Lecture outline

- Database searches
  - BLAST
  - FASTA
- Statistical Significance of Sequence Comparison Results
  - Probability of matching runs
  - Karin-Altschul statistics
  - Extreme value distribution

1

BLAST

- Basic Local Alignment Search Tool
  - Altschul *et al*. 1990,1994,1997
- Heuristic method for local alignment
- Designed specifically for database searches
- Idea: good alignments contain short lengths of exact matches

2

Steps of BLAST

1. *Query words* of length 4 (for proteins) or 11 (for DNA) are created from query sequence using a sliding window

```
MEFPGLGSLGTSEPLPQFVDPALVSS
MEFP
 EFPG
  FPGL
   PGLG
    GLGS
```

- Scan each database sequence for an exact match to *query word*s. Each match is a *seed* for an ungapped alignment.

3

Steps of BLAST

3. *(Original BLAST)* extend matching words to the left and right using ungapped alignments. Extension continues as long as score does not fall below a given threshold. This is an HSP (high scoring pair).

   (BLAST2) Extend the HSPs using gapped alignment.

4

Steps of BLAST

4. Using a cutoff score S, keep only the extended matches that have a score at least S.
5. Determine statistical significance of each remaining match.

5

Example BLAST run

- BLAST website:
  - http://www.ncbi.nlm.nih.gov/BLAST/

6

FASTA

- Derived from logic of the dot plot
  - compute best diagonals from all frames of alignment
- Word method looks for exact matches between words in query and test sequence
  - construct word position tables
  - DNA words are usually 6 bases
  - protein words are 1 or 2 amino acids
  - only searches for diagonals in region of word matches = faster searching

7

Steps of FASTA

1. Find k-tups in the two sequences (k=1-2 for proteins, 4-6 for DNA sequences)
2. Create a table of positions for those k-tups

8

The offset table

```
position 1 2 3 4 5 6 7 8 9 10 11
proteinA n c s p t a . . . . .
proteinB . . . . . a c s p r k
```

amino acid	position in protein A	position in protein B	offset pos A - posB
a	6	6	0
c	2	7	-5
k	-	11	
n	1	-	
p	4	9	-5
r	-	10	
s	3	8	-5
t	5	-	

```
Note the common offset for the 3 amino acids c,s and p
A possible alignment is thus quickly found -
protein 1 n c s p t a
            | | |
protein 2 a c s p r k
```

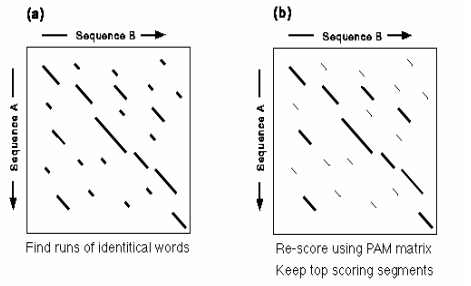
9

FASTA

3. Select top 10 scoring "local diagonals" with matches and mismatches but no gaps.
4. Rescan top 10 diagonals (representing alignments), score with PAM250 (proteins) or DNA scoring matrix. Trim off the ends of the regions to achieve highest scores.

10

FASTA Algorithm



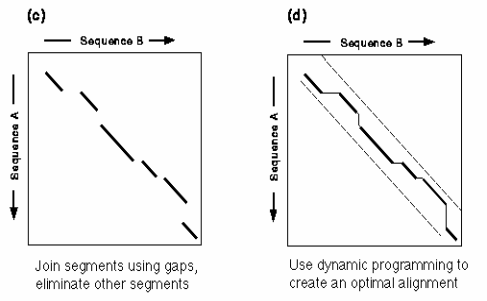
11

FASTA

5. After finding the best initial region, FASTA performs a DP global alignment centered on the best initial region.

12

FASTA Alignments



13

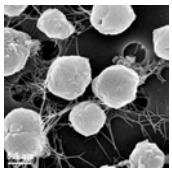
History of sequence searching

- 1970: NW
- 1981: SW
- 1985: FASTA
- 1990: BLAST
- 1997: BLAST2

14

The purpose of sequence alignment

- Homology
- Function identification
  - about 70% of the genes of *M. jannaschii* were assigned a function using sequence similarity (1997)




15

Similarity

- How much similar do the sequences have to be to infer homology?
- Two possibilities when similarity is detected:
  - The similarity is by chance
  - They evolved from a common ancestor – hence, have similar functions

16

Measures of similarity

- Percent identity:
  - 40% similar, 70% similar
  - problems with percent identity?
- Scoring matrices
  - matching of some amino acids may be more significant than matching of other amino acids
  - PAM matrix in 1970, BLOSUM in 1992
  - problems?

17

Statistical Significance

- Goal: to provide a universal measure for inferring homology
  - How different is the result from a random match, or a match between unrelated requences?
  - Given a set of sequences *not related* to the query (or a set of random sequences), what is the probability of finding a match with the same alignment score by chance?
- Different statistical measures
  - p-value
  - E-value
  - z-score

18

Statistical significance measures

- *p-value*: the probability that at least one sequence will produce the same score by chance
- *E-value*: expected number of sequences that will produce same or better score by chance
- *z-score*: measures how much standard deviations above the mean of the score distribution

19

How to compute statistical significance?

- Significance of a match-run
  - Erdös-Renyí
- Significance of local alignments without gaps
  - Karlin-Altschul statistics
  - Scoring matrices revisited
- Significance of local alignments with gaps
- Significance of global alignments

20

Analysis of coin tosses

- Let black circles indicate heads
- Let p be the probability of a "head"
  - For a "fair" coin, p = 0.5
- Probability of 5 heads in a row is (1/2)^5=0.031
- The expected number of times that 5H occurs in above 14 coin tosses is 10*0.031 = 0.31

21

Analysis of coin tosses

- The expected number of a length *l* run of heads in *n* tosses.

$$E(l) \cong np^l$$

- What is the expected length *R* of the longest match in *n* tosses?

$$1 = np^R \longrightarrow R = \log_{1/p}(n)$$

22

Analysis of coin tosses

- (Erdös-Rényi) If there are *n* throws, then the expected length *R* of the longest run of heads is

$$R = \log_{1/p}(n)$$

23

Example

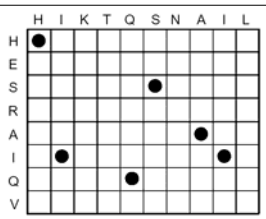
- Example: Suppose n = 20 for a "fair" coin

$$R=\log_2(20)=4.32$$

  - In other words: in 20 coin tosses we expect a run of heads of length 4.32, once.
- Trick is how to model DNA (or amino acid) sequence alignments as coin tosses.

24

Analysis of an alignment



- Probability of an individual match p = 0.05
- Expected number of matches: 10x8x0.05 = 4
- Expected number of two successive matches
  ≅10x8x0.05x0.05 = 0.2

25

Matching runs in sequence alignments

- Consider two sequences $a_{1..m}$ and $b_{1..n}$
- If the probability of occurrence for every symbol is p, then a match of a residue a_i with b_j is p, and a match of length l from a_i,b_j to a_{i+l-1},b_{j+l-1} is p^l.
- The head-run problem of coin tosses corresponds to the longest run of matches along the diagonals

26

Matching runs in sequence alignments

- There are m-l+1 x n-l+1 places where the match could start

$$E(l) \cong mnp^l$$

- The expected length of the longest match can be approximated as

$$R=\log_{1/p}(mn)$$

where m and n are the lengths of the two sequences.

27

Matching runs in sequence alignments

- So suppose $m = n = 10$ and we're looking at DNA sequences

$$R=\log_4(100)=3.32$$

- This analysis makes assumptions about the base composition (uniform) and no gaps, but it's a good estimate.

28

Length vs. Score

- We generally compute alignment scores instead of length of runs in sequence alignments
- Consider all mismatches receive a negative score of -∞ and a_ib_j match receives a positive score of $s_{i,j}$.
- What is the expected number of matching runs with a score x or higher?
  - $E(S>=x) \propto mnp^x$
  - or $E(S>=x) \propto mne^{-\lambda x}$ where $\lambda = -\ln p$

29

Statistics of local alignment

- Karin and Altschul provided an extension to the problem of runs where similarity is bounded by -∞
- However, a scoring matrix should satisfy the following constraint:
  - The expected score obtained by a scoring matrix should be negative to bound local alignments.

$$E(s_{i,j}) = \sum_{i,j} p_i p_j s_{i,j} < 0$$

  - Otherwise?
- If this requirement is met then the expected number of alignments with score S is:

$$E(S \geq x) = Kmne^{-\lambda x}$$

30

Statistics of local alignment

$$E(S \geq x) = Kmne^{-\lambda x}$$

- $K < 1$ is a proportionality constant that corrects the mn "space factor" for the fact that there are not really mn independent places that could have produced score $S \geq x$.
- K has little effect on the statistical significance of a similarity score
- λ is closely related to the scoring matrix used and it takes into account that the scoring matrices do not contain actual probabilities of cooccurence, but instead a scaled version of those values. To understand how λ is computed, we have to revisit the construction of scoring matrices.

31

Scoring Matrices

- In 1970s there were few protein sequences available. Dayhoff used a limited set of families of protein sequences multiply aligned to infer mutation likelihoods.

```
PGNPFATPLEILPEWYLYPVFQILRVLPNKLLGIACQGAIPLGLMMVPFIE
PANPFATPLEILPEWYFYPVFQILRTVPNKLLGVLAMAAVPVGLLTVPFIE
PANPMSTPAHIVPEWYFLPVYAILRSIPNKLGGVAAIGLVFVSLLALPFIN
PANPLVTPPHIKPEWYFLFAYAILRSIPNKLGGVLALLFSILMLLLVPFLH
PANPLSTPAHIKPEWYFLFAYAILRSIPNKLGGVLALLLSILVLIFIPMLQ
PANPLSTPPHIKPEWYFLFAYAILRSIPNKLGGVLALLLSILILIFIPMLQ
IANPMNTPTHIKPEWYFLFAYSILRAIPNKLGGVIGLVMSILIL..YIMIF
ESDPMMSPVHIVPEWYFLFAYAILRAIPNKVLGVVSLFASILVL..VVFVL
IVDTLKTSDKILPEWFFLYLFGFLKAIPDKFMGLFLMVILLFSL..FLFIL
```

32

Scoring Matrices

```
PGNPFATPLEILPEWYLYPVFQILRVLPNKLLGIACQGAIPLGLMMVPFIE
PANPFATPLEILPEWYFYPVFQILRTVPNKLLGVLAMAAVPVGLLTVPFIE
PANPMSTPAHIVPEWYFLPVYAILRSIPNKLGGVAAIGLVFVSLLALPFIN
PANPLVTPPHIKPEWYFLFAYAILRSIPNKLGGVLALLFSILMLLLVPFLH
PANPLSTPAHIKPEWYFLFAYAILRSIPNKLGGVLALLLSILVLIFIPMLQ
PANPLSTPPHIKPEWYFLFAYAILRSIPNKLGGVLALLLSILILIFIPMLQ
IANPMNTPTHIKPEWYFLFAYSILRAIPNKLGGVIGLVMSILIL..YIMIF
ESDPMMSPVHIVPEWYFLFAYAILRAIPNKVLGVVSLFASILVL..VVFVL
IVDTLKTSDKILPEWFFLYLFGFLKAIPDKFMGLFLMVILLFSL..FLFIL
```

- Dayhoff represented the similarity of amino acids as a log odds ratio:

$$s_{ij} = \log(q_{ij} / p_i p_j)$$

where q_{ij} is the observed frequency of co-occurrence, and p_i, p_j are the individual frequencies.

33

Example

- If M occurs in the sequences with 0.01 frequency and L occurs with 0.1 frequency. By random pairing, you expect 0.001 amino acid pairs to be M-L. If the observed frequency of M-L is actually 0.003, score of matching M-L will be
  - $\log_2(3)$=1.585 bits
- Since, scoring matrices are usually provided as integer matrices, these values are scaled by a constant factor. λ is approximately the inverse of the original scaling factor.

34