End of Semester Logistics ...

Final exam

- Thursday, December 14th,
- 1:30-4:30, WEH 7500
  - Cumulative, emphasis on 2nd half
  - Closed book, 2 pages of notes

- Review session:
  - Wednesday, Dec 13th
  - Time and Location, TBA

  Bring questions!

End of Semester Logistics ...

Homework

- **711-4 due Tomorrow**
- **Problem Set 7 due at 11:59pm on Monday, 12/11**

- 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 7

- Run 9 Blast searches with different parameter settings
- Record some results in Tables (excel worksheet)
- Interpret in terms of Blast heuristics and Karlin Altschul stats

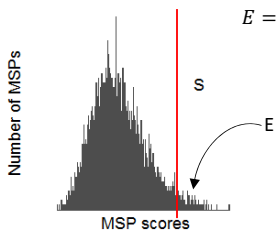
Recommendations:

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

PLEASE PLEASE PLEASE
FILL OUT FACULTY COURSE EVALUATIONS

4

Recall: BLAST (Karlin-Altschul) Statistics



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

Using bit score $S_b = \frac{\lambda S - lnK}{ln2}$

$E = mn2^{-S_b}$

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

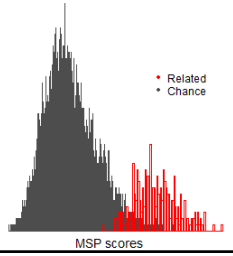
Maximal Segment Pair (MSP): an ungapped local alignment that cannot be improved by making it bigger or smaller.

- 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



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

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

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

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*

$$\sum_{\{H.T\}} q_i \log_2 \frac{q_i}{p_i} = q \log_2\left({}^{q}/_{0.5}\right) + (1-q)\log_2\left({}^{1-q}/_{0.5}\right)$$

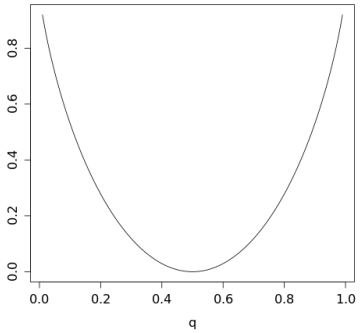
is the expected information available per coss of the coin to discriminate in favor of hypothesis $\hat{H}_a$ (biased coin) against hypothesis $\hat{H}_0$, (fair coin) if the coin is actually biased.

Relative Entropy for coin tosses

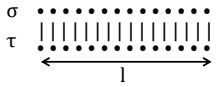
$$\sum_{\{H.T\}} q_i \log_2 \frac{q_i}{p_i} = q \log_2\left({}^{q}/_{0.5}\right) + (1-q)\log_2\left({}^{1-q}/_{0.5}\right)$$

Alternate hypothesis:
q = probability of H
1-q = probability of T

Null hypothesis:
probability of H = 0.5
probability of T = 0.5



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.

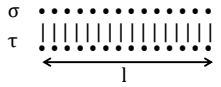
The relative entropy of a substitution matrix is given in bits per position and can be calculated from S^N using the equation

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

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%

15

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



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

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

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

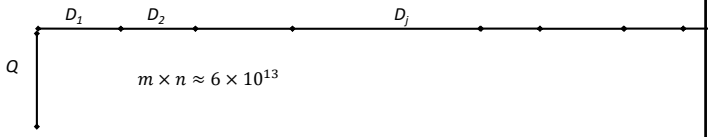
Suppose we seek matches with E values no greater than $E = 1$. Then, we require $\mathcal{S} \geq \log_2 m'n'$ bits. From this, we can estimate the minimum alignment length required to disinguish related from chance MSPs at N PAMS:

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

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

Given a query sequence, *Q,* of length *m,* and a database sequence, *D*, of length *n,* find all ungapped local alignments with score at least S_T

For sufficiently large *D,* dynamic programming is too slow.



	Non-redundant (nr) sequence database	
Sequences	Nucleic Acid	Amino Acid
Date:	Nov 24, 2023 7:56 AM	Nov 22, 2023 2:48 AM
Letters:	1,429,214,902,961	249,059,528,165
Sequences:	100,888,011	636,004,760

17

Implications

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

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

When $mn \approx 10^{13}$, *46* bits are required. Since the alignment cannot be longer than the query, a query sequence must be at least

46/2.57 = 18 residues long at ***30 PAMs***

46/0.70 = 68 residues long at ***160 PAMs***

46/0.36 = 127 residues long at ***250 PAMs***

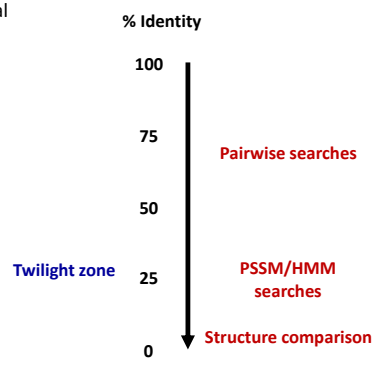
to distinguish significant HSP's from chance.

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 %

18

The "Twilight" Zone

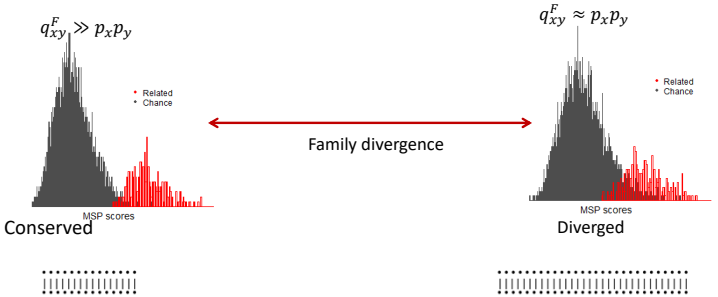
- The scale indicates % identity in local alignments (MSPs).
- The Twilight Zone
  - Around 20%-35% identity
  - Difficult to distinguish between MSPs in related sequences and "chance" alignments



19

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



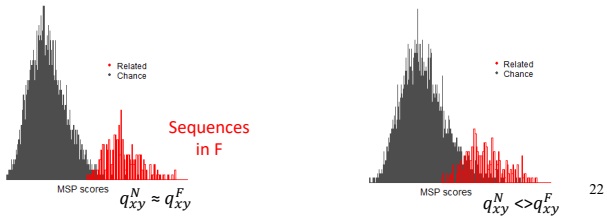
20

Choosing your scoring matrix

1. The lower the relative entropy, *H*, the longer the minimum alignment that is distinguishable from chance.
2. If your query is short, you will only be able to find closely related matches.
   - *Use PAM30*
3. *...*

21

- Each scoring matrix $S^N[x,y]$ has characteristic frequencies, q^N_{xy}
- Each protein family *F* has characteristic amino acid pair frequencies, q^F_{xy}
- The best results are obtained when q^N_{xy} is a good match for q^F_{xy}
- 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^N_{xy} is a good match for q^F_{xy}
- 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^N_{xy} is a good match for q^F_{xy}.



22

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
Y	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

Maxima highlighted in yellow

Best discrimination between related and chance MSPs :
Matrix divergence ~ Family divergence

23

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	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

= Efficiency ≥ 94%

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

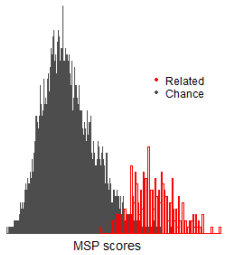
24

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

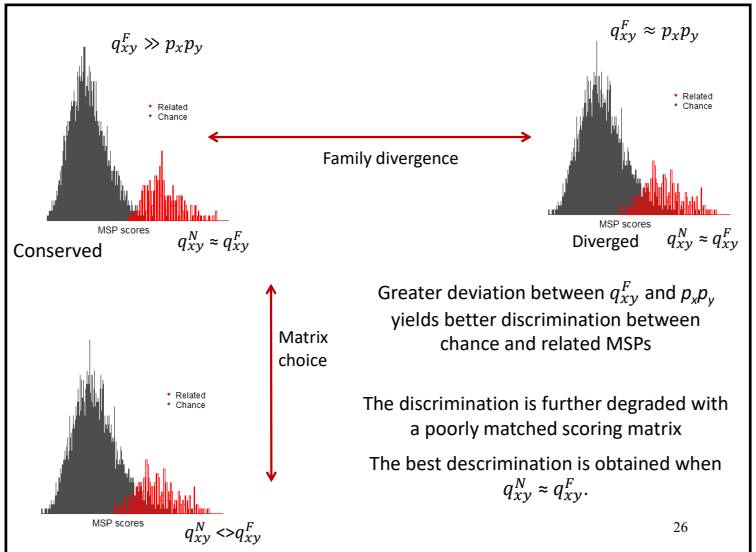
Implications

Scoring an alignment with a matrix that does not match the target frequencies characteristic of the query and sequences related to it, will result in lower MSP scores in related matches:

If the matrix does not match the target frequencies, the related (red) distribution will move to the left, increasing the overlap.



25



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 $q^N_{xy} \approx q^F_{xy}$.

26

Choosing your scoring matrix

1. The lower the relative entropy, *H*, the longer the minimum alignment that is distinguishable from chance.
2. If your query is short, you will only be able to find closely related matches.
   - *Use PAM30*
3. 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*

27