Poster I-68

Prediction of order and disorder at the domain level

Authors:
Zsuzsanna Dosztanyi (*Institute of Enzymology*)
Mark Sandor (*Institute of Enzymology*)
Istvan Simon (*Institute of Enzymology*)

Short Abstract: During target selection of structural genomics projects it is important to be able to discern ordered globular domains from disordered regions. A method based on the IUPred algorithm is suggested to locate ordered and disodered domains from amino acid sequences and was tested on datasets containing both order and disorder.

Long Abstract:
Intrinsically unstructured/disordered proteins (IUPs) lack a stable 3D structure, rather exist in highly flexible, unfolded structural state. Disordered regions often lead to difficulties in protein expression, purification and crystallization. Therefore, it is important to be able to discern ordered globular domains from disordered regions during the target selection process of structural genomics projects. Existing prediction methods assign order/disorder at the level of individual positions and their performance was tested using fully ordered or disordered regions. The task of predicting order and disorder at the level of domains, however, requires different type of datasets, algorithms, and evaluation criteria. Only a few methods have been tuned for prediction of order and disorder at domain level [1,2,3], but no systematic study is available.
The main challenge here is to find the boundaries between disordered and ordered domain sized regions, without predicting many short segments. Compared to the observed domain sizes, predictions often contain only short continuous stretches of ordered or disordered positions. Noise in the prediction methods can be one source of such unavoidable fragmentation. However, short regions can also occur when the prediction methods realistically capture some local tendencies for order or disorder. In most cases, these regions should be overlooked aiming at prediction at the domain level. For example, many X-ray structures contain regions with missing electron density, which do not prevent overall structure determination. Consequently, they should be merged into globular domains. The opposite case, short ordered regions within large disordered regions were also observed and were suggested to correspond to functional sites. Such regions are not capable of forming stable structures in themselves and should be considered as part of disordered regions. However, the exact requirements to form globular domains or disordered ones are not known.
The optimal rules for discerning domains and regions for both order and disorder were explored via various algorithms. The predictions were based on the IUPred algorithm [4] that predicts the energetic contribution of each residue from the amino acid composition of its neighborhood. It assumes that lack of 3D structure for IUPs is the result of their specific amino acid composition which does not allow sufficient favorable interactions to form. The strength of this approach is that its parameters are based on globular proteins only, without relying on datasets of unstructured proteins. Starting from IUPred predictions, various

algorithms have been implemented in order to tune the performance for recognizing ordered domains and long disordered segments. Each algorithm can be described by a few basic rules and required only a few parameters. Some of these were based on the length of regions. Taking advantage of the physical model behind IUPred, the energy of regions can also be used to decide which regions should be treated self-contained.
In the first algorithm we followed the procedure applied in GlobPlot [2]. This eliminates short segments based on their length only. Neighboring globular domains were merged if the intervening disordered region was too short, and ordered regions below some minimal size were omitted. In another algorithm an iterative procedure was implemented. Regions were ranked according to their lengths, and short regions were reassigned to their neighborhood. Different size limits were used for ordered and disordered regions. A similar algorithm used the energy instead of the length of segments for ranking and merging. In the last set of algorithms we tried to find directly the starting and end position of segments with the most favorable energy. The search was repeated until non-overlapping segments could be found in the sequence.
The evaluation of these algorithms requires a specific dataset containing long regions of both order and disorder, mixed the same way as observed in proteins. The performance of the algorithms was tested on two datasets. Only a small dataset could be collected from experimentally verified disordered regions which were juxtaposed by domains with known structure. A much larger dataset was generated using the mostly ordered domains of the Pfam database [5]. Regions containing both a Pfam family and an unassigned putative disordered region were selected for the purpose of optimizing the parameters and testing the performance of the various algorithms. Unlike in domain prediction, consecutive globular domains were treated as a single ordered unit. Although this larger dataset is less reliable, the performances and the optimal parameters were similar to that of the small dataset.

[1] Oldfield CJ, Ulrich EL, Chen Y, Dunker AK, Markley JL (2005) Proteins 59:444-453
[2] Linding R, Russell RB, Neduva V, Gibson TJ (2004) Nucleic Acids Res. 31: 3701-3708.
[3] Dosztányi Z, Csizmók V, Tompa P, Simon I (2005) Bioinformatics 21:3433-3434
[4] Dosztányi Z, Csizmók V, Tompa P, Simon I (2005) J. Mol. Biol. 347:827-839.
[5] Bateman A, Coin L, Durbin R, Finn RD, et al. (2004) Nucleic Acids Res. 32:D138-D141.