BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Hypothesis

Volume 8(20)

# Protein structure validation using a semi-empirical method

**Tapobrata Lahiri[1]*, Kalpana Singh[1], Manoj Kumar Pal[1] & Gaurav Verma[2]**

[1]Division of Applied Science and Indo-Russian Center for Biotechnology, Indian Institute of Information Technology, Deoghat, Jhalwa, Allahabad, India 211012; [2]School of Life Sciences, Jawaharlal Nehru University, New Delhi, India; Tapobrata Lahiri – Email: tlahiri@iiita.ac.in; Phone: 91-532-2922242, Fax: 91-532-2430006; *Corresponding author

Received September 19, 2012; Accepted October 01, 2012; Published October 13, 2012

**Abstract:**
Current practice of validating predicted protein structural model is knowledge-based where scoring parameters are derived from already known structures to obtain decision on validation out of this structure information. For example, the scoring parameter, Ramachandran Score gives percentage conformity with steric-property higher value of which implies higher acceptability. On the other hand, Force-Field Energy Score gives conformity with energy-wise stability higher value of which implies lower acceptability. Naturally, setting these two scoring parameters as target objectives sometimes yields a set of multiple models for the same protein for which acceptance based on a particular parameter, say, Ramachandran score, may not satisfy well with the acceptance of the same model based on other parameter, say, energy score. The confusion set of such models can further be resolved by introducing some parameters value of which are easily obtainable through experiment on the same protein. In this piece of work it was found that the confusion regarding final acceptance of a model out of multiple models of the same protein can be removed using a parameter Surface Rough Index which can be obtained through semi-empirical method from the ordinary microscopic image of heat denatured protein.

**Background:**
Protein structure validation is as important a task as to obtain its structure through either experiments like X-Ray Crystallography or NMR or by Homology or Threading based prediction methods. Importance and limitation of knowledge-based validation of protein structure is well documented in the review of Kihara *et al* (2009) **[1].** In this context, Semiempirical validation model for protein structure is indeed a new idea being introduced in this work. However, there exists reports on attempts based on semiempirical strategy to unearth structural information of many protein related events **[2]** worked on use of semiempirical methods for building geometric model of proteins. Möhle et al (2001) **[3]** showed utility of semi-empirical method to improve efficiency in deducing secondary structure of peptides and proteins. Paper of Khandogin and York (2004) **[4]** presented a set of macromolecular quantum descriptors for surface characterization of macro-biomolecules in solution, extraction of which needs modest computational cost because the method was backed by linear-scaling semi-empirical quantum / solvation methods. In a similar effort Raha and Merz (2005) **[5]** presented a scoring function that has been derived by using semi-empirical quantum mechanics to calculate the electrostatic interactions between protein and ligand and solvation free energy expected during complexation. Huey *et al* (2007) **[6]** claimed successful development and testing of semiempirical force field for incorporation in AutoDock4 formalism.

Giving due regards to these research works, it can further be noted that there is every possibility to end up with a set of multiple structural models for the same protein due to non-convergence of decision on single model based on different parameters. It indicates that there still remains requirement of a method that can remove above-described confusion. In this context, previous work of Mishra and Lahiri (2011) **[7]** and the references of works therein showed that extraction of structure parameter, Surface Roughness Index (SRI) of a protein whose structure was not known, was possible using semi-empirical

 

method. They also showed that the predicted SRI value of a protein well correspondence with its calculated counterpart obtained from its known PDB structure.

The present paper-work showed that the current practice of validation of protein structure can be further strengthened by introducing parameters that are easily and experimentally obtainable from the protein of interest only. Therefore, this approach can uniquely solve the confusion regarding acceptance of a particular model out of various models of the same protein.

**Methodology:**

***Selection of Proteins in the Study***

The proteins which are available in the market as well as listed in the PDB site were selected. For this pilot study, lysozyme, Cytochrome C, Ferritin and Albumin were chosen keeping in mind the diversity of secondary structure content for these proteins. All the proteins were obtained from Sigma Aldrich (USA).

***Experimental Design***

Each of the proteins is suspended in milipore water at concentration of 25 mg/cc and put in hot water bath having temperature 100°C for 15 minutes to obtain Heat Denatured Protein Aggregates (HDPAs). Suspension of HDPAs kept at hemocytometer slides (Model: Neubauer Chamber, Marienfeld, Germany) and covered with thin microscopic glass cover slip, was visualized at 400X magnification using phase contrast microscope (Leica Model DML-B2). Digital images of aggregates were captured using a camera (Canon PowerShot S50) at optical zoom 2X. Thus cumulative optical zoom of the microscope and camera was 800X. 50 images of different HDPAs were captured for each protein. This work was carried out following Mishra and Lahiri (2011) **[7].**

***Algorithmic Formalism***

All Multiple structural models were generated for each protein through homology model which also included its original PDB structure. Calculated SRI (CSRI) for each of these structural models were obtained following the method of Lahiri et al (2006) **[8]** and semi-empirically obtained predicted SRI (PSRI) values for the same protein were obtained using the method of Mishra and Lahiri (2011) **[7].** The target of this work was to check whether PSRI of a protein which can be obtained without being bothered about its original structure information, can be used to ascertain the final validation of a single model out of a confusion set of structural models of a protein. In these purpose Euclidian distances of CSRIs of all the structural models of a particular protein from that of CSRI of the original structure and PSRI of the same protein giving Distance Profile Vectors (Calculated), $\mathrm{DPVC} = \{dci\}_{i=1}^{N}$ and (Predicted), $\mathrm{DPVP} = \{dpi\}_{i=1}^{N}$ respectively for N number of models was measured. ***dci*** and ***dpi*** are the distances from CSRI of i-th model from that of original structure and PSRI of the same protein respectively. Our intention was to check whether both of these distance profiles are same which means the minimum distances obtained both from DPVC and DPVP are for the same structural model or not. If it is found for the same model then it indicates that in absence of experimentally obtained original protein structure, PSRI, that is obtainable through semi-empirical method, can be utilized for final validation and selection of a single-best structural model from a confusion set of multiple structural model. Flow chart of the algorithm is given in **(Figure 1)** for further clarification.

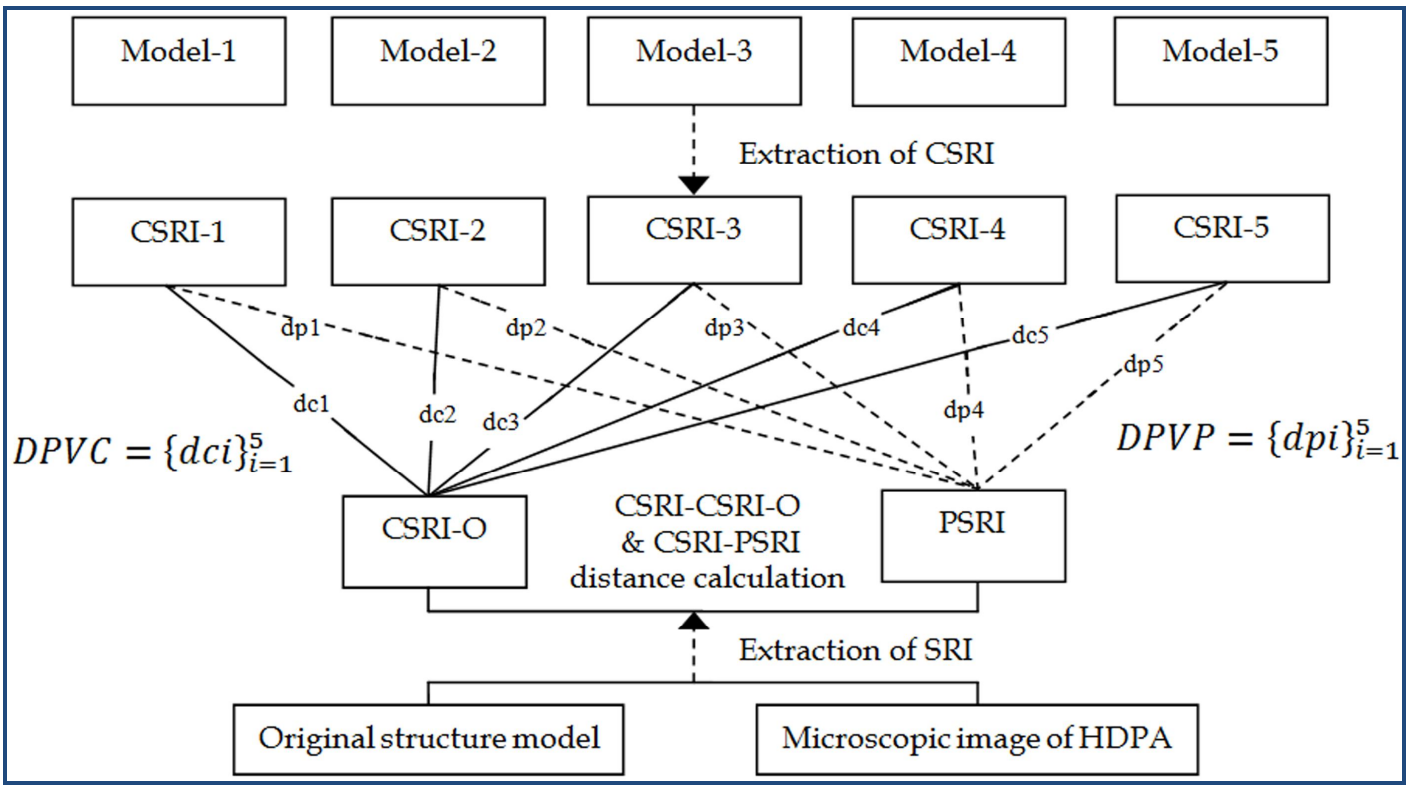


**Figure 1:** Calculation of distance profiles DPVC and DPVP for a set of 5 predicted structural models of a protein. Here, CSRI-O is SRI calculated from original structure, CSRI is calculated SRI from 5 models and PSRI is predicted SRI.

 

**Discussion:**

**Table 1 (see supplementary material)** shows results of structure validation for 5 best models generated from homology model for the same protein for which PDB structure was already available and its PDB structure was excluded from the homologue database to conform to the fact that prediction method had no prior knowledge of known structure. The result of validation from different validation methods, viz., Dope, Procheck, Verify3D and Errat shows confusion set of structure model for each of the protein indicating difficulty in accepting a single model as the best one. For example, for the protein Lysozyme, validation method Dope, Procheck, Verify3D and Errat shows best models as 1st, 5th, any one among five and 3rd model respectively indicating difficulty in forming a decision regarding acceptance of a single model as the best validated model.

The result shown in **Table 1** showed similar outcome for the other proteins too. In this situation the strategy given in method section and flowchart in **(Figure 1)** suggested to utilize the experimentally drawn information from the concerned protein only to help in zeroing in a single model for final acceptance. As described in Algorithmic Formalism of Method Section, the closeness of these models to the original structure can be checked by measuring distance between SRI deduced from original structure (CSRI) and SRIs deduced from these model structures. For example, dci values as referred in algorithmic section gives distance profile vector DPVC. As described by Lahiri *et al* (2006) **[8],** SRI is a surface roughness property marker of a protein and therefore can be utilized as an intrinsic property of a protein related to its surface. Hence distance profile values DPVC gives closeness of a model to the original structure from where the closest model can be accepted as final structural model. In this regard, DPVC is also helping us to eliminate confusion for acceptance of a model from a set of models.

**Conclusion:**

While utility of SRI to SRI distance of models from original structure can be understood for finalizing acceptance of a single model, the difficulty of this formalism is that it requires SRI derived from original structure. Therefore, in absence of experimentally evaluated original structure (which is the case for any protein for which we need to predict and finally validate its structure) we require reference SRI that can be derived by some other simpler means. In this direction Mishra and Lahiri (2011) **[7]** has given a semi-empirical method which uses a very simple experimental arrangement (vide section Material and Methods) to derive predicted SRI using non-parametric function and human cognition model referred as PSRI in our work. This work showed that PSRI is very close to SRI derived from the original structure for which we have decided to use PSRI as reference SRI. The result shown in Table 1 also conforms to the fact that SRI to SRI distance profile DPVC calculated using reference SRI derived from original structure matches well with that for reference SRI derived from semi-empirical method of Mishra and Lahiri (2011) **[7]** (referred as DPVP in this work). Therefore, this work indicates that semi-empirically drawn SRI of a protein can be used for final validation and acceptance of a single protein model out of a confusion set of multiple structural models obtained from homology and other prediction methods.

**Acknowledgement:**

We are thankful to Indian Council of Medical Research for financial aid in the form of external project (Grant No.52/8/2005-BMS, dated-04/02/2010) for funding this work.

**References:**

**[1]** Kihara D *et al. Current Protein and Peptide Science*. 2009 **I**: 10 [PMID: 19519452]
**[2]** Stewart JJP, *Journal of Molecular Structure: THEOCHEM*. 1997 **3**: 195
**[3]** Mohle K *et al. Journal of Computational Chemistry.* 2001 **22**: 5
**[4]** Khandogin J & York DM, *Bioinformatics.* 2004 **56:** 4 [PMID: 15281126]
**[5]** Raha K & Merz Jr KM, *Journal of Medical Chemistry.* 2005 **48**: 14 [PMID: 15999994]
**[6]** Huey R *et al. Journal of Computational Chemistry.* 2007 **28**: 6 [PMID: 17274016]
**[7]** Mishra H & Lahiri T, *Bioinformation.* 2011 **6:** 4 [PMID: 21572883]
**[8]** Singha S *et al. Online Journal of Bioinformatics*. 2006 **7:** 2

**Edited by P Kangueane**

**Citation: Lahiri *et al*.** Bioinformation 8(20): 984-987 (2012)

BIOINFORMATION

*open access*

## Supplementary material:

**Table 1:** Results of structure validation for 5 best models, where DC and DP are the distances from CSRI of a model from that of original Structure and PSRI of the same protein respectively

| Protein name | Model Number | Dope score | Procheck score | Verify 3D score | Errat score | DCs | DPs |
|---|---|---|---|---|---|---|---|
| **Lysozyme** | model1 | -13247 | 89.40% | 100% | 91.73 | 2.4613 | 2.5738 |
| | model2 | -13185 | 92% | 100% | 80.83 | 2.6249 | 2.6389 |
| | model3 | -12891 | 91.20% | 100% | 93.38 | **2.2671** | **2.3298** |
| | model4 | -13136 | 89.40% | 100% | 89.16 | 2.6715 | 2.7406 |
| | model5 | -13094 | 93.80% | 100% | 78.51 | 3.4831 | 3.4861 |
| **CytochromeC** | model1 | -3967 | 89.70% | 92.75% | 88.13 | 3.3985 | 4.0411 |
| | model2 | -4097 | 89.7%+1.7% | 89.86% | 67.79 | **3.1422** | **2.2847** |
| | model3 | -4037 | 93.10% | 82.61% | 76.27 | 3.9859 | 4.2918 |
| | model4 | -4080 | 89.7%+1.7% | 79.71% | 77.96 | 3.2232 | 3.9812 |
| | model5 | -4126 | 89.70% | 71.01% | 94.91 | 3.3669 | 2.8601 |
| **Ferritin** | model1 | -20346 | 97% | 68.85% | 91.37 | **3.1734** | **3.1132** |
| | model2 | -20076 | 96.40% | 58.47% | 91.95 | 3.2584 | 3.2188 |
| | model3 | -19913 | 95.80% | 67.21% | 84.48 | 4.3847 | 4.3345 |
| | model4 | -20140 | 94.60% | 96.72% | 81.03 | 4.0759 | 4.0145 |
| | model5 | -20053 | 95.80% | 77.60% | 85.05 | 3.4773 | 3.4585 |
| **Albumin** | model1 | -66914 | 93.10% | 92.32% | 91.68 | 19.3521 | 19.2177 |
| | model2 | -68092 | 93.60% | 91.30% | 91.85 | 19.2281 | 19.0856 |
| | model3 | -67802 | 93.20% | 91.30% | 92.37 | 19.6559 | 19.5344 |
| | model4 | -67408 | 93.10% | 89.08% | 94.45 | 20.6102 | 20.4679 |
| | model5 | -67508 | 93.10% | 96.76% | 94.45 | **18.9898** | **18.8635** |