Abstract
Experimental studies show that evolutionary information contained in an amino acid sequence is sufficient to specify functional positions and the tertiary structure of the protein. Statistical Coupling Analysis (SCA) and Direct-Coupling Analysis (DCA) are all very useful statistical models to estimate level for analyzing the coevolutionary information from Multiple Sequence Alignment (MSA). SCA quantifies how much the amino acid distribution at some position changes upon a perurbation of the amino acid distribution at another position. The resulting statistical coupling energy indicates the degree of evolutionary dependence between the residues with higher coupling energy corresponding to increased dependence. While DCA is a global statistical model to infer the coevolutionary coupling for each pair of amino acids among residues across homologous protein families. Here we highlight how co-evolutionary information contributes to protein folding, stability and design based on statistical analysis methods from protein multiple sequence alignment. We also perform detailed calculations and analysis on three different protein families. Results show a specific pattern of amino acid interactions are sufficient and necessary in specifying protein structure, stability and function by comparing to experimental studies. In addition, we utilize the evolutionary information captured from the statistical analysis method to design a more stable (hopefully functional) protein.

Evolutionary Analysis
Experimental studies show that evolutionary information contained in an amino acid sequence is sufficient to specify functional positions and the tertiary structure of the protein. (Note: Multiple Sequence Alignment, MSA)

Statistical Coupling Analysis (SCA)

\[
\begin{align*}
D_{ij}^{(a)} &= -\ln \left( C_{ij}^{(a)} \right) + 1 - \left( C_{ij}^{(a)} \right) \\
\Delta \Delta D_{ij}^{(a,b)} &= -\frac{1}{C_{ij}^{(a,b)}} 
\end{align*}
\]

Direct Coupling Analysis (DCA)

\[
\begin{align*}
D_{ij}^{(a)} &= -\ln \left( C_{ij}^{(a)} \right) + 1 - \left( C_{ij}^{(a)} \right) \\
\Delta \Delta D_{ij}^{(a,b)} &= -\frac{1}{C_{ij}^{(a,b)}} 
\end{align*}
\]

Statistical Model

\[
P(A_i A_j | A_k A_l) = \frac{1}{\sum_{C_{ij}^{(a)}}^{N}} \sum_{C_{ij}^{(a)}}^{N} \exp \left( \sum_{A_i A_j | A_k A_l}^{N} \eta_A (A_i A_j | A_k A_l) \right)
\]

Conclusion

- SCA and DCA can capture significant evolutionary information.
- Couplings from DCA can be used to assist protein design.
- Statistical analysis can be applied to proteins in neurobiology and brain science.

References

Results

Statistical Coupling Analysis (SCA) for Folding and Stability

Sequence Statistics

Maximal Entropy Model

\[
P(A_i A_j | A_k A_l) = \frac{1}{\sum_{C_{ij}^{(a)}}^{N}} \sum_{C_{ij}^{(a)}}^{N} \exp \left( \sum_{A_i A_j | A_k A_l}^{N} \eta_A (A_i A_j | A_k A_l) \right)
\]

Statistical Model

Conclusion