Poster I-36

DomainDiscovery: A Novel Algorithm for Protein Domain Boundary Assignment Using Support Vector Machine

Authors:
Abdur R Sikder (*Advanced Networks Research Group, School of Information Technologies, University of Sydney, NSW 2006*)
Stella Veretnik (*San Diego Supercomputer Center, University of California San Diego, 9500 Gilman Drive, La Jolla, CA*)
Albert Y Zomaya (*Advanced Networks Research Group, School of Information Technologies, University of Sydney, NSW 2006*)
Philip E Bourne (*Department of Pharmacology, University of California San Diego, 9500 Gilman Drive, La Jolla, CA 9209*)

Short Abstract: Knowledge of protein domain boundaries is critical for the characterisation and understanding of protein function, specifically in the post genomic era. The ability to identify structural domains without the knowledge of the structure – by using sequence information only – is essential step in many types of protein analysis. We present a novel method for domain identification from sequence-based information.

Long Abstract:
Background

Knowledge of protein domain boundaries is critical for the characterisation and understanding of protein function, specifically in the post genomic era. The ability to identify structural domains without the knowledge of the structure – by using sequence information only – is essential step in many types of protein analysis. We present a novel method for domain identification from sequence-based information. DomainDiscovery uses a Support Vector Machine (SVM) approach and a unique training dataset built on the principle of consensus among experts of protein structure.
Results
DomainDiscovery method is tested and compared with others on a structurally non-redundant dataset, as well as CASP5 targets. DomainDiscovery achieves above 52% accurate domain boundary identification for multi-domains protein chains from sequence information. DomainDiscovery is a machine learning approach to domain boundary prediction. We trained Support Vector Machine (SVM) using PSSM (Position Specific Scoring Matrix), Secondary Structure and Solvent accessibility information to detect possible domain boundaries for a target sequence.
Conclusions

We have presented a new protein domain boundary prediction method, DomainDiscovery, based on support vector machine (SVM) and training with structurally-defined domains based on consensus among experts.
In six-fold cross-validation technique using Benchmark_2 dataset we achieve 53% accuracy for the data that includes single-domain and multi-domain chains. Performance of

DomainDiscovery is comparable or better than other recent sequence-based methods, particularly with regards to its performance on multi-domain chains.
Additionally, a new evaluation method, Precision of Boundary Placement (PBP) is introduced and applied.