Combining rapid word searches with segment-to-segment alignment for sensitive similarity detection, domain identification and structural modelling.

Matej Lexa [1,2] and Giorgio Valle [1]

[1]CRIBI Biotechnology Centre, University of Padova, Via Ugo Bassi 58/b, 35131 Padova, Italy

[2]Laboratory of Functional Genomics and Proteomics, Faculty of Sciences, Masaryk University Brno, Kotlarska 2, 61137 Brno, Czech Republic

Keywords. Sequence alignment, similarity search, database mining, structural neighbors, active site.

INTRODUCTION

The most popular alignment and similarity search techniques are based on the classical Smith-Waterman scoring scheme. Conservation of a single structural or functional feature between proteins may be undetectable, because the similarities tend to persist only in the key areas, consisting of residues dispersed in a non-trivial manner. We propose a novel method that finds occurrences of short similar words common to the studied sequences and handles the identified matches in a manner similar to segment-to-segment alignment [2]. Our interest in this area stems from the development of programs for fast searches with mismatches in large biological databases [1]. As shown here, these programs can support large database searches that lead to automatic domain detection, sequence annotation. The use of this technique in fold-recognition and structure prediction is being studied.

RESULTS

1. SEARCH FOR SIMILARITIES

The recently developed PEPTIMEX server (http://bioinformatics.cribi.unipd.it/primex/) is approaching speeds that can support an exhaustive sequence-to-database search. The program is derived from PRIMEX [1] and can identify all approximate matches to an 8 aa word in a middle-sized protein database in about 1 s, a complete segment-to-segment relationship between a 500 aa protein and the database can be generated in a matter of minutes. We installed the program on our computers, serving the PDB, Arabidopsis and SwissProt protein databases. Perl scripts were used to collect the necessary information. The results are stored as Protein Similarity Records (PSRs) in an SQL database. The PSRs were processed by other programs to score the similarities between individual protein pairs, to find domains and to generate alignments and other data.

2. DOMAIN IDENTIFICATION

We propose a novel method for detecting domains, that does not require actual alignment of the sequences in the PSR file. For domain detection, we find all pairs of segments in the query sequence that are correlated in other proteins. The assumption is that if two words occur in the same domain, they have a higher chance of being found together in other proteins, than words belonging to different domains. We place domain boundaries in positions that break the minimal number of such correlations. Our preliminary results are very promising (Fig.1) and we are preparing to generate a full catalog of domains based on this approach. Moreover, we find that the best correlated segments are located close to each other in space in the native protein structure. We examine the possibility to predict crude structures from related PDB files or to define structural constraints that are based solely on sequence information and similarity data.

3. PROTEIN ANNOTATION

A logical extension of domain identification is automatic annotation of sequences. Common words can be extracted from Gene Ontology definitions or FASTA headers. We will present examples of automatic annotation based on PEPTIMEX searches.

RESULTS

We present the results of combining a proven method of alignment and similarity search with a rapid word search method developed in our laboratory. We show that segment-to-segment relationships between proteins can be used to predict important functional and structural properties of unknown sequences. The first comparisons of the described method with established procedures suggests that it has a potential to provide biologists with new information that could accelerate research in several areas.

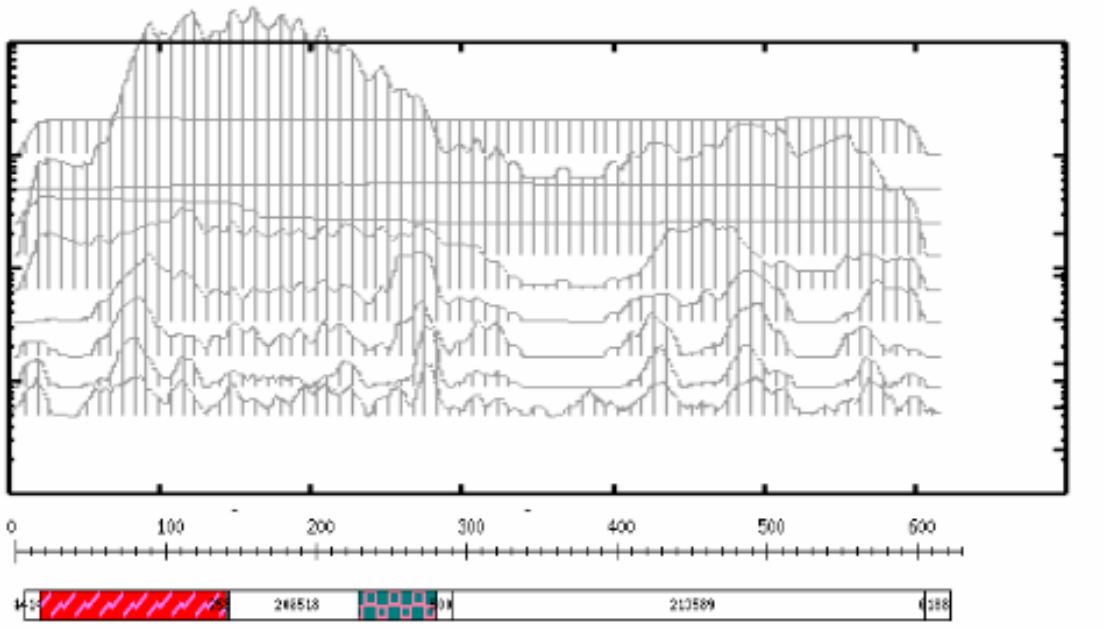

Fig. 1 - A graph showing the number of correlations spanning each position in the studied protein sequence (At5g07210) compared to PRODOM entries. The local minima of this function represent possible domain boundaries.

References

[1] M. Lexa and G. Valle, PRIMEX: Rapid identification of oligonucleotide matches in whole genomes. *Bioinformatics*, 19:2486-2488, 2003

[2] B. Morgenstern, DIALIGN2: improvement of the segment-to-segment approach to multiple sequence alignment. Bioinformatics, 15:211-218, 2003