Using sequence weighting to improve residue correlation analisys

Lucas Carrijo, Lucas Bleicher

Institute of Biological Sciences, Federal University of Minas Gerais, Brazil.

Analysing a multiple sequece alignment at the residue level, apart from the conserved positions, there are other patterns that are also indicative of functional importance and reflect functional divergence within a homologous protein family due to gene duplication. In families that have subfamilies with distinct functional specificities, some positions can be conserved only in a particular subfamily, or the conserved amino acid can be different for each of the subfamilies. This suggests that the role of this residue relates not to the global function of the family, but to functional specificities of that group. In these cases, it is reasonable that such specificities are not determined by the presence of a single residue, but by a group of residues, and this group will emerge from residue correlation analysis since a sufficient amount of proteins show the same specificities. However, some protein families have subfamilies less represented in terms of amount of sequences in the alignments. Meantime, this alignments use to come full of redundant sequences, many times mutants or variants of the same sequence, originary mainly from model organisms. This redundancy in the alignments tend to introduce bias to analysis with a statistical mean like the correlation methods. In this way, the present work has as objective to compare the effects of distinct approaches aiming the decreasing of redundancy in multiple sequence alignments: sequence weighting and filtering by maximum identity. Besides, this work also proposes approaches to make the correlation calculations compatible with sequence weighting, in order to improve analisys of residue conservation and correlation. Sequence weighting was capable of highlighting frequencies of amino acids specific of less sampled subfamilies, while decreasing the frequencies of amino acids present in redundant sequences. The adapted calculations were capable of detecting such differences, providing a good alternative to conservation and correlation analisys in alignments that are less representative of the actual protein diversity existent in nature.