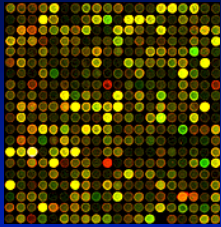


## Bioinformatics and Genomics Education



Jonathan Pevsner, Ph.D.  
ASMCUE  
Friday, June 3, 2011

## Outline

Overview of bioinformatics and genomics education

Approaches to teaching bioinformatics

Approaches to teaching functional genomics

Approaches to teaching genomics and the tree of life

## What are bioinformatics and genomics?

- Interface of biology and computers
- Analysis of proteins, genes and genomes using computer algorithms and computer databases
- Genomics is the analysis of genomes.  
The tools of bioinformatics are used to make sense of the billions of base pairs of DNA that are sequenced by genomics projects.

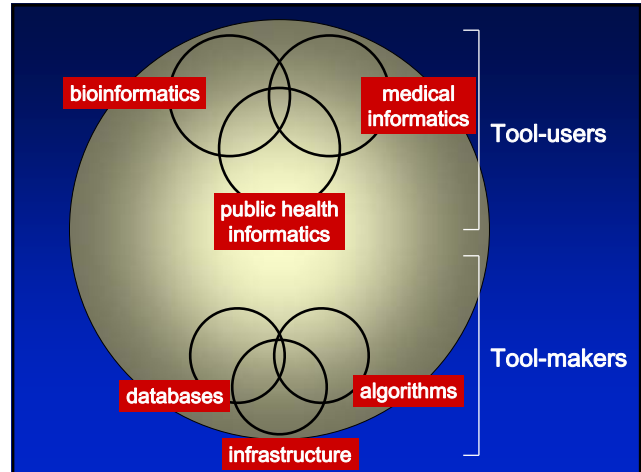
## On bioinformatics

"Science is about building causal relations between natural phenomena (for instance, between a mutation in a gene and a disease). The development of instruments to increase our capacity to observe natural phenomena has, therefore, played a crucial role in the development of science - the microscope being the paradigmatic example in biology. With the human genome, the natural world takes an unprecedented turn: it is better described as a sequence of symbols. Besides high-throughput machines such as sequencers and DNA chip readers, the computer and the associated software becomes the instrument to observe it, and the discipline of bioinformatics flourishes."

## On bioinformatics

"However, as the separation between us (the observers) and the phenomena observed increases (from organism to cell to genome, for instance), instruments may capture phenomena only indirectly, through the footprints they leave. Instruments therefore need to be calibrated: the distance between the reality and the observation (through the instrument) needs to be accounted for. This issue of *Genome Biology* is about calibrating instruments to observe gene sequences; more specifically, computer programs to identify human genes in the sequence of the human genome."

Martin Reese and Roderic Guigó, *Genome Biology* 2006 7(Suppl 1):S1, introducing EGASP, the Encyclopedia of DNA Elements (ENCODE) Genome Annotation Assessment Project

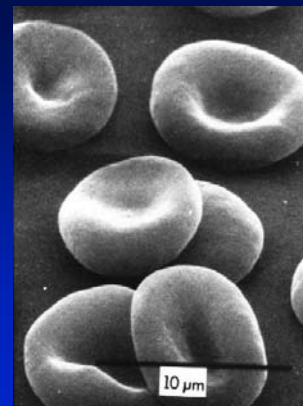


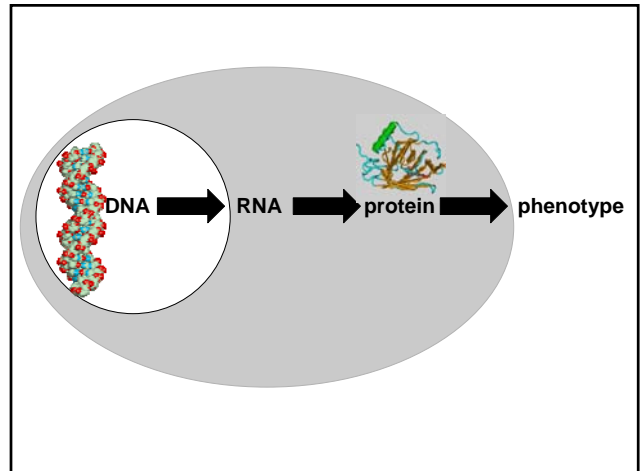
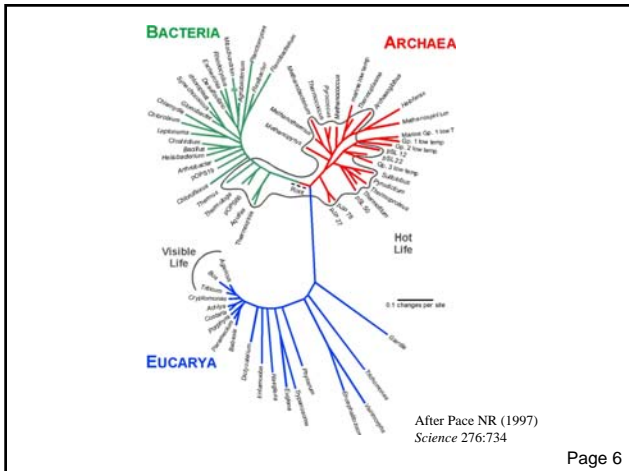
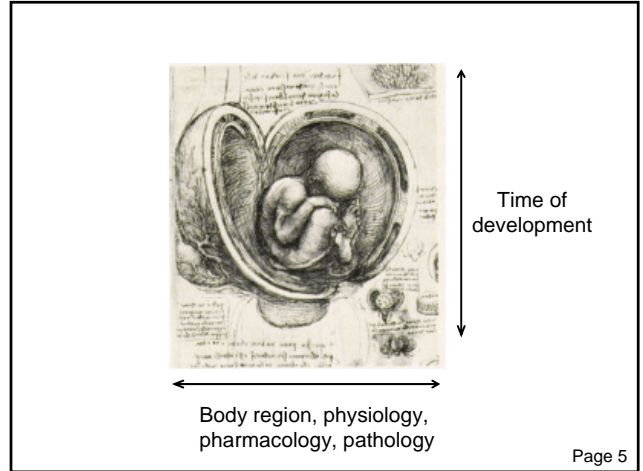
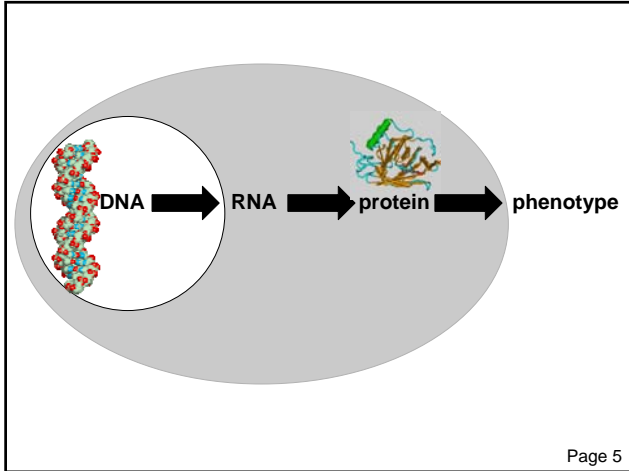
## Three perspectives on bioinformatics: the scope of the discipline

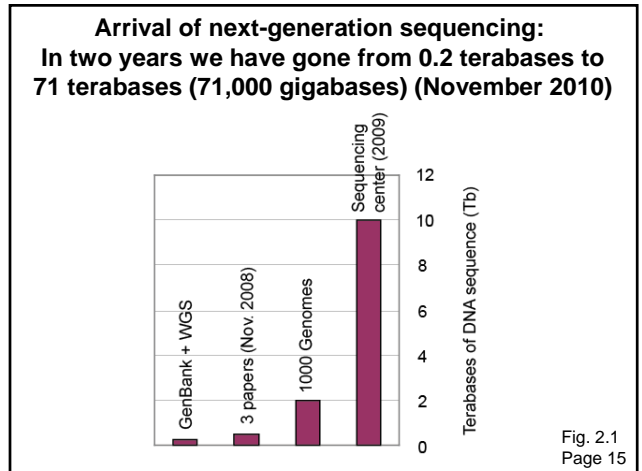
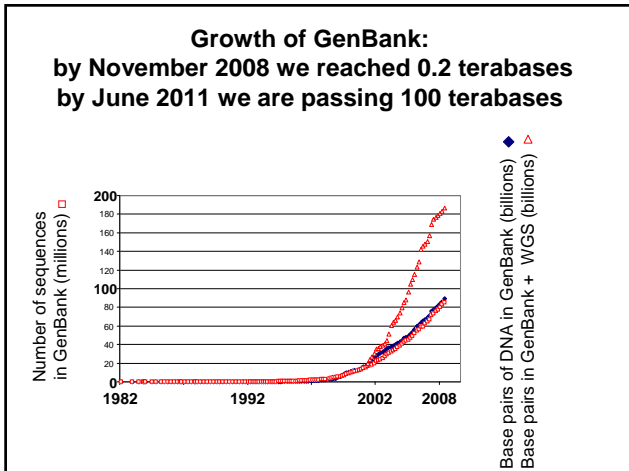
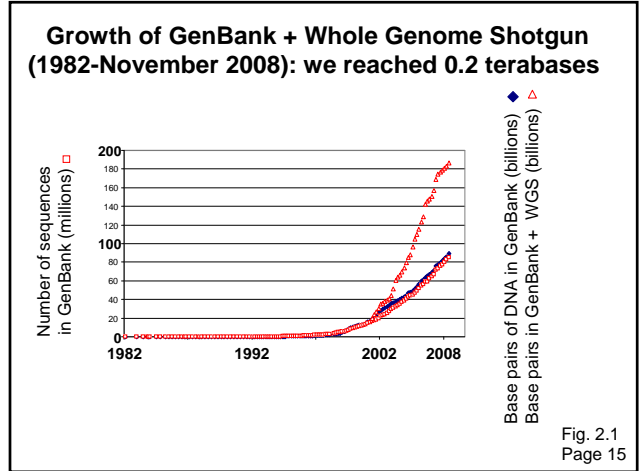
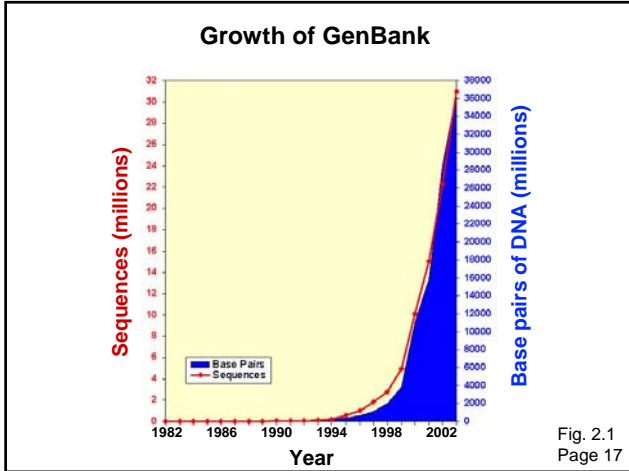
The cell

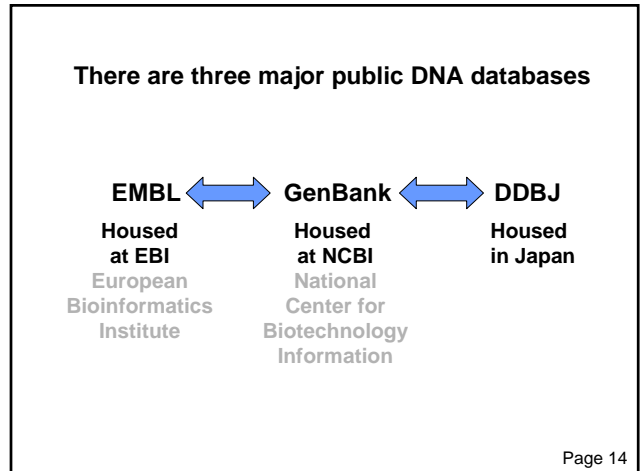
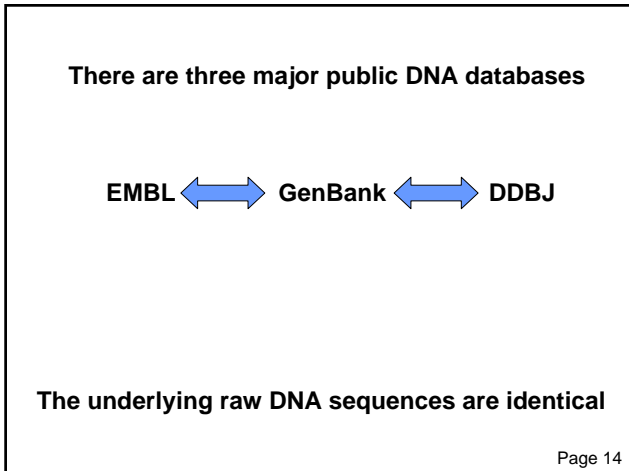
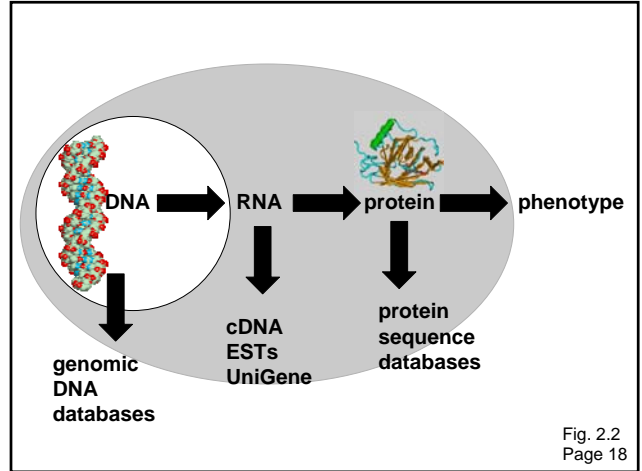
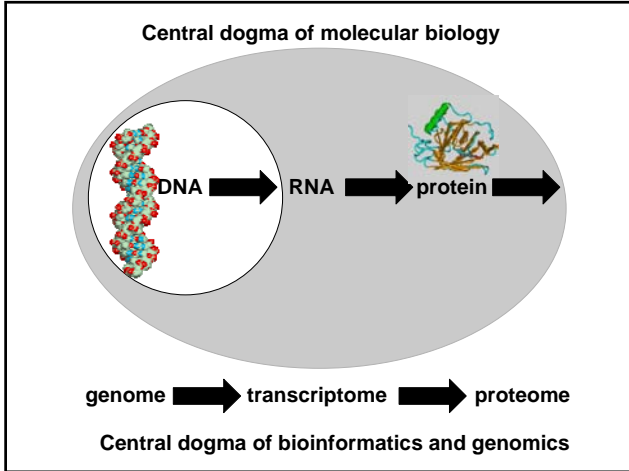
The organism

The tree of life









**Taxonomy at NCBI:**  
**>200,000 species are represented in GenBank**

Ranks:	higher taxa	genus	species	lower taxa	total
<a href="#">Archaea</a>	<a href="#">103</a>	<a href="#">116</a>	<a href="#">533</a>	<a href="#">162</a>	<a href="#">914</a>
<a href="#">Bacteria</a>	<a href="#">1126</a>	<a href="#">2041</a>	<a href="#">16236</a>	<a href="#">8938</a>	<a href="#">28341</a>
<a href="#">Eukaryota</a>	<a href="#">17050</a>	<a href="#">52186</a>	<a href="#">204884</a>	<a href="#">17318</a>	<a href="#">291436</a>
<a href="#">Fungi</a>	<a href="#">1211</a>	<a href="#">3675</a>	<a href="#">21980</a>	<a href="#">1486</a>	<a href="#">28350</a>
<a href="#">Metazoa</a>	<a href="#">12540</a>	<a href="#">32954</a>	<a href="#">92188</a>	<a href="#">8278</a>	<a href="#">145960</a>
<a href="#">Viridiplantae</a>	<a href="#">2012</a>	<a href="#">13339</a>	<a href="#">83523</a>	<a href="#">6466</a>	<a href="#">105340</a>
<a href="#">Viruses</a>	<a href="#">506</a>	<a href="#">344</a>	<a href="#">7020</a>	<a href="#">56107</a>	<a href="#">63977</a>
<a href="#">All taxa</a>	<a href="#">18810</a>	<a href="#">54691</a>	<a href="#">234123</a>	<a href="#">82560</a>	<a href="#">390186</a>

Page 16

10/10

<http://www.ncbi.nlm.nih.gov/Taxonomy/taxstat.cgi>

## Outline

Overview of bioinformatics and genomics education

Approaches to teaching bioinformatics

Approaches to teaching functional genomics

Approaches to teaching genomics and the tree of life

## Goals of teaching bioinformatics

- To provide an introduction to bioinformatics with a focus on the National Center for Biotechnology Information (NCBI), UCSC, and EBI
- To focus on the analysis of DNA, RNA and proteins
- To introduce students to the analysis of genomes
- To combine theory and practice to help students solve research problems

## Issues in teaching: textbooks

Some textbooks focus on the "big picture" and are more biological (less computational)

Baxevanis

Pevsner

Claverie & Notredame (Bioinformatics for Dummies)

Some textbooks are more computational

David Mount

Some emphasize programming languages

Tisdall (Beginning Perl for Bioinformatics)

Building Bioinformatics Solutions: with Perl, R, & MySQL

## Issues in teaching: textbooks

*Bioinformatics and Functional Genomics* (Wiley-Blackwell, 2<sup>nd</sup> edition 2009) has three parts:

### I. Bioinformatics

Sequence data; pairwise alignment, BLAST, multiple alignment, phylogeny and evolution

### II. Functional Genomics

RNA analyses (microarrays), proteomics

### III. Genomics: the tree of life

Viruses, bacteria, eukaryotes, the human genome

## Issues in teaching: powerpoints

A large collection of powerpoints is freely available at <http://bioinbook.org> along with data sets, website links, audio files, and other materials.

## Issues in teaching: websites

There are several websites that are most important:  
NCBI (<http://www.ncbi.nlm.nih.gov/>)  
UCSC (<http://genome.ucsc.edu>)  
EBI (<http://www.ebi.ac.uk/>)  
Each of these offers teaching resources.

Key databases (e.g. PFAM and UniProt for proteins, PDB for structures)

Key portals (e.g. ExpASY for proteomics)

Web-based course management systems

Moodle  
Sakai  
Blackboard

Google "moodle bioinformatics" to get here;  
Click "Bioinformatics" to sign in;  
The enrollment key you need is...

The screenshot shows the Moodle interface for the 'Bioinformatics and Genomics' course at Johns Hopkins University. The page includes a header with the course title and a navigation menu. A central section titled 'Bioinformatics and Genomics' provides an overview of the course and its resources. Below this, there is a section for 'Available Courses' with a red arrow pointing to the 'Bioinformatics' course link.

## A thematic approach to teaching: the globins

I use beta globin as a model gene/protein throughout my bioinformatics course. Globins including hemoglobin and myoglobin carry oxygen. We study globins in a variety of contexts including

- sequence alignment
- gene expression
- protein structure
- phylogeny
- homologs in various species

## Teaching bioinformatics: computer labs

Computer labs are highly recommended. Teachers must decide whether the scope includes just websites (“point-and-click”) or the use of command line functions. Commonly used languages include Perl and R (www.r-project.org).

The use of the Unix operating system and such languages is essential for the handling of large data sets.

For a course with **no** formal computer labs, regular quizzes can function as a computer lab. To solve the questions, students need to go to websites, use databases, and use software. These can be organized on a site such as moodle.

### Example of a quiz question requiring students to download and use MEGA, a phylogeny software program

- 3.4
- Marks: 1
- (1) Copy the 12 globin DNA sequences from problem 1 above.
  - (2) Get MEGA 4 software (from <http://www.megasoftware.net/>).
  - (3) Paste the sequences into MEGA (by going to the Alignment pull-down and selecting Alignment Explorer/Clustal, choose create a new alignment. And remember to choose DNA and not protein).
  - (4) After you have pasted in the 12 DNA sequences, perform a ClustalW alignment (under the Alignment pull-down choose Align by ClustalW, then click OK). Save as a mas file (optional) and meg file (required), and open the data in MEGA. (After you have used ClustalW in the Alignment Explorer, you can save the files using the Data pull-down. Or if you just close the data set you will be prompted to save your data in the mas and meg formats.)
  - (5) Note that a Sequence Data Explorer is open, as well as the main MEGA dialog box. Please note that question 6 (below) asks you to use this Sequence Data Explorer.
  - (6) Under Phylogeny on the main MEGA dialog box, choose "construct phylogeny". First make a neighbor-joining tree, and then make a maximum parsimony tree. For each, you can click "Compute" using the default settings. Note that the green boxes are clickable to choose various options, but you don't need to change them for these problems. True or false: the trees have different topologies.

Answer:  True  
 False

## Outline

Overview of bioinformatics and genomics education

Approaches to teaching bioinformatics

Approaches to teaching functional genomics

Approaches to teaching genomics and the tree of life

## Functional genomics: overview

Functional genomics refers to the genome-wide study of the function of DNA (including genes and nongenic elements) as well as the nucleic acid and protein products encoded by DNA.

Examples:

- studies of DNA (the genome)
- studies of RNA (the transcriptome)
- studies of proteins (the proteome)
- the use of high throughput screens
- perturbation of gene function
- relating genotype and phenotype

## Functional genomics: teaching considerations

The analysis of large-scale data requires...

- appropriate statistics for experimental design
- statistics for data analysis
- tools to handle large data sets (millions of rows)
- packages to perform analyses (e.g. preprocessing, hypothesis testing, exploratory analyses)
- understanding of the biological principles

Teachers must decide which aspects to cover.

## Outline

---

Overview of bioinformatics and genomics education

Approaches to teaching bioinformatics

Approaches to teaching functional genomics

Approaches to teaching genomics and the tree of life 

## Teaching genomics

---

The availability of complete genome sequences is revolutionizing our understanding of the tree of life.

For teaching purposes, I require students to select any one genome and write a report on it. Students are encouraged to organize the report according to five principles.

## Five approaches to genomics

---

As we survey the tree of life, consider these perspectives:

### Approach I: cataloguing genomic information

Genome size; number of chromosomes; GC content; isochores; number of genes; repetitive DNA; unique features of each genome

### Approach II: cataloguing comparative genomic information Orthologs and paralogs; COGs; lateral gene transfer

### Approach III: function; biological principles; evolution How genome size is regulated; polyploidization; birth and death of genes; neutral theory of evolution; positive and negative selection; speciation

### Approach IV: Human disease relevance

### Approach V: Bioinformatics aspects Algorithms, databases, websites

## Teaching genomics

---

As a second option, students can select one gene and study it in detail.

See <http://bioinfbook.org/genomics.php> for more information.

## Summary and conclusions

---

The discipline of bioinformatics allows one to study biological problems of interest. Key tools include sequence alignment, phylogeny, and high throughput functional measurements of cellular function. Genomics allows these tools to be applied to the tree of life.

Education in these areas is evolving rapidly because the pace of technological change is rapid.