VOL.4 NO.2 September 2015 - ISSN 2233 – 1859

Operations Research Society in Bosnia and Herzegovina

Southeast Europe Journal of Soft Computing

Available online: www.scjournal.com.ba

IUS Soft Computing Research Group

Secondary Structure Segments are Much More Conserved than Primary Sequence Segments

Faruk B. Akcesme, Mehmet Can

International University of Sarajevo, Faculty of Engineering and Natural Sciences, HrasnickaCesta 15, Ilidža 71210 Sarajevo, Bosnia and Herzegovina

fakcesme@ius.edu.ba; mcan@ius.edu.ba

Article Info

Article history:

Received on 17 Jul. 2015

Revised version received on 17 Aug. 2015

Keywords:

Key words: protein amino acid sequence; secondary structure; FIND-SIM;

Abstract

To be biologically functional, all proteins must adopt specific folded three-dimensional structures. Some believes in that the genetic information for the protein specifies only the primary structure, the linear sequence of amino acids in the polypeptide backbone, and most purified proteins can spontaneously refold in vitro after being completely unfolded, so the three-dimensional structure must be determined by the primary structure (Creighton, 1990). How this occurs has come to be known as 'the protein folding problem'. As a part of the protein folding problem, the existence of similar substrings in diverse proteins is remarkable. Some scientist call it “conserved core” which echoes the claim that all proteins diversified from a common ancestor protein, and these similar pieces of the two or several proteins are the substrings that resisted the pressure of the evolution. Due to naturally-occurring (DNA fails to copy accurately) and external influences just like ultraviolet radiation, electromagnetic fields, atomic radiations, protein coding genes and proteins may undergo some changes by the time in response to mutations. The rate of these mutations is strongly correlated to the intensity of the environmental conditions, and it is not possible to estimate a constant rate just in the case of radioactive decay. Also there is no much evidence that the diversity of proteins relies on only these mutations. For this reason we prefer the term "similar substrings". In this paper we focused in the relation between primary and secondary structure mismatches of the substrings of length seventeen residues. We have seen that the mismatches in the corresponding secondary structure sequence substrings of the same length lags behind primary mismatches. We constructed a conditional probability landscape that resembles the conditional probability of a certain secondary substring mismatch given the primary substring mismatch. This landscape shows that even when 6-7 mismatches exist in two primary substrings of length 17 that belong to the two different proteins, the probability of full match of corresponding secondary structure substrings is remarkable. We downloaded primary and secondary sequences of all 303,524 proteins of the PDB protein databank. Eliminating the duplicates and proteins of residue length less than 30, we have got a non redundant database of 80,592. We developed a search algorithm FIND-SIM to find similar primary sequence substrings in a query protein and target proteins. Some examples of full secondary structure matches of short substrings corresponding to short primary structure substrings with high mismatches are given.

1. INTRODUCTION

Time dependent changes of protein domain primary structures that become fixed in populations are mainly replacements of single amino acid residues and short insertions or deletions. Since most secondary and tertiary structures of proteins are partially determined by their amino acid sequences (Anfinsen, 1973), secondary and higher-order structure will also change along these changes Chan, and Dill,1990). The extent of higher-order structural change due to changes in the amino acid sequence depends on the type and location of these changes (Illerga°rd, et. al, 2009, Haspel, 2003). While some single changes completely disrupting higher order structure, the others that conserve the physicochemical properties of the protein may slightly affect the structure (Matthews, 1995). The structures of homologous proteins are generally better conserved than their sequences. This phenomenon is demonstrated by the prevalence of structurally conserved regions (SCRs) even in highly divergent protein families (Huang, et. a., 2012). However, the recent explosion of sequence and structure information accompanied by the development of powerful computational methods led to the accumulation of examples of homologous proteins with globally distinct structures (Grishin 2001).

between amino acid sequence and secondary structure develops is the topic of this study.

We test one of the claims ‘‘secondary and tertiary structures are more conserved than amino acid sequence’’ which is often used in discussions about proteins. This claim is supported by our observations of conditional probability landscape in Table 1 and in Figure 1.

In Table 1, probabilities in rows are conditional probabilities of mismatches of a query secondary substring of length 17, with and secondary substring of same length cut from the proteins from the database given the mismatches of corresponding secondary structure substrings in the first column. For example second column of the table gives the probabilities of secondary substring full matches given primary amino acid substring mismatches in the first column. The highest probabilities of 37% of the secondary structure full matches appear at the row of 3 mismatches of the primary substring.

Table 1 Probabilities in rows are conditional probabilities of mismatches of a query secondary substring of length 17, with and secondary substring of same length cut from the proteins from the database given the mismatches of corresponding secondary structure substrings in the first column.

mm	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
0	0.19	0.21	0.23	0.11	0.14	0.08	0.02	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
1	0.25	0.26	0.21	0.1	0.11	0.05	0.01	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
2	0.26	0.26	0.14	0.14	0.11	0.06	0.02	0.01	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
3	0.37	0.25	0.11	0.11	0.1	0.03	0.02	0.01	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
4	0.2	0.2	0.19	0.13	0.18	0.04	0.04	0.01	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
5	0.17	0.21	0.21	0.12	0.19	0.07	0.03	0.01	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
6	0.15	0.13	0.23	0.1	0.18	0.19	0.02	0.01	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
7	0.15	0.14	0.15	0.23	0.08	0.06	0.12	0.02	0.01	0.	0.01	0.01	0.	0.	0.	0.	0.	0.
8	0.16	0.12	0.08	0.11	0.11	0.09	0.04	0.05	0.04	0.03	0.04	0.03	0.02	0.03	0.02	0.01	0.01	0.01
9	0.05	0.05	0.04	0.04	0.05	0.06	0.06	0.06	0.07	0.06	0.07	0.07	0.08	0.07	0.05	0.04	0.03	0.04
10	0.01	0.02	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.08	0.09	0.08	0.08	0.08	0.07	0.06	0.04	0.03

BLAST is the most widely used sequence similarity search machine, and most reliable, strategy for characterizing newly determined sequences. Sequence similarity searches with BLAST can identify”homologous” proteins or genes by detecting excess similarity (Pearson 2013, Madej, 2007). This study goes beyond the global homology, and tries to find local similarities of small substrings to be able to use it in secondary structure prediction of new query proteins that does not have close homology in the PDB databank.

A large body of experimental and theoretical evidence suggests that local structural determinants are frequently encoded in short segments of protein sequence(Menke, 2009). Although the local structural information, once recognized,
is particularly useful in protein structural and functional analyses, it remains a difficult problem to identify embedded local structural codes based solely on sequence information (Yang, and Wang, 2003; Chothial, and Lesk, 1986). This paper spread some light to the question how local secondary structure evolves along with the time dependent primary sequence changes and how the relation

In Figure 1 High mountain represents the lag of secondary structure mismatches after primary sequence mismatches. The starting point of introducing an algorithm to predict secondary structure of query proteins using the local similarities of 17 residue short subsequences is this observation.

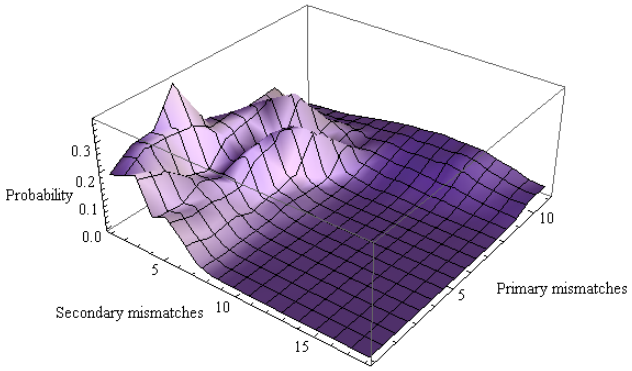

Figure 1. High Mountain represents the lag of secondary structure mismatches after primary sequence mismatches.

1. SIMILAR SUBSTRING FINDING ALGORITHM FIND-SIM

As of December 30, 2003, 23,000 solved protein structures had been deposited in the Brookhaven Protein Data Bank (PDB) (Berman et., al., 200). This number kept increasing, with 300 new entries added each month, and reached today the number of three hundred thousand. The size and completeness of the PDB is essential to the success of template-based approaches to protein structure prediction (Zhang, and Skolnick, 2005).

Observations in the above led us to introduce and algorithm that would find similar substrings to a substring.

We developed a search algorithm FIND-SIM to find similar primary sequence substrings from the database to a substring of the query protein (Soss, 2016, Sander, et., al., 2006). First we find an optimal window size. 17 is short enough to fit in local similarity regions, big enough to capture secondary structure objects α-helix regions and β-sheets. A code written in MATHEMATICA computer algebra package helped us visiting all 20,660,981 possible substrings of length 17 in 80,592 proteins. During this visit, addresses of amino acid substrings which make 0-10 mismatches with the query primary substring are put in eleven different baskets labeled by the number of the mismatches Guex, N. Each address consists of the number of the target protein, and the position of the first residue of the fitted substring. Then these addresses are used to visit the corresponding secondary sequence substrings, and mismatches of primary substrings an secondary structure substrings are compared. Frequencies of primary/secondary substring mismatches are stored in a matrix. The number in the *ij*position of this matrix is the number of cases with *i* primary mismatches and *j* secondary mismatches. This matrix is normalized such that all 11 rows have the same total of one. Table 1 depicts this matrix.

The most interesting part of this matrix is the first column (second column in the table), where full matches of secondary sequence substrings are observed at any number of primary sequence substring mismatches. In the next section we will supply some examples.

2. RESULTS

Full secondary structure matches vs. high amino acid mismatches

In the sequel some examples of full secondary structure matches of short substrings corresponding to short primary structure substrings with high mismatches are given. Each primary and secondary structure is represented and the similar segments are labeled. 3D images are generated by using Swiss-PdbViewer (Guex, N. and Peitsch, M.C. (1997)

a) The substring of length 17 of the protein 2SCP_B starting at the residue 7,

KMKTYFNRIDFDKDGAI

has six mismatches with the substring

KAITCFNTLDFNKNGQI

of protein *2HQ8_A*starting at the residue 139.

PDB code: 2SCP_B

SDLWVQKMKTYFNRIDFDKDGAITRMDFESMAERF
AKESEMKAEHAKVLMDSLTGVWDNFLTAVAGGKG
IDETTFINSMKEMVKNPEAKSVVEGPLPLFFRAVDT
NEDNNISRDEYGIFFGMLGLDKTMAPASFDAIDTNN
DGLLSLEEFVIAGSDFFMNDGDSTNKVFWGPLV

PDB code: 2HQ8_A

MPEITESERAYHLRKMKTRMQRVDVTGDGFISRED
YELIAVRIAKIAKLSAEKAEETRQEFLRVADQLGLAP
GVRISVEEAAVNATDSLLKMKGEEKAMAVIQSLIM
YDCIDTDKDGYVSLPEFKAFLQAVGPDLTDDKAITC
FNTLDFNKNGQISRDEFLVTVNDFLFGLEETALANA
FYGDLVD

While the corresponding subsequences of the secondary structures have a full match:

HHHHHHHHHCCCCCCEE
HHHHHHHHHCCCCCCEE

PDB code: 2SCP_B

CHHHHHHHHHHHHHHCCCCCCEECHHHHHHHHHH
HHHHHCCCCHHHHHHHHHHHHHHHHHHCHHHCCC
CCCEEHHHHHHHHHHHHCCCHHHCHHHHCCHHHH
HHHHCCCCCCEE**C**HHHHHHHHHHHCCCCCCCHHHH
HHHHCCCCCCCEEHHHHHHHHHHHHHHCCCCCHHH
HCCCCCC

PDBcode: 2HQ8_A

CCCCCHHHHHHHHHHHHHHHHCCCCCCCCEEEHH
HHHHHHHHHHHHHHCCCHHHHHHHHHHHHHHHHH
HCCCCCCEEEHHHHHHHHHHHHHHHCCCHHHHHCC
HHHHHHHHHHCCCCCCEEEHHHHHHHHHHHCCCC
CHHHHHHHHHHCCCCCCEEEHHHHHHHHHHHH
HCCCCCHHHHHHCCCCCC
HHHHHHHHHCCCCCCEEE

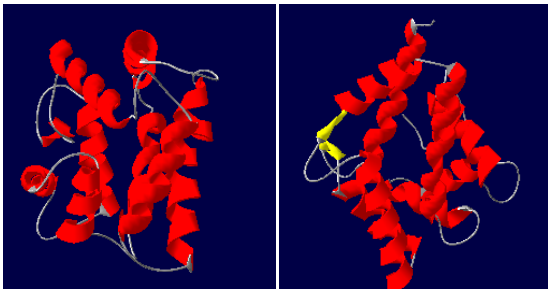

Image 1: 2SCP_B and *2HQ8_A*

b) The substring of length 20 of the protein 2SCP_B starting at the residue 8,

MKTYFNRIDFDKDGAITRMD

a) has 11 mismatches with the substring

AEAVFQRLDADRDGAITFQE

of protein *2PMY_A*starting at the residue 63.

PDB code: 2SCP_B

SDLWVQKMKTYFNRIDFDKDGAITRMDFESMAERF
AKESEMKAEHAKVLMDSLTGVWDNFLTAVAGGKG
IDETTFINSMKEMVKNPEAKSVVEGPLPLFFRAVDT
NEDNNISRDEYGIFFGMLGLDKTMAPASFDAIDTNN

DGLLSLEEFVIAGSDFFMNDGDSTNKVFWGPLV
PDB code: 2PMY_A
MHHHHHHSSGRENLYFQGADGDGEELARLRSVFA
ACDANRSGRLEREEFRALCTELRVRPADAEAVFQRL
DADRDGAITFQEFARGFLGSL

While the corresponding subsequences of length 20 of the secondary structures are in fullmatch:

HHHHHHHHCCCCCCEECHHH
HHHHHHHHCCCCCCEECHHH

PDB code: 2SCP_B
CHHHHHHHHHHHHHHCCCCCCEECHHHHHHHHH
HHHHHCCCCHHHHHHHHHHHHHHHHHHCHHHCCC
CCCEEHHHHHHHHHHHHCCCHHHCHHHHCCHHHH
HHHHCCCCCCEE**C**HHHHHHHHHHCCCCCCCHHHH
HHHHCCCCCCCEEHHHHHHHHHHHHHHCCCCCHHH
HCCCCCC
PDB code: 2PMY_A
CCCCCCCCCCCCCCCCCCCHHHHHHHHHHHHHHH
HHCCCCCCCEEHHHHHHHHHHCCCCHHHHHHHH
HHCCCCCCEECHHHHCHHHHHCC

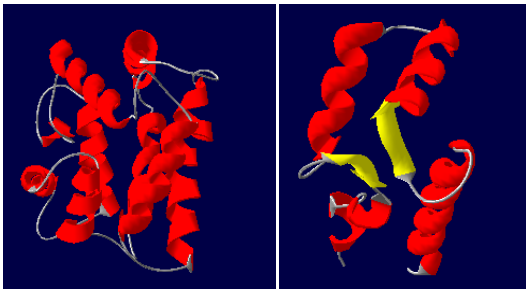

Image 2: 2SCP_B and*2PMY_A*

b) The substring of length 17 of the protein *2SCP_B* starting from the residue 96

LPLFFRAVDTNEDNNIS
has eight mismatches with the substring

LRDAFREFDTNGDGEIS

of protein *2K7C_A* starting from the residue 9

PDB code: 2SCP_B
SDLWVQKMKTYFNRIDFDKDGAITRMDFESMAERF
AKESEMKAEHAKVLMDSLTGVWDNFLTAVAGGKG
IDETTFINSMKEMVKNPEAKSVVEGPLPLFFRAVDT
NEDNNISRDEYGIFFGMLGLDKTMAPASFDAIDTNN
DGLLSLEEFVIAGSDFFMNDGDSTNKVFWGPLV

PDB code:2K7C_A
ADMIGVKELRDAFREFDTNGDGEISTSELREAMRKL
LGHQVGHRDIEEIIRDVDLNGDGRVDFEEFVRMMSR

While the corresponding subsequences of the secondary structures are in fullmatch:

HHHHHHHHCCCCCCEE**C**
HHHHHHHHCCCCCCEE**C**

PDB code: 2SCP_B
CHHHHHHHHHHHHHHCCCCCCEECHHHHHHHHH
HHHHHCCCCHHHHHHHHHHHHHHHHHHCHHHCCC
CCCEEHHHHHHHHHHHHCCCHHHCHHHHCCHHHH
HHHHCCCCCCEE**C**HHHHHHHHHHCCCCCCCHHHH
HHHHCCCCCCCEEHHHHHHHHHHHHHHCCCCCHHH
HCCCCCC
PDB code: 2K7C_A
CCCCCHHHHHHHHHHCCCCCCEECHHHHHHHHH
HCCCCCCCHHHHHHHHHHCCCCCCCCEEHHHHHH
HHH

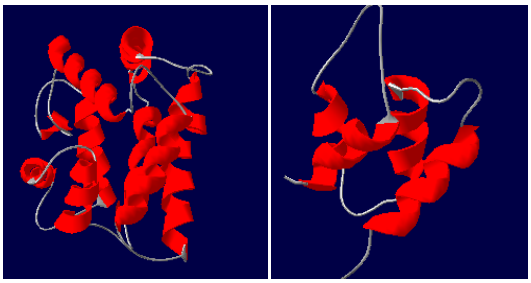

Image 3: 2SCP_B and *2K7C_A*

c) The substring of length 17 of the protein *2SCP_B* starting from the residue 96

LPLFFRAVDTNEDNNIS
has eight mismatches with the substring

LRDAFREFDTNGDGEIS

of protein *3OX6_A* starting from the residue 96

PDB code: 2SCP_B
SDLWVQKMKTYFNRIDFDKDGAITRMDFESMAERF
AKESEMKAEHAKVLMDSLTGVWDNFLTAVAGGKG
IDETTFINSMKEMVKNPEAKSVVEGPLPLFFRAVDT
NEDNNISRDEYGIFFGMLGLDKTMAPASFDAIDTNN
DGLLSLEEFVIAGSDFFMNDGDSTNKVFWGPLV

PDB code: 3OX6_A
MDRSLRPEEIEELREAFREFDKDKDGYINCRDLGNC
MRTMGYMPTEMELIELSQQINMNLGGHVDFDDFVE
LMGPKLLAETADMIGVKELRDAFREFDTNGDGEIST
SELREAMRALLGHQVGHRDIEEIIRDVDLNGDGRVD
FEEFVRMMSR

While the corresponding subsequences of the secondary structures are in fullmatch:

HHHHHHHHCCCCCCEEC
HHHHHHHHCCCCCCEEC

PDB code: 2SCP_B
CHHHHHHHHHHHHHHCCCCCCEECHHHHHHHHH
HHHHHCCCCHHHHHHHHHHHHHHHHHHCHHHCCC
CCCEEHHHHHHHHHHHHCCCHHHCHHHHCCHHHH
HHHHCCCCCCEE**C**HHHHHHHHHHCCCCCCCHHHH
HHHHCCCCCCCEEHHHHHHHHHHHHHHCCCCCHHH
HCCCCCC

PDB code: 3OX6_A
CCCCCCHHHHHHHHHHHHHHHCCCCCCCCHHHH
HHHHHHCCCCCCHHHHHHHHHHHHHCCCCCCCCH
HHHHHHHHHHHHHCCCHHHHCHHHHHHHHHHCC
CCCCEECHHHHHHHHHHHHHCCCCCHHHHHHHHH
HHCCCCCCCEEHHHHHHHCCC

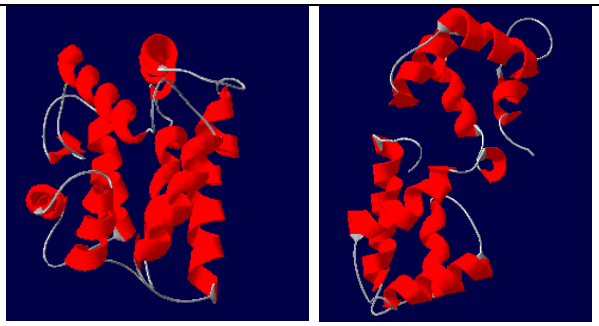

Image 4: 2SCP_B and 3OX6_A

d) The substring of length 17 of the protein *2SCP_B* starting from the residue 99

FFRAVDTNEDNNISRDE
has eight mismatches with the substring
AFREFDTNGDGEISTSE
of protein *2K7C_A*from the residue 12.

PDB code: 2SCP_B
SDLWVQKMKTYFNRIDFDKDGAITRMDFESMAERF
AKESEMKAEHAKVLMDSLTGVWDNFLTAVAGGKG
IDETTFINSMKEMVKNPEAKSVVEGPLPLFFRAVDT
NEDNNISRDEYGIFFGMLGLDKTMAPASFDAIDTNN
DGLLSLEEFVIAGSDFFMNDGDSTNKVFWGPLV

PDB code:2K7C_A
ADMIGVKELRDAFREFDTNGDGEISTSELREAMRKL
LGHQVGHRDIEEIIRDVDLNGDGRVDFEEFVRMMSR

While the corresponding subsequences of the secondary structures are in full match:

HHHHHCCCCCCEE**C**HHH

HHHHHCCCCCCEE**C**HHH

PDB code: 2SCP_B
CHHHHHHHHHHHHHHCCCCCCEECHHHHHHHHHH
HHHHHCCCCHHHHHHHHHHHHHHHHHHCHHHCCC
CCCEEHHHHHHHHHHHHCCCHHHCHHHHCCHHHH
HHHHCCCCCCEE**C**HHHHHHHHHHCCCCCCHHHH
HHHHCCCCCCCEEHHHHHHHHHHHHHHCCCCCHHH
HCCCCCC

PDB code:2K7C_A
CCCCCHHHHHHHHHHHHCCCCCCEECHHHHHHHHH
HCCCCCCCHHHHHHHHHHHCCCCCCCCEEHHHHHH
HHHC

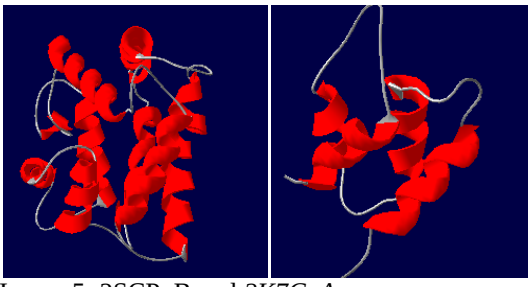

Image 5: 2SCP_B and *2K7C_A*

3. DISCUSSION

Conserved segments of protein sequences are important to assign function to proteins. Appropriate methods for computing them need to be developed in order to get benefits from conserved segments. Conserved segments are considered as one of the sign which shows how natural protein structures respond to evolutionary changes. In this study we aimed to find the similar segments of protein in two protein`s primary structures. We observed more similarity in their secondary structure in the same substrings. One of the reasons of higher similarity in secondary structure might be that the secondary structures of proteins are giving slower response to mutations than primary sequences. In this way proteins avoid losing certain function.

One of the other main approaches which make our study unique among the similar studies is that rather than using multiple sequence alignment for detecting conserved cores we use 17 long substrings to find similar non-consecutive matches in two sequences. We also observe that similar substrings found by our algorithm are classified in the same superfamily by SCOP. Superfamily is design to find proteins which have structural evidence to support a common evolutionary ancestor but may not have detectable sequence homology.

Furthermore, we believe that using similar substrings whose secondary structure similarity is higher, can help to predict secondary structure of proteins with higher accuracy.

4. REFERENCES

Anfinsen, C. (1973) Principles that govern the folding of protein chains. Science,181:223–230.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000) The Protein Data Bank, Nucleic Acids Res. 28, 235–242.

Chan, H.S., and Dill, K.A. (1990) Origins of structure in globular proteins, Proc. Nati. Acad. Sci. USA Vol. 87, pp. 6388-6392, August 1990 Chemistry

Chothial, C., and Lesk, A.M. (1986) The relation between the divergence of sequence and structure in proteins, The EMBO Journal vol.5 no.4 pp.823-826.

Creighton, T.E. (1990) Protein folding, Biochem. J. 270, 1-16

Grishin, N.V. (2001) Fold Change in Evolution of Protein Structures, Journal of Structural Biology 134, 167–185

Guex, N. and Peitsch, M.C. (1997) SWISS-MODEL **and the Swiss-PdbViewer: An environment for comparative protein modeling.** *Electrophoresis* **18**, 2714-2723.

Haspel, N. Tsai, C.-J., Wolfson, H., And Nussinov, R. (2003) Reducing the computational complexity of protein folding via fragment folding and assembly, Protein Science, 12:1177–1187.

Huang, I.K., Pei, J., and Grishin, N.V. (2012) Defining and predicting structurally conserved regions in protein superfamilies, Bioinformatics Advance Access published November 28, 2012.

Illerga°rd, K., David H. Ardell, D.H., and Elofsson, A. (2009) Structure is three to ten times more conserved than sequence—A study of structural response in protein cores, Proteins; 77:499–508.

Madej, T., Panchenko, A.R., Chen, J., and Bryant, S.H. (2007) Protein homologous cores and loops: important clues to evolutionary relationships between structurally similar proteins, BMC Structural Biology, 7:23 doi:10.1186/1472-6807-7-23

Matthews, B. (1995) Studies on protein stability with T4 lysozyme. Adv Protein Chem 1995;46:249–278.

Menke, M.E. (2009) Computational Approaches to Modeling the Conserved Structural Core Among Distantly Homologous Proteins, PhD Thesis, Massachusetts Institute of Technology.

Pearson, W.R. (2013) An Introduction to Sequence Similarity ("Homology") Searching, CurrProtoc Bioinformatics. June ; 0 3: . doi:10.1002/0471250953.bi0301s42.

Sander, O., Sommer, I., and Lengauer, T. (2006) Local protein structure prediction using discriminative models, BMC Bioinformatics 2006, 7:14 doi:10.1186/1471-2105-7-14.

Soss, M. (2016) Analysis of the structurally conserved core of a set of proteins, Chemical Computing Group.

Yan, R., Xu, D., Yang, J., Walker, S., and Zhang, Y. (2013) A comparative assessment and analysis of 20 representative sequence alignment methods for protein structure prediction, Scientific Reports, 3 : 2619, DOI: 10.1038/srep02619.

Yang, An-S., and Wang, Lu-Y. (2003) Local structure prediction with local structure-based sequence profiles, Bioinformatics Vol. 19 no. 10, pages 1267–1274.

Zhang, Y., and Skolnick, J. (, 2005) The protein structure prediction problem could be solved using the current PDB library, PNAS, January 25, 2005, vol. 102, no. 4, 1029–1034