# BIOC4004 - Industrial Biochemistry
# Lecture 12 - Wed Feb 11, 04

**Topics for the Day:**

- Bioinformatics
- Computational Biology
- Development of Bioinformatics
- Sequence Alignments

# What is Bioinformatics?

- The convergence of biology, computer science, and information technology
- Computational management of all kinds of biological information
- The search for and use of patterns to study the inherent structure in biological data
- The development of new methods for database access and queries

ie. use the tools from each of these separate disciplines to discover new biological concepts from complex biological data

# Computational Biology vs. Bioinformatics

- Is there a difference ? Yes….sort of…

- **Bioinformatics :** the development of informatics **tools** to handle, analyze and visualize complex biological information
    - NOT RESTRICTED TO MOLECULAR DATA !!!!
        - Epidemiological data
        - Environmental data
        - etc…etc...

- **Computational Biology**: the use of Bioinformatics tools to perform biological **studies**(analysis and interpretation)

**Is there a difference? yes, but there is PLENTY of overlap !!!!**

- Bioinformatics: the techie side of the equation
- Computational Biology: the science side of the equation

# Three Main Sub-Disciplines of Bioinformatics

1. The development of new algorithms and statistical methods
   assessing relationships among members of large data sets.
   ⇒ *statistics, mathematics, …*

2. The development and implementation of tools for storage, access and management of information
   software engineering, databases
   ⇒ *database management*, *information technology*

3. The visualization, analysis, and interpretation of various types of data
   analysis of nucleotide and amino acid sequences, protein domains, and protein structures, genome comparisons (or any other type of biological data)
   ⇒ *computational biology, data mining*

## Bioinformatics as Information Technology

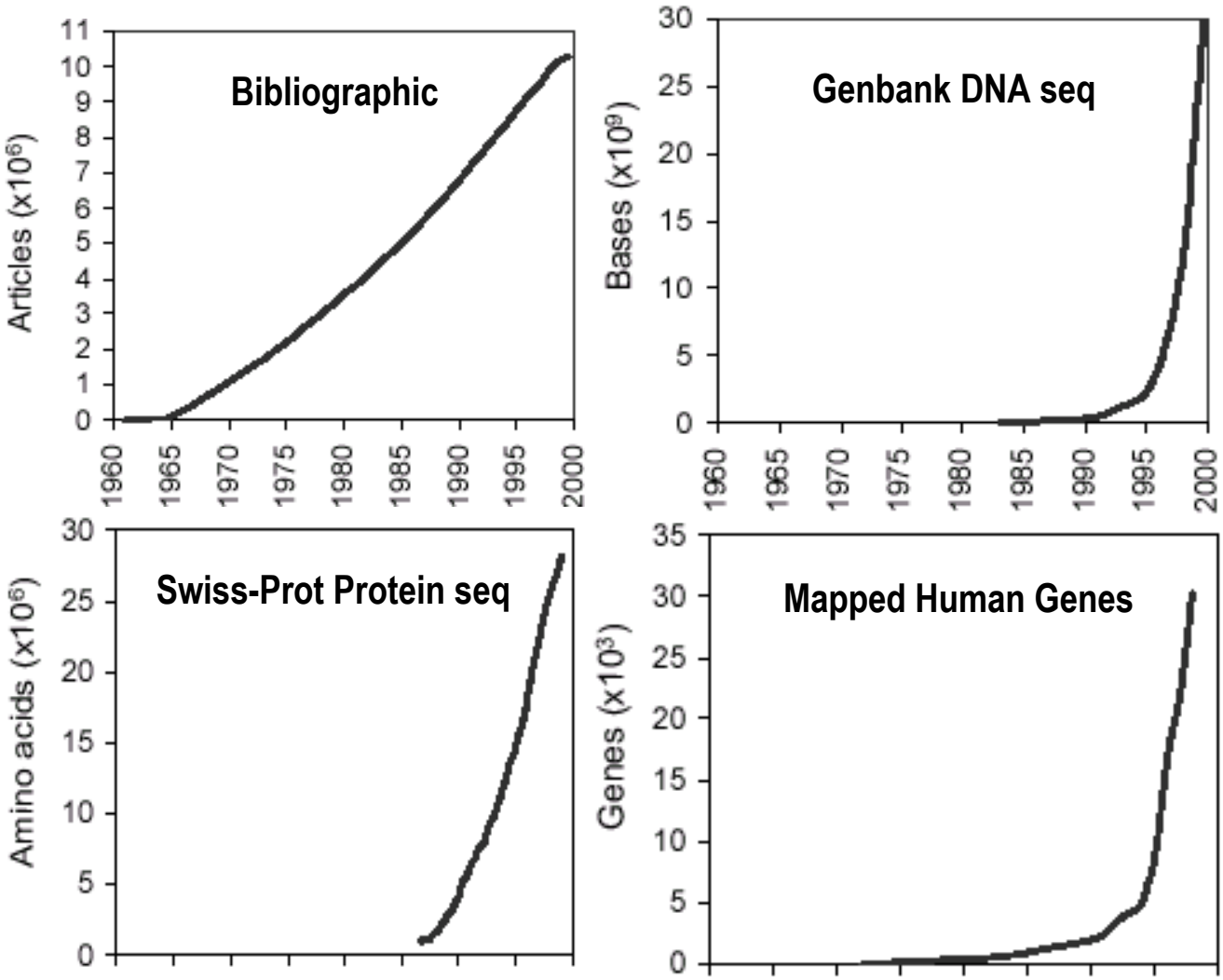
- Database: GenBank, SWISS-PROT, etc..
- Information retrieval: biomedical text analysis
- Algorithm: sequence alignment
- Pattern Recognition: 2D/3D image analysis
- Data mining: clustering, classification, association
- Hardware: supercomputing

# Why use bioinformatics?

- An explosive growth in the amount of biological information which necessitates
  - use of computers for information storage, cataloging, retrieval
  - tools for visualizing and analyzing the data

- A more global perspective in experimental design.
  - one scientist-one gene paradigm vs. consideration of whole organisms

- Data-mining - the process by which testable hypotheses are generated regarding the function or structure of a gene or protein of interest by identifying similar sequences in better characterized organisms.




**Molecular data being accumulated at an astounding rate !!!!**

# The Development of Bioinformatics

- With the development of programmable computers:
  - use computers to handle and store large amounts of data
    - statistics
    - mathematical transformations
    - "data crunching"
- Among the first to embrace computers :
  - X-ray crystallographers
  - Biochemists and molecular biologists

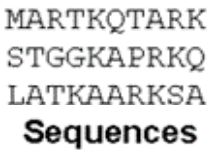
## Some milestones

- 1962 Pauling's theory of molecular evolution
- 1965 Margaret Dayhoff's Atlas of Protein Sequences
- 1970 Needleman-Wunsch algorithm
- 1977 DNA sequencing and software to analyze it (Staden)
- 1981 Smith-Waterman algorithm developed
- 1981 The concept of a sequence motif (Doolittle)
- 1982 GenBank Release 3 made public
- 1982 Phage lambda genome sequenced
- 1983 Sequence database searching algorithm (Wilbur-Lipman)
- 1985 FASTA : fast sequence similarity searching
- 1988 National Center for Biotechnology Information (NCBI) created at NIH/NLM
- 1990 BLAST : fast sequence similarity searching
- 1995 First bacterial genome completely sequenced (*H. influenzae* ~ 1.83 Mb)
- 1996 First eukaryotic (yeast) genome completely sequenced (13 Mb)
- 1997 PSI-BLAST
- 1998 First multicellular (C. elegans) genome completely sequenced (97 Mb)
- 1999 Fly genome completely sequenced (137 Mb)
- 2001 First plant (Arabidopsis ) genome completely sequenced (125 Mb)
- 2000 First Draft of Human genome [Public consortium & CELERA] (3000 Mb)

# Thank you Margaret Dayhoff !!!!

- development of computer aids to protein sequence determination
- the development of recognition and display programs for use in X-ray crystallography
- the development of computer methods for the comparison of protein sequences
- derivation of methods for evaluating evolutionary relationships from alignments of protein sequences.
    - Concept of protein superfamilies
    - Early pioneer of molecular phylogenetics (evolutionary relationships between organisms)
    - Early proponent of endosymbiont origin of organelles
- compiled the first Atlas of Protein Sequence and Structure (65 sequences)
- showed feasibility of on-line computer database and retrieval system for molecular sequences (early 80s)
- envisioned the early blueprints of GenBank database
    - sequence retrieval
    - sequence prediction or identification based on sequence
    - browsing of sequence information

# Integrating Biological Information



Sequences

**Biological information (II)**

• DNA

coding sequence, exon/intron, promoter, enhancer, genome, ...

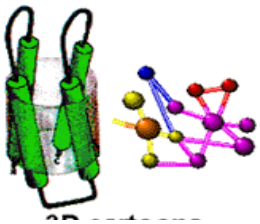
• RNA

coding sequence, poly(A) signal, Kozak, destabilizing signal, secondary structure, expression data (transcriptome), ...

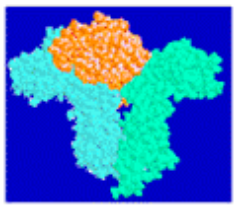
• Protein

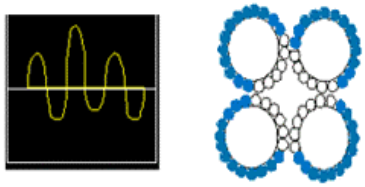
domain, motif, NLS, organellar targeting signal, modification sites (phophorylation, glycosylation, acetylation), secondary structure, fold, 3D structure, expression data (proteome), ...


Domain-cartoons (secondary structure cartoons)

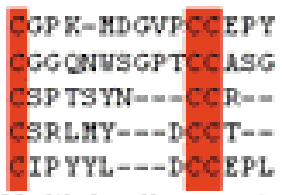


Extended sequences
(e.g. disulphide-topologies)


3D structures


Diagrams (hydrophobicity plots, helical circles)


Evolutionary trees



Multiple alignments

**Biological information (III)**

• Cell

protein-protein interaction, metabolic pathways, signal transduction, metabolome, ...

• Population/Evolution

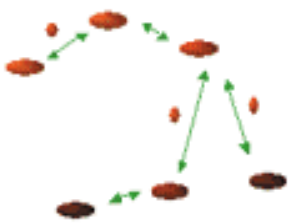
polymorphism (SNP, LP), mutation, gene frequency, ethnic groups, gene map, polymorphic marker, orthologs, paralogs, phylogeny, ...

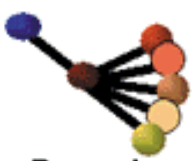
• Bibliography

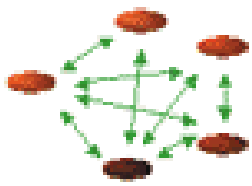
"gene A activates gene B", "protein X interacts with Y", ...


Genomes


Subunit structures, ligands


Metabolic pathways


Dynamic assemblies


Similarity groups or neighborhoods

# Sequence Alignments (Part I)

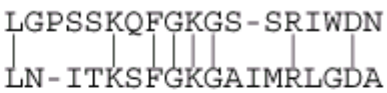
- A major concept in bioinformatics, allow the comparison of related sequences
- Provide a powerful way to compare sequences for either:
    - evolutionary relatedness
    - structural/functional relatedness.

- Pairwise alignment (2 sequences)
    - Similarity
    - Database search
    - Conserved sequences and patterns

- Multiple alignment (3 or more sequences)
    - Evolutionary relationship (orthologous/paralogous genes, organisms)
    - Functional domain/motif finding
    - Single-Nucleotide Polymorphism (SNP) search
    - DNA sequence assembly

- **Global** vs **Local** alignments

- global alignment : stretched over the entire sequence length to iclude as many matching amino acids as possible up to and including the sequence ends (orthologs, paralogs)




- local alignment : the alignment stops at the ends of regions of identity or strong similarity (patterns, domains, motifs)


**Depends on whether the sequences are presumed to be related over their entire lengths or to only share isolated regions of homology !**

- **Local is better: assume only small regions are fully alignable**

# Sequence Alignments (Part II)

• Sequence alignments maximize the **alignment scores** in calculating the best possible alignment between two sequences given some constraints:

- **scoring matrix**: weighted scheme for matches and mismatches
  - DNA: eg. +1.0 for **match**, -0.1 for **mismatch**

```
ACTGGATCTTGGCAATTTG
|||   ||   ||  |||||  |
ACTACAT--TGCCAATTAG
```

  - BLOSUM and PAM are two well known protein scoring matrices
    - necessary to account for **"similar"** (but not identical) residues

```
KLQKAKEILSNAESRARYD
 ||  |||      |  :||:|
QLQEAKE--CDPDKKARFD
```

• **gap penalties** are used to "penalize" against the introduction of gaps (indels: insertions / deletions)

  - if no penalty, can introduce a gazillion gaps to maximize the alignment

```
ATCGACGTTAGA        ATCGACGTTAGA
ACACTTGGACAG        A-C-AC-TT-G-...
```

  - since gaps are relatively rare events, evolutionarily speaking, need a penalty for opening and extending a gap
  - choose these parameters depending on the sequences being compared:

    large open and extend: closely related proteins

    large open and small extend: proteins where large loops between conserved domains are tolerated

# Sequence Alignments (Part III)

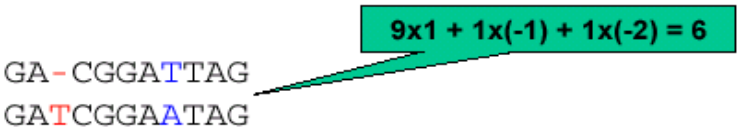
• the **alignment scores** in calculated by adding up the weighted score for matches, mismatches, and gap penalties (gap opening and gap extension)

  • in protein alignments, must also account for partial matches between "similar" residues (need a more sophisticated scoring matrix - see next page)

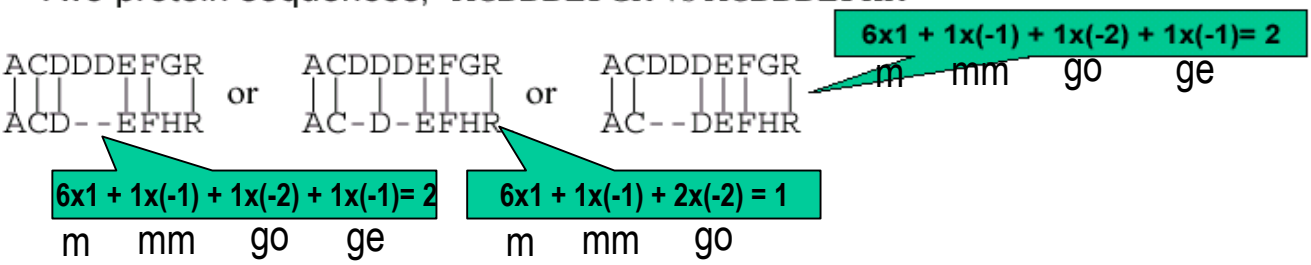
• the "best possible alignment" between two sequences is obtained by maximizing the alignment score (at least in theory)

• You have two DNA sequences, GACGGATTAG and GATCGGAATAG.

Are these two similar, and how much similar?




• Two protein sequences, ACDDDEFGR vs ACDDDEFHR




Match (m) : +1
Mismatch (mm) : -1
Gap or Indel : opening (go) -2, extension (ge) -1

• Note how the middle alignment suffers from the gap penalty associated with two separate gaps.
• In this case, we have two alternate alignments that have an identical score

# Aligning Proteins

• DNA is easier to align than protein: only 4 bases !!!

• in protein you have to consider similarity between residues and their relative abundance.

- tyrosine and tryptophan residues are uncommon and weighted heavier than alignments between common residues such as alanine and serine

- alignments between glutamate and aspartate are scored positively because of their similarity, whereas dissimilar residues are penalized

• if trying to align protein coding DNA, you should align the translated sequence…why ?

## About Protein Scoring Matrices… (part I)

- Scoring Matrices are used to calculate the "alignability" of two sequences

- Based on the propensity of an amino acid to mutate into another

**PAM Matrix (Dayhoff)**

• A matrix of Point Accepted Mutations

• 1972, analyzed Cytochrome C sequences (~1% divergence ie.1 mutation per 100 amino acids)

• aligned the sequences in the "best possible way"

• computed the frequency of "mutability" of one amino acid into another

  • ie. how likely that an Alanine will mutate into a Glycine

• generated a matrix for every possible amino acid pair

• the matrix is multiplied against itself N times to generate a PAM "N" matrix

• 1 PAM ~ 1 million years of divergence; use PAM matrix to suit the problem being studied (eg. Use PAM100 for sequences that diverged 100 MYA):

  • use low PAM matrices for highly similar proteins

  • use high PAM matrices for less similar proteins

# About Protein Scoring Matrices… (part II)

**BLOSUM (BLOcks Subtitution Matrix) (Hanikoff and Hanikoff, 1991)**

- Based on an analysis of conserved protein blocks (PROSITE)
- Much larger and more diverse set of proteins than that used in PAM matrices
    - ~ 2000 conserved protein BLOCKS
- computed the frequency of "mutability" of one amino acid into another
    - generated a matrix for every possible amino acid pair
- Whereas different PAM matrices are generated by multiplying the original PAM matrix to itself, BLOSUM matrices are computed from distinct sets of data:
    - BLOSUM 90 - use protein blocks with > 90% sequence identity
        - short alignement of highly similar sequences
    - BLOSUM 62 - use protein blocks with > 62% sequence identity
        - most general alignments (default BLOSUM matrix)
    - BLOSUM 30 - use protein blocks wih > 30% sequence identity
        - detection of weak local alignments

| | C | S | T | P | A | G | N | D | E | Q | H | R | K | M | I | L | V | F | Y | W | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| C | 9 | | | | | | | | | | | | | | | | | | | | C |
| S | −1 | 4 | | | | | | | | | | | | | | | | | | | S |
| T | −1 | 1 | 5 | | | | | | | | | | | | | | | | | | T |
| P | −3 | −1 | −1 | 7 | | | | | | | | | | | | | | | | | P |
| A | 0 | 1 | 0 | −1 | 4 | | | | | | | | | | | | | | | | A |
| G | −3 | 0 | −2 | −2 | 0 | 6 | | | | | | | | | | | | | | | G |
| N | −3 | 1 | 0 | −2 | −2 | 0 | 6 | | | | | | | | | | | | | | N |
| D | −3 | 0 | −1 | −1 | −2 | −1 | 1 | 6 | | | | | | | | | | | | | D |
| E | −4 | 0 | −1 | −1 | −1 | −2 | 0 | 2 | 5 | | | | | | | | | | | | E |
| Q | −3 | 0 | −1 | −1 | −1 | −2 | 0 | 0 | 2 | 5 | | | | | | | | | | | Q |
| H | −3 | −1 | −2 | −2 | −2 | −2 | 1 | −1 | 0 | 0 | 8 | | | | | | | | | | H |
| R | −3 | −1 | −1 | −2 | −1 | −2 | 0 | −2 | 0 | 1 | 0 | 5 | | | | | | | | | R |
| K | −3 | 0 | −1 | −1 | −1 | −2 | 0 | −1 | 1 | 1 | −1 | 2 | 5 | | | | | | | | K |
| M | −1 | −1 | −1 | −2 | −1 | −3 | −2 | −3 | −2 | 0 | −2 | −1 | −1 | 5 | | | | | | | M |
| I | −1 | −2 | −1 | −3 | −1 | −4 | −3 | −3 | −3 | −3 | −3 | −3 | −3 | 1 | 4 | | | | | | I |
| L | −1 | −2 | −1 | −3 | −1 | −4 | −3 | −4 | −3 | −2 | −3 | −2 | −2 | 2 | 2 | 4 | | | | | L |
| V | −1 | −2 | 0 | −2 | 0 | −3 | −3 | −3 | −2 | −2 | −3 | −3 | −2 | 1 | 3 | 1 | 4 | | | | V |
| F | −2 | −2 | −2 | −4 | −2 | −3 | −3 | −3 | −3 | −3 | −1 | −3 | −3 | 0 | 0 | 0 | −1 | 6 | | | F |
| Y | −2 | −2 | −2 | −3 | −2 | −3 | −2 | −3 | −2 | −1 | 2 | −2 | −2 | −1 | −1 | −1 | −1 | 3 | 7 | | Y |
| W | −2 | −3 | −2 | −4 | −3 | −2 | −4 | −4 | −3 | −2 | −2 | −3 | −3 | −1 | −3 | −2 | −3 | 1 | 2 | 11 | W |
| | C | S | T | P | A | G | N | D | E | Q | H | R | K | M | I | L | V | F | Y | W | |

**The BLOSUM 62 matrix is highly recommended for sequence alignment and database searching - based on "Structural Data", so better than PAM**

# Aligning Protein Sequences

## Pairwise Aligments

| | Alignment | | | |
|---|---|---|---|---|
| sequence A | Y | C | D | A |
| sequence B | F | M | E | G |
| BLOSUM62 matrix value | 3 | -1 | 2 | 0 |

Total score = 3-1+2+0 = 4

- pairs of sequences are aligned by maximizing the score

## Multiple Alignments (eg. CLUSTAL algorithm)

- **global alignment algorithm**
- **the closest pair of sequences aligned first**
- **next closest sequence is aligned to that alignment**
- **etc, etc, until all sequences have been added one by one**


**even the best program make mistakes (tweaking by eye !!!)**

# Searching of Molecular Databases

- Why ?

  To look for similarities between a sequence of interest and and the sequences in the database

  - could help elucidate the function of an unknown sequence
  - could help find conserved motifs shared by different sequences

- How ?

  Need to compare the sequence to every sequence in the DB

  - need to align the query sequence to every sequence in the DB
  - need to calculate some metric to assess the quality of the match
  - need to report the "good" matches

- Early sequence alignment programs were meant to perform the best possible alignment of a pair of, at most, a few (dozen) sequences
  - good results but computationally expensive
  - no way could you use these to screen a DB of thousands of sequences

- Design of "heuristic" methods to perform the searches:
  - an initial quick and dirty alignment between query and all the sequences
  - assessment of potentially matching sequences worth "revisiting"
  - a second, more accurate, alignment on the worthwhile sequences

- These heuristic methods are computationally economical:
  - fast and good
  - can handle lots of queries !!!
  - Provide a statistical assessment of the match (more on this later)

- The program used for searching the GenBank DB is BLAST (Basic Local Alignment Tool)

  [Some of the European DBs use a program called FASTA, however, GenBank cross-references entries in the European DBs anyway, so we'll mostly talk about GB]

# BLAST (Basic Local Alignment Tool)

- find it at http://www.ncbi.nlm.nih.gov/BLAST/
  also mirrored at the Canadian Bioinformatics Resource
  http://www.cbr.ca/blast/
- uses an approach based on:
  - matching short sequence fragments (HSPs: High Scoring Segment Pairs)
    - HSPs can be perfect matches or highly homologous
  - looks for clusters of HSPs that are close to each other
  - a "cut-off alignment score" is calculated to determine the minimum score that should be met by an alignment in order to be significant
  - the best local alignments between the query sequence and the database sequence(s) are calculated and reported

**BLAST VARIANTS**

| PROGRAM | QUERY | DB | COMMENTS |
|---|---|---|---|
| BLASTP | protein | protein | compares amino acid query against protein sequences |
| BLASTN | DNA | DNA | compares nucleotide query against DNA sequences |
| BLASTX | DNA | protein | compares 6X translations of nucleotide query against protein sequences |
| TBLASTN | protein | DNA | compares protein query against 6X translations of DNA sequences |
| TBLASTX | DNA | DNA | compares 6X translations of nucleotide query against 6X translations of DNA sequences |

- also :
  - PSI-BLAST: protein query and protein DB - different statistics used to detect weak similarities
  - PHI-BLAST: protein query and protein DB - used to look for protein patterns (small regions of conservation)
  - MEGA-BLAST: larges sets of long DNA sequences
  - CD Search - conserved domain detection
  - BLAST2: pairwise alignment of two sequences
  - Genome BLAST : alignment of DNA sequences to genome data
  - VecScreen: used to detect vector sequence within sequence data