DBT- CTEP SPONSORED

सत्यमेव जयते

Department of Biotechnology
Government of India

DBT-CTEP SPONSPORED

Two Days National Level Workshop On

BIOETHICS AND COMPUTATIONAL INTELLIGENCE IN HEALTH INFORMATICS

Jointly Organized with

DEPARTMENT OF INFORMATION TECHNOLOGY

Kongunadu
College of Engineering & Technology
Namakkal - Trichy Main Road, Tholurpatti (P.O.), Thottiyam (Tk), Trichy (Dt.) - 621 215.
(AICTE Approved, Affiliated to Anna University, Chennai & ISO 9001 : 2008 Certified Institution)

INDEX

DAY-I (05.02.2016)

DAY-II (06.02.2016)

INTRODUCTION TO BIO ETHICS

By,

PRAMOD P WANGIKAR,

IIT BOMBAY

OBJECTIVE:

To give a brief introduction about the various ethical systems people use when facing ethical dilemmas/decisions. It also aims to provide student with an opportunity to comfortably discuss their views and values on different bioethical issues. It the deals with the issues in the context of rules that govern the behaviour of the individual.

THEME:

To introduce key topics shaping the development of bioinformatics as a discipline, including commercialisation of innovations and intellectual property in biotechnology and software industries.To encourage reflection on the ethical concerns that affect the practice of bioinformatics both as a life science and an engineering discipline.To develop a number of professional skills including cross-discipline communication, teamwork, and job search and interview skills Study of the normative judgments related to problems of how to decide or how to find the best course of action in a variety of issues related to biomedical science.

- Difference between *descriptive* and *normative* studies.
- Bioethics is naturally an interdisciplinary field of study.
- Since medicine deals with the life and death of people, ethical problems naturally arise.
- Doctors face ethical problems everyday. Research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral or health data, including those to acquire, store, organize, archive, analyze, or visualize such data.”
- “Bioinformatics applies principles of information sciences and technologies to make the vast, diverse, and complex life sciences data more understandable and useful.

Discuss the factors that affect biotechnology and computer industries (especially at the start-up stage) such as intellectual property and funding.Develop an awareness of various views of ethical issues (including engineering ethics and bioethics) and a framework for reflecting on these issues.

The focus of this activity is to help students first recognize their way of thinking and then remove themselves from personal interest inorder to understand other types of reasoning. Bioinformatics, the application of computational techniques to analyse the information associated with biomolecules on a large-scale, has now firmly established itself as a discipline in molecular biology, and encompasses a wide range of subject areas from structural biology, genomics to gene expression studies. In this review we provide an introduction and overview of the current state of the field. We discuss the main principles that underpin bioinformatics analyses, look at the types of biological information and databases that are commonly used, and finally examine some of the studies that are being conducted, particularly with reference to transcription regulatory systems. The aims of bioinformatics are threefold. First, at its simplest bioinformatics organises data in a way that allows researchers to access existing information and to submit new entries as they are produced, eg the Protein Data Bank for 3D macromolecular structures . While data-curation is an essential task, the information stored in these databases is essentially useless until analysed. Thus the purpose of bioinformatics extends much further. The second aim is to develop tools and resources that aid in the analysis of data. For example, having sequenced a particular protein, it is of interest to compare it with previously characterised sequences. This needs more than just a simple text-based search and programs such as FASTA and PSI-BLAST [9] must consider what comprises a biologically significant match. Development of such resources dictates expertise in computational theory as well as a thorough understanding of biology. The third aim is to use these tools to analyse the data and interpret the results in a biologically meaningful manner. Traditionally, biological studies examined individual systems in detail, and frequently compared them with a few that are related. In bioinformatics, we can now conduct global analyses of all the available data with the aim of uncovering common principles that apply across many systems and highlight novel features. Teaching bioethics can serve as a way to teach science to students who otherwise might not be engaged with the subject. Bioethics provides a realworld context for introducing and underscoring the “need to know” science concepts. Case studies help students see the relevance of the science content they are learning and motivate them to apply their science understanding to issues of social relevance. Bioethics may also inspire students to gain a deeper understanding of the scientific facts so they can make well-reasoned ethical arguments. Bioethical issues interest students across a range of learning abilities and inclinations. Th e National Science Education Standards point to the need for students to understand the role of science in society and to recognize how science influences and is influenced by economic, political, and social issues.National standards also ask that students be able to understand and evaluate costs and benefits

associated with technological advances.Bioethics is a way to deepen students' understanding of medical research and its impact on society. Biomedical and clinical research has led to dramatic breakthroughs in the understanding of disease and disease prevention as well as new treatments. New knowledge requires a citizenry capable of making informed decisions to guide personal choices and public policy. This supplement gives students an opportunity to prepare for the scientifi c, medical, ethical, personal, and public-policy choices they will face as adults in the 21st century. Promote respectful dialogue among people with diverse views. Engaging in bioethics discussions helps develop students' ability for reasoned dialogue, especially among students with different perspectives. It also encourages students to think about choices from a variety of viewpoints and interests, thus facilitating respectful discussions of potentially contentious issues. These skills are fundamental for an effective democracy. Cultivate critical-reasoning skills. Bioethics activities emphasize the importance of justification, a process of giving reasons for views. Research indicates that people have more difficulty reasoning in the ethical domain than in any other. Even many adults tend to rely on rules and often resist delving deeply to consider the reasons for the rules, or to see whether there are ever appropriate exceptions. Others believe that moral truths are wholly subjective, resistant to reasoned analysis, and that any one opinion is as good as any other. Exploring Bioethics gives students the chance to develop their ethical reasoning skills so that they can critically analyze problems in a more careful and nuanced way.

The ethical system to be studied are:

1. Ethical relativism
2. Divine command theory
3. Utilitarianism
4. Deontology
5. Virtue ethics

The inherent tension between respecting an individual's choice not to be vaccinated and the need for widespread vaccination to ensure the health of the entire community; apply the ethical consideration of fairness to circumstances in which individuals who do not bear any potential burdens of vaccination still benefit from community immunity; and describe under what circumstances, if any, students believe vaccination

The basic principle of bio-ethics are:

AUTONOMY :**one should respect the right of individuals to make their own decision.**

The principle of autonomy is based on the Principle of Respect for Persons, which holds that individual persons have right to make their own choices and develop their own life . In a

health care setting, the principle of autonomy translates into the principle of informed consent: You shall not treat a patient without the informed consent of the patient or his or her lawful surrogate, except in narrowly defined exceptions.

Beneficence – one should take positive steps to help others

JUSTICE: The even distribution of benefits and risks throughout society .

The same applies to biomedical research - Groups (racial, ethnic, gender, etc) should potentially benefit and/or potentially be exposed to risk at the same rate as other groups, unless there is something else that justifies unequal distribution (refer to "acing the test")

NONMALEFICENCE: Do no harm

Example: While it might be interesting to see what happens to your kneecap if we smashed it with a crowbar, we know that would cause harm to you, so couldn't do that (even though it might be interesting scientifically) Experiments should not be performed on humans if we know they will produce harm. If it becomes apparent that the experiment is harming the human participants, the study must stop.

OUTCOME:

To understand about the typical controversial ethical problems emerging from new situations and possibilities brought about by advances in biology and medicine. It is also moral discernment as it relates to medical policy and practice.

ETHICAL ISSUES AND CHALLENGES IN GLOBAL WEALTH

By,

PRAMOD P WANGIKAR,

IIT BOMBAY

OBJECTIVE:

To describe about the issues that are available in developing a healthy nation, and also to overcome the factors that act as a barrier for the development of global wealth.

THEME

There is always a rich public debate about how the potential risks associated with biotechnology methods and bioindustry.The products should be assessed about how bioethics should influence public policy. A general structure for guiding such public policy is emerging in current situation but is not fully developed for the betterment of the nation. Groups started receiving risks that are differently depending on their culture, scientific background, perception of government, and other factors. Expert say that opinion may supports a wide range of positions in individual factors.Deeply and honestly held but often conflicting beliefs and values about nature, animals, and the community good animate the debate. The result is that biotechnology issues are often highly contentious and debated on both scientific and ethical grounds. There are some contemporary that are followed in experimenting the bioinformatics.

Two contemporary examples are:

- First, Does humans social benefit like living longer and leading more productive life is due to the improvement in biotechnology?.If so, at any cost will it harm the animal group or animals that are involved for experiment to produce benefits?
- Second, to apply for the insurance, does a company require the information about individual's genetic inheritance?

ISSUES

Bioethics examines broad issues such as animal rights and welfare, human testing, and the potential effects of genetically engineered species on other species and the environment. The Risk assessments analyze the relative risks posed by possible toxic, pathogenic, and ecological effects of biotechnology and bio industry.

There are three broad analytical approaches to risk assessment:

- To process the design and develop the new organism and their effect on environment.

- The various effect of the organism has to be studied.
- The inherent rights and its relationship to the environment have been determined.

States and the federal government generally focus on the second approach, the characteristics and environmental risks of the altered organism, and not on the processes used to produce it or on possible natural rights.

This "organism-in the environment" approach is to assess the risk that is involved in evaluation of any of the following.

- The probability of the survival and growth of genetically engineered organisms beyond intended environments;
- The extent to which genetically engineered organisms may be harmful to humans, the environment, or to other organisms or species that they come in contact with; and
- The extent to which genetically engineered organisms may exchange genetic material or information with other organisms, resulting in possible harmful effects.

USAGE OF HAZARDOUS MATERIAL

Waste Categories Hazardous waste can be broadly categorised into four categories:

- Chemical
- Radioactive
- biohazardous
- material that is sharp.

Each category has hazards which have an effect on safer handling and safe disposal practices, and a specific waste may have properties of more than one category.

Chemical waste:

Chemical wastes which are hazardous are disposed through a hazardous waste disposal program managed by the safety department. The term “hazardous” refers to materials or chemicals that are corrosive, flammable, reactive, explosive or toxic. The regulatory description of hazardous waste, in a broader sense, includes the majority of known chemicals when they are to be discarded. The waste disposal of hazardous chemicals is managed in accordance with regulation of the Cregon Department of Environmental Quality (DEQ) and the U.S. Environmental Protection Agency. These regulations suggest specific methods for disposal of different types of hazardous chemical wastes. Therefore, the safety department has specific guidelines which must be strictly followed with reference to packaging, labelling, and disposal of hazardous waste. Since generators are charged for costs

associated with waste disposal, guidelines have also been established by the safety department for recycling and waste minimizing methods.

Radioactive waste:

Radioactive substances are most toxic. As compared to organic poisons, infurious effects of radio-nuclides are exceedingly high. For example, radium is 25,000 times more lethal than arsenic. Nuclear war materials, test explosions, craze for power plants, radioisotope use in medicine, industry and research are the main source of radioactive pollution that could threaten our environmental security. There is no suitable and cheap method of disposal of radioactive waste (spent nuclear fuel gaseous effluents and low level wastes). At any time radioactivity is likely to escape from the waste in water bodies, concrete cases and salt formations in high mountains. The nuclear waste is thus likely to get leached into the biosphere. Pollution control boards and environmental protection agencies must evolve certain foolproof methods to prevent above mentioned pollution by handling radioactive wastes carefully.

Biohazardous waste:

Biological hazard or biohazard means infectious agents causing a risk of death, injury or illness to individuals who handle them. All waste materials which contain such agents must be autoclaved or chemically sterilized before disposing into the general trash. A control viz., sterilizer indicator tape has to be used to assure the effectiveness of treatment. Toxicity and radioactivity like hazards should not be ignored when disposing of sterilized materials. Provided sterlization is not practical, then biohazardous material must be incinerated in a DEQ-permitted infections waste incinerator.

Sharp materials:

Sharp materials including needles, broken glass, and razor blades provide danger both to initial users and to others who may come in contact with that. Besides causing physical damage, such materials, when contaminated, can provide an entry route into the body for toxic or infectious substances. Therefore, sharp materials should be enclosed in a rigid container and placed in garbage dumpsters.

CHALLENGES IN GLOBAL WEALTH

Wealth management is now emerging as a key core competency across universal banks and well-balanced financial services institutions. Firms are focusing on integrating business units to improve client satisfaction by offering more targeted services. While firms have attempted to integrate multiple business units in the past, the focus was more on capturing overall firm-level synergies than on providing client benefits. As wealth

management firms try to grow their businesses, they are also facing challenges from increased regulations. Many of these regulations are expected to alter the way investment products have traditionally been distributed by wealth management firms. Some regulations have also increased the burden on data reporting. Firms will now need to process more data and create more reports, potentially increasing their costs and negatively impacting their operational efficiencies. The industry is also witnessing dynamic changes as new firms enter the market and as clients become more demanding. Firms now feel the need to be more flexible and quicker in response to these changes to remain competitive in the industry and to better serve all stakeholders.

These changes have led to the emergence of the following key business-focused trends in the wealth management industry globally:

1. Stronger focus on leveraging enterprise value.
2. Increased investment in IT as regulations become an operational challenge
3. Increased use of Software-as-a-Service as a cost-effective
4. Adaptable solution to changing business needs

The main challenge that is related to the future development of Wealth is how to maintain a high level of growth at the global level, while simultaneously tackling the issue of higher wealth inequality. Before the Great Recession, wealth inequality was a topic of discussion and concern mainly in developing countries where inequality was historically high. Nevertheless, in the post-recession era, there is an increasing concern on topics related to wealth inequality in Developed countries, most notably in the USA and the Euro Zone.56% of people living in rich countries, believe the most pressing problem of the economy is inequality.

Another challenge is the need to reanalyse and review the role of capitalism in wealth creation and wealth distribution. Capitalism has been the engine behind wealth growth in the large majority of countries in the world since the industrial revolution. But, the model is currently under attack and an increasing proportion of the global population even in developed countries which believes capitalism has contributed to the global crisis without contributing to the search for a long-term solution. As a result, trust in capitalist societies as problem solvers, is at an historically low level. Even if the large majority of global leaders would agree that there is no better alternative to the creation and distribution of wealth, there is an increasing pressure to move to a new form of capitalism, one with a more human side to it. Finally, another major challenge that needs to be considered is the rapid growth in wealth which is taking place in developing countries, especially China and India. The increasing

proportion of citizens from those massively populated countries who now have access to higher levels of wealth, will have important consequences in terms of global supply chains, global prices, environmental issues, as well as the geopolitical implications, that have already began to become evident. It is clear, for example, that the position of geopolitical importance of China before and after the Great Recession has completely shifted in favour of the Asian giant. But as the importance of China is growing in a large number of global value chains, both as a main producer and consumer, there is increasing concern about how a potential downturn in that economy will affect the rest of world, still feeling the pinch from the last recession.

OUTCOME:

To overcome the ethical issues the nation facing interms of health and animal welfare by using various aspects.There are also certain challenges that are to be taken up by the individuals to safe the biotechnology and the products.

EVOLUTIONARY COMPUTATION TECHNIQUES

By,

PROF.M.KRISHNAN,

BHARATHIDASAN UNIVERSITY,

TRICHY.

OBJECTIVE

Bioinformatics and computational biology involve the comprehensive application of mathematics, statistics, science, and computer science to the understanding of living systems. Research and development in these areas require cooperation among specialists from the fields of biology, computer science, mathematics, statistics, physics, and related sciences. The objective of this book series is to provide timely treatments of the different aspects of bioinformatics spanning theory, new and established techniques, technologies and tools, and application domains. This series emphasizes algorithmic, mathematical, statistical, and computational methods that are central in bioinformatics and computational biology.

THEME

The analysis of micro-array data is also central to much research in computational systems biology, although here the emphasis is slightly different. A major concern of computational systems biology is the development of dynamic predictive models of biological (especially genetic and biochemical) processes. The first stage in this process is the identification of interacting partners (used in a loose sense).One approach to identifying gene–gene interactions is to attempt to use observed correlations in gene microarray data to infer networks of interaction.

Network inference:A variety of different approaches to network inference are possible, and many widely used techniques are fundamentally Bayesian in nature. Again, it is worth emphasising the apparent confusion between discrete Bayesian networks and more general Bayesian methods.

(1)Sequence alignment projects/techniques – articles in this area cite sequence alignment as a primary tool for the research. This might include pairwise and multiple sequence alignments as well as BLAST searches. Sequence alignment is often used by researchers to compare either the DNA or amino acid sequences of organisms to determine homology and generate phylogenetic relationships between them. There are actually two main types of sequence alignment, pairwise and multiple. Pairwise involves comparing two sequences to

each other, while multiple sequence alignment involves aligning several sequences to each other or to a single sequence. Without the ability to sequence DNA and proteins, much of the research generated in the last fifty years would not have been possible. Diabetes research is no exception. The only way to tell if a form of insulin was usable by humans was experimentation. The sequence of the protein itself was not known and could not be compared to the sequences from different organisms. Fred Sanger and O.E.P Thompson published their landmark paper in which they presented the entire sequence of the insulin protein, which was the first protein to be entirely sequenced . They discovered that human insulin has two chains, the A chain of 21 amino acids and the B chain, which has 30 amino acids and they are connected by disulfide bridges. The impact of this research was far-reaching not just for diabetes research, but also for genetics and sequencing research. Sanger and Thompson basically demonstrated that sequencing could be done in a reasonable amount of time for some proteins. Moreover, once sequencing was done, as a synthetic could. Pairwise sequence alignment and multiple sequence alignment techniques are often used by researchers in biology . Research was based on the premise that brain-derived neurotrophic factor (BDNF) controls the actions of several proteins including insulin, leptin, and gherlin and is significant to the pathobiology of type2 diabetes and obesity. To test their hypothesis, Rao et al. located genes and proteins that are commonly present in diabetics for both homo sapiens and mus musculus (house mouse). Using multiple sequence alignment, they aligned the sequences representing BDNF, MET66, CRP, Insulin, Leptin, and Ghrelin from each organism. Using those alignment scores, they generated a phylogenetic tree using the ClustalW ver 1.83 program. After working with this smaller dataset, they then aligned additional sequences of other proteins involved in diabetes and obesity which numbered around most was for the homo sapiens family and 59 for the mus musculus family. Results from both sets of alignment indicate there is a strong relationship between the two organisms and the related proteins that suggest BDNF could be used as a biomarker for diabetes and obesity in future studies or could be exploited as a target for drug development. Rao et al., was a straightforward use of sequence alignment and provided some insight on how this could be used for diabetes research. However, the article was not in depth in its description of its techniques and motivations. Therefore we looked for other articles that discussed sequence alignment and diabetes. It provides us a well described study that used sequence alignment and other bioinformatics tools to analyze the function of newly discovered gene GLUT10, which they believe has a role in Type 2 diabetes. The Dawson article is extremely detailed the methodology used to isolate the human and mouse genes used in the study. Since the process of isolating genes is not the focus here, we will only briefly describe it. Using a

partially identified transcript from the Sanger Center, the researchers searched NCBI for similar matches in. The information gathered from these searches were used to develop primers which were then used to amplify the GLUT10 transcript from the cDNA for both human and mouse. Here we see the importance of the NCBI databases in this type of research. Without a central repository for sequence information, many experiments would be impossible in the short term and difficult in the long run. In addition to isolating the sequences, the functional analysis of the gene was tested in Xenopus laevis subjects by removing the healthy oocytes and adding insulin to them to observe their reactions. After the data was collected the analysis was performed. Several bioinformatics tools were employed at this stage. First, the DNA sequences were translated to amino acid sequences and the resulting sequences of residues were aligned with the programs align and BOXSHADE.

(2) Gene Expression projects/techniques – articles in this area cite various methods to measure the expressions of genes in different organisms and conditions. Microarray analysis is also frequently mentioned in these papers. Most cells in the body contain a full set of chromosomes and identical genes. However, only certain genes are actually "turned on" for different cells that give the cell type particular abilities. The genes that are "turned on" are defined as expressed genes. For many studies investigating which genes are turned on is the focus of the research. In fact, gene expression techniques were the focus of almost half the recent articles we reviewed. Techniques varied from generating genetic linkage maps to microarray analysis. In this section we will present several papers that outlined the use of different gene analysis tools.

The tools they used in their study included:

(1) Prioritizer

(2) Endeavour

(3) DGP

(4) Geneseeker

(5) G2D

(6) PandS

Another popular tool for investigating gene expression is microarray analysis. In general microarray analysis involves hybridizing mRNA molecules to the DNA templates from which they originated. By measuring the amount of mRNA bound to each site of the array, they can ascertain how genes are expressed under different conditions, in different tissues, and in different organisms. These have become important tools because several thousand genes can be expressed at one time in one experiment. This facilitates the process of gene study considerably. To analyze microarrays, several statistical techniques are used in concert

including sample t tests, log2 ratios, normalization, and ANOVA tests. Reece et al. describe their use of microarray analysis investigate how maternal diabetes can affect the developing embryo. Since the physical process of developing a microarray is not the focus here, we will concern ourselves with statistical tools used for the analysis. The article specifically states the researchers used the 1-sample t test on log2 ratios to determine significant differences between expression levels. In addition, they used the student t test to compare the microarray data to qRT-PCR data. While this article did not name diabetes specifically, the focus was on pancreatic tissue and islets, which are responsible for insulin production. It is interesting to mention, that diabetes seems to be used often as prototype disease for developing and validating bioinformatics approaches. This is certainly the case with Collins et al., who write about high throughput biomarker discovery . The goal for their study was to search for biomarkers that might indicate a disease. For their test case, they searched for genes that are known to make organisms susceptible to Type 1 diabetes . By locating these markers, they hope to determine what other factors might be responsible for a person developing diabetes who possesses the susceptibility genes as having the gene alone does not indicate a person will develop the disease. Using traditional microarray experiments to generate data, they used t-tests to compare the data from T1D patients with those who were autoantibody positive (susceptible to develop diabetes, but do not currently have diabetes). The first part of their experiment involved comparing single markers in the different populations, which led to univariate statistical analysis. They reasoned however that using multiple markers would yield more conclusive results, so they employed several additional statistical tools in the discriminant analysis area including, parametric and nonparametric tests (kernel based and K nearest neighbor).

(3) Databases and database techniques – articles in this area cite various databases that were either used to assist with research or compiled to assist other researchers with their research. While sequence alignment and microarray analysis techniques are important in diabetes and other medical research, the storage and retrieval of biology related data is key to the exploitation of the data acquired during that research. Supporting that argument, several of the papers we reviewed dealt specifically with the topic of databases that were used or developed for research. In this section we will look at those articles to see how diabetes research is impacted by existence and development of databases. Perhaps the most common interaction with databases that researchers have is using them. We have already seen several examples of databases (BLAST, genome database, NCBI resources) that have been used during the normal course of research. They are often used to supplement data that cannot be reasonably collected directly by the researchers as was the case for Craig et al. . In their

study, they screened several exon regions from the DNA of 48 African Americans and 48 Caucasians from which they identified 21 single nucleotide polymorphisms. In addition to the 21 SNPs they observed, they used public databases (NCBI) to located other SNPs on which they could perform their study. By drawing from other sources, they were able to combine their data with a wider resource to get a better picture of the binding protein in which they were interested. Without such resources, this type research can be difficult and possibly less convincing because research completed with smaller datasets can be suspect depending on the context .

FEATURE SELECTION TECHNIQUES

As many pattern recognition techniques were originally not designed to cope with large amounts of irrelevant features, combining them with FS techniques has become a necessity in many applications . The objectives of feature selection are manifold, the most important ones being:

- to avoid overfitting and improve model performance, i.e. prediction performance in the case of supervised classification and better cluster detection in the caseof clustering,
- to provide faster and more cost-effective models,and
- to gain a deeper insight into the underlying processes that generated the data.

However, the advantages of feature selection techniques come at a certain price, as the search for a subset of relevant features introduces an additional layer of complexity in the modeling task. Instead of just optimizing the parameters of the model for the full feature subset, we now need to find the optimal model parameters for the optimal feature subset, as there is no guarantee that the optimal parameters for the full feature set are equally optimal for the optimal feature subset . As a result, the search in the model hypothesis space is augmented by anotherdimension: the one of finding the optimal subset of relevant features. Feature selection techniques differ from each other in the way they incorporate this search in the added space of feature subsets in the model selection.In the context of classification, feature selection techniques can be organized into three categories, depending on how they combine the feature selection search with the construction of the classification model: filter methods, wrapper methods, and embedded methods. This provides a common taxonomy of feature selection methods, showing for each technique the most prominent advantages and disadvantages, as well as some examples of the most influential techniques.

Filter techniques assess the relevance of features by looking only at the intrinsic properties of the data. In most cases a feature relevance score is calculated, and low scoring features are removed. Afterwards, this subset of features is presented as input to the classification

algorithm. Advantages of filter techniques are that they easily scale to very high-dimensional datasets, they are computationally simple and fast, and they are independent of the classification algorithm. As a result, feature selection needs to be performed only once, and then different classifiers can be evaluated. A common disadvantage of filter methods is that they ignore the interaction with the classifier (the search in the feature subset space is separated from the search in the hypothesis space), and that most proposed techniques are univariate.

It is a well-established paradigm with current systems having many of the characteristics of biological computers and capable of performing a variety of tasks that are difficult to do using conventional techniques. It is a methodology involving adaptive mechanisms and/or an ability to learn that facilitate intelligent behavior in complex and changing environments, such that the system is perceived to possess one or more attributes of reason, such as generalization, discovery, association and abstraction. The objective of this article is to present to the CI and bioinformatics research communities some of the state-of-the-art in CI applications to bioinformatics and motivate research in new trend-setting directions. In this article, we present an overview of the CI techniques in bioinformatics. We will show how CI techniques including neural networks, restricted Boltzmann machine, deep belief network, fuzzy logic, rough sets, evolutionary algorithms (EA), genetic algorithms (GA), swarm intelligence, artificial immune systems and support vector machines, could be successfully employed to tackle various problems such as gene expression clustering and classification, protein sequence classification, gene selection, DNA fragment assembly, multiple sequence alignment, and protein function prediction and its structure. We discuss some representative methods to provide inspiring examples to illustrate how CI can be utilized to address these problems and how bioinformatics data can be characterized by CI.

OUTCOME

To combine all the gene expression using various technique that are available and to avoid over fitting and improve model performance.The performance is predicted in the case of supervised classification and better cluster detection in finalized.

NEURAL COMPUTATION NETWORKS

By,

PROF.M.KRISHNAN,

BHARATHIDASAN UNIVERSITY,

TRICHY.

OBJECTIVE

TO discuss about the model that is used to attack a variety of biological modeling or engineering problems. It helps to learn the complicated non-linear related sets of various property that suits the detection of complicated trends in high-dimension data sets.

THEME

Many neural network computation technique has been proposed by researchers such as marr and albus to combine several aspects of neural models into computer stimulated vocal tract.It shows the unified account for a wide range of experimental observation concerning human speech.It also address the human cognition,movement control,vision,language and memory.This system helps to learn the complicated non-linear related sets of various property that suits the detection of complicated trends in high-dimension data sets. One challenging application for Neural Networks would be to try and actually mimic the behaviour of the system that has inspired their creation as computational algorithms. That is to use Neural Networks in order to simulate important brain functions. The computation attempt to do so, by proposing a Neural Network computational model for simulating visual selective attention, a specific aspect of human attention. The internal operation of the model is based on recent neurophysiologic evidence emphasizing the importance of neural synchronization between different areas of the brain. Synchronization of neuronal activity has been shown to be involved in several fundamental functions in the brain especially in attention. The investigation on this theory have been done by applying a correlation control module comprised by basic integrate and fire model neurons combined with coincidence detector neurons. Thus providing the ability to the model to capture the correlation between spike trains originating from endogenous or internal goals and spike trains generated by the saliency of a stimulus such as in tasks that involve top down attention . The theoretical structure of this model is based on the temporal correlation of neural activity as initially proposed by Niebur and Koch . More specifically; visual stimuli are represented by the rate and temporal coding of spiking neurons. The rate is mainly based on the saliency of each

stimuli while the temporal correlation of neural activity plays a critical role in a later stage of processing were neural activity passes through the correlation control system and based on the correlation, the corresponding neural activity is either enhanced or suppressed. In this way, attended stimulus will cause an increase in the synchronization as well as additional reinforcement of the corresponding neural activity and therefore it will "win" a place in working memory. It is successfully tested the model by simulating behavioural data from the "attentional blink". we suggest that a correlation control module responsible for comparing temporal patterns arising from top-down information and spike trains initiated by the characteristics of each incoming stimuli could be applied in the proposed computational model. If we extend this assumption based on relevant anatomical areas of the brain then the possible existence of such a correlation control module, would more ideally fit somewhere in the area V4 of the visual cortex were synchronization of neural activity. Artificial neural networks have been developed as generalizations of mathematical models of biological nervous systems. In a simplified mathematical model of the neuron, synapses are representedby connectionweights that modulate the effect of the associated input signals, and the nonlinear characteristic exhibited by neurons is represented by a transfer function. There are many transfer functions developed to process the weighted and biased inputs, among which four basic and widely adopted in the field transfer. The neuron impulse is computed as the weighted sum of the input signals, transformed by the transfer function. The learning capability of an artificial neuron is achieved by adjusting the weights in accordance to the chosen learning algorithm.

Most applications of neural networks fall into the following categories:

(1) Prediction: Use the input values to predict some output;

(2) Classification: Use the input values to determine the classification of the input;

(3) Data Association: Similar to classification, but also recognizes data containing errors; and

(4) Data conceptualization: Analyze the inputs so that grouping relationships can be inferred.

In the field of pattern recognition, clustering refers to the process of partitioning a dataset into a finite number of groups according to some similarity measure. Currently, it has become a widely used process in microarray engineering for understanding the functional relationship between groups of genes. Clustering was used, for example, to understand the functional differences in cultured primary epatocytes relative to the intact liver. In another study, clustering techniques were used on gene expression data for tumour and normal colon tissue probed by oligonucleotide arrays . A number of clustering algorithms, including hierarchical clustering, Principle Component Analysis (PCA) , genetic algorithms , and artificial neural networks, have been used to cluster gene expression data. However, in 2002, Yuhui et al.

proposed a new approach to analysis of gene expression data using Associative Clustering Neural Network (ACNN). ACNN dynamically evaluates similarity between any two gene samples through the interactions of a group of gene samples. It exhibits more robust performance than the methods with similarities evaluated by direct distances, which has been tested on the leukemia data set. The experimental results demonstrate that ACNN is superior in dealing with high dimensional data . The performance can be further enhanced when some useful feature selection methodologies are incorporated. It have been used the Self-Organizing Tree Algorithm (SOTA) for analysis of gene expression data coming from DNA array experiments, using an unsupervised neural network. DNA array technologies allow monitoring thousands of genes rapidly and efficiently. One of the interests of these studies is the search for correlated gene expression patterns, and this is usually achieved by clustering them. The result of the algorithm is a hierarchical cluster obtained with the accuracy and robustness of a neural network. SOTA clustering confers several advantages over classical hierarchical clustering methods. The clustering process is performed from top to bottom, i.e. the highest hierarchical levels are resolved before going to the details of the lowest levels. The growing can be stopped at the desired hierarchical level. Moreover, a criterion to stop the growing of the tree, based on the approximate distribution of probability obtained by randomisation of the original data set, is provided. In addition, obtaining average gene expression patterns is a built-in feature of the algorithm. Different neurons defining the different hierarchical levels represent the averages of the gene expression patterns contained in the clusters. A neural network has to be con- figured such that the application of a set of inputs produces the desired set of outputs. Various methods to set the strengths of the connections exist. One way is to set the weights explicitly, using a priori knowledge. Another way is to train the neural network by feeding it teaching patterns and letting it change its weights according to some learning rule. The learning situations in neural networks may be classified into three distinct sorts. These are supervised learning, unsupervised learning, and reinforcement learning. In supervised learning, an input vector is presented at the inputs together with a set of desired responses, one for each node, at the output layer. A forward pass is done and the errors or discrepancies, between the desired and actual response for each node in the output layer, are found. These are then used to determine weight changes in the network according to the prevailing learning rule. The term ‘supervised’ originates from the fact that the desired signals on individual output nodes are provided by an external teacher. The best-known examples of this technique occur in the backpropagation algorithm, the delta rule, and perceptron rule. In unsupervised learning (or self-organization) an output unit is trained to respond to clusters of patterns within the input. In this paradigm the system is

supposed to discover statistically salient features of the input population. Unlike the supervised learning paradigm, there is no a priori set of categories into which the patterns are to be classified; rather the system must develop its own representation of the input stimuli. Reinforcement learning is learning what to do–how to map situations to actions–so as to maximize a numerical reward signal. The learner is not told which actions to take, as in most forms of Machine Learning (ML), but instead must discover which actions yield the most reward by trying them. In the most interesting and challenging cases, actions may affect not only the immediate reward, but also the next situation and, through that, all subsequent rewards. These two characteristics, trial-and-error search and delayed reward are the two most important distinguishing features of reinforcement learning.

BACTERIAL COMPUTING

"A single *E. coli*bacterium will replicate itself approximately every 30 minutes. Growing a bacterium overnight results in approximately a billion identical, independent biologicalprocessors.DNA is composed of four nucleotides, represented by A, T, G, and C.
In double-stranded DNA, an A on one strand always pairs with a T on the other strand, and similarly, a G on one strand always pairs with a C on the other. These A-T and G-C interactions are called Watson-Crick base pairing.

HURDLES IN CREATING A BACTERIAL COMPUTER

Designing and constructing a bacterial computer to solve any math problem presents several distinct challenges.

- First, how will the components of the problem be encoded into the DNA of a bacterium?
- Second, how will the information be manipulated (a necessary component for computation)?
- Finally, how will the results of the computation be "displayed"?

This computing technique addressed each of these challenges to solve the HPP. The technique used in this design used living *E. Coli* to find the solution to the HPP, Many other iGEM projects have components of computation, simulation, and mathematical modeling. The results have application in medicine, the environment, and biofuels, as well as other benefits from data analysis and model building, to the testing of experimental designs." The iGEM community provides a supportive environment for teams wishing to engage in research. The applications of synthetic biology are widely varied, and projects can be tailored to individual interests. The projects carried out by team have a deliberate focus on

mathematics by applying biology to solve math problems, but many other iGEM projects have components of computation, simulation, and mathematical modeling. The results have application in medicine, the environment, and biofuels, as well as other benefits from data analysis and model building, to the testing of experimental designs.

The capability to establish adaptive relationships with the environment is an essential characteristic of living cells. Both bacterial computing and bacterial intelligence are two general traits manifested along adaptive behaviors that respond to surrounding environmental conditions. These two traits have generated a variety of theoretical and applied approaches. Since the different systems of bacterial signaling and the different ways of genetic change are better known and more carefully explored, the whole adaptive possibilities of bacteria may be studied under new angles. For instance, there appear instances of molecular "learning" along the mechanisms of evolution. More in concrete, and looking specifically at the time dimension, the bacterial mechanisms of learning and evolution appear as two different and related mechanisms for adaptation to the environment; in somatic time the former and in evolutionary time the latter. In the present chapter it will be reviewed the possible application of both kinds of mechanisms to prokaryotic molecular computing schemes as well as to the solution of real world problems.Adaptive behavior in bacteria depends on organized networks of proteins governing molecular processes within the cellular system. Bacteria are able to explore the environment within which they develop by utilizing the motility of their flagellar system as well as a biochemical navigation system that samples the environmental conditions surrounding the cell . In this, we described how in some important aspects proteins can be considered as processing elements or McCulloch-Pitts artificial neurons that transfer and process information from the bacterium's membrane surface to the flagellar motor. Every time a neuron receives a set of input signals, performs the weighted sum (with the weights associated with each line) obtaining a net_i value, finally deciding its state or output with a threshold function In general, the molecular systems involved in bacterial signaling (and in *M. tuberculosis*) are extremely diverse, ranging from very simple transcription regulators to the multi-component, multi-pathway signaling cascades that regulate crucial stages of the cell cycle, such as sporulation, biofilm formation, dormancy, pathogenesis, etc.

OUTCOME

Adaptive relationships with the environment is established with the living cells,so that each neuron reeives a set of valid input and a better computing is done.

COMPUTATIONAL HEALTHINFORMATICS

BY

DR.N.JEYAKUMAR

BHARATHIAR UNIVERSITY

COIMBATORE

OBJECTIVES:

- Distinguish the various types of healthcare information, including data, knowledge, sources, and importance of technology standards.
- Consumer Health Informatics (CHI) versus Medical Informatics (MI).
- Analyze obstacles and success factors for implementing and integrating information and decision technologies in healthcare.
- Define ethics and computer ethics Health Informatic learning objectives.

THEME:

Computational health informatics is the Informatics part of health informatics. Health informatics is defines as knowledge-based data which is information to make critical decision. We study computational mechanisms for enabling more efficient and easy-to-use healthcare delivery, and we study large-scale data analytics problems in healthcare. The computational health informatics field arises from the ready availability techniques like SNP genotyping. Which lead to a characterization of an individual's genetic profile. By combining this with environmental factors and disease outcomes, as well as array expression experiments, we hope to uncover the complex causal factors behind disease. To do this we make use of machine learning and statistical techniques.

The health domain provides an extremely wide variety of problems that can be tackled using computational techniques, and computer scientists are attempting to make a difference in medicine by studying the underlying principles of computer science that will allow for meaningful algorithms and systems to be developed. Thus, computer scientists working in computational health informatics and health scientists working in medical health informatics combine to develop the next generation of healthcare technologies.

Computational techniques:

- Artificial intelligence
- Algorithms (for architectures ranging from single CPU to massively parallel machines),
- Programming
- Object-oriented system design
- Databases, information retrieval
- Computer graphics and visualization
- Data mining, information extraction

Probability, statistics and decision science:

- Theory of probability
- Statistical inference
- Cost/risk-benefit analysis
- Probabilistic analysis, stochastic modeling
- Decision theory, statistical data analysis
- Probabilistic networks, pattern classification
- Statistical learning and modeling
- Statistical data mining

Applied mathematics:

- Graph theory, differential equations, optimization theory, wavelets, group theory.

Electrical engineering methods:

- Signal and image processing.

Domain knowledge:

- Art and cultural heritage, biology, chemistry, engineering, medicine, the World Wide Web.

User sciences:

- Design, human-computer interaction, evaluation.

Cyberinfrastructure for informatics:

- Search engines, digital repositories, storage.

Scientometrics, bibliometrics and economics:

- Science and policy evaluation, data mining and information extraction, knowledge discovery.

Social sciences:

- Social network analysis and metrics.

Big data is defined as high-volume, high-velocity, and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision-making. Healthcare data certainly fits the definition of big data. Large amount healthcare data is produced continually and store in different databases. With the wide adoption of electronic health records that has increased the amount of data available exponentially. Nevertheless, the healthcare providers has been slow to leverage the vast amount of data to improve heath care system or use data to improve efficiency to reduce overall cost of healthcare.

Healthcare use

It is estimated that in the US healthcare spending approximately, $75B to $265B is lost each year to healthcare fraud.

With the amount of healthcare fraud, the importance of identifying fraud and abuse in healthcare cannot be ignored; so healthcare providers must develop automated systems to identify fraud, waste and abuse to reduce its harmful impact on their business.

Algorithms

Develop statistical analysis, visualization, and machine learning tools to statistically analyze and develop predictive models for healthcare payment data and possibly detect irregularities and prevent healthcare payment fraud.

Dataset

The Healthcare dataset: Center for Medicare and Medicaid Services (CMS) , released in the dataset into the public domain known as “Medicare Part-B in 2014”. The dataset includes a set of records documenting about transactions between over 900,000 medical providers and CMS.

Possible Tools

Big Data:

- Apache Hadoop or Apache Spark, Apache HBase, Apache Mahout, apache Lucene/Solr, MLlib -Machine Learning Library for Apache Spark's

Visualization:

- D3 Visualization, Tableau visualization.

Development:

- Java, Python, Scala and JavaScript, JQuery

Here are 5 healthcare data solutions of Big Data and Hadoop:

1. Hadoop technology in Cancer Treatments and Genomics

Industry reports indicate that, there are about 3 billion base pairs that constitute the human DNA and it is necessary for such large amounts of data to be organized in an effective manner if we have to fight cancer.

The biggest reason why cancer has not been cured yet is because of the fact that cancer mutates in different patterns and reacts in different ways based on the genetic makeup of an individual. Hence, oncology researchers have come up with a solution that in order to cure cancer, patients will need to be given personalized treatment based on the type of cancer the individual patient's genetics make up. Leveraging Hadoop technology will offer great support for parallelization and help in mapping the 3 billion DNA base pairs using MapReduce programs.

The goal of using Hadoop in Healthcare, is to collect and analyze data that can do everything from assess public health trends in a region of millions of people to pinpoint treatment options for one cancer patient.

2. Hadoop technology in Monitoring Patient Vitals

There are several hospitals across the world that use Hadoop to help the hospital staff work efficiently with Big Data. Without Hadoop, most patient care systems could not even imagine working with unstructured data for analysis.

Children's Healthcare of Atlanta treats over 6,200 children in their ICU units. On average, the duration of stay in Pediatric ICU varies from a month to a year. Children's Healthcare of Atlanta used a sensor beside the bed that helps them continuously track patient signs such as blood pressure, heartbeat and the respiratory rate. These sensors produce large chunks of data, which using legacy systems cannot be stored for more than 3 days for analysis. The main motive of Children's Healthcare of Atlanta was to store and analyze the vital signs. If there is any change in pattern, then the hospital wanted an alert to be generated to a team of

doctors and assistants. All this was successfully achieved using Hadoop ecosystem components - Hive, Flume, Sqoop, Spark, and Impala.

3. Hadoop technology in the Hospital Network

A Cleveland Clinic spinoff company known as Explorys is making use of Big Data in healthcare to provide the best clinical support, reduce the cost of care measurement and manage the population of at-risk patients. Explorys has reportedly built the largest database in the healthcare industry with over a hundred billion data points all thanks to Hadoop.

Explorys uses Hadoop technology to help their medical experts analyze data bombardments in real time from diverse sources such as financial data, payroll data, and electronic health records. The analytics tool developed by Explorys is used for data mining so that it helps clinicians determine the deviations among patients and the effects treatments have on their health. These insights help the medical practitioners and health care providers find out the best treatment plans for a set of patient populations or for an individual patient.

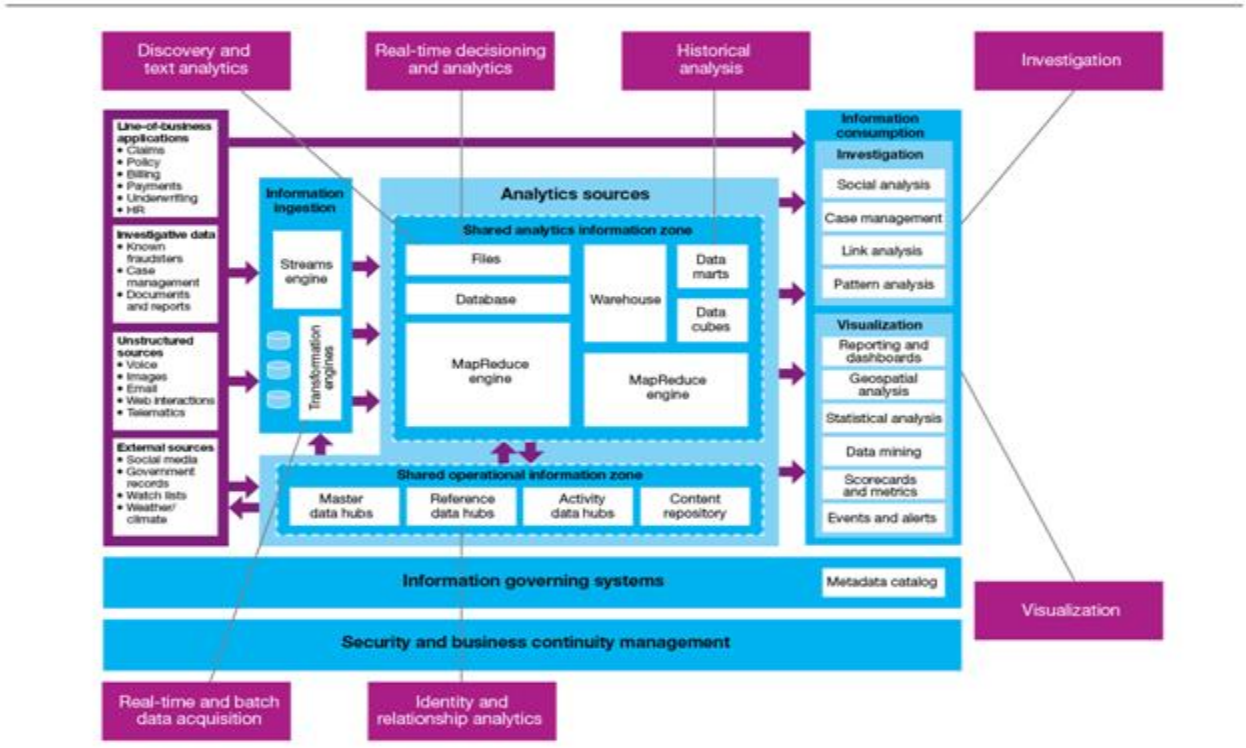
4. Hadoop technology in Healthcare Intelligence

Using Hadoop technology in Healthcare Intelligence applications helps hospitals, payers and healthcare agencies increase their competitive advantages by devising smart business solutions.

To gather desired age, insurance companies will have to process huge data sets to extract meaningful information such asmedicines, diseases, symptoms, opinions, geographic region detail etc. In this scenario, using Hadoop's Pig, Hive and MapReduce is the best solution to process such large datasets.

5. Hadoop technology in Fraud Prevention and Detection

Big Data Analytics helps healthcare insurance companies find different ways to identify and prevent fraud at an early stage. Using Hadoop technology, insurance companies have been successful in developing predictive models to identify fraudsters by making use of real-time and historical data of medical claims, weather data, wages, voice recordings, demographics, cost of attorneys and call center notes. Hadoop's capability to store large unstructured data sets in NoSQL databases and using MapReduce to analyze this data helps in the analysis and detection of patterns in the field of Fraud Detection.



Learning Outcomes:

- Exhibit knowledge in the underlying biological phenomena related to bioinformatics and computational biology.
- Evaluate statistical analyses which can be used to solve bioinformatics and computational biology problems.
- Demonstrate scholarly oral and written presentations
- Adhere to the professional and legal conduct standards of the field of clinical informatics.

HEALTH INFORMATICS TOOL

BY

DR.N.JEYAKUMAR,

BHARATHIAR UNIVERSITY,

COIMBATORE.

OBJECTIVES:

- Apply bioinformatics methods and tools related to genomics, proteomics, biology, and physiology in an academic setting.
- Adhere to the professional and legal conduct standards of the field of bioinformatics and computational biology.
- Produce solutions that address academic or industrial needs using bioinformatics and computational biology tools and knowledge.

THEME:

- Factors that must be taken into consideration when designing bioinformatics tools, software and programmes are:
- The end user (the biologist) may not be a frequent user of computer technology
- These software tools must be made available over the internet given the global distribution of the scientific research community

Major categories of Bioinformatics Tools :

There are both standard and customized products to meet the requirements of particular projects. There are data-mining software that retrieve data from genomic sequence databases and also visualization tools to analyze and retrieve information from proteomic databases. These can be classified as homology and similarity tools, protein functional analysis tools, sequence analysis tools and miscellaneous tools.

Homology and Similarity Tools:

Homologous sequences are sequences that are related by divergence from a common ancestor. Thus the degree of similarity between two sequences can be measured while their homology is a case of being either true of false. This set of tools can be used to identify

similarities between novel query sequences of unknown structure and function and database sequences whose structure and function have been elucidated.

Protein Function Analysis:

This group of programs allow you to compare your protein sequence to the secondary protein databases that contain information on motifs, signatures and protein domains. Highly significant hits against these different pattern databases allow you to approximate the biochemical function of your query protein.

Structural Analysis:

This set of tools allow you to compare structures with the known structure databases. The function of a protein is more directly a consequence of its structure rather than its sequence with structural homologs tending to share functions. The determination of a protein's 2D/3D structure is crucial in the study of its function.

Sequence Analysis:

This set of tools allows you to carry out further, more detailed analysis on your query sequence including evolutionary analysis, identification of mutations, hydropathy regions, CpG islands and compositional biases. The identification of these and other biological properties are all clues that aid the search to elucidate the specific function of your sequence.

Some examples of Bioinformatics Tools:

BLAST:

BLAST (Basic Local Alignment Search Tool) comes under the category of homology and similarity tools. It is a set of search programs designed for the Windows platform and is used to perform fast similarity searches regardless of whether the query is for protein or DNA.

Comparison of nucleotide sequences in a database can be performed. Also a protein database can be searched to find a match against the queried protein sequence. NCBI has also introduced the new queuing system to BLAST (Q BLAST) that allows users to retrieve results at their convenience and format their results multiple times with different formatting options.

- Depending on the type of sequences to compare, there are different programs:
- blastp compares an amino acid query sequence against a protein sequence database
- blastn compares a nucleotide query sequence against a nucleotide sequence database

- blastx compares a nucleotide query sequence translated in all reading frames against a protein sequence database
- tblastn compares a protein query sequence against a nucleotide sequence database dynamically translated in all reading frames
- tblastx compares the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database.

FASTA:

FAST homology search A ll sequences .An alignment program for protein sequences created by Pearsin and Lipman in 1988. The program is one of the many heuristic algorithms proposed to speed up sequence comparison. The basic idea is to add a fast prescreen step to locate the highly matching segments between two sequences, and then extend these matching segments to local alignments using more rigorous algorithms such as Smith-Waterman.

EMBOSS:

EMBOSS (European Molecular Biology Open Software Suite) is a software-analysis package. It can work with data in a range of formats and also retrieve sequence data transparently from the Web. Extensive libraries are also provided with this package, allowing other scientists to release their software as open source. It provides a set of sequence-analysis programs, and also supports all UNIX platforms.

Clustalw:

It is a fully automated sequence alignment tool for DNA and protein sequences. It returns the best match over a total length of input sequences, be it a protein or a nucleic acid.

RasMol:

It is a powerful research tool to display the structure of DNA, proteins, and smaller molecules. Protein Explorer, a derivative of RasMol, is an easier to use program.

PROSPECT:

PROSPECT (PROtein Structure Prediction and Evaluation Computer ToolKit) is a protein-structure prediction system that employs a computational technique called protein threading to construct a protein's 3-D model.

PatternHunter :

PatternHunter, based on Java, can identify all approximate repeats in a complete genome in a short time using little memory on a desktop computer. Its features are its advanced patented algorithm and data structures, and the java language used to create it.

COPIA :

COPIA (COnsensus Pattern Identification and Analysis) is a protein structure analysis tool for discovering motifs in a family of protein sequences. Such motifs can be then used to determine membership to the family for new protein sequences, predict secondary and tertiary structure and function of proteins and study evolution history of the sequences.

Application of Programmes in Bioinformatics:

JAVA in Bioinformatics:

Java is emerging as a key player in bioinformatics. Physiome Sciences' computer-based biological simulation technologies and Bioinformatics Solutions' PatternHunter are two examples of the growing adoption of Java in bioinformatics.

Perl in Bioinformatics:

String manipulation, regular expression matching, file parsing, data format interconversion etc are the common text-processing tasks performed in bioinformatics. Perl excels in such tasks and is being used by many developers.

Bioinformatics Projects:

BioJava:

The BioJava Project is dedicated to providing Java tools for processing biological data which includes objects for manipulating sequences, dynamic programming, file parsers, simple statistical routines, etc.

BioPerl:

The BioPerl project is an international association of developers of Perl tools for bioinformatics and provides an online resource for modules, scripts and web links for developers of Perl-based software.

BioXML:

A part of the BioPerl project, this is a resource to gather XML documentation, DTDs and XML aware tools for biology in one location.

Biocorba:

CORBA is a framework for interlanguage support, and the biocorba project is currently implementing a CORBA interface for bioperl. With biocorba, objects written within bioperl will be able to communicate with objects written in biopython and biojava.

Ensembl :

Ensembl is an ambitious automated-genome-annotation project at EBI. Much of Ensembl\'s code is based on bioperl, and Ensembl developers, in turn, have contributed significant pieces of code to bioperl. In particular, the bioperl code for automated sequence annotation has been largely contributed by Ensembl developers.

bioperl-db:

Bioperl-db is a relatively new project intended to transfer some of Ensembl's capability of integrating bioperl syntax with a standalone Mysql database to the bioperl code-base.. The object data such as sequences, their features, and annotations can be easily loaded into the databases, as in $loader->store($newid,$seqobj) Similarly one can query the database in a variety of ways and retrieve arrays of Seq objects. See biodatabases.pod, Bio::DB::SQL::SeqAdaptor, Bio::DB::SQL::QueryConstraint, and Bio::DB::SQL::BioQuery for examples.

Biopython and biojava:

Biopython and biojava are open source projects with very similar goals to bioperl. However their code is implemented in python and java, respectively. With the development of interface objects and biocorba, it is possible to write java or python objects which can be accessed by a bioperl script, or to call bioperl objects from java or python code.

Biopython

The Biopython Project is freely available Python tools for computational molecular biology. Python is an object oriented, interpreted, flexible language that is becoming increasingly popular for scientific computing. Python is easy to learn, has a very clear syntax and can easily be extended with modules written in C, C++ or FORTRAN.

Biopython features include parsers for various Bioinformatics file formats, access to online services, interfaces to common and not-so-common programs, a standard sequence class, various clustering modules, a KD tree data structure etc. and even documentation. Basically, we

just like to program in Python and want to make it as easy as possible to use Python for bioinformatics by creating high-quality, reusable modules and scripts.

Installing Biopython

Biopython runs on many platforms (Windows, Mac, and on the various flavors of Linux and Unix). For Windows we provide pre-compiled click-and-run installers, while for Unix and other operating systems you must install from source as described in the included README file. This is usually as simple as the standard commands:

- python setup.py build
- python setup.py test
- sudo python setup.py install

Working with sequences

Disputably , the central object in bioinformatics is the sequence. Thus, we'll start with a quick introduction to the Biopython mechanisms for dealing with sequences, the Seq object.Most of the time when we think about sequences we have in my mind a string of letters like 'AGTACACTGGT'. You can create such Seq object with this sequence as follows - the ">>>" represents the Python prompt followed by what you would type in:

```
>>> from Bio.Seq import Seq
>>> my_seq = Seq("AGTACACTGGT")
>>> my_seq
Seq('AGTACACTGGT', Alphabet())
>>> print(my_seq)
AGTACACTGGT
>>> my_seq.alphabet
Alphabet()
```

What we have here is a sequence object with a *generic* alphabet - reflecting the fact we have *not* specified if this is a DNA or protein sequence.In addition to having an alphabet, the Seq object differs from the Python string in the methods it supports. You can't do this with a plain string:

```
>>> my_seq
Seq('AGTACACTGGT', Alphabet())
>>> my_seq.complement()
Seq('TCATGTGACCA', Alphabet())
>>> my_seq.reverse_complement()
Seq('ACCAGTGTACT', Alphabet())
```

The next most important class is the SeqRecord or Sequence Record. This holds a sequence with additional annotation including an identifier, name and description. TheBio.SeqIO module for reading and writing sequence file formats works with SeqRecord objects.

Simple FASTA parsing example

The lady slipper orchids FASTA file ls_orchid.fasta in text editor starts like this:

```
>gi|2765658|emb|Z78533.1|CIZ78533 C.irapeanum 5.8S rRNA gene and ITS1 and
ITS2 DNA
CGTAACAAGGTTTCCGTAGGTGAACCTGCGGAAGGATCATTGATGAGACCG
TGGAATAAACGATCGAGTG
AATCCGGAGGACCGGTGTACTCAGCTCACCGGGGGCATTGCTCCCGTGGTG
ACCCTGATTTGTTGTTGGG
...
```

It contains 94 records, each has a line starting with “>” (greater-than symbol) followed by the sequence on one or more lines. Now in Python:

```
from Bio import SeqIO
for seq_record in SeqIO.parse("ls_orchid.fasta", "fasta"):
   print(seq_record.id)
   print(repr(seq_record.seq))
   print(len(seq_record))
```

Output will be like this

```
gi|2765658|emb|Z78533.1|CIZ78533
Seq('CGTAACAAGGTTTCCGTAGGTGAACCTGCGGAAGGATCATTGATGAGA
CCGTGG...CGC', SingleLetterAlphabet())
740
...
gi|2765564|emb|Z78439.1|PBZ78439
Seq('CATTGTTGAGATCACATAATAATTGATCGAGTTAATCTGGAGGATCTGT
TTACT...GCC', SingleLetterAlphabet())
592
```

Simple GenBank parsing example

Now let's load the GenBank file ls_orchid.gbk instead - notice that the code to do this is almost identical to the snippet used above for the FASTA file - the only difference is we change the filename and the format string:

```
from Bio import SeqIO
for seq_record in SeqIO.parse("ls_orchid.gbk", "genbank"):
    print(seq_record.id)
    print(repr(seq_record.seq))
    print(len(seq_record))
```

This should give:

```
Z78533.1
Seq('CGTAACAAGGTTTCCGTAGGTGAACCTGCGGAAGGATCATTGATG
AGACCGTGG...CGC', IUPACAmbiguousDNA())
740
...
Z78439.1
Seq('CATTGTTGAGATCACATAATAATTGATCGAGTTAATCTGGAGGAT
CTGTTTACT...GCC', IUPACAmbiguousDNA())
592
```

Learning Outcomes:

- Show competence in use of information technology tools
- Integrate knowledge in a specialized cognate area in order to form a foundation for future research in clinical informatics.
- Construct and deliver educational content in clinical informatics to the standards of the department and field.
- Conduct independent research which contributes new knowledge to the field of clinical informatics.

RECENT RESEARCH AND APPLICATIONS ON HEALTH INFORMATICS

BY

Prof.D.VELMURUGAN
UNIVERSITY OF MADRAS
CHENNAI

The Bioinformatics pathway focuses on three areas of research:

- Bioinformatics and computational biology
- Genetics and genomics
- Systems biology

1. Bioinformatics and computational biology

The fields of bioinformatics and computational biology at UCSF aim to investigate questions about biological composition, structure, function, and evolution of molecules, cells, tissues, and organisms using mathematics, informatics, statistics, and computer science.

Because these approaches allow large-scale and quantitative analyses of biological phenomena and data obtained from many disciplines, they can ask questions and achieve unique insights not imaginable before the genomic era.

Both bioinformatics and computational biology are frequently integrated in faculty laboratories, often with experimental studies as well, with bioinformatics emphasizing informatics and statistics, while computational biology emphasizes development of theoretical methods, mathematical modeling, and computational simulation techniques to answer these questions.

Examples of bioinformatics studies include analysis and integration of -omics data, prediction of protein function from sequence and structural information, and cheminformatics comparisons of protein ligands to identify off-target effects of drugs. Examples in computational biology include simulation of protein motion and folding and how proteins interact with each other.

Faculty members working in these areas include:

- Patricia Babbitt
- Sourav Bandyopadhyay

- Steven Brenner
- Thomas Ferrin
- Matthew Jacobson
- Ajay Jain
- Tanja Kortemme
- Katherine Pollard
- Andrej Sali
- Brian Shoichet

2. Genetics and genomics

Genetics is the study of DNA-based inheritance and variation of individuals, while genomics is the study of the structure and function of the genome. Both apply bioinformatics and computational techniques using data generated from methods such as DNA and RNA sequencing, microarrays, proteomics, and electron microscopy, or optical methods for nucleic acid structure determination.

Availability of these and many other new technologies, such as those that can conduct deep sequencing or sequencing of entire microbial communities, is generating massive amounts of data faster than informatics and computational methods can be developed to manage and query them. This opens opportunities for genetics and genomics scientists to develop and apply new cutting-edge technologies to analyze these data.

Faculty members working in genomics and genetics include:

- Nadav Ahituv
- Sourav Bandyopadhyay
- Bruce Conklin
- Michael Fischbach
- Kathy Giacomini
- Ryan Hernandez

3. Systems biology

Systems biology seeks to understand how cells, tissues, and organisms function from the perspective of the system as a whole. Computational systems biologists use mathematical modeling, simulation, and statistical analysis to gain a fundamental understanding of biological

processes such as maintenance of homeostasis, minimal requirements for function, system response to environmental perturbation, predicting response to system stressors, and dissecting protein and nucleic acid networks.

Biomedical research is being revolutionized by new technologies for generating high throughput data. For example, the mRNA counts contained in gene microarrays provide a global view of cellular activity by simultaneously recording the expression levels of thousands of genes. Similarly, new methods for measuring the expression of proteins in cells and tissues and mapping protein-protein interactions are providing rich sources of information for learning about disease mechanisms.

Application of BioInformatics:

In broad spectrum application of bio informatics is mainly used in the field of Medicine, Microbial Genome Applications and Agriculture.

1. Medicine

A.Drug Discovery

The idea of using X ray Crystallography in drug discovery emerged more than 30 years ago,when the first 3 dimensional structure of protein was determined. Within a decade. A radical change in drug design had begun, incorporating the knowledge of 3 dimensional structures of target protein into design process. Protein structure can influence drug discovery at every stage in design process. Classicaly, it is used in lead optimization, a process that uses structure to guide the chemical modification of a lead molecule to give an optimized fit in terms of shape, hydrogen bonds and other non – covalent interactions with the target.

B.Personal Medicine:

Personalized medicine is a medical model that proposes the customization of healthcare, with all decisions and practices being tailored to the individual patient by use of genetic or other information. Practical application outside of long established considerations like a patient's family history, social circumstances, environment and behaviors are very limited so far and practically no progress has been made in the last decade. Personalized medicine research attempts to identify individual solutions based on the susceptibility profile of each individual. It is hoped that these fields will enable new approaches to diagnosis, drug development and individualized therapy.

C.Preventive Medicine

Preventive medicine or preventive care consists of measures taken to prevent diseases, rather than curing them or treating their symptoms. This contrasts in method with curative and palliative medicine and in scope with public health methods

D.Gene therapy

Gene therapy is a novel form of drug delivery that enlists the synthetic machinery of the patient's cell to produce a therapeutic agent. It involves the efficient introduction of functional gene into the appropriate cells of the patient in order to produce sufficient amount of protein encoded by transferred gene so as to precisely and permanently correct the disorder. Strategies of gene therapy are following:

- Gene addition
- Removal of harmful gene by antisense nucleotide or ribozymes
- Control of gene expression

2. Microbial Genome Applications

In the field of Microbial Genome Applications, applications of bioinformatics are used for following areas

A.Waste Cleanup

In bioinformatics bacteria and microbes are identified which are helpful in cleaning waste. Deinococcus radiodurans is listed in the guinness book of world records as "the world's toughest bacterium". This bacterium has the ability to repair damaged dna and small fragments from chromosomes by isolating damage segments in a concentrated area.

B.Climate Change

Climate changes is caused by factors that include oceanic process, variations in solar radiation received by Earth, plate tectonics and volcanic eruptions and human induced alterations of the natural world.By studying microorganisms genome scientists can begin to understand these microbes at a very fundamental level and isolated the genes that give them their unique abilities to survive under extreme conditions.

C.Biotechnology

The wide concept of “biotech” encompasses a wide range of procedure for modifying living organisms according to human purposes, going back to domedtication of animal, cultivation of plants and “improvements” to these through breeding programs that employ artificial selection and hybridization. Modern usage also includes genetic engineering as well as tissue culture technologies. Biotechnology has identified organisms and microorganisms which can be very useful in dairy industry and food manufactures. Lactococcus latics is one of the most important microorganism.

D.Alternative Energy

Scientists are studying the genome of the microbe chlorobium tepidum which has an unusual capacity for generating energy from light. Chlorobium tepidum is a thermophilic gram negative green sulfur bacterium isolated from a hot spring in new Zealand in which it forms a dense mat. The bacterium carries out photosynthesis in ways that are different from plants and other bacteria. **Agriculture**

A.Crop Improvement

Comparative genetics of the plant genomes has shown that the organization of their genes has remained more conserved over evolutionary time than was previously believed.These findings suggest that information obtained from the model crop systems can be used to suggest improvements to other food crops.

B.Insect Resistance

Genes from Bacillus thuringiensis that can control a number of serious pests have been successfully transferred to cotton, maize and potatoes.This new ability of the plants to resist insect attack means that the amount of insecticides being used can be reduced and hence the nutritional quality of the crop is increased.

C.Improve nutritional Quality

Scientists have recently succeeded in transferring genes into rice to increase levels of Vitamin A, iron and other micronutrients. This work could have a profound impact in reducing occurrences of blindness and anaemia caused by deficiencies in Vitamin A and iron respectively.

INFORMATICS CERTIFICATIONS

BY

Prof.D.VELMURUGAN
UNIVERSITY OF MADRAS
CHENNAI

CCHIIM (Commission on Certification for Health Informatics and Information Management)

- The Commission on Certification for Health Informatics and Information Management (CCHIIM) manages and sets the strategic direction for the certifications.

AHIMA

Types of certification:

1. HIM Certification
    - Registered Health Information Administrator (RHIA):
    - Registered Health Information Technician (RHIT)
2. Coding Certification
    - Certified Coding Associate (CCA)
    - Certified Coding Specialist (CCS)
    - Certified Coding Specialist-Physician-based (CCS-P)
3. Specialty Certification
    - Certified Health Data Analyst (CHDA)
    - Certified in Healthcare Privacy and Security (CHPS)
    - Certified Documentation Improvement Practitioner (CDIP)
    - Certified Healthcare Technology Specialist (CHTS) Exams

Programmes:

Post Graduate Certificate in Medical Informatics (PGCMI)

Minimum Duration: 6 Months

Maximum Duration: 2 Years

Course Fee: Rs. 10,000

Minimum Age: No bar

Maximum Age: No bar

Eligibility:

Bachelor's Degree in any discipline from any recogniSed University/Institutions.

Courses:

CourseCode	Course Name
MAH-001	Introduction to Medical Informatics
MAH-002	Data Management
MAH-003	Information System –I
MAH-004	Information System –II

Programme overview:

Medical Informatics is the intersection of information science, computer science, and health care. It deals with the resources, devices, and methods required optimizing the acquisition, storage, retrieval, and use of information in health and biomedicine. Health informatics tools include not only computers but also clinical guidelines, formal medical terminologies, and information and communication systems.

Institutes Offering Courses

1. International Institute of Health Management Research (IIHMR) Plot No. 3, HAF Pocket, Sector 18A, Phase-II Dwarka Near Veer Awas/Kargil Appartment Sector- 12 Metro Station New Delhi- 110075 Website: www.iihmrdelhi.org	2. Indira Gandhi National Open University(IGNOU) Maidan Garhi, New Delhi-110068 Website: www.ignou.ac.in
3. Indian Institute of Public Health, Hyderabad Plot # 1, A N V Arcade Amar Co-operative Society, Kavuri Hills Madhapur Hyderabad- 500033 Website: www.phfi.org	4. Osmania University Shivam Road, Prashanti Nagar, Nallakunta Hyderabad, Andhra Pradesh- 500007 Website: www.osmania.ac.in
5. Foundation of Healthcare Technologies Society B/403 Somdatt Chamber-I, 5 Bhikaji Cama Place New Delhi-110066 Website: http://fhts.ac.in/	6. Medavarsity online Life Sciences Building, Apollo Hospitals Complex, Film Nagar, Hyderabad, Andhra Pradesh- 500096 Website: www.medvarsity.com
7. Indraprastha Apollo Hospital Saritha Vihar,Mathura Road New Delhi- 110076 Website: http://www.medvarsity.com	8. Bioinformatics Institute of India C-56A/28, Sector-62 Noida Uttar Pradesh- 201309 Website: www.bii.in
9. eHCF School of Medical Informatics B-5, 2nd Floor, Street 13 Madhu Vihar, IP Extension New Delhi-110092 Website: http://www.ehcfsmi.edu.in/	10. Department of Health Information Management School of Allied Health Sciences, Manipal University Manipal, Karnataka- 576104 Website: http://www.manipal.edu

Medical and Healthcare Informatics (6 Months)

Program Areas

- **Module I-** Introduction to Medical Informatics
- **Module II-** Healthcare Organization and Management
- **Module III-** Telemedicine
- **Module IV-** Hospital and Clinical Information Systems
- **Module V-** Biomedical Engineering

Job Opportunities

After the completion of the program the students would be eligible for the following types of jobs:

- Medical Consultants
- IT Health Advisor
- Healthcare Executive
- Healthcare Manager
- Technical Consultant
- IT Medical service Executive
- Clinical Information manager

Eligibility (Medical graduates, Science graduates)

The eligibility for the training program is graduation in any discipline. Highly interested participants in final year can also apply.The course is intended for:

- Doctors

- Clinicians
- Health Professionals
- Paramedical sciences Professionals
- Nurses
- IT Professionals
- Health Care Planners
- Technicians

Program Fee

The Program Participation Fee for this program is Rs 5,500 /-(for participants based in India / USD 500 (for overseas participants).

Examination Fee

Every student has to pay Rs.500/- per module as examination fees.

Kongunadu

College of Engineering & Technology

Namakkal - Trichy Main Road, Tholurpatti (P.O.), Thottiyam (Tk), Trichy (Dt.) - 621 215.

(AICTE Approved, Affiliated to Anna University, Chennai & ISO 9001 : 2008 Certified Institution)

VISION:

To become Internationally Renowned Institution in Technical Education, Research and Development by Transforming the students into Competent Professionals with Leadership skills and Ethical values.

MISSION:

- Providing the best resources and Infrastructure.
- Creating Learner - Centric Environment and Continuous Learning.
- Promoting Effective Links with Intellectuals and Industries.
- Enriching Employability and Entrepreneurial skills.
- Adapting to Changes for Sustainable Development.

QUALITY POLICY:

To Strive Continuously for Producing the best result in terms of Knowledge, Self-Discipline and application of the Knowledge acquired.