VOL 6, NO. 1 March 2017 - ISSN 2233 – 1859

Operations Research Society in Bosnia and Herzegovina

Southeast Europe Journal of Soft Computing

Available online: http://scjournal.ius.edu.ba

IUS Soft Computing Research Group

On the Accuracy of Sequence Similarity Based Protein 3D Prediction

Muhamed Adilovic, Faruk B. Akcesme, Mehmet Can

International University of Sarajevo,
Faculty of Engineering and Natural Sciences,
Hrasnicka Cesta 15, Ilidža 71210 Sarajevo,
Bosnia and Herzegovina
madilovic@ius.edu.ba; fakcesme@ius.edu.ba; mcan@ius.edu.ba

Article Info

Article history:
Article received on 19 Jan. 2017
Received in revised form24 Feb.2017

Keywords:
Protein Tertiary Structure Prediction; PDB, Homology

Abstract

In an article (Akcesme, and Can 2015), authors examined the relation between primary and secondary structure mismatches of the substrings of length seventeen residues from two different proteins. They have shown that the mismatches in the corresponding secondary structure sequence substrings of the same length mostly lag behind primary mismatches. In the PhD dissertation thesis (Akcesme 2016) author examined the possibility of secondary structure prediction by the use of smaller conserved segments and created a software AVISENNA that outperforms PSIPRED and all other available secondary structure prediction tools. In another article (Akcesme, et. al. 2017), the issue of how far secondary structure of proteins can be predicted based on hosts (larger proteins that contain the query protein as a subchain) of this protein in the set of solved structures currently deposited in PDB. It is seen that around 17% of proteins have hosts in PDB, and secondary structures of them can be predicted with a mean accuracy of 90.39 %. This accuracy of the host based secondary structure prediction set also an upper bound for the homology based tertiary structure predictions. In this article the impact of the mentioned inaccuracy on the homology based 3D structure predictions by the three predictors I-Tasser, Phyre2, and SwissModel are studied. Inaccuracies in predicted tertiary structures are seen in the visual comparison of the 3D structures of query proteins and their predicted 3D images by the three 3D predictors, and their counterparts in host proteins.

1. INTRODUCTION

All proteins adopt a specific folded three-dimensional structure and hence become biologically functional. It is widely believed in that the genetic information for the protein specifies only linear sequence of amino acids in the protein backbone, and many proteins in their native state can refold in vitro after being completely unfolded, so the three-dimensional structure must be determined by the primary structure (Anfinsen, 1973; Creighton, 1990; Chan, and Dill, 1990). But of course there are limitations of the level of unfolding, for example the process of cooking an egg can never be reversed. The protein folding problem' is the problem of how all proteins adopt a specific folded three-dimensional structure. As a part of this 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 (Chothial, and Lesk 1986, Huang, et., al., 2012, Illergard, et., al., 2009, Madej, et., al., 2007, Menke, 2009). Some others may also view it as the economy in creation. Due to failure of the DNA in proper replication, and external influences just like electromagnetic fields, ultraviolet, and 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 in chemistry. Also there is no much evidence that the diversity of proteins relies on only these mutations in amino acid sequences.

It is seen that around 17% of proteins in the PDB dataset have identical subsequences in other larger proteins of PDB dataset. When the secondary structures of proteins are assigned as the corresponding secondary structures of identical parts in other larger proteins, the average prediction accuracy is found to be 90.39 %.

In an article (Akcesme, et. al. 2017), the issue of how far secondary structure of proteins can be predicted based on homlogs of this protein in the set of solved structures currently deposited in Protein Data Bank (PDB) is studied (Berman, et. al. 2000, Zhang, and Skolnick, 2005). It is reported that around 17% of proteins among around 80 thousand proteins of PDB have homolog domains in PDB, and when secondary structures of these proteins predicted as the corresponding parts of secondary structures of hosts, the mean accuracy is be 90.39 %. The percentage of accuracies less than 50% is only 1%. This accuracy 90.39 % of the super chain (host) based secondary structure prediction, set also an upper bound for the homology based tertiary structure predictions. In this article the impact of the mentioned inaccuracy on the homology based 3D structure predictions are shown by the use of predictions by three predictors I-Tasser, Phyre2, and SwissModel.

2. LOW SECONDARY STRUCTURE SIMILARITIES OF IDENTICAL AMINO ACID SEQUENCES IMPOSE LOW 3D STRUCTURE SIMILARITIES

In the sequel some examples of pairs of proteins are given such that the amino acid sequence of the first protein (query) is a segment in the amino acid sequence of the second protein (host). They are chosen such that the secondary structure sequence of the query has high mismatches with the secondary structure sequence of the corresponding part of the host. For sampled pairs, primary and are presented and the identical segments are highlighted with red color. 3D images are generated by using Swiss-PdbViewer (Guex and Peitsch, 1997).

Secondary and tertiary structures of five query/ host pairs are as follows:

Pair 1; Query Length 86

QUERY; 1BNP

SKRKAPQETLNGGITDMLVELANFEKNVSQAIHKYN
AYRKAASVIAKYPHKIKSGAEAKKLPGVGTKIAEKID
EFLATGKLRKLEK

HOST; 1BPE

MSKRKAPQETLNGGITDMLVELANFEKNVSQAIHKY
NAYRKAASVIAKYPHKIKSGAEAKKLPGVGTKIAEKI
DEFLATGKLRKLEKIRRDDTSSSINFLTRVTGIGPSAA
RKLVDEGIKTLEDLRKNEDKLNHHQRIGLKYFEDFEK
RIPREEMLQMQDIVLNEVKKLDPEYIATVCGSFRRGA
ESSGDMDVLLTHPNFTSESSKQPKLLHRVVEQLQKVR
FITDTLSKGETKFMGVCQLPSENDENEYPHRRIDIRLIP
KDQYYCGVLYFTGSDIFNKNMRAHALEKGFTINEYTI
RPLGVTGVAGEPLPVDSEQDIFDYIQWRYREPKDRSE

Predicted Secondary Structure

Experimental

CCCCCCCCCCCCCHHHHHHHHHHHHHHCCCCCH
HHHHHHHHHHHHHCCCCCCCCHHHHCCCCCCC
HHHHHHHHHHCCCCCCCCC

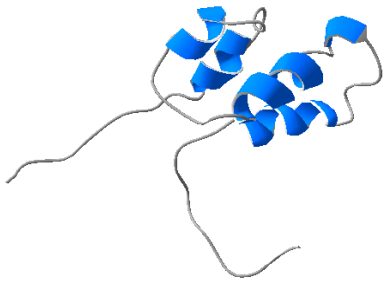
Predicted from host; Similarity 58.14%

CCCCCCCCCCCHHHHHHHHHHHCCCCCCCCCC
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
CCCCCCCHHHHCCCHHHHCC

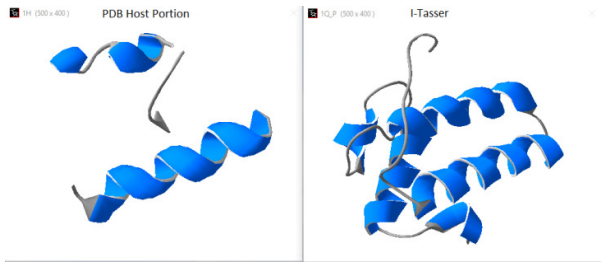
Inferred from predicted 3D; Accuracy %52

HHHHHHCCCCCHHHHHHHHHHHCCCHHHHHC

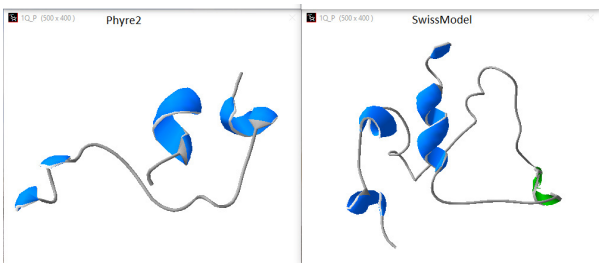
Predicted Tertiary Structure



(a)



(b) (c)



(d) (e)

Figure 1. 1NPB and 1BPE
a) Experimental b) Part of host c)I-Tasser d) Phyre2
e) SwissModel

Pair 2; Query Length 62

QUERY; 1CJG

MKPVTLYDVAEYAGVSYQTVSRVVNQASHVSAKTR
EKVEAAMAELNYIPNRVAQQLAGKQSL

HOST; 1JWL

MKPVTLYDVAEYAGVSYQTVSRVVNQASHVSAKTR
EKVEAAMAELNYIPNRVAQQLAGKQSLLIGVATSSLA
LHAPSQIVAAIKSRADQLGASVVVSMVERSGVEACKT
AVHNLLAQRVSGLIINYPLDDQDAIAVEAACTNVPAL
FLDVSDQTPINSIIFSHEDGTRLGVEHLVALGHQQIAL
LAGPLSSVSARLRLAGWHKYLTRNQIQPIAEREGDWS
AMSGFQQTMQMLNEGIVPTAMLVANDQMALGAMR
AITESGLRVGADISVVGYDDTEDSSCYIPPLTTIKQDFR
LLGQTSVDRLLQLSQGQAVKGNQLLPVSLVKRKTTL
APN

Predicted Secondary Structure

Experimental

CCCCCHHHHHHHHCCCHHHHHHHHCCCCCCCHH
HHHHHHHHHHHCCCCCCHHHHHHCCCCCC

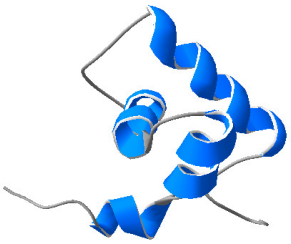
Predicted from host; Similarity. 53.23%

CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
CCCCCCCCCCCCCCCCCHHHHHHCCCCCC

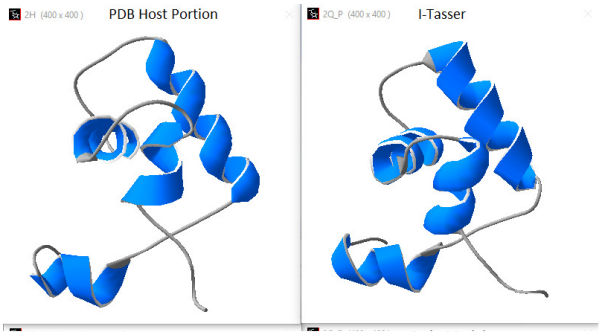
Inferred from predicted 3D; Accuracy %48

CCCCCHHHHHHHHCCCHHHHHHHHCCCCCCCHHH
HHHHHHHHHHHCCCCCHHHHHHCCCCCC

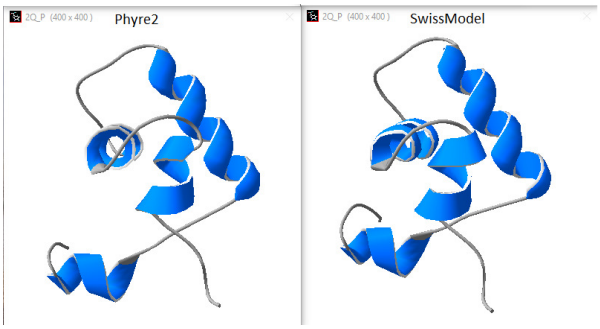
Predicted Tertiary Structure



(a)



(b) (c)



(d) (e)

Image 2: 1CJG, and 1JWL
a) Experimental b) Homolog c) I-Tasser d) Phyre2
e) SwissModel

Pair 3; Length 87

QUERY; 4IHT

MELRHLRYFVAVVEEQSFTKAADKLCIAQPPLSRQIQ
NLEEELGIQLLERGSRPVKTTPEGHFFYQYAIKLLSNV
DQMVSMTKRIAS

HOST: 3K1P

MELRHLRYFVAVVEEQSFTKAADKLCIAQPPLSRQIQ
NLEEELGIQLLERGSRPVKTTPEGHFFYQYAIKLLSNV
DQMVSMTKRIASVEKTIRIGFVGSLLFGLLPRIIHLYR
QAHPNLRIELYEMGTKAQTEALKEGRIDAGFGRLKIS
DPAIKRTLLRNERLMVAVHASHPLNQMKDKGVHLN
DLIDEKILLYPSSPKPNFSTHVMNIFSDHGLEPTKINEV
RKVQLALGLVAAGEGISLVPASTQSIQLFNLSYVPLLD
PDAITPIYIAVRNMEESTYIYSLYETIRQIYAYEGFTEPP
NWLEHHHHHH

Predicted Secondary Structure

Experimental

CCHHHHHHHHHHHHHCCHHHHHHHCCCCHHHHH
HHHHHHHHHHCCCCECCCCCCCCCECHHHHHHHH
HHHHHHHHHHHHHHHHHHHHC

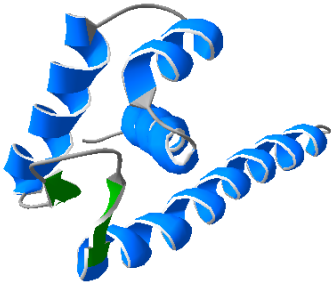
Predicted from host; Acc. 71.26%

CCCHHHHHHHHHHCCCCCCCCCCCCCCCCCCCCC
HHHHHHHHHHCCCCCCCCCCCCCCCCCHHHHHHH
HHHHHHHHHHHHHCCCCCCCC

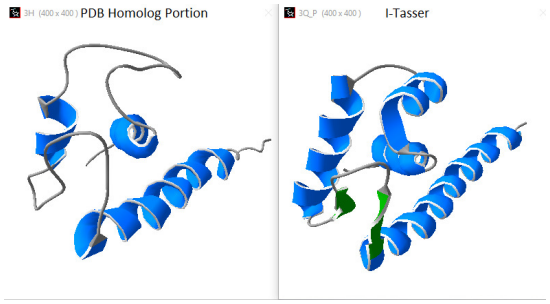
Inferred from predicted 3D; Accuracy %58

CCHHHHHHHHHHHHHHCCHHHHHHHHCCCCHHHHH
HHHHHHHHHHCCHHEEECCCCCEECHHHHHHHHHHH
HHHHHHHHHHHHHHHHHHCCC

Predicted Tertiary Structure



(a)



(b) (c)

(d) (e)

Image 3: 4IHT, and 3K1P
a) Experimental b) Part of host c)I-Tasser d) Phyre2
e) SwissModel

Pair 4; Length 160

QUERY; 3LRU

KRLGQLAKWKTAEEVAALIRSLPVEEQPKQIIVTRKG
MLDPLEVHLLDFPNIVIKGSELQLPFQACLKVEKFGDL
ILKATEPQMVLFNLYDDWLKTISSYTAFSRLILILRAL
HVNNDRAKVILKPDKTTITEPHHIWPTLTDEEWIKVE
VQLKDLILAD

HOST; 4JK9

GELFSNQIIWFVDDTNVYRVTIHKTFEGNLTTKPINGA
IFIFNPRTGQLFLKIIHTSVWAGQKRLGQLAKWKTAEE
VAALIRSLPVEEQPKQIIVTRKGMLDPLEVHLLDFPNI
VIKGSELQLPFQACLKVEKFGDLILKATEPQMVLFNL
YDDWLKTISSYTAFSRLILILRALHVNNDRAKVILKPD
KTTITEPHHIWPTLTDEEWIKVEVQLKDLILAD

Predicted Secondary Structure

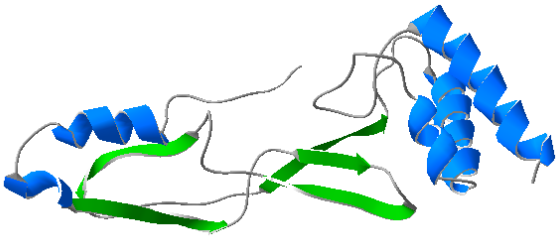
Experimental

CCCCCCCCCEEHHHHHHHHHCCCHHHHHHCEEE
EECCEEEEEEEECCCCCEEEEEEECCCCCCCEE
EEEEECCCCCCCCCCCCEEEEECCCCHHHCCCH
HHHHHHHHHHHHHHHHCHHHHHHHCCCCCCCCC
CCCCCCCCCCHHHHHHHHHHHHHHHHC

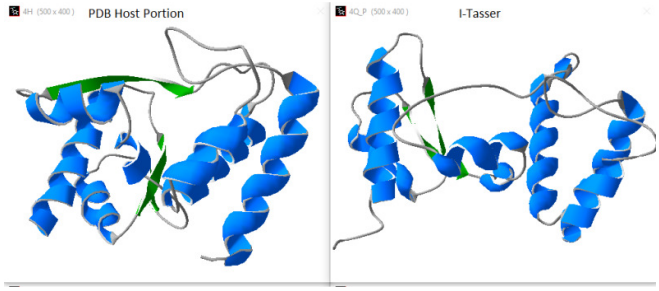
Predicted from host; Similarity 74.38%

CCCCCCCHHHHHHHHHHHHHHCCHHHCCCEEEEE
CCCCCHHHHHHHCCCCCCCEEEECCCCCCCHHHH
HHCHHHHHHHHHCCCCEEEEEECCCCHHHCCCH
HHHHHHHHHHHHHHHHHCHHHHHHHCCCCCCCCC
CCCCCCCCCCHHHHHHHHHHHHHHCCC

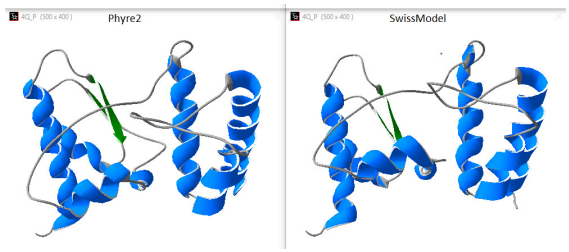
Predicted Tertiary Structure



(a)



(b) (c)



(d) (e)

Image 4: 3LRU, and 4JK9
a) Experimental b) Part of host c)I-Tasser d) Phyre2
e) SwissModel

Pair 5; Length 91

QUERY; 2HGF – query shorter than the original

RNTIHEFKKSAKTTLIKIDPALKIKTKKVNTADQCANR
CTRNKGLPFTCKAFVFDKARKQCLWFPFNSMSSGVK
KEFGHEFDLYENKDYIR

HOST; 4D3C

MGSSHHHHHHSQDPQRKRRNTIHEFKKSAKTTLIKID
PALKIKTKKVNTADQCANRCTRNKGLPFTCKAFVFD
KARKQCLWFPFNSMSSGVKKEFGHEFDLYENKDYIR
NCIIGKGGSYKGTVSITKSGIKCQPWSSMIPHEHSFLPS
SYRGKDLQENYCRNPRGEEGGPWCFTSNPEVRYEVC
DIPQCSEVE

Predicted Secondary Structure

Experimental

CCCHHHEEEEEEEEEECCCCCCCEEEECCCHH
HHHHHHHHCCCCCCCCCEEEEECCCCEEEEECC
CCCCCCEEEEEEEEEEEEEHHHCC

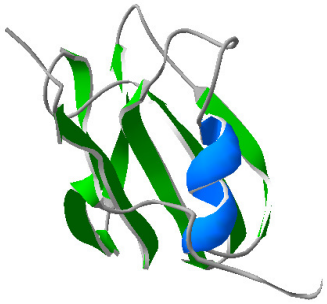
Predicted from host; Similarity 67.03%

CCCCCCEEEECCCCCCCCCCCCCCCCCCCCCH
HHHHHHHCCCCCCCCCCEEECCCCCCCCEEECC
CCCCCCCCCCCCCCCCCEEEEHHHCC

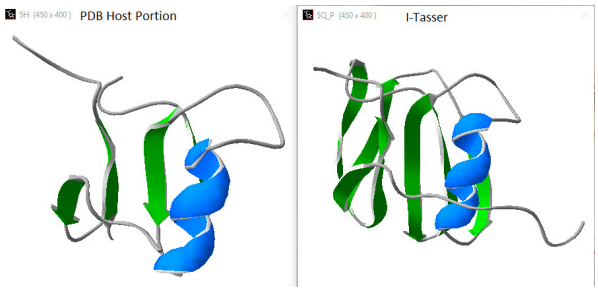
Inferred from predicted 3D; Accuracy %57

CCEEECCCCHHHHHHHHHCCCCCCCCEEEEEEECCC
CEEEEEECCCCCCCCEEECCCCEEEEECHHHCCC

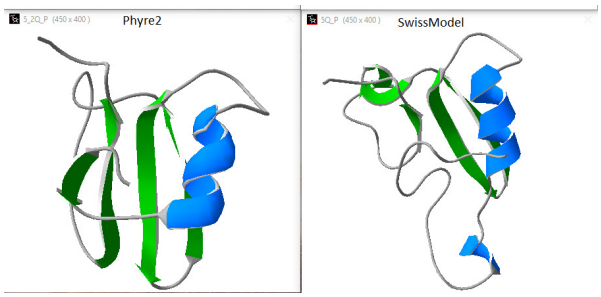
Predicted Tertiary Structure



(a)



(b) (c)



(d) (e)

Image 5: 2HGF, and 4D3C
a) Experimental b) Part of host 3D c)I-Tasser d) Phyre2
e) SwissModel

4. DISCUSSION

In an article (Akcesme, et. al. 2017), it is reported that around 17% of proteins among around 80 thousand proteins of PDB have host proteins in PDB which contains them as subsequences, and when secondary structures of these proteins predicted as the secondary structures of corresponding parts of hosts, the mean accuracy is 90.39%. The percentage of accuracies less than 50% is only 1%. This 90.39% accuracy of the host protein based secondary structure prediction set also an upper bound for the homology based tertiary structure predictions. For five query proteins with PDB codes 1BNP, 1CJG, 4IHT, 4D3C, and 2HGF, proteins with domains which have identical amino acid sequences are found in PDB with codes 1BPE, 1JWL, 3K1P, 4JK9, 4D3C respectively.

Secondary structure sequences predicted by hosts, and predicted 3D's are compared with the x-ray experimental secondary structure sequences of query proteins. Accuracies is found as seen in Table 1.

Table 1. Accuracies of host based predictions, and predictions from 3D structures for secondary structure sequences.

In %	1BNP	1CJG	4IHT	4D3C	2HGF
From Host	58	53	71	74	67
From Pr. 3D	52	48	58	-	57

The impact of the inaccuracies of host based secondary structure sequence predictions on the homology based 3D structure predictions can be seen in the above figures on the predictions made by the three predictors I-Tasser, Phyre2, and SwissModel.

Inaccuracies in predicted tertiary structures are revealed by the visual comparison of the 3D structures of query proteins and their predicted 3D images by the three 3D predictors, and their counterparts in host proteins. Also from Table 1. at the second row, it is seen that the similarity of secondary structure objects in predicted 3D's to secondary structure objects in query proteins is very poor, and always less than the similarity of the parts in hosts. This proves that when 3D structure predictors create these predictions relying on the homologs in PDB, the inaccuracies in host based secondary structure sequence predictions is carried into the homology based 3D structure predictions even by some amplification.

To some researchers claim that because of the tendency of preserving function, the inaccuracies in host based secondary structure sequence predictions would be carried into the homology based 3D structure predictions with some dissipation. But we did not observe this dissipation. This will be the topic of a further investigation wit more numerous proteins by the authors of this article.

REFERENCES

Akcesme F. B., and Can M. (2015) Secondary Structure Segments are Much More Conserved than Primary Sequence Segments, Southeast Europe Journal of Soft Computing Vol.4 No.2, pp60-65.

Akcesme, F.B. (2016) A Sequence Segments Similarity based Protein Secondary Structure Prediction Method by the Use of the Relationship between Primary and Secondary Structure of Proteins, International University of Sarajevo (PhD Thesis at IUS).

Akcesme, F.B, and Can, M. (2016a) Protein Secondary Structure Prediction Using Super-chains in PDB, SEJSC, Vol. 5, No1, 1-4.

Akcesme, F.B, and Can, M. (2016b) A Promising Similarity Based Secondary Structure Prediction Method, SEJSC, Vol. 5, No1, 15-18.

Akcesme F. B., Adilovic M., and Can M. (2017) Accuracy of Homology Based Protein Secondary Structure Prediction, Southeast Europe Journal of Soft Computing Vol.6 No.1, pp1-7.

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

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

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.

Illergard, 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

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

Miller, C. S., Eisenberg, D. (2008) Using inferred residue contacts to distinguish between correct and incorrect protein models, Bioinformatics Vol. 24 no. 14, pp. 1575–1582

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

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