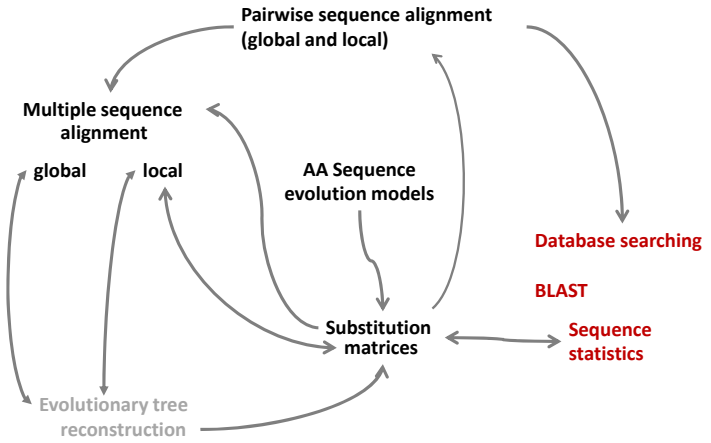
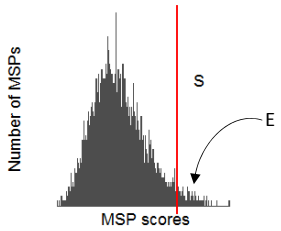
Blast review



1. BLAST (Karlin-Altschul) Statistics

When is a sequence retrieved in a database search statistically significant?



1. BLAST (Karlin-Altschul) Statistics

When is a sequence retrieved in a database search statistically significant?

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

$$E = Km'n' e^{-\lambda S}$$

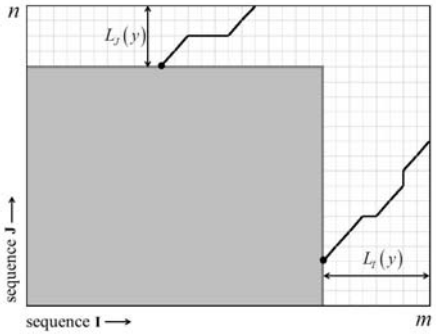
where K and λ are constants that depend on $S[i,j]$ and can be computed from the theory for any scoring function.

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

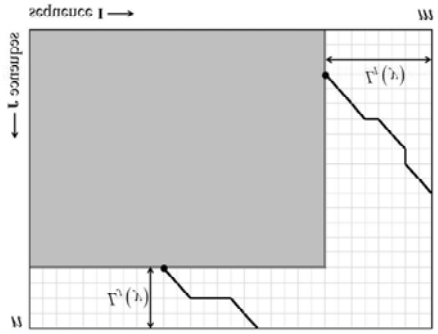
Length adjustment

Given a query sequence of length m and database of length n, to a first approximation, the size of the search space is mn.

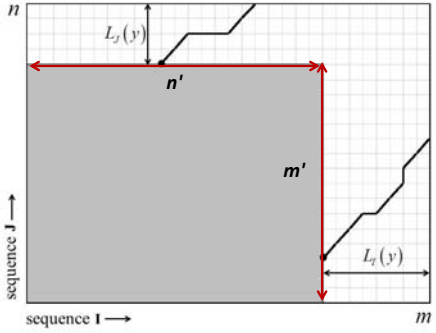
However, the database is not a single sequence but a set of concatenated sequences. An alignment cannot bridge the boundary between two database sequences. Furthermore, an alignment cannot start to close to the end of the query sequence. Both of these reduce the actual size of the search space.



"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



"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



"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

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

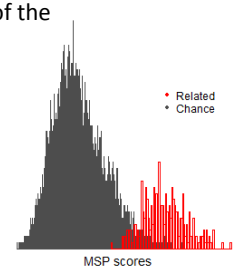
we obtain a "bit score," S_b, and an expression for *E* that is independent of *K*, λ and *S[i,j]*. $E = mn2^{-S_b'}$

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

3. Sensitivity and Specificity

FPs and FNs depend on the overlap of the distributions of

- Chance MSPs
- MSPs in related sequences



What factors influence this overlap?

3a. Target frequencies

Given sequences *Q* and *M:*

- Alternate hypothesis ($\hat{H}_a$)*:* *Q* and *M* are related with divergence…, *n*.
  - Residues *j* and *k* are aligned with "target" frequencies, q^Q_{jk}
- Null hypothesis ($\hat{H}_0$)*:* *s* and *b* are unrelated.
  - Residues *j* and *k* are aligned with background frequencies, $p_j p_k$

Note that the PAM and BLOSUM matrices were constructed by estimating q_{jk} from data. However, any scoring matrix (that satisfies the appropriate assumptions for Karlin Altschul statistics) can be expressed as a log odds matrix of the form

$$S^n[j,k] = \log \frac{q^n{}_{jk}}{p_j p_k}$$

"Theorem" (Karlin and Altschul, 1990)

The best scoring matrix for distinguishing significant alignments from chance alignments is the scoring matrix that gives the greatest difference in scores between related alignments and chance alignments. For sequences diverged by *n* PAMs, the best discrimination is obtained by

$$S^n[j,k] = \log \frac{q^n{}_{jk}}{p_j p_k}$$

the matrix corresponding to the characteristic target frequencies, q^Q_{jk} from related sequences at the evolutionary distance of interest.

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

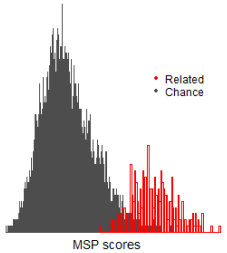
Maxima highlighted in yellow

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

Implications (1)

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.



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

= Efficiency ≥ 94%

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

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

Implications (2)

BLAST will give reasonable accuracy as long as the empirical target frequencies, q^{Q}_{jk}, in the alignments of interest do not deviate too far from the theoretical target frequencies for matrix *S*:

$$q_{jk} = p_j p_k e^{\lambda S[j,k]}$$

Reasonable accuracy can be achieved with two or three matrices.

3b. Information content of substitution matrices and alignments

How much information is there in an MSP that can be used to discriminate between

- related matches and
- chance matches?

This depends on

- The length of the MSP
- The target frequencies
- The scoring matrix

...

4. Information content of substitution matrices and alignments

Suppose you have a general framework with an alternate hypothesis $\hat{H}_a$ and a null hypothesis $\hat{H}_0$. Event *i* occurs with probability p_i under the null hypothesis and probability q_i under the alternate hypothesis The *relative entropy,*

$$H = \sum_i q_i \log \frac{q_i}{p_i}$$

is the expected discrimination information: the information available to discriminate in favor of hypothesis $\hat{H}_a$ against hypothesis $\hat{H}_0$, given $\hat{H}_a$ is true.

In the BLAST context, the relative entropy gives information available to distinguish related alignments at *n* PAMs ($\hat{H}_a$) from chance alignments ($\hat{H}_0$) given a particular substitution matrix $S^n[]$

$$H = \sum_{i,j} q^n_{ij} \log \frac{q^n_{ij}}{p_i p_j} = \sum_{i,j} q^n_{ij} S^n[i,j]$$

The relative entropy of a substitution matrix is given in bits per position and can be calculated from *S[]* using the equations

$$H = \sum_{j,k} q^n{}_{jk} S^n[j,k] \quad \text{and} \quad q^n{}_{jk} = p_j p_k e^{-\lambda S^n[j,k]}$$

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%

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***

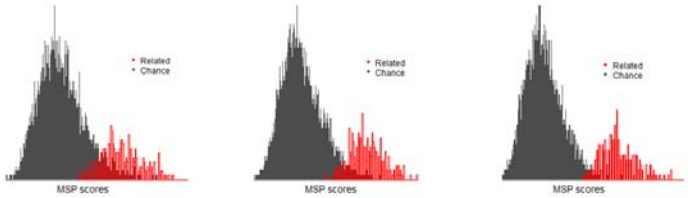
to distinguish significant HSP's from chance.

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 %

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.

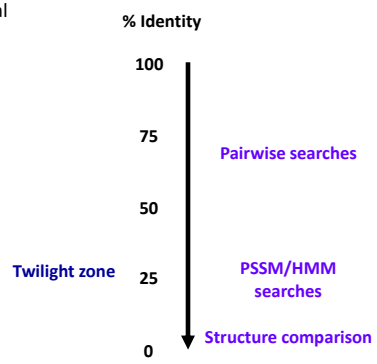
Greater deviation between q^n_{jk} and $p_j p_k$ yields fewer false positives and false negatives.

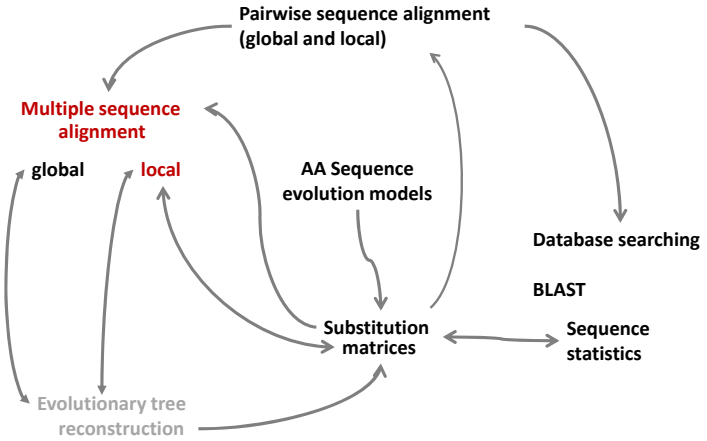


Diverged ⟷ Conserved

The "Twilight" Zone

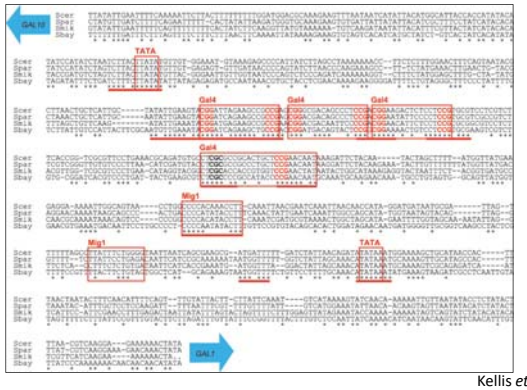
- The scale indicates % identity in local alignments (MSPs).
- The Twilight Zone
  - Around 20%-35% identity
  - Difficult to distinguish between MSPs in related sequences and "chance" alignments





Local MSA
Conserved patterns in biological sequences

Example: Transcription factor binding sites



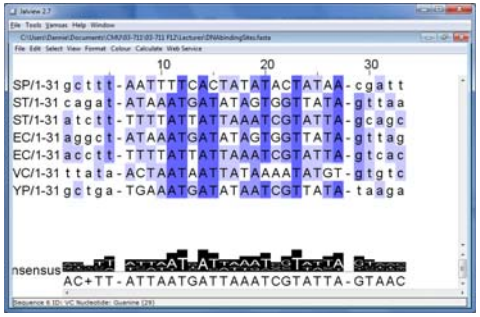
Kellis *et al, Nature*, 03

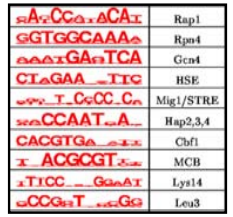
Local MSA
Conserved patterns in biological sequences

Example: Transcription factor binding sites

```
SP  ...gcttt AATTTTCACTATATACTATAA cgatt...
ST  ...cagat ATAAATGATATAGTGGTTATA gttaa...
ST  ...atctt TTTTATTATTAAATCGTATTA gcagc...
EC  ...aggct ATAAATGATATAGTGGTTATA gttag...
EC  ...acctt TTTTATTATTAAATCGTATTA gtcac...
VC  ...ttata ACTAATAATTATAAAATATGT gtgtc...
YP  ...gctga TGAAATGATATAATCGTTATA taaga...
```

…agcgagcctgagcactcgaggcatctctgcacatttagcatgggatgggccctgtatgcgcctgatga…



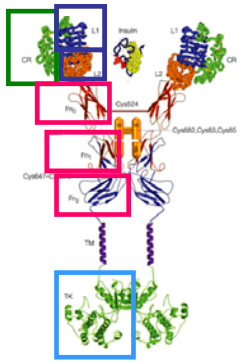


Some known binding site motifs

Applications of Local MSA
Conserved patterns in biological sequences

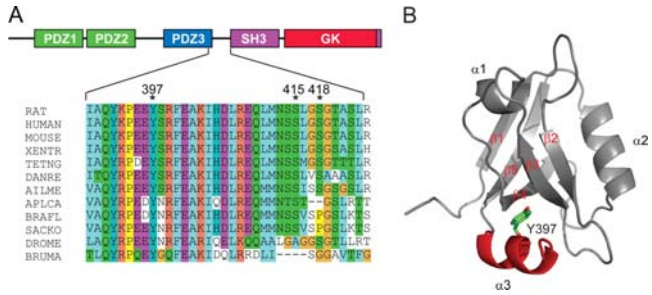
Example: Protein domains

- Fold independently
- Carry out specific functions
- Found in diverse contexts
- Conserved in evolution



Insulin receptor

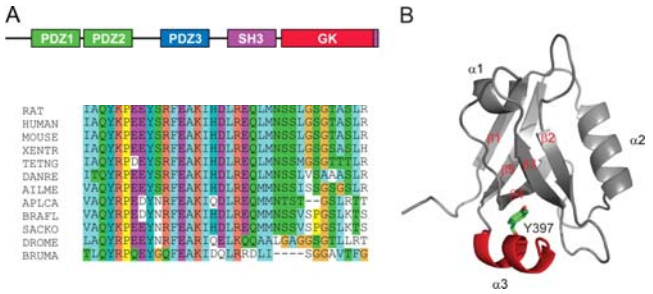
Domain architecture of PSD-95 and crystal structure of PDZ3.



Zhang J et al. J. Biol. Chem. 2011;286:41776-41785

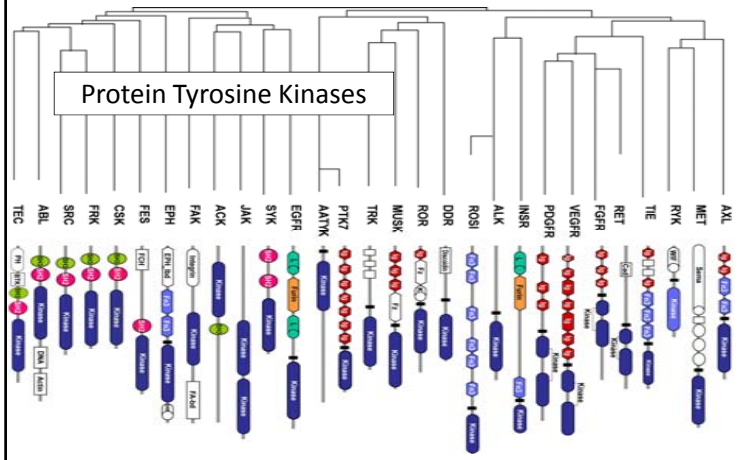
jbc

Domain architecture of PSD-95 and crystal structure of PDZ3.



Zhang J et al. J. Biol. Chem. 2011;286:41776-41785

jbc



Adapted from Robinson *et al*., 2000

Protein domain databases

Conserved Domain Database (CDD)

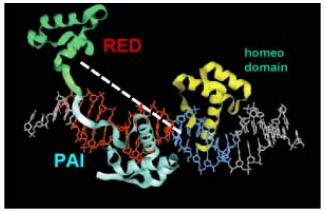
- Representation: Position specific scoring matrices (PSSMs)
- Structurally corrected local MSAs
- CDART: Conserved Domain Architecture Retrieval Tool

PFAM, SMART

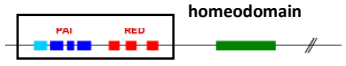
- Representation: Hidden Markov Models (HMM's)
- Curated local MSA's

Superfamily

- Curated HMMs based on SCOP structure database



Pax structure



paired domain

Pax domain architecture

http://www.gene-regulation.com/info/pax.html

```
  1 mphnsirsgh gglnqlggaf vngrplpevv rqrivdlahq gvrpcdisrq lrvshgcvsk
 61 ilgryyetgs irpgviggsk pkvatpkvve kigdykrqnp tmfaweirdr llaegvcdnd
121 tvpsvssinr iirtkvqqpf nlpmdscvat kslspghtli pssavtppes pqsdslgsty
181 singllgiaq pgndnkrkmd dsdqdscrls idsqssssgp rkhlrtdtfs qhhlealecp
241 ferqhypeay aspshtkgeq glyplpllns alddgkatlt ssntplgrnl sthqtypvva
301 dphspfaikq etpelsssss tpsslsssaf ldlqqvgsgg pagasvppfn afphaasvyg
361 qftgqallsg remvgptlpg ypphiptsgq gsyassaiag mvagseysgn ayshtpyssy
            421 seawrfpnss llsspyyyss tsrpsappts atafdhl
```

paired box gene 8 [Mus musculus]

gi|6754990|ref|NP_035170.1|[6754990]

CDART: Conserved Domain Architecture Retrieval Tool