Poster C-35

Phylogenetic studies on UniProtKB/Swiss-Prot families and data from complete proteomes

Authors:
Boeckmann (*SIB*)
Kriventseva (*University of Geneva*)
Amendolia (*SIB*)
Bairoch (*SIB & University of Geneva*)

Short Abstract: Swiss-Prot families and data from complete proteomes have recently been analyzed in various studies. Results suggest that the protein groupings are of high quality, but there is room for improvements. Classification issues could be addressed by the integration of relevant analysis methods into the annotation platform.

Long Abstract:

Introduction

The UniProtKB/Swiss-Prot Protein Knowledgebase [1,2] is well known for its high annotation standards including its presentation of accurate sequence data, which is a prerequisite for any meaningful sequence analysis. Similar proteins from related species are commonly collectively curated or updated, which ensures the same quality standards over homologous protein groups. Recently, such protein groupings have been subject to various studies.

Analysis

4 very different types of phylogenetic studies have been performed on Swiss-Prot: 1) a large-scale analysis of sequence data from human and mouse based on 7300 orthologous protein pairs retrieved from UniProtKB was conducted to investigate in protein evolution. In the course of this study, especially alternate isoforms and multigene families were analyzed in depth, resulting in the correction of 4% of the sequence data [3]; 2) a phylogenetic analysis workbench has been developed which includes various alignment and tree-building methods (BioNJ, PhyML, the Phylip package, TreePuzzle etc). Until now we analyzed about 1000 protein groups, testing distinct methods and parameter settings on randomly selected protein groups, which have been retrieved according their annotation; 3) HAMAP families of UniProtKB/Swiss-Prot were analyzed by the group of Gaston Gonnet (ETHZ), using a method for the detection of non-orthology [4]; randomly selected protein groups were subsequently studied by phylogenetic tree-building methods; 4) a complementary study is the analysis of coding genes from completely sequenced genomes through a large-scale classification of proteins into hierarchical protein groups using all-against-all sequence distance comparisons. We have identified group of orthologous genes in 11 recently completed vertebrate genomes, which are now analyzed using the phylogenetic workbench.

Conclusions and outlook

The results of these studies suggests that the Swiss-Prot protein groupings are of high quality, but there is room for improvements. Classification issues could be addressed by the integration of relevant analysis methods into the annotation platforms. The combination of large-scale horizontal analysis of data from complete genomes with the specific vertical

analysis of homologous protein groups during the annotation process seems to be a promising approach leading to a confident protein classification.

References

[1] Wu CH, Apweiler R, Bairoch A, Natale DA, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Mazumder R, O'Donovan C, Redaschi N, Suzek B. The Universal Protein Resource (UniProt): an expanding universe of protein information. NAR 34:D187-D191 (2006).

[2] Bairoch A, Boeckmann B, Ferro S, Gasteiger E. Swiss-Prot: juggling between evolution and stability. Brief Bioinform. 2004 Mar;5(1):39-55.

[3] Amendolia V. Molecular Evolution on Proteomes. Comparing human and mouse orthologous proteins. Diploma (2005), University of Pisa, Italy.

[4] Dessimoz C., Boeckmann B., Roth A., Gonnet G.H. Detecting Non-Orthology in the COG Database and Other Approaches Grouping Orthologs Using Genome-Specific Best Hits. NAR (accepted).