Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Hypothesis**

New tips for structure prediction by comparative modeling

Anwar Rayan*

QRC-Qasemi Research Center,Al-Qasemi Academic College, P.O.B. 124, Baka El-Garbiah 30100, Israel;
Email: a_rayan@qsm.ac.il; * Corresponding author

received December 13, 2008; accepted December 29, 2008; published January 12, 2009

Abstract:
Comparative modelling is utilized to predict the 3-dimensional conformation of a given protein (target) based on its sequence alignment to experimentally determined protein structure (template). The use of such technique is already rewarding and increasingly widespread in biological research and drug development. The accuracy of the predictions as commonly accepted depends on the score of sequence identity of the target protein to the template. To assess the relationship between sequence identity and model quality, we carried out an analysis of a set of 4753 sequence and structure alignments. Throughout this research, the model accuracy was measured by root mean square deviations of Cα atoms of the target-template structures. Surprisingly, the results show that sequence identity of the target protein to the template is not a good descriptor to predict the accuracy of the 3-D structure model. However, in a large number of cases, comparative modelling with lower sequence identity of target to template proteins led to more accurate 3-D structure model. As a consequence of this study, we suggest new tips for improving the quality of comparative models, particularly for models whose target-template sequence identity is below 50%.

Keywords: comparative modelling; homology modelling; model refinement

Background:

The 3D structure determination of a certain protein greatly helps unravelling its function and binding mechanisms. Such structural information can also aids in designing experiments in mutagenesis and even utilized for structure-guided drug development or virtual screening **[1, 2]**. Since experimental structures are available only for a small number of sequenced proteins, alternative strategies are required to predict reliable models for protein structures when X-ray diffraction or NMR are not yet available **[3]**. Among the different strategies currently used for constructing 3-D structures of certain proteins, we shall find the comparative modelling (termed also as homology modelling) as the most accurate method among the computational methods, yielding reliable models **[4, 5]**.

Another approach termed "ab-initio" modelling is not practical yet for the construction of reliable models **[6]**. Usually, in comparative modelling the template is chosen by virtue of having the highest level of sequence similarity with the target, and similar secondary and tertiary structure (belongs to the same "fold"). Baker and Sali **[7]** have shown that a comparative model for a protein at medium size at least and with sequence identity of less than 30% to the template crystal structure is unreliable. The rule of sequence identity score exceeding 30% does not specify how identity should be distributed along a sequence. The quality of the models is assessed by comparing predicted structures to X-ray solved structures via superimposition and atomic root mean square deviation assessment (RMSD). A model can be considered 'accurate' or 'reliable' model when its RMSD is less than 3-4 Å.

The comparative modelling procedure for protein structure prediction is built generally from few steps: after identification of the homologous protein with known 3-D structure, sequence alignment (based on score of identity or similarity) is performed. Usually, the structurally conserved regions (SCRs) are identified and coordinates for the core of the models are generated. Following the core generation, one predicts the conformations of the structurally variable regions (termed loops) **[8]** and adds the side chains **[9]**.Some approaches, align multiple known structures firstly, then, identifying structurally conserved regions to construct an average structure, for modelling these regions of the inquiry protein. The optimal homology-based model is obtained when the correct template is chosen and each residue pair correctly aligned in the target-template sequence alignment **[10]**.

In this communication, we carried out an analysis of a large set of 4753 sequence and structure alignments and tried to answer few questions: (1) Can we predict the accuracy of the modelled structure based on sequence identity score? (2) Is it always justified to select the protein with highest identity score as a template for comparative modelling? (3) How can we improve accuracy of homology-based models?

© 2009 Biomedical Informatics Publishing Group

Methodology:

We downloaded about 124 unique proteins which belong to serine protease family from the Brookhaven Protein Databank (PDB) **[11-12]**. Then, IMSA - Intelligent Multiple Sequence Alignment **[13]** (in-house software based on the Intelligent Learning Engine (ILE) optimization technology) was utilized to optimally align the whole set of all sequences. Accurate multiple sequence alignment (MSA) is important step that may improve the accuracy of pairwise sequence alignments, minimize misalignments and generate more accurate 3-D models **[14-16].** Sequence identity score was calculated for each pair of sequences. All residues from the multiple sequence alignment were found only on 98 proteins (Table 1, see supplementary material). Other twenty eight proteins lack coordinates of one residue at least in their 3-D experimentally determined structures. The alpha carbons (Cα) for residues of selected proteins were extracted from the PDB structures and structurally superimposed.

The quality of the models was assessed via superimposition of the predicted homology-based model (target protein) and the X-ray structure of the protein and then, measurement of the Cα root mean square deviation (Cα RMSD). We have defined 'highly accurate' model as the one having <=2 Å RMSD from the experimentally determined structure, while models having Cα RMSD above this threshold and <=4 Å were termed "reliable" models. Such reliable models could fit for designing mutagenesis experiments but not for drug design and binding affinity tests. BioLib software was used for performing structural alignment and for computing the Cα RMSD (BioLib is an open-environment developing toolkit developed by BioLog Technologies Ltd.).

The multiple sequence alignment matrix obtained from running our in-house software on the selected database of serine proteases, was processed as described below, in order to specify which parts of the whole set of sequences to select for comparative modelling. We use a "voting" approach, in which each amino acid contributes to the conservation at a sequence position according to its frequency in that particular position (equation 1, under supplementary material). These frequencies are measured in all sequences of the database.

Discussion:

In this study, we aim to assess models obtained by comparative modelling by analyzing a large set of sequence/structure alignments that belong to the same family of proteins (adopt the same "fold"). The pair-wise sequence alignments in our database produced sequence identity that ranges between 28% and 100% (Figure 1).

The sequence analysis of the indicated database revealed highly conserved amino acid residues distributed along the protein chain (Figure 2, for number of amino acids found above certain conservation thresholds). We postulate that the orientation of such residues within their spatial coordinates play an important role in the protein function and/or in stabilizing the protein folding (or conformation). Thus, the inter-residue distance matrix should be somehow similar in each protein. This could be assessed qualitatively by extracting those residues from the X-ray structures of the proteins and then performing pair-wise superposition. As depicted in Table 2 (under supplementary material), the Cα RMS deviation is very low in general and correlates well with the Positional Conservation Threshold (PCT). These findings reveal the correctness of the multiple sequence alignment and could be utilized in refinement of models.

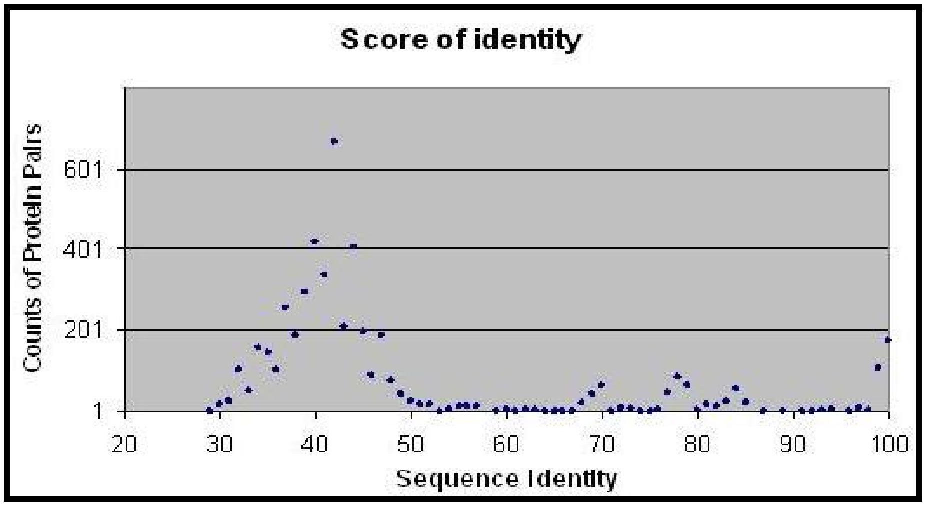

Figure 1: Spread of sequence identities in the database (4753 protein pairs in total).

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 3(6): 263-267 (2009)

Bioinformation, an open access forum
© 2009 Biomedical Informatics Publishing Group

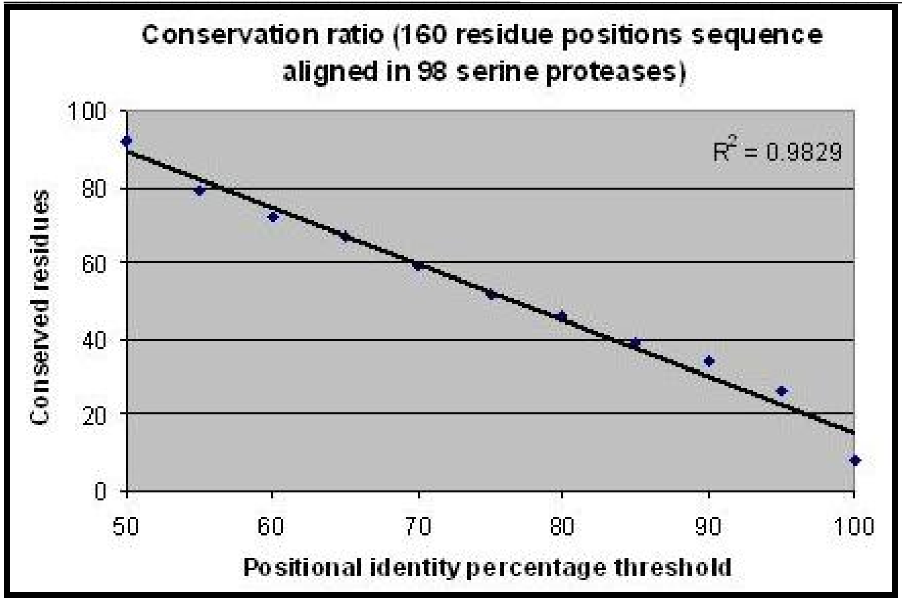

Figure 2: Analysis of positional conservations in the sequences of 98 unique serine proteases. Each protein has 160 residues and the multiple sequence alignment was performed without gaps.

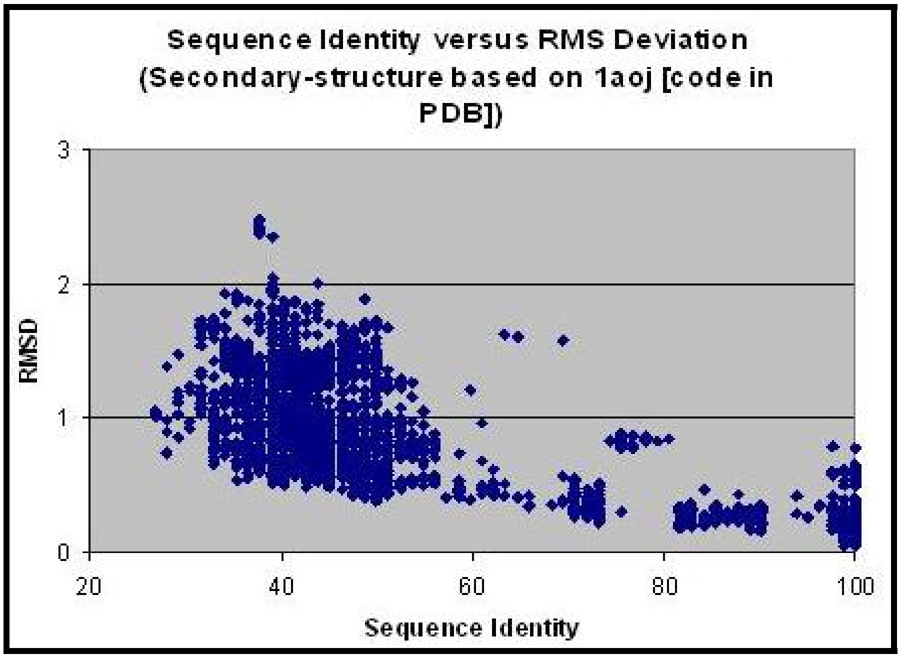

Figure 3: This plot describes the sequence identity between target and template sequences and the relative mean square deviation of the models from their corresponding experimental control structure (taking into account only secondary structure segments, based on 1a0j[18]).

4753 models of proteins have been generated and assessed (Figure 3). Models of proteins that were built based on templates that share a certain degree of sequence identity (> 28%) with the target are mostly accurate (<2 Å RMSD). Such models seem to be useful for drug design and docking experiments. However, when the degree of sequence identity is below 50%, the best template to thread on is not always the one with the highest identity score. To choose the best template for comparative modelling, other protein structures with lower sequence identity should be evaluated.

In comparative modelling, two important issues should be taken into consideration in order to avoid inaccuracy in model generation. First, choosing the proper modelling template, as most high deviations between model and experimental control structures can be traced back to the selected modelling templates. Second, conducting the right sequence alignment between the target and template. Any error introduced by the alignment algorithm will have profound effects on the model. Obtaining higher percentage of accuracy highly depends on choosing the correct protein as a template, performing the correct alignment, and choosing the correct stretches to remodel. Position conservation threshold may be used for further refinement of the model via applying molecular dynamics (MD), simulated annealing (SA), iterative stochastic elimination (ISE) or other optimization approaches [**17, 18**].

Conclusion:

We present sequence and structural analysis of 4753 pairs of proteins and raise few questions regarding the comparative modelling procedure. We may inquire the justification of the common accepted rule of choosing templates having the highest sequence similarity with the target for comparative modelling. Our findings show that sequence identity of the target to the template is not always a reliable descriptor to predict the accuracy of the 3-D structure model. In a large number of cases, comparative modelling with lower sequence identity of targets to templates led to better 3-D structure models. It is seen clearly when the sequence identity is below 50%. Employing position conservation threshold - PCT (data shown in Table 2, under supplementary material) to refine models is currently under evaluation in our lab. Preliminary results show that such usage is recommended as better homology-based models could be obtained.

Acknowledgment:

I gratefully acknowledge Rand Biotechnologies LTD for supporting this research.

References:

[01] A. Patny *et al.*, *Current Medicinal Chemistry*, 13: 1667 (2006)[PMID: 16787212]

[02] P. Ferrara and E. Jacoby, *J Mol Model*, 13: 897 (2007) [PMID: 17487515]

[03] H. Eszter and B. Zsolt, *Journal of Structural Biology*, 162: 63 (2008)[PMID: 18249138]

[04] N. Furnham *et al.*, *BMC Struct Biol.*, 8: 89 (2008)[PMID: 18237407]

[05] R. Sanchez and A. Sali, *PNAS*, 95: 13597 (1998)[PMID: 9811845]

[06] C. Hardin *et al.*, *Curr Opin Struct Biol.*, 12: 176 (2002) [PMID: 11959494]

[07] D. Baker and A. Sali, *Science*, 294: 93 (2001) [PMID: 11588250]

[08] A. Rayan *et al.*, *Current Medicinal Chemistry*, 11: 675 (2004)[PMID: 15032723]

[09] M. Glick *et al.*, *PNAS*, 99: 703 (2002)[PMID: 11792838]

[10] L. Roland and Jr. Dunbrack, *Current Opinion in Structural Biology*, 16: 374 (2006)[PMID: 16713709]

[11] H. M. Berman *et al.*, *Nucleic Acids Research*, 28: 235 (2000)[PMID: 10592235]

[12] K. Henrick *et al.*, *Nucleic Acids Res.*, 36: D426 (2008)[PMID: 18073189]

[13] A. M. Rayan and J. A. Raiyn, *Intelligent Learning Engine (ILE) Optimization Technology*, Provisional Patent (2008)

[14] S. Chakravaty *et al.*, *BMC Structural Biology*, 8: 31 (2008) [PMID: 18631402]

[15] M. S. Chang and S. A. Benner, *J Mol Bio.*, 341: 617 (2004)[PMID: 15276848]

[16] J. O. Wrabl and N. V. Grishin, *Proteins*, 54: 71 (2004) [PMID: 14705025]

[17] A. Rayan *et al.*, *J Mol Graph Model*, 22: 319 (2004) [PMID: 15099829]

[18] H. K. Schroder *et al.*, *Acta Crystallogr D Biol Crystallogr.*, 54: 780 (1998) [PMID: 9757092]

Edited by P. Kangueane

Citation: Rayan, Bioinformation 3(6): 263-267 (2009)

License statement: This is an open-access article, which permits unrestricted use, distribution, and reproduction in any medium, for non-commercial purposes, provided the original author and source are credited.

Supplementary material

Equation

$$\boxed{C_{ij} = \frac{n_{ij}}{k} * 100\%} \longrightarrow \quad (1)$$

$\mathbf{C_{ij}}$ is thus the conservation factor for residue type i at sequence position j. n_{ij} is the number of sequences, which have amino acid *i* at position *j* of the multiple alignment, and *k* is the total number of sequences in the database.

© 2009 Biomedical Informatics Publishing Group

Table 1: PDB codes of 98 serine proteases (the first four letters are the code of the protein in the PDB while the last letter is the chain ID).

1 AMHA 1ANB0 1ANC0 1AND0
1BRBE
1 CO7E 1DPO0 1F7ZA 1SLUB
1SLWB
3 TGJE 1QL9A 1J16A 1TRMA 1EZSC
1 F5RA 1FY8E 3TGKE 1AN1E
1MCTA
1 S83A 1TAWA 1UTNA 1OPHB
1V2OT
1 V2QT 1V2RT 1V2ST 1V2WT
1V2NT
1 V2LT 1H4WA 1TRNA 1UTMA
1HJ8A
1 MBQA 1BIT0 1A0JA 1DX5M
1JOUB
1 RD3B 1THPB 1C5LH 1H8DH
2THFB
1 H8IH 1Q3XA 1GPZA 1TQ7B
1SHHB
1 VR1H 1UCYK 1EUFA 1FI8A 1PJPA
1 NN6A 1KLT0 1IAUA 1GVKB
1HAXB
1 QNJA 1BRUP 1DST0 1BIO0 1RFNA
1 PFXC 1A0LA 1CGHA 1FXYA
1LO6A
1G2LA 1FAXA 1LTOA 1TON0
1NPMA
1MZAA 3RP2A 1AO5A 1KLIH 1KIGH
1AZZA 1EAXA 1GVZA 1PYTD
1OP8A
1ORFA 1RTFB 1AUTC 1P57B 1FIZA
1FIWA 1BQYA 1A5IA 1B7XB 1BTHH
1 MD8A 1EQ9A 1EKBB

Table 2: Correlation between Positional Conservation Threshold (PCT) and RMS deviation is given. * Positional Conservation Threshold (according to equation 1)– the residue should be conserved above this threshold in the certain position.

PCT*	Average RMSD	Median	Standard deviation
100	0.503	0.463	0.219
95	0.631	0.486	0.387
90	0.621	0.497	0.361
85	0.609	0.492	0.342
80	0.704	0.577	0.368
75	0.757	0.702	0.361
70	0.812	0.822	0.362
65	0.862	0.898	0.372
60	0.894	0.920	0.405
55	0.936	0.979	0.405
50	0.990	1.059	0.408