*International Journal of Current Research and Modern Education (IJCRME)*
*ISSN (Online): 2455 – 5428 & Impact Factor: 3.165*
*Special Issue, NCFTCCPS - 2016*

# DIAGNOSIS AND PROGNOSIS OF CANCER

**K. K. Manju* & G. Srinitys****
* PG Scholar, Department of Information Technology Bannari Amman Institute of Technology, Erode, Tamilnadu
** Assistant Professor, Department of Information Technology Bannari Amman Institute of Technology, Erode, Tamilnadu

**Abstract:**

*Cancer being a deadly disease is still the inscrutable problem. Diagnosis and Prognosis are the two main challenges which are to be addressed in treating cancer. From various solid cancers and metabolic systems Genomic, metabolic and clinical data can be used to identify patient subgroups for tailored therapy and monitoring. The information content is higher in integrated analysis than in any levels separately, and large number of statistical methods for the integration of 'omics' data have been emerged.*

**Key Words:** Diagnosis, Prognosis & Omics

## 1. Introduction:

Cancer is among the leading causes of death worldwide, and treatments for cancer range from clinical procedures such as surgery to complex combinations of drugs, surgery and chemo radiation. Most (if not all) cancers involve genetic aberrations in the germ line and/or at the somatic level. By producing a complete catalogue of inherited and acquired mutations, with functional consequences of each mutation with respect to tumour type, it is hoped that one can, for example, assess the metastatic potential of a tumour and suggest the most promising treatment. Although data are rapidly accumulating from various cancer-profiling projects, interpreting these data is not easy. The development and progression of a tumour is a dynamic biological and evolutionary process. It involves composite organ systems, with genomes shaped by gene aberrations, epigenetic changes, the cellular biological context, characteristics that are specific to the individual patient, and environmental influences 9, 10. Sophisticated statistical and mathematical techniques have been developed for the analysis, interpretation and validation of biological data, and novel computational techniques and tools are continuously emerging. In principle, mathematical modelling of pattern formation using methods from interacting particle systems, system dynamics and hierarchical models can be used to study tumour formation and growth.

## 2. Approaches:

### A. Sequential Analysis:

This approach allows the confirmation or refinement of findings based on one data type, with additional analyses of further omics data obtained from the same set of samples. In this case, at least two types of omics data are analysed - for example, copy-number alterations (CNAs) and gene expression level data. To integrate two different levels of omics data from the same set of breast cancer samples, Chin *et al.*[1] identified genes whose expression levels were significantly deregulated by CNAs, as well as genes that are associated with metastasis and reduced survival. Lando *et al.* used CNAs integrated with gene expression and gene ontology to identify genes representing five biological processes associated with poor outcome in cervical cancer after chemotherapy and radiotherapy. Moreover, Beroukhim *et al.*[2] combined data from 3,131 cancer specimens, which represented 26 different histological types of cancer, and identified 158 regions with focal CNAs that were significantly altered across all samples. Interestingly, 122 of these CNAs did not harbour a known cancer gene. Each of these papers used the approach in which an analysis of each data set is made

independently of the others and produces a list of interesting entities, which are then linked to each other.

**B. Latent Variable Analysis:**

Unsupervised clustering of omics data can be used to partition individuals or samples into subgroups of potential clinical relevance. In the iCluster package[3], for example, the clustering of individual samples is carried out by applying metrics (or noise structures) that are specific to each data type but using common latent labels among all data types, employing an expectation-maximization algorithm (EM algorithm). This method can be extended to supervise clustering when the data are continuous, such as for expression data, and it can accommodate any number of data types. The number of clusters is difficult to determine and is estimated by cross-validation methods.A further development has been suggested by Yuan *et al.*[4] Their Patient-Specific Data Fusion (PSDF) algorithm exploits the fact that the data to be integrated in individual samples might seem to be contradictory within the data pool; for example, a high copy number of a gene could be associated with a high expression of the same gene in *cis* in most, but not all, samples. Such a contradiction can be seen as a measurement error or biological variation due to the cell composition of a biopsy or patient characteristics. PSDF estimates a latent variable per patient, which helps to exclude (or minimize) contradictory samples. This idea could potentially be used beyond clustering, for other tasks of integrative analysis.

**C. Penalized Likelihood Analysis:**

The aim of integrative regression is to determine the genes (or entities) — using at least two different omics data types — that allow the best prediction of the outcome. Since the number of covariates mostly supersedes the number of samples, some form of variable selection or penalized regression is necessary[5]. When sparsity can be assumed (that is, when only a few entities are expected to actually be relevant for the outcome), Lasso[6,7] is a very useful penalization method, as it carries out variable selection. Cross-validation is used to determine an optimal level of penalization, which influences the sparsity of the solution.

**D. Gene Set Analysis:**

One of the earliest reported examples of an integrative approach for gene expression data was the use of GeneXPress to identify modules of genes that affect the activity of a tumour[8]. Segal *et al.*[8] analyzed data from 22 cancer types and found that distinct shared modules of gene activity, which probably represented common tumour progression mechanisms, characterized distinct tumour types. A different strategy involves initially defining a collection of gene sets (for example, gene ontology terms or pathways). This step typically involves the use of publically available databases that collect extensive annotation and knowledge (for example, Kyoto Encyclopedia of Genes and Genomes (KEGG), Reactome and WikiPathways[9]; see Further information). A score is calculated for each gene[10] (for example, a *P* value that reflects the degree of differential expression), and all gene sets that are 'enriched' or over-represented with high or low scores are identified. The scores can also be binary (0 or 1), thereby indicating, for example, membership in a group of differentially expressed genes. By combining gene ontology, gene expression and clinical data, Subramanian *et al.*[11] used gene set enrichment analysis (GSEA) to identify genes consistently associated with poor outcome in two independent cohorts of patients with lung cancer. Information based on known protein–protein interactions has been used to identify gene modules expressed in non-malignant bystander cells[12], associated with metastatic disease61 or associated with aggressive disease in lymphoma. Several alternative ways of scoring the abnormal

presence of specific pathways have also emerged, including Gene Microarray Pathway Profiler (also known as MicroArray Pathway Profile Finder (MAPPFinder))[13].

**E. Pairwise Correlation Analysis:**

In this type of analysis, for each pair of co-measured omics data, a correlation matrix is estimated[14]; with *P* values that are corrected for multiple testing and that therefore reflect the strength of association. This approach includes associations in *trans*. The structure in the matrix can be used to identify master regulators[15]. Correlation analysis does not directly facilitate the study of how entities (such as expression levels and CNAs) regulate outcomes of interest, but highly correlated entries can be used in further studies, such as in canonical correlation analysis[16]. There are multiple ways to extend the correlation analysis to more than two data types.

**F. Network Analysis:**

Networks are a representation of how genes or other entities collaborate in certain biological systems. A graph 'sums up' these effects over time, and two genes will be linked by an edge if they seem to interact in a specific process. Graphical algorithms that capture the interaction between differentially expressed genes by correlation include jActiveModules[17] and Graph-based iterative Group Analysis[18] (GiGA). JActive Modules integrates knowledge from protein–protein and protein–DNA interaction databases into mRNA expression data by assigning a Z-score for differentially expressed genes, and it searches for connected sub-networks by simulated annealing and greedy search algorithms[19]. Both simulated annealing and greedy search identify differentially expressed sub-networks; in the first case, in an optimal but computationally very intensive way; in the second case, more rapidly but less accurately. GiGA also ranks genes on the basis of differential expression levels and searches for sub-networks.

**G. Bayesian Analysis:**

Bayesian methods naturally facilitate the integration of biological knowledge through the design of appropriate prior distributions. In a Bayesian multiple testing setup, one can use a second type of omic data (for example, CNAs) to modulate the *a priori* probability that each test for a first data set (for example, expression levels) is likely to be rejected88. Bayesian networks are not new, and they were used in the early 2000s to incorporate various data[20]. As is true for every statistical method, Bayesian analysis is based on assumptions (both probabilities and prior assumptions) and models based on these have to be realistic and well designed so that they can be trusted.

**3. Integrative Analysis:**

Over the past decade, the accumulation of high-throughput molecular data from various cancer types has revealed an enormous range of alterations. Although subgroups of tumours with similarities in biological properties or clinical behaviour can be defined, the initial studies mainly analyzed one type of molecular data at a time. The access to large data sets that have been made available by the ICGC and TCGA has made it possible to compare the performance of some of the tools described above, on the same data set, as well as to compare the identified deregulated pathways between different cancer types. A pilot project from TCGA integrated DNA copy num-ber, gene expression and DNA methylation, as well as nucleotide sequence aberrations from glioblastoma samples[21]. Enrichment analysis revealed new roles for known cancer genes, as well as network activity. Later, the same data set was interrogated by Anduril[22] and by PARADIGM[23]. Both approaches suggested that amplification of the epidermal growth factor receptor (*EGFR*) was important in glioblastoma. Anduril, which can make use of DNA methylation data, also indicated DNA hypomethylation as a significant change that was evident in glioblastoma.

Data from the first pan-cancer analyses aim to iden-tify drivers of tumorigenesis that are common to multiple tumour types. For example, the aim of TCGA is to generate genomic data at all molecular layers in 10,000 tumours from 20 tumour types and to make these data available for the community. A recent endeavour to integrate somatic mutations, CNAs and DNA methylation was carried out in 3,299 tumours of 12 different cancer types. After integration with mRNA expression, a total of 479 candidate functional alterations were predicted, including 116 copy-number gains, 151 copy-number losses, 199 recurrently mutated genes and 13 epigenetically silenced genes. A hierarchical stratification was built using principles from network modularity[24]. Interestingly, on the basis of these analyses, tumours seemed to be driven either by somatic mutations or by CNAs — a phenomenon that the authors named 'the cancer genome hyperbola', owing to the inverse relationship between these events. However, some genes, such as *TP53* and phosphatidylinositol-4,5-bisphosphate 3-kinase, catalytic subunit-α (*PIK3CA*), can be subjected to both aberration modes, thereby leading to the deregulation of common pathways such as p53-mediated apoptosis, PI3K–AKT signaling and cell cycle control. Studying the relationship between the different genomic levels opens a debate over their explanatory weight and potential to discover drivers of cancer100. Ovaska *et al.*[25] found unexpectedly poor con-cordance between gene amplification, over expression of the genes from the amplicons and survival in patients with glioblastoma. Akavia *et al.*[26] showed that the expression of a driver drives a phenotype. The authors draw our attention to the fact that many of the current studies attempt to identify drivers only in genomic loci for which there is a good cor-relation between copy number and mRNA expression. So far, many current approaches have been based on linear correlation analysis. On the basis of knowledge of the enzyme kinetics and gene regulation, we expect nonlinear dependencies to occur in addition to linear effects. We recently proposed a statistical approach to investigate linear and nonlinear dependencies between CNA and mRNA expression.

**4. Working Principle:**

The discussion above has addressed the problem of inferring biological networks of relevance for translation into the clinic, based on a simple map of genes, transcripts and proteins. A paradigm shift is needed, from searching for single strong clinical markers to searching for a combined effect of multiple markers, as, in general, genes and proteins function by interacting with DNA, RNA and proteins, and these interactions might be specific for a given disease subclass. Many of the current targeted therapies focus on proteins that are involved in cell signaling pathways, which form a complex cellular communication system that governs basic cellular functions[28,29]. Established examples of targeted cancer treatment include EGFR-mutated non-small-cell lung cancer that can be treated with tyrosine kinase inhibitors (gefitinib or erlotinib), ERBB2 (also known as HER2)-directed therapy in breast cancer and melanomas with *BRAFV600E* mutations that can be targeted with vemurafenib. A major challenge in drug development is to precisely define the subset of cancer patients that are likely to respond. Within each pathway, a range of drugs may be available, and the optimal target (and, hence, the optimal drug) will be determined by the rate-limiting protein and the individual perturbations in the pathway. In colorectal cancer, EGFR-directed therapy with monoclonal antibodies has proven to be effective110. However, in the presence of a downstream activating KRAS mutation, the inhibition of EGFR is ineffective111. It seems likely that similar mechanisms are present in cases with resistance to other cancer treatments (both targeted and more traditional chemotherapeutic agents). Iadevaia *et al.* have proposed a computational procedure to generate experimentally

testable intervention strategies for the optimal use of available drugs in a cocktail. They used reverse phase protein array to evaluate the changes in the phosphorylation status of proteins after stimulation of the MDA-MB 231 breast cancer cell line with insulin-like growth factor, and they were able to conclude that the simultaneous inhibition of MAPK and PI3K–AKT pathways was sufficient to significantly halt cell proliferation. Future methods will require adding methylation and expression data to such integrative approaches. Introducing systematic clinical screenings for mutations that perturb these pathways is of great importance to identify the targets for targeted therapies and the patients that will respond to each treatment.

Outcome prediction that is based on genomic data is another central area of genomic research, and it has proven to be promising in breast cancer. One of the crucial issues in retrospective studies is that treatment selection is mostly based on the predicted risk of recurrence. Thus, treatment might be confounded by prognosis. This challenge the identification of pure prognostic markers, as the treatment interaction is not known. Even though the results from prospective validation trials, such as the Microarray In Node-negative and 1–3 positive lymph node Disease may Avoid Chemotherapy (MINDACT) trial and the Trial assigning individualized options for treatment (TailorX), are still pending, prediction tools based on gene expression are included in some clinical guidelines113,114. Optimal strategies for risk prediction are, however, not settled and remain controversial. Crowd sourcing strategies for problem solving, which were previously success-fully applied to biology in areas such as the prediction of protein folding and function[30], have been applied to this problem. In the DREAM BCC competition, participants competed to create an algorithm that could predict — more accurately than current benchmarks — the prognosis of patients with breast cancer from clinical information (age, tumour size and histological grade), genome-scale tumour mRNA expression data and DNA copy-number data from 1,980 patients. Integration of data was encouraged, and more than 1,400 models were submitted. The winners used a mathematical approach that was based on co-expression gene networks associated with tumour phenotype and functional characteristics to identify signature ‘attractor’ meta-genes, and this approach outperformed other models to predict outcome[31]. These examples support the notion that using the expertise of participants outside of traditional biological disciplines could be a powerful way to accelerate the translation of biomedical science into the clinic.

**5. Limitations:**

Integrative analyses are likely to become ever more important as computational strategies and tools are further improved and multilevel omics data sets become more abundant. The quest to understand the interplay within and between different molecular levels in cancer is no longer beyond our reach. It is important, however, to be aware of the limitations of the current methodologies. From a statistical perspective, the most fundamental challenge in integrative analyses is dimensionality: taking more levels into account in the analysis tends to increase the dimensionality of the problem. Adding more layers of data or increasing the resolution of measurements increases the dimension of unknown parameters, which are often difficult to estimate, thereby making the overall inference weaker. This might seem paradoxical, as the purpose of taking multiple levels into account is precisely the opposite — to use more observations to obtain a more accurate picture of the biological system under study. The way out of this apparent paradox is to realize that, first, one is able to infer more properties of a system with integrative approaches and, second, statistically efficient integrative methodologies can be constructed by actively using known properties of the

relationships between the molecular levels. The second point ensures that additional variables in the analysis are not, in effect, increasing the degrees of freedom of the underlying model but rather lending information to existing variables. In addition, at every step, there will be checkpoints of compatibility of the data, such as normalization to the same scale, sample selection from representative cohorts, adequate correction for technical batch effects and use of different platforms. Although numerous methods and tools are introduced to address these obstacles, it is still, so far, the case that large-scale true integration is possible within only a few projects worldwide, which have sufficient funding that allows all analyses to be carried out simultaneously and on the entire data set. Intuitively, it seems that as a 'gold standard', integration attempts are best carried out in supervised settings that are based on some priming biological knowledge or within the frame of defined biological hypotheses. Combining additional layers in unsupervised analyses might fail to contribute new information, as multiple use of the same data might artificially reduce variance or will increase the false discovery rate.

**6. Conclusions:**

A more fundamental understanding of the biological dynamics of cancer will enable us to better identify risk factors, refine cancer diagnosis, predict therapeutic effects and prognosis, and identify new targets for therapy. We are seeing a paradigm shift from large randomized clinical trials towards treatment modalities that are tailored for stratified patient groups, down to N-of-1 trials, in which data from a single patient represents an entire trial. This will fundamentally alter the way that we statistically model and evaluate treatment strategies, from identifying patient groups that have a response to treatment that is above random to identifying pathways and biological entities that are druggable and altered above random; and from evaluating the response in randomized arms, using the other arm as a control, to evaluating the response of experimental and control interventions in each individual, using the same individual as a control. The real challenge would be to develop statistical models to identify crucial, rate-limiting molecular targets for intervention, out of the wealth of information that next-generation sequencing uncovers, on the background of great redundancy of pathways and heterogeneity of tumours. As we are moving towards an era in which the amount of data produced every year is increasing exponentially, the biomedical community needs to embrace this complexity and find new methods of shared analysis.

**7. References:**

1. Chin, K. et al. Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. Cancer Cell 10, 529–541 (2006).J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68-73.
2. Lando, M. et al. Gene dosage, expression, and ontology analysis identifies driver genes in the carcinogenesis and chemoradioresistance of cervical cancer. PLoS Genet. 5, e1000719 (2009).
3. Shen, R. et al. Integrative subtype discovery in glioblastoma using iCluster. PLoS ONE 7, e35236 (2012).
4. Yuan, Y., Savage, R. S. & Markowetz, F. Patient-specific data fusion defines prognostic cancer subtypes. PLoS Comput. Biol. 7, e1002227 (2011).
5. Bøvelstad, H. M. et al. Predicting survival from microarray data—a comparative study. Bioinformatics 23, 2080–2087 (2007).
6. Tibshirani, R. Regression shrinkage and selection via the Lasso. J. R. Statist. Soc. Series B. 58, 267–288 (1996).
7. Nowak, G., Hastie, T., Pollack, J. R. & Tibshirani, R. A fused lasso latent feature

model for analyzing multi-sample aCGH data. Biostatistics 12, 776–791 (2011).
8. Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. J. R. Statist. Soc.: Series B (Statist. Methodol.) 67, 301–320 (2005).
9. Kelder, T. et al. WikiPathways: building research communities on biological pathways. Nucleic Acids Res. 40, D1301–D1307 (2012).
10. Rhee, S. Y., Wood, V., Dolinski, K. & Draghici, S. Use and misuse of the gene ontology annotations. Nature Rev. Genet. 9, 509–515 (2008).
11. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl Acad. Sci. USA 102, 15545–15550 (2005).
12. Dittrich, M. T., Klau, G. W., Rosenwald, A., Dandekar, T. Müller, T. Identifying functional modules in protein-protein interaction networks: an integrated exact approach. Bioinformatics 24, i223–i231 (2008).
13. Doniger, S. W. et al. MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data. Genome Biol. 4, R7 (2003).
14. Mayer, C.-D., Lorent, J. & Horgan, G. W. Exploratory analysis of multiple omics datasets using the adjusted RV coefficient. Stat. Appl. Genet. Mol. Biol. 10, Article 14 (2011).
15. Quigley, D. A. et al. Genetic architecture of mouse skin inflammation and tumour susceptibility. Nature 458, 505– 508 (2009).
16. Lê Cao, K.-A., González, I. & Déjean, S. integrOmics: an R package to unravel relationships between two omics datasets. Bioinformatics 25, 2855–2856 (2009).
17. Ideker, T., Ozier, O., Schwikowski, B. & Siegel, A. F. Discovering regulatory and signalling circuits in molecular interaction networks. Bioinformatics 18 (Suppl. 1), S233–240 (2002).
18. Breitling, R., Amtmann, A. & Herzyk, P. Graph-based iterative Group Analysis enhances microarray interpretation. BMC Bioinformatics 5, 100 (2004).
19. Dittrich, M. T., Klau, G. W., Rosenwald, A., Dandekar, T. Müller, T. Identifying functional modules in protein-protein interaction networks: an integrated exact approach. Bioinformatics 24, i223–i231 (2008).
20. Imoto, S. et al. Combining microarrays and biological knowledge for estimating gene networks via bayesian networks. J. Bioinform. Comput. Biol. 2, 77–98 (2004).
21. Cancer, Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature 455, 1061–1068 (2008).
22. Ovaska, K. et al. Large-scale data integration framework provides a coprehensive view on glioblastoma multiforme. Genome Med. 2, 65 (2010).
23. Vaske, C. J. et al. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. Bioinformatics 26, i237–i245 (2010).
24. Newman, M. E. J. Fast algorithm for detecting community structure in networks. Phys. Rev. E Stat. Nonlin Soft Matter Phys. 69, 066133 (2004).
25. Ovaska, K. et al. Large-scale data integration framework provides a comprehensive view on glioblastoma multiforme. Genome Med. 2, 65 (2010).
26. Ovaska, K. et al. Large-scale data integration framework provides a comprehensive view on glioblastoma multiforme. Genome Med. 2, 65 (2010).
27. Akavia, U. D. et al. An integrated approach to uncover drivers of cancer. Cell 143,

1005–1017 (2010).
28. Hoshino, D. et al. Network analysis of the focal adhesion to invadopodia transition identifies a PI3K-PKCα invasive signaling axis. Sci. Signal. 5, ra66 (2012).
29. Stronach, E. A. et al. DNA-PK mediates AKT activation and apoptosis inhibition in clinically acquired platinum resistance. Neoplasia 13, 1069–1080 (2011).
30. Cooper, S. et al. Predicting protein structures with a multiplayer online game. Nature 466, 756–760 (2010).
31. Cheng, W.-Y., Ou Yang, T.-H. & Anastassiou, D. Development of a prognostic model for breast cancer survival in an open challenge environment. Sci. Transl. Med. 5, 181ra50 –181ra50 (2013).