimum performance on unseen data using the Monte Carlo cross validation strategy. It is then parameterized to determine a weighting of the link from input probe to output probe. This link has both a sign and a magnitude. This process is repeated for the remaining 99 probes setting each probe as the output generating a total of 100 models.

The parameterizations of each of these models are then integrated to produce an interaction matrix. The 100 strongest interactions (representing the top

1% of interactions) were then selected for visualization. Using this multifactorial nonlinear ANN based inference approach has an advantage over simple linear regression based approaches in that the interaction may be nonlinear and that bi-directionality can be modelled.

4.4.2 Visualisation of network models

For the visualization of the gene interactions in this study, we have used a free standalone Cytoscape Ver 2.7.3 software (http://www.cytoscape.org), as mentioned in *Chapter 2, Section 2.8.6*. The genes are represented as nodes and the interaction intensities as edges, presented in either positive (green) or negative (red) with the magnitude of interaction represented by the width of the edge. The directionality is given from source to target and represented by arrows. The thickness of the edge was maintained proportional to the intensity of interactions by using a rescaling method of weights based on minimum and maximum values to squeeze the dimensionality of the edges to create more generalized models.

4.5.3 Clinical validation of ER-associated markers:

Tissue microarray sections of primary operable breast carcinoma patients from the Nottingham Tenovous Breast Cancer unit at the Nottingham NHS Hospital, collected between 1988 to 1998, were used to investigate the expression

of ER-associated markers by Immunohistochemistry. The patients were of age from 27 to 70 years with a median age being 55 years. The optimization of optimal antibody concentration and method was successfully carried out on tissue microarray of formalin fixed paraffin-embedded tissue sections.

4.5.3.1 Immunohistochemistry reagents:

Chemicals	**Supplier**
Absolute Ethanol	VWR
Citric acid	Sigma
DPX mountant	Sigma
Hydrochloric acid	Fischer Scientific
Magnesium chloride	Fischer Scientific
Sodium chloride	Fischer Scientific
Sodium hydroxide	Fischer Scientific
Tris anhydrate	Sigma
Tween 20	Sigma
Xylene	Acros organics

IHC reagents	**Supplier**	**Catalogue No.**
Anti-VAV3 antibody; polyclonal, raised in goat	Abcam	ab21208
Anti-DACH1 antibody; polyclonal, raised in rabbit	Sigma	HPA012672
Secondary anti-goat antibody, biotintylated	DAKO	
Secondary anti-rabbit antibody, biotintylated	DAKO	
IHC staining kit	DAKO	REAL™

Preparations:

- **10mM Sodium citrate buffer [pH 6.0]:** Dissolve 2.1g of Citric acid in about 900ml of distilled water. Add 25ml of 1M Sodium hydroxide. Make up the volume to 1000ml with distilled water simultaneously adjusting the pH to 6.0 with 1M Sodium chloride or 1M Hydrochloric acid.

- **Tris Buffered Saline (TBS) [pH 7.6]:** Dissolve 21g of Tris and 6g of Sodium chloride in about 900ml of distilled water. Make up the volume to 1000ml with distilled water simultaneously adjusting the pH to 7.6 with 1M Sodium chloride or 1M Hydrochloric acid.

- **1% acid alcohol solution:** Add 70% Ethanol (70ml absolute Ethanol in 30ml of distilled water) and 1ml of concentrated Hydrochloric acid.

- **Blueing solution:** Dissolve a pinch of magnesium chloride crystals in distilled water.

4.5.3.2 Immunohistochemistry protocol:

- Melt the sections on a hot plate or in a hot air oven at 58-60^{0}C for 10 minutes.

Dewaxing:

- Dewax the sections in Xylene 2 times for 10 minutes each.
- Pass through 3 changes of absolute Ethanol for 1 minute each.

Rehydration:

- Wash the slides in tap water for 1 minute.

Antigen Retrieval:

- Load the sections into a plastic tray and place in a plastic sandwich box holding 850 ml of 10mM Sodium citrate (pH 6.0) buffer.

- Place the box into the microwave and place the lid on the box so that there is a gap for steam to escape.
- Turn on the microwave and set it to 80% for 23 minutes. Press START.
- Periodically check the level of the citrate buffer so that the buffer level should not fall below the level of the slides.
- When the time has elapsed, remove the box from the microwave.
- Carefully lower the slide rack into a bowl of cold running water, in the sink.
- Leave for 20 minutes.
- Meanwhile, wash the container and clean the microwave for future use.

Washing:

- Incubate the sections in 0.1%Tween 20 in TBS at room temperature (RT) for 20 minutes.
- Incubate the sections in TBS at RT for 3 minutes.
- Drain and wipe round sections.

Blocking:

- Carefully layer the sections with Peroxidase Block (H_2O_2 Blocking solution) [1:50] and incubate for 5 minutes at RT.
- Drain the sections and wash in TBS for 5 minutes at RT with agitation.
- Drain and wipe round the sections.

Primary Antibody:

- Carefully load the sections with Primary antibodies and incubate for 30 minutes at RT
- Drain the sections and wash in TBS for 5 minutes at RT with agitation.

- Drain and wipe round the sections.

Secondary Antibody:

- Carefully load the sections with Secondary antibodies and incubate for 30 minutes at RT.
- Drain the sections and wash in TBS for 5 minutes at RT with agitation.
- Drain and wipe round the sections.

Peroxidase Conjugation:

- Carefully load the sections with Streptavidin (Peroxidase) conjugate reagent and incubate for 30 minutes at RT.
- Drain the sections and wash in TBS for 5 minutes at RT with agitation.
- Drain and wipe round the sections.

Chromogen Addition:

- Carefully load the sections with freshly prepared DAB Chromogen reagent and incubate for 10 minutes at RT.
- Drain the sections and wash in tap water for 5 minutes at RT.

Counter Staining:

- Counterstain the slides with Haematoxylin by immersing the rack slides for 5 minutes.
- Wash well in running tap water.

Acid Alcohol Treatment and Blueing:

- Dip in 1% acid alcohol for 15-30 seconds.
- Wash well in running tap water.
- Dip in Blueing Solution for 5-10 seconds.

- Wash well in running tap water.

Dehydration:

- Pass the slides through 3 changes of Ab. Ethanol for 2 minutes each.
- Pass the slides through 2 changes of Xylene for 2 minutes each.

Mounting:

- Mount the slide with DPX mountant with a coverslip.
- After proper drying, visualise under the light microscope.

4.5.3.3 Univariate and Multivariate statistical analysis

The protein expression of ER-associated markers and their association with other categorical biomarkers was studied by Chi square test to produce contingency tables by using SPSS 16.0 statistical software (SPSS Inc, Chicago, IL, USA) Normal cytoplasmic staining with presence or absence of tumour associated membranous staining was categorically scored as positive or negative irrespective of staining intensity. A Kaplan Meier plot with a log rank test was used to assess disease free survival, breast cancer specific survival and time for metastasis formation. The number of months from diagnosis to invasive local recurrence, local lymph node or distant relapse was defined as the disease free survival. Patients who died from reasons other than breast cancer were censored during the survival analysis. A Mantle Cox log rank test compares survival rates. Independent prognostic effect of variables was evaluated by multivariate Cox proportional hazard regression with 95% confidence interval. A p-value of <0.05 was considered to be significant.

4.5 Results

4.5.1 ER interactomes

In the gene-gene interaction model developed by considering the top 100 genes associated with ER-status in a total (278) population of samples is depicted in the Figure 16. For simpler visualization of the complex network model generated, the model with only top 100 interactions is shown here. The nodes are represented as light blue or pink according to their up-regulated or down-regulated status in the microarray dataset considered. The edges were scaled according to their minimum and maximum values to give a more generalized pattern of the interactions and to reduce the bias.

The genes with multiple connecting edges (of weights >5) with other genes are termed as **Hubs** in this study. The top 100 gene-gene interaction model generated showed *FOXA1* and *GATA3* as being the positive hubs with multiple edges joining with other genes, whereas *CDH3, TONDU, EGFR* and *SOX11* appeared to be negative hubs. It was interesting to note that some edges had bidirectional links with other genes representing multiple associations with the gene or gene variants. It was also interesting to note that some edges had multiple regulations suggesting the involvement of genes in reciprocal regulation or feed-back regulation with each other. Some genes were even found to be having self-regulatory effects such as *DACH1, MAPT* and *GATA3*. This phenomenon may be because genes such as *GATA3* act as transcriptional regulators for other genes and so are self- regulatory.

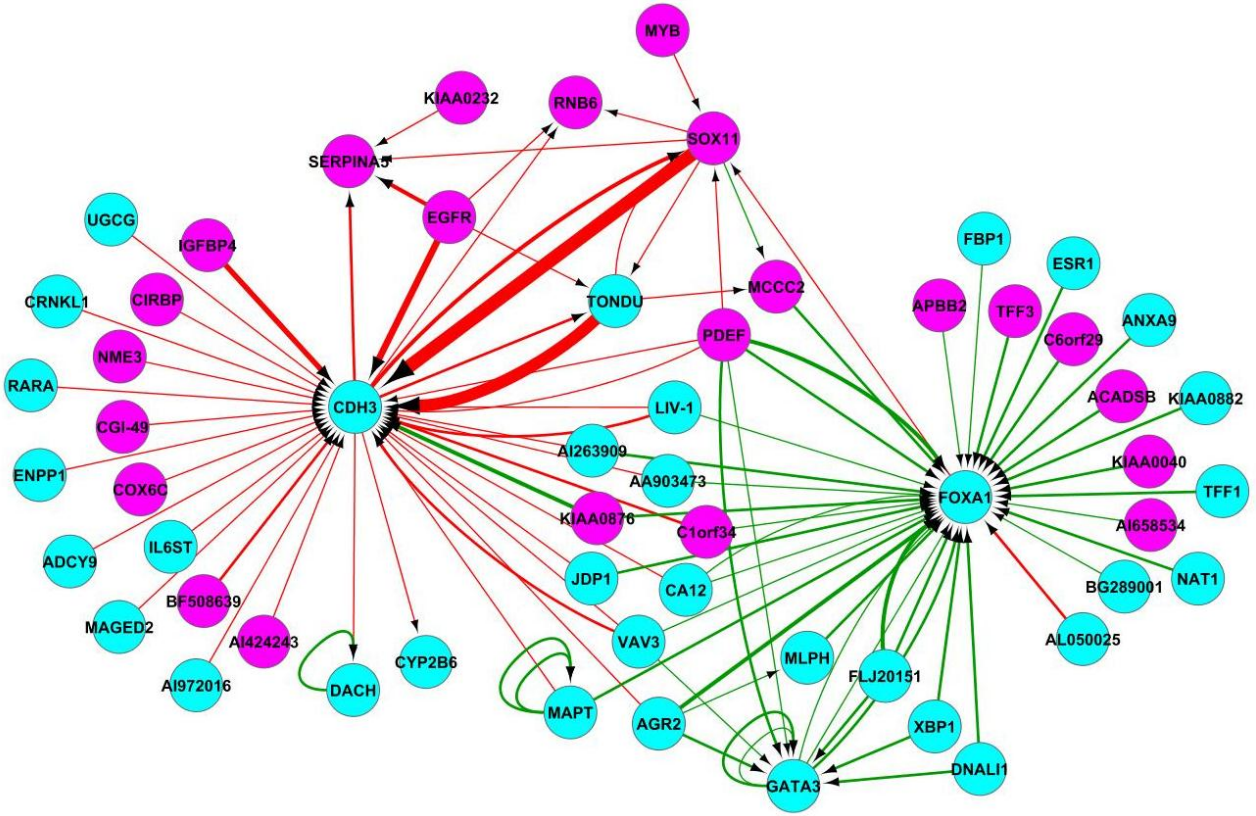


Figure 16: Positive and negative interactions. Interaction model of top 100 positive and negative interactions in breast cancer samples. The genes are represented as nodes and the interaction intensities as thickness of edges with directionality from source to target gene. The light blue and pink nodes represents up-regulation and down-regulation respectively. The green and red edges represents positive and negative interactions respectively.

4.5.2 ER-Positive and ER-Negative interactomes

In the later part of the study, we were interested in looking at the gene-gene interaction models when the samples were stratified based on their ER-status. Hence, a gene-gene interaction analysis was carried out on a separate cohort of ER-Positive (114 samples) and ER-Negative samples (164 samples).

When the network inference was undertaken on the different ER status groups, the network interaction model (top 100 interactions) showed some interesting results. In the interaction model from ER-Positive samples (Figure 17), the genes *CDH3, SOX11, EGFR* were found to be highly negatively regulated by other genes. The highest negative regulation was noticed between *CDH3* and *SOX11*. *DACH, SERPINA5, ANXA9* and *TFF3* formed the major positively regulated hubs. The highest positive interaction was found to be self-regulation of *DACH*. *RARA* was found to be acting as an ambiguous hub with both positive and negative interactions.

In the gene-gene interaction network model from ER-Negative samples (Figure 18), *VAV3, GAMT, ANXA9* appeared as positive hub. The negative hubs were again found to be *CDH3, SOX11* and *TONDU*. This demonstrate the consistency of negatively regulated genes among ER-Positive and ER-Negative samples.

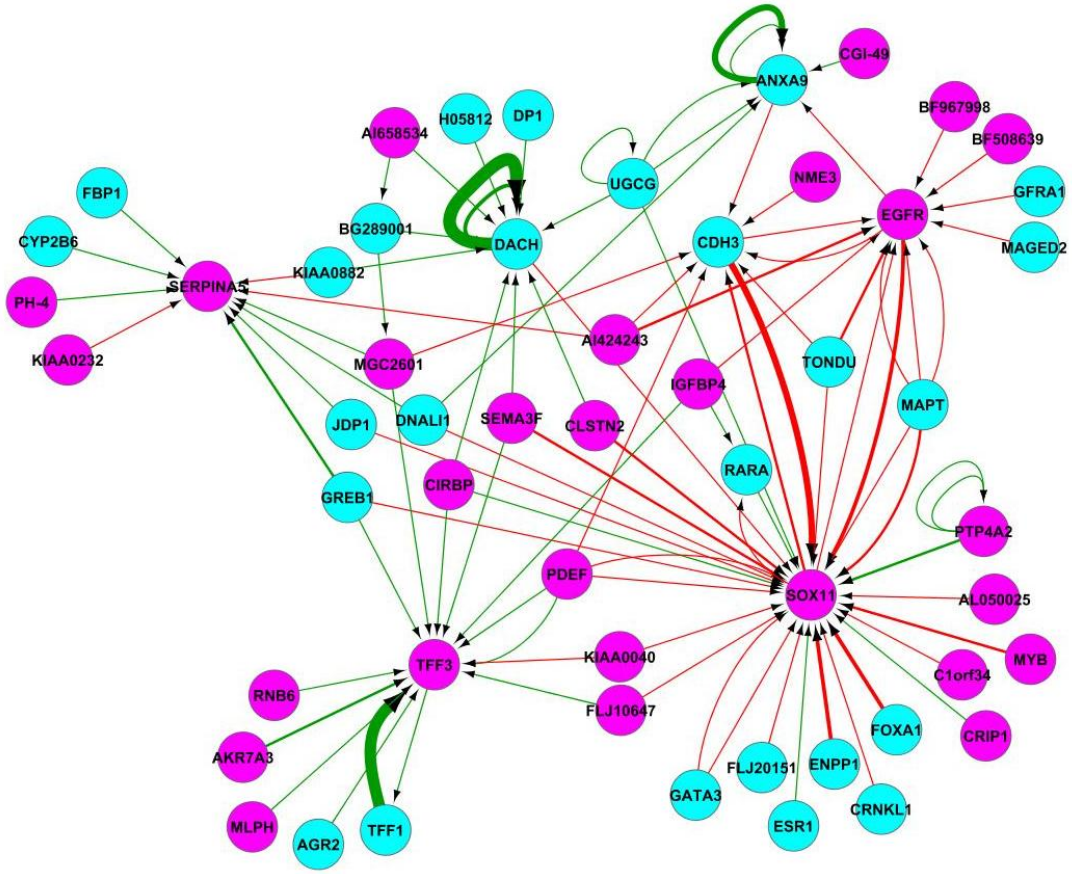


Figure 17: Interactions in ER-Positive samples. Interaction model of top 100 positive and negative interactions in ER-Positive breast cancer samples. The genes are represented as nodes and the interaction intensities as thickness of edges with directionality from source to target gene. The light blue and pink nodes represents up-regulation and down-regulation respectively. The green and red edges represents positive and negative interactions respectively.

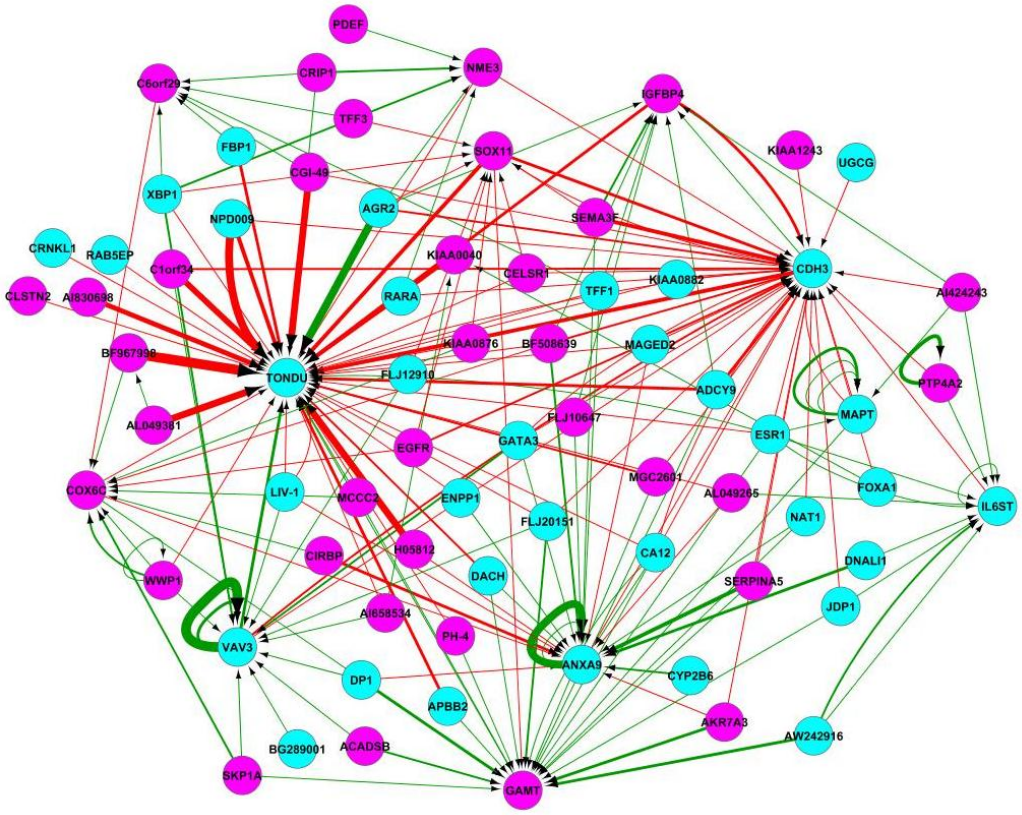


Figure 18: Interactions in ER-Negative samples. Interaction model of top 100 positive and negative interactions in ER-Negative breast cancer samples. The genes are represented as nodes and the interaction intensities as thickness of edges with directionality from source to target gene. The light blue and pink nodes represents up-regulation and down-regulation respectively. The green and red edges represents positive and negative interactions respectively.

The potential markers having multiple interactions from/with other genes act as hubs in these gene-gene interaction models. Further stratification of the samples based on the ER status and an investigation for markers acting as interacting hubs, identified a total of 17 potential hubs by considering a top one hundred interactions under each category of the samples. A list of 17 potential markers for the ER status group samples is tabulated in Table 11. The above results are published in IEEE-EMBS International Conference on Biomedical and Health informatics (Dhondalay et al., 2012) (Annexure 2).

Table 11: ER-associated gene-gene interaction hubs.

	ER (278 samples)	**ER-Positive samples (114 samples)**	**ER-Negative samples (164 samples)**
Positive Interactions	FOXA1	ANXA9	ANXA9
	GATA3	DACH1	C6orf29
		MGC2601	COX6C
		RARA	GAMT
		SERPINA5	IGFBP4
		TFF3	IL6ST
			NME3
			VAV3
Negative Interactions	CHD3	CHD3	CHD3
	SOX11	SOX11	SOX11
	TONDU	EGFR	TONDU

Summary of interactions acting as hubs in ER total sample population and ER status stratified samples. Hubs where found to be with multiple links converging or diverging at particular gene when taken as nodes in gene-gene interaction models. Red representing negative, Green representing positive and yellow represents ambiguous interaction hubs.

4.5.3 Clinical validation of ER-associated markers

Patient Selection and Tissue Microarray preparation:

Primary operable breast carcinoma sample, aged 70 or less, were selected from patients presenting at the Nottingham Breast Cancer Unit from 1988 to 1998 for Tissue MicroArray (TMA) preparation from Nottingham Tenovous series with additional biomarker data (Habashy et al., 2010).

For comparison of observed heterogeneity for VAV3 (N=739) and DACH1 (N=993) expression, paraffin processed TMA blocks were constructed from tissue core sampling from periphery avoiding areas of necrosis. A well characterised series was selected with information of patients' clinical and pathological data of age, tumour size, histologic type, mitotic index, lymph node status and histological grade, NPI and vascular invasion. Information regarding the tissue protein expression for relevant tumour biomarkers, pathological markers and follow up data about local, regional and distant metastasis with survival outcome was also present. Maintained on a prospective basis was information regarding disease free survival (DFS), distant metastasis (DM) and breast cancer specific survival (BCSS).

4.5.3.1 VAV3 protein expression:

Sections stained for anti-VAV3 primary antibody (ab21208, Abcam), raised in goat and anti-goat biotinylated secondary antibody (DAKO), were used for

immunohistochemistry. The negative control section was stained with an inappropriate antibody for CD68 of the same immunoglobulin class. Figure 19 shows the variation of VAV3 intensities classified as negative, weak, moderate and strong staining.

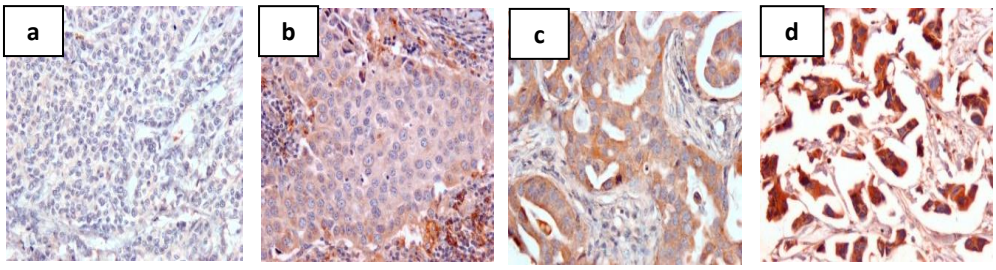


Figure 19: VAV3 protein expression by IHC. Immunohistochemical staining of Grade III invasive ductal carcinomas showing cytoplasmic staining for VAV3 expression.. VAV3 staining intensities as (a) negative, (b) weak (c) moderate and (d) strong staining pattern (x200).

4.5.3.1 Association with clinicopathological markers

Clinicopathological markers such as patient age, tumour size, tumour stage, tumour grade, tumour type, NPI, distance metastasis, Metastatic recurrence, vascular invasion, cellular proliferation as mitosis and menopausal status were investigated for association with VAV3 by using a Chi-square test. As a result, in the whole series of breast cancer patients compared for clinical parameters with VAV3, no significant association was found corresponding to any factor for the categories of negative and positive for VAV3 staining. Table 12 is the summary of the Chi-square test of VAV3 protein expression compared with clinicopathological markers of patients.

Table 12: VAV3 expression association with clinicopathological markers.

Marker	VAV3 Negative		VAV3 Positive		Chi-Square (χ^2)	p-value
	No. of cases	%	No. of cases	%		
Age					3.686	0.946
<40	51	6.9	51	6.9		
40 - 50	210	28.5	211	28.7		
51-60	251	34.1	250	34.0		
>60	251	30.40	224 (	30.4		
Tumour Size					1.604	0.206
<1.5cm	366	49.7	366	49.7		
>1.5 but < 2cm	370	50.3	370	50.3		
Tumour Stage					0.567	0.666
1	461	62.6	462	62.4		
2	208	28.2	209	28.4		
3	68	9.2	68	9.2		
Tumour Grade					0.195	0.827
1	115	15.6	115	15.6		
2	242	32.9	242	32.9		
3	379	51.5	379	51.5		
Nottingham Prognostic Index					0.16	0.763
Poor	131	17.8	131	17.8		
Moderate	393	53.4	394	53.5		
Good	212	28.8	211	28.7		
Distant Metastasis					0.06	0.898
No metastasis	508	69.9	508	69.9		
Metastasis present	219	30.1	219	30.1		
Recurrence					0.15	0.901
No recurrence	423	59.0	423	59.0		
Recurrence present	294	41.0	294	41.0		
Vascular Invasion					0.932	0.337
Negative	399	54.9	398	54.7		
Probable	103	14.2	104	14.3		
Definite	225	30.9	225	30.9		
Tumour Type					9.527	0.464
Ductal - NST	429	59.3	429	59.3		
Lobular	66	9.1	66	9.1		
Tubular and tubular mixed	144	19.9	144	19.9		
Medullary	24	3.3	24	3.3		
Other special types (a)	11	1.5	11	1.5		
Mixed NST and Lobular (b)	44	6.1	44	6.1		
DCIS	3	0.4	3	0.4		
MISCA	2	0.3	2	0.3		
Mitosis					0.466	0.667
1	23	32.9	231	32.9		
2	149	21.2	148	21.1		
3	322	45.9	323	46.0		
Menopause					0.38	0.846
Premenopause	270	36.7	270	36.7		
Postmenopause	466	63.3	466	63.3		

VAV3 expression association with clinicopathological markers of tumour tissue. Sample distribution Multivariate Chi-square test along with p-values are indicated.

4.5.3.1.1 Association with breast cancer markers

The association of VAV3 protein expression with breast cancer related markers was investigated. A detailed list of breast cancer markers and Chi-square with statistical significance are tabulated in Table 13. VAV3 showed no significant association with hormonal makers such as ER, PgR, AR, Her2, Her3 and Her4; or immunological markers such as CK5/6, CK7/8, and CD71. VAV3 was also not significantly related to other breast cancer related markers such as ALCAM, CA9, cyclin D1, DGP4B5, ERBB2, HERCEPT, muc1, p16, p27, TK1, VEGF, FOXA1, p53, REGR, TFF1 and EGFR. Interestingly, VAV3 was found to be significantly associated with E-cadherin, PELP1, FHIT and BUC11 with p-values of 0.01, 0.007, 0.003 and 0.000 respectively.

Table 13: VAV3 expression association with breast cancer markers.

Markers	VAV3 Negative		VAV3 Positive		Chi-Square (χ^2)	p-value
	No. of cases	%	No. of cases	%		
ALCAM					0.345	0.564
Negative	14	77.8	4	22.2		
Positive	36	70.6	15	29.4		
AR					0.084	0.772
Negative	170	70.8	70	29.2		
Positive	286	69.8	124	30.2		
BUC11 Cyt					17.08	0.000*
0	14	46.7	16	53.3		
1	49	58.3	35	41.7		
2	87	75.0	29	25.0		
3	44	81.5	10	18.5		
c-myc					0.913	0.341
Negative	39	73.6	14	26.4		
Positive	126	66.7	63	33.3		
CA 9					3.117	0.078
Negative	158	70.5	66	29.5		
Positive	17	89.5	2	10.5		
CD 71					0.154	0.696
Negative	55	67.1	27	32.9		
Positive	132	69.5	58	30.5		
CK 5/6					0.32	0.572
Negative	392	68.8	178	31.2		
Positive	172	71.7	68	28.3		
CK 7/8					0	0.984
Negative	7	70.0	3	30.0		
Positive	483	69.7	210	30.3		
Cyclin D1					0.896	0.347
Low	72	68.6	33	31.4		
Moderate	43	72.9	16	27.1		
High	26	76.5	8	23.5		
DGP4B5					2.2982	0.872
Negative	215	69.1	96	30.9		
Positive	14	63.6	8	36.4		
Mod. Positive	9	52.9	8	47.1		
Str. Positive	35	74.5	12	25.5		
E-cadherin					6.679	0.010*
Negative	161	63.6	92	36.4		
Positive	317	73.0	117	27.0		
EGFR					0.652	0.42
Negative	340	69.8	147	30.2		
Positive	89	73.6	32	26.4		
ER					0.034	0.853
Negative	147	69.3	65	30.7		
Positive	332	70.0	142	30.0		
ERBB2					0.082	0.775
Negative	301	69.4	133	30.6		
Positive	176	70.4	74	29.6		

(Continued)

(Table 13 continued)

Markers	VAV3 Negative		VAV3 Positive		Chi-Square (χ^2)	p-value
	No. of cases	%	No. of cases	%		
FHIT					9.011	0.003*
Negative	67	58.8	47	41.2		
Positive	365	73.0	135	27.0		
FOXA1					3.526	0.071
Negative	168	64.9	91	35.1		
Positive	173	72.4	66	27.6		
HER2					1.188	0.590
1	224	70.0	96	30.0		
2	13	59.1	9	40.9		
3	36	67.9	17	32.1		
HER3					1.797	0.181
Negative	32	62.7	19	37.3		
Positive	382	71.7	151	28.3		
HER4					0.049	0.826
Negative	78	69.0	35	31.0		
Positive	335	70.1	143	29.9		
HERCPT					0.527	0.469
Negative	410	69.3	182	30.7		
Positive	75	72.8	28	27.2		
MUC 1					1.792	0.182
Negative	20	83.3	4	16.7		
Positive	225	70.5	94	29.5		
P 16					1.034	0.311
Negative	72	63.2	42	36.8		
Positive	103	69.1	46	30.9		
P 27_50					0.909	0.342
Negative	50	74.6	17	25.4		
Positive	130	68.4	60	31.6		
P53					0.035	0.852
Negative	326	70.1	139	29.9		
Positive	152	69.4	67	30.6		
PELP1					8.222	0.007*
Low	51	60.7	33	39.3		
Moderate	226	66.3	115	33.7		
Strong	60	81.1	14	18.9		
PgR					0.029	0.866
Negative	210	69.1	94	30.9		
Positive	262	69.7	114	30.3		
RERG					0.345	0.558
Negative	237	66.9	117	33.1		
Positive	101	69.7	44	30.3		
TFF1					0.001	0.981
Negative	155	69.2	69	30.8		
Positive	152	69.1	68	30.9		
TK1					3.432	0.064
Negative	122	62.9	72	37.1		
Positive	153	71.5	61	28.5		
VEGF					3.011	0.084
Negative	12	57.1	9	42.9		
Positive	58	76.3	18	23.7		

VAV3 protein expression association with breast cancer markers. Significant associations are labelled with *.

4.5.3.1.2 Association with clinical outcome of patients

A multivariate Mantle-Cox regression analysis with minimal and maximal cut–off for 95% of confidence interval revealed that VAV3 was significantly associated with tumour stage and tumour grade for breast cancer specific survival, recurrence survival time and distance metastasis in breast cancer patient outcomes when the VAV3 expression absence against weak, moderate and strong staining intensities were considered. Table 14 summarizes the hazard ratio of VAV3 for a patient's outcome prediction.

Table 14: Mantle-Cox regression analysis of VAV3 for predicting patient outcome.

	Variable	p-value	Hazard Ratio	Lower 95 % CI	Upper 95% CI
Breast cancer specific survival	Tumour size	0.230	1.258	0.865	1.83
	Tumour stage	0.000*	1.894	1.408	2.547
	Tumour grade	0.000*	1.983	1.408	2.547
Recurrence survival time	Tumour size	0.150	1.23	0.928	1.631
	Tumour stage	0.000*	1.734	1.334	2.254
	Tumour grade	0.010*	1.398	1.083	1.805
Distance metastasis	Tumour size	0.198	1.242	0.893	1.728
	Tumour stage	0.000*	2.079	1.558	2.774
	Tumour grade	0.001*	1.675	1.234	2.275

Prediction of patient outcome using VAV3 expression by Mantle-Cox regression analysis. CI = Confidence Interval. The significant associations are indicated with *.

4.5.3.2 DACH1 protein expression

Four micron thick formalin fixed paraffin-processed TMA and full face sections were subjected to microwave antigen retrieval in citrate buffer (pH 6.0), and then immunohistochemically stained with a goat polyclonal antibody against DACH1 (Sigma HPA012672) using a streptavidin biotin technique (Dako, Cambridge, UK). The DACH1 antibody was optimised at a working dilution of 1:250. Sections were counterstained in haematoxylin and mounted using DPX mounting medium. Negative controls comprising omission of the primary antibody or substitution with an inappropriate primary antibody of similar immunoglobulin class was used.

The immunohistochemically stained TMA sections were scored by observers blinded to the clinicopathological features of tumours and patients' outcome. Nuclear staining intensity and the percentage of cells stained was assessed in unequivocal malignant epithelium using the H-score (histochemical score)(McCarty et al., 1985). Staining intensity was scored from 0, 1, 2 or 3 and the percentage of positive cells at each intensity subjectively estimated to produce a final score in the range 0–300 as depicted in Figure 20. DACH-1 H-score data was categorised using an H-score ≥200 to define positivity based on the histogram of frequencies. Damaged tissue cores and those that did not contain invasive carcinoma were censored.

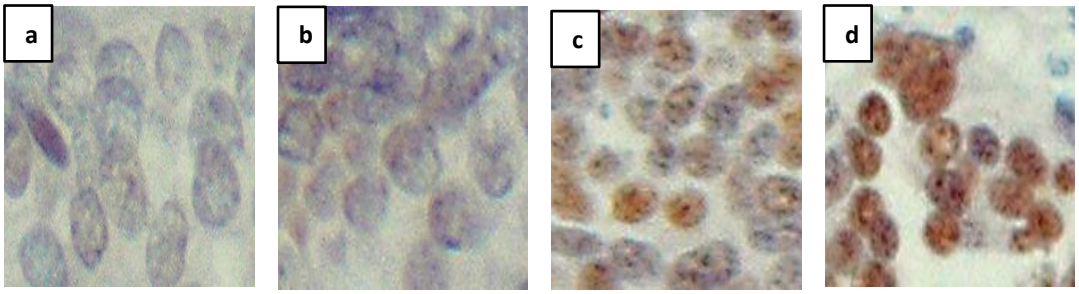


Figure 20: DACH1 protein expression by IHC. Immunohistochemical staining of Grade III invasive ductal carcinomas showing nuclear staining for DACH1 expression. DACH1 staining intensities as (a) negative as 0 (b) weak as 1 (c) moderate as 2 and (d) strong staining pattern as 3 for H-score.

4.5.3.2 Association with clinicopathological markers

To test the clinical relevance in breast cancer, the association of the DACH1 protein with clinicopathology features was investigated in a well characterised patient cohort. The median age of the patients was 55 years (range 27-70). Nuclear DACH1 protein expression was significantly increased in post-menopausal patients with lobular and tubular cancer types but in contrast, was rarely seen in patients with medullary cancer. DACH1 expression showed no significant association with tumour size, tumour stage, development of metastasis, tumour recurrence, or vascular invasion. DACH1 expression was significantly increased in tumours of low grade, good NPI and candidacy for hormonal therapy (Table 15).

Table 15: Association of DACH1 protein expression with clinicopathological markers.

Clinical Parameter	DACH1 absent		DACH1 present		Chi-square (χ^2)	p-value
	N	%	N	%		
Age group					12.505	0.006*
<40	40	10.31	35	5.79		
40-50	128	32.99	176	29.09		
51-60	124	31.96	197	32.56		
>60	96	24.74	197	32.56		
Menopause					8.912	0.003*
Premenopausal	174	44.85	214	35.37		
Postmenopausal	214	55.15	391	64.63		
Tumour Size					2.283	0.131
≤1.5 cm	178	46.23	307	51.17		
>1.5 cm	207	53.77	293	48.83		
Tumour Stage					0.413	0.813
1	241	62.27	362	60.23		
2	112	28.94	183	30.45		
3	34	8.79	56	9.32		
Tumour Grade					69.335	<0.001*
1	35	9.09	134	22.33		
2	94	24.42	226	37.67		
3	256	66.49	240	40.00		
Nottingham Prognostic Index					22.571	<0.001*
Good	75	19.48	200	33.28		
Moderate	233	60.52	309	51.41		
Poor	77	20.00	92	15.31		
Tumour type					57.194	<0.001*
Ductal - Non Specific Type (NST)	260	68.60	314	53.04		
Lobular (Classical and variants)	28	7.39	85	14.36		
Tubular & Tubular mixed	50	13.19	136	22.97		
Medullary	20	5.28	3	0.51		
Special type (Mucinous, Cribriform and Invasive papillary)	4	1.06	14	4.36		
Mixed NST with Lobular and special types	17	4.49	40	6.76		
Distant metastasis formation					0.349	0.555
Absent	268	69.43	425	71.19		
Present	118	30.57	172	28.81		
Tumour recurrence					0.078	0.780
Absent	231	60.63	353	59.73		
Present	150	39.37	238	40.27		
Vascular invasion					5.345	0.069
Negative	222	57.81	325	54.53		
Probable	33	8.59	80	13.42		
Definite	129	33.59	191	32.05		
Endocrine Therapy					9.085	0.003*
Untreated	261	71.12	331	61.41		
Treated	106	28.88	208	38.59		

Association of DACH1 protein expression with clinicopathological markers of breast cancer patients. Statistical significant associations are indicated with *.

4.5.3.2.1 Association with breast cancer markers

Nuclear DACH1 expression was highly significantly increased in patients with ER-alpha positive tumours expressing PgR, and epithelial CK18/19 cytokeratins. Significant positive associations were seen with 'luminal-like' markers of good prognosis including FOXA1 and RERG. In contrast, highly significant inverse associations were found with luminal markers indicating poor prognosis including CD71 and TK1 (Table 16 showing association with other biomarkers).

Supporting its association with good prognosis, tumour's DACH1 expression significantly correlated with low cell proliferation (MIB1) and the functional apoptosis marker Bcl2. Low DACH1 frequency and expression was seen in tumours bearing markers of poor prognosis including the basal-like markers CK14/5/6 and EGFR, as well as the HER2 and p53 positivity.

4.5.3.2.2 Association with clinical outcome of patients

Nuclear DACH1 positivity showed a significant association in predicting cancer specific survival ($\chi2=11.96$, $p<0.001$), disease free interval ($\chi2=15.33$, $p<0.001$), tumour recurrence ($\chi2=16.49$, $p<0.001$) and distant metastasis ($\chi2=16.31$, $p<0.001$) over 5 years post diagnosis (Figure 21). However, the level of significance lessened for predicting cancer specific survival ($\chi2=2.31$, $p=0.13$),

Table 16: Association of DACH1 protein expression with breast cancer markers.

Markers	DACH1 absent		DACH1 present		Chi-square (χ^2)	p-value
	N	%	N	%		
ER					142.867	<0.001*
Absent	181	49.45	78	13.66		
Present	185	50.55	493	86.34		
PgR					55.671	<0.001*
Absent	212	58.56	191	33.69		
Present	150	41.44	376	66.31		
CK18					54.282	<0.001*
Absent	86	24.86	39	7.21		
Present	260	75.14	502	92.79		
CK19					5.786	0.016*
Absent	50	13.51	50	8.61		
Present	320	86.49	531	91.39		
HER2					6.595	0.010*
Absent	311	83.38	524	89.12		
Present	62	16.62	64	10.88		
E-cadherin					0.853	0.356
Absent	145	40.06	213	37.04		
Present	217	59.94	362	62.96		
EGFR					6.371	0.012*
Absent	249	76.62	425	83.66		
Present	76	23.38	83	16.34		
CK5/6					66.158	<0.001*
Absent	267	71.97	534	91.75		
Present	104	28.03	48	8.25		
CK14					11.671	0.001*
Absent	304	82.83	518	90.40		
Present	63	17.17	55	9.60		
p53					33.999	<0.001*
Absent	227	62.71	457	80.04		
Present	135	37.29	114	19.96		
MIB1					28.563	<0.001*
Absent	59	29.95	154	54.61		
Present	138	70.05	128	45.39		
BCL2					10.510	0.001*
Absent	21	19.63	21	7.89		
Present	86	80.37	245	92.11		
FOXA1					26.495	<0.001*
Absent	178	62.9	174	43.0		
Present	105	37.1	231	57.0		
TFF1					1.563	0.211
Absent	133	53.6	174	48.5		
Present	115	46.4	185	51.5		
CD71					25.926	<0.001*
Absent	90	32.4	220	51.9		
Present	188	67.6	204	48.1		
TK1					9.172	0.002*
Absent	94	41.6	183	54.6		
Present	132	58.4	152	45.4		
PELP1(170)					0.375	0.540
Absent	250	87.4	369	85.8		
Present	36	12.6	61	14.2		
RERG					4.291	0.038*
Absent	214	78.7	306	71.7		
Present	58	21.3	121	28.3		

Association of DACH1 with breast cancer biomarkers. N=number of samples. Statistically significant p-values are indicated with *.

disease free interval ($\chi 2$=1.75, p=0.17), tumour recurrence ($\chi 2$=2.11, p=0.15) and distant metastasis ($\chi 2$=3.74, p=0.053) over 10 years.

The effect of endocrine therapy on the ability of DACH1 to predict breast cancer specific survival using Kaplan-Meier modelling was considered. DACH1 positivity was associated with good survival in patients treated with tamoxifen (χ^2=8.30, p=0.004) and in addition, also showed a strong trend in patients not receiving tamoxifen (χ^2=3.7, p=0.055).

The predictive independence of DACH1 was tested using multivariate Cox regression models. DACH1 remained an independent predictor of specific survival (Hazard ratio (HR))=0.69, *p*=0.012, 95% CI=0.52 – 0.92) when hormonal treatment was considered, but not when tumour grade (HR=1.006, p=0.97, 95% CI=0.77 – 1.31) was taken into account.

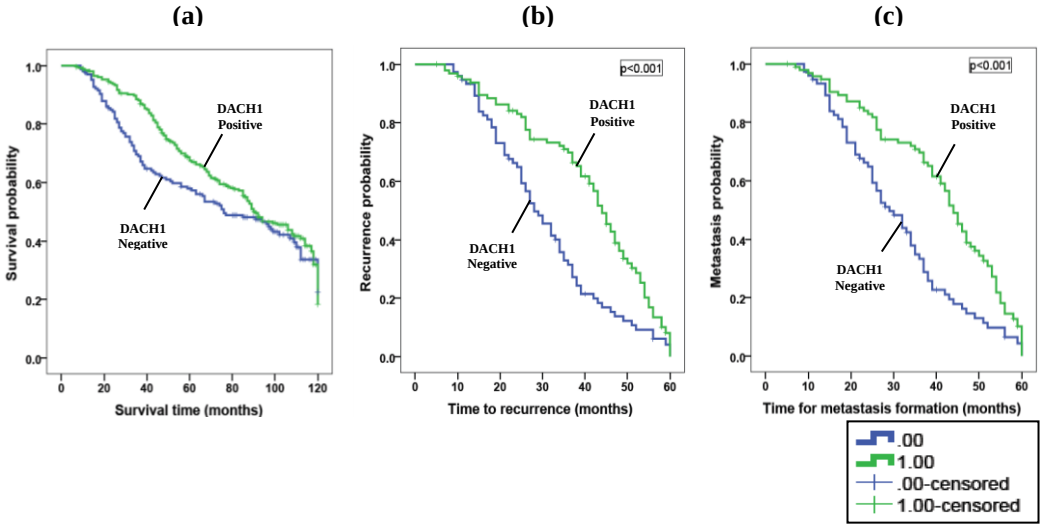


Figure 21: Association of DACH1 protein expression with clinical outcome of patients. Association of DACH1 protein expression with clinical outcome of patients using Kaplan-Meier plots for 5 year post-diagnostic for (a) Survival, (b) Recurrence and (c) Distant metastasis.

4.6 Discussion

The aim of this chapter was to investigate the interactions between ER-associated genes in breast cancer microarray dataset, utilising an ANN-based network inference method. The analysis was successful in quantifying the interactions. Promisingly most of the interacting hubs recorded were found to be relevant with respect to breast cancer or tumorigenesis process of disease progression i.e., *FOXA1*, *GATA3, CA12* and *CDH3*. Surprisingly, when comparing the interacting hubs identified from the total population of sample without ER stratification and after stratification, there were some distinct observations noticed. We could notice that the negative hubs were found to be more consistent than the positive hubs. Below are some of the most prominent genes found to be highly interacting with others and their literature corroboration.

4.6.1 Literature validation of ER-associated markers

FOXA1:

The forkhead box A1, *FOXA1* gene is localised on 14q21.1. The gene codes for a protein which is a forkhead class of DNA binding proteins acting as transcriptional activators. *FOXA1* is involved in embryonic development, differentiation and tissue specific gene expression regulation (Lupien et al., 2008). *FOXA1* binding with consensus sequences on DNA and interacting with histones, brings about relaxation of chromatin and facilitates enhancer or promoter site

availability to other transcriptional regulators (Taube et al., 2010). Supporting our findings, *FOXA1* is found to be involved in ESR1-mediated transcription (Hurtado et al., 2011). Specifically in breast cancer, *FOXA1* is necessary for ESR1-NKX2-1 promoter binding (Minoo et al., 2007), ESR1-induced repression of RPRM (Malik et al., 2010), and also facilitates apoptosis by repressing BCL2 expression (Song et al., 2009). Furthermore, *FOXA1* has also been considered as a prognostic indicator (Habashy et al., 2008), of disease-free survival in luminal A subtype in breast cancer (Badve et al., 2007).

From ANN-based interaction inferencing, the *FOXA1* gene which appeared as a positive hub in total samples, was not found to be an hub in ER-stratified samples. This suggests that *FOXA1* expression is controlled by other genes. Conformingly, *FOXA1* expression was found to be directly dependent on *GATA3* expression in ER positive samples (Albergaria et al., 2009). Furthermore, the reason *FOXA1* is unable to act as interacting hub during sample stratification may be because of the expression pattern of the gene itself. *FOXA1* was found to have an up regulated expression in both ER-Positive and ER-Negative samples with mean expression of 10.52 and 8.44 respectively. The population distribution of *FOXA1* revealed that there was clear step-function in the expression pattern of the *FOXA1* that hindered an active classification of samples due to lack of learning of the pattern by the Stepwise-ANN.

GATA3:

The *GATA3* gene, encodes a protein belonging to GATA family of transcription factors and is found to be vital for T-cell development since it binds

to 5´-AGATAG-3´ consensus sequence of enhancer of T-cell receptor α and β (Wei et al., 2011). GATA3 is related with Barakat syndrome which is characterised by hypothyroidism, renal dysplasia and sensorineural deafness (Zahirieh et al., 2005), and rheumatoid arthritis (Kawashima and Miossec, 2005).

In breast cancer, *GATA3* expression is associated with ER (Hoch et al., 1999), and exhibits coregulatory activity with ESR1 (Eeckhoute et al., 2007). In non-invasive (MCF-7) and in invasive (MDA-MB-231) breast cancer cell lines, GATA3 is found to be associated with a reversal of epithelial to mesenchymal transition, thus inhibiting metastasis (Yan et al., 2010). GATA3 as a marker of hormonal response (Fang et al., 2009), is found to be a prognostic marker in TMA meta-analysis (Mehra et al., 2005). Even though, GATA3 has a strong association with ESR1, it lacks prognostic value independently (Voduc et al., 2008), but serves as a prognostic indictor in hormone-negative tumours along with FOXA1 (Albergaria et al., 2009). Supporting to this, in our ANN-based interaction inference model, GATA3 is found to positively interact with FOXA1 in the breast cancer samples.

CA12:

CA12, a transmembrane carbonic anhydrase isoform is activated in response to low oxygen as a marker of hypoxia (Wykoff et al., 2001). CA12 regulates intra and extracellular pH and is an important therapeutic molecule targetable by carbonic anhydrase inhibitors (Pastorekova et al., 2008). Hypoxia inducible growth factor brings about a change in cellular expression programme by activating genes involved in angiogenesis, anaerobic glycolysis, deadhesion and

invasion. This potentiates aggressive, metastatic spread and poor response to therapy in tumour cells. Hypoxia through physiological stress causes accumulation of acidic products and induction of *CA12* expression in breast cancer cells (Wykoff et al., 2000).

In breast tumour cells, cellular response to hypoxia is mediated by *Hypoxia Inducible Factor* 1 (*HIF1*) whose expression is indicative of poor prognosis. In a study by Tureci el at., 77 out of 103 ER positive breast cancer patients showed *CA12* expression associated with negative EGFR, absence of necrosis and low grade of tumor (Tureci et al., 1998). In another study, 84% DCIS, and 71% invasive breast cancers show *CA12* expression with low grade and good prognosis (Potter and Harris, 2003). Tumors positive for CA12 showed lower relapse rate, better survival and better prognosis in invasive breast cancer (Watson et al., 2003).

IGFBP4:

The insulin like growth factor binding protein (*IGFBP*) 4 belongs to Insulin like growth factor system consists of two peptides (IGF I & II), two receptors (IGF IR & IIR) which are homologous to each other and pro-insulin (Douglas et al., 2010), six insulin like growth factor binding proteins (1-6) and four insulin like growth factor binding protein related peptides. The mechanism of action of IGFBP4 involves binding with IGF I & II and then bringing about a change in their biological activity (Durai et al., 2006). Over expression of IGFBP4 inhibits cell cycle proliferation and induces apoptosis. Reduced expression of IGFBP4

increases the availability of IGF I, affects vascular growth response and presents anti-proliferative action which is important in cancer biology.

In breast cancer IGF I and II act as potent mitogens stimulating growth. In a study by Douglas et al., on breast cancer patients, inhibited binding of IGF to IGFR1 indicating that IGF I and II stimulated breast cancer cell growth (Douglas et al., 2010). In ER α positive breast cancers, estradiol combines with IGF to promote cell growth through ligand independent activation of ER occurring by activation of MAPK phosphorylation (Wood and Yee, 2000). *IGFBP4* is also one of the highly expressed genes in the ER-Positive group with tamoxifen failure compared to the control study group (Vendrell et al., 2008). IGFBP4 high expression was found in the nucleus and cytoplasm in breast tumour cells. The raised nuclear expression was induced by estrogen which stimulated ER positive breast cancer cell growth (Yoshida et al., 2004). IGFBP4 formed an independent prognostic factor of disease free survival in ER positive breast cancer patients having a high IGFBP4 expression (Mita et al., 2007).

CDH3:

The placental (P)-cadherin 3, *CDH3* gene encodes a calcium dependent cell-cell adhesion glycoprotein consisting of five extracellular repeats, a transmembrane and a highly conserved cytoplasmic trail peptide. *CDH3* is localised at 16q22.1 within six cadherin cluster. *CDH3* expression is highly associated with differentiation of lung carcinoma cells (Shimoyama et al., 1989). Imai et al., found CDH3 to be having tumour antigen-specific immunotherapy potential in colorectal and pancreatic cancers (Imai et al., 2008).

In breast cancer, P-cadherin inhibits invasion and metastasis functioning as a pro-adhesive and anti-invasive molecule (Paredes et al., 2004). Moreover, P-cadherin indicates poor prognosis in patients due to hypomethylation of its promoter region (Paredes et al., 2005).

VAV3:

The *VAV3* oncogene is a member of the VAV family of guanine nucleotide exchange factors associated with nucleotide free-state of guanocine triphosphatase, cytoskeleton rearrangement and transcriptional alterations. The VAV family consists of 3 members; *VAV, VAV2 and VAV3* present on chromosome 19, 9 and 1 respectively in humans. Among the members, the *VAV* gene is expressed in the hematopoietic cells causing stimulation of Interleukin-2. *VAV2* expression is ubiquitous and *VAV3* has a broader expression profile (Moores et al., 2000). In the case of the *VAV3* gene, the zinc finger domain region increases VAV3's binding and activation ability (Bustelo, 2001). Movilla et al., has confirmed VAV3's participation in protein tyrosine kinase pathway (Movilla and Bustelo, 1999).

VAV3 also causes androgen receptor transactivation by PI3K/AKT pathway and thus plays an important role in cancer development and progression (Dong et al., 2006). In tumour cells, activation and hypersensitivity is mediated by PI3K/AKT signalling which causes progression and antiestrogen resistance in breast cancer cells. Overexpression of the VAV3 protein is noticed in poorly differentiated tumours. The SH2, PH and DH domains of VAV3 interact with the

receptor protein tyrosine kinase, PIP3 and ERα respectively thus potentiating regulation of cell differentiation cell growth and ERα activation (Lee et al., 2008).

Through our immunostaining of VAV3 in the 739 TMAs from Nottingham Tenovous Series of breast cancer patients, we did not observe association of VAV3 protein expression with any clinicopathological markers and patient clinical outcome. But we did find VAV3 significantly associated with cell adhesion marker E-cadherin (p=0.010), luminal markers FHIT (p=0.003) and PELP1 (p=0.007).

DACH1:

The Drosophila *DACH* gene, is a member of the retinal determination gene network family of nuclear proteins. It plays an important role in the development of Drosophila eye and limb, and regulates cell fate programming by encoding a chromatin associated protein (Hammond et al., 1998). In humans, *DACH1* expression is associated with breast, ovary, renal, and gastric carcinomas.

In breast cancer, DACH1 plays a vital role. DACH1 physically associates with ERα to inhibits ER ligand dependent activity and *DACH1* expression correlates with ERα (Popov et al., 2009). DACH1 inhibits cellular migration and metastasis by inhibiting interleukin-8 (Wu et al., 2008). Moreover, DACH1 is associated with cell fate determination and cellular proliferation by inhibiting cyclin D1 activity, wherein cyclin D1 acts as a check point for transition from the G1 to S phase of the cell cycle (Wu et al., 2006). Many cancer cells over express cyclin D1 which correlates with their DACH1 expression (Nan et al., 2009).

Cellular proliferation and migration in hormone dependent breast and prostate cancer repressed by DACH1 (Popov et al., 2009). *DACH1* expression is found to be greater in normal cells compared to metastatic cells in breast cancer TMA samples (Wu et al., 2008). Reduced DACH1 and increased FOX expression can cause deregulation of the genes associated with tumorigenesis (Zhou et al., 2010).

Through our ANN based interaction modelling method in ER-positive samples, we found *DACH1* to be one of the positive hubs and positively interacting with other ER associated genes. On the contrary, *DACH1* did not appear as an interacting hub in ER-negative samples by network inference as well as immunostaining methods, supporting its relevance with ER positivity.

Immunostaining for DACH1 has revealed significantly reduced expression in invasive cancer compared to normal breast epithelium. DACH1 expression was also found inversely proportional to mitosis and cyclin D1 expression in breast cancer patient samples (Wu et al., 2006). More recently, DACH1 expression is also correlated with reduced expression of IL-8 and other related chemokines, thus inhibiting cellular migration and invasion in MCF10A cells (Wu et al., 2008); as well as in endometrial cancers (Nan et al., 2009). It is known that DACH1 homozygous deletion stimulates tumorigenesis in glioma cells (Watanabe et al., 2011). Furthermore, it is also found that DACH1 overexpression represents poor prognosis in ovarian cancers (Liang et al., 2012).

Significant association of DACH1 with cellular proliferative markers such as MIB1 and Bcl2 in our studies is supported with the fact that DACH1

suppresses cellular proliferation, growth and survival through TGF-beta signalling pathway by binding with Smad4 (Wu et al., 2003), and DNA synthesis inhibition by DACH1 in breast epithelial cells (Wu et al., 2006). Contrastingly, DACH1 loses its tumour suppressor activity in cancer cells grown with IGF-1 (DeAngelis et al., 2011).

The hormonal responsiveness of DACH1 in breast cancer patients was also confirmed by significant association of DACH1 expression with ER (also through our DACH1 interactome), PgR and HER2. Moreover, the resistance in anti-EGFR treatment using Cetuximab is also found (DeAngelis et al., 2011). But with respect to tamoxifen treatment, DACH1 was significantly associated with good prognosis.

Through immunostaining of the DACH1 protein, we investigated DACH1 protein expression association in 993 samples from the Nottingham Tenovous series TMA. DACH1 protein expression was significantly associated with clinicopathological factors such as age (p=0.006), menopause status (p=0.003), tumour grade (p=<0.001), tumour type (p=<0.001), NPI (p=<0.001) and endocrine treatment (p=0.003). DACH1 protein expression was also significantly associated with breast cancer markers such as hormonal markers ER (p=<0.001) and PR (p=<0.001); HER2 (p=0.010); growth factor EGFR (p=0.012); immunological markers CK5/6 (p=<0.001), CK14 (p=0.001), CK18 (p=<0.001) and CK19 (p=0.016); cell cycle marker p53 (p=<0.001); cellular proliferation markers MIB1 (p=<0.001) and BCL2 (p=0.001); luminal markers FOXA1

(p=<0.001), CD71 (p=<0.001), TK1 (p=0.002) and RERG (p=0.038). DACH1 was found to be an independent predictor of specific survival (HR=0.69, p=0.012) with hormonal therapy.

4.7 Summary and Conclusion

Breast cancer is a complex and heterogenous disease, ER protein alone is unable to provide extensive information for biology of breast cancer. However, ER-associated genes and their contribution towards the ER phenotype in the breast cancer may provide sufficient information. To investigate the interaction among ER-associated genes, an ANN-based gene-gene network inferencing method was adopted. This inference method was successful in quantifying the interaction magnitudes depending on the weights used to predict the expression of the target gene from the source gene expression. The representation of genes as nodes and interaction as edges enabled the visualisation of the network models. Positive weights representing positive regulation and negative weights representing negative regulation of interaction was also applied. The directionality of interaction was assigned from source to target gene.

ANN-based network inference analysis of breast cancer sample expression profile was helpful in identification of interacting hubs. There were quite distinct differences in interacting hubs between the total breast cancer samples and ER-status based stratified samples. The ANN method was found to be promising in pulling out most of the breast cancer associated genes such as *FOXA1, GATA3,*

CDH3, CA12 successfully and inferring involvement of them in breast cancer development and disease prognosis.

Clinical validation of the DACH1 protein expression in breast tumour tissues suggested that DACH1 could act as an independent predictor for breast cancer specific survival with hormonal therapy.

Chapter 5

SYSTEMS BIOLOGY OF ER RELATING TO BREAST CANCER

Chapter abstract: The systems biology approach to modelling biological system comprises a dynamic interconnection of elements. In the case of complex diseases such as breast cancer, investigation of cross-talk between components of ER (ERα and ERβ) might give information on disease aggressiveness and prognosis. To investigate contenders of ER system and cross talk between them, in this study we have identified and utilised *ESR1* (coding for ERα) and *ESR2* (coding for ERβ) predictive markers with ANN-based modelling approaches. The bespoke ANN-based network inference algorithm was also used for construction of interactomes of *ESR1* and *ESR2*. Simulation of the ER system from known interacting molecules of *ESR1* and *ESR2* using ANN-based network inference revealed close association between ER-associated markers such as *PGR, AR, EGFR, FOXA1* and *RARA*, validating biological relation between ER and PR along with AR. ER system simulation also validated *ERBB2* cluster as a separate entity apart from ER and *PGR*.

5.1 Introduction

Systems biology is a science dealing with the study of biology and medicine incorporating computational mathematical approaches to model biological systems

through prerequisite understanding of how the elements are interconnected or interrelated, in both a structural and dynamic way (Kitano, 2002b, Kitano, 2002a).

Broadly, there are two approaches adopted by researchers to understand the interconnections.

- ***Pathway-centric*** approaches are governed by the associated pathway of the marker or a process studied. Apart from primitive single-pathway models, recently the cross-talk between the kinase inhibitors influencing the migration of cells with overexpression of Her2 in mammary epithelial cells was investigated adopting a multi-pathway model (Lee et al., 2006).
- ***Interactomic*** approaches are the consideration of all possible cross-talk or interactions of the marker of study with protein-DNA or protein-protein associations (Rual et al., 2005).

There has been an exponential increase in the application of 'systems biology approaches' from nearly one hundred reported publications in 2000 to over one thousand in 2012 in NCBI's PubMed library. Examples of their successful applications include: - Drug discovery (Cho et al., 2006); predicting oncogenes and molecular perturbations in B-cell lymphoma (Mani et al., 2008); neurodegenerative disorders such as multiple sclerosis and Alzheimer's (Goni et al., 2008); for pathway modelling and discovery (Bosl, 2007). These dealt with various aspects of disease and effective treatment discovery.

5.1.1 Systems biology in breast cancer

It is currently accepted that breast cancer is not a single disease, but instead it is a highly heterogenous disease, with complex pathology. It shows dysregulation of multiple pathways governing cellular proliferation, differentiation, migration and death. It could arise from multi-pathway dysregulation. In addition Perou et al. has formulated "intrinsic subtypes of breast cancer" identifying 5 distinct tumour types and a normal breast-like group, to predict patient relapse, overall survival and response to endocrine and chemotherapy (Caldas and Aparicio, 2002, Perou, 2011, Perou et al., 2000).

Predicting clinical outcome in patients is very necessary but highly challenging. There are two computational algorithms that have been developed to predict the clinical outcome in patients using patterns of gene expression data available, namely Mammaprint® and OncotypeDX RS® (Kaklamani, 2006, Slodkowska and Ross, 2009). Gene expression profiling using DNA microarray has contributed considerably to our understanding of complex pathways (Itadani et al., 2008), thus can promisingly help us to understand the molecular heterogeneity of breast cancer formation, progression and recurrence.

Recently there has been an increase in biomarker identification using genome-wide expression profile analysis (Golub et al., 1999, Ramaswamy and Perou, 2003). This has enabled researchers to attempt systems biology approaches in breast cancer. Chuang et al. have investigated markers associated with breast cancer metastasis by using protein network-based methods. Here individual markers having differential

expression between metastasis groups were considered as sub-networks of a protein-protein interaction network rather than as conventional individual markers (Chuang et al., 2007). Furthermore, Faratian et al. has proposed a novel cell signalling kinetic model to identify the role of PTEN in trastuzumab resistance in breast cancer (Faratian et al., 2009).

Even though systems biology approaches have seen significant applications, current published approaches possess some limitations. In the majority of the studies, investigators have considered only a small subset of highly predictive or highly discriminative markers associated with their question, later extending their studies to validation of findings based on Gene Ontology pathway-based analysis (Chuang et al., 2007). One of the prime hurdles to pathway-based analysis is the absence of information for the majority of human genes in relation to a given pathway. Furthermore, the overlapping of genes with other different pathways makes interpretation highly challenging.

In the case of complex diseases such as breast cancer, little information is known about predictive and prognostic markers. Even less is known about the cross-talk between the ER counterparts such as ERα (coding gene *ESR1*) and ERβ (coding gene *ESR2*). Knowledge of the interaction between *ESR1* and *ESR2* could be beneficial for determining more contenders of aggressive disease.

Discovery of biomarkers in the ER pathway holds promise for finding new druggable targets potentially inhibiting ER. Use of reductionist approaches through knock out techniques to characterise the functionality of biomarkers is a time and effort consuming task. Moreover, the functionality is a complex result of collective

influence of all the possible contenders at the same time and space in a biological system. Hence, non-reductionist computational approaches to systems biology have the potential to represent the complexity of biological system.

ANNs have been shown to be able to deal with the non-linearity in biological datasets, making them successful in class prediction (Lancashire et al., 2009), finding markers of biological relevance. For example ER class prediction (Dhondalay et al., 2011) and gene-gene network inferencing in breast cancer (Dhondalay et al., 2012, Lemetre et al., 2009). From a computational viewpoint, systems biology has inspired network modelling approaches promising need-oriented methodologies specific to data associated with given question (Kreeger and Lauffenburger, 2010). In this study, ANN is used for application in systems biology.

5.2 Aims and objectives

- As both the components of ER system i.e., ERα (*ESR1*) and ERβ (*ESR2*) play an important role in determining the aggressiveness of disease, the first aim of work is the identification of predictive markers of ESR1 and ESR2.

- Following this, the investigation of cross-talk between *ESR1*-associated and *ESR2*-associated markers through ANN-based network inferencing method highlights the influence of associated markers on overall ER system.

- Finally, ANN-based network inference contenders are validated with known contenders of ER system.

5.3 Materials and methods

5.3.1 Breast cancer microarray dataset

The E-GEOD-20194 dataset (Shi et al., 2010), a pre-normalised cDNA microarray breast cancer dataset developed on the HG-U133A platform, was obtained from EMBL-EBI database library (http://www.ebi.ac.uk/arrayexpress). The dataset consisted of 22283 probesets across 278 primary breast cancer samples. The samples were classified based on the median expression values of *ESR1* and *ESR2* probes as *positive* for samples having expression values greater than median values and as *negative* for samples having expression values lesser than median values of the population.

5.3.2 ANN architecture and model development

Selection of predictive markers:

A Stepwise-ANN coupled with back-propagation of error algorithm as described earlier (Lancashire et al., 2009), was used to predict *ESR1* and *ESR2* class. All available 22283 probesets from the array were used as single input variables for training each model. The algorithm parameters consists of momentum of 0.5, learning rate of 0.1 and 3000 epochs with a threshold window of 1000. The models were subjected to MCCV partition as 60% for training, 20% for testing and remaining 20% for validation. The whole process was repeated for 10 independent iterations. The probesets were selected based on the predictiveness of each probeset

by considering averaged error in the test subset of samples. From this ranking, 100 highly predictive probesets were selected for network inferencing.

Network inference algorithm:

The 100 highly predictive probesets for *ESR1* and *ESR2* derived from the stepwise-ANN method were analysed to model interactions by a gene-gene network inference algorithm (Lemetre et al., 2009). Here, the expression of a given probeset was predicted one at a time from all the other probesets iteratively i.e., 99 used to predict 1. The samples are subjected to MCCV strategy to optimise performance on unseen data in training. Weights of the trained models were used to determine the magnitude network. Once all probes had been considered as output, a matrix of interactions was produced. The weight of the link with respect to input and output probes is determined along with sign and magnitude. This complete process was repeated 10 times independently to develop a generalised models and averaged across all 10 to increase consistency and averaged accuracy. All interactions with *ESR1* and *ESR2* were selected to construct *ESR1* and *ESR2* interactomes respectively.

Visualisation of network interactomes:

As described earlier in Chapter 2 (*Section 2.8.6*), we have utilised Cytoscape Ver 2.7.3 software to visualise the interactome models of both *ESR1* and *ESR2*. In the interaction models, the genes were represented as nodes and the interaction intensities as edges. The directionality of the edges as an arrow is maintained from source to target. The edges were either positive (green) or negative (red) along with

the magnitude of interaction as the thickness of the edges. The thickness of edge was scaled based on minimum and maximum interaction values across all interactions to squeeze the dimensionality of edges to create more generalized models.

5.3.3 ESR1 and ESR2 interactome simulation and validation

All previously reported (known) interactions of *ESR1* and *ESR2* were pooled from both genomic and proteomic interaction databases namely; VEGA (Vertebrate Genome Annotation database) (Wilming et al., 2008), Ensemble from Welcome Trust Sanger Institute, Uniprot KB (UniProtein Knowledgebase), HPRD (Human Protein Reference Database) (Peri et al., 2003), HGNC (HUGO Gene Nomenclature Committee) (www.genenames.org), OMIM (Online Mendelian Inheritance in Man) from John Hopkins University, Entrez gene, GeneBank and RefSeq from National Centre for Biotechnology Information, STRING (Szklarczyk et al., 2011), MINT (Molecular INTeraction database) (Chatr-aryamontri et al., 2007) and Gene Cards (www.genecards.org). The BioGRID (Biological General repository for Interaction Datasets) platform was used to merge interactions from the above mentioned databases (Breitkreutz et al., 2003).

5.3.4 ER system simulation

Both, the ERα (*ESR1*) and ERβ (*ESR2*) subunits were considered and combined to construct a simulation of the ER system. After merging 438 interactions

of *ESR1* and 422 interaction of *ESR2*, a total of 412 unique molecules were considered to be within the putative ER system. The gene expression profiles of all 412 genes were selected from the E-GEOD-20104 breast cancer microarray dataset. As described earlier, an ANN gene-gene network inference method was used for network modelling and simulation. 10 independent model simulations of 412 genes across 278 samples were performed for the generalisation of models.

5.4 Results:

5.4.1 ESR1 and ESR2 predictive markers

A bespoke Stepwise-ANN based predictive modelling approach (Chapter 2, section 2.7) was used to identify predictive probes for *ESR1* and *ESR2* gene expression-based classification. The probes were ranked based on the averaged error developed by the probe in the test subset of samples. Table 17 and Table 18 are the tabulation of the top 100 predictive probes for ESR1 and ESR2 classification.

5.4.2 ESR1 and ESR2 interactomes

The ESR1 and ESR2 interactome network inference model was created using the expression profile of top 100 highly predictive probesets from the *ESR1* and *ESR2*-based classification model separately. From the bespoke ANN-based network inference algorithm (Chapter 2, Section 2.8), out of 9100 potential interactions computed from 100 predictive probesets of *ESR1* and *ESR2*, only 75 unique

Table 17: The top 100 predictive probes for ESR1 classification.

Rank Order	Probeset ID	Gene Symbol	p-value	Rank Order	Probeset ID	Gene Symbol	p-value
1	205225_at	ESR1	4.49E-05	51	205597_at	SLC44A4	2.29E-03
2	204623_at	TFF3	8.98E-05	52	201983_s_at	EGFR	2.33E-03
3	219414_at	CLSTN2	1.35E-04	53	212195_at	IL6ST	2.38E-03
4	205186_at	DNALI1	1.80E-04	54	209696_at	FBP1	2.42E-03
5	218806_s_at	VAV3	2.24E-04	55	212441_at	KIAA0232	2.47E-03
6	219051_x_at	METRN	2.69E-04	56	205472_s_at	DACH1	2.51E-03
7	209459_s_at	ABAT	3.14E-04	57	41660_at	CELSR1	2.56E-03
8	204508_s_at	CA12	3.59E-04	58	204863_s_at	IL6ST	2.60E-03
9	209623_at	MCCC2	4.04E-04	59	219197_s_at	SCUBE2	2.65E-03
10	205354_at	GAMT	4.49E-04	60	210735_s_at	CA12	2.69E-03
11	214440_at	NAT1	4.94E-04	61	218211_s_at	MLPH	2.74E-03
12	208615_s_at	PTP4A2	5.39E-04	62	206469_x_at	AKR7A3	2.78E-03
13	203963_at	CA12	5.83E-04	63	212496_s_at	JMJD2B	2.83E-03
14	218259_at	MKL2	6.28E-04	64	214404_x_at	SPDEF	2.87E-03
15	200670_at	XBP1	6.73E-04	65	209602_s_at	GATA3	2.92E-03
16	205471_s_at	DACH1	7.18E-04	66	210085_s_at	ANXA9	2.96E-03
17	212638_s_at	WWP1	7.63E-04	67	213419_at	APBB2	3.01E-03
18	218976_at	DNAJC12	8.08E-04	68	208617_s_at	PTP4A2	3.05E-03
19	204862_s_at	NME3	8.53E-04	69	212151_at	PBX1	3.10E-03
20	205066_s_at	ENPP1	8.98E-04	70	212196_at	IL6ST	3.14E-03
21	219913_s_at	CRNKL1	9.42E-04	71	204914_s_at	SOX11	3.19E-03
22	201826_s_at	SCCPDH	9.87E-04	72	217838_s_at	EVL	3.23E-03
23	205355_at	ACADSB	1.03E-03	73	204881_s_at	UGCG	3.28E-03
24	209460_at	ABAT	1.08E-03	74	212148_at	PBX1	3.32E-03
25	215867_x_at	AL050025	1.12E-03	75	200711_s_at	SKP1A	3.37E-03
26	201754_at	COX6C	1.17E-03	76	212446_s_at	LASS6	3.41E-03
27	218195_at	C6orf211	1.21E-03	77	203928_x_at	MAPT	3.46E-03
28	208682_s_at	MAGED2	1.26E-03	78	203749_s_at	RARA	3.50E-03
29	204798_at	MYB	1.30E-03	79	216381_x_at	AKR7A3	3.55E-03
30	215729_s_at	VGLL1	1.35E-03	80	212956_at	TBC1D9	3.59E-03
31	221710_x_at	C1orf78	1.39E-03	81	209603_at	GATA3	3.64E-03
32	214164_x_at	CA12	1.44E-03	82	205862_at	GREB1	3.68E-03
33	35666_at	SEMA3F	1.48E-03	83	200810_s_at	CIRBP	3.72E-03
34	212492_s_at	JMJD2B	1.53E-03	84	203628_at	IGF1R	3.77E-03
35	205081_at	CRIP1	1.57E-03	85	206754_s_at	CYP2B6	3.81E-03
36	221765_at	UGCG	1.62E-03	86	203929_s_at	MAPT	3.86E-03
37	201508_at	IGFBP4	1.66E-03	87	209604_s_at	GATA3	3.90E-03
38	210652_s_at	C1orf34	1.71E-03	88	208873_s_at	REEP5	3.95E-03
39	213712_at	ELOVL2	1.75E-03	89	211000_s_at	IL6ST	3.99E-03
40	209173_at	AGR2	1.80E-03	90	220192_x_at	SPDEF	4.04E-03
41	212442_s_at	LASS6	1.84E-03	91	212960_at	TBC1D9	4.08E-03
42	214552_s_at	RABEP1	1.88E-03	92	218807_at	VAV3	4.13E-03
43	205696_s_at	GFRA1	1.93E-03	93	203627_at	IGF1R	4.17E-03
44	212637_s_at	WWP1	1.97E-03	94	212099_at	RHOB	4.22E-03
45	205009_at	TFF1	2.02E-03	95	202089_s_at	SLC39A6	4.26E-03
46	204667_at	FOXA1	2.06E-03	96	211712_s_at	ANXA9	4.31E-03
47	203256_at	CDH3	2.11E-03	97	203144_s_at	KIAA0040	4.35E-03
48	209443_at	SERPINA5	2.15E-03	98	222125_s_at	PH_4	4.40E-03
49	204497_at	ADCY9	2.20E-03	99	215552_s_at	ESR1	4.44E-03
50	206401_s_at	MAPT	2.24E-03	100	202088_at	LIV_1	4.49E-03

Tabulation of top 100 ranked gene probes found associated with ESR1 classification using Stepwise-ANN analysis. The average of test subset error for each gene under each independent 10 repetitions was considered for ranking. The probes with minimum error is being ranked highest in the table.

Table 18: The top 100 predictive probes for ESR2 classification.

Rank Order	Probeset ID	Gene symbol	p-value	Rank	Probeset ID	Gene symbol	p-value
1	211119_at	ESR2	4.49E-05	51	216187_x_at	AF222691	2.29E-03
2	r2_Ec_bioD_3_at	K01391	8.98E-05	52	220222_at	PRO1905	2.33E-03
3	CreX_3_at	K01391	1.35E-04	53	220385_at	JPH2	2.38E-03
4	r2_P1_cre_3_at	K01391	1.80E-04	54	208448_x_at	IFNA16	2.42E-03
5	CreX_5_at	K01391	2.24E-04	55	211835_at	AJ225093	2.47E-03
6	222207_x_at	LOC389517	2.69E-04	56	215593_at	AK023918	2.51E-03
7	BioDn_5_at	K01391	3.14E-04	57	204428_s_at	LCAT	2.56E-03
8	r2_Ec_bioD_5_at	K01391	3.59E-04	58	215320_at	LOC645961	2.60E-03
9	216707_at	AL162044	4.04E-04	59	215179_x_at	PGF	2.65E-03
10	BioDn_3_at	K01391	4.49E-04	60	206899_at	NTSR2	2.69E-03
11	203716_s_at	DPP4	4.94E-04	61	213586_at	CDKN2D	2.74E-03
12	220804_s_at	TP73	5.39E-04	62	220601_at	C16orf70	2.78E-03
13	222086_s_at	WNT6	5.83E-04	63	213155_at	WSCD1	2.83E-03
14	203948_s_at	MPO	6.28E-04	64	216789_at	AK026439	2.87E-03
15	219699_at	LGI2	6.73E-04	65	206320_s_at	SMAD9	2.92E-03
16	211788_s_at	TREX2	7.18E-04	66	210324_at	C8G	2.96E-03
17	hum_alu_at	K01391	7.63E-04	67	206820_at	HRBL	3.01E-03
18	210344_at	OSBPL7	8.08E-04	68	206148_at	IL3RA	3.05E-03
19	219796_s_at	MUPCDH	8.53E-04	69	219011_at	PLEKHA4	3.10E-03
20	216646_at	L11372	8.98E-04	70	211030_s_at	SLC6A6	3.14E-03
21	r2_Ec_bioB_5_at	K01391	9.42E-04	71	206865_at	HRK	3.19E-03
22	207685_at	CRYBB3	9.87E-04	72	210879_s_at	RAB11FIP5	3.23E-03
23	216617_s_at	MAG	1.03E-03	73	214637_at	OSM	3.28E-03
24	207377_at	I_4	1.08E-03	74	205314_x_at	SNTB2	3.32E-03
25	216957_at	USP22	1.12E-03	75	221921_s_at	CADM3	3.37E-03
26	207288_at	FLJ13330	1.17E-03	76	216796_s_at	AK026847	3.41E-03
27	211314_at	CACNA1G	1.21E-03	77	217669_s_at	AKAP6	3.46E-03
28	210565_at	GCGR	1.26E-03	78	206328_at	CDH15	3.50E-03
29	216790_at	AK026465	1.30E-03	79	220611_at	DAB1	3.55E-03
30	216116_at	NCKIPSD	1.35E-03	80	220818_s_at	TRPC4	3.59E-03
31	217715_x_at	BE045142	1.39E-03	81	208088_s_at	CFHR5	3.64E-03
32	215140_at	AL109667	1.44E-03	82	222356_at	FLJ32069	3.68E-03
33	205183_at	HNRPD	1.48E-03	83	206013_s_at	ACTL6B	3.72E-03
34	BioB_5_at	K01391	1.53E-03	84	216572_at	FOXL1	3.77E-03
35	r2_P1_cre_5_at	K01391	1.57E-03	85	221406_s_at	MSH5	3.81E-03
36	206657_s_at	MYOD1	1.62E-03	86	215652_at	AK024382	3.86E-03
37	215861_at	RP4_724E16.2	1.66E-03	87	208494_at	SLC6A7	3.90E-03
38	211915_s_at	TUBB4Q	1.71E-03	88	219642_s_at	PEX5L	3.95E-03
39	208560_at	KCNA10	1.75E-03	89	216204_at	COMT	3.99E-03
40	208354_s_at	SLC12A3	1.80E-03	90	39650_s_at	KIAA0435	4.04E-03
41	210122_at	PRM2	1.84E-03	91	211225_at	FUT5	4.08E-03
42	216630_at	AL110190	1.88E-03	92	208331_at	BPY2	4.13E-03
43	211604_x_at	HAP1	1.93E-03	93	215996_at	GNAI2	4.17E-03
44	214780_s_at	MYO9B	1.97E-03	94	211461_at	CSPG4LYP1	4.22E-03
45	205902_at	KCNN3	2.02E-03	95	220152_at	C10orf95	4.26E-03
46	216900_s_at	CHRNA4	2.06E-03	96	222198_at	AA447740	4.31E-03
47	BioB_3_at	K01391	2.11E-03	97	215628_x_at	AL049285	4.35E-03
48	220071_x_at	CEP27	2.15E-03	98	211398_at	FGFR2_G	4.40E-03
49	r2_Ec_bioB_3_at	K01391	2.20E-03	99	214715_x_at	ZNF160	4.44E-03
50	206647_at	HBZ	2.24E-03	100	206270_at	PRKCG	4.49E-03

Tabulation of top 100 ranked gene probes found associated with ESR2 classification using Stepwise-ANN analysis. The average of test subset error for each gene under each independent 10 repetitions was considered for ranking. The probes with minimum error is being ranked highest in the table.

interactions (8 negative and 67 positive) for *ESR1* and 85 unique interactions (85 negative) for *ESR2* were identified after removing non-significant and multiple interactions. Figure 22 and Figure 23 represents the *ESR1* and *ESR2* interactome developed from the ANN-based network inference algorithm respectively.

5.4.3 Modelling known ER pathway elements using ANN

Modelling of both components of the ER system, i.e., *ESR1* and *ESR2* was achieved by considering possible interacting molecules from the *in silico* interaction databases. 438 interacting molecules with *ESR1* and 422 interacting molecules with *ESR2* were derived from *in silico* curated information using BioGRID and GeneCards databases. A complete list of interacting molecules with ESR1 (Appendix 1) and ESR2 (Appendix 2) are tabulated. After removing duplicate and repeated interactions, 698 interacting molecules from both ESR1 and ESR2 were identified (Appendix 3). All the probes representing the interacting molecules were identified. The gene expression profile of each of them across 278 samples (from E-GEOD-20194) dataset were selected and input into the ANN-based interaction inference algorithm. Ten independent interaction models yielded 169390 interactions making complex inferences. Pre-selection of strongest 100 positive and 100 negative interactions were considered for simple visualisation and inferencing. Figure 24 represents the ER system (ERα and ERβ) interaction models simulated by the ANN-based network inference method.

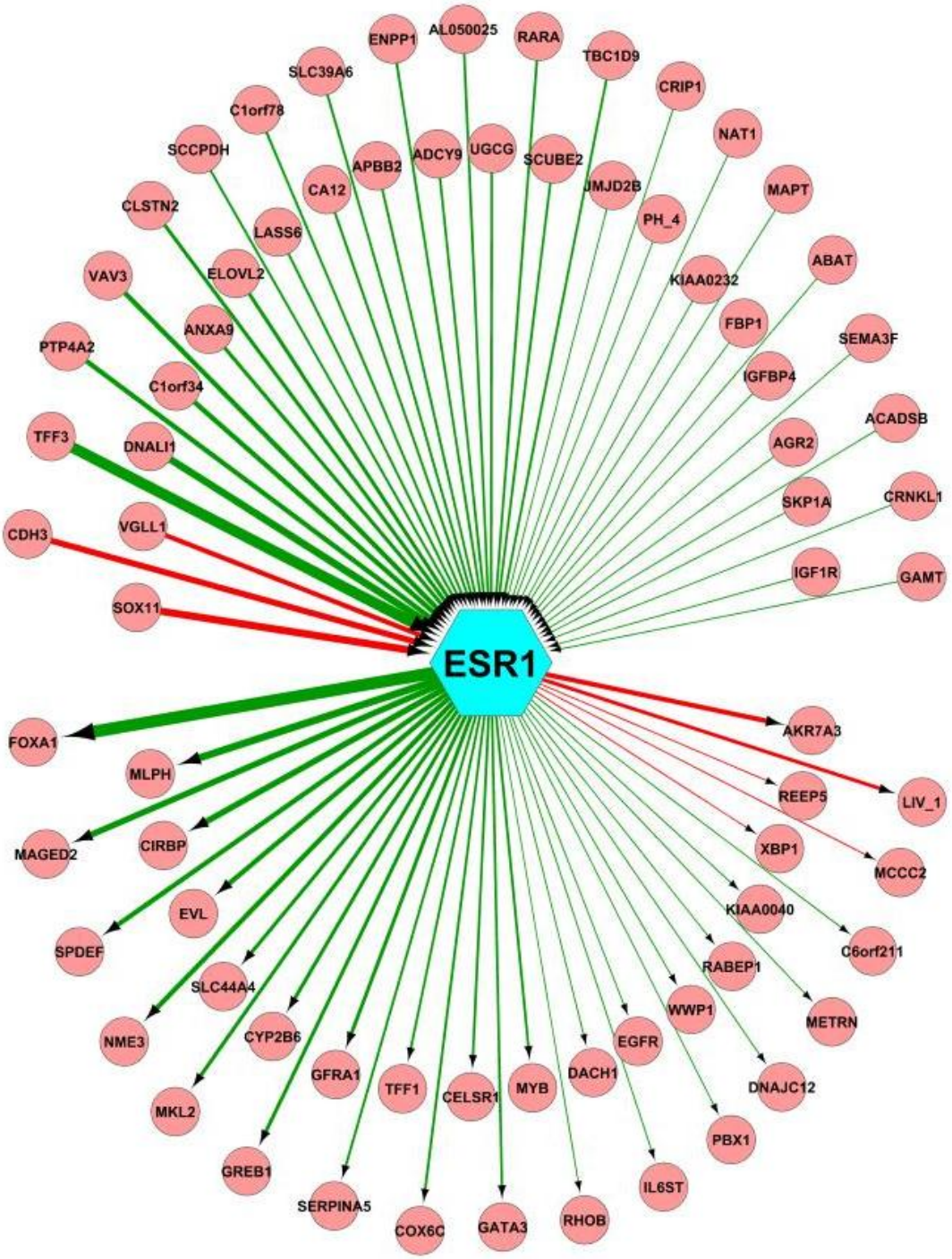


Figure 22: *ESR1* interactome from ANN. *ESR1* interactome using ANN network inference method. The interactome is constructed by selecting all interactions with *ESR1* in total (278) samples from E-GEOD-20194 microarray dataset. The arrows of edges indicate the direction of interaction from source to target. The green edges represents positive interactions and red representing negative interactions. The thickness of edges are scaled based on the population of interactions.

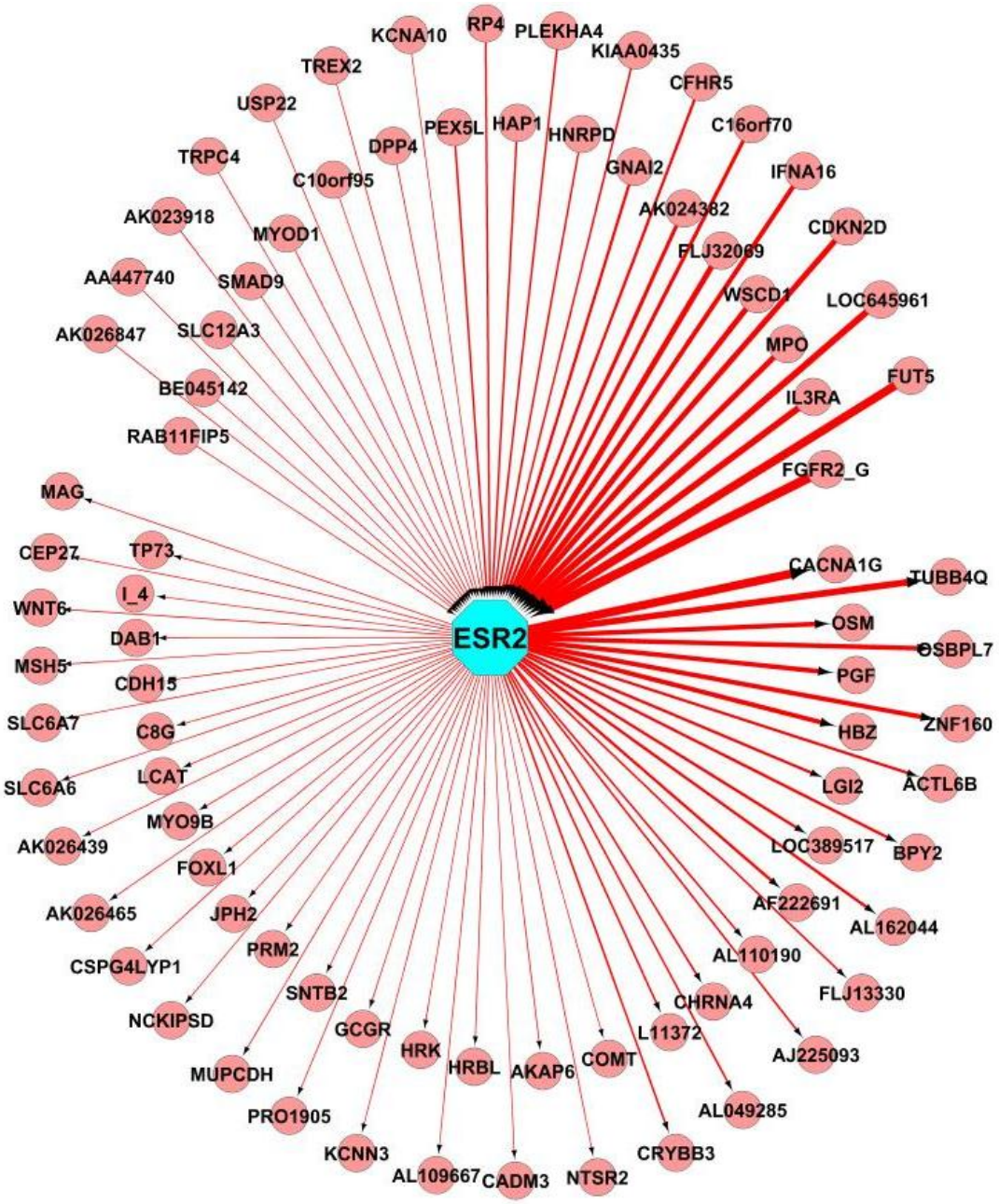


Figure 23: *ESR2* interactome from ANN. *ESR2* interactome using ANN network inference method. The interactome is constructed by selecting all interactions with *ESR2* in total (278) samples from E-GEOD-20194 microarray dataset. The arrows of edges indicate the direction of interaction from source to target. The red edges represents negative interactions. The thickness of edges are scaled based on the population of interactions.

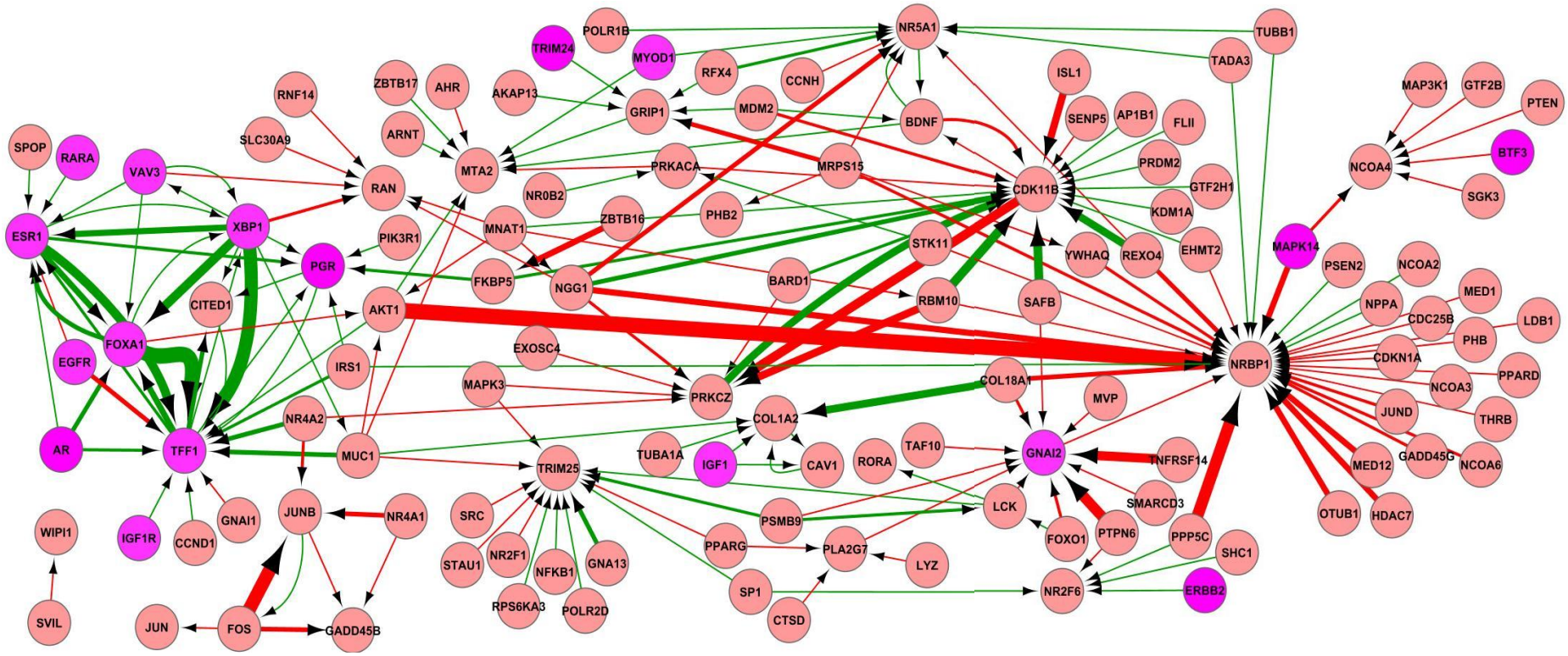


Figure 24: ER system (ERα and ERβ) interaction simulation by ANN-based network interaction inference. The simulation is done on gene expression profile of unique molecules from E-GEOD-20194 breast cancer microarray dataset. The arrows of edges indicate the direction of interaction from source to target. The green edges represents positive interactions and red representing negative interactions. The thickness of edges are scaled based on the population of interactions. The pink nodes are overlapping genes from predictive gene sets of ESR1 and ESR2. The interaction model is constructed by selecting strongest100 positive and 100 negative interactions for simple visualisation.

For validation of interactions, the genes were compared with the predictive gene sets of *ESR1* and *ESR2* by taking 100 highly predictive genes for the respective classes. Upon comparing, 17 genes were found to be overlapping with the predictive genes. These are; Androgen Receptor (*AR*), Basic Transcription Factor 3 (*BTF3*), *EGFR*, Erythroblastic leukemia viral oncogene homolog 2 (*ERBB2/HER2/neu*), *ESR1*, *FOXA1*, Guanine Nucleotide binding protein (G-Protein) alpha 12 (*GNA12*), Insulin-like Growth Factor 1 (*IGF1*), *IGF1R*, Mitogen Activated Protein Kinase 14 (*MAPK14*), *MYOD1*, *PGR*, *RARA*, *TFF1*, Tripartite motif containing 24 (*TRIM24*), *VAV3* and *XBP1*.

5.5 Discussion

Aiming to identify predictive probes responsible for *ESR1* and *ESR2* classification and to investigate the cross-talk between the predictive probes, we have used ANN-based predictive modelling and network inferencing approaches.

ERα (*ESR1*) and ERβ (*ESR2*) are components of the ER system, both playing a role in mammary development, female reproduction and in the breast cancer progression system. Investigation of each and every component of the system is necessary for a complete understanding and functionality of the system as a whole. The functionality of *ESR1* and *ESR2* could depend on the collective functionality of the interacting molecules within it. In an attempt to investigate the interacting molecules, here computational determination of the interacting

molecules is achieved by considering the predictive markers for the respective components of the ER system.

ESR1*(ERα) *interactome:

Among the 7 genes () found overlapping in the *ESR1* interactome with the genes cumulated from various interaction datasets, all were found to be directly or indirectly related in breast cancer. This proves the ANN's predictability with respect to *ESR1*. Supporting this, EGFR upon transactivated by estradiol, is found to be negatively associated with ER expression in breast cancer (Levin, 2003). As a result of bidirectional cross-talk between ER and *EGFR*, in breast cancer patients with antiestrogen resistance, the loss of *ESR1* expression is linked to overexpression of *EGFR* (Hoffmann and Sommer, 2005). The FOXA1 transcription factor is a member of the ER signalling pathway and is found to be strongly correlated with the luminal subtype of breast cancer and prognosis (Habashy et al., 2008). *FOXA1* is also associated with ESR1 expression regulation and ductal morphogenesis in mammary cells (Bernardo et al., 2010). Upon activation of the IGF1R pathway by ERα, *IGF1R* overexpression is found to be associated with ESR1 in primary breast cancer and lymph node metastasis (Koda et al., 2003). *RARA* and ERα shows a feed-back mechanism wherein, RARA competitively binds to the ERE region affecting the transactivation of ERα (Song et al., 1998), and as a result, ERα is found to induce overexpression of *RARA* in MCF-7 breast cancer cell line (Laganiere et al., 2005), *TFF1* expression is induced by estradiol and plays a transcriptional role in breast cancer (Sun et al., 2005). Chen et al. in their study have found *TFF1* as one among the other 20

differentially expressed ER-related genes in breast cancer samples through microarray (Chen et al., 2011). VAV3 is a guanine nucleotide exchange factor protein found to be an oncogene, and its overexpression aberrantly enhances ERα activity, thus playing a role in breast cancer development and progression (Lee et al., 2008). XBP1 being a member of stress signalling pathway, is involved in the impairment of cell cycle regulation associated with breast carcinogenesis (Andres and Wittliff, 2011).

ESR2* (ERβ) *interactome:

ESR2 interactome validation showed one gene overlapping i.e., *MYOD1*. MYOD1 encodes a nuclear protein belonging to the helix-loop-helix family of transcription factors, responsible for myogenic initiation and regulation of muscle cell differentiation and regeneration (Maguire et al., 2012). MYOD1 is found to exhibit anti-proliferative effect in MCF-7 breast cancer cell lines (Chen et al., 2010). Furthermore, MYOD1 along with ERβ is found to attenuate ERα-mediated BRCA2 expression (Jin et al., 2008a).

The ESR2 interactome of the ER system is computationally distinct from ERα. This is due to the fact that the predictive genes of *ESR2* are separate from *ESR1* and is amplified by the pre-selection of only 100 highly predictive probes. A broader consideration of predictive probes may be beneficial for finding commonalities between the *ESR1* and *ESR2* subsets.

Although the *in silico* interaction database was useful in identifying interacting molecules with ESR1 and ESR2, the information was found to be

incomplete. For example; the GATA3 transcription factor was not found interacting with either ESR1 or ESR2. This is in spite of the fact that, the association of *GATA3* with *ESR1* (Voduc et al., 2008), and the association of *FOXA1* and *GATA3* in breast cancer is well documented (Albergaria et al., 2009). Moreover, the interaction information was often repeated under different physical or experimental setups. Furthermore, text mining approaches used to generate interaction information in databases may look at mere mentioning of molecules in text. On the other hand, in spite of lower overlapping of molecules, the ANN network inferencing model was successful in identifying biological relevant interactions.

ER system simulation:

Simulation of the ER system by combining *ESR1* and *ESR2* interacting molecules in an ANN-based network inference method was able to pull out some of the ER-associated genes missed by considering individual components. Even though the filtering of the top100 positive and negative interaction was applied, the interaction network was complex. Apart from the ESR1 interacting molecules such as RARA, EGFR, FOXA1, VAV3, TFF1, IGF1R, XBP1 and the ESR2 interacting molecule MYOD1 previously found, other molecules were also found to be associated with ER and/or breast cancer.

ESR1 encodes the ERα subunit and acts as a transcription factor for ER-associated genes. *PGR* encodes for the PR which plays a central role in establishment and maintenance of pregnancy (Gadkar-Sable et al., 2005). PR

being able to directly bind to DNA and regulate target gene transcription, acts as a biomarker in breast development, tumorigenesis, progression and clinical outcome in luminal and basal tumour subtype along with ER (Kimura et al., 1988). ERBB2, also known as Her2/neu is also another prognostic marker in breast cancer along with ER and PR.

AR is found to be directly associated with growth in breast cancer cell lines. Apart from being over-expressed in breast tumour cells, AR is found to be associated with good clinical prognosis in ER-positive post-menopausal women and acts as a secondary biomarker for clinical outcome (Hu et al., 2011). TRIM24 protein mediates transcriptional control of nuclear receptors including ER and RARA. TRIM24 also plays a role in overexpression of ligand-dependent transcription activation of AR and ESR1. TRIM24/TIF-1α is a negative regulator of transcription factors and its overexpression is associated with poor prognosis and worse survival in breast cancer (Chambon et al., 2011). IGF1 shares similar functionality to IGF1R, as discussed previously (Chapter 4, Section 4.6.1). GNA12 as a member of the G proteins, acting as a modulator or transducer in transmembrane signalling. Upon activation by ligand-receptor binding, these proteins initiates secondary messenger signalling cascades linked to cellular regulatory processes. Along with polycystin-1, GNA12 shares a binding region which is necessary for apoptosis regulation (Yu et al., 2010). MAPK14 also called as p38, encodes a protein of family MAP kinases which are involved in cellular processes such as proliferation, differentiation, and transcription regulation. MAPK14 have been found to be have a role in various cancers such as lung, colon

ovarian and even in breast. In breast cancer, cross-talk between ER, HER2 and p38 have proved to contribute to tamoxifen resistance (Gutierrez et al., 2005). BTF3 forming a complex with RNA polymerase IIB is responsible for transcription initiation and regulates transcription in pancreatic cancer (Kusumawidjaja et al., 2007). In breast tumour cells, BTF3 is found to interact with NH2-terminal of ESR1 to initiate transcription of ER-associated genes (Green et al., 2007).

Even though filtering of the top 200 (100 positive and 100 negative) interactions which is 0.11% of total 169390 interactions, the ANN-based network inference algorithm was able to pull out distinctive interactions emphasising the prominent interactions across all interacting molecules. Most importantly, the interaction network clustered around *ESR1* and *PGR* together with other genes such as *AR, EGFR, FOXA1, TFF1, XBP1, VAV3* and *RARA*, making them separated from *ERBB2* which is clinically valid. Other genes like *GNA12, MAPK14, BTF3, IGF1* and *TRIM24* would be involved in the basic functionality of breast cancer as MAPK14 is a transducer of many signalling pathways and TRIM24 is associated with transcription initiation.

5.6 Conclusion

A Stepwise-ANN method for predictive probesets of *ESR1* and *ESR2* were found to be biologically relevant with overlapping of markers from curated

databases. Construction of the interaction network using ANN-based network inferencing method was informative in determining the causal interaction and quantitative measurement of interactions computationally.

The ANN-based network inferencing method for *ESR1* was supported by the interactions of ESR1 with ER-associated expression of *FOXA1, VAV3* and *TFF1*.

The ER system pathway simulation using ANN-based network inference method was found to be biologically relevant by clustering interactions among the genes *PGR, AR, EGFR, FOXA1, VAV3, TFF1*. The systems biology approach to ER could be extended at the cost of complexity of the interactions.

Chapter 6

NON-REDUCTIONAL SYSTEMS BIOLOGY APPROACH TO ESR1 WITH BROADER CONSIDERATION AND RELATIONSHIP TO PGR AND ERBB2

Chapter abstract: Breast cancer with varied clinically responsive subtypes represents a complex system. The behaviour of the system is governed by the cross-talk between the components of the system such as ER, PR and Her2. Computationally, the identification of predictive markers of the component system would benefit understanding of the behaviour of the system. In this study, we have adopted ANN-based predictive modelling to identify predictive markers in a non-reductional approach and network inferencing method to investigate the cross-talk between markers. Subsequent interactomes are also constructed and later combined to represent a systems biology of breast cancer. Our results identified a closed relation between *ESR1* and *PGR* interactomes with sharing of predictive markers, suggesting them as links between the systems. The *ERBB2* system showed a discrete network among its predictive markers and was found to be a separate entity with very few connections with the *ESR1* system.

6.1 Introduction

6.1.1 ER, PR and Her2

Breast cancer is a complex disease comprising of distinct biological subtypes with a diverse spectrum of molecular, pathological and clinical features exhibiting different prognostic and therapeutic implications. The former classification of breast cancer patients is solely based on ER expression and is less discriminatory in terms of prognosis. Furthermore, the independent prognostic and predictive role of Progesterone receptor (PR/*PGR*) expression has been a subject of controversy and it is well known that the PR gene is regulated by the oestrogen pathway (Clarke et al., 2005). Even though, further sub-classification of patients based on PR and Human epidermal growth factor receptor 2 (*ERBB2*/*ErbB-2/HER2/neu*) expression has provided enhanced discrimination, a complete characterisation of clinically beneficial breast cancer patients is ever expanding. Recently, researchers have focussed their efforts on identifying additional biomarkers for therapeutic guidance.

Current identification and investigation of the functionality of biomarkers is extensively based on conventional knock-down approaches wherein a single marker is investigated at a time. Correlation of a marker with respect to the known parameters is studied using linear methods. As the number of parameters increases, there is a considerable increase in the complexity of the system being studied and this makes it extremely difficult and challenging to analyse using the conventional reductionist approaches. A major drawback of current reductionist approach is the time taken to model such systems, as a single marker is investigated at a time.

Furthermore, as biological systems consists of various molecules acting at a single time instance, a lack of information of the cumulative effect and complex interaction among molecules over time widens the gap making interpretation of the whole system extremely unfeasible and complex.

6.1.2 Non-reductional systems biology approach

A **Non-reductional** or **additive** approach is about putting together rather than taking apart i.e., integration rather than reduction. In prior chapters, the addition of markers was predetermined or restricted to the markers of the study utilising only the top 100 associated to a given question. This poses a potential bias from a systems biology view. Alternatively, scanning of predictive markers associated with a given marker across the whole genome overcomes the bias of predetermined or restricted addition. This may benefit the information concerned in biological systems. Computationally, this non-reductionist approach is a serial addition of highly predictive markers to their primary marker. The subsequent construction of a pathway or network facilitates the study of a pathway with markers derived from the whole probeset.

6.2 Aims and objectives

- To identify predictive probes for median-based gene expression classification for *ESR1*, *PGR* and *ERBB2*.
- To apply a non-reductionist approach to *ESR1* interactome using ANN.

- To expand the approach to the *PGR* and *ERBB2* interactomes.
- To investigate the common markers across the *ESR1*, *PGR* and *ERBB2* interactomes.

6.3 Materials and methods

6.3.1 Predictive gene identification using stepwise-ANN

The E-GEOD-20194 microarray dataset, consisting of 22283 probes developed on HG-U133A Affymetrix array across 278 primary breast cancer, was chosen for the computational predictive gene identification.

The stepwise-ANN algorithm, mentioned earlier in Chapter 2 (*Section 2.7*), was used for identification of predictive genes for *ESR1*, *PGR* and *ERBB2*. Briefly, the classification of samples for training was based on the median-based gene expression levels. The algorithm was trained with 60% of total samples and tested on another 20%. A 20% fraction of sample was used for validation of prediction using a Monte Carlo Cross Validation strategy coupled in a stepwise-ANN with back-propagation of error algorithm. The prediction models were trained with a momentum of 0.5 and learning rate of 0.1 with 3000 epochs. The complete algorithm was iterated to generate 10 independent predictive models and predictiveness across matched probes was averaged for accurate prediction for each probe. The predictive rank was assigned based on the error generated by a test subset of samples for every probe.

6.3.2 Non-reductionist approach

A non-reductionist approach is a serial addition of markers to the primary markers, thereby increasing the complexity of the network. In this chapter, we have adopted gene-median expression based classification using a stepwise-ANN to identify predictive markers. The level of complexity is increased by scanning the entire genome for top 10 predictive markers every time i.e., at initial level, top 10 predictive markers are identified and later extended to the another level by identifying the top 10 additional markers for the earlier identified ones. A schematic representation of this non-reductional approach is depicted in Figure 25.

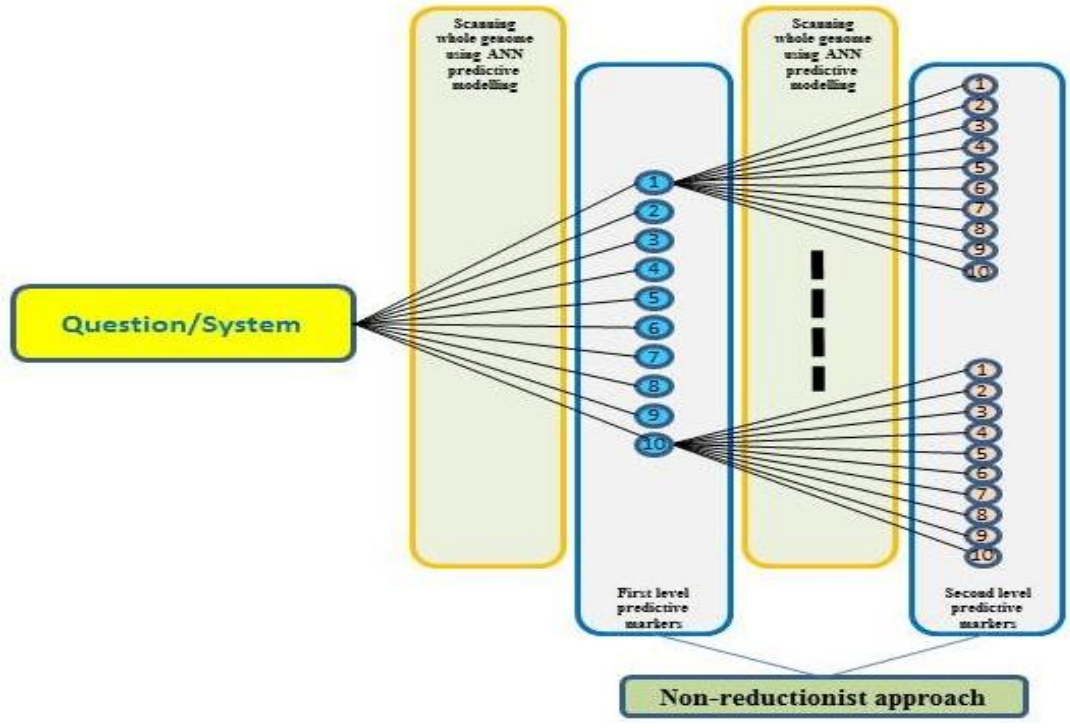


Figure 25: Schematic representation of non-reductional approach. Non-reductional approach initiate with identification of predictive markers for a particular question or system. Then serial addition of predictive markers for prior identified markers is carried out. There by exponentially increasing the number of markers identified from 10 to 100 in two step additive approach.

In this chapter, we have addressed serial additional of predictive markers for breast cancer components such as the *ESR1* (coding for ERα), *PGR* (coding for PgR) and *ERBB2* (coding for Her2) genes. A median expression based classification was used to identify the top 10 predictive genes for these systems separately. Later, the entire genome was scanned for predictive genes for the prior identified genes in a serial additive manner. This increases the level of complexity in an exponential form from 10 to 100 markers. Due to the exponential increase in the number of markers, here, for convenient and simple inferencing, two additive levels are carried out.

6.3.3 ANN interaction network inferencing

An ANN-based interaction network inferencing algorithm (discussed in Chapter 2 Section 2.8) was used to decipher the interconnections between the predictive genes. Briefly, the algorithm iteratively predicts every gene's expression as target by considering the remaining gene's expressions as inputs, until all the genes are considered as targets as well as providing possible interactions for that gene. At every step, the weights are recorded representing a relation the relationship between input and target. The algorithm is run for ten independent iterations and weights averaged. The weights are then scaled based in minimum and maximum values to generalise the weights across all genes.

The interaction models based on weights are created in Cytoscape Ver 2.7.3 software (www.cytoscape.org). Figure 26 is a schematic overview of non-reductional systems biology approach used in this chapter.

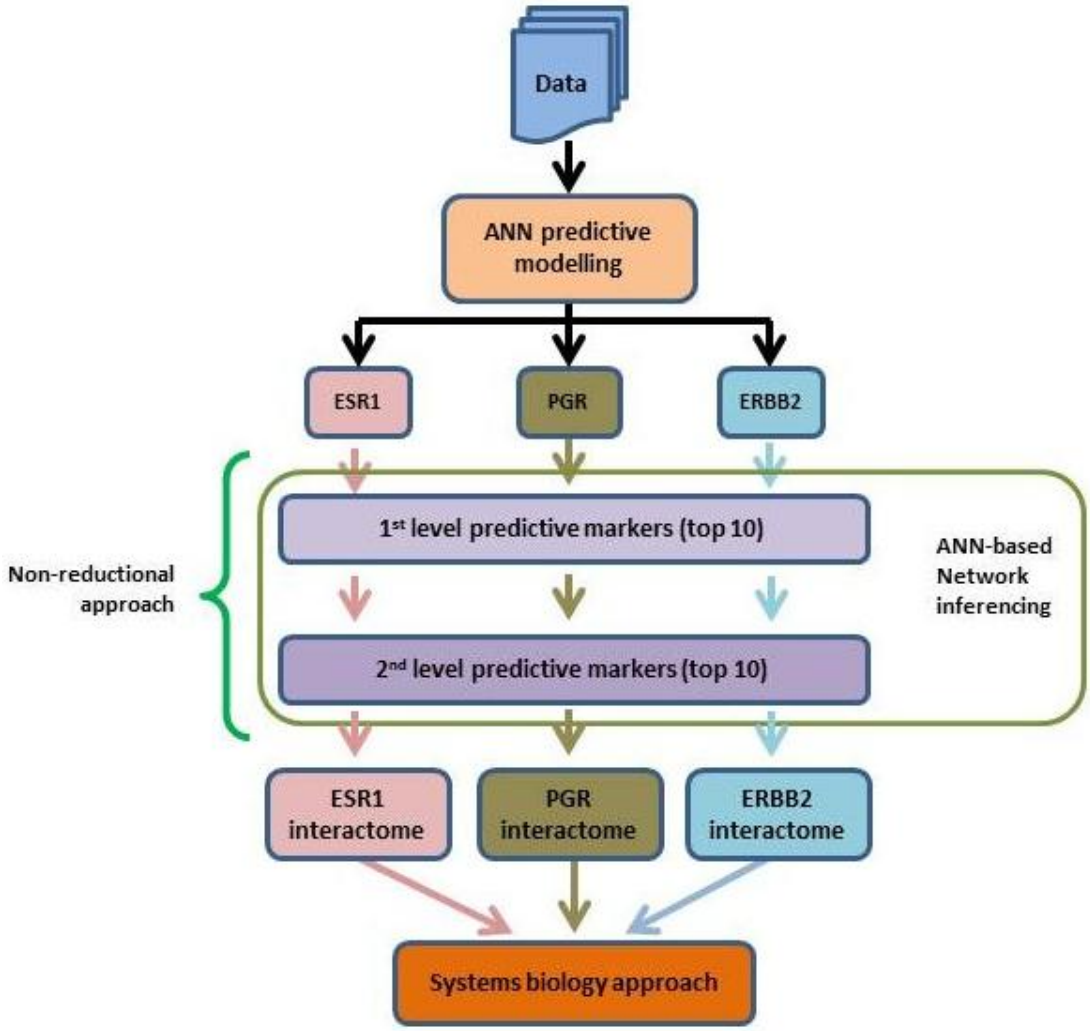


Figure 26: Overview of non-reductional systems biology approach. The identification of markers for *ESR1*, *PGR* and *ERBB2* is done separately using Stepwise-ANN predictive modelling method till second level of 100 marker identification (top 10 of top 10s). Later these markers are subjected to ANN-based network inferencing whereby construction of ESR1, PGR and ERBB2 interactomes is carried out to represent systems biology approach.

6.4 Results

6.4.1 Non-reductional ESR1 Interactome

Non-reductional interactome of *ESR1* (yellow node) was constructed taking the top ten most predictive genes for median-based gene expression classification of *ESR1*. Figure 27 is a *ESR1* non-reductional interactome. Blue nodes represent the first level of predictive genes whilst the associated pink nodes are the second level predictive genes for the level one marker.

Genes such as *GATA3, CA12, NAT1, TBC1D9* and others were found to be highly predictive for *ESR1* gene expression based classification. All of the first level predictive genes were found to be predictive for others, forming highly interconnected edges, suggesting a clustered and mutual interaction among them. Genes such as *XPBP1* and *LIV_1* were found to be predictive for more than two genes, representing epicentres of the network. The highest negative interaction was *TBC1D9* to *DNALI1* and the highest positive interaction was found to be between *GATA3* and *FOXA1*.

6.4.2 Extension to PGR and ERBB2 interactomes

A non-reductional interactome of *PGR* and *ERBB2* (yellow nodes) were also constructed taking the highly predictive ten genes from median-based gene expression classification of *PGR* and *ERBB2* respectively, represented as blue nodes. All pink nodes are the second level predictive genes of earlier genes.

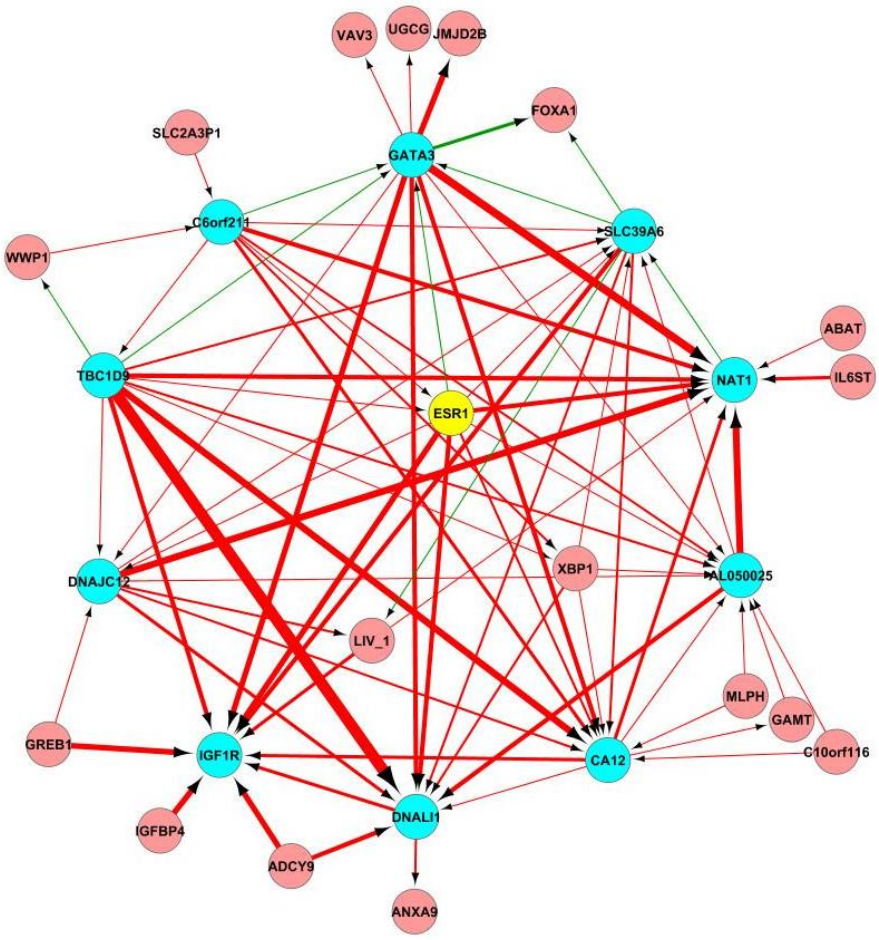


Figure 27: Non-reductional *ESR1* interaction model. The genes are represented as nodes and interactions as edges. The directionality of interactions is represented as arrows, from source to target. The positive and negative interactions are represented as Green and Red edges respectively. Blue nodes are the first level predictive genes of *ESR1* and act as primary genes for second level predictive genes.

In the *PGR* (yellow node) non-reductional interactome (Figure 28), genes such as *SCUBE2*, *MAPT* and *IL6ST* was found to be highly interactive with other predictive genes of *PGR*. When compared with *ESR1*, the PGR interactome was less compact suggesting a scattered network of genes. The highest negative interaction was found to be between *PTGER3* and *MYBL2* and highest positive interaction was from *PTGER3* to *STC2*.

In the *ERBB2* (yellow node) non-reductional interactome (Figure 29), the interactions were much more scattered and less clustered compared to the *ESR1* and *PGR* interactomes. Gene *GRP65* formed a distinct cluster of its predictive genes totally isolated from the others. The highest negative interaction was found to be between *PSMD3* and *PPARBP* and the highest positive interaction was between *ERBB2* and *SPDEF*.

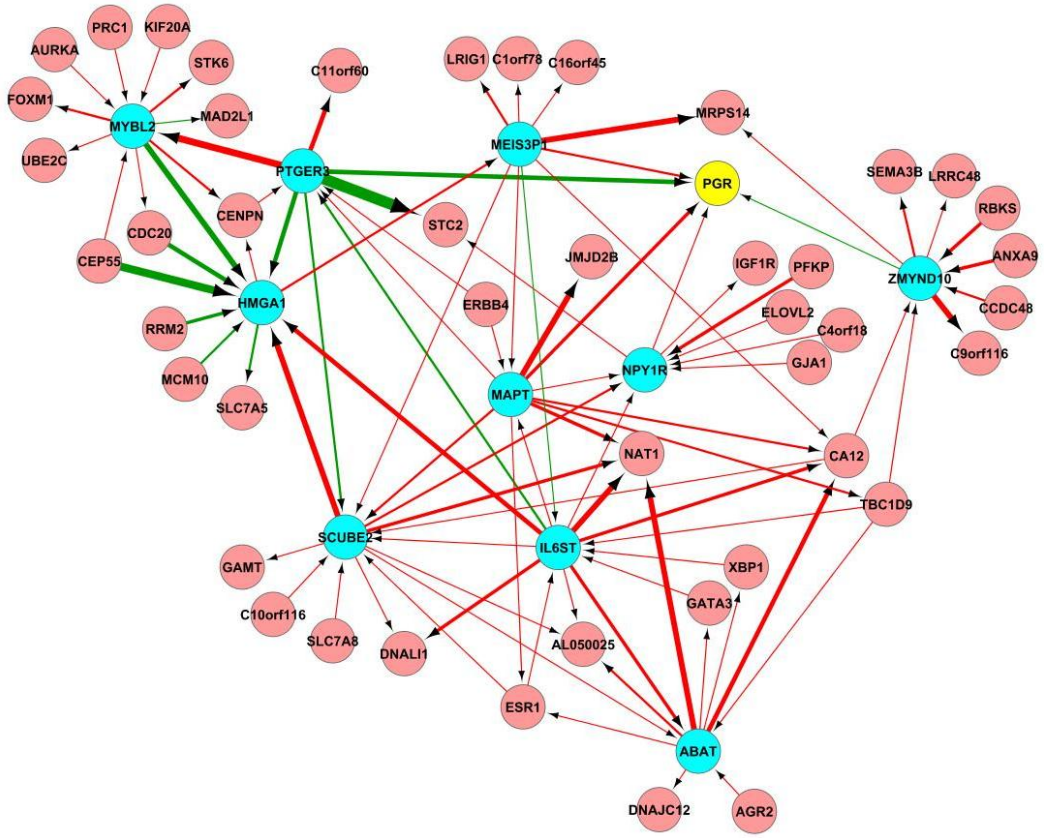


Figure 28: Non-reductional *PGR* interaction model. The genes are represented as nodes and interactions as edges. The directionality of interactions is represented as arrows, from source to target. The positive and negative interactions are represented as Green and Red edges respectively. Blue nodes are the first level predictive genes of *PGR* and act as primary genes for second level predictive genes.

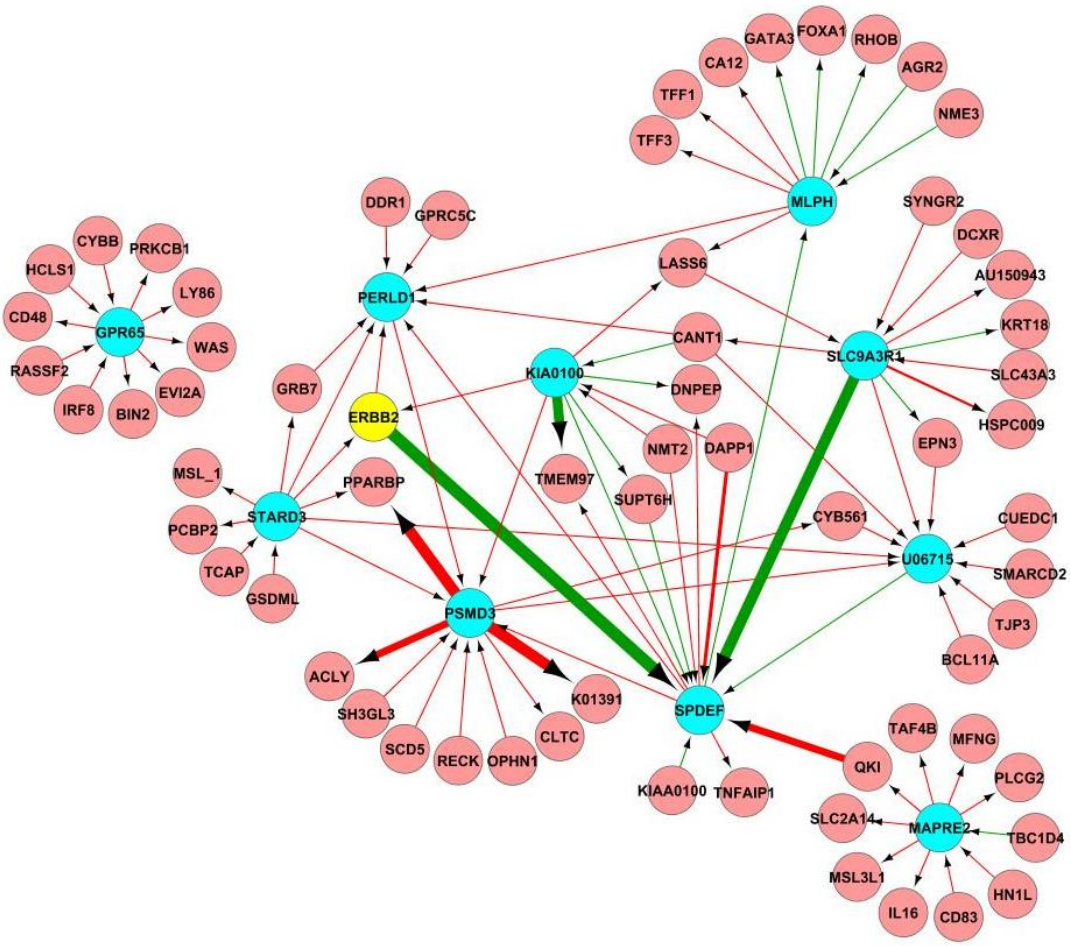


Figure 29: Non-reductional *ERBB2* interaction model. The genes are represented as nodes and interactions as edges. The directionality of interactions is represented as arrows, from source to target. The positive and negative interactions are represented as Green and Red edges respectively. Blue nodes are the first level predictive genes of *ERBB2* and act as primary genes for second level predictive genes.

6.4.3 Cumulative ESR1, PGR and ERBB2 non-reductional model

The cumulative interactomes of *ESR1*, *PGR* and *ERBB2* (yellow nodes) is shown in Figure 30. The first level predictive genes of *ERS1*, *PGR* and *ERBB2* are represented with green, orange and blue nodes respectively. The interactions are represented as edges with arrows indicating the directionality of interaction from

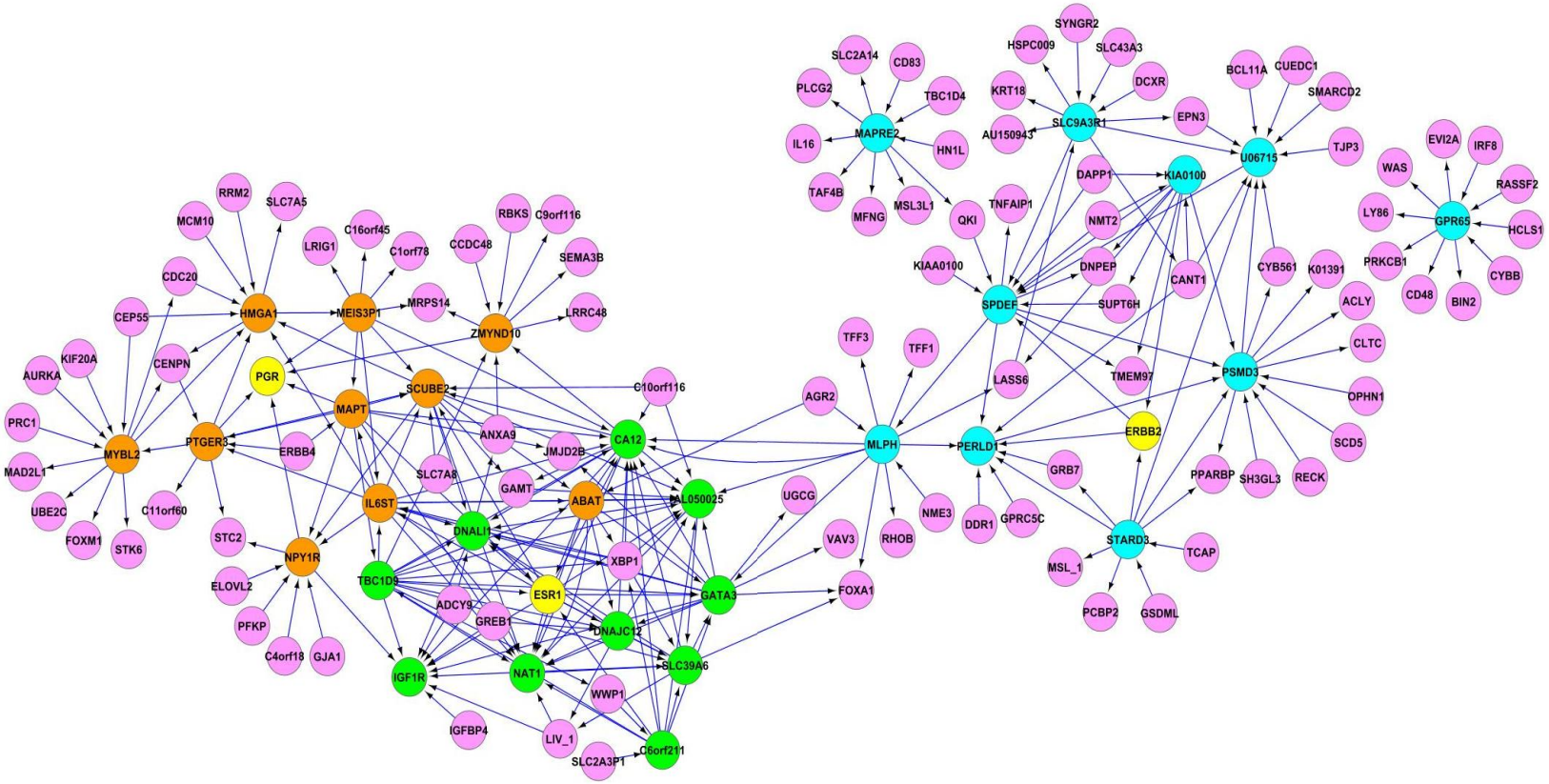


Figure 30: Cumulative *ESR1, PGR* and *ERBB2* non-reductional systems model. Non-reductional systems model comprising *ESR1*, *PGR* and *ERBB2* interactomes. Genes are represented as nodes and interactions as edges. Yellow nodes are the starting points of interactomes. First level predictive markers of *ESR1*, *PGR* and *ERBB2* are represented with green, orange and blue nodes respectively. All other secondary markers are represented with pink nodes. The directionality of interactions is represented with arrows from source to target.

source to target. A complete list of predictive genes for *ESR1*, *PGR* and *ERBB2* interactomes are listed in Table 19.

The genes *AGR2* and *FOXA1* were found to be the only linking genes between the *ESR1* and *ERBB2* interactomes. However, extensive bridging genes were found between the *ESR1* and *PGR* interactomes. The gene *GPR65*, a predictive gene for *ERBB2*, and its predictive genes remained as an isolated cluster from rest of the genes. It was promising to get *ESR1* and *PGR* interactomes highly connected sharing many interlinks, but the *ERBB2* interactome was clustered distinctly.

6.5 Discussion

This chapter describes to the identification of predictive markers for *ESR1*, *PGR* and *ERBB2*, using a non-reductional systems biology approach for these predictive markers in an serial additive manner.

ANNs were found to be efficient machine learning method for classification of samples and predictive gene selection. The application of this method for systems biology parse is established in this study. The construction of interactomes with an additive or non-reductionist approach for predictive markers using a stepwise-ANN is extensively investigated in this study. The ANN-based network inferencing approach is well established and documented (Dhondalay et al., 2011, Lemetre et al., 2009).

Table 19: Predictive gene list of *ESR1, PGR* and *ERBB2* non-reductional interactomes.

Predictive gene list of ESR1 non-reductional interactome

RANK	1	2	3	4	5	6	7	8	9	10
	CA12	**GATA3**	**TBC1D9**	**AL050025**	**NAT1**	**C6orf211**	**IGF1R**	**DNALI**	**DNAJC12**	**SLC39A6**
1	AL050025	ESR1	GATA3	CA12	CA12	ESR1	ESR1	GATA3	ESR1	LIV_1
2	GATA3	CA12	ESR1	GATA3	ESR1	GATA3	IGFBP4	CA12	TBC1D9	ESR1
3	ESR1	TBC1D9	CA12	ESR1	GATA3	TBC1D9	GATA3	ESR1	GATA3	NAT1
4	NAT1	AL050025	SLC39A6	TBC1D9	AL050025	XBP1	GREB1	TBC1D9	NAT1	TBC1D9
5	TBC1D9	DNALI1	NAT1	NAT1	TBC1D9	CA12	CA12	AL050025	SLC39A6	CA12
6	XBP1	NAT1	AL050025	XBP1	SLC39A6	SLC2A3P1	TBC1D9	XBP1	CA12	GATA3
7	GAMT	FOXA1	XBP1	DNALI	ABAT	SLC39A6	SLC39A6	ANXA9	AL050025	XBP1
8	DNALI1	JMJD2B	C6orf211	GAMT	DNALC12	NAT1	DNALI	ADCY	DNALI	FOXA1
9	MLPH	VAV3	WWP1	MLPH	LIV_1	AL050025	ADCY9	TBC1D9	GREB1	AL050025
10	C10orf116	UGCG	DNALI1	C10orf116	IL6ST	WWP1	LIV_1	SLC39A6	LIV_1	DNALC12

Predictive gene list of PGR non-reductional interactome

RANK	1	2	3	4	5	6	7	8	9	10
	PTGER3	**NPY1R**	**ABAT**	**HMGA1**	**ZMYND10**	**MEIS3P1**	**MAPT**	**MYBL2**	**SCUBE2**	**IL6ST**
1	SCUBE2	GJA1	ESR1	MYBL2	C9orf116	IL6ST	TBC1D9	CEP55	ESR1	ESR1
2	MAPT	ELOVL2	GATA3	CDC20	TBC1D9	MAPT	ESR1	UBE2C	NAT1	TBC1D9
3	PGR	SCUBE2	DNAJC12	SLC7A5	MRPS14	C1orf78	PGR	CDC20	IL6ST	CA12
4	STC2	PGR	TBC1D9	RRM2	LRRC48	CA12	IL6ST	STK6	GAMT	NAT1
5	MYBL2	IGF1R	CA12	SCUBE2	CA12	PGR	SCUBE2	CENPN	AL050025	GATA3
6	IL6ST	MAPT	NAT1	RRM2	CCDC48	C16orf45	CA12	MAD2L1	CA12	DNALI1
7	HMGA1	C4orf18	IL6ST	CENPN	RBKS	HMGA1	ERBB4	KIF20A	SLC7A8	XBP1
8	CENPN	STC2	XBP1	MCM10	PGR	SCUBE2	PTGER3	PRC1	C10orf116	ABAT
9	C11orf60	PFKP	AL050025	CEP55	SEMA3B	LRIG1	NAT1	AURKA	DNALI1	MAPT
10	ERBB4	IL6ST	AGR2	IL6ST	ANXA9	MRPS14	JMJD2B	FOXM1	ABAT	AL050025

Predictive gene list of ERBB2 non-reductional interactome

RANK	1	2	3	4	5	6	7	8	9	10
	MAPRE2	**GPR65**	**STRAD3**	**U06715**	**SLC9A3R1**	**MLPH**	**SPDEF**	**KIA0100**	**PERLD**	**PSMD3**
1	PLCG2	HCLS1	ERBB2	CYB561	SPDEF	LASS6	KIAA0100	ERBB2	PERLD1	CYB561
2	TAF4B	EVI2A	PERLD1	CANT1	CANT1	SPDEF	ERBB2	SPDEF	ERBB2	RECK
3	IL16	LY86	PSMD3	EPN3	KRT18	CA12	DAPP1	SUPT6H	STARD3	SCD5
4	TBC1D4	CYBB	GRB7	SLC3A3R1	SLC43A3	NME3	SUPT6H	PSMD3	MLPH	U06715
5	SLC2A14	RASSF2	U06715	TJP3	DCXR	FOXA1	PSMD3	DAPP1	GRB7	OPHN1
6	HN1L	IRF8	GSDML	SMARCD2	AU150943	AGR2	NMT2	CANT1	PSMD3	ACLY
7	QKI	CD48	PPARBP	CUEDC1	EPN3	RHO	TMEM97	DNPEP	SPDEF	K01391
8	MFNG	PRKCB1	TCAP	SPDEF	LASS6	TFF1	TNFAIP1	TMEM97	CANT1	SH3GL3
9	MSL3L1	WAS	MSL	BCL11A	HSPC009	TFF3	DNPEP	LASS6	DDR1	PPARBP
10	CD83	BIN2	PCBP2	OPTN	SYNGR2	GATA3	QKI	NMT2	DNPEP	CLTC

Predictive gene list of *ESR1*, *PGR* and *ERBB2* non-reductional interactomes. The primary predictive genes for each category are listed in bold grey highlighted boxes and the rank order of their predictive genes are listed underneath.

The primary predictive genes from the *ESR1* non-reductional interactome was considered promising due to the fact that the genes such as *GATA3*, *NAT1*, *CA12*, *IGF1R* had been reported to play role with ER and/or in breast cancer, as previously discussed in chapter 3 and 4. They were also found to be highly predictive for ESR1 classification. Additionally, *ESR1* appearing as predictive gene for all primary genes suggest a close interaction and association of *ESR1* with other predictive genes. The gene *DNALI1* (Dynein, axonemal, light intermediate chain 1) is an homolog of *Chlamydomonas* p28 gene and the actual function of it in human is unclear, but is found to be down regulated in breast carcinomas (Parris et al., 2010). The gene *SLC39A6*, coding for a zinc transporter protein is found to be associated in breast cancer and even found to be overexpressed in meta-analysis of microarray datasets (Srour et al., 2008). A positive interaction of *SLC39A6* with another zinc transporter, *LIV_1* was found to be supporting the functionality and role in breast cancer. Furthermore, the relation of *GATA3* with *FOXA1* was found to be positively associated with *ESR1*. The complete characterisation of gene *C6orf211* and *AL050025* mRNA clone sequence could lead to more information about their functionality in breast cancer.

The *PGR* non-reductional interactome was moderately discrete with few inter connections between the primary predictive genes. The *PGR* gene was found to be directly associated with *MEIS3P1*, *PTGER2*, *MAPT*, *NPY1R* and *ZMYND10* genes. MYBL2 (v-myb myeloblastosis viral oncogene homolog (avian)-like 2) is a transcription factor involved in cell cycle progression and is often amplified or over expressed in many tumours. *MYBL2* being associated with tumour

progression, is also found to be associated in basal-like breast tumour specific transcription regulation in mammary epithelial cells (Thorner et al., 2009). As a transcription factor, MYBL2 is found to regulate other breast cancer related genes such as *BCL2* and *TOP2A* (Shi et al., 2012). Lower expression of the gene *MAPT* (Microtubule-Associated Protein Tau) has been found to be associated with paclitaxel sensitivity in breast cancer samples (Rouzier et al., 2005). Recently, it has been found that MAPT protein expression is directly related with ER expression (Ikeda et al., 2010). Interleukin 6 signal transducer (*IL6ST*) gene is related to enhanced cellular proliferation in breast cancer and in combination with *BLC2* alters proliferation or imbalance in apoptosis leading to neoplastic tumour development (Garcia-Tunon et al., 2005). Promisingly in the *PGR* interactome, *ESR1* was found to be linked with *ABAT, SCUBE2, IL6ST* and *MAPT*.

The *ERBB2* non-reductional interactome was also discrete and scattered compared to *ESR1* and *PGR*. The primary predictive genes were less interconnected suggesting a sparse network of genes in the *ERBB2* cluster. Gene such as *GPR65* were totally isolated from rest of the network implicating distinct diverged interactions. The melanophilin (*MLPH*) gene is associated with bladder cancer, but in breast cancer the involvement with *ERBB2* is still unclear. The G protein-coupled receptor 65 (*GPR65*) belongs to the family of G proteins which are often overexpressed in tumoregenesis associated with apoptosis in ovarian cancers (Sin et al., 2004). The per1-like domain 1 (*PERLD1*) protein involved in lipid remodelling process, supporting our findings and is found to be associated with Her-2 over-expressive subclass of breast cancer indicating a poor prognosis

(Staaf et al., 2010). Furthermore, in gastric tumour cell lines, PERLD1 is found up regulated along with ERBB2 (Junnila et al., 2010). *STARD3* gene encodes StAR-related lipid transfer (START) domain containing 3 proteins related to lipid trafficking and steroidogenesis with an implication in hormone-receptor-positive breast tumours (Andrechek et al., 2003). Furthermore *STARD3* is localised with *ERBB2* on 17q12 in humans (Glynn et al., 2010). The *PSMD3* gene encodes Proteosome 26S subunit 3 protein is known to regulate cell cycle check points and may have role in cell progression and tumorogenesis. SPDEF protein is a transcription factor highly expressed in prostate epithelial cells. The over-expression of *SPDEF* is also found in breast, brain, lung and ovarian carcinomas. In breast cancer, high expression of *SPDEF* is correlated with poor overall survival of patients in ER+ tumours and also to luminal Her-2+ tumours (Sood, 2009). *MAPRE2* gene codes for Microtubule associated protein RP/EB family member 2 protein which is responsible for spindle symmetry and polymerisation during mitosis associated with cell division and migration. In pancreatic cancers, over-expression of *MAPRE2* is associated with reduced post-operative survival of patients (Abiatari et al., 2009).

Computational cumulative integration of non-reductional *ESR1, PGR* and *ERBB2* interactomes to represent systems biology was an attempt to mimic a biological breast cancer scenario. Promisingly *ESR1, PGR* and *ERBB2* formed distinct clusters with their predictive genes. Upon integration, parallel to clinical findings by other researchers, *ESR1* and *PGR* shared a complex network of interactions among their predictive genes. The genes such as *IGF1R, TBC1D9,*

IL6ST, DNALI1 and *SCUBE2* acted as bridging genes between the *ESR1* and *PGR* clusters indicating the shared functionality of these genes in the ER and PR systems. The *ERBB2* cluster was distinct and separated from the *ESR1* and *PGR* clusters indicating *ERBB2* as a separate entity. There were no genes found linking *ERBB2* and *PGR* clusters, but this is expected as they have different modalities. However, genes such as *AGR2, MLPH, GATA3* and *FOXA1* were found to link the *ESR1* and *ERBB2* clusters indicating that these genes might be having role in regulation of *ERBB2* expression by ER.

Consideration of a two level predictive gene non-reductional additive approach even though being limited, was successful in addressing patterns of interactions shared by common genes with common functionality. The increase in higher level additive approaches poses challenges for computational time consumption, network complexity and inferencing.

6.6 Conclusion

The non-reductional addition of predictive gene approach using ANN is promising in highlighting biologically relevant associations or interactions among the predictive genes. The construction of additive interactomes of *ESR1, PGR* and *ERBB2* were unique in their topology and interaction complexity. *ESR1* being highly complex, shared genes with the *PGR* predictive cluster more extensively than the *ERBB2* cluster. The *ERBB2* cluster being sparse and isolated from *ESR1* and *PGR* represented a biologically different entity.

The application of ANNs in systems biology is promising. Development of object-oriented algorithms will be beneficial for reducing computational time and will provide a standardised methodology for addressing hypothesis driven investigations.

Chapter 7

OVERALL SUMMARY AND FUTURE WORK

Breast cancer being a complex and heterogenous disease, shows a varied response to endocrine and hormonal treatment. The steroidal hormone, oestrogen plays a vital role in mammary tissue development and prognosis of breast cancer. The ER pathway remains unclear, giving an opportunity for the investigation of ER-associated markers. Although, prognostic indicators such as ER, PR and Her2 are well established and documented in breast cancer, efforts in identification of novel biomarkers and classification of favourable subgroups have enabled molecular subtyping of patients based on their gene expression profiles. Recent technological advances in high throughput techniques have facilitated the possibility of investigating gene expression profiles of the entire human genome. These techniques show promise, but also pose challenges for efficient analysis and interpretation of data with respect to their dimensionality, non-linearity and multivariate nature. Computational researchers have developed machine learning approaches to combat these challenges. One such machine learning approach which is gaining recognition in biosciences is Artificial Neural Network (ANN).

ANNs are a form of supervised machine learning methods motivated from physiological information processing by biological neurons. An in-house developed variant of ANN, a stepwise-ANN, has been successfully applied and validated in genomic data analysis (Lancashire et al., 2009) and network inferencing (Lemetre et al., 2009).

This thesis demonstrates the application of this ANN in a systems biology focussing on the ER systems to provide a novel non-reductional approach in systems biology of breast cancer. The studies follow the aims using the ANN;

- To classify breast cancer microarray samples for ER status.
- To identify predictive genes associated with ER status.
- To validate ER-associated markers in clinical breast tumour samples.
- To investigate cross-talk between ER-associated genes using a ANN-based network inference approach.
- To investigate predictive markers of the ER system (*ESR1* coding for ERα and *ESR2* coding for ERβ).
- To investigate cross-talk between ER systems associated markers using a ANN-based network inference approach.
- As novelty, to investigate predictive markers in a non-reductional approach in systems biology of breast cancer.

Chapter 3 demonstrated the successful application of stepwise-ANN model for ER status classification in breast cancer microarray dataset. The ANN showed a consistence classification performance in test and validation subsets of samples

with a mean of 90.45% and 89.45% respectively under ten independent runs. The classification accuracy of the ANN for ER status was found to be 87.72%. It was promising to find the *ESR1* gene (coding for ERα) as the best predictor for ER status in samples. The classification accuracy of ANN was improved to 90.97% when a gene's median expression based classification was performed for the ESR1 gene, with a sensitivity and specificity of 96.31% and 83.33% respectively. It was also promising to find breast cancer related markers in the high rank order of predictability. The highly predictive 100 markers were investigated for gene ontological information and found to be dominated with binding activity and metabolic process. Finally, our findings were also validated on the original study of dataset and on independent breast cancer datasets.

Chapter 4 demonstrates the application of the ANN in network inferencing to investigate cross-talk between markers. For this study, we have used 100 highly predictive markers of ER status derived from a stepwise-ANN method. The ANN network inference method identified interaction hubs specific for ER status and associated with breast cancer. Markers such as *DACH1, TFF3, RARA, EGFR, SERPINA5* in ER-positive samples and *VAV3, IL6ST, NME3, GAMT* in ER-negative samples. Among the identified interacting hubs, we validated the protein expression of VAV3 and DACH1 in clinical breast tumour samples from Nottingham Tenovous series. Our finding confirmed that VAV3 was not associated with clinicopathological markers, breast cancer specific markers and patient outcome. On the other hand, DACH1 was found highly associated with clinicopathological markers, breast cancer specific markers and clinical outcome

in patients with respect to tumour recurrence and metastasis formation in 5 year follow-up data.

Chapter 5 focussed on the systems biology of ER using the ANN network inference approach. To investigate cross-talk in the ER system, we considered *ESR1* and *ESR2* components. Initially the *ESR1* and *ESR2* interactomes were constructed using the top 100 predictive markers specific for *ESR1* and *ESR2* status in breast cancer microarray samples. Furthermore, in an attempt to model known contenders of the ER pathway, we identified *ESR1* and *ESR2* interacting molecules from *in silico* interaction databases. The interacting molecules were modelled using an ANN-based network inference method. Our finding clustered *ESR1* and *PGR* together along with other breast cancer biomarkers such as *AR*, *FOXA1*, *RARA*, *TFF1* and *EGFR*, confirming the biological relevance of our findings.

Finally in *Chapter 6*, a novel non-reductional systems biology approach was developed for *ESR1* and later extended to *PGR* and *ERBB2*. In this study, we have utilised predictive markers in a serial additive manner with increasing level of markers and complexity. In the first level, 10 predictive markers associated with the system were identified. Later, in a second level, 10 predictive markers associated with each of the first level markers were added. At every level, the predictive marker identification was achieved by considering the gene median expression based classification and scanning the entire genome array across all samples from dataset. Our finding showed more complex interactions among the

ESR1 predictive markers than *PGR* and *ERBB2* markers. Cumulative addition of *ESR1*, *PGR* and *ERBB2* systems to represent a complete biology of breast cancer resulted in overlap among the *ESR1* and *PGR* system markers, confirming the biological association between the systems. *ERBB2* showed no overlap with any other systems, suggesting a separate biological entity in breast cancer.

Overall, in this research work, the ANN model has been successfully applied to microarray dataset for the classification of samples based on ER status and identification of ER-associated markers. Protein expressions of identified ER-associated markers, VAV3 and DACH1, have been validated in clinical breast tumour tissue samples by immunohistochemistry. The ANN-based network inferencing method has been used to investigate the cross-talk in ER system. The systems biology approach to ER system is structured in consideration with its components, *ESR1* and *ESR2*, as interactomes. Furthermore, to investigate the cross-talk between multi-system involvement in breast cancer, as a novelty, the integration of ER, PR and Her2 systems have been achieved with respect to *ESR1*, *PGR* and *ERBB2* associated genes in a non-reductional systems biology approach.

Advantages:

The non-reductional systems biology approach described here bring novelties to existing approaches. One of the main advantages of ANNs in systems biology is their ability to efficiently infer cross-talks between the set of markers rather than single markers in a pairwise manner.

Additionally, a non-reductional systems biology approach, as contrary to conventional reductionist approach, gives the advantage of serial addition of markers without neglecting the influence of prior markers in an additive approach.

Furthermore, this novel approach, contrary to most alternatives, the potential to be applied to pathway decipherization using several types of data from different high throughput platforms of proteomics and metabolomics is possible.

Limitations:

Despite the biologically relevant findings, nonetheless, the approach may have limitations. Firstly, one of the central limitations may come from the source of data itself. Even though we restricted our study to whole genome arrays, the interactions can only be relative to all others within the dataset and none of the interactions can be complete, because of ever increasing panel of genes in array.

Another important limitation is the validation of interactions. In spite of using well documented multiple *in silico* interaction databases, some interactions of *ESR1* with *FOXA1* and *GATA3* were missing despite being well documented and associated with *ESR1* and breast cancer. Knowing the fact that only direct interactions were considered in the *in silico* databases considered here, the inference of indirect interactions may be considered as an advantage.

Additionally, preselecting only 100 predictive gene set may impose a restricted investigation of contenders involved, highlighting the limitation of the method, but not the approach. On the other hand, increasing the gene set number

will proportionally increase the complexity of networks of interactomes and may introduce difficulty in inferencing.

Furthermore, the gene expression levels considered as output in network inferencing approach using ANN is modelled for over or under the median expression level across whole population of samples i.e., the inputs are used to predict the output in a binary fashion at an expense of losing information during conversion of continuous data points into discrete output.

FUTURE WORK:

- As the selection of 100 predictive genes was one among the limitations of network inferencing approach, it would be interesting to utilise other statistical methods such as principle component analysis to determine the number of significant predictive genes to be used with removal of outliers. Adding on to this, calculating the percentage of population explained by each gene will also be interesting.
- Development of network topology featuring distance between the nodes is another interesting aspect of network inference. The inherent errors encountered during prediction of output from the inputs could be exploited to localise the nodes correspondingly i.e., the output nodes with higher errors could be placed at further distance from source nodes than nodes with lower errors.

- Development of an algorithm to calculate 'interaction coefficient' of nodes which could be the product of the number of interactions of a node and the cumulative errors encountered across all interactions from the same node. This could help in identifying and selection of the most significant node with highest interaction coefficient.
- Further validation is required in order to confirm the findings. Especially, the *TONDU* (or *VGLL1*) identified as a negative interaction hub in ER-negative samples through ANN-network inferencing approach (*Chapter 4*), requires validation in clinical breast tumour samples with the expectation of serving as a biomarker in triple negative (adverse clinical responsive) patients.
- Finally, incorporation of protein-protein interaction information may provide more meaningful network inferencing.

Another aspect for expansion in future work could be in the direction of ER-beta subunit of ERs. Recently, ER-beta 1 and ER-beta 2 isoforms have been found to be clinically important and prognostic indicators in breast cancer studies (Speirs, 2008, Speirs et al., 2008, Speirs and Shaaban, 2009, Honma et al., 2008, Novelli et al., 2008). The investigation of cross-talk between ER-beta-1 and ER-beta-2 using ANN-based network inferencing methods could help in characterising the role of ER-beta in breast cancer.

Additionally, as this entire research work was concentrated towards female breast cancer, investigations in rare and important male breast cancer is also feasible. As indicated by White et. al. that male breast cancer requires more

awareness and information (White et al., 2011), the application of ANN based classification and network inferencing methods might be helpful in understanding the disease.

Furthermore, using other neural network models such as recurrent neural networks and Kohonen self-organising neural networks will be interesting. Furthermore, the application of ANN-based network inferencing approach to other cancers and disease is believed to have potential in identifying biomarkers and contenders of pathway.

BIBLIOGRAPHY

www.cancerresearchuk.org [Online].
www.cytoscape.org [Online].
www.ebi.ac.uk/arrayexpress [Online].
www.genenames.org [Online].
ABBASS, H. A. 2002. An evolutionary artificial neural networks approach for breast cancer diagnosis. *Artif Intell Med. ,* 25**,** 265-81.
ABIATARI, I., GILLEN, S., DEOLIVEIRA, T., KLOSE, T., BO, K., GIESE, N. A., FRIESS, H. & KLEEFF, J. 2009. The microtubule-associated protein MAPRE2 is involved in perineural invasion of pancreatic cancer cells. *Int J Oncol,* 35**,** 1111-6.
AGATONOVIC-KUSTRIN, S. & BERESFORD, R. 2000. Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. *J Pharm Biomed Anal. ,* 22**,** 717-27.
AHR, A., KARN, T., SOLBACH, C., SEITER, T., STREBHARDT, K., HOLTRICH, U. & KAUFMANN, M. 2002. Identification of high risk breast-cancer patients by gene expression profiling. *Lancet,* 359**,** 131-2.
ALBERGARIA, A., PAREDES, J., SOUSA, B., MILANEZI, F., CARNEIRO, V., BASTOS, J., COSTA, S., VIEIRA, D., LOPES, N., LAM, E. W., LUNET, N. & SCHMITT, F. 2009. Expression of FOXA1 and GATA-3 in breast cancer: the prognostic significance in hormone receptor-negative tumours. *Breast Cancer Res,* 11**,** R40.
ALBERTS, B., JOHNSON, A., LEWIS, J., RAFF, M., ROBERTS, K. & WALTER, P. 2002. *Molecular Biology of the Cell,* New York, Garland Science.
ALMEIDA, J. S. 2002. Predictive non-linear modeling of complex data by artificial neural networks. *Current Options in Biotechnology,* 13**,** 72-76.
ANDERSON, E. 2002. The role of oestrogen and progesterone receptors in human mammary development and tumorigenesis. *Breast Cancer Res,* 4**,** 197-201.
ANDERSON, J. A. 1995. *An Introduction to Neural Networks*, MIT Press.
ANDRE, F. & PUSZTAI, L. 2006. Molecular classification of breast cancer: implications for selection of adjuvant chemotherapy. *Nat Clin Pract Oncol,* 3**,** 621-32.
ANDRECHEK, E. R., LAING, M. A., GIRGIS-GABARDO, A. A., SIEGEL, P. M., CARDIFF, R. D. & MULLER, W. J. 2003. Gene expression profiling of neu-induced mammary tumors from transgenic mice reveals genetic and morphological similarities to ErbB2-expressing human breast cancers. *Cancer Res,* 63**,** 4920-6.
ANDRES, S. A. & WITTLIFF, J. L. 2011. Relationships of ESR1 and XBP1 expression in human breast carcinoma and stromal cells isolated by laser capture microdissection compared to intact breast cancer tissue. *Endocrine,* 40**,** 212-21.
ARKIN, M. R., RANDAL, M., DELANO, W. L., HYDE, J., LUONG, T. N., OSLOB, J. D., RAPHAEL, D. R., TAYLOR, L., WANG, J., MCDOWELL, R. S., WELLS, J. A. & BRAISTED, A. C. 1997. Binding of small molecules to an adaptive protein-protein interface. *Proc Natl Acad Sci U S A,* 100**,** 1603-8.
ARNOULD, L., ROGER, P., MACGROGAN, G., CHENARD, M. P., BALATON, A., BEAUCLAIR, S. & PENAULT-LLORCA, F. 2012. Accuracy of HER2 status determination on breast core-needle biopsies (immunohistochemistry, FISH, CISH and SISH vs FISH). *Mod Pathol,* 25**,** 675-82.

ASLAN, K., BOZDEMIR, H., SAHIN, C., OGULATA, S. N. & EROL, R. 2008. A radial basis function neural network model for classification of epilepsy using EEG signals. *J Med Syst. ,* 32**,** 403-8.

BADVE, S., TURBIN, D., THORAT, M. A., MORIMIYA, A., NIELSEN, T. O., PEROU, C. M., DUNN, S., HUNTSMAN, D. G. & NAKSHATRI, H. 2007. FOXA1 expression in breast cancer--correlation with luminal subtype A and survival. *Clin Cancer Res,* 13**,** 4415-21.

BALL, G., MIAN, S., HOLDING, F., ALLIBONE, R. O., LOWE, J., ALI, S., LI, G., MCCARDLE, S., ELLIS, I. O., CREASER, C. & REES, R. C. 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**,** 395-404.

BARABASI, A.-L. & OLTVAI, Z. N. 2004. Network biology: understanding the cell's functional organization. *Nat Rev Genet,* 5**,** 101-113.

BASHEER, I. A. & HAJMEER, M. 2000. Artificial neural networks: fundamentals, computing, design, and application. *J Microbiol Methods,* 43**,** 3-31.

BASKAR, S. S., AROCKIAM, L., ARUL KUMAR, V. & JEYASIMMAN, L. 2010. Brief Survey of Application of Data Mining Techniques to Agriculture. *Agricultural Journal,* 5**,** 116-118.

BASSO, K., MARGOLIN, A. A., STOLOVITZKY, G., KLEIN, U., DALLA-FAVERA, R. & CALIFANO, A. 2005. Reverse engineering of regulatory networks in human B cells. *Nat Genet,* 37**,** 382-90.

BEAL, M. J., FALCIANI, F., GHAHRAMANI, Z., RANGEL, C. & WILD, D. L. 2005. A Bayesian approach to reconstructing genetic regulatory networks with hidden factors. *Bioinformatics,* 21**,** 349-56.

BEBIS, G. & GEORGIOPOULOS, M. 1994. Feed-forward neural networks. *Potentials, IEEE,* 13**,** 27-31.

BECK, H. C., NIELSEN, E. C., MATTHIESEN, R., JENSEN, L. H., SEHESTED, M., FINN, P., GRAUSLUND, M., HANSEN, A. M. & JENSEN, O. N. 2006. Quantitative proteomic analysis of post-translational modifications of human histones. *Mol Cell Proteomics,* 5**,** 1314-25.

BERNARDO, G. M., LOZADA, K. L., MIEDLER, J. D., HARBURG, G., HEWITT, S. C., MOSLEY, J. D., GODWIN, A. K., KORACH, K. S., VISVADER, J. E., KAESTNER, K. H., ABDUL-KARIM, F. W., MONTANO, M. M. & KERI, R. A. 2010. FOXA1 is an essential determinant of ERalpha expression and mammary ductal morphogenesis. *Development,* 137**,** 2045-54.

BISHOP, C. M. & HINTON, G. 1995. *Neural Networks for Pattern Recognition*, Clarendon Press.

BOCCHINFUSO, W. P. & KORACH, K. S. 1997. Mammary gland development and tumorigenesis in estrogen receptor knockout mice. *J Mammary Gland Biol Neoplasia.,* 2**,** 323-34.

BOSL, W. J. 2007. Systems biology by the rules: hybrid intelligent systems for pathway modeling and discovery. *BMC Syst Biol,* 1**,** 13.

BREITKREUTZ, B. J., STARK, C. & TYERS, M. 2003. The GRID: the General Repository for Interaction Datasets. *Genome Biol,* 4**,** :R23.

BURRASCANO, P. 1991. Learning vector quantization for the probabilistic neural network. *Neural Networks, IEEE Transactions on,* 2**,** 458-461.

BUSTELO, X. R. 2001. Vav proteins, adaptors and cell signaling. *Oncogene,* 20**,** 6372-81.

BUYSE, M., LOI, S., VAN'T VEER, L. J., VIALE, G., DELORENZI, M., GLAS, A. M., D'ASSIGNIES, M. S., BERGH, J., LIDEREAU, R., ELLIS, P., HARRIS, A., BOGAERTS, J., THERASSE, P., FLOORE, A., AMAKRANE, M., PIETTE, F., RUTGERS, E., SOTIRIOU, C., CARDOSO, F. & PICCART, M. J. 2006. Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer. *J Natl Cancer Inst,* 98**,** 1183-92.

BUZA-VIDAS, N., DUARTE, S., LUC, S., BOURIEZ-JONES, T., WOLL, P. S. & JACOBSEN, S. E. 2011. GATA3 is redundant for maintenance and self-renewal of hematopoietic stem cells. *Blood,* 118**,** 1291-3.

CALDAS, C. & APARICIO, S. A. 2002. The molecular outlook. *Nature,* 415**,** 484-5.

CALZA, S., HALL, P., AUER, G., BJOHLE, J., KLAAR, S., KRONENWETT, U., LIU, E. T., MILLER, L., PLONER, A., SMEDS, J., BERGH, J. & PAWITAN, Y. 2006. Intrinsic molecular signature of breast cancer in a population-based cohort of 412 patients. *Breast Cancer Res Treat,* 8**,** R34.

CHAKRAVARTY, D., TEKMAL, R. R. & VADLAMUDI, R. K. 2010. PELP1: A novel therapeutic target for hormonal cancers. *IUBMB Life,* 62**,** 162-9.

CHAMBON, M., ORSETTI, B., BERTHE, M. L., BASCOUL-MOLLEVI, C., RODRIGUEZ, C., DUONG, V., GLEIZES, M., THENOT, S., BIBEAU, F., THEILLET, C. & CAVAILLES, V. 2011. Prognostic significance of TRIM24/TIF-1alpha gene expression in breast cancer. *Am J Pathol,* 178**,** 1461-9.

CHATR-ARYAMONTRI, A., CEOL, A., PALAZZI, L. M., NARDELLI, G., SCHNEIDER, M. V., CASTAGNOLI, L. & CESARENI, G. 2007. MINT: the Molecular INTeraction database. *Nucleic Acids Res,* 35(Database issue)**,** D572-4.

CHEANG, M. C., CHIA, S. K., VODUC, D., GAO, D., LEUNG, S., SNIDER, J., WATSON, M., DAVIES, S., BERNARD, P. S., PARKER, J. S., PEROU, C. M., ELLIS, M. J. & NIELSEN, T. O. 2009. Ki67 index, HER2 status, and prognosis of patients with luminal B breast cancer. *J Natl Cancer Inst. ,* 101**,** 736-50.

CHEN, C. H., KUO, S. C., HUANG, L. J., HSU, M. H. & LUNG, F. D. 2010. Affinity of synthetic peptide fragments of MyoD for Id1 protein and their biological effects in several cancer cells. *J Pept Sci,* 16**,** 231-41.

CHEN, K. C., WANG, T. Y., TSENG, H. H., HUANG, C. Y. & KAO, C. Y. 2005. A stochastic differential equation model for quantifying transcriptional regulatory network in Saccharomyces cerevisiae. *Bioinformatics,* 21**,** 2883-90.

CHEN, Y., CHEN, C., YANG, B., XU, Q., WU, F., LIU, F., YE, X., MENG, X., MOUGIN, B., LIU, G., SHEN, Z., SHAO, Z. & WU, J. 2011. Estrogen receptor-related genes as an important panel of predictors for breast cancer response to neoadjuvant chemotherapy. *Cancer Lett,* 302**,** 63-8.

CHI, C.-L., STREET, W. N. & WOLBERG, W. H. Application of Artificial Neural Network-Based Survival Analysis on Two Breast Cancer Datasets. AMIA, 2007. 130-134.

CHIDANANDA, G. K. & RAVI, T. V. 1995. Divisive clustering of symbolic objects using the concepts of both similarity and dissimilarity. *Pattern Recognition,* 28**,** 1277-1282.

CHIN, K., DEVRIES, S., FRIDLYAND, J., SPELLMAN, P. T., ROYDASGUPTA, R., KUO, W. L., LAPUK, A., NEVE, R. M., QIAN, Z., RYDER, T., CHEN, F., FEILER, H., TOKUYASU, T., KINGSLEY, C., DAIRKEE, S., MENG, Z., CHEW, K., PINKEL, D., JAIN, A., LJUNG, B. M., ESSERMAN, L., ALBERTSON, D. G., WALDMAN, F. M. & GRAY, J. W. 2006.

Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. *Cancer Cell,* 10**,** 529-41.

CHO, C. R., LABOW, M., REINHARDT, M., VAN OOSTRUM, J. & PEITSCH, M. C. 2006. The application of systems biology to drug discovery. *Curr Opin Chem Biol,* 10**,** 294-302.

CHUANG, H. Y., LEE, E., LIU, Y. T., LEE, D. & IDEKER, T. 2007. Network-based classification of breast cancer metastasis. *Mol Syst Biol,* 3**,** 140.

CHURCHLAND, P. & SEJNOWSKI, T. J. 1994. *The Computational Brain*, MIT Press.

CIMINO, D., FUSO, L., SFILIGOI, C., BIGLIA, N., PONZONE, R., MAGGIOROTTO, F., RUSSO, G., CICATIELLO, L., WEISZ, A., TAVERNA, D., SISMONDI, P. & DE BORTOLI, M. 2008. Identification of new genes associated with breast cancer progression by gene expression analysis of predefined sets of neoplastic tissues. *Int J Cancer,* 123**,** 1327-38.

CLARKE, R. B., SPENCE, K., ANDERSON, E., HOWELL, A., OKANO, H. & POTTEN, C. S. 2005. A putative human breast stem cell population is enriched for steroid receptor-positive cells. *Dev Biol,* 277**,** 443-56.

COOPER, C. S. 2001. Applications of microarray technology in breast cancer research. *Breast Cancer Res,* 3**,** 158-75.

COPPEDE, F., GROSSI, E., MIGHELI, F. & MIGLIORE, L. 2010. Polymorphisms in folate-metabolizing genes, chromosome damage, and risk of Down syndrome in Italian women: identification of key factors using artificial neural networks. *BMC Med Genomics. ,* 3**,** 42.

CORDERA, F. & JORDAN, V. C. 2006. Steroid receptors and their role in the biology and control of breast cancer growth. *Semin Oncol.,* 33**,** 631-41.

COUSE, J. F., LINDZEY, J., GARDIEN, K., GUSTAFSSON, J. A. & KORACH, K. S. 1997. Tissue Distribution and Quantitative Analysis of Estrogen Receptor- alplha and Estrogen Receptor- beta Messenger Ribonucleic Acid in the Wild-Type and ER -Knockout Mouse. *Endocrinology,* 138**,** 4613-4621.

CRACIUN, G., KIM, J., PANTEA, C. & REMPALA, G. A. 2013. Statistical Model for Biochemical Network Inference. *Commun Stat Simul Comput,* 42**,** 121-137.

CRUZ, J. A. & WISHART, D. S. 2007. Applications of machine learning in cancer prediction and prognosis. *Cancer Inform.,* 2**,** 59-77.

CUMMINGS, S. R., ECKERT, S., KRUEGER, K. A., GRADY, D., POWLES, T. J., CAULEY, J. A., NORTON, L., NICKELSEN, T., BJARNASON, N. H., MORROW, M., LIPPMAN, M. E., BLACK, D., GLUSMAN, J. E., COSTA, A. & JORDAN, V. C. 1999. The effect of raloxifene on risk of breast cancer in postmenopausal women: results from the MORE randomized trial. Multiple Outcomes of Raloxifene Evaluation. *JAMA,* 281**,** 2189-97.

D'HAESELEER, P., WEN, X., FUHRMAN, S. & SOMOGYI, R. 1999. Linear modeling of mRNA expression levels during CNS development and injury. *Pac Symp Biocomput***,** 41-52.

DE JONG, H. 2002. Modeling and simulation of genetic regulatory systems: a literature review. *J Comput Biol,* 9**,** 67-103.

DEANGELIS, T., WU, K., PESTELL, R. & BASERGA, R. 2011. The type 1 insulin-like growth factor receptor and resistance to DACH1. *Cell Cycle,* 10**,** 1956-1959.

DEROO, B. J. & KORACH, K. S. 2006. Estrogen receptors and human disease. *J Clin Invest,* 116**,** 561-70.

DESMEDT, C., PIETTE, F., LOI, S., WANG, Y., LALLEMAND, F., HAIBE-KAINS, B., VIALE, G., DELORENZI, M., ZHANG, Y., D'ASSIGNIES, M. S., BERGH, J., LIDEREAU, R., ELLIS, P., HARRIS, A. L., KLIJN, J. G., FOEKENS, J. A., CARDOSO, F., PICCART, M. J., BUYSE, M. & SOTIRIOU, C. 2007. Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the TRANSBIG multicenter independent validation series. *Clin Cancer Res.,* 13**,** 3207-14.

DHILLON, V. S. & DHILLON, I. K. 1995. Genotoxicity evaluation of estradiol. *Mutation Research,* 345**,** 87-95.

DHONDALAY, G. K., LEMETRE, C. & BALL, G. R. Modeling estrogen receptor pathways in breast cancer using an Artificial Neural Networks based inference approach. Biomedical and Health Informatics (BHI), 2012 IEEE-EMBS International Conference on, 5-7 Jan. 2012 2012. 948-951.

DHONDALAY, G. K., TONG, D. L. & BALL, G. R. Estrogen receptor status prediction for breast cancer using artificial neural network. Machine Learning and Cybernetics (ICMLC), 2011 International Conference on, 10-13 July 2011 2011. 727-731.

DOANE, A. S., DANSO, M., LAL, P., DONATON, M., ZHANG, L., HUDIS, C. & GERALD, W. L. 2006. An estrogen receptor-negative breast cancer subset characterized by a hormonally regulated transcriptional program and response to androgen. *Oncogene,* 25**,** 3994-4008.

DONG, Z., LIU, Y., LU, S., WANG, A., LEE, K., WANG, L. H., REVELO, M. & LU, S. 2006. Vav3 oncogene is overexpressed and regulates cell growth and androgen receptor activity in human prostate cancer. *Mol Endocrinol,* 20**,** 2315-25.

DOTZLAW, H., LEYGUE, E., WATSON, P. H. & MURPHY, L. C. 1997. Expression of estrogen receptor-beta in human breast tumors. *J Clin Endocrinol Metab.,* 82**,** 2371-4.

DOUGLAS, J. B., SILVERMAN, D. T., POLLAK, M. N., TAO, Y., SOLIMAN, A. S. & STOLZENBERG-SOLOMON, R. Z. 2010. Serum IGF-I, IGF-II, IGFBP-3, and IGF-I/IGFBP-3 molar ratio and risk of pancreatic cancer in the prostate, lung, colorectal, and ovarian cancer screening trial. *Cancer Epidemiol Biomarkers Prev,* 19**,** 2298-306.

DOWSETT, M. & DUNBIER, A. K. 2008. Emerging biomarkers and new understanding of traditional markers in personalized therapy for breast cancer. *Clin Cancer Res.,* 14**,** 8019-26.

DREISEITL, S., OHNO-MACHADO, L., KITTLER, H., VINTERBO, S., BILLHARDT, H. & BINDER, M. 2001. A comparison of machine learning methods for the diagnosis of pigmented skin lesions. *J Biomed Inform.,* 34**,** 28-36.

DUFFY, M. J. 2005. Predictive markers in breast and other cancers: a review. *Clin Chem. ,* 51**,** 494-503.

DUFFY, M. J., O'DONOVAN, N. & CROWN, J. 2011. Use of molecular markers for predicting therapy response in cancer patients. *Cancer Treat Rev. ,* 37**,** 151-9.

DURAI, R., DAVIES, M., YANG, W., YANG, S. Y., SEIFALIAN, A., GOLDSPINK, G. & WINSLET, M. 2006. Biology of insulin-like growth factor binding protein-4 and its role in cancer (review). *Int J Oncol,* 28**,** 1317-25.

EBCTCG 2005. Effects of chemotherapy and hormonal therapy for early breast cancer on recurrence and 15-year survival: an overview of the randomised trials. *The Lancet,* 365**,** 1687-1717.

EECKHOUTE, J., KEETON, E. K., LUPIEN, M., KRUM, S. A., CARROLL, J. S. & BROWN, M. 2007. Positive cross-regulatory loop ties GATA-3 to estrogen receptor alpha expression in breast cancer. *Cancer Res,* 67**,** 6477-83.

ELSTON, C. W. 2005. Classification and grading of invasive breast carcinoma. *Verh Dtsch Ges Pathol.,* 89**,** 35-44.

ENMARK, E. & GUSTAFSSON, J. A. 1999. Oestrogen receptors - an overview. *J Intern Med,* 246**,** 133-8.

FABIAN, C. J. & KIMLER, B. F. 2005. Selective estrogen-receptor modulators for primary prevention of breast cancer. *J Clin Oncol,* 23**,** 1644-55.

FANG, S. H., CHEN, Y. & WEIGEL, R. J. 2009. GATA-3 as a marker of hormone response in breast cancer. *J Surg Res,* 157**,** 290-5.

FARATIAN, D., GOLTSOV, A., LEBEDEVA, G., SOROKIN, A., MOODIE, S., MULLEN, P., KAY, C., UM, I. H., LANGDON, S., GORYANIN, I. & HARRISON, D. J. 2009. Systems biology reveals new strategies for personalizing cancer medicine and confirms the role of PTEN in resistance to trastuzumab. *Cancer Res,* 69**,** 6713-20.

FARMER, P., BONNEFOI, H., BECETTE, V., TUBIANA-HULIN, M., FUMOLEAU, P., LARSIMONT, D., MACGROGAN, G., BERGH, J., CAMERON, D., D., G., DUSS, S., NICOULAZ, A. L., BRISKEN, C., FICHE, M., DELORENZI, M. & IGGO, R. 2005. Identification of molecular apocrine breast tumours by microarray analysis. *Oncogene,* 24**,** 4660-71.

FAUSETT, L. V. 1994. *Fundamentals of Neural Networks: Architectures, Algorithms, and Applications*, Prentice-Hall.

FIRPI, H. A., UCAR, D. & TAN, K. 2010. Discover regulatory DNA elements using chromatin signatures and artificial neural network. *Bioinformatics,* 26**,** 1579-86.

FOEKENS, J. A., ATKINS, D., ZHANG, Y., SWEEP, F. C., HARBECK, N., PARADISO, A., CUFER, T., SIEUWERTS, A. M., TALANTOV, D., SPAN, P. N., TJAN-HEIJNEN, V. C., ZITO, A. F., SPECHT, K., HOEFLER, H., GOLOUH, R., SCHITTULLI, F., SCHMITT, M., BEEX, L. V., KLIJN, J. G. & WANG, Y. 2006. Multicenter validation of a gene expression-based prognostic signature in lymph node-negative primary breast cancer. *J Clin Oncol,* 24**,** 1665-71.

GADKAR-SABLE, S., SHAH, C., ROSARIO, G., SACHDEVA, G. & PURI, C. 2005. Progesterone receptors: various forms and functions in reproductive tissues. *Front Biosci,* 10**,** 2118-30.

GARCIA-TUNON, I., RICOTE, M., RUIZ, A., FRAILE, B., PANIAGUA, R. & ROYUELA, M. 2005. IL-6, its receptors and its relationship with bcl-2 and bax proteins in infiltrating and in situ human breast carcinoma. *Histopathology,* 47**,** 82-9.

GLYNN, R. W., MILLER, N. & KERIN, M. J. 2010. 17q12-21 - the pursuit of targeted therapy in breast cancer. *Cancer Treat Rev,* 36**,** 224-9.

GOHARI, M. R., BIGLARIAN, A., BAKHSHI, E. & POURHOSEINGHOLI, M. A. 2011. Use of an artificial neural network to determine prognostic factors in colorectal cancer patients. *Asian Pac J Cancer Prev. ,* 12**,** 1469-72.

GOKMEN-POLAR, Y., NAKSHATRI, H. & BADVE, S. 2011. Biomarkers for breast cancer stem cells: the challenges ahead. *Biomark Med. ,* 5**,** 661-71.

GOLUB, T. R., SLONIM, D. K., TAMAYO, P., HUARD, C., GAASENBEEK, M., MESIROV, J. P., COLLER, H., LOH, M. L., DOWNING, J. R., CALIGIURI, M. A., BLOOMFIELD, C. D. & LANDER, E. S. 1999. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science,* 286**,** 531-7.

GONI, J., ESTEBAN, F. J., DE MENDIZABAL, N. V., SEPULCRE, J., ARDANZA-TREVIJANO, S., AGIRREZABAL, I. & VILLOSLADA, P. 2008. A computational analysis of protein-protein interaction networks in neurodegenerative diseases. *BMC Syst Biol,* 2**,** 52.

GONZALEZ, E., LILJENSTROM, H., RUIZ, Y. & LI, G. 2010. A biologically inspired model for pattern recognition. *J Zhejiang Univ Sci B,* 11**,** 115-26.

GOWN, A. M. 2008. Current issues in ER and HER2 testing by IHC in breast cancer. *Mod Pathol,* 21 Suppl 2**,** S8-S15.

GREEN, C. D., THOMPSON, P. D., JOHNSTON, P. G. & EL-TANANI, M. K. 2007. Interaction between transcription factor, basal transcription factor 3, and the NH2-terminal domain of human estrogen receptor alpha. *Mol Cancer Res,* 5**,** 1191-200.

GRUVBERGER, S., RINGNER, M., CHEN, Y., PANAVALLY, S., SAAL, L. H., A., B., FERNO, M., PETERSON, C. & MELTZER, P. S. 2001. Estrogen receptor status in breast cancer is associated with remarkably distinct gene expression patterns. *Cancer Res.,* 61**,** 5979-84.

GUTIERREZ-RIOS, R. M., ROSENBLUETH, D. A., LOZA, J. A., HUERTA, A. M., GLASNER, J. D., BLATTNER, F. R. & COLLADO-VIDES, J. 2003. Regulatory network of Escherichia coli: consistency between literature knowledge and microarray profiles. *Genome Res,* 13**,** 2435-43.

GUTIERREZ, M. C., DETRE, S., JOHNSTON, S., MOHSIN, S. K., SHOU, J., ALLRED, D. C., SCHIFF, R., OSBORNE, C. K. & DOWSETT, M. 2005. Molecular changes in tamoxifen-resistant breast cancer: relationship between estrogen receptor, HER-2, and p38 mitogen-activated protein kinase. *J Clin Oncol,* 23**,** 2469-76.

HABASHY, H. O., POWE, D. G., RAKHA, E. A., BALL, G., PAISH, C., GEE, J., NICHOLSON, R. I. & ELLIS, I. O. 2008. Forkhead-box A1 (FOXA1) expression in breast cancer and its prognostic significance. *Eur J Cancer,* 44**,** 1541-51.

HABASHY, H. O., POWE, D. G., STAKA, C. M., RAKHA, E. A., BALL, G., GREEN, A. R., ALESKANDARANY, M., PAISH, E. C., DOUGLAS, M. R., NICHOLSON, R. I., ELLIS, I. O. & GEE, J. M. 2010. Transferrin receptor (CD71) is a marker of poor prognosis in breast cancer and can predict response to tamoxifen. *Breast Cancer Res Treat,* 19**,** 283-93.

HAMMOND, K. L., HANSON, I. M., BROWN, A. G., LETTICE, L. A. & HILL, R. E. 1998. Mammalian and Drosophila dachshund genes are related to the Ski proto-oncogene and are expressed in eye and limb. *Mech Dev,* 74**,** 121-31.

HAPPEL, B. L. & MURRAE, M. J. 1994. The Design and Evolution of Modular Neural Network Architectures. *Neural NEtworks,* 7**,** 985-1004.

HART, C. E., MJOLSNESS, E. & WOLD, B. J. 2006. Connectivity in the yeast cell cycle transcription network: inferences from neural networks. *PLoS Comput Biol,* 2**,** e169.

HARTEMINK, A. J. 2005. Reverse engineering gene regulatory networks. *Nat Biotechnol,* 23**,** 554-5.

HARTEMINK, A. J., GIFFORD, D. K., JAAKKOLA, T. S. & YOUNG, R. A. 2002. Bayesian methods for elucidating genetic regulatory networks. *Intelligent Systems, IEEE,* 17**,** 37-43.

HAWKINS, D. M. 2004. The problem of overfitting. *J Chem Inf Comput Sci.,* 44**,** 1-12.

HAYASHI, S. I., EGUCHI, H., TANIMOTO, K., YOSHIDA, T., OMOTO, Y., INOUE, A., YOSHIDA, N. & YAMAGUCHI, Y. 2003. The expression and function of estrogen

receptor alpha and beta in human breast cancer and its clinical application. *Endocr Relat Cancer.,* 10**,** 193-202.

HAYETE, B., GARDNER, T. S. & COLLINS, J. J. 2007. Size matters: network inference tackles the genome scale. *Mol Syst Biol,* 3**,** 77.

HAYKIN, S. S. & GWYNN, R. 2009. *Neural Networks and Learning Machines,* Upper Saddle River/US.

HEDENFALK, I. A., RINGNER, M., TRENT, J. M. & BORG, A. 2002. Gene expression in inherited breast cancer. *Adv Cancer Res. ,* 84**,** 1-34.

HENNESSY, B. T., GONZALEZ-ANGULO, A. M., STEMKE-HALE, K., GILCREASE, M. Z., KRISHNAMURTHY, S., LEE, J. S., FRIDLYAND, J., SAHIN, A., AGARWAL, R., JOY, C., LIU, W., STIVERS, D., BAGGERLY, K., CAREY, M., LLUCH, A., MONTEAGUDO, C., HE, X., WEIGMAN, V., FAN, C., PALAZZO, J., HORTOBAGYI, G. N., NOLDEN, L. K., WANG, N. J., VALERO, V., GRAY, J. W., PEROU, C. M. & MILLS, G. B. 2009. Characterization of a naturally occurring breast cancer subset enriched in epithelial-to-mesenchymal transition and stem cell characteristics. *Cancer Res. ,* 69**,** 4116-24.

HERNANDEZ-ESPINOSA, C. & FERNANDEZ-REDONDO, M. On the design of constructive training algorithms for multilayer feedforward. Neural Networks, 2002. IJCNN '02. Proceedings of the 2002 International Joint Conference on, 2002 2002. 890-895.

HERSCHKOWITZ, J. I. 2010. Breast cancer stem cells: initiating a new sort of thinking. *Dis Model Mech. ,* 5**,** 257-8.

HIETER, P. & BOGUSKI, M. 1997. Functional Genomics: It's All How You Read It. *Science,* 278**,** 601-602.

HOCH, R. V., THOMPSON, D. A., BAKER, R. J. & WEIGEL, R. J. 1999. GATA-3 is expressed in association with estrogen receptor in breast cancer. *Int J Cancer,* 84**,** 122-8.

HOFFMANN, J. & SOMMER, A. 2005. Steroid hormone receptors as targets for the therapy of breast and prostate cancer--recent advances, mechanisms of resistance, and new approaches. *J Steroid Biochem Mol Biol,* 39**,** 191-200.

HONMA, N., HORII, R., IWASE, T., SAJI, S., YOUNES, M., TAKUBO, K., MATSUURA, M., ITO, Y., AKIYAMA, F. & ., S. G. 2008. Clinical importance of estrogen receptor-beta evaluation in breast cancer patients treated with adjuvant tamoxifen therapy. *J Clin Oncol,* 26**,** 3727-34.

HU, R., DAWOOD, S., HOLMES, M. D., COLLINS, L. C., SCHNITT, S. J., COLE, K., MAROTTI, J. D., HANKINSON, S. E., COLDITZ, G. A. & TAMIMI, R. M. 2011. Androgen receptor expression and breast cancer survival in postmenopausal women. *Clin Cancer Res,* 17**,** 1867-74.

HUANG, E., CHENG, S. H., DRESSMAN, H., PITTMAN, J., TSOU, M. H., HORNG, C. F., BILD, A., IVERSEN, E. S., LIAO, M., CHEN, C. M., WEST, M., NEVINS, J. R. & HUANG, A. T. 2003. Gene expression predictors of breast cancer outcomes. *The Lancet,* 361**,** 1590-1596.

HURTADO, A., HOLMES, K. A., ROSS-INNES, C. S., SCHMIDT, D. & CARROLL, J. S. 2011. FOXA1 is a key determinant of estrogen receptor function and endocrine response. *Nat Genet,* 43**,** 27-33.

HUSMEIER, D. 2003. Reverse engineering of genetic networks with Bayesian networks. *Biochem Soc Trans,* 31**,** 1516-8.

IKEDA, H., TAIRA, N., HARA, F., FUJITA, T., YAMAMOTO, H., SOH, J., TOYOOKA, S., NOGAMI, T., SHIEN, T., DOIHARA, H. & MIYOSHI, S. 2010. The estrogen receptor influences microtubule-associated protein tau (MAPT) expression and the selective estrogen receptor inhibitor fulvestrant downregulates MAPT and increases the sensitivity to taxane in breast cancer cells. *Breast Cancer Res,* 12**,** R43.

IMAI, K., HIRATA, S., IRIE, A., SENJU, S., IKUTA, Y., YOKOMINE, K., HARAO, M., INOUE, M., TSUNODA, T., NAKATSURU, S., NAKAGAWA, H., NAKAMURA, Y., BABA, H. & NISHIMURA, Y. 2008. Identification of a novel tumor-associated antigen, cadherin 3/P-cadherin, as a possible target for immunotherapy of pancreatic, gastric, and colorectal cancers. *Clin Cancer Res,* 14**,** 6487-95.

ITADANI, H., MIZUARAI, S. & KOTANI, H. 2008. Can systems biology understand pathway activation? Gene expression signatures as surrogate markers for understanding the complexity of pathway activation. *Curr Genomics,* 9**,** 349-60.

JAGLA, B. & SCHUCHHARDT, J. 2000. Adaptive encoding neural networks for the recognition of human signal peptide cleavage sites. *Bioinformatics.,* 16**,** 245-50.

JEMAL, A., BRAY, F., CENTER, M. M., FERLAY, J., WARD, E. & FORMAN, D. 2011. Global cancer statistics. *CA Cancer J Clin,* 61**,** 69-90.

JIANG, X., BARMADA, M. M. & VISWESWARAN, S. 2011. Identifying genetic interactions in genome-wide data using Bayesian networks. *Genet Epidemiol. ,* 34**,** 575-81.

JIN, Q., HUANG, Y. & FAN, N. 2012. Learning images using compositional pattern-producing neural networks for source camera identification and digital demographic diagnosis. *Pattern Recognition Letters,* 33**,** 381-396.

JIN, W., CHEN, Y., DI, G. H., MIRON, P., HOU, Y. F., GAO, H. & SHAO, Z. M. 2008a. Estrogen receptor (ER) beta or p53 attenuates ERalpha-mediated transcriptional activation on the BRCA2 promoter. *J Biol Chem,* 283**,** 29671-80.

JIN, X., FURBER, S. B. & WOODS, J. V. Efficient modelling of spiking neural networks on a scalable chip multiprocessor. Neural Networks, 2008. IJCNN 2008. (IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on, 1-8 June 2008 2008b. 2812-2819.

JUNNILA, S., KOKKOLA, A., KARJALAINEN-LINDSBERG, M. L., PUOLAKKAINEN, P. & MONNI, O. 2010. Genome-wide gene copy number and expression analysis of primary gastric tumors and gastric cancer cell lines. *BMC Cancer,* 10**,** 73.

KAK, S. 1999. Better web searches and prediction with instantaneously trained neural networks. *IEEE Intelligent Systems***,** 78-81.

KAKLAMANI, V. 2006. A genetic signature can predict prognosis and response to therapy in breast cancer: Oncotype DX. *Expert Rev Mol Diagn,* 6**,** 803-9.

KARLEBACH, G. & SHAMIR, R. 2008. Modelling and analysis of gene regulatory networks. *Nat Rev Mol Cell Biol,* 9**,** 770-780.

KASABOV, N. 2001. Evolving fuzzy neural networks for supervised/unsupervised online knowledge-based learning. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on,* 31**,** 902-918.

KAVZOGLU, T. & MATHER, P. M. Assessing Artificial Neural Network Prining Algorithms. 24th Annual Conference and Exhibition of the Remote Sensing Society, 1998 Greenwich, UK. 603-609.

KAWASHIMA, M. & MIOSSEC, P. 2005. mRNA quantification of T-bet, GATA-3, IFN-gamma, and IL-4 shows a defective Th1 immune response in the peripheral

blood from rheumatoid arthritis patients: link with disease activity. *J Clin Immunol,* 25**,** 209-14.

KHAN, J. I. 1998. Characteristics of multidimensional holographic associative memory in retrieval with dynamically localizable attention. *Neural Networks, IEEE Transactions on,* 9**,** 389-406.

KHAN, S. S. & AHMAD, A. 2004. Cluster center initialization algorithm for K-means clustering. *Pattern Recognition Letters,* 25**,** 1293-1302.

KIMURA, M., KOIDA, T., SHIMIZU, Y., FUKUDA, T., SAKAUE, S. & NAKAMURA, S. 1988. [Hormone receptors (ER, PgR) as a prognostic factor in breast cancer]. *Gan No Rinsho,* 34**,** 1083-8.

KITANO, H. 2002a. Computational systems biology. *Nature,* 420**,** 206-10.

KITANO, H. 2002b. Systems biology: a brief overview. *Science,* 295**,** 1662-4.

KLINGE, C. M. 2001. Estrogen receptor interaction with estrogen response elements. *Nucleic Acids Res,* 29**,** 2905-2919.

KODA, M., SULKOWSKI, S., GAROFALO, C., KANCZUGA-KODA, L., SULKOWSKA, M. & SURMACZ, E. 2003. Expression of the insulin-like growth factor-I receptor in primary breast cancer and lymph node metastases: correlations with estrogen receptors alpha and beta. *Horm Metab Res,* 35**,** 794-801.

KOHL, P. & NOBLE, D. 2009. Systems biology and the virtual physiological human. *Mol Syst Biol. ,* 2**,** 292.

KOHONEN, T. 1990. The self-organizing map. *Proceedings of the IEEE,* 78**,** 1464-1480.

KREEGER, P. K. & LAUFFENBURGER, D. A. 2010. Cancer systems biology: a network modeling perspective. *Carcinogenesis,* 31**,** 2-8.

KUSUMAWIDJAJA, G., KAYED, H., GIESE, N., BAUER, A., ERKAN, M., GIESE, T., HOHEISE, J. D., FRIESS, H. & KLEEFF, J. 2007. Basic transcription factor 3 (BTF3) regulates transcription of tumor-associated genes in pancreatic cancer cells. *Cancer Biol Ther,* 6**,** 367-76.

LAGANIERE, J., DEBLOIS, G. & GIGUERE, V. 2005. Functional genomics identifies a mechanism for estrogen activation of the retinoic acid receptor alpha1 gene in breast cancer cells. *Mol Endocrinol,* 19**,** 1584-92.

LANCASHIRE, L., SCHMID, O., SHAH, H. & BALL, G. 2005. Classification of bacterial species from proteomic data using combinatorial approaches incorporating artificial neural networks, cluster analysis and principal components analysis. *Bioinformatics,* 21**,** 2191-9.

LANCASHIRE, L. J., LEMETRE, C. & BALL, G. R. 2009. An introduction to artificial neural networks in bioinformatics--application to complex microarray and mass spectrometry datasets in cancer studies. *Brief Bioinform.,* 10**,** 315-29.

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. 2010. A validated gene expression profile for detecting clinical outcome in breast cancer using artificial neural networks. *Breast Cancer Res Treat. ,* 120**,** 83-93.

LANCASHIRE, L. J., REES, R., C. & BALL, G. R. 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**,** 99-111.

LEE, K., JESSOP, H., SUSWILLO, R., ZAMAN, G. & LANYON, L. 2003. Bone adaptation requires oestrogen receptor-alpha. *Nature,* 424**,** 389-390.

LEE, K., LIU, Y., MO, J. Q., ZHANG, J., DONG, Z. & LU, S. 2008. Vav3 oncogene activates estrogen receptor and its overexpression may be involved in human breast cancer. *BMC Cancer,* 8**,** 158.

LEE, T. J., POULIOT, Y., WAGNER, V., GUPTA, P., STRINGER-CALVERT, D. W. J., TENENBAUM, J. D. & KARP, P. D. 2006. BioWarehouse: a bioinformatics database warehouse toolkit. *BMC Bioinformatics,* 7**,** 170.

LEMETRE, C., LANCASHIRE, L. J., REES, R. C. & BALL, R. G. Artificial neural network based algorithm for biomedical interactions modeling. *In:* CABESTANY, J., ed. Interanaltion Work-Conference on Artificial Neural Networks, 2009 Salamanca, Spain. Springer, 877-885.

LEMIEUX, P. & FUQUA, S. 1996. The role of the estrogen receptor in tumor progression. *J Steroid Biochem Mol Biol.,* 55**,** 87-91.

LEVIN, E. R. 2003. Bidirectional signaling between the estrogen receptor and the epidermal growth factor receptor. *Mol Endocrinol,* 17**,** 309-17.

LEVIN, E. R. 2009. Membrane oestrogen receptor alpha signalling to cell functions. *J Physiol,* 587**,** 5019-23.

LI, X., HUANG, J., YI, P., BAMBARA, R. A., HILF, R. & MUYAN, M. 2004. Single-chain estrogen receptors (ERs) reveal that the ERalpha/beta heterodimer emulates functions of the ERalpha dimer in genomic estrogen signaling pathways. *Mol Cell Biol,* 24**,** 7681-94.

LIANG, F., LU, Q., SUN, S., ZHOU, J., POPOV, V. M., LI, S., LI, W., LIU, Y., JIANG, J. & KONG, B. 2012. Increased expression of dachshund homolog 1 in ovarian cancer as a predictor for poor outcome. *Interanational Journal of Gynecolopgical Cancer,* 22**,** 386-393.

LIPSHUTZ, R. J., FODOR, S. P., GINGERAS, T. R. & LOCKHART, D. J. 1999. High density synthetic oligonucleotide arrays. *Nat Genet. ,* 21**,** 20-4.

LISBOA, P. J. 2002. A review of evidence of health benefit from artificial neural networks in medical intervention. *Neural Networks* 15**,** 11-39.

LISBOA, P. J. & TAKTAK, A. F. 2006. The use of artificial neural networks in decision support in cancer: a systematic review. *Neural Networks,* 19**,** 408-15.

LIU, Q., SUNG, A. H., CHEN, Z., LIU, J., HUANG, X. & DENG, Y. 2009. Feature selection and classification of MAQC-II breast cancer and multiple myeloma microarray gene expression data. *PLoS One,* 4**,** e8250.

LOPEZ-TARRUELLA, S. & SCHIFF, R. 2007. The dynamics of estrogen receptor status in breast cancer: re-shaping the paradigm. *Clin Cancer Res.,* 13**,** 6921-5.

LUPIEN, M., EECKHOUTE, J., MEYER, C. A., WANG, Q., ZHANG, Y., LI, W., CARROLL, J. S., LIU, X. S. & BROWN, M. 2008. FoxA1 translates epigenetic signatures into enhancer-driven lineage-specific transcription. *Cell,* 132**,** 958-70.

MA, J. & JEMAL, A. 2013. Breast Cancer Statistics. *In:* AHMAD, A. (ed.) *Breast Cancer Metastasis and Drug Resistance.* Springer New York.

MAGUIRE, R. J., ISAACS, H. V. & POWNALL, M. E. 2012. Early transcriptional targets of MyoD link myogenesis and somitogenesis. *Dev Biol,* 371**,** 256-68.

MAHADEVAN, R. & PALSSON, B. O. 2005. Properties of metabolic networks: structure versus function. *Biophys J,* 88**,** L07-9.

MALIK, S., JIANG, S., GAREE, J. P., VERDIN, E., LEE, A. V., O'MALLEY, B. W., ZHANG, M., BELAGULI, N. S. & OESTERREICH, S. 2010. Histone deacetylase 7 and FoxA1 in estrogen-mediated repression of RPRM. *Mol Cell Biol,* 30**,** 399-412.

MANI, K. M., LEFEBVRE, C., WANG, K., LIM, W. K., BASSO, K., DALLA-FAVERA, R. & CALIFANO, A. 2008. A systems biology approach to prediction of oncogenes and molecular perturbation targets in B-cell lymphomas. *Mol Syst Biol,* 4**,** 169.

MARINE, J. & WINOTO, A. 1991. The human enhancer-binding protein Gata3 binds to several T-cell receptor regulatory elements. *Proc. Natl. Acad. Sci.,* 88**,** 7284-7288.

MATHAROO-BALL, B., HUGHES, C., LANCASHIRE, L. J., TOOTH, D., BALL, G., CREASER, C., ELGASIM, M., REES, R., LAYFIELD, R. & ATIOMO, W. 2007. Characterization of biomarkers in polycystic ovary syndrome (PCOS) using multiple distinct proteomic platforms. *J Proteome Res.,* 6**,** 3321-8.

MCCARTY, K. S., MILLER, L. S., COX, E. B., KONRATH, J. & MCCARTY, K. S., SR. 1985. Estrogen receptor analyses. Correlation of biochemical and immunohistochemical methods using monoclonal antireceptor antibodies. *Arch Pathol Lab Med,* 109**,** 716-21.

MCCULLOCH, W. & PITTS, W. 1943. A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biology,* 5**,** 115-133.

MEHRA, R., VARAMBALLY, S., DING, L., SHEN, R., SABEL, M. S., GHOSH, D., CHINNAIYAN, A. & KLEER, C. G. 2005. Identification of GATA3 as a breast cancer prognostic marker by global gene expression meta-analysis. *Cancer Res,* 65**,** 11259-64.

MENDES, K. B., FIUZA, R. M. & STEINER, M. T. A. 2010. Diagnosis of Headache using Artificial Neural Networks. *International Journal of Computer Science and Network Security,* 10**,** 172-178.

MIGNOTTE, M. 2008. Segmentation by fusion of histogram-based k-means clusters in different color spaces. *IEEE Trans Image Process,* 17**,** 780-7.

MILLER, L. D., SMEDS, J., GEORGE, J., VEGA, V. B., VERGARA, L., PLONER, A., PAWITAN, Y., HALL, P., KLAAR, S., LIU, E. T. & BERGH, J. 2005. An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. *Proc Natl Acad Sci U S A.,* 120**,** 13550-5.

MILO, M., CACCIABUE-RIVOLTA, D., KNEEBONE, A., VAN DOORNINCK, H., JOHNSON, C., LAWOKO-KERALI, G., NIRANJAN, M., RIVOLTA, M. & HOLLEY, M. 2009. Genomic analysis of the function of the transcription factor gata3 during development of the mammalian inner ear. *PLoS One,* 4**,** e7144.

MINOO, P., HU, L., XING, Y., ZHU, N. L., CHEN, H., LI, M., BOROK, Z. & LI, C. 2007. Physical and functional interactions between homeodomain NKX2.1 and winged helix/forkhead FOXA1 in lung epithelial cells. *Mol Cell Biol,* 27**,** 2155-65.

MITA, K., ZHANG, Z., ANDO, Y., TOYAMA, T., HAMAGUCHI, M., KOBAYASHI, S., HAYASHI, S., FUJII, Y., IWASE, H. & YAMASHITA, H. 2007. Prognostic significance of insulin-like growth factor binding protein (IGFBP)-4 and IGFBP-5 expression in breast cancer. *Jpn J Clin Oncol,* 37**,** 575-82.

MITCHELL, T. M. 1997. *Machine Learning*, Mcgraw-Hill International.

MOKBEL, K. 2004. Anastrozole and Breast Cancer. *Breast cancer reviews***,** 2-9.

MOORES, S. L., SELFORS, L. M., FREDERICKS, J., BREIT, T., FUJIKAWA, K., ALT, F. W., BRUGGE, J. S. & SWAT, W. 2000. Vav family proteins couple to diverse cell surface receptors. *Mol Cell Biol,* 20**,** 6364-73.

MORRIS, P. 1993. The breakout method for escaping from local minima. *Proceedings of the eleventh national conference on Artificial intelligence.* Washington, D.C.: AAAI Press.

MOVILLA, N. & BUSTELO, X. R. 1999. Biological and regulatory properties of Vav-3, a new member of the Vav family of oncoproteins. *Mol Cell Biol,* 19**,** 7870-85.

MUNI, D. P., PAL, N. R. & DAS, J. 2006. Genetic programming for simultaneous feature selection and classifier design. *IEEE Trans Syst Man Cybern B Cybern. ,* 36**,** 106-17.

MURPHY, D. 2002. Gene expression studies using microarrays: principles, problems, and prospects. *Adv Physiol Educ. ,* 26**,** 256-70.

NAN, F., LU, Q., ZHOU, J., CHENG, L., POPOV, V. M., WEI, S., KONG, B., PESTELL, R. G., LISANTI, M. P., JIANG, J. & WANG, C. 2009. Altered expression of DACH1 and cyclin D1 in endometrial cancer. *Cancer Biol Ther,* 8**,** 1534-9.

NICHOLSON, R. I., HUTCHESON, I. R., HISCOX, S. E., KNOWLDEN, J. M., GILES, M., BARROW, D. & GEE, J. M. W. 2005. Growth factor signalling and resistance to selective oestrogen receptor modulators and pure anti-oestrogens: the use of anti-growth factor therapies to treat or delay endocrine resistance in breast cancer. *Endocr Relat Cancer.,* 12**,** S29-36.

NIKOLIC, Z. J. & FU, K. S. 1986. An algorithm for learning without external supervision and its application to learning control systems. *IEEE Trans Pattern Anal Mach Intell. ,* 8**,** 304-12.

NOVELLI, F., MILELLA, M., MELUCCI, E., DI BENEDETTO, A., SPERDUTI, I., PERRONE-DONNORSO, R., PERRACCHIO, L., VENTURO, I., NISTICO, C., FABI, A., BUGLIONI, S., NATALI, P. G. & MOTTOLESE, M. 2008. A divergent role for estrogen receptor-beta in node-positive and node-negative breast cancer classified according to molecular subtypes: an observational prospective study. *Breast Cancer Res,* 10**,** R74.

OKAMOTO, T., SUZUKI, T. & YAMAMOTO, N. 2000. Microarray fabrication with covalent attachment of DNA using bubble jet technology. *Nat Biotechnol. ,* 18**,** 438-41.

OKOH, V., DEORAJ, A. & ROY, D. 2011. Estrogen-induced reactive oxygen species-mediated signalings contribute to breast cancer. *Biochim Biophys Acta,* 1815**,** 115-33.

OLDENHUIS, C. N., OOSTING, S. F., GIETEMA, J. A. & DE VRIES, E. G. E. 2008. Prognostic versus predictive value of biomarkers in oncology. *Eur J Cancer. ,* 44**,** 946-53.

PAREDES, J., ALBERGARIA, A., OLIVEIRA, J. T., JERONIMO, C., MILANEZI, F. & SCHMITT, F. C. 2005. P-cadherin overexpression is an indicator of clinical outcome in invasive breast carcinomas and is associated with CDH3 promoter hypomethylation. *Clin Cancer Res,* 11**,** 5869-77.

PAREDES, J., STOVE, C., STOVE, V., MILANEZI, F., VAN MARCK, V., DERYCKE, L., MAREEL, M., BRACKE, M. & SCHMITT, F. 2004. P-cadherin is up-regulated by the antiestrogen ICI 182,780 and promotes invasion of human breast cancer cells. *Cancer Res,* 64**,** 8309-17.

PARRIS, T. Z., DANIELSSON, A., NEMES, S., KOVACS, A., DELLE, U., FALLENIUS, G., MOLLERSTROM, E., KARLSSON, P. & HELOU, K. 2010. Clinical implications of gene dosage and gene expression patterns in diploid breast carcinoma. *Clin Cancer Res,* 16**,** 3860-74.

PARRY, R. M., JONES, W., STOKES, T. H., PHAN, J. H., MOFFITT, R. A., FANG, H., SHI, L., OBERTHUER, A., FISCHER, M., TONG, W. & WANG, M. D. 2010. k-Nearest neighbor models for microarray gene expression analysis and clinical outcome prediction. *Pharmacogenomics J,* 10**,** 292-309.

PASCUAL-LEONE, A., AMEDI, A., FREGNI, F. & MERABET, L. B. 2005. The plastic human brain cortex. *Annu Rev Neurosci,* 28**,** 377-401.

PASTOREKOVA, S., RATCLIFFE, P. J. & PASTOREK, J. 2008. Molecular mechanisms of carbonic anhydrase IX-mediated pH regulation under hypoxia. *BJU Int,* 101 Suppl 4.

PERI, S., NAVARRO, J. D., KRISTIANSEN, T. Z., AMANCHY, R., SURENDRANATH, V., MUTHUSAMY, B., GANDHI, T. K., CHANDRIKA, K. N., DESHPANDE, N., SURESH, S., RASHMI, B. P., SHANKER, K., PADMA, N., NIRANJAN, V., HARSHA, H. C., TALREJA, N., VRUSHABENDRA, B. M., RAMYA, M. A., YATISH, A. J., JOY, M., SHIVASHANKAR, H. N., KAVITHA, M. P., MENEZES, M., CHOUDHURY, D. R., GHOSH, N., SARAVANA, R., CHANDRAN, S., MOHAN, S., JONNALAGADDA, C. K., PRASAD, C. K., KUMAR-SINHA, C., DESHPANDE, K. S. & PANDEY, A. 2003. Human protein reference database as a discovery resource for proteomics. *Nucleic Acids Res,* 32(Database issue)**,** D497-501.

PEROU, C. M. 2011. Molecular stratification of triple-negative breast cancers. *Oncologist,* 16**,** 61-70.

PEROU, C. M., SORLIE, T., EISEN, M. B., VAN DE RIJN, M., JEFFREY, S. S., REES, C. A., POLLACK, J. R., ROSS, D. T., JOHNSEN, H., AKSLEN, L. A., FLUGE, O., PERGAMENSCHIKOV, A., WILLIAMS, C., ZHU, S. X., LONNING, P. E., BORRESEN-DALE, A. L., BROWN, P. O. & BOTSTEIN, D. 2000. Molecular portraits of human breast tumours. *Nature,* 104**,** 747-52.

POPOV, V. M., ZHOU, J., SHIRLEY, L. A., QUONG, J., YEOW, W. S., WRIGHT, J. A., WU, K., RUI, H., VADLAMUDI, R. K., JIANG, J., KUMAR, R., WANG, C. & PESTELL, R. G. 2009. The cell fate determination factor DACH1 is expressed in estrogen receptor-alpha-positive breast cancer and represses estrogen receptor-alpha signaling. *Cancer Res,* 69**,** 5752-60.

POTTER, C. P. & HARRIS, A. L. 2003. Diagnostic, prognostic and therapeutic implications of carbonic anhydrases in cancer. *Br J Cancer,* 89**,** 2-7.

POZHARISSKI, K. M. 2005. Pathology of tumours in laboratory animals. Tumours of the rat. Tumours of the intestines. *ARC Sci Publ.,* 99**,** 159-98.

RAMASWAMY, S. & PEROU, C. M. 2003. DNA microarrays in breast cancer: the promise of personalised medicine. *Lancet,* 361**,** 1576-7.

RASMUSSEN, B. B., THORPE, S. M., NORGAARD, T., RASMUSSEN, J., AGDAL, N. & ROSE, C. 1989. Immunohistochemical steroid receptor detection in frozen breast cancer tissue. A multicenter investigation. *Acta Oncol.,* 27**,** 757-60.

RAYCHAUDHURI, S., STUART, J. M. & ALTMAN, R. B. 2000. Principal components analysis to summarize microarray experiments: application to sporulation time series. *Pac Symp Biocomput.* **,** 455-66.

RICKETTS, D., TURNBULL, L., RYALL, G., BAKHSHI, R., RAWSON, N. S., GAZET, J. C., NOLAN, C. & COOMBES, R. C. 1991. Estrogen and progesterone receptors in the normal female breast. *Cancer Res.,* 51**,** 1817-22.

RILEY, M. J. W., THOMPSON, C. P. & JENKINS, K. W. Improving the Performance of Cascade Correlation Neural Networks on Multimodal Functions. World Congress on Engineering, 2010 London, UK.

ROJAS, R. 1996. *Neural networks : a systematic introduction,* Berlin; New York, Springer-Verlag.

ROSENBLATT, F. 1962. *Principles of neurodynamics: perceptrons and the theory of brain mechanisms*, Spartan Books.

ROUZIER, R., PEROU, C. M., SYMMANS, W. F., IBRAHIM, N., CRISTOFANILLI, M., ANDERSON, K., HESS, K. R., STEC, J., AYERS, M., WAGNER, P., MORANDI, P., FAN, C., RABIUL, I., ROSS, J. S., HORTOBAGYI, G. N. & PUSZTAI, L. 2005. Breast cancer molecular subtypes respond differently to preoperative chemotherapy. *Clin Cancer Res,* 11**,** 5678-85.

RUAL, J. F., VENKATESAN, K., HAO, T., HIROZANE-KISHIKAWA, T., DRICOT, A., LI, N., BERRIZ, G. F., GIBBONS, F. D., DREZE, M., AYIVI-GUEDEHOUSSOU, N., KLITGORD, N., SIMON, C., BOXEM, M., MILSTEIN, S., ROSENBERG, J., GOLDBERG, D. S., ZHANG, L. V., WONG, S. L., FRANKLIN, G., LI, S., ALBALA, J. S., LIM, J., FRAUGHTON, C., LLAMOSAS, E., CEVIK, S., BEX, C., LAMESCH, P., SIKORSKI, R. S., VANDENHAUTE, J., ZOGHBI, H. Y., SMOLYAR, A., BOSAK, S., SEQUERRA, R., DOUCETTE-STAMM, L., CUSICK, M. E., HILL, D. E., ROTH, F. P. & VIDAL, M. 2005. Towards a proteome-scale map of the human protein-protein interaction network. *Nature,* 437**,** 1173-8.

RUMELHART, D. E., HINTON, G. E. & WILLIAMS, R. J. 1986. Learning representations by back-propagating errors. *Nature,* 323**,** 533-536.

SAMAAN, N. A., BUZDAR, A. U., ALDINGER, K. A., SCHULTZ, P. N., YANG, K.-P., ROMSDAHL, M. M. & MARTIN, R. 1981. Estrogen Receptor: A Prognostic Factor in Breast Cancer. *Cancer,* 47**,** 554-560.

SARGENT, D. J. 2001. Comparison of artificial neural networks with other statistical approaches: results from medical data sets. *Cancer. ,* 91**,** 1636-42.

SAUER, U., HEINEMANN, M. & ZAMBONI, N. 2007. Genetics. Getting closer to the whole picture. *Science,* 316**,** 550-1.

SCHIFF, R., MASSARWEH, S. A., SHOU, J., BHARWANI, L., MOHSIN, S. K. & OSBORNE, C. K. 2004. Cross-talk between estrogen receptor and growth factor pathways as a molecular target for overcoming endocrine resistance. *Clin Cancer Res,* 10**,** 331S-6S.

SHAO, R., FENG, Y., ZOU, S., WEIJDEGARD, B., WU, G., BRANNSTROM, M. & BILLIG, H. 2012. The role of estrogen in the pathophysiology of tubal ectopic pregnancy. *Am J Transl Res,* 4**,** 269-78.

SHAPIRO, C. L. & RECHT, A. 2001. Side effects of adjuvant treatment of breast cancer. *N Engl J Med. ,* 344**,** 1997-2008.

SHAW, J. A., UDOKANG, K., MOSQUERA, J. M., CHAUHAN, H., JONES, J. L. & WALKER, R. A. 2002. Oestrogen receptors alpha and beta differ in normal human breast and breast carcinomas. *J Pathol.,* 198**,** 450-7.

SHENOUDA, E. 2006. A Quantitative Comparison of Different MLP Activation Functions in Classification. *In:* WANG, J., YI, Z., ZURADA, J., LU, B.-L. & YIN, H. (eds.) *Advances in Neural Networks - ISNN 2006.* Springer Berlin / Heidelberg.

SHI, H., BEVIER, M., JOHANSSON, R., ENQUIST-OLSSON, K., HENRIKSSON, R., HEMMINKI, K., LENNER, P. & FORSTI, A. 2012. Prognostic impact of polymorphisms in the

MYBL2 interacting genes in breast cancer. *Breast Cancer Res Treat,* 131**,** 1039-47.

SHI, L., CAMPBELL, G., JONES, W. D., CAMPAGNE, F., WEN, Z., WALKER, S. J., SU, Z., CHU, T. M., GOODSAID, F. M., PUSZTAI, L., SHAUGHNESSY, J. D. J., OBERTHUER, A., THOMAS, R. S., PAULES, R. S., FIELDEN, M., BARLOGIE, B., CHEN, W., DU, P., FISCHER, M., FURLANELLO, C., GALLAS, B. M., GE, X., MEGHERBI, D. B., SYMMANS, W. F., WANG, M. D., ZHANG, J., BITTER, H., BRORS, B., BUSHEL, P. R., BYLESJO, M., CHEN, M., CHENG, J., CHOU, J., DAVISON, T. S., DELORENZI, M., DENG, Y., DEVANARAYAN, V., DIX, D. J., DOPAZO, J., DORFF, K. C., ELLOUMI, F., FAN, J., FAN, S., FAN, X., FANG, H., GONZALUDO, N., HESS, K. R., HONG, H., HUAN, J., IRIZARRY, R. A., JUDSON, R., JURAEVA, D., LABABIDI, S., LAMBERT, C. G., LI, Y., LI, Z., LIN, S. M., LIU, G., LOBENHOFER, E. K., LUO, J., LUO, W., MCCALL, M. N., NIKOLSKY, Y., PENNELLO, G. A., PERKINS, R. G., PHILIP, R., POPOVICI, V., PRICE, N. D., QIAN, F., SCHERER, A., SHI, T., SHI, W., SUNG, J., THIERRY-MIEG, D., THIERRY-MIEG, J., THODIMA, V., TRYGG, J., VISHNUVAJJALA, L., WANG, S. J., WU, J., WU, Y., XIE, Q., YOUSEF, W. A., ZHANG, L., ZHANG, X., ZHONG, S., ZHOU, Y., ZHU, S., ARASAPPAN, D., BAO, W., LUCAS, A. B., BERTHOLD, F., BRENNAN, R. J., BUNESS, A., CATALANO, J. G., CHANG, C., CHEN, R., CHENG, Y., CUI, J., CZIKA, W., et al. 2010. The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. *Nat Biotechnol,* 28**,** 827-38.

SHIMOYAMA, Y., HIROHASHI, S., HIRANO, S., NOGUCHI, M., SHIMOSATO, Y., TAKEICHI, M. & ABE, O. 1989. Cadherin cell-adhesion molecules in human epithelial tissues and carcinomas. *Cancer Res,* 49**,** 2128-33.

SHOEMAKER, B. A. & PANCHENKO, A. R. 2007. Deciphering protein-protein interactions. Part II. Computational methods to predict protein and domain interaction parters. *Computational Biology,* 3**,** 595-601.

SIMON, R. 2003. Diagnostic and prognostic prediction using gene expression profiles in high-dimensional microarray data. *Br J Cancer.,* 89**,** 1599-604.

SIN, W. C., ZHANG, Y., ZHONG, W., ADHIKARAKUNNATHU, S., POWERS, S., HOEY, T., AN, S. & YANG, J. 2004. G protein-coupled receptors GPR4 and TDAG8 are oncogenic and overexpressed in human cancers. *Oncogene,* 23**,** 6299-303.

SINHA, N. K., GUPTA, M. M. & RAO, D. H. Dynamic neural networks: an overview. Industrial Technology 2000. Proceedings of IEEE International Conference on, 19-22 Jan. 2000 2000. 491-496 vol.2.

SLODKOWSKA, E. A. & ROSS, J. S. 2009. MammaPrint 70-gene signature: another milestone in personalized medical care for breast cancer patients. *Expert Rev Mol Diagn,* 9**,** 417-22.

SONG, L., WEI, X., ZHANG, B., LUO, X., LIU, J., FENG, Y. & XIAO, X. 2009. Role of Foxa1 in regulation of bcl2 expression during oxidative-stress-induced apoptosis in A549 type II pneumocytes. *Cell Stress Chaperones,* 14**,** 417-25.

SONG, M. R., LEE, S. K., SEO, Y. W., CHOI, H. S., LEE, J. W. & LEE, M. O. 1998. Differential modulation of transcriptional activity of oestrogen receptors by direct protein-protein interactions with retinoid receptors. *Biochem J,* 336 (Pt 3)**,** 711-7.

SOOD, A. K. 2009. PDEF and PDEF-induced proteins as candidate tumor antigens for T cell and antibody-mediated immunotherapy of breast cancer. *Immunol Res,* 46**,** 206-15.

SORLIE, T., PEROU, C. M., TIBSHIRANI, R., AAS, T., GEISLER, S., JOHNSEN, H., HASTIE, T., EISEN, M. B., VAN DE RIJN, M., JEFFREY, S. S., THORSEN, T., QUIST, H., MATESE, J. C., BROWN, P. O., BOTSTEIN, D., LONNING, P. E. & BORRESEN-DALE, A. L. 2001. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci U S A,* 98**,** 10869-74.

SORLIE, T., TIBSHIRANI, R., PARKER, J., HASTIE, T., MARRON, J. S., NOBEL, A., DENG, S., JOHNSEN, H., PESICH, R., GEISLER, S., DEMETER, J., PEROU, C. M., LONNING, P. E., BROWN, P. O., BORRESEN-DALE, A.-L. & BOSTEIN, D. 2003. Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc Natl Acad Sci U S A.,* 100**,** 8418-23.

SOTIRIOU, C. & DESMEDT, C. 2006. Gene expression profiling in breast cancer. *Ann Oncol. ,* 10**,** 259-62.

SOTIRIOU, C., NEO, S.-Y., MCSHANE, L. M., KORN, E. L., LONG, P. M., JAZAERI, A., MARTIAT, P., FOX, S. B., HARRIS, A. L. & LIU, E. T. 2003. Breast cancer classification and prognosis based on gene expression profiles from a population-based study. *Proc Natl Acad Sci U S A.,* 100**,** 10393-8.

SOTIRIOU, C. & PUSZTAI, L. 2009. Gene-expression signatures in breast cancer. *N Engl J Med.,* 360**,** 790-800.

SPEIRS, V. 2008. The evolving role of oestrogen receptor beta in clinical breast cancer. *Breast Cancer Res,* 10**,** 111.

SPEIRS, V., GREEN, A. R., HUGHES, T. A., ELLIS, I. O., SAUNDERS, P. T. & SHAABAN, A. M. 2008. Clinical importance of estrogen receptor beta isoforms in breast cancer. *J Clin Oncol,* 26**,** 5825.

SPEIRS, V. & SHAABAN, A. M. 2009. Role of ERβ in Clinical Breast Cancer. *Cancer Treat Res,* 147**,** 1-20.

SPIRIN, V. & MIRNY, L. A. 2003. Protein complexes and functional modules in molecular networks. *Proc Natl Acad Sci U S A.,* 100**,** 12123-8.

SROUR, N., REYMOND, M. A. & STEINERT, R. 2008. Lost in translation? A systematic database of gene expression in breast cancer. *Pathobiology,* 75**,** 112-8.

STAAF, J., JONSSON, G., RINGNER, M., VALLON-CHRISTERSSON, J., GRABAU, D., ARASON, A., GUNNARSSON, H., AGNARSSON, B. A., MALMSTROM, P. O., JOHANNSSON, O. T., LOMAN, N., BARKARDOTTIR, R. B. & BORG, A. 2010. High-resolution genomic and expression analyses of copy number alterations in HER2-amplified breast cancer. *Breast Cancer Res,* 12**,** R25.

STEFFEN, M., PETTI, A., AACH, J., D'HAESELEER, P. & CHURCH, G. 2002. Automated modelling of signal transduction networks. *BMC Bioinformatics,* 3**,** 34.

STRYHN, H. & CHRISTENSEN, J. 2013. The analysis-Hierarchical models: Past, present and future. *Prev Vet Med***,** 1873-1716.

SUN, J. M., SPENCER, V. A., LI, L., YU, C. H., YU, J. & DAVIE, J. R. 2005. Estrogen regulation of trefoil factor 1 expression by estrogen receptor alpha and Sp proteins. *Exp Cell Res,* 302**,** 96-107.

SUN, L. V., CHEN, L., GREIL, F., NEGRE, N., LI, T. R., CAVALLI, G., ZHAO, H., VAN STEENSEL, B. & WHITE, K. P. 2003. Protein-DNA interaction mapping using genomic tiling path microarrays in Drosophila. *Proc Natl Acad Sci U S A,* 100**,** 9428-33.

SZKLARCZYK, D., FRANCESCHINI, A., KUHN, M., SIMONOVIC, M., ROTH, A., MINGUEZ, P., DOERKS, T., STARK, M., MULLER, J., BORK, P., JENSEN, L. J. & VON MERING, C.

2011. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res,* 39(Database issue)**,** D561-8.

TANOS, T., ROJO, L. J., ECHEVERRIA, P. & BRISKEN, C. 2012. ER and PR signaling nodes during mammary gland development. *Breast Cancer Res.,* 14**,** 210.

TAUBE, J. H., ALLTON, K., DUNCAN, S. A., SHEN, L. & BARTON, M. C. 2010. Foxa1 functions as a pioneer transcription factor at transposable elements to activate Afp during differentiation of embryonic stem cells. *J Biol Chem,* 285**,** 16135-44.

TEFFERI, A., BOLANDER, M. E., ANSELL, S. M., WIEBEN, E. D. & SPELSBERG, T. C. 2002. Primer on Medical Genomics Part III: Microarray Experiments and Data Analysis. *Mayo Clinic Proceedings,* 77**,** 927-940.

THOMAS, P. D., KEJARIWAL, A., CAMPBELL, M. J., MI, H., DIEMER, K., GUO, N., LADUNGA, I., ULITSKY-LAZAREVA, B., MURUGANUJAN, A., RABKIN, S., VANDERGRIFF, J. A. & DOREMIEUX, O. 2003. PANTHER: a browsable database of gene products organized by biological function, using curated protein family and subfamily classification. *Nucleic Acids Res,* 31**,** 334-41.

THORNER, A. R., HOADLEY, K. A., PARKER, J. S., WINKEL, S., MILLIKAN, R. C. & PEROU, C. M. 2009. In vitro and in vivo analysis of B-Myb in basal-like breast cancer. *Oncogene,* 28**,** 742-51.

TONG, D. L. & SCHIERZ, A. C. 2011. Hybrid genetic algorithm-neural network: feature extraction for unpreprocessed microarray data. *Artif Intell Med,* 53**,** 47-56.

TSAI, K. Y. & WANG, F. S. 2005. Evolutionary optimization with data collocation for reverse engineering of biological networks. *Bioinformatics,* 21**,** 1180-8.

TURECI, O., SAHIN, U., VOLLMAR, E., SIEMER, S., GOTTERT, E., SEITZ, G., PARKKILA, A.-K., SHAH, G. N., GRUBB, J. H., PFREUNDSCHUH, M. & SLY, W. S. 1998. Human carbonic anhydrase XII: cDNA cloning, expression, and chromosomal localization of a carbonic anhydrase gene that is overexpressed in some renal cell cancers. *Proc. Natl. Acad. Sci.,* 95**,** 76078-7613.

VADREVU, S. H. & MURTY, S. U. 2010. A novel tool for classification of epidemiological data of vector-borne diseases. *J Glob Infect Dis. ,* 2**,** 35-8.

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. 2002. Gene expression profiling predicts clinical outcome of breast cancer. *Nature. ,* 415**,** 530-6.

VAN DE VIJVER, M. J., HE, Y. D., VAN'T VEER, L. J., DAI, H., HART, A. A. M., VOSKUIL, D. W., SCHREIBER, G. J., PETERSE, J. L., ROBERTS, C., MARTON, M. J., PARRISH, M., ATSMA, D., WITTEVEEN, A., GLAS, A., DELAHAYE, L., VAN DER VELDE, T., BARTELINK, H., RODENHUIS, S., RUTGERS, E. T., FRIEND, S. H. & BERNARDS, R. 2002. A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med,* 347**,** 1999-2009.

VAN REGENMORTEL, M. H. 2004. Reductionism and complexity in molecular biology. Scientists now have the tools to unravel biological and overcome the limitations of reductionism. *EMBO Rep,* 5**,** 1016-20.

VAZQUEZ, A., FLAMMINI, A., MARITAN, A. & VESPIGNANI, A. 2003. Global protein function prediction from protein-protein interaction networks. *Nat Biotechnol,* 21**,** 697-700.

VENDRELL, J. A., ROBERTSON, K. E., RAVEL, P., BRAY, S. E., BAJARD, A., PURDIE, C. A., NGUYEN, C., HADAD, S. M., BIECHE, I., CHABAUD, S., BACHELOT, T., THOMPSON, A. M. & COHEN, P. A. 2008. A candidate molecular signature associated with tamoxifen failure in primary breast cancer. *Breast Cancer Res,* 10**,** R88.

VODUC, D., CHEANG, M. & NIELSEN, T. 2008. GATA-3 expression in breast cancer has a strong association with estrogen receptor but lacks independent prognostic value. *Cancer Epidemiol Biomarkers Prev,* 17**,** 365-73.

WANG, R. S., WANG, Y., ZHANG, X. S. & CHEN, L. 2007. Inferring transcriptional regulatory networks from high-throughput data. *Bioinformatics,* 23**,** 3056-64.

WANG, Y., KLIJN, J. G., ZHANG, Y., SIEUWERTS, A. M., LOOK, M. P., YANG, F., TALANTOV, D., TIMMERMANS, M., MEIJER-VAN GELDER, M. E., YU, J., JATKOE, T., BERNS, E. M., ATKINS, D. & FOEKENS, J. A. 2005. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet,* 365**,** 671-9.

WATANABE, A., OGIWARA, H., EHATA, S., MUKASA, A., ISHIKAWA, S., MAEDA, D., UEKI, K., INO, Y., T., T., TAMADA, Y., FUKAYAMA, M., SAITO, N., MIYAZONO, K. & ABURATANI, H. 2011. Homozygously deleted gene DACH1 regulates tumor-initiating activity of glioma cells. *PNAS,* 108**,** 12384-12389.

WATSON, P. H., WYKOFF, C. C., HAN, C., LEEK, R. D., SLY, W. S., GATTER, K. C., RATCLIFFE, P. & HARRIS, A. L. 2003. Carbonic anhydrase XII is a marker of good prognosis in invasive breast carcinoma. *Br J Cancer,* 88**,** 1065-70.

WEBSTER, L. R., BILOUS, A. M., WILLIS, L., BYTH, K., BURGEMEISTER, F. C., SALISBURY, E. L. C., CLARKE, C. L. & BALLEINE, R. L. 2005. Histopathologic indicators of breast cancer biology: insights from population mammographic screening. *Br J Cancer.,* 92**,** 1366-71.

WEI, G., ABRAHAM, B. J., YAGI, R., JOTHI, R., CUI, K., SHARMA, S., NARLIKAR, L., NORTHRUP, D. L., TANG, Q., PAUL, W. E., ZHU, J. & ZHAO, K. 2011. Genome-wide analyses of transcription factor GATA3-mediated gene regulation in distinct T cell types. *Immunity,* 35**,** 299-311.

WEST, M., BLANCHETTE, C., DRESSMAN, H., HUANG, E., ISHIDA, S., SPANG, R., ZUZAN, H., OLSON, J. A., JR., MARKS, J. R. & NEVINS, J. R. 2001. Predicting the clinical status of human breast cancer by using gene expression profiles. *Proc Natl Acad Sci U S A,* 98**,** 11462-7.

WHITE, J., KEARINS, O., DODWELL, D., HORGAN, K., HANBY, A. M. & SPEIRS, V. 2011. Male breast carcinoma: increased awareness needed. *Breast Cancer Res,* 13**,** 219.

WIDROW, B. & HOFF, M. E. 1960. Adaptive Switching Circuits.

WILCOX, R. H. 1961. Adaptive control processes—A guided tour, by Richard Bellman, Princeton University Press, Princeton, New Jersey, 1961, 255 pp., $6.50. *Naval Research Logistics Quarterly,* 8**,** 315-316.

WILLIAMS, R. J. & ZIPSER, D. 1989. A Learning Algorithm for Continually Running Fully Recurrent Neural Networks. *Neural Computation,* 1**,** 270-280.

WILMING, L. G., GILBERT, J. R., HOWE, K., TREVANION, S., HUBBARD, T. & HARROW, J. L. 2008. The vertebrate genome annotation (Vega) database. *Nucleic Acids Res,* 36(Database issue)**,** D753-60.

WILSON, B. J. & GIGUERE, V. 2008. Meta-analysis of human cancer microarrays reveals GATA3 is integral to the estrogen receptor alpha pathway. *Mol Cancer,* 4**,** 49.

WOOD, T. L. & YEE, D. 2000. Introduction: IGFs and IGFBPs in the normal mammary gland and in breast cancer. *J Mammary Gland Biol Neoplasia.,* 5**,** 1-5.

WU, K., KATIYAR, S., LI, A., LIU, M., JU, X., POPOV, V. M., JIAO, X., LISANTI, M. P., CASOLA, A. & PESTELL, R. G. 2008. Dachshund inhibits oncogene-induced breast cancer cellular migration and invasion through suppression of interleukin-8. *Proc Natl Acad Sci U S A,* 105**,** 6924-9.

WU, K., LI, A., RAO, M., LIU, M., DAILEY, V., YANG, Y., DI VIZIO, D., WANG, C., LISANTI, M. P., SAUTER, G., RUSSELL, R. G., CVEKL, A. & PESTELL, R. G. 2006. DACH1 is a cell fate determination factor that inhibits cyclin D1 and breast tumor growth. *Mol Cell Biol,* 26**,** 7116-29.

WU, K., YANG, Y., WANG, C., DAVOLI, M. A., D'AMICO, M., LI, A., CVEKLOVA, K., KOZMIK, Z., LISANTI, M. P., RUSSELL, R. G., CVEKL, A. & PESTELL, R. G. 2003. DACH1 inhibits transforming growth factor-beta signaling through binding Smad4. *J Biol Chem,* 278**,** 51673-84.

WYKOFF, C. C., BEASLEY, N., WATSON, P. H., CAMPO, L., CHIA, S. K., ENGLISH, R., PASTOREK, J., SLY, W. S., RATCLIFFE, P. & HARRIS, A. L. 2001. Expression of the hypoxia-inducible and tumor-associated carbonic anhydrases in ductal carcinoma in situ of the breast. *Am J Pathol,* 158**,** 1011-9.

WYKOFF, C. C., BEASLEY, N. J., WATSON, P. H., TURNER, K. J., PASTOREK, J., SIBTAIN, A., WILSON, G. D., TURLEY, H., TALKS, K. L., MAXWELL, P. H., PUGH, C. W., RATCLIFFE, P. J. & HARRIS, A. L. 2000. Hypoxia-inducible expression of tumor-associated carbonic anhydrases. *Cancer Res,* 60**,** 7075-83.

XIONG, H. & CHEN, X.-W. 2006. Kernel-based distance metric learning for microarray data classification. *BMC Bioinformatics,* 7**,** 299.

XU, Q.-S. & LIANG, Y.-Z. 2001. Monte Carlo cross validation. *Chemometrics and Intelligent Laboratory Systems,* 56**,** 1-11.

XU, X., WANG, L. & DING, D. 2004. Learning module networks from genome-wide location and expression data. *FEBS Lett,* 578**,** 297-304.

YAN, W., CAO, Q. J., ARENAS, R. B., BENTLEY, B. & SHAO, R. 2010. GATA3 inhibits breast cancer metastasis through the reversal of epithelial-mesenchymal transition. *J Biol Chem,* 285**,** 14042-51.

YOSHIDA, N., OMOTO, Y., INOUE, A., EGUCHI, H., KOBAYASHI, Y., KUROSUMI, M., SAJI, S., SUEMASU, K., OKAZAKI, T., NAKACHI, K., FUJITA, T. & HAYASHI, S. 2004. Prediction of prognosis of estrogen receptor-positive breast cancer with combination of selected estrogen-regulated genes. *Cancer Sci,* 95**,** 496-502.

YU, K., LEE, C. H., TAN, P. H., HONG, G. S., WEE, S. B., WONG, C. Y. & TAN, P. 2004. A molecular signature of the Nottingham prognostic index in breast cancer. *Cancer Res.,* 64**,** 2962-8.

YU, W., KONG, T., BEAUDRY, S., TRAN, M., NEGORO, H., YANAMADALA, V. & DENKER, B. M. 2010. Polycystin-1 protein level determines activity of the Galpha12/JNK apoptosis pathway. *J Biol Chem,* 285**,** 10243-51.

ZAHIRIEH, A., NESBIT, M. A., ALI, A., WANG, K., HE, N., STANGOU, M., BAMICHAS, G., SOMBOLOS, K., THAKKER, R. V. & PEI, Y. 2005. Functional analysis of a novel GATA3 mutation in a family with the hypoparathyroidism, deafness, and renal dysplasia syndrome. *J Clin Endocrinol Metab,* 90**,** 2445-50.

ZHANG, X., TWOROGER, S. S., ELIASSEN, A. H. & HANKINSON, S. E. 2013. Postmenopausal plasma sex hormone levels and breast cancer risk over 20 years of follow-up. *Breast Cancer Res Treat.,* 137**,** 1573-7217 (Electronic).

ZHAO, G. Y., LI, Z. Y., ZOU, H. L., HU, Z. L., SONG, N. N., ZHENG, M. H., SU, C. J. & DING, Y. Q. 2008. Expression of the transcription factor GATA3 in the postnatal mouse central nervous system. *Neurosci Res,* 61**,** 420-8.

ZHOU, J., WANG, C., WANG, Z., DAMPIER, W., WU, K., CASIMIRO, M. C., CHEPELEV, I., POPOV, V. M., QUONG, A., TOZEREN, A., ZHAO, K., LISANTI, M. P. & PESTELL, R. G. 2010. Attenuation of Forkhead signaling by the retinal determination factor DACH1. *Proc Natl Acad Sci U S A,* 107**,** 6864-9.

ZHOU, Z. H., JIANG, Y., YANG, Y. B. & CHEN, S. F. 2002. Lung cancer cell identification based on artificial neural network ensembles. *Artif Intell Med. ,* 24**,** 25-36.

ZWART, W., THEODOROU, V. & CARROLL, J. S. 2011. Estrogen receptor-positive breast cancer: a multidisciplinary challenge. *Wiley Interdiscip Rev Syst Biol Med.***,** 216-30.

APPENDIX 1

Curated list of interacting molecules with ESR1 interaction databases:

Sl. No.	Gene Symbol	Database	Reference
1	ABCC5	GeneCards	I2D: score=0.14
2	ACTB	BioGrid	Ambrosino C (2010)
3	ACTC1	BioGrid	Ambrosino C (2010)
4	ACTN4	BioGrid	Khurana S (2011)
5	ACTR2	BioGrid	Ambrosino C (2010)
6	ACTR3	BioGrid	Ambrosino C (2010)
7	AHR	BioGrid	Klinge CM (2000)
8	AHRR	BioGrid	Klinge CM (2000)
9	AKAP13	BioGrid	Rubino D (1998)
10	AKT1	BioGrid	Campbell RA (2001)
11	AKT2	BioGrid	Sun M (2001)
12	AP1B1	GeneCards	STRING: ENSP00000350199
13	AR	GeneCards	I2D: score=1.00 STRING: ENSP00000363822
14	ARG1	GeneCards	I2D: score=0.09
15	ARHGDIA	BioGrid	Barone I (2011)
16	ARHGEF15	GeneCards	I2D: score=0.19
17	ARNT	GeneCards	I2D: score=1.00 STRING: ENSP00000351407
18	ASF1A	GeneCards	EBI-78473,EBI-749553 STRING: ENSP00000229595
19	ASH2L	BioGrid	Shi L (2011)
20	ASPG	GeneCards	STRING: ENSP00000389003
21	ATAD2	BioGrid	Zou JX (2007)
22	ATAD3C	GeneCards	I2D: score=0.22
23	ATP5J	GeneCards	STRING: ENSP00000389649
24	BAG1	BioGrid	Zeiner M (1995)
25	BARD1	BioGrid	Dizin E (2010)
26	BAZ1B	BioGrid	Kitagawa H (2003)
27	BCAR1	BioGrid	Cabodi S (2004)
28	BCAS2	GeneCards	I2D: score=0.41 STRING: ENSP00000358554
29	BDNF	GeneCards	I2D: score=0.98 STRING: ENSP00000414303
30	BLCAP	GeneCards	EBI-78473,EBI-3895726
31	BLOC1S1	BioGrid	Mishra SK (2003)
32	BRCA1	BioGrid	Wang C (2005)
33	BTF3	GeneCards	I2D: score=0.70 STRING: ENSP00000369965
34	C16orf53	GeneCards	MINT-7013584
35	CALM1	BioGrid	Garcia Pedrero JM (2002)

36	CALM2	GeneCards	I2D: score=0.99
37	CALM3	GeneCards	I2D: score=0.99
38	CAV1	BioGrid	Schlegel A (2001)
39	CBLL1	BioGrid	Gong EY (2010)
40	CCNC	BioGrid	Kang YK (2002)
41	CCND1	BioGrid	Zwijsen RM (1997)
42	CCNH	BioGrid	Talukder AH (2003)
43	CCNT1	GeneCards	I2D: score=0.83 STRING: ENSP00000261900
44	CDC25B	BioGrid	Ma ZQ (2001)
45	CDK11B	BioGrid	Wang Y (2009)
46	CDK7	BioGrid	Wang Y (2009)
47	CDK8	BioGrid	Wang Y (2009)
48	CDKN1A	BioGrid	Maynadier M (2008)
49	CEBPA	BioGrid	Boruk M (1998)
50	CEBPB	BioGrid	Stein B (1995)
51	CHD3	BioGrid	Okada M (2008)
52	CHD9	GeneCards	I2D: score=0.71 STRING: ENSP00000381522
53	CHUK	GeneCards	I2D: score=0.98 STRING: ENSP00000359424
54	CITED1	BioGrid	Yahata T (2001)
55	COBRA1	BioGrid	Aiyar SE (2004)
56	COL18A1	GeneCards	STRING: ENSP00000352798
57	COL1A2	GeneCards	STRING: ENSP00000297268
58	COPS5	BioGrid	Callige M (2005)
59	CPEB1	GeneCards	STRING: ENSP00000414187
60	CREBBP	BioGrid	Heery DM (1997)
61	CRIPAK	GeneCards	I2D: score=0.44 STRING: ENSP00000323978
62	CSNK2B	GeneCards	I2D: score=0.98 STRING: ENSP00000365025
63	CTNNB1	BioGrid	Kouzmenko AP (2004)
64	CTSD	GeneCards	STRING: ENSP00000236671
65	CUEDC2	BioGrid	Pan X (2011)
66	CUL3	BioGrid	Byun B (2008)
67	CUL4B	GeneCards	I2D: score=0.33 STRING: ENSP00000384109
68	DAP3	BioGrid	Hulkko SM (2000)
69	DDX17	BioGrid	Fujita T (2003)
70	DDX3X	BioGrid	Fujita T (2003)
71	DDX5	BioGrid	Fujita T (2003)
72	DDX54	BioGrid	Rajendran RR (2003)
73	DNM1L	GeneCards	EBI-78473,EBI-724571
74	DNTTIP2	BioGrid	Bu H (2004)
75	DSCAM	GeneCards	STRING: ENSP00000383303
76	DUT	BioGrid	Albers M (2005)

77	E2F1	BioGrid	Liu W (2012)
78	EBAG9	GeneCards	I2D: score=0.12
79	EGFR	GeneCards	MINT-6169667 I2D: score=1.00 STRING: ENSP00000275493
80	EHMT2	BioGrid	Purcell DJ (2011)
81	EIF3I	GeneCards	I2D: score=0.33 STRING: ENSP00000362688
82	ELAVL1	GeneCards	I2D: score=0.21
83	EP300	BioGrid	Fujita T (2003)
84	ERBB2	BioGrid	Yang Z (2004)
85	EREG	GeneCards	STRING: ENSP00000244869
86	ESR1	BioGrid	Ikeda M (2002)
87	ESR2	BioGrid	Monroe DG (2003)
88	ESRRA	GeneCards	I2D: score=0.99 STRING: ENSP00000000442
89	ESRRB	GeneCards	STRING: ENSP00000261532
90	ESRRG	GeneCards	STRING: ENSP00000386171
91	EXOSC4	GeneCards	I2D: score=0.12
92	EZH2	GeneCards	I2D: score=0.35
93	FHL1	GeneCards	I2D: score=0.68 STRING: ENSP00000359724
94	FHL2	BioGrid	Kobayashi S (2004)
95	FKBP4	BioGrid	Nair SC (1996)
96	FKBP5	BioGrid	Nair SC (1996)
97	FLII	BioGrid	Jeong KW (2009)
98	FOS	BioGrid	Baron S (2007)
99	FOXA1	GeneCards	STRING: ENSP00000250448
100	FOXL2	GeneCards	I2D: score=0.90 STRING: ENSP00000333188
101	FOXM1	GeneCards	STRING: ENSP00000342307
102	FOXO1	BioGrid	Zhao HH (2001)
103	FOXO3	GeneCards	I2D: score=0.98 STRING: ENSP00000339527
104	FOXO4	BioGrid	Schuur ER (2001)
105	GADD45A	BioGrid	Yi YW (2000)
106	GADD45B	BioGrid	Yi YW (2000)
107	GADD45G	BioGrid	Yi YW (2000)
108	GNA13	GeneCards	STRING: ENSP00000400717
109	GNA14	GeneCards	STRING: ENSP00000365807
110	GNA15	GeneCards	STRING: ENSP00000262958
111	GNAI1	GeneCards	I2D: score=0.13 STRING: ENSP00000343027
112	GNAI2	GeneCards	STRING: ENSP00000312999
113	GNAL	GeneCards	STRING: ENSP00000334051
114	GNAO1	GeneCards	STRING: ENSP00000262493
115	GNAZ	GeneCards	STRING: ENSP00000248996
116	GNG2	GeneCards	STRING: ENSP00000334448
117	GRB2	GeneCards	I2D: score=1.00 STRING: ENSP00000339007

118	GRIP1	BioGrid	Yi YW (2000)
119	GSN	BioGrid	Yi YW (2000)
120	GTF2B	BioGrid	Yi YW (2000)
121	GTF2H1	BioGrid	Yi YW (2000)
122	HAX1	GeneCards	EBI-78473,EBI-357001
123	HDAC1	BioGrid	Kawai H (2003)
124	HDAC2	BioGrid	Itoh Y (2007)
125	HDAC3	BioGrid	Liu XF (2004)
126	HDAC4	BioGrid	Leong H (2005)
127	HDAC5	BioGrid	van Rooij E (2010)
128	HDAC7	BioGrid	Malik S (2010)
129	HDAC9	BioGrid	van Rooij E (2010)
130	HEXIM1	GeneCards	I2D: score=0.32 STRING: ENSP00000328773
131	HNF4A	GeneCards	I2D: score=1.00 STRING: ENSP00000312987
132	HNF4G	GeneCards	STRING: ENSP00000379701
133	HSP90AA1	BioGrid	Nair SC (1996)
134	HSP90AB1	GeneCards	I2D: score=1.00 STRING: ENSP00000325875
135	HSPA4	BioGrid	Nair SC (1996)
136	HSPA8	BioGrid	Tateishi Y (2004)
137	IGF1	GeneCards	STRING: ENSP00000302665
138	IGF1R	BioGrid	Santen RJ (2005)
139	IL25	GeneCards	I2D: score=0.21
140	ING1	BioGrid	Toyama T (2003)
141	IRS1	BioGrid	Morelli C (2003)
142	IRS2	BioGrid	Morelli C (2003)
143	ISL1	BioGrid	Gay F (2000)
144	ITGB3BP	GeneCards	I2D: score=0.90 STRING: ENSP00000271002
145	JUN	BioGrid	Teyssier C (2001)
146	JUNB	GeneCards	I2D: score=0.98 STRING: ENSP00000303315
147	JUND	GeneCards	I2D: score=0.92 STRING: ENSP00000252818
148	KAT2A	BioGrid	Oishi H (2006)
149	KAT2B	GeneCards	STRING: ENSP00000263754
150	KAT5	BioGrid	Gaughan L (2001)
151	KAT6A	BioGrid	Yin H (2007)
152	KDM1A	BioGrid	Garcia-Bassets I (2007)
153	KDM4B	BioGrid	Shi L (2011)
154	KDM5A	BioGrid	Chan SW (2001)
155	KDM6B	GeneCards	MINT-8286415 MINT-8286424
156	KIF11	GeneCards	STRING: ENSP00000260731
157	KIF1A	BioGrid	Albers M (2005)
158	KLF5	BioGrid	Guo P (2010)

159	LCK	GeneCards	I2D: score=0.99 STRING: ENSP00000337825
160	LCOR	BioGrid	Fernandes I (2003)
161	LDB1	BioGrid	Johnsen SA (2009)
162	LINC00312	GeneCards	MINT-50571 MINT-50573 MINT-50906 I2D: score=0.04
163	LMO4	BioGrid	Singh RR (2005)
164	LOC100505603	GeneCards	MINT-14693 MINT-14694 I2D: score=0.87
165	LRIF1	BioGrid	Li HJ (2007)
166	LYZ	GeneCards	I2D: score=0.12
167	MAD2L1	GeneCards	STRING: ENSP00000296509
168	MAP3K1	GeneCards	MINT-50709 I2D: score=0.28 STRING: ENSP00000382423
169	MAPK1	BioGrid	Metivier R (2002)
170	MAPK11	GeneCards	STRING: ENSP00000333685
171	MAPK14	GeneCards	I2D: score=0.98 STRING: ENSP00000229794
172	MAPK3	GeneCards	I2D: score=0.99 STRING: ENSP00000263025
173	MBD2	BioGrid	Chatagnon A (2010)
174	MDM2	BioGrid	Duong V (2007)
175	MED1	BioGrid	Fujita T (2003)
176	MED10	BioGrid	Fujita T (2003)
177	MED12	BioGrid	Kang YK (2002)
178	MED13	BioGrid	Kang YK (2002)
179	MED14	BioGrid	Burakov D (2000)
180	MED16	BioGrid	Kang YK (2002)
181	MED17	BioGrid	Kang YK (2002)
182	MED20	BioGrid	Kang YK (2002)
183	MED21	BioGrid	Kang YK (2002)
184	MED23	BioGrid	Kang YK (2002)
185	MED24	BioGrid	Kang YK (2002)
186	MED25	BioGrid	Lee HK (2007)
187	MED30	GeneCards	I2D: score=0.92
188	MED4	GeneCards	I2D: score=0.92
189	MED6	BioGrid	Lee HK (2007)
190	MED7	BioGrid	Lee HK (2007)
191	MEN1	BioGrid	Dreijerink KM (2006)
192	MGMT	BioGrid	Teo AK (2001)
193	MKNK2	GeneCards	I2D: score=0.97 STRING: ENSP00000250896
194	MLL2	BioGrid	Shi L (2011)
195	MMS19	GeneCards	I2D: score=0.98 STRING: ENSP00000359818
196	MNAT1	BioGrid	Talukder AH (2003)
197	MPG	BioGrid	Likhite VS (2004)
198	MRPS15	GeneCards	I2D: score=0.02
199	MRPS17	GeneCards	I2D: score=0.05

200	MRPS2	GeneCards	I2D: score=0.03
201	MRPS22	GeneCards	I2D: score=0.25
202	MRPS27	GeneCards	I2D: score=0.07
203	MRPS31	GeneCards	I2D: score=0.28
204	MRPS35	GeneCards	I2D: score=0.03
205	MRPS6	GeneCards	I2D: score=0.04
206	MSH2	BioGrid	Wada-Hiraike O (2005)
207	MSH6	BioGrid	Oishi H (2006)
208	MTA1	BioGrid	Mishra SK (2003)
209	MTA2	BioGrid	Okada M (2008)
210	MTA3	BioGrid	Okada M (2008)
211	MTCH2	BioGrid	Albers M (2005)
212	MUC1	BioGrid	Wei X (2006)
213	MVP	BioGrid	Abbondanza C (1998)
214	MYC	BioGrid	Cheng AS (2006)
215	MYL6	BioGrid	Ambrosino C (2010)
216	MYLK2	GeneCards	I2D: score=0.08
217	MYO1C	BioGrid	Ambrosino C (2010)
218	MYOD1	BioGrid	Jin W (2008)
219	MYSM1	GeneCards	STRING: ENSP00000418734
220	NCOA1	BioGrid	Kraichely DM (2000)
221	NCOA2	BioGrid	Kraichely DM (2000)
222	NCOA3	BioGrid	Suen CS (1998)
223	NCOA4	BioGrid	Alen P (1999)
224	NCOA6	BioGrid	Mahajan MA (2000)
225	NCOA7	BioGrid	Lazennec G (1997)
226	NCOR1	BioGrid	Fujita T (2003)
227	NCOR2	BioGrid	Fujita T (2003)
228	NDRG2	GeneCards	EBI-78473,EBI-3895741
229	NDUFV3	GeneCards	STRING: ENSP00000346196
230	NFKB1	GeneCards	I2D: score=1.00 STRING: ENSP00000226574
231	NGG1	BioGrid	Benecke A (2002)
232	NOS3	GeneCards	I2D: score=0.96
233	NPM1	BioGrid	Ambrosino C (2010)
234	NPPA	BioGrid	Albers M (2005)
235	NR0B1	GeneCards	I2D: score=1.00 STRING: ENSP00000368253
236	NR0B2	BioGrid	Seol W (1998)
237	NR1D1	GeneCards	STRING: ENSP00000246672
238	NR1D2	GeneCards	STRING: ENSP00000310006
239	NR1H2	GeneCards	STRING: ENSP00000253727
240	NR1H4	GeneCards	I2D: score=0.97 STRING: ENSP00000315442

241	NR1I2	BioGrid	Seol W (1998)
242	NR2C1	BioGrid	Hu YC (2002)
243	NR2C2	BioGrid	Shyr CR (2002)
244	NR2C2AP	GeneCards	STRING: ENSP00000332823
245	NR2E1	GeneCards	STRING: ENSP00000357982
246	NR2F1	BioGrid	Metivier R (2002)
247	NR2F6	BioGrid	Zhu XG (2000)
248	NR4A1	GeneCards	STRING: ENSP00000243050
249	NR4A2	GeneCards	STRING: ENSP00000344479
250	NR5A1	GeneCards	STRING: ENSP00000362690
251	NRBP1	GeneCards	STRING: ENSP00000233557
252	NRIP1	BioGrid	Heery DM (1997)
253	NSD1	BioGrid	Huang N (1998)
254	OTUB1	BioGrid	Stanisic V (2009)
255	PABPC5	GeneCards	I2D: score=0.24
256	PAK1	GeneCards	I2D: score=0.98 STRING: ENSP00000278568
257	PAK6	BioGrid	Lee SR (2002)
258	PARP1	BioGrid	Ju BG (2006)
259	PDLIM1	BioGrid	Johnsen SA (2009)
260	PELP1	BioGrid	Vadlamudi RK (2001)
261	PGC	BioGrid	Bourdoncle A (2005)
262	PGR	BioGrid	Giulianelli S (2012)
263	PHB	BioGrid	Giulianelli S (2012)
264	PHB2	BioGrid	Martini PG (2003)
265	PIAS1	BioGrid	Kobayashi S (2004)
266	PIAS2	BioGrid	Kotaja N (2000)
267	PIAS3	GeneCards	I2D: score=0.99 STRING: ENSP00000376765
268	PIK3CA	GeneCards	I2D: score=0.96 STRING: ENSP00000263967
269	PIK3R1	BioGrid	Castoria G (2001)
270	PIK3R2	BioGrid	Cabodi S (2004)
271	PIK3R3	GeneCards	I2D: score=0.96 STRING: ENSP00000262741
272	PLA2G7	GeneCards	I2D: score=0.05
273	PNRC2	BioGrid	Albers M (2005)
274	POLR1B	GeneCards	I2D: score=0.40 STRING: ENSP00000263331
275	POLR2A	BioGrid	Ambrosino C (2010)
276	POLR2D	GeneCards	I2D: score=0.16
277	POU2F1	BioGrid	Prefontaine GG (1999)
278	POU2F2	BioGrid	Prefontaine GG (1999)
279	POU4F1	BioGrid	Budhram-Mahadeo V (1998)
280	POU4F2	BioGrid	Budhram-Mahadeo V (1998)
281	PPARA	GeneCards	STRING: ENSP00000262735

282	PPARD	GeneCards	STRING: ENSP00000310928
283	PPARG	BioGrid	DiRenzo J (1997)
284	PPARGC1A	BioGrid	DiRenzo J (1997)
285	PPARGC1B	BioGrid	DiRenzo J (1997)
286	PPID	BioGrid	Nair SC (1996)
287	PPP1CC	BioGrid	Flores-Delgado G (2007)
288	PPP5C	GeneCards	EBI-78473,EBI-716663 I2D: score=0.05 STRING: ENSP00000012443
289	PRDM2	BioGrid	Abbondanza C (2000)
290	PRKACA	GeneCards	I2D: score=0.98 STRING: ENSP00000309591
291	PRKCZ	BioGrid	Yi P (2008)
292	PRKDC	BioGrid	Ju BG (2006)
293	PRMT1	BioGrid	Koh SS (2001)
294	PRMT2	BioGrid	Qi C (2002)
295	PRPF6	GeneCards	I2D: score=0.99 STRING: ENSP00000266079
296	PSEN2	GeneCards	STRING: ENSP00000355747
297	PSMB9	GeneCards	MINT-3291999 I2D: score=0.07
298	PSMC5	BioGrid	vom Baur E (1996)
299	PSMD1	GeneCards	I2D: score=0.96 STRING: ENSP00000309474
300	PTCD3	GeneCards	I2D: score=0.05
301	PTEN	BioGrid	Lin HK (2004)
302	PTGES3	BioGrid	Nair SC (1996)
303	PTMA	BioGrid	Martini PG (2003)
304	PTPN1	GeneCards	MINT-6761899 I2D: score=0.99 STRING: ENSP00000360683
305	PTPN6	GeneCards	MINT-6761906 I2D: score=0.98 STRING: ENSP00000391592
306	RABGEF1	BioGrid	La Rosa P (2011)
307	RAC3	GeneCards	EBI-78473,EBI-767084
308	RAN	BioGrid	Castoria G (2012)
309	RARA	GeneCards	I2D: score=1.00 STRING: ENSP00000254066
310	RARB	GeneCards	STRING: ENSP00000332296
311	RARG	GeneCards	STRING: ENSP00000332695
312	RBBP4	BioGrid	Creekmore AL (2008)
313	RBBP5	BioGrid	Shi L (2011)
314	RBBP6	BioGrid	Peidis P (2010)
315	RBBP7	BioGrid	Creekmore AL (2008)
316	RBFOX2	BioGrid	Norris JD (2002)
317	RBM10	GeneCards	STRING: ENSP00000366829
318	RBM23	GeneCards	I2D: score=0.62 STRING: ENSP00000352956
319	RBM39	BioGrid	Jung DJ (2002)
320	RELA	BioGrid	Nettles KW (2008)
321	REXO4	BioGrid	Montano MM (2000)
322	RFX4	GeneCards	I2D: score=0.35 STRING: ENSP00000376585

323	RGS3	GeneCards	I2D: score=0.11 STRING: ENSP00000259406
324	RLIM	BioGrid	Johnsen SA (2009)
325	RNF14	GeneCards	I2D: score=0.95 STRING: ENSP00000324956
326	RNF4	BioGrid	Moilanen AM (1998)
327	RORA	GeneCards	STRING: ENSP00000261523
328	RORB	GeneCards	STRING: ENSP00000366093
329	RORC	GeneCards	STRING: ENSP00000327025
330	RPL18	BioGrid	Moilanen AM (1998)
331	RPL36AL	GeneCards	I2D: score=0.08
332	RPL7	BioGrid	Ambrosino C (2010)
333	RPL7A	BioGrid	Ambrosino C (2010)
334	RPLP0	BioGrid	Ambrosino C (2010)
335	RPLP0P6	GeneCards	I2D: score=0.02
336	RPS11	GeneCards	I2D: score=0.08
337	RPS18	GeneCards	I2D: score=0.10
338	RPS20	GeneCards	I2D: score=0.22
339	RPS4X	BioGrid	Ambrosino C (2010)
340	RPS6KA1	BioGrid	Joel PB (1998)
341	RPS6KA3	BioGrid	Clark DE (2001)
342	RPS8	BioGrid	Ambrosino C (2010)
343	RPS9	BioGrid	Ambrosino C (2010)
344	RXRA	BioGrid	DiRenzo J (1997)
345	RXRG	GeneCards	STRING: ENSP00000352900
346	SAFB	BioGrid	Oesterreich S (2000)
347	SAFB2	BioGrid	Oesterreich S (2000)
348	SENP5	GeneCards	EBI-78473,EBI-3895753
349	SETD7	BioGrid	Subramanian K (2008)
350	SF1	BioGrid	Zhou D (2001)
351	SGK3	BioGrid	Wang Y (2011)
352	SHC1	BioGrid	Song RX (2002)
353	SIN3A	BioGrid	Ellison-Zelski SJ (2009)
354	SIRT1	BioGrid	Elangovan S (2011)
355	SKI	BioGrid	Elangovan S (2011)
356	SKIL	BioGrid	Band AM (2011)
357	SLC30A9	GeneCards	I2D: score=0.92 STRING: ENSP00000264451
358	SMAD2	BioGrid	Ito I (2010)
359	SMAD3	BioGrid	Ito I (2010)
360	SMAD4	GeneCards	I2D: score=1.00 STRING: ENSP00000341551
361	SMARCA2	BioGrid	Ichinose H (1997)
362	SMARCA4	BioGrid	Ichinose H (1997)
363	SMARCB1	BioGrid	Okada M (2008)

364	SMARCC2	BioGrid	Okada M (2008)
365	SMARCD1	BioGrid	Hsiao PW (2003)
366	SMARCD3	GeneCards	I2D: score=0.93 STRING: ENSP00000262188
367	SMARCE1	BioGrid	Kiskinis E (2006)
368	SMURF1	BioGrid	Ito I (2010)
369	SMYD3	BioGrid	Kim H (2009)
370	SOD1	GeneCards	STRING: ENSP00000270142
371	SOS1	GeneCards	I2D: score=0.95 STRING: ENSP00000384675
372	SOS2	BioGrid	Yang Z (2004)
373	SP1	BioGrid	Petz LN (2004)
374	SP2	BioGrid	He S (2005)
375	SP3	BioGrid	Dong J (2006)
376	SPOP	BioGrid	Byun B (2008)
377	SRA1	BioGrid	Watanabe M (2001)
378	SRC	BioGrid	Monroe DG (2003)
379	SREBF1	GeneCards	I2D: score=0.96 STRING: ENSP00000348069
380	STAT3	BioGrid	Wang LH (2001)
381	STAT5A	BioGrid	Faulds MH (2001)
382	STAU1	GeneCards	I2D: score=0.30
383	STK11	GeneCards	STRING: ENSP00000324856
384	STRN	GeneCards	I2D: score=0.84 STRING: ENSP00000263918
385	STUB1	BioGrid	Tateishi Y (2004)
386	SUMO1	BioGrid	Sentis S (2005)
387	SVIL	BioGrid	Ting HJ (2002)
388	TAB2	BioGrid	Cutrupi S (2012)
389	TADA3	BioGrid	Li CW (2010)
390	TAF10	GeneCards	I2D: score=1.00 STRING: ENSP00000299424
391	TAF1A	BioGrid	Loven MA (2003)
392	TAF1B	BioGrid	Loven MA (2003)
393	TAF2	BioGrid	Suen CS (1998)
394	TAP1	GeneCards	I2D: score=0.93 STRING: ENSP00000346206
395	TBP	BioGrid	Lazennec G (1997)
396	TCF20	GeneCards	I2D: score=0.97 STRING: ENSP00000352463
397	TDG	BioGrid	Chen D (2003)
398	TFF1	BioGrid	Yi P (2008)
399	THRA	GeneCards	I2D: score=1.00 STRING: ENSP00000264637
400	THRB	GeneCards	I2D: score=1.00 STRING: ENSP00000348827
401	TNFRSF14	GeneCards	I2D: score=0.96 STRING: ENSP00000347948
402	TOP2B	BioGrid	Ju BG (2006)
403	TP53	BioGrid	Liu G (2000)
404	TRAM1	BioGrid	Kraichely DM (2000)

405	TRIM24	BioGrid	vom Baur E (1996)
406	TRIM25	BioGrid	Nakajima A (2007)
407	TRIM28	GeneCards	I2D: score=0.97 STRING: ENSP00000253024
408	TRIM59	GeneCards	I2D: score=0.87 STRING: ENSP00000311219
409	TRIP4	BioGrid	Kim HJ (1999)
410	TRRAP	BioGrid	Fujita T (2003)
411	TSC2	GeneCards	I2D: score=0.94 STRING: ENSP00000219476
412	TUBA1A	BioGrid	Benecke A (2002)
413	TUBA1B	GeneCards	I2D: score=0.14 STRING: ENSP00000336799
414	TUBB	GeneCards	I2D: score=0.04
415	TUBB1	BioGrid	Benecke A (2002)
416	TUBB2A	GeneCards	STRING: ENSP00000369703
417	UBC	BioGrid	Stanisic V (2009)
418	UBE2I	BioGrid	Kobayashi S (2004)
419	UBE3A	BioGrid	Sun J (2012)
420	UIMC1	BioGrid	Yan J (2007)
421	USF1	GeneCards	I2D: score=0.75 STRING: ENSP00000356999
422	UTP14A	GeneCards	I2D: score=0.81
423	VAV3	GeneCards	I2D: score=0.98 STRING: ENSP00000359073
424	VDR	GeneCards	STRING: ENSP00000229022
425	WDR5	BioGrid	Shi L (2011)
426	WDR5B	GeneCards	I2D: score=0.18 STRING: ENSP00000330381
427	WIPI1	GeneCards	I2D: score=0.88 STRING: ENSP00000262139
428	XBP1	BioGrid	Ding L (2003)
429	XPO1	BioGrid	Castoria G (2012)
430	XRCC5	BioGrid	Ju BG (2006)
431	XRCC6	BioGrid	Ju BG (2006)
432	YWHAH	GeneCards	I2D: score=0.98 STRING: ENSP00000248975
433	YWHAQ	BioGrid	Zilliacus J (2001)
434	ZBTB16	BioGrid	Martin PJ (2003)
435	ZBTB17	GeneCards	I2D: score=0.95 STRING: ENSP00000364895
436	ZC3HAV1	GeneCards	I2D: score=0.17
437	ZNF398	BioGrid	Conroy AT (2002)
438	ZNHIT3	GeneCards	I2D: score=0.96

APPENDIX 2

Curated list of interacting molecules with ESR2 interaction databases:

Sl. No.	Gene Symbol	Database	Reference
1	ACAD11	GeneCards	I2D: score=0.01
2	ACSM2B	GeneCards	I2D: score=0.13
3	ACTB	GeneCards	I2D: score=0.06
4	ACTN1	GeneCards	I2D: score=0.07
5	ACTN3	GeneCards	I2D: score=0.09
6	ACTR3	GeneCards	I2D: score=0.06
7	AKAP13	GeneCards	I2D: score=0.98 STRING: ENSP00000354718
8	AKAP8	GeneCards	I2D: score=0.15
9	AKAP9	GeneCards	I2D: score=0.12
10	AKT1	GeneCards	I2D: score=0.07 STRING: ENSP00000270202
11	ANAPC5	GeneCards	I2D: score=0.06
12	ANKFY1	GeneCards	I2D: score=0.15
13	AP2A1	GeneCards	I2D: score=0.07
14	ARID5A	BioGRID	Georgescu SP (2005)
15	ARPC2	GeneCards	I2D: score=0.08
16	ASB2	GeneCards	I2D: score=0.08
17	ATG2A	GeneCards	I2D: score=0.03
18	BAIAP2L1	GeneCards	I2D: score=0.03
19	BAZ1A	GeneCards	I2D: score=0.29
20	BCAS2	GeneCards	I2D: score=0.07 STRING: ENSP00000358554
21	BCLAF1	GeneCards	I2D: score=0.19
22	BICD1	GeneCards	I2D: score=0.05
23	BRD8	GeneCards	I2D: score=0.66
24	C10orf47	GeneCards	I2D: score=0.02
25	C14orf43	GeneCards	I2D: score=0.63
26	C19orf68	GeneCards	I2D: score=0.02
27	C22orf28	GeneCards	I2D: score=0.03
28	C9orf114	GeneCards	I2D: score=0.03
29	CALM1	GeneCards	I2D: score=0.77 STRING: ENSP00000349467
30	CALM2	GeneCards	I2D: score=0.77
31	CALM3	GeneCards	I2D: score=0.77
32	CAND1	BioGRID	Bennett EJ (2010)
33	CCAR1	GeneCards	I2D: score=0.50
34	CCDC106	GeneCards	I2D: score=0.05

35	CCNC	BioGRID	Kang YK (2002)
36	CCNC	GeneCards	I2D: score=0.15 STRING: ENSP00000358222
37	CDC16	GeneCards	I2D: score=0.76
38	CDC5L	GeneCards	I2D: score=0.06
39	CDK8	BioGRID	Kang YK (2002)
40	CGN	GeneCards	I2D: score=0.02
41	CHM	GeneCards	I2D: score=0.05
42	CLASP2	GeneCards	I2D: score=0.07
43	CLCF1	GeneCards	I2D: score=0.01
44	CNN1	GeneCards	I2D: score=0.06
45	CNTRL	GeneCards	I2D: score=0.07
46	COIL	GeneCards	I2D: score=0.06
47	COPB2	GeneCards	I2D: score=0.02
48	COPG2	GeneCards	I2D: score=0.03
49	CPSF6	GeneCards	I2D: score=0.12
50	CPT1A	GeneCards	I2D: score=0.07
51	CREBBP	GeneCards	I2D: score=0.99 STRING: ENSP00000262367
52	CSE1L	GeneCards	I2D: score=0.07
53	CTNNB1	GeneCards	I2D: score=0.89
54	CTTN	GeneCards	I2D: score=0.05
55	CWC15	GeneCards	I2D: score=0.10
56	CWC27	GeneCards	I2D: score=0.03
57	DAP3	GeneCards	I2D: score=0.11 STRING: ENSP00000341692
58	DBN1	GeneCards	I2D: score=0.02
59	DDX17	GeneCards	I2D: score=0.93 STRING: ENSP00000380033
60	DDX39A	GeneCards	I2D: score=0.08
61	DDX3X	GeneCards	I2D: score=0.10
62	DDX41	GeneCards	I2D: score=0.07
63	DDX5	GeneCards	I2D: score=0.80 STRING: ENSP00000225792
64	DDX50	GeneCards	I2D: score=0.05
65	DDX54	BioGRID	Rajendran RR (2003)
66	DLGAP4	GeneCards	I2D: score=0.08
67	DNTTIP1	GeneCards	I2D: score=0.04
68	DNTTIP2	GeneCards	I2D: score=0.12 STRING: ENSP00000352137
69	DOK6	GeneCards	I2D: score=0.02
70	EDC4	GeneCards	I2D: score=0.15
71	EEF1A2	GeneCards	I2D: score=0.04
72	EFTUD2	GeneCards	I2D: score=0.05
73	EIF2A	GeneCards	I2D: score=0.04
74	EIF3I	BioGRID	Albers M (2005)
75	EP300	GeneCards	I2D: score=0.99 STRING: ENSP00000263253

76	EPB41L5	GeneCards	I2D: score=0.02
77	EPS15L1	GeneCards	I2D: score=0.06
78	ESR1	GeneCards	I2D: score=1.00 STRING: ENSP00000206249
79	ESR2	BioGRID	Powell E (2010)
80	ESRRA	GeneCards	STRING: ENSP00000000442
81	ESRRB	GeneCards	STRING: ENSP00000261532
82	ESRRG	GeneCards	STRING: ENSP00000386171
83	EXOSC10	GeneCards	I2D: score=0.03
84	EXOSC4	GeneCards	I2D: score=0.04
85	FAM120A	GeneCards	I2D: score=0.02
86	FAM179B	GeneCards	I2D: score=0.04
87	FAM82A2	GeneCards	I2D: score=0.07
88	FBLL1	GeneCards	I2D: score=0.11
89	FIP1L1	GeneCards	I2D: score=0.08
90	FLG	GeneCards	I2D: score=0.04
91	FLNB	GeneCards	I2D: score=0.03
92	FLNC	GeneCards	I2D: score=0.03
93	FMR1	GeneCards	I2D: score=0.06
94	FOXL2	GeneCards	I2D: score=0.07
95	FTSJ3	GeneCards	I2D: score=0.07
96	FXR2	GeneCards	I2D: score=0.09
97	GCC2	GeneCards	I2D: score=0.04
98	GFAP	GeneCards	I2D: score=0.07
99	GIGYF1	GeneCards	I2D: score=0.03
100	GNA14	GeneCards	STRING: ENSP00000365807
101	GNA15	GeneCards	STRING: ENSP00000262958
102	GNAL	GeneCards	STRING: ENSP00000334051
103	GNB2L1	GeneCards	I2D: score=0.08
104	GNB3	GeneCards	I2D: score=0.10
105	GNG2	GeneCards	STRING: ENSP00000334448
106	GNL2	GeneCards	I2D: score=0.05
107	GNL3	GeneCards	I2D: score=0.06
108	GOLGA3	GeneCards	I2D: score=0.07
109	GOLGA4	GeneCards	I2D: score=0.04
110	GSN	GeneCards	I2D: score=0.02
111	GSPT1	GeneCards	I2D: score=0.04
112	GTF3C1	GeneCards	I2D: score=0.12
113	GTF3C2	GeneCards	I2D: score=0.10
114	GTF3C3	GeneCards	I2D: score=0.14
115	GTF3C4	GeneCards	I2D: score=0.03
116	GTF3C5	GeneCards	I2D: score=0.11

117	GTPBP4	GeneCards	I2D: score=0.04
118	H2BFS	GeneCards	I2D: score=0.05
119	HCFC1	GeneCards	I2D: score=0.90
120	HCFC2	GeneCards	I2D: score=0.03
121	HEATR4	GeneCards	I2D: score=0.05
122	HIST1H1C	GeneCards	I2D: score=0.17
123	HIST1H4A	GeneCards	I2D: score=0.10
124	HIST1H4B	GeneCards	I2D: score=0.10
125	HIST1H4C	GeneCards	I2D: score=0.10
126	HIST1H4D	GeneCards	I2D: score=0.10
127	HIST1H4E	GeneCards	I2D: score=0.10
128	HIST1H4F	GeneCards	I2D: score=0.10
129	HIST1H4H	GeneCards	I2D: score=0.10
130	HIST1H4I	GeneCards	I2D: score=0.10
131	HIST1H4J	GeneCards	I2D: score=0.10
132	HIST1H4K	GeneCards	I2D: score=0.10
133	HIST1H4L	GeneCards	I2D: score=0.10
134	HIST2H4A	GeneCards	I2D: score=0.10
135	HIST2H4B	GeneCards	I2D: score=0.10
136	HIST4H4	GeneCards	I2D: score=0.10
137	HNF4A	GeneCards	STRING: ENSP00000312987
138	HNF4G	GeneCards	STRING: ENSP00000379701
139	HNRNPCL1	GeneCards	I2D: score=0.12
140	HNRNPH2	GeneCards	I2D: score=0.08
141	HP1BP3	GeneCards	I2D: score=0.12
142	HR	GeneCards	I2D: score=0.03
143	HSP90AA1	BioGRID	Powell E (2010)
144	HSPA1A	GeneCards	I2D: score=0.29
145	HSPA1B	GeneCards	I2D: score=0.29
146	HSPA6	GeneCards	I2D: score=0.26
147	HSPA8	GeneCards	I2D: score=0.24
148	HTATSF1	GeneCards	I2D: score=0.09
149	JMJD1C	BioGRID	Iannone MA (2001)
150	KANK2	GeneCards	I2D: score=0.06
151	KAT5	BioGRID	Gaughan L (2001)
152	KATNAL2	GeneCards	I2D: score=0.04
153	KHDRBS1	GeneCards	I2D: score=0.06
154	KLF5	BioGRID	Nakajima Y (2011)
155	KRI1	GeneCards	I2D: score=0.04
156	LACTB	GeneCards	I2D: score=0.02
157	LAD1	GeneCards	I2D: score=0.02

158	LAMB1	GeneCards	I2D: score=0.02
159	LARP1	GeneCards	I2D: score=0.09
160	LAS1L	GeneCards	I2D: score=0.08
161	LCK	GeneCards	I2D: score=0.76 STRING: ENSP00000337825
162	LIMA1	GeneCards	I2D: score=0.32
163	LOC100505603	GeneCards	I2D: score=0.09
164	LOC100507025	GeneCards	I2D: score=0.10
165	LONP1	GeneCards	I2D: score=0.08
166	LRRC1	GeneCards	I2D: score=0.04
167	LRRC59	GeneCards	I2D: score=0.03
168	LYAR	GeneCards	I2D: score=0.09
169	MAD2L1	GeneCards	I2D: score=0.96 STRING: ENSP00000296509
170	MAPK1	BioGRID	Tremblay A (2001)
171	MAPK11	GeneCards	I2D: score=0.97 STRING: ENSP00000333685
172	MDM2	BioGRID	Sanchez M (2012)
173	MED1	GeneCards	I2D: score=0.98 STRING: ENSP00000300651
174	MED10	BioGRID	Kang YK (2002)
175	MED12	GeneCards	I2D: score=0.88 STRING: ENSP00000363193
176	MED13	GeneCards	I2D: score=0.93
177	MED14	GeneCards	I2D: score=0.93 STRING: ENSP00000323720
178	MED16	GeneCards	I2D: score=0.86 STRING: ENSP00000325612
179	MED17	GeneCards	I2D: score=0.86 STRING: ENSP00000251871
180	MED20	GeneCards	I2D: score=0.02 STRING: ENSP00000265350
181	MED21	GeneCards	I2D: score=0.04 STRING: ENSP00000282892
182	MED23	GeneCards	I2D: score=0.75 STRING: ENSP00000357047
183	MED24	GeneCards	I2D: score=0.96 STRING: ENSP00000348610
184	MED27	GeneCards	I2D: score=0.34
185	MED4	GeneCards	I2D: score=0.87
186	MED6	GeneCards	I2D: score=0.38 STRING: ENSP00000256379
187	MED7	GeneCards	I2D: score=0.03 STRING: ENSP00000286317
188	MKI67	GeneCards	I2D: score=0.25
189	MKNK2	BioGRID	Slentz-Kesler K (2000)
190	MLLT1	GeneCards	I2D: score=0.04
191	MMS19	GeneCards	I2D: score=0.89 STRING: ENSP00000359818
192	MRPL44	GeneCards	I2D: score=0.03
193	MRPL45	GeneCards	I2D: score=0.02
194	MRPS15	GeneCards	I2D: score=0.03
195	MRPS2	GeneCards	I2D: score=0.03
196	MRPS21	GeneCards	I2D: score=0.02
197	MRPS22	GeneCards	I2D: score=0.05
198	MRPS27	GeneCards	I2D: score=0.01

199	MRPS31	GeneCards	I2D: score=0.24
200	MRPS5	GeneCards	I2D: score=0.05
201	MRPS9	GeneCards	I2D: score=0.04
202	MSH2	BioGRID	Wada-Hiraike O (2005)
203	MTA2	GeneCards	I2D: score=0.37
204	MYBBP1A	GeneCards	I2D: score=0.07
205	MYH13	GeneCards	I2D: score=0.05
206	MYH14	GeneCards	I2D: score=0.08
207	MYH4	GeneCards	I2D: score=0.05
208	MYH7	GeneCards	I2D: score=0.12
209	MYH7B	GeneCards	I2D: score=0.04
210	MYL6	GeneCards	I2D: score=0.05
211	MYLK2	GeneCards	I2D: score=0.05
212	MYO1C	GeneCards	I2D: score=0.09
213	MYO1D	GeneCards	I2D: score=0.06
214	MYO3B	GeneCards	I2D: score=0.08
215	MYOD1	GeneCards	I2D: score=0.83 STRING: ENSP00000250003
216	NAT10	GeneCards	I2D: score=0.03
217	NBPF8	GeneCards	I2D: score=0.03
218	NCBP1	GeneCards	I2D: score=0.04
219	NCOA1	GeneCards	I2D: score=1.00 STRING: ENSP00000320940
220	NCOA2	BioGRID	Iannone MA (2001)
221	NCOA3	GeneCards	I2D: score=1.00 STRING: ENSP00000361066
222	NCOA6	GeneCards	I2D: score=1.00 STRING: ENSP00000351894
223	NCOA7	GeneCards	I2D: score=0.98 STRING: ENSP00000357341
224	NCOR1	GeneCards	I2D: score=0.99 STRING: ENSP00000268712
225	NCOR2	GeneCards	I2D: score=0.97 STRING: ENSP00000348551
226	NEBL	GeneCards	I2D: score=0.04
227	NEFH	GeneCards	I2D: score=0.05
228	NFKBIB	BioGRID	Iannone MA (2001)
229	NKRF	GeneCards	I2D: score=0.02
230	NOLC1	GeneCards	I2D: score=0.05
231	NOP56	GeneCards	I2D: score=0.15
232	NR0B1	GeneCards	I2D: score=1.00 STRING: ENSP00000368253
233	NR0B2	GeneCards	I2D: score=0.99 STRING: ENSP00000254227
234	NR1D1	GeneCards	STRING: ENSP00000246672
235	NR1D2	GeneCards	STRING: ENSP00000310006
236	NR1H2	GeneCards	STRING: ENSP00000253727
237	NR2C1	GeneCards	STRING: ENSP00000333275
238	NR2C2	GeneCards	STRING: ENSP00000320447
239	NR2C2AP	GeneCards	STRING: ENSP00000332823

240	NR2E1	GeneCards	STRING: ENSP00000357982
241	NR2F1	GeneCards	STRING: ENSP00000325819
242	NR2F6	GeneCards	STRING: ENSP00000291442
243	NR4A1	GeneCards	STRING: ENSP00000243050
244	NR4A2	GeneCards	STRING: ENSP00000344479
245	NR5A1	GeneCards	STRING: ENSP00000362690
246	NRBP1	GeneCards	STRING: ENSP00000233557
247	NRIP1	GeneCards	I2D: score=0.99
248	NUP205	GeneCards	I2D: score=0.10
249	OXT	GeneCards	I2D: score=0.95 STRING: ENSP00000217386
250	PABPC1L	GeneCards	I2D: score=0.02
251	PABPC3	GeneCards	I2D: score=0.04
252	PABPC5	GeneCards	I2D: score=0.06
253	PBRM1	GeneCards	I2D: score=0.81
254	PELP1	GeneCards	I2D: score=0.98 STRING: ENSP00000301396
255	PES1	GeneCards	I2D: score=0.09
256	PIAS1	GeneCards	I2D: score=1.00 STRING: ENSP00000249636
257	PIAS2	GeneCards	I2D: score=0.98 STRING: ENSP00000381648
258	PIAS3	GeneCards	I2D: score=0.75 STRING: ENSP00000376765
259	PIK3R4	GeneCards	I2D: score=0.05
260	PIP	GeneCards	I2D: score=0.02
261	PKP3	GeneCards	I2D: score=0.09
262	PNRC1	GeneCards	I2D: score=0.20
263	PNRC2	BioGRID	Albers M (2005)
264	PNRC2	GeneCards	I2D: score=0.09
265	POLR1E	GeneCards	I2D: score=0.05
266	POM121	GeneCards	I2D: score=0.04
267	POM121C	GeneCards	I2D: score=0.03
268	PPARA	GeneCards	STRING: ENSP00000262735
269	PPARD	GeneCards	STRING: ENSP00000310928
270	PPARG	GeneCards	STRING: ENSP00000287820
271	PPARGC1A	GeneCards	I2D: score=0.99
272	PPP5C	GeneCards	EBI-78505,EBI-716663 I2D: score=0.07 STRING: ENSP00000012443
273	PPRC1	GeneCards	I2D: score=0.19
274	PRDX3	GeneCards	I2D: score=0.06
275	PRDX4	GeneCards	I2D: score=0.06
276	PREPL	GeneCards	I2D: score=0.03
277	PRKDC	GeneCards	I2D: score=0.78
278	PRKRA	GeneCards	I2D: score=0.18
279	PRMT2	GeneCards	I2D: score=0.79 STRING: ENSP00000347906
280	PRPF31	GeneCards	I2D: score=0.09

281	PRPH	GeneCards	I2D: score=0.11
282	PRRC2C	GeneCards	I2D: score=0.02
283	PSMC3IP	GeneCards	I2D: score=0.09 STRING: ENSP00000377384
284	PSMC5	GeneCards	I2D: score=0.90
285	PSMD1	GeneCards	I2D: score=0.18
286	PSMD12	GeneCards	I2D: score=0.15
287	PTCD3	GeneCards	I2D: score=0.04
288	PTPN23	GeneCards	I2D: score=0.04
289	PURA	GeneCards	I2D: score=0.03
290	PURB	GeneCards	I2D: score=0.42
291	PWP1	GeneCards	I2D: score=0.06
292	RALY	GeneCards	I2D: score=0.17
293	RARA	GeneCards	STRING: ENSP00000254066
294	RARB	GeneCards	STRING: ENSP00000332296
295	RARG	GeneCards	STRING: ENSP00000332695
296	RBFOX2	GeneCards	I2D: score=0.88 STRING: ENSP00000413035
297	RBM10	GeneCards	I2D: score=0.04
298	RBM14	GeneCards	I2D: score=0.93
299	RBM25	GeneCards	I2D: score=0.09
300	RBM28	GeneCards	I2D: score=0.17
301	RBM39	GeneCards	I2D: score=0.99 STRING: ENSP00000253363
302	RBM43	GeneCards	I2D: score=0.04
303	RBM47	GeneCards	I2D: score=0.06
304	RBMXL2	GeneCards	I2D: score=0.05
305	REXO4	BioGRID	Montano MM (2000)
306	REXO4	GeneCards	I2D: score=0.89 STRING: ENSP00000361010
307	RFC1	GeneCards	I2D: score=0.03
308	RFC2	GeneCards	I2D: score=0.02
309	RFC4	GeneCards	I2D: score=0.03
310	RIMS2	GeneCards	I2D: score=0.06
311	RORA	GeneCards	STRING: ENSP00000261523
312	RORB	GeneCards	STRING: ENSP00000366093
313	RORC	GeneCards	STRING: ENSP00000327025
314	RPF2	GeneCards	I2D: score=0.07
315	RPL10A	GeneCards	I2D: score=0.02
316	RPL17	GeneCards	I2D: score=0.03
317	RPL18A	GeneCards	I2D: score=0.02
318	RPL23	GeneCards	I2D: score=0.02
319	RPL23A	GeneCards	I2D: score=0.05
320	RPL29	GeneCards	I2D: score=0.02
321	RPL32	GeneCards	I2D: score=0.02

322	RPL36	GeneCards	I2D: score=0.02
323	RPL36AL	GeneCards	I2D: score=0.04
324	RPL3L	GeneCards	I2D: score=0.08
325	RPL7	GeneCards	I2D: score=0.05
326	RPL7A	GeneCards	I2D: score=0.23
327	RPLP0	GeneCards	I2D: score=0.02
328	RPS11	GeneCards	I2D: score=0.02
329	RPS13	GeneCards	I2D: score=0.02
330	RPS16	GeneCards	I2D: score=0.05
331	RPS17	GeneCards	I2D: score=0.02
332	RPS18	GeneCards	I2D: score=0.05
333	RPS24	GeneCards	I2D: score=0.05
334	RPS25	GeneCards	I2D: score=0.05
335	RPS26	GeneCards	I2D: score=0.02
336	RPS27A	GeneCards	I2D: score=0.02
337	RPS27L	GeneCards	I2D: score=0.03
338	RPS3	GeneCards	I2D: score=0.05
339	RPS4X	GeneCards	I2D: score=0.03
340	RPS4Y1	GeneCards	I2D: score=0.02
341	RPS6	GeneCards	I2D: score=0.03
342	RPS7	GeneCards	I2D: score=0.02
343	RPS9	GeneCards	I2D: score=0.03
344	RRBP1	GeneCards	I2D: score=0.06
345	RRP1	GeneCards	I2D: score=0.04
346	RRP12	GeneCards	I2D: score=0.13
347	RRP1B	GeneCards	I2D: score=0.15
348	RXRA	GeneCards	STRING: ENSP00000419692
349	RXRG	GeneCards	STRING: ENSP00000352900
350	SCN1A	GeneCards	I2D: score=0.01
351	SEMG2	GeneCards	I2D: score=0.05
352	SENP3	GeneCards	I2D: score=0.06
353	SERBP1	GeneCards	I2D: score=0.76
354	SERPINH1	BioGRID	Iannone MA (2001)
355	SF3A1	GeneCards	I2D: score=0.07
356	SF3B3	GeneCards	I2D: score=0.09
357	SHROOM3	GeneCards	I2D: score=0.03
358	SIN3A	GeneCards	I2D: score=0.03
359	SIPA1L2	GeneCards	I2D: score=0.02
360	SLC7A6	GeneCards	I2D: score=0.04
361	SMAD2	GeneCards	I2D: score=0.95 STRING: ENSP00000262160
362	SMAD3	GeneCards	I2D: score=0.88 STRING: ENSP00000332973

363	SMAD4	GeneCards	I2D: score=0.95 STRING: ENSP00000341551
364	SMARCA5	GeneCards	I2D: score=0.31
365	SMARCC2	GeneCards	I2D: score=0.86
366	SMARCE1	GeneCards	I2D: score=0.99 STRING: ENSP00000323967
367	SMC1A	GeneCards	I2D: score=0.03
368	SMN1	GeneCards	I2D: score=0.08
369	SMN2	GeneCards	I2D: score=0.08
370	SMTN	GeneCards	I2D: score=0.10
371	SND1	GeneCards	I2D: score=0.10
372	SP1	BioGRID	Jin W (2008)
373	SP1	GeneCards	I2D: score=0.93 STRING: ENSP00000329357
374	SPRR1B	GeneCards	I2D: score=0.07
375	SPTBN2	GeneCards	I2D: score=0.04
376	SRC	GeneCards	I2D: score=0.91 STRING: ENSP00000350941
377	SREBF1	GeneCards	I2D: score=0.95 STRING: ENSP00000348069
378	SRRM2	GeneCards	I2D: score=0.14
379	SRSF2	GeneCards	I2D: score=0.12
380	SRSF5	GeneCards	I2D: score=0.18
381	SSR1	GeneCards	I2D: score=0.04
382	STAT5A	BioGRID	Faulds MH (2001)
383	STAU1	GeneCards	I2D: score=0.10
384	STK31	GeneCards	I2D: score=0.10
385	SURF6	GeneCards	I2D: score=0.05
386	SYNGAP1	GeneCards	I2D: score=0.05
387	TAF15	GeneCards	I2D: score=0.11
388	TCOF1	GeneCards	I2D: score=0.12
389	TEX10	GeneCards	I2D: score=0.03
390	THOC6	GeneCards	I2D: score=0.01
391	THRA	GeneCards	STRING: ENSP00000264637
392	THRB	GeneCards	STRING: ENSP00000348827
393	TRAM1	BioGRID	Kraichely DM (2000)
394	TRIM24	GeneCards	I2D: score=0.95 STRING: ENSP00000340507
395	TRIM59	GeneCards	I2D: score=0.80 STRING: ENSP00000311219
396	TRIP4	BioGRID	Iannone MA (2001)
397	TTC21B	GeneCards	I2D: score=0.04
398	UBC	BioGRID	Picard N (2008)
399	UBC	GeneCards	STRING: ENSP00000344818
400	UBE3A	BioGRID	Picard N (2008)
401	UBN1	GeneCards	I2D: score=0.12
402	UBR3	GeneCards	I2D: score=0.02
403	VDR	GeneCards	STRING: ENSP00000229022

404	VIM	GeneCards	I2D: score=0.14
405	VMAC	GeneCards	I2D: score=0.02
406	VPS41	GeneCards	I2D: score=0.10
407	WDR18	GeneCards	I2D: score=0.01
408	WDR62	GeneCards	I2D: score=0.04
409	WIPI1	GeneCards	I2D: score=0.53 STRING: ENSP00000262139
410	WNK4	GeneCards	I2D: score=0.14
411	WWP1	BioGRID	Nakajima Y (2011)
412	YWHAH	GeneCards	I2D: score=0.76 STRING: ENSP00000248975
413	YWHAQ	BioGRID	Zilliacus J (2001)
414	YWHAQ	GeneCards	I2D: score=0.57 STRING: ENSP00000238081
415	ZBTB17	GeneCards	I2D: score=0.93 STRING: ENSP00000364895
416	ZC3H18	GeneCards	I2D: score=0.03
417	ZC3HAV1	GeneCards	I2D: score=0.09
418	ZFP91	GeneCards	I2D: score=0.10
419	ZFR	GeneCards	I2D: score=0.07
420	ZNF512B	GeneCards	I2D: score=0.09
421	ZNF638	GeneCards	I2D: score=0.09
422	ZNHIT3	BioGRID	Iannone MA (2001)

.

APPENDIX 3

Curated list of interacting molecules from both ESR1 and ESR2 for ER system simulation:

Sl. No.	Gene Symbol	Database	Reference
1	ABCC5	GeneCards	I2D: score=0.14
2	ACAD11	GeneCards	I2D: score=0.01
3	ACSM2B	GeneCards	I2D: score=0.13
4	ACTB	BioGrid	Ambrosino C (2010)
5	ACTC1	BioGrid	Ambrosino C (2010)
6	ACTN1	GeneCards	I2D: score=0.07
7	ACTN3	GeneCards	I2D: score=0.09
8	ACTN4	BioGrid	Khurana S (2011)
9	ACTR2	BioGrid	Ambrosino C (2010)
10	ACTR3	BioGrid	Ambrosino C (2010)
11	AHR	BioGrid	Klinge CM (2000)
12	AHRR	BioGrid	Klinge CM (2000)
13	AKAP13	BioGrid	Rubino D (1998)
14	AKAP8	GeneCards	I2D: score=0.15
15	AKAP9	GeneCards	I2D: score=0.12
16	AKT1	BioGrid	Campbell RA (2001)
17	AKT2	BioGrid	Sun M (2001)
18	ANAPC5	GeneCards	I2D: score=0.06
19	ANKFY1	GeneCards	I2D: score=0.15
20	AP1B1	GeneCards	STRING: ENSP00000350199
21	AP2A1	GeneCards	I2D: score=0.07
22	AR	GeneCards	I2D: score=1.00 STRING: ENSP00000363822
23	ARG1	GeneCards	I2D: score=0.09
24	ARHGDIA	BioGrid	Barone I (2011)
25	ARHGEF15	GeneCards	I2D: score=0.19
26	ARID5A	BioGRID	Georgescu SP (2005)
27	ARNT	GeneCards	I2D: score=1.00 STRING: ENSP00000351407
28	ARPC2	GeneCards	I2D: score=0.08
29	ASB2	GeneCards	I2D: score=0.08
30	ASF1A	GeneCards	STRING: ENSP00000229595
31	ASH2L	BioGrid	Shi L (2011)
32	ASPG	GeneCards	STRING: ENSP00000389003
33	ATAD2	BioGrid	Zou JX (2007)

34	ATAD3C	GeneCards	I2D: score=0.22
35	ATG2A	GeneCards	I2D: score=0.03
36	ATP5J	GeneCards	STRING: ENSP00000389649
37	BAG1	BioGrid	Zeiner M (1995)
38	BAIAP2L1	GeneCards	I2D: score=0.03
39	BARD1	BioGrid	Dizin E (2010)
40	BAZ1A	GeneCards	I2D: score=0.29
41	BAZ1B	BioGrid	Kitagawa H (2003)
42	BCAR1	BioGrid	Cabodi S (2004)
43	BCAS2	GeneCards	I2D: score=0.41 STRING: ENSP00000358554
44	BCLAF1	GeneCards	I2D: score=0.19
45	BDNF	GeneCards	I2D: score=0.98 STRING: ENSP00000414303
46	BICD1	GeneCards	I2D: score=0.05
47	BLCAP	GeneCards	EBI-78473,EBI-3895726
48	BLOC1S1	BioGrid	Mishra SK (2003)
49	BRCA1	BioGrid	Wang C (2005)
50	BRD8	GeneCards	I2D: score=0.66
51	BTF3	GeneCards	I2D: score=0.70 STRING: ENSP00000369965
52	C10orf47	GeneCards	I2D: score=0.02
53	C14orf43	GeneCards	I2D: score=0.63
54	C16orf53	GeneCards	MINT-7013584 I2D: score=0.39
55	C19orf68	GeneCards	I2D: score=0.02
56	C22orf28	GeneCards	I2D: score=0.03
57	C9orf114	GeneCards	I2D: score=0.03
58	CALM1	BioGrid	Garcia Pedrero JM (2002)
59	CALM2	GeneCards	I2D: score=0.99
60	CALM3	GeneCards	I2D: score=0.99
61	CAND1	BioGRID	Bennett EJ (2010)
62	CAV1	BioGrid	Schlegel A (2001)
63	CBLL1	BioGrid	Gong EY (2010)
64	CCAR1	GeneCards	I2D: score=0.50
65	CCDC106	GeneCards	I2D: score=0.05
66	CCNC	BioGrid	Kang YK (2002)
67	CCND1	BioGrid	Zwijsen RM (1997)
68	CCNH	BioGrid	Talukder AH (2003)
69	CCNT1	GeneCards	I2D: score=0.83 STRING: ENSP00000261900
70	CDC16	GeneCards	I2D: score=0.76
71	CDC25B	BioGrid	Ma ZQ (2001)
72	CDC5L	GeneCards	I2D: score=0.06
73	CDK11B	BioGrid	Wang Y (2009)
74	CDK7	BioGrid	Wang Y (2009)

75	CDK8	BioGrid	Wang Y (2009)
76	CDKN1A	BioGrid	Maynadier M (2008)
77	CEBPA	BioGrid	Boruk M (1998)
78	CEBPB	BioGrid	Stein B (1995)
79	CGN	GeneCards	I2D: score=0.02
80	CHD3	BioGrid	Okada M (2008)
81	CHD9	GeneCards	I2D: score=0.71 STRING: ENSP00000381522
82	CHM	GeneCards	I2D: score=0.05
83	CHUK	GeneCards	I2D: score=0.98 STRING: ENSP00000359424
84	CITED1	BioGrid	Yahata T (2001)
85	CLASP2	GeneCards	I2D: score=0.07
86	CLCF1	GeneCards	I2D: score=0.01
87	CNN1	GeneCards	I2D: score=0.06
88	CNTRL	GeneCards	I2D: score=0.07
89	COBRA1	BioGrid	Aiyar SE (2004)
90	COIL	GeneCards	I2D: score=0.06
91	COL18A1	GeneCards	STRING: ENSP00000352798
92	COL1A2	GeneCards	STRING: ENSP00000297268
93	COPB2	GeneCards	I2D: score=0.02
94	COPG2	GeneCards	I2D: score=0.03
95	COPS5	BioGrid	Callige M (2005)
96	CPEB1	GeneCards	STRING: ENSP00000414187
97	CPSF6	GeneCards	I2D: score=0.12
98	CPT1A	GeneCards	I2D: score=0.07
99	CREBBP	BioGrid	Heery DM (1997)
100	CRIPAK	GeneCards	I2D: score=0.44 STRING: ENSP00000323978
101	CSE1L	GeneCards	I2D: score=0.07
102	CSNK2B	GeneCards	I2D: score=0.98 STRING: ENSP00000365025
103	CTNNB1	BioGrid	Kouzmenko AP (2004)
104	CTSD	GeneCards	STRING: ENSP00000236671
105	CTTN	GeneCards	I2D: score=0.05
106	CUEDC2	BioGrid	Pan X (2011)
107	CUL3	BioGrid	Byun B (2008)
108	CUL4B	GeneCards	I2D: score=0.33 STRING: ENSP00000384109
109	CWC15	GeneCards	I2D: score=0.10
110	CWC27	GeneCards	I2D: score=0.03
111	DAP3	BioGrid	Hulkko SM (2000)
112	DBN1	GeneCards	I2D: score=0.02
113	DDX17	BioGrid	Fujita T (2003)
114	DDX39A	GeneCards	I2D: score=0.08
115	DDX3X	BioGrid	Fujita T (2003)

116	DDX41	GeneCards	I2D: score=0.07
117	DDX5	BioGrid	Fujita T (2003)
118	DDX50	GeneCards	I2D: score=0.05
119	DDX54	BioGrid	Rajendran RR (2003)
120	DLGAP4	GeneCards	I2D: score=0.08
121	DNM1L	GeneCards	EBI-78473,EBI-724571
122	DNTTIP1	GeneCards	I2D: score=0.04
123	DNTTIP2	BioGrid	Bu H (2004)
124	DOK6	GeneCards	I2D: score=0.02
125	DSCAM	GeneCards	STRING: ENSP00000383303
126	DUT	BioGrid	Albers M (2005)
127	E2F1	BioGrid	Liu W (2012)
128	EBAG9	GeneCards	I2D: score=0.12
129	EDC4	GeneCards	I2D: score=0.15
130	EEF1A2	GeneCards	I2D: score=0.04
131	EFTUD2	GeneCards	I2D: score=0.05
132	EGFR	GeneCards	EBI-4309277, MINT-6169656
133	EHMT2	BioGrid	Purcell DJ (2011)
134	EIF2A	GeneCards	I2D: score=0.04
135	EIF3I	GeneCards	I2D: score=0.33 STRING: ENSP00000362688
136	ELAVL1	GeneCards	I2D: score=0.21
137	EP300	BioGrid	Fujita T (2003)
138	EPB41L5	GeneCards	I2D: score=0.02
139	EPS15L1	GeneCards	I2D: score=0.06
140	ERBB2	BioGrid	Yang Z (2004)
141	EREG	GeneCards	STRING: ENSP00000244869
142	ESR1	BioGrid	Ikeda M (2002)
143	ESR2	BioGrid	Monroe DG (2003)
144	ESRRA	GeneCards	I2D: score=0.99 STRING: ENSP00000000442
145	ESRRB	GeneCards	STRING: ENSP00000261532
146	ESRRG	GeneCards	STRING: ENSP00000386171
147	EXOSC10	GeneCards	I2D: score=0.03
148	EXOSC4	GeneCards	I2D: score=0.12
149	EZH2	GeneCards	I2D: score=0.35
150	FAM120A	GeneCards	I2D: score=0.02
151	FAM179B	GeneCards	I2D: score=0.04
152	FAM82A2	GeneCards	I2D: score=0.07
153	FBLL1	GeneCards	I2D: score=0.11
154	FHL1	GeneCards	I2D: score=0.68 STRING: ENSP00000359724
155	FHL2	BioGrid	Kobayashi S (2004)
156	FIP1L1	GeneCards	I2D: score=0.08

157	FKBP4	BioGrid	Nair SC (1996)
158	FKBP5	BioGrid	Nair SC (1996)
159	FLG	GeneCards	I2D: score=0.04
160	FLII	BioGrid	Jeong KW (2009)
161	FLNB	GeneCards	I2D: score=0.03
162	FLNC	GeneCards	I2D: score=0.03
163	FMR1	GeneCards	I2D: score=0.06
164	FOS	BioGrid	Baron S (2007)
165	FOXA1	GeneCards	STRING: ENSP00000250448
166	FOXL2	GeneCards	I2D: score=0.90 STRING: ENSP00000333188
167	FOXM1	GeneCards	STRING: ENSP00000342307
168	FOXO1	BioGrid	Zhao HH (2001)
169	FOXO3	GeneCards	I2D: score=0.98 STRING: ENSP00000339527
170	FOXO4	BioGrid	Schuur ER (2001)
171	FTSJ3	GeneCards	I2D: score=0.07
172	FXR2	GeneCards	I2D: score=0.09
173	GADD45A	BioGrid	Yi YW (2000)
174	GADD45B	BioGrid	Yi YW (2000)
175	GADD45G	BioGrid	Yi YW (2000)
176	GCC2	GeneCards	I2D: score=0.04
177	GFAP	GeneCards	I2D: score=0.07
178	GIGYF1	GeneCards	I2D: score=0.03
179	GNA13	GeneCards	STRING: ENSP00000400717
180	GNA14	GeneCards	STRING: ENSP00000365807
181	GNA15	GeneCards	STRING: ENSP00000262958
182	GNAI1	GeneCards	I2D: score=0.13 STRING: ENSP00000343027
183	GNAI2	GeneCards	STRING: ENSP00000312999
184	GNAL	GeneCards	STRING: ENSP00000334051
185	GNAO1	GeneCards	STRING: ENSP00000262493
186	GNAZ	GeneCards	STRING: ENSP00000248996
187	GNB2L1	GeneCards	I2D: score=0.08
188	GNB3	GeneCards	I2D: score=0.10
189	GNG2	GeneCards	STRING: ENSP00000334448
190	GNL2	GeneCards	I2D: score=0.05
191	GNL3	GeneCards	I2D: score=0.06
192	GOLGA3	GeneCards	I2D: score=0.07
193	GOLGA4	GeneCards	I2D: score=0.04
194	GRB2	GeneCards	I2D: score=1.00 STRING: ENSP00000339007
195	GRIP1	BioGrid	Yi YW (2000)
196	GSN	BioGrid	Yi YW (2000)
197	GSPT1	GeneCards	I2D: score=0.04

198	GTF2B	BioGrid	Yi YW (2000)
199	GTF2H1	BioGrid	Yi YW (2000)
200	GTF3C1	GeneCards	I2D: score=0.12
201	GTF3C2	GeneCards	I2D: score=0.10
202	GTF3C3	GeneCards	I2D: score=0.14
203	GTF3C4	GeneCards	I2D: score=0.03
204	GTF3C5	GeneCards	I2D: score=0.11
205	GTPBP4	GeneCards	I2D: score=0.04
206	H2BFS	GeneCards	I2D: score=0.05
207	HAX1	GeneCards	EBI-78473,EBI-357001
208	HCFC1	GeneCards	I2D: score=0.90
209	HCFC2	GeneCards	I2D: score=0.03
210	HDAC1	BioGrid	Kawai H (2003)
211	HDAC2	BioGrid	Itoh Y (2007)
212	HDAC3	BioGrid	Liu XF (2004)
213	HDAC4	BioGrid	Leong H (2005)
214	HDAC5	BioGrid	van Rooij E (2010)
215	HDAC7	BioGrid	Malik S (2010)
216	HDAC9	BioGrid	van Rooij E (2010)
217	HEATR4	GeneCards	I2D: score=0.05
218	HEXIM1	GeneCards	I2D: score=0.32 STRING: ENSP00000328773
219	HIST1H1C	GeneCards	I2D: score=0.17
220	HIST1H4A	GeneCards	I2D: score=0.10
221	HIST1H4B	GeneCards	I2D: score=0.10
222	HIST1H4C	GeneCards	I2D: score=0.10
223	HIST1H4D	GeneCards	I2D: score=0.10
224	HIST1H4E	GeneCards	I2D: score=0.10
225	HIST1H4F	GeneCards	I2D: score=0.10
226	HIST1H4H	GeneCards	I2D: score=0.10
227	HIST1H4I	GeneCards	I2D: score=0.10
228	HIST1H4J	GeneCards	I2D: score=0.10
229	HIST1H4K	GeneCards	I2D: score=0.10
230	HIST1H4L	GeneCards	I2D: score=0.10
231	HIST2H4A	GeneCards	I2D: score=0.10
232	HIST2H4B	GeneCards	I2D: score=0.10
233	HIST4H4	GeneCards	I2D: score=0.10
234	HNF4A	GeneCards	I2D: score=1.00 STRING: ENSP00000312987
235	HNF4G	GeneCards	STRING: ENSP00000379701
236	HNRNPCL1	GeneCards	I2D: score=0.12
237	HNRNPH2	GeneCards	I2D: score=0.08
238	HP1BP3	GeneCards	I2D: score=0.12

239	HR	GeneCards	I2D: score=0.03
240	HSP90AA1	BioGrid	Nair SC (1996)
241	HSP90AB1	GeneCards	I2D: score=1.00 STRING: ENSP00000325875
242	HSPA1A	GeneCards	I2D: score=0.29
243	HSPA1B	GeneCards	I2D: score=0.29
244	HSPA4	BioGrid	Nair SC (1996)
245	HSPA6	GeneCards	I2D: score=0.26
246	HSPA8	BioGrid	Tateishi Y (2004)
247	HTATSF1	GeneCards	I2D: score=0.09
248	IGF1	GeneCards	STRING: ENSP00000302665
249	IGF1R	BioGrid	Santen RJ (2005)
250	IL25	GeneCards	I2D: score=0.21
251	ING1	BioGrid	Toyama T (2003)
252	IRS1	BioGrid	Morelli C (2003)
253	IRS2	BioGrid	Morelli C (2003)
254	ISL1	BioGrid	Gay F (2000)
255	ITGB3BP	GeneCards	I2D: score=0.90 STRING: ENSP00000271002
256	JMJD1C	BioGRID	Iannone MA (2001)
257	JUN	BioGrid	Teyssier C (2001)
258	JUNB	GeneCards	I2D: score=0.98 STRING: ENSP00000303315
259	JUND	GeneCards	I2D: score=0.92 STRING: ENSP00000252818
260	KANK2	GeneCards	I2D: score=0.06
261	KAT2A	BioGrid	Oishi H (2006)
262	KAT2B	GeneCards	STRING: ENSP00000263754
263	KAT5	BioGrid	Gaughan L (2001)
264	KAT6A	BioGrid	Yin H (2007)
265	KATNAL2	GeneCards	I2D: score=0.04
266	KDM1A	BioGrid	Garcia-Bassets I (2007)
267	KDM4B	BioGrid	Shi L (2011)
268	KDM5A	BioGrid	Chan SW (2001)
269	KDM6B	GeneCards	MINT-8286415 MINT-8286424
270	KHDRBS1	GeneCards	I2D: score=0.06
271	KIF11	GeneCards	STRING: ENSP00000260731
272	KIF1A	BioGrid	Albers M (2005)
273	KLF5	BioGrid	Guo P (2010)
274	KRI1	GeneCards	I2D: score=0.04
275	LACTB	GeneCards	I2D: score=0.02
276	LAD1	GeneCards	I2D: score=0.02
277	LAMB1	GeneCards	I2D: score=0.02
278	LARP1	GeneCards	I2D: score=0.09
279	LAS1L	GeneCards	I2D: score=0.08

280	LCK	GeneCards	I2D: score=0.99 STRING: ENSP00000337825
281	LCOR	BioGrid	Fernandes I (2003)
282	LDB1	BioGrid	Johnsen SA (2009)
283	LIMA1	GeneCards	I2D: score=0.32
284	LINC00312	GeneCards	MINT-50571 MINT-50573 MINT-50906 I2D: score=0.04
285	LMO4	BioGrid	Singh RR (2005)
286	LOC100505603	GeneCards	MINT-14693 MINT-14694 I2D: score=0.87
287	LOC100507025	GeneCards	I2D: score=0.10
288	LONP1	GeneCards	I2D: score=0.08
289	LRIF1	BioGrid	Li HJ (2007)
290	LRRC1	GeneCards	I2D: score=0.04
291	LRRC59	GeneCards	I2D: score=0.03
292	LYAR	GeneCards	I2D: score=0.09
293	LYZ	GeneCards	I2D: score=0.12
294	MAD2L1	GeneCards	STRING: ENSP00000296509
295	MAP3K1	GeneCards	MINT-50709 I2D: score=0.28 STRING: ENSP00000382423
296	MAPK1	BioGrid	Metivier R (2002)
297	MAPK11	GeneCards	STRING: ENSP00000333685
298	MAPK14	GeneCards	I2D: score=0.98 STRING: ENSP00000229794
299	MAPK3	GeneCards	I2D: score=0.99 STRING: ENSP00000263025
300	MBD2	BioGrid	Chatagnon A (2010)
301	MDM2	BioGrid	Duong V (2007)
302	MED1	BioGrid	Fujita T (2003)
303	MED10	BioGrid	Fujita T (2003)
304	MED12	BioGrid	Kang YK (2002)
305	MED13	BioGrid	Kang YK (2002)
306	MED14	BioGrid	Burakov D (2000)
307	MED16	BioGrid	Kang YK (2002)
308	MED17	BioGrid	Kang YK (2002)
309	MED20	BioGrid	Kang YK (2002)
310	MED21	BioGrid	Kang YK (2002)
311	MED23	BioGrid	Kang YK (2002)
312	MED24	BioGrid	Kang YK (2002)
313	MED25	BioGrid	Lee HK (2007)
314	MED27	GeneCards	I2D: score=0.34
315	MED30	GeneCards	I2D: score=0.92
316	MED4	GeneCards	I2D: score=0.92
317	MED6	BioGrid	Lee HK (2007)
318	MED7	BioGrid	Lee HK (2007)
319	MEN1	BioGrid	Dreijerink KM (2006)
320	MGMT	BioGrid	Teo AK (2001)

321	MKI67	GeneCards	I2D: score=0.25
322	MKNK2	GeneCards	I2D: score=0.97 STRING: ENSP00000250896
323	MLL2	BioGrid	Shi L (2011)
324	MLLT1	GeneCards	I2D: score=0.04
325	MMS19	GeneCards	I2D: score=0.98 STRING: ENSP00000359818
326	MNAT1	BioGrid	Talukder AH (2003)
327	MPG	BioGrid	Likhite VS (2004)
328	MRPL44	GeneCards	I2D: score=0.03
329	MRPL45	GeneCards	I2D: score=0.02
330	MRPS15	GeneCards	I2D: score=0.02
331	MRPS17	GeneCards	I2D: score=0.05
332	MRPS2	GeneCards	I2D: score=0.03
333	MRPS21	GeneCards	I2D: score=0.02
334	MRPS22	GeneCards	I2D: score=0.25
335	MRPS27	GeneCards	I2D: score=0.07
336	MRPS31	GeneCards	I2D: score=0.28
337	MRPS35	GeneCards	I2D: score=0.03
338	MRPS5	GeneCards	I2D: score=0.05
339	MRPS6	GeneCards	I2D: score=0.04
340	MRPS9	GeneCards	I2D: score=0.04
341	MSH2	BioGrid	Wada-Hiraike O (2005)
342	MSH6	BioGrid	Oishi H (2006)
343	MTA1	BioGrid	Mishra SK (2003)
344	MTA2	BioGrid	Okada M (2008)
345	MTA3	BioGrid	Okada M (2008)
346	MTCH2	BioGrid	Albers M (2005)
347	MUC1	BioGrid	Wei X (2006)
348	MVP	BioGrid	Abbondanza C (1998)
349	MYBBP1A	GeneCards	I2D: score=0.07
350	MYC	BioGrid	Cheng AS (2006)
351	MYH13	GeneCards	I2D: score=0.05
352	MYH14	GeneCards	I2D: score=0.08
353	MYH4	GeneCards	I2D: score=0.05
354	MYH7	GeneCards	I2D: score=0.12
355	MYH7B	GeneCards	I2D: score=0.04
356	MYL6	BioGrid	Ambrosino C (2010)
357	MYLK2	GeneCards	I2D: score=0.08
358	MYO1C	BioGrid	Ambrosino C (2010)
359	MYO1D	GeneCards	I2D: score=0.06
360	MYO3B	GeneCards	I2D: score=0.08
361	MYOD1	BioGrid	Jin W (2008)

362	MYSM1	GeneCards	STRING: ENSP00000418734
363	NAT10	GeneCards	I2D: score=0.03
364	NBPF8	GeneCards	I2D: score=0.03
365	NCBP1	GeneCards	I2D: score=0.04
366	NCOA1	BioGrid	Kraichely DM (2000)
367	NCOA2	BioGrid	Kraichely DM (2000)
368	NCOA3	BioGrid	Suen CS (1998)
369	NCOA4	BioGrid	Alen P (1999)
370	NCOA6	BioGrid	Mahajan MA (2000)
371	NCOA7	BioGrid	Lazennec G (1997)
372	NCOR1	BioGrid	Fujita T (2003)
373	NCOR2	BioGrid	Fujita T (2003)
374	NDRG2	GeneCards	EBI-78473,EBI-3895741
375	NDUFV3	GeneCards	STRING: ENSP00000346196
376	NEBL	GeneCards	I2D: score=0.04
377	NEFH	GeneCards	I2D: score=0.05
378	NFKB1	GeneCards	I2D: score=1.00 STRING: ENSP00000226574
379	NFKBIB	BioGRID	Iannone MA (2001)
380	NGG1	BioGrid	Benecke A (2002)
381	NKRF	GeneCards	I2D: score=0.02
382	NOLC1	GeneCards	I2D: score=0.05
383	NOP56	GeneCards	I2D: score=0.15
384	NOS3	GeneCards	I2D: score=0.96
385	NPM1	BioGrid	Ambrosino C (2010)
386	NPPA	BioGrid	Albers M (2005)
387	NR0B1	GeneCards	I2D: score=1.00 STRING: ENSP00000368253
388	NR0B2	BioGrid	Seol W (1998)
389	NR1D1	GeneCards	STRING: ENSP00000246672
390	NR1D2	GeneCards	STRING: ENSP00000310006
391	NR1H2	GeneCards	STRING: ENSP00000253727
392	NR1H4	GeneCards	I2D: score=0.97 STRING: ENSP00000315442
393	NR1I2	BioGrid	Seol W (1998)
394	NR2C1	BioGrid	Hu YC (2002)
395	NR2C2	BioGrid	Shyr CR (2002)
396	NR2C2AP	GeneCards	STRING: ENSP00000332823
397	NR2E1	GeneCards	STRING: ENSP00000357982
398	NR2F1	BioGrid	Metivier R (2002)
399	NR2F6	BioGrid	Zhu XG (2000)
400	NR4A1	GeneCards	STRING: ENSP00000243050
401	NR4A2	GeneCards	STRING: ENSP00000344479
402	NR5A1	GeneCards	STRING: ENSP00000362690

403	NRBP1	GeneCards	STRING: ENSP00000233557
404	NRIP1	BioGrid	Heery DM (1997)
405	NSD1	BioGrid	Huang N (1998)
406	NUP205	GeneCards	I2D: score=0.10
407	OTUB1	BioGrid	Stanisic V (2009)
408	OXT	GeneCards	I2D: score=0.95 STRING: ENSP00000217386
409	PABPC1L	GeneCards	I2D: score=0.02
410	PABPC3	GeneCards	I2D: score=0.04
411	PABPC5	GeneCards	I2D: score=0.24
412	PAK1	GeneCards	I2D: score=0.98 STRING: ENSP00000278568
413	PAK6	BioGrid	Lee SR (2002)
414	PARP1	BioGrid	Ju BG (2006)
415	PBRM1	GeneCards	I2D: score=0.81
416	PDLIM1	BioGrid	Johnsen SA (2009)
417	PELP1	BioGrid	Vadlamudi RK (2001)
418	PES1	GeneCards	I2D: score=0.09
419	PGC	BioGrid	Bourdoncle A (2005)
420	PGR	BioGrid	Giulianelli S (2012)
421	PHB	BioGrid	Giulianelli S (2012)
422	PHB2	BioGrid	Martini PG (2003)
423	PIAS1	BioGrid	Kobayashi S (2004)
424	PIAS2	BioGrid	Kotaja N (2000)
425	PIAS3	GeneCards	I2D: score=0.99 STRING: ENSP00000376765
426	PIK3CA	GeneCards	I2D: score=0.96 STRING: ENSP00000263967
427	PIK3R1	BioGrid	Castoria G (2001)
428	PIK3R2	BioGrid	Cabodi S (2004)
429	PIK3R3	GeneCards	I2D: score=0.96 STRING: ENSP00000262741
430	PIK3R4	GeneCards	I2D: score=0.05
431	PIP	GeneCards	I2D: score=0.02
432	PKP3	GeneCards	I2D: score=0.09
433	PLA2G7	GeneCards	I2D: score=0.05
434	PNRC1	GeneCards	I2D: score=0.20
435	PNRC2	BioGrid	Albers M (2005)
436	POLR1B	GeneCards	I2D: score=0.40 STRING: ENSP00000263331
437	POLR1E	GeneCards	I2D: score=0.05
438	POLR2A	BioGrid	Ambrosino C (2010)
439	POLR2D	GeneCards	I2D: score=0.16
440	POM121	GeneCards	I2D: score=0.04
441	POM121C	GeneCards	I2D: score=0.03
442	POU2F1	BioGrid	Prefontaine GG (1999)
443	POU2F2	BioGrid	Prefontaine GG (1999)

444	POU4F1	BioGrid	Budhram-Mahadeo V (1998)
445	POU4F2	BioGrid	Budhram-Mahadeo V (1998)
446	PPARA	GeneCards	STRING: ENSP00000262735
447	PPARD	GeneCards	STRING: ENSP00000310928
448	PPARG	BioGrid	DiRenzo J (1997)
449	PPARGC1A	BioGrid	DiRenzo J (1997)
450	PPARGC1B	BioGrid	DiRenzo J (1997)
451	PPID	BioGrid	Nair SC (1996)
452	PPP1CC	BioGrid	Flores-Delgado G (2007)
453	PPP5C	GeneCards	EBI-78473,EBI-716663 I2D: score=0.05 STRING: ENSP00000012443
454	PPRC1	GeneCards	I2D: score=0.19
455	PRDM2	BioGrid	Abbondanza C (2000)
456	PRDX3	GeneCards	I2D: score=0.06
457	PRDX4	GeneCards	I2D: score=0.06
458	PREPL	GeneCards	I2D: score=0.03
459	PRKACA	GeneCards	I2D: score=0.98 STRING: ENSP00000309591
460	PRKCZ	BioGrid	Yi P (2008)
461	PRKDC	BioGrid	Ju BG (2006)
462	PRKRA	GeneCards	I2D: score=0.18
463	PRMT1	BioGrid	Koh SS (2001)
464	PRMT2	BioGrid	Qi C (2002)
465	PRPF31	GeneCards	I2D: score=0.09
466	PRPF6	GeneCards	I2D: score=0.99 STRING: ENSP00000266079
467	PRPH	GeneCards	I2D: score=0.11
468	PRRC2C	GeneCards	I2D: score=0.02
469	PSEN2	GeneCards	STRING: ENSP00000355747
470	PSMB9	GeneCards	MINT-3291999 I2D: score=0.07
471	PSMC3IP	GeneCards	I2D: score=0.09 STRING: ENSP00000377384
472	PSMC5	BioGrid	vom Baur E (1996)
473	PSMD1	GeneCards	I2D: score=0.96 STRING: ENSP00000309474
474	PSMD12	GeneCards	I2D: score=0.15
475	PTCD3	GeneCards	I2D: score=0.05
476	PTEN	BioGrid	Lin HK (2004)
477	PTGES3	BioGrid	Nair SC (1996)
478	PTMA	BioGrid	Martini PG (2003)
479	PTPN1	GeneCards	MINT-6761899 I2D: score=0.99 STRING: ENSP00000360683
480	PTPN23	GeneCards	I2D: score=0.04
481	PTPN6	GeneCards	MINT-6761906 I2D: score=0.98 STRING: ENSP00000391592
482	PURA	GeneCards	I2D: score=0.03
483	PURB	GeneCards	I2D: score=0.42
484	PWP1	GeneCards	I2D: score=0.06

485	RABGEF1	BioGrid	La Rosa P (2011)
486	RAC3	GeneCards	EBI-78473,EBI-767084
487	RALY	GeneCards	I2D: score=0.17
488	RAN	BioGrid	Castoria G (2012)
489	RARA	GeneCards	I2D: score=1.00 STRING: ENSP00000254066
490	RARB	GeneCards	STRING: ENSP00000332296
491	RARG	GeneCards	STRING: ENSP00000332695
492	RBBP4	BioGrid	Creekmore AL (2008)
493	RBBP5	BioGrid	Shi L (2011)
494	RBBP6	BioGrid	Peidis P (2010)
495	RBBP7	BioGrid	Creekmore AL (2008)
496	RBFOX2	BioGrid	Norris JD (2002)
497	RBM10	GeneCards	STRING: ENSP00000366829
498	RBM14	GeneCards	I2D: score=0.93
499	RBM23	GeneCards	I2D: score=0.62 STRING: ENSP00000352956
500	RBM25	GeneCards	I2D: score=0.09
501	RBM28	GeneCards	I2D: score=0.17
502	RBM39	BioGrid	Jung DJ (2002)
503	RBM43	GeneCards	I2D: score=0.04
504	RBM47	GeneCards	I2D: score=0.06
505	RBMXL2	GeneCards	I2D: score=0.05
506	RELA	BioGrid	Nettles KW (2008)
507	REXO4	BioGrid	Montano MM (2000)
508	RFC1	GeneCards	I2D: score=0.03
509	RFC2	GeneCards	I2D: score=0.02
510	RFC4	GeneCards	I2D: score=0.03
511	RFX4	GeneCards	I2D: score=0.35 STRING: ENSP00000376585
512	RGS3	GeneCards	I2D: score=0.11 STRING: ENSP00000259406
513	RIMS2	GeneCards	I2D: score=0.06
514	RLIM	BioGrid	Johnsen SA (2009)
515	RNF14	GeneCards	I2D: score=0.95 STRING: ENSP00000324956
516	RNF4	BioGrid	Moilanen AM (1998)
517	RORA	GeneCards	STRING: ENSP00000261523
518	RORB	GeneCards	STRING: ENSP00000366093
519	RORC	GeneCards	STRING: ENSP00000327025
520	RPF2	GeneCards	I2D: score=0.07
521	RPL10A	GeneCards	I2D: score=0.02
522	RPL17	GeneCards	I2D: score=0.03
523	RPL18	BioGrid	Moilanen AM (1998)
524	RPL18A	GeneCards	I2D: score=0.02
525	RPL23	GeneCards	I2D: score=0.02

526	RPL23A	GeneCards	I2D: score=0.05
527	RPL29	GeneCards	I2D: score=0.02
528	RPL32	GeneCards	I2D: score=0.02
529	RPL36	GeneCards	I2D: score=0.02
530	RPL36AL	GeneCards	I2D: score=0.08
531	RPL3L	GeneCards	I2D: score=0.08
532	RPL7	BioGrid	Ambrosino C (2010)
533	RPL7A	BioGrid	Ambrosino C (2010)
534	RPLP0	BioGrid	Ambrosino C (2010)
535	RPLP0P6	GeneCards	I2D: score=0.02
536	RPS11	GeneCards	I2D: score=0.08
537	RPS13	GeneCards	I2D: score=0.02
538	RPS16	GeneCards	I2D: score=0.05
539	RPS17	GeneCards	I2D: score=0.02
540	RPS18	GeneCards	I2D: score=0.10
541	RPS20	GeneCards	I2D: score=0.22
542	RPS24	GeneCards	I2D: score=0.05
543	RPS25	GeneCards	I2D: score=0.05
544	RPS26	GeneCards	I2D: score=0.02
545	RPS27A	GeneCards	I2D: score=0.02
546	RPS27L	GeneCards	I2D: score=0.03
547	RPS3	GeneCards	I2D: score=0.05
548	RPS4X	BioGrid	Ambrosino C (2010)
549	RPS4Y1	GeneCards	I2D: score=0.02
550	RPS6	GeneCards	I2D: score=0.03
551	RPS6KA1	BioGrid	Joel PB (1998)
552	RPS6KA3	BioGrid	Clark DE (2001)
553	RPS7	GeneCards	I2D: score=0.02
554	RPS8	BioGrid	Ambrosino C (2010)
555	RPS9	BioGrid	Ambrosino C (2010)
556	RRBP1	GeneCards	I2D: score=0.06
557	RRP1	GeneCards	I2D: score=0.04
558	RRP12	GeneCards	I2D: score=0.13
559	RRP1B	GeneCards	I2D: score=0.15
560	RXRA	BioGrid	DiRenzo J (1997)
561	RXRG	GeneCards	STRING: ENSP00000352900
562	SAFB	BioGrid	Oesterreich S (2000)
563	SAFB2	BioGrid	Oesterreich S (2000)
564	SCN1A	GeneCards	I2D: score=0.01
565	SEMG2	GeneCards	I2D: score=0.05
566	SENP3	GeneCards	I2D: score=0.06

567	SENP5	GeneCards	EBI-78473,EBI-3895753
568	SERBP1	GeneCards	I2D: score=0.76
569	SERPINH1	BioGRID	Iannone MA (2001)
570	SETD7	BioGrid	Subramanian K (2008)
571	SF1	BioGrid	Zhou D (2001)
572	SF3A1	GeneCards	I2D: score=0.07
573	SF3B3	GeneCards	I2D: score=0.09
574	SGK3	BioGrid	Wang Y (2011)
575	SHC1	BioGrid	Song RX (2002)
576	SHROOM3	GeneCards	I2D: score=0.03
577	SIN3A	BioGrid	Ellison-Zelski SJ (2009)
578	SIPA1L2	GeneCards	I2D: score=0.02
579	SIRT1	BioGrid	Elangovan S (2011)
580	SKI	BioGrid	Elangovan S (2011)
581	SKIL	BioGrid	Band AM (2011)
582	SLC30A9	GeneCards	I2D: score=0.92 STRING: ENSP00000264451
583	SLC7A6	GeneCards	I2D: score=0.04
584	SMAD2	BioGrid	Ito I (2010)
585	SMAD3	BioGrid	Ito I (2010)
586	SMAD4	GeneCards	I2D: score=1.00 STRING: ENSP00000341551
587	SMARCA2	BioGrid	Ichinose H (1997)
588	SMARCA4	BioGrid	Ichinose H (1997)
589	SMARCA5	GeneCards	I2D: score=0.31
590	SMARCB1	BioGrid	Okada M (2008)
591	SMARCC2	BioGrid	Okada M (2008)
592	SMARCD1	BioGrid	Hsiao PW (2003)
593	SMARCD3	GeneCards	I2D: score=0.93 STRING: ENSP00000262188
594	SMARCE1	BioGrid	Kiskinis E (2006)
595	SMC1A	GeneCards	I2D: score=0.03
596	SMN1	GeneCards	I2D: score=0.08
597	SMN2	GeneCards	I2D: score=0.08
598	SMTN	GeneCards	I2D: score=0.10
599	SMURF1	BioGrid	Ito I (2010)
600	SMYD3	BioGrid	Kim H (2009)
601	SND1	GeneCards	I2D: score=0.10
602	SOD1	GeneCards	STRING: ENSP00000270142
603	SOS1	GeneCards	I2D: score=0.95 STRING: ENSP00000384675
604	SOS2	BioGrid	Yang Z (2004)
605	SP1	BioGrid	Petz LN (2004)
606	SP2	BioGrid	He S (2005)
607	SP3	BioGrid	Dong J (2006)

608	SPOP	BioGrid	Byun B (2008)
609	SPRR1B	GeneCards	I2D: score=0.07
610	SPTBN2	GeneCards	I2D: score=0.04
611	SRA1	BioGrid	Watanabe M (2001)
612	SRC	BioGrid	Monroe DG (2003)
613	SREBF1	GeneCards	I2D: score=0.96 STRING: ENSP00000348069
614	SRRM2	GeneCards	I2D: score=0.14
615	SRSF2	GeneCards	I2D: score=0.12
616	SRSF5	GeneCards	I2D: score=0.18
617	SSR1	GeneCards	I2D: score=0.04
618	STAT3	BioGrid	Wang LH (2001)
619	STAT5A	BioGrid	Faulds MH (2001)
620	STAU1	GeneCards	I2D: score=0.30
621	STK11	GeneCards	STRING: ENSP00000324856
622	STK31	GeneCards	I2D: score=0.10
623	STRN	GeneCards	I2D: score=0.84 STRING: ENSP00000263918
624	STUB1	BioGrid	Tateishi Y (2004)
625	SUMO1	BioGrid	Sentis S (2005)
626	SURF6	GeneCards	I2D: score=0.05
627	SVIL	BioGrid	Ting HJ (2002)
628	SYNGAP1	GeneCards	I2D: score=0.05
629	TAB2	BioGrid	Cutrupi S (2012)
630	TADA3	BioGrid	Li CW (2010)
631	TAF10	GeneCards	I2D: score=1.00 STRING: ENSP00000299424
632	TAF15	GeneCards	I2D: score=0.11
633	TAF1A	BioGrid	Loven MA (2003)
634	TAF1B	BioGrid	Loven MA (2003)
635	TAF2	BioGrid	Suen CS (1998)
636	TAP1	GeneCards	I2D: score=0.93 STRING: ENSP00000346206
637	TBP	BioGrid	Lazennec G (1997)
638	TCF20	GeneCards	I2D: score=0.97 STRING: ENSP00000352463
639	TCOF1	GeneCards	I2D: score=0.12
640	TDG	BioGrid	Chen D (2003)
641	TEX10	GeneCards	I2D: score=0.03
642	TFF1	BioGrid	Yi P (2008)
643	THOC6	GeneCards	I2D: score=0.01
644	THRA	GeneCards	I2D: score=1.00 STRING: ENSP00000264637
645	THRB	GeneCards	I2D: score=1.00 STRING: ENSP00000348827
646	TNFRSF14	GeneCards	I2D: score=0.96 STRING: ENSP00000347948
647	TOP2B	BioGrid	Ju BG (2006)
648	TP53	BioGrid	Liu G (2000)

649	TRAM1	BioGrid	Kraichely DM (2000)
650	TRIM24	BioGrid	vom Baur E (1996)
651	TRIM25	BioGrid	Nakajima A (2007)
652	TRIM28	GeneCards	I2D: score=0.97 STRING: ENSP00000253024
653	TRIM59	GeneCards	I2D: score=0.87 STRING: ENSP00000311219
654	TRIP4	BioGrid	Kim HJ (1999)
655	TRRAP	BioGrid	Fujita T (2003)
656	TSC2	GeneCards	I2D: score=0.94 STRING: ENSP00000219476
657	TTC21B	GeneCards	I2D: score=0.04
658	TUBA1A	BioGrid	Benecke A (2002)
659	TUBA1B	GeneCards	I2D: score=0.14 STRING: ENSP00000336799
660	TUBB	GeneCards	I2D: score=0.04
661	TUBB1	BioGrid	Benecke A (2002)
662	TUBB2A	GeneCards	STRING: ENSP00000369703
663	UBC	BioGrid	Stanisic V (2009)
664	UBE2I	BioGrid	Kobayashi S (2004)
665	UBE3A	BioGrid	Sun J (2012)
666	UBN1	GeneCards	I2D: score=0.12
667	UBR3	GeneCards	I2D: score=0.02
668	UIMC1	BioGrid	Yan J (2007)
669	USF1	GeneCards	I2D: score=0.75 STRING: ENSP00000356999
670	UTP14A	GeneCards	I2D: score=0.81
671	VAV3	GeneCards	I2D: score=0.98 STRING: ENSP00000359073
672	VDR	GeneCards	STRING: ENSP00000229022
673	VIM	GeneCards	I2D: score=0.14
674	VMAC	GeneCards	I2D: score=0.02
675	VPS41	GeneCards	I2D: score=0.10
676	WDR18	GeneCards	I2D: score=0.01
677	WDR5	BioGrid	Shi L (2011)
678	WDR5B	GeneCards	I2D: score=0.18 STRING: ENSP00000330381
679	WDR62	GeneCards	I2D: score=0.04
680	WIPI1	GeneCards	I2D: score=0.88 STRING: ENSP00000262139
681	WNK4	GeneCards	I2D: score=0.14
682	WWP1	BioGRID	Nakajima Y (2011)
683	XBP1	BioGrid	Ding L (2003)
684	XPO1	BioGrid	Castoria G (2012)
685	XRCC5	BioGrid	Ju BG (2006)
686	XRCC6	BioGrid	Ju BG (2006)
687	YWHAH	GeneCards	I2D: score=0.98 STRING: ENSP00000248975
688	YWHAQ	BioGrid	Zilliacus J (2001)
689	ZBTB16	BioGrid	Martin PJ (2003)

690	ZBTB17	GeneCards	I2D: score=0.95 STRING: ENSP00000364895
691	ZC3H18	GeneCards	I2D: score=0.03
692	ZC3HAV1	GeneCards	I2D: score=0.17
693	ZFP91	GeneCards	I2D: score=0.10
694	ZFR	GeneCards	I2D: score=0.07
695	ZNF398	BioGrid	Conroy AT (2002)
696	ZNF512B	GeneCards	I2D: score=0.09
697	ZNF638	GeneCards	I2D: score=0.09
698	ZNHIT3	GeneCards	I2D: score=0.96

COMMUNICATIONS

Peer reviewed publications

Gopal K Dhondalay, Christophe Lemetre and Graham R Ball. **Modelling estrogen receptor pathways in breast cancer using an Artificial Neural Networks based inference approach.** *Proceedings of the IEEE-EMBS International conference on Biomedical and Health Informatics*. Hong Kong, Shenzhen, 5-7 January, 2012.
(http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6211745&tag=1**)**

Gopal K Dhondalay, Dong L Tong, Graham R Ball**. Estrogen receptor status prediction for breast cancer using artificial neural network.** *Proceedings of the 2011 International Conference on Machine Learning and Cybernetics*, Guilin, 10-13 July, 2011.
(http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6016771**)**

Oral presentation and conferences

Nottingham Trent University seminar (final year talk)
Nottingham, UK January 2013

Neural Network 2012 (Summer School)
Oporto, Portugal July 2012

Cancer Bioinformatics Workshop
Cambridge, UK September 2010

British Breast Cancer Research Conference
Nottingham, UK September 2010

The Nottingham Trent University Annual Research School Conference
Nottingham, UK May 2010

Posters

Dhondalay, G. K.
What I did last summer (for PhD student travel bursary award)
Nottingham Trent University
Nottingham, UK September 2012

Dhondalay, G. K., Lemetre, C., Allen, T., Ball, G. R.
Multi dataset analysis for ER-associated genes in breast cancer
The Nottingham Trent University Annual Research School Conference
Nottingham, UK May 2011

Dhondalay, G. K., Lawrence, K.,Hoare, M., McCoy, R., Ball, G.R.
Use of artificial neural network technologies to visualise multidimensional influences on cell stress markers
National Stem Cell Network
University of York, UK April 2011

Dhondalay, G. K., Lemetre, C., Burnett, A., Lancashire, L. J., Ellis, I. O., Martin, S., Ball, G. R.
Computational interaction analysis for Redox reaction pathway
British Breast Group
Cambridge, UK March 2011

Dhondalay, G. K., Lemetre, C., Burnett, A., Lancashire, L. J., Caldas, C., Ellis, I. O., Ball, G. R.
Artificial neural network in prediction of lymphovascular invasion in primary breast cancer
British Breast Group
Cambridge, UK March 2011

Dhondalay, G. K., Lemetre, C., Burnett, A., Lancashire, L. J., Ellis, I. O., Martin, S., Ball, G. R.
Deciphering Redox reaction pathway using computational interaction analysis
British Breast Cancer Research Conference
Nottingham, UK September 2010

ANNEXURE 1:

Gopal K Dhondalay, Dong L Tong, Graham R Ball**. Estrogen receptor status prediction for breast cancer using artificial neural network.** *Proceedings of the 2011 International Conference on Machine Learning and Cybernetics*, Guilin, 10-13 July, 2011.
(http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6016771**)**

ANNEXURE 1

ESTROGEN RECEPTOR STATUS PREDICTION FOR BREAST CANCER USING ARTIFICIAL NEURAL NETWORK

Gopal K. Dhondalay, Dong L. Tong, Graham R. Ball

John van Geest Cancer Research Centre, School of Science and Technology, Nottingham Trent University, Nottingham, UK NG11 8NS
E-MAIL:
gopal.dhondalay@ntu.c.uk

Abstract:

The status of estrogen receptor (ER) has been profoundly associated with breast cancer. Numerous studies have been conducted to identify informative genes that are associated to ER status. However, the integrity of the reported genes is still inconclusive as the results are derived from small cohort of breast cancer patients (< 200 samples).

In this paper, we studied gene signatures from a cohort of 278 breast cancer samples, labelled in ER positive and ER negative classes, using artificial neural network (ANN). Our model has showed its efficacy for selecting significant genes compared to the previous study. The result also showed that the highly ranked genes have been previously reported in association to the breast cancer development.

Keywords: Estrogen receptor; Breast cancer; Microarray data; Artificial neural network; Prediction

1. Introduction

Estrogen receptor (ER) is a type of hormone receptor protein that acts as a stabiliser in hormonal system. From the literature, ER status, i.e. ER positive (over-expression) or ER negative (suppression), has been associated to breast cancer development, in which it has been normally over-expressed in breast cancer. For instance, Yaich et al [16] examined the influence on the alterations of ER-associated gene in 26 primary breast cancers. Chanrion et al [4] studied the expression of 47 genes in the molecular association to a cohort of 199 primary breast tumours. They identified 10 genes which showed the over-expression characteristics scattered in different ER positive molecular subgroups. Chen et al [5] in an approach to find novel predictive gene signatures of chemotherapy response, found genes to be significantly related to ER status. In another recent study Budczies et al [3] found 346 genes in a signature correlated to ER status in paraffin embedded breast cancer core biopsies.

There are numerous data mining approaches that have been proposed for analysing breast gene microarray data. For instance, Bloom et al [2] proposed a standard 3-layered backpropagation neural network to classify oligonucleotide-based breast microarray data and the average classification accuracy of 67% was achieved in their study. Meanwhile, Liu, Krishnan and Mondry [10] compared the prediction performance of greedy search and simulated annealing (SA) to classify 49 breast cancer microarrays into ER positive and ER negative classes. They reported 31 significant genes with the prediction accuracies of 89.8% in both greedy and SA approaches. Zhou, Wang and Dougherty [17], on the other hand, used linear regression method to select significant genes that are associated to BRCA1 and BRCA2 mutation-positive breast cancer. They used 22 breast tumour samples obtained from 21 patients.

Although, ER status in breast cancer has been studied in the past, however, the reported results are mostly based on small cohort of samples. Little studies in the ER status using large cohort of breast cancer patients. Thus, the objective of this paper is to study the ER status in a large cohort of breast cancer samples (278 samples) and to identify gene signatures that are responsible for the expression in the ER status using artificial neural network (ANN).

For the remaining of this paper, we first describe our ANN method. We then describe the data set and the comparison results. Finally, we conclude the paper.

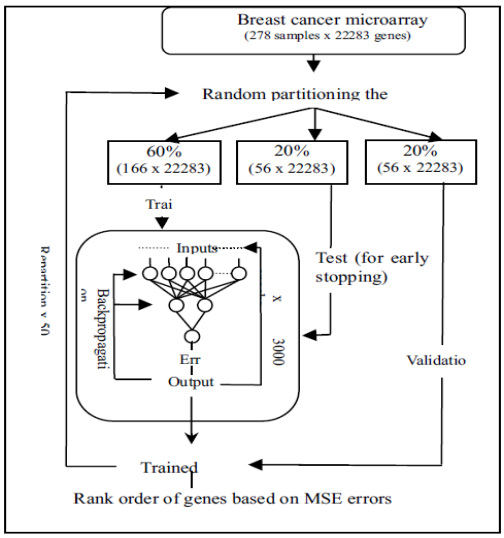
2. Artificial neural networks (ANNs)

A bespoke ANN model has been developed aiming at the ion identification from high throughput proteomics data in our previous studies [1, 8, 13]. This similar model has been recently applied to study the gene profiles in microarray data [9]. In this paper, we use the same model to predict ER status for breast cancer patients. Figure 1 illustrates the workflow of our model. For the detail of our algorithm, please refer to our previous study [13].

In brief, the data are first divided into 3 sample subsets, i.e. training, test and validation sets. The sample partition ratio is 60:20:20, i.e. 60% of the data is used to train the network, 20% act as the early stopping criteria for the network and the remaining 20% is used to validate the network model. A 3-layered ANN with backpropagation learning was constructed for predicting the training samples. In the network training process, stepwise approach is adopted, in which the number of input nodes (i.e. genes) will be increase by one each time the network is trained. In other words, we perform an exhaustive search throughout the whole gene population to find potential candidate genes in solving our problem. The search will iterate until the termination criteria is met. The validation set is then used to further examine the significance of the identified genes. The whole data partition, training and validation processes are repeated for 50 times. Figure 1 is a diagrammatic representation of ANN model.

We set the network epochs to 3000, 1000 threshold window times if no further improvement on mean squared error (MSE) with threshold of 0.01, learning rate of 0.1, momentum rate of 0.5 and random cross-validation sampling of 50 times. We use backpropagation learning algorithm with an exhaustive stepwise search and sigmoidal activation function. The MSE error was used to compute the network error as it is the most widely used error measurement for ANN.

Figure 1. Diagrammatic representation of ANN model



3. Breast cancer gene microarray data

The cDNA data set, labelled as E-GEOD-20194 (MAQC II), was downloaded from the EMBL-EBI database library (http://www.ebi.ac.uk/arrayexpress/). The dataset contains 278 samples categorized into 2 classes. One-hundred-sixty-four samples were confirmed as estrogen receptor positive (ER+) class and the remaining 114 samples in estrogen receptor negative (ER-). Each sample in the data set is associated with 22283 genes.

It is worth to note that this is a relatively new microarray data that is submitted by MicroArray Quality Control (MAQC) consortium [12]. This is the only publicly available breast gene microarray data that comprised of large cohort of breast cancer patients, as we are aware of. Most of the existing breast gene microarray analysis was performed based on a small number of samples (< 200 samples).

4. Results

We compared our results with the result reported by Liu and his co-workers [11], a study carried out by the MAQC consortium emphasizing on this data set.

To show an unbiased result comparison, we compared the top 100 genes selected by our model with the top 100 genes reported by Liu et al. [11]. Due to the limitation of space here, we only present the top 20 genes in this paper. Table 1 below summarises the comparison between the top
100 genes. The top 20 genes selected by our model is presented in Table 2.

A total 9 genes were overlapped with the top 100 genes reported by Liu et al. These overlapping genes are EGFR (probe set ID 201983_s_at), ESR1 (probe set ID 205225_at), IL6ST (probe set IDs 211000_s_at & 212195_at), MAPT (probe set ID 203929_s_at), DACH (probe set ID 205471_s_at), GATA3 (probe set ID 209602_s_at), GREB1 (probe set ID 205862_at) and SERPINA5 (probe set ID 209443_at). Among these overlapping genes, gene ESR1 is the highest ranked gene in our model (see Table 2). We are not able to compare the rank order of the genes as it is not supplied in the original study.

The highly ranked genes by our model have revealed significant association with the ER status. Most of the identified genes, specifically genes ESR1, GATA3 and CA12, have been previously reported in the association to breast cancer [6, 7, 14, 15].

Gene ESR1 (probe set ID 205225_at) produces estrogen receptors which are ligand activated transcription factors belongs to the nuclear hormone superfamily. It plays a role in regulating the expression of specific genes that involve in the cellular processes, such as growth, differentiation and function of reproductive systems.

Gene GATA3 (gene probe set I D s 209602_at, 209603_at, 209604_at), on the other hand, encodes transcription factors of GATA family which act as the regulators of T-cell and endothelial cell development.

Meanwhile, gene CA12 (gene probe set ids 215867_x_at, 203963_at, 210735_s_at) participates in the processes, such as respiration, bore resorption, formation of saliva, gastric acids and celebrospinal fluid. CA12 gene produces a type-1 membrane protein which is highly expressed in normal tissues, including colon, pancreas and kidney. However, it also shows a relative high expression in renal carcinomas.

ANNEXURE 1:

Gopal K Dhondalay, Dong L Tong, Graham R Ball**. Estrogen receptor status prediction for breast cancer using artificial neural network.** *Proceedings of the 2011 International Conference on Machine Learning and Cybernetics*, Guilin, 10-13 July, 2011.
(http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6016771**)**

Table 1. The comparison result based on the genes selected in the data set

Studies	No. of selected genes	Overlapped genes	Non-overlapping genes
ANN	100	9	91
Liu et al [11]	100	-	-

Table 2. The top 20 genes selected by ANN.

Rank Order	Probe set ID	Accession No.	Gene Symbol
1	205225_at	NM_000125	ESR1
2	209603_at	AI796169	GATA3
3	212956_at	AI348094	KIAA0882
4	214164_x_at	BF752277	FLJ20151
5	215867_x_at	AL050025	CA12
6	209602_s_at	AI796169	GATA3
7	214440_at	NM_000662	NAT1
8	204508_s_at	BC001012	FLJ20151
9	209604_s_at	BC003070	GATA3
10	203963_at	NM_001218	CA12
11	218195_at	NM_024573	FLJ12910
12	212960_at	BE646554	KIAA0882
13	209173_at	AF088867	AGR2
14	210735_s_at	BC000278	CA12
15	203628_at	H05812	IGF1R
16	211712_s_at	BC005830	ANXA9
17	218976_at	NM_021800	JDP1
18	212196_at	AW242916	IL6ST
19	205009_at	NM_003225	TFF1
20	202089_s_at	NM_012319	LIV-1

An ANN prediction model was constructed based on the top 100 selected genes. Using the validation set, our model achieves 77.62% of classification accuracy with sensitivity of 96.32% and specificity of 58.78% for ER positive class. For ER negative class, the sensitivity of 58.78% and specificity of 96.32% were achieved. This showed that our model is robust for identifying the most significant set of genes from high dimensional gene microarray data.

We also performed the receiver operating sample subset is depicted in Figure 2. The y-axis of the curve was plot based on the equation 1 and the x-axis of the curve is based on the equation 2. The area under the ROC curve for the validation sample subset is 0.935.

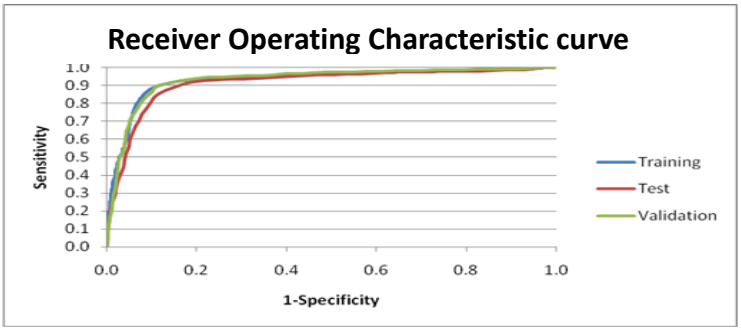
Sensitivity = (Positive correctly classified samples)/(Total positive samples in the class) (1)

Specificity = (Negative correctly classified samples)/(Total negative samples in the class) (2)

ANNEXURE 1:

Gopal K Dhondalay, Dong L Tong, Graham R Ball**. Estrogen receptor status prediction for breast cancer using artificial neural network.** *Proceedings of the 2011 International Conference on Machine Learning and Cybernetics*, Guilin, 10-13 July, 2011.
(http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6016771**)**

Figure 2: The averaged ROC curve based on 50 random sampling for each sample set



4. Discussion and Conclusions

Our model has shown the ability to identify the most significant gene subset from the breast gene microarray data. Although only 9 out of 100 selected genes by our model are overlapped with the top 100 genes reported by Liu et al. [11], our model has significantly selected 3 most important predictor genes which have been previously associated with breast cancer. The ESR1 gene which is highest ranked by ANN shows that it is a strong predictor to discriminate ER status in breast cancer. Gene ESR1 has been profoundly used as a general marker for cancer prognosis in the literature [7].

Furthermore, we apply random cross-validation on the data set to obtain the unbiased top 100 genes from the data set. We use exhaustive search method coupled with backpropagation of errors to perform a thorough search on the entire gene population to ensure that the most significant gene subset will always be selected by our model. Liu et al, on the other hand, used Naive Bayesian approach to identify their top 100 genes. Consequently, some of the "true" marker genes may be overlooked in their search process.

Based on the top 100 genes selected by our model, we achieved an average classification accuracy of 77.62% in the validation set and high sensitivity value (96.32%) in the ER positive class. However, low sensitivity value (58.78%) in the ER negative class is obtained. This might due to a lack of standard threshold used by the immunohistochemistry to label ER status on the samples. The low sensitivity in ER negative class may suggest that high possibility of misclassifying true negative samples as positives in the pathological screening. Therefore, study on the boundary measures of the threshold for immunohistochemistry assay will be carried out in future. In addition, we will also conduct the same experiments with different data mining approaches to investigate the implications of different mining algorithms on the classification performance and gene selection. This future study will help in formulating the most significant and overlapped gene markers for ER status in breast gene microarray data.

We believe our approach is robust in providing the most relevant set of genes that are highly associated to the ER status in the breast cancer. We use an exhaustive search method and the universal computation power of ANN to identify the most significant genes from high dimension (>22000 genes) and large cohort of >200 breast cancer patients.

Acknowledgements: This paper is supported by the John and Lucille van Geest Foundation.

References

[1] G. Ball, S. Mian, F. Holding, R.O. Allibone, J. Lowe, S. Ali, G. Li, S McCardle, I.O. Ellis, C. Creaser and R. Rees, "An integrated approach utilizing artificial neural networks and SELDI mass spectrometry for the classification

ANNEXURE 1:

Gopal K Dhondalay, Dong L Tong, Graham R Ball**. Estrogen receptor status prediction for breast cancer using artificial neural network.** *Proceedings of the 2011 International Conference on Machine Learning and Cybernetics*, Guilin, 10-13 July, 2011.
(http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6016771**)**

of human tumours and rapid identification of potential biomarkers", Bioinformatics Vol 18, No. 3, pp. 395-404, 2002.

[2] G. Bloom, Yang IV, D. Boulware, K.Y. Kwong, D. Coppola, S. Eschrich, J. Quackenbush and T.J. Yeatman, "Multi-Platform, Multi-Site, Microarray-Based Human Tumor Classification", Am J Pathol, Vol 164, No. 1, pp. 9-16, 2004.

[3] J. Budczies, W. Weichert, A. Noske, B.M. Müller, C. Weller, T. Wittenberger, H.P. Hofmann, M. Dietel, C. Denkert and V. Gekeler, "Genome-wide gene expression profiling of formalin-fixed paraffin-embedded breast cancer core biopsies using microarrays", J Histochem Cytochem, Vol 59, No. 2, pp. 146-157, 2011.

[4] M. Chanrion, H. Fontaine, C. Rodriguez, V. Negre, F. Bibeau, C. Theillet, A. Hénaut and J-M. Darbon, "A new molecular breast cancer subclass defined from a large scale real-time quantitative RT-PCR study", BMC Cancer, Vol 7, No. 39, 2007.

[5] Y. Chen, C. Chen, B. Yang, Q. Xu, F. Wu, F. Liu, X. Ye, X. Meng, B. Mougin, G. Liu, Z. Shen, Z. Shao and J. Wu, "Estrogen receptor-related genes as an important panel of predictors for breast cancer response to neoadjuvant chemotherapy", Cancer Lett., Vol 302, No. 1, pp. 63-68, 2011.

[6] R.V. Hoch, D.A. Thompson, R.J. Baker and R.J. Weigel, "GATA-3 is expressed in association with estrogen receptor in breast cancer", Int. J. Cancer (Pred. Oncol.), Vol. 84, pp. 122–128, 1999.

[7] F. Holst, P.R. Stahl, C. Ruiz, O. Hellwinkel, Z. Jehan, M. Wendland, A. Lebeau, L. Terracciano, K. Al-Kuraya, F. Jänicke, G. Sauter and R. Simon, "Estrogen receptor alpha (ESR1) gene amplification is frequent in breast cancer", Nat Genet, No. 39, pp. 655-660, 2007.

[8] L.J. Lancashire, O. Schmid, H. Shah and G. Ball, "Classification of bacterial species from proteomic data using combinatorial approaches incorporating artificial neural networks, cluster analysis and principal components analysis", Bioinformatics, Vol 21, No. 10, pp. 2191-2199, 2005.

[9] L.J. Lancashire, C. Lemetre, G.R. Ball, "An introduction to artificial neural networks in bioinformatics – application to complex microarray and spectrometry dataset in cancer studies", Briefings in Bioinformatics, Vol 10, No. 3, pp. 315-329, 2009.

[10] X. Liu, A. Krishnan and A. Mondry, "An Entropy-based gene selection method for cancer classification using microarray data", BMC Bioinformatics, Vol 6, No. 76, 2005.

[11] Q. Liu, A.H. Sung, Z. Chen, J. Liu, X. Huang and Y. Deng, "Feature Selection and classification of MAQC-II Breast Cancer and Multiple Myeloma Microarray Gene Expression Data", PLoS One, Vol 4, No. 12, pp. e8250, 2009.

[12] MAQC Consortium, "The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models", Nat Biotechnol., Vol 28, No. 8, pp. 827-838, 2010.

[13] 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 Clin Appl., Vol 1, No. 6, pp. 605-20, 2007.

[14] R. Mehra, S. Varambally, L. Ding, R. Shen, M.S. Sabel, D. Ghosh, A.M. Chinnaiyan and C.G. Kleer, "Identification of GATA3 as a breast cancer prognostic marker by global gene expression meta-analysis", Cancer Res, Vol. 65, No. 24, pp. 11259-1164, 2005.

[15] P.H. Watson, S.K. Chia, C.C. Wykoff, C. Han, R.D. Leek, W.S. Sly, K.C. Gatter, P. Ratcliffe and A.L. Harris, "Carbonic anhydrase XII is a marker of good prognosis in invasive breast carcinoma", British Journal of Cancer, Vol. 88, pp. 1065–1070, 2003.

[16] L. Yaich, W.D. Dupont, D.R. Cavener and F.F. Parl, "Analysis of the PvuII Restriction Fragment-length Polymorphism and Exon Structure of the Estrogen Receptor Gene in Breast Cancer and Peripheral Blood", Cancer Res, Vol. 52, pp. 77:83, 1992.

[17] X. Zhou, X. Wang and E.R. Dougherty, "Gene Selection Using Logistic Regressions Based on AIC, BIC, and MDL Criteria", New Mathematics and Natural Computation, Vol 1, No. 1, pp. 129.

ANNEXURE 2:

Gopal K Dhondalay, Christophe Lemetre and Graham R Ball. **Modelling estrogen receptor pathways in breast cancer using an Artificial Neural Networks based inference approach.** *Proceedings of the IEEE-EMBS International conference on Biomedical and Health Informatics*. Hong Kong, Shenzhen, 5-7 January, 2012.
(http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6211745&tag=1**)**

ANNEXURE 2

MODELING ESTROGEN RECEPTOR PATHWAYS IN BREAST CANCER USING AN ARTIFICIAL NEURAL NETWORKS BASED INFERENCE APPROACH

Gopal K Dhondalay, Christophe Lemetre and Graham R Ball

Gopal K. Dhondalay: The John van Geest Cancer Research Centre, Nottingham Trent University, Clifton Lane, Nottingham. NG11 8NS UK
Christophe Lemetre: Albert Einstein College of Medicine, 1301 Morris Park Avenue, Price 353, Bronx, New York 10461 USA.
Graham R Ball (corresponding author): The John van Geest Cancer Research Centre, Nottingham Trent University, Clifton Lane, Nottingham. NG11 8NS. UK (email: Graham.balls@ntu.ac.uk)

Manuscript received November 10th, 2011. This work was supported in part by the van Geest Trust and by the Higher Education Funding Council for England.

Abstract:

Estrogen receptor (ER) status is an important consideration in the prognosis and management of breast cancer patients, dictating treatment and patient management. While the prognosis of ER positive patients is generally poorer because of treatments such as Tamoxifen this situation has been reversed. Some detail is known of the ER pathway, however this has been based on reductionist studies of small numbers of markers. Here we present an Artificial Neural Network (ANN) using a feed forward back-propagation algorithm applied to a three layer multi-layer perceptron based approach that facilitates a wider more holistic approach to the identification of genes associated with ER status and the modeling of their interactions with one another in the context of a pathway.

I. Introduction

The treatment of breast cancer covers a wide spectrum of heterogeneity from locally advanced to metastatic spread of disease[1] and a range of molecular subtypes have been identified. Surgical treatment includes axillary lymph node resection, radical mastectomy which aids control locoregional
disease progression causing distant metastasis. Radiotherapy is another integral part of treatment given to breast chest and axilla to prevent brain lung and bone metastasis.

It is well known that estrogen receptor (ER) plays a pivotal role in the biology of breast cancer [2]. It is an commonly used prognostic factor predicting a favourable response to treatment and better survival prospects. Hormonal treatment is given for breast cancers expressing positive estrogens and progesterone hormone receptors. Antiestrogens, progestins, aromatase inhibitors, gonadotropin releasing hormone analogues and ablative endocrine procedures can be included under hormonal treatment. Selective estrogen receptor modulators (SERM) like Tamoxifen and Raloxifen and aromatase inhibitors like Letrozole and Anastrazole are the two main classes of drugs used in ER positive breast cancer [3].

Current investigations of the ER pathway have largely focused on analysis of annotated literature based databases such as ingenuity pathways analysis.

Artificial Neural Networks (ANNs) are a form of machine learning capable of accurately modeling biological systems and identifying biomarkers. They are capable producing models that discriminate well between multiple classes for blind data [4], [5] while allowing parameterization

to determine the influence of causal agents within the system. ANNs have been applied to a number of gene micro-array studies in cancer in a number of ways [6], [7]. Many of the studies conducted pre-filter the probe-set before ANN analysis. A number of studies have indicated the approach can produce generalized models with a greater accuracy than conventional statistical techniques in medical diagnostics [4],[8] without relying on predetermined relationships as in other modeling techniques.

Network inference is a branch of systems biology wherein interactions among the contender of the system are logically connected to each other representing behavior of the system. The simplest network in systems graph is a pairwise relation between Nodes (called vertices) through Edges (called links). The nodes of cellular systems may be genes or mRNA, protein or other molecules. Edges are comprised with a source node and a target node representing the relation between nodes. Depending on the availability of information edges can have positive or negative values, representing activation or inhibition; or weights quantifying confidence levels, strength, or reaction speeds.

Here we apply an ANN based network inference approach to model and identify the key interactions between identified genes associated with the Estrogen Receptor pathway. From these interactions we aim to:- Identify probes and thus genes that interact with ER from a gene expression array study through an ANN based data mining approach; identify the strongest interactions between these probes through an ANN based network inference approach and evaluate the biological relevance of the findings.

II. Computational approaches

A. Data sets used for analysis.

A pre-normalized microarray breast cancer cDNA expression profile was downloaded from freely available EMBL-EBI database library (http://www.ebi.ac.uk/arrayexpress). The cDNA microarray profile was found to be developed on Affymetrix HG-133U array from E-GEOD-20194 study. The breast cancer microarray dataset contained 22283 gene probes in 278 arrays of primary breast cancer samples before neo-adjuvant therapy. The samples were categorized into 2 classes based on immune-histochemical Estrogen Receptor (ER) status as ER positive (ER+; 164 samples) class and the remaining as ER negative (ER-; 114 samples).

B. Phase 1 Identification of biomarker probes associated with Estrogen Receptor status.

The analysis and modeling of gene-gene interactions from a whole microarray probe set is an extreme undertaking. If a whole Affymetrix probe-set (having in excess of 22,000 probes was to be analyzed using a bidirectional interaction modeling then in excess of $9x10_8$ interactions would have to be modeled, analyzed and filtered. The computational requirements coupled with the potential for false discovery in such an interaction model would be both very large making analysis very difficult. To reduce these constraints the probe-set used for network inference was first analyzed to identify the top 100 most appropriate genes. This was achieved by coupling a filter and a wrapper method with early stopping and a Monte Carlo cross validation approach (utilizing a 60:20:20 randomized split of the data), analyzing single probes based on their ability to predict for an unseen data set. This process was repeated 10 times and the results averaged and compared.

This approach has been used successfully in a number of studies to identify genes showing high sensitivity, specificity and biological relevance that are associated with a given clinical question [4], [9],[10]. A more detailed description of this approach is presented in [4]. P-values for each of the top genes are quoted for further validation.

C. Phase 2 Artificial Neural Network based Network inference.

The top 100 genes identified in phase 1 of the analysis were applied to an ANN based network inference algorithm [11]. This approach utilizes a feed forward back propagation algorithm to predict the expression level of a given probe from the other 99 probes in the set. The model is trained to optimum performance on unseen data using Monte Carlo cross validation. It is then parameterized to determine a weighting of the link from input probe to output probe. This link has both a sign and a magnitude. This process is repeated for the remaining 99 probes setting each probe as the output generating a total of 100 models.

The parameterizations of each of these models are then integrated to produce an interaction matrix. The 100 strongest interactions (representing the top 1% of interactions) were then further selected for visualization. Using this multifactorial nonlinear ANN based inference approach has an advantage over simple linear regression based approaches in that the interaction may be nonlinear and that bi-directionality can be modeled.

D. Visualization

The gene interaction values from the software generates tab delimitated text formatted file. The format of the output file comprises three columns represented as source and target, sandwiching the interaction values in terms of weighted vectors in between. For the visualization of the gene interactions in this study, we have used a free standalone Cytoscape Ver 2.7.3 software (www.cytoscape.org). Interactions (edges) are represented as either positive (green) or negative (red) with magnitude of interaction represented by the width of the line and directionality represented by arrows.

III. RESULTS

A. Probe selection

Examination of the rank order of probes indicated good concordance between the 10 repeats of the analysis. The mean RMS error for seen data for classification between ER positive and negative was 0.113 +/- 0.008. The mean RMS error for blind data was 0.118 +/- 0.009. In all 10 repeats the highest performing probe was ESR1 representing the gene coding for Estrogen Receptor alpha. Receiver operator characteristic curves (ROC) were plotted for the models and a mean area of 0.930 was produced for seen data and 0.917 for blind data. The top 100 predictive probes were taken forward for further analysis. Here the top 20 only are presented for the sake of brevity (Table 1).

The genes found in the top 100 were :-ACADSB, ADCY9, AGR2, AKR7A3(x2), ANXA9(x2), APBB2(x2), C1orf34, C6orf29, CA12(x3), CDH3, CELSR1, CGI-49, CIRBP, LSTN2, COX6C, CRIP1, CRNKL1, CYP2B6, DACH(x2), DNALI1, DP1, EGFR, ELOVL2, ENPP1, ESR1(x2), FBP1, FLJ10647, FLJ12910, FLJ20151(Also CA12)(x2), FOXA1, GAMT, GATA3(x3), GFRA1, GREB1, IGF1R(x2), IGFBP4, IL6ST(x4), JDP1, KIAA0040, KIAA0232, KIAA0876(x2), KIAA0882(x2), KIAA1243, LASS6(x2), LIV-1(x2), MAGED2, MAPT(x3), MCCC2, MGC2601, MLPH, MYB, NAT1, NME3, NPD009(x2), PBX1(x2), PDEF(x2), PH-4, PTP4A2, RAB5EP, RARA, RHOB, RNB6, SCUBE2, SEMA3F, SERPINA5, SKP1A, TFF1, TFF3, TONDU, UGCG(x2), VAV3(x2), WWP1(x2) and XBP1. Only 73 genes are listed as some probes were repeated multiple times for the same gene (numbers of times indicated in parenthesis).

B. Network inference.

The top 100 interactions are illustrated in Figure 1. Red interactions indicate a negative interaction, Green interactions indicate a positive interaction. Examination of the network map

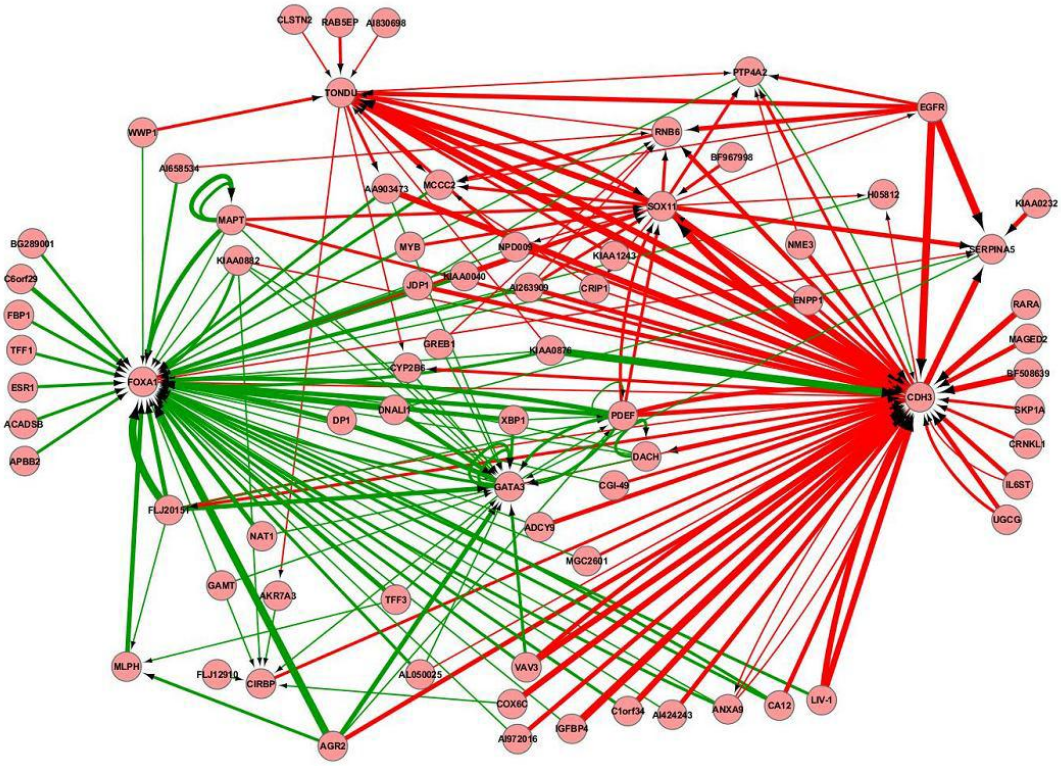
Gopal K Dhondalay, Christophe Lemetre and Graham R Ball. **Modelling estrogen receptor pathways in breast cancer using an Artificial Neural Networks based inference approach.** *Proceedings of the IEEE-EMBS International conference on Biomedical and Health Informatics*. Hong Kong, Shenzhen, 5-7 January, 2012.
(http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6211745&tag=1**)**

indicates a strong connectivity to GATA3 (GATA binding protein 3) with positive influence, FOXA1(forkhead box A1) with positive influence, TONDU (transcription cofactor vestigial-like protein 1) with negative influence, CDH3 (cadherin 3, type 1, Pcadherin) with negative influence and SOX11 (transcription factor SOX-11)with negative influence. The strongest negative interaction in the network was between CDH3 and SOX11. The strongest positive interaction was between FLJ20151(a variant of Carbonic Anhydrase 12) and FOXA1.

Table 1: Summary of top ranked 20 genes from stepwise ANN

Rank Order	Probe set ID	Gene Symbol	p-value	Regulation
1	205225_at	ESR1	4.49E-05	Up
2	209603_at	GATA3	8.98E-05	Up
3	212956_at	KIAA0882	1.35E-04	Up
4	214164_x_at	FLJ20151	1.80E-04	Up
5	215867_x_at	CA12	2.24E-04	Up
6	209602_s_at	GATA3	2.69E-04	Up
7	214440_at	NAT1	3.14E-04	Up
8	204508_s_at	FLJ20151	3.59E-04	Up
9	209604_s_at	GATA3	4.04E-04	Up
10	203963_at	CA12	4.49E-04	Up
11	218195_at	FLJ12910	4.94E-04	Up
12	212960_at	KIAA0882	5.39E-04	Down
13	209173_at	AGR2	5.83E-04	Up
14	210735_s_at	CA12	6.28E-04	Up
15	203628_at	IGF1R	6.73E-04	Up
16	211712_s_at	ANXA9	7.18E-04	Up
17	218976_at	JDP1	7.63E-04	Up
18	212196_at	IL6ST	8.08E-04	Up
19	205009_at	TFF1	8.53E-04	Up
20	202089_s_at	LIV-1	8.98E-04	Up

Figure 1. Inferred network for ER status in breast cancer.



IV. DISCUSSION.

The aims of this study were to identify probes and thus genes that interact with ER from a gene expression array study through an ANN based data mining approach; to identify the strongest interactions between these probes through an ANN based network inference approach and to identify the biological relevance of the findings.

The probe set selected from the array shows a lot of potential in its representation of the biology of ER status. Firstly the top probe consistently identified was ESR1 the gene that codes for estrogen receptor alpha. The presence of this probe in the highest ranked position indicates the other top ranked probes are also of potential biological importance. This conclusion is further strengthened by the fact that different probes for the same genes also occur multiple times in the top 100. In some cases all of the probes in the top 100 are represented on the array.

From the combination of data mining and network inference the key genes seem to be GATA3, FOXA1, TONDU, CDH3, SOX11 and CA12. Of these FOXA1, GATA, TONDU and SOX11 are known transcription factors and thus would be likely to be influential in the coregulation of other genes.

GATA3 is very well known to be associated with ER phenotype in breast cancer and is thought to regulate genes critical to the hormone response phenotype of breast cancer [2],[12]. Furthermore a cross regulatory loop has been identified between ESR1 and GATA3 [13]. GATA3 has
previously been found to be associated with ER status in breast cancer through ANN data mining of expression array data [10].

FOXA1 is also known to have a strong role in the Estrogen receptor phenotype [14], [15] and is likely to influence other genes in the pathway. It has also been associate with poor prognostic outcome in breast cancer [9]. In this study it also showed a significant association with ER and P- Cadherin.

SOX11has not directly been shown to relate to ER status in breast cancer. It has been shown to associate with prognosis in ovarian cancer [16] and could potentially have a role in breast cancer. It has been shown to have a role in tumorigenesis and was found by Christodoulou et al [17] to be associated with ER status.

CDH3 (P-Cadherin) has been associated with the basal phenotype in breast cancer [18] and with ER independent growth [19]. Furthermore using ANN based techniques [7] identified it as associated with ER status in breast cancer.

TONDU has no currently identified role in cancer.

CA12 is associated with hypoxia and poor prognosis in breast cancer. Tumor tissue in hypoxic condition is subjected to physiological stress causing accumulation of acidic products of glycolytic metabolism. Hypoxia inducible growth factor brings about a change in cellular expression programme by activating genes involved in angiogenesis, anaerobic glycolysis, adhesion and invasion. It has been shown to associate with ER status in breast cancer through a distil enhancer [20].

ESR1 is not a strong hub in the interaction network which is somewhat surprising on first consideration. However given that all of the elements included in the pathway are strongly associated with ER some of the variance associated with the marker is likely to have been removed by inclusion of these factors. This has the effect of emphasising the interaction between factors associated with ER status but not with ER directly.

ANNEXURE 2:

Gopal K Dhondalay, Christophe Lemetre and Graham R Ball. **Modelling estrogen receptor pathways in breast cancer using an Artificial Neural Networks based inference approach.** *Proceedings of the IEEE-EMBS International conference on Biomedical and Health Informatics*. Hong Kong, Shenzhen, 5-7 January, 2012.
(http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6211745&tag=1**)**

V. CONCLUSIONS

Utilizing the combined approach of ANN based data mining and ANN based network inference we have identified 6 genes that associate with the ER system in breast cancer. Of these 4 are known to be strongly associated with ER status. Through network inference we have identified interactions that with further evaluation have the potential to provide further insights into the etiology of Estrogen Receptor in breast cancer.

ACKNOWLEDGMENT: The authors would like to acknowledge the support of the van Geest foundation and the Higher Education Funding Council for England that enabled this research to be conducted.

REFERENCES

[1] Osborne CK, Elledge RM, Brown PH, Hilsenbeck SG. BRCA1 in clinical breast cancer. Breast Dis. 1998 Apr;10(1-2):77-88.

[2] Jason S Carroll, Clifford A Meyer, Jun Song, Wei Li, Timothy R Geistlinger, Je´roˆme Eeckhoute, Alexander S Brodsky, Erika Krasnickas Keeton, Kirsten C Fertuck, Giles F Hall, Qianben Wang, Stefan Bekiranov, Victor Sementchenko, Edward A Fox, Pamela A Silver, Thomas R Gingeras, X Shirley Liu & Myles Brown. 2006. Genome-wide analysis of estrogen receptor binding sites Nature Genetics Volume 38:11, 1289-1297.

[3] Jordan VC, Brodie AM. Development and evolution of therapies targeted to the estrogen receptor for the treatment and prevention of breast cancer. Steroids. 2007 Jan;72(1):7-25

[4] 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. (2010) A validated gene expression profile for detecting clinical outcome in breast cancer using artificial neural networks. Breast Cancer Res Treat. Volume 120, Number 1, 83-93

[5] Ball, G., Mian, S., Holding, F., Allibone, RO., Lowe, J., Ali, S., Li, G., Mccardle, S., Ellis, IO., Creaser, C. And Rees, RC., 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. vol 18 (3) , pp. 395-404.

[6] 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.

[7] Sofia Gruvberger, Markus Ringne´r, Yidong Chen, Sujatha Panavally, Lao H. Saal, Åke Borg, Mårten Ferno Carsten Peterson, and Paul S. Meltzer. 2001. Estrogen Receptor Status in Breast Cancer Is Associated with Remarkably Distinct Gene Expression Patterns. Cancer Research 61, 5979–5984.

[8] Paulo J. Lisboa, Azzam F.G. Taktak. 2006. The use of artificial neural networks in decision support in cancer: A systematic review J Neural Networks Volume 19, Issue 4, Pages 408-415

[9] Hany Onsy Habashy, Desmond G Powe, Emad A Rakha, Graham Ball, Julia Gee, Robert I Nicholson, And Ian O Ellis. 2008. Forkhead-box A1 (FOXA1) expression in breast cancer and its prognostic significance. European Journal of Cancer, 44:11, 1541-1551.

[10] Lancashire LJ., Rees RC And 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. Artificial Intelligence in Medicine. 2008 43, 99- 111.

[11] C., Lancashire L. J., Rees R. C., And Ball G. R. 2009. Artificial Neural Network Based Algorithm for Biomarker Interactions Modeling. Lecture notes in Computer Science 5517, Bio-Inspired Systems: Computational and Ambient Intelligence. Eds. Cabestany, Sandoval, Prieto and Corchado. pp 877-885

[12] Renee V. Hoch, Devon A. Thompson, Robin J. Baker and Ronald J. Weigel, 1999, GATA-3 Is Expressed In Association With Estrogen Receptor In Breast Cancer. *Int. J. Cancer (Pred. Oncol.):* **84,** 122–128 (1999).

[13] Jérôme Eeckhoute, Erika Krasnickas Keeton, Mathieu Lupien, Susan A. Krum, Jason S. Carroll, and Myles Brown, 2007. Positive Cross- Regulatory Loop Ties GATA-3 to Estrogen Receptor α Expression in Breast Cancer. Cancer Res 67; 6477-6483

[14] Antoni Hurtado, Kelly A Holmes, Caryn S Ross-Innes, Dominic Schmidt & Jason S Carroll. 2011. FOXA1 is a key determinant of estrogen receptor function and endocrine response. Nature Genetics. Volume: 43, 27–33.

[15] Josée Laganière, Geneviève Deblois, Céline Lefebvre, Alain R. Bataille, François Robert, and Vincent Giguère. 2005. Location analysis of estrogen receptor α target promoters reveals that FOXA1 defines a domain of the estrogen response. PNAS 102:33, 11651-11656

[16] Donal J. Brennan, Sara Ek, Emma Doyle, Thomas Drew, Michael Foley, Grainne Flannelly, Darran P. O'Connor, William M. Gallagher, Sami Kilpinen, Olli-Pekka Kallioniemie, Karin Jirstrom, Colm O'Herlihy, Carl A.K. Borrebaeck. 2009. The transcription factor Sox11 is a prognostic factor for improved recurrence-free survival in epithelial ovarian cancer European Journal of Cancer. Volume 45, Issue 8, Pages 1510-1517 .

[17] E. Christodoulou, M. Ioannou, M. Kafousi, E. Sanidas, G. Papagiannakis, V. Danilatou, G. Tsiliki, T. Margaritis, H. Kondylakis, D. Manakanatas, L. Koumakis, A. Kanterakis, S. Vassilaros, M. Tsiknakis, A. Analyti, G. Potamias, D. Tsiftsis, E. Stathopoulos, D. Kafetzopoulos.2008. A new gene expression signature related to breast cancer Estrogen Receptor status. Proceedings of the 8th IEEE International Conference on BioInformatics and BioEngineering, BIBE 2008. Pp 1 – 7

ANNEXURE 2:

Gopal K Dhondalay, Christophe Lemetre and Graham R Ball. **Modelling estrogen receptor pathways in breast cancer using an Artificial Neural Networks based inference approach.** *Proceedings of the IEEE-EMBS International conference on Biomedical and Health Informatics*. Hong Kong, Shenzhen, 5-7 January, 2012.
(http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6211745&tag=1**)**

[18] Joana Paredes, Fernanda Milanezi, Jorge Sergio Reis-Filho, Dina Leitão, Daniel Athanazio, Fernando Schmitt. 2002. Aberrant P-Cadherin Expression: Is it Associated with Estrogen-Independent Growth in Breast Cancer? Pathology Research and Practice, Volume 198, Number 12 795-801.

[19]Sousa B, Paredes J, Milanezi F, Lopes N, Martins D, Dufloth R, Vieira D, Albergaria A, Veronese L, Carneiro V, Carvalho S, Costa JL, Zeferino L, Schmitt F. 2010. P-cadherin, vimentin and CK14 for identification of basal-like phenotype in breast carcinomas: an immunohistochemical study. Histol Histopathol. 25(8):963-74.

[20] Daniel H. Barnett, Shubin Sheng, Tze Howe Charn, Abdul Waheed, William S. Sly, Chin-Yo Lin, Edison T. Liu, and Benita S. Katzenellenbogen. 2008 Estrogen Receptor Regulation of Carbonic Anhydrase XII through a Distal Enhancer in Breast Cancer. Cancer Res 68; 3505.