

# Alignment Statistics and Substitution Matrices

BMI/CS 576

[www.biostat.wisc.edu/bmi576/](http://www.biostat.wisc.edu/bmi576/)

Mark Craven

[craven@biostat.wisc.edu](mailto:craven@biostat.wisc.edu)

Fall 2011

## Probabilistic model of alignments

- we'll focus on protein alignments without gaps
- given an alignment, we can consider two possibilities
  - R**: the sequences are related by evolution
  - U**: the sequences are unrelated
- How can we distinguish these possibilities?
- How is this view related to amino-acid substitution matrices?

## Model for *unrelated* sequences

- we'll assume that each position in the alignment is sampled randomly from some distribution of amino acids
- let  $q_a$  be the probability of amino acid  $a$
- the probability of an  $n$ -character alignment of  $x$  and  $y$  is given by

$$\Pr(x, y | U) = \prod_{i=1}^n q_{x_i} \prod_{i=1}^n q_{y_i}$$

## Model for *related* sequences

- we'll assume that each pair of aligned amino acids evolved from a common ancestor
- let  $p_{ab}$  be the probability that evolution gave rise to amino acid  $a$  in one sequence and  $b$  in another sequence
- the probability of an alignment of  $x$  and  $y$  is given by

$$\Pr(x, y | R) = \prod_{i=1}^n p_{x_i y_i}$$

## Probabilistic model of alignments

- How can we decide which possibility ( $U$  or  $R$ ) is more likely?
- one principled way is to consider the relative likelihood of the two possibilities

$$\frac{\Pr(x, y | R)}{\Pr(x, y | U)} = \frac{\prod_i p_{x_i y_i}}{\prod_i q_{x_i} \prod_i q_{y_i}} = \frac{\prod_i p_{x_i y_i}}{\prod_i q_{x_i} q_{y_i}}$$

- taking the log, we get

$$\log \frac{\Pr(x, y | R)}{\Pr(x, y | U)} = \sum_i \log \left( \frac{p_{x_i y_i}}{q_{x_i} q_{y_i}} \right)$$

## Probabilistic model of alignments

- the score for an alignment is thus given by:

$$S = \sum_i s(x_i, y_i) = \log \frac{\Pr(x, y | R)}{\Pr(x, y | U)}$$

- the substitution matrix score for the pair  $a, b$  should thus be given by:

$$s(a, b) = \log \left( \frac{p_{ab}}{q_a q_b} \right)$$

# Substitution matrices

- two popular sets of matrices for protein sequences
  - PAM matrices [Dayhoff *et al.*, 1978]
  - BLOSUM matrices [Henikoff & Henikoff, 1992]
- both try to capture the the relative substitutability of amino acid pairs in the context of evolution

## Blosum 62 matrix

BLOSUM62																					
A	4																				
R	-1	5																			
N	-2	0	6																		
D	-2	-2	1	6																	
C	0	-3	-3	-3	9																
Q	-1	1	0	0	-3	6															
E	-1	0	0	2	-4	2	5														
G	0	-2	0	-1	-3	-2	-2	6													
H	-2	0	1	-1	-3	0	0	-2	8												
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4											
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4										
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5									
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5								
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6							
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7						
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4					
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5				
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11			
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7		
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4	
X	0	-1	-1	-1	-2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-2	0	0	-2	-1	-1	-1
A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	X	

## Substitution matrices

- the substitution matrix score for the pair  $a, b$  is given by:

$$s(a, b) = \log\left(\frac{p_{ab}}{q_a q_b}\right)$$

- but how do we get values for  $p_{ab}$  (probability that  $a$  and  $b$  arose from a common ancestor)?
- it depends on how long ago sequences diverged

diverged recently:

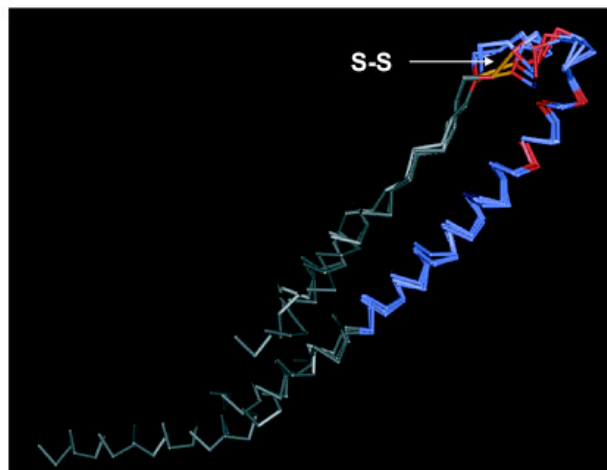
$$p_{ab} \approx 0 \text{ for } a \neq b$$

diverged long ago:

$$p_{ab} \approx q_a q_b$$

## Substitution matrices

- key idea: trusted alignments of related sequences provide information about biologically permissible mutations
- protein structure similarity provides the gold standard for which alignments are trusted



```

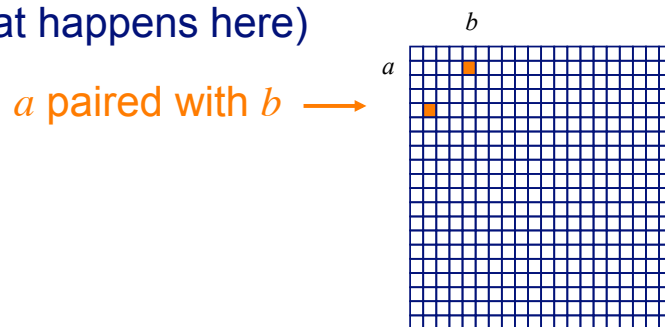
D2/CEL/OR sdvqAISSTIQDLQDQVDSLAEVVLQNRRLDLLTAEQGGIILALQEKfcfyank
MMLV  ddlrEVEKSI SNLEKSLTSLSEVVLQNRRLDLLFLKEGGLSALKEEcfayad~
HTLV-1 kdisQLTQAIVKNHKNLLKIAQYAAQNRRLDLLFWEQGGIILKALQECcflnit
Ebola  qlanETTQALQLFLRATTELRTFSILNRKAIDFLLQRWGGTTHILGPDoriephd
  
```

# BLOSUM matrices

- [Henikoff & Henikoff, *PNAS* 1992]
- probabilities estimated from “blocks” of sequence fragments that represent *structurally* conserved regions in proteins
- transition frequencies observed directly by identifying blocks that are at least
  - 45% identical (BLOSUM-45)
  - 50% identical (BLOSUM-50)
  - 62% identical (BLOSUM-62)
  - etc.

# BLOSUM matrices

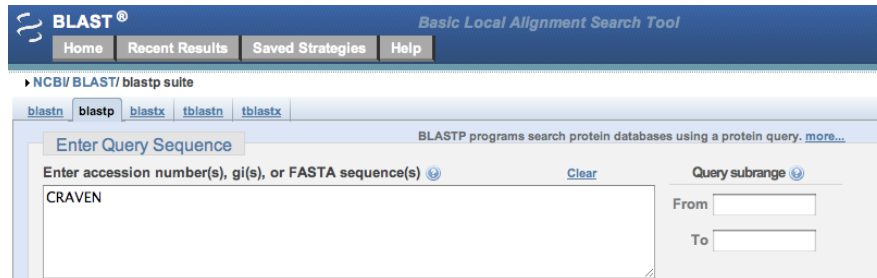
- given: a set of sequences in a block
- fill in matrix  $A$  with number of observed substitutions (we won't worry about details of some normalization that happens here)



$$p_{ab} = \frac{A_{ab}}{\sum_{c,d} A_{cd}} \quad q_a = \frac{\sum_b A_{ab}}{\sum_{c,d} A_{cd}}$$

# Statistics of alignment scores

Any BLAST query will return some results (local alignments). How do we assess whether an alignment provides good evidence for homology?



## Sequences producing significant alignments:

Accession	Description	Max score	Total score	Query coverage	E value	Links
<a href="#">XP_002258225.1</a>	hypothetical protein in Plasmodium species [Plasmodium kno	<a href="#">23.1</a>	23.1	100%	224	<a href="#">G</a>
<a href="#">EG162193.1</a>	Toll-like protein 2 [Acromyrmex echinator]	<a href="#">23.1</a>	23.1	100%	225	
<a href="#">YP_001144220.1</a>	hypothetical protein ASA_P4G174 [Aeromonas salmonicida su	<a href="#">23.1</a>	23.1	100%	225	<a href="#">G</a>
<a href="#">ACG58392.1</a>	envelope glycoprotein [Human immunodeficiency virus 1] >g	<a href="#">23.1</a>	23.1	100%	225	
<a href="#">ADI62543.1</a>	envelope glycoprotein [Human immunodeficiency virus 1]	<a href="#">23.1</a>	23.1	100%	225	
<a href="#">ADZ36241.1</a>	envelope glycoprotein [Human immunodeficiency virus 1]	<a href="#">23.1</a>	23.1	100%	225	
<a href="#">XP_003388655.1</a>	PREDICTED: zygotic DNA replication licensing factor mcm6-B	<a href="#">23.1</a>	23.1	100%	225	<a href="#">GM</a>
<a href="#">XP_001544753.1</a>	predicted protein [Ajellomyces capsulatus NAM1] >gb EDN05	<a href="#">23.1</a>	23.1	100%	225	
<a href="#">XP_002473859.1</a>	predicted protein [Postia placenta Mad-698-R] >gb EED80970	<a href="#">23.1</a>	23.1	100%	225	<a href="#">G</a>
<a href="#">XP_001542370.1</a>	predicted protein [Ajellomyces capsulatus NAM1] >gb EDN05	<a href="#">23.1</a>	42.8	100%	225	<a href="#">G</a>
<a href="#">ZP_08133473.1</a>	2-isopropylmalate synthase [Kingella denitrificans ATCC 3339	<a href="#">23.1</a>	23.1	100%	226	
<a href="#">ZP_04603208.1</a>	hypothetical protein GCWU000324_02693 [Kingella oralis ATC	<a href="#">23.1</a>	23.1	100%	226	
<a href="#">XP_001612415.1</a>	adrenodoxin reductase [Plasmodium vivax SaI-1] >gb EDL42	<a href="#">23.1</a>	23.1	100%	226	<a href="#">G</a>
<a href="#">XP_001615509.1</a>	adrenodoxin reductase [Plasmodium vivax SaI-1] >gb EDL45	<a href="#">23.1</a>	23.1	100%	226	<a href="#">G</a>

# Statistics of alignment scores

**Q:** How do we assess whether an alignment provides good evidence for homology?

**A:** determine how likely it is that such an alignment score would result from chance.

3 ways to calculate chance; look at alignment scores for

- real but non-homologous sequences
- real sequences shuffled to preserve compositional properties
- sequences generated randomly based upon a DNA/protein sequence model

# Statistics of alignment scores

BLAST returns an *E value*: the expected number of alignments with score at least *S* due to chance



Sequences producing significant alignments:

Accession	Description	Max score	Total score	Query coverage	E value	Links
<a href="#">XP_002258225.1</a>	hypothetical protein in Plasmodium species [Plasmodium kn...	<a href="#">23.1</a>	23.1	100%	224	<a href="#">G</a>
<a href="#">EGI62193.1</a>	Tolloid-like protein 2 [Acromyrmex echinator]	<a href="#">23.1</a>	23.1	100%	225	
<a href="#">YP_001144220.1</a>	hypothetical protein ASA_P4G174 [Aeromonas salmonicida s...	<a href="#">23.1</a>	23.1	100%	225	<a href="#">G</a>
<a href="#">ACG58392.1</a>	envelope glycoprotein [Human immunodeficiency virus 1] >g...	<a href="#">23.1</a>	23.1	100%	225	
<a href="#">ADI62543.1</a>	envelope glycoprotein [Human immunodeficiency virus 1]	<a href="#">23.1</a>	23.1	100%	225	
<a href="#">ADZ36241.1</a>	envelope glycoprotein [Human immunodeficiency virus 1]	<a href="#">23.1</a>	23.1	100%	225	
<a href="#">XP_003388655.1</a>	PREDICTED: zygotic DNA replication licensing factor mcm6-B	<a href="#">23.1</a>	23.1	100%	225	<a href="#">GM</a>
<a href="#">XP_001544753.1</a>	predicted protein [Ajellomyces capsulatus NAM1] >gb EDN03...	<a href="#">23.1</a>	23.1	100%	225	<a href="#">G</a>
<a href="#">XP_002473859.1</a>	predicted protein [Postia placenta Mad-698-R] >gb EED80970...	<a href="#">23.1</a>	23.1	100%	225	<a href="#">G</a>
<a href="#">XP_001542370.1</a>	predicted protein [Ajellomyces capsulatus NAM1] >gb EDN05...	<a href="#">23.1</a>	42.8	100%	225	<a href="#">G</a>
<a href="#">ZP_08133473.1</a>	2-isopropylmalate synthase [Kingella denitrificans ATCC 3339...	<a href="#">23.1</a>	23.1	100%	226	
<a href="#">ZP_04603208.1</a>	hypothetical protein GCWU000324_02693 [Kingella oralis ATCC...	<a href="#">23.1</a>	23.1	100%	226	
<a href="#">XP_001612415.1</a>	adrenodoxin reductase [Plasmodium vivax SaI-1] >gb EDL42...	<a href="#">23.1</a>	23.1	100%	226	<a href="#">G</a>
<a href="#">XP_001615509.1</a>	adrenodoxin reductase [Plasmodium vivax SaI-1] >gb EDL45...	<a href="#">23.1</a>	23.1	100%	226	<a href="#">G</a>

# Statistics of alignment scores

Another example query...

Sequences producing significant alignments:

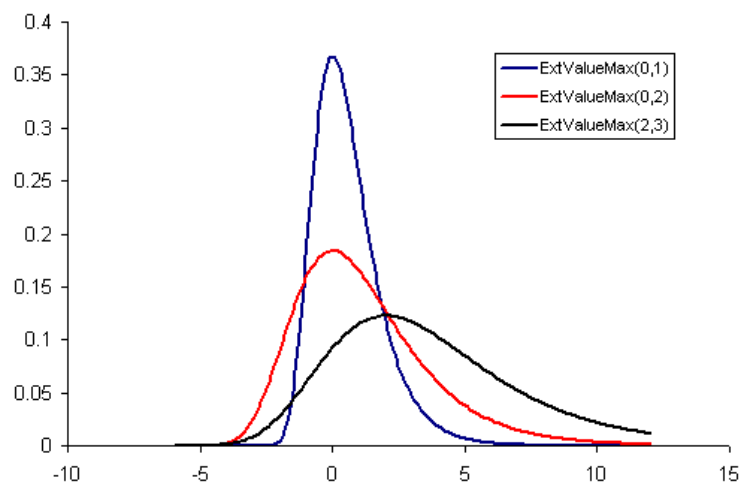
Accession	Description	Max score	Total score	Query coverage	E value	Links
<a href="#">EGI62193.1</a>	Tolloid-like protein 2 [Acromyrmex echinator]	<a href="#">70.6</a>	70.6	100%	6e-14	
<a href="#">EFN76162.1</a>	Tolloid-like protein 2 [Harpegnathos saltator]	<a href="#">35.4</a>	35.4	90%	0.16	
<a href="#">EGC46999.1</a>	predicted protein [Ajellomyces capsulatus H88]	<a href="#">33.7</a>	33.7	85%	0.62	
<a href="#">EER41112.1</a>	predicted protein [Ajellomyces capsulatus H143]	<a href="#">33.7</a>	33.7	85%	0.62	
<a href="#">ABG37112.1</a>	ataxia-telangiectasia mutated protein splice variant 7 [Sus scrofa]	<a href="#">31.6</a>	31.6	57%	3.5	<a href="#">GM</a>
<a href="#">EFY99627.1</a>	putative ZIP zinc transporter [Metarhizium anisopliae ARSEF...	<a href="#">31.6</a>	31.6	42%	3.5	
<a href="#">NP_001116552.1</a>	serine-protein kinase ATM [Sus scrofa] >sp Q6PQD5.2 ATM_...	<a href="#">31.6</a>	71.1	100%	3.5	<a href="#">UGM</a>
<a href="#">YP_001567419.1</a>	DNA-directed RNA polymerase subunit beta' [Petrotoga mobi...	<a href="#">31.2</a>	31.2	90%	5.0	<a href="#">G</a>

## Scores from random alignments

- suppose we assume
  - sequence lengths  $m$  and  $n$
  - a particular substitution matrix and amino-acid frequencies
- and we consider generating random sequences of lengths  $m$  and  $n$  and finding the best alignment of these sequences
- this will give us a distribution over alignment scores for random pairs of sequences

## The extreme value distribution

- but we're picking the best alignments, so we want to know what the distribution of max scores for alignments against a random set of sequences looks like
- this is given by an *extreme value distribution*



## Distribution of scores

- the expected number of alignments,  $E$ , with score at least  $S$  is given by:

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

- $S$  is a given score threshold
- $m$  and  $n$  are the lengths of the sequences under consideration
- $K$  and  $\lambda$  are constants that can be calculated from
  - the substitution matrix
  - the frequencies of the individual amino acids

## Statistics of alignment scores

- to generalize this to searching a database, have  $n$  represent the summed length of the sequences in the DB (adjusting for edge effects)
- the NCBI BLAST server does just this
- theory for *gapped* alignments not as well developed
- computational experiments suggest this analysis holds for gapped alignments (but  $K$  and  $\lambda$  must be estimated from data)