Introduction to Bioinformatics

Prof. Dr. Nizamettin AYDIN

naydin@yildiz.edu.tr

http://www3.yildiz.edu.tr/~naydin

Course Details

- Course Code : **BLM 3810**
- Course Name : **Introduction to Bioinformatics**
- Credit : **3**
- Course Level : **Undergradute**
- Course web page:
  http://www3.yildiz.edu.tr/~naydin/na_I2B.htm
- Instructors : Nizamettin AYDIN
  Room: D-128
  Email: naydin@yildiz.edu.tr, nizamettinaydin@gmail.com

Assesement

Quiz	:	10%
Midterm	:	25%
Homework	:	20%
Final	:	40%
Attendance & participation	:	05%

(The requirement for attendance is 70%)

Rules of the Conduct

- No eating /drinking in class
  - *except water*
- Cell phones must be kept outside of class or switched-off during class
  - *If your cell-phone rings during class or you use it in any way, you will be asked to leave and counted as unexcused absent.*
- No web surfing and/or unrelated use of computers,
  - when computers are used in class or lab.

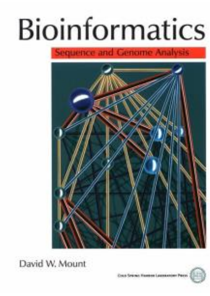
Rules of the Conduct

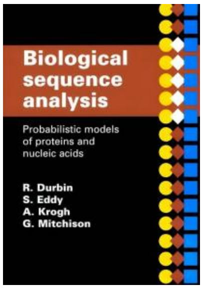
- You are responsible for checking the class web page often for announcements.
  - http://www.yildiz.edu.tr/~naydin/na_I2B.htm
- Academic dishonesty and cheating
  - will not be tolerated
  - will be dealt with according to university rules and regulations
    - http://www.yok.gov.tr/content/view/475/
    - Presenting any work that does not belong to you is also considered academic dishonesty.

Attendance Policy

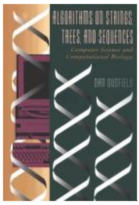
- The requirement for attendance is 70%
  - Hospital reports are not accepted to fulfill the requirement for attendance.
  - The students, who fail to fulfill the attendance requirement, will be excluded from the final exams and the grade of F0 will be given.
- Link for the rules and regulations:
  http://www.ogi.yildiz.edu.tr/category.php?id=17
  http://www.yok.gov.tr/content/view/544/230/lang,tr_TR/

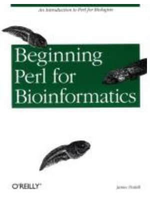
Recommended Texts

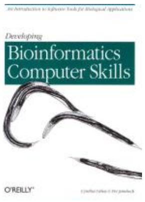


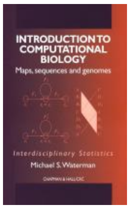

Biological sequence analysis
Probabilistic models of proteins and nucleic acids
R. Durbin
S. Eddy
A. Krogh
G. Mitchison

Recommended Texts - 2

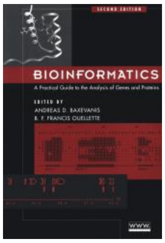

ALGORITHMS ON STRINGS, TREES, AND SEQUENCES
Computer Science and Computational Biology
DAN GUSFIELD

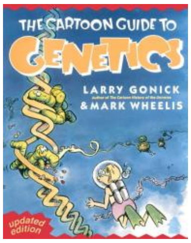

An Introduction to Perl for Biologists
Beginning Perl for Bioinformatics
O'REILLY
James Tisdall

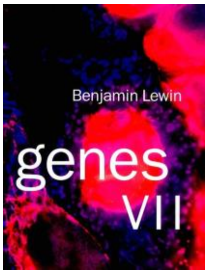

An Introduction to Software Tools for Biological Applications
Developing
Bioinformatics Computer Skills
O'REILLY


INTRODUCTION TO COMPUTATIONAL BIOLOGY
Maps, sequences and genomes
Interdisciplinary Statistics
Michael S. Waterman
CHAPMAN & HALL/CRC

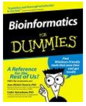
Recommended Texts - 3



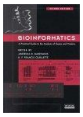




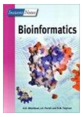
Recommended Texts - 4



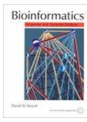
Bioinformatics for Dummies
Jean Claverie, Cedric Notredame



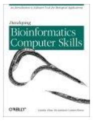
Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins
Andreas D. Baxevanis, B. F. Ouellette, Ouellette B. F. Francis.



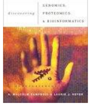
Instant Notes in Bioinformatics
D. R. Westhead, Richard M. Twyman, J. H. Parish



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

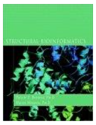


Developing Bioinformatics Computer Skills
Cynthia Gibas, Per Jambeck, Lorrie LeJeune (Editor)

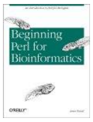


Discovering Genomics, Proteomics, and Bioinformatics
A. Malcolm Campbell, Laurie J. Heyer

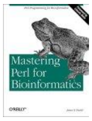
Recommended Texts - 5



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

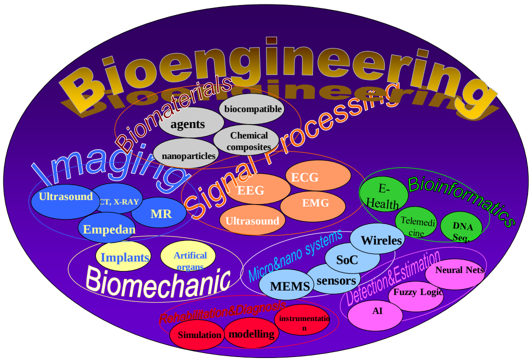


Beginning Perl for Bioinformatics
James Tisdall

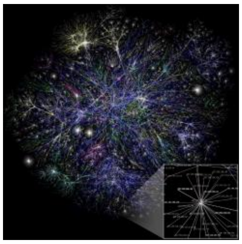


Mastering Perl for Bioinformatics
James D. Tisdall

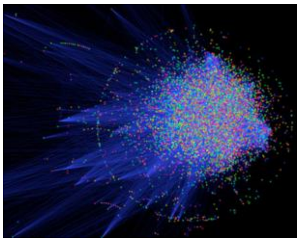
Bioengineering



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.

What is Bioinformatics?...

- A quick google search with the keyword bioinformatics yields about
  - About 37 100 000 results (0.65 sec.) (02. 03.2022)

- **Synonyms:**
  - Computational Biology
  - Computational Molecular Biology
  - Biocomputing

… What is Bioinformatics?...

- Bioinformatics
  - the study of how information is represented and transmitted in biological systems, starting at the molecular level

  is a discipline that does not need a computer!
  - An ink pen and a supply of traditional laboratory notebooks could be used to record results of experiments.
  - However, to do so would be like foregoing the use of a computer and word-processing program in favor of pen and paper to write a novel.

… What is Bioinformatics?...

- According to a National Institutes of Health (NIH) definition, bioinformatics is
  - "research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral, or health data, including those to acquire, store, organize, analyze, or visualize such data."
    - The related discipline of computational biology is "the development and application of data-analytical and theoretical methods, mathematical modeling, and computational simulation techniques to the study of biological, behavioral, and social systems."

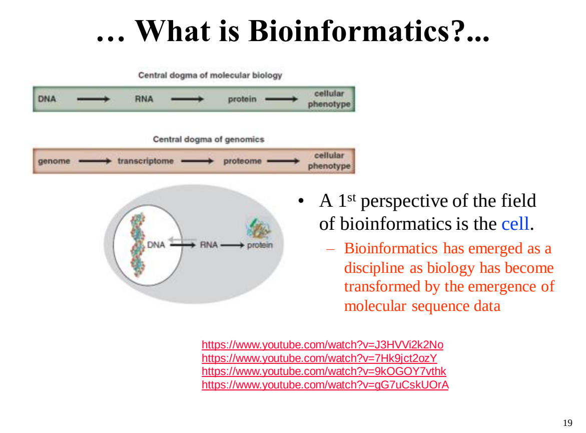
…What is Bioinformatics?...

- **From Webopedia:**
  - The application of computer technology to the management of biological information.
  - Specifically, it is the science of developing computer databases and algorithms to facilitate and expedite biological research.
  - Bioinformatics is being used largely in the field of human genome research by the Human Genome Project that has been determining the sequence of the entire human genome (about 3 billion base pairs) and is essential in using genomic information to understand diseases.
  - It is also used largely for the identification of new molecular targets for drug discovery.

… What is Bioinformatics?...

- Another definition from the National Human Genome Research Institute (NHGRI) is that
  - "Bioinformatics is the branch of biology that is concerned with the acquisition, storage, display, and analysis of the information found in nucleic acid and protein sequence data."
- Russ Altman (1998) and Altman and Dugan (2003) offer two definitions of bioinformatics.
  - The first involves information flow following the central dogma of molecular biology (next slide)
  - The second definition involves information flow that is transferred based on scientific methods. This definition includes problems such as
    - designing, validating, and sharing software;
    - storing and sharing data;
    - performing reproducible research workflows;
    - interpreting experiments.

Altman, R.B. 1998. Bioinformatics in support of molecular medicine. *Proceedings of AMIA Symposium* **1998**, 53–61. PMID: 9929182.
Altman, R.B., Dugan, J.M. 2003. Defining bioinformatics and structural bioinformatics. *Methods of Biochemical Analysis* **44**, 3–14. PMID: 12647379.

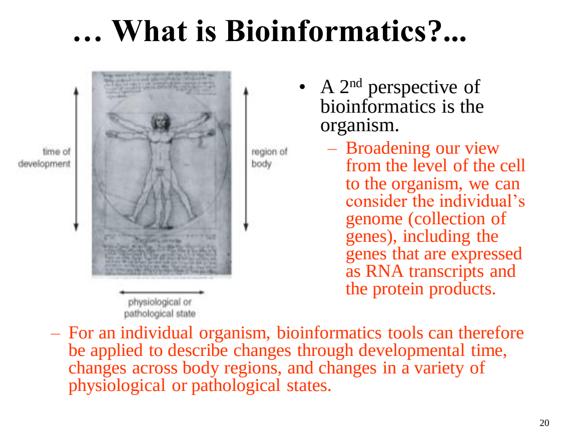


… What is Bioinformatics?...

Central dogma of molecular biology
DNA → RNA → protein → cellular phenotype
Central dogma of genomics
genome → transcriptome → proteome → cellular phenotype
DNA → RNA → protein

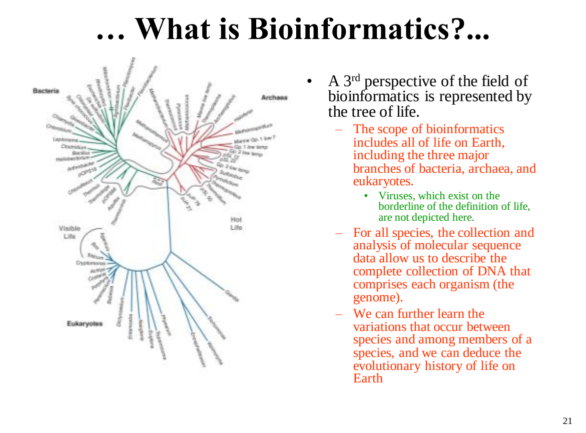
- A 1st perspective of the field of bioinformatics is the cell.
  - Bioinformatics has emerged as a discipline as biology has become transformed by the emergence of molecular sequence data

https://www.youtube.com/watch?v=J3HVVi2k2No
https://www.youtube.com/watch?v=7Hk9jct2ozY
https://www.youtube.com/watch?v=9kOGOY7vthk
https://www.youtube.com/watch?v=gG7uCskUOrA



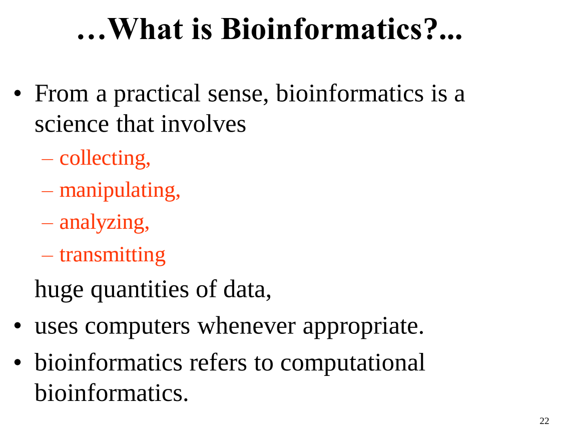
… What is Bioinformatics?...

- A 2nd perspective of bioinformatics is the organism.
  - Broadening our view from the level of the cell to the organism, we can consider the individual’s genome (collection of genes), including the genes that are expressed as RNA transcripts and the protein products.
  - For an individual organism, bioinformatics tools can therefore be applied to describe changes through developmental time, changes across body regions, and changes in a variety of physiological or pathological states.



… What is Bioinformatics?...

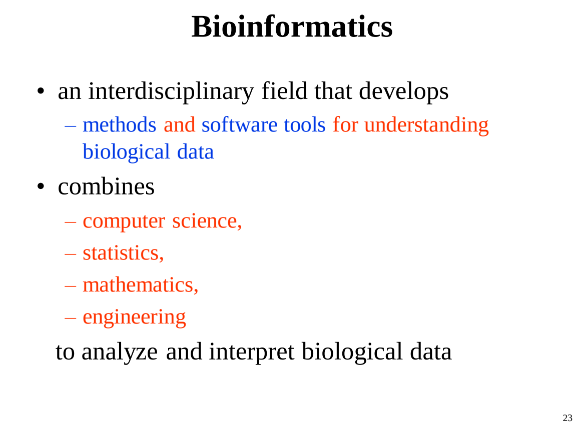
- A 3rd perspective of the field of bioinformatics is represented by the tree of life.
  - The scope of bioinformatics includes all of life on Earth, including the three major branches of bacteria, archaea, and eukaryotes.
    - Viruses, which exist on the borderline of the definition of life, are not depicted here.
  - For all species, the collection and analysis of molecular sequence data allow us to describe the complete collection of DNA that comprises each organism (the genome).
  - We can further learn the variations that occur between species and among members of a species, and we can deduce the evolutionary history of life on Earth



…What is Bioinformatics?...

- From a practical sense, bioinformatics is a science that involves
  - collecting,
  - manipulating,
  - analyzing,
  - transmitting

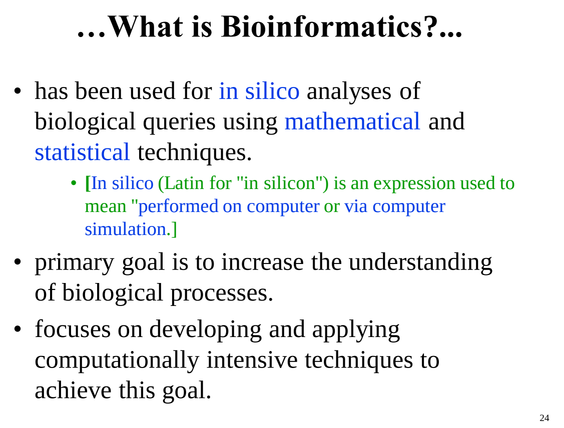
  huge quantities of data,
- uses computers whenever appropriate.
- bioinformatics refers to computational bioinformatics.



Bioinformatics

- an interdisciplinary field that develops
  - methods and software tools for understanding biological data
- combines
  - computer science,
  - statistics,
  - mathematics,
  - engineering

  to analyze and interpret biological data



…What is Bioinformatics?...

- has been used for in silico analyses of biological queries using mathematical and statistical techniques.
  - [In silico (Latin for "in silicon") is an expression used to mean "performed on computer or via computer simulation.]
- primary goal is to increase the understanding of biological processes.
- focuses on developing and applying computationally intensive techniques to achieve this goal.

…What is Bioinformatics?...

- Techniques used include
  - pattern recognition, data mining, machine learning algorithms, and visualization
- Analyzing biological data to produce meaningful information involves writing and running software programs that use algorithms from
  - graph theory, artificial intelligence, soft computing, data mining, signal processing, image processing, and computer simulation.

…What is Bioinformatics?...

- The algorithms in turn depend on theoretical foundations such as
  - discrete mathematics
  - control theory
  - system theory
  - information theory
  - statistics

...What is Bioinformatics?...

- Bioinformatics derives knowledge from computer analysis of biological data that can consist of the information stored in the
  - genetic code,
  - experimental results from various sources,
  - patient statistics,
  - scientific literature.
- Research in bioinformatics includes method development for
  - storage,
  - retrieval,
  - analysis

  of the data.

...What is Bioinformatics?...

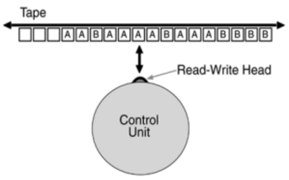
- Bioinformatics
  - a rapidly developing branch of biology
  - highly interdisciplinary,
  - using techniques and concepts from
    - informatics,
    - statistics,
    - mathematics,
    - chemistry,
    - biochemistry,
    - physics,
    - linguistics.

…What is Bioinformatics?...

- The relationship between computer science and biology is a natural one for several reasons.
  - 1st,
    - the phenomenal rate of biological data being produced provides challenges:
      - massive amounts of data have to be stored, analysed, and made accessible.
  - 2nd,
    - the nature of the data is often such that a statistical method, and hence computation, is necessary.
      - This applies in particular to the information on the building plans of proteins and of the temporal and spatial organisation of their expression in the cell encoded by the DNA.
  - 3rd,
    - there is a strong analogy between the DNA sequence and a computer program
      - it can be shown that the DNA represents a Turing Machine.

The Turing Machine

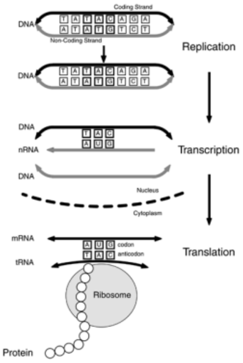
- The Turing Machine
  - can simulate any computing system
  - consists of three basic elements:
    - a control unit,
    - a tape,
    - a read-write head.



- The read-write head moves along the tape and transmits information to and from the control unit.

The Central Dogma of Molecular Biology

- DNA is transcribed to messenger RNA in the cell nucleus, which is in turn translated to protein in the cytoplasm.
- The Central Dogma, shown here from a structural perspective, can also be depicted from an information flow perspective



Path to the Bioinformatics

- 1st,
  - Learn Biology.
- 2nd,
  - Decide and pick a problem that interests you for experiment.
- 3rd,
  - Find and learn about the Bioinformatics tools.
- 4th,
  - Learn the Computer Programming Languages.
    - Perl, Pyton, R, Java, etc.
- 5th,
  - Experiment on your computer and learn different programming techniques.

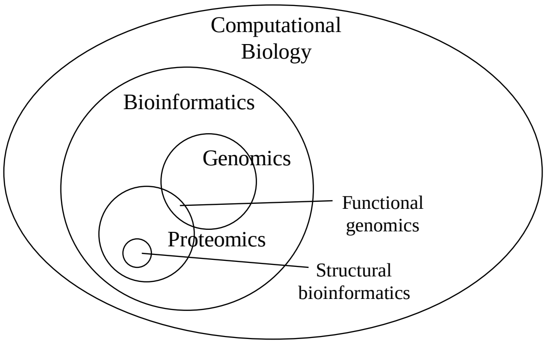
Why is Bioinformatics Important?

- 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!

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 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.

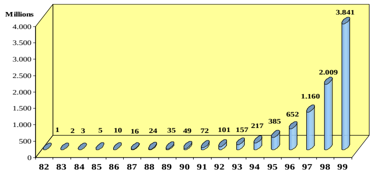
Why is bioinformatics hot?

- 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

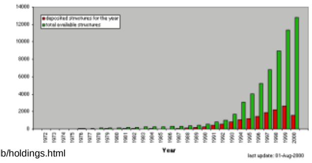
The Role of *Computational* Biology

GenBank BASEPAIR GROWTH

Source: GenBank



3D Structures Growth:

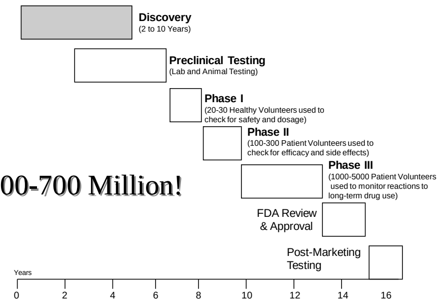


Source: http://www.rcsb.org/pdb/holdings.html

Fighting Human Disease

- Genetic / Inherited
  - Diabetes
- Viral
  - Flu, common cold
- Bacterial
  - Meningitis, Strep throat

Drug Development Life Cycle



Discovery (2 to 10 Years)

Preclinical Testing (Lab and Animal Testing)

Phase I (20-30 Healthy Volunteers used to check for safety and dosage)

Phase II (100-300 Patient Volunteers used to check for efficacy and side effects)

Phase III (1000-5000 Patient Volunteers used to monitor reactions to long-term drug use)

FDA Review & Approval

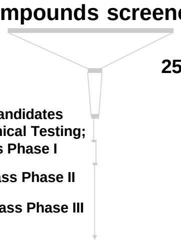
Post-Marketing Testing

$600-700 Million!

7 – 15 Years!

Drug lead screening

5,000 to 10,000 compounds screened



250 Lead Candidates in Preclinical Testing

5 Drug Candidates enter Clinical Testing; 80% Pass Phase I

30%Pass Phase II

80% Pass Phase III

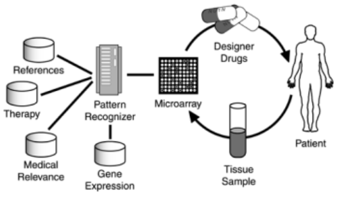
One drug approved by the FDA

Killer application

- In the biotechnology industry, every researcher and entrepreneur hopes to develop or discover the next "killer app“
  - the one application that will bring the world to his or her door and provide funding for R&D, marketing, and production.
    - For example, in general computing, the electronic spreadsheet and the desktop laser printer have been the notable killer apps.
    - The spreadsheet not only transformed the work of accountants, research scientists, and statisticians, but the underlying tools formed the basis for visualization and mathematical modeling.
    - The affordable desktop laser printer created an industry and elevated the standards of scientific communications, replacing rough graphs created on dot-matrix printers with high-resolution images.

Killer application

- "What might be the computer-enabled 'killer app' in bioinformatics?“
- Although there are numerous military and agricultural opportunities, one of the most commonly cited examples of the killer app is in personalized medicine, as illustrated in Figure



personalized medicine: the custom, just-in-time delivery of medications (popularly called "designer drugs") tailored to the patient'scondition.

Killer application

- Instead of taking a generic or over-the-counter drug for a particular condition, a patient would submit a tissue sample, such as a mouth scraping, and submit it for analysis.
  - A microarray would then be used to analyze the patient's genome and the appropriate compounds would be prescribed.
- The drug could be a cocktail of existing compounds, much like the drug cocktails used to treat cancer patients today.
- Alternatively, the drug could be synthesized for the patient's specific genetic markers—as in tumor specific chemotherapy, for example.
  - This synthesized drug might take a day or two to develop, unlike the virtually instantaneous drug cocktail.
  - The tradeoff is that the drug would be tailored to the patient's genetic profile and condition, resulting in maximum response to the drug, with few or no side effects.

Killer application

- How will this or any other killer app be realized?
  - The answer lies in addressing the molecular biology, computational, and practical business aspects of proposed developments such as custom medications.
- A practical system would include:
  - High throughput screening
    - The use of affordable, computer-enabled microarray technology to determine the patient's genetic profile. The issue here is affordability, in that microarrays costs tens of thousands of dollars

Killer application

- Medically relevant information gathering
  - Databases on gene expression, medical relevance of signs and symptoms, optimum therapy for given diseases, and references for the patient and clinician must be readily available.
  - The goal is to be able to quickly and automatically match a patient's genetic profile, predisposition for specific diseases, and current condition with the efficacy and potential side effects of specific drug-therapy options.
- Custom drug synthesis
  - The just-in-time synthesis of patient-specific drugs, based on the patient's medical condition and genetic profile, presents major technical as well as political, social, and legal hurdles.
  - For example, for just-in-time synthesis to be accepted by the FDA, the pharmaceutical industry must demonstrate that custom drugs can skip the clinical-trials gauntlet before approval.

Killer application

- Achieving this killer app in biotech is highly dependent on
  - computer technology,
    - especially in the use of computers to speed the process testing-analysis-drug synthesis cycle, where time really is money.
- For example, consider that for every 5,000 compounds evaluated annually by the U.S. pharmaceutical R&D laboratories, 5 make it to human testing, and only 1 of the compounds makes it to market.

Killer application

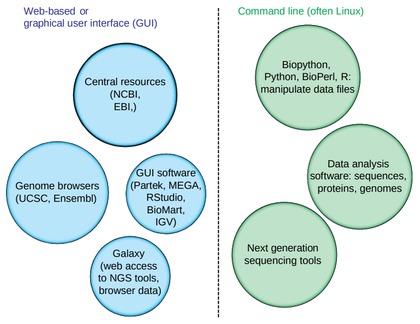
- In addition, the average time to market for a drug is over 12 years,
  - including several years of pre-clinical trials followed by a 4-phase clinical trial.
- These clinical trials progress from
  - safety and dosage studies in Phase I,
  - to effectiveness and side effects in Phase II,
  - to long-term surveillance in Phase IV,

  with each phase typically lasting several years.

Killer application

- Most pharmaceutical companies view computerization as the solution to creating smaller runs of drugs focused on custom production.
- Obvious computing applications range from
  - predicting efficacy and side effects of drugs based on genome analysis,
  - to visualizing protein structures to better understand and predict the efficacy of specific drugs,
  - to illustrating the relative efficacy of competing drugs in terms of quality of life and cost, based on the Markov simulation of likely outcomes during Phase IV clinical trials.

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

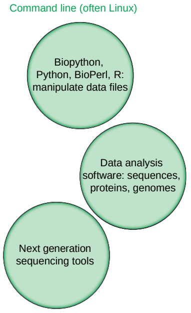
- Many bioinformatics tools and resources are available on the internet, such as major genome browsers and major portals (NCBI, Ensembl, UCSC).
- These are:
  - – accessible (requiring no programming expertise)
  - – easy to browse to explore their depth and breadth
  - – very popular
  - – familiar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

Bioinformatics Software: Two Cultures

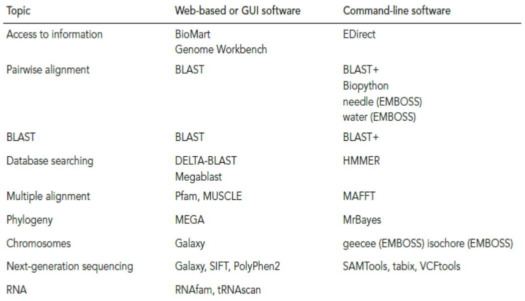
- Many bioinformatics tools and resources are available on the command-line interface (sometimes abbreviated CLI).
  - – These are often on the Linux platform (or other Unix-like platforms such as the Mac command line).
  - – They are essential for many bioinformatics and genomics applications.
  - – 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?

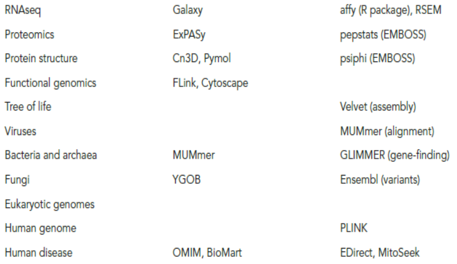

Command line (often Linux)
Biopython, Python, BioPerl, R: manipulate data files
Data analysis software: sequences, proteins, genomes
Next generation sequencing tools

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



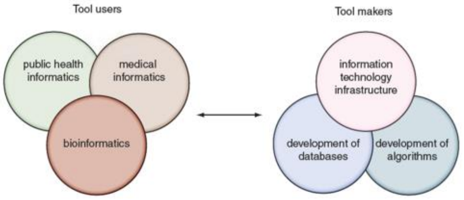
Topic	Web-based or GUI software	Command-line software
Access to information	BioMart Genome Workbench	EDirect
Pairwise alignment	BLAST	BLAST+ Biopython needle (EMBOSS) water (EMBOSS)
BLAST	BLAST	BLAST+
Database searching	DELTA-BLAST Megablast	HMMER
Multiple alignment	Pfam, MUSCLE	MAFFT
Phylogeny	MEGA	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

Tool makers and tool users across informatics disciplines


Tool users
public health informatics
medical informatics
bioinformatics
Tool makers
information technology infrastructure
development of databases
development of algorithms

- Many informatics disciplines have emerged in recent years.
- Bioinformatics is distinguished by its particular focus on DNA and proteins (impacting its databases, its tools, and its entire culture).

Learning Programming for Bioinformatics

- In addition to available books and courses, many websites offer online training in the forms of tutorials or courses.
- Rules for online learning include:
  - make a plan;
  - be selective;
  - organize your learning environment;
  - do the readings;
  - do the exercises;
  - do the assessments;
  - exploit the advantages (e.g., convenience);
  - reach out to others;
  - document your achievements;
  - be realistic about your expectations for what you can learn.
- These rules also apply to reading a textbook.

https://doi.org/10.1371/journal.pcbi.1002631
https://doi.org/10.1371/journal.pcbi.1000589

Reproducible Research in Bioinformatics

- Science by its nature is cumulative and progressive.
- Whether you use web-based or command-line tools, research should be conducted in a way that is reproducible by the investigator and by others.
- This facilitates the cumulative, progressive nature of your work.

Reproducible Research in Bioinformatics

- A workflow should be well documented.
  - This may include keeping text documents on your computer in which you can copy and paste complex commands, URLs, or other forms of data.
- To facilitate your work, information stored on a computer should be well organized.
- Data should be made available to others.
  - Repositories are available to store high-throughput data in particular.
    - Examples are Gene Expression Omnibus (GEO) and Sequence Read Archive (SRA) at NCBI and ArrayExpress and European Nucleotide Archive (ENA) at EBI.

Reproducible Research in Bioinformatics

- Metadata can be equally as crucial as data.
  - Metadata refers to information about datasets.
    - For a bacterial genome that has been sequenced, the metadata may include the location from which the bacterium was isolated, the culture conditions, and whether it is pathogenic.
- Databases that are used should be documented.
  - Since the contents of databases change over time, it is important to document the version number and the date(s) of access.
- Software should be documented.
  - For established packages, the version number should be provided.
    - Further documenting the specific steps you use allows others to independently repeat your analyses.
      - In an effort to share software, many researchers use repositories such as GitHub.

https://doi.org/10.1371/journal.pcbi.1000424

Bioinformatic Research Areas

- Bioinformatics
- Genomics
- Transcriptomics
- Proteomics
- Epigenetics
- Interactomics
- Protein Analysis and Structure Prediction
- Next Generation Sequencing Technology
- Comparative Sequence Analysis
- Systems Biology
- Text Mining and Information Extraction

Where Can I Learn More?

- ISCB: http://www.iscb.org/
- NBCI: http://ncbi.nlm.nih.gov/
- http://www.bioinformatics.org/
- Books
- Journals
- Conferences

Where Can I Learn More?

- https://www.codeschool.com/
- https://www.codecademy.com/
- https://www.datacamp.com/
- https://software-carpentry.org/
- https://github.com/