Probabilistic context-free grammars for classification of helix-helix contact sites and recognition of amyloidogenic peptides

Witold Dyrka, Florence Thirion, Jean-Christophe Nebel, Malgorzata Kotulska

Probabilistic context-free grammars for classification of helix-helix contact sites and recognition of amyloidogenic peptides

Witold Dyrka [a, b *], Florence Thirion [a], Jean-Christophe Nebel [c], Malgorzata Kotulska [a]

[a] Institute of Biomedical Engineering and Instrumentation, Wroclaw University of Technology, Poland
[b] Inria Centre de Research Sud-Ouest, Bordeaux, France
[c] School of Computing and Information Systems, Faculty of Science, Engineering and Computing, Kingston University, London
*e-mail: witold.dyrka@pwr.wroc.pl

Keywords:
probabilistic context-free grammar, grammar inference, helix-helix pairs, amyloidogenic peptides

Hidden Markov Models power many state-of-the-art tools in the field of protein bioinformatics. While excelling in their tasks, these methods of protein analysis do not convey directly information on medium and long-range residue-residue interactions. This requires an expressive power of at least context-free grammars. However, application of more powerful grammar formalisms to protein analysis has been surprisingly limited. We have developed a probabilistic grammatical framework for problem-specific protein languages, which has been already successfully applied to recognition of ligand binding sites. The core of the model consists of a probabilistic context-free grammar (PCFG), automatically inferred by a genetic algorithm from only a generic set of expert-based rules and positive training sequences. Here, we show that the PCFG approach matches state-of-the-art performance in two other tasks: classification of transmembrane helix-helix pairs and recognition of amyloidogenic peptides. First, the framework was applied to produce grammar descriptors of four classes of transmembrane helix-helix contact sites. The highest performance of the classifiers reached AUC ROC of 0.70. Second, the analogous approach was used to distinguish between amyloidogenic and non-amyloidogenic protein fragments. It yielded good results whether these fragments were isolated or within an entire protein (AUC ROC up to 0.80). Finally, an attempt to model pairing amyloidogenic fragments resulted in classifiers reaching AUC ROC of 0.70. A significant feature of the PCFG method is that grammar rules and parse trees are human-readable, and thus could provide biologically meaningful information.