Discovering Sequence-Structure Patterns in Proteins with Variable Secondary Structure

Tom Milledge, Gaolin Zheng, and Giri Narasimhan

Bioinformatics Research Group (BioRG), School of Computer Science,
Florida International University, Miami, Florida, 33199, USA
{tmille01, gzhen001, giri}@cs.fiu.edu

Abstract. Proteins that share a similar function often exhibit conserved sequence patterns. Sequence patterns help to classify proteins into families where the exact function may or may not be known. Research has shown that these domain signatures often exhibit specific three-dimensional structures. We have previously shown that sequence patterns combined with structural information, in general, have superior discrimination ability than those derived without structural information. However in some cases, divergent backbone configurations and/or variable secondary structure in otherwise well-aligned proteins make identification of conserved regions of sequence and structure problematic. In this paper, we describe improvements in our method of designing biologically meaningful sequence-structure patterns (SSPs) starting from a seed sequence pattern from any of the existing sequence pattern databases. Improved pattern precision is achieved by including conserved residues from coil regions that are not readily apparent from examination of multiple sequence alignments alone. Pattern recall is improved by systematically comparing the structure of all known true family members and to include all the allowable variations in the pattern residues.

1 Introduction

Databases such as PROSITE [1] [2], eMOTIF [3] [4], PRINTS [5], and SPAT [6] have been created as repositories for sequence patterns that describe and distinguish functional and structural domains in proteins. Kasuya et al. [7] systematically investigated the three-dimensional structures of protein fragments whose sequences contain a specific PROSITE pattern. They observed that in a large number of cases, the three-dimensional conformations of the residues from the PROSITE pattern were nearly identical in all the true positives (i.e., proteins belonging to the family and containing the sequence pattern). The main drawback with the approach followed by existing databases to generate sequence patterns is that they base their computations on multiple sequence alignments, which are often inaccurate, especially when the sequences exhibit considerable diversity. We have previously described a method [8] which uses both sequence and structure information to construct patterns consisting of a sequence component (a "PROSITE-style" regular expression pattern) and a structure component (a structure template). In our method, sequence-structure patterns (SSPs) are generated by starting from "seed" sequence patterns from PROSITE, eMOTIF, PRINTS, or other sources, and improving them using a novel method that alternates between sequence and structure alignment of proteins, while using the knowledge of

V.N. Alexandrov et al. (Eds.): ICCS 2006, Part II, LNCS 3992, pp. 702 – 709, 2006.

substitution groups [9]; protein structures are obtained from the Protein Data Bank (PDB) [10]. We say that a protein has a sequence match with the SSP if it contains the sequence component of the SSP. The SSPs are evaluated with regard to their specificity (TP/(TP+FP)) and sensitivity (TP/(TP+FN)), where TP is the set of true positive sequence matches, while FP and FN are the sets of false positive and false negative sequence matches with respect to their membership in a SCOP (Structural Classification of Proteins) protein family. The SCOP database is a comprehensive classification of all proteins of known structure [11]. The basic classification unit in SCOP is the domain, a unit of the protein that is either observed isolated in nature or in more than one context in different multi-domain proteins. A related database is the ASTRAL Compendium, which provides sequences and structures for all domains filtered according to percentage sequence similarity [5]. The ASTRAL 40% database (version 1.65) contains a subset of proteins from the PDB database with less than 40% sequence identity to each other, and this database will be referred to as ASTRAL40 in this paper. We also refer to ASTRAL95 and ASTRAL100 (or full PDB) to refer to the corresponding databases with 95% and 100% sequence identity respectively. For our purposes, the ASTRAL40 database was used to generate SSPs for families that were well represented in the PDB and the ASTRAL95 database was used to generate SSPs for protein families with fewer PDB examples. In both cases, the ASTRAL100 database was used for testing.

2 SSP Method Improvements

Fig. 1 gives a brief description of our algorithm for generating SSPs and is reproduced from our earlier work. It takes as input a "seed" PROSITE-style pattern along with a training set database (in our case, we use ASTRAL40 unless there are not enough structures in it, in which case we use ASTRAL95). It produces as output a SSP, which is a pair <P, T>, where P is a sequence pattern, and T is a structure template for the sequence pattern. As mentioned earlier, it also produces sequence and structure alignments of proteins with this SSP. We first identified the PROSITE patterns with the highest number of hits in the ASTRAL40 database. The algorithm was then experimentally tested on these protein families. In the case where an SSP for a SCOP family already had a corresponding PROSITE sequence pattern, the PROSITE pattern was improved about 90% of the time with respect to the SSP sequence pattern. This represents an average improvement of specificity of +27.3% and an average improvement of sensitivity of +16.2%. Although patterns generated by this method were shown to have higher precision and recall values than comparable patterns in PROSITE, problems were encountered when variations in protein backbone configuration among the true members of a protein family did not allow structurally conserved residues to be detected in the corresponding sequence alignment. In this paper, we describe methods for improving the SSP discovery method in cases where variability in the protein backbone and/or secondary structure obscures instances of residue conservation.

2.1 Zinc Finger Example

The function of the C2H2 (SCOP family G.37.1.1) zinc finger proteins is zinc-dependent DNA or RNA binding, where the first pair of zinc coordinating residues

SSP ALGORITHM

Input: **(a)** A database of protein structures, and associated protein sequences, **N,**
(b) A PROSITE-style sequence pattern , **P.**

Output: **(a)** Sequence-structure pattern **<P', T>**,
(b) Structure alignment **S** of proteins with pattern **P'**, and
(c) Sequence alignment **Q** of proteins with pattern **P'**.

1. Search for pattern **P** in database **N** to generate a list of candidate proteins **C**.
2. Pick a "cluster" **L** of proteins from **C** that belong to the same SCOP family.
3. Create a structure alignment **S** for **L** using the residues of pattern **P**.
4. Extract sequence alignment **Q** from structure alignment **S**.
5. Identify all positions in sequence alignment **Q** that have residues from a substitution group.
6. If stopping condition is not satisfied, then create a new structure alignment **S** for **L** using the positions identified in Step 6. Then go to Step 5.
7. Construct a PROSITE-style sequence-structure pattern **P'** and template T from the positions in **Q**.
8. Iterate the whole process if new candidates from database **N** are matched.

Fig. 1. SSP Algorithm

are cysteines and the second pair are histidines. The PROSITE pattern for this family, PS00028, is **C-x(2,4)-C-x(3)-[LIVMFYWC]-x(8)-H-x(3,5)-H** (Fig. 2a). The pattern generated by the SSP algorithm, SSP91022, is **[AHITFVY]-x(1)-C-x(2,5)-C-x(8,12)-[RIMFYL]-x(2)-H-x(3,5)-H** (Fig. 2b). Only the fully conserved zinc-coordinating cysteine and histidine residues are common to both patterns**.** In the PROSITE pattern, the first histidine is required to follow exactly 13 positions after the second cysteine. In the ASTRAL40, the C2H2 protein 1rmd (shown in Fig. 2c in alignment with 1a1i) has a gap between these two residues of eleven positions and 2gli (Fig. 2d) has a gap of 15. In addition to the C2H2 residues of the PROSITE pattern, the zinc finger SSP output by the SSP algorithm (SSP91022) includes a partially aliphatic position two residues before the first cysteine and another partially aliphatic position three residues before the first histidine. The first of these locations is usually occupied by an aromatic residue (phenylalanine or tyrosine) 70% of the time. The exceptions of alanine, histidine, isoleucine, threonine, and valine appear to form a closed set as no other residues are observed at this position. The second of these locations is occupied by a long-chain aliphatic group consisting of [FILMRY]. Of these residues, leucine is found in 55% of the ASTRAL40 proteins at this position. Valine is the only aliphatic residue not seen here and arginine is the only non-aliphatic residue found to be tolerated at this position.

With the adjustment made for the gap length between the second cysteine and first histidine as described above, the SSP of **[AHITFVY]-x-C-x(2,5)-C-x(8,12)-[RIMFYL]-x(2)-H-x(3,5)-H** generated by the original method was found to match all 102 known C2H2 proteins in the ASTRAL100, SCOP family G.37.1.1, thus providing for 100% recall with respect to the ASTRAL100 database. However, due to this increased variable gap length, the precision of SSP91022 is 71.8% compared to the precision of 80.5% for PROSITE pattern PS00028. In order to increase the precision

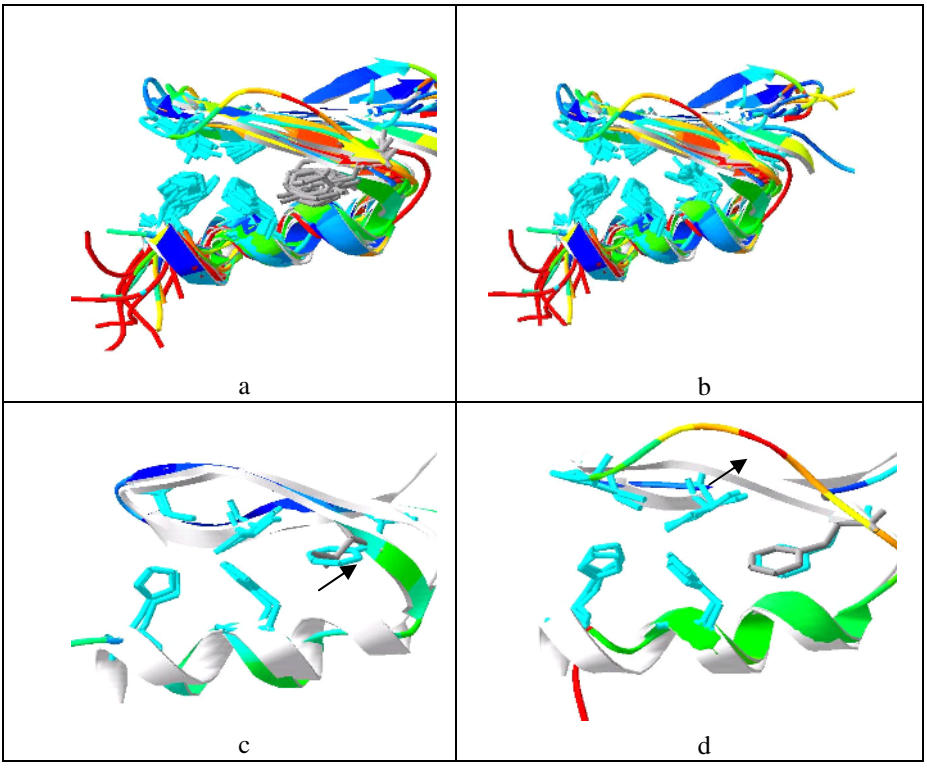

Fig. 2. (a): Multiple alignment of zinc finger domain proteins from ASTRAL40 showing PROSITE pattern PS00028 resides. (b): Multiple alignment showing SSP91022 residues. The tetrahedral C-C-H-H motif is on the left side of the domain with the cysteines above the histidines below. Arrows in the bottom two figures indicate region of (c) shortened backbone region, and (d) lengthened backbone region of the zinc finger domain.

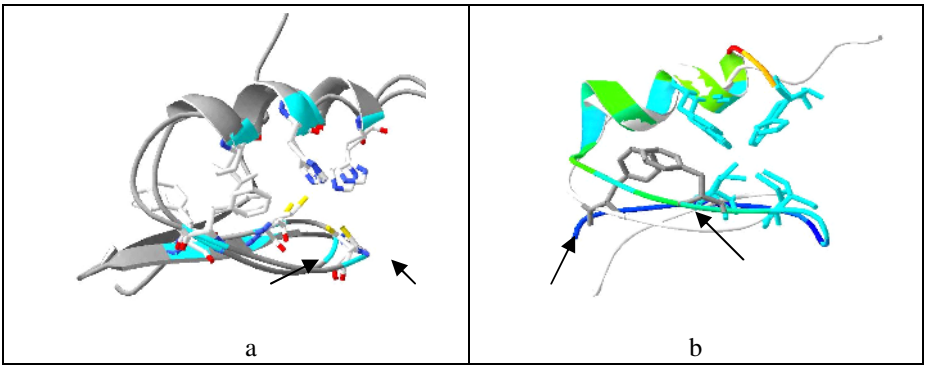

Fig. 3. (a): Alignment of zinc finger proteins 2adr and 1ncs. Arrows indicate regions of backbone variability in zinc finger domain resulting in sequence, but not structure, alignment. (b): Alignment of two zinc finger proteins (2adr_and 1tf6) showing structurally conserved phenyalanine sidechains that are offset three positions in the corresponding sequence alignment.

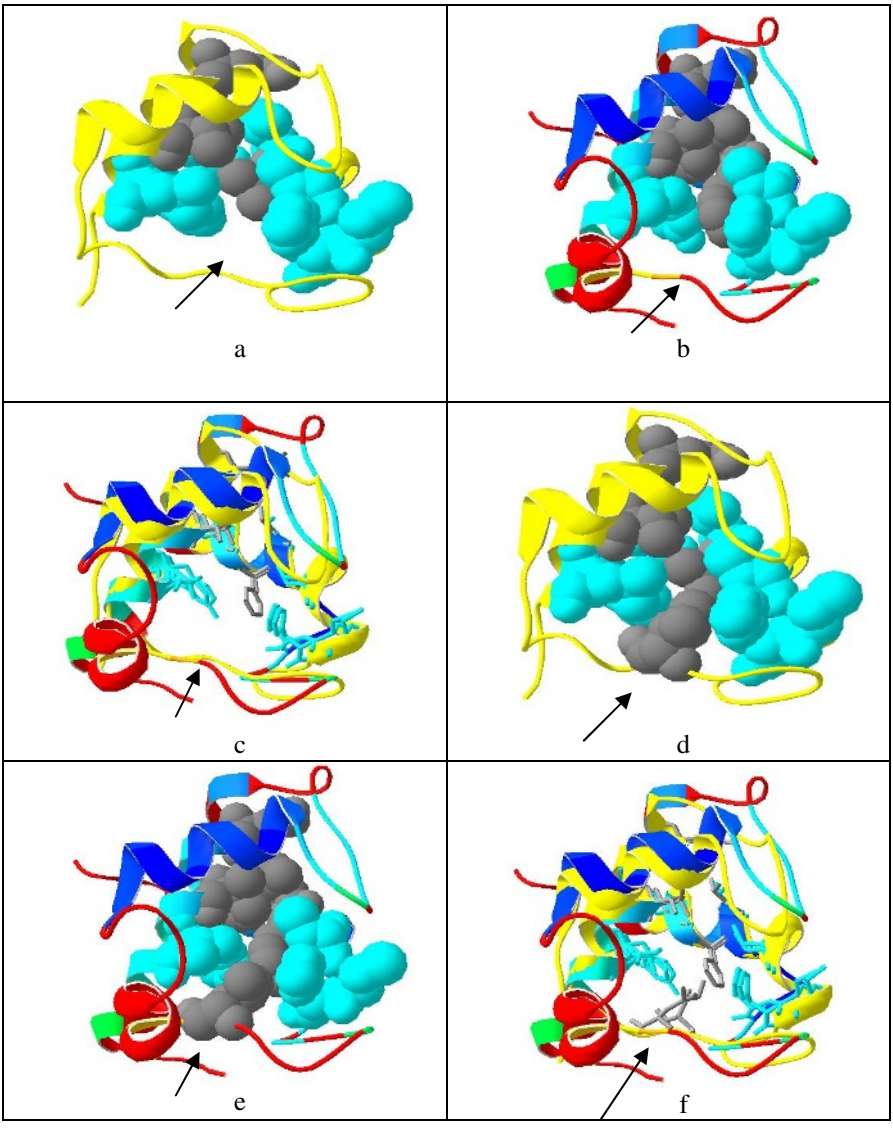

Fig. 4. (a): Cytochrome c protein 1c75 with pattern residue sidechains filled. Arrow shows region of "hole" in the pattern scaffold. (b): Protein 1ctj pattern residues showing hole in the same region. (c): Multiple alignment of 1c75 and 1ctj. The arrow shows variable loop region adjoining the pattern "scaffold". (d): 1c75 showing addition of an isoleucine filling "hole". (e): 1ctj with added leucine. (f): Multiple alignment of 1c75 and 1ctj showing added residues from coil region.

and recall of this SSP, additional positions need to be added to the pattern. Unfortunately, although there are additional conserved residues in the zinc finger domain, variations in the backbone and secondary structure does not allow all of the alpha

carbon atoms of these residues to be structurally aligned. In Fig. 3a, the region between the arrows shows a backbone variation of the domain around the second cysteine residues. Although the sidechains of both cysteines are clearly aligned, the sequence alignment produced from this structure alignment does not have these two cysteines in the same position. In Fig. 3b, the arrows indicate two Phenylalanine residues that are separated by three positions in the sequence alignment, but nevertheless occupy the same position in the structure of the zinc finger domain. Recognition of these offset residues as legitimately conserved elements of the domain allows the creation of the new SSP (SSP95022): **[AIHTYFV]-x(1)-C-x(2,5)-C-x(1,6)-[IYF]-x(5,7)-[RFYMIL]-x(2)-H-x(3,5)-H**. This pattern has a precision of 96.2% and a recall of 98.0% with respect to the ASTRAL100, which represents a significant improvement over the 80.5% precision and 93.1% recall of the PROSITE zinc finger signature.

2.2 Cytochrome C Example

In the case of the cytochrome c family of proteins, the original SSP method identified a number of positions in the domain that showed a high degree of conservation in sequence and structure in addition to those in the short PROSITE pattern. However, the resulting SSP (SSP91008) did not match all known cytochrome c proteins in the ASTRAL100 (see Table 1) and, unlike in the zinc finger example above, the pattern region was bordered by a region of variable secondary structure. In particular, a long and variable length coil between the first and second alpha helices of the domain interposed itself between the pattern residues. In some cases this coil folded upon itself to form a beta strand. Since the backbone of this coil region did not display a high degree of conformation among the cytochrome c proteins in the multiple structural alignment, conserved residues in this region did not, in general, share the same positions in the corresponding multiple sequence alignment.

Table 1. Cytochrome c (SCOP Family A.3.1.1) sequence patterns. Pattern residue position **[ILPVY]** is added to SSP91008 to improve the pattern precision. Other, less conserved, positions are removed to improve recall to 100% with respect to the ASTRAL100.

Accn. num.	Pattern	Precision (TP/TP+FP)	Recall (TP/TP+FN)
PS00190	C-{CPWHF}-{CPWR}-C-H-{CFYW}.	41.7% (196/470)	100% (196/196)
SSP91001	[PSGA]-x(2,3)-[FELKIV]-[MAYFV]-x(2,13)-C-x(2)-C-H-x(41,90)-[KRTEDNGA]-[ESFATLDV]-[MKVILA]-[EKTGNA]-[AFNHYW]-[MTIVL].	94.6% (175/185)	89.3% (175/196)
SSP91008	[FELKIV]-[MAYFV]-x(2,13)-C-x(2)-C-H-x(3,19)-**[ILPVY]**-x(31,77)-[EFTALV]-x(2)-[AKFHYW]-[MTYIVL].	65.3% (196/300)	100% (196/196)

However upon visual inspection of the van der Waals radii of the original SSP, a higher level of correspondence between sequence and structure became apparent. The sidechains of the SSP pattern residues formed a scaffold around the cytochrome C binding between the first and third alpha helices. For each cytochrome c structure, this scaffold formed a contiguous section of the domain tertiary structure with the exception of a "hole" adjoining the coil region (Fig. 4a-c). By identifying the residues whose sidechains filled this space, a new SSP (SSP91001) was created (Fig. 4d-f). Like the original PROSITE pattern, the recall is complete with respect to the ASTRAL100. And like the original SSP, the precision of SSP91001 was significantly improved over the shorter PROSITE pattern (see Table 1).

3 Discussion and Conclusion

As we described previously, the addition of information on conserved residue types gained from multiple structure alignment greatly improves the accuracy of multiple sequence alignment and that new information derived from the improved sequence alignment can likewise improve structure alignment such that both sequence and structure alignments can be improved simultaneously in an iterated manner. Although the precision and recall of sequence patterns derived from the SSP method are usually quite good for protein domains with a well-ordered secondary structure, pattern discovery is more challenging for proteins with variable loop regions where residue type may be well conserved, but nevertheless the "position" of the residue is not well defined. In some cases we have discovered ways to extract this "hidden regularity", as we describe in the examples above. The key concept is determining when a residue type within a structure is playing the same role in all members of the protein family. Typically the protein backbone will be offset among proteins in the structure alignment, however the residue sidechain at the offset position will be rotated such that all family members have a residue of that type in the same location in the structure. Occasionally a conserved residue type will exist at one of several nearby positions in the alignment where the sidechain will also demonstrate this form of "structural compensation". Using the methods we described above, the identification of these variable positions and incorporating the residue types in the protein family signatures can greatly increase their performance.

Acknowledgement

This research was supported in part by NIH grant P01 DA15027-01.

References

1. Falquet, L., M. Pagni, P. Bucher, N. Hulo, C.J.A. Sigrist, K. Hofmann, and A. Bairoch, *The PROSITE database, its status in 2002.* Nucl. Acids. Res., 2002. **30**(1): p. 235-238.
2. Sigrist, C.J., L. Cerutti, N. Hulo, A. Gattiker, L. Falquet, M. Pagni, A. Bairoch, and P. Bucher, *PROSITE: a documented database using patterns and profiles as motif descriptors.* Brief Bioinform, 2002. **3**(3): p. 265-74.

3. Nevill-Manning, C.G., T.D. Wu, and D.L. Brutlag, *Highly specific protein sequence motifs for genome analysis.* Proc Natl Acad Sci U S A, 1998. **95**(11): p. 5865-71.
4. Huang, J.Y. and D.L. Brutlag, *The EMOTIF database.* Nucl. Acids. Res., 2001. **29**(1): p. 202-204.
5. Attwood, T.K., *The PRINTS database: a resource for identification of protein families.* Brief Bioinform, 2002. **3**(3): p. 252-63.
6. Hart, R., A. Royyuru, G. Stolovitzky, and A. Califano, *Systematic and Fully Automatic Identification of Protein Sequence Patterns.* J. Comput. Biol., 2000. **7**((3/4)): p. 585-600.
7. Kasuya, A. and J.M. Thornton, *Three-dimensional structure analysis of PROSITE patterns1.* Journal of Molecular Biology, 1999. **286**(5): p. 1673-1691.
8. Milledge, T., S. Khuri, X. Wei, C. Yang, G. Zheng, and G. Narasimhan, *Sequence-Structure Patterns: Discovery and Applications.* 6th Atlantic Symposium on Computational Biology and Genome Informatics (CBG), 2005: p. 1282-1285.
9. Wu, T.D. and D.L. Brutlag. *Discovering Empirically Conserved Amino Acid Substitution Groups in Databases of Protein Families*. in *ISMB-96*. 1996.
10. Berman, H.M., J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne, *The Protein Data Bank.* Nucleic Acids Res, 2000. **28**(1): p. 235-42.
11. Brenner, S.E., C. Chothia, T.J.P. Hubbard, and A.G. Murzin, *Understanding protein structure: Using SCOP for fold interpretation.* Methods in Enzymology, 1996: p. 635-643.