Improved and automated residue function determination from multiple sequence alignment

Kai Ye[*‡], Gert Vriend[†] and Adriaan P. IJzerman[*]

[*]Division of Medicinal Chemistry, Leiden/Amsterdam Center for Drug Research, Leiden University, Einsteinweg 55, 2333CC, Leiden, The Netherlands; and [†]CMBI, Radboud University Nijmegen, Toernooiveld 1, 6525 ED, Nijmegen, The Netherlands.

ABSTRACT

Multiple sequence alignments (MSAs) have shown to hold information about many aspects of proteins, ranging from their phylogenetic origin till the function of individual amino acids. A series of techniques exists that relate residue positions with function. These techniques tend to be based on measures of conservation or correlated mutation. One powerful method to determine the function of individual residue is the so-called two-entropies analysis[1] that is based on a scatter plot of the entropy over entire column in the MSA versus the average of the entropy values at this same position in each subfamily. This so-called two-entropies analysis (TEA1) is powerful in identifying both conserved and specificity positions, but comes with the disadvantage that a human expert must decide on the subdivision of all sequences in families. In this study, we have overcome the subjective human interference step in the TEA1 by simply performing the whole analysis once for every possible objective division of the sequences in subfamilies according to a phylogenetic tree, and average all resulting two-entropies plots. The resulting, single, objective TEA2 plot (TEA2, two-entropies analysis 2) classifies residues with a known function equally well, or sometimes even a bit better than previous methods. In addition to the conserved and specificity positions, TEA2 also identify positions provide communication between conserved and specificity positions.

[1]Ye, K., Lameijer, E. W., Beukers, M. W. & IJzerman, A. P. (2006) Proteins 63, 1018-30.