ATLANTA
120th ASEE Annual Conference & Exposition
FRANKLY, WE DO GIVE A D*MN
June 23-26, 2013

Paper ID #5919

Genomics, Signal Processing, and Bioinformatics

Prof. Maurice F. Aburdene, Bucknell University

Maurice Felix Aburdene is a Professor of Electrical Engineering and Professor of Computer Science at Bucknell University. His teaching and research interests include control systems, parallel algorithms, simulation of dynamic systems, and signal processing.

Dr. Marie Catherine Pizzorno, Department of Biology, Bucknell University

I received BA in Biology and Chemistry from Whittier College in 1985. I received Ph.D. from the Biochemistry, Cellular and Molecular Biology program at Johns Hopkins School of Medicine in 1991. I did postdoctoral work at Princeton University from 1991 to 1995. I taught in the Biology Department of Vassar College from 1995-1996. I became an Assistant Professor in the Biology Department at Bucknell University in 1996 and was promoted to Associate Professor in 2001.My primary research interests are the molecular biology of viruses that infect eukaryotes, specifically viruses that infect honeybees.

Mr. Alexander P Thompson, Bucknell University

I am a senior Electrical Engineering student at Bucknell University. I have been studying genomic signal processing for the last two years as part of a research project through the Bucknell Presidential Fellows program. My primary role in the development of the genomic signal processing elective course is to evaluate it from the perspective of an enrolled student. In addition, I may provide assistance to Professor Aburdene in the areas of study related to my research.

Genomics, Signal Processing, and Bioinformatics

Abstract

We present our approach to teaching an elective course, ELEC 402: Genomics Signal Processing, focusing on genomics, signal processing, and bioinformatics for biology, biochemistry, electrical engineering, and computer engineering students. The course was offered for the first time in the Fall of 2012. The objective of the course was to examine current issues in mathematical biology, computational biology, and bioinformatics from an engineering perspective.

Introduction

Since today's biologists are not only scientists but also "engineers" due to their ability to "build systems," our course emphasized a multidisciplinary view of research and problem solving. This will be useful with the study of any topic, not just genomic signal processing. The content was broad and required a background in biology, mathematics, physics, chemistry, statistics, computer science and signal processing. Additionally, the course drew from the resources of several academic departments and visiting lecturers to provide fundamental concepts from each field.

A major goal of this course was to introduce students studying biology or biochemistry to the concepts and analytical techniques of engineering. Likewise, engineers were exposed to the terminology and concepts of molecular biology and genomics. One outcome of this effort was that the term projects, which constituted a major assignment in the course, were completed by mixed pairs of students from biology and engineering.

The course outline was ambitious and is expected to evolve, adapt, mutate, and crossover to other topics in the future. The prerequisite for the course was calculus II or permission of instructor. The final grade was based on homework assignments and a final term project.

Course Outcomes

The course outcomes were based on ABET's General Criteria 3[1]. We focused on the following student outcomes:

(a) an ability to apply knowledge of mathematics, science, and engineering

(d) an ability to function on multidisciplinary teams

(g) an ability to communicate effectively

(h) the broad education necessary to understand the impact of engineering solutions in a global, economic, environmental, and societal context

(i) a recognition of the need for, and an ability to engage in life-long learning

(j) a knowledge of contemporary issues

(k) an ability to use the techniques, skills, and modern engineering tools necessary for engineering practice.

Course Syllabus

The course syllabus focused on the following topics:

Introduction to the Course and Objectives
Genetic Algorithms
Genomic Companies
DNA Sequencing Methods
Bioinformatics, Introduction to Matlab® Bioinformatics Tool box
DNA Data Bases and Sequence Search Tools
Basic Statistical Computations on DNA Sequences
Using Large Electronic Datasets for Observational Clinical Studies
DNA Indicator Sequences, Periodic Signals
Fourier Transforms, Discrete Fourier Transforms (DFT), Fast Fourier Transforms (FFT)
Using Fourier Methods to Find Protein Coding Regions
Using Correlation to compare DNA sequences
DNA Filter design (Digital Filtering)
Introduction to Nuclear Magnetism and Nuclear Magnetic Resonance (NMR) basics
Fundamentals of Magnetic Resonance Imaging (MRI)
Personal Genomics: Some of the Ethical, Legal and Social Issues (ELSI) Involved With Personal Genomics.

These topics will be addressed in detail, including some homework assignments, in the next section.

Course Content

DNA has been known to be the molecule composing cellular genomes for over half a century, yet the details of exactly how the bases of DNA (adenine, guanine, thymine , and cytosine) code for all of the traits observed in living organisms are still being elucidated. The majority of the topics discussed during the class allowed both the engineering and biology students to explore various mathematical and computer techniques for analyzing DNA sequences and determining their functions.

We viewed the DNA as complementary sequences of a four character alphabet (AGTC) representing the nucleotides, as shown in Figure 1.

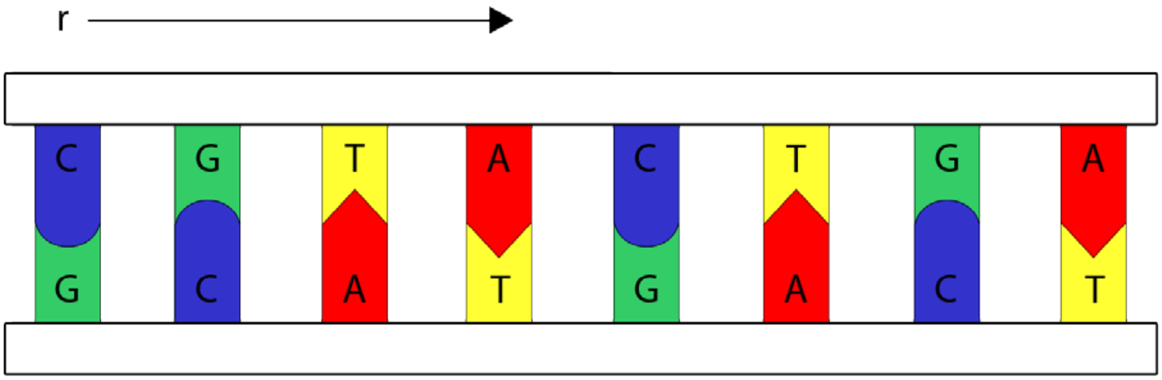


Figure 1: A representation of a flattened segment of a double-sided DNA strand

Meeting 1: Introduction[2, 3, 4, 5, 6]

This meeting consisted of an introduction to the course, of students, course objectives, and course grading. The promise of genomics was discussed by referencing articles from both the popular press and technical literature. Additional discussion points included fundamental tools, microscopes and spectrum analyzers, advances in computer chips, sequencing methods, and bioinformatics. Lastly, the students were engaged in an introductory exercise on what operations can be performed on a sequence with four symbols: A, T, G, and C.

Homework Assignment:

1) Would you please read the following article?
Hall S.S., "Revolution Postponed", Scientific American, 2010, 303, no. 4: 60-67.

2) Would you please summarize the article (in your own words) and identify what you think is significant? Indicate which points you agree with and which ones you disagree with? Assume the audience is you.

3) Would you please browse the web for comments/reviews/responses to "Revolution Postponed"?

4) Would you please submit your summary? In addition, please prepare a 3-4 slide power point presentation of your summary to share with the class.

Excerpts of a student response to Homework Assignment

"Revolution Postponed" is a critical analysis and review of the Human Genome Project. In this article, published in Scientific American, author Stephen S. Hall compares the current progress of the endeavor to its projected outcome and discusses its future potential. The overall sentiment of the article is that the extensive amount of research and resources applied to the project has produced too few valuable results.

When the Human Genome Project commenced in 2000, many anticipated that it would transform medicine and essentially lead to the cure of most human diseases within ten years. That ten year window has closed, and while the project has been hugely successful in amassing information and discoveries about DNA in general, it has not met its goal.

Hall suggests that the research has failed to deliver thus far because scientists have approached it with the wrong strategy. The primary hypothesis until now has been that people with a certain disease all have “common variants,” or similar small differences in a gene sequence. Now, the scientific community is split; some believe the common variant hypothesis is wrong while others believe it will prove successful as technology advances. According to the article, geneticist Joseph D. Terwilliger believes that common variants only result in minor biological effects and that they would have been eliminated by natural selection if they cause serious illnesses.

Hall then goes on to discuss the future of genomic research. One of the most successful studies correlating genetics to disease susceptibility was done by Helen H. Hobbs and Jonathon C. Cohen using a strategy different from common variants. They found people with a certain mutation in the PCSK9 gene, which regulates cholesterol metabolism, experienced dramatic reductions in their risk for heart disease. This is one of the most significant links between genetic variants and disease risk to date and gives promise to the future of the Human Genome Project.”

Meeting 2: DNA: Structure, Replication, and Repair
Visiting Lecturer -Professor Mitchell I. Chernin, Professor of Biology

Professor Chernin presented an introduction to the structure and functionality of DNA, an essential topic for this course, especially for the engineers who have had very little biology coursework.

Meetings 3 and 4: Genetic Algorithms[7, 8, 9, 10]

The meeting focused on optimization of functions using biologically inspired algorithms and built on the presentation “DNA: Structure, Replication, and Repair.” The concepts of population, reproduction, mutation, and crossover are used to find the optimal value of a function.

Homework Assignment:

1) Would you please find the value of x that maximizes $f(x) = 31x-x^2$ with $0 \le x \le 30$ using a genetic algorithm? Show your work and please feel free to work in teams.

2) Would you please prepare a list of five genomic/biotechnology companies and identify their: main products, stock symbols, and job advertisements. In addition, please prepare a 3-4 slide power point presentation of your summary to share with the class. The purpose of this assignment is to prepare students for meeting 6.

Meeting 5: Student Presentations

Students presented their earlier homework assignments and the class discussed the article entitled “The Survival of the Fittists: Understanding the role of replication in research is crucial for the interpretation of scientific advances”.[11]

Meeting 6: Student Presentations

Each student presented a brief overview of five genomic/biotechnology companies. The objective of this meeting was to have students see how many of these companies exist and what they do. Having students give presentations is a great way to cover this information while also motivating students to do some research. In the future, it might be best to have each student research three companies, rather than five, and focus on doing a more in-depth study of their goals, services, history, and future product development.

Meeting 7: DNA Sequencing Technologies
Visiting Lecturer- Professor Marie C. Pizzorno, Associate Professor of Biology and Cell Biology and Biochemistry Program

The presentation focused on the development of sequencing techniques, including the Sanger dideoxy method and more recent methods such as 454 pyrosequencing, Ion Torrent, and SOLiD. The presentation concluded with a comparison of the strengths and weaknesses of each method and a discussion of how these recent technologies have resulted in the ability to sequence genomes much more quickly and cheaply than ever before.

Homework Assignment:

Suggested Biology Reading (NCBI)
(http://www.ncbi.nlm.nih.gov/Class/minicourses/reading.html)
NCBI Home Page (http://www.ncbi.nlm.nih.gov/)
NCBI 4-Pack: Practical Web-based Analysis
(http://www.ncbi.nlm.nih.gov/Class/PowerTools/pack/course.html)
The GENSCAN Web Server at MIT (http://genes.mit.edu/GENSCAN.html)
IBM DNA Sequencing (http://www.youtube.com/watch?v=wvclP3GySUY&feature=relmfu)

Meetings 8 and 9: Bioinformatics: Chargaff’s Patterns (Laws, Rules) and DNA Data Bases, Sequence Search Tools, and Matlab’s Bioinformatics Tool Box[3, 12]

This discussion presented some interesting implications about the nature of genomes that students had never thought about. In the future, it might make sense to present this topic as one of the first lectures following the DNA introduction because it is simple and helps to introduce students to bioinformatics. In addition, information about the distribution of nucleotides could be represented using histograms.

Chargaff’s rules are:

1) %G and %C are equal
2) (%A)/(%T) and (%G)/(%C) ratios can differ
3) %A and %T are equal for each organism or species

Class Exercise

The following is a class exercise used to expose the students to the NCBI database
Please download the power point slide for class meetings 8 and 9.
Open another browser.
Open MATLAB.
http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi
Explore and look at direct links to some of the organisms commonly used in molecular research projects.

1) Search for an organism of interest to you. For example, search for “oryza sativa”. Get the accession number and write it down, as you will need it later.

2) Click on “nucleotide direct links” along the right hand side of the screen.

3) Click on “GenBank” in the first entry listed. Study this page.

4) Click on “FASTA”. Study this page.

5) Click on “Graphics”. Seek help from the biology students or professor in the class to interpret this graph!

6) Click on “GenBank”. We are going back to step 4 to see the material there. Advance to the next slide. Take a break and enjoy the power point picture

7) Download the Excel spreadsheet. Update the Excel with your organism and accession number to prepare for your homework assignment.

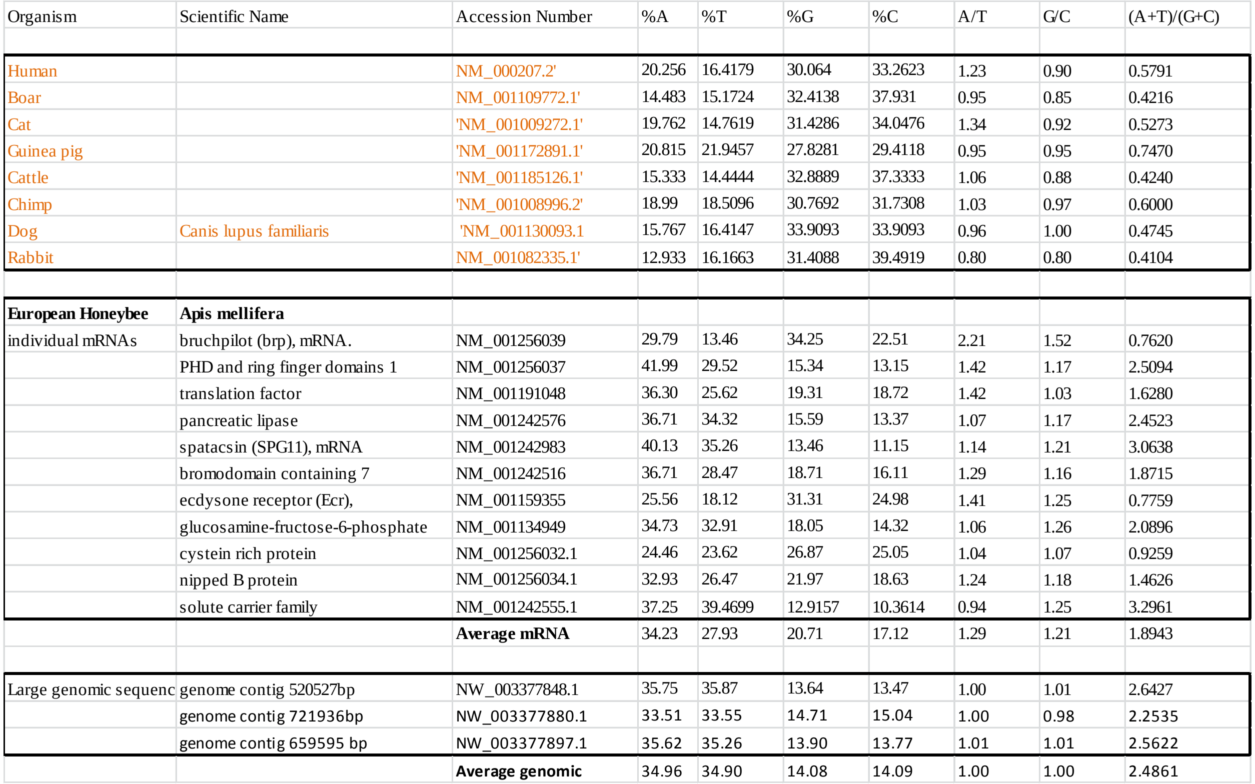
8) Download the MATLAB program, BioInfo_First_Program.m, and update it to include your accession number. Look at the output of the MATLAB program and look at the text file. You can use the text file with other programs, such as Excel to perform whatever processing you might be interested in, instead of MATLAB.

9) Download the MATLAB programs, nucleotide_counts_percents_gui.m and nucleotide_counts_percents_gui.fig and run nucleotide_counts_percents_gui.m. Enter an accession number and look at the results.

Homework Assignment:

1) Would you please complete the table below and add ten genes of one organism?

2) Would you please summarize your results? Please look up Chargaff’s patterns, laws, rules, and Waclaw Szybalski observations. Are your results consistent with these laws and observations?



Organism	Scientific Name	Accession Number	%A	%T	%G	%C	A/T	G/C	(A+T)/(G+C)
Human		NM_000207.2'	20.256	16.4179	30.064	33.2623	1.23	0.90	0.5791
Boar		NM_001109772.1'	14.483	15.1724	32.4138	37.931	0.95	0.85	0.4216
Cat		'NM_001009272.1'	19.762	14.7619	31.4286	34.0476	1.34	0.92	0.5273
Guinea pig		'NM_001172891.1'	20.815	21.9457	27.8281	29.4118	0.95	0.95	0.7470
Cattle		'NM_001185126.1'	15.333	14.4444	32.8889	37.3333	1.06	0.88	0.4240
Chimp		'NM_001008996.2'	18.99	18.5096	30.7692	31.7308	1.03	0.97	0.6000
Dog	Canis lupus familiaris	'NM_001130093.1	15.767	16.4147	33.9093	33.9093	0.96	1.00	0.4745
Rabbit		NM_001082335.1'	12.933	16.1663	31.4088	39.4919	0.80	0.80	0.4104
European Honeybee	**Apis mellifera**								
individual mRNAs	bruchpilot (brp), mRNA.	NM_001256039	29.79	13.46	34.25	22.51	2.21	1.52	0.7620
	PHD and ring finger domains 1	NM_001256037	41.99	29.52	15.34	13.15	1.42	1.17	2.5094
	translation factor	NM_001191048	36.30	25.62	19.31	18.72	1.42	1.03	1.6280
	pancreatic lipase	NM_001242576	36.71	34.32	15.59	13.37	1.07	1.17	2.4523
	spatacsin (SPG11), mRNA	NM_001242983	40.13	35.26	13.46	11.15	1.14	1.21	3.0638
	bromodomain containing 7	NM_001242516	36.71	28.47	18.71	16.11	1.29	1.16	1.8715
	ecdysone receptor (Ecr),	NM_001159355	25.56	18.12	31.31	24.98	1.41	1.25	0.7759
	glucosamine-fructose-6-phosphate	NM_001134949	34.73	32.91	18.05	14.32	1.06	1.26	2.0896
	cystein rich protein	NM_001256032.1	24.46	23.62	26.87	25.05	1.04	1.07	0.9259
	nipped B protein	NM_001256034.1	32.93	26.47	21.97	18.63	1.24	1.18	1.4626
	solute carrier family	NM_001242555.1	37.25	39.4699	12.9157	10.3614	0.94	1.25	3.2961
		Average mRNA	34.23	27.93	20.71	17.12	1.29	1.21	1.8943
Large genomic sequenc	genome contig 520527bp	NW_003377848.1	35.75	35.87	13.64	13.47	1.00	1.01	2.6427
	genome contig 721936bp	NW_003377880.1	33.51	33.55	14.71	15.04	1.00	0.98	2.2535
	genome contig 659595 bp	NW_003377897.1	35.62	35.26	13.90	13.77	1.01	1.01	2.5622
		Average genomic	34.96	34.90	14.08	14.09	1.00	1.00	2.4861

3) Would you please submit your table in an Excel spreadsheet and summary? In addition, please prepare a PowerPoint presentation of your summary to share with the class.

Meetings 10-13: Introduction to Probability and Statistics13, 14, 15

These meetings consisted of an overview of a few probability concepts including random variables, probability density function, probability distribution function, mean, standard deviation, and variance. In the future, more time might be devoted to relating these concepts to applications in genomics, sequence analysis, and similarity of genes.

Class Exercise

1) Let us consider the length of the gene as our random variable, X. Based on your earlier homework assignment, what values does X take?

2) Obtain the histogram of X. What is the average value of X? What is the variance of X? Based on your earlier homework assignment, what is the probability of obtaining an A, T, G, and C for your longest gene?

3) Using the probabilities above, what is the probability of getting all A's in a gene of length N? Similarly, what is the probability for getting all T's? What assumptions did you make?

4) What is the probability of getting A, only at the beginning of the DNA sequence of length N? (i.e. A and then any combination not including A)

5) What is the probability of getting AA, only at the beginning of the DNA sequence of N? (i.e. AA and then any combination not including A)

6) What is the probability of getting AAA, only at the beginning of the DNA sequence of length

N? (i.e. AAA and then any combination not including A)
7) Let the length of the DNA sequence be N. What is the probability of getting k A's at the beginning of the sequence, where k is less than or equal to N?
8) Let the length of the DNA sequence be N. What is the probability of getting an A at the kth position of the sequence, where k is less than or equal to N? (Geometric distribution, memoryless discrete distribution)
9) Let the random variable X be the position of the first A in the DNA sequence. Plot the probability density function of X (pdf).
10) Using the probabilities found in the homework, suppose we have a three nucleotide sequence.
11) What is the probability of getting one A only? What is the probability of getting two A's only? What is the probability of getting three A's only? Assuming the random variable is the number of A's, plot the probability density function of X (pdf). What assumptions did you make?
12) Assume we have a four nucleotide sequence, and repeat exercise 11.
Assume we have an N nucleotide sequence, and repeat exercise 11. (Binomial distribution)
13) Based on your earlier homework assignment, what missing nucleotide would you place in the following sequence? AA-TG…?

Homework Assignment:

Last week you completed the table on the preceding page using ten genes of one organism.
Let us assume that A, T, G, and C are the outcomes of your experiment are the A, T, G, and Cs obtained from a gene. Please assign values to g(A) = x_1, g(T) = x_2, g(G) = x_3, and g(C) = x_4. The random variable X takes on values {x_1, x_2, x_3, x_4}.

1) Would you please find the probability density function of X for your organism?
2) Would you please find the probability distribution function of X for your gene?
3) Would you please find the expected value of the random variable X?
4) Find the variance of the random variable X.
5) Would you please repeat the above for a gene from the same organism?
6) Would you please repeat the above for a gene from a different organism?
7) Would you please submit your results and summary of your findings?

Meeting 14: Using Large Electronic Datasets for Observational Clinical Studies
Visiting Lecturer- Dr. Jove Graham, Operations Manager, Biostatistics and Research Data Core, Geisinger Center for Health Research

The presentation focused on the Geisinger Health System (Anatomy 101), working with clinical databases, addressing confounding in "Big Data," and a case study on hip fractures. Although this presentation did not focus on genomics, the methodology and approach to "Big Data" was very informative and new to students.

Meeting 15: Genome Analysis with Inter-nucleotide Distances[16]

Internucleotide distances are an important metric used in comparing gene sequences, so they definitely need to be discussed prior to topics like the Euclidean distance. However, in the future, this topic it could easily be combined with the Euclidean distance presentation and students would still comprehend the material well enough.

Homework Assignment:

Reading: Afreixo V., Bastos C.A.C., Pinho A.J., Garcia S.P., and Ferreira P.J.S.G. 2009. "Genome Analysis with Inter-Nucleotide Distances". Bioinformatics. 25, no. 23: 3064-3070.

Puzzle

As an amusing exercise to keep your mind on genomics over the break, identify all the possible common English words that can be made, using only the letters A, C, T, and G, with repeats of letters allowed, i.e. gaga. Please list all the words you find. Have fun!

For even more fun: Search the web and find the number of words in a particular English dictionary. What is the probability of finding a two-letter word using the letter A, C, T, and G? What is the probability of finding: three-letter words; four letter words; five letter words?

Meetings 16-28 Genomic Signal Processing[4, 5, 6, 17, 18]

These meetings focused on the topics of sequences, periodic sequences, energy and power, digital filters, discrete Fourier series, discrete Fourier transform, fast Fourier transforms (FFT), FFT spectrograms, correlation, convolution, and using Euclidean distances.

These topics are very complex. In a class with many non-engineers who have not taken any signal processing, teaching the Fourier transform is challenging. If the goal is to get everyone in the class to understand how these methods work mathematically, then more time needs to be devoted to the subject. If, however, the goal is to give everyone a basic understanding of how the methods work, then the formulas and procedures do not need to be discussed so thoroughly. If this is the case, it might be better, for example, to explain in general how the Fourier transform is used to identify the primary frequencies that make up a signal and then spend more time talking about applications of this method in genomic signal processing.

The following is an example of an in-class exercise used to demonstrate the use of various signal processing techniques in genome analysis.

Class Exercise

Why do we need all this math? The objective is to look for gene similarity and protein coding regions.

1) Let x[n]=[1 1 1 1 0 0 0] and y[n]=[0 0 0 0 1 1 0 0]
2) Plot x[n] vs. n.

3) Plot x[n+1] vs. n.
4) Plot x[n-1] vs. n
5) Correlation: $w[n] = \sum_{k=0}^{7} \mathrm{x[k]}y[n+k]$
6) Plot w[n] vs. n.
7) Convolution: $z[n] = \sum_{k=0}^{7} \mathrm{x[k]}y[n-k]$
8) Plot z[n] vs. n.
9) Find the Fourier transform of x[n], X[k]. fft()
10) Reverse y[n], $y[n]^R$=[1 1 0 0 0 0 0 0]
11) Find the Fourier transform of $y[n]^R$, $Y^R[k]$.
12) Multiply (point by point): W[k]=X[k]* $Y^R[k]$.
13) Find the inverse of W[k]. ifft()
14) Discuss your results.
Try the above using two similar genes.

Meeting 28: Introduction to Basic Local Alignment Search Tool (BLAST)[19]

The alignment search tool finds regions of local similarity between sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches. BLAST can be used to infer functional and evolutionary relationships between sequences , as well as help identify members of gene families. Appendix A contains a short tutorial.

Meetings 29-30: Bioinformatics
Visiting Lecturer- Professor Brian King, Assistant Professor Computer Science Department

The presentation focused on:

1) What is "bioinformatics"?
2) Motivation for Sequence Alignment
3) Scoring a Sequence Alignment
4) Substitution matrices
5) Common methods: Dot matrix analysis, Dynamic programming algorithms, and Word or *k*-tuple methods

In addition, this presentation showed how the alignment methods can be applied to both DNA an amino acid sequences to investigate evolutionary or functional relationships between sequences from different organisms.

Meeting 31: Network Analysis of Biological Systems
Visiting Lecturer- Steven Steinway, MD/PhD Candidate, Penn State College of Medicine, Hershey, PA

The presentation focused on terminology used to describe networks, graphs, protein interaction networks, and how network models may be translated into a set of Boolean functions. In addition, Dr. Steinway discussed a specific cancer-related protein network to which he is applying network analysis to understand in greater detail.

Meeting 32: Molecular Evolution and Ecology of Aquatic Insects, from Bora Bora to Blackfoot Glacier
Visiting Lecture- Professor Steve Jordan, Associate Professor of Biology

The presentation focused on using DNA sequences from both mitochondrial genes and the somatic gene EF-1α determine the phylogenetics of two insect populations: damselflies in the Hawaiian islands and stoneflies in Montana. The data presented demonstrated how DNA sequences can be used to show evolution of species across both time and geography.

Meetings 33-35: Review of Technical Articles and Peer Review of Term Papers/Projects

These sessions were used for peer review of drafts of term project papers and oral presentations. Each paper was reviewed by multiple students and faculty who provided both technical and editorial comments.

Homework Assignment:

Would you please read the following articles?

Anastassiou, D., "Genomic signal processing," Signal Processing Magazine, IEEE , vol.18, no.4, pp.8-20, Jul 2001.

Afreixo, Vera, et al. "Genome analysis with inter-nucleotide distances." *Bioinformatics* 25.23 (2009): 3064-3070.

Would you please summarize your readings? Please prepare a PowerPoint presentation of your summary to share with the class.

Class Discussion

Anastassiou, D. , "Genomic signal processing," Signal Processing Magazine, IEEE , vol.18, no.4, pp.8-20, Jul 2001.

a) "Genomic signal processing (GSP) creates a paradigm shift." What does Anastassiou mean by it?
b) "Scientific and technological will be related to processing and interpretation of vast amount of data." Did Anastassiou predict BIG DATA initiatives and Analytics?
c) "Genomic information is digital in the real sense." What does Anastassiou mean by it?
d) "GSP did not have a significant impact on "genomics/biology/biomolecular" is because it deals with numerical sequences rather than strings". What solution(s) does Anastassiou propose?
e) "Fourier transforms can be used to predict important features. " What features does Anastassiou list?

f) Main concepts of biology. What is the nature of biomolecular sequences? What governs protein synthesis? How many possible triplets are there? How many amino acids are there? Why is there a difference in these numbers? What is a start codon? What is a STOP codon? What do we mean by gene regulation? What governs the rate of activation of genes?
g) Databases. Where are the data bases located? What is the National Institute of Health? Who supports it? What does NCBI refer to?
h) Character strings described by Numerical Sequences. How do we change a string to a numerical sequence? How do we represent the DNA sequence of length N? How would we represent the complementary DNA sequence?
i) Assume we used the complex conjugate notation of a, t, g, and c. What would be the sum of a and t? Reflecting on our class work? What would be the sum of a DNA sequence?
j) A digital filter is used to classify a codon. How is it done?
k) What is a DFT? What does it tell us? What is an FFT? What does it tell us?
l) Spectrograms. What do spectrograms tell us?
m) Let the length of a DNA sequence be N. “The DFT at k=N/3 corresponds to ta period of three samples the length of each codon.” What does this mean?
n) Reading frames. What do we mean by reading frames? Why do different reading frames exhibit different statistical characteristics? Why do different reading frames have different phase characteristics?
o) What are the advantages of DSP tools compared with other computational tools in the genomics/biomolecular area?

Afreixo, Vera, et al. "Genome analysis with inter-nucleotide distances." Bioinformatics 25.23 (2009): 3064-3070.

a) What do we mean by inter-nucleotide distance? What is the probability distribution of inter-nucleotide distances? What assumptions did the authors make?
b) How did the authors facilitate the visual comparison of various distance distributions with the theoretical distribution? What is relative error?
c) It is known that the coding regions have different characteristics than the complete genome. How did the authors demonstrate the difference?

Meetings 36-37: Introduction to Nuclear Magnetism and NMR Basics
Visiting Lecturer- Professor David Rovnyak, Associate Professor of Chemistry

The presentation focused on an introduction to nuclear magnetism and NMR basics as well as the use of FFTs. This included the modification of NMR for use in medical MRI technology and how it can be used to detect various body tissues and brain functions.

Meeting 38:Personal Genomics
Visiting Lecturer- Warren C. Lathe III, Ph.D. (Trey), AAAS Science and Technology Policy Fellow, National Science Foundation, Computer and Information Science and Engineering

The presentation focused on some of the ethical, legal and social issues involved with personal genomics. These topics included informed consent for those obtaining their genome sequence, the ethical issues involved as our understanding of the human genome changes over time, and legal issues that could arise from patenting gene sequences that are associated with specific diseases.

The remainder of the semester was devoted to student presentations of their term projects and course evaluation.

Term Projects

Five term projects were completed by submitting a written paper and an oral presentation. Three projects were done by mixed pairs of students majoring in biology/biochemistry and electrical/computer engineering. The other two pairs were composed only of electrical and computer engineering students. The titles of the projects were:

1. Biomedical Applications of Nanoparticles (Biochemistry and computer engineering students)
2. Viral DNA Comparisons (Biochemistry and computer engineering students)
3. Comparative Study of FIV and HIV Viruses (Biology and electrical engineering students)
4. Codon Sequencing/Heat Maps (Electrical engineering students)
5. Performance Comparison of Genetic Algorithms to Other Global Optimization (Electrical engineering students)

Assessment

Table I presents the results of an anonymous survey at the end of the semester. A total of 6 students responded to the survey and rated each course outcome on a scale of 1 to 5. A rating of 1 indicates that the students “disagree strongly” and rating of 5 indicates the students “agree strongly” in meeting the course outcome.

Table I: Results of the end-of-semester anonymous survey

	1	**2**	**3**	**4**	**5**	**Mean**	**Median**
The course has increased my ability to apply knowledge of mathematics, science, and engineering.				4	2	4.33	4
The course has increased my ability to function on multidisciplinary teams		1		4	1	3.83	4
The course has increased my ability to communicate effectively		1	3	2		3.17	3
The course has increased my ability to value the broad education necessary to understand the impact of engineering /biology solutions in a global, economic, environmental, and societal context				1	5	4.83	5
The course has increased my recognition of the need for, and an ability to engage in life-long earning.				3	3	4.5	4.5
The course has increased my knowledge of contemporary issues.				2	4	4.67	5
The course has increased my ability to use the techniques, skills, and modern engineering/biology tools necessary for engineering/biology practice.				4	2	4.33	4

Summary

We presented our approach to the course “ELEC 402: Genomics Signal Processing” including some suggestions for improvements. Signal processing, genomics, and bioinformatics are becoming an important component of both engineers and biologists’ education.

Acknowledgments

With many thanks we acknowledge to the contributions of the following guest speakers for their exciting presentations: Professor Mitchell I. Chernin, Professor Erik Cheever, Professor Brian R. King, Professor Steve Jordan, Professor David Rovnyak, Dr. Warren C. Lathe III, Dr. Jove Graham, and Mr. Steven Steinway.

References

1. ABET, http://www.abet.org/DisplayTemplates/DocsHandbook.aspx?id=3149 (accessed January 3, 2013).
2. Hall S.S., “Revolution Postponed”, Scientific American, 303(4), pp. 60-67, 2010.
3. Ian Stewart, The Mathematics of Life, Joat Enterprises, 2011.
4. Vaidyanathan, P.P., “Genomics and Proteomics: A Signal Processor's Tour,” Circuits and Systems Magazine, IEEE. 4(4), pp.1, Fourth Quarter 2004.
5. Dougherty, E.R.; Datta, A.; Sima, C. , “Research issues in genomic signal processing,” Signal Processing Magazine, IEEE. 22(6), pp. 46- 68, Nov. 2005.
6. Anastassiou, D. , “Genomic signal processing,” Signal Processing Magazine, IEEE 18(4), pp.8-20, Jul 2001
7. Shubha, K. R., and Maurice F. Aburdene. “Parameter Estimation for Second-Order Systems Using a Genetic Algorithm Technique.” In JCIS, pp. 180-183, 2002.
8. Mitchell, Melanie. An Introduction to Genetic Algorithms. Cambridge, Mass: MIT Press, 1996. <http://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=1337>.
9. Sivanandam, S. N., and S. N. Deepa, Introduction to Genetic Algorithms. Berlin: Springer, 2007.
10. Holland, J. H., “Genetic algorithms”, Scientific American, 267(1), pp. 66-72, 1992.
11. Wainer, H. “The Survival of the Fittists.” American Scientist. 100(5), pp.358-361, 2012.
12. Wayne M. Becker, W. M., Kleinsmith, L.J., and Hardin, J., The World of the Cell, 5th Edition, Benjamin Cummings, 2002.
13. Yates, R.D., and Goodman, D.J., Probability and Stochastic Processes: A Friendly Introduction for Electrical and Computer Engineers, Hoboken, NJ: John Wiley & Sons, 2005.
14. Lipschutz, S., and Schiller, J.S. Schaum's Outline of Theory and Problems of Introduction to Probability and Statistics, New York: McGraw Hill, 1998.
15. Aburdene, M.F. Computer Simulation of Dynamic Systems, Dubuque, Iowa: Wm. C. Brown, 1988.
16. Afreixo, V., C.A.C. Bastos, A.J. Pinho, S.P. Garcia, and P.J.S.G. Ferreira. “Genome Analysis with Inter-Nucleotide Distances.” Bioinformatics, 25(23), pp. 3064-3070, 2009.
17. Hsu, Hwei P. Schaum's outline of theory and problems of signals and systems, McGraw-Hill, 1995.
18. Blanford, D. and Parr, J., Introduction to Digital Signal Processing, Pearson Education, Inc., Pearson Education, Inc., 2012.
19. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J., "Basic local alignment search tool." J. Mol. Biol. 215, pp. 403-410, 1990.

Appendix A: Introduction to BLAST

Source: http://blast.ncbi.nlm.nih.gov/Blast.cgi
BLAST: Additional Information

Accepted Input Formats

Query sequence(s) to be used for a BLAST search should be pasted in the **'Search'** text area. It accepts a number of different types of input and automatically determines the format or the input. To allow this feature there are certain conventions required with regard to the input of identifiers (e.g., accessions or gi's). These are described in 3) below. Accepted input types are FASTA, bare sequence, or sequence identifiers.

1. **FASTA**

A sequence in FASTA format begins with a single-line description, followed by lines of sequence data. The description line (defline) is distinguished from the sequence data by a greater-than (">") symbol at the beginning. It is recommended that all lines of text be shorter than 80 characters in length. An example sequence in FASTA format is:

```
>gi|129295|sp|P01013|OVAX_CHICK GENE X PROTEIN (OVALBUMIN-RELATED)
QIKDLLVSSSTDLDTTLVLVNAIYFKGMWKTAFNAEDTREMPFHVTKQESKPVQMMC
MNNSFNVATLPAEKMKILELPFASGDLSMLVLLPDEVSDLERIEKTINFEKLTEWTNPNT
MEKRRVKVYLPQMKIEEKYNLTSVLMALGMTDLFIPSANLTGISSAESLKISQAVHGAF
MELSEDGIEMAGSTGVIEDIKHSPESEQFRADHPFLFLIKHNPTNTIVYFGRYWSP
```

Blank lines are not allowed in the middle of FASTA input.

Sequences are expected to be represented in the standard IUB/IUPAC amino acid and nucleic acid codes, with these exceptions: lower-case letters are accepted and are mapped into upper-case; a single hyphen or dash can be used to represent a gap of indeterminate length; and in amino acid sequences, U and * are acceptable letters (see below). Before submitting a request, any numerical digits in the query sequence should either be removed or replaced by appropriate letter codes (e.g., N for unknown nucleic acid residue or X for unknown amino acid residue). The nucleic acid codes supported are:

A adenosine	C cytidine	G guanine
T thymidine	N A/G/C/T (any)	U uridine
K G/T (keto)	*S G/C (strong)*	*Y T/C (pyrimidine)*
M A/C (amino)	*W A/T (weak)*	*R G/A (purine)*
B G/T/C	*D G/A/T*	*H A/C/T*
V G/C/A	- *gap of indeterminate length*	

For those programs that use amino acid query sequences (BLASTP and TBLASTN), the accepted amino acid codes are:

A alanine	P proline
B aspartate/asparagine	Q glutamine
C cystine	R arginine
D aspartate	S serine
E glutamate	T threonine
F phenylalanine	*U selenocysteine*
G glycine	V valine

H	histidine	W	tryptophan
I	isoleucine	X	any
K	lysine	Y	tyrosine
L	leucine	Z	glutamate/glutamine
M	methionine	*	translation stop
N	asparagine	-	*gap of indeterminate length*

NOTE:
[1] The degenerate nucleotide codes in red are treated as mismatches in nucleotide alignment. Too many such degenerate codes within an input nucleotide query will cause *blast.cgi* to reject the input. For protein queries, too many nucleotide-like code (A,C,G,T,N) may also cause similar rejection.
[2] For protein code, U is replaced by X first before the search since it is not specified in any scoring matrices.
[3] *blast.cgi* will not take "-" in the query. To represent gaps, use a string of N or X instead.

2. **Bare Sequence**

This may be just lines of sequence data, without the FASTA definition line, e.g.:

```
QIKDLLVSSSTDLDTTLVLVNAIYFKGMWKTAFNAEDTREMPFHVTKQESKPVQMMC
MNNSFNVATLPAEKMKILELPFASGDLSMLVLLPDEVSDLERIEKTINFEKLTEWTNPNT
MEKRRVKVYLPQMKIEEKYNLTSVLMALGMTDLFIPSANLTGISSAESLKISQAVHGAF
MELSEDGIEMAGSTGVIEDIKHSPESEQFRADHPFLFLIKHNPTNTIVYFGRYWSP
```

It can also be sequence interspersed with numbers and/or spaces, such as the sequence portion of a GenBank/GenPept flatfile report:

```
1   qikdllvsss tdldttlvlv naiyfkgmwk tafnaedtre mpfhvtkqes kpvqmmcmnn
61  sfnvatlpae kmkilelpfa sgdlsmlvll pdevsdleri ektinfeklt ewtnpntmek
121 rrvkvylpqm kieekynlts vlmalgmtdl fipsanltgi ssaeslkisq avhgafmels
181 edgiemagst gviedikhsp eseqfradhp flflikhnpt ntivyfgryw sp
```

Blank lines are not allowed in the middle of bare sequence input.