Local Alignment Statistics

Stephen Altschul

Central Issues in Biological Sequence Comparison

Definitions: What is one trying to find or optimize?

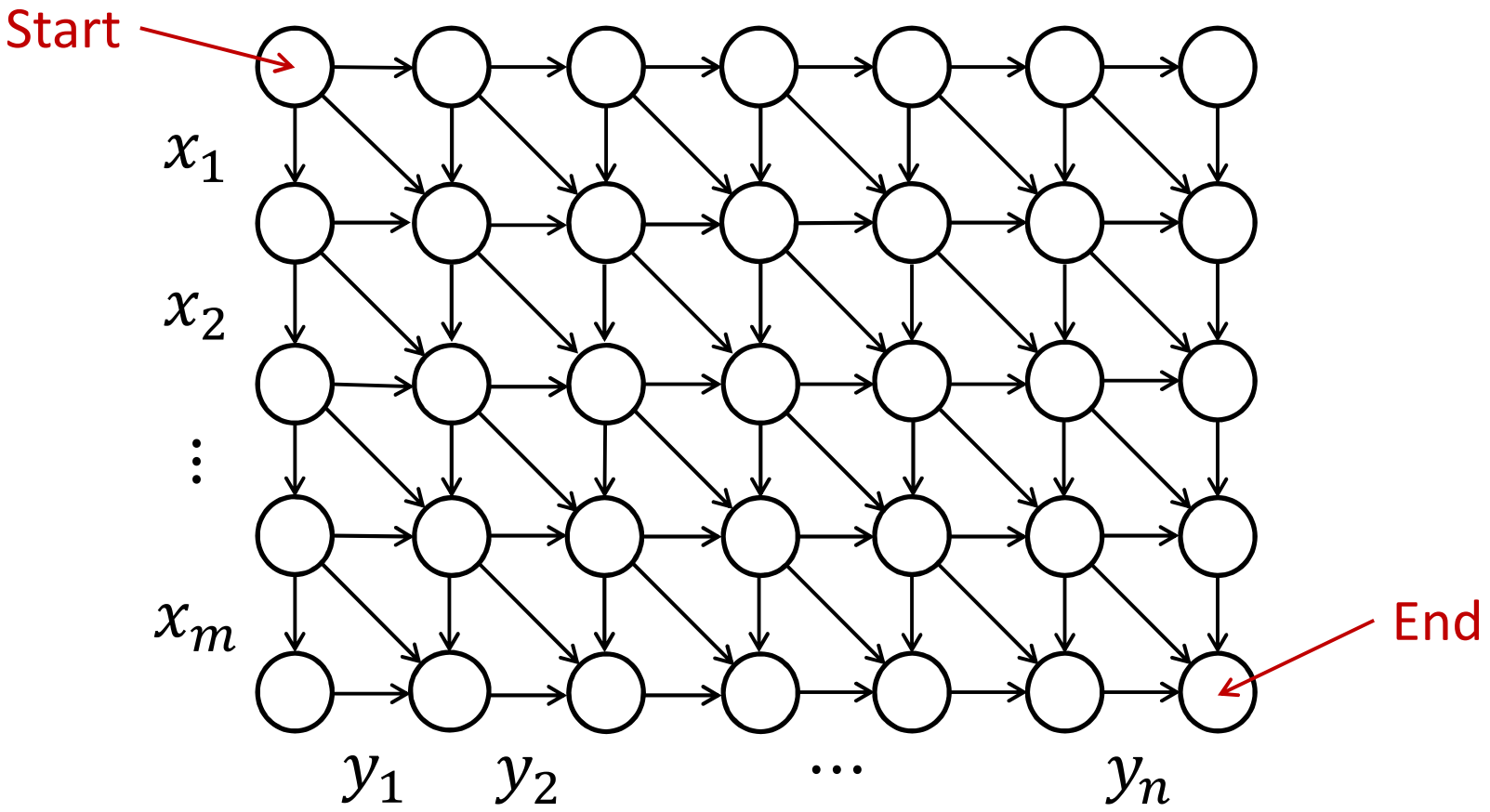
Algorithms: Can one find the proposed object optimally or in reasonable time optimize?

Statistics: Can one's result be explained by chance?

In general there is a tension between questions. A definition that is too simple may allow efficient algorithms, but may not yield results of biological interest. However, a definition that includes most of the relevant biology may entail intractable algorithms and statistics. The most successful approaches find a balance between these considerations.

Path graphs

A global alignment may be viewed as a path through a directed *path graph* which begins at the upper left corner and ends at the lower right. Diagonal steps correspond to substitutions, while horizontal or vertical steps correspond to indels. Scores are associated with each edge, and the score of an alignment is the sum of the scores of the edges it traverses. Each alignment corresponds to a unique path, and vice versa.



Ungapped Local Alignments

When two sequences are compared, how great are the local alignment scores that can be expected to arise purely by chance? In other words, when can a local alignment be considered statistically significant? We will first develop the theory for local alignments *without gaps*.

Our simplified model of chance: The various amino acids occur randomly and independently with the respective *background probabilities*

$$p_1 , \ p_2 , \ \ldots , \ p_i , \ \ldots , \ p_{20}.$$

Our scoring system: The substitution score for aligning amino acids i and j is $s_{i,j}$. A substitution *score matrix* then consists of the scores

$$s_{1,1} , \ s_{1,2} , \ \ldots , \ s_{i,j} , \ \ldots , \ s_{20,20}.$$

The BLOSUM-62 Substitution Score Matrix

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4																			
R	−1	5																		
N	−2	0	6																	
D	−2	−2	1	6																
C	0	−3	−3	−3	9															
Q	−1	1	0	0	−3	5														
E	−1	0	0	2	−4	2	5													
G	0	−2	0	−1	−3	−2	−2	6												
H	−2	0	1	−1	−3	0	0	−2	8											
I	−1	−3	−3	−3	−1	−3	−3	−4	−3	4										
L	−1	−2	−3	−4	−1	−2	−3	−4	−3	2	4									
K	−1	2	0	−1	−3	1	1	−2	−1	−3	−2	5								
M	−1	−1	−2	−3	−1	0	−2	−3	−2	1	2	−1	5							
F	−2	−3	−3	−3	−2	−3	−3	−3	−1	0	0	−3	0	6						
P	−1	−2	−2	−1	−3	−1	−1	−2	−2	−3	−3	−1	−2	−4	7					
S	1	−1	1	0	−1	0	0	0	−1	−2	−2	0	−1	−2	−1	4				
T	0	−1	0	−1	−1	−1	−1	−2	−2	−1	−1	−1	−1	−2	−1	1	5			
W	−3	−3	−4	−4	−2	−2	−3	−2	−2	−3	−2	−3	−1	1	−4	−3	−2	11		
Y	−2	−2	−2	−3	−2	−1	−2	−3	2	−1	−1	−2	−1	3	−3	−2	−2	2	7	
V	0	−3	−3	−3	−1	−2	−2	−3	−3	3	1	−2	1	−1	−2	−2	0	−3	−1	4

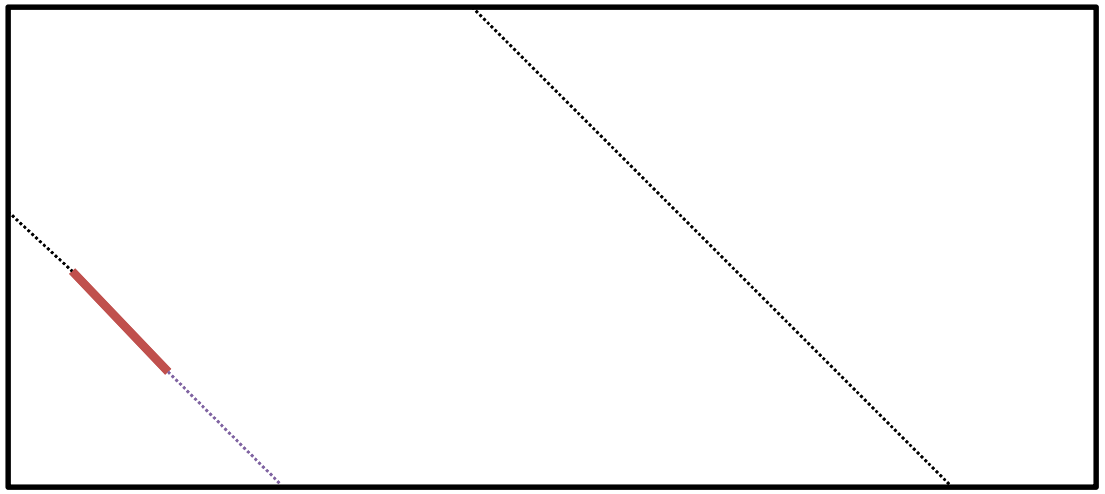
Henikoff, S. & Henikoff, J.G. (1992) *Proc. Natl. Acad. Sci. USA* **89**:10915-10919.

Negative Expected Score

Score matrices used to seek local alignments of variable length should have a negative expected score:

$$\sum_{i,j} p_i p_j s_{i,j} < 0.$$

Otherwise, alignments representing true homologies will tend to be extended with biologically meaningless noise:



Log-odds Scores

The scores of *any* substitution matrix (with a negative expected value and at least one positive score) can be written in the form

$$s_{i,j} = \left(\ln \frac{q_{i,j}}{p_i p_j} \right) / \lambda = \log \frac{q_{i,j}}{p_i p_j}$$

where λ is a positive *scale parameter*, and the $q_{i,j}$ are a set of positive numbers that sum to 1, called the *target frequencies* for aligned amino acid pairs. Conversely, a non-trivial matrix constructed in this way will have a negative expected value and at least one positive score.

Karlin, S. & Altschul, S.F. (1990) *Proc. Natl. Acad. Sci. USA* **87**:2264-2268.

Proof

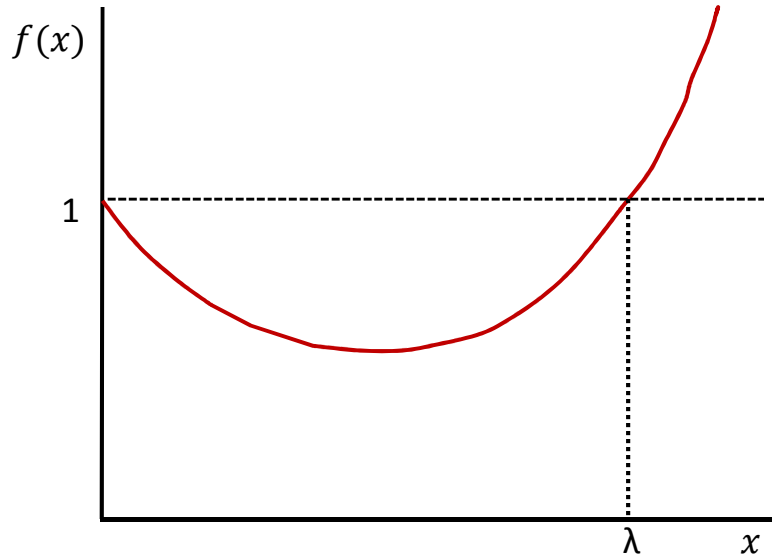
Define $f(x) = \sum_{i,j} p_i p_j e^{s_{i,j}x}$.

Then $f(0) = \sum_{i,j} p_i p_j = 1$.

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

and $f''(x) = \sum_{i,j} p_i p_j {s_{i,j}}^2 e^{s_{i,j}x} > 0$.

In addition, because at least one $s_{i,j}$ is positive, $f(x)$ diverges for large x.



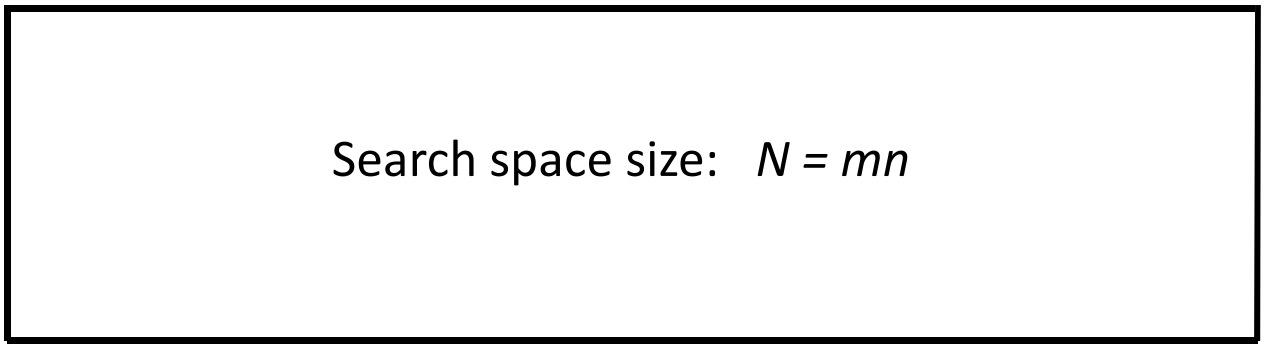
These facts imply that $f(x) = 1$ has a unique positive solution λ, which is easily calculated. Now define $q_{i,j} = p_i p_j e^{\lambda s_{i,j}}$. It is clear that all the $q_{i,j}$ are positive, and furthermore that they sum to 1, because $\sum_{i,j} q_{i,j} = f(\lambda) = 1$.

Finally, solving for $s_{i,j}$ yields: $s_{i,j} = \left(\ln \frac{q_{i,j}}{p_i p_j}\right) / \lambda$.

Search Space Size

Subject sequence (or database) length: *n* residues

Query sequence length: *m* residues



Question: Given a particular scoring system, how many *distinct* local alignments with score ≥ *S* can one expect to find by chance from the comparison of two random sequence of lengths *m* and *n*? The answer, *E(S,m,n),* should depend upon *S,* and the lengths of the sequences compared.

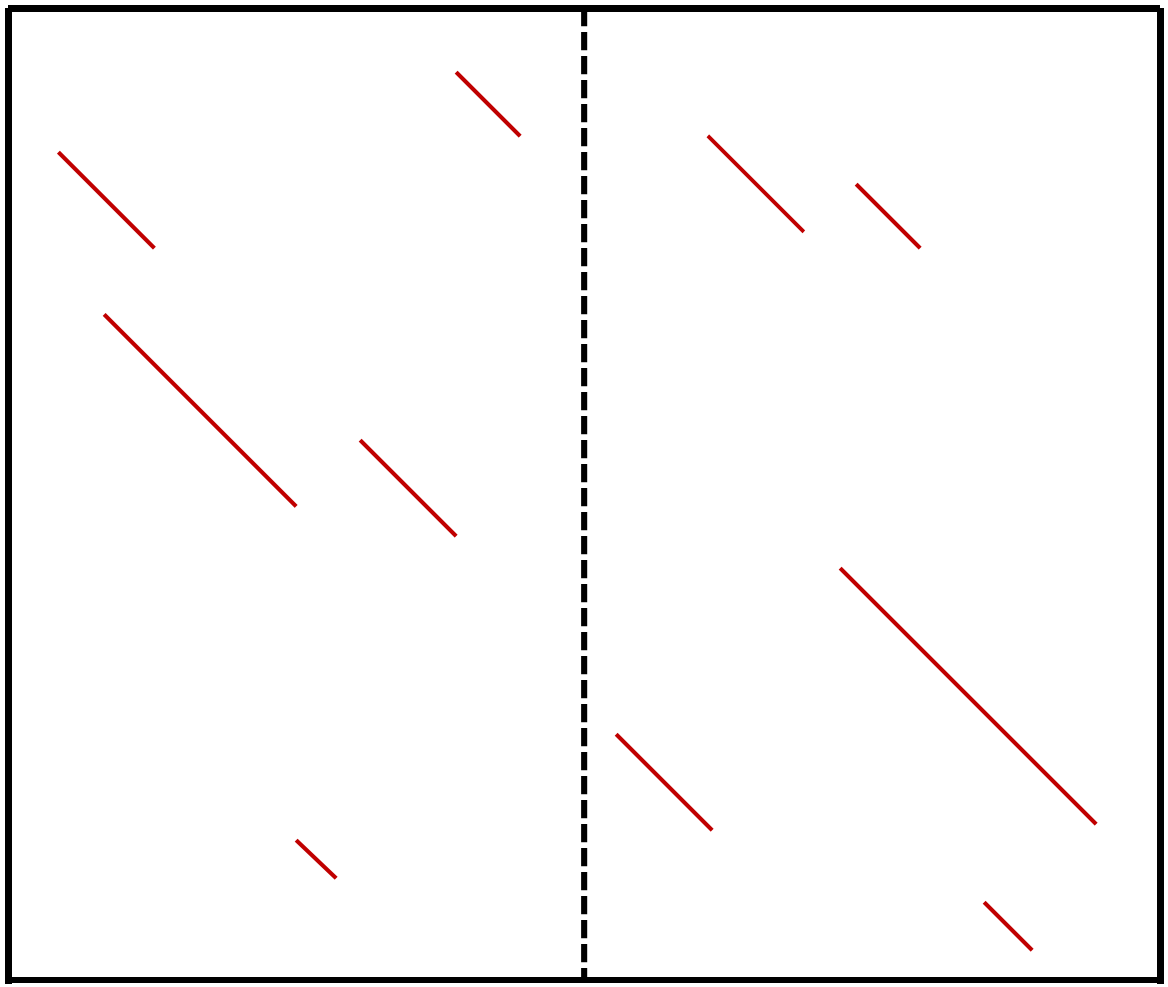
Note: We may define two local alignments to be distinct if they do not align any residue pairs in common. Thus, the slight trimming or extension of a high-scoring local alignment does not yield a distinct high-scoring local alignment.

The Number of Random High-scoring Alignments Should be Proportional to the Search Space Size

Doubling the size of the search space, i.e. by doubling the length of one sequence, should result in approximately twice as many random high-scoring alignments.

Doubling the length of both sequence should yield about four times as many random high-scoring alignments.

In other words, in the limit of large m and n, $E(S, m, n) \propto mn$.



The Number of Random Alignments with Score $\geq S$ Should Decrease Exponentially with S.

Consider a series of coin flips: **HHHTTHTTHTTTTTHHHTH . . . .**
The probability that it begins with a run of $\geq h$ heads is $(\frac{1}{2})^h = e^{-(\ln 2)h}$.

A substitution matrix with scores $+1$ along the main diagonal, and scores $-\infty$ off the main diagonal, yields as its high-scoring alignments runs of exact matches. If the probability of a match is p, then the probability that, starting at a particular position in each sequence, there are $\geq h$ matches is $p^h = e^{-(\ln \frac{1}{p})\, h}$.

Given an arbitrary scoring system, and for local alignments starting at an arbitrary position in each of two sequence, the probability that the highest scoring score is $\geq S$ should decrease exponentially with S.

This can be understood to imply that, for some positive parameter a, $E(S, m, n) \propto e^{-aS}$.

The Expected Number of High-Scoring Alignments

From the comparison of two random sequences of lengths m and n, the *expected number* of distinct local alignments with raw score at least S is approximately

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

where K is a calculable positive parameter which, like λ, is dependent on the substitution matrix and background letter frequencies. This is called the *E-value* associated with the score S.

The number of such high-scoring alignments is Poisson distributed, with expected value E, so the probability of finding 0 alignments with score $\geq S$ is e^{-E}. Thus the probability of finding *at least one* alignment with score $\geq S$ is

$$p = 1 - e^{-E}.$$

This is called the *p-value* associated with S. When $E \leq 0.1, p \approx E$.

Karlin, S. & Altschul, S.F. (1990) *Proc. Natl. Acad. Sci. USA* **87**:2264-2268.

Dembo, A., Karlin, S. & Zeitouni, O. (1994) *Ann. Prob.* **22**:2022-2039.

Normalized Scores

To know the *E*-value associated with a score, one needs to know the relevant statistical parameters λ and K. However, these parameters may be folded into the score using the equation

$$S' = (\lambda S - \ln K) \,/\, \ln 2$$

to yield a *normalized score* S', expressed in *bits*. When this is done, the formula for the *E*-value reduces to the extremely simple

$$E = N \,/\, 2^{S'}.$$

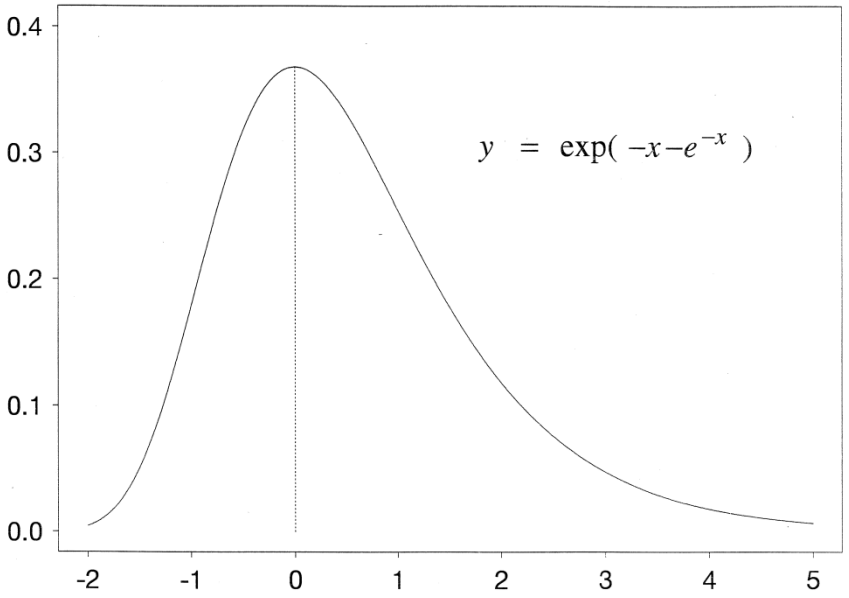
Example: Comparing a protein sequence of length 250 residues to a database of length one billion residues, how many local alignments with normalized score ≥ 35 bits can one expect to find by chance? The search space size is approximately $2^8 \times 2^{30} = 2^{38}$, so $E \approx 2^{38} / 2^{35} = 2^3 = 8$. The number of alignments with score ≥ 45 bits one can expect to find by chance is 0.008.

Sidelight: The Extreme Value Distribution

Almost all the relevant statistics for local alignment scores can be understood in terms of E-values. However, sometimes people are interested instead in the *distribution of optimal scores* from the comparison of two random sequences.

Analysis of the p-values described above shows that the distribution of these scores follows an *extreme value distribution* (e.v.d.). Just as the *sum* of a large number of independent random variables tends to follow a normal distribution, so the *maximum* of a large number of independent random variables tends to follow an e.v.d.

Like the normal distribution, the e.v.d. has two parameters which describe its offset and spread. However, it is easiest to describe an e.v.d. not by its mean and standard deviation but rather by its "characteristic value" u, and "scale" λ. In brief, the probability density function of an e.v.d. is given by $\exp[-\lambda(x-u)-e^{-\lambda(x-u)}]$. A graph of the density of the standard e.v.d., with $u=0$ and $\lambda=1$, is shown here.



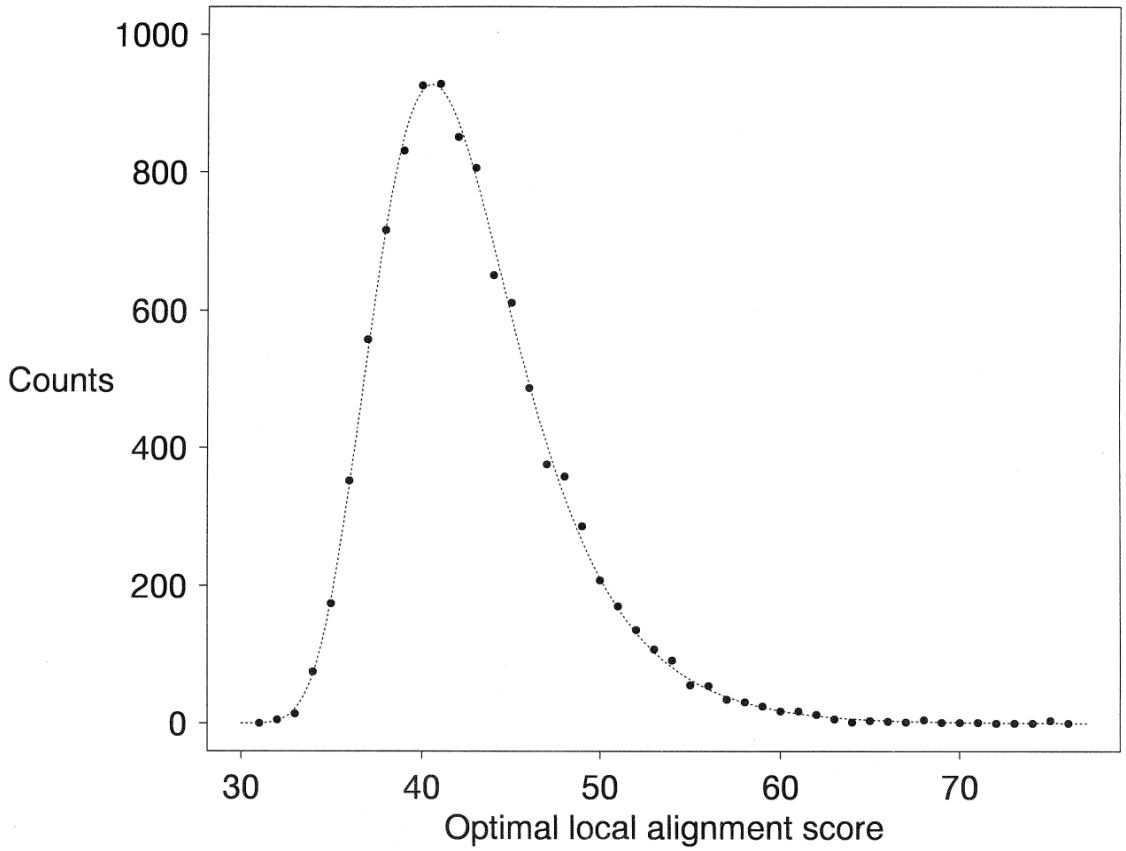
For optimal local alignment scores, the scale parameter of the e.v.d. is equivalent to the statistical parameter λ discussed previously. The characteristic value u is the score whose E-value is 1, and is given by $u=(\ln Kmn)/\lambda$.

Gap Costs for Local Alignment

Our statistical theory is provably valid only for local alignments *without gaps*. However, although no formal proof is available, random simulation suggests the theory remains valid when gaps are allowed, with sufficiently large gap costs.

In this case, no analytic formulas for the statistical parameters λ and K are available, but these parameters may be estimated by random simulation.

Here, 10,000 pairs of "random" protein sequences, each of length 1000, are compared using the BLOSUM-62 substitution scores, in conjunction with gap scores of $-11-k$ for a gap of length k. A histogram of the optimal local alignment scores from all comparisons is shown, as is the maximum-likelihood extreme value distribution fit to these scores. The estimated statistical parameters are $\lambda \approx 0.27$ and $K \approx 0.04$.



Local Alignment Substitution Matrices Are Log-Odds Matrices

The scores of *any* local substitution matrix can be written in the form:

$$s_{i,j} = \log \frac{q_{i,j}}{p_i p_j}$$

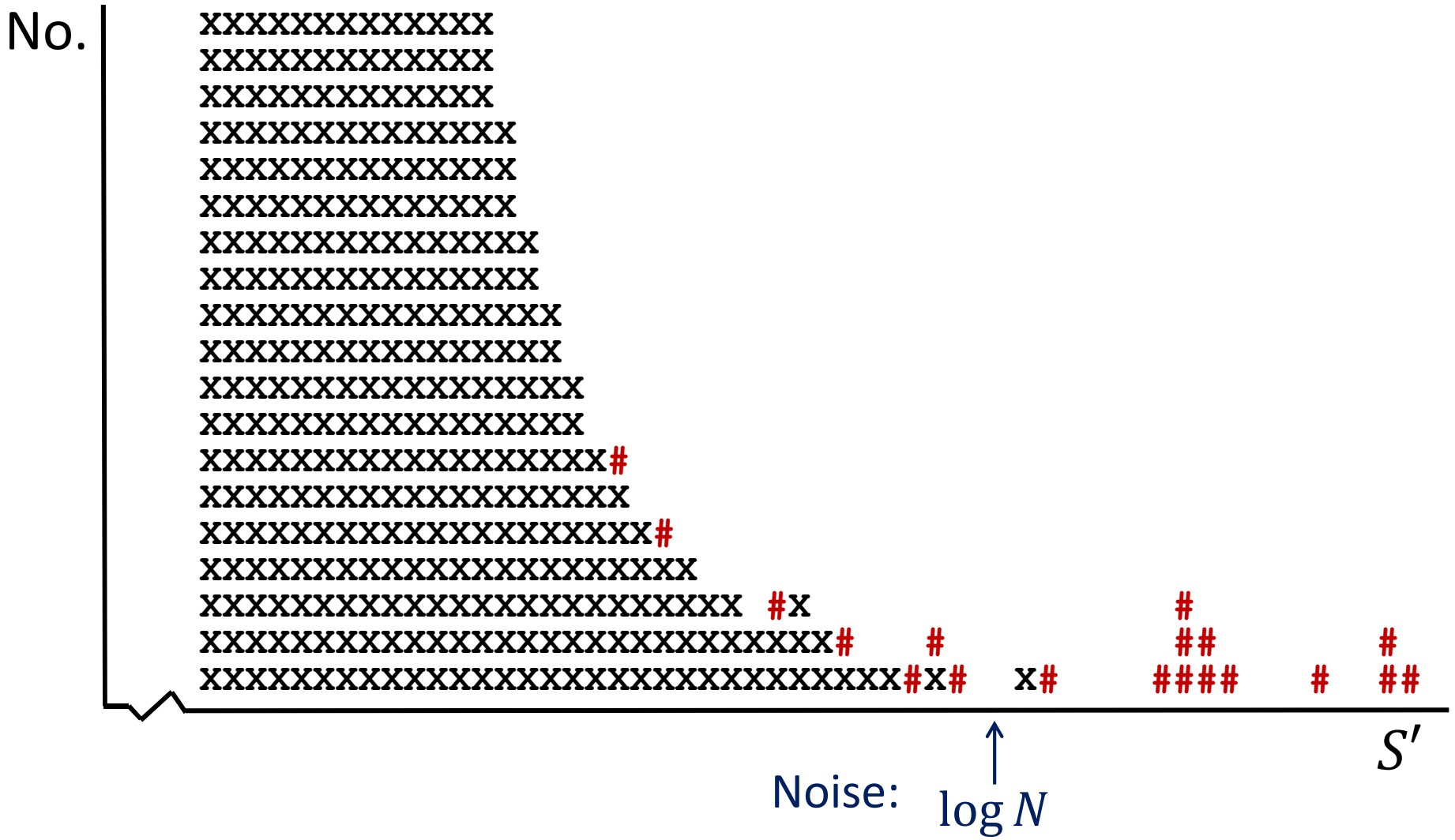
where the $q_{i,j}$ are target frequencies for the aligned amino acid pairs.

Question: What is the optimal way to choose these target frequencies?

Karlin, S. & Altschul, S.F. (1990) "Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes." *Proc. Natl. Acad. Sci. USA* **87**:2264-2268.

Altschul, S.F. (1991) "Amino acid substitution matrices from an information theoretic perspective." *J. Mol. Biol.* **219**:555-565.

A Schematic Database Search



Optimal Target Frequencies

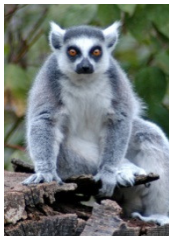
If the aligned pairs of amino acids within the set of *true alignments* occur with average frequencies $q_{i,j}$, then the normalized scores of these alignments will tend to be maximized by substitution scores that have the $q_{i,j}$ as target frequencies.

Selecting an optimal substitution matrix reduces to estimating the $q_{i,j}$ that characterize true alignments.

Karlin, S. & Altschul, S.F. (1990) *Proc. Natl. Acad. Sci. USA* **87**:2264-2268.

Altschul, S.F. (1991) *J. Mol. Biol.* **219**:555-565.

Alignments of Human Beta-Globin to Other Globins



```
Human beta-globin                 VHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGN
                                    LTPEE   VT LWGKVNV  VGGEALGRLLVVYPWTQRFFESFGDLS PDA MGN
Ring-tailed lemur beta-globin     TFLTPEENGHVTSLWGKVNVEKVGGEALGRLLVVYPWTQRFFESFGDLSSPDA MGN

PKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH
PKVKAHGKKVL AFS GL HLDNLKGTFA LSELHC  LHVDPENF LLGNVLV VLAHHFG  F P  QAA QKVV GVANALAHKYH
PKVKAHGKKVLSAFSEGLHHLDNLKGTFAQLSELHCVALHVDPENFKLLGNVLVIVLAHHFGNDFSPQTQAAFQKVVIGVANALAHKYH
```

```
Human beta-globin                 VHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNP
                                  V  T   E SA   LWGK N DE G  AL R L VYPWTQR F  FG LS P A MGNP
Goldfish beta-globin              VEWTDAERSAIIGLWGKLNPDELGPQALARCLIVYPWTQRYFATFGNLSSPAAIMGNP

KVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFG-KEFTPPVQAAYQKVVAGVANALAHKYH
KV AHG  V G          DN K T A LS  H  KLHVDP NFRLL        A  FG   F   VQ A QK    V  AL   YH
KVAAHGRTVMGGLERAIKNMDNIKATYAPLSVMHSEKLHVDPDNFRLLADCITVCAAMKFGPSGFNADVQEAWQKFLSVVVSALCRQYH
```

```
Human beta-globin           VHLTPEEKSAVTALW----GKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKA
                              L       V   W    G  N   VG E    L         F  F   S        P V
Bloodworm globin IV         MGLSAAQRQVVASTWKDIAGSDNGAGVGKECFTKFLSAHHDIAAVF-GFSGAS-------DPGVAD

HGKKVLGAFSDGLAHL-DNLKGTFATLSELHCDK----LHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH
 G KVL    D   HL D  K             K     H   E F  LG  L     H  G   T     A     A    AL
LGAKVLAQIGVAVSHLGDEGKMVAEMKAVGVRHKGYGYKHIKAEYFEPLGASLLSAMEHRIGGKMTAAAKDAWAAAYADISGALISGLQ
```

```
Human beta-globin             VHLTPEEKSAVTALWG--KVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK
                              V  T      V       K N            L   P     F        P     NPK
Soybean leghemoglobin         VAFTEKQDALVSSSFEAFKANIPQYSVVFYTSILEKAPAAKDLFSFLANGVDPT----NPK

VKAHGKKVLGAFSDGLAHLDNLKGTFA--TLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH
   H  K      D    L       A   L   H  K   DP  F      L        G         A        A A
LTGHAEKLFALVRDSAGQLKASGTVVADAALGSVHAQKAVTDPQ-FVVVKEALLKTIKAAVGDKWSDELSREWEVAYDELAAAIKKA--
```

The PAM Model of Protein Evolution: A Summary

A Markov model of protein evolution: during a given period of time, amino acid i has the probability $p_{i\to j}$ of mutating into amino acid j.

"1 PAM" of evolution corresponds to a single substitution, on average, per 100 amino acids.

The substitution probabilities $p_{i\to j}$ corresponding to 1 PAM of evolution are derived from the analysis of a large number of accurately aligned, homologous proteins that are ≥ 85% identical. By construction, $p_i p_{i\to j} = p_j p_{j\to i}$, although there is no biological reason this need be the case.

Given the $p_{i\to j}$ for 1 PAM, one may infer by matrix multiplication the $p_{i\to j}$ for any PAM distance, and therefore the probability $q_{i,j} = p_i p_{i\to j}$ of amino acid i corresponding to amino acid j in accurately aligned, homologous proteins diverged by this amount of evolution.

The PAM score for aligning amino acids i and j is $s_{i,j} = \log\frac{q_{i,j}}{p_i p_j} = \log\frac{p_{i\to j}}{p_j}$. By the construction of the asymmetric $p_{i\to j}$, the target frequencies $q_{i,j}$ and scores $s_{i,j}$ are symmetric.

The amino acid at a given position may mutate multiple times, and perhaps return to the original residue. Thus 100 PAMs actually corresponds to proteins that are about 43% identical, while 250 PAMs corresponds to proteins that are about 20% identical.

There is no uniform scale relating PAM distance to evolutionary time, because different protein families can evolve at greatly differing rates.

Dayhoff, M.O., Schwartz, R.M. & Orcutt, B.C. (1978) "A model of evolutionary change in proteins." In *Atlas of Protein Sequence and Structure, vol. 5, suppl. 3,* M.O. Dayhoff (ed.), pp. 345-352, Natl. Biomed. Res. Found., Washington, DC.

The BLOSUM Substitution Matrices

One criticism of the PAM matrices is that their extrapolation of substitution probabilities to distantly related proteins may be inaccurate.

In 1992, the Henikoffs proposed mitigating this problem by estimating the target frequencies $q_{i,j}$ directly from alignments of distantly related proteins.

A challenge for this approach is obtaining accurate alignments.

The Henikoffs considered only conserved "blocks" from alignments involving multiple protein sequences. The additional information available from multiple related proteins permits the accurate alignment even of sequences that are greatly diverged.

Varying degrees of divergence are dealt with by clustering sequences that are more than a given percentage identical, and counting substitutions only between distinct clusters, not within them. The widely-used BLOSUM-62 matrix clusters sequences that are ≥ 62% identical, and is roughly equivalent to the PAM-180 matrix.

Somewhat confusingly, the numbers for the PAM and BLOSUM matrices run in opposite directions. Specifically, low-number PAM matrices but high-number BLOSUM matrices are tailored for closely related proteins.

Henikoff, S. & Henikoff, J.G. (1992) "Amino acid substitution matrices from protein blocks." *Proc. Natl. Acad. Sci. USA* **89**:10915-10919.

Relative Entropy: The Expected Per-Position Alignment Score for Related Sequences

Consider an accurate alignment of two related sequences that are diverged by a known amount, so that the appropriate target frequencies and substitution scores are also known.

One may ask the question: *What is the expected substitution score per position?*

It is easy to write down a formula for this quantity:

$$H = \sum_{i,j} q_{i,j}\, s_{i,j} = \sum_{i,j} q_{i,j} \log \frac{q_{i,j}}{p_i p_j}.$$

If the scores $s_{i,j}$ are expressed in bits, then H too has the unit of bits.

H is a well-known quantity from information theory, called the *relative entropy* of the probability distributions $q_{i,j}$ and $p_i p_j$, but in the present context it has the simple interpretation given above.

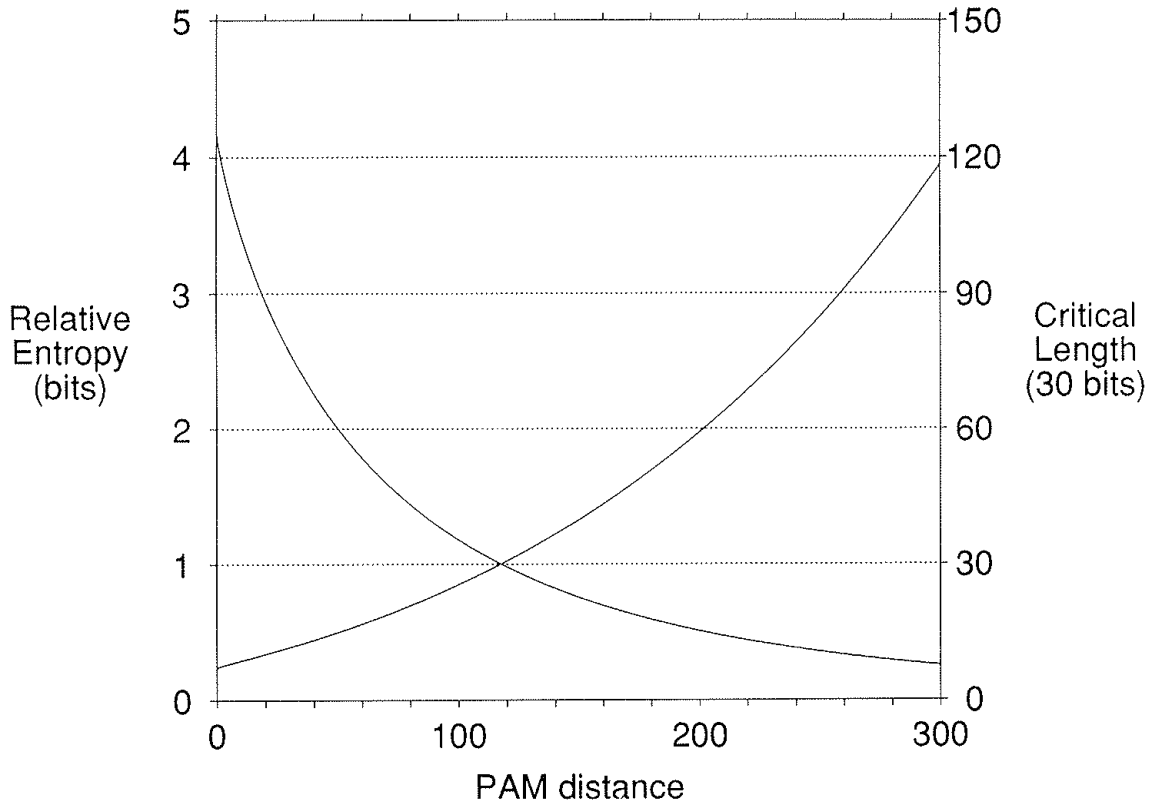
It is possible to show that H must always be positive, in contrast to the expected per-position alignment score for *unrelated* sequences, which we require to be negative.

Altschul, S.F. (1991) *J. Mol. Biol.* **219**:555-565.

Relative Entropy as a Function of PAM Distance

Given the PAM model of protein evolution, it is easy to calculate the relative entropy of a PAM substitution matrix as a function of PAM distance; the curve is shown here. The further sequences diverge, the less information one can expect to obtain per position.

If one requires a certain alignment score to rise above background noise, one can then calculate the minimum length required, on average, for an alignment to achieve this score. The graph here shows such critical lengths for assumed background noise of 30 bits.

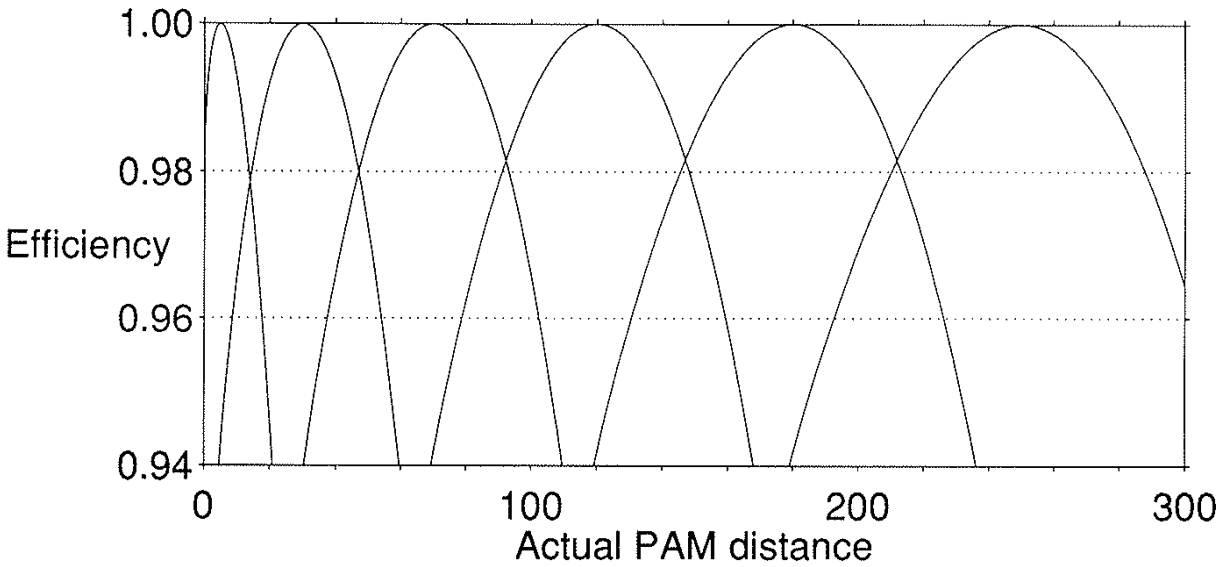


Substitution Matrix Efficiency

In general one does not know *a priori* the evolutionary distance separating two sequences, so one has to use a matrix that may not be optimal. How much information is lost by using the wrong matrix?

One may define the *efficiency* of the PAM-y matrix at PAM distance x as: $\left(\sum_{i,j} {q_{i,j}}^{x} {s_{i,j}}^{y}\right) / \left(\sum_{i,j} {q_{i,j}}^{x} {s_{i,j}}^{x}\right)$.

The graph shows efficiency curves for the PAM-5, PAM-30, PAM-70, PAM-120, PAM-180 and PAM-250 matrices. Each curve has maximum value 1.0 at its corresponding PAM distance.



DNA Substitution Scores

One may extend the PAM model to DNA sequences. Assuming uniform nucleotide frequencies and uniform substitution rates one may derive the PAM scores shown here.

One may assume an alternative substitution model in which transitions (A ↔ G and C ↔ T) are more likely than transversions. This implies mismatch scores that depend upon whether the mismatch is a transition or a transversion. It also implies ungapped relative entropies that differ from those shown here. The next slide assume such an alternative model.

PAM Distance	Percent Conserved	Match Score (bits)	Mismatch Score (bits)	Absolute Score Ratio	Relative Entropy (bits)
0	100.0	2.00	-∞	0.00	2.00
1	99.0	1.99	-6.24	0.32	1.90
2	98.0	1.97	-5.25	0.38	1.83
5	95.2	1.93	-3.95	0.49	1.64
10	90.6	1.86	-3.00	0.62	1.40
15	86.4	1.79	-2.46	0.73	1.21
20	82.4	1.72	-2.09	0.82	1.05
25	78.7	1.66	-1.82	0.91	0.92
30	75.3	1.59	-1.60	0.99	0.80
35	72.0	1.53	-1.42	1.07	0.70
40	69.0	1.46	-1.27	1.15	0.62
45	66.2	1.40	-1.15	1.22	0.54
50	63.5	1.34	-1.04	1.29	0.47
60	58.7	1.23	-0.86	1.43	0.37
70	54.5	1.12	-0.72	1.56	0.28
80	50.8	1.02	-0.61	1.68	0.22
90	47.6	0.93	-0.52	1.80	0.17
100	44.8	0.84	-0.44	1.90	0.13
110	42.3	0.76	-0.38	2.01	0.10
120	40.1	0.68	-0.33	2.10	0.08

States, D.J. *et al.* (1991) "Improved sensitivity of nucleic acid database searches using application-specific scoring matrices." *Methods* **3**:66-70.

Protein Comparison vrs. DNA Comparison

For protein-coding DNA sequences, is it better to compare the DNA sequences directly, or their encoded proteins?

It is often said that protein comparisons are more sensitive, but this needs qualification.

Protein PAM distance	Information per residue (bits)	DNA PAM distance	Information per codon (bits)	Information ratio
0	4.17	0	6.00	1.44
10	3.43	8	4.53	1.32
20	2.95	16	3.63	1.23
30	2.57	24	2.95	1.15
40	2.26	32	2.43	1.08
50	2.00	40	2.02	1.01
60	1.79	48	1.69	0.94
70	1.60	56	1.42	0.89
80	1.44	64	1.19	0.83
90	1.30	72	1.01	0.78
100	1.18	80	0.86	0.73
110	1.08	88	0.73	0.68
120	0.98	96	0.62	0.63
130	0.90	104	0.53	0.59
140	0.82	112	0.46	0.56
150	0.76	120	0.39	0.51

DNA sequences contain all the amino acid information, so how can comparing their encoded proteins be more sensitive?

When DNA sequences are compared directly, they are usually compare base-to-base, ignoring the genetic code. It is here that information is lost.

Even then, protein comparison is more sensitive only at greater than 50 PAMs. However, by 150 PAMS, where many protein relationships are easily found, naïve DNA comparison misses half the available information.

States, D.J. *et al.* (1991) *Methods* **3**:66-70.