Dr. Rosemary Renaut, renaut@asu.edu
Director, Computational Biosciences
www.asu.edu/compbiosci
OUTLINE

– THE CBS PROGRAM AT ASU- QUICK OVERVIEW
– CBS CURRICULUM
– REQUIREMENTS
– SOME HISTORY
– FUTURE
– PROJECTS – WHAT DO THEY INVOLVE
– OUR CASE STUDIES COURSE(S)
  • INTRODUCING BASIC MATHEMATICS TO LS/CSE STUDENTS
A Professional Science Masters Program
Mathematics and Statistics
The School of Life Sciences
Computer Science Engineering
W. P Carey School Of Business
Introduction to Structural and Molecular Biology (4 credit hours)

Multivariate Statistical Analysis (3 credit hours)

Modeling and Computational Biology (4 credit hours)

Elective one (3 credit hours)

Experimental Design (3 credit hours)

Case Studies in Computational Biosciences (4 credit hours)

Elective two (3 credit hours)

Pharmaceutical Biotechnology and Medical Technology Industries (2 credit hours)

Internship (6 credit hours)

Biotechnology Venture Leadership and Entrepreneurship (2 credit hours)

Elective three (3 credit hours)

Biotechnology Venture Management and Ethics (2 credit hours)

Elective four (3 credit hours)

Complete the electives if not completed in 3 semesters or complete internship

Bioinformatics Track I (Dr. Kumar)
- Bioinformatics (3 credit hours)
- Functional Genomics (2 credit hours)
- Data Mining (3 credit hours)
- Databases (3 credit hours)
- Mathematical methods of genetic analysis (3 credit hours)

Medical Imaging Track III (Dr. Renaut)
- Mathematics of Medical Imaging (3 credit hours)
- Numerical Optimization (3 credit hours)
- Spectral and Fourier Methods (3 credit hours)
- Mathematical Biology (3 credit hours)
- Advanced Statistics (3 credit hours)

Comp. Mol. Biology Track II (Dr. Fresch)
- Functional Genomics (2 credit hours)
- Cell Biology (3 credit hours)
- Biomembranes (3 credit hours)
- Mathematical methods of genetic analysis (3 credit hours)
- Mathematical aspects of Biotechnology (3 credit hours)
- Current Topics in Genetics (2 credit hours)

Mathematical Physiology Track IV (Dr. Crook)
- Neural Modeling (3 credit hours)
- Modeling and Simulation of Physiological Systems (3 credit hours)
- Cellular Physiology (3 credit hours)
- Bioelectric Phenomena (3 credit hours)
- Mathematical Cell Physiology (3 credit hours)

Example Elective Tracks:
These are examples of available course selections. Students can tailor make a program to fit their interests.
CORE REQUIREMENTS

- Scientific Computing for Biosciences (4)
- Case Studies/ Projects in Biosciences (4)
- Structural and Molecular Biology (4)
- Statistics and Experimental Design (6)
- Business Practice and Ethics (6)
- Internship and Applied Project (6)
ELECTIVE TRACKS

– Genomics/Proteomics
– Data Mining Data Bases,
– Medical Imaging
– Molecular/Functional Genomics
– Microarray Analysis
– Individualized to meet student needs
PRE-REQUISITES

- Calculus and Differential Equations
- Basic Statistics (junior)
- Discrete Algorithms and Data Structures
- Programming skills (C++/Java)
- Cell biology, genetics (junior level)
- Organic and Bio Chemistry (junior)
- Motivation, creativity, determination!
– Interdisciplinary Training/Team Work
– Internship/Applied Project Report
– Business, Management and Ethics
– (Health Services Administration MBA)
– Small Groups/Close Faculty Involvement
– Computer Laboratory

– Extensive Project work/Consulting
DATA

• Year 3: total 62 students, enrolled now 40
• Graduates: 15 (10 left without graduating)
• Internships: NIH, ASU, Tgen, AZ Game and Fish, US Water conservation lab, AZ biodesign
• Jobs: Tgen, ASU, Independent company, Medical record keeping
• PhD programs: Biology, Computer Science, Biochemistry
OTHER DEVELOPMENTS

– **Undergraduate: NIH MARC:**
– **Quantitative Skills (sophomore) spring 05**
– **Modeling Comp Bio (Junior) spring 05**
– **Doctoral Program Computational Biosciences**

• **Molecular Cellular Biology / Mathematics**
WHAT DO WE DO: SPRING 2004

- Database Construction/Mining of Pathology Specimens (Tgen)
- Gegenbauer high resolution reconstruction for MRI, ASU
- TLS-SVM for Feature Extraction of Microarray Data, ASU
- Automated video analysis for cell behavior. Tgen
- EST DB for Marine Dinoflagellate Cryptothecodinium cohnii, ASU
- Data mining for microsatellites in ESTS from arabidopsis thaliana and brassica species (US Water Conservation Laboratory)
- The Genome Assembler- Tgen
- A user interface to support navigation for scientific discovery ASU
- Cell Migration Software Tool Tgen
WHAT DO WE DO: SPRING 2005

• EVALUATION OF BIOINFORMATICS RESOURCES (Tgen/NIH/ASU)
• Pattern recognition Automated Cytoskeleton Reconstruction (ASU)
• Develop workable database on crop Lesquerella using Integrated Crop Information Systems (ICIS) (US Water Conservation Laboratory)
• Investigation of *Xylella fastidiosa* Within an Almond Tree Population: A Model System for Golden Death (ASU: Mathecology, AZ)
• Search for Epigenetic Properties of DNA and RNA involved in X Chromosome Inactivation , (Codon Solutions LLC)
WHAT DID WE NEED FOR THESE PROJECTS

Image Analysis
Data Mining
Fourier Analysis
Modeling Differential Equations
Sequence Comparisons
Mathematics for Genetic Analysis
Statistics
Data base development: for BIOLOGICAL APPLICATIONS
Geographic Information Systems
PERL/BIOPERL/MATLAB/MYSQL
Bioinformatics: Managing Scientific Data tackles this challenge head-on by discussing the current approaches and variety of systems available to help bioinformaticians with this increasingly complex issue. The heart of the book lies in the collaboration efforts of eight distinct bioinformatics teams that describe their own unique approaches to data integration and interoperability. Each system receives its own chapter where the lead contributors provide precious insight into the specific problems being addressed by the system, why the particular architecture was chosen, and details on the system's strengths and weaknesses. In closing, the editors provide important criteria for evaluating these systems that bioinformatics professionals will find valuable.
Computational Modeling Skills Motivated by Case Studies

**Phylogenetics and Tree Building** (for the data make the tree)

<table>
<thead>
<tr>
<th></th>
<th>Human(A)</th>
<th>Chimp(B)</th>
<th>Gorilla(C)</th>
<th>Orang-Utan(D)</th>
<th>Gibbon(E)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human(A)</td>
<td>-</td>
<td>.09190</td>
<td>.1082</td>
<td>.1790</td>
<td>.2057</td>
</tr>
<tr>
<td>Chimp(B)</td>
<td>.0919/.0821</td>
<td>-</td>
<td>.1134</td>
<td>.1940</td>
<td>.2168</td>
</tr>
<tr>
<td>Gorilla(C)</td>
<td>.1057/.1083</td>
<td>.1161/.1330</td>
<td>-</td>
<td>.1882</td>
<td>.2170</td>
</tr>
<tr>
<td>Orang-Utan(D)</td>
<td>.1806/.1838</td>
<td>.1910/.1838</td>
<td>.1895/.1838</td>
<td>-</td>
<td>.2172</td>
</tr>
<tr>
<td>Gibbon(E)</td>
<td>.2067/.2142</td>
<td>.2171/.2142</td>
<td>.2156/.2142</td>
<td>.2172/.2142</td>
<td>-</td>
</tr>
</tbody>
</table>
All additive trees with 5 branches which is the correct one?
For this tree we can calculate the patristic distances between sequences:

\[ p_{BD} = e_2 + e_6 + e_7 + e_4 \]

This should match the distance from the measured data.

We do a goodness of fit for all distances:

\[ \| p - d \|^2 = \| A e - d \|^2 \]

What is \( A \), what is \( e \)? Any conditions on \( e \)?

Repeat for all trees:
- Use matlab
- Understand
- Least Squares
- Nonnegative constraints
- Constrained LS
- Exhaustive search
- Genetic Algorithms
Computational Modeling Skills Motivated by Case Studies

**Phylogenetics and Tree Building** (for the data make the tree)

<table>
<thead>
<tr>
<th></th>
<th>Human(A)</th>
<th>Chimp(B)</th>
<th>Gorilla(C)</th>
<th>Orang-Utan(D)</th>
<th>Gibbon(E)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human(A)</td>
<td>-</td>
<td>79</td>
<td>92</td>
<td>144</td>
<td>162</td>
</tr>
<tr>
<td>Chimp(B)</td>
<td>79</td>
<td>-</td>
<td>95</td>
<td>154</td>
<td>169</td>
</tr>
<tr>
<td>Gorilla(C)</td>
<td>92</td>
<td>96.7</td>
<td>-</td>
<td>150</td>
<td>169</td>
</tr>
<tr>
<td>Orang-Utan(D)</td>
<td>149.3</td>
<td>154</td>
<td>152.1</td>
<td>-</td>
<td>169</td>
</tr>
<tr>
<td>Gibbon(E)</td>
<td>166.2</td>
<td>170.9</td>
<td>169</td>
<td>169</td>
<td>-</td>
</tr>
</tbody>
</table>
An ultrametric tree: what are the distances $e_i$?
Solve the linear programming problem

$$\min L(e) = \min \sum e_i,$$

where this is the total length of the tree. Moreover each length is positive, and the total lengths are preserved

$$e_1 = e_2, \quad e_4 + e_8 = e_1 + e_6 + e_7$$

LP problem with constraints

$$\max c^T x \text{ with } Ax \leq b \quad x \geq 0$$

Students identify $x$, $c$, $b$, $A$? Use matlab: linprog
BUT THERE ARE MANY DIFFERENT TREE SHAPES AND WHICH IS CORRECT? WE NEED EXHAUSTIVE SEARCH GENETIC ALGORITHMS?
HOW WAS THIS USEFUL?

Introduction to
• data fitting,
• optimization, genetic algorithms, exhaustive search
• matlab routines,
• Realistic solutions (positive branch lengths)
• Start on some multivariable calculus to derive normal equations

OTHER APPLICATIONS USING SIMILAR TECHNIQUES

Neural networks for classification – how do they learn?
Data mining – k-means clustering – minimize energy
Gradient Descent
A Novel Genome Assembler: Using K–mers to Indirectly Perform $N^2$ Comparisons in $O(N)$

Ho-Joon Lee, Stephanie Rogers and Maulik Shah
Advisor: Jeffrey Touchman, Arizona State University and Translational Genomics Research Institute, Phoenix

The novel, exhaustive approach to genome assembly aims to eliminate traditional heuristics and indirectly compare each sequence fragment to all the others in less than $O(N^2)$ running time. The first step in the algorithm involves building a k–mer library; accomplished by scanning through each sequence fragment using a sliding window of size $k$ and cataloging each of these k–mers and the sequence fragment in which it occurs. It is assumed that neighboring fragments from the genome will share k–mers. From this k–mer table, an adjacency table is built; cataloging each sequence fragment and its neighbors. This adjacency table is generated in $O(N)$ and represents all $N^2$ comparisons. Finally, with the information in the adjacency table, multiple breadth-first searches to collect and separate the connected fragments are performed. It is assumed that there exist disjoint graphs in the adjacency table and that each such graph represents an entire contig. This process was performed on two datasets, a simulated set and a real set: both having >40,000 sequence fragments of ~1,000 kb and 9–fold coverage. In both instances, the majority of the fragments were assigned to one contig.
New technology allows hundreds of pathology specimens from human diseases to be sampled as .6mm punches of tissues that are arrayed into new “TMA” paraffin blocks; these blocks are then sectioned with microtomes to produce hundreds of slides containing hundreds of human tissue specimens (tissue microarrays, TMAs). Databases to support analysis of these high throughput TMAs will include information on diagnosis, treatment, disease response, and multiple images from follow-on studies linked to the coordinates of each of the hundreds of punches on the TMA. Data mining from the results of TMA experiments will allow text mining and image feature extraction. In this project, we present the requirements, design, and a prototype of a web based TMA database application.
During embryo development, it is essential that relatively homogenous groups of cells undergo differentiation to form spatially different patterns and eventually take on many different functions. Intercellular communication and morphogen gradients are two aspects that have been shown to play roles in determining cell fate. To understand better how these activities result in pattern formation, we utilized the NetLogo programming environment to simulate these processes. We were able to observe visually the possible pattern formation of gene expression from activation and inhibition of genes, intercellular interactions (top figure), and the exposure to morphogen gradients (middle and bottom figures). The three developmental phenomena of zebrafish embryos we studied were neurogenesis, somitogenesis, and morphogenesis during anterior/posterior patterning. The model for neurogenesis examines how autocatalysis and lateral inhibition (notch signaling) are required to form stable patterns. In addition, we investigated to what extent the geometry of the domains and the initial noise in the ‘her’ gene expression help determine a cell’s fate. To study somitogenesis, we explored how transcription and translation delays coupled with notch pathway and independent moving wave-front activity of fgf may contribute to the oscillation and synchronization of gene expression during formation of somites. Lastly, we used our models to examine how time and concentration of a morphogen gradient may signal gene activation and eventually form patterns of cells with stable and differing gene expression. Although, there is much room to study in more detail the systems of equations and the numerical analysis we used, overall this method is a good addition to the traditional methods of studying how gene-expression patterns may develop.
A variety of image artifacts are routinely observed on MRI images. We concentrate on the Gibbs Ringing that manifests itself as bright or dark rings seen at the borders of abrupt intensity change on the images. Gegenbauer High Resolution Reconstruction method has been previously shown to eliminate the undesirable ringing at the jump discontinuities in MRI. Prior work concentrated on applying this reconstruction method on frequency data obtained from reconstructed images and not on raw K-space data obtained from the MRI scanning machine. Our project work concentrates on using the raw k-space data for reconstruction and aims at comparing the reconstructed images obtained to those reconstructed from other commonly used reconstruction methods.
Clustering has recently been demonstrated to be an important preprocessing step prior to parametric estimation from dynamic PET images. Clustering, as a form of segmentation, is useful in improving the accuracy of voxel level quantification in PET images. Classical clustering algorithms such as hierarchical clustering and K-means clustering can be applied to dynamic PET data using an appropriate weighting technique. New variants of hierarchical clustering with different preprocessing criteria were developed by Dr. Guo recently. Our research focus is to validate these different algorithms with respect to their efficiency and accuracy. Different inter and intra cluster measures and statistical tests are considered to assess the quality of the different cluster results.
The goal was learn about Support Vector Machines, and explore *use of the Regularized Total Least Squares statistical method in Support Vector Machine classification of microarray data*. SVM is a special case of a Neural Net algorithm. It eliminates the rows (patient cases) of data items least valuable in determining the hyperplane until only key data items are left. Then these points are weighted, with heavier weights being given to those points that are close to the hyperplane boundary. These points are called Support Vectors. The new reduced, weighted space is called Feature Space. Finally, the trained program is given test data to see how well it can classify new patients as sick or normal.
We are interested in using PET to image brain activity in patients with Alzheimer's disease (AD). In AD studies, one way to measure disease progression is by measuring Flouro-Deoxy-Glucose (FDG), which is an analog of glucose uptake in the brain. Studies which determine a local cerebral metabolic rate (LCMR) of FDG uptake in a region of interest have proved successful in understanding AD progression. More specific information may be obtained by estimating the individual kinetic parameters which describe FDG metabolism. In particular, it is believed that the individual FDG kinetic parameters may be used for early detection of AD. We had developed an application that is used to estimate the kinetic parameters in order to be able to focus towards understanding the spatial distribution of kinetic parameters in AD, and as well as towards developing a precise measure for utilization in the early detection of AD. It uses dynamic PET data obtained from one-dimensional, two-dimensional or three-dimensional measurements. It also allows the user to compare results with respect to the computational and estimation methods, filters, constraints, and input sources chosen by the user. Comparing the results could help find out what are the optimal estimation methods, what are the constraints or what is the best filtering technique that provides optimal results. Results could be compared with expected results according to theoretical information, and an educated decision can be made on what are the optimal computational methods to use for every situation.
Answering biological queries involves the navigation of numerous richly interconnected scientific data sources. The BioNavigation system supports the scientist in exploring these sources and paths. Scientific queries can be posed at the conceptual level rather than being restricted to particular data sources. The BioNavigation interface provides a scientist with information about the available data sources and the scientific classes they represent. The user can graphically create a navigational query. BioNavigation will evaluate the query and will return a list of suitable paths through the data sources identified by the ESearch algorithm. ESearch searches a graph for paths satisfying a regular expression and ranks them using benefit and cost metrics. BioNavigation thus complements the traditional mediation approach and provides scientists with much needed guidance in selecting data sources and navigational paths.
The study of reaction center proteins from both the photosystems and the primordial reaction centers from bacteria reveal the conservation of certain amino acids. The multiple sequence alignment and phylogenetic trees created from the proteins show high degree of conserved regions in photosystem–II and bacterial reaction center–II, implying common genealogy. Also, the similarity between photosystem–I heterodimers and reaction center I homodimer proteins, indicate them having a single precursor. It is seen that even though L–M and D1–D2 show similar evolution with gene duplication, L–M proteins show step-by-step diversification whereas the other branch bifurcates into D1s and D2s just at the end. The reaction center I homodimer is placed nearly at the center between the photosystem I and II portions of the tree, suggesting it to be an ancestral type of reaction center. The structural alignment of these proteins depicts five well aligned \( \alpha \)-carbon helices. Their sequences show good amount of similarity in the hydrophobic domains forming the transmembrane helices, which are the main functional regions.

Figures:
(top) phylogenetic tree 
(bottom left) structural alignment 
(bottom right) hydropathy plots