ISSN No. 0976-5697

Volume 6, No. 3, May 2015 (Special Issue)

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Bioinformatics - A case study of Selection of Relevant Genes in Cancer Cell

V. Bhaskara Murthy
Assoc. Professor, Padmasri Dr. BVRICE
Vishnupur, Bhimavaram, W.G.Dt. A.P.
email: murthyvb@gmail.com

Dr. G. Pardha Saradhi Varma
Professor & Director of PG Courses
Head, Department of IT, S.R.K.R. Engineering
Chinamiram, Bhimavaram., W.G.Dt. A.P
email: gpsvarma@yahoo.

***Abstract*:** This paper gives introduction to cell biology and Bioinformatics and computational methods to get significant genes from cancer Micro Array Datasets.

Keywords : Cell Division—Aging—Feature Selection—Max. Dependency and Maximum Relevance and Significance

I. INTRODUCTION

About Cell Biology—functions of cell-Cell Division and role of medicines followed by introduction to Bioinformatics feature selection depending on Maximum Dependency and Maximum Relevance cum significance using cancer data sets.

a. ROLE OF A CELL:

Human body contains many different cell types, each customized for a particular role. Red blood cells carry life-giving oxygen to every corner of human body. White blood cells kill germ invaders. Intestinal cells squirt out chemicals that chisel away food so one can absorb its nutrients. Nerve cells sling chemical and electrical messages that allow to think and move. Heart cells constantly pump blood, enabling life itself.

Knowledge of the inner workings of cells underpins our understanding of health and disease. Cell structures called organelles. Like the internal organs in human body, organelles in the cell each have a unique biological role to play.

The nucleus—basically the cell's . The nucleus is the most prominent organelle and can occupy up to 10 percent of the space inside a cell. It contains the equivalent of the cell's gray matter—its genetic material, or DNA. In the form of genes, each with a host of helper molecules, DNA determines the cell's identity, masterminds its activities, and is the official cookbook for the body's proteins. Virtually all forms of life fall into one of two categories: eukaryotes or prokaryotes

b. EUKARYOTIC CELLS Vs PROKARYOTIC

EUKARYOTIC CELLS	*PROKARYOTIC CELLS*
The cells of "complex" organisms, including all plants and animals	*"Simple" organisms, including bacteria and blue-green algae*
Contain a nucleus and many other organelles, each surrounded by a membrane (the nucleus and mitochondrion have two membranes)	*Lack a nucleus and other membrane-encased organelles*
Can specialize for certain functions, such as absorbing nutrients from food or transmitting nerve impulses; groups of cells can form large, multicellular organs and organisms	*Usually exist as single, virtually identical cells*
Most animal cells are 10–30 micrometers across, and most plant cells are 10–100 micrometers across	*Most are 1–10 micrometers across*

c. ROLE OF MEDICINE:

All cellular organisms, including bacteria, have ribosomes. And all ribosomes are composed of proteins and ribosomal RNA. But the precise shapes of these biological machines differ in several very specific ways between humans and bacteria. That's a good thing for researchers trying to develop bacteria-killing medicines called antibiotics because it means that scientists may be able to devise therapies that knock out bacterial ribosomes (and the bacteria along with them) without affecting the human hosts.

Several antibiotic medicines currently on the market work by inhibiting the ribosomes of bacteria that cause infections. Because many microorganisms have developed resistance to these medicines, we urgently need new antibiotics to replace those that are no longer effective in fighting disease.

Now, computers are allowing scientists to examine many factors involved in cellular neighbors and decisions all at the same time. The field of computational biology blossomed with the advent of high-end computers. For example, sequencing the 3.2 billion base pairs of the human genome, which was

CONFERENCE PAPER
4th National Conference on Recent Trends in Information Technology 2015 on 25/03/2015
Organized by Dept. of IT, Prasad V. Potluri Siddhartha Institute of Technology, Kanuru, Vijayawada-7 (A.P.) India

completed in 2003, depended on computers advanced enough to tackle the challenge.

d. CELL DIVISION

Each human began as a single cell. This cell couldn't move, think, see, or do things like laugh and talk. But the one thing it could do, and do very well, was dividing—and divide it did. The lone cell became two, and then four, then eight and so on, in time becoming the amazing person that is human. Think of how far one has come. One can laugh at a joke, stand on their head, read a book, eat an ice cream cone, hear a symphony, and do countless other things.

There are two kinds of cell division: mitosis and meiosis. Mitosis is essentially a duplication process: It produces two genetically identical "daughter" cells from a single "parent" cell. Human grew from a single embryonic cell to the person one is now through mitosis. Even after humans are grown, mitosis replaces cells lost through everyday wear and tear. The constant replenishment of human skin cells, for example, occurs through mitosis. Mitosis takes place in cells in all parts of the body, keeping tissues and organs in good working order. Meiosis, on the other hand, is quite different. It shuffles the genetic deck, generating daughter cells that are distinct from one another and from the original parent cell. Although virtually all of human cells can undergo mitosis, only a few special cells are capable of meiosis: those that will become eggs in females and sperm in males. So, basically, mitosis is for growth and maintenance, while meiosis is for sexual reproduction.

Human body carefully controls which cells divide and when they do so by using molecular stop and go signals. For example, injured cells at the site of a wound send go signals to the surrounding skin cells, which respond by growing and dividing and eventually sealing over the wound. Conversely, stop signals are generated when a cell finds itself in a nutrient-poor environment. Sometimes, however, go signals are produced when they shouldn't be, or stop signals aren't sent or heeded. Both scenarios can result in uncontrolled cell division and cancer. Mitosis then becomes a weapon turned against the body, spurring the growth of invasive tumors. Fortunately, it takes more than one mistaken stop or go signal for a cell to become cancerous. Because our bodies are typically quite good at protecting their essential systems, it usually requires a one-two punch for healthy cells to turn malignant. The punches come in the form of errors, or mutations, in DNA that damage a gene and result in the production of a faulty protein. Sunlight, X rays and other radiation, toxins such as those found in cigarette smoke and air pollution, and some viruses can cause such mutations. People also can inherit mutations from their parents, which explains why some families have higher rates of certain cancers: The first punch is delivered at conception. Subsequent mutations can then push a cell down the path toward becoming cancerous. Ironically, mitosis itself can cause mutations too, because each time a cell's DNA is copied, errors are made. (Fortunately, almost all of these errors are corrected by one's extremely efficient DNA repair systems.) So there's an inherent trade off in mitosis: It allows us to grow to maturity and keeps us healthy, but it's also the source of potentially damaging DNA mutations. A number of environmental factors cause DNA mutations that can lead to cancer: toxins in cigarette smoke, sunlight and other radiation, and some viruses.

e. GENE COPY – SIBLINGS

Even members of the same family, who share much of their genetic material, can be dramatically different from one another. If one can ever been to a family reunion, one can seen living proof of this. How can the incredible diversity that one can see in their own families, let alone in the world at large, be explained? Imagine 23 couples participating in a dance. In fact, the number of possible arrangements is 223, or more than 8 million! This means that a single set of parents can produce over 64 trillion different zygotes!

Some family members are exactly the same (genetically, at least): identical twins. Identical twins arise when the embryo splits early in development and creates two genetically identical babies. Fraternal twins, the more common type, are genetically no more similar than siblings. They develop from two different eggs, each fertilized by a different sperm.

II. DIAGNOSIS

The diagnosis of disease involves several levels of uncertainty and imprecision, and it is inherent to medicine. A single disease may manifest itself quite differently, depending on the patient, and with different intensities. A single symptom may correspond to different diseases. On the other hand, several diseases present in a patient may interact and interfere with the usual description of any of the diseases. The best and most precise description of disease entities uses linguistic terms that are also imprecise and vague. Moreover, the classical concepts of health and disease are mutually exclusive and opposite.

Everybody is healthy to some degree h and ill to some degree i. If you are totally healthy, then of course $h = 1, I = 0$. Usually, everybody has some minor health problems and $h < 1$, but $h + I = 1$.

III. BIOINFORMATICS

Bioinformatics derives knowledge from computer analysis of biological data. This data can consist of the information stored in the genetic code, and also experimental results (and hence imprecision) from various sources, patient statistics, and scientific literature. Bioinformatics combines computer science, biology, physical and chemical principles, and tools for analysis and neighbor of large sets of biological data, the managing of chronic diseases, the study of molecular computing, cloning, and the development of training tools of bio-computing systems [16]. Bioinformatics is a very active and attractive research field with a high impact in new technological development [17]. Molecular biologists are currently engaged in some of the most impressive data collection projects. Recent genome sequencing projects are generating an enormous amount of data related to the function and the structure of biological molecules and sequences. Other complementary high-throughput technologies, such as DNA

microarrays, are rapidly generating large amounts of data that are too overwhelming for conventional approaches to biological data analysis. We have at our disposal a large number of genomes, protein structures, genes with their corresponding expressions monitored in experiments, and single-nucleotide polymorphisms

IV. FUZZY LOGIC

Fuzzy logic and fuzzy technology are now frequently used in bioinformatics. The following are some examples.

- To increase the flexibility of protein motifs [1].
- To study differences between poly nucleotides [2].
- To analyze experimental expression data [3] using fuzzy adaptive resonance theory.
- To align sequences based on a fuzzy recast of a dynamic programming algorithm [4].
- DNA sequencing using genetic fuzzy systems [5].
- To cluster genes from microarray data [6].
- To predict proteins subcellular locations from their dipeptide composition [7] using fuzzy k-nearest eighbours algorithm.
- To simulate complex traits influenced by genes with fuzzy-valued effects in pedigreed populations [8].
- To attribute cluster membership values to genes [9] applying a fuzzy partitioning method, fuzzy C-means.
- To map specific sequence patterns to putative functional classes since evolutionary comparison leads to efficient functional characterization of hypothetical proteins [10]. The authors used a fuzzy alignment model.
- To analyze gene expression data [11].
- To unravel functional and ancestral relationships between proteins via fuzzy alignment methods [12], or using a generalized radial basis function neural network architecture that generates fuzzy classification rules [13].
- To analyze the relationships between genes and decipher a genetic network [14].
- To process complementary deoxyribonucleic acid (cDNA) microarray images [15]. The procedure should be automated due to the large number of spots and it is achieved using a fuzzy vector filtering framework.

V. FEATURE SELECTION

In real-data analysis, data set may contain a number of redundant and insignificant features with low relevance to the classes.[18]

- The presence of such redundant, insignificant, and non-relevant features leads to a reduction in the useful information.
- Ideally, the selected features should have high significance and relevance with classes, while redundancy among them would be as low as possible.
- Expected to be able to predict the classes of the samples.
- Hence, to assess the effectiveness of the features, relevance, significance, and redundancy need to be measured quantitatively.
- Rough sets, entropy, mutual information, f-information

VI. MAX-DEPENDENCY CRITERION

Max-Dependency: Select a subset of features (condition attributes) which jointly have the largest dependency on the target class (decision attribute)[19]

- Problem of Max-Dependency for real life applications:
- Not sufficient for selecting highly discriminative features
- Hard to generate resultant equivalence classes in the high-dimensional space: the number of samples is often insufficient and the generation of resultant equivalence classes is usually an ill-posed problem.
- Slow computational speed

VII. MAX-RELEVANCE-MAX-SIGNIFICANCE CRITERION

Dependency with the mean value of all dependency values between individual feature and target class label

- Features selected according to Max-Relevance could have rich redundancy, that is, dependency among them could be large.
- When two features highly depend on each other, the respective class
- Discriminative power would not change much if one of them were removed.
- Max-Significance: Select mutually exclusive features

VIII. RESULTS

In publicly available six cancer and two arthritis data sets are used. Since binary classification is a typical and fundamental issue in diagnostic and prognostic prediction of cancer and arthritis, different methods are compared using the following eight binary class data sets.

1) Breast Cancer
2) Leukemia
3) Colon Cancer
4) Lung Cancer
5) Prostate Cancer
6) Rheumatoid Arthritis versus Osteoarthritis (RAOA)
7) Rheumatoid Arthritis versus Healthy Controls (RAHC)

CONFERENCE PAPER
4th National Conference on Recent Trends in Information Technology 2015 on 25/03/2015
Organized by Dept. of IT, Prasad V. Potluri Siddhartha Institute of Technology, Kanuru, Vijayawada-7 (A.P.) India

The following graphs show Max. Dependence, Max. Relevance (figure: 1) and Max. Relevance and Max. Significance (figure: 2)

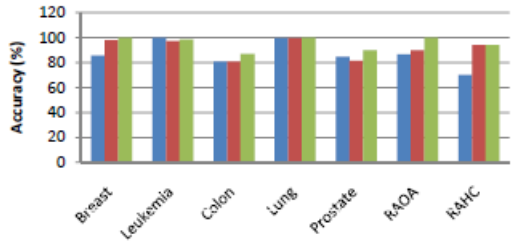

Figure 1.

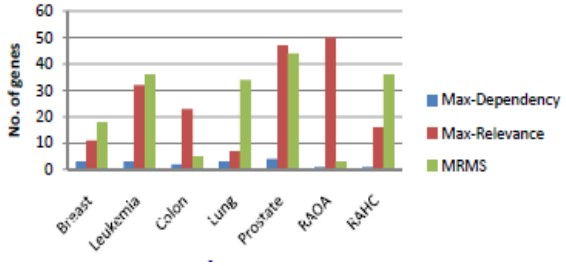

Figure 2.

IX. Acknowledgment

In This paper, Feature Selection is based on Max. Significance and Max. Relevance genes to be expected for estimating the feature cancer disease with a prior knowledge to prevent it. Any modifications are welcome.

X. References

[1] Chang BC, Halgamuge SK. Protein motif extraction with neuro-fuzzy optimization. Bioinformatics. 002;18(8):1084–1090.

[2] Torres A, Nieto JJ. The fuzzy polynucleotide space: basic properties. Bioinformatics. 2003;19(5):587–592.

[3] Tomida S, Hanai T, Honda H, Kobayashi T. Analysis of expression profile using fuzzy adaptive resonance theory. Bioinformatics.2002;18(8):1073–1083.

[4] Schlosshauer M, Ohlsson M. A novel approach to local reliability of sequence alignments. Bioinformatics.2002; 18(6):847–854.

[5] Cord´on O, Gomide F, Herrera F, Hoffmann F, Magdalena L.Ten years of genetic fuzzy systems: current framework and new trends. Fuzzy Sets and Systems. 2004;141(1):5.31.

[6] Belacel N, C˘ uperlovic´-Culf M, Laflamme M, Ouellette R.Fuzzy J-Means and VNS methods for clustering genes from microarray data. Bioinformatics. 2004;20(11):1690–1701.

[7] Huang Y, Li Y. Prediction of protein subcellular locations using fuzzy k-NN method. Bioinformatics. 2004;20(1):21– 28.[8]

[8] Carleos C, Rodriguez F, Lamelas H, Baro JA. Simulating complextraits influenced by genes with fuzzy-valued effects in pedigreed populations. Bioinformatics. 2003; 19(1):144–148.

[9] Demb´el´e D, Kastner P. Fuzzy C-means method for Clustering microarray data. Bioinformatics. 2003; 19(8) :973– 980.

[10] Heger A, Holm L. Sensitive pattern discovery with 'fuzzy' alignments of distantly related proteins. Bioinformatics.2003;19(suppl 1):i130–i137.

[11] Woolf PJ, Wang Y. A fuzzy logic approach to analyzing gene expression data. Physiological Genomics. 2000;3(1):9–15.

[12] [12] Blankenbecler R, Ohlsson M, Peterson C, Ringn´er M. Matching protein structures with fuzzy alignments. Proceedings of theNational Academy of Sciences of the United States of America. 2003;100(21):11936–11940.

[13] Wang DH, Lee NK, Dillon TS. Extraction and optimization of fuzzy protein sequence classification rules using GRBF neural networks. Neural InformationProcessing—Letters and Reviews. 2003;1(1):53–59.

[14] Ressom H, Reynolds R, Varghese RS. Increasing the efficiency of fuzzy logic-based gene expression data analysis. Physiological Genomics. 2003;13(2):107–117.

[15] Lukac R, Plataniotis KN, Smolka B, Venetsanopoulos AN cDNA microarray image processing using fuzzy vector filtering framework. Fuzzy Sets and Systems. 2005;152(1):17–35.

[16] Bourbakis NG. Bio-imaging and bio-informatics. IEEE Transactions on Systems, Man and Cybernetics, Part B: Cybernetics.2003;33(5):726–727.

[17] Fuchs R. From sequence to biology: the impact on bioinformatics.

[18] C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," J. Bioinformatics Comput. Biol., vol. 3, no. 2, pp. 185–205, Apr. 2005.