PROTEIN LOCAL STRUCTURE PREDICTION BASED ON THE PHYSICO-CHEMICAL PROPERTIES OF THE SEQUENCE

Milchevskyi Yu.V., Nikitin A.M., Lukshin S.A., Milchevskaya V.Yu.[1], Tumanian V.G.

Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, Russia, 119991, Moscow, Vavilov str, 32, milch@eimb.ru
[1]European Molecular Biology Laboratory, Meyerhofstraße 1, 69117 Heidelberg, Germany Rastatter strasse 10, Heidelberg 69126, Germany

Most recent prediction methods (Nature 2016) [1] can obtain as much as 84% Q3 accuracy and 72% Q8 accuracy. Nevertheless, it is not possible to recover the Cartesian coordinates of the protein fragment under study even if the correct prediction is available.

We present a new approach to predict local structure in proteins, which allows to recover the Cartesian coordinates of the protein fragment. First, we describe a new definition of structural clusters (referred to as ”general coordinate functions”), that serve as a basis in the space of protein structures. We explain the minimal number of the basis coordinate functions necessary to recover the Cartesian coordinates for protein fragments of different length. Additionally, several assumptions are considered: with the dihedral angles fixed/ flexible, valence angles fixed/flexible, different bond distances allowed. Further we focused on prediction of the structure for 5-bp long peptides. To obtain general coordinate functions by clustering all 5-bp long fragments of the protein structures from PDB, which resulted in 30 clusters, i.e. 30 basis structures. With the the proposed general coordinates, one can reduce the multi-dimensional problem of sequence-based local structure prediction to several univariate problems. Same as in Cartesian coordinates the location of a point is defined by distances from the basis vectors, here the structure is defined by distances from the general coordinate functions. Based on the sequence of a protein fragment, we predict its distance to each general coordinate, and thus connect physico-chemical properties of the sequence to distances from the structural clusters [2].

We implemented a server milch.eimb.ru that predicts local structure of a protein based on its sequence. The prediction obtains 72% accuracy for the 30 clusters proposed above.

References.

1. Wang, S. et al. Protein Secondary Structure Prediction Using Deep Convolutional Neural Fields. Sci. Rep. 6, 18962; doi: 10.1038/srep18962 (2016).
2. http://www.genome.jp/aaindex/