

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	MVLSGEDKSNIKAAWGKIGGHGAEYGAEALERMFAFPTTKTYFPHFDVSHGSAQVKGHG	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

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

For specific group of sequences

Weighted amino acid frequencies

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

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

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

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 $(20 - i)/20$.

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

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

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

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

Conservation index

Entropy based measure

$$C^e(i) = \sum f_a(i) \cdot \ln f_a(i), a = 1, 20$$

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

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

f_a : overall frequency for amino acid **a** in the alignment

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

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

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

Conservation index

Sum of pairs measure

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

S_{ab} : amino acid scoring matrix

Conservation score will be higher for the positions occupied by similar 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

$$\begin{aligned} c(i) &= 1 \ln 1 \\ &= 0 \end{aligned}$$

Position 6

$$\begin{aligned} c(i) &= 0.8 \ln 0.8 + 0.1 \ln 0.1 + 0.1 \ln 0.1 \\ &= 0.8 * -0.223 + 0.1 * -2.302 + 0.1 * -2.302 \\ &= -0.1784 - 0.23 - 0.23 \\ &= -0.638 \end{aligned}$$

AL2CO sequence conservation analysis server

The AL2CO program calculates positional conservation for a multiple sequence alignment.
[\[Documentation\]](#)

DATA INPUT

Enter protein sequence alignment in [CLUSTAL](#) format:

[Clear sequences](#)

```
sp| P69905| MVLSPADKTNVKAAWGKVGAHAGEYGAELERMFLSFPTTKTYFPHFDLSHGSAQVKGHG 60
sp| P69907| MVLSPADKTNVKAAWGKVGAHAGEYGAELERMFLSFPTTKTYFPHFDLSHGSAQVKGHG 60
sp| P06635| MVLSPADKTNVKTAWGKVGAHAGDYGAELERMFLSFPTTKTYFPHFDLSHGSAQVKDHG 60
sp| P01958| MVLSAADKTNVKAAWSKVGGHAGEYGAELERMFLGFPTTKTYFPHFDLSHGSAQVKAHG 60
sp| P01959| MVLSAADKTNVKAAWSKVGGNAGEFGAEALERMFLGFPTTKTYFPHFDLSHGSAQVKAHG 60
sp| P01965| -VLSAADKANVKAAWGKVGGQAGAHGAEALERMFLGFPTTKTYFPHFNLSHGSDQVKAHG 59
sp| P01966| MVLSAADKGNVKAAGKVGGHAAEYGAELERMFLSFPTTKTYFPHFDLSHGSAQVKGHG 60
sp| P60529| -VLSPADKTNIKSTWDKIGGHAGDYGGEALDRFTQSFPPTTKTYFPHFDLSPGSAQVKAHG 59
sp| P01942| MVLSGEDKSNIAAWGKIGGHGA EYGAELERMFASFPTTKTYFPHFDVSHGSAQVKGHG 60
sp| P01946| MVLSADDKTNIKNCWGKIGGHGGEYGEELQRMFAAFPTTKTYFSHIDVSPGSAQVKAHG 60
```

Or upload a file [Choose File](#) No file chosen

DATA SUBMIT

Enter **email** to receive the result (optional):

Enter a **job name** (optional): [Submit](#) [Reset](#)

PARAMETERS

- [sequence weighting scheme](#): ☐ henikoff-henikoff ☐ independent count ☒ unweighted
- [conservation calculation method](#): ☒ entropy ☐ variance ☐ sum-of-pairs
- For sum-of-pairs method only:
[scoring matrix](#):
 - ☒ BLOSUM62 matrix ☐ identity matrix
- [scoring matrix transformation](#):
 - ☒ no transformation ☐ normalization ☐ adjustment
- [normalize conservation values](#): ☒ True ☐ False

<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	MVLSPADKTNVKAANGKVGAGHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
sp P69907	MVLSPADKTNVKAANGKVGAGHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
sp P06635	MVLSPADKTNVKTAWGKVGAGHAGDYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKDHG
sp P01958	MVLSAADKTNVKAANGSKVGGHAGEYGAEALERMFLGFPTTKTYFPHFDLSHGSAQVKAHG
sp P01959	MVLSAADKTNVKAANGSKVGGNAGEFGAEALERMFLGFPTTKTYFPHFDLSHGSAQVKAHG
sp P01965	-VLSAADKANVKAANGKVGQGAGAHGAEALERMFLGFPTTKTYFPHFNLSHGSDQVKAHG
sp P01966	MVLSAADKGNVKAANGKVGGHAAEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
sp P60529	-VLSPADKTNIKSTWDKIGGHAGDYGGEALDRTFQSFPPTTKTYFPHFDLSPGSAQVKAHG
sp P01942	MVLSGEDKSNIAANGKIGGHGAIEYGAEALERMFASFPTTKTYFPHFDVSHGSAQVKGHG
sp P01946	MVLSADDKTNIKNCWGKIGGHGGEYGEEALQRMFAAFPTTKTYFSHIDVSPGSAQVKAHG

9: conserved
0: variable

The ConSurf Server

Method of calculating the amino acid conservation scores (obligatory)

Method: Max. Likelihood (ML) ▼

Protein Structure (obligatory)

Enter the PDB ID 4lyz

Or

Chain Identifier

A

Enter your own PDB file

Browse...

Please enter your e-mail address: michael-gromiha@aist.go.jp

We will use this address in case your job has exceeded its [maximal running time](#).

☒ Send a link to the results by e-mail

Providing your e-mail address ensures that you don't lose the URL to your results.

Submit

Clear


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

>2LZM:A|PDBID|CHAIN|SEQUENCE

MNIFEMLRIDEGLRLKIYKDTEGYTIGIGHLLTKSPSLNAAKSELDKAIGRNCNGVITKDEAEKLFN
 QDVDAAVRGILRNAKLKPVYDSLDAVRRRCALINMVFMGETGVAGFTNSLRMLQQKRWDEAAVN
 LAKSRWYNQTPNRAKRVITTFRTGTWDA YKNL

ConSurf Color-Coded MSA

Input_pdb_SEQRES_A	M	N	I	F	E	M	L	R	I	D	E	G	L	R	L	K	I	Y	K	D	T	E	G	Y	Y	T	I	G	I	G	H	L	L	T	K	S	-	P	S	L	N	A	A	K	-	-	S	E	L	D
UniRef90_Q7Y2B5_1_156	-	-	-	-	-	M	L	R	N	D	E	G	L	R	L	T	L	Y	K	D	T	E	G	F	W	T	I	G	I	G	H	L	V	T	K	N	-	P	S	L	A	V	A	K	-	-	A	E	L	D
UniRef90_A8R9C2_1_162	M	D	I	F	D	M	L	R	Q	D	E	G	L	D	L	N	L	Y	K	D	T	E	G	Y	W	T	I	G	I	G	Q	L	V	T	K	N	-	P	S	K	D	V	A	R	-	-	A	E	L	D
UniRef90_Q556F2_4_170	-	S	I	K	D	M	L	K	Y	D	E	G	E	K	L	E	M	Y	K	D	T	E	G	Y	Y	T	I	G	I	G	H	L	I	T	R	I	-	K	E	R	N	A	A	I	-	-	L	S	L	E
UniRef90_Q86AA1_7_170	-	-	-	-	D	M	L	K	Y	D	E	G	E	K	L	E	M	Y	K	D	T	E	G	N	Y	T	I	G	I	G	H	L	I	T	K	N	-	K	D	K	N	E	A	I	-	-	K	I	L	E
UniRef90_Q56EM5_184_346	V	T	I	E	D	M	L	R	Y	D	E	G	I	R	V	V	V	Y	W	D	S	E	G	Y	P	T	V	G	I	G	H	L	I	I	R	E	-	K	T	K	N	M	S	R	I	N	S	L	L	S
UniRef90_Q19CF2_1_164	M	T	L	E	D	M	L	I	Y	D	E	G	R	V	L	K	V	Y	W	D	H	L	G	Y	P	T	I	G	I	G	H	L	I	I	P	Q	-	K	T	T	D	M	A	L	I	N	H	T	L	S
UniRef90_Q8SDG3_181_342	-	-	I	E	K	M	L	R	G	D	E	G	Y	R	E	K	W	Y	L	D	S	E	G	Y	P	T	I	G	I	G	H	L	I	I	Y	K	-	K	T	S	D	L	G	I	I	N	N	E	L	S
UniRef90_D5JFK5_180_341	-	-	I	E	K	M	L	K	Q	D	E	G	I	R	T	R	W	Y	T	D	S	E	G	Y	P	T	I	G	I	G	H	L	L	I	R	E	-	K	T	R	D	T	A	K	I	N	A	A	I	S
UniRef90_Q56EE1_1_164	M	T	L	E	D	M	L	V	Y	D	E	G	R	K	L	K	V	Y	W	D	H	L	G	Y	P	T	V	G	I	G	H	L	I	V	L	R	-	E	T	K	D	M	G	V	I	N	H	M	L	G
UniRef90_P16009_174_339	M	S	M	A	E	M	L	R	R	D	E	G	L	R	L	K	V	Y	W	D	T	E	G	Y	P	T	I	G	I	G	H	L	I	M	K	Q	-	P	V	R	D	M	A	Q	I	N	K	V	L	S
UniRef90_Q76YA6_4_163	-	-	-	-	Q	M	L	K	Q	D	E	G	Y	K	E	S	V	Y	W	D	T	E	G	Y	P	T	I	G	I	G	H	L	I	L	R	K	-	K	T	K	D	M	G	E	I	N	R	E	L	S
UniRef90_Q19CN7_158_320	V	T	I	E	D	M	L	R	Y	D	E	G	I	R	V	S	V	Y	W	D	S	E	G	Y	P	T	V	G	I	G	H	L	I	V	H	E	-	K	T	R	N	M	T	R	I	N	Q	L	L	S
UniRef90_Q56BJ4_180_343	-	-	I	E	K	M	I	R	G	D	E	G	I	R	L	T	W	Y	Y	D	V	K	G	Y	-	T	I	G	I	G	H	F	F	L	T	A	P	Q	G	T	D	P	A	V	V	N	A	A	L	S
UniRef90_C4MZP0_176_339	-	-	I	E	N	M	L	H	R	D	E	G	L	R	L	K	V	Y	W	D	T	E	G	Y	P	T	I	G	I	G	H	L	I	T	P	Q	-	P	I	R	D	M	N	Q	I	N	K	I	L	S
UniRef90_Q7Y4Y4_176_339	-	-	I	T	E	M	L	R	R	D	E	G	L	R	D	K	V	Y	W	D	H	L	G	Y	P	T	V	G	I	G	H	L	I	V	M	E	-	K	T	R	D	M	S	R	I	N	K	L	L	S
UniRef90_Q76YN5_188_349	-	-	I	E	K	M	L	V	Q	D	E	G	V	R	T	K	W	Y	L	D	S	E	G	Y	P	T	I	G	I	G	H	L	I	I	R	E	-	R	T	S	N	L	V	T	I	N	S	I	L	S
UniRef90_A4SN28_23_148	-	-	-	-	-	-	-	F	D	N	G	M	F	L	R	-	Y	K	D	S	L	G	Y	W	T	I	G	Y	G	H	L	I	K	P	N	-	E	S	Y	-	-	-	-	-	-	-	-	-		
UniRef90_A9I970_8_140	-	-	-	-	-	L	L	R	G	D	E	G	E	V	L	H	A	Y	R	D	H	L	G	Y	L	T	I	G	V	G	R	L	I	D	K	R	-	K	G	-	-	-	-	-	-	-	-	-	-	

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.

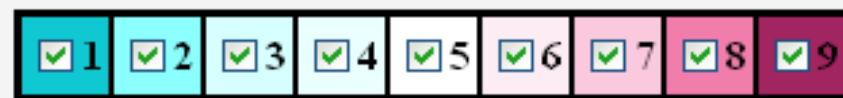
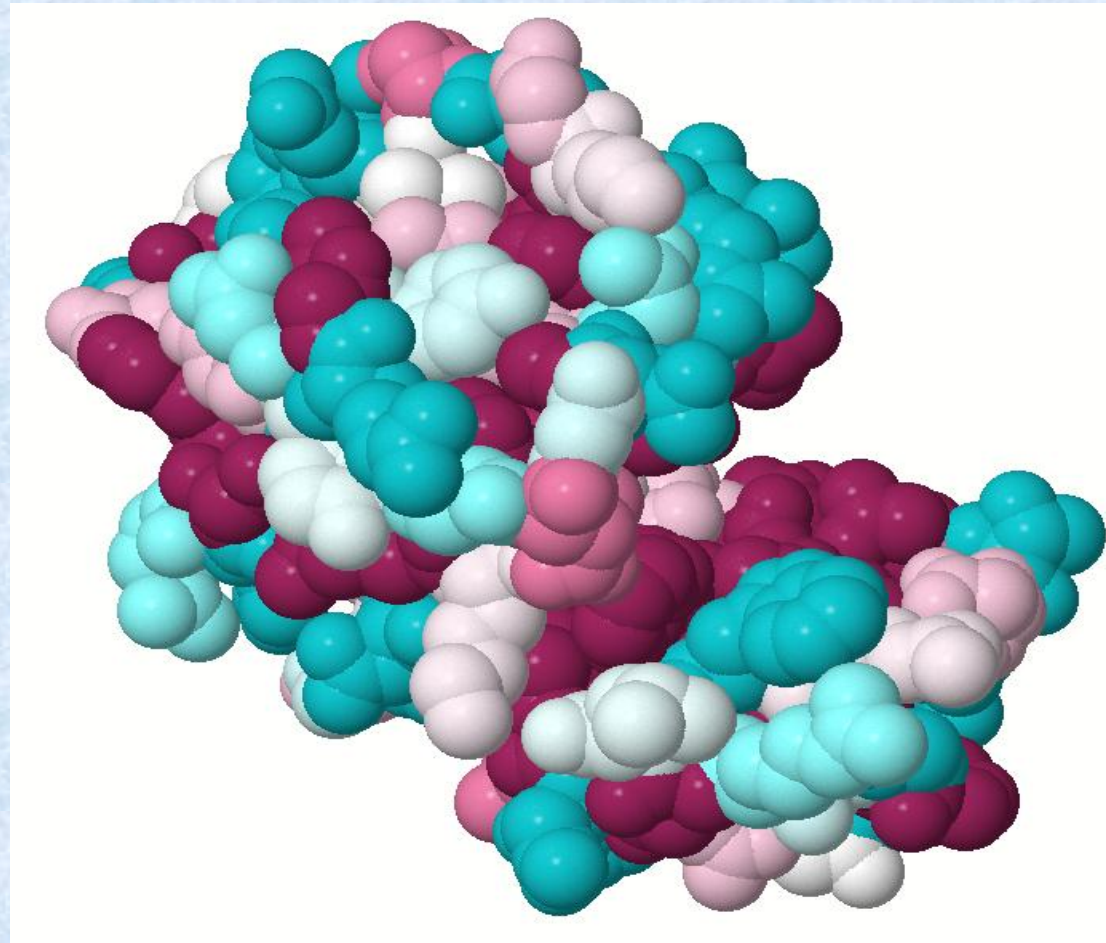
SCORE: The normalized conservation scores.

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



Variable

Average

Conserved