Algorithms for Sequence Alignment

- Previous lectures
  - Global alignment (Needleman-Wunsch algorithm)
  - Local alignment (Smith-Waterman algorithm)
- Heuristic method
  - BLAST
- Statistics of BLAST scores

Dynamic programming

Dynamic programming matrix:

i		j →	0	1 T	2 G	3 C	4 T	5 C	6 G	7 T	8 = N A
			0	−6	−12	−18	−24	−30	−36	−42	−48
1	T		−6	5	−1	−7	−13	−19	−25	−31	−37
2	T		−12	−1	3	−3	−2	−8	−14	−20	−26
3	C		−18	−7	−3	8	2	3	−3	−9	−15
4	A		−24	−13	−9	2	6	0	1	−5	−4
5	T		−30	−19	−15	−4	7	4	−2	6	0
M = 6	A		−36	−25	−21	−10	1	5	2	0	11

(sequence x) ↓

Optimum alignment scores 11:

T	−	−	T	C	A	T	A
T	G	C	T	C	G	T	A
+5	−6	−6	+5	+5	−2	+5	+5

x = TTCATA
y = TGCTCGTA

Scoring system:
+5 for a match
-2 for a mismatch
-6 for each indel

BLAST
Basic Local Alignment Search Tool

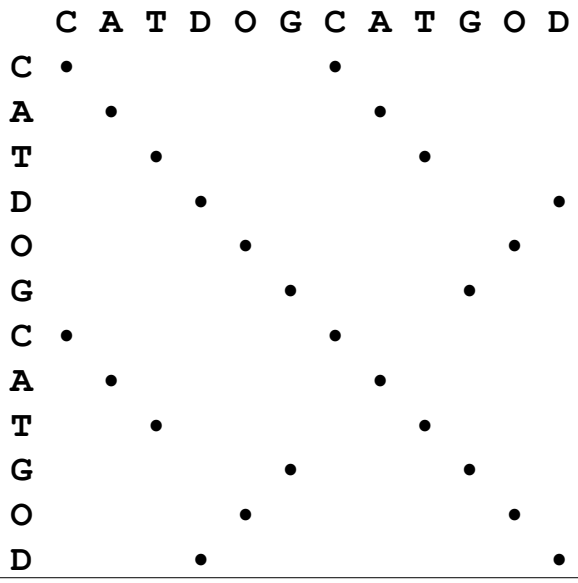
A Fast Pair-wise Alignment and Database Searching Tool

Dot Plot

- Quick detection of high similarity
- Identify internal repeats and inversions of a new sequence
- Use a sliding window to filter out noise from random matches
  - A dot is recorded at window positions where the number of matches is greater than or equal to the stringency
- Global alignment strategy that is also useful for visualizing local matches

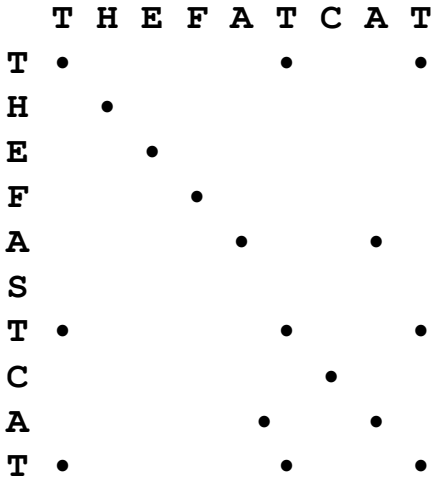
Dot Plot

Identify Internal Repeats and Inversions



Dot Plot

Compare Two Sequences

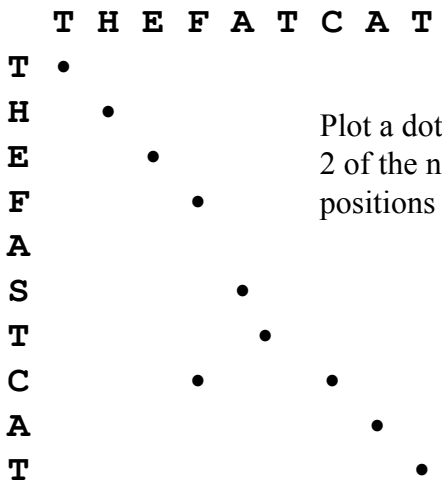


J-M. Claverie and C. Notredame. *Bioinformatics for Dummies*. Wiley Publishing, Inc.: Indianapolis, IN, 2003.

Dot Plot

Sliding Window

Stringency = 2, Window = 3



Plot a dot only if 2 of the next 3 positions are identical.

The BLAST Algorithms

- Original BLAST

  Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic Local Alignment Search Tool. *J. Mol. Biol.*, **215**(3):403-410, 1990.

  - Makes local gapless alignments between sequences

- Gapped BLAST (BLAST 2.0)

  Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a New Generation of Protein Database Search Programs. *Nucleic Acids Res.*, **25**(17):3389-3402, 1997.

  - Allows gaps in alignments
  - 3 times faster than original BLAST

BLAST Example

1. **Divide all database sequences into overlapping constituent words (size w)**

Database

```
123456
NLNYTPW
NL
 LN
  NY
   YT
    TP
     PW
```

w = 2

BLAST Example

2. **Convert every word in the database into a hash score (HS) between 0 and 20^w-1**

For w=2, hash scores are between 0 and 399
HS(AA) = 0 and HS(YY) = 399

A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20

HS(NL) = 20^1 * (Score of N -1) + 20^0 * (Score of L -1)
= 20^1 * (12-1) + 20^0 * (10-1)
= 229

BLAST Example

3. Sort all the hash scores within each database sequence

Before Sorting

words	HS	Location
NL	229	1
LN	191	2
NY	239	3
YT	396	4
TP	332	5
PW	258	6

Database sequence: NLNYTPW

After Sorting

HS	Location
1	-1
2	-1
...	-1
191	2
...	-1
229	1
...	-1
239	3
...	-1
258	6
...	-1
332	5
...	-1
396	4
...	-1

BLAST Example

4. Divide the query sequence into overlapping constituent words and convert them into hash scores

Query	HS
1234567	
QLNFSAGW	
QL	269
LN	191
NF	224
FS	95
SA	300
AG	5
GW	118

BLAST Example

5. Identify all synonyms of the query words

2-letter words that score at least T=8 when aligned with the query 2-letter word using the PAM 120 matrix:

```
  1234567              123456
query: QLNFSAGW; database: NLNYTPW

1 QL QL=11, QM=9, HL=8
2 LN LN=9
3 NF NF=12, AF=8, NY=8, DF=10, QF=8, EF=9, GF=8, HF=10,
     KF=9, SF=9, TF=8, BF=11, ZF=8
4 FS FS=12, FA=9, FN=9, FD=8, FG=9, FP=9, FT=10, YS=8
5 SA none
6 AG AG=8
7 GW AW=13, RW=8, NW=12, DW=12, QW=9, EW=11, GW=17,
HW=8,
     IW=8  KW=9  MW=8  PW=10  SW=13  TW=11  VW=10
```

PAM-120 Substitution Matrix

```
A  3 -3 -1  0 -3 -1  0  1 -3 -1 -3 -2 -2 -4  1  1  1 -7 -4
0
R -3  6 -1 -3 -4  1 -3 -4  1 -2 -4  2 -1 -5 -1 -1 -2  1 -5 -
3
N -1 -1  4  2 -5  0  1  0  2 -2 -4  1 -3 -4 -2  1  0 -4 -2 -
3
D  0 -3  2  5 -7  1  3  0  0 -3 -5 -1 -4 -7 -3  0 -1 -8 -5 -
3
C -3 -4 -5 -7  9 -7 -7 -4 -4 -3 -7 -7 -6 -6 -4  0 -3 -8 -1 -
3
Q -1  1  0  1 -7  6  2 -3  3 -3 -2  0 -1 -6  0 -2 -2 -6 -5 -
3
E  0 -3  1  3 -7  2  5 -1 -1 -3 -4 -1 -3 -7 -2 -1 -2 -8 -5 -
3
G  1 -4  0  0 -4 -3 -1  5 -4 -4 -5 -3 -4 -5 -2  1 -1 -8 -6 -
2
H -3  1  2  0 -4  3 -1 -4  7 -4 -3 -2 -4 -3 -1 -2 -3 -3 -1 -
3
I -1 -2 -2 -3 -3 -3 -3 -4 -4  6  1 -3  1  0 -3 -2  0 -6 -2
3
L -3 -4 -4 -5 -7 -2 -4 -5 -3  1  5 -4  3  0 -3 -4 -3 -3 -2
1
K -2  2  1 -1 -7  0 -1 -3 -2 -3 -4  5  0 -7 -2 -1 -1 -5 -5 -
```

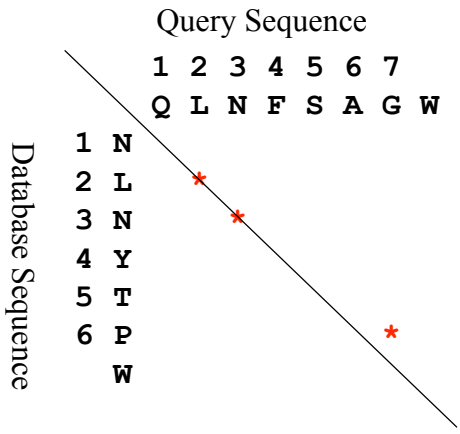
BLAST Example

6. Match the hash scores

Locate the hits by looking up the synonyms of the query words in the sorted database hash table.

HS	words	Location in query	Syn	HS	Location in database	Hits Diagonal
269	QL	1				
191	LN	2	LN	191	2	0
224	NF	3	NY	239	3	0
95	FS	4				
300	SA	5				
5	AG	6				
118	GW	7	PW	258	6	-1

Identify Word Matches



BLAST Example

7. Extension of Hits

Original BLAST: Each hit is extended in both directions until the running alignment's score has dropped more than X below the maximum score yet attained

BLAST 2.0: If two non-overlapping hits are found within distance A of one another on the **same diagonal**, then merge the hits into an alignment and extend the alignment in both directions until the running alignment's score has dropped more than X below the maximum score yet attained

If an extended alignment has a score above S then it is a "high-scoring segment pair" or HSP

Execution Time

- The extension step in the original version of BLAST usually accounts for >90% of the execution time
- Since BLAST 2.0 requires two hits rather than one to invoke an extension, the threshold parameter T must be lowered to retain comparable sensitivity
  - Many more single hits are found but only a small fraction have an associated second hit on the same diagonal that triggers an extension
  - The computation saved by requiring fewer extensions more than offsets the extra computation required to process the larger number of hits

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a New Generation of Protein Database Search Programs. *Nucleic Acids Res.*, **25**(17):3389-3402, 1997.

BLAST Example

8. BLAST 2.0: Evoke a gapped alignment for any HSP exceeding score S_g

- Dynamic Programming is used to find the optimal gapped alignment
- Only alignments that drop in score no more than X_g below the best score yet seen are considered
- A gapped extension takes much longer to execute than an ungapped extension but S_g is chosen so that no more than about one extension is invoked per 50 database sequences
- The resulting gapped alignment is reported only if it has an E-value low enough to be of interest

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a New Generation of Protein Database Search Programs. *Nucleic Acids Res.*, **25**(17):3389-3402, 1997.

Gapped BLAST

Default blastp Parameters

- word size: w = 3
- threshold parameter: T = 11
- window length for extending two hits: A=40
- amino acid substitution matrix: BLOSUM62

BLAST:
Reported Quantities

- **Nominal Score:** The raw score which is the sum of the similarity scores and gap penalties. This is dependent on the query, the database and the scoring scheme.

- **Bit Score S:** Normalized score of the final gapped alignment. This is still dependent on the lengths of the query and the database, but presumably is independent of the scoring scheme.

- **E-value:** Expected number of times of finding such a score (or better) by chance.

- **P-value:** Probability of finding such a score (or better) by chance.

BLAST
Protein Sequences

- **blastp:** compares a protein sequence with a protein database
  - learn something about the structure and function of a protein

- **tblastn:** compares a protein sequence with a nucleotide database
  - discover genes that encode a protein

BLAST
DNA Sequences

- **blastn:** compares a DNA sequence with a DNA database
  - compare very similar DNA sequences
- **tblastx:** compares a translated DNA sequence with a translated DNA database
  - discover new proteins
- **blastx:** compares translated DNA with a protein database
  - analyze the query DNA sequence

BLAST web server at NCBI

http://www.ncbi.nlm.nih.gov/blast/Blast.cgi

The statistics of BLAST scores

Statistical Significance

Equivalent Questions:

- I obtained a score *S* from my pair-wise sequence alignment using BLAST. How significant is this score *S*?
- What is the probability of obtaining a score *S* or better from chance alone?
  - When I align a pair of biological but non-homologous sequences.
  - When I align a pair of shuffled sequences that preserve compositional properties of biological sequences.
  - When I align a pair of sequences that have been computationally generated, based upon a statistical model of DNA or protein sequences.

Empirical Simulations

- Generate many random sequence pairs of the appropriate length and composition
- Calculate the optimal alignment score for each pair using a specific scoring scheme
- If 100 random alignments have score inferior to the alignment of interest, the *P*-value in question is likely less than 0.01.
- However one must take into account multiple testing in database searching. When many alignments are generated between the query sequence and the database sequences, the significance of the best must be discounted accordingly. An alignment with *P*-value 0.0001 in the context of a single trial may be assigned a *P*-value of only 0.1 if it was selected as the best among 1000 independent trials.

Global vs. Local Alignments

- Little is known about the random distribution of optimal global alignment scores.
- The tail behavior of global alignment scores is unknown.
- The statistics for the scores of local alignments (especially ungapped alignment) is well understood.

A Model of Random Sequences

Required Information:

- $\{a_1, a_2, ..., a_r\}$ Amino acid or nucleotide alphabet (r=20 or 4 respectively)
- $\{p_1, p_2, ..., p_r\}$ & $\{p'_1, p'_2, ..., p'_r\}$ The abundances of the two input sequences.
- $\{s_{11}, s_{12}, ...s_{1r}, s_{21}, s_{22}, ..., s_{2r}, ..., s_{r1}, s_{r2}, ..., s_{rr}\}$ s_{ij} is the similarity score between amino acid types a_i and a_j. It can be the elements of a "log-likelihood-ratio" scoring matrix (such as PAM and BLOSUM): $s_{ij} = \log (q_{ij} / p_i p_j)$

Ungapped Local Alignment

- HSP: High-scoring Segment Pairs
- In the limit of sufficiently large sequence lengths ***m*** and ***n***, the statistics of HSP scores are characterized by two parameters, *K* and ***l*** .
- Most simply, the expected number of HSPs with score at least *S* (The E-value of S) is given by the formula:

$$E = Kmn\, e^{-\ddot{e}S}$$

l and K Parameters

l and K can be thought of simply as natural scales for the search space size and the scoring system respectively.

They can be computed from

$\{p_1, p_2, ..., p_r\}, \{p_1', p_2', ..., p_r'\}$ and $\{s_{11}, s_{12}, ..., s_{rr}\}$.

λ is the solution of the equation $\sum_{i=1}^{r}\sum_{j=1}^{r} p_i p_j e^{\lambda s_{ij}} = 1$

The way to compute K can be found in Karlin & Altschul (1990), PNAS, 87:2264

Constraints on $\{s_{ij}\}$

- At least one of the s_{ij} is positive. Otherwise you will end up with (on average) 0-length HSP.
- The expected score for aligning a random pair of amino acid is required to be negative. Were this not the case, long alignments would tend to have high score independently of whether the segments aligned were related, and the statistical theory would break down.

$$\sum_{i=1}^{r} p_i p_j s_{ij} < 0 \qquad \text{For protein sequences, } r = 20$$

- Log-likelihood-ratio scores naturally satisfy the above two constraints.

Bit Score

Bit score S' (in bits) subsumes the statistical essence of the scoring system employed, so that to calculate significance one needs to know in addition only the size of the search space

$$S' = \frac{\lambda S - \ln K}{\ln 2}$$

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

The Number of Random HSPs

The number of HSPs with score $\geq S$ is approximately Poisson distributed, with mean as the E-value of S.

$$P(a \text{ HSPs}) = \frac{e^{-E} (E)^a}{a!}$$

Specifically the chance of finding zero HSP with score $\geq S$ (a=0) is e^{-E}.
So the probability of finding at least one such HSP is
$P = 1 - e^{-E}$
This is the P-value associated with the score S.

Extreme Value Distribution (EVD)

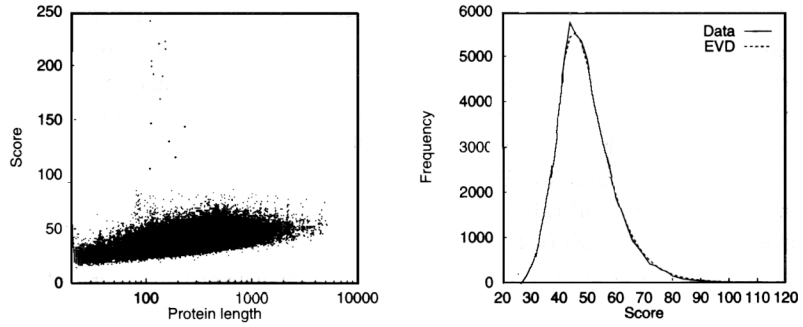


Figure 2.13 *Left, a scatter plot of the distribution of local match scores obtained from comparing human cytochrome C (*SWISS-PROT *accession code P000001) against the* SWISS-PROT34 *protein database with the Smith–Waterman implementation* SSEARCH *[Pearson 1996]. Right, the corresponding length-normalised distribution of scores, showing the fit to an EVD distribution.*

The Extreme Value Distribution

$P=1-e^{-Kmn\, e^{-\lambda S}}$

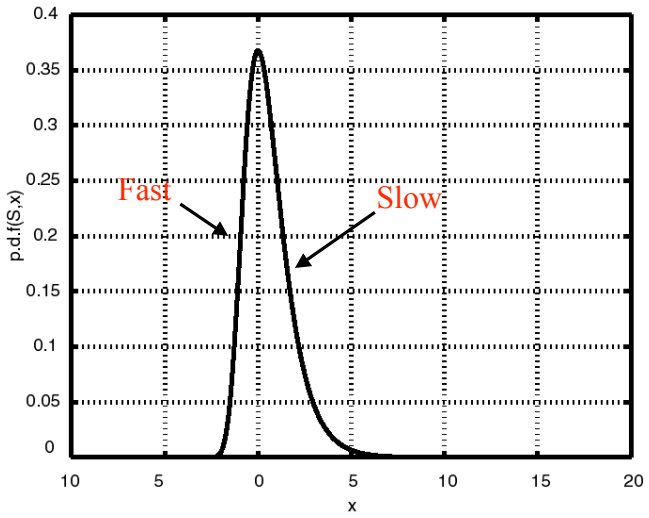
Set $X=\lambda S-\ln(Kmn)$ to simplify the equation:

$$f(x) = p.d.f(X,x) = e^{-x}e^{-e^{-x}}$$

$$F(x) = c.d.f(X,x) = P(X \le x) = e^{-e^{-x}}$$

=> Gumbel distribution

The Extreme Value Distribution



EVD for large x

For large value of x:

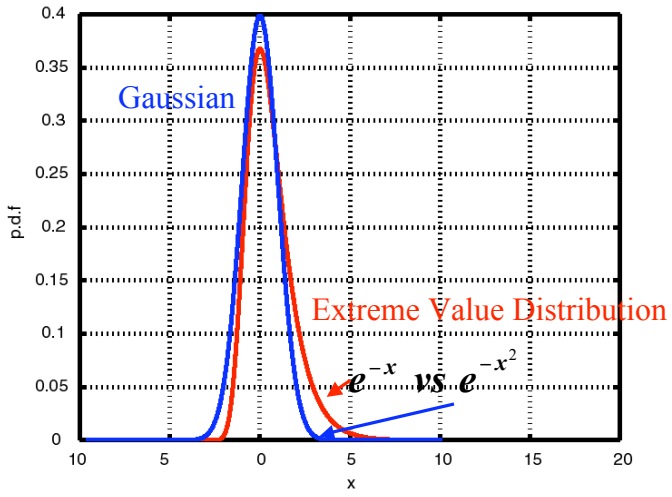
$$p.d.f(X,x) = e^{-x}e^{-e^{-x}}$$

$$\approx e^{-x}(1 - e^{-x} + \frac{e^{-2x}}{2!} - \frac{e^{-3x}}{3!} + ...) \approx e^{-x}$$

Compare to a Gaussian Distribution with mean 0 and standard deviation 1:

$$p.d.f(T,x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$$

For large x: EVD vs. Gaussian



Example

Question: When we align two protein sequences of lengths m and n respectively, how high does the HSP score *S* need to be in order to achieve the 0.01-level significance?

Solution:

$1-e^{-Kmn\, e^{-\lambda S}} = 0.01$

To solve for S, we must know K, λ, m and n

Database Searches

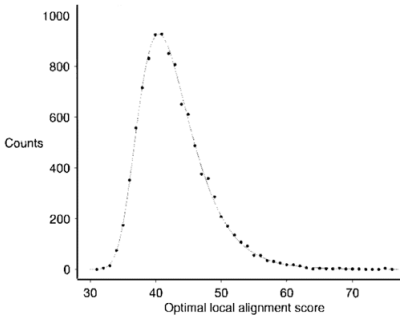
- If we assume the *a priori* chance of relatedness is proportional to sequence length, then the pairwise *E*-value involving a database sequence of length *n* should be multiplied by *N/n*, where *N* is the total length of the database in residues.
- This can be accomplished simply by treating the database as a single long sequence of length *N*.

Gapped Local Alignments

- For ungapped alignments, the statistical parameters can be calculated, using analytic formulas, from the substitution scores and the background residue frequencies of the sequences being compared (as explained in the previous lecture).
- For gapped alignments, these parameters must be estimated from a large-scale comparison of "random" sequences (i.i.d. of background residue frequencies).
- BLAST uses a random sequence model and pre-computes the λ and K parameters for a selected set of substitution matrices and gap costs by fitting to the Extreme Value Distribution (EVD).

l and K for Gapped Local Alignments

Using BLOSUM-62 amino acid substitution scores and affine gap costs in which a gap of length g is assigned a score of -(10 + g), 10,000 pairs of length-1000 random protein sequences are generated and the Smith-Waterman algorithm is used to calculate 10,000 optimal local alignment scores. From these scores, l is estimated at 0.252 and K at 0.035. A plot of local alignment scores and the fitted EVD curve is shown on the right.



http://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-3.html

Erdos-Renyi Law for the Longest Run of Hs

- A N-amino-acid-long random sequence is made up of two types of amino acids: hydrophobic (H) and polar (P). The occurrence of Hs or Ps at each position of the sequence is random, with probabilities P(H) = p and P(P) = 1-p.
- What is the expected length L of the longest run of Hs in this random sequence?

An Intuitive Explanation for the Expected Length of Longest Run of Hs

HPPHHHPHHPHHHPPPHHHHHHHHPPPPHHP

Randomly pick a position in the random sequence (N=31 in the above example). The probability that L residues (L=5 in the above example) from that point on are all Hs is: p^L. There are (N-(L-1)) ways such a picking can be done, therefore the frequency of observing such a run is p^L *(N-(L-1)). A theory (the Erdos-Renyi law) indicates when a frequency is a small value, it can be treated as a probability. Since a probability can at most be 1, by setting p^L *(N-(L-1)) to 1, we solve L to be:

$$L=\log_{\frac{1}{p}} N \qquad \text{For a large N}$$

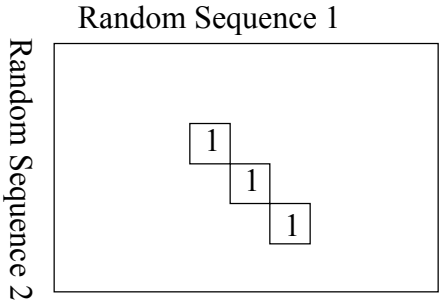
The Erdos-Renyi Law

$$P(\lim_{N\to\infty} \frac{L}{\log_{\frac{1}{p}} N}=1)=1$$

P( ): The probability of an event

Or: $L=\lim_{N\to\infty} (\log_{\frac{1}{p}} N)$

The Expected Length L of the Longest Run of Matches between Two Random Sequences



Let us make a dot plot. A match has a score of 1 and a mismatch has a score of 0. The probability of having a single match is p. A run of 3 consecutive matches is shown in the above plot.

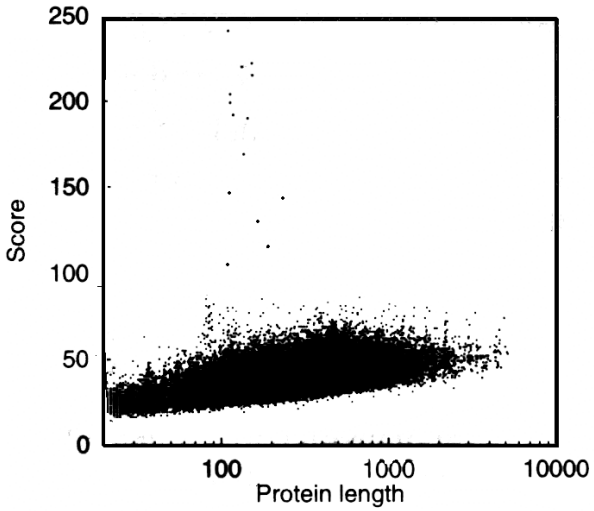
The Expected Length L of the Longest Run of Matches between Two Random Sequences

If the lengths of the two sequences are N_1 and N_2, the frequency of a L-long match is $p^L*(N_1-(L-1))*(N_2-(L-1))$. By setting it to 1, we obtain:

$$L = \log_{\frac{1}{p}}(N_1 * N_2) \qquad \text{For large } N_1, N_2$$

Length & Alignment Score

Local alignment scores of human cytochrome C against the Swiss-Prot database



References

- Karlin, S., et al., 1983, PNAS, 80: 5660-5664
- Karlin, S. & Altschul, SF. 1990, PNAS, 87:2264-2268
- Altschul, SF. 1991, JMB, 219:555-565
- Karlin, S. & Altschul, SF. 1993, PNAS, 90:5873-5877
- Altschul, SF et al., 1994, Nature Genetics, 6:119-129

Acknowledgement

- The slides on blast are due to Zhiping Weng.