

CS 124 Spring 2001, Lecture 1 March 30, D. Gusfield

### Outline of First Lecture

1. Logistical Matters, registration, auditing, computer accounts, texts, prereqs for the course, URL for course materials, password for website, first handout, reserve books. To jumpstart the course accounts, use Mothra to give yourself access to the vis lab machines. The Perl book by Johnson is late to the MU, try Amazon.com

### 2. What is Bioinformatics?

Bioinformatics and Computational Molecular Biology are generally concerned with the use of computation and large, cross-species, organized databases of biological information to augment or leverage traditional laboratory-based biology. Bioinformatics has become critical in biology due to recent changes in our ability and determination to acquire massive biological data sets (for example through genomic-level sequencing, and through expression monitoring of multiple mRNAs and even proteins), and due to the ubiquitous (almost routine) successful biological insights that have come from the comparison of the already available sequence data. Quoting from the proposal to establish a UC Life Sciences Informatics Program:

The capabilities to acquire this information have become so powerful that the volume of new data generated in some studies is growing exponentially ... There is general agreement that significant new understanding of basic biology and biomedical concerns will come from analyzing the vast amounts of data being generated. So important has this domain become that it has acquired its own name – bioinformatics.

The most mature example of the power of bioinformatics is from the accumulation, curation, and use of DNA and protein *sequence* data, and we will first use this example to illustrate the paradigm and potential of bioinformatics.

a) Bioinformatics is the intersection of biology, computer science, and math/stat. for the acquisition, maintenance, comparison, analysis and exploitation of large-scale biological data; for the prediction of structure and/or function of biomolecules; for the deduction and exploitation of phylogenetic histories; for the specification, prediction and analysis of larger-scale biological structures such as metabolic pathways and gene regulatory networks.

b) Sequence analysis and sequence data is the most mature prototype of bioinformatics (many large-scale sequencing projects) and huge accumulation of sequence data in many databases; but sequence data and analysis is just the beginning. Structure data (RNA and protein) are also active; expression monitoring of mRNA is becoming common; (proteomics) large-scale and dynamic monitoring of protein activity is nearly here; large-scale protein structure projects are approaching; exhaustive protein-protein interaction projects are in the works; many other omics to come: phenomics, phylogenomics, ....

c) Bioinformatics predates genomics, but its importance has been dramatically increased due to genomics and the many other omics that are coming on.

Biology has had until recently three major, interconnected modalities - "observation-driven biology", intense observation and collection in nature; "theory-driven biology", development of theories to explain some of the observations; "hypothesis/experiment-driven biology" intense laboratory experimentation to try to understand some of the mechanisms behind the observations, or to try to resolve theoretical questions. The key point about the laboratory modality is that most experiments are hypothesis driven, focusing on some well-articulated, specific phenomenon or question.

In contrast, we are now at the beginnings of a fourth modality: large-scale, even exhaustive, molecular data accumulation, coupled with exploitation of that data. The key point is that this is systematic collection and organization of data without any driving hypothesis or question. In a way, it is similar to the first modality, but the focus and scale are different. This is "(molecular) data-driven biology". It exists because technology has been developed for large-scale accumulation of certain kinds of molecular data (DNA sequence data particularly), and because of initial successes in exploiting the data, and the general belief (a minority view until recently) that having this data leads to great power over and insight into biological systems of importance. A huge quantitative change of scale that is believed (by many) to have important qualitative consequences.

Genomics - systematic accumulation of DNA data, full sequences and gene fragments (ESTs) etc. - is the first and most mature example of this "data-driven biology". But genomics is just the start. Next comes large-scale, systematic accumulation of DNA polymorphism data from populations, of DNA expression data under different conditions and in different cells, of time-

course data, of protein synthesis data, of protein structure data, of protein-protein interaction data, of protein-DNA interaction data, of phenotype data, of metabolic data, of carbohydrate data, etc. etc. Hence proteomics, phenomics, expresionomics, populationomics, and many many more omics's to come. And for each one there will be a "functional omics" because understanding function is the real payoff one wants from the data and from the technology development. For many of the "omics" there will also be "comparative omics" because comparison is a technique that will likely be (as it has been in genomics) an extremely valuable way to extract understanding from the data.

"Genomics" has become the generic term for data-driven biology, even when the data is not about genes or genomes. The shift to data-driven biology, the accumulation and exploitation of large-scale data, has led to the need for new computational technology (machines, software, algorithms, theory) and for people who develop and use it. A greater need is for people who can both develop that technology and focus it in creative ways to attack biological problems. There is a small core of professional biologists who do just that. Check out *Methods in Enzymology*, vol. 266: *Computer methods in Macromolecular Sequence Analysis*, edited by R.F. Doolittle. More than 60 authors are represented, most with affiliations in biology labs or university biology departments. Similarly, there are people from computer science or mathematics or statistics who now focus primarily on data-driven biology, and some have even been hired into traditional biology departments.

Whatever you call those people and the field they work in, the driving need world-wide is for people who develop and exploit techniques that facilitate data-driven biology - people whose work addresses the change in scale occurring in biology. It is those kinds of computationally oriented people who are being hired into "bioinformatics units" at genomics and pharmaceutical companies and national labs, into "bioinformatics programs" at the universities that have started such programs.

<http://linkage.rockefeller.edu/wli/bioinfocourse.html>

for links to about 70 bioinformatics courses around the world. The field is pretty well defined by the common topics in those courses (summarized in the editorial linked at the start of that page), and the materials they rely on.

### 3. Introduction to the course

a) We will concentrate on sequence-based bioinformatics, genomics and phylogenomics but also discuss a bit of structure-based bioinformatics and

perhaps gene regulatory networks if time permits. We will also emphasize sequence analysis and exploitation over sequence acquisition, since acquisition is becoming more routine. See the syllabus for more detail on specific topics covered in the course.

b) I don't know how successful this will be, but I want to emphasize, for each specific analysis topic, that existing software tools are the end-result of a chain-of-reasoning, a process that proceeds through a series of stages. Some of the stages may be fuzzy, missing or hidden, and different stages are often done by different players. Still, we will attempt to identify each stage for each analysis topic.

The chain starts with a Knowledge Stage, where we (the general "we") have, or expect to have, a corpus of Biological Data and Biological Knowledge (observations and experimental results), and have some set of Biological Questions that we want to use the data to help resolve; The next stage is a Biological Modeling stage, where some general rules, or simplifying statements about the data are developed; next is a Mathematical Modeling stage, where Biological Insight and Models are translated into, or reflected by, Mathematical Statements or Formulas; the next stage is Computational Model stage, where mathematics is translated into Computational Problems; next is the Algorithm Stage, where Algorithms (general solution methods) are developed for the Computational Problems; finally we have the Program Stage, where working programs are written that reflect the whole process. These programs are used by working biologists (and others) to analyze data and help to answer the biological questions posed at the start of the process.

The thing I want to emphasize is that many compromises and simplifications are made at almost every stage of the process. Moreover, the effort at each stage is not always successful. The final program, software-tool, reflects all of these compromises and partial successes. The field of sequence analysis is therefore dynamic as people develop and try out different models, compromises, algorithms, etc. at each stage. The process is also not completely linear, as work in one stage can effect work in proceeding stages, and the results one gets from running the programs can effect what data you next try to obtain.

In order to better understand the available software tools , and to help in influencing future tool development, you have to know something about the entire chain-of-reasoning that leads to tools, and where the weak links are.

An example of a biological rule, or generalization that comes out of a

corpus of biological knowledge or observations, is the Central Dogma: The flow of genetic information goes from DNA to RNA to protein, and not the opposite direction. Of course, today we know that this is not correct, as flow from RNA to DNA is possible, and indeed common and significant in many phenomena.

Another example of biological rule is that the most conserved regions in a protein sequence (over evolution) are the functional domains of the protein. In fact, in bioinformatics, we extensively exploit this assumption in reverse as follows: to find functional domains, look for regions that are conserved homologous proteins found in different species. However, before there was much comparative protein data, some people made a reasonable case for the opposite: that the most conserved regions in a protein sequence should be the non-functional parts of the protein sequence. Can you flesh out such an argument? The point here is that we work with simplified rules rather than a mass of raw data, and those rules are subject to change.

c) Given the above, the course is really an interweaving of four subcourses: theory, programming, mastery of existing software, and exploitation of software and data to solve biological problems. However, given my background, emphasis in lectures will be on the first two subcourses; labs will emphasize the second and third; and the fourth subcourse will only be illustrated.