BIO 285/CSCI 285/MATH 285
Bioinformatics
Programming Lecture 8
Pairwise Sequence Alignment 2
And Python Function

Instructor: Lei Qian
Fisk University

Alignment

Measures of Sequence Similarity

Alignment with dot plot can just give us a general impression of the similarity between sequences. Defining and calculating quantitative measurement of sequences similarity and difference are important.

Measurement of Similarity:

For two sequences s1 and s2, we need to define their distance.
d(s1, s2)
-- The greater the distance, the less similar between these two sequences.
-- d(s,s) = 0

Alignment

Hamming Distance:
Defined between two strings of equal length, is the number of positions with mismatching characters.

```
agtc          Hamming distance = 2
cgta

agtctgtca     Hamming distance = 5
gatctctgc
```

Alignment

Hamming Distance:

Python code to calculate Hamming Distance:

```
s1 = "agtctgtca"
s2 = "gatctctgc"

distance = 0
for i in range(len(s1)):
      if s1[i]!=s2[i]:  #compare i-th letter
                        #of s1 and s2
            distance += 1

print distance
```

Python Functions

Hamming Distance:

Using function to calculate Hamming Distance:

```
def hamming_distance(s1, s2):
    distance = 0
    for i in range(len(s1)):
        if s1[i]!=s2[i]:  #compare i-th letter
                          #of s1 and s2
            distance += 1
    return distance

d = hamming_distance("agtctgtca", "gatctctgc")
print d
print hamming_distance("attgctg", "atgcctg")
```

Alignment

Levenshtein Distance:
Also called **edit distance**

Defined between two strings of not necessarily equal length, is the minimal number of "edit operations" required to change one string into the other. An edition operation can be *deletion, insertion, or alteration* of a single character. A given sequence of edit operations induces a unique alignment, but not vice versa.

Example:
agtcc and **cgctca**

```
agtcc
cgtcc      alteration
cgctcc     insertion
cgctca     alteration

ag-tcc
cgctca
Levenshtein distance = 3.
```

Alignment

Scoring Schemes:
For applications to molecular biology, recognize that certain changes are more likely to occur naturally than others.

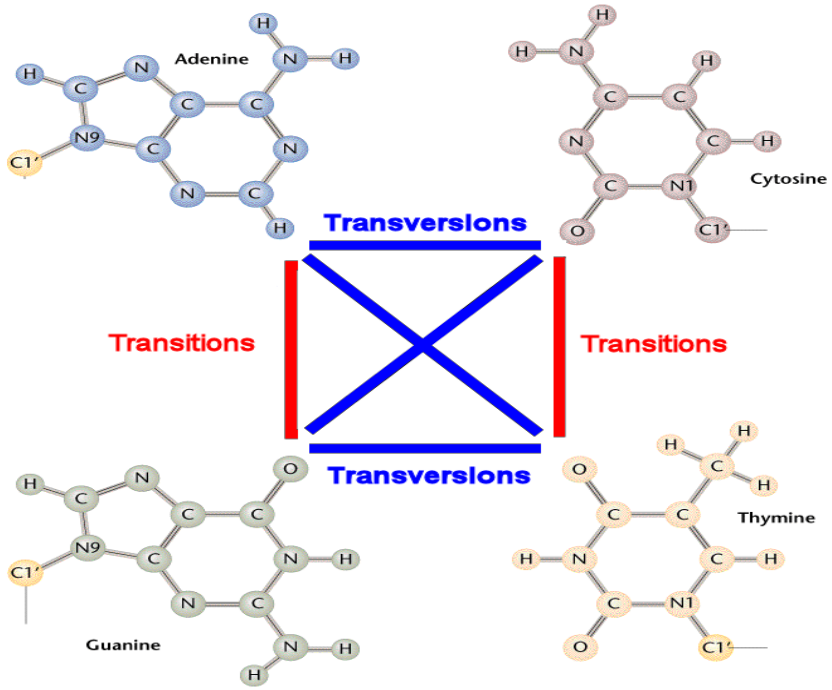
For example, amino acid substitutions tend to be conservative: the replacement of one amino acid by another with similar size or physicochemical properties is more likely to have occurred than its replacement by another amino acid with greater difference in their properties. Or, the deletion of a succession of contiguous bases or amino acids is a more probable event than the independent deletion of the same number of bases or amino acids at non-contiguous positions in the sequences.

We may wish to assign variable weights to different edit operations.

Alignment

Example:
Transition mutations (a↔g and t↔c) are more common than transversions ((a, g)↔(t, c)). Suggest a substitution matrix that reflects this.

	a	g	t	c
a	20	10	5	5
g	10	20	5	5
t	5	5	20	10
c	5	5	10	20



Alignment

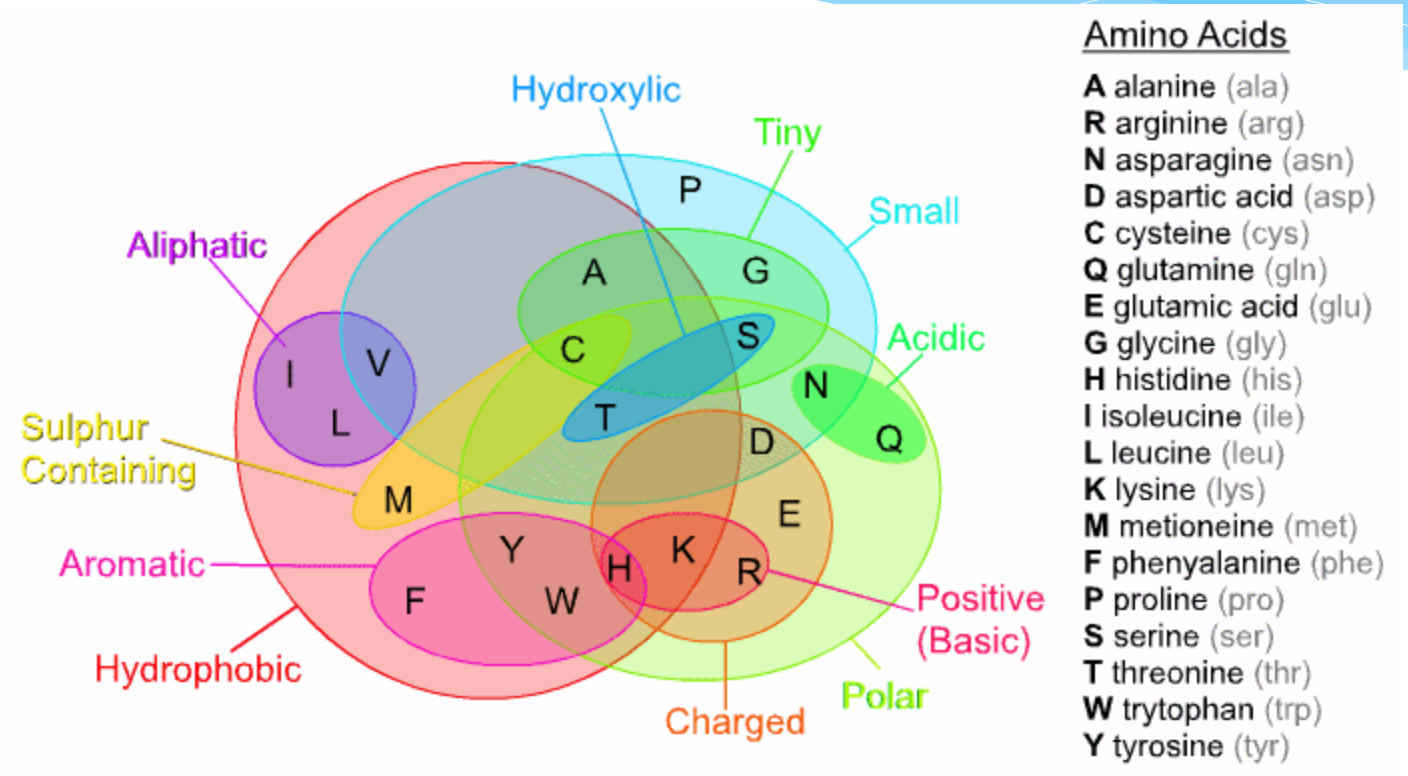
Calculate scores of an alignment:

```
def score(s1, s2):
    sc = 0
    for i in range(len(s1)):
        if s1[i]==s2[i]:
                sc+= 20
        elif (s1[i]=='A' and s2[i]=='G') or \
        (s1[i]=='G' and s2[i]=='A'):
                sc+=10
        #.....
    return sc

strand1 = "ACTCCG"
strand2 = "CGACGC"
print score(strand1, strand2)
```

Alignment

For proteins, the definition of similarity would be much more complicate. Amino acids were not born equally:



Alignment

Substitution Matrix

- Scoring matrix S
  - 20x20 for protein alignment (Amino-acid)

- $S_{i,j}$ represents the gain/penalty due to substituting AA_j by AA_i (**i – line** , **j – colomn**)
  - Based on likelihood this substitution is found in nature
  - Computed differently in PAM and BLOSUM

Alignment

Substitution Matrix

PAM - Point Accepted Mutations

Based on closely related proteins (X% divergence)

Matrices for comparison of divergent proteins computed

BLOSUM - Blocks Substitution Matrix

Based on conserved blocks bounded in similarity (at least X% identical)

Matrices for divergent proteins are derived using appropriate X%

Alignment

PAM (Point Accepted Mutation) Substitution Matrix
Measurement of the relative probability of any particular substitution.

To measure the relative probability of any particular substitution, for instance Serine→Threonine, we can count the number of Serine→ Threonine changes in pairs of aligned homologous sequences. We could use the relative frequencies of such changes to form a scoring matrix for substitutions. A likely change should score higher than a rare one.

Alignment

1 PAM = 1 Percent Accepted Mutation.
Thus, two sequences 1 PAM apart have 99% identical residues. For pairs of sequences within the 1 PAM level of divergence, it is likely that there has been no more than one change at 100 positions. Collecting statistics from pairs of sequences as closely related as this, and correcting for different amino acid abundances, produces the 1 PAM substitution matrix.

To produce a matrix appropriate for more widely divergent sequences, we can take powers of this matrix.

$$PAM_n = (PAM_1)^n$$

The PAM250 level, corresponding to ~20% overall sequence identity, is the lowest sequence similarity for which we can hope to produce a correct alignment by sequence analysis alone. It is therefore the appropriate level to choose for practical work

Alignment

PAM-1 Matrix (probability*100)

PAM1 Mutation Matrix

1 PAM evolutionary distance

	Ala A	Arg R	Asn N	Asp D	Cys C	Gln Q	Glu E	Gly G	His H	Ile I	Leu L	Lys K	Met M	Phe F	Pro P	Ser S	Thr T	Trp W	Tyr Y	Val V
Ala A	9867	2	9	10	3	8	17	21	2	6	4	2	6	2	22	35	32	0	2	18
Arg R	1	9913	1	0	1	10	0	0	10	3	1	19	4	1	4	6	1	8	0	1
Asn N	4	1	9822	36	0	4	6	6	21	3	1	13	0	1	2	20	9	1	4	1
Asp D	6	0	42	9859	0	6	53	6	4	1	0	3	0	0	1	5	3	0	0	1
Cys C	1	1	0	0	9973	0	0	0	1	1	0	0	0	0	1	5	1	0	3	2
Gln Q	3	9	4	5	0	9876	27	1	23	1	3	6	4	0	6	2	2	0	0	1
Glu E	10	0	7	56	0	35	9865	4	2	3	1	4	1	0	3	4	2	0	1	2
Gly G	21	1	12	11	1	3	7	9935	1	0	1	2	1	1	3	21	3	0	0	5
His H	1	8	18	3	1	20	1	0	9912	0	1	1	0	2	3	1	1	1	4	1
Ile I	2	2	3	1	2	1	2	0	0	9872	9	2	12	7	0	1	7	0	1	33
Leu L	3	1	3	0	0	6	1	1	4	22	9947	2	45	13	3	1	3	4	2	15
Lys K	2	37	25	6	0	12	7	2	2	4	1	9926	20	0	3	8	11	0	1	1
Met M	1	1	0	0	0	2	0	0	0	5	8	4	9874	1	0	1	2	0	0	4
Phe F	1	1	1	0	0	0	0	1	2	8	6	0	4	9946	0	2	1	3	28	0
Pro P	13	5	2	1	1	8	3	2	5	1	2	2	1	1	9926	12	4	0	0	2
Ser S	28	11	34	7	11	4	6	16	2	2	1	7	4	3	17	9840	38	5	2	2
Thr T	22	2	13	4	1	3	2	2	1	11	2	8	6	1	5	32	9871	0	2	9
Trp W	0	2	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	9976	1	0
Tyr Y	1	0	3	0	3	0	1	0	4	1	1	0	0	21	0	1	1	2	9945	1
Val V	13	2	1	1	3	2	2	3	3	57	11	1	17	1	3	2	10	0	2	9901

[top row shows original amino acid; left column shows replacement amino acid]

Alignment

PAM-250 Probability Matrix

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	13	6	9	9	5	8	9	12	6	8	6	7	7	4	11	11	11	2	4	9
R	3	17	4	3	2	5	3	2	6	3	2	9	4	1	4	4	3	7	2	2
N	4	4	6	7	2	5	6	4	6	3	2	5	3	2	4	5	4	2	3	3
D	5	4	8	11	1	7	10	5	6	3	2	5	3	1	4	5	5	1	2	3
C	2	1	1	1	52	1	1	2	2	2	1	1	1	1	2	3	2	1	4	2
Q	3	5	5	6	1	10	7	3	7	2	3	5	3	1	4	3	3	1	2	3
E	5	4	7	11	1	9	12	5	6	3	2	5	3	1	4	5	5	1	2	3
G	12	5	10	10	4	7	9	27	5	5	4	6	5	3	8	11	9	2	3	7
H	2	5	5	4	2	7	4	2	15	2	2	3	2	2	3	3	2	2	3	2
I	3	2	2	2	2	2	2	2	2	10	6	2	6	5	2	3	4	1	3	9
L	6	4	4	3	2	6	4	3	5	15	34	4	20	13	5	4	6	6	7	13
K	6	18	10	8	2	10	8	5	8	5	4	24	9	2	6	8	8	4	3	5
M	1	1	1	1	0	1	1	1	1	2	3	2	6	2	1	1	1	1	1	2
F	2	1	2	1	1	1	1	1	3	5	6	1	4	32	1	2	2	4	20	3
P	7	5	5	4	3	5	4	5	5	3	3	4	3	2	20	6	5	1	2	4
S	9	6	8	7	7	6	7	9	6	5	4	7	5	3	9	10	9	4	4	6
T	8	5	6	6	4	5	5	6	4	6	4	6	5	3	6	8	11	2	3	6
W	0	2	0	0	0	0	0	0	1	0	1	0	0	1	0	1	0	55	1	0
Y	1	1	2	1	3	1	1	1	3	2	2	1	2	15	1	2	2	3	31	2
V	7	4	4	4	4	4	4	5	4	15	10	4	10	5	5	5	7	2	4	17

Alignment

PAM-250 log odds scoring Matrix

$S(a,b) = 10 \log_{10} (M_{ab}/P_b)$

PAM 250

	G	A	V	L	I	P	S	T	D	E	N	Q	K	R	H	F	Y	W	M	C	B	Z	X	*	
G	5																								G
A	1	2																							A
V	-1	0	4																						V
L	-4	-2	2	6																					L
I	-3	-1	4	2	5																				I
P	0	1	-1	-3	-2	6																			P
S	1	1	-1	-3	-1	1	2																		S
T	0	1	0	-2	0	0	1	3																	T
D	1	0	-2	-4	-2	-1	0	0	4																D
E	0	0	-2	-3	-2	-1	0	0	3	4															E
N	0	0	-2	-3	-2	0	1	0	2	1	2														N
Q	-1	0	-2	-2	-2	0	-1	-1	2	2	1	4													Q
K	-2	-1	-2	-3	-2	-1	0	0	0	0	1	1	5												K
R	-3	-2	-2	-3	-2	0	0	-1	-1	-1	0	1	3	6											R
H	-2	-1	-2	-2	-2	0	-1	-1	1	1	2	3	0	2	6										H
F	-5	-3	-1	2	1	-5	-3	-3	-6	-5	-3	-5	-5	-4	-2	9									F
Y	-5	-3	-2	-1	-1	-5	-3	-3	-4	-4	-2	-4	-4	-4	0	7	10								Y
W	-7	-6	-6	-2	-5	-6	-2	-5	-7	-7	-4	-5	-3	-2	-3	0	0	17							W
M	-3	-1	2	4	2	-2	-2	-1	-3	-2	-2	-1	0	0	-2	0	-2	-4	6						M
C	-3	-2	-2	-6	-2	-3	0	-2	-5	-5	-4	-5	-5	-4	-3	-4	0	-8	-5	12					C
B	0	0	-2	-3	-2	-1	0	0	3	3	2	1	1	-1	1	-4	-3	-5	-2	-4	3				B
Z	0	0	-2	-3	-2	0	0	-1	3	3	1	3	0	0	2	-5	-4	-6	-2	-5	2	3			Z
X	-1	0	-1	-1	-1	-1	0	0	-1	-1	0	-1	-1	-1	-1	-2	-2	-4	-1	-3	-1	-1	-1		X
*	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	1	*
	G	A	V	L	I	P	S	T	D	E	N	Q	K	R	H	F	Y	W	M	C	B	Z	X	*	

Alignment

In PAM 250 table, the score is calculated by log-odds values. The score of mutation i<-->j is:

$$\log_{10} \frac{Observed\ i \leftrightarrow j\ mutation\ rate}{Expected\ Mutation\ Rate} * 10$$

For example, if the value is 2, then the actual value before scaling (by 10) is 0.2. The value $0.2 = \log 1.6$ $(\text{or } 10^{0.2} = 1.6)$. So the probability of this mutation is about 1.6 times more than random.

Alignment

BLOSUM matrices

Steven Henikoff and Jorja Henikoff developed the family of BLOSUM matrices for scoring substitutions in amino acid sequence comparisons. Their goal was to replace the Dayhoff matrix with one that would perform best in identifying distant relationships, making use of the much larger amount of data that had become available since Dayhoff's work.

The BLOSUM matrices are based on the BLOCKS database of aligned protein sequences; hence the name BLOcks SUbstitution Matrix. From regions of closely-related proteins alignable without gaps, Henikoff calculated the ratio, of the number of observed pairs of amino acids at any position, to the number of pairs expected from the overall amino acid frequencies. As in the Dayhoff matrix, the results are expressed as log-odds.

In order to avoid overweighting closely-related sequences, the Henikoffs replaced groups of proteins that have sequence identities higher than a threshold by either a single representative or a weighted average. The threshold 62% produces the commonly used BLOSUM62 substitution matrix. This is offered by all programs as an option and is the default in most.

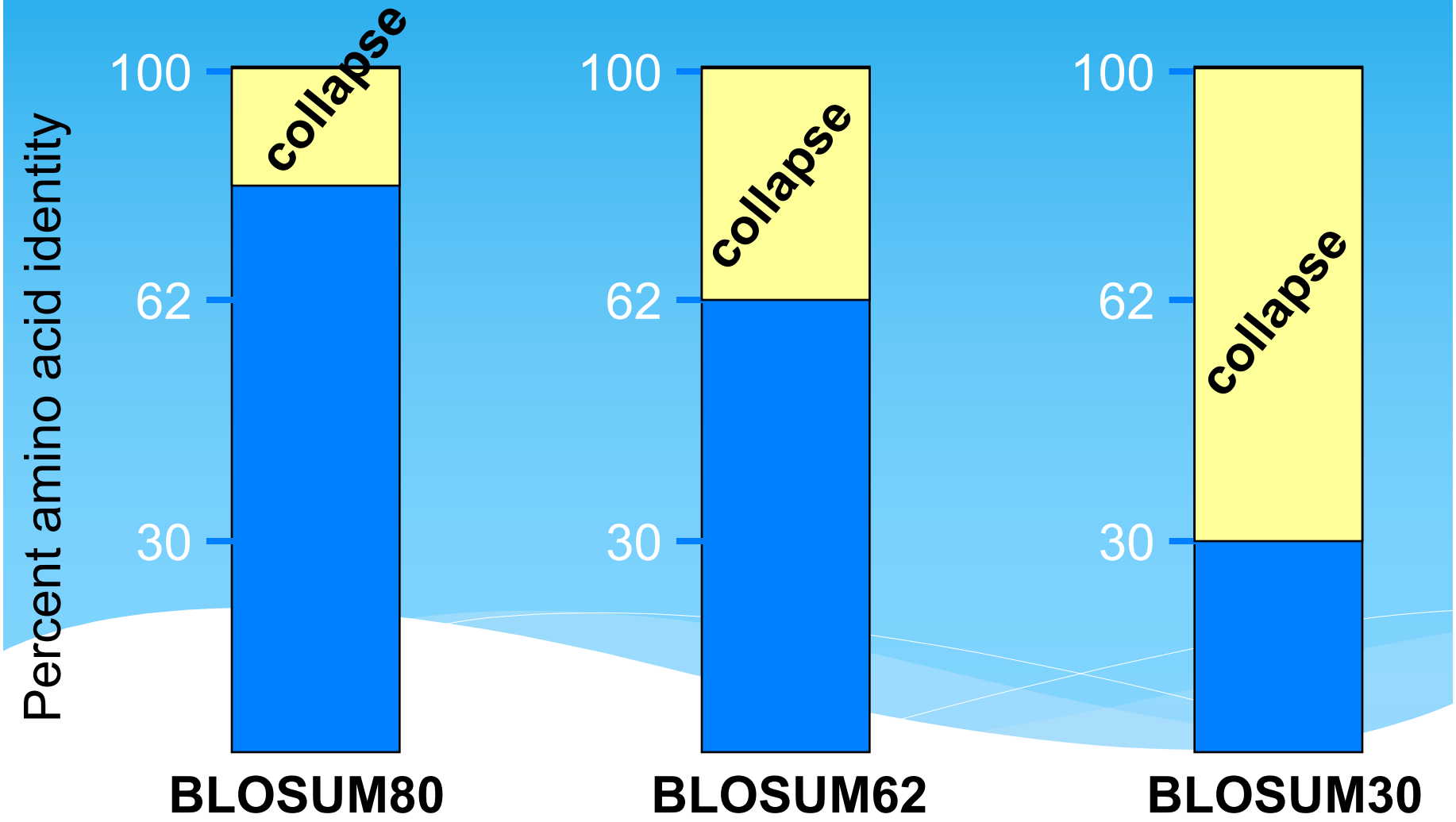
BLOSUM matrices have largely replaced the Dayhoff matrix for most applications.

Alignment

BLOSUM 62 matrices

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	4																			
Arg	-1	5																		
Asn	-2	0	6																	
Asp	-2	-2	1	6																
Cys	0	-3	-3	-3	9															
Gln	-1	1	0	0	-3	5														
Glu	-1	0	0	2	-4	2	5													
Gly	0	-2	0	-1	-3	-2	-2	6												
His	-2	0	1	-1	-3	0	0	-2	8											
Ile	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
Leu	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
Lys	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
Met	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
Phe	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
Pro	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
Ser	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
Thr	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
Trp	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Tyr	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
Val	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4
	Ala	**Arg**	**Asn**	**Asp**	**Cys**	**Gln**	**Glu**	**Gly**	**His**	**Ile**	**Leu**	**Lys**	**Met**	**Phe**	**Pro**	**Ser**	**Thr**	**Trp**	**Tyr**	**Val**

Alignment

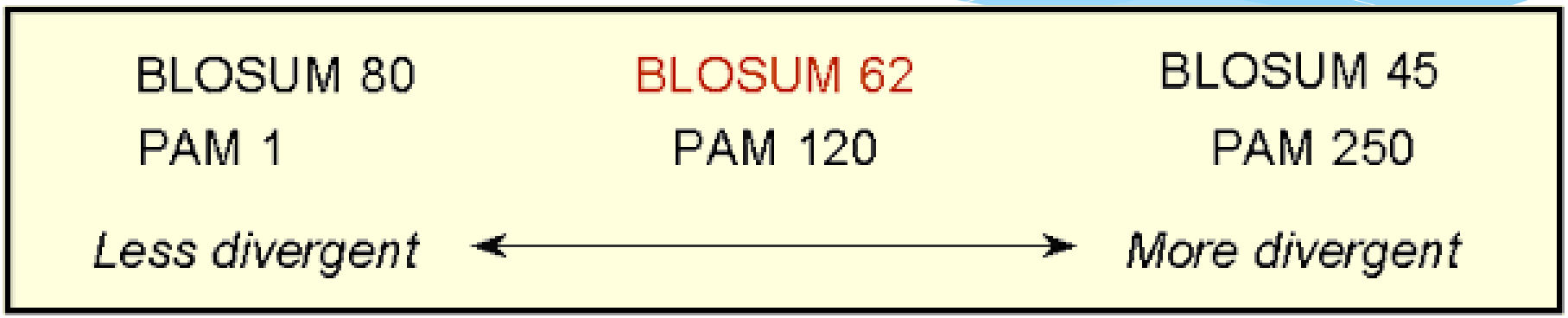

Percent amino acid identity
100
62
30
collapse
BLOSUM80
100
62
30
collapse
BLOSUM62
100
62
30
collapse
BLOSUM30

Alignment

PAM vs BLOSUM Comparison

PAM	BLOSUM
Built from global alignments	Built from local alignments
Built from small amout of Data	Built from vast amout of Data
Counting is based on minimum replacement or maximum parsimony	Counting based on groups of related sequences counted as one
Perform better for finding global alignments and remote homologs	Better for finding local alignments
Higher PAM series means more divergence	Lower BLOSUM series means more divergence

Alignment



Rat versus mouse RBP

Rat versus bacterial lipocalin