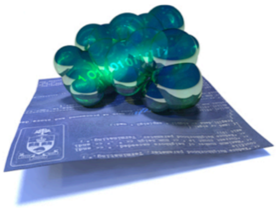
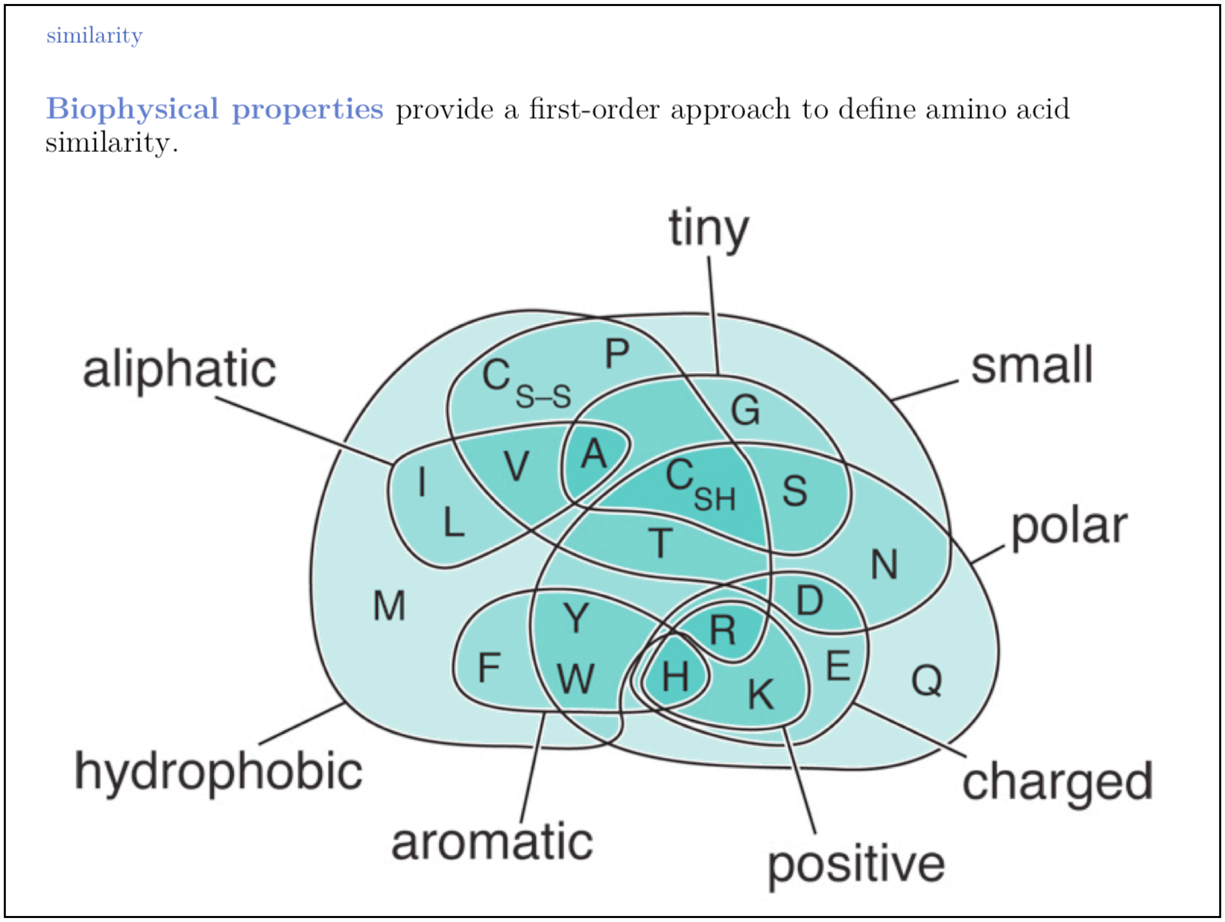
A
Bioinformatics
Course

Amino Acid Similarity



Boris Steipe

department of biochemistry – department of molecular genetics
University of Toronto



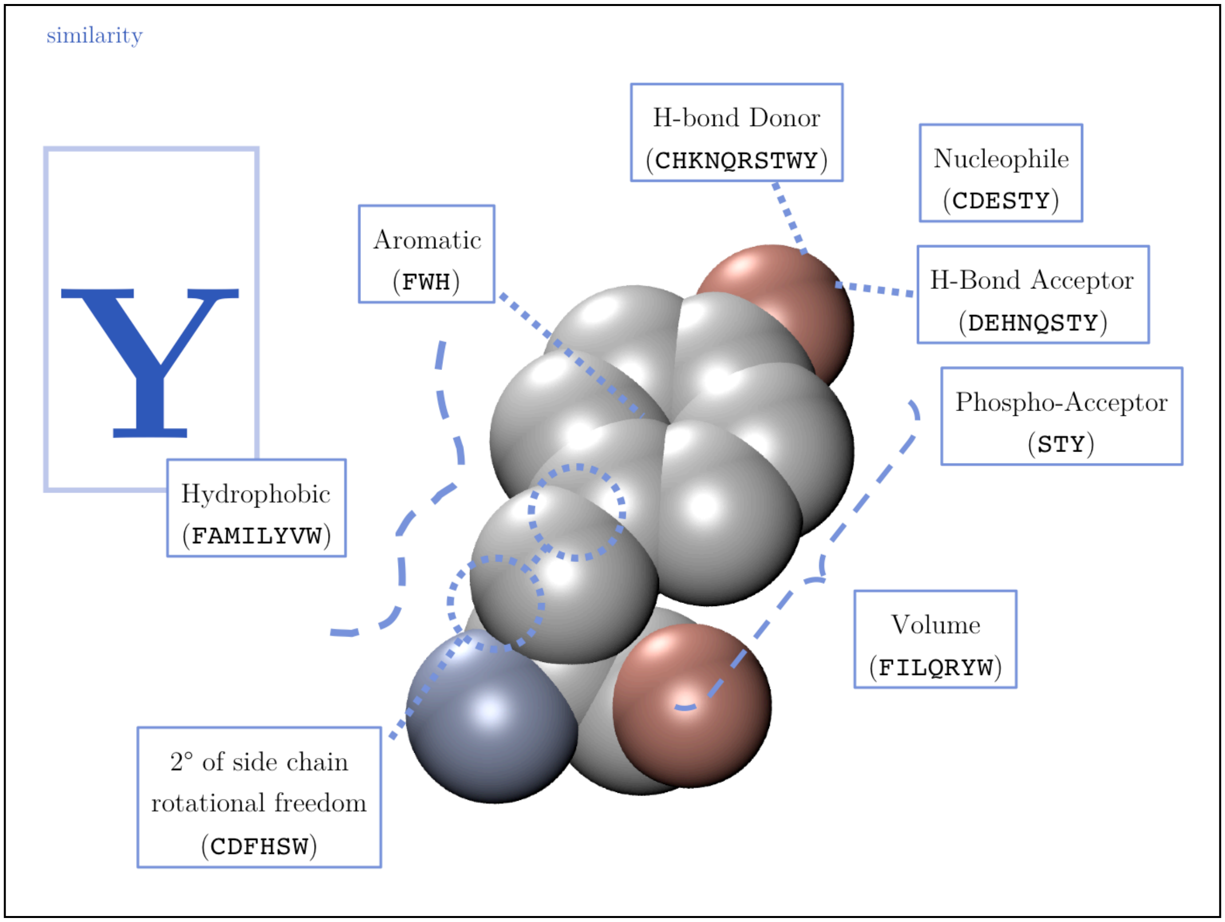
similarity

Biophysical properties provide a first-order approach to define amino acid similarity.

To measure the quality of a sequence alignment, we need to define some way to quantify the similarity of two amino acids. Whether two amino acids contribute similar stability or function to a folded protein depends on their precise context.

This Venn diagram (originally going back to Willie Taylor) provides a good first aproximation to summarize shared sidechain properties and to estimate amino acid similarity.

Note that "C" appears twice in this sketch: once as cysteine (C_{SH}) with its free thiol function, once as the disulfide bonded cystine ($C_{S\text{-}S}$). These two forms have very different properties.



similarity

Y

Aromatic (FWH)

H-bond Donor (CHKNQRSTWY)

Nucleophile (CDESTY)

H-Bond Acceptor (DEHNQSTY)

Phospho-Acceptor (STY)

Hydrophobic (FAMILYVW)

Volume (FILQRYW)

2° of side chain rotational freedom (CDFHSW)

As an example consider which amino acids are “similar” to tyrosine.

Which amino acid(s) we regard as being similar to tyrosine depends on which property we are considering. There are many properties that one can quantify, all of them imply a different set of “similar” amino acids, and no obvious strategy exists how to combine properties such as eg. hydrophobicity and volume into a single metric, as a similarity score for an amino acid pair.

The problem:

Amino acids can have multiple functions.
Which function is important, is determined by **context**.
What is more, context may influence the function.

Quantifying similarity in *sequences* implies a measure based only on pairs of amino acids, independent of the context!

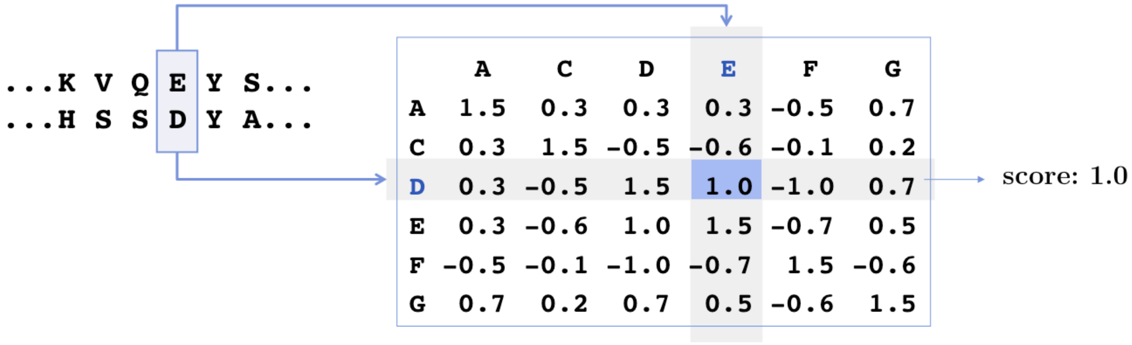
Example:
pK of Side Chain: charge is determined by environment.
E.g. a-helix dipole can easily shift pK by ± 2 pH units ...

pK	AA	
3.9	D	ASP
4.4	E	GLU
6.5	H	HIS
9.2	C	CYS
10.1	Y	TYR
10.5	K	LYS
12.0	R	ARG

(TJ Creighton, Proteins.
2.ed. Freeman, NY 1993)

Models of amino acid similarity can be quantified in a scoring matrix

Scoring matrix:
define similarity of each amino acid with each possible aligned amino acid ...
(also: "similarity matrix", "mutation matrix", "substitution matrix" ...)

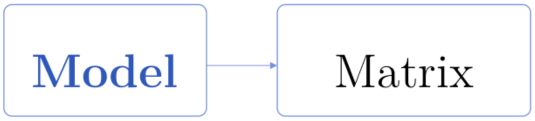


```
...K V Q E Y S...
...H S S D Y A...
```

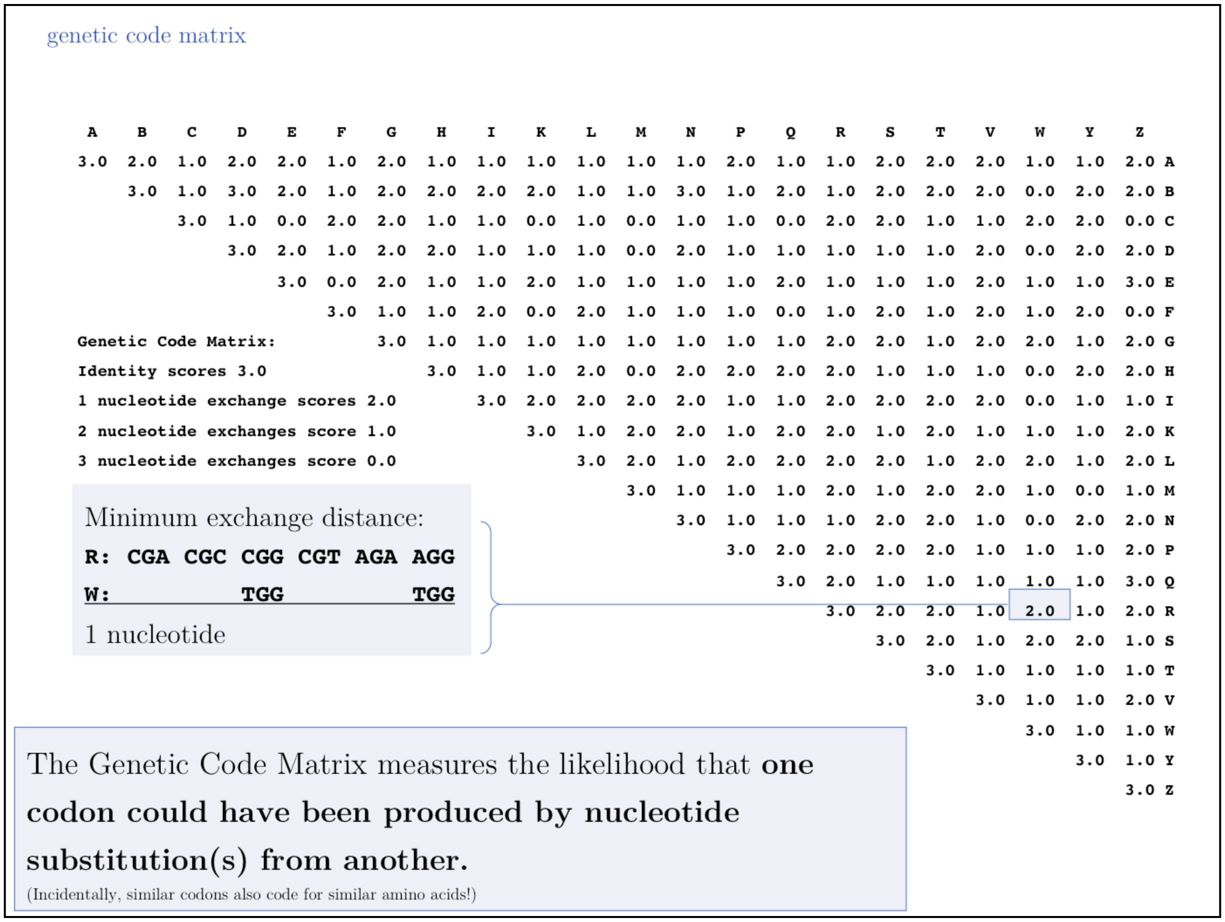
	A	C	D	E	F	G
A	1.5	0.3	0.3	0.3	-0.5	0.7
C	0.3	1.5	-0.5	-0.6	-0.1	0.2
D	0.3	-0.5	1.5	1.0	-1.0	0.7
E	0.3	-0.6	1.0	1.5	-0.7	0.5
F	-0.5	-0.1	-1.0	-0.7	1.5	-0.6
G	0.7	0.2	0.7	0.5	-0.6	1.5

score: 1.0

A scoring matrix is a formal representation of a specific model of similarity !



Ignore similarity, use only identity	**Identity Matrix**
Biophysical similarity	**Biophysical Similarity Matrix**
Required nucleotide substitutions	**Genetic Code Matrix**
Let nature decide: evolution by point mutations	**PAM (Dayhoff) Matrices**
Let nature decide: amino acids in similar context	**BLOSUM Matrices**



genetic code matrix

```
 A     B     C     D     E     F     G     H     I     K     L     M     N     P     Q     R     S     T     V     W     Y     Z
3.0   2.0   1.0   2.0   2.0   1.0   2.0   1.0   1.0   1.0   1.0   1.0   1.0   2.0   1.0   1.0   2.0   2.0   2.0   1.0   1.0   2.0 A
      3.0   1.0   3.0   2.0   1.0   2.0   2.0   2.0   2.0   1.0   1.0   3.0   1.0   2.0   1.0   2.0   2.0   2.0   0.0   2.0   2.0 B
            3.0   1.0   0.0   2.0   2.0   1.0   1.0   0.0   1.0   0.0   1.0   1.0   0.0   2.0   2.0   1.0   1.0   2.0   2.0   0.0 C
                  3.0   2.0   1.0   2.0   2.0   1.0   1.0   1.0   0.0   2.0   1.0   1.0   1.0   1.0   1.0   2.0   0.0   2.0   2.0 D
                        3.0   0.0   2.0   1.0   1.0   2.0   1.0   1.0   1.0   1.0   2.0   1.0   1.0   1.0   2.0   1.0   1.0   3.0 E
                              3.0   1.0   1.0   2.0   0.0   2.0   1.0   1.0   1.0   0.0   1.0   2.0   1.0   2.0   1.0   2.0   0.0 F
Genetic Code Matrix:                3.0   1.0   1.0   1.0   1.0   1.0   1.0   1.0   1.0   2.0   2.0   1.0   2.0   2.0   1.0   2.0 G
Identity scores 3.0                       3.0   1.0   1.0   2.0   0.0   2.0   2.0   2.0   2.0   1.0   1.0   1.0   0.0   2.0   2.0 H
1 nucleotide exchange scores 2.0                3.0   2.0   2.0   2.0   2.0   1.0   1.0   2.0   2.0   2.0   2.0   0.0   1.0   1.0 I
2 nucleotide exchanges score 1.0                      3.0   1.0   2.0   2.0   1.0   2.0   2.0   1.0   2.0   1.0   1.0   1.0   2.0 K
3 nucleotide exchanges score 0.0                            3.0   2.0   1.0   2.0   2.0   2.0   2.0   1.0   2.0   2.0   1.0   2.0 L
                                                                  3.0   1.0   1.0   1.0   2.0   1.0   2.0   2.0   1.0   0.0   1.0 M
                                                                        3.0   1.0   1.0   1.0   2.0   2.0   1.0   0.0   2.0   2.0 N
                                                                              3.0   2.0   2.0   2.0   2.0   1.0   1.0   1.0   2.0 P
                                                                                    3.0   2.0   1.0   1.0   1.0   1.0   1.0   3.0 Q
                                                                                          3.0   2.0   2.0   1.0   2.0   1.0   2.0 R
                                                                                                3.0   2.0   1.0   2.0   2.0   1.0 S
                                                                                                      3.0   1.0   1.0   1.0   1.0 T
                                                                                                            3.0   1.0   1.0   2.0 V
                                                                                                                  3.0   1.0   1.0 W
                                                                                                                        3.0   1.0 Y
                                                                                                                              3.0 Z
```

Minimum exchange distance:

```
R:  CGA CGC CGG CGT AGA AGG
W:          TGG         TGG
```

1 nucleotide

The Genetic Code Matrix measures the likelihood that **one codon could have been produced by nucleotide substitution(s) from another.**

(Incidentally, similar codons also code for similar amino acids!)

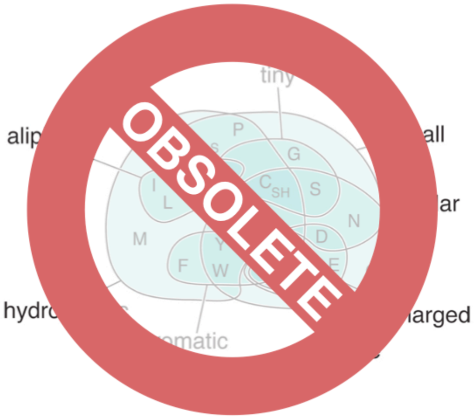
A scoring matrix can be used to quantify how well a given model is represented in two aligned sequences. Here the model says: two amino acids are similar, if it is easy to change one codon into the other by single nucleotide substitutions. For very closely related sequences, this is actually not a bad metric. And it captures an intriguing property of the genetic code: being robust against mutations in the sense that the biophysical properties tend to be conserved between similar codons.

Any biophysical property of amino amino acids can be turned into such a scoring matrix. However, whether amino acids are likely to be paired in a correct alignment of natural sequences is not well described by any single biophysical property, and there is no abvious way how to weight their combinations.

similarity: Dayhoff model

M. Dayhoff

A quantitative model of evolution:



Similarity can be defined as the empirical probability that two amino acids can substitute for each other during evolution!

This makes speculations about about amino acid similarity based on first principles unnecessary.

The Dayhoff model of evolution postulates a quantitative model of the likelihood of specific amino acid substitutions as a consequence of evolution, based on the empirical observation of variation in related protein sequences. This rejects a definition of amino acid similarity from first principles in favor of an empirical approach.

similarity: Dayhoff model

M. Dayhoff (1978): A quantitative model of evolution:

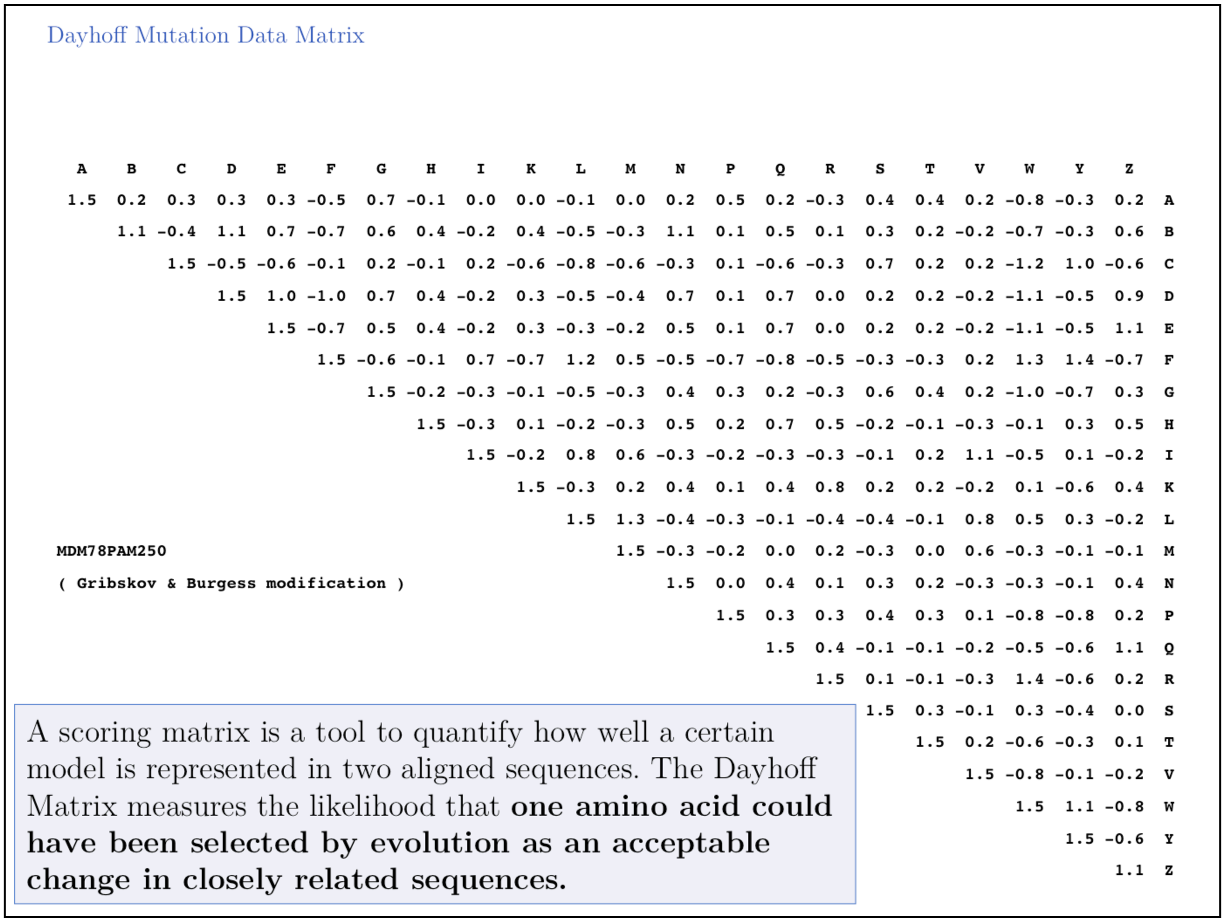
1. Construct a complete **phylogenetic tree** including all ancestral sequences.
2. For each of the observed and inferred sequences, the amino acid pair exchanges are tabulated into a 20x20 matrix (**Pair Exchange Frequency Matrix**).
3. Compile table of global **frequencies of occurrence**.
4. Tabulate, how often an amino acid is mutated into a different one (**relative mutabilities**).
5. Calculate the expectation value for the event that amino acid i will be replaced by amino acid j in a natural mutation (**Mutation Probability Matrix**).
6. **Scale** for evolutionary distance. (Necessary, because the matrix represents mutation probabilities derived from exactly one mutation per sequence.)
7. Compare P(event through mutation) with P(event by random chance) (**Relatedness Odds Matrix**).
8. Calculate **Log Odds Matrix** (Useful, because sequential independent events are as probable as the product of their individual probabilities, thus scores from sums of log-probabilites represent aggregate probabilities.)

When we apply the resulting matrix, **the sum of amino acid pair-scores in an alignment quantifies the probability that the sequences are related!**

The model takes into account the observed changes in a set of closely related sequences for which all current and ancestral states can be inferred. It then normalizes the observed frequency of change with the overall likelihood of mutation, which is different for different amino acids – due to their unique properties as well as their unequal number of codons. This gives – for any observed change – the probability that the change has occurred in the sample of related sequences, i.e. as a **consequence of evolution**.

We can also calculate the probability that a change has occurred **due to random chance**: this is simply governed by the frequency of the target amino acid. For example a *random* change from leucine to methionine (2.4% of database residues) is almost three times less likely than a change to glutamic acid (6.8% of database residues).

Comparing the likelihood of an evolutionary change with the likelihood of a random change gives us the “odds” that the two sequences in which the change was observed are related. For example the *mutation probability* of Met to Glu is quite low since these amino acids have very different properties.



Dayhoff Mutation Data Matrix

A	B	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	Z	
1.5	0.2	0.3	0.3	0.3	-0.5	0.7	-0.1	0.0	0.0	-0.1	0.0	0.2	0.5	0.2	-0.3	0.4	0.4	0.2	-0.8	-0.3	0.2	A
	1.1	-0.4	1.1	0.7	-0.7	0.6	0.4	-0.2	0.4	-0.5	-0.3	1.1	0.1	0.5	0.1	0.3	0.2	-0.2	-0.7	-0.3	0.6	B
		1.5	-0.5	-0.6	-0.1	0.2	-0.1	0.2	-0.6	-0.8	-0.6	-0.3	0.1	-0.6	-0.3	0.7	0.2	0.2	-1.2	1.0	-0.6	C
			1.5	1.0	-1.0	0.7	0.4	-0.2	0.3	-0.5	-0.4	0.7	0.1	0.7	0.0	0.2	0.2	-0.2	-1.1	-0.5	0.9	D
				1.5	-0.7	0.5	0.4	-0.2	0.3	-0.3	-0.2	0.5	0.1	0.7	0.0	0.2	0.2	-0.2	-1.1	-0.5	1.1	E
					1.5	-0.6	-0.1	0.7	-0.7	1.2	0.5	-0.5	-0.7	-0.8	-0.5	-0.3	-0.3	0.2	1.3	1.4	-0.7	F
						1.5	-0.2	-0.3	-0.1	-0.5	-0.3	0.4	0.3	0.2	-0.3	0.6	0.4	0.2	-1.0	-0.7	0.3	G
							1.5	-0.3	0.1	-0.2	-0.3	0.5	0.2	0.7	0.5	-0.2	-0.1	-0.3	-0.1	0.3	0.5	H
								1.5	-0.2	0.8	0.6	-0.3	-0.2	-0.3	-0.3	-0.1	0.2	1.1	-0.5	0.1	-0.2	I
									1.5	-0.3	0.2	0.4	0.1	0.4	0.8	0.2	0.2	-0.2	0.1	-0.6	0.4	K
										1.5	1.3	-0.4	-0.3	-0.1	-0.4	-0.4	-0.1	0.8	0.5	0.3	-0.2	L
											1.5	-0.3	-0.2	0.0	0.2	-0.3	0.0	0.6	-0.3	-0.1	-0.1	M
												1.5	0.0	0.4	0.1	0.3	0.2	-0.3	-0.3	-0.1	0.4	N
													1.5	0.3	0.3	0.4	0.3	0.1	-0.8	-0.8	0.2	P
														1.5	0.4	-0.1	-0.1	-0.2	-0.5	-0.6	1.1	Q
															1.5	0.1	-0.1	-0.3	1.4	-0.6	0.2	R
																1.5	0.3	-0.1	0.3	-0.4	0.0	S
																	1.5	0.2	-0.6	-0.3	0.1	T
																		1.5	-0.8	-0.1	-0.2	V
																			1.5	1.1	-0.8	W
																				1.5	-0.6	Y
																					1.1	Z

MDM78PAM250

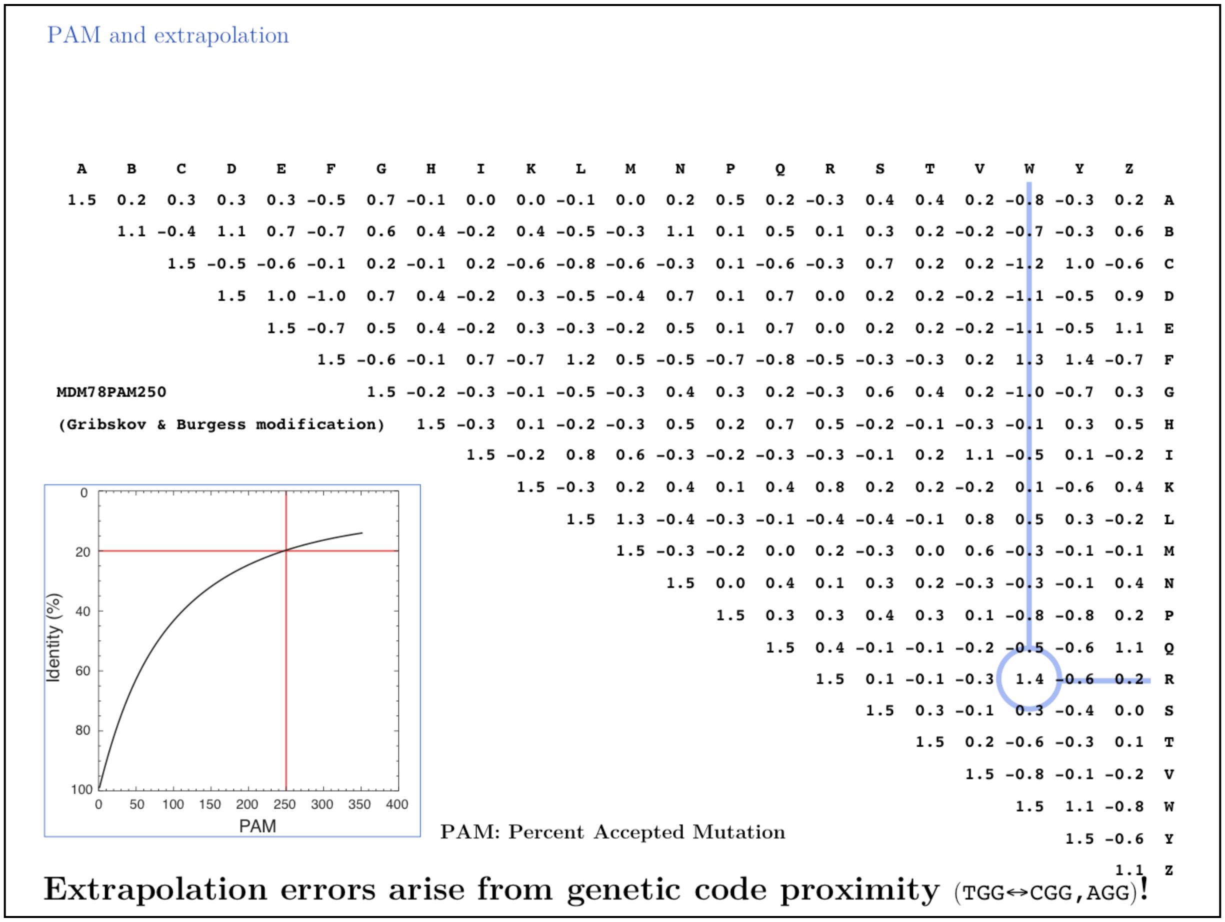
(Gribskov & Burgess modification)

A scoring matrix is a tool to quantify how well a certain model is represented in two aligned sequences. The Dayhoff Matrix measures the likelihood that **one amino acid could have been selected by evolution as an acceptable change in closely related sequences.**

MDM78PAM250 is a frequently used mutation data matrix. It is the Margret Dayhoff Model of 1978, extrapolated to a Percent Accepted Mutation rate of 250.

But the matrix as used in many alignment tools does not actually give the original numbers: it has been modified to score all identities the same (i.e. 1.5, which is IMO a big source of alignment problems), and it has been abbreviated to easily map to integers – both changes were done to speed up computation which was a big concern at the time these matrices were written.

This approach has been superseded.



PAM and extrapolation

MDM78PAM250

(Gribskov & Burgess modification)

A	B	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	Z	
1.5	0.2	0.3	0.3	0.3	-0.5	0.7	-0.1	0.0	0.0	-0.1	0.0	0.2	0.5	0.2	-0.3	0.4	0.4	0.2	-0.8	-0.3	0.2	A
	1.1	-0.4	1.1	0.7	-0.7	0.6	0.4	-0.2	0.4	-0.5	-0.3	1.1	0.1	0.5	0.1	0.3	0.2	-0.2	-0.7	-0.3	0.6	B
		1.5	-0.5	-0.6	-0.1	0.2	-0.1	0.2	-0.6	-0.8	-0.6	-0.3	0.1	-0.6	-0.3	0.7	0.2	0.2	-1.2	1.0	-0.6	C
			1.5	1.0	-1.0	0.7	0.4	-0.2	0.3	-0.5	-0.4	0.7	0.1	0.7	0.0	0.2	0.2	-0.2	-1.1	-0.5	0.9	D
				1.5	-0.7	0.5	0.4	-0.2	0.3	-0.3	-0.2	0.5	0.1	0.7	0.0	0.2	0.2	-0.2	-1.1	-0.5	1.1	E
					1.5	-0.6	-0.1	0.7	-0.7	1.2	0.5	-0.5	-0.7	-0.8	-0.5	-0.3	-0.3	0.2	1.3	1.4	-0.7	F
						1.5	-0.2	-0.3	-0.1	-0.5	-0.3	0.4	0.3	0.2	-0.3	0.6	0.4	0.2	-1.0	-0.7	0.3	G
							1.5	-0.3	0.1	-0.2	-0.3	0.5	0.2	0.7	0.5	-0.2	-0.1	-0.3	-0.1	0.3	0.5	H
								1.5	-0.2	0.8	0.6	-0.3	-0.2	-0.3	-0.3	-0.1	0.2	1.1	-0.5	0.1	-0.2	I
									1.5	-0.3	0.2	0.4	0.1	0.4	0.8	0.2	0.2	-0.2	0.1	-0.6	0.4	K
										1.5	1.3	-0.4	-0.3	-0.1	-0.4	-0.4	-0.1	0.8	0.5	0.3	-0.2	L
											1.5	-0.3	-0.2	0.0	0.2	-0.3	0.0	0.6	-0.3	-0.1	-0.1	M
												1.5	0.0	0.4	0.1	0.3	0.2	-0.3	-0.3	-0.1	0.4	N
													1.5	0.3	0.3	0.4	0.3	0.1	-0.8	-0.8	0.2	P
														1.5	0.4	-0.1	-0.1	-0.2	-0.5	-0.6	1.1	Q
															1.5	0.1	-0.1	-0.3	1.4	-0.6	0.2	R
																1.5	0.3	-0.1	0.3	-0.4	0.0	S
																	1.5	0.2	-0.6	-0.3	0.1	T
																		1.5	-0.8	-0.1	-0.2	V
																			1.5	1.1	-0.8	W
																				1.5	-0.6	Y
																					1.1	Z

PAM: Percent Accepted Mutation

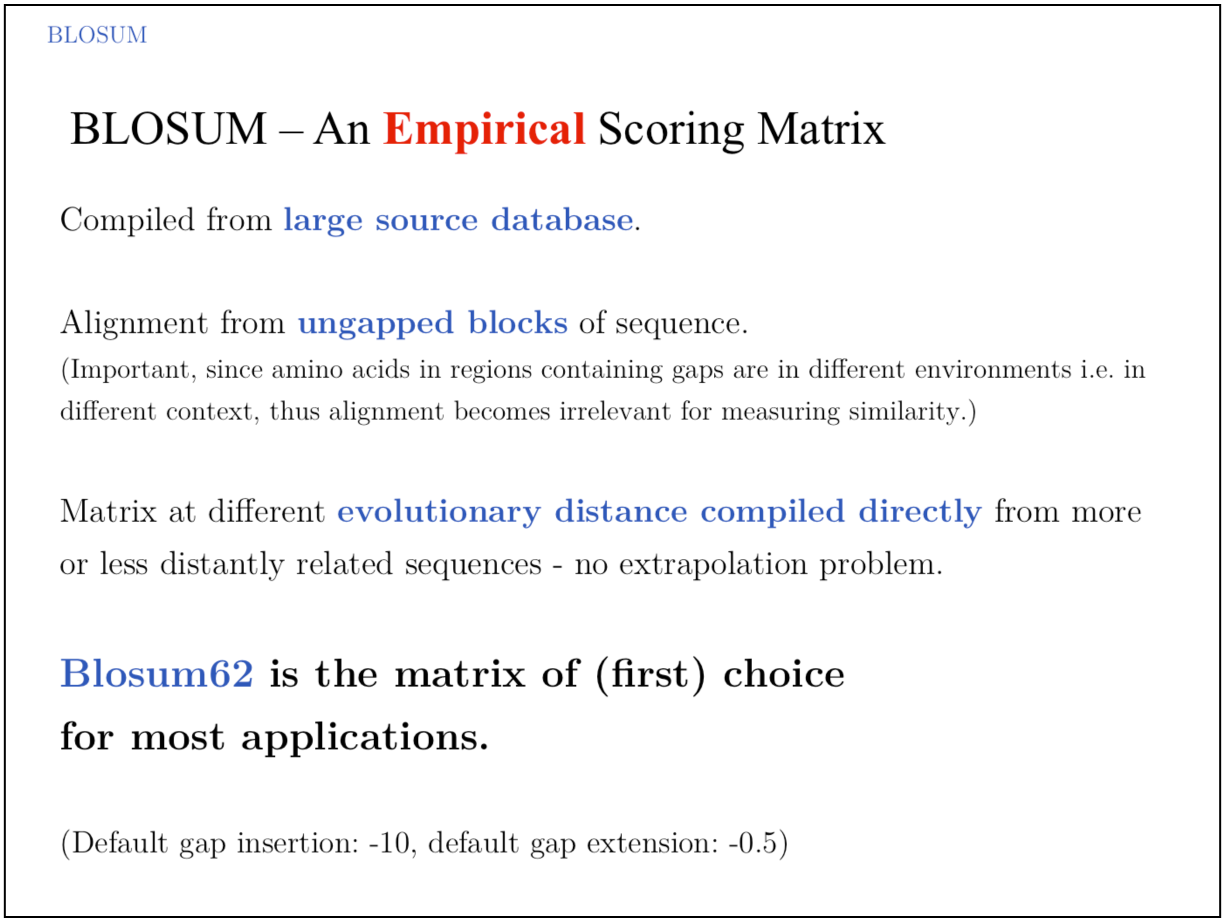
Extrapolation errors arise from genetic code proximity (`TGG`↔`CGG,AGG`)**!**

PAM 250 means: 250 accepted changes in the evolution of 100 amino acids of sequence: Percent Accepted Mutations. It expresses the evolutionary distance for which the matrix best describes the likelihood of relatedness. But how can the value of Percent Accepted Mutations be more than 100?

Mutations are located randomly in the sequence, therefore some amino acids may be hit several times and others never at all. Moreover, once an amino acid is changed, it may still revert to its original state through a second mutation. It is easy to see that even with very, very many mutations it is virtually impossible to arrive at a sequence that is 100% different from the original sequence.

As the graph inset shows, PAM250 corresponds to about sequence 20% identity.

Extrapolation to large PAM distances has problems. For example, since Arg and Trp have similar codons (`_GG`), an R→W mutation is quite likely at the very close evolutionary distances of the proteins in the Dayhoff dataset. It is also quite likely that evolution will favor secondary mutations at that site, to introduce an amino acid that is biophysically more compatible, and theR→W becomes unlikely in more distantly related pairs. But in the Dayhoff model, where large evolutionary distances are extrapolated by repeatedly multiplying the matrix with itself, that discrepancy gets amplified and as a result the pairscore of R→W is almost as high as an identity.



BLOSUM – An **Empirical** Scoring Matrix

Compiled from **large source database**.

Alignment from **ungapped blocks** of sequence.
(Important, since amino acids in regions containing gaps are in different environments i.e. in different context, thus alignment becomes irrelevant for measuring similarity.)

Matrix at different **evolutionary distance compiled directly** from more or less distantly related sequences - no extrapolation problem.

Blosum62 is the matrix of (first) choice for most applications.

(Default gap insertion: -10, default gap extension: -0.5)

To address the extrapolation problem, Steve Henikoff compiled matrices directly from blocks of ungapped alignments of sequences at given evolutionary distances, once a sufficient number of such sequences were available in the databases. These are the BLOSUM matrices.

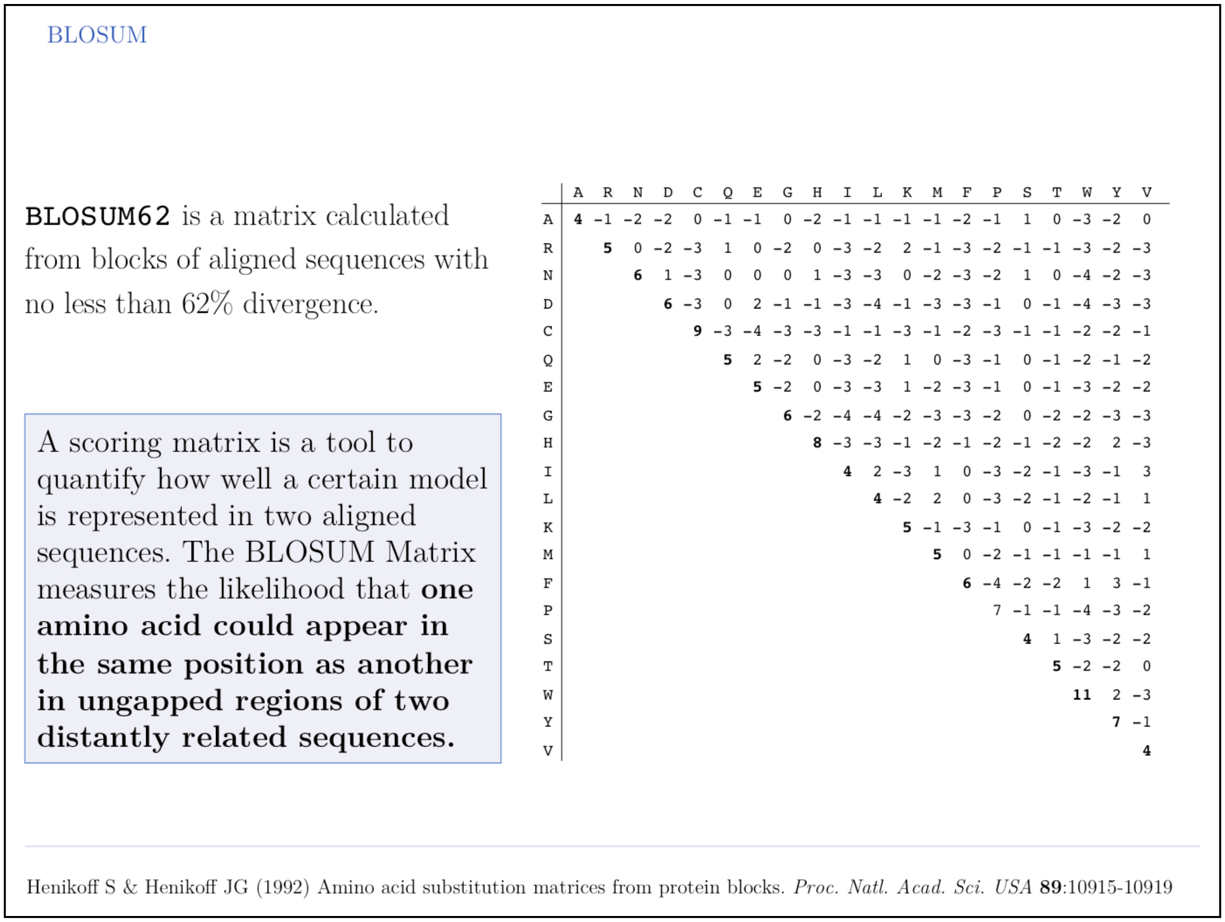
BLOSUM62 is a matrix compiled from sequences of not more than 62% identity. It corresponds approximately to a PAM160 matrix and appears to be the most sensitive choice to search for just barely detectably related sequence pairs.

Use BLOSUM62 unless you have a well understood reason not to.

Henikoff, S.; Henikoff, J.G. (1992). Amino Acid Substitution Matrices from Protein Blocks. PNAS **89**:10915–10919.

Eddy, S: (2004), *Nat Biotechnol.* **8**:1035-1036

See also: http://en.wikipedia.org/wiki/BLOSUM (Good article!)



BLOSUM

BLOSUM62 is a matrix calculated from blocks of aligned sequences with no less than 62% divergence.

A scoring matrix is a tool to quantify how well a certain model is represented in two aligned sequences. The BLOSUM Matrix measures the likelihood that **one amino acid could appear in the same position as another in ungapped regions of two distantly related sequences.**

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

Henikoff S & Henikoff JG (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA* **89**:10915-10919

Note that the R→W pairscore of BLOSUM62 is very much more in line with our biological intuition.

The matrix has been scaled to integers, for ease of computation. Also, its overall expectation value is negative, so we can't increase alignment scores by randomly adding pairs. This is important for *local alignments*. Finally, as we would expect, the score of residue identities depends on the nature of the residue: e.g. C, H, or W identities are (and should be) more significant than A or L.

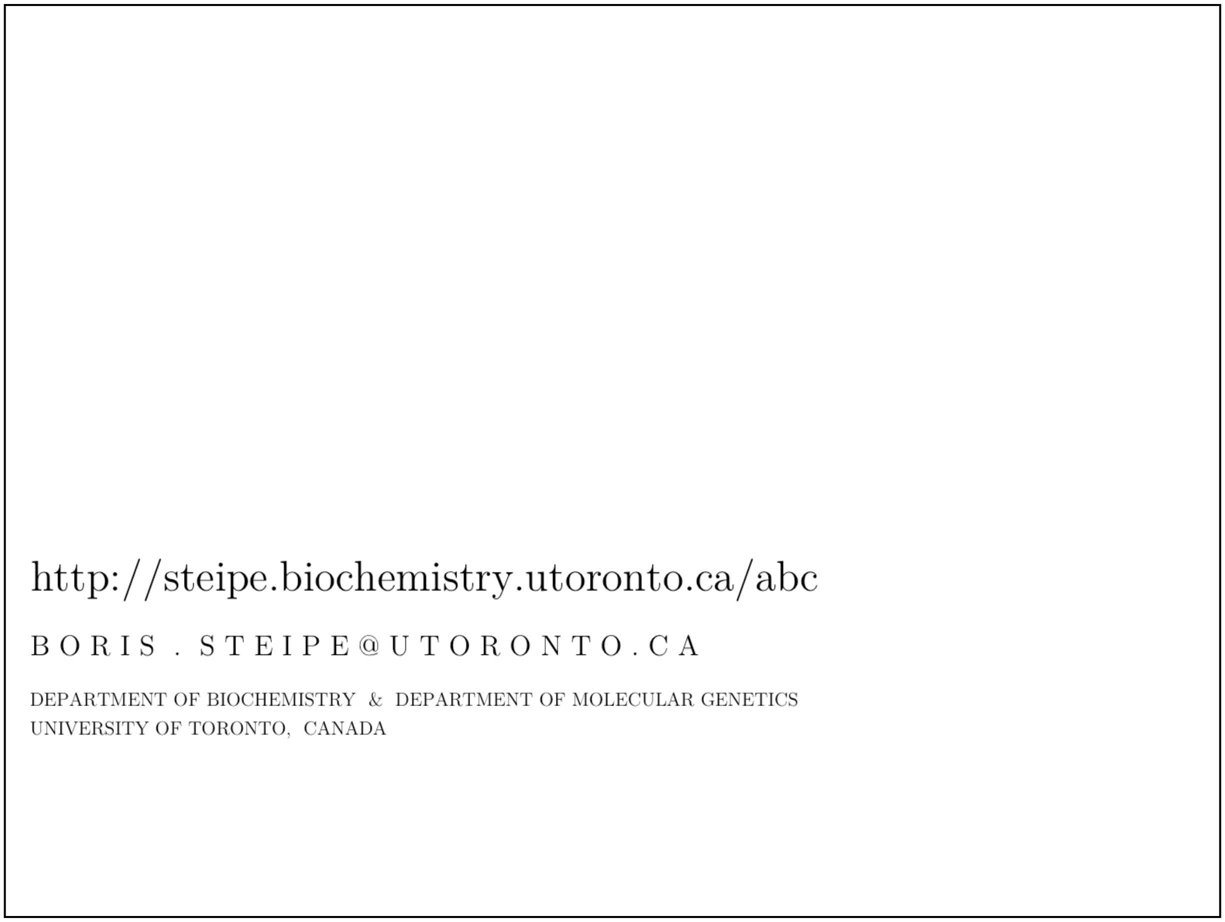
To repeat:

A scoring matrix represents a model of amino acid relatedness.

PAM Matrices measure the likelihood that one amino acid could have been selected by evolution as an acceptable change in closely related sequences.

BLOSUM matrices measure the likelihood that one amino acid could appear in the same position as another in ungapped regions of two distantly related sequences.

That is not exactly the same.



http://steipe.biochemistry.utoronto.ca/abc

BORIS.STEIPE@UTORONTO.CA

DEPARTMENT OF BIOCHEMISTRY & DEPARTMENT OF MOLECULAR GENETICS
UNIVERSITY OF TORONTO, CANADA