Conservation score

In biology, **conserved sequences** are similar or identical sequences that occur within nucleic acid or protein sequences.

The conservation score at a site corresponds to the site's evolutionary rate.

The rate of evolution is not constant among amino acid sites: some positions evolve slowly and are commonly referred to as "conserved", while others evolve rapidly and are referred to as "variable".

The conservation score for all the residues in a protein can be obtained by comparing the sequence of a PDB chain with the proteins deposited in Swiss-Prot/Uniprot and finds the ones that are homologous to the PDB sequence.

Algorithm development

Multiple sequence alignment

```
sp|P69905|     MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG 60
sp|P69907|     MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG 60
sp|P06635|     MVLSPADKTNVKTAWGKVGAHAGDYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKDHG 60
sp|P01958|     MVLSAADKTNVKAAWSKVGGHAGEYGAEALERMFLGFPTTKTYFPHFDLSHGSAQVKAHG 60
sp|P01959|     MVLSAADKTNVKAAWSKVGGNAGEFGAEALERMFLGFPTTKTYFPHFDLSHGSAQVKAHG 60
sp|P01965|     -VLSAADKANVKAAWGKVGGQAGAHGAEALERMFLGFPTTKTYFPHFNLSHGSDQVKAHG 59
sp|P01966|     MVLSAADKGNVKAAWGKVGGHAAEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG 60
sp|P60529|     -VLSPADKTNIKSTWDKIGGHAGDYGGEALDRTFQSFPTTKTYFPHFDLSPGSAQVKAHG 59
sp|P01942|     MVLSGEDKSNIKAAWGKIGGHGAEYGAEALERMFASFPTTKTYFPHFDVSHGSAQVKGHG 60
sp|P01946|     MVLSADDKTNIKNCWGKIGGHGGEYGEEALQRMFAAFPTTKTYFSHIDVSPGSAQVKAHG 60
```

Steps

1. Position specific amino acid frequencies

(a) Unweighted amino acid frequencies

(b) Weighted amino acid frequencies

(c) Estimated independent counts

2. Calculation of conservation scores

(a) Entropy based measure

(b) Variance based measure

(c) Sum of pairs measure

Bioinformatics 17, 700-711 (2001)

Unweighted amino acid frequencies

$f_a^u(i) = n_a(i)/n(i)$

$n_a(i)$:number of sequences in which position i is occupied by amino acid a

n(i): total number of aligned sequences

$= \Sigma\, n_a(i)$, i=1,20

For specific group of sequences

Weighted amino acid frequencies

$f_a^w(i) = \Sigma\ \delta(a,k,i)\ w_k/\Sigma w_k,\ k=1,n(i)$

w_k: weight of a sequence k

$\delta(a,k,i) = 1$, if amino acid a is in sequence k at position i

$\delta(a,k,i) = 0$, otherwise

Correct for unequal distances between different sequence pairs

Estimated independent counts

$f_a^{ic}(i) = n_a^{ic}(i)/n^{ic}(i)$

Probable independent observation with respect to random distribution; E.g. probability of a at position i (F)

20 amino acid residues with random distribution,

$F = 20\,(1-0.95^N)$

$N_{eff} = \ln(1-F/20)/\ln 0.95$,

Any F, amino acid a at position i in a single sequence, $n_a^{ic}(i) = 1$

More sequences: $n_a^{ic}(i) = 1$ (identical)

$$n_a^{ic}(i) = n_a(i)$$

Correlation between aligned sequences

$F = 20\,(1-0.95^N)$

If, N= 1, F = 1; True

Assume N = n,

f(n,i) probability of i different amino acids to occur at the position

$F = \Sigma f(n,i)i = 20(1-0.95^n)$

If N=n+1, number of amino acid is the same with the probability of i/20 (adding additional one with existing one) or i+1 with probability (20-i)/20.

$F = \Sigma[i(i/20)+(i+1)(20-i)/20]\, f(n,i)$

$= \Sigma\,(1+0.95i)\, f(n,i)$

$= 1+0.95*20(1-0.95^n)$

$= 20\,(1-0.95^{n+1})$

Conservation index

Entropy based measure

$$C^e(i) = \Sigma\, f_a(i).\ln f_a(i),\ a = 1,20$$

Order of a system can be measured with entropy

Measure for sequence variability

Maximal if all 20 amino acids have equal frequencies

Not biased with amino acid composition or similarities among amino acids

Conservation index

Variance based measure

$$C^v(i) = \{\Sigma\, [f_a(i) - f_a]^2\}^{0.5},\ a = 1,20$$

f_a: overall frequency for amino acid a in the alignment

$= \Sigma\, n_a(i)/\Sigma n(i)$, i=1,l , l: total number of aligned positions (unweighted)

$= \Sigma\, \Sigma\, \delta(a,k,i)\, w_k/\Sigma\Sigma w_k$, i=1,l , k=1,n(i)(weighted)

Advantages: overall amino acid frequencies, which differ for different protein families.

Conservation index

Sum of pairs measure

$$C^p(i) = \{\Sigma\ \Sigma\ [f_a(i)\ f_b(i)\ \ S_{ab}]^2\}^{0.5},\ a = 1,20;\ b=1,20$$

S_{ab}: amino acid scoring matrix

Conservation score will be higher for the positions occupied by simiar amino acids

Depends on amino acid type

Normalization

$$C_n(i) = (C(i) - C')/\sigma_c$$

$$C' = \Sigma C(i)/l,\ i=1,l$$

$$\sigma_C = [\Sigma(C(i)-C')^2/(l-1)]^{0.5}$$

Bioinformatics 17, 700-711 (2001)

Example

Unweighted frequency

Position 3

$f_a(L) = 10/10 = 1$

Position 6

$f_a(A) = 8/10 = 0.8$

$f_a(D) = 1/10 = 0.1$

$f_a(E) = 1/10 = 0.1$

Entropy based measure

Position 3

$c(i) = 1 \ln 1$

$= 0$

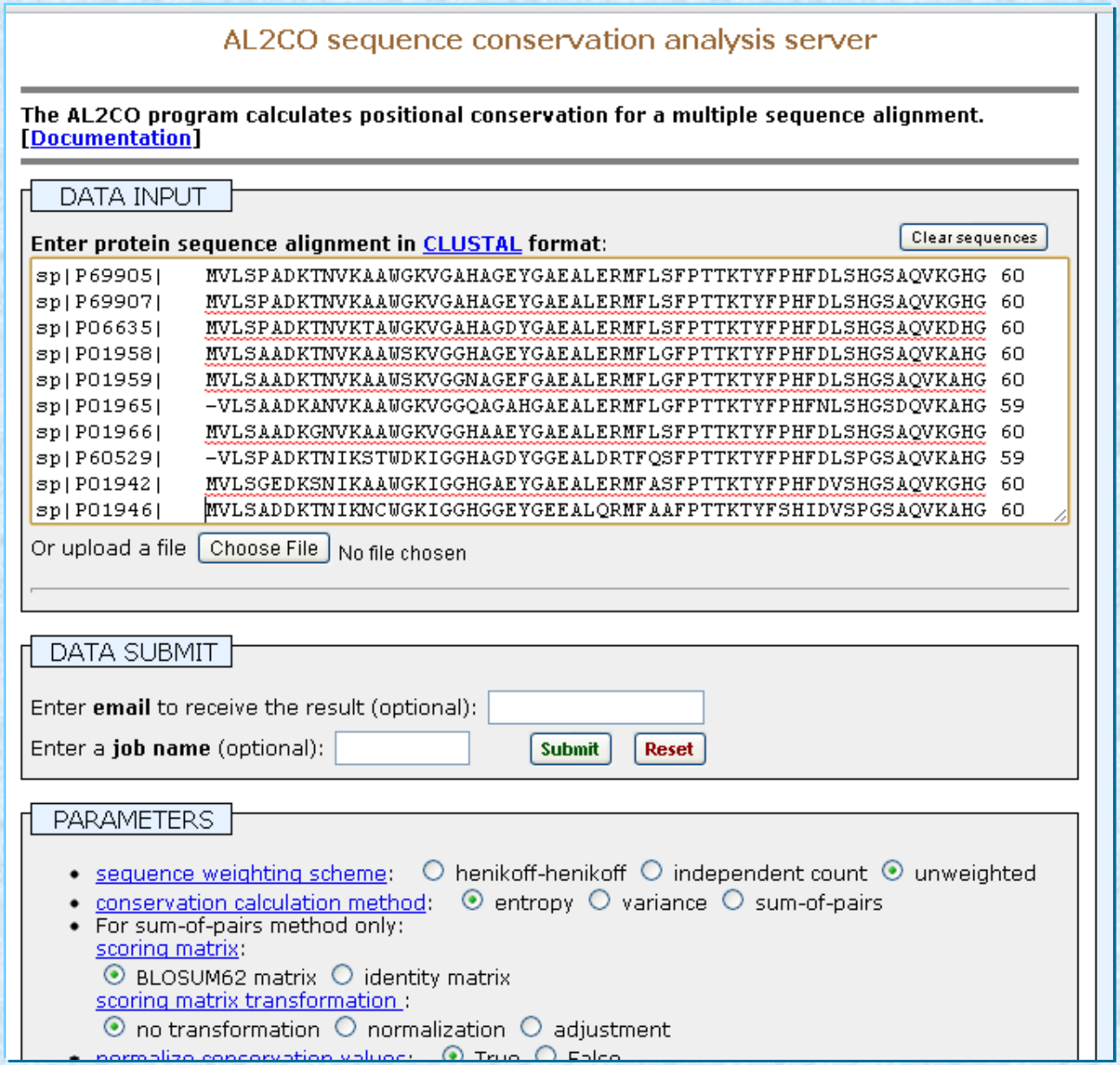
Position 6

$c(i) = 0.8 \ln 0.8 + 0.1 \ln 0.1 + 0.1 \ln 0.1$

$= 0.8 * -0.223 + 0.1 * -2.302 + 0.1 * -2.302$

$= -0.1784 - 0.23 - 0.23$

$= -0.638$



http://prodata.swmed.edu/al2co/al2co.php

AL2CO alignment conservation server

RESULTS:

- The list of positional conservation values is here.
- The alignment with integer conservation indices is here.

INPUTS:

- Input alignment is here.
- Input pdb file: none.

1	M	0.000
2	V	0.000
3	L	0.000
4	S	0.000
5	P	-0.943
6	A	-0.639
7	D	0.000
8	K	0.000
9	T	-0.940
10	N	0.000
11	V	-0.611
12	K	0.000
13	A	-0.940
14	A	-0.639
15	W	0.000
16	G	-0.802
17	K	0.000
18	V	-0.611
19	G	0.000
20	A	-0.611
21	H	-0.639
22	A	-0.500
23	G	-0.500
24	E	-0.802
25	Y	-0.639
26	G	0.000
27	A	-0.639
28	E	0.000
29	A	0.000
30	L	0.000

1	M	0.796
2	V	0.796
3	L	0.796
4	S	0.796
5	P	-1.958
6	A	-1.070
7	D	0.796
8	K	0.796
9	T	-1.950
10	N	0.796
11	V	-0.987
12	K	0.796
13	A	-1.950
14	A	-1.070
15	W	0.796
16	G	-1.545
17	K	0.796
18	V	-0.987
19	G	0.796
20	A	-0.987
21	H	-1.070
22	A	-0.665
23	G	-0.665
24	E	-1.545
25	Y	-1.070
26	G	0.796
27	A	-1.070
28	E	0.796
29	A	0.796
30	L	0.796

Weighted

1	M	0.000
2	V	0.000
3	L	0.000
4	S	0.000
5	P	-0.956
6	A	-0.815
7	D	0.000
8	K	0.000
9	T	-0.966
10	N	0.000
11	V	-0.689
12	K	0.000
13	A	-1.121
14	A	-0.875
15	W	0.000
16	G	-0.815
17	K	0.000
18	V	-0.689
19	G	0.000
20	A	-0.498
21	H	-0.663
22	A	-0.613
23	G	-0.482
24	E	-0.897
25	Y	-0.663
26	G	0.000
27	A	-0.875
28	E	0.000
29	A	0.000
30	L	0.000

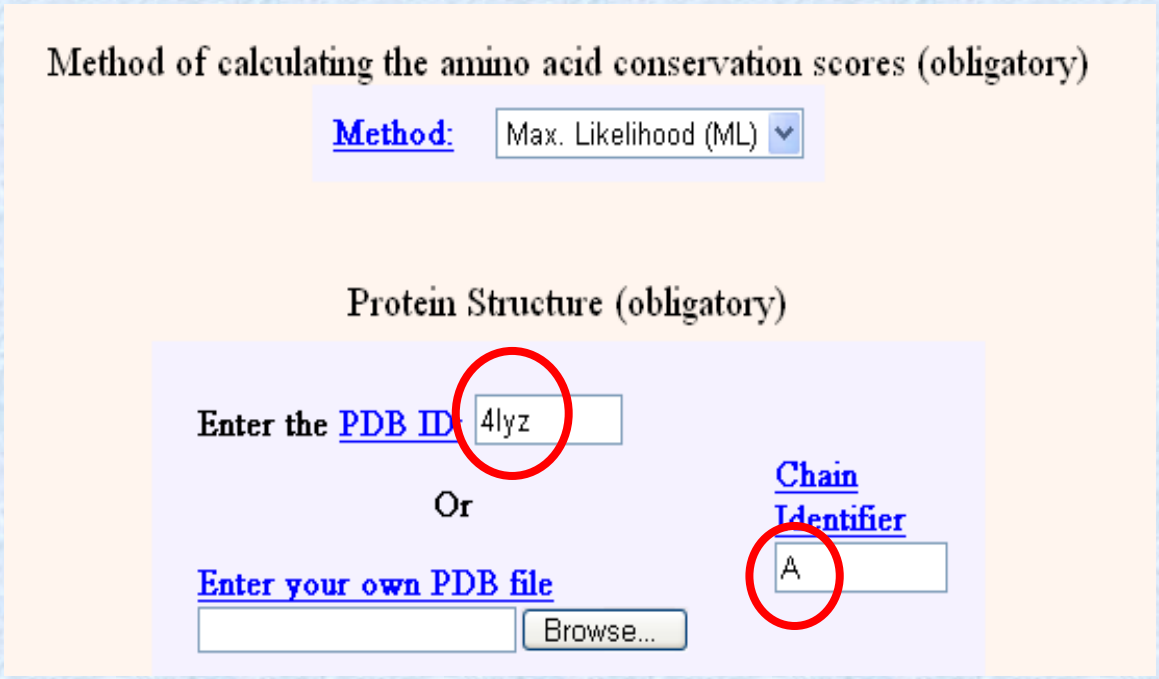
$$(x-X')/\sigma$$

```
Conservation:       999903990939039193933441393999396910999999996966494996999099
sp|P69905|          MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
sp|P69907|          MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
sp|P06635|          MVLSPADKTNVKTAWGKVGAHAGDYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKDHG
sp|P01958|          MVLSAADKTNVKAAWSKVGGHAGEYGAEALERMFLGFPTTKTYFPHFDLSHGSAQVKAHG
sp|P01959|          MVLSAADKTNVKAAWSKVGGNAGEFGAEALERMFLGFPTTKTYFPHFDLSHGSAQVKAHG
sp|P01965|          -VLSAADKANVKAAWGKVGGQAGAHGAEALERMFLGFPTTKTYFPHFNLSHGSDQVKAHG
sp|P01966|          MVLSAADKGNVKAAWGKVGGHAAEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
sp|P60529|          -VLSPADKTNIKSTWDKIGGHAGDYGGEALDRTFQSFPTTKTYFPHFDLSPGSAQVKAHG
sp|P01942|          MVLSGEDKSNIKAAWGKIGGHGAEYGAEALERMFASFPTTKTYFPHFDVSHGSAQVKGHG
sp|P01946|          MVLSADDKTNIKNCWGKIGGHGGEYGEEALQRMFAAFPTTKTYFSHIDVSPGSAQVKAHG
```

9: conserved

0: variable

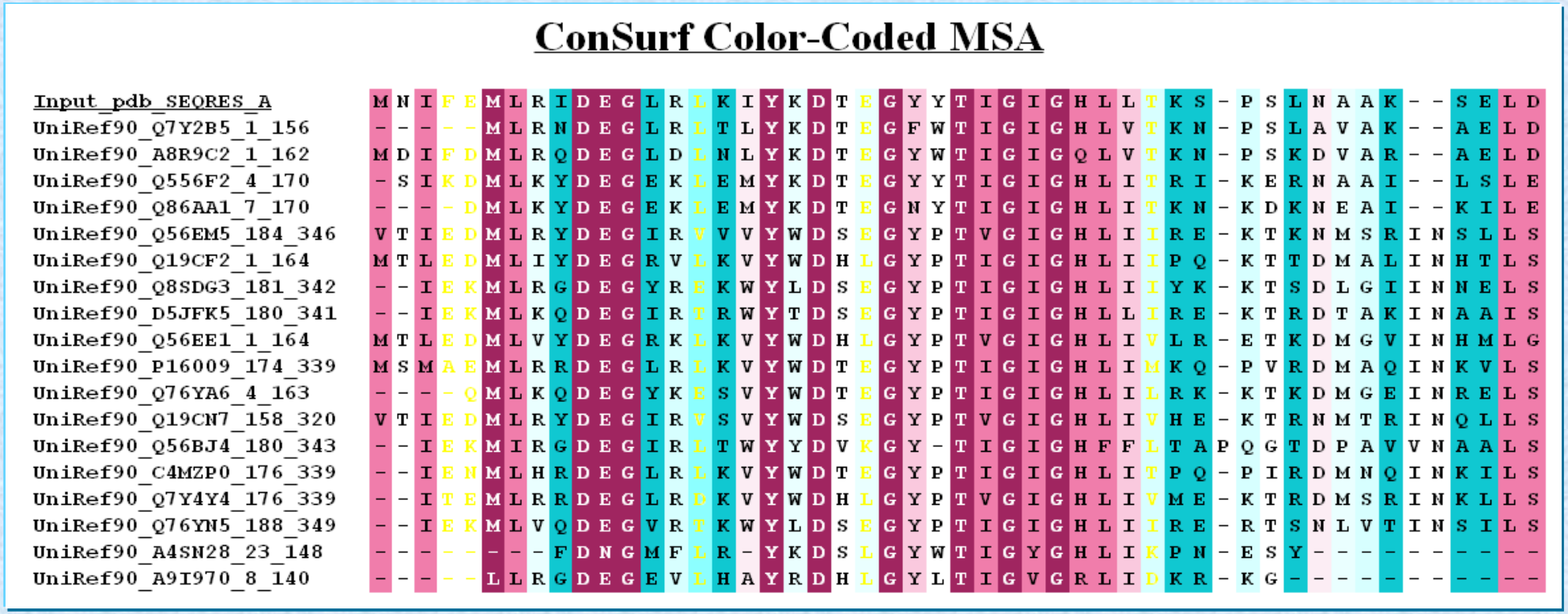
The ConSurf Server



http://consurf.tau.ac.il/.

>2LZM:A|PDBID|CHAIN|SEQUENCE
MNIFEMLRIDEGLRLKIYKDTEGYYTIGIGHLLTKSPSLNAAKSELDKAIGRNCNGVITKDEAEKLFN
QDVDAAVRGILRNAKLKPVYDSLDAVRRCALINMVFQMGETGVAGFTNSLRMLQQKRWDEAAVN
LAKSRWYNQTPNRAKRVITTFRTGTWDA YKNL

ConSurf Color-Coded MSA



POS: The position of the AA in the SEQRES derived sequence.

SEQ: The SEQRES derived sequence in one letter code.

3LATOM: The ATOM derived sequence in three letter code, including the AA's positions as they appear in the PDB file and the chain identifier.

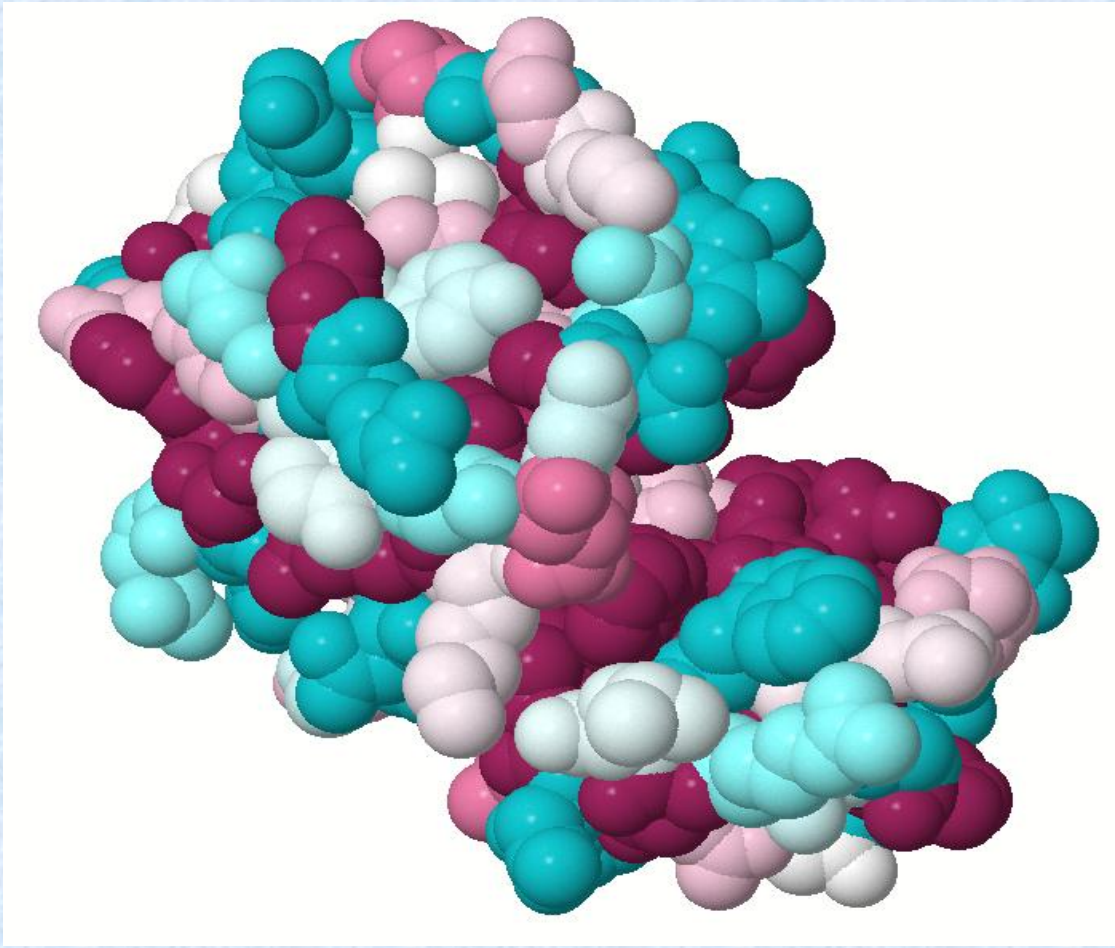
SCORE: The normalized conservation scores.

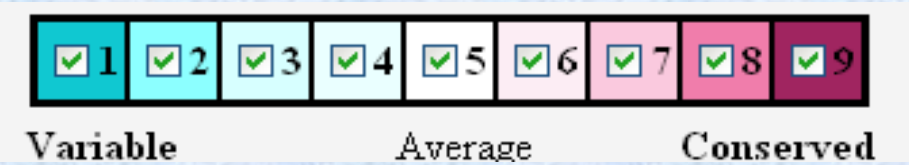
COLOR: The color scale representing the conservation scores (9 - conserved, 1 - variable).

MSA DATA: The number of aligned sequences having an amino acid (non-gapped) from the overall number of sequences at each position. –

RESIDUE VARIETY: The residues variety at each position of the multiple sequence alignment.

POS	SEQ	3LATOM	SCORE (normalized)	COLOR	MSA DATA	RESIDUE VARIETY
1	K	LYS1:A	-0.877	9	49/50	K
2	V	VAL2:A	0.436	3	49/50	I,K,T,V
3	F	PHE3:A	0.763	1	50/50	F,Y
4	G	GLY4:A	1.212	1	50/50	D,E,G,K,Q,S,T
5	R	ARG5:A	-0.673	8	50/50	Q,R
6	C	CYS6:A	-0.877	9	50/50	C
7	E	GLU7:A	-0.877	9	50/50	E
8	L	LEU8:A	0.174	4	50/50	A,F,L,W
9	A	ALA9:A	-0.877	9	50/50	A
10	A	ALA10:A	-0.477	7	50/50	A,K,R




1
2
3
4
5
6
7
8
9
Variable
Average
Conserved