Predicting disease phenotypes based on the molecular networks with Condition-Responsive Correlation

Sejoon Lee

Department of Bio and Brain Engineering,
KAIST, Daejeon, Republic of Korea
E-mail: sejoon@biosoft.kaist.ac.kr

Eunjung Lee

Department of Medicine,
Brigham and Women's Hospital,
Harvard Medical School,
Boston, MA 02115, USA
E-mail: ejalice.lee@gmail.com

Kwang H. Lee* and Doheon Lee*

Department of Bio and Brain Engineering,
KAIST, Daejeon, Republic of Korea
E-mail: khlee@biosoft.kaist.ac.kr
E-mail: dhlee@biosoft.kaist.ac.kr
*Corresponding authors

Abstract: Network-based methods using molecular interaction networks integrated with gene expression profiles have been proposed to solve problems, which arose from smaller number of samples compared with the large number of predictors. However, previous network-based methods, which have focused only on expression levels of proteins, nodes in the network through the identification of condition-responsive interactions. We propose a novel network-based classification, which focuses on both nodes with discriminative expression levels and edges with Condition-Responsive Correlations (CRCs) across two phenotypes. We found that modules with condition-responsive interactions provide candidate molecular models for diseases and show improved performances compared conventional gene-centric classification methods.

Keywords: molecular module; CRC; condition-responsive correlation; network-based phenotype classification.

Reference to this paper should be made as follows: Lee, S., Lee, E., Lee, K.H. and Lee, D. (2011) 'Predicting disease phenotypes based on the molecular networks with Condition-Responsive Correlation', *Int. J. Data Mining and Bioinformatics*, Vol. 5, No. 2, pp.131–142.

Biographical notes: S. Lee received his BS in Computer Science from Soongsil University, and MS in the Department of Bio and Brain Engineering from KAIST. He is currently a Doctoral Student in the Department of Bio and Brain Engineering at KAIST. His research interests include bioinformatics, systems biology and text mining.

E. Lee received her BS and MS in Computer Science and PhD from the Department of Bio and Brain Engineering, KAIST. She is currently a Postdoctoral Research Fellow in Brigham and Women's Hospital and Harvard Medical School, USA. Her research interests include integrative pathway and network analysis.

K.H. Lee received his DEA and Dr. Ing. Degrees from the Department of Computer Science, INSA de Lyon University, France, in 1982 and 1985, respectively, and the Dr. Etat. Degree from the Department of Computer Science, INSA de Lyon University, France, in 1988. He is a Professor of the Department of Bio and Brain Engineering, and a Chair Professor of Mirae Corporation at KAIST. His research interests include fuzzy systems, artificial intelligence and bioinformatics.

D. Lee received the BS, MS and PhD in Computer Science from KAIST, Daejeon, Korea, in 1990, 1992 and 1995, respectively. He has conducted visiting researches in the University of Texas, Austin, and National Institutes of Health, Bethesda, in 1999 and 2002. He is now a Professor of the Department of Bio and Brain Engineering at KAIST. He is an Associate Editor for ACM Transactions on Internet Technology and Computers in Biology and Medicine. His research interests include bio-data mining, bio-system modelling and bioinformatics.

1 Introduction

Genome-wide expression profiles of diseased samples have been exploited to predict various disease states (Alizadeh et al., 2000; Golub et al., 1999; Ramaswamy et al., 2003; Wang et al., 2005). Golub et al. (1999) suggested a formalised strategy for discovering and predicting cancer classes based on gene expression monitoring. Alizadeh et al. (2000) showed that genome-wide expression-profiles-based molecular classification of tumours can identify undetected and clinically significant subtypes of cancers. In recent years, studies have addressed the question of classification problem about metastasis of cancer based on relatively large number of genome-wide expression profiles. Ramaswamy et al. (2003) found 17 gene signatures associated with metastasis, and Wang et al. (2005) identified a set of 70 gene markers for breast cancer metastasis using 286 gene expression profiles. However, the fact that typical microarray expression profiles have large number of predictors, i.e., expression measurements of tens of thousands of genes, compared with relatively small number (at most a few hundreds) of samples makes it difficult to obtain satisfactory accuracies in some classification problems especially related with complex diseases such as cancer. To address this challenge, extracting smaller number of predictors, called marker genes, by measuring the discriminative powers of individual predictors across two phenotypes has been widely incorporated in the classification procedure (Antoniadis et al., 2003).

Yet, another challenge of expression-based classification arose from weak signals of individual predictors due to cellular heterogeneity within tumour tissues and genetic heterogeneity across patients. Distinct subtypes of cancers have distinct gene expression patterns, and individual samples have different aberration components even in the same pathways (Wood et al., 2007). These genetic heterogeneities of tumours are more likely to be the basis not only for wide variations in tumour behaviours and responsiveness to therapies, but also for the weak signals of individual genes in their mRNA levels. To cope with this heterogeneity, it has been proposed to combine the expression measurements for genes in the same functional modules extracted from Gene Ontology (GO), curated pathways, or protein–protein interaction networks, and use the combined activity levels of the modules as more informative predictors for disease classification (Chuang et al., 2007; Guo et al., 2005; Lee et al., 2008). Guo et al. (2005) suggested a method using a simple mean or median expression level of all member genes in each pathway based on GO. However, this method did not show significant improvement over the individual gene-based approach. This might be due to added noise from genes non-responsive in mRNA level in the pathways extracted from GO.

Chuang et al. (2007) proposed a network-based classification of breast cancer metastasis, which not only improved prediction accuracy and reproducibility, but also provided a novel hypothesis of underlying mechanism. The method extracted subnetworks from protein interaction networks by integrating with functional expression profiles of breast cancer patients. While Chuang et al. (2007) only focused on expression levels of proteins, which are nodes in the networks, other groups focused on edges in the networks to capture interactions, which occur responsively to conditions of interest (Guo et al., 2007; Mani et al., 2008). However, these edge-based approaches have not yet been applied for the classification to our knowledge.

Here, we propose a novel network-based classification strategy, which focuses on both condition-responsive proteins (nodes) and interactions (edges) in the network to extract functional modules. The interactions between each pair of two proteins are screened to have CRCs in their expression levels across two phenotypes, and the activity level of the module is inferred from a subset of member genes in the module whose combined expression levels deliver the maximal discriminative power. The proposed method has successfully identified subnetworks altered in a specific phenotype in terms of interactions suggesting candidate pathogenic processes, and their activities inferred from a subset of member genes serve as better predictors in classification compared with the conventional gene-centric procedures.

This paper is organised as follows: In Section 2, we describe our five data sets related to diseases and the proposed method in detail, and how to evaluate the performance using Area Under ROC Curve (AUC). In Section 3, we explain CTC modules in five data sets, and interpret modules in the case of prostate cancer and breast cancer metastasis profiles. Finally, we estimate the classification performance by comparing with the conventional gene-based approach. In Section 5, we conclude with summary of our work.

2 Data sets and method

2.1 Human interaction networks

We constructed a protein–protein and protein–DNA interaction network, which consists of 57235 interactions among 11203 proteins curated from public databases including HPRD, BIND and REACTOME (Bader et al., 2001; Peri et al., 2004; Vastrik et al., 2007), and several literatures (Ramani et al., 2005; Rual et al., 2005; Stelzl et al., 2005), and 1538 interactions among 333 transcription factors and 556 target genes extracted from Transfac database (Knuppel et al., 1994).

2.2 Gene expression profiles

We obtained five previously published mRNA expression datasets, which were divided into two populations of distinct phenotypes as per the original publications as shown in Table 1. The expression profiles were downloaded from NCBI GEO microarray repository or authors' websites.

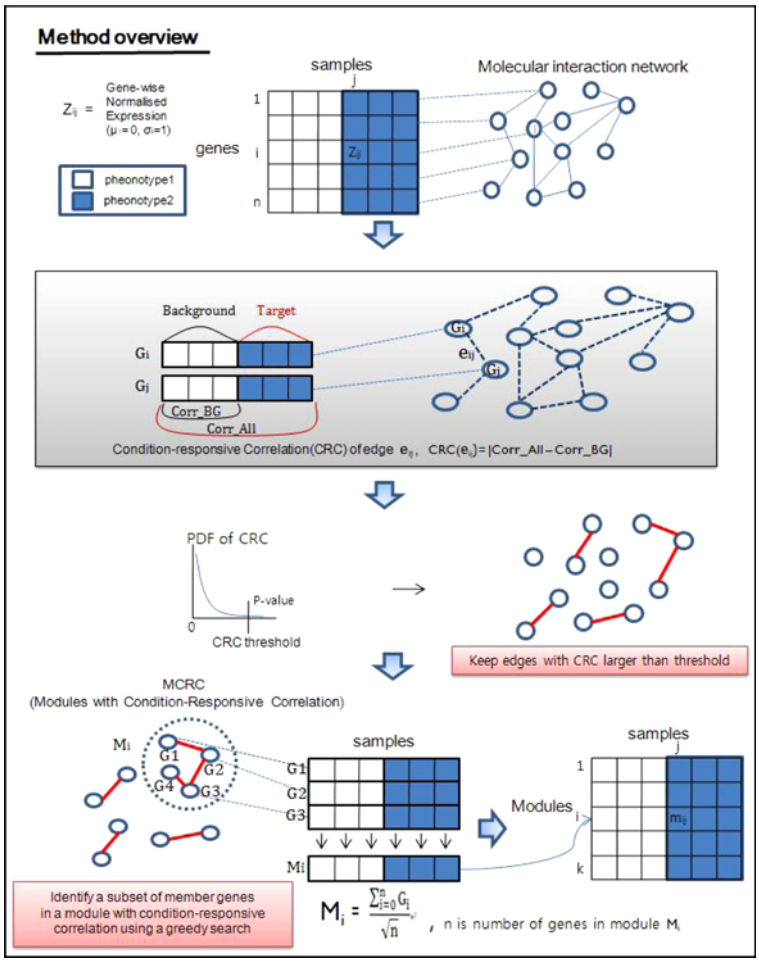
Table 1 The five data sets used in method evaluation

Study	*Phenotypes and samples*
Brain cancer (Nutt et al., 2003)	14 classic glioblastoma vs. 7 anaplastic oligodendrogliomas
Prostate cancer (Singh et al., 2002)	50 normal prostate samples vs. 52 prostate tumour samples
Leukaemia (Yeoh et al., 2002)	79 TEL-AML1 leukaemia samples vs. 64 leukaemia samples with HH (hyperdiploid hyperdip) > 50
Lung cancer prognosis (Bhattacharjee et al., 2001)	31 primary lung tumours with poor prognosis vs. 31 primary lung tumours with good prognosis
Breast cancer metastasis (Wang et al., 2005)	106 metastatic primary breast tumours vs. 180 non-metastatic primary breast tumours

2.3 Extracting molecular Modules with Condition Responsive Correlations (MCRCs)

To extract molecular Modules with Condition-Responsive Correlations (MCRCs) in a specific phenotype of our interest, we first overlaid the expression levels of each gene onto its corresponding protein in the network as shown in Figure 1. For each interaction, edge in the network, the Pearson correlation coefficient for gene expression levels of two interacting proteins was calculated using all samples (Corr_All), and also using samples of a background phenotype (Corr_BG). The degree of CRC for each edge was defined by the magnitude of difference between Corr_All and Corr_BG as proposed by Mani et al. (2008). MCRCs were defined as a set of remaining connected components after removing edges with CRC values lower than a specific threshold. In this work, we used the statistical significance of CRC score (p value) less than 0.001 or 0.0025 as a CRC cut-off, and removed edges with CRC p value larger than a p value threshold. The p value of a CRC score was calculated by indexing the score on the null distribution of CRC scores from all the edges in the network.

Figure 1 Schematic diagram for extracting CRC molecular modules with Condition-Responsive Correlations (CRCs), and inferring their activities based on a subset of discriminative member genes (see online version for colours)



2.4 Inferring module activities

Given the identified set of MCRCs from the previous module extraction step, for each MCRC, we applied a greedy search similar to the method proposed by Lee et al. (2008) to identify a subset of member genes in the module whose combined expression levels deliver maximal discriminative power across two phenotypes. Because the original method by Lee et al. (2008) was invented for a set of genes, pathways, we modified it to consider network connectivity.

For each MCRC, we chose a protein with the highest discriminative power in its gene expression levels as a starting node of greedy expansion. In the next step, the search considers the addition of a protein from the neighbours of the starting node. An addition that yields the increase in discriminative score of inferred module activity (M_i in Figure 1) is adopted. This step iterates until any neighbour protein of already selected proteins cannot increase the discriminative score of module activity.

Finally, we produced a matrix of module activity levels across patients, which was utilised to build a classifier. In this study, we used the *t*-score to measure the discriminative power between two different phenotypes.

2.5 Classification evaluation

Naïve Bayesian classifier (Bishop, 1995) was trained on both the activity matrix of MCRCs and the original gene expression matrix for comparison. The Naïve Bayesian classifier technique is based on the Bayesian theorem and is particularly suited when the dimensionality of the inputs is high. The assumption of the Naïve Bayesian classifier is that every feature should be independent. The advantage of using simple Naïve Bayesian classification technique is that we do not need to suffer from optimising classification parameters since our goal is comparing two different types of feature sets (module activities vs. gene expressions). To get the unbiased classification performance from the data with small number of samples, we applied the Leave One Out Cross Validation (LOOCV) scheme, and reported the averaged AUC on test samples as a final classification performance. We strictly utilised only training samples in the steps of MCRC extraction, and module activity inference to avoid the inflow of any information from a test sample including the class label.

Instead of using all the features of either module activities or individual gene expressions together in classification, we built classifiers using varying number of features by sequentially adding one by one in decreasing order of their discriminative power. Gene-based classifiers, for comparison, used the matched number of top discriminative genes to the number of unique genes included in modules to keep the same amount of information content to be used.

3 Results and discussion

3.1 MCRC from five data sets

Table 2 shows the number of extracted MCRCs and the number of unique genes included in those modules under the given CRC *p* value cut-off for five data sets. Through the integration of a molecular network and expression profiles, the greatly reduced number of features was extracted from the original gene expression matrix. For example, 28 activity features of modules composed of 66 different genes were extracted from 8478 individual gene expression features in brain cancer profiles.

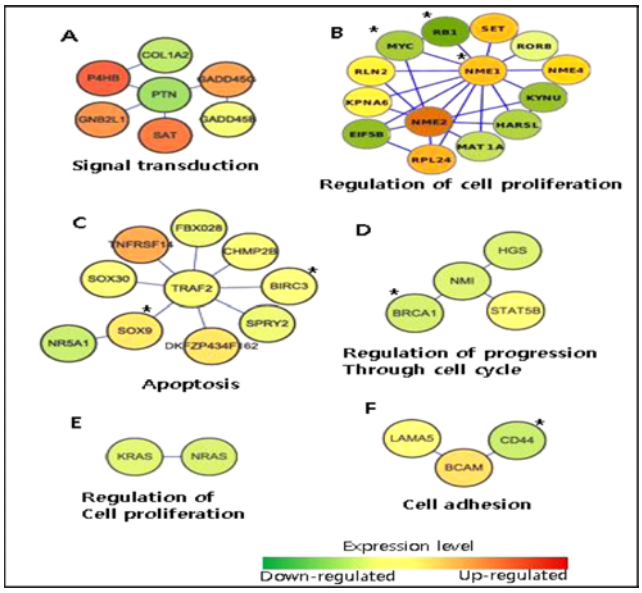
Table 2 Condition-responsive correlated modules from five data sets

Dataset	*#Modules*	*# Genes*	*CRC cut-off* (*p value*)
Brain cancer	28	66	0.001
Prostate cancer	19/30	57/116	0.001/0.0025
Leukaemia	17	42	0.001
Lung cancer	16	53	0.001
Breast cancer	34	84	0.001

3.2 MCRC associated with prostate cancer

Many of 30 identified modules with CRC $p < 0.0025$ from the prostate cancer dataset were enriched for proteins functioning in a common pathway (using a hypergeometric test on Biological Process annotation from the GO database) while 19 modules of CRC $p < 0.001$ were not. It might be because important oncogenic processes with altered interactions in prostate cancer might not be captured under the strict CRC cut-off $p < 0.001$. Figure 2 shows some of the modules enriched with important functions such as cell proliferation, apoptosis and cell adhesion associated with tumorigenesis. Many of known prostate cancer causing genes such as MYC, RB1, NME1 in cell proliferation, BRCA1 in cell cycle control, BIRC3 and SOX9 in apoptosis and CD44 in cell adhesion module were identified (Li et al., 2003). Mutations in NME1, RB1, BRCA1, and amplification of MYC gene are known to cause prostate cancer (Li et al., 2003), and SOX9 is known to function both in the development and the maintenance of normal prostate, and indicated for the contribution to tumour growth and invasion (Wang et al., 2008).

Figure 2 Modules with Condition-Responsive Correlation from prostate cancer dataset. Known prostate cancer-related genes are marked by an asterisk (see online version for colours)

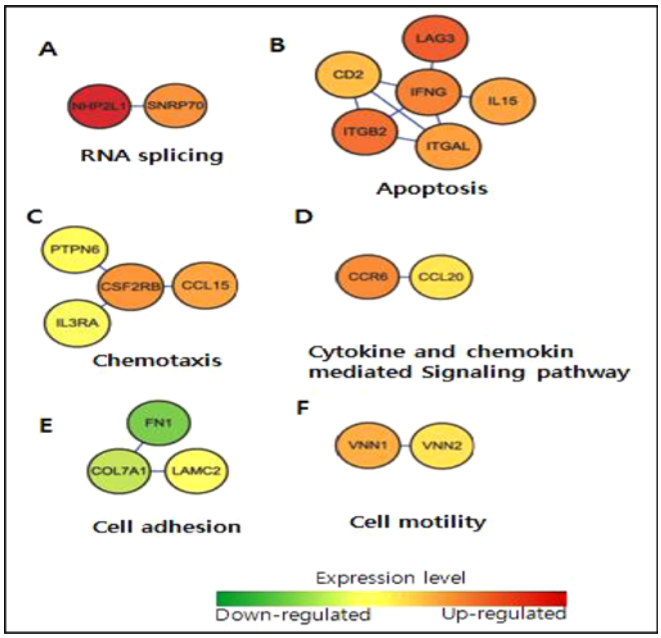


3.3 MCRC associated with breast cancer metastasis

A total of 34 modules were identified from the gene expression profiles of metastasis vs. non-metastatic lymphnode-negative primary breast tumour samples. Proteins functioning in the biological processes associated with tumour invasion and metastasis such as apoptosis, chemotaxis, cell adhesion and cell motility were enriched in the modules (Figure 3) (Yu et al., 2007). Especially aberrations in cell adhesion molecules such as FN1 (fibronectin) and LAMC2 were associated with metastasis and invasion (Sisci et al., 2004; Yuen et al., 2005). While the conventional gene expression analysis,

which focuses on discriminative genes, cannot capture LAMC2 due to its minor change in expression level, our approach identified it because the correlation between LAMC2 and COL7A1 was different in metastatic and non-metastatic tumours.

Figure 3 Modules with Condition-Responsive Correlation associated with breast cancer metastasis (see online version for colours)

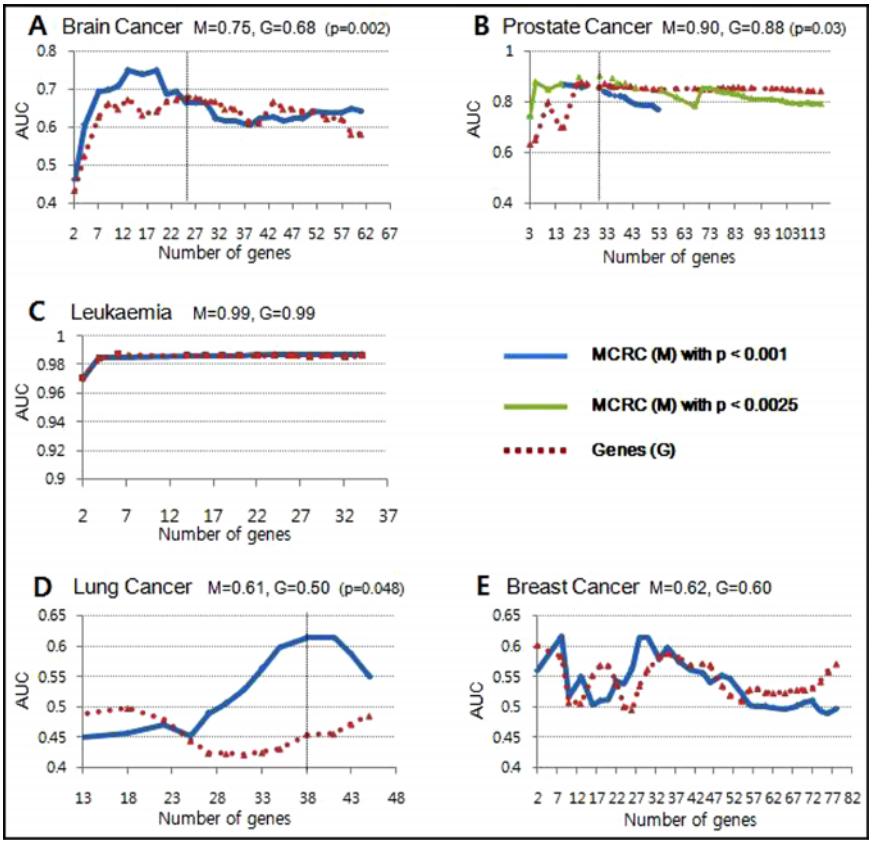


3.4 Classification performance

We evaluated the classification performance of our MCRC method, and the conventional gene-based method in five data sets. As shown in Figure 4, classification based on activities of MCRC modules outperformed the conventional individual gene-based method significantly in brain cancer, prostate cancer and lung cancer ($p = 0.002$, 0.03 and 0.048, respectively), and showed comparable performance in leukaemia, and breast cancer metastasis. The p value was calculated using Wilcoxon signed rank test of two sequences of AUCs from the left-hand start to the dotted vertical line from both methods. The vertical line represents the maximum value from the number of genes used to generate a best AUC using each method. For example, in the brain cancer, MCRC showed the best AUC using modules consisting of 13 genes, and the gene-based methods using 25 genes. In this case, we compared the AUCs using 1–25 genes to get p value.

While MCRC with $p < 0.0025$, in prostate cancer classification, showed better performance than genes, MCRC with $p < 0.001$ showed only comparable result ($M = 0.87$, $G = 0.88$, $p = 0.63$). Also as shown in the previous GO enrichment analysis, MCRCs with $p < 0.0025$ were significantly enriched with biological processes associated with cancer progression while MCRC with $p < 0.001$ were not. This implies that a systematic way to decide CRC cut-off, which enables the identification of more biologically meaningful MCRCs, needs to be developed as a further work.

Figure 4 Classification performance of MCRC and the conventional method. The maximum AUC using each method, and the p value through Wilcoxon signed rank test to measure the significance of difference in AUCs up to the dotted vertical line from both methods were denoted next to the name of data set. The vertical line denotes the maximum value from the number of genes used to generate a best AUC using each method (see online version for colours)



5 Conclusion

This paper has proposed a novel network-based classification approach, which focuses on both condition-responsive proteins (nodes), and interactions (edges) in the network to extract condition-responsive functional modules. The interactions between each pair of two proteins are screened to have CRCs in their expression levels across two phenotypes, and then the activity level of a module is inferred from a subset of member genes in the module whose combined expression levels deliver the maximal discriminative power. Finally, we estimate the classification performance by comparing with gene-based conventional approach using Naive Bayesian classifier.

We have demonstrated that exploiting both condition-responsive interactions and discriminative genes in the molecular network can identify novel functional modules associated with key pathogenic processes, and improve classification performance. Condition-responsive interactions with gain or loss of correlation in cancer samples can

especially provide clues for identifying oncogenic aberrations. In addition, the improved performance in disease classification support that compressing the expression levels of multiple genes, which have closely correlated expression due to their roles in the same biological processes into an activity level effectively reduces redundant signals into a better predictor.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea Government (MEST) (2009-0086964), and the World Class University (WCU) program through the Korea Science and Engineering Foundation funded by the Ministry of Education, Science and Technology (R32-2009-000-10218-0), and the Development of Large Volume Multi-scale Systems Dynamics Interpretation Technology for Virtual Cell Application Platforms research funded by the Korea Institute of Science and Technology Information (KISTI).

References

Alizadeh, A.A., Eisen, M.B., Davis, R.E., Ma, C., Lossos, I.S., Rosenwald, A., Boldrick, J.C., Sabet, H., Tran, T., Yu, X., Powell, J.I., Yang, L., Marti, G.E., Moore, T., Hudson Jr., J., Lu, L., Lewis, D.B., Tibshirani, R., Sherlock, G., Chan, W.C., Greiner, T.C., Weisenburger, D.D., Armitage, J.O., Warnke, R., Levy, R., Wilson, W., Grever, M.R., Byrd, J.C., Botstein, D., Brown, P.O. and Staudt, L.M. (2000) 'Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling', *Nature*, Vol. 403, pp.503–511.

Antoniadis, A., Lambert-Lacroix, S. and Leblanc, F. (2003) 'Effective dimension reduction methods for tumor classification using gene expression data', *Bioinformatics*, Vol. 19, pp.563–570.

Bader, G.D., Donaldson, I., Wolting, C., Ouellette, B.F., Pawson, T. and Hogue, C.W. (2001) 'BIND – the biomolecular interaction network database', *Nucleic Acids Res.*, Vol. 29, pp.242–245.

Bhattacharjee, A., Richards, W.G., Staunton, J., Li, C., Monti, S., Vasa, P., Ladd, C., Beheshti, J., Bueno, R., Gillette, M., Loda, M., Weber, G., Mark, E.J., Lander, E.S., Wong, W., Johnson, B.E., Golub, T.R., Sugarbaker, D.J. and Meyerson, M. (2001) 'Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses', *Proc. Natl. Acad. Sci. USA*, Vol. 98, pp.13790–13795.

Bishop, C.M. (1995) *Neural Networks for Pattern Recognition*, Oxford University Press, pp.385–424.

Chuang, H.Y., Lee, E., Liu, Y.T., Lee, D. and Ideker, T. (2007) 'Network-based classification of breast cancer metastasis', *Mol. Syst. Biol.*, Vol. 3, p.140.

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. and Lander, E.S. (1999) 'Molecular classification of cancer: class discovery and class prediction by gene expression monitoring', *Science*, Vol. 286, pp.531–537.

Guo, Z., Li, Y., Gong, X., Yao, C., Ma, W., Wang, D., Li, Y., Zhu, J., Zhang, M., Yang, D. and Wang, J. (2007) 'Edge-based scoring and searching method for identifying condition-responsive protein-protein interaction sub-network', *Bioinformatics*, Vol. 23, pp.2121–2128.

Guo, Z., Zhang, T., Li, X., Wang, Q., Xu, J., Yu, H., Zhu, J., Wang, H., Wang, C., Topol, E.J., Wang, Q. and Rao, S. (2005) 'Towards precise classification of cancers based on robust gene functional expression profiles', *BMC Bioinformatics*, Vol. 6, p.58.

Knuppel, R., Dietze, P., Lehnberg, W., Frech, K. and Wingender, E. (1994) 'TRANSFAC retrieval program: a network model database of eukaryotic transcription regulating sequences and proteins', *J. Comput. Biol.*, Vol. 1**,** pp.191–198.

Lee, E., Chuang, H.Y., Kim, J.W., Ideker, T. and Lee, D. (2008) 'Inferring pathway activity toward precise disease classification', *PLoS Comput. Biol.*, Vol. 4, p.e1000217.

Li, L. C., Zhao, H., Shiina, H., Kane, C.J. and Dahiya, R. (2003) 'PGDB: a curated and integrated database of genes related to the prostate', *Nucleic Acids Res.*, Vol. 31, pp.291–293.

Mani, K.M., Lefebvre, C., Wang, K., Lim, W.K., Basso, K., Dalla-Favera, R. and Califano, A. (2008) 'A systems biology approach to prediction of oncogenes and molecular perturbation targets in B-cell lymphomas', *Mol. Syst. Biol.*, Vol. 4, p.169.

Nutt, C.L., Mani, D.R., Betensky, R.A., Tamayo, P., Cairncross, J.G., Ladd, C., Pohl, U., Hartmann, C., Mclaughlin, M.E., Batchelor, T.T., Black, P.M., Von Deimling, A., Pomeroy, S.L., Golub, T.R. and Louis, D.N. (2003) 'Gene expression-based classification of malignant gliomas correlates better with survival than histological classification', *Cancer Res.*, Vol. 63, pp.1602–1607.

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. and Pandey, A. (2004) 'Human protein reference database as a discovery resource for proteomics', *Nucleic Acids Res*., Vol. 32, pp.D497–D501.

Ramani, A.K., Buneu, R.C., Moey, R.J. and Maotte, E.M. (2005) 'Consolidating the set of known human protein-protein interactions in preparation for large-scale mapping of the human interactome', *Genome Biol.*, Vol. 6, p.R40.

Ramaswamy, S., Ross, K.N., Lander, E.S. and Golub, T.R. (2003) 'A molecular signature of metastasis in primary solid tumors', *Nat. Genet.*, Vol. 33, pp.49–54.

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. and Vidal, M. (2005) 'Towards a proteome-scale map of the human protein-protein interaction network', *Nature*, Vol. 437, pp.1173–1178.

Singh, D., Febbo, P.G., Ross, K., Jackson, D.G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A.A., D'amico, A.V., Richie, J.P., Lander, E.S., Loda, M., Kantoff, P.W., Golub, T.R. and Sellers, W.R. (2002) 'Gene expression correlates of clinical prostate cancer behavior', *Cancer Cell*, Vol. 1, pp.203–209.

Sisci, D., Aquila, S., Middea, E., Gentile, M., Maggiolini, M., Mastroianni, F., Montanaro, D. and Ando, S. (2004) 'Fibronectin and type IV collagen activate ERalpha AF-1 by c-Src pathway: effect on breast cancer cell motility', *Oncogene*, Vol. 23, pp.8920–8930.

Stelzl, U., Worm, U., Lalowski, M., Haenig, C., Brembeck, F.H., Goehler, H., Stroedicke, M., Zenkner, M., Schoenherr, A., Koeppen, S., Timm, J., Mintzlaff, S., Abraham, C., Bock, N., Kietzmann, S., Goedde, A., Toksoz, E., Droege, A., Krobitsch, S., Korn, B., Birchmeier, W., Lehrach, H. and Wanker, E.E. (2005) 'A human protein-protein interaction network: a resource for annotating the proteome', *Cell*, Vol. 122, pp.957–968.

Vastrik, I., D'eustachio, P., Schmidt, E., Gopinath, G., Croft, D., De Bono, B., Gillespie, M., Jassal, B., Lewis, S., Matthews, L., Wu, G., Birney, E. and Stein, L. (2007) 'Reactome: a knowledge base of biologic pathways and processes', *Genome Biol*, Vol. 8, p.R39.

Wang, H., Leav, I., Ibaragi, S., Wegner, M., Hu, G.F., Lu, M.L., Balk, S.P. and Yuan, X. (2008) 'SOX9 is expressed in human fetal prostate epithelium and enhances prostate cancer invasion', *Cancer Res*, Vol. 68, pp.1625–1630.

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. and Foekens, J.A. (2005) 'Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer', *Lancet*, Vol. 365, pp.671–679.

Wood, L.D., Parsons, D.W., Jones, S., Lin, J., Sjoblom, T., Leary, R.J., Shen, D., Boca, S.M., Barber, T., Ptak, J., Silliman, N., Szabo, S., Dezso, Z., Ustyanksky, V., Nikolskaya, T., Nikolsky, Y., Karchin, R., Wilson, P.A., Kaminker, J.S., Zhang, Z., Croshaw, R., Willis, J., Dawson, D., Shipitsin, M., Willson, J.K., Sukumar, S., Polyak, K., Park, B.H., Pethiyagoda, C.L., Pant, P.V., Ballinger, D.G., Sparks, A.B., Hartigan, J., Smith, D.R., Suh, E., Papadopoulos, N., Buckhaults, P., Markowitz, S.D., Parmigiani, G., Kinzler, K.W., Velculescu, V.E. and Vogelstein, B. (2007) 'The genomic landscapes of human breast and colorectal cancers', *Science*, Vol. 318, pp.1108–1113.

Yeoh, E.J., Ross, M.E., Shurtleff, S.A., Williams, W.K., Patel, D., Mahfouz, R., Behm, F.G., Raimondi, S.C., Relling, M.V., Patel, A., Cheng, C., Campana, D., Wilkins, D., Zhou, X., Li, J., Liu, H., Pui, C.H., Evans, W.E., Naeve, C., Wong, L. and Downing, J.R. (2002) 'Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling', *Cancer Cell*, Vol. 1, pp.133–143.

Yu, J.X., Sieuwerts, A.M., Zhang, Y., Martens, J.W., Smid, M., Klijn, J.G., Wang, Y. and Foekens, J.A. (2007) 'Pathway analysis of gene signatures predicting metastasis of node-negative primary breast cancer', *BMC Cancer*, Vol. 7, p.182.

Yuen, H.W., Ziober, A.F., Gopal, P., Nasrallah, I., Falls, E.M., Meneguzzi, G., Ang, H.Q. and Ziober, B.L. (2005) 'Suppression of laminin-5 expression leads to increased motility, tumorigenicity, and invasion', *Exp Cell Res.*, Vol. 309, pp.198–210.