End of Semester Logistics ...

Final exam

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

- Study guide: bottom of syllabus page
- Review session:
  - Sunday, Dec 11th
  - 2pm -5pm
  - 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.

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:

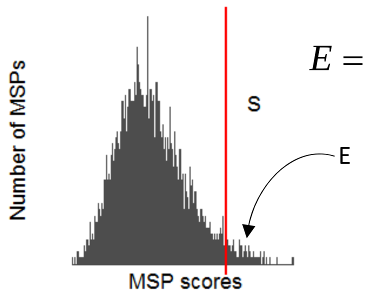
- Run all five searches in one session
- Record results immediately
- Interpret results at your leisure

PLEASE PLEASE PLEASE
FILL OUT FACULTY COURSE EVALUATIONS

4

BLAST (Karlin-Altschul) Statistics

E = Expected number of matches with score at least *S* under the null model



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

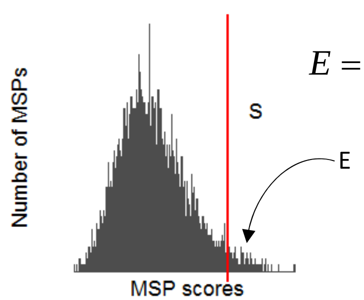
E-values depend on K and λ, which in turn depend on the scoring matrix, *S[i,j]*.

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

5

BLAST (Karlin-Altschul) Statistics

E = Expected number of matches with score at least *S* under the null model



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

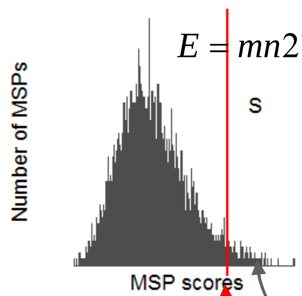
By normalizing the alignment scores

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

we obtain a **"bit score"** S_b and an expression for *E* that is independent of *K* and λ

6

BLAST (Karlin-Altschul) Statistics



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

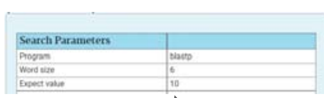
E = number of MSPs with scores > *S.*

☑	putative cysteine proteinase inhibitor 7 [Oryza sativa J...	Oryza s...	35.8	35.8	74%	2.3	47.37%	117	XP_
☑	cysteine proteinase inhibitor 5-like [Aegilops tauschii s...	Aegilops...	35.0	35.0	78%	4.9	48.78%	121	XP_
☑	cysteine proteinase inhibitor 5-like [Triticum aestivum]	Triticum ...	35.0	35.0	78%	5.6	48.78%	118	XP_

Bit scores E values

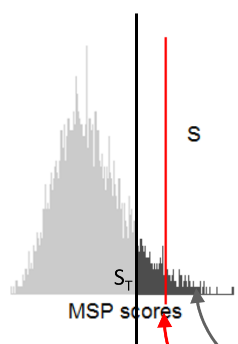
7

BLAST (Karlin-Altschul) Statistics



Search Parameters	
Program	blastp
Word size	6
Expect value	10

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



$$E = m'n'2^{-S_b}$$

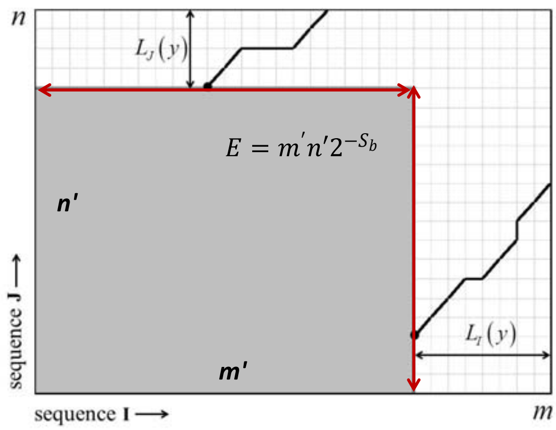
number of MSPs with scores > *S.*

☑	putative cysteine proteinase inhibitor 7 [Oryza sativa J...	Oryza s...	35.8	35.8	74%	2.3	47.37%	117	XP_
☑	cysteine proteinase inhibitor 5-like [Aegilops tauschii s...	Aegilops...	35.0	35.0	78%	4.9	48.78%	121	XP_
☑	cysteine proteinase inhibitor 5-like [Triticum aestivum]	Triticum ...	35.0	35.0	78%	5.6	48.78%	118	XP_

Bit scores E values

8

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

- How much information is available to distinguish between chance MSPs and MSPs in related sequences? — Today
  - Information content of substitution matrices
  - Information content of alignments
- Which substitution matrix will maximize precision and recall? — Thursday

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.*

13

Relative Entropy

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

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

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, $X=\{H,T\}$

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

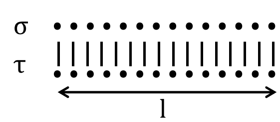
the *relative entropy*

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

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

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.

Relative Entropy – ungapped local alignments



- Alternate hypothesis ($\hat{H}_a$)*:* σ and τ are related at *N* PAMs divergence. Amino acids *x* and *y* are aligned with frequency, q^N_{xy}
- Null hypothesis ($\hat{H}_0$)*:* σ and τ are unrelated. Amino acids *x* and *y* are aligned with background frequencies, $p_x p_y$

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

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

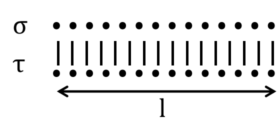
The relative entropy of a substitution matrix is given in bits per position and can be calculated from S^N using the equation

$$\mathcal{H}^N = \sum_{\{xy\}} q_{xy}^N \log_2 S^N[x,y]$$

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%

17

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, *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 %

19