Lecture outline

- Sequence alignment
  - Why do we need to align sequences?
  - Evolutionary relationships
- Gaps and scoring matrices
- Dynamic programming
  - Global alignment (Needleman & Wunsch)
  - Local alignment (Smith & Waterman)
- Database searches
  - BLAST
  - FASTA

1

Alignment

Procedure of comparing two (pairwise) or more (multiple) sequences by searching for a series of individual characters that are in the same order in the sequences

```
GCTAGTCAGATCTGACGCTA
|  ||||  |||||  |||
TGGTCACATCTGCCGC
```

2

Alignment

Procedure of comparing two (pairwise) or more (multiple) sequences by searching for a series of individual characters that are in the same order in the sequences

```
VLSPADKTNVKAAWGKVGAHAGYEG
|||   |    |    | || |     ||
VLSEGDWQLVLHVWAKVEADVAGEG
```

3

Sequence alignment

- Comparing DNA/protein sequences for
  - Similarity
  - Homology
- Prediction of function
- Construction of phylogeny
- Shotgun assembly
  - End-space-free alignment / overlap alignment
- Finding motifs

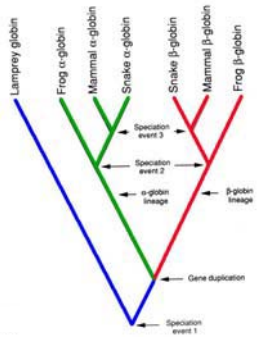
4

Homology

- Orthologs
  - Divergence follows speciation
  - Similarity can be used to construct phylogeny between species
- Paralogs
  - Divergence follows duplication
- Xenologs
- Article on terminology
- ISMB tutorial on protein sequence comparison

5

Orthologs and paralogs



6

Sources of variation

- Nucleotide substitution
  - Replication error
  - Chemical reaction
- Insertions or deletions (indels)
  - Unequal crossing over
  - Replication slippage
- Duplication
  - a single gene (complete gene duplication)
  - part of a gene (internal or partial gene duplication)
    - Domain duplication
    - Exon shuffling
  - part of a chromosome (partial polysomy)
  - an entire chromosome (aneuploidy or polysomy)
  - the whole genome (polyploidy)

7

Differing rates of DNA evolution

- Functional/selective constraints (particular features of coding regions, particular features in 5' untranslated regions)
- Variation among different gene regions with different functions (different parts of a protein may evolve at different rates).
- Within proteins, variations are observed between
  - surface and interior amino acids in proteins (order of magnitude difference in rates in haemoglobins)
  - charged and non-charged amino acids
  - protein domains with different functions
  - regions which are strongly constrained to preserve particular functions and regions which are not
  - different types of proteins -- those with constrained interaction surfaces and those without

8

Common assumptions

- All nucleotide sites change independently
- The substitution rate is constant over time and in different lineages
- The base composition is at equilibrium
- The conditional probabilities of nucleotide substitutions are the same for all sites, and do not change over time
- Most of these are not true in many cases…

9

Lecture outline

- Sequence alignment
  - Why do we need to align sequences?
  - Evolutionary relationships
- Gaps and scoring matrices
- Dynamic programming
  - Global alignment (Needleman & Wunsch)
  - Local alignment (Smith & Waterman)
- Database searches
  - BLAST
  - FASTA

10

A simple alignment

- Let us try to align two short nucleotide sequences:
  - AATCTATA and AAGATA
- Without considering any gaps (insertions/deletions) there are 3 possible ways to align these sequences

```
AATCTATA    AATCTATA    AATCTATA
AAGATA        AAGATA      AAGATA
```

- Which one is better?

11

Scoring the alignments

- We need to have a scoring mechanism to evaluate alignments
  - match score
  - mismatch score
- We can have the total score as:

$$\sum_{i=1}^{n} \text{match or mismatch score at position i}$$

- For the simple example, assume a match score of 1 and a mismatch score of 0:

```
AATCTATA    AATCTATA    AATCTATA
AAGATA      AAGATA      AAGATA
4           1           3
```

12

Good alignments require gaps

- Maximal consecutive run of spaces in alignment
  - Matching mRNA (cDNA) to DNA
  - Shortening of DNA/protein sequences
  - Slippage during replication
  - Unequal crossing-over during meiosis
  - …
- We need to have a scoring function that considers gaps also

13

Simple alignment with gaps

- Considering gapped alignments vastly increases the number of possible alignments:

AATCTATA	AATCTATA	AATCTATA	more?
AAG-AT-A	AA-G-ATA	AA--GATA	
1	3	3	

- If gap penalty is -1 what will be the new scores?

14

More complicated gap penalties

- Nature favors small number of long gaps compared to large number of short gaps.
- How do we adjust our scoring scheme to account for this fact above?

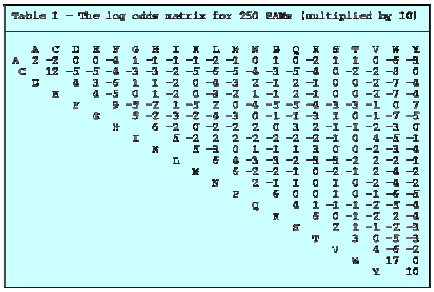
  By having different gap opening and gap extension penalties.
- Choices of gap penalties
  - Linear
  - Affine
    - Gap open penalty
    - Gap extension penalty
  - Arbitrary

15

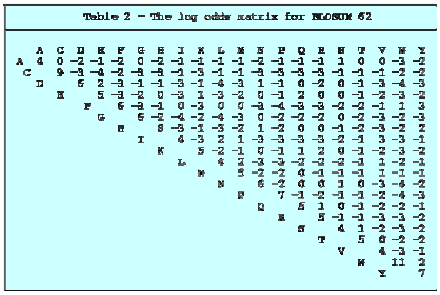
Score matrix

- Instead of having a single match/mismatch score for every pair of nucleotides or amino acids, consider chemical, physical, evolutionary relationships:
  - E.g.
    - alanine vs. valine or alanine vs. lysine? Alanine and valine are both small and hydrophobic, but lysine is large and charged.
    - which substitutions occur more in nature?
- Assign scores to each pair of symbol
  - Higher score means more similarity

16



17



18

Major Differences between PAM and BLOSUM

PAM	BLOSUM
Built from global alignments	Built from local alignments
Built from small amout of Data	Built from vast amout of Data
Counting is based on minimum replacement or maximum parsimony	Counting based on groups of related sequences counted as one
Perform better for finding global alignments and remote homologs	Better for finding local alignments
Higher PAM series means more divergence	Lower BLOSUM series means more divergence

19

Typical score matrix

- DNA
  - Match = +1
  - Mismatch = -3
  - Gap penalty = -5
  - Gap extension penalty = -2
- Protein sequences
  - Blossum62 matrix
  - Gap open penalty = -11
  - Gap extension = -1

20