Comparing pairs of sequences

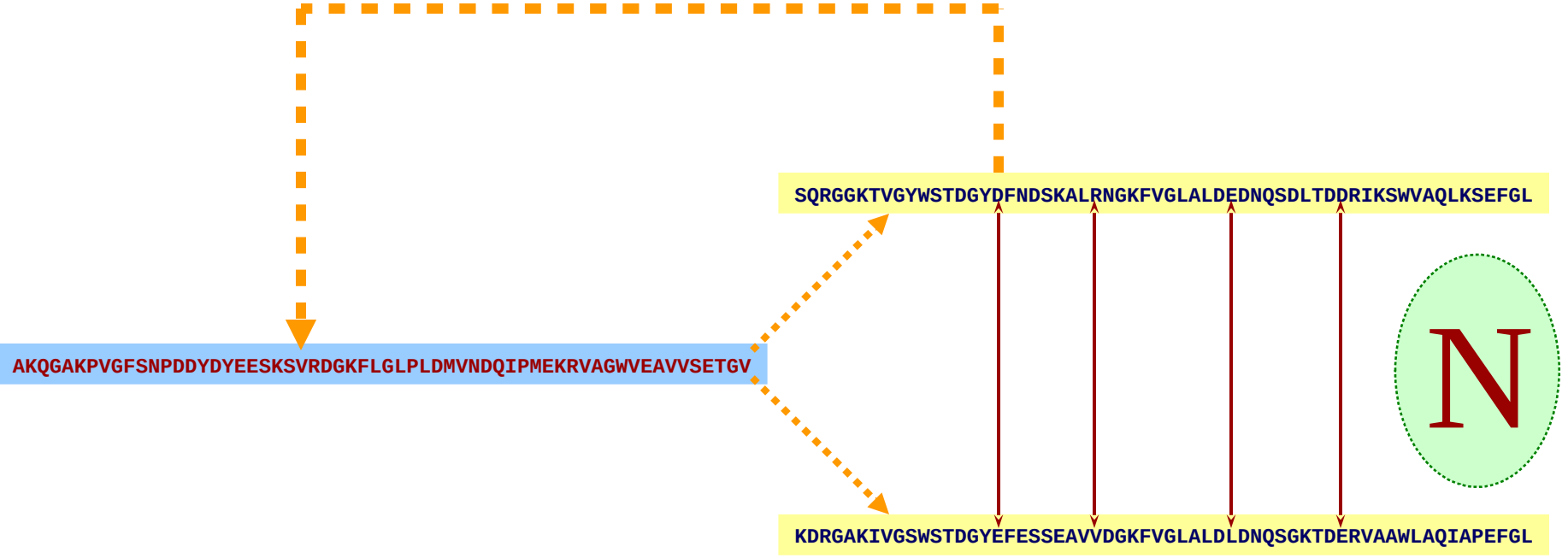
Objectives: To align two sequences in a biologically logical fashion.

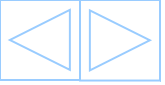
To generate a score representing the "quality" of the alignment.

{Sadly, the computed score has little absolute meaning}

Assumption: The proteins being compared are homologous.

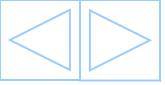
{so differences between the proteins are exclusively due to evolutionary processes}





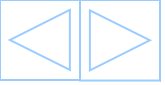
ACBEERGYALEDILAGERAFGSTOUTFAWATERM

ABEERNALEDLAGERDFWGALSTOUTWRARWATERA



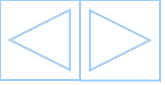
ACBEERGYALEDILAGERAFGSTOUTFAWATERM

ABEERNALEDLAGERDFWGALSTOUTWRARWATERA

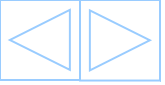


ACBEERGYALEDILAGERAFGSTOUTFAWATERM

ABEERNALEDLAGERDFWGALSTOUTWRARWATERA




ACBEERGYALEDILAGERAFGSTOUTFAWATERM
ABEERNALEDLAGERDFWGALSTOUTWRARWATERA

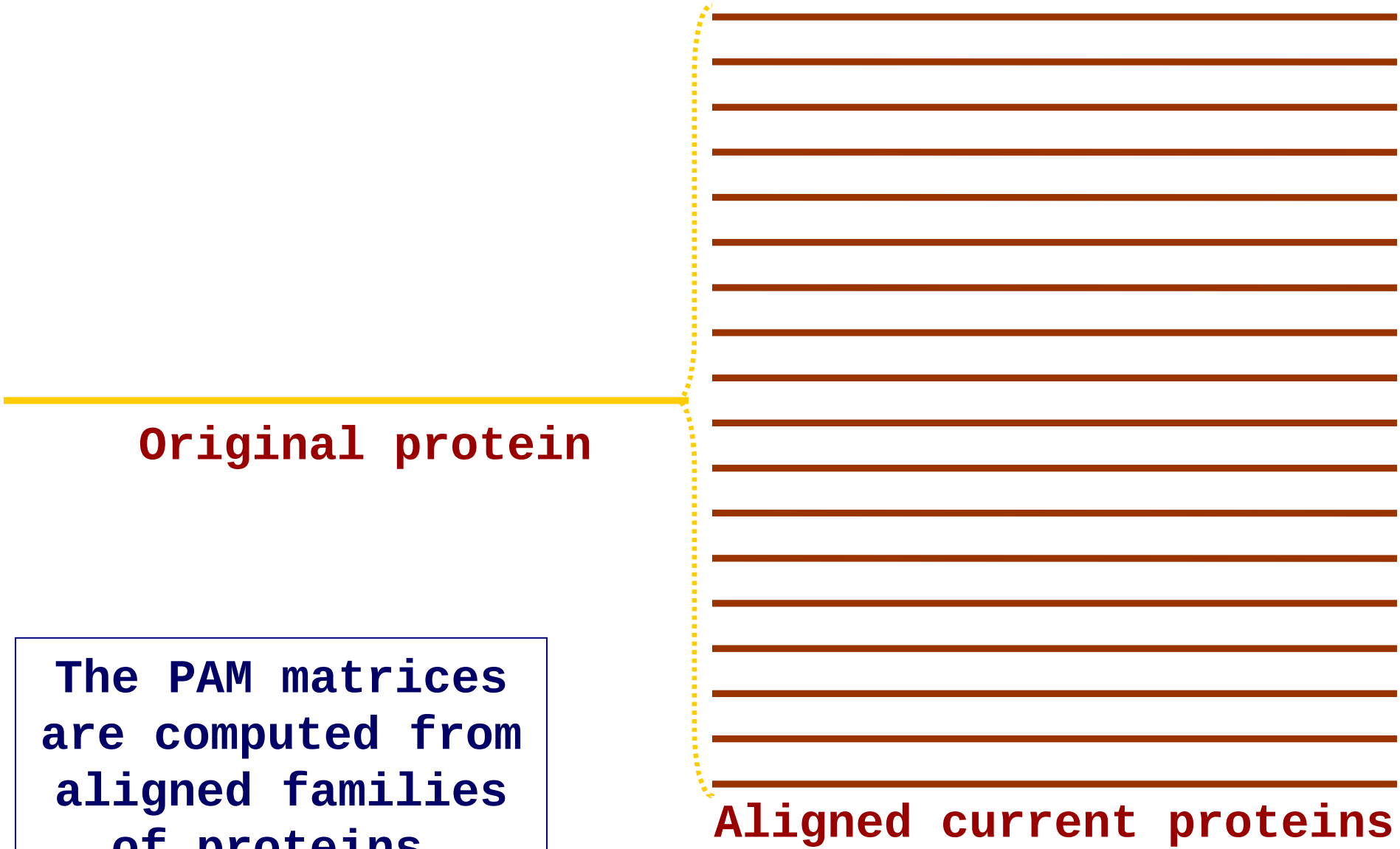


ACBEERGYALEDILAGERAFGSTOUTFAWATERM

ABEERNALEDLAGERDFWGALSTOUTWRARWATERA

Dayhoff PAM 250 Matrix

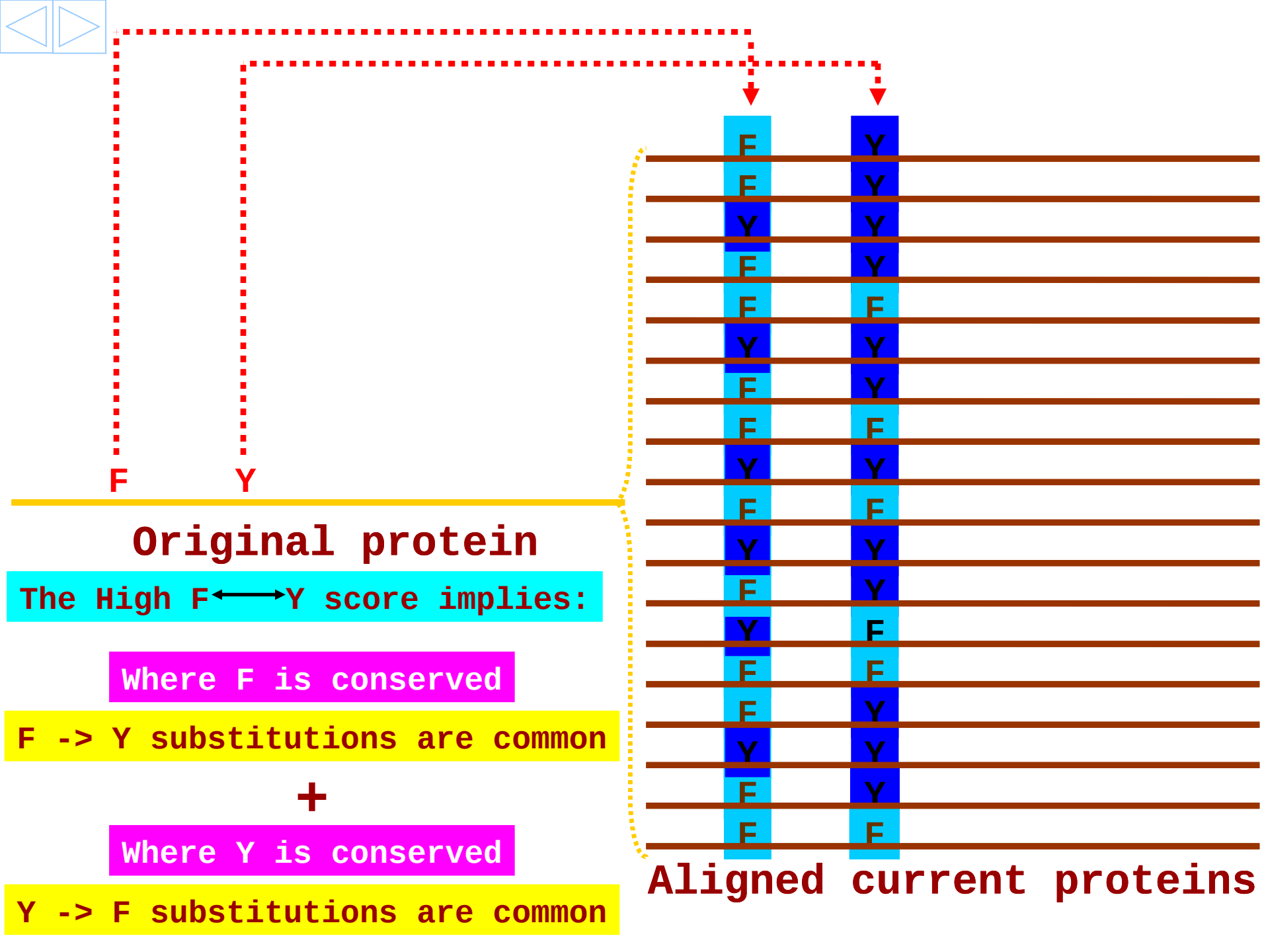
	A	B	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	Z
A	**2**	0	-2	0	0	-4	1	-1	-1	-1	-2	-1	0	1	0	-2	1	1	0	-6	-3	0
B	0	**2**	-4	3	2	-5	0	1	-2	1	-3	-2	2	-1	1	-1	0	0	-2	-5	-3	2
C	-2	-4	**12**	-5	-5	-4	-3	-3	-2	-5	-6	-5	-4	-3	-5	-4	0	-2	-2	-8	0	-5
D	0	3	-5	**4**	3	-6	1	1	-2	0	-4	-3	2	-1	2	-1	0	0	-2	-7	-4	3
E	0	2	-5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



The PAM matrices are computed from aligned families of proteins.

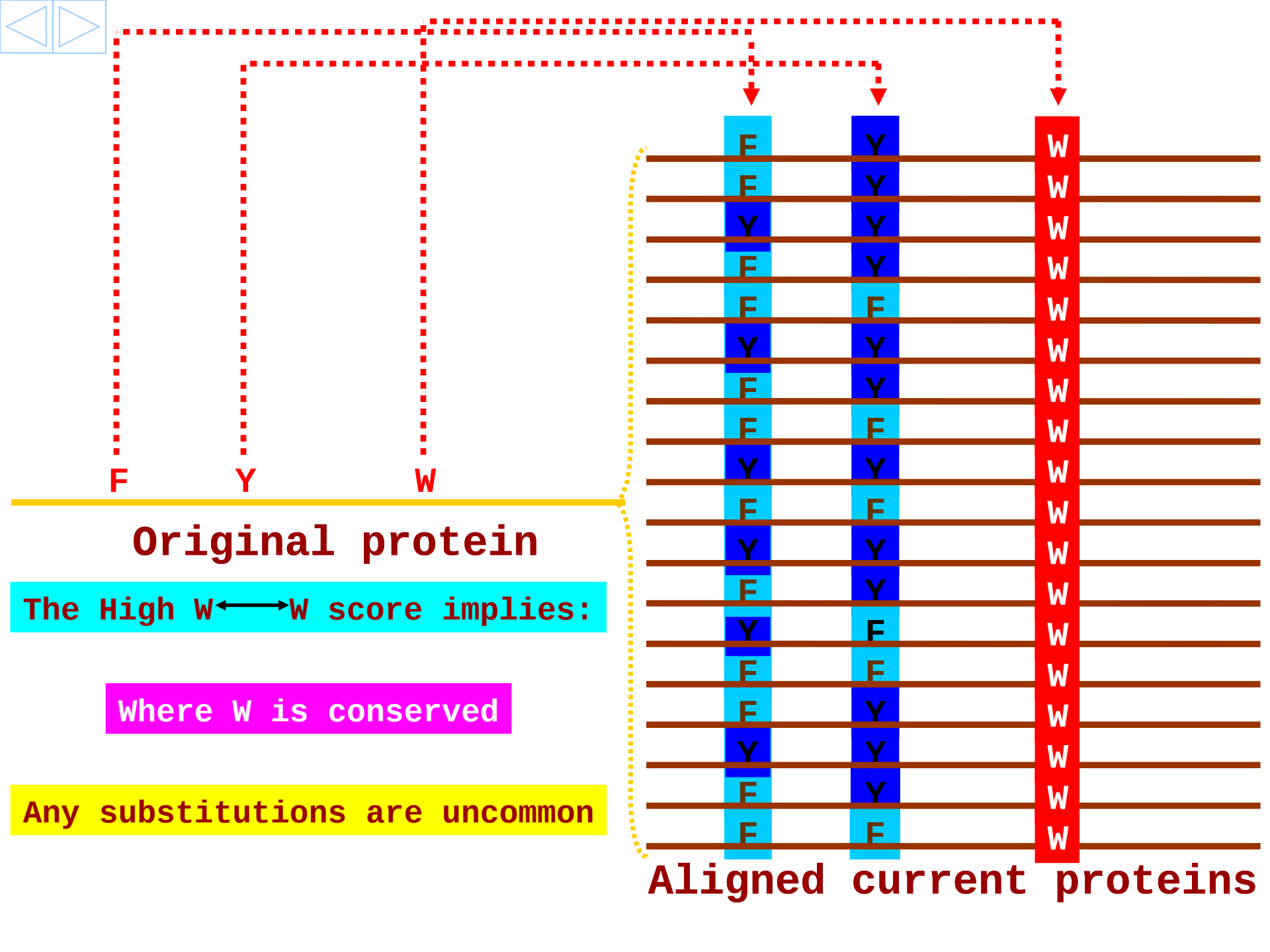
Dayhoff PAM 250 Matrix

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


F
Y
Original protein
The High F⟷Y score implies:
Where F is conserved
F -> Y substitutions are common
+
Where Y is conserved
Y -> F substitutions are common
Aligned current proteins

Dayhoff PAM 250 Matrix

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

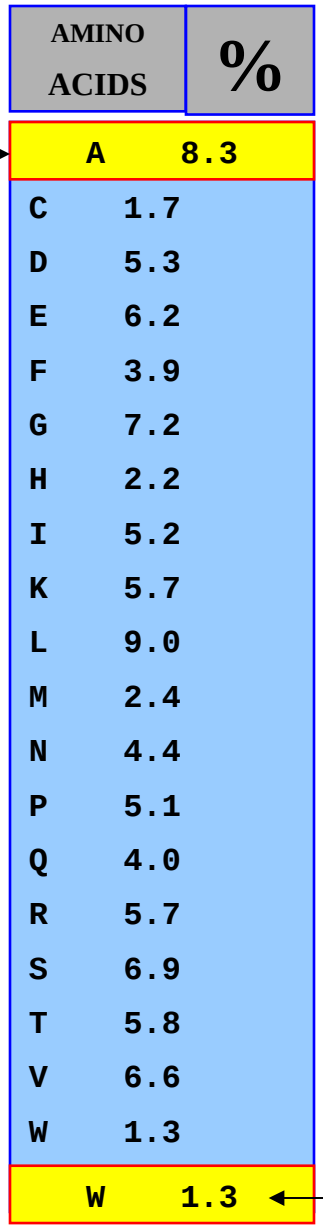

F
Y
W
Original protein
The High W⟷W score implies:
Where W is conserved
Any substitutions are uncommon
Aligned current proteins

Dayhoff PAM 250 Matrix

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

Typical Amino Acid composition

{according to Argos and McCaldon}

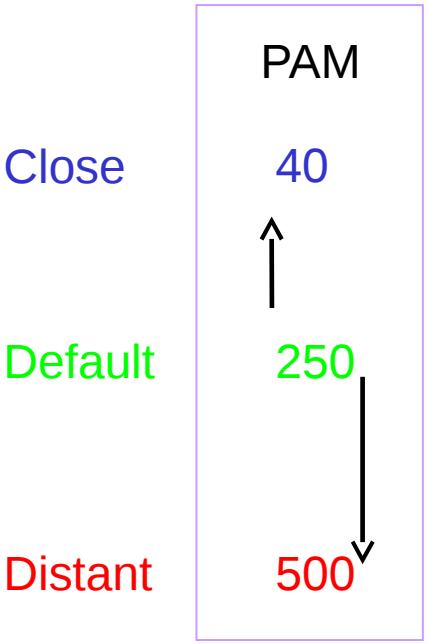


AMINO ACIDS	%
A	8.3
C	1.7
D	5.3
E	6.2
F	3.9
G	7.2
H	2.2
I	5.2
K	5.7
L	9.0
M	2.4
N	4.4
P	5.1
Q	4.0
R	5.7
S	6.9
T	5.8
V	6.6
W	1.3
W	1.3

PAM – Point Accepted Mutation

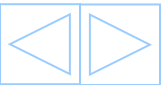
PAM – is a measure of evolutionary distance

An evolutionary distance of 1 PAM indicates the probability of a residue mutating during a distance in which 1 point mutation was accepted per 100 residues.

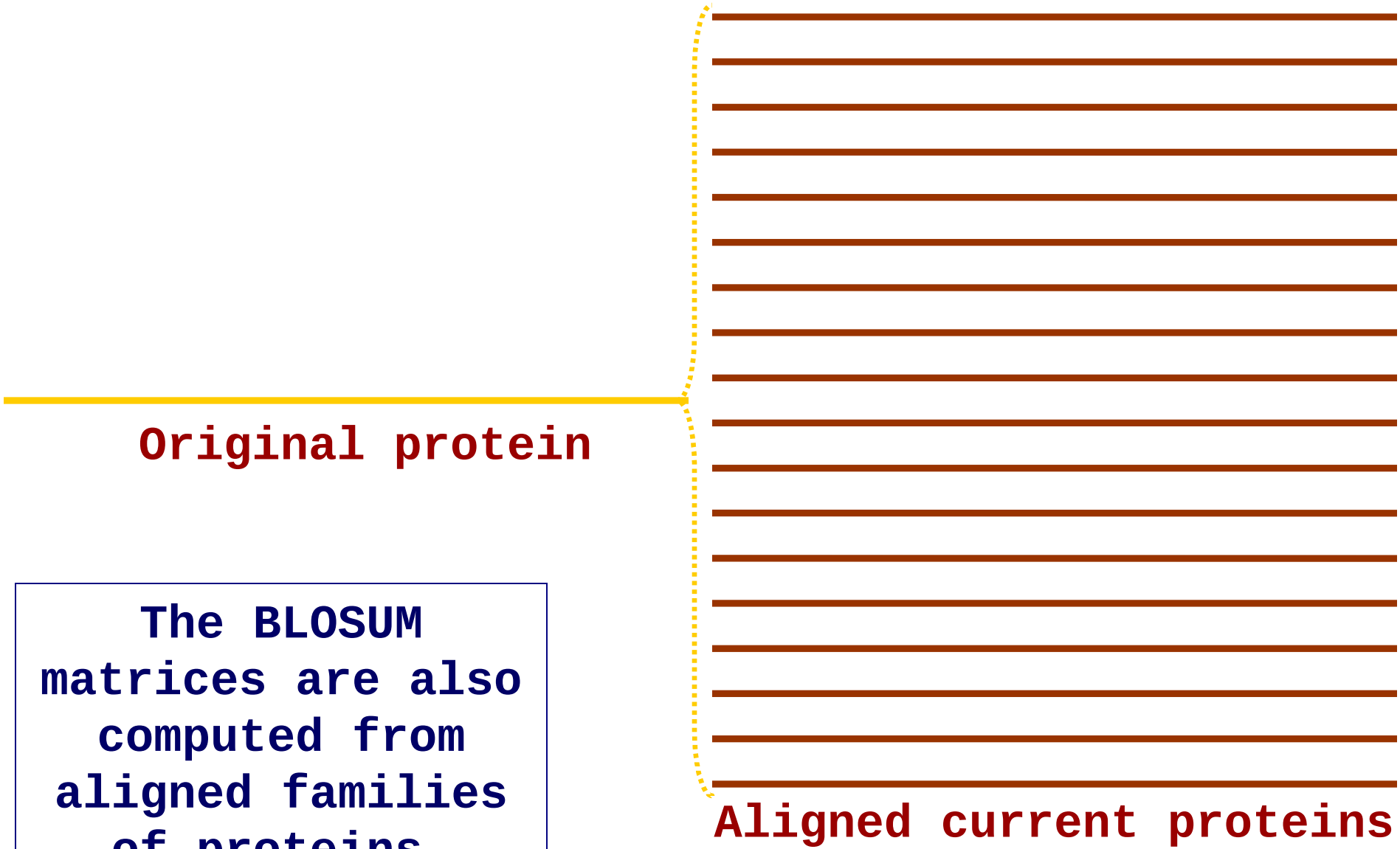

PAM
Close
40
Default
250
Distant
500

Relationship between Observed Identity and PAM value.

Observed % Identity	Evolutionary Distance (PAMs)
99	1
90	11
80	23
70	38
60	56
50	80
40	112
30	159
20	246

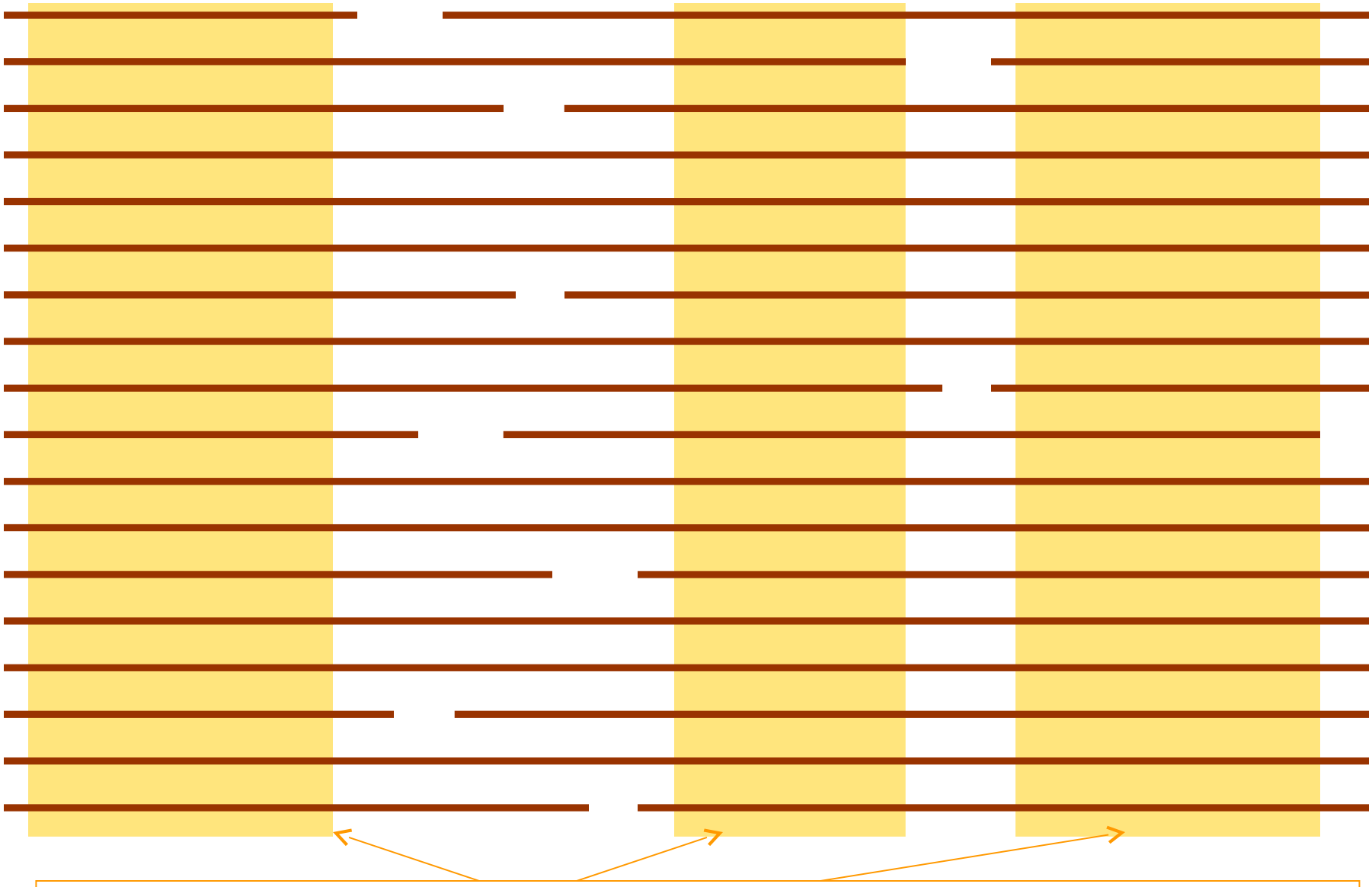


The BLOSUM scoring matrices

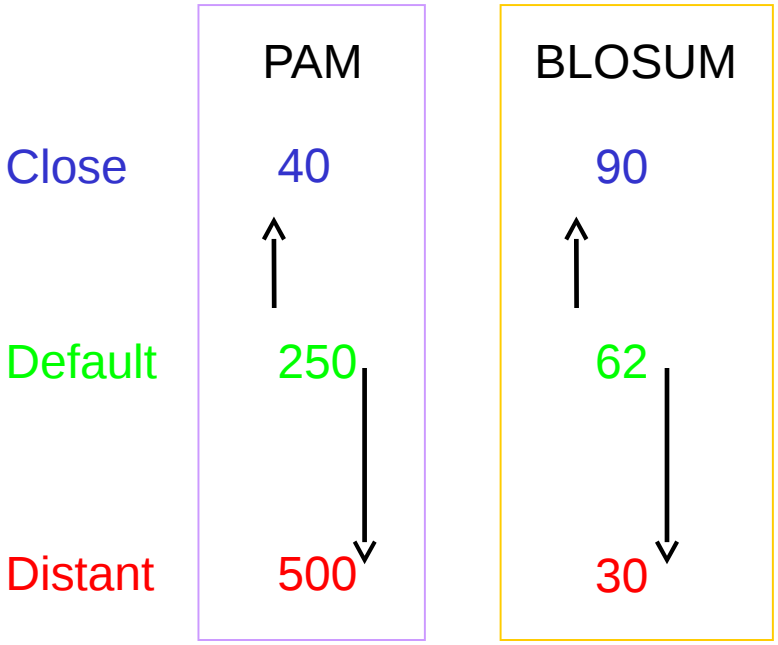


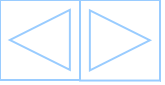
The BLOSUM matrices are also computed from aligned families of proteins.

The BLOSUM scoring matrices



Regions of high conservation, stored in the BLOCKS database


PAM
BLOSUM
Close
40
90
Default
250
62
Distant
500
30



Insertions/Deletions => GAPS

ACBEERGYALEDILAGERAFGSTOUTFAWATERM

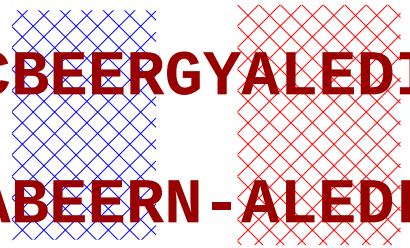
ABEERNALEDLAGERDFWGALSTOUTWRARWATERA

Insertions/Deletions => GAPS

ACBEERGYALEDILAGERAFGSTOUTFAWATERM

ABEERNALEDLAGERDFWGALSTOUTWRARWATERA

Insertions/Deletions => GAPS



ACBEERGYALEDILAGERAFGSTOUTFAWATERM

ABEERN-ALEDLAGERDFWGALSTOUTWRARWATERA

Insertions/Deletions => GAPS

ACBEERGYALEDILAGERAFGSTOUTFAWATERM

ABEERN-ALED-LAGERDFWGALSTOUTWRARWATERA

Insertions/Deletions => GAPS

Insertions/Deletions => GAPS

ACBEERGYALEDILAGERAFG--STOUTFAWATERM

ABEERN-ALED-LAGERDFWGALSTOUTWRARWATERA

Insertions/Deletions => GAPS

Insertions/Deletions => GAPS

Insertions/Deletions => GAPS

Insertions/Deletions => GAPS

?????? – but adds 12 to the score

G<->W = -7
G<->G = +5

The need for penalising GAPs.

?????? – but adds 12 to the score

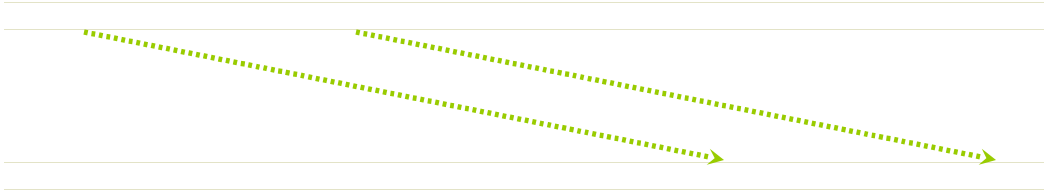
G<->W = -7
G<->G = +5

?????? – but adds 4 to the score

A<->C = -2
A<->A = +2

LOCAL and GLOBAL implementations of pairwise alignment

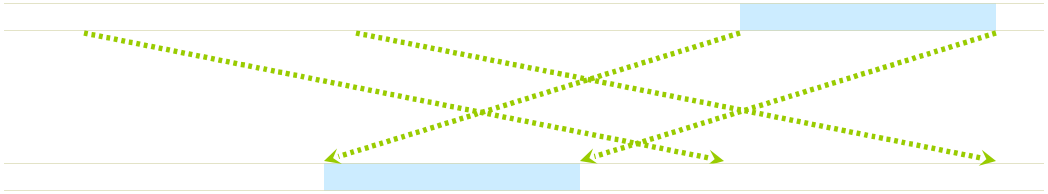
LOCAL Alignment



LOCAL and GLOBAL implementations of pairwise alignment

LOCAL Alignment

GCG : **bestfit**
Staden : **spin**
Emboss : **water/matcher**



GLOBAL Alignment

GCG : **gap**
Staden : **spin**
Emboss : **needle/stretcher**

The End.