End of Semester Logistics ...

Final exam

- Monday, December 12th,
- 8:30-11:30, Hamerschlag B131
  - Cummulative, emphasis on 2^{nd} half
  - Closed book, 2 pages of notes

- Study guide: bottom of syllabus page
- Review session:
  - Sunday, Dec 11^{th}
  - 2pm – 4pm
  - BH 255A (here)

End of Semester Logistics ...

Homework

- Solutions 1-6 online
- Solution 7 online tomorrow.
- **Problem Set 8 due at midnight on Friday**
- 711-6 canceled

- Late homework recieves a zero score once the solution sets have been posted.
- In calculating your final score, your lowest homework score will be dropped provided that all assignments have been submitted by the last day of classes.
- REMEMBER to note who you worked with!

End of Semester Logistics ...

Problem set 8

- Run 5 Blast searches with different parameter settings
- Record some results in Tables 1 & 2 (excel worksheet)
- Interpret in terms of Blast heuristics and Karlin Altschul stats

Recommendations:

- Run all five searches in one session
- Record results immediately
- Interpret results at your leisure

PLEASE PLEASE PLEASE
FILL OUT FACULTY COURSE EVALUATIONS

4

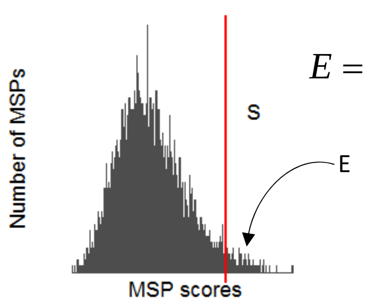
- How much information is available to distinguish between chance MSPs and MSPs in related sequences?
  - Information content of substitution matrices
  - Information content of alignments

Tuesday

- Which substitution matrix will maximize precision and recall?

Today

Recall: BLAST (Karlin-Altschul) Statistics



$$E = Kmne^{-\lambda S}$$

Using bit score $S_b = \frac{\lambda S - \ln K}{\ln 2}$

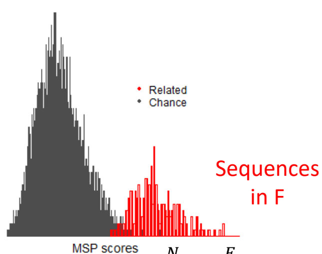
$$E = mn2^{-S_b}$$

E = number of MSPs with scores > *S.*

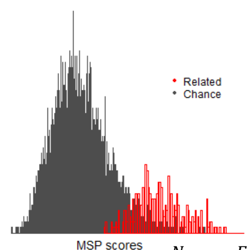
Maximal Segment Pair (MSP): an ungapped local alignment that cannot be improved by making it bigger or smaller.

6

- Each scoring matrix S^N *[x,y]* has characteristic frequencies, q_{xy}^N
- Each protein family *F* has characteristic amino acid pair frequencies, q_{xy}^F
- The best results are obtained when q_{xy}^N is a good match for q_{xy}^F
- Given a query Q and a database sequence D_j both from family F, the alignment of Q and D_j will result in the highest MSP score when q_{xy}^N is a good match for q_{xy}^F
- In a database search with query Q, the best descrimination between chance MSPs and MSPs in alignments from related sequences in family *F* will be obtained when q_{xy}^N is a good match for q_{xy}^F.



$q_{xy}^N \approx q_{xy}^F$



$q_{xy}^N <> q_{xy}^F$

9

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

(Calculated by simulation)

PAM matrix		Actual PAM distance *n*							
		40	80	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
Y	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

Maxima highlighted in yellow

Best discrimination between related and chance MSPs :

Matrix divergence ~ Family divergence

10

Altschul SF, J. Mol. Biol., 219, 555-565 (1991)

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

(Calculated by simulation)

PAM matrix *Y*	Actual PAM distance *n*: 40	80	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**

□ = Efficiency ≥ 94%

$$\text{Efficiency} = \frac{\text{Score with PAM } Y}{\text{Score with PAM } n}$$

11

Altschul SF, J. Mol. Biol., 219, 555-565 (1991)

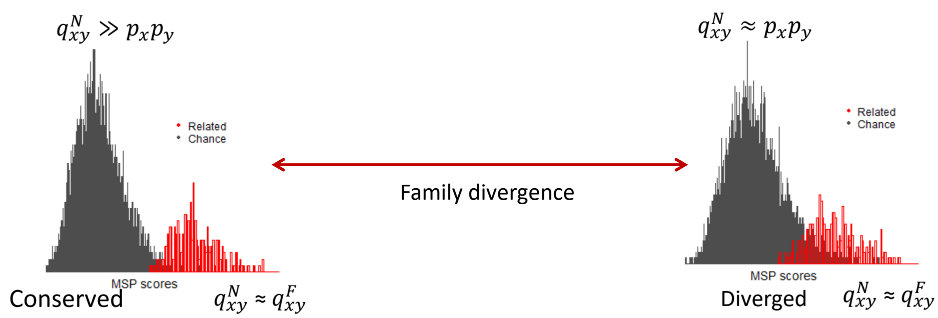
Choosing your scoring matrix

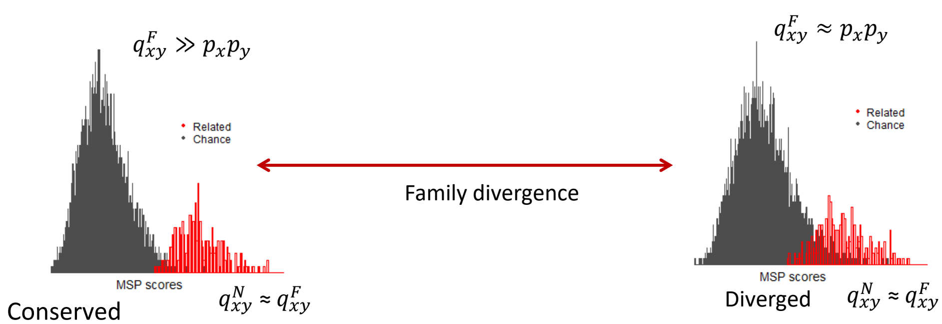
1. BLAST will give reasonable accuracy as long as the empirical target frequencies do not deviate too far from the theoretical target frequencies
   - *Use PAM40, BLOSUM62 & BLOSUM45, or BLOSUM62 & BLOSUM45*

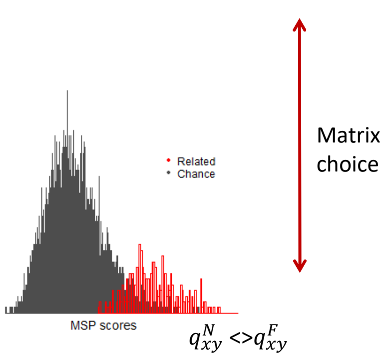
13

The more conserved the protein family, the greater the difference between pair frequencies in the family (q^F_{xy}) and background pair frequencies ($p_x p_y$).

Greater deviation between q^F_{xy} and $p_x p_y$ yields better discrimination between chance and related MSPs







Greater deviation between q^F_{xy} and $p_x p_y$ yields better discrimination between chance and related MSPs

The best descrimination is obtained when $q^N_{xy} \approx q^F_{xy}$.

A warm-up thought experiment: How much information is available in a sequence of coin tosses to determine if the coin is fair or biased?

Alternate Hypothesis (H_A): Coin is biased

– $pr(H|H_A) = q$, $pr(T|H_A) = (1-q)$, *where* $q \neq 0.5$

Null Hypothesis (H_0): : Coin is fair

– $p(H|H_0) = p$, $p(T|H_0) = (1-p)$, *where* $p = 0.5$

- If q >>0.5 (e.g., q = 0.8), then a short series of coin tosses is sufficient to convince us that H_A is true.
- If q ≈ 0.5 (e.g., q = 0.5001), then we require a much longer series of coin tosses is sufficient to convince us that p(H) ≠ *0.5.*

16

Relative Entropy

Given two probability distributions, P and Q, defined on the same event space, $X=\{E_1, E_2, \ldots E_N\}$

- $P = pr(X|\hat{H}_0) = \{p_1, p_2, \ldots p_N\}$
- $Q = pr(X|\hat{H}_A) = \{q_1, q_2, \ldots q_N\}$

the *relative entropy* or *Kullback-Leibler Divergence*

$$\mathcal{H} = \sum_{X} q_i \log_2 \frac{q_i}{p_i}$$

is the expected information provided by each observation to discriminate in favor of hypothesis $\hat{H}_a$ against hypothesis $\hat{H}_0$, when $\hat{H}_a$ is true.

Note: the KL Divergence is not symmetric and therefore not a distance.

Relative Entropy – coin toss example

Given two probability distributions, P and Q, defined on the same event space, *X={H,T}*

- P = pr(X |$\hat{H}_0$) = *{0.5, 0.5}*
- Q = pr(X |$\hat{H}_A$) = *{q, (1-q)}), where q ≠ 0.5*

the *relative entropy*

$$\sum_{\{H.T\}} q_i \log_2 \frac{q_i}{p_i}$$

$$q \log_2 \frac{q}{p} + (1-q) \log_2 \frac{(1-q)}{p}$$

is the expected information available per coss of the coin to discriminate in favor of hypothesis $\hat{H}_a$ (biased coin) against hypothesis $\hat{H}_{0,}$ (fair coin) if the coin is actually biased.

A warm-up thought experiment: How much information is available in a sequence of coin tosses to determine if the coin is fair or biased?

Alternate Hypothesis (H_A): Coin is biased

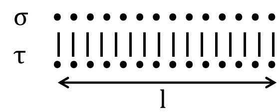
– pr(H|H_A) = *q*, pr(T|H_A) = *(1-q), where q ≠ 0.5*

Null Hypothesis (H_0): : Coin is fair

– p(H|H_0) = *p*, p(T|H_0) = *(1-p), where p = 0.5*

- If q >>0.5 (e.g., q = 0.8), then a short series of coin tosses is sufficient to convince us that H_A is true.
- If q ≈ 0.5 (e.g., q = 0.5001), then we require a much longer series of coin tosses is sufficient to convince us that p(H) *≠ 0.5.*

Relative Entropy – ungapped local alignments



Alternate Hypothesis (H_A): σ and τ are related at *N* PAMs divergence.

– Amino acids *x* and *y* are aligned with frequency, q^N_{xy}

– Each alignment column is an observation, similar to a coin toss

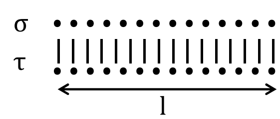
Null Hypothesis (H_0): σ and τ are unrelated

– Amino acids *x* and *y* are aligned with background frequencies, $p_x p_y$

- If q^N_{xy} very different from $p_x p_y$, then only a few observations are sufficient to convince us that H_A is true (i.e., short alignment).
- If the family is diverged, (q^N_{xy} more similar to $p_x p_y$), then we require more observations (i.e., a longer alignment).

20

Relative Entropy – ungapped local alignments



Alternate Hypothesis (H_A): σ and τ are related at *N* PAMs divergence.

– Amino acids *x* and *y* are aligned with frequency, q^N_{xy}

– Each alignment column is an observation, similar to a coin toss

Null Hypothesis (H_0): σ and τ are unrelated

– Amino acids *x* and *y* are aligned with background frequencies, $p_x p_y$

Relative entropy $\mathcal{H}^N = \sum_{\{xy\}} q^N_{xy} \log_2 \frac{q^N_{xy}}{p_x p_y} = \sum_{\{xy\}} q^N_{xy} S^N[x, y]$

gives the number of bits per position available to distinguish chance MSPs from MSPs in related sequences with *N* PAMs of divergence.

21

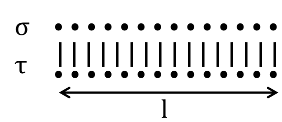
The *average* relative entropy of a substitution matrix is given in bits per position and can be calculated from S^N using the equation

$$\mathcal{H}^N = \sum_{\{xy\}} q_{xy}^N S^N[x,y]$$

BLOSUM		PAM		Sequence identity
	bits/site		bits/site	
		20	2.95	83%
		30	2.57	
		60	2.00	63%
		70	1.60	
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%

22

An ungapped alignment of length *l* between two sequences separated *N* PAMs divergence contains $l\mathcal{H}^N$ bits of discriminatory information, on average.



How many bits of information are needed to find a related match in a database search?

$$E = m'n'2^{-\mathcal{S}}$$

$$\mathcal{S} = \log_2 \frac{m'n'}{E}$$

Suppose we seek matches with E values no greater than $E = 1$. Then, we require $\mathcal{S} \geq \log_2 m'n'$ bits. From this, we can estimate the minimum alignment length required to disinguish related from chance MSPs at N PAMS:

$$l\mathcal{H}^N = \log_2 m'n'$$

$$l = \frac{\log_2 m'n'}{\mathcal{H}^N}$$

Implications

The lower the relative entropy, $\mathcal{H}^N$, the longer the minimum alignment that is distinguishable from chance.

$$l = \frac{\log_2 m'n'}{\mathcal{H}^N}$$

In a data base of length *n = 50* billion, $\log_2 m'n'$ = *44* bits are required. Since the alignment cannot be longer than the query, a query sequence must be at least

44/2.57 = 17 residues long at ***30 PAMs***
44/0.70 = 62 residues long at ***160 PAMs***
44/0.36 = 121 residues long at ***250 PAMs***

to distinguish significant HSP's from chance.

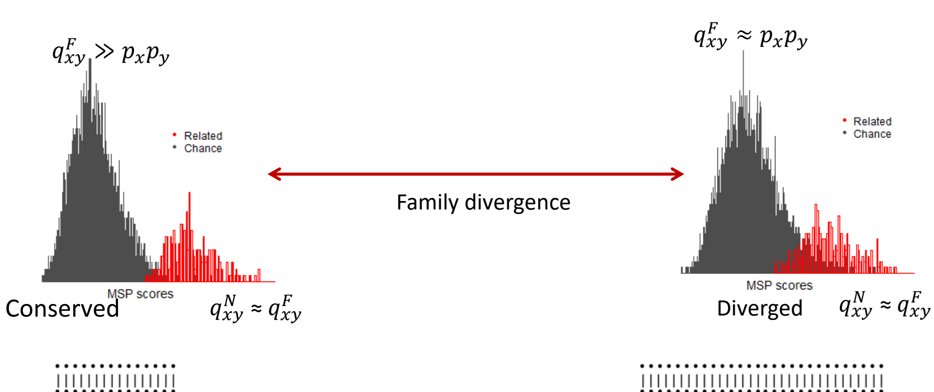
PAM		Seq Id
30	2.57	
100	1.18	43 %
120	0.98	38%
160	0.70	30 %
200	0.51	25%
250	0.36	20 %

Note that $\mathcal{H}^N$ is an average over scoring matrix S^N.
A shorter alignment may encode enough information if it contains many high-scoring pairs; alternatively, you may need a longer alignment if there are many low-scoring pairs.

24

Greater deviation between q^F_{xy} and $p_x p_y$ yields better discrimination between chance and related MSPs

Lower $\mathcal{H}$ requires longer alignments for the same discrimination

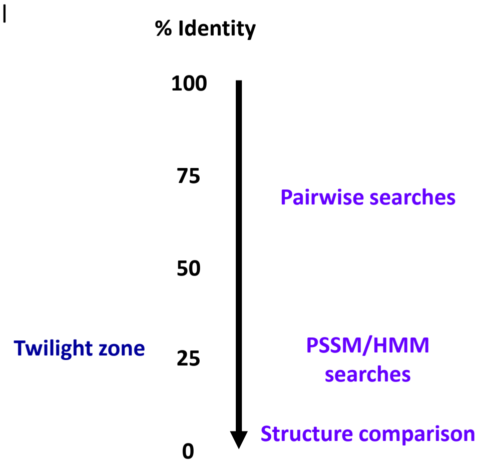


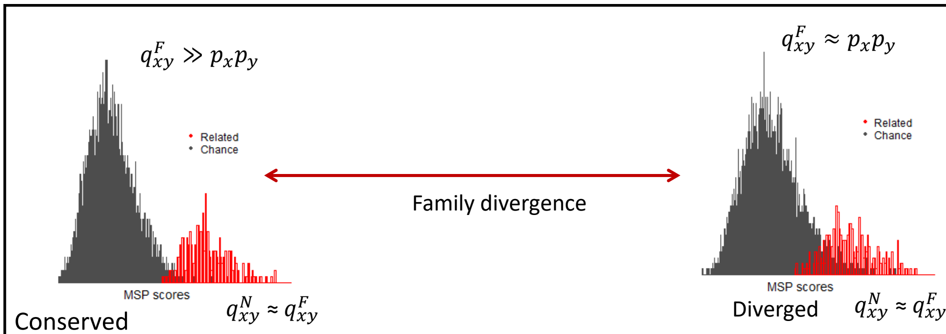
25

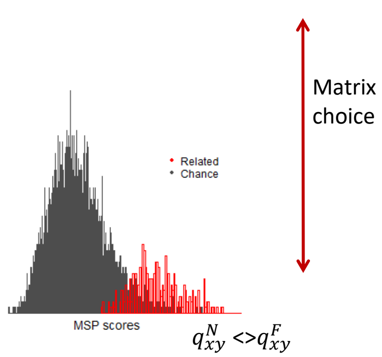
The "Twilight" Zone

- The scale indicates % identity in local alignments (MSPs).

- The Twilight Zone
  - Around 20%-35% identity
  - Difficult to distinguish between MSPs in related sequences and "chance" alignments







Greater deviation between q^F_{xy} and $p_x p_y$ yields better discrimination between chance and related MSPs

The discrimination is further degraded with a poorly matched scoring matrix

The best descrimination is obtained when $q^N_{xy} \approx q^F_{xy}$.

Choosing your scoring matrix

1. BLAST will give reasonable accuracy as long as the empirical target frequencies do not deviate too far from the theoretical target frequencies
   - *Use PAM40, BLOSUM62 & BLOSUM45, or BLOSUM62 & BLOSUM45*
2. The lower the relative entropy, *H*, the longer the minimum alignment that is distinguishable from chance.
3. If your query is short, you will only be able to find closely related matches.
   - *Use PAM30*

28

DATABASE SEARCHING RECAP

29

Searching a sequence database

Input:
- query Q of length *m*
- database D=D1 D2 D3... D_N of length *n*

Search:

for j = 1 to N
- Find best local aligment of Q with Dj
- If "good alignment", add Dj to *Results*

Output: *Results*

PROBLEMS
- Too slow
- What is a "good" alignment?
- Which matrix should you use?
- Which results are trustworthy?
- Can you find all related sequences in the database?

30

Basic Local Alignment Search Tool

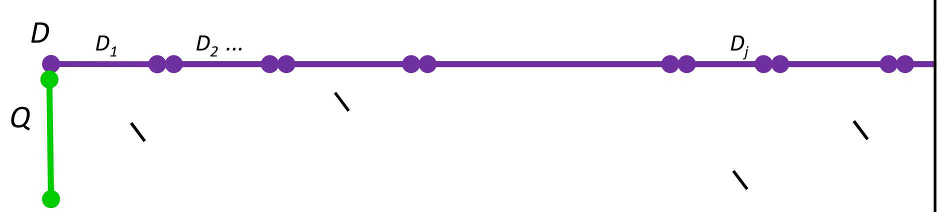
Altschul *et al*, 90

Construct hash table *L*
- Find all strings of length *w* that align with a w-mer in Q with score ≥ *T*

Scan database *D* for *hits* – instances of words in *L*

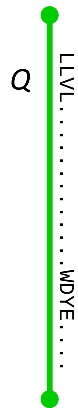
Extend hits to find MSPs

If D_j contains MSP with score $S > S_T$ report D_j



Not all w-mers in Q will be included in L
Some w-mers not in W will be included in L

Suppose that w = 4 and T = 17 and
word scores are calculated with BLOSUM 62



L L V L
L L V L
4 4 4 4   16 < T

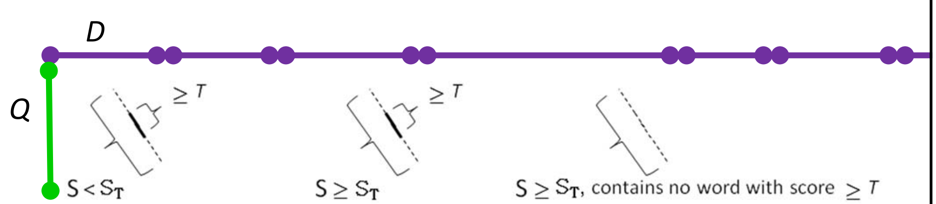
LLVL will not be included in L, because when aligned with itself the word score is lower than T

W D Y E
W E F E
11 2 3 5   21 > T

WEFE will be included in L because it aligns with a word in Q (WDYE) with a score greater than T

33

Problems with Blast 90: Accuracy and running time



Unnecessary extension: Contains a hit but MSP score is too low
- W-mer false positive

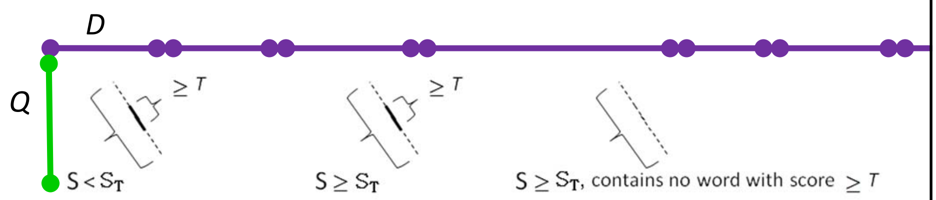
success

Misses too many related sequences
- W-mer false negative

- Two many unnecessary extensions
- Only finds ungapped alignments.

34

Problems with Blast 90: Accuracy and running time



Unnecessary extension: Contains a hit but MSP score is too low
- W-mer false positive

success

Misses too many related sequences
- W-mer false negative

- Two many unnecessary extensions
- Only finds ungapped alignments.

Fail to find matching sequence if *S* and *S'* are significant together, but neither *S* nor *S'* is significant alone,

35

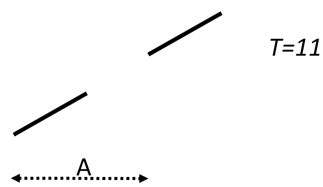
```
43 FSFLKDSAGVVDSPKLGAHAEKVFGMVRDSAVQLRATGEVV--LDGKDGS------ 90
   F  L +    V+ +PK+ AH +KV          L + GE V LD    G+
45 FGDLSNPGAVMGNPKVKAHGKKV----------LHSFGEGVHHLDNLKGTFAALSE 90

91 IHIQKGVLDP-HFVVVKEALLKTIKEASGDKWSEELSAAWEVAYDGLATAI 140
   +H  K  +DP +F ++    L+  +     G  ++ EL A+++     G+A A+
91 LHCDKLHVDPENFRLLGNVLVVVLARHFGKDFTPELQASYQKVVAGVANAL 141
```

Altschul *et al*, 97

An example: This alignment has two conserved regions connected by gapped region

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***



- Misses fewer significant MSPs
- Fewer unnecessary extensions

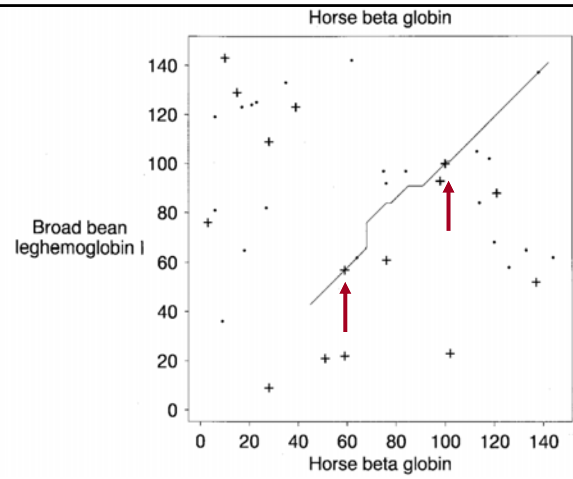
Gapped, 2-hit Blast

Altschul *et al*, 97

1. Find hits of length *w* with similarity threshold *T*.
2. If D_J has
   two hits
   on *same diagonal*
   separated by a *distance of at most A*,
   perform an *ungapped* extension to obtain MSP
3. If MSP score $S_1 > S_g$, perform a *gapped extension* with dynamic programming
4. If gapped extension score $S_2 > S_T$, report D_J as a match.

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

Only two extensions were performed!

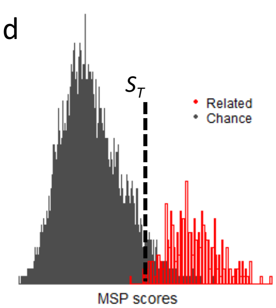



```
43 FSFLKDSAGVVDSPKLGAHAEKVFGMVRDSAVQLRATGEVV--LDGKDGS------ 90
   F  L +    V+ +PK+ AH +KV          L + GE V  LD    G+
45 FGDLSNPGAVMGNPKVKAHGKKV---------LHSFGEGVHHLDNLKGTFAALSE 90

91 IHIQKGVLDP-HFVVVKEALLKTIKEASGDKWSEELSAAWEVAYDGLATAI 140
   +H  K  +DP +F ++    L+  +     G   ++ EL A+++     G+A A+
91 LHCDKLHVDPENFRLLGNVLVVVLARHFGKDFTPELQASYQKVVAGVANAL 141
```

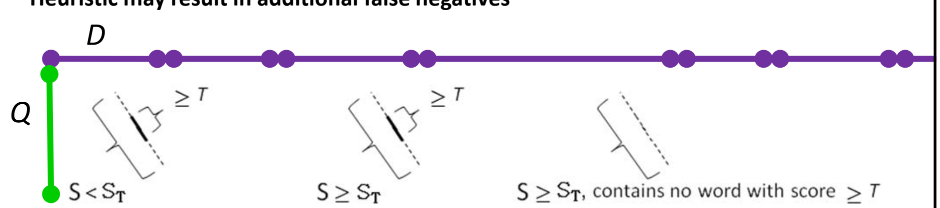
Altschul et al, 97

False positives and false negatives depend on the overlap of the distributions of

- Chance MSPs
- MSPs in related sequences



Heuristic may result in additional false negatives



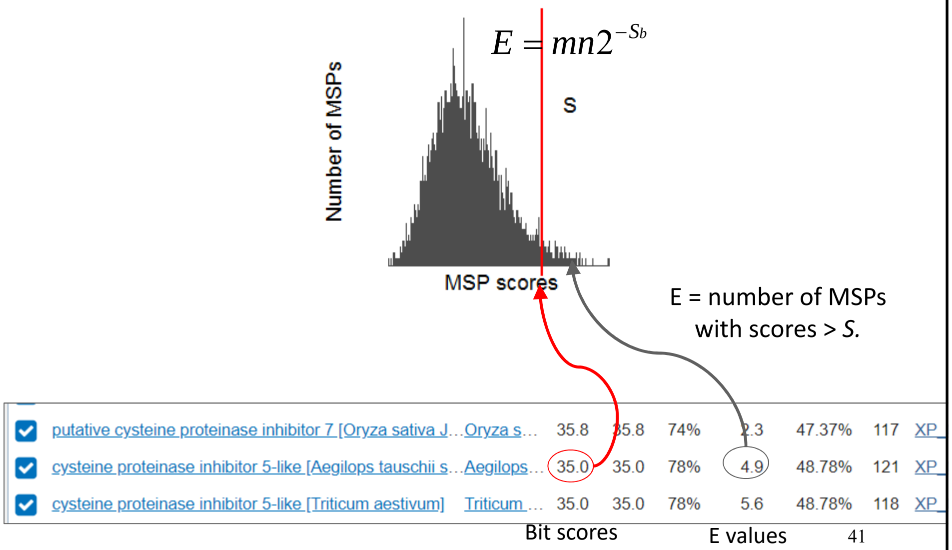
unnecessary extension

success

heuristic fails to find a related sequence

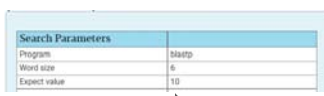
40

BLAST (Karlin-Altschul) Statistics

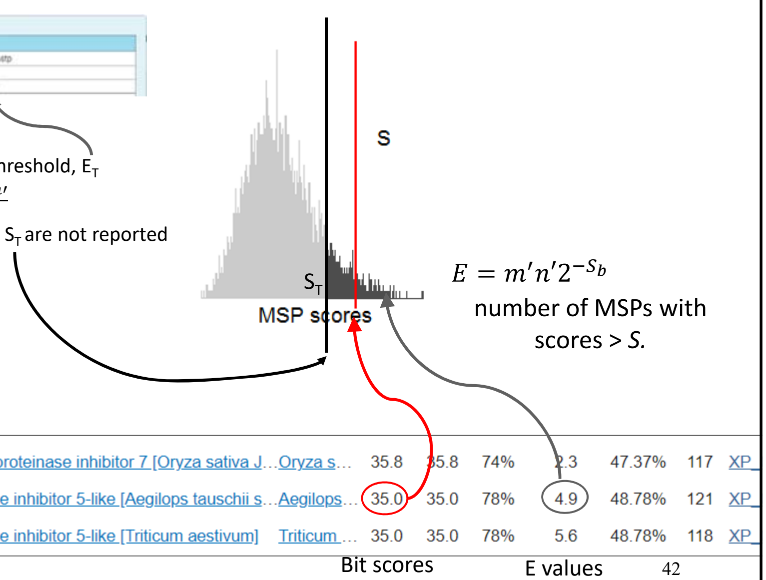


41

BLAST (Karlin-Altschul) Statistics

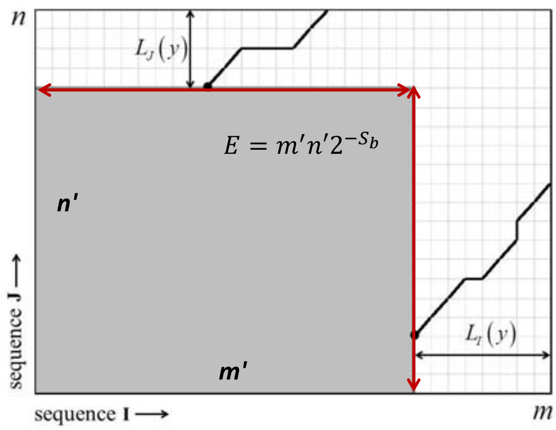


- User selects threshold, E_T
- $S_T = \log_2 \frac{m'n'}{E_T}$
- MSPs with S < S_T are not reported



42

The distances *m′* and *n′* include an "edge correction"



Note: the NCBI papers show alignment matrices starting in the lower LH corner, not the upper LH corner as we use in this class.

"On average", an alignment must start within the gray box to accrue a score of at least S_T before reaching the end of the sequence

New finite-size correction for local alignment score distributions. Park, Sheetlin, Ma, Madden, Spouge* *BMC Research Notes* 2012, 5:286 doi:10.1186/1756-0500-5-286

40