BINF 730

Lecture 2

Sequence Alignment

DNA Sequence Alignment – Why?

Recognition sites might be common – restriction enzymes, start sequences, stop sequences, other regulatory sequences

Homology – evolutionary common progenitor

Mutations

-Deletions

-Insertions

-Transitional Substitution (purine-purine A-G, pyr-pyr T-C)

-Translational Substitution (pur-pyr, pyr-pur)

Example

Start with ACGTACGT after 9540 generations with the following probabilities:

Deletion 0.0001

Insertion 0.001

Transitional substitution 0.00008

Translational substitution 0.00002

Example

The two new sequences are:

- - ACG –T -A - - - CG -T - - - -
ACACGGTCCTAATAATGGCC

- - - AC - GTA- C- - G - T - -
CAG - GAAGATCTTAGTTC

Example

However if we align the two sequences by superposition

- ACAC- GGTCCTAAT--AATGGCC
CAG- GAA- G- AT- - CTTAGTTC- -

Example

or using Gotoh's algorithm with mismatch penalty 3 and gap penalty function g(k) = 2+2k for length k gap

ACACG - - GTCCTAATAATGGCC
- CAGGAAGATCT - - TAGTT - - C

The alignment depends on algorithm used!

Protein sequence alignment

A. Homologous proteins
   i. Evolutionary common origin
   ii. Structural similarity
   iii. Functional similarity

B. Conserved regions
   i. Functional domains
   ii. Evolutionary similarity
   iii. Structural motif

Example 3.2

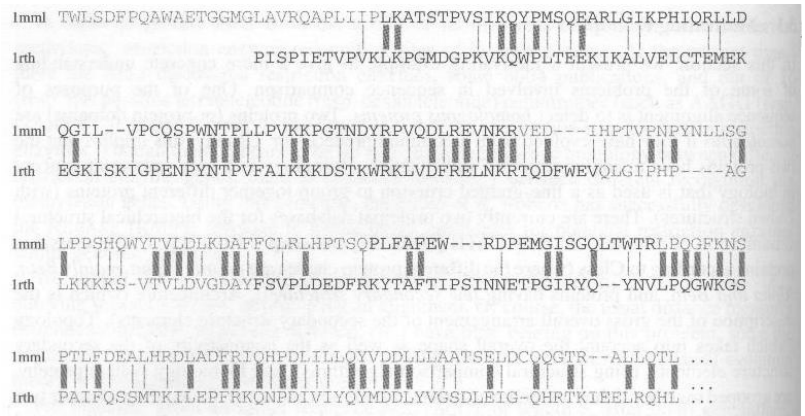

Figure 3.2 Part of an alignment of Mmlv reverse transcriptase (PDB code 1MML) and HIV-1 reverse transcriptase (PDB code 1RTH). The alignments were produced using the FASTA package [PL88] with the Blosum50 substitution matrix [HH92]. The sequence of the first domains of both chains are printed in black, while all other amino acids are printed in gray. Thicker gray bars indicate an exact matching of amino acids, whereas the thinner gray bars indicate amino acids that are similar according to the Blossum50 similarity matrix.

Choosing the best alignment

- Every alignment has a score
- Chose alignment with highest score
- Must choose appropriate scoring function
- Scoring function based on evolutionary model with insertions, deletions, and substitutions
- Use substitution score matrix – contains an entry for every amino acid pair

Substitution score matrix

	A	D	K
S	$s_{S,A}$	$s_{S,D}$	$s_{S,K}$
R	$s_{R,A}$	$s_{R,D}$	$s_{R,K}$
K	$s_{K,A}$	$s_{K,D}$	$s_{K,K}$

Scoring Strategies

- Ad hoc method – a biologist can set up a score matrix that gives a good alignment
- Use physical/chemical properties
- Statistical approach

Statistical approach

- Let s and s' be two amino acid sequences of length n that we want to compute an alignment score
- Assume only substitutions occur (no insertions or deletions)
- Works for local alignment
- Odds Ratio and Log Odds Ratio

Odds Ratio and Log Odds Ratio

The score for aligning s and s' is based on the comparison of the hypothesis that the two sequences are generated randomly with the hypothesis that they come from a common ancestor.

Assume q_A is the probability of producing amino acid A in model R (based on the relative frequency at which A is found in proteins). The probability for the null hypothesis (that s and s' do not stem from a common ancestor) is

$$P(s, s' \mid R) = \prod_{1 \le i \le n} q_{s,i} \prod_{1 \le i \le n} q_{s',i} = \prod_{1 \le i \le n} q_{s,i}\, q_{s',i}$$

Odds Ratio and Log Odds Ratio

The second hypothesis (homologous hypothesis) that s and s' arise from a common ancestor sequence r, of length n, is based on the evolutionary model (E). The probability that the amino acids A and B are aligned and hence have been derived from an ancestor amino acid C is given by $p_{A,B}$ is given by

$$P(s,s' \mid E) = \prod_{1 \le i \le n} p_{s,is',i}$$

How this probability is determined will be explained later.

Odds Ratio and Log Odds Ratio

The odds ratio compares the homologous hypothesis with the null hypothesis

$$\frac{P(s,s' \mid E)}{P(s,s' \mid R)} = \frac{\prod_{1 \le i \le n} p_{s,is',i}}{\prod_{1 \le i \le n} q_{s,i}\, q_{s',i}} = \prod_{1 \le i \le n} \frac{p_{s,is',i}}{q_{s,i}\, q_{s',i}}$$

To achieve a scoring function that is additive rather that multiplicative, the log odds ratio can be used

$$s_{A,B} = \log \frac{p_{AB}}{q_A q_B}$$

Point Accepted Mutation (PAM) and Amino Acid Pair Probabilities

We mentioned that we must choose an appropriate evolutionary model $E((p_{AB})_{AB})$ for the homologous hypothesis, ie we have to find p_{AB} for each pair of amino acids A and B. Since we are using a statistical approach, this has to be estimated from data. If we know that two sequences s and s' are homologous, we could estimate p_{AB} by finding the value of p_{AB} that would maximize

$$P(E((p_{AB})_{AB})|s,s')$$

PAM and Amino Acid Pair Probabilities

This can be done by using the maximum likelihood approach (section 2.1.6 pp 52-53, Clote and Backofen)

Lagrange Multipliers (Section 2.2 Clote and Backofen)

Appendix (Chapter 3 Clote and Backofen)

PAM and Amino Acid Pair Probabilities

We now have which is the relative frequency of a pair (A,B) in the alignment of s and s' where $n_{AB}(s,s')$ is the number of times the amino acids A and B are aligned in one column in the alignment of s and s' and n is the length of s and s'.

To find a value for n_{AB}, some homologous sequences are needed. To do this Dayhoff and co-workers used local sequence alignment.

PAM and Amino Acid Pair Probabilities

Problem – They used sequence alignment to find a substitution matrix (substitution score matrix) for sequence alignment – which comes first, the chicken or the egg?

Answer – Use only very closely related sequence (sequences differ in at most 15% of the amino acid.

Caveat – The substitution matrix is only valid for closely related protein sequences