

Application of a Novel Structured Pattern Database to Identifying Sequence Motif Conservation in Protein Families

Zayed I. Albertyn, Tay Liang Chung, M. Ramachandran, and Robert G. Hercus.

The discovery of conserved sequence motifs from families of aligned protein sequences has seen the development of multiple resources for assigning functional information to uncharacterized proteins. A novel utilization of a proprietary technology for managing genomics data, SynaBASE™, has been successfully applied for identification of sequence conservation in protein families. Given an unaligned set of nucleotide or protein data, SynaBASE generates a set of patterns or motifs and maintains their relationships in an efficient manner. These relationships are described by shared sequence motifs in the source data set of sequences. Pattern SIGNIFICANCE in SynaBASE was evaluated by mining known conserved sequences from the BLOCKS+ database against a protein SynaBASE built from SWISS-PROT Release 43.0. Results indicated that patterns that were grouped into the substitution groups of amino acids in protein families correlate strongly with published data relating to conservation of protein family motifs. These findings suggest that patterns with high-SIGNIFICANCE, built from unaligned protein datasets match the motifs described in the BLOCKS database. This approach may be well suited for searching entire proteomes or assigning biological functions to proteins encoded by newly sequenced genomes.

Introduction

The identification and annotation of new genes in a newly sequenced genome can be achieved by the use of highly specific sequence comparison methods. A systematic and objective approach to mining genomic data for functional information has been the application of motifs belonging to gene/protein families [1]. The basic premise is that a significant degree of sequence similarity between two sequences implies a conserved function. Sequence motifs specific to a particular family could then be used to discover new family members based on a known conserved signatures.

Protein sequences from multiple family members can be aligned to each other and blocks of sequence conservation can then be extracted for defining conserved sequence patterns in a family. Many protein family or motif databases such as BLOCKS, Prints, Pfam and Prosite collate this information for identification of functional elements in a suspected gene sequence [2-5]. The BLOCKS database contains short, un-gapped regions that are highly conserved, according to sequence characteristics [1-2]. BLOCKS defines conserved and consensus patterns for protein families from these detailed alignments [2].

To refine the process of defining conserved sequence motifs in protein families, a structured database of protein patterns, SynaBASE™, has been constructed from all proteins in the SWISS-PROT database [6]. SynaBASE automatically calculates and stores associations between sequences based on shared patterns. Adding more source data to SynaBASE makes it more efficient as redundancy is minimised. Pattern frequency and SIGNIFICANCE i.e. the probability of predecessor and successor characters, are already known, the process of identifying blocks of sequence conservation can be done without the need for computationally intensive multiple sequence alignments [7]. Scoring the SIGNIFICANCE of patterns stored in SynaBASE provides a novel approach for analysis of protein sequence data (as well as other genomics data) for conserved pattern information and improving upon traditional sequence comparison methods.

In order to demonstrate that SynaBASE stores conserved motifs in protein families as significant patterns, signatures from the BLOCKS+ database were used as queries against a SWISS-PROT version of SynaBASE. A SWISS-PROT SynaBASE was also built using a simplified alphabet system to improve detection of conserved motifs in protein families.

Methods

Blocks Cobbler sequences (referred to as IPBXXXXX) are consensus sequences from the BLOCKS+ database (<http://blocks.fhcrc.org>) that only contain aligned segments. A sample set of these sequences from a few chosen protein families were queried against SynaBASE, built from SWISS-PROT,43.0, and a simplified alphabet version of SynaBASE SWISS-PROT [6].

For the simplified alphabet database, the 20 amino acids are partitioned into 6 substitution groups as listed below: [FILMVYWH],[EKQRDN],[AST], C,G and P. These groups were identified according to their propensity to occur in the same alignment columns in the BLOCKS and HSSP database [1,8]. The simplified alphabet treats each member of a group as a single character, thereby allowing for variability and increasing the probability of a character match in an alignment.

Results & Discussion

An analysis of the consensus BLOCK for alpha tubulin (IPB002452) against SynaBASE SWISS-PROT showed a significant match along the entire length of the sequence (Figure 1 below). An almost identical result was found by querying the sequence against the simplified alphabet database (data not shown). The run times for SWISS-PROT SynaBASE and the simplified alphabet were 198 and 321 milliseconds respectively. Therefore alpha tubulin BLOCKS signatures correlated strongly with SynaBASE patterns by consensus sequences having high SIGNIFICANCE values in SynaBASE.

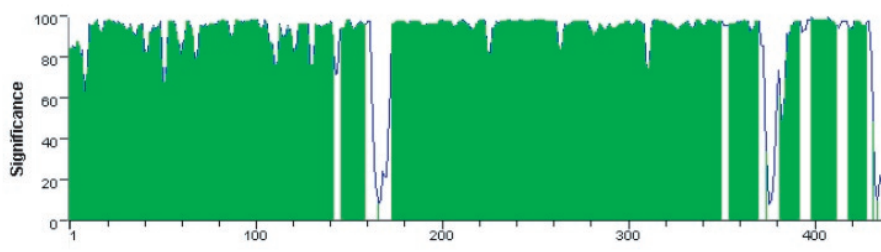


Figure 1 Synamine analysis of the alpha tubulin consensus BLOCK (IPB002452). Green areas indicate SWISS-PROT keywords linked to SynaBASE patterns. The query sequence consisted of all aligned segments from proteins in this family. High-scoring SIGNIFICANCE reflects a high degree conservation of these patterns in SynaBASE.

SynaBASE pattern SIGNIFICANCE is the probability of predicting a sequence of characters given the occurrence of these patterns in SynaBASE. In light of sequence motifs it can be seen as a measure of sequence conservation based solely on pattern

data. Therefore if sections of conserved amino acid motifs defined by the BLOCKS database have high SIGNIFICANCE scores in the SWISS-PROT SynaBASE scale, it then provides evidence that these blocks are being identified by SynaBASE.

SynaMine™ results from analysis of a putative zinc finger domain (IPB007808) showed an increase in SynaBASE pattern coverage based on simplified alphabets of substitution groups (Figure 2 below). The run time for the normal and simplified alphabet SynaBASE was 59- and 900 milliseconds respectively. The difference in results could be seen where a higher SIGNIFICANCE was observed for patterns matching the simplified alphabet SynaBASE. Pattern frequency also confirms motif conservation in that specific patterns defined by BLOCKS alignments occur frequently in SynaBASE.

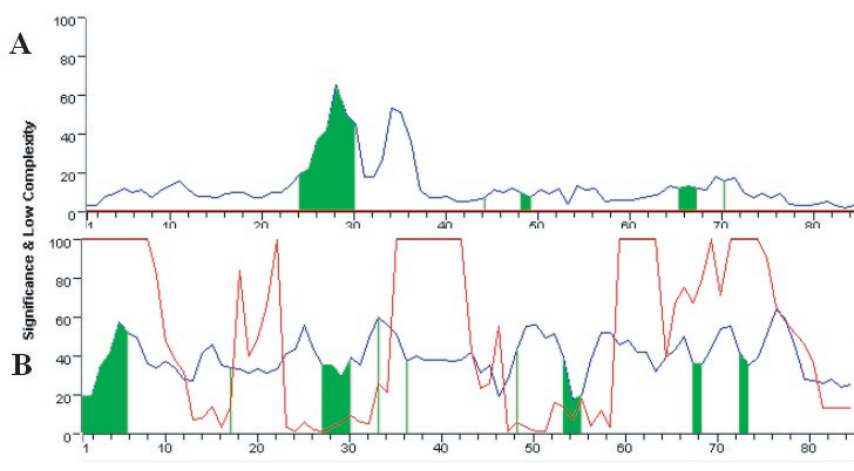


Figure 2 Increase motif coverage in SynaBASE by using a simplified alphabet of amino acid substitution groups. Results are shown for a putative zinc finger domain (IPB007808) BLOCKS+ entry mined against a) SynaBASE SWISS-PROT and b) SynaBASE simplified alphabet of amino acid substitution groups. The red and blue line indicate relative frequency and SIGNIFICANCE respectively. Patterns with keywords are shaded in green.

Another example of the increased coverage of conserved motifs was observed in the receptor tyrosine kinase, class II (IPB002011) example shown in Figure 3 on the next page. Therefore the use of guided rules describing amino acid substitution groups can be used to improve coverage of patterns in SynaBASE for enhanced motif discovery.

Conclusion

SynaBASE SIGNIFICANCE is able to identify and quantify sequence similarity without prior information of protein families or conventional alignments. This feature may be used to complement and improve the quality of information in existing

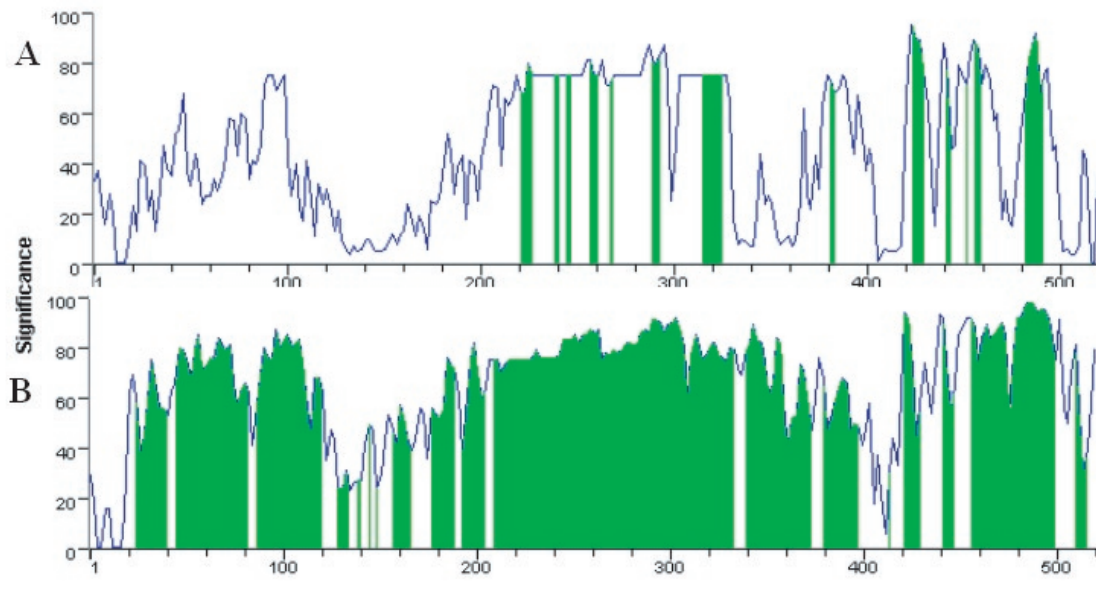


Figure 3 Increase motif coverage in SynaBASE by using a simplified alphabet. A receptor tyrosine kinase class II consensus BLOCK IPB002011 was queried against a) SynaBASE SWISS-PROT and b) SynaBASE simplified alphabet of amino acid substitution groups. SynaBASE SIGNIFICANCE is shown as a blue line and patterns with SWISS-PROT keywords are indicated by areas shaded in green. Elapsed time for queries a) and b) were 81 milliseconds and 3 seconds respectively.

protein family databases such as BLOCKS, Pfam and Prosite.

Sequence conservation in SynaBASE, calculated as pattern SIGNIFICANCE, was identified in known examples of the most common protein families. In addition to SynaBASE patterns storing all conserved protein signatures for family identification, the use of a simplified alphabet of amino acid substitution groups improved the quality of motif discovery.

SWISS-PROT Patterns stored in SynaBASE exhibit the most value in assigning function to newly sequence proteins, improving protein family definitions in motif databases and identifying protein coding regions in newly sequenced genomes. The scalability of current SynaBASE applications permits the analysis of single proteins to whole mammalian chromosomes ranging from milliseconds to minutes, respectively. SynaBASE patterns can hence be used to assign function to proteins based on shared motifs in the absence of any homology identifiable by traditional sequence comparison methods.

REFERENCES

1. Nevill-Manning CG, Wu TD, Brutlag DL. (1998). Highly specific protein sequence motifs for genome analysis. *Proc Natl Acad Sci* 95, 5865-71.
2. Henikoff JG, Pietrokovski S, McCallum CM, Henikoff S (2000). Blocks-based methods for detecting protein homology. *Electrophoresis* 21,1700-6
3. Attwood TK (2002). The PRINTS database: a resource for identification of protein families. 3, 252-63.
4. Bateman A, Coin L, Durbin R, Finn RD, Hollich V,

Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL, Studholme DJ, Yeats C, Eddy SR. The Pfam protein families database. *Nucleic Acids Res.*1,32.

5. Sigrist CJ, Cerutti L, Hulo N, Gattiker A, Falquet L, Pagni M, Bairoch A, Bucher P (2002). PROSITE: a documented database using patterns and profiles as motif descriptors. *Brief Bioinform.* 3, 265-74.
6. Boeckmann B., Bairoch A., Apweiler R., Blatter M.-C., Estreicher A., Gasteiger E., Martin M.J., Michoud K., O'Donovan C., Phan I., Pilbout S., Schneider M (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL. *Nucleic Acids Research* 31, 365-370.
7. Albertyn, ZI, Wong CS, Tay, L and Hercus, RG (2004) A new Approach to Genome-wide Annotation Based upon Calculation of SIGNIFICANCE from a Structured Pattern Database. Available online at www.synamatix.com
8. Dodge C, Schneider R, Sander C (1998). The HSP database of protein structure-sequence alignments and family profiles. *Nucleic Acids Res.* 26, 313-5.

To request detailed technical information or feedback please send an email to tech@synamatix.com

This Research Newsletter is for distribution to Synamatix members, associate members and mailing list subscribers only. The contents are provided for personal, non-commercial purposes and are protected by various national and international intellectual property laws, conventions and treaties. All title and intellectual property rights in and to Synamatix, SynaBASE, SynaSearch and the accompanying printed materials are owned by Synamatix sdn bhd. Other trademarks or names are used only in an editorial fashion and to the benefit of the respective trademark owner with no intention of the infringement of the trademark. All trademarks or service marks are the property of their respective owners.