PROMISE OF BIOINFORMATICS IN PHARMACEUTICS AND DIAGNOSTICS

ABUL B.M.M.K. ISLAM

Department of Genetic Engineering and Biotechnology, University of Dhaka, Dhaka 1000, Bangladesh
Phone: +880-2-9661920 Ext. 7825. Fax: +880-2-8615583. Email: khademul@du.ac.bd

ABSTRACT

The combination of the high-throughput technologies with information technologies has produced an enormous amount of information related to biomedicine. However, extensive growth in biomedical data generation has not yet been translated proportionately for clinical returns. Bioinformatics holds immense promise in this area by developing new tools to efficiently capture, curate and analyze these huge data and thereby help to diagnose diseases, to identify drug targets and to development new medicines.

Integration of multi-dimentional genomic data, and GenomeWide Association Studies (GWAS) may contribute profoundly to explore the mechanism of complex diseases. However, it requires correct record of phenotypic information. Bioinformaticians can play role to develop softwares for collecting, integrating and extracting clinical information; database development and data management. They further can develop a database of novel mutations/SNPs and their associations to drug responses. Such high-throughput studies with Bangladeshi patients, however, are still very limited and should be explored.

While conventional diagnosis may sometimes be erroneous, bioinformatic methods/tools may help us to find out disease-specific gene signature or biomarkers for accurate and specific molecular diagnosis. This method may also classify diseases, find new disease genes, delineate the pathogenic pathways, predict patients' survival time, predict functional consequences of mutations, identify the mode of action of candidate drugs and improve therapy by detecting and clustering important disease subtypes.

Address of Corresp0ndence:

Department of Genetic Engineering and Biotechnology,
University of Dhaka,Dhaka-1000,Bangladesh,

Phone:- 880-2-9661920 Ext. 7825.Fax: +880-2-8615583.
Email: khademul@du.ac.bd

Recently, the Open Source Drug Discovery (OSDD) initiative for system level understanding has drawn much attention. Computer aided drug design can dramatically reduce time and cost of effective biopharmaceuticals. Fortunately we have enormous varieties of plants having medicinal value, whose active compounds can be tested computationally for rational drug design. Recently an epigenetic enzyme has drawn much attention as drug target due to its reversible nature. Wealth of data generated world-wide can be analyzed to find epigenetic drug target, predict clinical outcome and possible side-effects. Correlation and network based analysis becoming more promising for designing combinational therapy. As the new era of personalized medicine is approaching, development of new bioinformatic systems and databases are needed for individualized therapies. Therefore, translational bioinformatic applications in genome medicine is expected to generate a great hope for future medicine.

INTRODUCTION

The code of life is embedded in the DNA sequences of any living organism, be it a life-threatening smallest virus/bacteria, a waste-utilizing friendly bacteria, our favorite food crops and other plants, animals or we humans ourselves. Therefore, to diagnose and combat diseases caused by microbes or by other genetic alterations, to use microorganisms/plants/animals for producing various products to use as medicine or other useful products we need to be able to analyze DNA sequences and understand function of its protein product from system level.[1-2] With the advent of sequencing technology, DNA/RNA/protein sequences of many organisms started becoming available from the beginning of the 21st century.[3] Currently sequences of hundreds of organisms are available with more than 126 billion bases of data.[4] Therefore the biggest challenge facing scientists today is to analyze, interpret and make sense of the wealth of data that has been produced by genome sequencing projects. Sequence generation, and its subsequent storage, interpretation and analysis are dependent on the computer-which has given rise to the Science of Bioinformatics.[5]

Bioinformatics can be defined as the utilization of computational tools and techniques using mathematical, statistical, and biophysical approaches to address complex biological/biomedical questions. Bioinformatics is an interdisciplinary science that combines knowledge of biology/medicine and computer science together with other fields such as mathematics, statistics, physics and chemistry to study complex biological processes and generate new knowledge of biology and medicine.[5] Over the years bioinformatics have proven competence in generating hypotheses and simulating experiments in easy and novel approaches that would otherwise demand the application of long, tedious and expensive wet laboratory procedures. Examples of these successes include biomarker discoveries for different cancers,[6-7] identification of molecular basis of diseases, design of more effective drugs, etc. Bioinformatics have shaped the modern structure of research by bringing together the researchers from different disciplines, like computer scientists, molecular and developmental biologists, biochemists, doctors, pharmacologists, and even mathematicians. It has created a common ground where knowledge from different disciplines can combine to bring more benefits to people.[5]

Understanding of and building expertise in bioinformatics hold the key to solve many of the problems that Bangladesh faces today. We need to have a better understanding of the available biological information to ensure effective and benefiting health-care system. For example, bioinformatic tools can be used to search new targets to design new drugs as well as to discover new biomarkers disease diagnosis, classification and predict prognosis.

Drug discovery and development is highly complex and, lengthy and process. For a new ligand to reach in the market as a potent drug must develop by the process which is commonly known as developmental chain or "pipeline: and consists of a number of distinct stages.(8) Generally it can be grouped under two stages, early pharmaceutical research and late pharmaceutical R&D. Early pharmaceutical research involves two steps process in which identification and modeling of the biological target within the body (the protein) is the first step, followed by the second step of identification of lead compound (ligand) that shows drug-like properties with respect to this protein. Later, the drug goes through many phases of clinical development in humans.

However, failure of most candidate molecule can occur as a result of combination of reasons such as poor pharmacodynamics, poor pharmacokinetics, side effects, lack of efficacy and commercial reasons. Also the cost and time required for the development of new drug from conventional method is very high. From conventional methods it would take approx $1.8 billion and 15 years from the initial stage to the successful marketing of a new drug.(9) Most drugs are discovered by either modifying the structure of known drugs, by screening compound libraries or by developing proteins as therapeutic agents.(9) With the advent of genomics, proteomics, bioinformatics and technologies like, NMR, crystallography, the structures of more and more protein targets are becoming available. So, there is an urge need for novel computational tools for drug designing that can identify and analyze active sites and suggest potential drug molecule that can bind to these sites and not only shorten the R&D time cycle but also reduce the ever increasing cost involved in the drug discovery process. *In silico* drug design or CADD (Computer Aided Drug Design) method fills this research requirement. Moreover, the overall cost of drug development could be reducing by as much as 50% through extensive use of *in silico* technologies in drug discovery procedure.(10) Also, studies from quantum mechanics, molecular dynamics, molecular dockings, QSAR to ADMET prediction including dissolution studies are performed *in silico*.

Francis Crick in 1958 first enunciated the phrase 'Central dogma of molecular biology' which states that RNA is transcribed from DNA and protein is translated from RNA; since then this simple relationship has gotten more complex and complicated. The more genes, RNAs and proteins are discovered, the more complex interaction networks are evolving.(11) Owing to this complexity, the study of individual genes and proteins, although necessary, cannot explain the systematic operation of biological systems, specially mechanisms of complex diseases. Therefore, genome-wide studies are becoming an integral part of any biological research.(12) With the advancement of technology, genome-wide studies of genetic and epigenetic landscapes/molecular networks are gradually emerging as a new way to study biological function as well as disease mechanisms.(13-14) There are several genome-wide high throughput experiments, such as sequencing, ChIP-seq, RNA-seq, Methylseq, ChIP-on-chip, microarray etc., producing enormous quantities of data that need careful bioinformatic analysis to extract the biological information to be utilized as biomarkers for disease diagnosis or as a target for new drug development.(15)

Epigenetic research uses a wide range of molecular biology techniques including genome-wide studies which require detailed bioinformatics analysis in order to extract biologically meaningful information. Therefore, the field of computational epigenetics is becoming more and more popular. The important role of epigenetic defects in cancer opens up new opportunities for improved diagnosis and therapy.(16) These active areas of research reveal several opportunities for bioinformatic analysis. Firstly, given a list of genomic locations exhibiting epigenetic differences between tumor cells and normal cells, can we detect co-regulated patterns or find evidence of a functional relationship of these regions to cancer? Secondly, is it possible to do a functional classification with the aforementioned sets of genes and by relating this to other experimental data, such as expression data, can we identify active gene modules that would be affected by drugs targeted to the epigenetic effectors? Thirdly, can we integrate several sets of epigenetic modification data to better explain the mechanisms of cell growth regulation in cancer? Fourth, can we use bioinformatic methods in order to find suitable drug targets and improve diagnosis and therapy by detecting and clustering important disease subtypes?

In our studies we used both bioinformatic and experimental biological techniques; and high throughput data both available publicly and generated by our laboratory to address these questions and show that bioinformatic analyses of microarray data identified deregulated histone modifying enzymes, specially KDM5A in several cancers, including breast, that can be used as drug target, diagnosis. *In silico* drug design can be a promising strategy to reduce time. Also, signature gene can predicted from genome-wide analyses that are able to predict disease prognosis and also can be used as disease classification.

METHODS

INTEGRATED EXPRESSION DATA EXTRACTION

We extracted “experimental level” data from “IntOGen”(17) database (v1) for 17 different cancer types and visualized the expression of epigenetic enzymes histome methyl transferases (HMT) and histone demethylases (HDM) in Gitools.(18)

EXPRESSION MICROARRAY DATA ANALYSIS

Raw Agilent platform microarray data on breast cancer downloaded from GEO.(19) For dataset GSE21974 all samples normalize together using Bioconductor(20) package Limma(21) and used method "vsn" for normalization of the single color microarray data. We also extracted GSE20194 dataset from GEO. Since this study does not have control samples before treatment, we will use cancer samples from another study GSE2034. Normalize all CEL files together using Bioconductor package "Affy"(22) and “Limma”.(21) Tumor expression data from neoadjuvant trial of cisplatin monotherapy in triple negative breast cancer patients were downloaded from GEO dataset GSE18864. Breast cancer data with survival information were downloaded from GSE3494 and GSE4922. These arrays were also normalized using Bioconductor package "Affy". Probe IDs were converted to Ensembl gene using Biomart portal. Where more probes were corresponded to same Ensembl genes, values were averaged.

SURVIVAL ANALYSIS

Public microarray data on colon cancer (GEO accession GSE3494 and GSE4922) were normalized as described above. We then used Gitools for Sample Level Enrichment Analysis (SLEA)(23) with modular Z-score enrichment statistics.(18) Positive and negative z-scores indicated significantly higher or lower expression levels of genes in the module (CAF signature genes)(24). We grouped the colon cancer samples according to their modular expression; and Kaplan-Meier survival curves and the Cox Hazard Ratio (HR) were then calculated with the R statistical program package, “survival” and “survplot”.

RESULTS

IDENTIFICATION OF BIOMARKERS FOR CANCER DIAGNOSIS

With the advancement of high-throughput technologies growing interest in cancer biomarkers have been observed. Bioinformatic analyses of genome-wide alteration data and integration studies such as expression, mutation and epigenetic signals have shown to identify novel markers of several disease types including cancer.(6,25) Such molecular biomarkers has the potentially to accurately predict cancer subtype and subsequently, the aid in therapy. Translational bioinformatics has helped in improving biomarker identification, specially application of prior knowledge to design algorithm that can rank the relevant genes.(26)

In this study we explored IntOGen database,(17) which system for integration and data mining of multidimensional oncogenomic data, for the altered expression of HMTs and HDMs in several cancer types. Several enzymes were found to be upregulated in different cancer types (Figure 1) including KDM5A in hematologic malignancies, ovary, stomach and breast cancer. KDM5A/RBP2 is one such histone-modifying enzyme whose activity can cause gene suppression by removing methyl groups from histone H3 on Lysine 4 (H3K4) residues. KDM5A is located in the retinoblastoma (RB) pathway – a critical pathway for cell cycle regulation.

FIG. 1

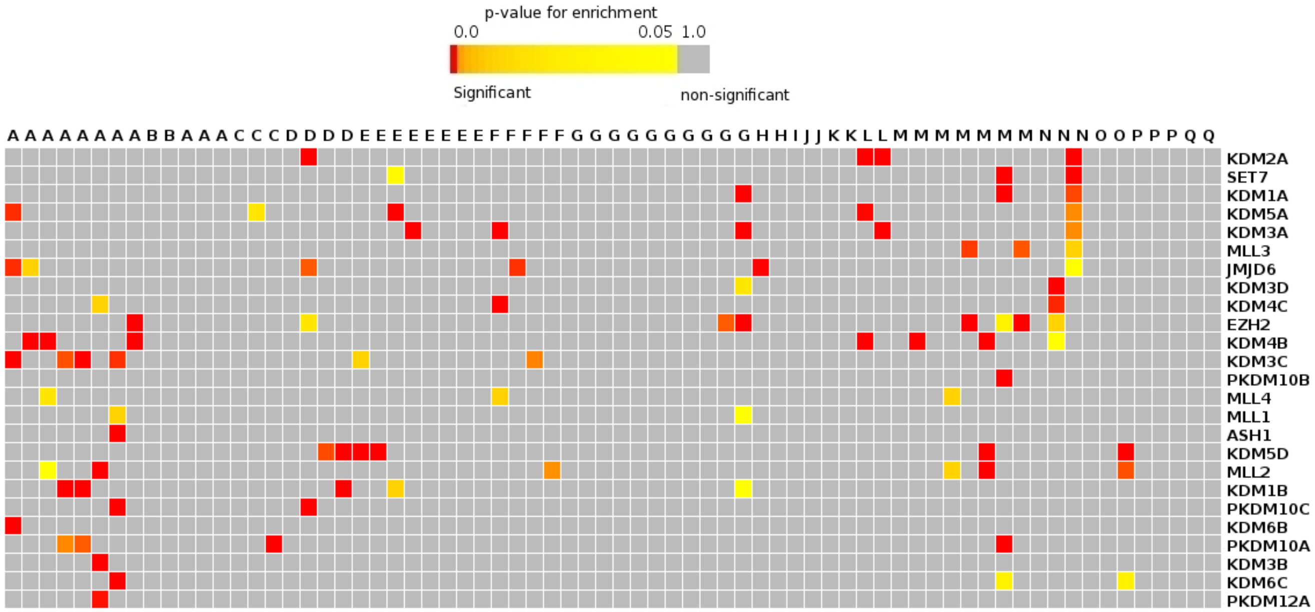


Figure 1: Upregulated expression of HMTs and HMDs in various cancer types. P-value of expression plotted in heatmap. Columns are different cancer types. Rows are genes. The more red color, the more significance of upregulated expression compare to normal counterpart. A = Breast, B= Brain, C= Cervix, D = Colon, E = Hematopoietic and reticuloendothelial, F = Liver, G = Lung, H = Lymph Node, I = Mediastinum, J = Mouth, K = Ovary, L = Pancrease, M = Prostate, N = Stomach

The retinoblastoma tumor susceptibility gene product (pRB), was discovered 25 years ago. It is now known that pRB directly interacts with KDM5A, promoting differentiation by inhibiting the activity of KDM5A. Deregulation of this interaction leads to cancer by hampering normal differentiation and accelerating proliferation through activation of other transcription factors.(27-28) We also showed that the pathway enrichment analysis with the target genes of KDM5A in mouse ES cells suggested involvement of KDM5A targets in several disease pathways including Huntington's disease, Perkinson's disease and Alzheimer's disease.(29)

NEW DRUG IS REQUIRED TO REDUCE KDM5A EXPRESSION IN CANCER TREATMENT

Here we first check the effect of some currently available drugs use to treat breast cancer patients on the reduction of KDM5A expression. In breast cancer dataset GSE21974 a cohort study of 32 women with primary invasive breast cancer is used to treat with epirubicine 90mg/m2 and cyclophosphamide 600mg/m2 every 3 weeks, followed by 4 cycles of docetaxel 100mg/m2. Expression of KDM5A and other HDMs and HMTs remain mostly unchanged before and after treatment (Figure 2a).

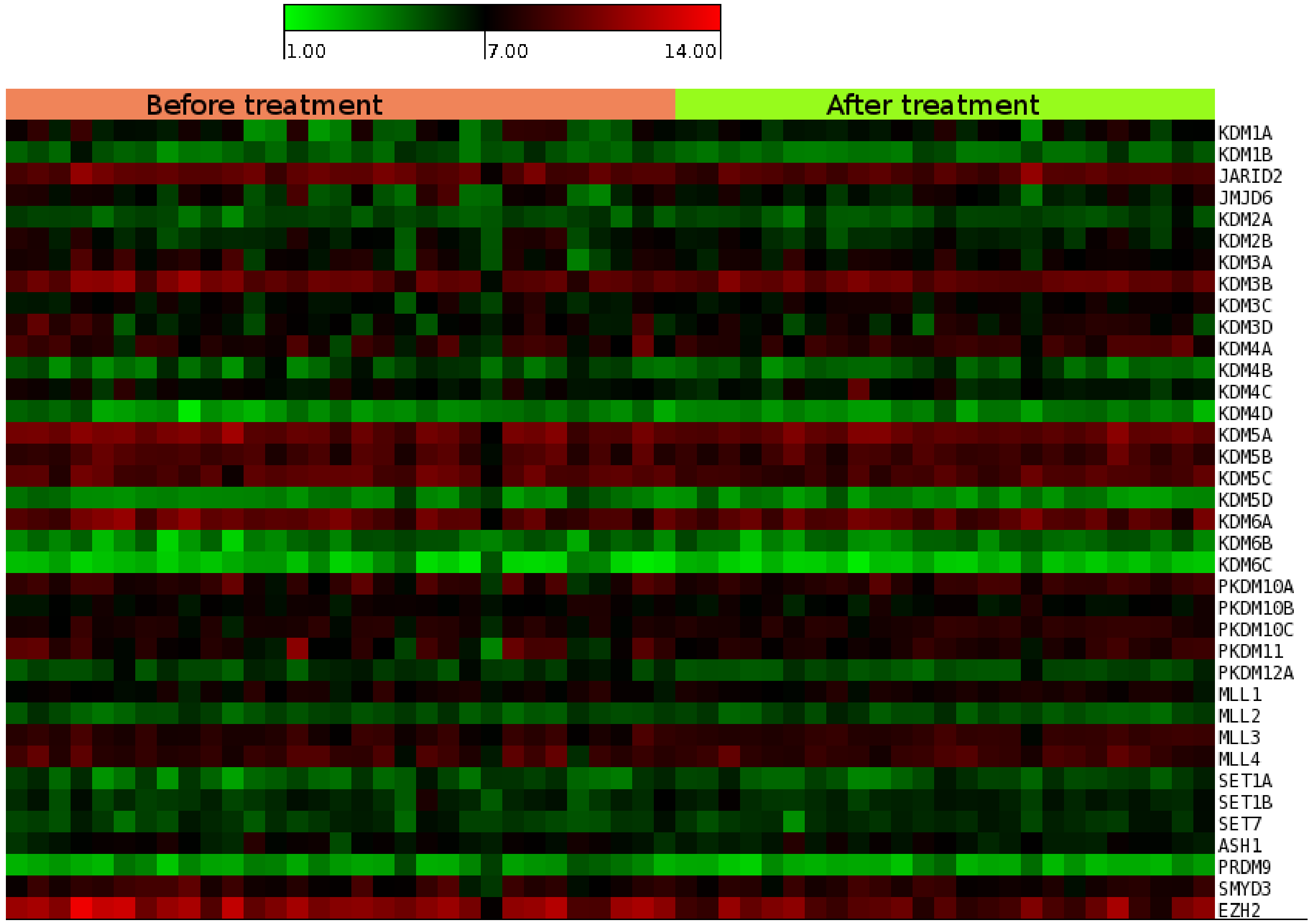
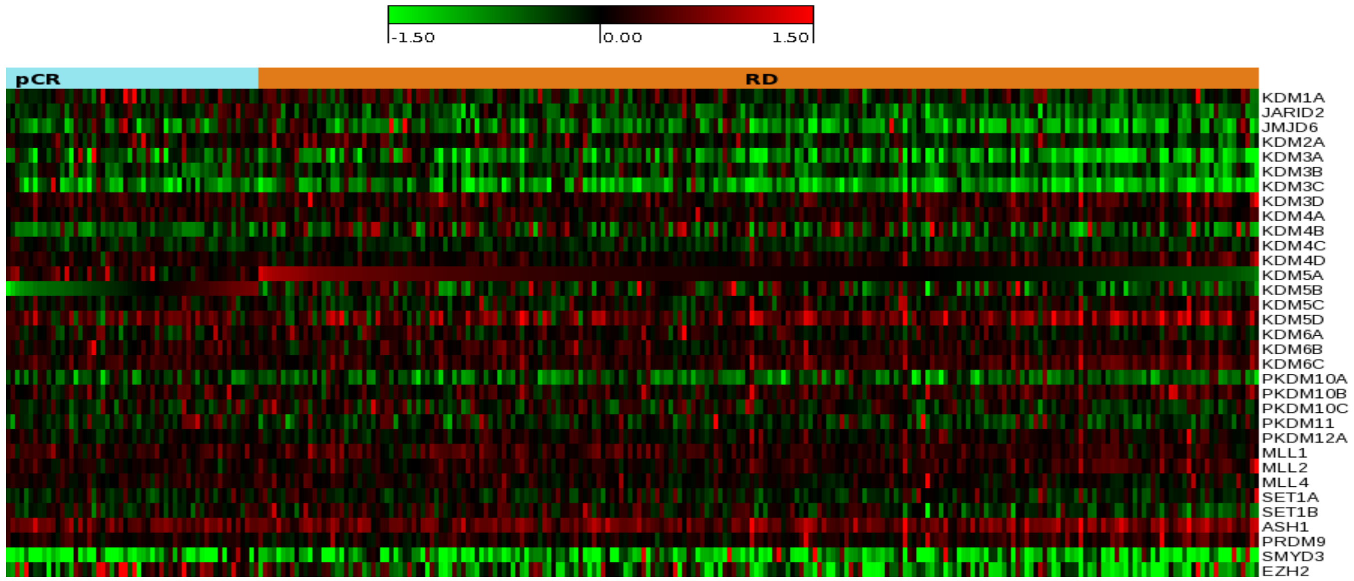


FIG. 2a

Similar phenomena was observed in case of dataset GSE20194. Here patients received 6 months of preoperative (neoadjuvant) chemotherapy including paclitaxel, 5-fluorouracil, cyclophosphamide and doxorubicin followed by surgical resection of the cancer. Response to preoperative chemotherapy was categorized as a pathological complete response (pCR = no residual invasive cancer in the breast or lymph nodes) or residual invasive cancer (RD) (Figure 2b).



, O = testis, P = Thyroid, Q = Urinary.

FIG. 2b

Similary, Cisplatin treatment had no effect of KDM5A or some other multi-drug resistant (MDR) target genes as the breast cancer dataset GSE18864 (N=24) showed (Figure 2c).

FIG. 2C

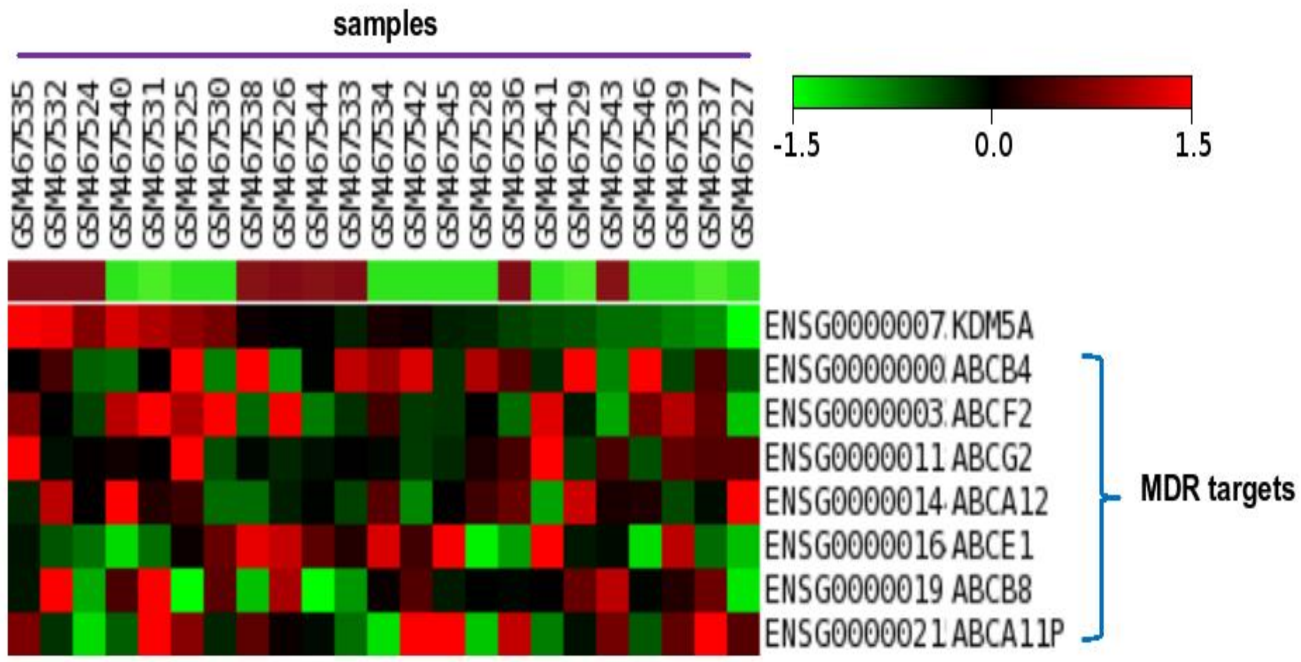


Figure 2: Expression of HMTs and HDMs before and after chemotherapy. Scale indictes Log2 expression values. **(A)** GEO dataset dataset GSE21974, **(B)** dataset GSE20194, (pCR = no residual invasive cancer in the breast or lymph nodes; RD = residual invasive cancer RD); (**C)** dataset GSE18864, patients with lower KDM5A level will have lower level of several implicated MDR genes, and therefore, better clinical response. The green horizontal bar indicates samples before treatment and maroon bar indicates after treatment.

These analyses suggest that new drug may be needed to target KDM5A in breast cancer. Bioinformatic approaches can be taken to virtual screening and identify putative targets in short possible time.

BIOINFORMATIC ANALYSES FOR THE DETERMINATION OF DISEASE PROGNOSIS

Bioinformatic analyses of high-throughput data, especially microarray expression data have lead to the identification of "signature" genes for several cancer types. In our previous study with Cancer-associated fibroblasts (CAF) gene expression profiles were analyzed by microarray to identify deregulated genes in different promigratory CAFs.[(24)] The gene expression signature, derived from the most protumorogenic CAFs, was identified. Interestingly, this "CAF signature" showed a remarkable prognostic value for the clinical outcome of patients with colon cancer. Here we explore the possibility that if the same gene signature can be used to predict the prognosis of breast cancer. Our analyses on two different breast cancer dataset (GSE3494 and GSE4922) suggest that the molecular gene expression signature also has prognostic capabilities in breast cancer (Figure 3a, 3b).

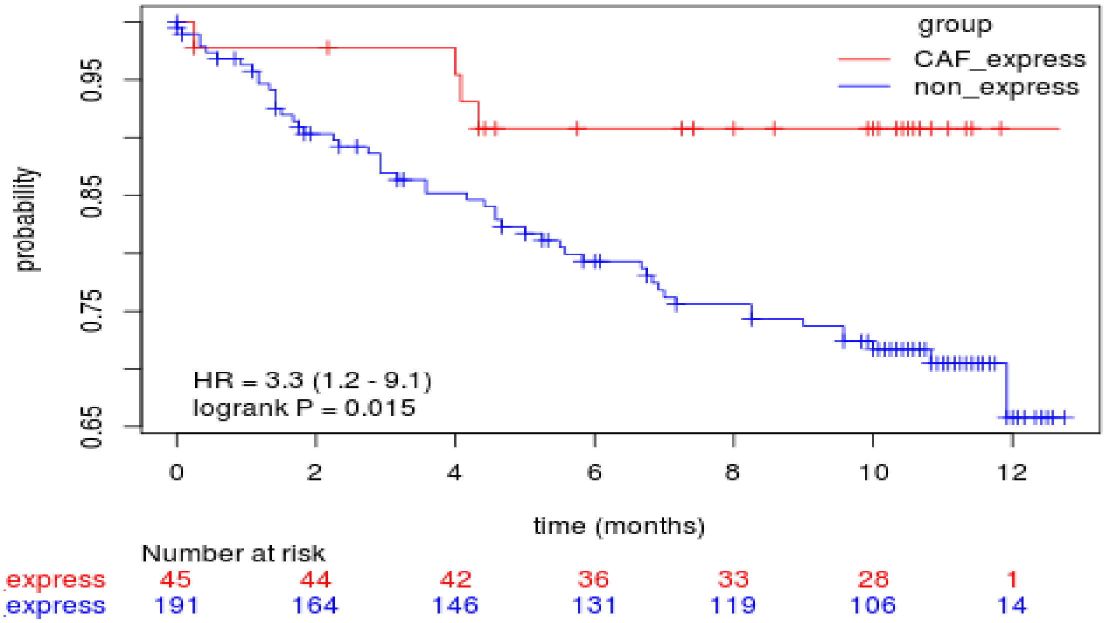

group
CAF_express
non_express
probability
0.65
0.75
0.85
0.95
HR = 3.3 (1.2 - 9.1)
logrank P = 0.015
0
2
4
6
8
10
12
time (months)
Number at risk
express 45 44 42 36 33 28 1
express 191 164 146 131 119 106 14

FIG. 3a

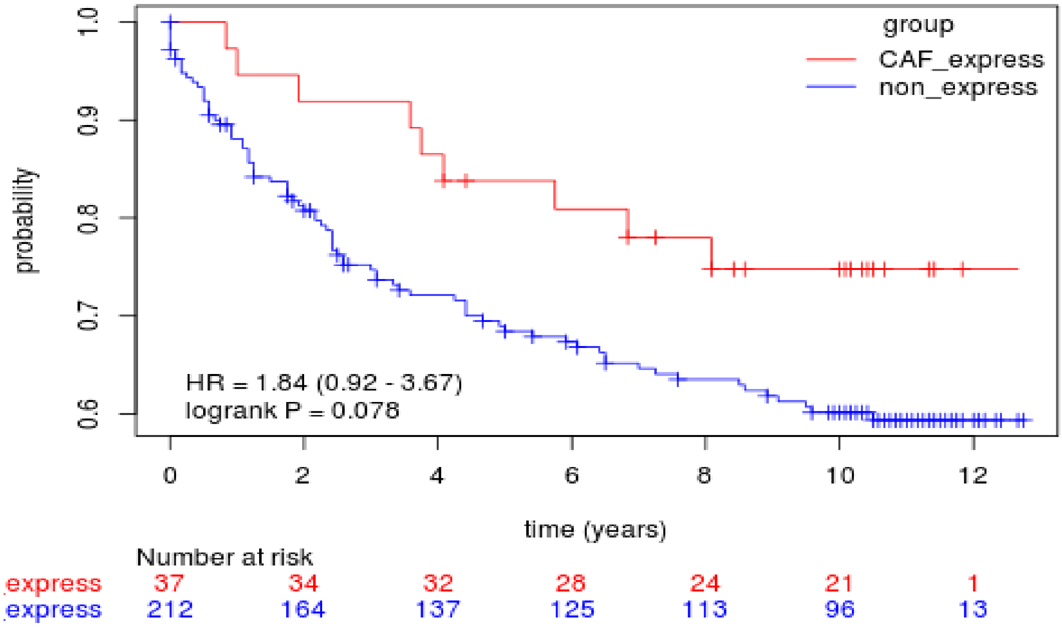

group
CAF_express
non_express
probability
0.6
0.7
0.8
0.9
1.0
HR = 1.84 (0.92 - 3.67)
logrank P = 0.078
0
2
4
6
8
10
12
time (years)
Number at risk
express 37 34 32 28 24 21 1
express 212 164 137 125 113 96 13

FIG. 3b

Figure 3: Capacity of CAF gene signature to identify prognostic outcome. Survivability of breast cancer patients expressing higher and lower expression of CAF signature genes (patients grouped on the basis of the Z score of the CAF signature gene module on median centered expression breast dataset; see Materials and Methods for details). The Hazard Ratio (HR) is based on the Cox model. GEO dataset (a) GSE3494, (b) GSE 4922.

DISCUSSION

The Bioinformatic field has arisen and received much attention at present due to the recent advent of technologies in biological experiments that generate enormous quantities of data requiring interrogation and interpretation. Such experiments include whole-genome sequencing, expression arrays, genome-wide association studies, proteomics, transcriptomics, metabolomics, among others. With the advancement of sequencing technology, and world-wide projects like the 1000 Genomes Project,(30) HapMap,(31) the International Cancer Genome Consortium,(32) and ENCODE,(33) we are currently generating larger volumes of data than ever before. Although these approaches have become popular due to their abilities to handle large quantities of data, bioinformatic applications are not limited to dealing with large datasets that are otherwise not possible to analyze manually. Computational searches using biological data, even on a small scale, have been proven to drive hypotheses, simulate experiments and predict results that would otherwise demand the application of long, tedious and expensive wet laboratory procedures. Therefore, at present bioinformatic studies are not limited to theoretical computational prediction only, but, working closely with a biological scientist, bioinformaticians are playing roles in designing experimental procedures, delineating hypotheses, large scale data analyses, modeling/simulation/prediction, suggesting experimental validation, and data interpretation.

The vast amount of large scale genetic and epigenetic data, especially in the field of tumor biology, being generated everyday from heterogeneous sources encounters a challenge to interpret and extract biologically meaningful information.(5) However, data generated by individual experiments cannot explain complicated biological processes. An integrative analysis of various kinds of experimental data is needed to understand the underlying molecular mechanisms.(17) For example, location analysis of various histone modifications independently would only tell which of the genes are receiving these modification. But only when combining the information of various modifications, can it illustrate the interaction of one with another; explain the necessity of such bi- and multi-lateral controls for cells; and the logic behind the relative distribution of various modifications. To this end, combinations of expression data would tell us spatial and temporal control mechanisms. Moreover, integration of these genetic changes would clarify how the epigenetic switching of signals transmitted to genetic processes.

Furthermore, addition of functional clustering would tell which groups of genes and signals were required for particular processes.(34)

Once the target is identified, next is drug design. Pharmaceutical industries will be able to design new drugs, find new drug targets computationally. Also, biosimilar research will be boosted when the copyright of many drugs from developed country pharmaceuticals expire.

Drug discovery and development is a time-consuming and expensive Process. Estimated that it takes 10–15 years and US $500–800 million to introduce a drug into the market.(35-36) The development of new methods in the field of molecular biology and computer science, has improved the tools for drug design significantly.(37) The field of bioinformatics has become a major part of the drug design that plays a key role for validation drug targets. Bioinformatics also can help in understanding of complex biological processes and to improve drug discovery. It supports CADD research by virtual High-Throughput Screening (vHTS), sequence Analysis, homology modeling with known protein, similarity Searches, drug Lead Optimization, and drug Bioavailability and Bioactivity. Overall, the basic of drug design at bioinformatics lies at structure's based drug design.(37)

ACKNOWLEDGEMENT

I thank Dr. Nuria Lopez-Bigas, Universitat Pompeu Fabra, Spain and Dr. Elizabeta V. Benevolenskaya, University of Illinois at Chicago, USA for their valuable suggestion.

REFERENCES

1. Ideker T, Galitski T, Hood L. A new approach to decoding life: Systems Biology. Annual Review of Genomics and Human Genetics. 2001. 2: 343-372.
2. Hood L, Heath JR, Phelps ME, Lin B. Systems Biology and New Technologies Enable Predictive and Preventative Medicine. Science. 2004. 306(5696): 640-643.
3. Pareek CS, Smoczynski R, Tretyn A. Sequencing technologies and genome sequencing. J Appl Genet. Nov 2011. 52(4): 413–435.
4. Gyles C. The DNA revolution. Can Vet J. Aug 2008. 49(8): 745–746.
5. Concept of Bioinformatics. http://www.iasri.res.in/ebook/CAFT_sd/Concepts%20of%20Bioinformatics. pdf
6. Kosanam H, Prassas I, Chrystoja CC, Soleas I, Chan A, Dimitromanolakis A, Blasutig IM, Rückert F, Gruetzmann R, Pilarsky C, Maekawa M, Brand R, Diamandis EP. LAMC2: A

promising new pancreatic cancer biomarker identified by proteomic analysis of pancreatic adenocarcinoma tissues. Mol Cell Proteomics. 2013. Oct. 12(10): 2820-32.

7. Yap YL, Zhang X. Bioinformatics in Cancer Biomarker Discovery. Landes Bioscience. Chapter 10. Page: 117-127. http://www.landesbioscience.com/pdf/10Zhang_Yap.pdf

8. Antre RV, Pranita P. Computer-Aided Drug Design: An Innovative Tool for Modeling. Open Journal of Medicinal Chemistry. 2012. 2: 139-148.

9. Bharath ES, Manjula, Vijaychand A. In silico drug design-tool for overcoming the innovation deficit in the drug discovery process. Chemistry. 2011. 18(23.3): 10.

10. White S. Pharma 2020: The Vision–Which path will you take? SA Pharmaceutical Journal. 2007. 74(9): 40-41.

11. Furlong LI. Human diseases through the lens of network biology. Cell. 2013. 29(3): 150–159.

12. Kaski S, Saloj¨arvi J, Leen G, Klami A, Peltonen J, Caldas J, Ermolov A, Faisal A, Huopaniemi I, Lahti L, Parkkinen J, Tripathi A. Chapter 5, Bioinformatics, page: 97-106. http://research.ics.aalto.fi/airc/ reports/R0809/bioinf.pdf

13. AmmalKaidery N, Tarannum S, Thomas B. Epigenetic landscape of Parkinson's disease: emerging role in disease mechanisms and therapeutic modalities. Neurotherapeutics. 2013. Oct. 10(4): 698-708.

14. Sharma S, Kelly TK, Jones PA. Epigenetics in cancer. Carcinogenesis. 2010; 31(1): 27-36.

15. Simmons D. Epigenetic influence and disease. Nature Education, 2008. 1(1): 6.

16. Miyamoto K, Ushijima T. Diagnostic and therapeutic applications of epigenetics. Jpn J Clin Oncol. 2005. Jun. 35(6): 293-301.

17. Gundem G, Perez-Llamas C, Jene-Sanz A, Kedzierska A, Islam A, Deu-Pons J, Furney SJ, Lopez-Bigas N. Intogen: integration and data mining of multidimensional oncogenomic data. Nat Methods. Feb. 2010. 7(2): 92–93.

18. Perez-Llamas C, Lopez-Bigas N. Gitools: analysis and visualisation of genomic data using interactive heat-maps. PLoS One. 2011. 6(5): e19541.

19. GEO (http://www.ncbi.nlm.nih.gov/geo)

20. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JY, Zhang J. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol., 2004/,, 5:

R80.

21. Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol., 2004, 3:Article3.

22. Gautier L, Cope L, Bolstad BM, Irizarry RA. Affy--analysis of Affymetrix GeneChip data at the probe level. Bioinformatics 2004, 20: 307-315.

23. Gundem G, Lopez-Bigas N. Sample-level enrichment analysis unravels shared stress phenotypes among multiple cancer types. Genome Med. 2012, 4: 28.

24. Herrera M, Islamm ABMMK, Herrera A, Martín P, Casal I, García de Herreros A, Bonilla F, Peña C. Heterogeneity of cancer-associated fibroblast from human colon tumor shows specific prognostic gene expression signature. Clinical Cancer Research. 2013, 19(21): 5914–26.

25. de la Rica L, Urquiza JM, Gómez-Cabrero D, Islam ABMMK, López-Bigas N, Tegnér J, Toes REM, Ballestar E. Identification of novel markers in rheumatoid arthritis through integrated analysis of DNA methylation and microRNA expression. J. of Autoimmunity. 2013, Mar, 41: 6-16.

26. Furney SJ, Calvo B, Larranaga P, Lozano JA, Lopez-Bigas N. Prioritization of candidate cancer genes - an aid to oncogenomic studies. Nucl. Acids Res. 2008,36(18): e115.

27. Benevolenskaya EV. Histone h3k4 demethylases are essential in development and differentiation. Biochem Cell Biol.,2007, Aug. 85(4): 435–443.

28. Lopez-Bigas N, Kisiel TA, Dewaal DC, Holmes KB, Volkert TL, Gupta S, Love J, Murray HL, Young RA, Benevolenskaya EV. Genome-wide analysis of the h3k4 histone demethylase rbp2 reveals a transcriptional program controlling differentiation. Mol Cell. 2008. Aug. 31(4): 520–530.

29. Beshiri M, Holmes K, Richter W, Hess S, Islam ABMMK, Yan Q, Plante L, Gévry N, Lopez-Bigas N, Kaelin W, Benevolenskaya EV. Coordinated repression of cell cycle genes by KDM5A and E2F4 during differentiation. PNAS. 2012. Nov., 109(45): 18499-18504.

30. 1000 Genomes Project: http://www.1000genomes.org/

31. HapMap: http://hapmap.ncbi.nlm.nih.gov/

32. International Cancer Genome Consortium: https://icgc.org/

33. ENCODE: http://genome.ucsc.edu/ENCODE/

34. Chen Y, Hao J, Jiang W, He T, Zhang X, Jiang T, Jiang R. Identifying potential cancer driver genes by genomic data integration. Nature Scientific Report. 2013. Article number: 3538; doi:10.1038/srep03538.

35. Workman P. How much gets there and what does it do?: The need for better pharmacokinetic and pharma codynamic endpoints in contemporary drug discovery anddevelopment. Curr Pharm Des. 2003. 9: 891–902.

36. Brown D, Superti-Furga G. Rediscovering the sweet spot in drug discovery. Drug Discov Today. 8: 1067–1077.

37. Gomeni R, Bani M, D'Angeli C, Corsi M, Bye A. Computer-assisted drug development (CADD): an emerging technology for designing first-time-in-man and proof-of-concept studies from preclinical experiments. Eur J Pharm Sci. 2001. 13: 261–270.