Introduction to Bioinformatics

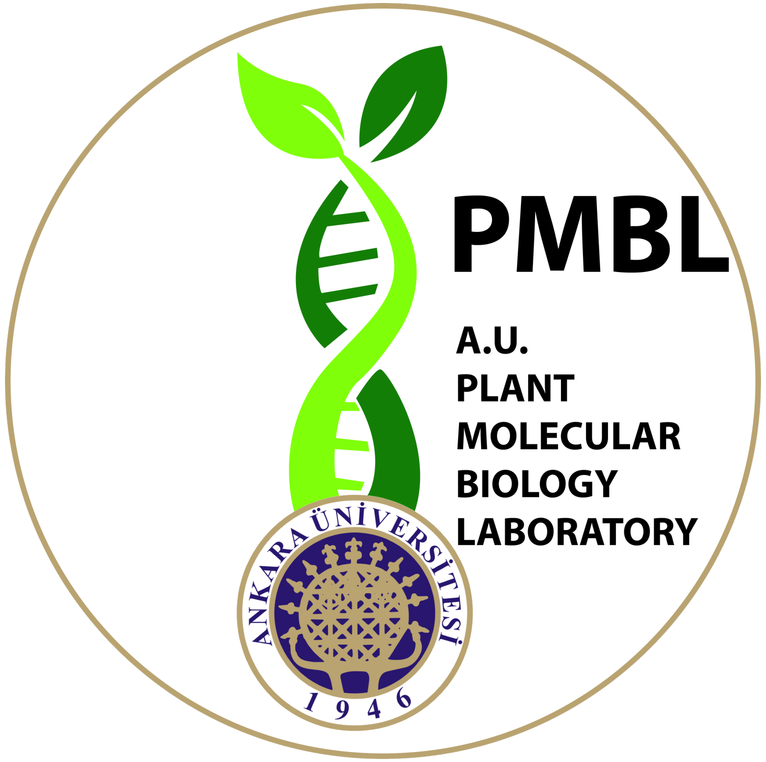
Assoc. Prof. Ilker BUYUK

About me

- Assoc. Prof. Dr. Ilker BUYUK
- E-mail: ilker.buyuk@ankara.edu.tr
- Office: Z67

BMBL

A.Ü.
Bitki Moleküler
Biyolojisi
Laboratuvarı



PMBL

A.U.
PLANT
MOLECULAR
BIOLOGY
LABORATORY

http://bmbl.ankara.edu.tr

Course Details

Course Code	: **BIO 212**
Course Name	: **Introduction to Bioinformatics**
Credit	: **2**
Course Level	: **Undergradute**
Instructor	: **İlker BÜYÜK**

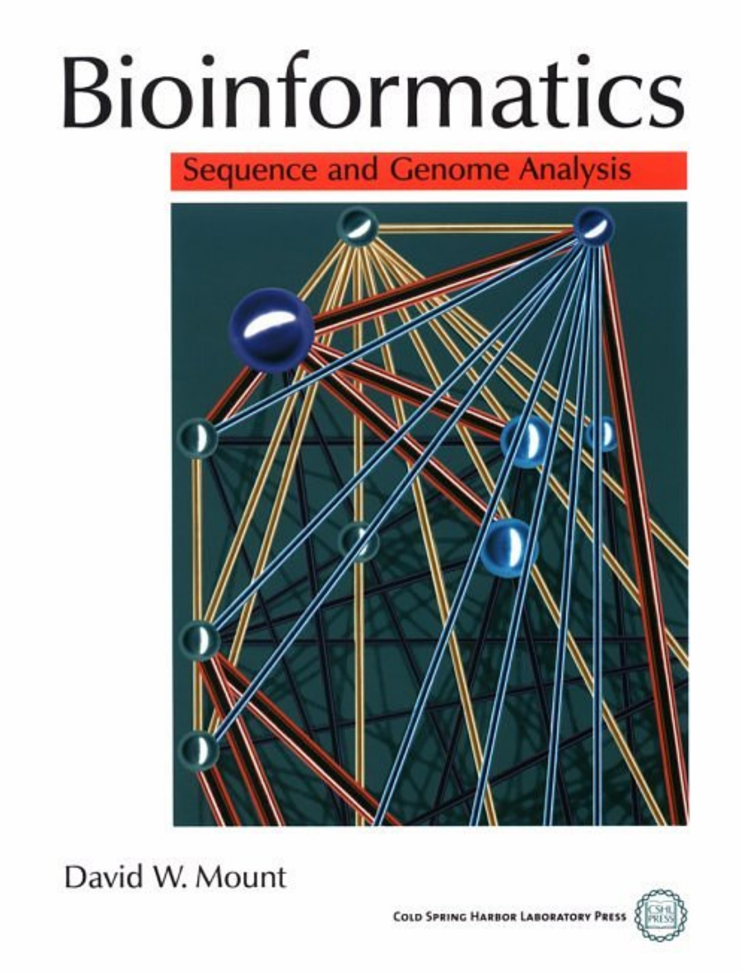
Room: **Online / www.ekampus.ankara.edu.tr**

Email: **buyuki@ankara.edu.tr**

Assesment

Midterm	:	30%
Homework	:	20%
Final	:	40%
Attendance & participation	:	10 %

Recommended Texts

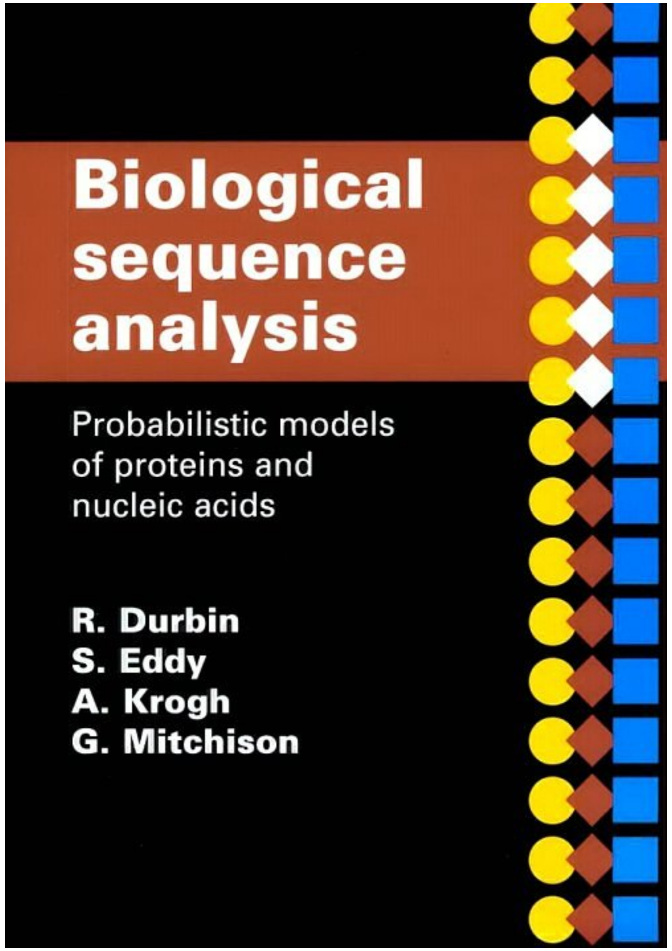


Bioinformatics

Sequence and Genome Analysis

David W. Mount

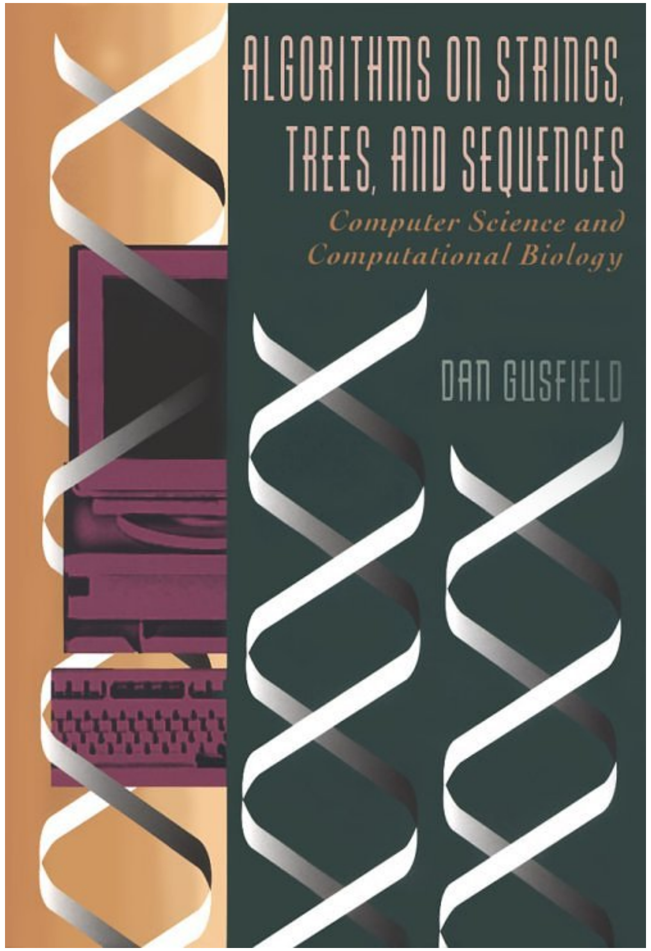
Cold Spring Harbor Laboratory Press



Biological sequence analysis

Probabilistic models of proteins and nucleic acids

R. Durbin
S. Eddy
A. Krogh
G. Mitchison

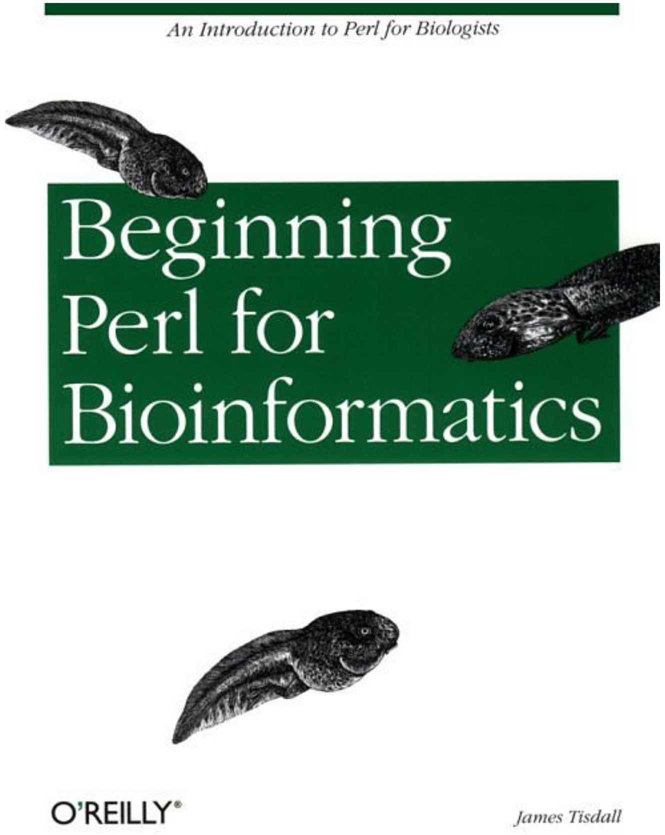


Algorithms on Strings, Trees, and Sequences

Computer Science and Computational Biology

Dan Gusfield

Recommended Texts-2

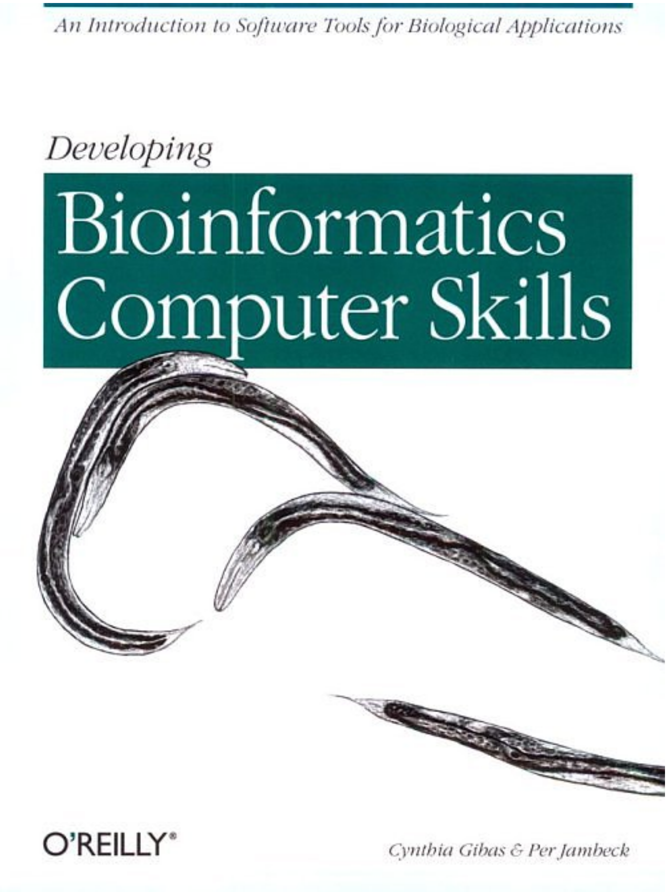


An Introduction to Perl for Biologists

Beginning Perl for Bioinformatics

O'REILLY®

James Tisdall

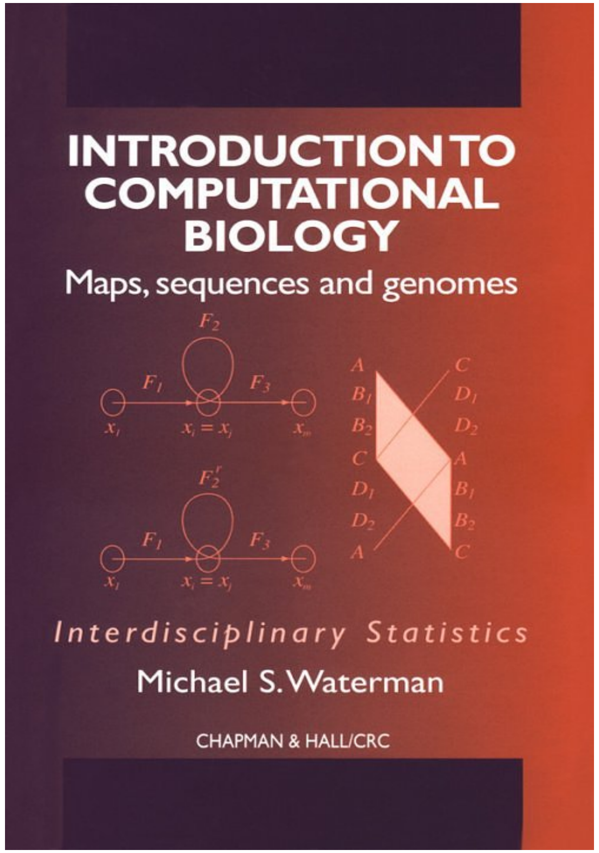


An Introduction to Software Tools for Biological Applications

Developing Bioinformatics Computer Skills

O'REILLY®

Cynthia Gibas & Per Jambeck



INTRODUCTION TO COMPUTATIONAL BIOLOGY

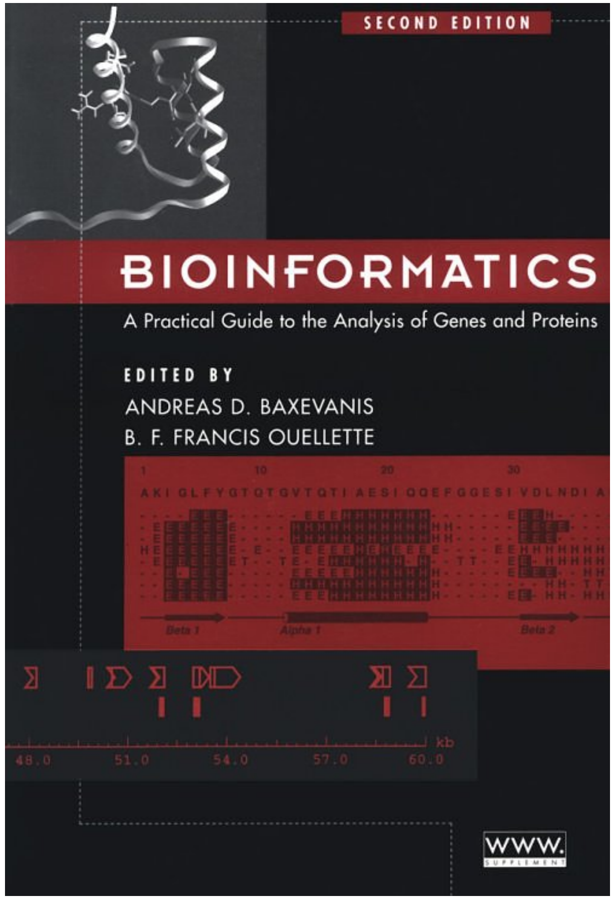
Maps, sequences and genomes

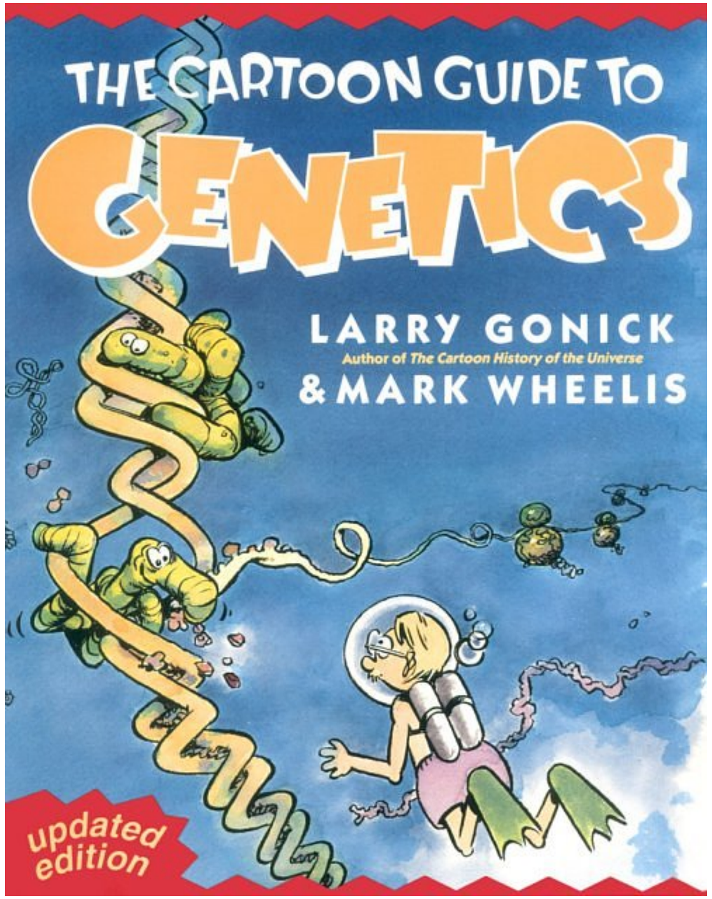
Interdisciplinary Statistics

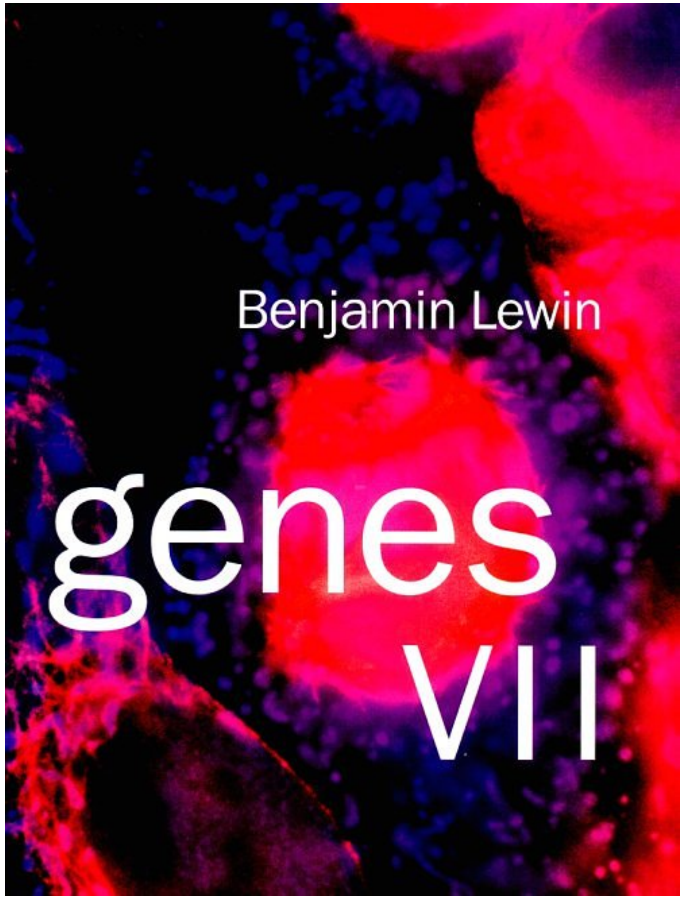
Michael S. Waterman

CHAPMAN & HALL/CRC

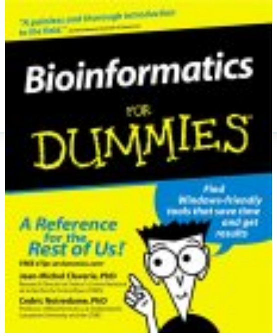
Recommended Texts-3



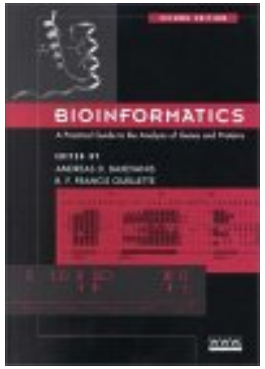




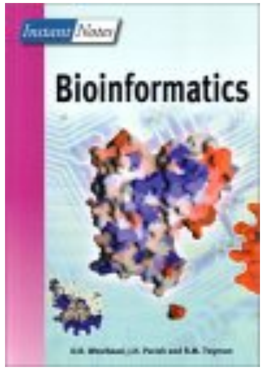
Recommended Texts-4



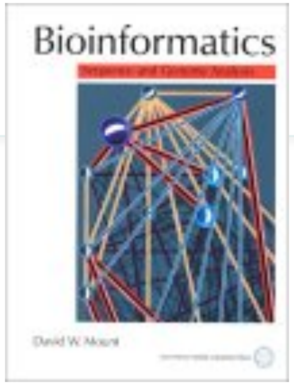
Bioinformatics for Dummies
Jean Claverie, Cedric Notredame



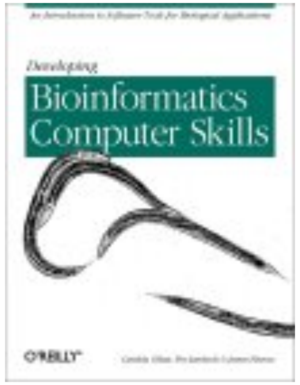
Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins
Andreas D. Baxevanis, B. F. Ouellette, Ouellette B. F. Francis.



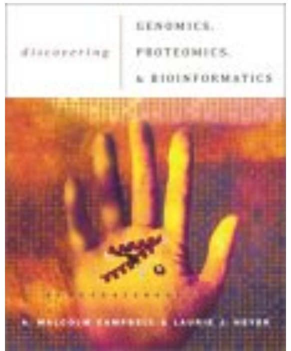
Instant Notes in Bioinformatics
D. R. Westhead, Richard M. Twyman, J. H. Parish



Bioinformatics: Sequence and Genome Analysis, Vol. 5
David W. Mount, David Mount

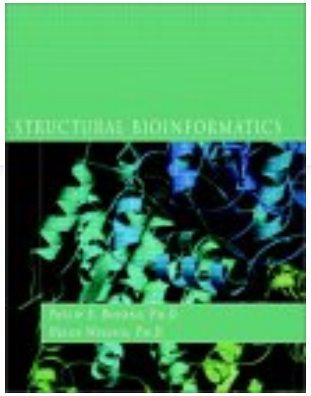


Developing Bioinformatics Computer Skills
Cynthia Gibas, Per Jambeck, Lorrie LeJeune (Editor)

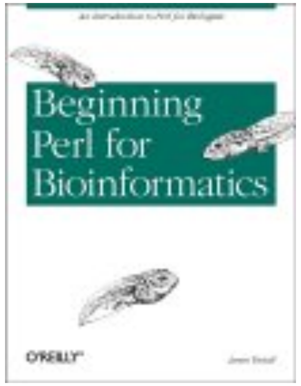


Discovering Genomics, Proteomics, and Bioinformatics
A. Malcolm Campbell, Laurie J. Heyer

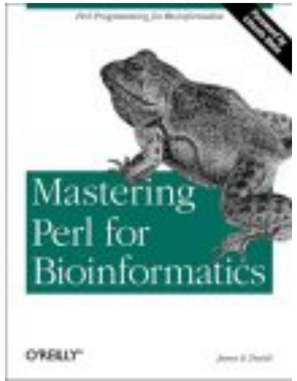
Recommended Texts-5



Structural Bioinformatics
Philip E. Bourne (Editor),
Helge Weissig

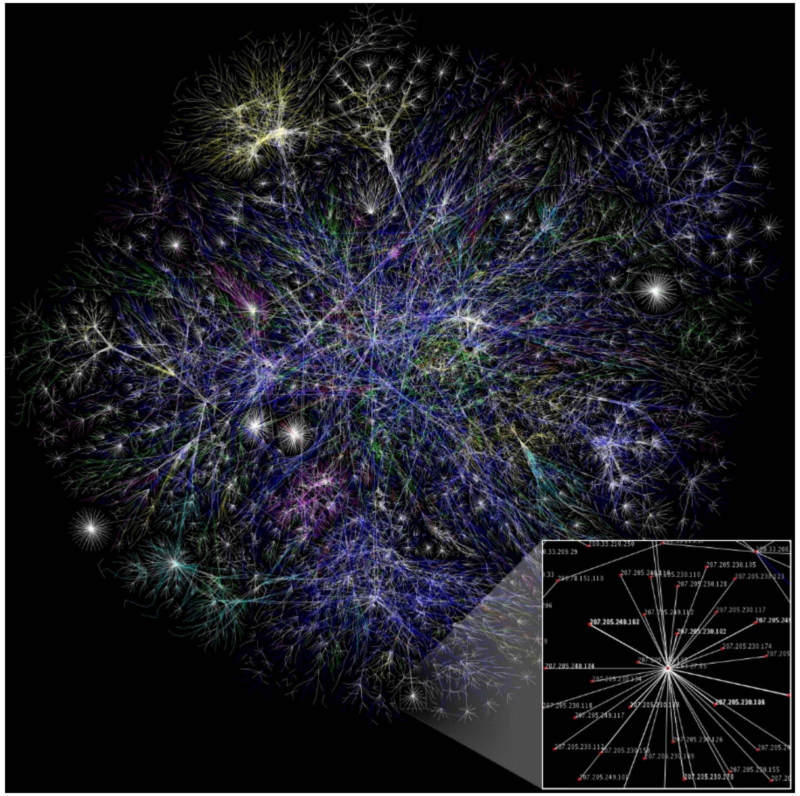


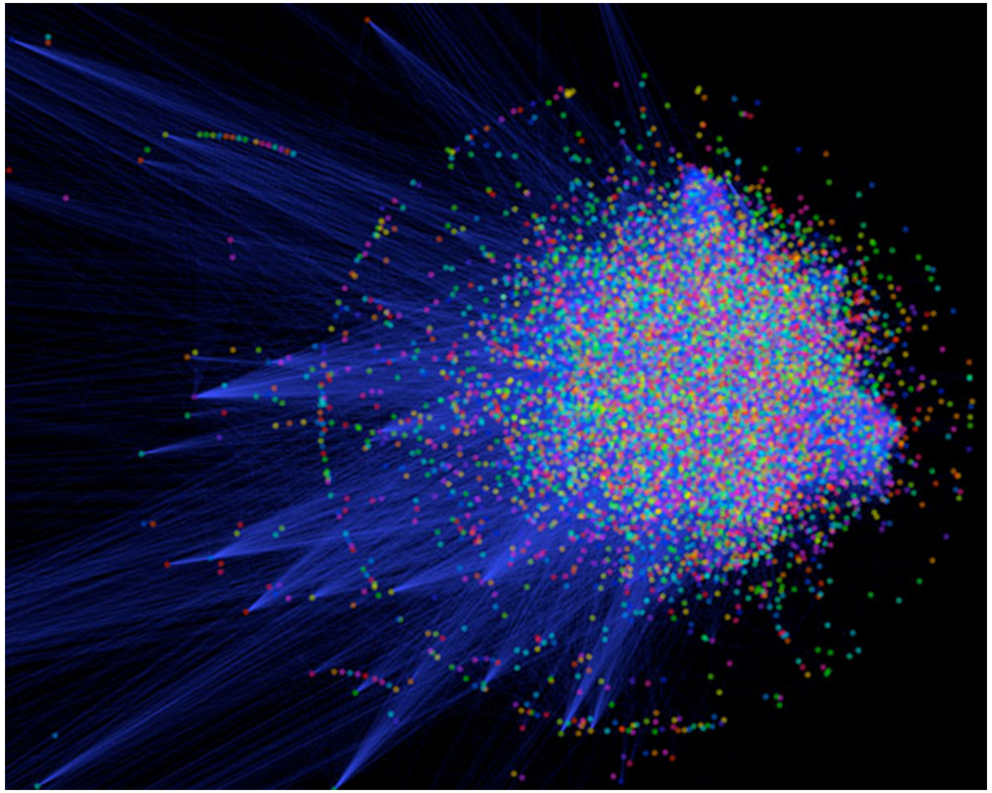
Beginning Perl for Bioinformatics
James Tisdall



Mastering Perl for Bioinformatics
James D. Tisdall

Introduction





- The connectivity of the internet (from the Wikipedia entry for “internet”)
- A map of human protein interactions (from the Wikipedia entry for “Protein–protein interaction”).

- We seek to understand biological principles on a genome-wide scale using the tools of bioinformatics.

Bioinformatics?

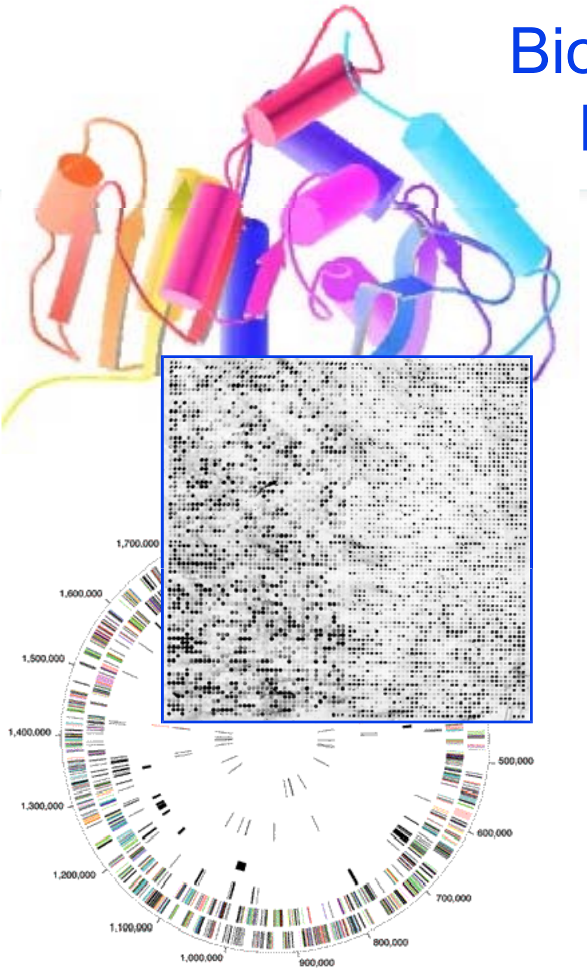
A quick google search with the keyword bioinformatics yields about **40.800.000** results !!!

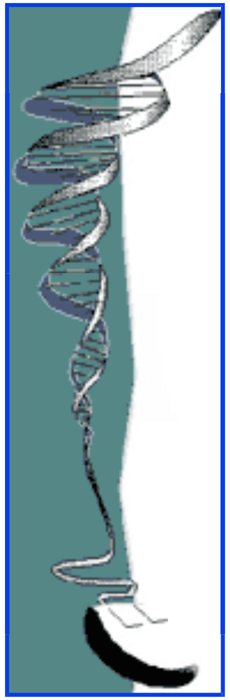
Synonyms:

- Computational Biology
- Computational Molecular Biology
- Biocomputing

Bioinformatics: A simple view

Biological Data + Computer Calculations





A marriage between Biology and Computers!

What is Bioinformatics?

(Molecular) **Bio** - **informatics**

One idea for a definition?

Bioinformatics is conceptualizing **biology in terms of molecules** (in the sense of physical- chemistry) and then applying **“informatics” techniques** (derived from disciplines such as applied math, CS, and statistics) to understand and **organize the information associated** with these molecules, **on a large-scale.**

Bioinformatics is a practical discipline with many **applications**.

Computing *versus* Biology

- what computer science is to molecular biology is like what mathematics has been to physics ......
  - -- Larry Hunter, ISMB’94

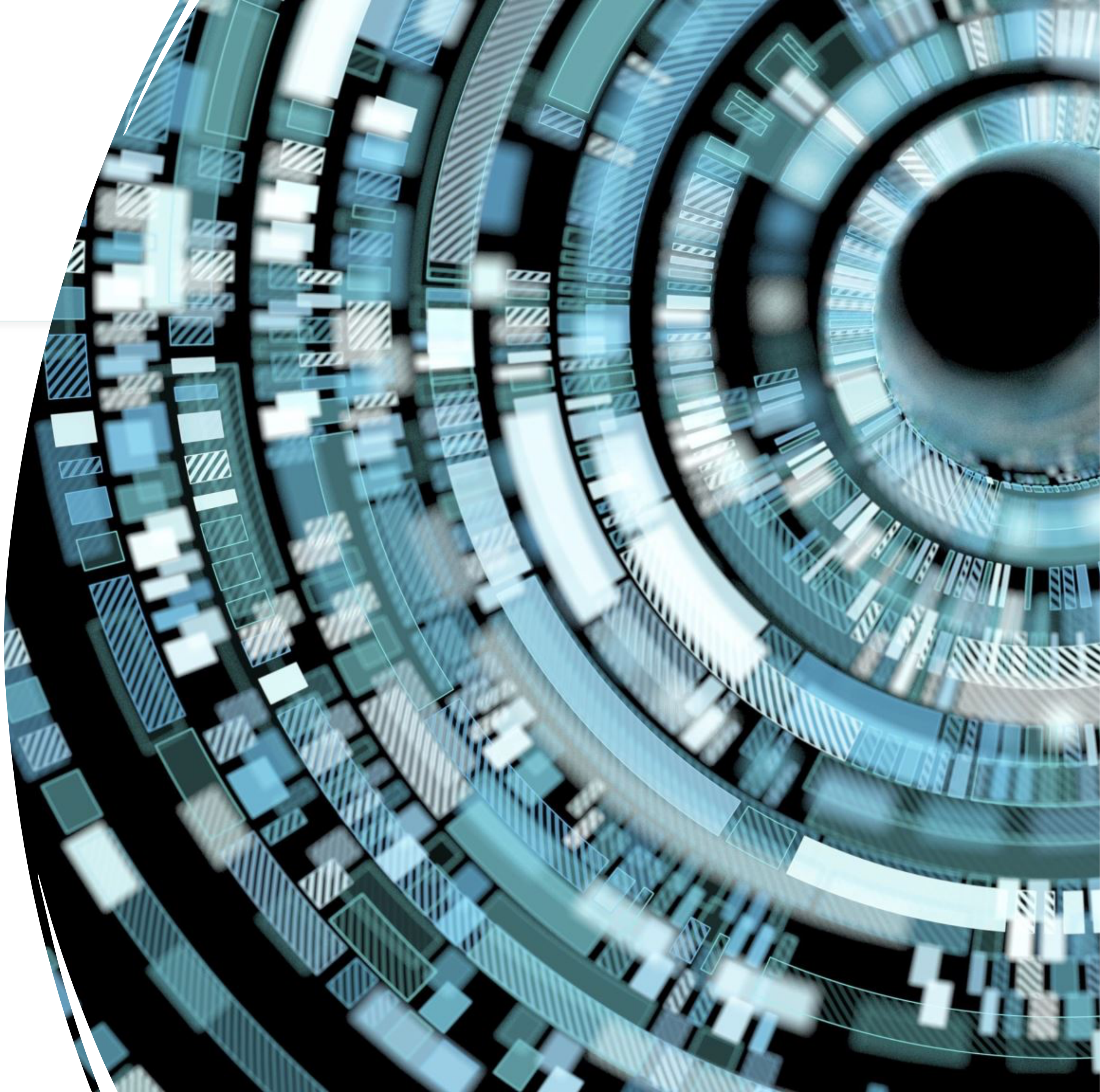
- molecular biology is (becoming) an information science.......
  - -- Leroy Hood, RECOMB’00

- bioinformatics ... is the research domain focused on linking the behavior of biomolecules, biological pathways, cells, organisms, and populations to the information encoded in the genomes
  - --Temple Smith, 2002,
  - Current Topics in Computational Molecular Biology

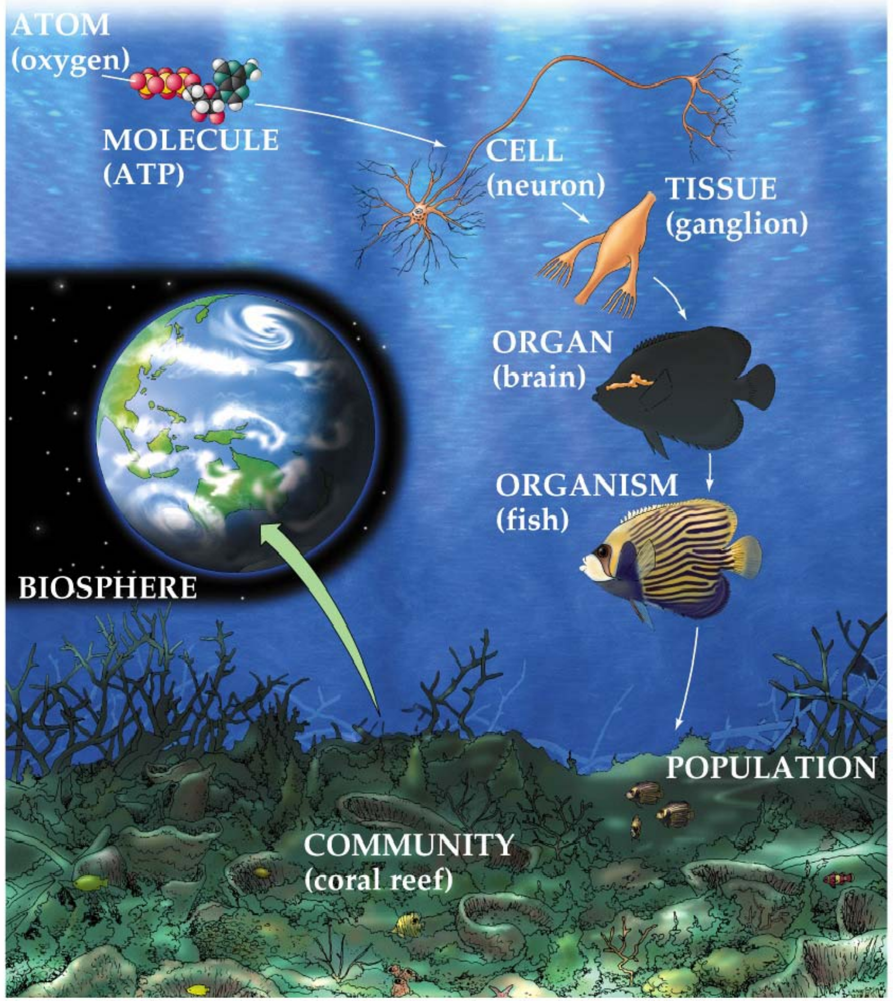
Computing *versus* Biology looking into the future

- Like physics, where general rules and laws are taught at the start, biology will surely be presented to future generations of students as a set of basic systems
- ……. duplicated and adapted to a very wide range of cellular and organismic functions, following basic evolutionary principles constrained by Earth's geological history.

--Temple Smith, 2002, Current Topics in Computational Molecular Biology

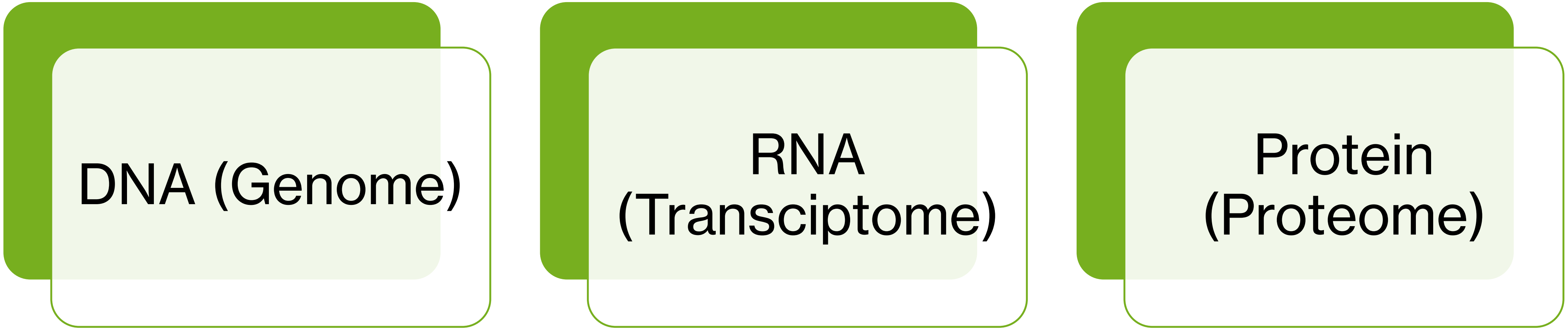


Scales of life



© 2001 Sinauer Associates, Inc.

Examples of biological data used in bioinformatics



DNA (Genome)

RNA (Transciptome)

Protein (Proteome)

What is done in bioinformatics?

Analysis and interpretation

Development of new algorithms and statistics

Development and implementation of tools

Why is Bioinformatics Important?

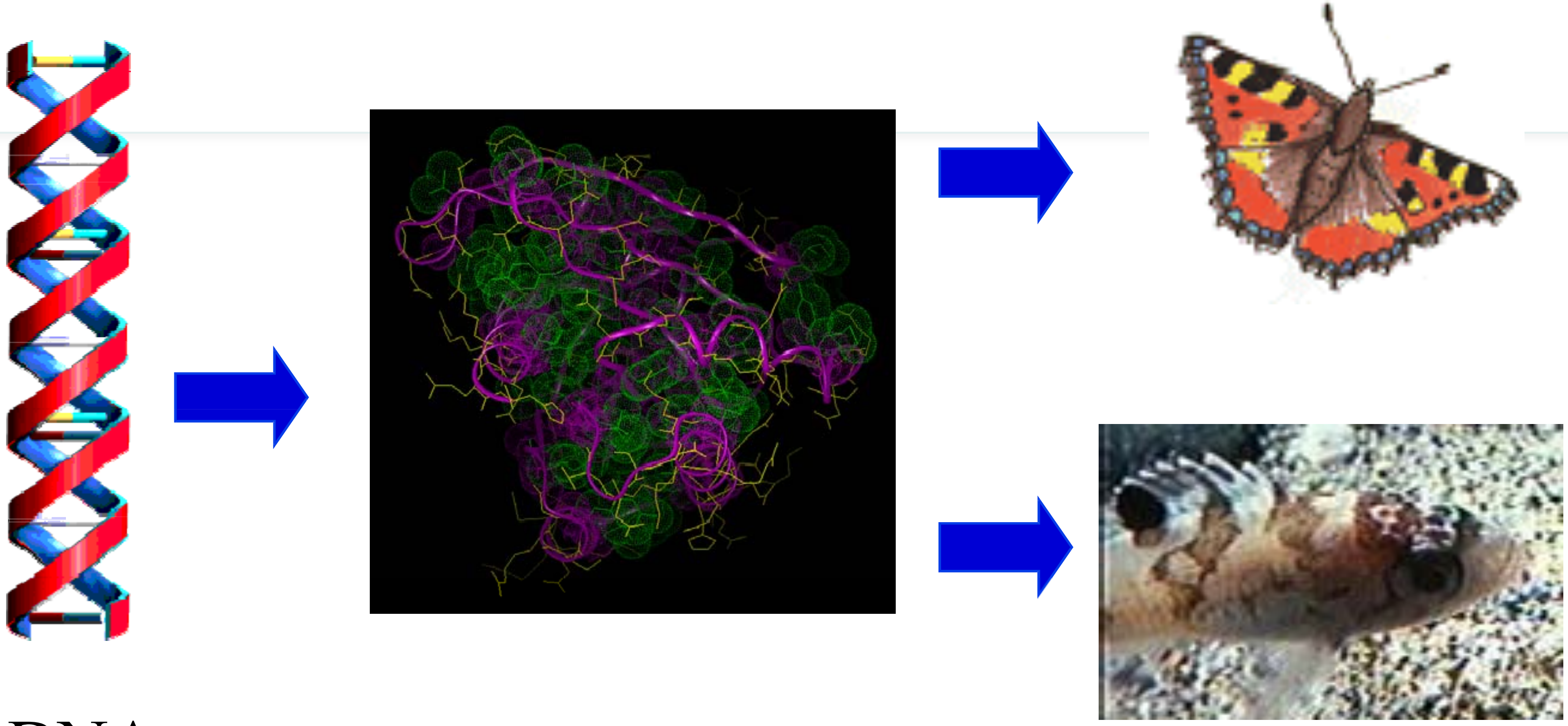
- Applications areas include
    - Medicine
    - Pharmaceutical drug design
    - Toxicology
    - Molecular evolution
    - Biosensors
    - Biomaterials
    - Biological computing models
    - DNA computing

What skills are needed?

- Well-grounded in one of the following areas:
    - Computer science
    - Molecular biology
    - Statistics

- Working knowledge and appreciation in the others!

Introductory Biology



DNA
(Genotype)

Protein

Phenotype

Molecular Biology Information - DNA

- Raw DNA Sequence
  - Coding or Not?
  - Parse into genes?
  - 4 bases: AGCT
  - ~1 Kb in a gene, ~2 Mb in genome
  - ~3 Gb Human

```
atggcaattaaaattggtatcaatggttttggtcgtatcggccgtatcgtattccgtgca
gcacaacaccgtgatgacattgaagttgtaggtattaacgacttaatcgacgttgaatac
atggcttatatgttgaaatatgattcaactcacggtcgtttcgacggcactgttgaagtg
aaagatggtaacttagtggttaatggtaaaactatccgtgtaactgcagaacgtgatcca
gcaaacttaaactggggtgcaatcggtgttgatatcgctgttgaagcgactggtttattc
ttaactgatgaaactgctcgtaaacatatcactgcaggcgcaaaaaaagttgtattaact
ggcccatctaaagatgcaacccctatgttcgttcgtggtgtaaacttcaacgcatacgca
ggtcaagatatcgtttctaacgcatcttgtacaacaaactgtttagctcctttagcacgt
gttgttcatgaaactttcggtatcaaagatggtttaatgaccactgttcacgcaacgact
gcaactcaaaaaactgtggatggtccatcagctaaagactggcgcggcggccgcggtgca
tcacaaaacatcattccatcttcaacaggtgcagcgaaagcagtaggtaaagtattacct
gcattaaacggtaaattaactggtatggctttccgtgttccaacgccaaacgtatctgtt
gttgatttaacagttaatcttgaaaaaccagcttcttatgatgcaatcaaacaagcaatc
aaagatgcagcggaaggtaaaacgttcaatggcgaattaaaaggcgtattaggttacact
gaagatgctgttgtttctactgacttcaacggttgtgctttaacttctgtatttgatgca
gacgctggtatcgcattaactgattctttcgttaaattggtatc . . .
```

```
.  .  .  caaaaatagggttaatatgaatctcgatctccattttgttcatcgtattcaa
caacaagccaaaactcgtacaaatatgaccgcacttcgctataaagaacacggcttgtgg
cgagatatctcttggaaaaactttcaagagcaactcaatcaactttctcgagcattgctt
gctcacaatattgacgtacaagataaaatcgccattttttgcccataatatggaacgttgg
gttgttcatgaaactttcggtatcaaagatggtttaatgaccactgttcacgcaacgact
acaatcgttgacattgcgaccttacaaattcgagcaatcacagtgcctatttacgcaacc
aatacagcccagcaagcagaatttatcctaaatcacgccgatgtaaaaattctcttcgtc
ggcgatcaagagcaatacgatcaaacattggaaattgctcatcattgtccaaaattacaa
aaaattgtagcaatgaaatccaccattcaattacaacaagatcctctttcttgcacttgg
```

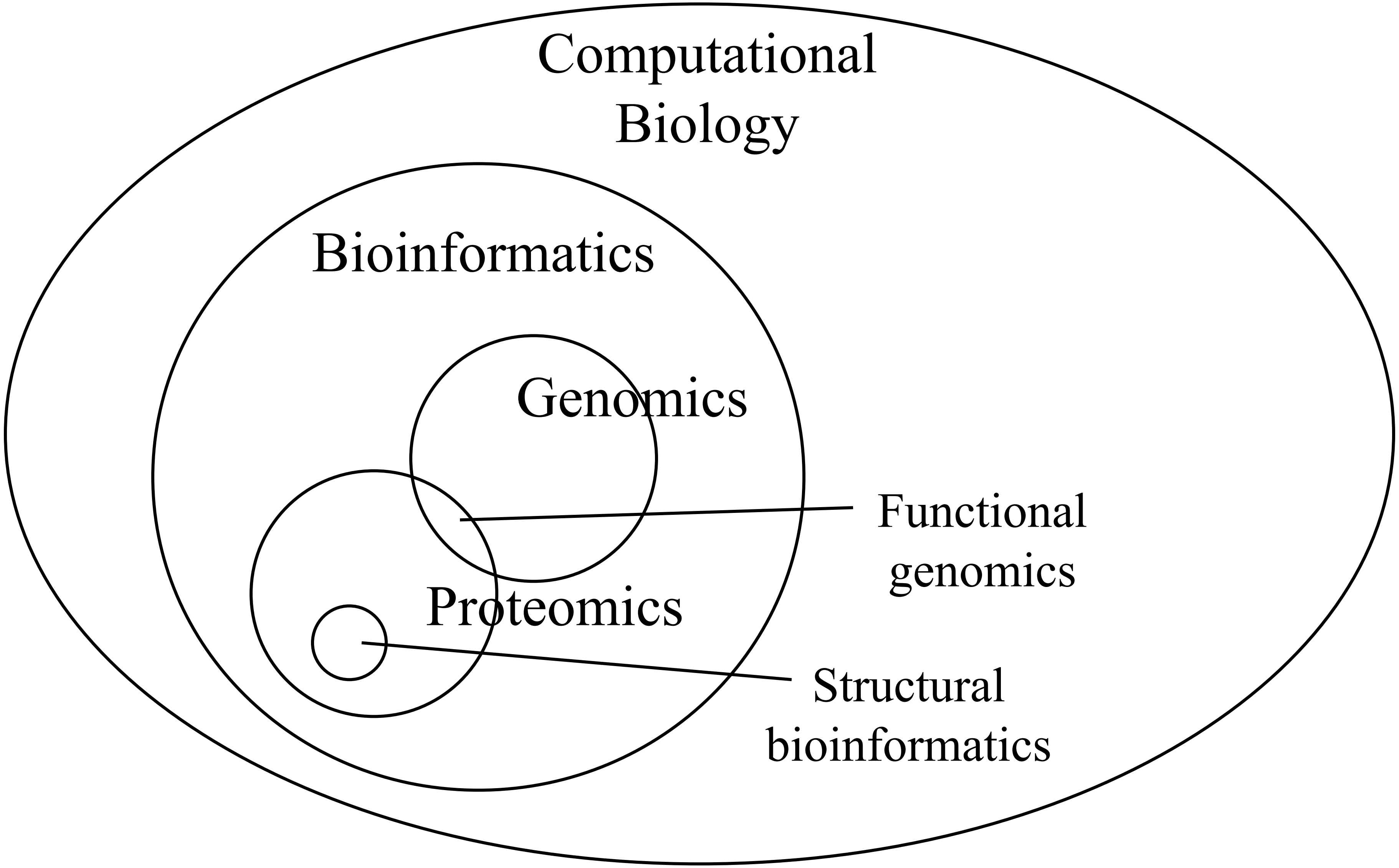
Molecular Biology Information: Protein Sequence

- 20 letter alphabet
  - ACDEFGHIKLMNPQRSTVWY   but not BJOUXZ
- Strings of ~300 aa in an average protein (in bacteria),
  ~200 aa in a domain
- ~13M known protein sequences, 500 000 well annotated.

```
d1dhfa_  LNCIVAVSQNMGIGKNGDLPWPPLRNEFRYFQRMTTTSSVEGKQ-NLVIMGKKTWFSI
d8dfr    LNSIVAVCQNMGIGKDGNLPWPPLRNEYKYFQRMTSTSHVEGKQ-NAVIMGKKTWFSI
d4dfra_  ISLIAALAVDRVIGMENAMPWN-LPADLAWFKRNTL--------NKPVIMGRHTWESI
d3dfr_  TAFLWAQDRDGLIGKDGHLPWH-LPDDLHYFRAQTV--------GKIMVVGRRTYESF

d1dhfa_  LNCIVAVSQNMGIGKNGDLPWPPLRNEFRYFQRMTTTSSVEGKQ-NLVIMGKKTWFSI
d8dfr    LNSIVAVCQNMGIGKDGNLPWPPLRNEYKYFQRMTSTSHVEGKQ-NAVIMGKKTWFSI
d4dfra_  ISLIAALAVDRVIGMENAMPW-NLPADLAWFKRNTLD--------KPVIMGRHTWESI
d3dfr_  TAFLWAQDRNGLIGKDGHLPW-HLPDDLHYFRAQTVG--------KIMVVGRRTYESF
```

Scope of Computational Biology



Genomics

- The study of the genome,
  - which is the complete set of the genetic material or DNA present in an organism.
- studies all genes and their inter relationships in an organism, so as to identify their combined influence on its growth and development.
- The field of genomics attracted worldwide attention in the late 1990s with the race to map the human genome.
  - The Human Genome Project (HGP), completed in April 2003, made available for the first time the complete genetic blueprint of a human being.

Proteomics

- large-scale study of proteomes,
  - which is a set of proteins produced in an organism, system, or biological context.
    - We may refer to, for instance, the proteome of a species (eg, Homo sapiens) or an organ (eg, the liver).
  - The proteome is not constant;
    - it differs from cell to cell and changes over time.
  - To some degree, the proteome reflects the underlying transcriptome.
    - However, protein activity (often assessed by the reaction rate of the processes in which the protein is involved) is also modulated by many factors in addition to the expression level of the relevant gene.

Proteomics

- is used to investigate:
  - when and where proteins are expressed;
  - rates of protein production, degradation, and steady-state abundance;
  - how proteins are modified (for example, post-translational modifications (PTMs) such as phosphorylation);
  - the movement of proteins between subcellular compartments;
  - the involvement of proteins in metabolic pathways;
  - how proteins interact with one another.
- can provide significant biological information for many biological problems, such as:
  - Which proteins interact with a particular protein of interest (for example, the tumor suppressor protein p53)?
  - Which proteins are localized to a subcellular compartment (for example, the mitochondrion)?
  - Which proteins are involved in a biological process (for example, circadian rhythm)?

Structural bioinformatics/genomics

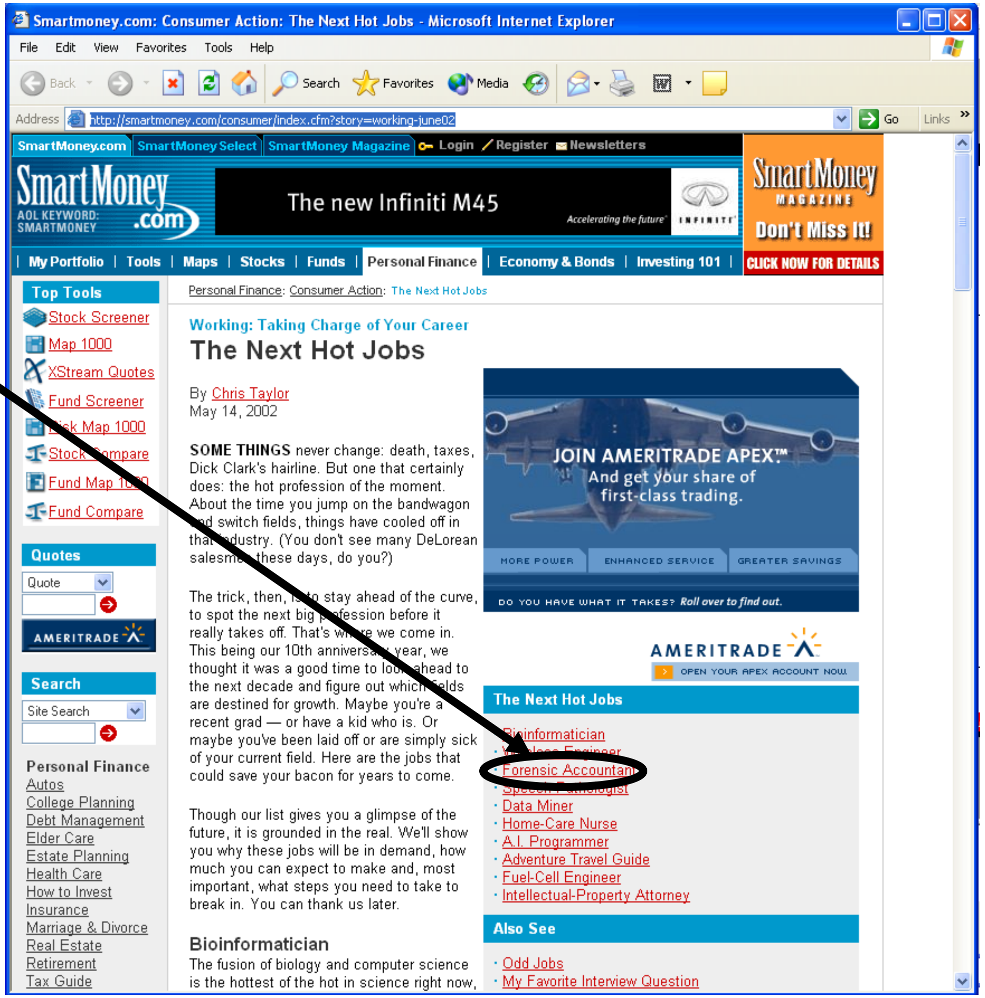
- is the branch of bioinformatics
  - which is related to the analysis and prediction of the three-dimensional structure of biological macromolecules such as proteins, RNA, and DNA.
- deals with generalizations about macromolecular 3D structure such as comparisons of overall folds and local motifs, principles of molecular folding, evolution, and binding interactions, and structure/function relationships, working both from experimentally solved structures and from computational models.

Functional genomics

- is a field of molecular biology,
  - which attempts to make use of the vast wealth of data given by genomic and transcriptomic projects (such as genome sequencing projects and RNA sequencing) to describe gene (and protein) functions and interactions.
    - Unlike structural genomics, it focuses on the dynamic aspects such as gene transcription, translation, regulation of gene expression and protein–protein interactions, as opposed to the static aspects of the genomic information such as DNA sequence or structures.
- attempts to answer questions about the function of DNA at the levels of genes, RNA transcripts, and protein products.

Why should I care?

- Bioinformatics ranks among #10 HotJobs
- Jobs available, exciting research potential
- Important information waiting to be decoded!



Personal Finance: Consumer Action: The Next Hot Jobs

Working: Taking Charge of Your Career

The Next Hot Jobs

By Chris Taylor
May 14, 2002

SOME THINGS never change: death, taxes, Dick Clark's hairline. But one that certainly does: the hot profession of the moment. About the time you jump on the bandwagon and switch fields, things have cooled off in that industry. (You don't see many DeLorean salesmen these days, do you?)

The trick, then, is to stay ahead of the curve, to spot the next big profession before it really takes off. That's where we come in. This being our 10th anniversary year, we thought it was a good time to look ahead to the next decade and figure out which fields are destined for growth. Maybe you're a recent grad — or have a kid who is. Or maybe you've been laid off or are simply sick of your current field. Here are the jobs that could save your bacon for years to come.

Though our list gives you a glimpse of the future, it is grounded in the real. We'll show you why these jobs will be in demand, how much you can expect to make and, most important, what steps you need to take to break in. You can thank us later.

Bioinformatician

The fusion of biology and computer science is the hottest of the hot in science right now,

The Next Hot Jobs

- Bioinformatician
- Wireless Engineer
- Forensic Accountant
- Speech Pathologist
- Data Miner
- Home-Care Nurse
- A.I. Programmer
- Adventure Travel Guide
- Fuel-Cell Engineer
- Intellectual-Property Attorney

Also See

- Odd Jobs
- My Favorite Interview Question

Why is bioinformatics hot?

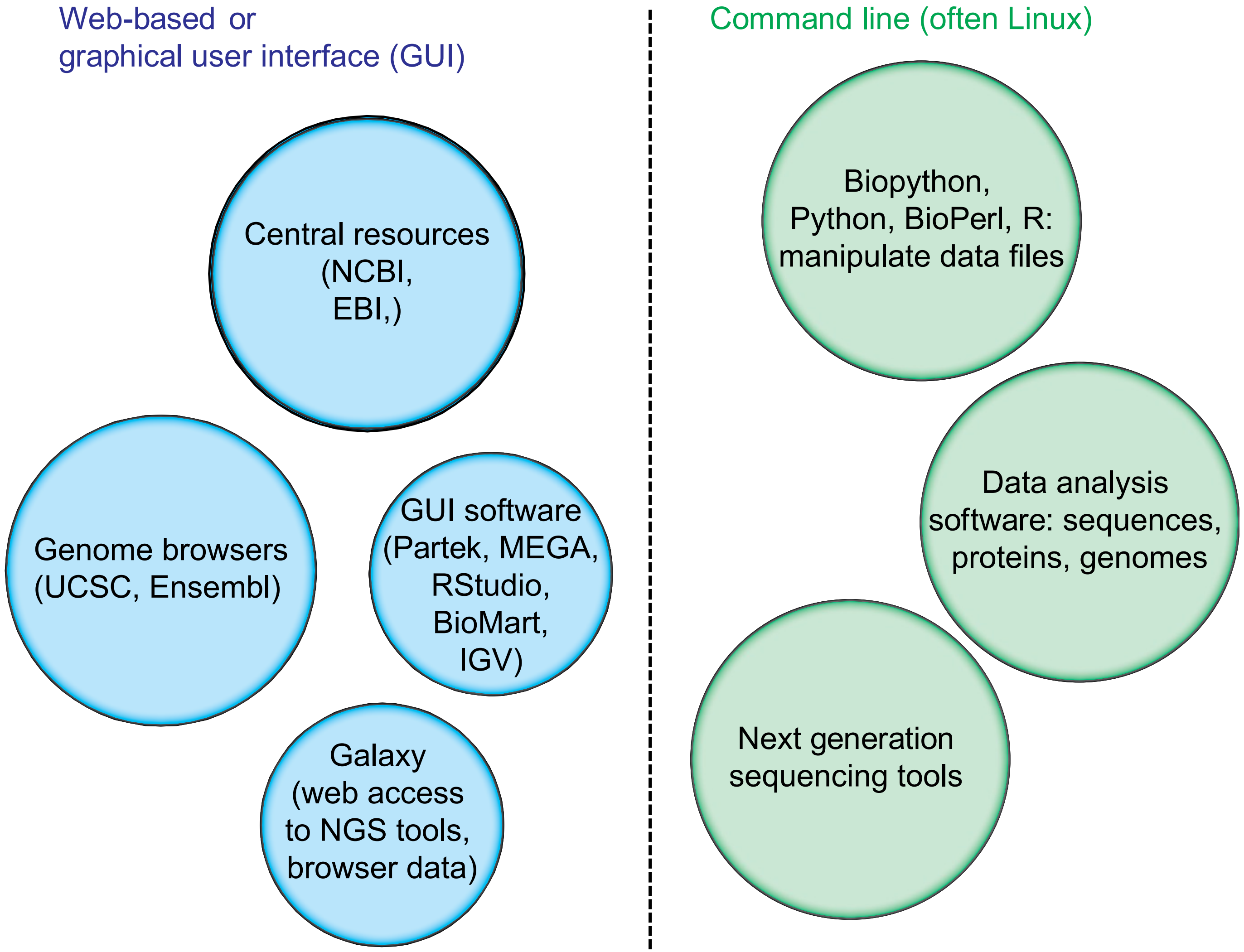
- Supply/demand: few people adequately trained in both biology and computer science

- Genome sequencing, microarrays, etc lead to large amounts of data to be analyzed

- Leads to important discoveries

- Saves time and money

Bioinformatics Software: Two Cultures



Web-based or
graphical user interface (GUI)

Central resources (NCBI, EBI,)

Genome browsers (UCSC, Ensembl)

GUI software (Partek, MEGA, RStudio, BioMart, IGV)

Galaxy (web access to NGS tools, browser data)

Command line (often Linux)

Biopython, Python, BioPerl, R: manipulate data files

Data analysis software: sequences, proteins, genomes

Next generation sequencing tools

Bioinformatics Software: Two Cultures

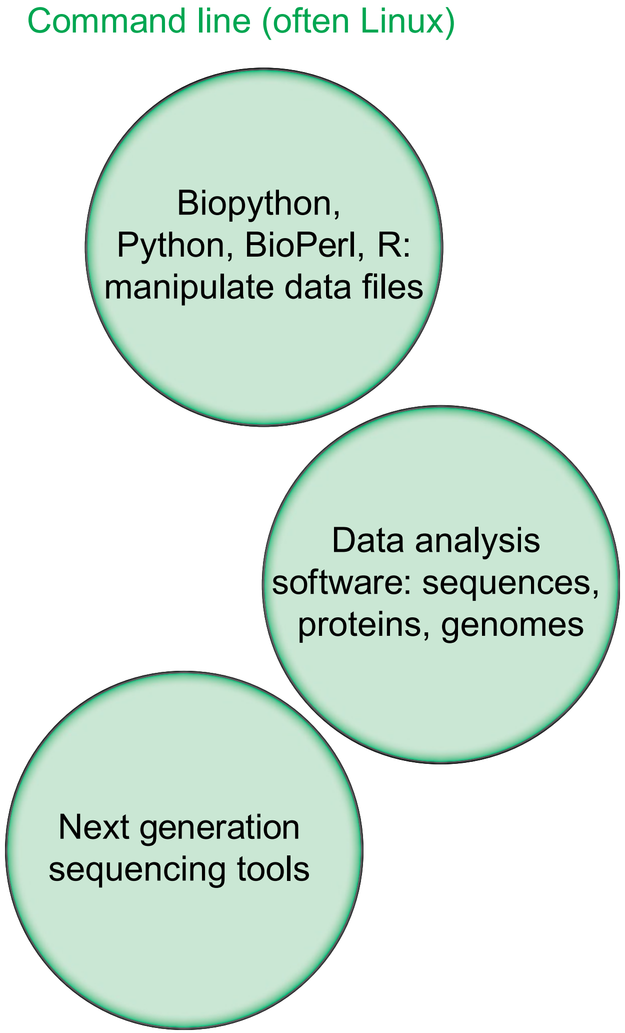
- Many bioinformatics tools and resources are available on the internet, such as major genome browsers and major portals (NCBI, Ensembl, UCSC).
- These are:
  - accessible (requiring no programming expertise)
  - easy to browse to explore their depth and breadth
  - very popular
  - familiar (available on any web browser on any platform)

Bioinformatics Software: Two Cultures

- Many bioinformatics tools and resources are available on the command-line interface (sometimes abbreviated CLI).
  - These are often on the Linux platform (or other Unix-like platforms such as the Mac command line).
  - They are essential for many bioinformatics and genomics applications.
  - Most bioinformatics software is written for the Linux platform.
    - Many bioinformatics datasets are so large (e.g. high throughput technologies generate millions to billions or even trillions of data points) requiring command-line tools to manipulate the data.

CLI

- Should you learn to use the Linux operating system?
  - Yes, if you want to use mainstream bioinformatics tools.
- Should you learn Python or Perl or R or another programming language?
  - It’s a good idea if you want to go deeper into bioinformatics, but also, it depends what your goals are.
  - Many software tools can be run in Linux on the command-line without needing to program.
- Think of this figure like a map.
  - Where are you now?
  - Where do you want to go?



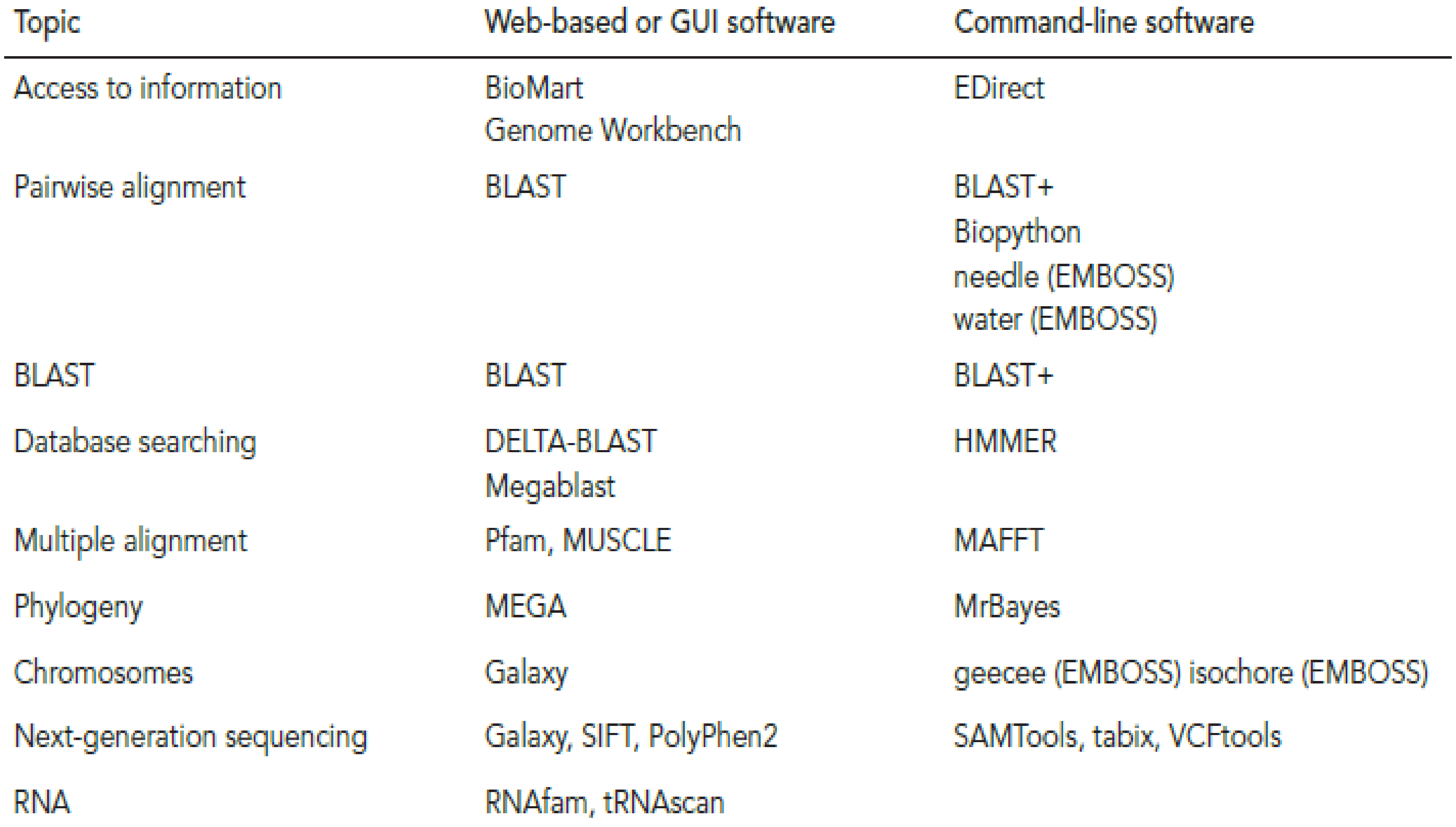
Command line (often Linux)

Biopython, Python, BioPerl, R: manipulate data files

Data analysis software: sequences, proteins, genomes

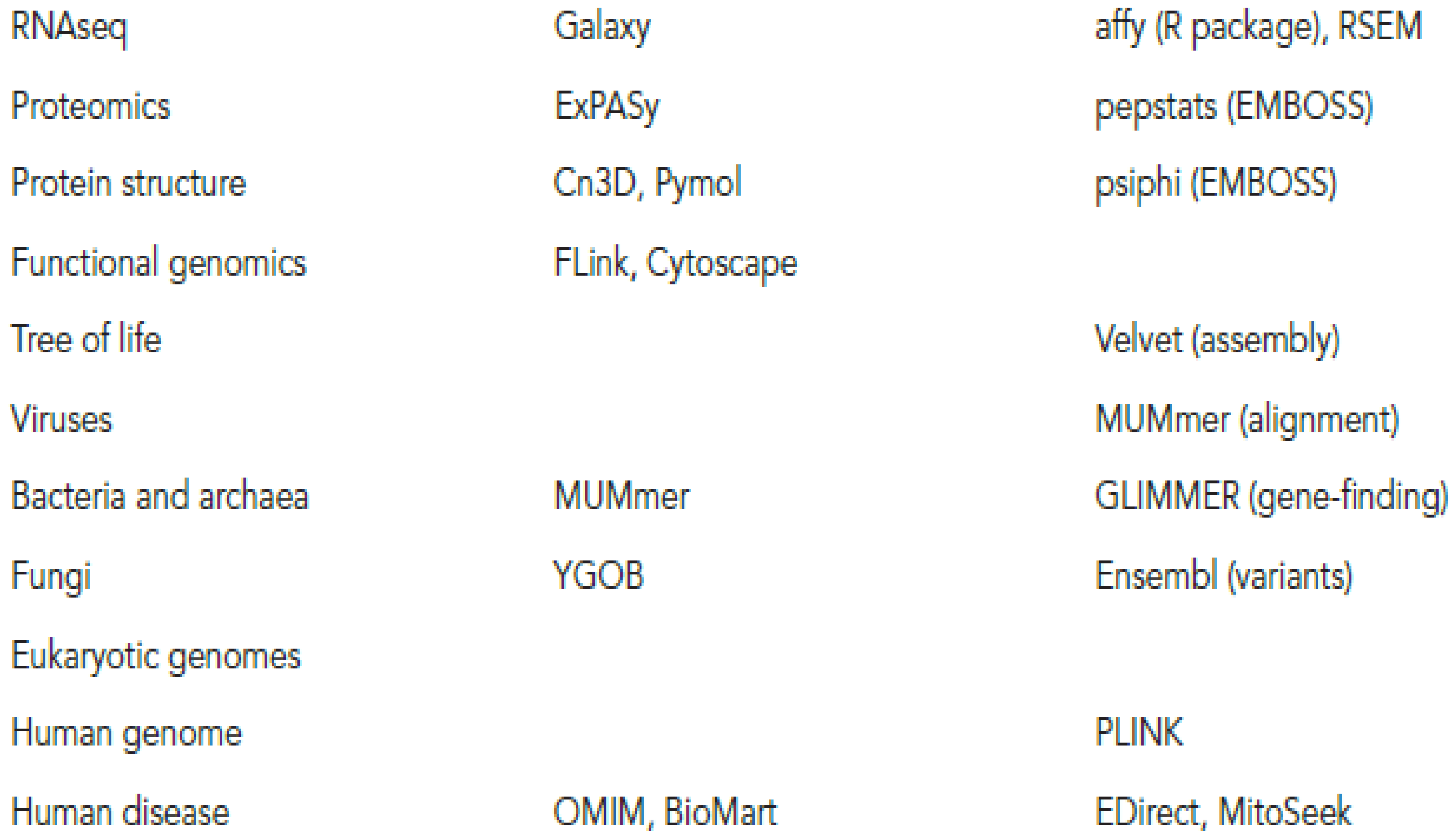
Next generation sequencing tools

Some web-based (GUI) and command-line (CLI) software



Topic	Web-based or GUI software	Command-line software
Access to information	BioMart Genome Workbench	EDirect
Pairwise alignment	BLAST	BLAST+ Biopython needle (EMBOSS) water (EMBOSS)
BLAST	BLAST	BLAST+
Database searching	DELTA-BLAST Megablast	HMMER
Multiple alignment	Pfam, MUSCLE	MAFFT
Phylogeny	MEGA	MrBayes
Chromosomes	Galaxy	geecee (EMBOSS) isochore (EMBOSS)
Next-generation sequencing	Galaxy, SIFT, PolyPhen2	SAMTools, tabix, VCFtools
RNA	RNAfam, tRNAscan	

Some web-based (GUI) and command-line (CLI) software



RNAseq	Galaxy	affy (R package), RSEM
Proteomics	ExPASy	pepstats (EMBOSS)
Protein structure	Cn3D, Pymol	psiphi (EMBOSS)
Functional genomics	FLink, Cytoscape	
Tree of life		Velvet (assembly)
Viruses		MUMmer (alignment)
Bacteria and archaea	MUMmer	GLIMMER (gene-finding)
Fungi	YGOB	Ensembl (variants)
Eukaryotic genomes		
Human genome		PLINK
Human disease	OMIM, BioMart	EDirect, MitoSeek