Identifying Sequence-Structure Patterns

Tom Milledge[1], Chengyong Yang[1], Gaolin Zheng[1], Xintao Wei[1], Sawsan Khuri [2], and Giri Narasimhan[1],
[1]Bioinformatics Research Group (BioRG), School of Computer Science, Florida International University, Miami, FL.
[2]The Dr. John T. Macdonald Foundation Center for Medical Genetics, University of Miami School of Medicine, Miami, FL

Abstract

Proteins that share a similar function often exhibit conserved "sequence patterns" or "signatures" or "motifs". Such sequence signatures are derived from multiple sequence alignments and have been collected in databases such as PROSITE, PRINTS, and eMOTIF. Recent research has shown that these domain signatures often exhibit specific three-dimensional structures (Kasuya *et al.*, 1999; Mondal *et al.*, 2003). We, therefore, hypothesized that sequence patterns derived from structural information would have superior discrimination ability than those derived by other methods.

Here we show how to start with a sequence signature and use it to design meaningful sequence-structure patterns (SSPs) from a combination of sequence and structure information. Given a seed signature from one of the current databases, a set of structurally related proteins was generated via a pattern search of the protein structures compiled at the ASTRAL web site. After performing a multiple structure alignment based on the pattern residues, improved SSPs were obtained by including aligned positions containing either a single conserved residue or a context-specific substitution group (Wu and Brutlag, 1996). The patterns were further enhanced by looking for association rules generated by application of the APRIORI algorithm to the sequence alignment. These association rules indicate structurally adjacent residue positions in the protein that are mutually constrained and therefore correlated. By focusing on small core regions of the protein in which a high packing density constrains the substitution of one residue for another, we generated improved SSPs that outperformed existing profiles in the identification of a number of functional domains. The quality of our improved SSPs were evaluated by computing the sensitivity (TP/TP+FN) and precision (TP/TP+FP). Several examples of the resulting SSPs are discussed.

SSP Algorithm

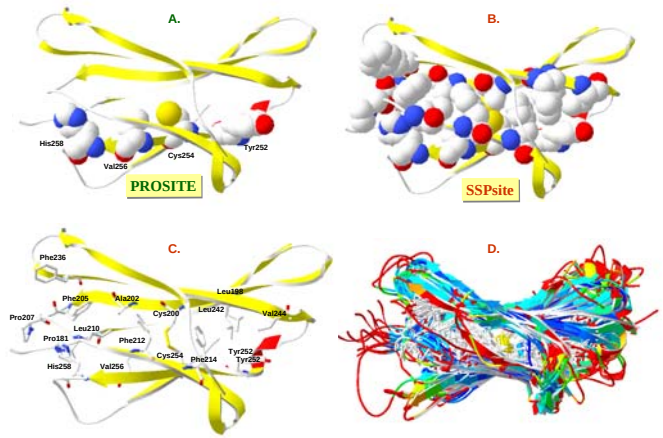
Input: A PROSITE-type sequence pattern, P, of length m.
A Database of protein structures, and associated sequences, N.

Output: One or more SSPs.

a. Find list C of candidate proteins in N that contain sequence pattern P and that align structurally at the pattern residues.
b. Create a sequence alignment and a structure alignment for the list C.
c. Compute a sequence-structure pattern (SSP) consisting of residues in positions that align well in the sequence alignment and in the structure alignment and that satisfy the following criteria:
   i. The majority of the residues at the aligned position are conserved, i.e., they are of the same type (e.g. all Gly), or the majority of the residues at the aligned position belong to a "substitution group" (Wu, Brutlag 1996).
   ii. Every residue interacts with one or more other residues in the pattern and occupy a connected three-dimensional region.
   iii. The residues have similarly oriented side chains.
   iv. The residues in question have a small RMSD value when aligned with a template for this pattern.
   v. The pattern has at least five residues and is present in at least 80% of the candidate proteins C.
d. Evaluate the SSP by computing precision and sensitivity.
e. Improve the SSP by deleting or adding residues in order to increase its precision and sensitivity.
f. If necessary, split the SSP into more than one fragment to improve precision and sensitivity.

Results

Example 1: Immunoglobin/Major Histocompatability Complex Proteins



A. Space-filled model of residues in protein 3frua from PROSITE signature, PS00290; B. Space-filled model of residues in 3frua from SSPsite signature SSP15290 residues; C. Ball & Stick model of residues in 3frua from SSPsite signature SSP15290 residues; D. Structural alignment of 25 proteins with signature SSP15290, colored by RMSD.

Immunoglobin/Major Histocompatability Complex Proteins. The original PROSITE pattern IG_MHC (PS00290) had a number of matches that were not immunoglobins. SSP15290 improves on this pattern.

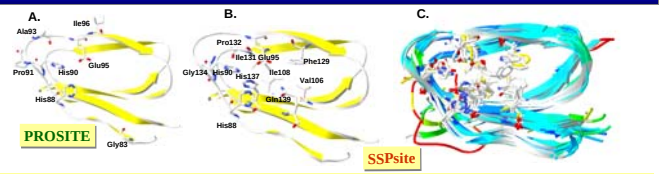
PROSITE PS00290:[FY]-x-C-x-[VA]-x-H.**

- Precision: 98.71% (993/1006) , Sensitivity: 70.78% (993/1403) (SCOP Family: B.1.1.2) *

SSPsite SSP15290: [FPV]-x(9,20)-[FILVY]-x-C-x-[AILMTV]-x(1,2)-[DGFILVY]-x(1,3)-[DPS]-x(1,5)-[AILMV]-x-[FILMTV]-x-[FLWY]-x(19,31)-[AFLWY]-x(5)-[AFGLTV]-x-[FILMSTVY]-x(5,11)-[FHILVY]-x-C-x-[ALMV]-x-[FHNSY].

- Precision: 100% (1319/1319) , Sensitivity 94.01% (1319/1403) (SCOP Family: B.1.1.2) *

Example 2: Germin (Cupin) Family Proteins



A. Ball & Stick model of residues in protein 1f12a from PROSITE signature PS00725; B. Model of residues in 1f12a from SSPsite signature SSP59821; C. Structural alignment of 10 proteins with signature SSP59821, colored by RMSD.

Germin-Like Protein (GLP) Family. The original PROSITE pattern was overly specific and had a high false negative rate.

PROSITE PS00725: G-x(4)-H-x-H-P-x-[AGS]-x-E-[LIVM].**

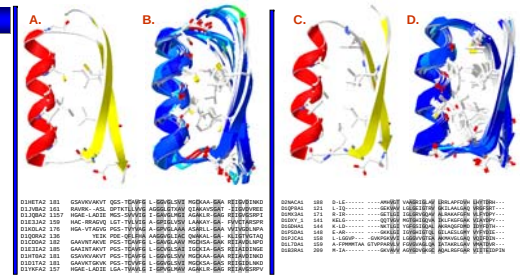
- Precision: 100% (1/1) , Sensitivity: 1.64% (1/61) (SCOP Family: B.82.1.2) *

SSPsite SSP59821: [HQS]-X-[AHNSTY]-X(3,4)-[EFLQV]-X(10)-[FILV]-X-[ILMV]-X(16,39)-[FIQW]-X-[FILV]-[PQ]-X-[AGNS]-X(2)-[FHVWY]-X-[ILMQV].

- Precision: 100% (57/57) , Sensitivity: 93.44% (57/61) (SCOP Family: B.82.1.2) *

* ASTRAL SCOP 1.63 PDB SEQRES records (Current); ** PROSITE Release 18.0 of 12-Jul-2003 (Current).

Protein Family Distinction



A. 1het: SSP09052 residues; C. 2nac: SSP09062 residues;
B. SSP09052 alignment; D. SSP09062 alignment.

One motif (Strand-Helix-Strand), resulting in 2 SSPs.

Alcohol/glucose dehydrogenases: SCOP Family C.2.1.1

SSP09052: [CFLV]-x-[FV]-x-[AG]-x(1,2)-G-[ACGP]-x-G-x(2)-[AGSV]-[ACIV]-x(2)-[AC]-x(3,4)-G-A-x(1,2)-[ILV]-x-[ACGV]-x-[ADGV].

- Precision: 100% (138/138), Sensitivity: 100% (138/138) *

Formate/glycerate dehydrogenases, SCOP Family C.2.1.4

SSP09062: [ALV]-x-[IVT]-x(2)-[AFGLTVY]-G-x(2)-G-x(2)-[AFCLV]-[AGILM]-x(2)-[AFLM]-x(4)-[AFMV]-x-[ILV]-x-[AFGITY]-x-[DSE].

- Precision: 100% (71/71), Sensitivity: 100% (71/71) *

Protein Family Consolidation

A. 1h98 (SCOP D.58.1.2): SSP02198 residues; B. 1vjw (SCOP D.58.1.4): SSP02198 residues; C. SSP02198 alignment.

Ferredoxins: SCOP Families D.58.1.1, D.58.1.2, D.58.1.3, D.58.1.4

SSP02198: [ADEKNQSV]-C-[AEGIKV]-[AENPRS]-[AEILV]-x(4,5)-[IFHLY]-x(18,31)-[AC]-x(3)-C-P.

- Precision: 100% (61/61), Sensitivity: 100% (61/61) *
- SSPs can also be used to group together structurally diverse groups such as the Ferredoxins. Although functionally related proteins may vary in residue composition, they often have specific regions with close residue side chain conformation. This allows the group as a whole to be characterized by an SSP.

Conclusions

Based on our experience with several protein families, our improved SSPs for several PROSITE-style signature patterns :

- Contained more residues covering a greater length of the protein sequence,
- Contained a larger number of variable length gaps,
- Contained higher contact order (CO) patterns , and
- Exhibited higher sensitivity (TP/TP+FN) and precision (TP/TP+FP) .

SSPsite Online: **www.cs.fiu.edu/sspsite**