CISC 889 Bioinformatics (Spring 2004)

Sequence pairwise alignment (II)

Heuristic algorithms
- –FASTA
- –BLAST

Score statistics
- –Bayesian
- –Extreme value distribution
- –E-value

Heuristic alignment algorithms

- motivation: speed

  sequence DB ~ O(100,000,000) basepair

  query sequence 1000 basepair

  O(nm) time complexity => 10^{11} matrix cells in dynamic programming table

  if 10,000,000 cells/second => 10000 seconds ~ 3 hours.

  O(n+m) time => ~ 10 seconds

- heuristic versus rigorous

FASTA [Pearson & Lipman 1988]

1. k-tup (k=6 for DNA, 2 for proteins)

   rule of thumb: the smaller the word, the slower and more sensitive the search is

2. build a lookup table (also called as hash table or dictionary)

word	positions_in_the_query
AAAAAA	xxx
AAAAAT	xxx
AAAAAC	xxx
....	….

FASTA cont'd

3. scan through the database: for each every word of size k, look up in the lookup table in step 2. The offsets between the positions of the word in the query and the database entry are calculated and saved.

(implication: same offsets suggest segment of similarity.)

4. join nearby *contiguous* stretches of similarity (diagonal) =>scores init1.

5. join adjacent diagonals into a single long region (by introducing gaps) => scores initN.

6. do a dynamic programming algorithm for regions with high initN score to determine the OPT score.

Q: what is the FASTA format?

Basic **L**ocal **A**lignment **S**earch **T**oolkit [Altschul et al, 1990]

1. A list of neighborhood words of fixed length (3 for protein and 11 for DNA) that match the query with score > a threshold.
2. Scan the database sequences and look for words in the list; once find a spot, try a "hit extension" process to extend the possible match as an ungapped alignment in both directions, stopping at the maximum scoring extension.

Variants of BLAST search

- BLASTP: protein vs. protein
- BLASTN: nucleotide vs. nucleotide
- BLASTX: nucleotide (translated to protein) vs. protein
- TBLASTN: protein vs. nucleotide (translated to protein)
- TBLASTX: nucleotide (translated to protein) vs. nucleotide (translated to protein)

Note: Since proteins are strings of 20 alphabets the odds of having false positive matches is significantly lower than that of DNA sequences, which are strings of 4 alphabets.

Significance of scores

Pairwise alignment goals:

(1) whether and

(2) how two sequences are related.

It is rare that you have just two particular sequences to compare. More often, you have one query sequence and a large database of sequences.

Database searching: find all sequences in the database that are related to the query sequence.

Solution:

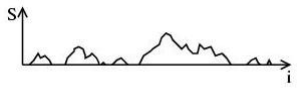
(1) For each sequence in the database, use Smith-Waterman/FASTA/BLAST to align with the query sequence and return the score of the optimal alignment.

(2) Rank the sequences by the score.

Q: how good is a score?

Score statistics [Karlin & Altschul 1990]

-The score of an ungapped alignment is $H_{i,j} = \max\{H_{i-1,j-1} + s(x_i,y_i), 0\}$



- $\sum_{a,b \in \text{alphabet}} s(a,b)p(a)b(b) < 0 \Rightarrow$ most regions receive zero score.

-The scores of individual sites are independent.

-The landscape of non-zero zero regions are "islands" in the sea.

-The optimal alignment score S is the global maximum of these island peaks:

$$S = \max\{\sigma_1, \sigma_2, \sigma_1, \ldots \sigma_\kappa\}$$

- The island peak σ_i satisfies a Poisson distribution:

$$\Pr(\sigma_i > x) = (\text{constant}) \exp(-\lambda x)$$

Score Statistics (cont'd)

The probability that the maximum S is smaller than x is

$$P(S < x) = \prod_i [1 - Pr(\sigma_i > x)] \rightarrow \exp[-\kappa e^{-\lambda x}] \text{ when } \kappa \rightarrow \infty.$$

This is a form of **Extreme Value Distribution**.

Definition: Take N samples from the density g(u), the probability that the largest amongst them is less than x is G(x)N, where $G(x) = \int_{-\infty}^{x} g(u)\,du$. The density is Ng(x)G(x)(N-1) . The limit for large N of Ng(x)G(x)(N-1) is called the EVD for g(x). (Draw a EVD curve).

Once the score distribution is known, for a given score, the percentage of the curve is to its right gives the p-value of that score.

p-value = probability of at least one sequence scoring above S in the given database.

$$P(S < x) = 1 - \exp[-\kappa e^{-\lambda x}].$$

E-value = expected number of scores better than S in a database search.

$$E(S) = kmn\, e^{-\lambda S}.$$

Notes:

- all of the above discussions only applicable to local alignments.
- For gapped local alignments, same statistics, although not proved, are believed to apply.
- The trick is to learn λ and K. These values depend upon the substitution matrix and can be estimated from randomly generated data. The parameter λ is the positive root of the equation

$$\sum_{a,b \in \text{alphabet}} s(a,b)p(a)b(b) = 1$$

- score statistics for global alignments are not well known.

Q: what is a bit score in the blast search result?

A: bit score is defined as $S' = (\lambda S - \ln K)/\ln 2$

it is then convenient to calculate the e-value

$$E(S) = mn\, 2^{-S'}$$

Question: What threshold to use?

Answer: No absolute answer, it varies.

E.g., E-value < 1.

$E(S) = kmn\, e^{-\lambda S}$. $\Rightarrow S > T + \log(mn)/\lambda$.

Bayesian perspective

$$P(M|x,y) = P(x,y|M)P(M)/P(x,y)$$

$$= P(x,y|M)P(M) / [P(x,y|R)P(R) + P(x,y|M)P(M)]$$

$$= \frac{[P(x,y|M)P(M)/P(x,y|R)P(R)]}{[1+ P(x,y|M)P(M)/P(x,y|R)P(R)]}$$

if

$S = \log[P(x,y|M)/P(x,y|R)]$, this is the score returned from a SW alignment.

$S' = S + \log[P(M)/P(R)]$

then

$P(M|x,y) = \exp(S') / [1 + \exp(S')]$

(draw a curve and explain)

Gap penalties

Linear

$$\gamma(g) = - g\, d$$

where g is the gap length and d is the penalty for a gap of one base

Affine

$$\gamma(g) = - d - (g-1)e$$

where d is gap-open penalty and e is gap-extension penalty

Note: gap penalty is a sort of gray area due to less knowledge about gap distribution.