Bioinformatics

#1

Computer uses in biology:

- Data management
  - acquisition
  - storage, annotation
  - interpretation, analysis, data-mining
- Modelling, simulation

Service

Narrow def.

Research "biocomputing"

Broad def.

Narrow def.:

Science of biological data, mostly molecular biology. Description, management, interpretation.

Broad def.:

Science of biological knowledge. All computer applications in (molecular) biology including modeling (simulation of behavior)

#2

Structure: a set of elements or entities connected with relationships (molecular structures have many different representations)

Entities: can be defined as a thing capable of an independent existence that can be uniquely identified.

Relationship: connection between the entities (express the dependencies & requirements between them)

Examples

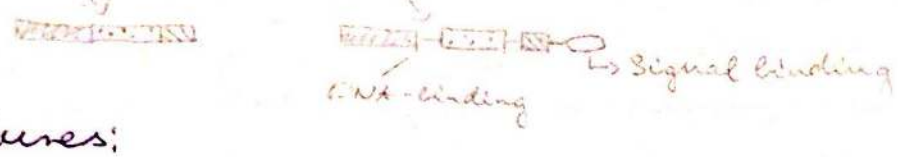
System (structure)	Entities	Relationships
Molecules	Atoms	Atomic interact.
Pathways	Enzymes	Chem. reactions
Genetic network	Genes	Co-regulation
Prot. sequence	Amino acid	Sequential linking

#3

Core data-types: Sequences, 3D structure, Networks, Text

- all have simplified and extended (annotated) versions
- we put all of them into database records

Sequences:
- Logical structure -> chemical formula (series of amino-acids or nucleotides)
- Standard description -> series of characters
- Simplified and/or extended visualization



3D-Structures:
- Logical struct.: 3D-chemical structures
- Standard description -> 3D coordinates + subunit descrip.
- (atomic, amino-acid, nucleotide)
- Simplified & ....

Networks:
- Entities, relationships
- Standard description -> graphs
- Simplified & ...

Texts - scientific publications
- Logical struct: Standard description:

Question	⟶	Introduction
Answer	⟶	Results (what), Methods (How)
Conclusion	⟶	Discussion

Structure definitions are hierarchical (atom - aa - protein - pathway - cell - tissue etc.)
For a given problem it is convenient to choose a standard description or "core structural level" (Eg. DNA sequences are the standard level for molecular biology problems.)
For a standard or core desc. we always have an underlying logical structure + various additional simplified & annotated views
-> annotation means extending ~~the~~ with external information

t4

Data representation types -> depending if we know or want to use the internal structure

- Unstructured
  - We know nothing about internal structure
  - only the properties are known, can be discreet or continuous
  - best described as vectors, sometimes a large number of dimensions -> vector operations are fast (Vector types - binary, non-binary
- Structured
  - We know the internal structure in terms of entites and relationships (EAV - Entity - Attributes - Values)
  - Information rich, allows detailed comparisons
  - Need alignment (matching) for comperison
  - character strings, graphs
- Mixed or Composition like
  - decompose an object to parts of known structure, count the parts
  - the result is a vector -> fast operation, alignment is not necessary
  - the info. content of the vector depends on the granularity of the parts. (Atomic composition of proteins or of people is not informative)

Granularity means the resolution of the description.

t5

Substitution matrices (Scoring matrices)

- contains costs of for amino acids identities & substitutions in an alignment (for amino acids it is a 20×20 symmetrical matrix)
- groups of related sequences can be organized into a multiple alignment for calculation of the matrix elements

S/T denotes that S is aligned with T or T with S

$$\Rightarrow M(S/T) = \log\left(\frac{f(S/T)}{f(S) \times f(T)}\right)$$

- The values are calculated from many MA. The log odd values in the matrix are then normalized to a giv range depending on the application.

PAM matrices (Percent Accepted Mutation)

- unit of evolutionary change for protein sequences
- calculated from related sequences organized into "accepted" evolutionary trees

BLOSUM matrices (Blocks Substitution Matrix)

- most often used
- derived from the 'blocks' database
- developed by Henikoff & Henikoff

PAM	vs	BLOSUM
- assumed a Markov Model of evolution		- no evolutionary model assumed
- derived from small, closely related proteins (~15% divergence)		- much larger, more diverse set of proteins (30% - 90%)
- higher PAM numbers to detect more remote sequences		- lower BLOSUM numbers ...
- lower PAM numbers to detect high similarities		- higher BLOSUM ...

A6

We usually compare one unknown sequence or less known sequence ('query') with a known sequence ('subject' or the 'database entry')

We can classify the alignment problems according to what the query is & what the subject or entry is.

What do we align?

① Two gene/prot. sequence

② One gene/prot seq vs a database of many prot/gene seq.

③ Many short DNA seq. vs one long DNA seq.

④ Many very short seq. vs many very short seq.

Classifying the alignment tasks

one-to-one
one-to-many
many-to-many
} Pairwise alignment problem

many-to-many → Multiple alignment problem

Examples

① basic exercise

② database searching

③ identifying mutations

④ diagnosing known disease mutations in selected human genes, bacterial metagenomics

*≈

Pairwise Alignment

- we have 2 sequences originating from a common ancestor
- the purpose of the alignment is to line up all positions that originates from the same position in the ancestral sequences 2) residues that were derived from the same residue pos. in the ancestral gene or protein in two sequences

How we do it?

→ Global alignment
- align 2 sequences from 'head to toe'
- exhaustive algorithm (all cases tested so the result is guaranteed to be optimal) → Needleman, Wunsch

→ Local alignment
- locate regions with high degree of similarity in 2 sequences
- exhaustive algorithm → Smith-Waterman

Multiple Alignment

- the goal is to present conserved sequence motifs within a group of related proteins, prot. sites or nucleic acid sites

- a method to visualize a group of sequences, a group description

When can we build one?

-> single domain sequences (if they are sufficiently similar)

-> sub-sequences that are short

-> DNA / RNA sequences, mostly with short sites

When can we not build?

-> multidomain proteins, unrelated prot-s., long DNA sequences

-> in all cases where point mutations are not the predominant cause of variations

Why not exhaustive?

- since we analyse short sites or single domain proteins Needleman-Wunsch is often used ...

A MA can be decomposed into pairwise alignments (PA)

- delete all lines but 2 & delete all columns that are empty (contain only gaps)
- this PA will still contain the 'traces' of the MA. A simple PA may give a diff. result.
- A MA can't always be constructed from pairwise alignments

PA -> MA 1. Additive

① compare every single seq. to every other (using PA)
② record the resulting gaps in vectors
③ combine vectors for each pair
④ construct MA

2. Iterative

① PA of all seq-s.
② record the resulting similarity scores
③ construct a guide tree from the matrix, containing the pairwise comparison values (Neighbor-joining)
④ construct MA

A8

The protein universe as a sequence similarity network

- contains many tightly connected 'similarity groups' these are the protein families that share structure and function
- if a sequence is > 90% identical with members of very little similarity between the groups
- most prot. similarities are trivial, so classification can be almost always can done by alignment

The rest is difficult

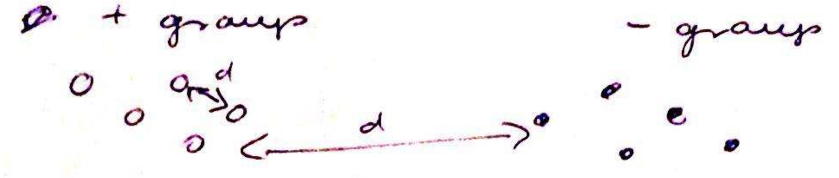
-> proteins consisting of a single domain from +/- clearcut groups (families), but deeper analysis may reveal subgroups

-> Multidomain proteins are connected to many diff. families and are difficult do deal with because of the shared domains

Group similarities

-> an 'easy case' of sim.

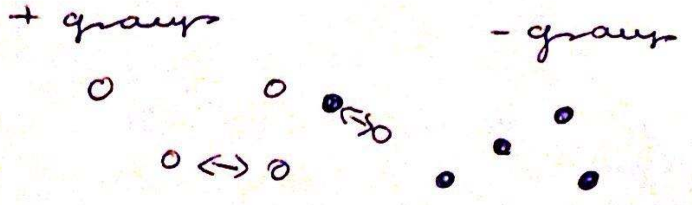
?



$d_{within} << d_{btw.}$

$similarity_{within} >> sim._{btw}$

-> 'difficult' case'



$d_{within} \sim d_{btw}$

$sim_{w} \sim sim_{b}$

- neighborhoods of + members include - members

A9

Similarity searching

- given a query & a database, find the entry in the database that is most similar to the query in terms of a numerical similarity measure (distance, similarity score)

- in contrast: retrieval looks for an exact match to the query

→ Example: - Is John on the playlist?

Retrieval: Yes/No, based on exact match

Similarity search: His brother Joe Brown is.

-> so we classify Joh into the Brown family, based on approximate matching.

-> How it can help to understand the prot. universe?

Class Annotated Databases

-> COG - database of orthologous groups, ~~NCBI - all sequences~~ *

-> SBASE - library of protein domain sequences

-> domains, local homology assignments

- clusters defined as searchable sequence collections similar to COG

* COG clustering

- detect triangles of best hits two genomes
- merge triangles with a common side to form COG
- case-by-case manual analysis, examination of large COG's

Annotated vs.	Non-Annotated (like Uniref)
- manually curated, homogenous, same function	- automatically maintained sometimes heterogenous
- searched as a separated database	- searched as part of a standard sequence dbase (Uniprot)
- can be used automatically, but difficult to maintain	- searches need to be checked manually

How can they help?

=> 06_BLAST.ppt slides 54-57 !

A 10

Maximum Likelihood algorithm

- a Model based phylogenetical tree generating algorithm
- mostly for nucleotides (few sequences)
- it shows what is the probability of seeing the observed data (D) given a model/theory (T) $P_r(D|T)$

Bayesian inference

- - 11 -
- for many/long nucleotide sequences
- what is the probability that the model/theory is correct given the observed data? $P_r(T|D)$

A11

Main steps of...

- getting sample (tissue, blood, hair)
- breaking up the cells
- isolation of DNA
- DNA quality control (UV spectrophotometer)
- DNA fragmentation -> sequencing

Factor V Leiden

-> Blood clotting is regulated by multiple pathways, in which Factor V has a critical role

-> Mutation in factor V (Leiden) leads to increased risk of thrombotic events

-> Factor V Leiden can be detected by sanger venom, PCR-RFLP and sequencing

BRCA

-> cancer is a genetic disease originating in a single cell

-> BRCA 1/2 are tumor supressor genes

-> Germ line mutation in BRCA 1/2 leads to early breast cancer

-> „A gene is not patent eligible merely because it has been isolated"

KRAS

-> oncogene-addiction: tumors depend on a single signal transduction pathway

-> The ERBB/KRAS has critical role in solid tumors

-> R. Patients with a KRAS mutation do not respond to a therapy targeting a higher member in the pathway

-> methods for KRAS mutation status detection are either PCR or sequencing based