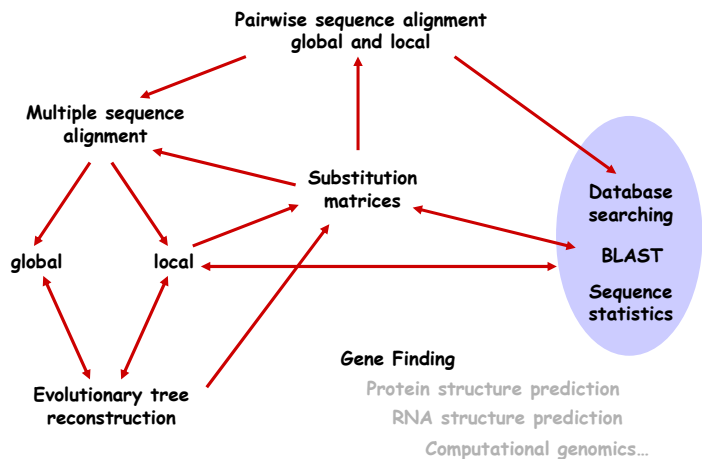
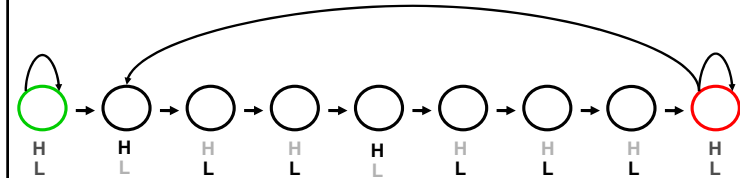


Logistics

- Problem Set 4 returned today
- Problem Set 5 out today, due Dec. 5
- Today: no office hours
- Thursday: Thanksgiving – no class
- Next week: Gene finding
- Dec 7: Last class, project presentations
- Final exam: Dec 18th

Coiled Coil HMM



Today

1. BLAST Statistics
 2. Introduction to information theory
 3. Information content of alignments
 4. How to select a substitution matrix
 4. Normalized bit scores
 5. Gapped BLAST.
- } Review

1. Relating S and E

Expected number of *ungapped* alignments with score S found with random sequences is:

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

where K is a constant that depends on $S[i,j]$ and can be computed from the theory for any scoring function.

The parameter λ is specified by the equation

$$1 = \sum p_i p_j e^{-\lambda S[i,j]}$$

Note that E is proportional to the size of the search space, mn , and decreases exponentially with the score, S

2. Introduction to information theory

Concepts from Shannon Information theory:

- Uncertainty; also called the *surprisal*
- Entropy
- Relative Entropy

Let's consider some examples:

- Tossing a coin
 - $N = 2$ states: H, T
 - Uncertainty: what is the outcome of the next toss?
- Telegraph
 - $N = |\Sigma|$ states: set of symbols
 - Uncertainty: what is the outcome of the next letter?

Uncertainty associated with state, i : $u(i) = -\log p_i$
 where p_i is the probability of state i .

Note:

- Since $p_i \leq 1$, $u(i) \geq 0$
- $u(i)$ approaches infinity as p_i approaches 0.
- $u(i)$ approaches 0 as p_i approaches 1.
- Uncertainty is additive: $u(i \text{ and } j) = u(i) + u(j)$.

Units depend on the base of the logarithm:

- Using \log_2 gives uncertainty in bits.
- In this case, uncertainty \sim number of yes/no questions needed to determine the state

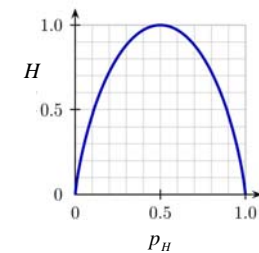
The **Entropy** is the expected uncertainty:

$$H = \sum_i p_i u(i) = -\sum_i p_i \log p_i$$

Coin toss example:

$$H = -p_H \log_2 p_H - (1 - p_H) \log_2 (1 - p_H)$$

Note: The entropy is maximal when all states are equally likely.



Relative Entropy aka Kullback-Leibler Divergence

$$D(Q || P) = \sum_i q_i \log \frac{q_i}{p_i}$$

Interpretations:

- If Q is the prior distribution (before seeing the data) and P is the posterior distribution, the KL divergence measures the decrease in uncertainty from seeing the data
- Expected discrimination information: Information available to discriminate in favor of hypothesis H_a against hypothesis H_0 , given H_a is true.

Note: D(Q|P) is not symmetric and therefore not a distance.

Relative Entropy for ungapped local alignments.

- Given sequences *a* and *b*:
- Alternate hypothesis (H_a): *a* and *b* are related at *n* PAMs divergence. *i* and *j* are aligned with “target” frequencies, q_{ij}^n
- Null hypothesis (H_0): *a* and *b* are unrelated. *i* and *j* are aligned with background frequencies, $p_i p_j$

The relative entropy $D_{KL}(Q || P) = \sum_{i,j} q_{ij}^n \log \frac{q_{ij}^n}{p_i p_j} \propto \sum_{i,j} q_{ij}^n S[i, j]$ gives the number of bits per position available to distinguish related alignments from chance at *n* PAMs.

To calculate the relative entropy, we need to know the target frequencies

How to determine the target frequencies, q_{ij} ?

1. From PAM transition probability: $q_{ij} = p_i P^n[i, j]$
2. Generate “random” sequences from background probabilities
Find MSPs in pairs of random sequences
Count target frequencies in MSPs

3. From theory: $q_{ij}^n = p_i p_j e^{-\lambda S^n[i, j]}$

Karlin & Altschul, 90

The information available in the alignment depends on $S[i, j]$ and is given by the *relative entropy*:

$$D_{kl} = \sum P(H_A) \log_2 \frac{P(H_A)}{P(H_0)} = \sum q_{ij} \log_2 \frac{q_{ij}}{p_i p_j}$$

BLOSUM		PAM		Sequence identity
	bits/site		bits/site	
		20	2.95	83%
		30	2.57	
		60	2.00	63%
90	1.18	100	1.18	43%
80	0.99	120	0.98	38%
60	0.66	160	0.70	30%
50	0.52	200	0.51	25%
45	0.38	250	0.36	20%

3. Information content of alignments

Given the expected number of false positives:

$$E = K m n e^{-\lambda S}$$

If we choose $\lambda = \ln 2$, then with some algebraic manipulation

$$S \approx \log_2 mn$$

Interpretation:

the minimum score need to distinguish MSPs from chance is equivalent to the number of bits required to specify the starting position of the alignment.

How many bits are required to find meaningful alignments in today's databases?

$$S \approx \log_2 mn$$

For a typical amino acid sequence of length $m = 250$,

–if $n = 1$ billion, then a minimum of **38 bits** are required to distinguish significant MSP's from chance.

The information available in the alignment depends on $S^n[i,j]$ and is given by the *relative entropy*

$$D_{kl} = \sum q^n_{ij} \log_2 \frac{q^n_{ij}}{p_i p_j}$$

Implications

The lower the relative entropy, H , the longer the minimum alignment that is distinguishable from chance.

$$\frac{\text{minimum number of bits}}{\text{bits per position}} = \text{minimum query sequence length}$$

In a data base of length 1 billion, 38 bits are required.

A query sequence must be at least

$38/2.57 = 15$ residues long at **30 PAMs**

$38/0.70 = 54$ residues long at **160 PAMs**

$38/0.36 = 105$ residues long at **250 PAMs**

to distinguish significant HSP's from chance.

	PAM	Seq Id
30	2.57	
100	1.18	43 %
120	0.98	38%
160	0.70	30 %
200	0.51	25%
250	0.36	20 %

$$\frac{\text{minimum number of bits}}{\text{bits per position}} = \text{minimum query sequence length}$$

Today

1. BLAST Statistics
 2. Introduction to information theory
 3. Information content of alignments
 4. How to select a substitution matrix
 4. Normalized bit scores
 5. Gapped BLAST.
- } Review

4. How to choose $S[i,j]$?

The best scoring matrix for distinguishing significant alignments by chance is the matrix corresponding to the q_{ij} from related sequences at the evolutionary distance of interest [KA90].

$$S[i, j] = \log_2 \frac{q_{ij}}{p_i p_j}$$

Proof by contradiction

- Let $S^n[]$ be the optimal matrix, i.e., the matrix that best distinguishes significant alignments from chance alignments, at n PAMs.
- Then MSPs in sequences n PAMs diverged will have maximum scores when scored with $S^n[]$.
- Suppose $S^n[]$ does not correspond to $q_{ij} = p_i p_j e^{-\lambda S[i, j]}$
- Then there must be some x, y in Σ that occurs more frequently than q_{ij} .
- We can increase the score of the MSPs by increasing $S^n[x, y]$.
- $S^n[]$ is not optimal.

Implications

BLAST will give reasonable accuracy as long as the empirical target frequencies, q_{ij} , in the alignments of interest do not deviate too far from the theoretical target frequencies:

$$q_{ij} = p_i p_j e^{-\lambda S[i, j]}$$

Reasonable accuracy can be achieved with two or three matrices.

The average score (in bits) per alignment position when using a PAM M matrix to compare sequences in fact separated by D PAMs

(Calculated by simulation)

PAM matrix M employed	40	80	Actual PAM distance D of segments					
			120	160	200	240	280	320
40	2.26	1.31	0.62	0.10	-0.30	-0.61	-0.86	-1.06
80	2.14	1.44	0.92	0.53	0.23	-0.02	-0.21	-0.37
120	1.93	1.39	0.98	0.67	0.42	0.22	0.06	-0.07
160	1.71	1.28	0.95	0.70	0.50	0.33	0.20	0.09
200	1.51	1.16	0.90	0.68	0.51	0.38	0.26	0.17
240	1.32	1.05	0.82	0.65	0.51	0.39	0.29	0.21
280	1.17	0.94	0.75	0.60	0.48	0.38	0.30	0.23
320	1.03	0.84	0.68	0.56	0.46	0.37	0.30	0.24

$$\text{Efficiency} = \frac{\text{Score with PAM } M}{\text{Score with PAM } D}$$

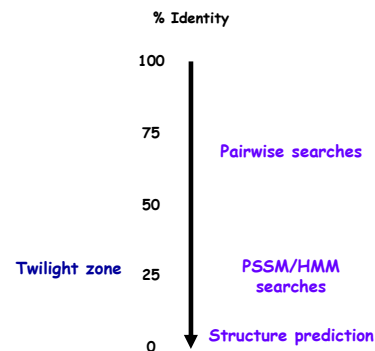
□ = 94% efficiency

Choosing your scoring matrix

- BLAST will give reasonable accuracy as long as the empirical target frequencies do not deviate too far from the theoretical target frequencies
 - Use PAM40, PAM120 & PAM240, or PAM120 & PAM250
- The lower the relative entropy, H , the longer the minimum alignment that is distinguishable from chance.

The "Twilight" Zone

- The scale indicates % identity between aligned sequences
- The Twilight Zone
 - Around 20%-35% identity
 - "Random" alignments begin to appear



5. Normalized bit scores

E-values depend on K and λ , $E = Kmne^{-\lambda S}$ which in turn depend on the scoring matrix, $S[i,j]$.

By normalizing the alignment scores

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

we obtain E values that are independent of K , λ and $S[i,j]$.

$$E = mn2^{-S'}$$

With *normalized bit scores*, E values from database searches with different scoring matrices can be compared.

6. Gapped, two-hit BLAST

Altschul *et al*, 97

Recall BLAST 90:

0. Select S based on desired E
1. Construct L : a list of high scoring words of length w with score $\geq T$
Choose $w = 3$, $T = 13$ empirically (1997)
2. Find *hits* – instances of words in L in database
3. Extend hits to find HSPs with score $> S$.
Stop extension if ungapped score drops below X

90%

Problems with BLAST 90

- Only finds ungapped alignments. *Solution*: find several HSP's and merge them:



- However, if $P(S1\&S2)$ is significant but neither $P(S1)$, nor $P(S2)$ is significant alone, we need to find hits in both $S1$ and $S2$ to find this alignment.
- To increase probability that both $S1$ and $S2$ are found, decrease word threshold to $T=11$.
- This will increase the number of hits found in step two and the number of unnecessary extensions in step 3.

Gapped, two-hit BLAST

Altschul *et al*, 97

Innovations:

1. Gapped extensions
2. Two hit BLAST
3. PSI-BLAST not covered in class
Apply BLAST iteratively to build a PSSM

Gapped Extensions

- Increase T to 13, yielding fewer hits.
- Find HSP's using *ungapped* extensions
Stop if ungapped alignment score drops below X
- If HSP score $> S_1$, perform a *gapped* extension.
Stop if gapped alignment score drops below X_g
- Report match if S_2 exceeds threshold



S_g : gapped extension threshold X : ungapped extension cutoff
 X_g : gapped extension cutoff

Gapped alignment statistics

- Test $mn2^{-S'} \leq E_s$, $S' = \frac{\lambda S - \ln K}{\ln 2}$
- For ungapped alignments, γ_u and K_u can be determined from theory.
- For gapped alignments, γ_g and K_g must be determined by simulation in advance.
- Gapped BLAST cannot be used with arbitrary scoring matrices.

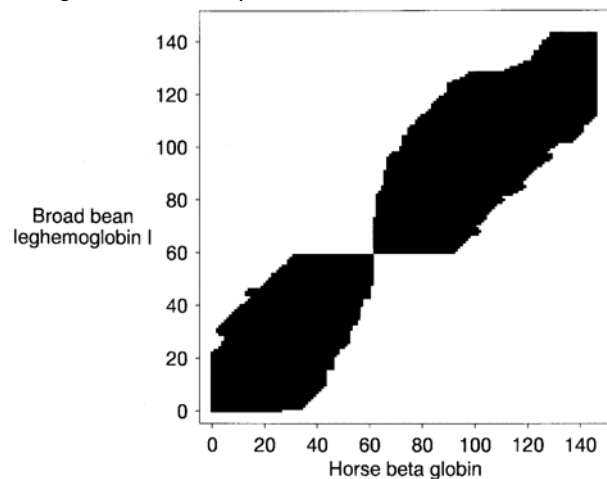
```

43 F S F L K D S A G V V D S P K L G A H A E K V F G M V R D S A V Q L R A T G E V V L D G K D G S ----- 90
   F L + V + + P K + A H + K V L + G E V L D G +
45 F G D L S N P G A V M G N P K V K A H G K K V ----- L H S F G E G V H L D N L K G T F A A L S E 90

91 I H I Q K G V L D P H F V V V K E A L L K T I K E A S G D K W S E E L S A A W E V A Y D G L A T A I 140
   + H K + D P + F + + L + + G + + E L A + + + G + A A +
91 L H C D K L H V D P E N F R L L G N V L V V L A R H F G K D F T P E L Q A S Y Q K V V A G V A N A L 141
    
```

Altschul *et al*, 97

- cutoff X_g limits search space



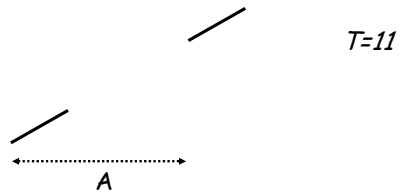
Altschul *et al*, 97

Gapped Extensions: Performance

1. Increase T to 13 yielding fewer hits.
 2. Find HSP's, using *ungapped* extensions
 3. If HSP score $> S_1$, perform a *gapped* extension.
 4. If gapped extension score $> S_2$, report match.
- Ungapped extensions reduced by 2/3
 - Gapped extensions cost 500 times ungapped extensions
 - One gapped extension per 4000 ungapped extensions
 - Reduce running time by more than a factor of 2.

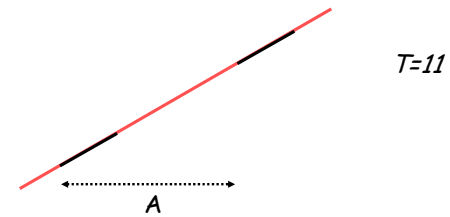
Two Hit BLAST

- Reduce threshold T to obtain *more* hits
- Only trigger an ungapped extension if there are *two hits* on the *same diagonal* within distance A



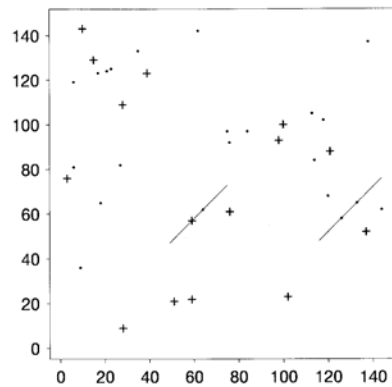
Two Hit BLAST

- Reduce threshold T to obtain *more* hits
- Only trigger an ungapped extension if there are *two hits* on the *same diagonal* within distance A



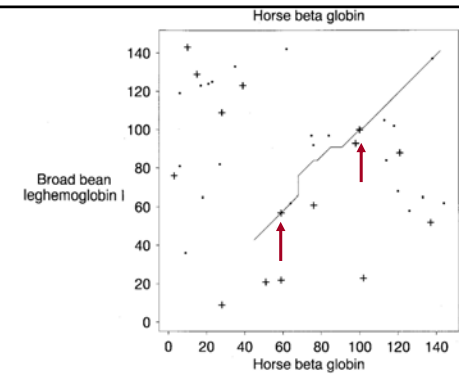
- + 15 hits with score ≥ 13
- 37 hits with score ≥ 11

But, only two extensions were performed!



Altschul *et al*, 97

- + 15 hits with score ≥ 13
- 37 hits with score ≥ 11



```

43 FSFLKDSAGVVDSPKLGAAAEKVFQMVDRSAVQLRATGEVVLDGKDGSD----- 90
   F L + V+ +PK+ AH +KV L + GE V LD G+
45 FGDLSNPGAVMGNPKVKAHGKVV-----LHSFGEGVHLDNLKGTFAALSE 90

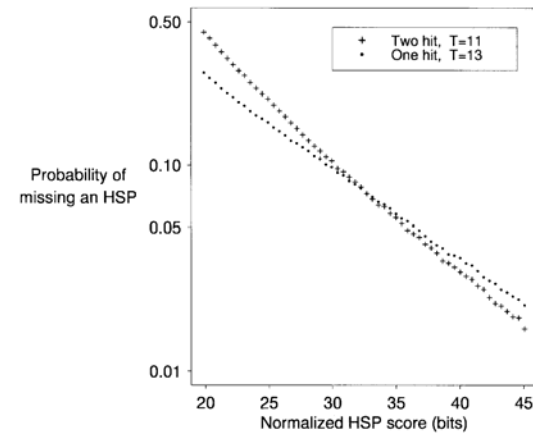
91 IHIQGVLDP-HFVVVKEALLKTIKASGDKWSELSAAWEVAYDGLATAI 140
   +H K +DP +F ++ L+ + G ++ EL A+++ G+A A+
91 LHCDKLHVDPEFRLLGNLVVVLARHFGKDFTPELQASYQKVVAGVANAI 141
    
```

Altschul *et al*, 97

Two Hit BLAST: Performance

- Reduce threshold T to obtain *more* hits
- Only trigger an ungapped extension if there are *two hits* on the *same diagonal* within distance A
- For $w=3$, $T=11$, $A=40$,
 - 3.2 times as many hits
 - 0.14 times as many extensions
 - speed up $\sim 2X$

Two hit BLAST has better sensitivity above ~ 34 bits



Altschul *et al*, 97

Putting it all together

1. Choose tolerated false positive rate, E
2. Find hits of length w with similarity threshold T .
3. If there are
 - two* word pairs
 - on *same diagonal*
 - separated by a *distance of at most A*,perform an *ungapped* extension to obtain an HSP using cutoff, $X1$.
4. If HSP score $\geq S_1$, computed from E using λ_u and K_u perform a *gapped extension* using dynamic programming with cutoff $X2$.

Putting it all together

5. If gapped local alignment score $\geq S_2$, calculated using λ_g and K_g , realign with dynamic programming using cutoff $X3 > X2$ to get traceback.
6. Report
 - Match
 - Normalized score (in bits)
 - Significance expressed as an "E-value".

Gapped BLAST statistics

- Ungapped parameters: λ_u and K_u
Computed from theory
- Gapped parameters: λ_g and K_g
Computed using simulations
- You cannot easily use your own substitution matrix when using gapped BLAST

Comparing BLAST parameters

Parameters	BLAST 90	BLAST 97
matrix	$S[i,j]$	$S[i,j]$
threshold	E or S	E or S_2, S_1
word size	$w=3$	$w=3$
word threshold	$T=13$ or 11	$T=11$
max distance btw hits		$A=40$
extension cutoff	X	$X1, X2, X3$

