Sequence Alignment and Structural Disorder: A Substitution Matrix for an Extended Alphabet

Zoran Obradovic

Director, Center for Information Science and Technology, Temple University, Philadelphia, USA

ABSTRACT: In protein sequence alignment algorithms, a substitution matrix of 20x20 alignment parameters is used to describe the rates of amino acid substitutions over time. Development and evaluation of most substitution matrices including the BLOSUM family was based almost entirely on fully structured proteins. Structurally disordered proteins (i.e. proteins that lack structure, either in part or as a whole) that have been shown to be very common in nature have a significantly different amino acid composition than ordered (i.e. structured) proteins. Furthermore, the sequence evolution rate is higher in unstructured as compared to structured regions of proteins containing both structured and unstructured regions. These results cast doubt on appropriateness of the BLOSUM substitution matrices for alignment of structurally disordered proteins. To address this problem, we take into the account the concept of structural disorder by extending the alphabet for sequence representation to 2x20=40 symbols, 20 for amino acids in disordered regions and 20 for amino acids in ordered regions. A 40x40 substitution matrix is required for alignment of sequences represented in the extended alphabet. Such an expanded matrix contains 20x20 submatrices that correspond to matching ordered-ordered, ordered-disordered, and disordered-disordered pairs of residues. In this talk we will describe an iterative procedure that we used to estimate such a 40x40 substitution matrix. The iterative procedure converged with stable results with respect to the choice of the sequences in the dataset. In the obtained 40x40 matrix we found substantial differences between the 20x20 submatrices corresponding to ordered-ordered, ordered-disordered, and disordered-disordered region matching. These differences provide evidence that for alignment of protein sequences that contain disordered segments, the discovered substitution matrix is more appropriate than the BLOSUM substitution matrices. At the same time, the new substitution matrix is applicable for sequence alignment of fully ordered proteins as its order-order submatrix is very similar to a BLOSUM matrix.

Reported results were obtained in collaboration with Uros Midic and A. Keith Dunker and will appear at *Proc. Workshop on Statistical and Relational Learning and Mining in Bioinformatics at the 15th ACM SIGKDD Conf. Knowledge Discovery and Data Mining*, Paris, France, June 2009.

SPEAKER: Zoran Obradovic, professor of computer and information sciences and the director of the Center for Information Science and Technology at Temple University in Philadelphia is an internationally recognized leader in data mining and bioinformatics. He has published more than 200 articles addressing data mining challenges in health informatics, the social sciences, environmental management and other domains. His group's pioneering research on the prediction and functional analysis of intrinsically disordered regions in proteins has provided new insight into how protein structure establishes function and the program his team developed was the best rated predictor of intrinsic disorder at three consecutive international competitions organized by protein structure prediction assessment community (CASP 5-7). Obradovic was the program chair at five, track chair at seven and program committee member at about 40 international conferences on data mining. He currently serves as an editorial board member at seven journals.