Sequence Reweighting and Pseudocounts

In order to control for sequence bias in our MSA, sets of sequences that exceed a certain identity threshold are down-weighted as a group (Weigt et al. 2009; Marks et al. 2011; Morcos et al. 2011; Hopf et al. 2012). For every sequence \( m \) in an MSA, the number of “identical” sequences \( k_m \) is defined as

\[
k_m = \sum_{n=1}^{M} \theta \left( \sum_{i=1}^{L} \delta(A_i^m, B_i^n) - xL \right)
\]

where \( \theta \) is a step function equal to one if its argument is greater than or equal to zero and zero if the summation is negative, \( \delta \) is the Kronecker symbol used for counting, which is equal to one if \( A_i^m \) equals \( B_i^n \) and to zero otherwise, and \( x \) is the identity threshold, defined here as 0.7. When counting pair and single amino acid frequencies, the contribution of sequence \( m \) is down-weighted by \( 1/k_m \). The effective number of sequences in an alignment is therefore not \( M \) but \( M_{\text{eff}} \), where

\[
M_{\text{eff}} = \sum_{m=1}^{M} \frac{1}{k_m}
\]

Pair and single amino acid frequencies are then calculated according to the relationships

\[
f_i(A) = \frac{1}{\lambda + M_{\text{eff}}} \left( \frac{\lambda}{q} + \sum_{m=1}^{M} \frac{1}{k_m} \delta(A_i^m, A) \right)
\]

\[
f_{ij}(A,B) = \frac{1}{\lambda + M_{\text{eff}}} \left( \frac{\lambda}{q^2} + \sum_{m=1}^{M} \frac{1}{k_m} \delta(A_i^m, A) \delta(B_j^m, B) \right)
\]

where \( \lambda \) is a pseudocount term used to ameliorate statistical noise due to underrepresented amino acids and pairs. Here we set \( \lambda \) equal to \( M_{\text{eff}} \). Note that the empirical correlation matrix is not invertible before pseudocounts are incorporated.

DCA

According to DCA, the coupling between columns \( i \) and \( j \) in an MSA is given by the direct information, \( DI_{ij} \), score according to the relationship

\[
DI_{ij} = \sum_{A,B=1}^{q} P_{ij}(A,B) \ln \left( \frac{P_{ij}(A,B)}{f_i(A)f_j(B)} \right)
\]
where \( P_{ij}(A,B) \) represents the inferred probability of finding amino acid pair \((A,B)\) at positions \(i\) and \(j\) in the absence of interactions with other residues, \( f_i(A) \) and \( f_j(B) \) represent the single amino acid frequencies of \(A\) and \(B\) at positions \(i\) and \(j\), and the summation is evaluated over all 441 pairs \((A,B)\) possible for a \(q = 21\) state system, where the states represent the twenty amino acids and a gap. \( P_{ij}(A,B) \) is itself a function of the inferred coupling energy \( e_{ij}(A,B) \) and the inferred single residue energies \( \tilde{h}_i(A) \) and \( \tilde{h}_j(B) \) of amino acids \(A\) and \(B\) at positions \(i\) and \(j\) according to

\[
P_{ij}(A,B) = \frac{1}{Z_{ij}} \left\{ f_i(A) f_j(B) e_{ij}(A,B) + \tilde{h}_i(A) + \tilde{h}_j(B) \right\}
\]  

where \( Z_{ij} \) is the partition function. The coupling energies \( e_{ij}(A,B) \) are determined as described below by inverting an empirical correlation matrix, \( C \).

The empirical correlation matrix \( C \) is determined from the MSA according to the relationships

\[
C_{ij}(A,B)_{i \neq j} = f_{ij}(A,B) - f_i(A) f_j(B)
\]  

\[
C_{ij}(A,B)_{i=j, A \neq B} = f_i(A) (1 - f_i(A))
\]  

where \( f_i(A) \) is the frequency of amino acid \(A\) in MSA column \(i\), \( f_j(B) \) is the frequency of amino acid \(B\) in MSA column \(j\), and \( f_{ij}(A,B) \) is the frequency of amino acid pair \((A,B)\) in columns \(i\) and \(j\). Calculation of correlations \( C_{ij}(A,B) \) where \(i = j\) but \(A \neq B\) is carried out according to Equation S6. Note that pair frequencies \( f_{ij}(A,B) \) are set to zero for these entries (despite having a finite value based on pseudocounts, as described below to reflect the fact that no protein sequence contains two different amino acids at a single site. The empirical correlation matrix has the dimensions 20\(L\) by 20\(L\) despite the fact that we employ a \(q = 21\) state model. This is because one amino acid, in our case the gap, is left out of the analysis in order to serve as a reference energy.

The global nature of the DCA algorithm derives from inversion of the empirical correlation matrix (or the composite matrix \( C^* \) described below), which results in the coupling energy matrix, \( e \):

\[
e = -C^{-1}.
\]  

The fields \( \tilde{h}_i(A) \) and \( \tilde{h}_j(B) \) from Equation S5 are calculated numerically along with the partition function \( Z_{ij} \) so that the pair probabilities recapitulate the single amino acid frequencies, \( f_i(A) \) and \( f_j(B) \), observed in the MSA:

\[
\sum_{B=1}^{q} P_{ij}(A,B) = f_i(A)
\]  

\[
\sum_{A=1}^{q} P_{ij}(A,B) = f_j(B).
\]
Once field and coupling energies have been determined, direct information $D_{ij}$ scores can be evaluated using Equations S4 and S5. The result is a list of $D_{ij}$ scores representing the direct information between every pair of positions.
Supporting Literature Cited


