Introduction to Bioinformatics

Dr. Taysir Hassan Abdel Hamid
Lecturer, Information Systems Department
Faculty of Computer and Information
Assiut University
taysirhs@aun.edu.eg
taysir_soliman@hotmail.com

Agenda

• Definition of Bioinformatics
• The need for Bioinformatics
• Distinction between important terminologies
• Sequence Alignment
• Protein Structures
• Useful Books
Bioinformatics… A Definition

- **Bioinformatics** is the:
  - recording,
  - annotation,
  - storage,
  - analysis, and
  - searching/retrieval of

- Nucleic acid sequence (genes and RNAs), protein sequence and structural information.

Bioinformatics … (Cont…)

- Roughly, **Bioinformatics** describes use of computers to handle biological information.
- A tight definition (Fredj Tekaia):

  "The mathematical, statistical and computing methods that aim to solve biological problems using DNA and amino acid sequences and related information."

Bioinformatics Goal

• The ultimate goal of the field is to enable the discovery of new biological insights as well as to create a global perspective from which unifying principles in biology can be discerned.

• This includes databases of:
  – Literature
  – sequences and
  – structural information

as well methods to access, search, visualize and retrieve the information.

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!

http://smartmoney.com/consumer/index.cfm?story=working-june02
What skills are needed?

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

• Working knowledge and appreciation in the others!

So…

• Our genome encodes an enormous amount of information about our beings
  – our looks
  – our size
  – how our bodies work
  – …. 
  – our health
  – our behaviors
  – … who we are!
Human Genome Initiative

• The Human Genome Project main Goal is Sequencing the Human Genome
• The Human Genome Project arose from two key insights (1980):
  - Accelerate biomedical research through a global view of genomes
  - Require a communal effort in infrastructure building

Progress of the Human Genome Project
Samples of DNA Sequences

- GTGTCGGTGGAACTTTGGCAGCA
  GTGCGTGGATTTCTCCGCGATG
  AGCTGCCGTCCCAACAATTCAAC
  ACTGGAT

- AACGTCCAGGTCGAAGGTGCGCA
  GAAGCACACACCAGCTATCTCAACC
  GTACCTTCACCTTCGAGAACTTCTG
  TCGAGG

A --> Adenine  C --> Cytosine
G --> Guanine  T --> Thymine
As technologies improve...

- We are able to extract more and more information encoded in a genome

potential synergies and competition between medical informatics (MI) and bioinformatics (BI) J Am Med Inform Assoc. 2003 November; 10 (6): 515–522
Bioinformatics versus Computational Biology

- Finding the genes in the DNA sequences of various organisms
- Developing methods to predict the structure and/or function of newly discovered proteins and structural RNA sequences.
- Clustering protein sequences into families of related sequences and the development of protein models.
- Aligning similar proteins and generating phylogenetic trees to examine evolutionary relationships.

Bioinformatics versus Health Informatics

**Problem:** Health care generates mountains of unstructured data.

**Solution:** Storing data in Computer-Based Patient Records that could be stored in data warehouses, shared, and mined.
System Structure

Patient data

Patient History

MRI

X-rays

Cat Scan

Feature Extraction

Association, Classification and Clustering

Recommendations, Predictions, Similar cases, etc.

Mining Visual Data

(a) X-ray chest scan  (b) HRCT chest scan  (c) MRI brain scan

The heritage sector, including art galleries, museums and libraries; Newspapers and other media organizations maintaining image archives
System interface that uses content-based retrieval for aid of diagnosis of chest diseases

Biomedical Informatics

Biomedical mining can enhance existing techniques of:
- Predicting various kinds of diseases
- Providing early treatment for diseases
Examples of Diseases

2. Corona Virus, that causes Severe Acute Respiratory Syndrome (SARS).
3. HCV (Hepatitis B Virus) – Infectious Virus in Liver.
4. HCV (Hepatitis C Virus) – Causing Cancer in liver.
5. HCV (Hepatitis C Virus) – Type 1.
6. HCV (Hepatitis C Virus) – Type 2.
7. HCV (Hepatitis C Virus) – Type 3a.
8. HCV (Hepatitis C Virus) – Type 3b.
9. HCV (Hepatitis C Virus) – Type 4.
10. Mental Illness
11. Hypertension
12. Heart Disease
13. Colon Cancer
15. Alzheimer
Agricultural Bioinformatics

Find additional resistance genes for
• Different plants (tomato, potato, rice, and wheat),
• Understand these biochemical processes that lead to resistance

• In the future we may learn how to modify them to make these genes more strong and avoid the toxic effects of singlet oxygen.

How about other species?
Biology is a data-rich science
Bioinformatics works at:

• DNA level:
  - DNA sequence alignment; gene prediction; gene evolution;...

• RNA level:
  - Study of gene expression; transcription mechanism; post-transcription modification;...

Bioinformatics works at:

• Protein level:
  - protein 2D and 3D structure prediction;
  - protein active site prediction;
  - protein-protein interactions;
  - protein-DNA interactions;...
Bioinformatics works at:

- **Genome**
  - (gene-to-gene interactions)

- **Proteome**
  - (protein-protein interactions)

- **System level**
  - (pathways, networks)

The revolution of ‘omics’ World

- Proteomics
- Functional Genomics
- Structural Genomics
• **Proteomics** is the subdivision of genomics concerned with analyzing the complete protein complement,

• It includes studying the proteome of organisms, both within and between different organisms.

---

**Functional genomics**

• Studies biological functions of proteins, complexes, pathways based on the analysis of genome sequences. It includes functional assignments for protein sequences.
Structural Bioinformatics

- “Structural bioinformatics is a subset of bioinformatics concerned with the use of biological structures: proteins, DNA, RNA, ligands etc. and different complexes to extend our understanding of biological systems.”

- http://biology.sdsc.edu/strucb.html

Protein Structures

- **Primary Structure**: Linear Amino Acid sequence of a protein.

- **Secondary Structure**: Regular structures includes:
  - α-Helices
  - β-Sheets
  - Coils
Three-dimensional protein structure = atomic coordinates in 3D space

Measured in Angstrom:
Conversion into metric measurement:
Unit
Angstrom $\times 10^{-8}$ cm
$\times 0.1$ nm

And Quaternary Structure

---

Structural Bioinformatics

- [http://biology.sdsc.edu/strucb.html](http://biology.sdsc.edu/strucb.html)
Role of Bioinformatics/Computational Biology in Proteomics Research

Sequence → Function

Sequence Alignment

- **Sequence comparison** is one of the most fundamental problems of computational biology, which is usually solved with a technique known as **sequence alignment**.

- **Sequence comparison** can be defined as the problem of finding which parts of the sequences are similar and which parts are different.

- Sequence alignment leads to identify similar functionality, structural similarity and finding important regions in a genome.

- Then, given an appropriate **scoring scheme**, their similarity can be computed.
Alignments Considered

- **Global Alignment:**
  align full length of both sequences

- **Local Alignment:**
  find best partial alignment of two sequences

- **Pair wise Alignment:**
  consists of two aligned sequences.

- **Multiple Alignment:**
  consists of three or more aligned sequences.
Scoring schemes

• Once the alignment is produced, a score can be assigned to each pair of aligned letters, according to a chosen scoring scheme.

• The similarity of two sequences can be defined as the best score among all possible alignments between them.

• Scoring schemes:

  ➢ Fixed scores were given for matches, mismatches and gap penalties (for DNA and protein sequence alignment).
  ➢ Alphabet-Weight scoring schemes, and is usually implemented by a substitution matrix (for protein sequence alignment).

• Sequence edits: **Fixed Score**

  - Mutations: \[ A G G A C T C \]
  - Insertions: \[ A G G G C C T \]
  - Deletions: \[ A G G . C T C \]

• Scoring Function:

  Match: \(+m\)  
  Mismatch: \(-s\)  
  Gap: \(-d\)

  \[ \text{Score } (F) = (\# \text{ matches}) \times m - (\# \text{ mismatches}) \times s - (\# \text{ gaps}) \times d \]
Example 1:

sequences $A=ACAAGACAGCGT$ and $B=AGAACAAGGCGT$.

$A = ACAAGACAG-CGT$

$B = AGAACA-AGGCGT$

An insertion of a character from the second sequence into the first one

A deletion of a character of the first sequence

using a scoring scheme that gives:

+1 value to matches

−1 to mismatches and gaps

alignment scores $= 9 * (1) + 2 * (-1) + 2 * (-1) = 5.$

BLAST… As an example of sequence alignment approach

• Basic Local Alignment Search Tool.
• Can be resumed into several steps:

  ✔ A list containing every three-letter word of the initial sequence is constructed.

  • Each word in the previously constructed list is then compared to every sub-word of length 3 of the database sequences using the BLOSUM62 substitution matrix.

  • Alignments having a score higher or equal to a threshold $T$ (usually 13), called hits, are conserved.

  • These hits are then placed into a very efficient search tree that will be used in the third step.
The set of hits found in the previous step (corresponding to the positions \( i = 1,2,3 \ldots \) of the initial sequence) are aligned against every database sequence.

- The three-letter alignments are then extended in both directions in order to obtain higher scoring alignments.
- This procedure ends when the elongation no longer improves the alignment score.

And a huge number of algorithms for handling different operations in Bioinformatics
Eukaryotic Genomes

- **Yeast**: //genome-www.stanford.edu/Saccharomyces
- **Fly**: //flybase.bio.indiana.edu:7081
- **Worm**: //www.sanger.ac.uk/Projects/C_elegans
- **Mouse**: //www.ensembl.org/Mus_musculus
- **Puffer Fish**: //www.ensembl.org/Fugu_rubripes
- **Mosquito**: //www.ensembl.org/Anopheles_gambiae

Useful Textbooks

- [Bioinformatics: Sequence and Genome Analysis](https://www.amazon.com/Bioinformatics-Sequence-Genome-Analysis-David-Mount/dp/0199734902)
- [Biological sequence analysis: Probabilistic models of proteins and nucleic acids](https://www.amazon.com/Biological-sequence-analysis-Probabilistic-nucleic/dp/0521585292)
Other Useful Textbooks

Other reference books
Thank You