A BioInformatics Survey
... just a taste, with an emphasis on the GCG suite.

Steven M. Thompson
Florida State University School of
Computational Science and Information Technology (CSIT)

Summary
What is bioinformatics, genomics, sequence analysis, computational molecular biology . . .
Reverse Biochemistry & Evolution.
Database growth & cpu power.
Very brief ‘show-and-tell,’ ‘how-to,’ e.g:
NCBI Resources, phylogenetics, GCG’s SeqLab.
High quality training is essential!
Graduates need to be competitive on a world biotechnology market.

My definitions
Biocomputing and computational biology are synonymous and describe the use of computers and computational techniques to analyze any biological system, from molecules, through cells, tissues, and organisms, all the way to populations.
Bioinformatics describes using computational techniques to access, analyze, and interpret the biological information in any of the available biological databases.
Sequence analysis is the study of molecular sequence data for the purpose of inferring the function, mechanism, interactions, evolution, and perhaps structure of biological molecules.
Genomics analyzes the context of genes or complete genomes (the total DNA content of an organism) within and across genomes.
Proteomics is the subdivision of genomics concerned with analyzing the complete protein complement, i.e. the proteome, of organisms, both within and between different organisms.
The reverse biochemistry analogy from a 'virtual' DNA sequence to actual molecular physical characterization, not the other way 'round. Using bioinformatics tools, you can infer all sorts of functional, evolutionary, and structural insights into a gene product, without the need to isolate and purify massive amounts of protein! Eventually you can go on to clone and express the gene based on that analysis using PCR techniques. The computer and molecular databases are an essential part of this process.

The exponential growth of molecular sequence databases & cpu power

<table>
<thead>
<tr>
<th>Year</th>
<th>BasePairs</th>
<th>Sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>1982</td>
<td>403,010</td>
<td>904</td>
</tr>
<tr>
<td>1983</td>
<td>227,429</td>
<td>2417</td>
</tr>
<tr>
<td>1984</td>
<td>2,008,795</td>
<td>4,170</td>
</tr>
<tr>
<td>1985</td>
<td>5,204,420</td>
<td>5,700</td>
</tr>
<tr>
<td>1986</td>
<td>16,132,713</td>
<td>8,710</td>
</tr>
<tr>
<td>1987</td>
<td>55,107,764</td>
<td>14,934</td>
</tr>
<tr>
<td>1988</td>
<td>200,380,955</td>
<td>20,579</td>
</tr>
<tr>
<td>1989</td>
<td>347,423,253</td>
<td>28,791</td>
</tr>
<tr>
<td>1990</td>
<td>591,674,864</td>
<td>36,522</td>
</tr>
<tr>
<td>1991</td>
<td>914,479,926</td>
<td>53,037</td>
</tr>
<tr>
<td>1992</td>
<td>1,555,049,856</td>
<td>78,005</td>
</tr>
<tr>
<td>1993</td>
<td>2,571,004,492</td>
<td>140,486</td>
</tr>
<tr>
<td>1994</td>
<td>3,737,604,940</td>
<td>215,270</td>
</tr>
<tr>
<td>1995</td>
<td>5,080,949,953</td>
<td>550,849</td>
</tr>
<tr>
<td>1996</td>
<td>6,818,728,864</td>
<td>1,002,521</td>
</tr>
<tr>
<td>1997</td>
<td>11,063,906,907</td>
<td>1,736,897</td>
</tr>
<tr>
<td>1998</td>
<td>20,067,152,189</td>
<td>2,575,007</td>
</tr>
<tr>
<td>1999</td>
<td>34,911,583,111</td>
<td>4,508,470</td>
</tr>
<tr>
<td>2000</td>
<td>111,901,062,984</td>
<td>10,016,820</td>
</tr>
<tr>
<td>2001</td>
<td>184,991,747,486</td>
<td>14,997,162</td>
</tr>
<tr>
<td>2002</td>
<td>258,539,016,384</td>
<td>22,210,803</td>
</tr>
</tbody>
</table>


Doubling time - 1 year!

Database growth (cont.)

The Human Genome Project and numerous other genome projects have kept the data coming at alarming rates. As of April 2003, 50 years after the Watson-Crick double-helix!16 Archaea, 128 Bacteria, and 10 Eukaryote complete, finished genomes; and 4 Vertebrate and 5 Plant essentially complete genome maps are publicly available for analysis; not counting all the virus and viroid genomes available.

The International Human Genome Sequencing Consortium announced the completion of a "Working Draft" of the human genome in June 2000, independently that same month, the private company Celera Genomics announced that it had completed the first assembly of the human genome. Both articles were published mid-February 2001 in the journals Science and Nature.
Some neat stuff from those papers
We, *Homo sapiens*, aren't nearly as special as we had once hoped we were. Of the 3.2 billion base pairs in our DNA —

*Traditional*, text-book estimates of the number of genes were often in the 100,000 range; turns out we've only got about twice as many as a fruit fly, between 25,000 and 35,000!  
The protein coding region of our genome is only about 1% or so, much of the remainder 'junk' is 'jumping,' 'selfish DNA' of which much may be involved in regulation and control. Understanding this network is a huge challenge.

100-200 genes were transferred from an ancestral bacterial genome to an ancestral vertebrate genome! (Later shown to be not true by more extensive analyses, and to be due to gene loss rather than transfer.)

What are primary sequences?

(Central Dogma: DNA —> RNA —> protein)  
Primary refers to one dimension — all of the 'symbol' information written in sequential order necessary to specify a particular biological molecular entity, be it polypeptide or nucleotide.

The symbols are the one letter alphabetic codes for all of the biological nitrogenous bases and amino acid residues and their ambiguity codes. Biological carbohydrates, lipids, and structural information are not included within this sequence, however, much of this type of information is available in the reference documentation sections associated with primary sequences in the databases.

What are sequence databases?

These databases are an organized way to store the tremendous amount of sequence information that accumulates from laboratories worldwide. Each database has its own specific format. Three major database organizations around the world are responsible for maintaining most of this data; they largely 'mirror' one another.

Also Georgetown University's NBRF Protein Identification Resource: PIR & NRL 3D.

Europe: European Molecular Biology Laboratory (also EBI & ExpPep): EMBL & Swiss-Prot.

Asia: The DNA Data Bank of Japan (DDBJ).
Content & organization

Most sequence database installations are examples of complex ASCII/Binary databases, but they usually are not Oracle or SQL or Object Oriented (proprietary ones often are). They often contain several very long text files containing different types of information all related to particular sequences, such as all of the sequences themselves, versus all of the title lines, or all of the reference sections. Binary files often help ‘glue together’ all of these other files by providing index functions.

Software is usually required to successfully interact with these databases and access is most easily handled through various software packages and interfaces, either on the World Wide Web or otherwise. Nucleic acid databases are split into subdivisions based on taxonomy (historical). Protein databases are often organized into sections by level of annotation.

What are other biological databases?

Three dimensional structure databases:
- the Protein Data Bank and Rutgers Nucleic Acid Database.

Still more; these can be considered ‘non-molecular’:
- Reference Databases: e.g.
  - OMIM — Online Mendelian Inheritance in Man
  - PubMed/Medline — over 11 million citations from more than 4 thousand bio/medical scientific journals.
- Phylogenetic Tree Databases: e.g. the Tree of Life.
- Metabolic Pathway Databases: e.g. BIR (What Is There) and Japan’s GenomeNet KEGG (the Kyoto Encyclopedia of Genes and Genomes).
- Population studies data — which strains, where, etc.

And then databases that most biocomputing folk don’t even usually consider:
- e.g. GIS/GPS/remote sensing data, medical records, census counts, mortality and birth rates . . .

What are the primary algorithms used?

- Dot matrix approaches;
- The dynamic programming algorithm;
- Heuristics based, hashing methods, for similarity searching;
- Multiple sequence alignment;
- Consensus and weight matrix descriptors, including HMM’s;
- Phylogenetic inference methodology;
- Structure estimation and homology modeling.

Common Thread: Inference through homology is a fundamental principle of biology!

What is homology — in this context it is similarity great enough such that common ancestry is implied. Walter Fitch, a famous molecular evolutionist, likes to relate the analogy — homology is like pregnancy; you either are or you’re not; there’s no such thing as 65% pregnant!
### So how do you do bioinformatics?

<table>
<thead>
<tr>
<th>Site Name</th>
<th>URL</th>
<th>Contact</th>
</tr>
</thead>
<tbody>
<tr>
<td>EEI (Empower User Interface)</td>
<td><a href="http://www.eei.ad.jp/eei">http://www.eei.ad.jp/eei</a></td>
<td>databases/analytics/software metabolic sequence databases</td>
</tr>
<tr>
<td>WIT Metabolism</td>
<td><a href="http://wit.metabolism.ad.jp">http://wit.metabolism.ad.jp</a></td>
<td>databases/analytics/software protein sequence databases</td>
</tr>
<tr>
<td>EMBL-EBI (European Bioinformatics Institute)</td>
<td><a href="http://www.ebi.ac.uk/">http://www.ebi.ac.uk/</a></td>
<td>databases/analytics/software sequencing databases</td>
</tr>
<tr>
<td>European Bioinformatics Institute</td>
<td><a href="http://www.ebi.ac.uk/">http://www.ebi.ac.uk/</a></td>
<td>databases/analytics/software sequencing databases</td>
</tr>
<tr>
<td>Swiss-Prot@EBI</td>
<td><a href="http://www.expasy.ch/">http://www.expasy.ch/</a></td>
<td>databases/analytics/software protein sequence databases</td>
</tr>
<tr>
<td>Prosite</td>
<td><a href="http://www.expasy.org/prosite">http://www.expasy.org/prosite</a></td>
<td>databases/analytics/software protein sequence databases</td>
</tr>
<tr>
<td>Database of Protein/Structure/Function</td>
<td><a href="http://www.ncbi.nlm.nih.gov/">http://www.ncbi.nlm.nih.gov/</a></td>
<td>databases/analytics/software nucleic acid sequence databases</td>
</tr>
<tr>
<td>The Swiss-Prot Project</td>
<td><a href="http://www.expasy.org/">http://www.expasy.org/</a></td>
<td>databases/analytics/software protein sequence databases</td>
</tr>
<tr>
<td>The Human Genome Project</td>
<td><a href="http://www.hgmd.org/">http://www.hgmd.org/</a></td>
<td>databases/analytics/software nucleic acid sequence databases</td>
</tr>
<tr>
<td>The Gene Ontology Project</td>
<td><a href="http://www.geneontology.org/">http://www.geneontology.org/</a></td>
<td>databases/analytics/software nucleic acid sequence databases</td>
</tr>
<tr>
<td>The Web of Science</td>
<td><a href="http://www.wos.com/">http://www.wos.com/</a></td>
<td>databases/analytics/software nucleic acid sequence databases</td>
</tr>
<tr>
<td>PDB</td>
<td><a href="http://www.pdb.org/">http://www.pdb.org/</a></td>
<td>databases/analytics/software nucleic acid sequence databases</td>
</tr>
<tr>
<td>The International Protein Database</td>
<td><a href="http://www.rcsb.org/ipdb">http://www.rcsb.org/ipdb</a></td>
<td>databases/analytics/software nucleic acid sequence databases</td>
</tr>
</tbody>
</table>

### But large datasets become intractable.

**What other resources are available?**

- **Desktop software solutions** — public domain programs are available, but . . . complicated to install, configure, and maintain. User must be pretty computer savvy. So,

  - commercial software packages are available, e.g. MacVector, DS Gene, DNAStar, etc.,
  - but . . . license hassles, big expense per machine, and Internet and/or CD database access all complicate matters!

### Therefore, UNIX server-based solutions

- Public domain solutions also exist, but now a very cooperative systems manager needs to maintain everything for users, so,
- commercial products, e.g. the Accelrys GCG Wisconsin Package and the SeqLab Graphical User Interface, simplify matters for administrators and users.

One license fee for an entire institution and very fast, convenient database access on local server disks. Connections from any networked terminal or workstation anywhere!

**Operating system:** UNIX command line operation hassles; communications software — telnet, ssh, and terminal emulation; X graphics; file transfer — ftp, and scp/sftp; and editors — vi, emacs, pico (or desktop word processing followed by file transfer [save as “text only”]).
The Genetics Computer Group — The Accelrys Wisconsin Package for Sequence Analysis

Begun in 1982 in Oliver Smithies' lab at the Genetics Dept. at the University of Wisconsin, Madison, then a private company for over 10 years, then acquired by the Oxford Molecular Group U.K., and now owned by Pharmacopeia Inc. U.S.A. under the new name Accelrys.

The suite contains almost 150 programs designed to work in a "toolbox" fashion. Several simple programs used in succession can lead to sophisticated results.

Also 'internal compatibility,' i.e. once you learn to use one program, all programs can be run similarly, and, the output from many programs can be used as input for other programs.

Used all over the world by more than 30,000 scientists at over 530 institutions in 35 countries, so learning it here will most likely be useful anywhere else you may end up.

To answer the always perplexing GCG question — "What sequence(s)? . . ."

Specifying sequences, GCG style:

in order of increasing power and complexity:

The sequence is in a local GCG format single sequence file in your UNIX account. (GCG Reformat and all From & To programs)

The sequence is in a local GCG database in which case you 'point' to it by using any of the GCG database logical names. A colon, ':', always sets the logical name apart from either an accession number or a proper identifier name or a wildcard expression and they are case insensitive.

The sequence is in a GCG format multiple sequence file, either an MSF (multiple sequence format) file or an RSF (rich sequence format) file. To specify sequences contained in a GCG multiple sequence file, supply the file name followed by a pair of braces, '{ }', containing the sequence specification, e.g. a wildcard — '{ }'.

Finally, the most powerful method of specifying sequences is in a GCG "list" file. It is merely a list of other sequence specifications and can even contain other list files within it. The convention to use a GCG list file in a program is to precede it with an at sign, '@'. Furthermore, one can supply attribute information within list files to specify something special about the sequence.

'Clean' GCG format single sequence file after 'reformat' (or any of the From... programs)

This is a small example of GCG single sequence format. Always put some documentation on top, so in the future you can figure out what it is you're dealing with! The line with the two periods is converted to the checkmark line.

SeqLab's Editor mode can also "Import" native GenBank format and ABI or LI-COR trace files!
The List File Format

remember the @ sign!

An example GCG list file of many elongation factors follows. As with all GCG data files, two periods separate documentation from data.

my-special.pep begin:24 end:134
SwissProt:EfTu_Ecoli
EfTa-Tu.msf(*)
/usr/accounts/test/another.rsf(efla_*)
@another.list

The 'way' SeqLab works!
## SeqLab — GCG's X-based GUI!

SeqLab is the merger of Steve Smith’s Genetic Data Environment and GCG’s Wisconsin Package Interface:

\[
\text{GDE + WPI = SeqLab}
\]

Requires an X-Windowing environment — either native on UNIX computers (including LINUX, but not included by Apple in Mac OS X [v.10+]) but see Apple’s X11 package and XDarwin), or emulated with X-Server Software on personal computers.

## Conclusions

Gunnar von Heijne in his old but quite readable treatise, *Sequence Analysis in Molecular Biology; Treasure Trove or Trivial Pursuit* (1987), provides a very appropriate conclusion:

"Think about what you’re doing; use your knowledge of the molecular system involved to guide both your interpretation of results and your direction of inquiry; use as much information as possible, and do not blindly accept everything the computer offers you."

He continues:

"... if any lesson is to be drawn... it surely is that to be able to make a useful contribution one must first and foremost be a biologist, and only second a theoretician... We have to develop better algorithms, we have to find ways to cope with the massive amounts of data, and above all we have to become better biologists. But that’s all it takes."

## FOR MORE INFO...

Visit my Web page:

http://bio.fsu.edu/~stevet/cv.html

Contact me (stevet@bio.fsu.edu) for specific bioinformatics assistance and/or long distance collaboration.

Many fine texts are also starting to become available in the field.

To ‘honk-my-own-horn’ a bit, check out the new —

Current Protocols in Bioinformatics from John Wiley & Sons, Inc; [http://www.does.org/cp/bioinfo.html](http://www.does.org/cp/bioinfo.html). They asked me to contribute a chapter on multiple sequence analysis using GCG software.

Humana Press, Inc. also asked me to contribute. I’ve got two chapters in their —

Introduction to Bioinformatics:

A Theoretical And Practical Approach

http://www.humanapress.com/Product.pasp?TextCatalog=HumanaBooks&TextCategory=4&stProductID=1-58829-241-0&stVariant=0

Both volumes are now available.