End of Semester Logistics

Last day of class: Thursday Dec 5th

PS7/711-4: due Saturday, Dec 7, 11:59pm

Review session: Sunday, Dec 8th, 3pm to 7pm, WH 5407

Problem Set 7 review

Your questions!

Final exam: Monday, Dec 9th, 1pm-4pm

– Cummulative, emphasis last 3rd of the semester
– Closed book, 4 pages of notes

End of Semester Logistics ...

Homework

- **Problem Set 7 due at 11:59pm on Sat 12/7**

- Late homework receives 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

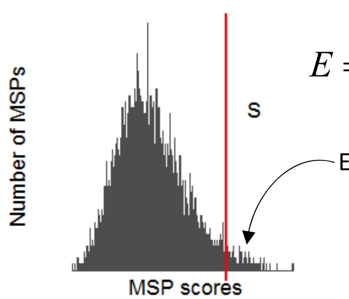
- Run 6 Blast searches with different parameter settings
- Record some results in Tables (excel worksheet)
- Interpret in terms of Blast heuristics and Karlin Altschul stats

Recommendations:

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

BLAST (Karlin-Altschul) Statistics

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



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

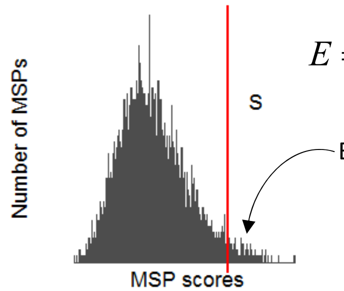
E-values depend on K and λ, which in turn depend on the scoring matrix, *S[i,j]*.

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 raw 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

E ~ the number of potential starting points for a high-scoring local alignment.

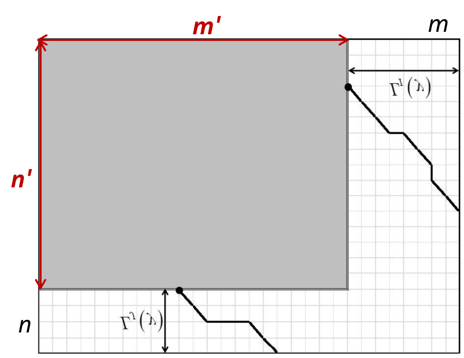
~~$\sim m \times n$, the number of cells in the alignment matrix~~

Not quite!

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

An alignment that starts too close to the end of s_1 or s_2 will not achieve a score of at least S_b

An alignment must start within the gray box to accrue a score of at least S_T before reaching the end of the sequence.



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

where the *m'* and *n'* are the "effective" lengths that reflect this edge correction

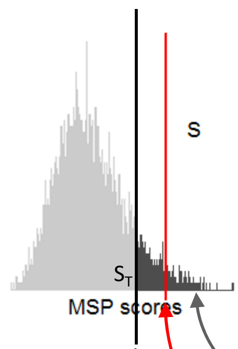
Blast software estimates the effective lengths automatically

7

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

BLAST (Karlin-Altschul) Statistics

Search Parameters	
Program	blastp
Word size	5
Expect value	0.05



- User selects threshold, E_T
- Score threshold is a function of 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.*

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	cysteine proteinase inhibitor 5 [Medicago truncatula]	Medicago…	94.0	94.0	98%	4e-22	43.59%	116	XP_013454866
☑	putative cysteine proteinase inhibitor 7 [Medicago truncatula]	Medicago…	88.2	[illegible]	[illegible]	8e-[illegible]	[illegible]	115	XP_003617412
☑	cysteine proteinase inhibitor 1-like [Vicia villosa]	Vicia villosa	86.7	86.7	98%	4e-19	44.44%	115	XP_058727946
☑	cysteine proteinase inhibitor 5-like [Momordica charantia]	Momordi…	87.0	87.0	76%	4e-19	46.15%	127	XP_022142402

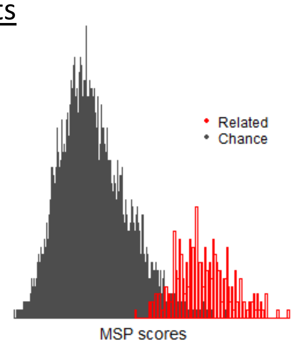
Bit score

E value

9

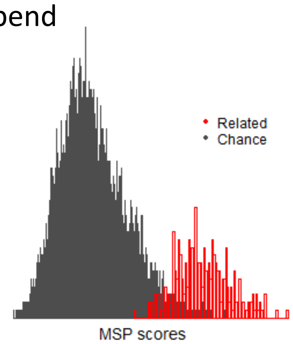
How much information is available to distinguish between chance MSPs and MSPs in related sequences?

- Information content of substitution matrices
- Which substitution matrix will maximize precision and recall?
- <u>Information content of alignments</u>



How much information is available to distinguish between chance MSPs and MSPs in related sequences?

False postives and false negatives depend on the overlap of the distributions of
- Chance MSPs
- MSPs in related sequences



What factors influence this overlap?

11

A warm-up thought experiment:

Alternate Hypothesis: Coin is biased
- $pr(Heads|H_A) = q,\ \ pr(Tails|H_A) = (1-q),$ *where* $q \neq 0.5$

Null Hypothesis: : Coin is fair
- $pr(Heads|H_0) = p,\ \ pr(Tails|H_0) = (1-p),$ *where* $p = 0.5$

How many coin tosses are required to decide if the coin is biased?

- If $q \gg 0.5$, then a short series of coin tosses is sufficient
- 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 an event space $\mathcal{E} = \{E_1 E_2 \cdots E_N\}$ and
probability distributions, P and Q, defined on $\mathcal{E}$:

$$Q = pr\{\mathcal{E}|H_A\} = \{q_1 q_2 \cdots q_N\}$$
$$P = pr\{\mathcal{E}|H_0\} = \{p_1 p_2 \cdots p_N\}$$

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

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

is the expected information provided by each observation to discriminate in favor of hypothesis H_A against hypothesis $H_{A,}$ when H_A is true.

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

Relative Entropy – coin toss example

Given an event space $\mathcal{E} = \{\boldsymbol{H}eads, \boldsymbol{T}ails\}$ and
probability distributions, P and Q, defined on $\mathcal{E}$:

$$Q = pr\{\mathcal{E}|H_A\} = \{q, 1-q\}, q \neq 0.5$$
$$P = pr\{\mathcal{E}|H_0\} = \{0.5, 0.5\}$$

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

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

is the expected information provided by each coin toss to discriminate in favor of hypothesis H_A *(bias)* against hypothesis H_0(fair) when H_A is true.

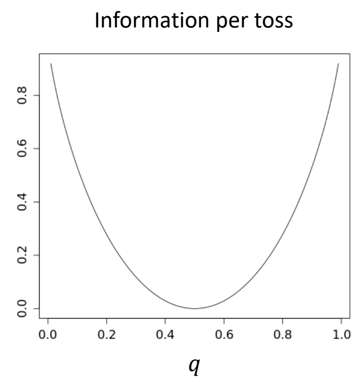
Note: the KL Divergence is not symmetric and therefore not a distance.

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 (H_A):
- probability of Heads: $q \neq 0.5$
- probability of Tails: $1 - q$

Null hypothesis (H_0):
- probability of Heads: 0.5
- probability of Tails: 0.5



Ungapped local alignments

```
T  Q  S  S  R  A  A  K  R  Y  S  V  C  S  L  ...
   |  +  |  +  +  |  |  |  +  |  +        |
E  Q  A  S  Q  S  A  K  R  W  S  L  A  G  L  ...
```

- Alternate hypothesis: Sequences are related at *N* PAMs divergence. Amino acids *x* and *y* are aligned with frequency, $q^N{}_{xy}$
- Null hypothesis: Sequences are unrelated. Amino acids *x* and *y* are aligned with background frequencies, $p_x p_y$

How many aligned sites are required to decide if the sequences are related or share chance similarity?

Relative Entropy – ungapped local alignments

```
T  Q  S  S  R  A  A  K  R  Y  S  V  C  S  L  ...
   |  +  |  +  +  |  |  |  +  |  +           |
E  Q  A  S  Q  S  A  K  R  W  S  L  A  G  L  ...
```

- Alternate hypothesis: Sequences are related at *N* PAMs divergence (q^N_{xy})
- Null hypothesis: Sequences are unrelated ($p_x p_y$)

The relative entropy

$$\mathcal{H} = \sum_{\{xy\}} q^N_{xy} \log_2 \frac{q^N_{xy}}{p_x p_y}$$

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

gives the number of bits per position available tzo 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%

20

Since PAM 30 has 2.57 bits of information per site, should you always score your alignments with PAM30?

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 %

NO!

27

Target frequencies q_{xy}^N

Empirical "target" frequencies:

frequency of x aligned with y in sequences related to the query

```
Query   ....... x ...... y ....... x .....
        ....... y ...... x ....... y .....
```

Theoretical "target" frequencies:

frequency of x aligned with y used to construct the matrix

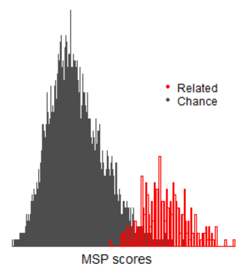
$$S^N[x, y] = \log_2 \frac{q_{xy}^N}{p_x p_y}$$

28

The highest alignment scores are obtained when the matrix target frequencies match the empirical target frequencies.

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.



29

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

30

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%

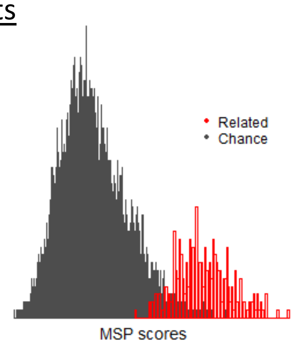
$$\text{Efficiency} = \frac{\text{Score with PAM } Y}{\text{Score with PAM } n}$$

32

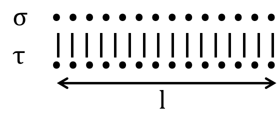
Altschul SF, J. Mol. Biol., 219, 555-565 (1991)

How much information is available to distinguish between chance MSPs and MSPs in related sequences?

- Information content of substitution matrices
- Which substitution matrix will maximize precision and recall?
- <u>Information content of alignments</u>



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



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

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

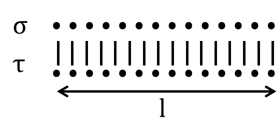
$$\mathcal{S} = \log_2 \frac{m'n'}{E}$$

Given the number of bits required, we can estimate the minimum alignment length required to disinguish related from chance MSPs at N PAMS:

$$l \cdot \mathcal{H}^N = \log_2(m'n'/E)$$

$$l = \frac{1}{\mathcal{H}^N} \log_2 \frac{m'n'}{E}$$

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



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

We obtain a rough estimate of the number of bits required as follows

- A typical protein query length is *m* = 150 amino acids
- As of Nov 24, the *nr* database contains roughly 322 billion amino acids
- The current default E value threshold is 0.05

Ignoring the edge correction for effective lengths, yields

$$\log_2 \frac{150 \cdot 322 \times 10^9}{0{,}05} \approx 50 \text{ bits}$$

2

Implications

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

$$l = \frac{1}{\mathcal{H}^N} \log_2 \frac{m'n'}{E}$$

Roughly *50* bits are required to find sequences related to a query. Since the alignment cannot be longer than the query, a query sequence must be at least

50/2.57 = 19 residues long at ***30 PAMs***

50/0.66 = 75 residues long at ***BLOSUM62***

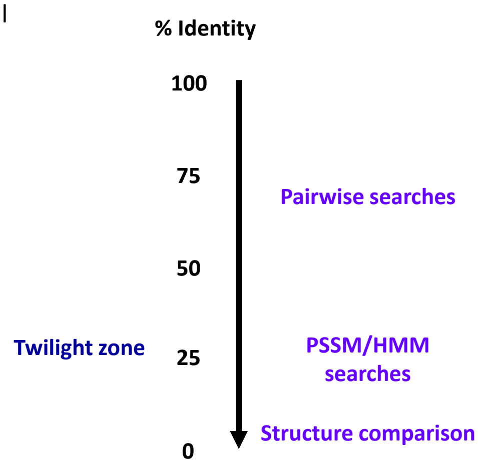
50/0.36 = 138 residues long at ***250 PAMs***

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 %

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



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*

49