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."

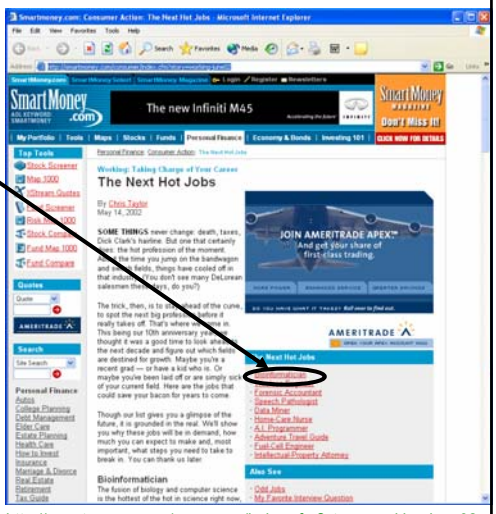
Bioinformatics Goal

- 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!



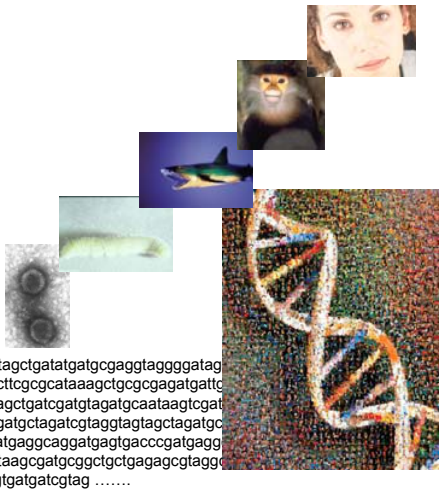
http://smartmoney.com/consumer/index.cfm?story=working-june02

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!



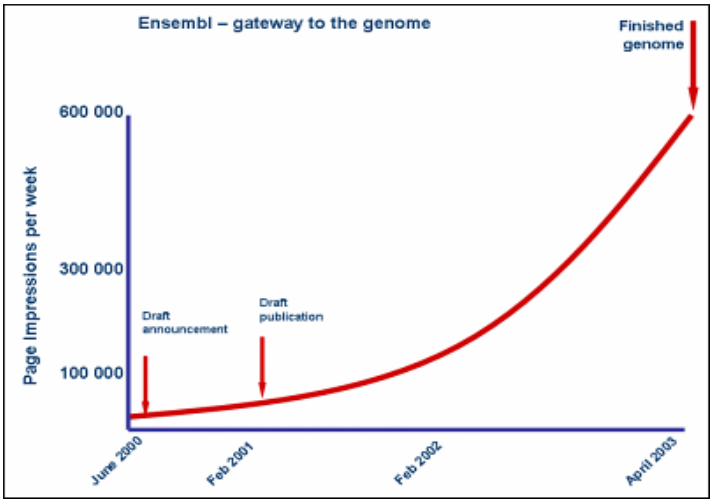
gcgtacgtacgtagagtgctagtctagtcgtagcgccgtagtcgatcgtgtgggtagtagctgatatgatgcgaggtaggggatag
gatgagcggatgctgagtgcagtggcatgcgatgtcgatgatagcggtaggtagacttcgcgcataaagctgcgcgagatgattg
agatgagctgatgctagaggtcagtgactgatgatcgatgcatgcatggatgatgcagctgatcgatgtagatgcaataagtcgat
atgctagatgatagctagatgtgatcgatggtaggtaggatggtaggtaaattgatagatgctagatcgtaggtagtagctagatgc
acacggaggcgagtgatcggtaccgggctgaggtgttagctaatgatgagtacgtatgaggcaggatgagtgacccgatgagg
ggatggatcgatgatcgatgcatggtgatgcgatgctagatgatgtgtgtcagtaagtaagcgatgcggctgctgagagcgtaggg
gagatgtaggaggaaggtttgatggtagttgtagatgattgtgtagttgtagctgatagtgatgatcgtag …….

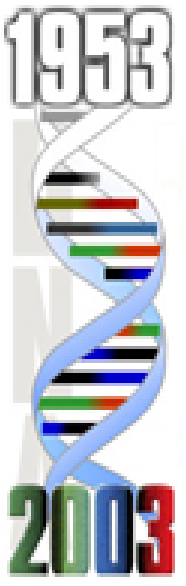
Human Genome Initiative

- The Human Genome Project main Goal is Sequencing the Human Genome
- The Human Genome Project arose from two key insights (1980):

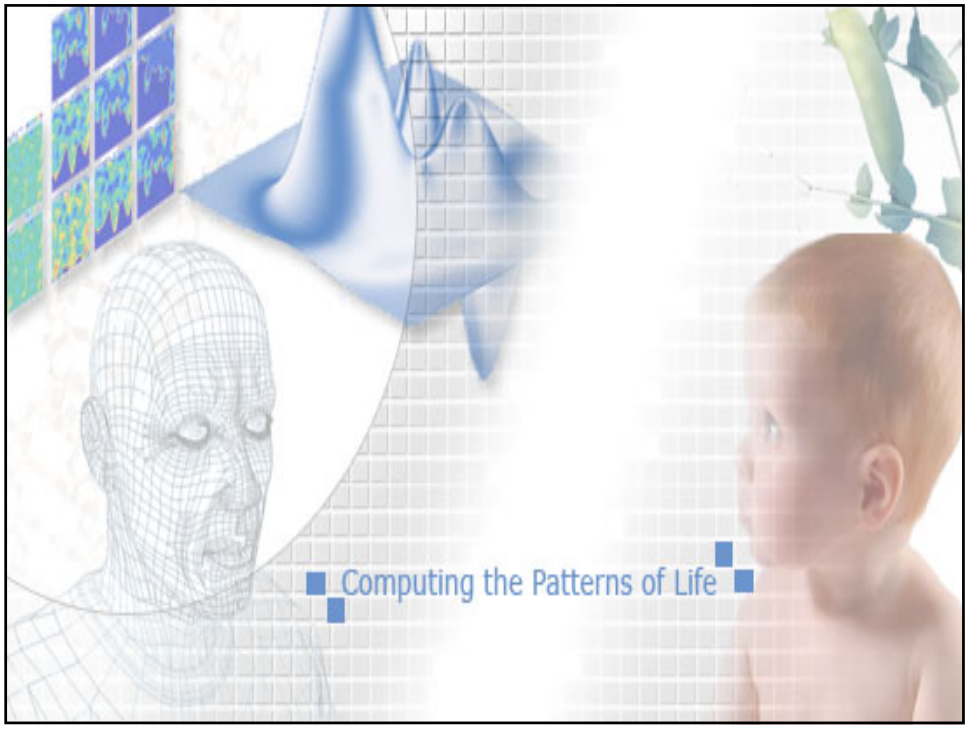
Accelerate biomedical research Through a global view of genomes

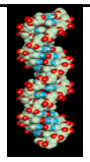
Require a communal effort in Infrastructure building





Progress of the Human Genome Project





Samples of DNA Sequences

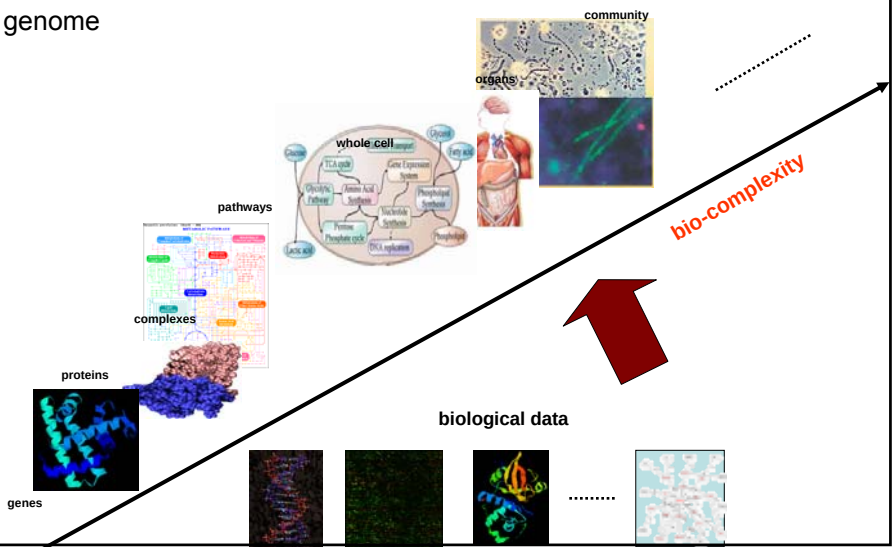
- GTGTCCGTGGAACTTTGGCAGCA GTGCGTGGATCTTCTCCGCGATG AGCTGCCGTCCCAACAATTCAAC ACTGGAT
- AACGTCCAGGTCGAAGGTGCGCT GAAGCACACCAGCTATCTCAACC GTACCTTCACCTTCGAGAACTTCG TCGAGG

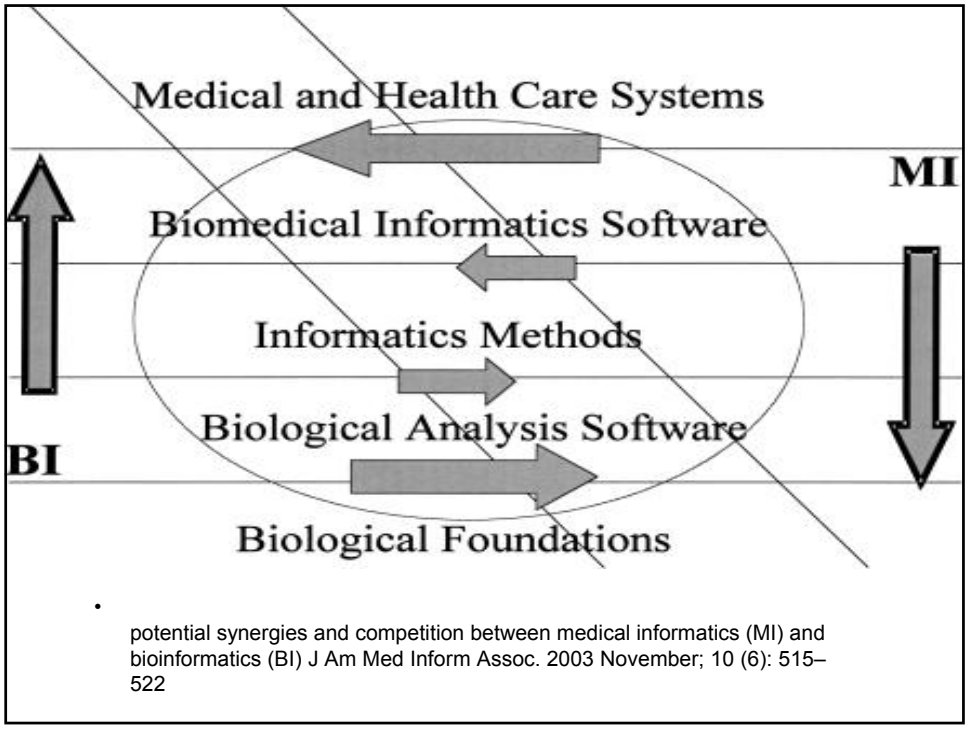
A --> Adenine C --> Cytosine

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

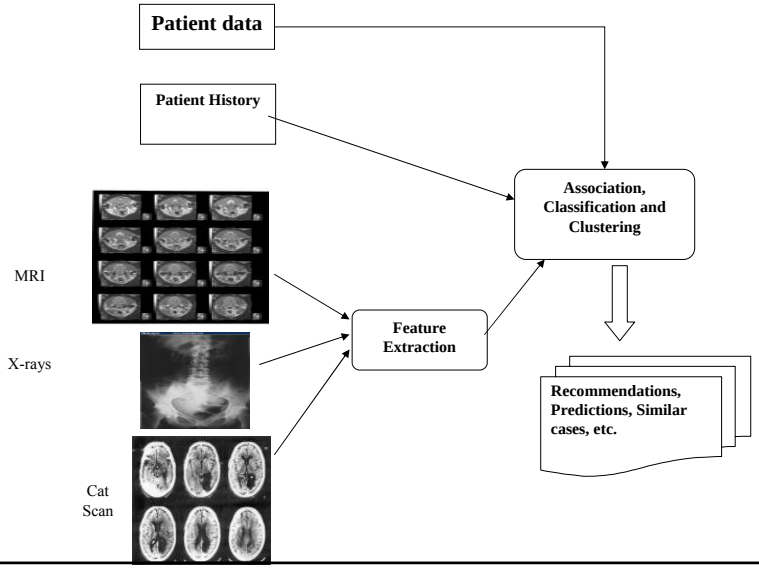
- Finding the genes in the DNA sequences of various organisms
- Developing methods to predict the structure and/or function of newly discovered proteins and structural RNA sequences.
- Clustering protein sequences into families of related sequences and the development of protein models.
- Aligning similar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

Bioinformatics versus Health Informatics

Problem: Health care generates mountains of unstructured data.

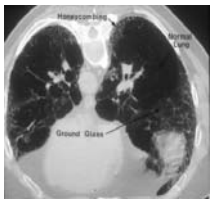
Solution: Storing data in Computer-Based Patient Records that could be stored in data warehouses, shared, and mined.

System Structure

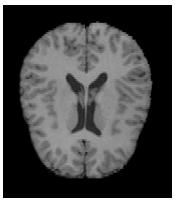


Mining Visual Data

(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

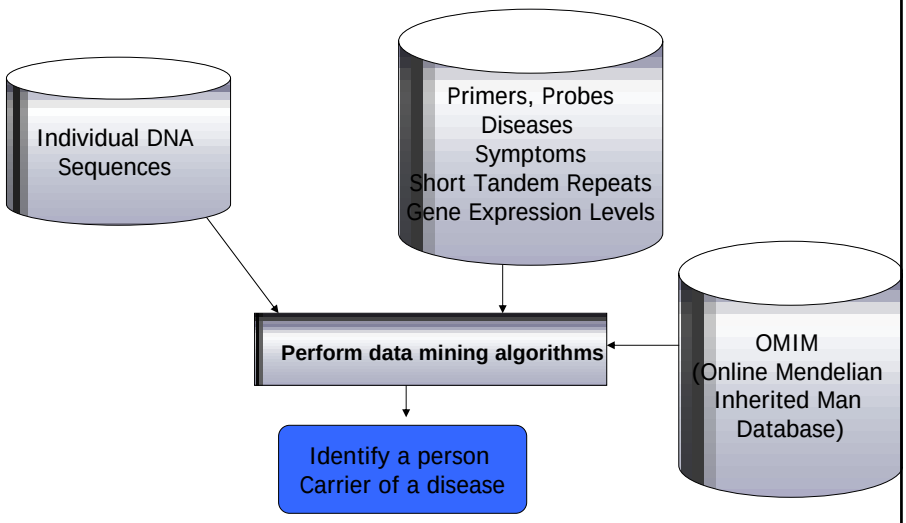
System interface that uses content-based retrieval for aid of diagnosis of chest diseases

Biomedical Informatics

Biomedical mining can enhance existing techniques of:

- Predicting various kinds of diseases
- Providing early treatment for diseases

Biomedical Product:



The same can be done for animals

Examples of Diseases

1. Breast Cancer.
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
14. Leukemia – Human Blood Cancer.
15. Alzheimer

Agricultural Bioinformatics

Potato infected with *Phytophthora infestans*

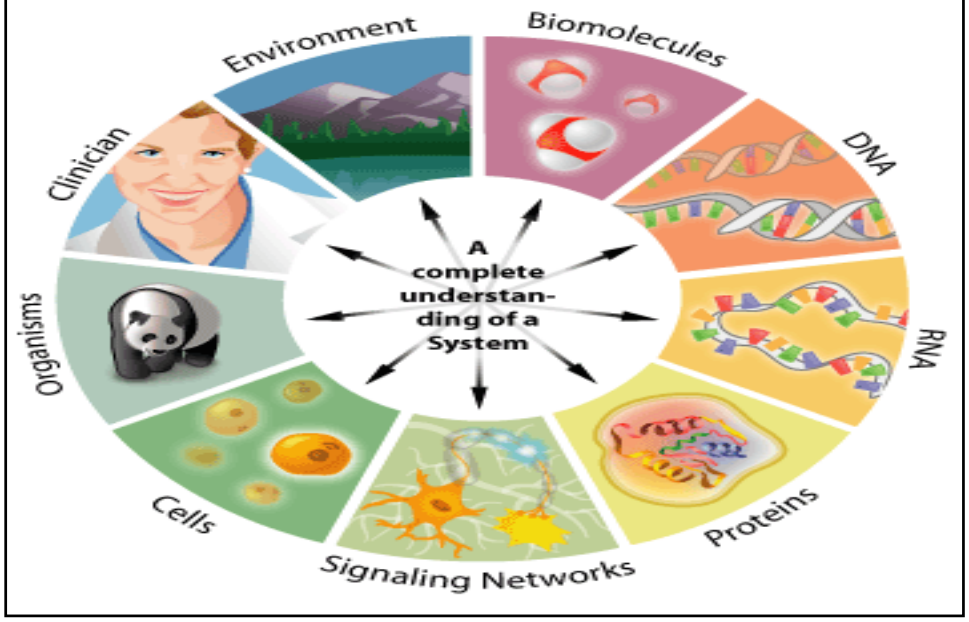
Find additional resistance genes for

- 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 ?

Bioinformatics



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

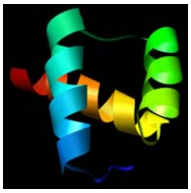
- "Structural bioinformatics is a subset of bioinformatics concerned with the use of biological structures: proteins, DNA, RNA, ligands etc. and different complexes to extend our understanding of biological systems."

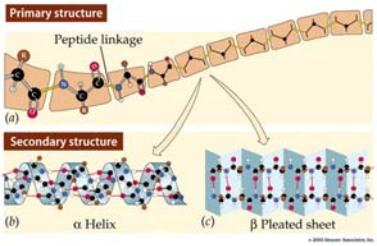
- *http://biology.sdsc.edu/strucb.html*

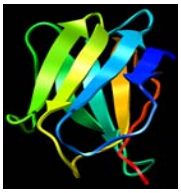
Protein Structures

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

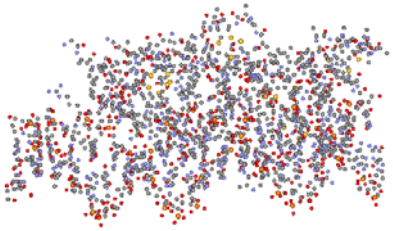
- **Secondary Structure:** Regular structures includes:
  - α-Helices
  - B-Sheets
  - Coils

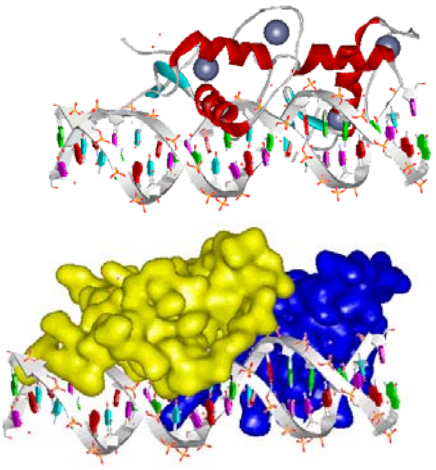






Three-dimensional protein structure = atomic coordinates in 3D space





Measured in Angstrom:

Conversion into metric measurement:

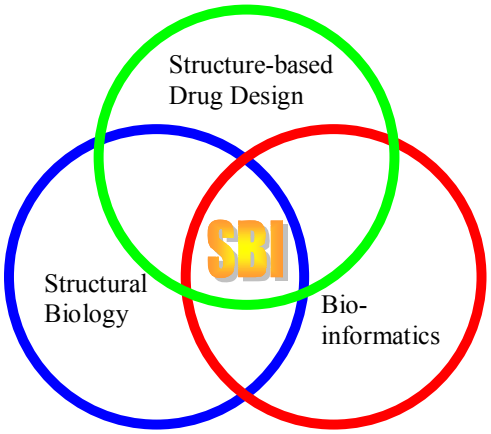
Unit

Angstrom x 10^{-8} cm

x 0.1 nm

And Quaternary Structure

Structural Bioinformatics



- *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.

NCBI Sample GenBank Record

PubMed Entrez BLAST OMIM Taxonomy Structure

GenBank Flat File Format

Click on any link in this sample record to see a detailed description of that data element or field. All of the descriptions are included on this page, so it can be printed as a single document. You can also return to the Alphabetical Quicklinks Table or Resource Guide

```
LOCUS       SCU49845     5028 bp    DNA             PLN       21-JUN-1999
DEFINITION  Saccharomyces cerevisiae TCP1-beta gene, partial cds, and Axl2p
            (AXL2) and Rev7p (REV7) genes, complete cds.
ACCESSION   U49845
VERSION     U49845.1  GI:1293613
KEYWORDS    .
SOURCE      Saccharomyces cerevisiae (baker's yeast)
  ORGANISM  Saccharomyces cerevisiae
            Eukaryota; Fungi; Ascomycota; Saccharomycotina; Saccharomycetes;
            Saccharomycetales; Saccharomycetaceae; Saccharomyces.
REFERENCE   1  (bases 1 to 5028)
            .
            .
            .
ORIGIN
        1 gatcctccat atacaacggt atctccacct caggtttaga tctcaacaac ggaaccattg
       61 ccgacatgag acagttaggt atcgtcgaga gttacaagct aaaacgagca gtagtcagct
      121 ctgcatctga agccgctgaa gttctactaa gggtggataa catcatccgt gcaagaccaa
      181 gaaccgccaa tagacaacat atgtaacata tttaggatat acctccgaaaa tcaataaaccg
      241 ccacactgtc attattataa ttagaaacag aacgcaaaaa ttatccacta tataattcaa
      301 agacgcgaaa aaaaaagaac aacgcgtcat agaacttttg gcaattcgcg tcacaaataa
      361 attttggcaa cttatgtttc ctcttcgagc agtactcgag ccctgtctca agaatgtaat
      421 aatacccatc gtaggtatgg ttaaagatag catctccaca acctcaaagc tccttgccga
      481 gagtcgccct cctttgtcga gtaattttca cttttcatat gagaacttat tttcttattc
      541 tttactctca catcctgtag tgattgacac tgcaacagcc accatcacta gaagaacaga
      601 acaattactt aatagaaaaa ttatatcttc ctcgaaacga tttcctgctt ccaacatcta
      661 cgtatatcaa gaagcattca cttaccatga cacagcttca gatttcatta ttgctgacag
      721 ctactatatc actactccat ctagtagtgg ccacgcccta tgaggcatat cctatcggaa
      781 aacaataccc cccagtggca agagtcaatg aatcgtttac atttcaaatt tccaatgata
      841 cctataaatc gtctgtagac aagacagctc aaataacata caattgcttc gacttaccga
      901 gctggctttc gtttgactct agttctagaa cgttctcagg tgaacctttct tctgacttac
      961 tatctgatgc gaacaccacg ttgtatttca atgtaatact cgagggtacg gactctgccg
     1021 acagcacgtc tttgaacaat acataccaat ttgtttgttac aaaccgtcca tccatctcgc
     1081 tatcgtcaga tttcaatcta ttggcgttgt taaaaaacta tggttatact aacggcaaaa
     1141 acgctctgaa actagatcct aatgaagtct tcaacgtgac ttttgaccgt tcaatgttca
     1201 ctaacgaaga atccattgtg tcgtattacg gacgttctca gttgtataat gcgccgttac
     1261 ccaattggct gttcttcgat tctggcgagt tgaagtttac tgggacggca ccggtgataa
     1321 actcggcgat tgctccagaa acaagctaca gttttgtcat catcgctaca gacattgaag
```

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).

Fixed Score

- Sequence edits:

  A G G C C T C

  – 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

  Score (F) = (# matches) × m - (# mismatches) × s – (#gaps) × d

- Example1:

sequences A=ACAAGACAGCGT And B=AGAACAAGGCGT.

using a scoring scheme that gives :

+1 value to matches
−1 to mismatches and gaps

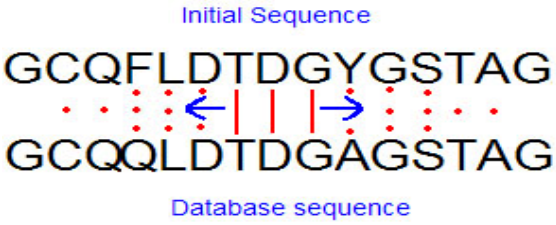
alignment scores = 9 * (1) + 2 * (−1) + 2 *(−1) = 5.

BLAST… As an example of sequence alignment approach

- **B**asic **L**ocal **A**lignment **S**earch **T**ool.
- 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… 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.

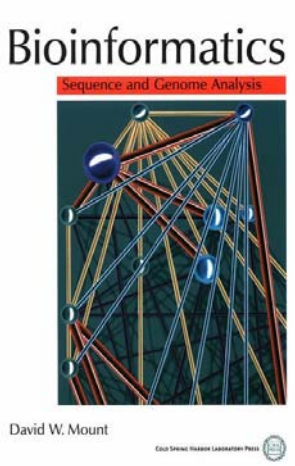


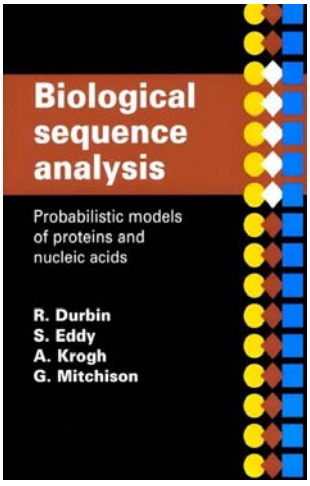
And a huge number of algorithms for handling different operations in Bioinformatics

Eukaryotic Genomes

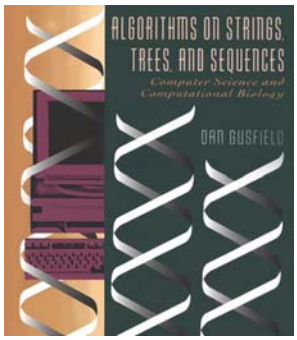
- Yeast : //genome-www.stanford.edu/Saccharomyces
- Fly : //flybase.bio.indiana.edu:7081
- Worm : //www.sanger.ac.uk/Projects/C_elegans
- Mouse : //www.ensembl.org/Mus_musculus
- Puffer Fish : //www.ensembl.org/Fugu_rubripes
- Mosquito : //www.ensembl.org/Anopheles_gambiae

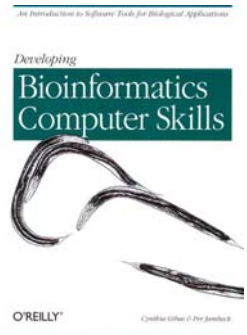
Useful Textbooks

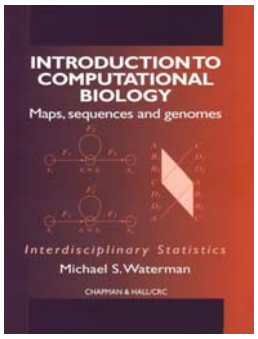




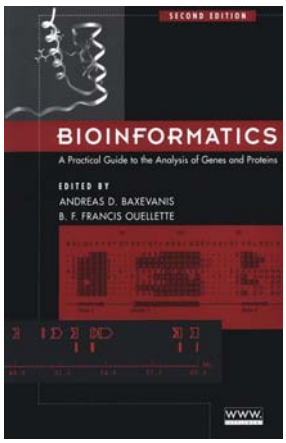
Other Useful Textbooks

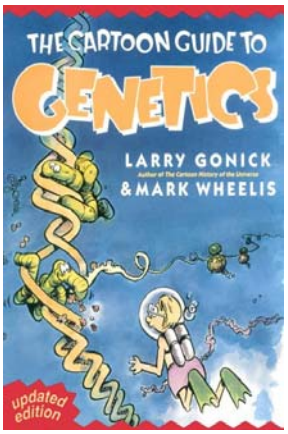


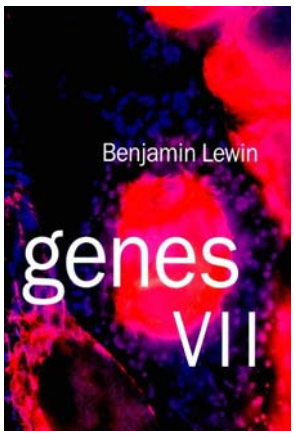




Other reference books







Thank You