Bioinformatics – Lecture Notes

Announcement – Seminar
Understanding Proteins from Physical Principles and Evolutionary Information
Dr. Huan-Xiang Zhou, Florida State University
Thursday, January 24, 1:30-3:30 PM
Student Union, Galaxy Room A

Class 3

1. 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

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

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

    However if we align the two sequences by superposition

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

    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!

2. Protein sequence alignment
    A. Homologous proteins
        i. Evolutionary common origin
        ii. Structural similarity
        iii. Functional similarity
    B. Conserved regions
        iv. Functional domains
        v. Evolutionary similarity
        vi. Structural motif

    Example Figure 3.2 page 84

3. As shown before there are many possible alignments – which is correct?
    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
4. 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	$s_{K,A}$	$s_{K,D}$	$s_{K,K}$

    Ad hoc method – a biologist can set up a score matrix that gives good alignment
    Use physical/chemical properties
    Statistical approach
5. 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
        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

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

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

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

How this probability is determined will be explained later.

The odds ratio compares the homologous hypothesis with the null hypothesis

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

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

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

6. 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

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

This can be done by using the maximum likelihood approach (section 2.1.6 pp 52-53)
Lagrange Multipliers (Section 2.2)
Appendix (Chapter 3)

7. Global Alignment (next)