Volume 10, Number 2 *2006*

Allied Academies
International Conference

Reno, Nevada
October 19-21, 2006

Academy of Information and Management Sciences

PROCEEDINGS

Volume 10, Number 2 *2006*

Table of Contents

A COMPARISON OF NEURAL NETWORK AND TIME-SERIES FORECASTS FOR STOCK MARKET INDEX: SOME KOREAN EVIDENCE

Kyungjoo Lee, Cheju, National University
mrlove@cheju.ac.kr.
Sehwan Yoo, University of Maryland Eastern
syoo@umes.edu
John Jongdae Jin, University of Maryland-Eastern Shore
jongdaejin@hotmail.com

INTRODUCTION

The ability to forecast the capital market price index is critical to individual investors and financial analysts as well. Among many forecasting models for stock prices and market price index, the neural network (NN) model has been gaining its popularity in recent years (E.G., Ansari et. al. (1994), Hamid et. al. (2004), Huang et. al. (2005), Kumar et. al. (2006), Malik et. al. (2006), Stansell et. al. (2004), Trinkle et. al. (2005)). Major reasons for the NN model's popularity in capital market forecast are twofold. First, the NN model is data driven method which learns from sample data and hence does not require any underlying assumptions about the data. Thus, the model is known as a universal functional approximate without severe model misspecification problems due to wrong assumptions (Hornik et. al. (1989)). The model is also outstanding in processing large amount of fuzzy, noisy, and unstructured data. For example, Hutchinson et. al. (1994) examine stock option price data and show that the NN model is computationally less time consuming and more accurate non-parametric forecasting method, especially when the underlying asset pricing dynamics are unknown or when the pricing equation cannot be solved analytically. Second, stock price data are large, highly complex and hard to model because the pricing dynamics are unknown, which suits the NN model.

Korean stock market is considered more volatile than its US counterpart and hence has fuzzier and unstructured price data, which suits the NN model well.

Thus, it is meaningful endeavor to examine how well the NN model performs in forecasting more volatile Korean market data relative to a conventional seasonal autoregressive integrated moving average (SARIMA) model which is one of the most popular forecasting models in capital market studies. The purpose of this study is to compare the ability of the NN model and that of SARIMA model in forecasting Korean Stock Price Index (KOSPI) and its returns. Weekly data of KOSPI are analyzed in this study.

The remainder of this paper is organized as follows. Sample data and methodology are discussed in the next chapter, which is followed by discussions on empirical results. The concluding remarks are presented in the last chapter.

DATA AND METHODOLOGY

Index Data

The data used in this study are KOSPI for closing prices from the Korean Stock Exchange (KSE) data base. The data series span from 4th January 1999 to 29th May 2006, totaling 390 weeks (89 months) of observations. The returns data (Rt) are defined as the continuously compounded returns on the price in the following way: $R_t = \ln (P_t/P_{t-1})$, where Pt is KOSPI in period t.

The data are divided into two sub-periods, one for the estimation and the other for the forecasting. We use four different forecasting periods to examine the potential impact of forecasting horizons on the forecasting accuracy. Forecasting horizons used are 20% (long range), 13% and 8% (mid range), and 6% (short range) of the total number of observations.

2.2 Neural Network Forecasting

In this study, one of the widely used NN model called the back-propagation neural network (BPNN) is used for time series forecasting. BPNN can be trained using the historical data of a time series in order to capture the non-linear characteristics of the specific time series. The model parameters will be adjusted iteratively by a process of minimizing the forecasting errors. Fore time series forecasting, the relationship between output (y_t) and the inputs (y_{t-1} $y_{t-2,\ldots}$, yt_{-p}) can be described by the following mathematical formulae.

$$y_t = a_o + \sum_{j=1}^{q} a_j f(w_{oj} + \sum_{i=1}^{p} w_{ij} y_{t-i}) + e_t \quad (1)$$

Where a_j (j = 0,1,2,…,q) is a bias on the jth unit, and w_{ij} (i = 0,1,2,…,p; j = 0,1,2.,,,q) is the connection weights between layers of the model, f(is the transfer function of the hidden layer, p is the number of input nodes and q is the number of hidden nodes. The BPNN model performs a nonlinear functional mapping from the past observation (y_{t-1} $y_{t-2,\ldots}$ y_{t-p}), to the future value (y_t), i.e.,

$$y_t = \varphi(y_{t-1}\ y_{t-2,\ldots}\ y_{t-p}) + e_t \quad (2)$$

Where w is a vector of all parameter and φ is a function determined by the network structure and connection weights.

Time-series Forecasting

To obtain the KOSPI forecasts from the SARIMA model, we adopted the Box and Jenkins' method which uses the following three-stage approach to select an appropriate model for the purpose of estimating and forecasting a time-series data.

Identification: we used the SARIMA procedure in SAS statistical software to determine plausible models. The SARIMA procedure uses standard diagnostics such as plots of the series, autocorrelation function (ACF), inverse autocorrelation function, and partial autocorrelation function (PACF).

Estimation: Each of the tentative models is fit and the various coefficient estimates are examined. The estimated models are compared using standard criteria such as Akaike Information Criteria and the significance level of coefficients.

Diagnostic checking: SARIMA procedure is used to check if the residuals from the different models are white noise. The procedure uses diagnostics tests such as ACF, PACF, and Ljung-Box Q-statistics for serial correlation.

Applying these steps, SARIMA (110) (12) for the KOSPI price series, and SARIMA (011) (12) for the KOSPI return series are selected as forecasting models.

Measurement of Forecast Accuracy

Forecast error (FE) is determined by subtracting forecasted value from actual value of KOSPI (price or return), and then deflating the difference by the absolute value of actual data as follows:

$$FEt = (At - Ft)/|At|, \quad (3)$$

where At = actual value of KOSPI in period t.
Ft = forecasted value of KOSPI in period t.

The reason for using the absolute value of actual KOSPI as deflator is to correct for negative values.

Accuracy of a forecast model is measured by sign-neutral forecast error metrics such as absolute forecast errors (AFE) and squared forecast errors (SFE).

RESULTS

The SARIMA model provides smaller mean (median) values of Forecasting Errors (FE) than the NN model for every forecasting horizon. The differences are statistically significant (at $\alpha<0.001$) only for mid range (31-50 weeks ahead) forecasting horizons, using both the t-test and the nonparametric Wilcoxon test. With respect to FE for KOSPI returns, the SARIMA model has larger mean and standard deviation than the NN model does.

Overall, these results indicate that the SARIMA model provides more accurate forecasts of KOSPI than the NN model, while NN model does better in KOSPI returns than SARIMA model does.

CONCLUSIONS

The purpose of this study is to compare the forecasting performance of a neural network (NN) model and a time-series (SARIMA) model in Korean Stock Exchange. In particular, we investigate which is the better model between the back-propagation neural network (BPNN) model the SARIMA model in forecasting the Korea Composite Stock Price Index (KOSPI) and its returns. Forecasting performance is measured by the forecast accuracy metrics such as absolute forecasting errors and square forecasting errors of each model.

KOSPI data and its return data over the 390 week (89 month) period extending from January 1999 to May 2006 are analyzed. We find the followings: first, the SARIMA model generally provides more accurate forecasts for the KOSPI than the BPNN model does. This relative superiority of the SARIMA model over the BPNN model is pronounced for the mid-range forecasting horizons. Second, the BPNN model is generally better than the SARIMA model in forecasting the KOSPI returns. However, the difference in forecasting accuracies of the two models is not statistically significant. These results are robust across different measures of forecast accuracy.

REFERENCES: Available upon request.

ENDNOTES

1 - KOSPI data is available since the opening of the Korean Options Exchange for stock price index in July 1997. We exclude two year data (1997-1998) because the Korean stock market had suffered the severe financial crisis called IMF crisis during this period.
2 - Although there is no negative price, returns data could have negative values.
3 - AFE assumes that the user loss function is linear, while SFE assumes quadratic loss function.

MODELING THE ACADEMIC PUBLICATION PIPELINE

Steven E. Moss, Georgia Southern University
smoss@georgiasouthern.edu
Xiaolong Zhang, Georgia Southern University
xzhang@georgiasouthern.edu
Mike Barth, Georgia Southern University
mbarth@georgiasouthern.edu

ABSTRACT

In recent years, many academic institutions have implemented more stringent academic qualification standards by increasing research requirements for faculty. This paper analyzes the impact of changing research requirements in terms of faculty research output. We model the academic research pipeline including review and revision processes as a queue network and study the stationary behavior of the research pipeline. Both a theoretical model of the publication process and a simulation showing numerous combinations of submission strategies and publication requirements are presented. Within this framework, research requirements can be analyzed via probability constraints on the output process, and the research effort that satisfies this constraint is derived. We also study the transitional behavior of the publication queue to understand the convergence towards the stationary solution. Our analyses shows that submission requirements substantially exceed publication requirements if a faculty member is to maintain a required number of publications over any given time interval.

HOLE LOTTA TRADIN' GOIN' ON: A LOOK AT THE EFFECTS OF ONLINE INVESTING ON INVESTORS AND BROKERS

Anna Turri, Sam Houston State University
annamblack@yahoo.com
Balasundram Maniam, Sam Houston State University
maniam@shsu.edu

ABSTRACT

Online trading is one of the latest ways to invest that has gained popularity in recent years as investors become savvier. This study discusses online trading's effects on investors and brokers. It also evaluates what characteristics most online investors share. It then examines how foreign investors are handling trading online in their own countries. Finally, it observes the future of online investing. Overall, with the convenience of online investing, it looks like it will increase in popularity. However, investors will have to weigh the benefits versus the costs and see if online investing is right for them. Also, brokers must change their services to accommodate this new breed of investors.

AN EXPERIMENTAL DESIGN COMPARISON OF FOUR HEURISTIC APPROACHES FOR BATCHING JOBS IN PRINTED CIRCUIT BOARD ASSEMBLY

Susan K. Williams, Northern Arizona University
SK.Williams@NAU.EDU
Michael J. Magazine, University of Cincinnati
Mike.Magazine@uc.edu

ABSTRACT

The goal of the printed circuit board (PCB) job-batching problem is to minimize the total manufacturing time required to process a set of printed circuit board jobs on a pick-and place machine. Specifically, to determine which jobs should be processed with the same setup so that the total number of setups is reduced without increasing the required processing time such that the reduction in setup time is offset. Since PCB manufacturers assemble thousands to millions of boards each year, even modest time savings per job are useful.

In this paper, we perform an experiment designed to compare four heuristic approaches to solve the PCB job-batching problem: cluster analysis, a bin-packing approach, a sequencing genetic algorithm, and a grouping genetic algorithm. We developed these heuristic approaches in previous work. However, based on that work, there was not overall best heuristic. These results show that the cluster analysis and binpacking approaches have fast execution time but do not find optimal solutions while the two genetic algorithms are slower but often find optimal solutions. Results describe the problem characteristics for which each heuristic performs best. Based on the results described here, a user can decide which heuristic is most appropriate based on PCB job characteristics and execution time requirements.

Keywords:

(*Heuristic algorithms for combinatorial optimization; PCB assembly)*

MATHEMATICAL PREDICTIONS OF ORGANIZOLOGY, THE NEW SCIENCE OF ORGANIZATION

Andrei Aleinikov, Defense Language Institute
Andrei.Aleinikov@monterey.army.mil
Ralucca Gera, Naval Postgraduate School
rgera@nps.edu

ABSTRACT

This presentation is the next step in the explication of the fundamental concepts of Organizology, the new science of organization. Organizology is of great importance to any information and management specialists because they deal with knowledge management systems (organizing the training process), data bases (organization) and people (organizations). Since the ultimate mission of any science is to create the science laws, after the introduction of Organizology (Aleinikov, 2004) and its basic units (Aleinikov, 2005), the task was to formulate the fundamental laws of organization. Due to the fact that Organizology is a highly abstract (vs. empirical) science, actually deduced from the previous mathematical breakthroughs (Bartini, 1965, 1966) and made open to public only in the 90's, we continue to explore the mathematical predictions of these breakthroughs and interpret the fields left dormant or unexplored by Bartini himself.

INTRODUCTION

By establishing truth through rigorous deduction from properly chosen axioms and definitions, mathematical theory has proven to be a powerful tool in developing scientific research. Mathematics, not a natural or social science in itself (but sometimes called a formal science) unveils the deeply hidden patterns of nature and society and is extremely vital for making any field of enquiry look and operate as a science—not just an observation or descriptive field.

It may be an unusual case in the history of science, but in our research mathematics is used for the role different from that of just a tool within a science: it is used as a tool of creating new sciences. Organizology and Intensiology are the sciences discovered at the tip of the pen after 20 years plus of rethinking of modern scientific classifications and practices. Mathematics, therefore, brings organization into the field of seemingly accidental discoveries and inventions, and, by unveiling new sciences, serves as a heuristic instrument in the field of scientology.

It is our understanding that many leading research institutions would like to have such an instrument. This is because a new science or field of research not only exposes the never approached horizons and highly profitable trends, but also predicts the otherwise unforeseen dangers of the future world, thus saving humanity. Mathematical precision, more powerful than the language precision (language is often vague and ambiguous), serves as the basis for the language descriptions. The latter can change and may be improved, while the mathematical foundations of the predicted new laws of organization are like the pillars for the new building of the new science.

PHYLOGENETIC TREE RECONSTRUCTION AND ANALYSIS USING DISTANCE AND CHARACTER BASED METHODS

Allam Appa Rao, Andhra University
allamapparao@gmail.com
M. Vijay Kumar, Andhra University
K. Chakraborty, Andhra University

ABSTRACT

Bioinformatics is a new science, which combines DNA sequencing data, computer applications, statistics, and mathematics in the study of the life sciences. It also involves the compilation, handling, analysis, and interpretation of massive DNA sequencing data, which are essential to the discovery of new drugs, vaccines, and cures that save lives. The problem of analyzing nucleotide sequences in genes to estimate the evolutionary relatedness also known as phylogenetic trees is one of the important aspects in Bioinformatics. Phylogenetics Analysis is the Study of ancestral-descendent relationships among species and/or genes. Using phylogeny we can answer questions like "which species is the closest relative to other species and when did Man separate from it". This research presents "A Software Tool for Phylogenetic Tree Reconstruction and Analysis Using Distance and Character Based Methods." with the main task of applying it to diabetes related enzyme ButyrylCholineterase(BChE), which may cause to Diabetics.

INTRODUCTION

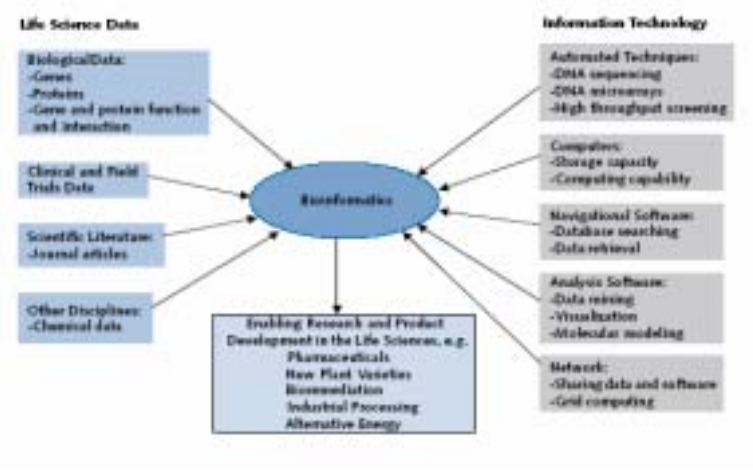
Computational biology and bioinformatics are multidisciplinary fields, involving researchers from different areas of specialty, including (but in no means limited to) statistics, computer science, physics, biochemistry, genetics, molecular biology and mathematics. The goal of these two fields is as follows (NCB, 2006 and BIOINFO, 2006):

Bioinformatics: Typically refers to the field concerned with the collection and storage of biological information. All matters concerned with biological databases are considered bioinformatics.

Computational biology: Refers to the aspect of developing algorithms and statistical models necessary to analyze biological data through the aid of computers.

Fig 1 depicts overview of BioInformatics from National Institute of Health website.

Fig 1: Overview of Bioinformatics



Why is bioinformatics important?

In a field that has been dominated by structural biology for the last 20-30 years, now witnessing a dramatic change of focus towards sequence analysis, spurred on by the advent of the genome projects and the resultant sequence/structure deficit. The central challenge of bioinformatics is the rationalization of the mass of sequence information, with a view not only to deriving more efficient means of data storage, but also to designing more incisive analysis tools. The imperative that drives this analytical process is the need to convert sequence information into biochemical and biophysical knowledge; to decipher the structural, functional and evolutionary clues encoded in the language of biological sequences.

It is clear that mere acquisition of sequences conveys little more about the intricate biology of the systems from which they are derived a company phone directory can reveal about the complexities of the company's business. To extract biological meaning from sequence information is an exacting science. In essence, faced with the task of decoding an unknown language. This language may be decomposed into sentences (proteins), words (motifs), and letters-its alphabet-(amino acids) and the code may be tackled at a variety of these levels. By themselves, the letters have no higher meaning, but their particular combination into words is important. Sometimes the most subtle of changes, a single letter within a word perhaps, can change its meaning (e.g., hog-hag), and hence the meaning of the entire sentence; so it is vital to decipher the code correctly. Consider, for example, the single base change in the human hemoglobin. A chain codon for glutamic acid (GAA) to valine (GUA); in homozygous individuals, this minute difference results in a change from normal healthy state to fatal sickle cell anemia. Ultimately, the aim is to be able to understand the words in a sequence sentence that form a particular protein structure, and perhaps one day to be able to write sentences (design proteins) of our own. Today, application of computational methods allows recognizing words that form characteristic patterns or signatures, but we do not yet understand the intricate syntax required to piece the patterns together and build complete protein structures.

In investigating the meaning of sequences, two distinct analytical themes have emerged: in the first approach, pattern recognition techniques are used to detect similarity between sequences and hence to infer related structures and functions; in the second, ab initio prediction methods are used to deduce 3D structure, and ultimately to infer function, directly from the linear sequence. The development of more powerful pattern recognition and structure prediction techniques will continue

to be dominant themes in Bioinformatics research while the number of experimentally determined protein structures remains small.

PHYLOGENETIC ANALYSIS

Phylogenies is the study of the origin, development and death of a taxon. It is a useful tool in conservation of species. P**hylogeny** is the evolutionary process of an organism from its initial occurrence to a concrete geological time. It includes the evolutionary process and also the organism itself and its descendants. **Phylogenetics** is a research field to study the phylogeny of organisms. It is impossible for a reseacher to study the phylogeny of all extinct and existing organisms. **Phylogenetists** generally study only the origin, development and death of one taxon .

Evolutionary history = Phylogenic relationship

Phylogentic tree construction methods

There are three main classes of Phylogentic methods for constructing Phylogenies from sequence data:

1. Methods directly based on sequences:
   a) Parsimony
      Find tree that requires minimum no. of changes to explain the data.
   b) Maximum likelihood
      Find tree that maximizes likelihood of data.
2. Methods indirectly based on sequences:
   a) Distance matrices (FITCH-MARGOLIASH, NEIGHBOR JOINING method)
      Find tree that accounts for estimated evolutionary distances.

RESEARCH PURPOSE and SCOPE

The purpose of this research is to present an integrated software product for Phylogenetic tree construction from given sequences. This product allows to arrange sequences in tree order. The similar groups in the tree are clustered. User-friendly graphical user interface reduces burden on evaluating the results. This research provides the functional, performance, design and verification of the software to be developed.

The scope of this system for input is limited to text format files only. The input format for the DNA sequence programs is standard: the data have A's, G's, C's and T's (or U's). The first line of the input file contains the number of species and the number of sites. The first 10 characters of that line are the species name. There then follows the base sequence of that species.

Function Requirements
Exact Sequence of Operations

The product takes data in text format sequence as input files and produces the phylogentic tree as output.
The sequence of operations is:

- Validating the data in a file with text format.
- Selecting the algorithm to construct the tree.
- Calculate the distance matrix for sequences.
- Construct the phylogenetic tree.

Relationship of outputs to inputs
Input/Output Sequences

Input	:	file name
Process	:	Phylogentic tree construction
Output	:	tree constructed
Input	:	FASTA sequences
Process	:	distance matrix calculation
Output	:	matrix calculated.

Performance requirements

- The number of terminals to be supported: 1
- The number of simultaneous users to be supported: 1
- Type of information to be handled: input files in text format only.

RESULTS

Using simulated data, comparing four methods of phylogenetic tree estimation: parsimony, maximum likelihood, Fitch-Margoliash, and neighbor joining. For each combination of substitution rates and sequence length, 100 data sets were generated for each of 50 trees, for a total of 5,000 replications per condition. Accuracy was measured by two measures of the distance between the true tree and the estimate of the tree, one measure sensitive to accuracy of branch lengths and the other not. The distance-matrix methods (Fitch-Margoliash and neighbor joining) performed best when they were constrained from estimating negative branch lengths; all comparisons with other methods used this constraint. Parsimony and compatibility had similar results, with compatibility generally inferior; Fitch-Margoliash and neighbor joining had similar results, with neighbor joining generally slightly inferior. Maximum likelihood was the most successful method overall, although for short sequences Fitch-Margoliash and neighbor joining were sometimes better. Bias of the estimates was inferred by measuring whether the independent estimates of a tree for different data sets were closer to the true tree than to each other.

CONCLUSIONS AND FURTHER RESEARCH

The system is based on the operation of computational Biology researcher's perspective .It is developed to integrate all individual operation those can be applied on a Phylogenetic tree in a single window. This system assumes that the section contains the only FASTA format files. It may be extended to various protein file formats with little bit of modification. The current developed in ANSI C under linux platform, which is open source and support for WEB BASED applications. In future this system can be easily upgraded to a web based system with little amount of effort

REFERENCES

NCBI (2006). http://www.ncbi.nlm.nih.gov

BIOINFO (2006). http://www.bioinformatics.org

A SOFTWARE ENVIRONMENT FOR PROTEIN STRUCTURE VISUALIZATION

Allam Appa Rao, Andhra University
allamapparao@gmail.com
A. Sanyoshukriya, Andhra University
B. Swapna, Andhra University
K. Soujanya, Andhra University

ABSTRACT

During the last decade, molecular biology has witnessed an information revolution in computer-based technologies. The broad term coined to encompass computer applications in biological sciences is bioinformatics. The main challenge of bioinformatics is rationalization of the mass of sequence information and to design more incisive analysis tools. The aim is to understand the elements in a sequence that forms a particular protein structure. Ultimately the goal is to develop strategies for drug discovery or disease analysis. This field of protein visualization is important because the structure of a protein is intrinsically related to its function. Experimental structure determination, aids the elucidation of protein function; conversely, synthetic protein sequences might be designed so that the protein performs a desired function. The study of protein structure is therefore not only of fundamental scientific interest in terms of understanding biochemical processes, but can also produce valuable practical benefits. Protein structures were originally determined by X-ray diffraction and neutron-diffraction studies of crystallized proteins, and more recently by nuclear magnetic resonance (NMR) spectroscopy Protein structures thus determined, are stored in Protein Data Bank (PDB) file formats. The Protein Data Bank is the single worldwide archive of structural data of biological macromolecules. All data in the archive have been validated and are accessible in a uniform archive. In this research we developed a Computer Program which takes the Protein Data Bank file as input to the program, it computes the data and visualizes the 3D structure of the protein and allows us to perform action on that structure.

INTRODUCTION

Bioinformatics has evolved into a full-fledged Trans Disciplinary subject that integrates developments in information technology (IT). Bioinformatics uses computer software tools for database creation, data management, data warehousing, data mining and global communication networking. Bioinformatics is the recording, annotation, storage, analysis, and searching/retrieval of nucleic acid sequence (genes and RNAs), protein sequence and structural information. This includes databases of the sequences and structural information as well methods to access, search, visualize and retrieve the information.

Bioinformatics concern the creation and maintenance of databases of biological information whereby researchers can both access existing information and submit new entries. Function genomics, bio-molecular structure, protease analysis, cell metabolism, biodiversity, downstream processing in chemical engineering, drug and vaccine design are some of the areas in which Bioinformatics is an integral component. The term Bio-Informatics has been commandeered by several disciplines to mean rather different things. In its broadest sense, the term can be considered to mean information technology applied to the management and analysis of biological data; this has implications in diverse areas, ranging from Artificial Intelligence and Robotics to genome analysis.

In the context of genome initiatives, the term was originally applied to the computational manipulation and analysis of Biological sequence data (DNA and/or Protein). However, in view of the recent rapid accumulation of available protein structures, the term now tends also to be used to embrace the manipulation and analysis of three-dimension (3-D) structural data.

The central challenge of Bio-Informatics is the rationalization of the mass of sequence information, with a view not only to deriving more efficient means of data storage, but also to designing more incisive analysis tools. The imperative that derives this analytical process is the need to convert sequence information into biochemical and biophysical knowledge; to decipher the structural, functional and evolutionary clues encoded in the language of biological sequences.

In investigating the meaning of sequences, two distinct analytical themes have emerged; in the first approach, pattern recognition techniques are used to detect similarity between sequences and hence to infer related structures and functions; in the second, ab initio prediction methods are used to deduce 3-D structure, and ultimately to infer function, directly from the linear sequence. The development of more powerful pattern recognition and structure prediction techniques will continue to be dominant themes in Bioinformatics. Research while the number of experimentally determined protein structures remains small.

PROTEIN STRUCTURE

Proteins are the polymers of L-alpha-amino acids. The structure of proteins is rather complex which can be divided into four levels of organization.

1. Primary Structure: The linear sequence of amino acid forming the backbone of proteins (polypeptides).
2. Secondary Structure: The special arrangement of protein by twisting of the polypeptide chain.
3. Tertiary Structure: The three-dimension structure of a functional protein.
4. Quaternary Structure: Some proteins are composed of two or more polypeptide chains referred to as sub-units. The special arrangement of the sub-units is known as quaternary structure.

VISUALIZATION

Variations in individual responses to environmental factors can be understood in terms of protein structure, especially of active and genetically altered (mutated) sites on enzymes and other proteins. By manipulating computer-generated, structural, 3-D models of proteins, investigators can focus on relevant features. However, structural modeling requires the skilled use of sophisticated graphics software and the resulting data can be difficult to evaluate without specialized training and experience. There are different models in which a protein can be visualized:

Star Model:	This model consists of small stars representing atoms in the protein. Each atom can be identified by specific color assigned to it. Colors can be changed depending on the structure or temperature factors, so that its properties can be identified.
Wire Frame Model:	This model consists of line segments linking each atom based on how atoms in each amino acid, and how amino acid bond together. In implementation we set two colors to each line segment. Each color represents the color of the atom or amino acid.
Ball & Stick Model:	This model consists of both line segments linking each bonded atoms and spheres representing atoms.
Ribbon Model:	This model represents the protein fold. This model is developed by connecting all the alpha carbon atoms.

Using above models a biologist can identify the cause of disease due to change in part of the protein or a pathologist can design a drug to target the protein if it is involved in the disease.

PROBLEM STATEMENT

The aim of this research is to develop a software environment which visualizes the structure of protein. The software being developed reads and converts the Protein Data Bank (PDB) text file into a three-dimensional structure which can be rotated in all three axes. Protein Data Bank (PDB) format is a standard for files containing atomic coordinates. Structures deposited in the Protein Data Bank at the Research Collaboratory for Structural Bioinformatics (RCSB) are written in this standardized format.

PROJECT GOALS

The main goals of the project are to visualize protein structure with following functionality:

Create a 3D protein structure from a given Protein Data Bank file.
Visualize 3D protein structures using several visualization techniques:
Rotate the opened 3D protein structure in all three axes..
Code each amino acid in the structure with specific color.
Code each atom in the structure with specific color.
Transport the image of the protein structure in Bitmap file.

RESULTS AND BENEFITS

The ability to visualize protein structures in 3D is critical to many aspects of biology. The Product being developed can be use to view the protein in a number of different ways. Visualization of protein structures helps in elucidation of protein functions. It allows the scientists to investigate:

Look at the cause of disease of it is due to change in part of the protein.
The effects of changing parts of protein.
Design drugs to target the protein if it's involved with the disease.
See how a protein may function.
See how two proteins interact.

CONCLUSIONS AND FURTHER RESEARCH

The software tool that is developed is graphical user interface that is easy to use. This allows convenient analysis of the protein functionality based on its structure.. Two or more protein structures can be compared and differences identified.

The product can be further developed to increase its functionality in following ways:

User defined colors can be given.
Specific amino acid can be highlighted.
Only a part of the protein can be visualized.
All the four models can be visualized at the same time.

REFERENCES

Attwood, T.K. & Parry-Smith, D.J. (1999) *Introduction to Bioinformatics.* Pearson Education.

Gibas, C. & Jambeck (2001) *Developing Bioinformatics Computer Skills.* O'Reilly Publications.

Westhead, D.R. & Twyman, R.M. (2003). *Bioinformatics Instant Notes.* Viva books publications.

SAMPLE OUTPUT

Figure 1: Visualization of Breast Cancer-Associated Protein (1JNX) in Star Model

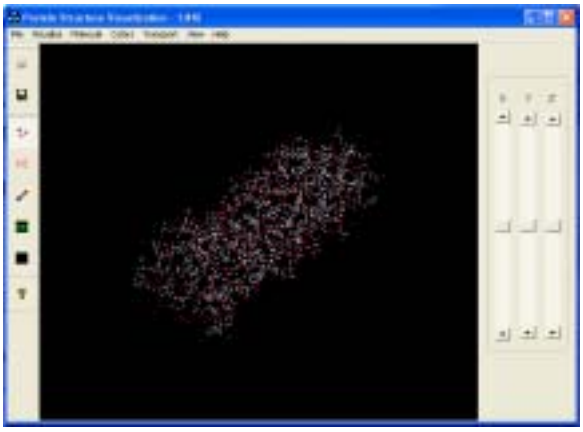
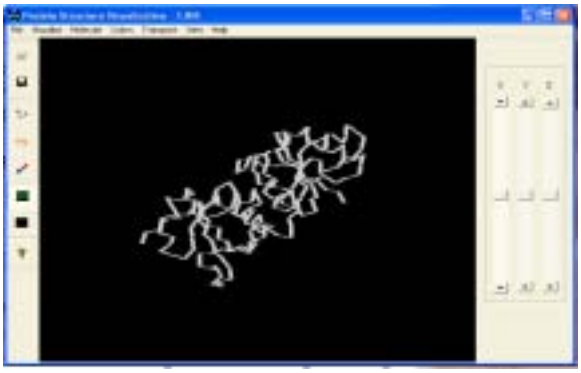


Figure 2: Visualization of Breast Cancer-Associated Protein (1JNX) in Ribbon Model



INFORMATION AND COMMUNICATION TECHNOLOGIES WITHIN ETHIOPIA: SOCIOPERSONAL FACTORS AFFECTING ADAPTATION AND USE

Melesse Asfaw, Walden University
masfaw@waldenu.edu
Raghu B. Korrapati, Walden University
rkorrapa@waldenu.edu

ABSTRACT

This mixed method descriptive study addressed the lack of adaptation and use of information and communication technologies (ICT) among males and females in Ethiopia. The purpose of the study was to examine the possible relationship between sociopersonal factors as barriers to the adaptation and use of ICT and male and female nonusers within Ethiopia. These factors included interest level, awareness, acceptance, and understanding/knowledge of ICT. The study findings revealed that sociopersonal barriers have a major impact on ICT usage and adoption by male and female nonusers within Ethiopia. The result also found that a lack of interest is the most significant factor hindering the adaptation and use of ICT. In addition, there were no significant differences between male and female nonusers. This study also identified four fundamental needs of ICT nonusers (education, training and support, motivation, and policy) that may enhance ICT adaptation and use in the future. The study is important to other researchers, the government, business organizations, educators, universities, private organizations, ICT manufacturers, Internet companies, and others.

INTRODUCTION

In developing countries, a diverse number of differences exist between male and female nonusers of ICT, including socioeconomic and sociopersonal status. Research on issues emerging from the adoption of new technology has focused on socioeconomic characteristics, the perceived attributes of innovations, technology clusters, situational factors, and the characteristics of innovations that influence adoption [1, 2]. Other studies have suggested that those who adopt new ICT are more highly educated and more affluent than those who do not use such technology [1, 3, 4, 5, 6].

Although it is difficult to determine whether socioeconomic or sociopersonal factors are more salient in explaining the digital divide, Morales-Gomez and Melesse [7] argued that sociopersonal factors such as interest level, awareness, understanding, and acceptance of ICT present greater significance than socioeconomic factors within developing countries. Enders and Seekins [8] supported this idea in their analysis of the digital divide. They found that 49% of adults living within developing countries who had never used ICT claimed that their reason was a lack of interest. The Office of National Statistics [9] reported a similar figure of 43% for disinterested nonusers of ICT. In a parallel study, Gary [10] documented that 39% of participating nonusers within developing countries stated that nothing would encourage them to use such technology.

Taschek [11] suggested that combining knowledge of socioeconomic barriers with a more methodical understanding of sociopersonal factors would provide a more holistic understanding of

the variables restricting the adaptation and use of ICT and how they might be overcome within developing countries such as Ethiopia. Therefore, further research is critical to enable a deeper understanding of the impact of sociopersonal factors as a barrier to ICT adaptation and use within Ethiopia.

PROBLEM STATEMENT

The problem addressed in the study was the lack of adaptation and use of ICT among males and females within Ethiopia. A considerable amount of literature exists regarding the underlying causes of the digital divide within such developing countries [12, 13, 14, 15, 16]. Most studies have indicated that socioeconomic factors such as low income [12], low education level [15], low skill level [14], unemployment [13], lack of access to technology, and lack of computer skills are the root causes of the information gap [16]. However, in the literature review conducted for this research, to the best of this researcher's knowledge, no research was found that explored sociopersonal factors as obstacles to the adaptation and use of ICT within Ethiopia.

According to Ulfeder [17], a substantial number of resources have been invested in achieving a more thorough understanding of socioeconomic barriers and in developing robust policies capable of overcoming them. However, the literature review conducted for this study [18, 19, 20, 21, 22] suggested that complementing knowledge of socioeconomic barriers with a more thorough understanding of sociopersonal barriers will provide a more holistic understanding of the factors restricting the use of ICT and may facilitate the development of possible solutions.

PURPOSE OF STUDY

The purpose of this study was to investigate the significance and implications of sociopersonal barriers to the adaptation and use of ICT within Addis Ababa City, Ethiopia. The focus was on attitudinal and behavioral issues such as awareness, interest level, understanding/knowledge, and acceptance as they relate to the adaptation and use of ICT by various social groups.

More specifically, the goals of this research were to:

1. Examine the relationship between sociopersonal barriers to the adaptation and use of ICT by male and female ICT nonusers in Addis Ababa City, Ethiopia.
2. Uncover which sociopersonal barriers most influence the decision against the use of ICT and how these issues can be addressed.
3. Determine which gender is more adversely affected by the digital divide and how this population can be encouraged to reverse this scenario.
4. Make recommendations on how government officials may formulate and implement policies toward the reduction of sociopersonal barriers.
5. Make recommendations on how policy makers integrate and give a high priority to the use of ICT effectively for a more equitable and pluralistic development of their education systems.
6. Make recommendations on how industrial experts and consultants, bank representatives, and selected others involved in the process of purchasing and distributing ICT products may contribute to the reduction in sociopersonal barriers to their adaptation and use.

METHODOLOGY

This mixed-method descriptive study was designed to collect data from participants in order to measure the socio-personal barriers affecting non-users of ICT in Ethiopia. Data from this population base has not been measured in past national surveys to determine the socio-personal factors affecting the adaptation and use of ICT. The study used a sample of 183 male and female ICT non-users (age >18 yrs., able to read and write English, and never used ICTs) in Ethiopia. Data

was collected through the administration of a survey to gather information about attitudinal and behavioral issues related to ICT non-users. The responses were utilized to measure the impact of sociopersonal factors to the adaptation and use of ICT process in Ethiopia. The data was limited to information supplied by survey respondents.

RESULTS

The study was guided by five research questions:

1. Which sociopersonal factors most significantly hinder adaptation and use of ICT within Ethiopia?
2. Which specific gender (male or female) within Ethiopia consists of the greatest number of computer and Internet nonusers?
3. What are the greatest needs of nonusers of ICT that are current barriers to increasing their technology use?
4. What are the attitudes, prejudices, and expectations of nonusers of ICT?
5. What are relevant and effective channels toward greater involvement of nonusers of ICT in future technology development?

A series of analyses were performed to answer the five research questions undergirding this study. In the subsequent presentation, the analyzed survey data were applied to answer the five research questions. Research question 1 asked, "Which sociopersonal factors most significantly hinder adaptation and use of ICT within Ethiopia?" It was found that:

1. A higher percentage of male and female participants agreed or strongly agreed that they are not interested in using computers and/or the Internet. This factor represented one reason why nonusers are not using and adopting ICT within Addis Ababa City, Ethiopia.
2. A higher percentage of male and female participants agreed or strongly agreed that they have no need to use a computer and/or the Internet. More importantly, the responses showed that the participants have little interest or need to use the computer or the Internet in their work.
3. A higher percentage of male and female participants agreed or strongly agreed that the computer and the Internet are far complicated to use. This may be one explanation for their reporting a lack of interest and need to use and adopt ICT.
4. A higher percentage of male and female participants agreed or strongly agreed that they have been scared to use a computer and the Internet. This was identified as another reason why nonusers were reluctant to use and adopt ICT in their lives.
5. In order to answer the first research question, composite scores were created, and the ranked ordered means and standard deviations for awareness, acceptance, understanding/knowledge, and interest factors were analyzed. The ranked ordered means and standard deviation analysis revealed that a lack of interest, with a mean score of 1.91 and a standard deviation of 0.87, is the most significant factor.

In summary, the analyses showed that a lack of interest in computers and the Internet is the most significant factor hindering the adaptation and use of ICT within Ethiopia. Both male and female participants held less interest and had more negative attitudes and perceptions about ICT use and adaptation. Diffusion theory also was evidenced here by the fact that most of the participants have been exposed to the innovation but lack complete information about it. This lack of information has made them uninterested in the new idea of technology and unwilling to seek more additional information about it.

Research question 2 asked, "Which specific gender (male or female) within Ethiopia consists of the greatest number of computer and Internet nonusers?" A total of 183 individuals participated in the survey; 78 (43%) were females and 105 (57%) were males. Chi-square analysis for each variable for gender differences revealed the following statistical difference:

1. ICT awareness: Females reported greater exposure and perception than did the males ($X^2 = 12.8$, $p = .01$).
2. ICT acceptance: Female reported greater considering ICT as a pleasant experience than did the males ($X^2 = 12.5$, $p = .01$).

3. ICT understanding/knowledge: Female reported greater that they would not feel comfortable using ICT than did the males ($X^2 = 11.4$, $p = 0.02$); on the two questions dealing with confidence, the females reported less confidence to use ICT with no instruction ($X^2 = 17.7$, $p < .01$) and less confidence to learn ICT on their own ($X^2 = 11.8$, $p = .02$).
4. ICT interest: A higher percentage of females than males disagreed that there is no need to use a computer ($X^2 = 31.7$, $p = < .01$) or the Internet ($X^2 = 18.13$, $p < .01$).
5. To answer research question 2, statistical analyses with ANOVA and MANOVA were conducted. The MANOVA was not significant $F(4,177) = 1.86$. The MANOVA results suggested no differences between the males and the females on the combination of the four survey subsections. Four individual ANOVAs were also conducted to assess if mean differences existed between the males and the females on any of the four survey subsections. The *F* proportions were lower than the critical values, so the researcher concluded that there were no significant differences between the IV and the DV.

Thus, the findings suggested that there were no significant differences between male and female nonusers of ICT. It also was evidenced that both male and female nonusers continue to endure the impact of illiteracy, poverty, lack of training, and sociocultural restrictions, which hampers their ability to benefit from opportunities offered through ICT.

Research question 3 asked, "What are the greatest needs of nonusers of ICT that are current barriers to increasing their technology use?" This research identified four fundamental needs for ICT nonusers to enhance their technology adaptation and use:

1. Education: Low level of formal education attendance and achievement can be a barrier to ICT use. In this study, it was noted that access associated with education marked the greatest disparities. None of the 69 high school graduates reported access to the computer and/or the Internet at work or at home. When the participants were asked to express their opinion about conditions that would inspire their understanding and knowledge of ICT, 45% of the males and 53% of the females responded that digital literacy would enhance their adoption and usage of ICT. They pointed out that the need for computer literacy courses in all level of education is crucial and that developing ICT skills at school is mandatory and providing sufficient stimulus for the less educated to appreciate the benefits of ICT and enhance their desire to learn how to access and use ICT are much more needed.
2. Training and Support: An important finding was that ICT nonusers believe that lack of training and/or support is one of the reasons for not using ICT. When they were asked to express their opinion about conditions that will inspire their understanding/knowledge of ICT, 31% of the males and 22% of the females indicated the need for basic computing training, and (b) awareness of ICT, 53% of the males and 51% of the females indicated the need for workshops and/or seminars. From the survey responses, it was evident that training and/or support, as well as workshops and/or seminars, is vital for nonusers or new users that should probably be implemented along with other initiatives.
3. Motivation: According to the findings derived from this study, 40% of the males and 35% of the females indicated no interest in adapting and using ICT. Twenty-nine percent of the male nonusers and 27% of the female nonusers claimed that nothing would encourage them to use the Internet. These statistics clearly indicated that even if access to ICT is widened, a considerable proportion of nonusers will remain as nonusers because of their lack of interest or negative attitude toward ICT. This finding identified the crucial need to inform individuals of the benefits that can be derived from ICT use. Motivation of the targeted groups through workshops, seminars, and the use of traditional media channels to introduce the possibilities of ICT in a local language were among the essential needs raised by the nonusers.
4. Policy: ICT policies need to take a more holistic approach and inform nonusers about the benefits of using and adapting ICT. A higher percentage of male and female participants indicated the need for new policies and promotional campaigns to make nonusers aware of the benefits of ICT. It was the participants' main concern that promotional campaigns should take a citizen-centric approach rather than continue the existing technocratic approach. They also pointed that the current national ICT policy environment must be modernized and revised to address potential users' needs.

The diffusion of innovation theory was evidenced here by noting the participants' needs and necessities. This theory argues that social needs and necessities play an important role in behavior and attitude changes. The role of needs and necessities in a community, acting as agents for behavior and attitude changes, is a key element of this theory.

Research question 4 asked, "What are the attitudes, prejudices, and expectations of nonusers of ICT?" This study investigated the relationship between nonusers' attitude and prejudices (as a personal factor) and the use and adaptation of ICT in Addis Ababa City, Ethiopia. As the literature review indicated, the components of attitude and prejudice that make ICT operationalization possible were cognitive, affective, and behavioral factors. Measures were made of the respondents' knowledge of the capabilities of ICT use (cognition) and their feelings about them (affect), which were then compared to their actual use of this technology (behavior). The respondents justified their negative attitude toward ICT by referring to their lack of training, lack of experience, lack of time, and lack of computer access.

It was found that a higher percentage of males and females were concerned with the value of ICT in their lives. It was also noted that male and female participants generally had negative attitudes and resistance toward using and adapting ICT. Forty percent of the male nonusers and 35% of the female nonusers in Addis Ababa City who participated in this survey are not interested in using or adapting ICT. This study also indicated that a lack of interest and a negative attitude toward ICT are probably major constraints on the use and adaptation of ICT by nonusers. Although several nonusers of ICT had definite excuses why they should not use such technologies, most of them reacted positively and expressed the prospect of using ICT in the future if training and support are offered, language barriers are resolved, affordable access becomes available, strong and meaningful policies are implemented, and ICT benefits are outlined.

Research question 5 asked, "What are relevant and effective channels toward greater involvement of nonusers of ICT in future technology development?" This study identified a solid pool of nonusers who believe they have no need for ICT; the resisters, who simply do not want to use ICT; and those who have no intention of ever using the Internet. A lack of understanding about the clear benefits of ICT appears to be a contributory factor in their decision, but almost 50% of the nonuser group cited a lack of interest or a lack of access as the main explanations for their nonuse of ICT. This finding accorded with results of the Pew Project, where the majority of non-Internet users (56%) did not think that they would ever go online. They felt no need or desire to use the Internet [23]. Additional concerns cited by a plurality of nonusers in this study included worry about safety, the cost, a lack of time, finding computers and the Internet too complicated and hard to understand, embarrassment by their lack of computer skills, no computer or Internet connection, as well as language skills and cultural roles that acted as barriers to ICT use.

To overcome these barriers, this researcher identified the need for core policies and program objectives concerned with such issues as affordable access, technological and social literacy, social capacity and application, and indigenous social and cultural content development as effective channels toward the involvement of ICT nonusers. In addition, multiple means of access to, and distribution of, information at comparable levels of quality and service must continue to be made available.

In summary, the result of this study revealed that sociopersonal barriers such as lack of acceptance, awareness, knowledge/understanding, and interest have a major impact on ICT usage and adoption by nonusers within Ethiopia. The result also demonstrated that lack of interest is the most significant factor hindering the adaptation and use of ICT. There were no significant differences between the male and the female respondents on any of the four survey subsections. Furthermore, this study identified four fundamental needs of ICT nonusers (education, training and support, motivation, and policy) that may enhance nonusers' technology adaptation and use in the future. The result of this study also suggested that motivation of the targeted population through workshops, seminars, and the use of traditional media channels to introduce the possibilities of ICT in a local language are essential needs of nonusers. Finally, this study uncovered the need for core ICT policies and programs, affordable access, technological and social literacy, social capacity and application, and indigenous social and cultural content development as effective channels toward the involvement of ICT nonusers.

CONCLUSION AND RECOMMENDATIONS

A number of recommendations for further study are presented based on the findings and conclusions of this study. The recommendations can be addressed by the following five research projects:

1. It is recommended that follow-up studies be conducted with a larger sample size and a broader diversity of the sample groups included in the population.
2. A research study is required that will examine the attitudes, acceptance, and understanding of ICT, distinguished by factors such as ethnicity, income, age, and gender.
3. A research study is required to examine, at a fine scale of geographical detail, the level of ICT use and nonuse by households throughout Ethiopia.
4. A workplace study focusing on ICT use by employees would provide details about current use and the potential for business support of ICT work-based initiatives.
5. A study of public access points and ICT training centers would be beneficial.

The dominant result from this research confirmed that sociopersonal barriers will not be addressed only through universal physical access to ICT. A large number of Ethiopians lack the motivation and interest to use and adopt ICT. This research suggested that the need for well-planned, targeted, and integrated programs, strategies, and policies should accompany the rollout of ICT implementation if high levels of ICT use and adaptation for local community benefit are to be obtained.

REFERENCES

[1] Rogers, E. (1995). *Diffusion of innovations* (4th ed.). New York: Free Press.

[2] Zhu, J. H., & Zhou, H. (2002). Diffusion, use, and impact of the Internet in Hong Kong: A chain process model. *Journal of Computer Mediated Communication, 7*(2), 1-26.

[3] Dutton, W. H., Rogers, E. M., & Jun, S. H. (1987). Diffusion and social impacts of personal computers. *Communication Research, 14*(2), 219-250.

[4] Garramone, G., Harris, A., & Anderson, R. (1986). Uses of political bulletin boards. *Journal of Broadcasting and Electronic Media, 30*, 325-339.

[5] James, M. L., Wotring, C. E., & Forest, E. J. (1995). An exploratory study of the perceived benefits of electronic bulletin board use and their impact on other communication activities. *Journal of Broadcasting and Electronic Media, 39*, 30-50.

[6] Lin, C. A. (1998). Exploring personal computer adaptation dynamics. *Journal of Broadcasting and Electronic Media*, *42*, 95-112.

[7] Morales-Gomez, D., & Melesse, M. (1998). Utilizing information and communication technologies for development: The social dimension. *Information Technology for Development, 8*, 3-13.

[8] Enders, A., & Seekins, T. (1999). Telecommunication success for rural Americans with disabilities. *Rural Development perspectives, 14*(3), 14-20.

[9] Office for National Statistics. (2000). *Internet access.* Retrieved April 18, 2005, from http://www.statistics.gov.uk/statbase/Product.asp?vlnk=5672&More=N

[10] Gary, V. (2005). *ITU world telecommunication indicators: Data collection and dissemination.* Paper presented at the Capacity-Building Workshop on Information Society Measurements: Core Indicators, Statistics, and Data Collection, Beirut, Lebanon.

[11] Taschek, J. (1999). Crossing the great digital divide. *PC Week, 16*(29), 65.

[12] Choudhury, S., & Wolf, S. (2002, May 10-11). Use of ICTs and economic performance of small and medium scale enterprises in East Africa. Paper presented at the WIDER conference on the New Economy in Development, Helsinki, Finland.

[13] Fink, C., Mattoo, A., & Rathindran, R. (2002). *An assessment of telecommunication reform in developing countries. Policy Research Paper No. 2909.* Washington, DC: World Bank.

[14] Kenny, C. J. (2000). Expanding Internet access to the rural poor in Africa. *Information Technology for Development, 9,* 25-31.

[15] Venables, A. J. (2001). *Geography and international inequalities: The impact of new technologies*. Paper presented at the Annual Bank Conference on Development Economics, Washington, DC.

[16] Warschauer, M. (2003). *Technology and social inclusion: Rethinking the digital divide.* Cambridge, MA: MIT Press.

[17] Ulfeder, J. (2002). Into the breach tackling the digital divide. *World Link, 15*(1), 12-25.

[18] Albaugh, P. (1997). *The role of skepticism in preparing teachers for the use of technology*. Paper presented at the Education for Community Conference, Westerville, OH.

[19] Foley, P., Alfonso, X., & Ghani, S. (2002). *The digital divide in a world city*. London: Greater London Authority.

[20] Jeffres, L., & Atkins, D. (1996). Predicting use of technologies for communication and consumer needs. *Journal of Broadcasting and Electronic Media, 40*, 318-330.

[21] Norris, P. (2001). *Digital divide: Civic engagement, information poverty and the Internet in democratic societies.* New York: Cambridge University Press.

[22] Wilhelm, A. (2004). *Digital divide: Toward an inclusive information society.* Cambridge, MA: MIT Press.

[23] Lenhart, A., Horrigan, J., Rainie, L., Allen, K., Boyce, A., Madden, M., et al. (2003). *The ever-shifting Internet population. A new look at Internet access and the digital divide. The Pew Internet and American Life Project.* Retrieved May 23, 2006, from http://www.pewinternet.org

REFLECTIVE LEARNING FOR STUDENTS' DATA MODELING

I-Lin Huang, Langston University
ihuang@lu-ns01.lunet.edu

ABSTRACT

Accurate data models are well known as prerequisites for the quality of the final information systems. However, data modeling remains a complex and error-prone process for student database designers. Empirical studies have showed that the student database designers have difficulties in modeling relationships correctly. Especially, the degree and connectivity in a relationship are two major sources for errors

Reflection learning has long been recognized in the field of learning research as a strategy that can improve students' cognitive abilities to solve complex problems. In order to improve students' cognitive abilities of data modeling, this research argued that student database designers should be trained to incorporate reflective learning mechanism into their data modeling process. On the basis of the theories on human cognition, this research proposed a reflective learning process for data modeling to stimulate student database designers to perform effective reflection and to achieve a higher-level of correctness of data models.

PROPERTIES OF SHARED KNOWLEDGE – APPLICATION OF HIGHLY INTEGRATED INFORMATION SHARING SYSTEMS IN PUBLIC EDUCATION

Robert Konopka, Walden University
rkono001@waldenu.edu
Raghu Korrapati, Walden University
rkorrapa@waldenu.edu

ABSTRACT

This research presents a business project the development of information sharing management system (ISMS) as a possible solution to the problem of delivering administrative services in a charter school. Charter schools are based on using technology in curriculum delivery, application delivery, and communication. Such multilayered use of technology must connect and integrate multiple locations without affecting academic performance or other aspects of school operations. Technology can indirectly increase the quality of managing the process of providing education, and positively influence the quality of education delivery. The paper follows a business project to improve the quality of information communication and data sharing system. The critical element of the project was a separation between strictly academic services – focused on students and delivering education to students, and administrative services – focused on providing business like services to all administrative employees. This principle was used with a strong support from application of business process reengineering (BPR). The development of highly integrated data communication and sharing model was later introduced as Information Sharing Management System (ISMS). This research finds a positive relationship between a centralized information sharing system and the quality of administrative services offered in a charter school. It demonstrates that school administration can be managed and improved by using widely available management theories and techniques.

INTRODUCTION

The opportunities in the education and knowledge industries have been consistently growing as modern society values knowledge and information as one of its most important assets. Early analysis of U.S. spending on education was estimated at more than $600 billion in 1997 [1]. The educational industry is measured to be the second largest industry, after health care. The market is utilizing all available technology products and service providers specializing in the education and knowledge markets. The business opportunities in the K-12 sector have been growing and the state-wide initiatives to allow parents a choice when selecting public school have attributed to growth of charter schools in education.

California Open School (COS) is an educational management company that oversees a number of charter school systems in the state of California. COS manages over 30 remote learning center locations throughout southern California that are connected in single virtual information sharing network. The IT department as a business unit evolved from the secondary to the primary role of electronic communication, curriculum delivery, electronic security, remote system access, identity protection, and business optimization. COS already created a framework for solid

information technology and secured system improvement funding. It is important to mention that all funding comes from the state education budget and the company must follow strict funding guidelines and complete detailed audits every year.

BUSINESS GROWTH AND CHALLENGES

California Open School (COS) realized that when technology was available as a service tool, it would not guarantee a success and was often cited as a failure cause. The opportunity to apply IT can be misguided by poor understanding or poor implementation. It is important to mention that computers in the academic environment cannot be classified as a strategic use of IT. Such a strategy must include fundamental properties of information technology as the administrative aspect of delivering education and to be as important as education itself. COS also established a baseline for benchmarking how education is delivered to the student population.

Using and placing computers in the classroom without the fundamental change in the way the organization provides an educational service does not influence the success rate. McCredie [4] pointed out that any strategy to use information technology would include personal computers but that would be the final aspect of the entire policy. Other authors pointed at the industry wide problems when discussing use of information technology [3] and used the following examples of poor use of technology in the modern education system:

- Lack of vision in application of modern IT.
- Lack of consistency in applying any strategy or chosen vision.
- Lack of depth in understanding how available tools can be used.
- Lack of adaptability in the use of IT as such is constantly evolving and changing.
- Lack of understanding how the IT works among educational practitioners and teachers.

All above mentioned problems are significantly magnified as a charter school must rely on IT to connect all traditionally static elements. It is not the single computer that enhances the educational service; it is the entire information sharing system with all elements of networking, software, hardware elements, and application used in the process that creates a common framework of operations. Such groundbreaking strategy is the only way that charter schools can succeed on a large, global scale [2].

The operational budget is directly related to the number of students. The volume of students dictates the capability of the business process. COS is calculating business growth at 15% next year and 10% each consecutive year. Information-driven growth requires the company to tightly integrate the Information Technology as a primary driver toward quality improvements and growth potential. Because of historical deficiencies in educational technology understanding and practice, there is a need to point out that such a systematic approach helps to integrate the curriculum and model the technology for a broader range of teaching. The problem needs to be identified as a lack of unified data retention and management policy that can interact with all available information about students, curriculum, administrative aspects of providing academic services, and school administration as business unit.

TECHNOLOGICAL ADVANCEMENTS AND OPPORTUNITIES

The selection of Michael Porter's 5 Force Model [5] was used to help the company to look at the problem from an outside the box perspective and treat the educational market like any other industry. It was applied to understand industry power relations and organizing industry research. Drawing from microeconomic theory, the model was applied to identify five forces that influence

the ability of California Open School (COS) to set business expectations. The patterns of forces had a dual purpose: to shape an industry and constrain company strategic choices within the industry and the company strategies choices and desire to change.

The next step was to analyze all known and assumed barriers to entry as discovered during the historical data analysis and presented in Table 1. At that time, COS was struggling to provide any services. The idea of quality of services was not yet considered and the majority of the efforts were focused on fixing existing administrative problems. Lack of planning and lack of understanding as the root cause of the problem was destroying any real efforts by the IT team.

Table 1: Perceived system weakness and barriers to entry

Type of Barrier	**Relationship to Information Technology**	**Business Technology Impact**
IT Infrastructure (Reliable, stable, secure)	The essential element of Information Technology Infrastructure, The foundation that is required for any future planning and growth.	Prerequisite to any business planning, Will require a significant capital investments, Will require change in internal IT department, Will require high level of technical expertise and skill set.
Application Development and Support	Corporate applications can be separated by their function to: Network Core, Business Core. and Education Core,	Each element will be analyzed separately (unless required to be combined with other element) improvements will focus on the overall network/system/application improvement. Current academic application provider will be used. No changes are planned or will be evaluated within the next 24-36 months. Might add significant cost to the project (application upgrade, licensing cost, support and maintenance cost). .
Centralized technology management	Will significantly change the decision center at IT department. Will add partner in any business discussions and will change how technology is perceived and valued at the highest level of management and control.	IT outsourcing will become a viable option when planning IT growth. Strong IT management will be required to influence future IT initiatives and development. Will add a significant value to the administrative services offered by the charter school managing company.
Industry regulations	Will influence future of technology as related to privacy, security, and characteristics of a K-12 environment.	Will require a constant evaluation of current/available/future technology. Federal regulations will follow the same trends as already observed in public school systems.

DATA SHARING AND ACCESS IMPROVEMENTS

The second phase of the project was focused on following objectives:

- Reduce data system complexity.
- Increase data availability.
- Increase data safety at all levels of data access.
- Increase data management.
- Reduce cost associated with data lifecycle business process.

The early business objective to: *complete all work as soon as possible and focus on cost to performance ratio*, was later changed to: *understand and improve the complex relationship between data acquisition, communication, manipulation, exchange, and sharing processes*. The question was

finally asked: *What was the connection between the scope of the data access and the creation of entirely new concept of Information Sharing Management System network infrastructure?* One interesting element was discovered: data protection was classified at the lover level than data access. Over 80% of responders listed unlimited data acquisition and access as the first objective of the new system. Data security and protection was listed by 58% of responders. The low level of technical knowledge of the responders indicated that they were concerned if the access to all information was going to be reduced or the system created an additional level of complexity. A typical corporate user had six to seven network data repositories that were used daily. A typical teacher had five to six network data repositories that were used for the academic purpose. Data Lifecycle infrastructure initiative was not only an opportunity for a future growth; it was also a business necessity to survive unnecessary system complexity and to increase data management.

Access to any information was possible from any remote or corporate location but was often unpredictable in the quality of service or uptime. System backup was performed in each physical location as a separate process. Each backup application had to be individually licensed and managed. The data management element required a part time employee to make sure all backup and archiving processes were conducted and completed successfully. Before this phase of the project was officially approved, the company requested a complete *disaster recovery drill* that was focused on recovering all data. The results of a *disaster recovery drill* were later used as one of the project justifications elements.

One important element was identified as a critical data access factor. Data access and usage is related to the business function of an employee, not the employee's function inside the corporate structure. Two administrative employees were identified as data power users (accounting and payroll services). To understand all limitations and requirements, extensive interviews were conducted with all layers of corporate management. The system benchmarking was focused on how the new information sharing system could add value to the overall IT Infrastructure and provide more tangible level of quality to administrative services. Information sharing performance benchmarking was based on two distinct categories: (a) system protection and disaster recovery and (b) data lifecycle access and manipulation. Results of the system benchmarking are presented in Table 2.

Table 2: Data Access Results Summary

Information Sharing Element	**Distributed Data Access System (old)**	**Centralized ISMS (new)**	**Comments**
Total Number of data server	43	13	Reduction of 30 servers (330.75% improvement)
Data Repository Servers	27	3	Reduction 24 servers (900% improvement)
Access to Academic Data 25 locations 3 locations 2 locations	 6 repositories 8+ repositories 10+ repositories	 2 repositories 2 repositories 2 repositories	 Reduction of 4 Reduction of 6 Reduction of 8
Network Storage	1,258 GB (gigabytes)	485 GB (gigabytes)	Reduction in space requirements 773 GB (385% improvement)
Total active storage space	3,287.25 GB (gigabytes)	658 GB (gigabytes)	Reduction in space requirements 2,629.25 GB (499.58% improvement)

INTEGRATED SYSTEM

That idea that all available information could be perceived in its organic matter (is being born, grows into adulthood, and is finally retired or recycled) allowed creation of three distinct levels of data access as a complete lifecycle of the process as presented in Figure 1.

THE IMPACT OF FAIRNESS ON USER'S SATISFACTION WITH THE IS DEPARTMENT

Obyung Kwun, Emporia State University
Khaled Alshare, Emporia State University

ABSTRACT

This study utilized the justice theory to study the effect of fairness of information systems development process on user satisfaction with the IS department. To validate the research model, partial least square (PLS) analysis was used to analyze the data that were collected from 123 middle-level managers who have participated in the IS development. The findings showed that interactional justice and distributive justice, but not procedural justice, had positive impacts on user satisfaction with IS department. Additionally, interactional justice had the strongest impact on user satisfaction with the IS department.

AN EVALUATIVE CASE STUDY OF DISTANCE LEARNER EXPECTATIONS FOR TECHNOLOGY-ENABLED SUPPORT SERVICES

Kathleen O. Simmons, Walden University
ksimmons@waldenu.edu
Raghu B. Korrapati, Walden University
rkorrapa@waldenu.edu

ABSTRACT

The purpose of this study will be to explore learner expectations for technology-enabled support services in undergraduate and graduate programs at one distance-based university. An evaluative case study method will be used as the research design. A survey instrument, telephone interview, and examination of historical documents will be used as the data collection techniques. The survey instrument will be built based on the SERVQUAL model, a survey instrument employed in the commercial sector to test consumer expectation against service performance. The participants for the survey element of this study will be chosen from approximately 20,000 current students at the school. The target sample of former students that will be interviewed is six. It is planned to obtain roughly 30 exit survey/interview documents for analysis. The results of this study may offer service providers at distance education institutions a framework for the further study of student support service quality. Positive social change will be the result of the unique position of distance-based institutions to reach underrepresented populations through technology-enabled programs and services.

INTRODUCTION

Higher education institutions that offer distance-based programs have not placed enough emphasis on providing technology-enabled support services to distance learners. Attrition rises when support services are not delivered using technology to enable remote access. Attrition must be managed because distance programs suffer higher drop out rates than do traditional programs and it is estimated to cost $7,000 to replace just one student.

The purpose of this study will be to investigate learner expectations for technology-enabled support services in undergraduate and graduate programs at a distance-based university. This study is also intended to establish if there is a perceived gap between expectation for service and actual service performance by a sample of students at this school. With this knowledge the service provider would be able to evaluate its method of delivering support, develop a strategy for change to improve the quality of support services, and manage attrition. Gender, age, program affiliation, and time in program will be explored as factors that might be useful in helping service providers plan service improvement programs that best match the demographic make up of a particular student population.

NATURE OF THE STUDY

For this study, an evaluative case study method will be used as the research design. According to Leedy and Ormrod (2005) qualitative research intends to uncover the complexity of an observed phenomenon with the goal of gaining a better understanding of a particular occurrence or experience. Cooper and Schindler (2003) agree and go on to say that, often, qualitative research

is exploratory in nature providing the researcher with greater insight into whether or not the supposed problem exists and warrants further study. One type of qualitative research, "the case study method refers to descriptive research based on a real-life situation, problem, or incident and situations calling for analysis, planning, decision making, or action with boundaries established by the researcher" (Simon, 2006, p. 48). For a qualitative study, the researcher examines a small group comprised of subjects who seem to have knowledge about or insight into the study area. The researcher starts with questions about some phenomenon that he or she wishes to answer. The data collection process, while not completely unplanned, is open-ended and leaves room for going in new directions given the results.

SIGNIFICANCE OF THE STUDY

The significance of the study will be a contribution to the SERVQUAL knowledge domain and to the literature related to remote student support. Also, it will allow service providers to assess support delivery methods and develop strategies for change to improve the quality of technology-enabled support services. A student satisfaction survey is often used to understand learner perception about the support capabilities of the institution. However, these surveys usually test service performance alone. Testing only the outcome of the service transaction leaves a gap in understanding the level of importance learners place on how services are delivered.

SERVQUAL is a survey instrument employed in the commercial sector to test consumer expectation against service performance. It will be used in this study to understand what learners expect from the service provider and determine if there is a gap in service quality. An importance factor will be added to the model as a way to gain insight into the level of importance distance learners place on how support services are delivered. Knowing this could help service administrators prioritize service needs and focus resources on delivering the services that are most important to the students in a way that meets their expectations.

RESEARCH QUESTIONS

Fundamental to the discovery and analysis phases of this study are a set of research questions designed to explore distance student expectation for technology-based student support services and determine if there is a perceived gap between expectation for service and actual service performance.

Question 1: How can the expectations of distance learners for technology-enabled support services be characterized?
Question 2: How do distance learners view the service performance of the university relative to their expectations?
Question 3: How does the introduction of an importance factor affect distance learners' perception about the service performance of the institution?
Question 4: How do expectations about technology-enabled support services differ based on age, gender, program (undergraduate vs. graduate), and time in program?
Question 5: How does importance placed on technology-enabled support services differ based on age, gender, program (undergraduate vs. graduate), and time in program?

SUMMARY

The method of delivery of student support services is a high priority issue for higher education institutions that offer distance-based programs because it contributes to distance student persistence and overall academic success (U.S. Department of Education National Center for Education Statistics, 2000). Failure of distance learners to persist through degree completion results

in attrition. While the exact rate of attrition from distance education programs has not been determined on a national level (Howell et al., 2003), Parchoma (2003) references the outcomes of three studies where "20-50% attrition rates in distance education programs in United States colleges". (p. 36) were reported. Studying the perceptions of students about the delivery of student support services is one way for administrators to gain a better understanding of their needs and to design strategies to better serve the learner. In our next published paper, a more extensive review will be presented on: the driving forces for change in higher education; the importance of student support services for distance learners; service quality concepts; business-based theories on customer satisfaction; and the link between technology and the management of distance student support needs. This is a work in progress study as part of Doctoral Degree.

REFERENCES

Cooper, D. R., & Schindler, P. S. (2003). *Business Research Methods*. (Eighth ed.). New York, NY: McGraw Hill.

Howell, S. L., Williams, P. B., & Lindsay, N. K. (2003). Thirty-two Trends Affecting Distance Education: An Informed Foundation for Strategic Planning. *Online Journal of Distance Learning Administration*, 6(3). Retrieved August 8, 2004, from http://www.westga.edu/~distance/ojdla/fall63/howell63.html.

Leedy, P. D., & Ormrod, J. E. (2005). *Practical Research: Planning and Design*. (Eighth ed.). Upper Saddle River, NJ: Prentice-Hall, Inc.

Parchoma, G. (2003). Learner-centered design: Two Examples of Success. Global Knowledge Enterprise - Learning Without Walls, 4th Distance Learning and the Internet Conference, sponsored by the Association of Pacific Rim Universities, December 1-2, 2003, Singapore.

Simon, M. K. (2006). *Dissertation & Scholarly Research: A Practical Guide to Start & Complete Your Dissertation, Thesis, or Formal Research Project*. Dubuque, IA: Kendall/Hunt Publishing Company.

U.S. Department of Education National Center for Education Statistics. (2000). Lifelong Learning NCES Task Force: Final Report, Volume I.

A JAVA BASED TOOL FOR IMPLEMENTING THE PAIR-WISE ALIGNMENT ALGORITHMS

Allam Appa Rao, Andhra University
allamapparao@gmail.com
G. Prakash Gupta, Andhra University
M. Rajesh Babu, Andhra University
P. Sateesh Chandra, Andhra University
D.V. Phaneendra Teja, Andhra University

ABSTRACT

Bio-Informatics is an inter-disciplinary subject spanning a range of specialties that include Computer Science, Molecular Biology and Mathematics. It makes use of scientific and technological advances in the areas of Computer Science, Information Technology and Computational Biology to solve complex problems in Life Sciences, particularly problems in Disease Diagnosis and Drug Discovery. The relationship between a query sequence, commonly termed as probe and other sequence, known as subject can be quantified and their similarity can be assessed. This similarity can be used to identify the evolutionary relationship between a newly determined sequence and a known gene family. When the degree of similarity is low, the relationship must remain computative, until evidence has been gathered. This research uses the various pairwise alignment algorithms, like Needleman-Wunsch global alignment, Smith-Waterman algorithm to find the optimum alignment (including gaps) of two sequences. Dynamic programming methods ensure the optimal global alignment by exploring all possible alignments and choosing the best. The user will be provided a good user-interface for giving the input sequences and opting for the required algorithm. The algorithm uses the BLOSUM 50 substitution matrix for computation, and outputs the Functional matrix(F-matrix), Optimal Alignment and the score. The total alignment score is calculated as a function of the identity between the aligned residues and the gap penalties incurred. The input sequences can be taken from a file stored on the disk.

PURPOSE OF STUDY

The purpose of this research is to implement various methods for pairwise alignment techniques and design a tool with a good user -interface for aligning two sequences and output the score of the alignment. The sequence alignment is a linear comparison of amino-acid sequence in which insertions are made in order to bring equivalent positions in adjacent sequences into the correct register.

Implementation Phases

The implementation part has three phases. They are

1 Input the Sequences.
2 Align the sequences using the algorithms.
3 Output the optimal alignment and score.

Input

The two sequences are read from the user and are feed to the selected algorithm. The sequences may be taken from a file stored on the disk also.

Alignment

The sequences are aligned for global and local alignment using Needle-man and wunsch algorithm, Smith-Waterman algorithm, repeated matches with simple gap costs, and overlap matches with simple gap costs. The functional matrices are computed along with the score and optimal alignment.

Ouput

The optimal alignment, score and the functional matrices are printed on their respective fields.

Alignment algorithms

Given a scoring scheme, we need to have an algorithm that computes the highest-scoring alignment of two sequences.We will discuss alignment algorithms based on dynamic programming. Dynamic programming algorithms play a central role in computational sequence analysis. They are guaranteed to find the optimal scoring alignment.

However, for large sequences they can be too slow and heuristics (such as BLAST, FASTA, MUMMER etc) are then used that usually perform very well, but will miss the best alignment for some sequence pairs.

Depending on the input data, there are a number of different variants of alignment that are considered, among them global alignment, local alignment and overlap alignment.

We will use two short amino acid sequences for illustration: HEAGAWGHEE and PAWHEAE.

To score the alignment we will use the BLOSUM50 matrix and a gap cost of d = 8. (Later, we will also use affine gap costs.)

Here they are arranged to show a matrix of corresponding BLOSUM50 values:

	H	E	A	G	A	W	G	H	E	E
P	-2	-1	-1	-2	-1	-4	-2	-2	-1	-1
A	-2	-1	**5**	0	**5**	-3	0	-2	-1	-1
W	-3	-3	-3	-3	-3	**15**	-3	-3	-3	-3
H	**10**	0	-2	-2	-2	-3	-3	**10**	0	0
E	0	**6**	-1	-3	-1	-3	-3	0	**6**	**6**
A	-2	-1	**5**	0	**5**	-3	0	-2	-1	-1
E	0	**6**	-1	-3	-1	-3	-3	0	**6**	**6**

Gap penalties

Gaps are undesirable and thus penalized. The standard cost associated with a gap of length g is given either by a linear score

$$\gamma(g) = -gd$$

or an affine score

$$\gamma(g) = -d-(g-1)e,$$

where d is the gap open penalty and e is the gap extension penalty.

Usually, e < d, with the result that less isolated gaps are produced, as shown in the following comparison:

Global alignment: Needleman-Wunsch algorithm

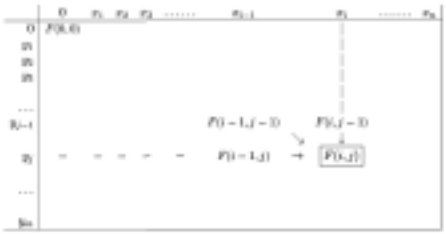
Obtaining the best global alignment of two sequences. The Needleman-Wunsch algorithm is a dynamic program that solves this problem.

Idea: Build up an optimal alignment using previous solutions for optimal alignments of smaller substrings.

Given two sequences $x = (x1, x2, \ldots x_n)$ and $y = (y1, y2, \ldots y_m)$ We will compute a matrix. $F : \{1, 2, \ldots, n\} \times \{1, 2, \ldots, m\} \Rightarrow R$

in which $F(i,j)$ equals the best score of the alignment of the two prefixes $(x1, x2, \ldots xi)$ and $(y1, y2, \ldots yj)$

This will be done recursively by setting $F(0,0) = 0$ and then computing $F(i, j)$ from $F(i - 1, j - 1)$, $F(i-1,j)$ and $F(i,j-1)$:



The Recursion

There are three ways in which an alignment can be extended up to (i,j):

x_i aligns to y_j:	x_i aligns to a gap:	y_j aligns to a gap:
I G A x_i	A I G A x_i	S G A x_i -
L G V y_j	A G V y_j -	S L G V y_j

We obtain F(i,j) as the largest score arising from these three options:

$$F(i,j) := \max \begin{cases} F(i-1, j-1) + s(x_i, y_j) \\ F(i-1, j) - d \\ F(i, j-1) - d. \end{cases}$$

This is applied repeatedly until the whole matrix F(i, j) is filled with values.

To complete the description of the recursion, we need to set the values of F(i,0) and F(0,j) for $i \neq 0$ and $j \neq 0$:

The final value F(n, m) contains the score of the best global alignment between x and y. To obtain an alignment corresponding to this score, we must find the path of choices that the recursion made to obtain the score. This is called a traceback.

Algorithm

Input: two sequences x and y
Output: optimal alignment and score α
Initialization: Set $F(i,0) := -i \cdot d$ for all $i = 0, 1, 2, \ldots, n$
Set $F(0,j) := -j \cdot d$ for all $j = 0, 1, 2, \ldots, m$
For $i = 1, 2, \ldots, n$ **do:**
  For $j = 1, 2, \ldots, m$ **do:**

$$\text{Set } F(i,j) := \max \begin{cases} F(i-1, j-1) + s(x_i, y_i) \\ F(i-1, j) - d \\ F(i, j-1) - d \end{cases}$$

    Set backtrace $T(i,j)$ to the maximizing pair (i', j')
The best score is $\alpha := F(n,m)$
Set $(i,j) := (n,m)$

repeat
  if $T(i,j) = (i-1, j-1)$ **print** $\binom{x_{i-1}}{y_{j-1}}$
  else if $T(i,j) = (i-1, j)$ **print** $\binom{x_{i-1}}{-}$ **else print** $\binom{-}{y_{j-1}}$
  Set $(i,j) := T(i,j)$
until $(i,j) = (0,0)$.

Complexity

We need to store (n + 1) x (m + 1) numbers. Each number takes a constant number of calculations to compute: three sums and a max.

Hence, the algorithm requires O(nm) time and memory.

For biological sequence analysis, we prefer algorithms that have time and space requirements that are linear in the length of the sequences. Quadratic time algorithms are a little slow, but feasible. O(n3) algorithms are only feasible for very short sequences.

Something to think about: if we are only interested in the best score, but not the actual alignment, then it is easy to reduce the space requirement to linear.

Local alignment: Smith-Waterman algorithm:

Global alignment is applicable when we have two similar sequences that we want to align from end-to-end, e.g. two homologous genes from related species.

Often, however, we have two sequences x and y and we would like to find the best match between substrings of both. For example, we may want to find the position of a fragment of DNA in a genomic sequence:

The best scoring alignment of two substrings of x and y is called the best local alignment. The Smith-Waterman local alignment algorithm is obtained by making two simple modifications to the global alignment algorithm.

In the main recursion, we set the value of F(i,j) to zero, if all attainable values at position (i, j) are negative:

$$F(i,j) = \max \begin{cases} 0, \\ F(i-1, j-1) + s(x_i, y_j), \\ F(i-1, j) - d, \\ F(i, j-1) - d. \end{cases}$$

The value F(i,j) = 0 indicates that we should start a new alignment at (i,j). This is because, if the best alignment up to (i,j) has a negative score, then it is better to start a new one, rather than to extend the old one.

Note that, in particular, we have F(i,0) = 0 and F(0,j) = 0 for all i = 0,1, 2,... , n and j = 0,l,2,...,m.

Instead of starting the traceback at (n,m), we start it at the cell with the highest score, argmax F(i, j). The traceback ends upon arrival at a cell with score 0, with corresponds to the start of the alignment.

For this algorithm to work, we require that the expected score for a random match is negative, i.e. that

$$\sum_{a,b \in \Sigma} q_a \cdot q_b \cdot s(a,b) < 0,$$

where qa and qb are the probabilities for the seeing the symbol a or b at any given position, respectively. Otherwise, matrix entries will tend to be positive, producing long matches between random sequences.

Algorithm

Input: two sequences x and y
Output: optimal local alignment and score α
Initialization: Set $F(i,0) := 0$ for all $i = 0, 1, 2, \ldots, n$
Set $F(0,j) := 0$ for all $j = 1, 2, \ldots, m$
For $i = 1, 2, \ldots, n$ **do:**
  For $j = 1, 2, \ldots, m$ **do:**

$$\text{Set } F(i,j) := \max \begin{cases} 0 \\ F(i-1, j-1) + s(x_i, y_i) \\ F(i-1, j) - d \\ F(i, j-1) - d \end{cases}$$

  Set backtrace $T(i,j)$ to the maximizing pair (i', j')
Set $(i,j) := \arg\max\{F(i,j) \mid i = 1, 2, \ldots, n, j = 1, 2, \ldots, m\}$
The best score is $\alpha := F(i,j)$
repeat
  if $T(i,j) = (i-1, j-1)$ **print** $\binom{x_{i-1}}{y_{j-1}}$
  else if $T(i,j) = (i-1, j)$ **print** $\binom{x_{i-1}}{-}$ **else print** $\binom{-}{y_{j-1}}$

  Set $(i,j) := T(i,j)$
until $F(i,j) = 0$.

Output Screens

The User -Interface .

The following sequences were taken as an example for implementing the algorithm.

SEQUENCE 1:

AGGCTCAGAACGCGTCCAGAAATCAGGGGAAGGAGACCCCTATCTGTCCTTCTTCTGGAAGAG
CTGGAAA

SEQUENCE 2:

ATGGGTGACT GGGGCTTCCT GGAGAAGTTG CTGGACCAGG CCAGGAGCA CTCGACCGTG

The output screens for various algorithms are shown below.

a) Optimal Alignment Using Needleman - Wunch Algorithm:

b) Optimal Alignment Using Smith - Waterman Algorithm:

c) Optimal Alignment Using Repeated - Match Algorithm:

d) Optimal Alignment Using Overlap - match algorithm:

Clearing the previous sequences and their results. A predefined example:

CONCLUSIONS AND FURTHER RESEARCH

Bio-Informatics is the study of complex biological information using computational techniques. The role of the computers in Bio-Informatics is required for their processing speed of complex data and for their problem solving power. Bio-Informatics include Molecular Biology, Bio-Physics and Computer Science, Mathematics and Statistics. This requires solidarity among Biologists, Mathematicians, Engineers and Computer scientists to provide effective solutions for scientific problems. Accelerating Biological research projects using computer databases and algorithms. In this project we have implemented the various pairwise alignment algorithms like Needleman and wunsch algorithm, Smith-waterman algorithm and designed a tool for the user through which he can input the sequences that are to be aligned, select the a particular algorithm and compute the optimal alignment along with the functional matrix and score. The results are displayed on the screen

REFERENCES

BIOINFO (2006). http://www.bioinformatics.org

SAMPLE PROJECT CODE in JAVA

```
import java.io.*;
import java.awt.*;
import java.awt.event.*;
import javax.swing.*;
import java.util.*;
abstract class Substitution {
  public int[][] score;
  void buildscore(String residues, int[][] residuescores) {
    // Allow lowercase and uppercase residues (ASCII code <= 127):
    score = new int[127][127];
    for (int i=0; i<residues.length(); i++) {
      char res1 = residues.charAt(i);
      for (int j=0; j<=i; j++) {
        char res2 = residues.charAt(j);
        score[res1][res2] = score[res2][res1]
              = score[res1][res2+32] = score[res2+32][res1]
              = score[res1+32][res2] = score[res2][res1+32]
              = score[res1+32][res2+32] = score[res2+32][res1+32]
              = residuescores[i][j];
      }
    }
  }
```

```
  abstract public String getResidues();
}
class Blosum50 extends Substitution {
  private String residues = "ARNDCQEGHILKMFPSTWYV";
  public String getResidues()
  { return residues; }
```