CAP5510 Introduction to Bioinformatics

Fall 2009

Term Projects

Bioinformatics is a rapidly growing field. Several new research areas emerged just within the last 5 to 10 years. Some of these topics have already trickled down to graduate curriculum in several universities. With limited time in a semester, it is not possible to cover all these important topics.

In the term project, several options will be given and you are supposed to choose one of them, as follows:

A. Tutorial Paper: Choose one of the following topics and write a self-contained paper.

1. Proteomics and Protein Analysis
2. Genomic Analysis of RNA
3. Functional Genomics
4. Comparative Genomics and Genome Rearrangements
5. Human Genome
6. Structural Alignments
7. Bacteria and Archaea Genomes
8. Gene Expression: Microarray Data Analysis
9. Protein Sequencing and Mass Spectroscopy technology
10. High Throughput Genomic Sequence
11. SNPs (Single Nucleotide Polymorphism) and Variation

B. Tools: Choose one of the major topics in Bioinformatics and find all the tools (except the tools that we have already discussed in the class like BLAST) available from important internet sites. You must include descriptions of the algorithms underlying these tools and need to research the papers where such descriptions were published. Explain the algorithms and then write a brief "user's manual" describing how to use the tools and interpret the results. You must provide test runs (identify the relevant databases) of the tools with currently available biological databases for all the tools identified.

C. Research Paper: Choose a research topic in Bioinformatics that might interest you, write a critical review of this field and then identify an open problem or an approach to improve the idea or the tool and then implement your own suggested improvement. The paper should be written in a professional style as if you are writing a chapter of a book or a technical paper for a conference, taking 10-15 pages (excluding Figures), in electronic form and a set of slides for presentation in the class based on the paper.

CAUTION: You should try to write the chapter or the paper using your own language as much as possible. If you have to quote something or adopt a Figure from some source, you must acknowledge the source. Copying large parts from a source and pasting it in your paper can be detected easily. It is also illegal because it violates copyright laws and if such activities are found, your grade will be heavily penalized.

Deadlines: **November 2, 2009**: Submit a proposal outlining your project; identify the sources that you will use and a plan to execute the project.(maximum of two pages)

November 16, 2009: Submit a first draft of your report electronically.

November 18, 2009: The draft will be retuned to you with comments.

Class presentation: November 23 and November 25, 2009. You have to sign up; send me email giving your preferred date (you may or may not get your preferred date). The final report is due one day before your presentation date. We may have to allocate additional time beyond our regular class time to cover all the projects.**Teams:** You may team up with another student in the class if you can justify such a proposal by showing the enlarged scope of the project and the amount of work. The report size must then be 15-20 pages and each member must write approximately 50% of the project. (Include a statement giving who wrote what sections.)

A set of starting references are given below that includes a couple of books and a few journal papers. You should consult at least two books and three journal papers. Another great source of resources is the Internet and you can obtain huge amount of basic or advanced research information browsing the Internet.

General References Textboks

1. Parts of chapters from the text by Jones and Pevzner, "Bioinformatics Algorithms".
2. **Jonathan Pevsner, "Bioinformatics and Functional Genomics", Wiley-Blackwell, 2009**
3. N. Cristianini and Matthew W. Hahn, "Introduction to Computational Genomics: A Case Studies Approach" G. Gibson and S.P. Muse, "A Primer on Genome Science" Sinauer Publishing, 2002. (You may borrow my personal copy.)
4. Xiaohua Hu and Yi Pan(Editors), "Knowledge Discovery in Bioinformatics", John Wiley, 2007
5. Shuba Gopal, Anne Haake, Rhys Price Jones, Paul Tyman, "Bioinformatics: A Computing Perspective" McGraw Hill, 2008
6. Thomas A. Creighton, "Proteins: Structure and Molecular Properties", Freeman and Co.,, latest edition, 1993.

Proteomics and Functional Genomics

Proteomics is concerned with the study of structure of protein (secondary and tertiary structures due to folding), protein expression and protein-protein interactions in biological pathways (Systems Biology). Functional Proteomics is concerned with gene annotation and assessing the gene function using the information provided by the structures.

Relevant references are 1, 3, 5, and 6 and the following sites:
http://www.nature.com/nature/insights/6928.html
http://www.ncbi.nlm.nih.gov/pubmed/15231748

The first site points to several basic publications published in *Nature* with full text (pdf) posted.

This project can be undertaken by a tem of two students.

Structural Alignment

Structural alignments can help detect distant evolutionary relationships that are hard or impossible to discern from protein sequences. The problem has been formulated in terms of structural alignment of proteins and expressed as a family of optimization problems.

Structural alignment
Approximate protein structural alignment in polynomial time
Rachel Kolodny*,â€ and Nathan Linialâ€¡

link: http://www.pnas.org/content/101/33/12201

There are several references that are cited in this paper and full text is posted.
Also, general references 3, 5 and 6 are relevant.

Protein Sequencing and mass spectroscopy technology

Mass spectroscopy is the most advanced high speed and very accurate technology for sequencing protein and identifying protein from its spectral graph.

SEQUEST
Jimmy K. Eng, Ashley L. McCormack, and John R. Yates, III (1994). "An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database". J Am Soc Mass Spectrom 5: 976â€“989

link:
http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6TH2-44FNFDS4&user=4429&_rdoc=1&_fmt=&_orig=search&_sort=d&view=c&_version=1&_urlVersion=0&_userid=4429&md5=9a7e8a00c8180131eadbe129a99f9ae5

Daniel Liebler, Introduction to Proteomics : Tools for New Biology (posted on line)

http://books.google.com/books?id=akCti9PoMtEC&pg=PA100&lpg=PA100&dq=Jimmy+Eng+and+John+Yates+SEQUEST&source=web&ots=2018Bap6uE&sig=4r1z22Ahv85nsucjzX9ftwCRiFw&hl=en&sa=X&oi=book_result&resnum=6&ct=result#PPP1,M1

The book by Liebler is posted on-line at the above link. This referece is also useful for projects 1 and 2.

SNPs (Single Nucleotide Polymorphism) and Variation

A single nucleotide **polymorphism**, or SNP (pronounced *snip*), is a DNA sequence variation occurring when a single nucleotide - A, T, C, or G - in the genome differs between members of a species (or between paired chromosomes in an individual). For example, two sequenced DNA fragments from different individuals, AAGCCTA to AAGCTTA, contain a difference in a single nucleotide. In this case we say that there are two *alleles* : C and T. Almost all common SNPs have only two alleles.

Within a population, SNPs can be assigned a minor allele frequency - the ratio of chromosomes in the population carrying the less common variant to those with the more common variant. It is important to note that there are variations between human populations, so a SNP allele that is common in one geographical or ethnic group may be much rarer in another. In the past, single nucleotide polymorphisms with a minor allele frequency of ≥ 1% (or 0.5%, etc.) were given the title "SNP," an unwieldy definition. Single nucleotide polymorphisms may fall within coding sequences of genes, non-coding regions of genes, or in the intergenic regions between genes. SNPs within a coding sequence will not necessarily change the amino acid sequence of the protein that is produced, due to degeneracy of the genetic code. A SNP in which both forms lead to the same polypeptide sequence is termed *synonymous* (sometimes called a silent mutation) - if a different polypeptide sequence is produced they are *non-synonymous*. SNPs that are not in protein-coding regions may still have consequences for gene splicing, transcription factor binding, or the sequence of non-coding RNA.

http://www.ncbi.nlm.nih.gov/About/primer/phylo.html

Chapter 5 of general reference 3 (Many references cited at the end of the chapter)

Comparative Genomics
From Wikipedia, the free encyclopedia

Comparative genomics is the study of the relationship of genome structure and function across different biological species or strains. Comparative genomics is an attempt to take advantage of the information provided by the signatures of selection to understand the function and evolutionary processes that act on genomes. While it is still a young field, it holds great promise to yield insights into many aspects of the evolution of modern species. The sheer amount of information contained in modern genomes (750 megabytes in the case of humans) necessitates that the methods of comparative genomics are automated. Gene finding is an important application of comparative genomics, as is discovery of new, non-coding functional elements of the genome.

Comparative genomics exploits both similarities and differences in the proteins, RNA, and regulatory regions of different organisms to infer how selection has acted upon these elements. Those elements that are responsible for similarities between different species should be conserved through time (stabilizing selection), while those elements responsible for differences among species should be divergent (positive selection). Finally, those elements that are unimportant to the evolutionary success of the organism will be unconserved (selection is neutral).

http://en.wikipedia.org/wiki/Comparative_genomics

1. Cristianini, N. and Hahn, M. *Introduction to Computational Genomics*, Cambridge University Press, 2006. (ISBN-13: 9780521671910 | ISBN-10: 0521671914) [General ref.no.2)

- Kellis M, Patterson N, Endrizzi M, Birren B, Lander E (2003). Sequencing and Comparison of yeast species to identify genes and regulatory elements. *Nature*, pp. 241-254 (15 May 2003).
- Filipski A, Kumar S (2005). Comparative genomics in eukaryotes. In The Evolution of the Genome (ed. T.R. Gregory), pp. 521-583. Elsevier, San Diego.
- Gregory TR, DeSalle R (2005). Comparative genomics in prokaryotes. In The Evolution of the Genome (ed. T.R. Gregory), pp. 585-675. Elsevier, San Diego.

Term Project B

If you choose a "tools" project, go to the homepage of references 2 and 3
http://www.computational-genomics.net/

Also, the NCBI site is the storehouse of all tools.

http://www.ncbi.nlm.nih.gov/

You can go to any one of the general references and many of the tools are either mentioned or discussed briefly.

Term Project C: Talk to me first before you undertake such a project.