Introduction to Bioinformatics

Prof. Dr. Nizamettin AYDIN

naydin@yildiz.edu.tr

Introduction

Learning objectives

• After studying these materials you should be able to do the following:
  – define the terms bioinformatics;
  – explain the scope of bioinformatics;
  – explain why globins are a useful example to illustrate this discipline; and
  – describe web-based versus command-line approaches to bioinformatics.

What is Bioinformatics?...

• A quick google search with the keyword bioinformatics
  – yields about 1.480.000 results (06.10.2008)
  – yields about 24.900.000 results (07.02.2016)
  – yields about 30.400.000 results (21.02.2018)

• 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.

- **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."

  - 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.

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

- **A 2nd perspective of bioinformatics is the organism.**

- A 3rd perspective of the field of bioinformatics is the tree of life.

- **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. These can consist of the information
  - stored in the genetic code,
  - experimental results from various sources,
  - patient statistics,
  - and 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.

1. The phenomenal rate of biological data being produced provides challenges:
   - Massive amounts of data have to be stored, analyzed, and made accessible.

2. 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 organization of their expression in the cell encoded by the DNA.

3. 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

1. Learn Biology.
2. Decide and pick a problem that interests you for experiment.
3. Find and learn about the Bioinformatics tools.
4. Learn the Computer Programming Languages.
   - Perl, Python, R, Java, etc.
5. 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 interrelationships 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?

• SmartMoney ranks Bioinformatics as #1 among next HotJobs
• Business Week 50 Masters of Innovation
• Jobs available, exciting research potential
• Important information waiting to be decoded!

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

Fighting Human Disease

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

Drug Development Life Cycle

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
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.

"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 1.

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.

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.

- Custom drug synthesis
  - 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.

- 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.

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.

**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.

**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.

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

<table>
<thead>
<tr>
<th>Topic</th>
<th>Web-based or GUI software</th>
<th>Command-line software</th>
</tr>
</thead>
<tbody>
<tr>
<td>Access to information</td>
<td>BioMart</td>
<td>EDGext</td>
</tr>
<tr>
<td>Protein similarity</td>
<td>Genome Workbench</td>
<td></td>
</tr>
<tr>
<td>Homology</td>
<td>BLAST</td>
<td>BLAST+</td>
</tr>
<tr>
<td>Database searching</td>
<td>DICTA-BLAST</td>
<td>HAMMER</td>
</tr>
<tr>
<td>Multiple alignment</td>
<td>MEGA</td>
<td>MAFFT</td>
</tr>
<tr>
<td>Bioinformatics</td>
<td>Galaxy</td>
<td>Genome EMBOSS (source EMBOSS)</td>
</tr>
<tr>
<td>Next-generation sequencing</td>
<td>Galaxy, SFF, PolyPhen2</td>
<td>SAMtools, tabi, IChoox</td>
</tr>
<tr>
<td>Rfam</td>
<td>RNAmem, RNAmem</td>
<td></td>
</tr>
</tbody>
</table>

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

- Phalanx
- Galaxy
- pf (Perl package), PSM
- Prosite
- DNAly
- pepid (EMBOSS)
- Protein structure
- OME, Pimp
- pepid (EMBOSS)
- Functional genomics
- Funk, Cytopatica
- Tree of life
- Venn (assembly)
- Viro
- MObiter (alignment)
- Bacteria and archaea
- NuMoR
- OMDRED (gene-finding)
- Voom
- VIGOR
- Esvor, Pandanal
- Zebrafish genomics
- FUNK
- Human genome
- OWL, Bullret
- EDGext, MltiSeq

Tool makers and tool users across informatics disciplines

- 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.
- In the realm of bioinformatics this means the following.

- 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 ArrayExpress Archive (AER) at EBI.
  - 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.

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

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 ArrayExpress Archive (AER) at EBI.
  - 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.

https://doi.org/10.1371/journal.pcbi.1000424
Where Can I Learn More?

- ISCB: http://www.iscb.org/
- 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/