Introduction to bioinformatics

Lecture 3

High-throughput Biological Data

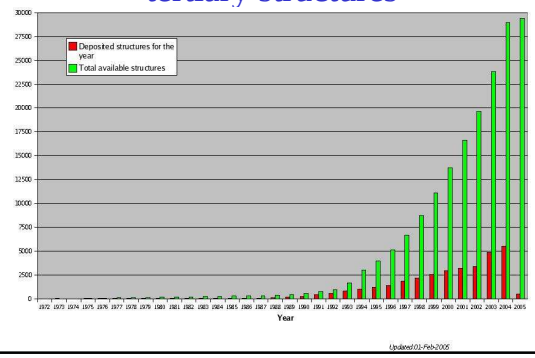
-data deluge, bioinformatics algorithms-

and evolution

Last lecture:

- Many different genomics datasets:
  - Genome sequencing: more than 500 species completely sequenced and data in public domain (i.e. information is freely available), virus genome can be sequenced in a day
  - Gene expression (microarray) data: many microarrays measured per day, new techniques for improved measurement (via sequencing) are being developed
  - Proteomics: Protein Data Bank (PDB) - as of Tuesday February 07, 2006 there are 35026 Structures. http://www.rcsb.org/pdb/
  - Protein-protein interaction data: many databases worldwide
  - Metabolic pathway, regulation and signaling data, many databases worldwide

Growth in number of protein tertiary structures



The data deluge

Although a lot of tertiary structural data is being produced (preceding slide), there is the

SEQUENCE-STRUCTURE-FUNCTION GAP

The gap between sequence data on the one hand, and structure or function data on the other, is widening rapidly: Sequence data grows much faster

High-throughput Biological Data
The data deluge

- Hidden in all these data classes is information that reflects
  - existence, organization, activity, functionality …… of biological machineries at different levels in living organisms

Utilising and analysing this information computationally and biologically effective is essential for Bioinformatics

Data issues: from data to distributed knowledge

- Data collection: getting the data
- Data representation: data standards, data normalisation …..
- Data organisation and storage: database issues …..
- Data analysis and data mining: discovering "knowledge", patterns/signals, from data, establishing associations among data patterns
- Data utilisation and application: from data patterns/signals to models for bio-machineries
- Data visualization: viewing complex data ……
- Data transmission: data collection, retrieval, …..
- ……

Databases in Bioinformatics

- *First practical*: **Sequence Retrieval System (SRS)**
- Developed by Thure Etzold, as an undergrad student in Cologne, then as a PhD student at the European Molecular Biology Laboratory (EMBL) in Heidelberg
- Many databases are connected and intergrated within SRS
- Bioinformatics analysis tools can be run from within SRS
- It has a special internal language to easily link in your own or otherwise new databases

Bio-Data Analysis and Data Mining

- Analysis and mining tools exist and are developed for:
  - DNA sequence assembly
  - Genetic map construction
  - Sequence comparison and database searching
  - Gene finding
  - Gene expression data analysis
  - Phylogenetic tree analysis, e.g. to infer horizontally-transferred genes
  - Mass spectrometry data analysis for protein complex characterization
  - ……

Bio-Data Analysis and Data Mining

- As the amount and types of data and their cross connections increase rapidly
- the number of analysis tools needed will go up "exponentially" if we do not reuse techniques
  - blast, blastp, blastx, blastn, … from BLAST family of tools (we will cover BLAST later)
  - gene finding tools for human, mouse, fly, rice, cyanobacteria, …..
  - tools for finding various signals in genomic sequences, protein-binding sites, splice junction sites, translation start sites, …..

Bio-Data Analysis and Data Mining

Many of these data analysis problems are fundamentally the same problem(s) and can be solved using the same set of tools

e.g.

•clustering or

•optimal segmentation by Dynamic Programming

We will cover both of these techniques in later lectures

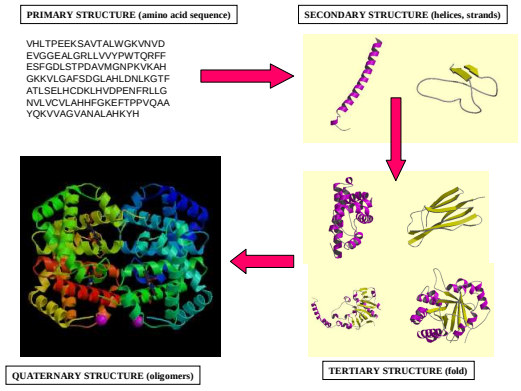
Bio-data Analysis, Data Mining and Integrative Bioinformatics

To have analysis capabilities covering a wide range of problems, we need to discover the common fundamental structures of these problems;

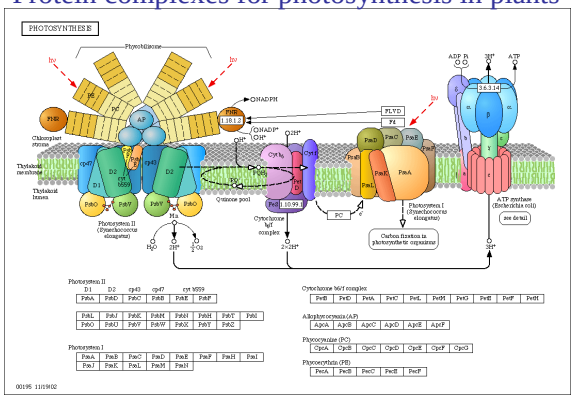
HOWEVER in biology one size does NOT fit all…

An important goal of bioinformatics is development of a data analysis infrastructure in support of Genomics and beyond

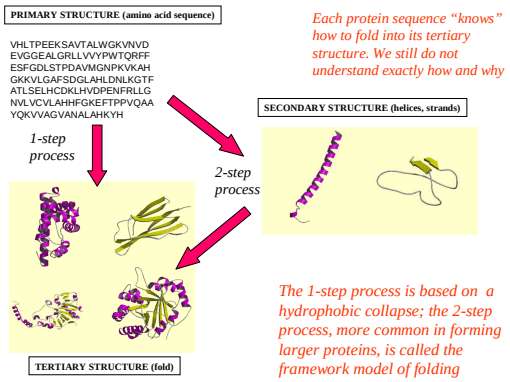
Protein structure hierarchical levels



Protein complexes for photosynthesis in plants



Protein folding problem



Each protein sequence "knows" how to fold into its tertiary structure. We still do not understand exactly how and why

The 1-step process is based on a hydrophobic collapse; the 2-step process, more common in forming larger proteins, is called the framework model of folding

Protein folding: step on the way is secondary structure prediction

- Long history -- first widely used algorithm was by Chou and Fasman (1974)
- Different algorithms have been developed over the years to crack the problem:
  - Statistical approaches
  - Neural networks (first from speech recognition)
  - K-nearest neighbour algorithms
  - Support Vector machines

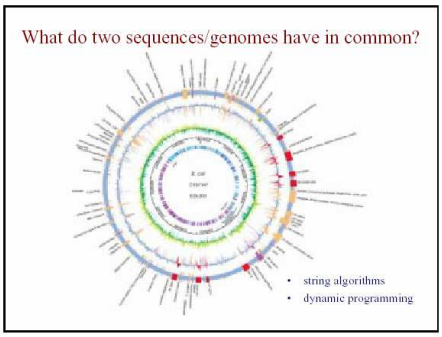
Algorithms in bioinformatics (recap)

- Sometimes the same basic algorithm can be re-used for different problems (1-method-multiple-problem)
- Normally, biological problems are approached by different researchers using a variety of methods (1-problem-multiple-method)

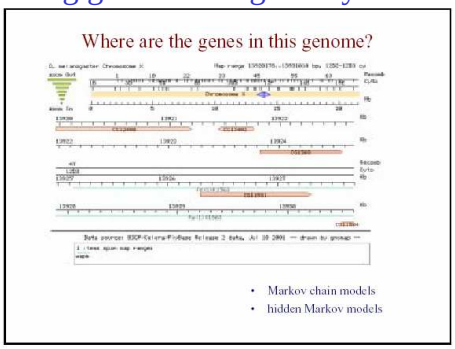
Algorithms in bioinformatics

- string algorithms
- dynamic programming
- machine learning (Neural Netsworks, k-Nearest Neighbour, Support Vector Machines, Genetic Algorithm, ..)
- Markov chain models, hidden Markov models, Markov Chain Monte Carlo (MCMC) algorithms
- molecular mechanics, e.g. molecular dynamics, Monte Carlo, simplified force fields
- stochastic context free grammars
- EM algorithms
- Gibbs sampling
- clustering
- tree algorithms
- text analysis
- hybrid/combinatorial techniques and more…

Sequence analysis and homology searching

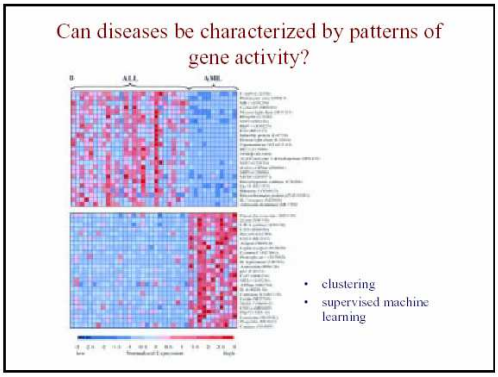


Finding genes and regulatory elements

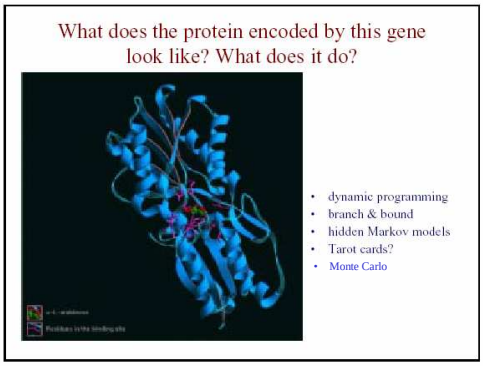


There are many different regulation signals such as start, stop and skip messages hidden in the genome for each gene, but what and where are they?

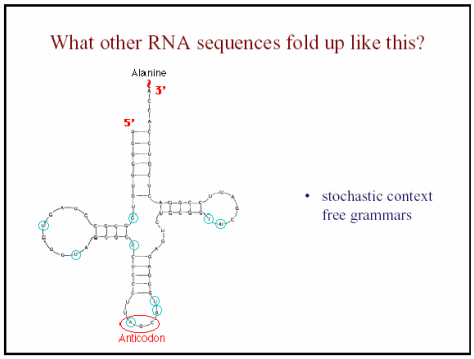
Expression data



Functional genomics



Protein translation



What is life?

- NASA astrobiology program:

  "Life is a self-sustained chemical system capable of undergoing Darwinian evolution"

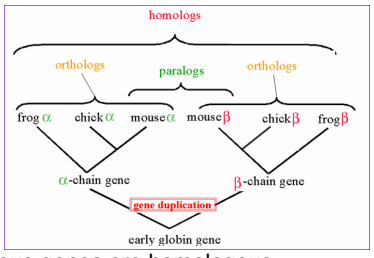
Evolution

Four requirements:

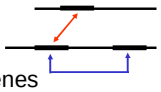
- Template structure providing stability (DNA)
- Copying mechanism (meiosis)
- Mechanism providing variation (mutations; insertions and deletions; crossing-over; etc.)
- Selection: some traits lead to greater fitness of one individual relative to another. Darwin wrote "*survival of the fittest*"

Evolution is a conservative process: the vast majority of mutations will not be selected (i.e. will not make it as they lead to worse performance or are even lethal) – this is called *negative* (or *purifying*) selection

Orthology/paralogy



Orthologous genes are homologous (corresponding) genes in different species



Paralogous genes are homologous genes within the same species (genome)

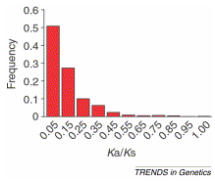
Changing molecular sequences

- **Mutations**: changing nucleotides ('letters') within DNA, also called '*point mutations*'
- A & G: purines, C & T/U: pyrimidines:
  - **Transition**: purine -> purine or pyrimidine -> pyrimidine
  - **Transversion**: purine -> pyrimidine or pyrimidine -> purine

Types of point mutation

- **Synonymous mutation**: mutation that does not lead to an amino acid change (where in the codon are these expected?)
- **Non-synonymous mutation**: does lead to an amino acid change
  - **Missense mutation:** one a.a replaced by other a.a
  - **Nonsense mutation**: a.a. replaced by stop codon (what happens with protein?)

Ka/Ks ratios



The frequency of different values of *Ka*/*Ks* for 835 mouse–rat orthologous genes. Figures on the *x* axis represent the middle figure of each bin; that is, the 0.05 bin collects data from 0 to 0.1

Ka/Ks Ratios

- **Ks** is defined as the number of synonymous nucleotide substitutions per synonymous site
- **Ka** is defined as the number of nonsynonymous nucleotide substitutions per nonsynonymous site
- The **Ka/Ks** ratio is used to estimate the type of selection exerted on a given gene or DNA fragment
- Need aligned orthologous sequences to do calculate **Ka/Ks** ratios (we will talk about alignment later).

Ka/Ks ratios

Three types of selection:

1. Negative (purifying) selection -> Ka/Ks < 1
2. Neutral selection (Kimura) -> Ka/Ks ~= 1
3. Positive selection -> Ka/Ks > 1