Conservation and allowed residues

Andrew C.R. Martin

21st December 2022

1 Introduction

We have previously looked at the problem of conservation in columns in a multiple sequence alignment (MSA), and generally have ways of scoring conservation including our version of 'scorecons'. We have also looked at weighting the conservation by species-level sequence conservation and have implemented a statistical cutoff for significant conservation in an alignment-dependent manner (ImPACT).

However, this does not address the problem of whether a *specific* mutation is allowed. Suppose we have 100 sequences where 90 of them have a serine at a given position and the remaining 10 have a threonine. Since serine and threonine are very similar amino acids and score highly in things like BLOSUM or PAM scoring matrices, this will be indicated as a very conserved position. Using our 'scorecons' and the default pet91 matrix, this would obtain a conservation score of 0.891. Using the grouped entropy scoring, it obtains a score of 1.0.

Now suppose we have a native sequence having a serine at this position, if a mutation is made to a very different amino acid then this is very likely to be problematic. However, if the mutation is to a threonine, on the basis of conservation, this will still be scored as problematic although we know that this is a very similar amino acid and is allowed in other sequences.

A further problem is the concept of 'compensated pathogenic mutations' (CPDs). Here a mutation known to be pathogenic in (say) humans results in the native residue found in another species (say mice). Clearly there are other accompanying mutations that compensate for this change. Thus the mutation in our species of interest is damaging, yet the resulting mutant residue is seen in an MSA at this position.

2 Pre-requisites

The following pre-requisites and assumptions are made below:

1. Any scoring matrix may be used, but it must be normalized such that a) all values are positive, b) the values on the diagonal are equal and maximal, c) the values on the diagonal are equal to 1. The Valdar scorecons paper explains how to do this.

2. An MSA is available for Proposals 4 and 5.

3. A method is available for scoring sequence identity or sequence similarity (between two pre-aligned sequences) for Proposals 4 and 5. In both cases, the value should be expressed as a number between 0 and 1.

3 Proposal 1 — Conservation with respect to a reference sequence

The first proposal is that, instead of calculating an overall conservation score for a position in an alignment, we should only look at conservation in a pairwise manner from a reference sequence of interest. In 'scorecons' we calculate a score for a particular column as:

$$S_c = \sum_{i,j>i}^{N} s_{i,j}/N_c \tag{1}$$

where S_c is the score for this column, $s_{i,j}$ is a score for residues at positions i and j in the column, maximal and equal to 1), N is the number of sequences in the alignment and N_c is the number of comparisons made ($N_c = N(N-1)/2$).

Instead, we would calculate:

$$S_r = \sum_{j>1}^{N} s_{1,j}/(N-1) \tag{2}$$

i.e. S_r is simply the conservation score *with respect to the reference sequence.*

4 Proposal 2 — minimum conservation

The second proposal is that, instead of calculating the mean conservation, we should calculate the minimum score for a substituted residue. In other words what is the minimum conservation (i.e. maximum diversity) allowed at this position?

$$S_m = \min_{j>1}^{N} s_{1,j} \tag{3}$$

This could then be used as a threshold for suggesting whether a particular mutation is likely to be damaging. i.e. we would calculate the difference between a score for a given mutation from the native sequence and S_m:

$$\Delta_m = s_{1,m} - S_m \tag{4}$$

where $s_{1,m}$ is the score for mutated residue m against the reference sequence residue and S_m is the thresold calculated in Equation 3. A negative value would indicate an unfavourable mutation.

5 Proposal 3 — minimum conservation by S.D.

Proposal 3 refines Proposal 2 by selecting a score that is d standard deviations lower than the mean rather than simply selecting the minimum. This accounts

for the distribution of mutated residues rather than simply taking the most different. The standard deviation is calculated as:

$$\sigma = \sqrt{\frac{\sum_{j>1}^{N}(s_{1,j} - S_r)^2}{N-1}} \tag{5}$$

where S_r is the mean conservation with respect to the reference sequence as calculated in Equation 2. A threshold score would then be calculated as:

$$S_\sigma = S_r - d\sigma \tag{6}$$

where d would have to be determined empirically, but typically could be 2.5 or 3.0.

As before, we could then calculate the difference between a score for a given mutation from the native sequence and S_σ:

$$\Delta_\sigma = s_{1,m} - S_\sigma \tag{7}$$

where $s_{1,m}$ is the score for mutated residue m against the reference sequence residue and S_σ is the thresold calculated in Equation 6. A negative value would indicate an unfavourable mutation.

6 Proposal 4 — weighted minimum conservation

Proposal 4 attempts to allow for CPDs. CPDs become more common as the sequences become more diverse. i.e. it is often small cummulative effects rather than a single compensatory mutation (see our paper). Thus the idea would be to weight allowed diversity higher for sequences that are overall very similar and lower for sequences that are less similar.

For example, suppose the reference sequence has a serine at a particular position and another sequence is identical except for an asparagine at that position. Clearly that asparagine is going to be allowed as a mutation with no effect on the function of the protein (unless the compensatory mutation is in another protein). On the other hand, if two sequences are quite diverse, there may be compensatory effects going on.

In this case, it is simpler to work in terms of diversity rather than conservation of residues. Assuming (as above) that residue similarity from a scoring matrix is scored in the range of $0 \ldots 1$ (where 1 represents an identical residue), we can simply define the diversity (v) as:

$$v_{i,j} = 1 - s_{i,j} \tag{8}$$

We can now revisit Proposal 2, weighting by the sequence identity (or similarity — both expressed as a fractional value, P) between the two sequences:

$$V_{mw} = \max_{j>1}^{N}(v_{1,j} \times P_{1,j}) \tag{9}$$

where $v_{1,j}$ is the diversity between the residue in the reference sequence residue and in sequence j as defined in Equation 8 and $P_{1,j}$ is the overall fractional

sequence identity (or similarity) between sequences the reference sequence and sequence j. As before, this could then be used as a threshold for suggesting whether a particular mutation is likely to be damaging:

$$\Delta_{mw} = V_{mw} - (v_{1,M} \times P_{1,M}) \tag{10}$$

where $v_{1,M}$ is the diversity score for mutated residue M against the reference sequence residue and V_{mw} is the threshold calculated in Equation 9. $P_{1,M}$ is the sequence identity (or similarity) between the reference sequence and the mutated sequence (which clearly will be very high, normally only one mutation being present). A negative value would indicate an unfavourable mutation.

7 Proposal 5 — weighted minimum conservation by SD

Proposal 5 is simply the ideas of Proposal 4 applied to Proposal 3 instead of Proposal 2, i.e. selecting a diversity score that is d standard deviations above the mean rather than simply selecting the maximum.

The mean weighted diversity score is calculated with resepct to the reference sequence (analagous to Equation 2):

$$V_r = \sum_{j>1}^{N} (v_{1,j} \times P_{1,j})/(N-1) \tag{11}$$

The standard deviation is calculated as (analagous to Equation 5):

$$\sigma_w = \sqrt{\frac{\sum_{j>1}^{N}((v_{1,j} \times P_{1,j}) - V_r)^2}{N-1}} \tag{12}$$

A threshold score would then be calculated as:

$$V_{\sigma w} = V_r + d\sigma_w \tag{13}$$

where d would have to be determined empirically, but typically could be 2.5 or 3.0.

As before, we could then calculate the difference between a score for a given mutation from the native sequence and $S_{\sigma w}$:

$$\Delta_{\sigma w} = V_{\sigma w} - (v_{1,M} \times P_{1,M}) \tag{14}$$

where $v_{1,M}$ is the diversity score for mutated residue M against the reference sequence residue, $V_{\sigma w}$ is the thresold calculated in Equation 13 and $P_{1,M}$ is the sequence identity (or similarity) between the reference sequence and the mutated sequence (which clearly will be very high, normally only one mutation being present). A negative value would indicate an unfavourable mutation.

8 Proposal 6 — weighted minimum conservation (Version 2)

Proposal 5 does not account for having a sequence that ends up being identical to the mutant version of the first sequence.

We have three things to consider when looking at the sequences in the MSA:

1. If the sequence identity between between the mutated sequence and a sequence in the MSA is high, then any mutation observed in the MSA (irrespective of what it is) must be allowed. i.e. the impact of the mutation must be low.

2. Conversely, if the sequence identity between the mutated sequence and sequence in the MSA is low, then the appearance of the mutated residue in that sequence is less informative as it is more likely that a CPD could be responsible.

3. If we have overall high diversity at a given position (i.e. a low conservation score), then the mutation is less likely to have an impact.

We can address Point 1 by taking the diversity score for mutated residue M and dividing it by the overall sequence identity between the mutant sequence and sequence j, ($P_{M,j}$):

$$S_{M,j} = v_{M,j}/P_{M,j} \tag{15}$$

such that any variability at a given position is scaled down if the overall sequences are very similar and scaled up if the sequences are very different.

Suppose we have a sequence which is identical to the mutant sequence, then $v_{M,j} = 0.0$ and $P_{M,j} = 1.0$, so $S_{M,j} = 0.0$

However, this does not address Point 2. Suppose we have a sequence in the MSA which is very different to the mutant sequence (say 10% sequence identity), but contains the mutated residue, then $v_{M,j} = 0.0$ and $P_{M,j} = 0.1$, so $S_{M,j} = 0.0$

This is clearly *not* what we want, as it isn't addressing the CPD issue. A further problem with this approach is that we can end up with values of $S_{M,j} > 1.0$ (when both overall sequence identity and residue similarity are low)

As an alternative we could scale the *conservation* score before calculating the diversity. In this case we calculate the corrected conservation score and variability score as:

$$\begin{aligned} s^c_{M,j} &= s_{M,j} \times P_{M,j} \\ v^c_{M,j} &= 1 - s^c_{M_j} \end{aligned} \tag{16}$$

and $S_{M,j}$ then simply becomes:

$$S_{M,j} = v^c_{M,j} \tag{17}$$

Looking at the same scenarios:

- With a sequence which is identical to the mutant sequence, then $s_{M,j} = 1.0$ and $P_{M,j} = 1.0$, so $s^c_{M,j} = 1.0$ and $S_{M,j} = v^c_{M,j} = 0.0$ which, as before, is what we want.

- With a 10% identity sequence in the MSA, but containing the mutated residue, we have $s_{M,j} = 1.0$ and $P_{M,j} = 0.1$, so $s^c_{M,j} = 0.1$ and $S_{M,j} = v^c_{M,j} = 0.9$. This is better, we are penalizing this residue as it is in a very diverse sequence.

- In addition, with a 50% identity sequence in the MSA, but containing the mutated residue, we have $s_{M,j} = 1.0$ and $P_{M,j} = 0.5$, so $s^c_{M,j} = 0.5$ and $S_{M,j} = v^c_{M,j} = 0.5$.

This may be too severe a penalty, so it may be better to use

$$\begin{aligned} s^c_{M,j} &= s_{M,j} \times \sqrt{P_{M,j}} \\ v^c_{M,j} &= 1 - s^c_{M_j} \end{aligned} \tag{18}$$

which would result in values of $1-(1.00\times\sqrt{1.0}) = 0.0$, $1-(1.00\times\sqrt{0.1}) = 0.684$, $1-(1.00\times\sqrt{0.5}) = 0.293$ respectively. The power of the root could be a tunable parameter.

In summary these approaches have the following characteristics:

	Residue similarity								
	Same (s=1)			Different (s=0)			Medium (s=0.5)		
Sequence ID	1.0	0.5	0.1	1.0	0.5	0.1	1.0	0.5	0.1
Approach									
Division	0.0	0.0	0.0	1.0	2.0	10.0	0.5	1.0	5.0
Similarity	0.0	0.5	0.9	1.0	1.0	1.0	0.5	0.75	0.95
$\sqrt{}$Similarity	0.0	0.293	0.683	1.0	1.0	1.0	0.5	0.646	0.842

i.e. If the residues are the same then this becomes less informative as being acceptable as the overall sequence identity drops. If the residues are very different, then it is never accepted (at high sequence ID, we haven't seen this residue and at low sequence ID we have no information owing to CPDs). At medium residue similarity, the score is gradually penalized as the overall sequence ID drops.