

Guest editorial:

***IN SILICO* PROTEIN MODELING: POSSIBILITIES AND LIMITATIONS**

Chhedi Lal Gupta¹, Salman Akhtar², Preeti Bajpai^{1*}

¹ Department of Biosciences, Integral University, Lucknow, 226026, India

² Department of Bioengineering, Integral University, Lucknow, 226026, India

* Corresponding author: Preeti Bajpai, Department of Biosciences, Integral University, Lucknow, 226026, India, Mobile: +91-9918040777, E-mail: preeti2874@gmail.com, pbajpai@iul.ac.in

Rapid advancements in the post-genomic era along with the introduction of novel sequencing technologies provided an even platform for the researchers around the world to sequence new protein and nucleotide sequences in a faster and efficient manner. April 16, 2014 release of UniProtKB database reports about 544996 protein sequence entries for SwissProt (<http://web.expasy.org/docs/relnotes/relstat.html>) while 54958551 protein sequence entries for TrEMBL (<http://www.ebi.ac.uk/uniprot/TrEMBLstats>). However this rate is unmatched with the way the structures of proteins have been deciphered experimentally during the same tenure. April 16, 2014 data of the Protein data bank reports around 99472 macromolecular structures being determined and submitted in the database with its 92.6 % share of protein structures (<http://www.rcsb.org/pdb/home/home.do>). Of which 88.5 % were deciphered by X-ray crystallography (XRD) followed by 10.4 % using Solution NMR and rest with other techniques like Electron Microscopy, Neutron Diffraction etc.

The fundamental biological concept of “Sequence implies the Structure and Structure implies the Function” deciphers that this increase in the amount of sequence knowledge does not reflect any biological significance until the structure of the protein is identified. Criticality the biological

function of a protein is totally dependent on its native 3D structure. Frequently applied protein structure determination techniques viz. XRD or NMR are nevertheless quite accurate but highly expensive and overwhelming venture (Schmidt and Lamzin, 2002). Furthermore the technical limitations of proteins resisting the purification process or upholding their native state after crystallization projects towards the pressing need for predicting protein structures computationally (Aloy and Russell, 2006). Different methods have been employed for predicting the 3D structures of protein which can be broadly categorized as a) Homology modeling b) Threading or Fold recognition and c) *ab initio* methods. Homology modeling also designated as Comparative modeling constructs the unknown structure of the target protein by comparing and utilizing the available information of its ≥ 50 % homologous protein sequence (Sali and Blundell, 1993). The method is highly reliable on the sequence similarity with limited errors in side chains and loop positioning. Homology build structures are analogous to typically resolved structures by NMR. Threading based methods employ the sequence – structure alignment strategy and fold assignments methods when the sequence similarity falls below the desired range of homology modeling technique (Wu and Zhang, 2007). Selected protein

structures from databases such as PDB, FSSP, SCOP or CATH after removing proteins with high sequence similarity act as structural templates for the alignment. The unavailability of suitable template for modeling dictates towards the use of *ab initio* methods where thermodynamic and molecular energy parameters are functional at the atomistic level of each of the amino acid with its congregation to propose a 3D conformation of entire protein with minimum entropy and maximum stability (Wu et al., 2007; Simons et al., 2001). The complexity and uncertainty due to the limitations to model within the range of 100 amino acid acids only restricts the use of this method extensively. Different offline and online tools for the computational prediction available at ease are Modeller, Swiss Model Workspace, Nest, SegMOD, Geno-3d, Threader, Rossetta etc with good web interface. ModBase, Swiss Model Repository and Protein model database (PMDb) have been specifically designed in the recent past to store the *in silico* modeled 3D structures of proteins.

Indeed current modeling techniques have achieved structure prediction accuracy at such a level that they are being frequently used in drug design, virtual screening, protein engineering and site-directed mutagenesis applications. At the molecular level it has also helped the researchers to further reveal the sequence to structure relationship, unveil the novel pathway of protein folding and characterizing the active and catalytic sites in the proteins. However despite of significant progress researchers in their ability to further increase the quality of the models to the experimental level have been delimited by several challenges and shortcomings. The computational models often represent only fractions of the full-length of desired protein leaving behind the unresolved questions in template-based modeling to combine information from multiple templates, viz., different structural domains, into larger complex assemblies. The development of consistent, accurate and

progressive methods for improvement of models by shifting the coordinates parallel to the native state is one of the burning issues (MacCallum et al., 2011; Wass et al., 2011). To some extent the largest possibilities in the escalation of the models came from more experimentally determined structures which allow better conceivable templates for the targets. Similarly employment of PSI BLAST algorithm as compared to normal BLAST may provide optimal template selections especially in distant evolutionary cases (Altschul et al., 1997). Improved sequence to structure alignment residuals with better energy functions for evaluating the fit may allow precise fold recognition and alignment in threading studies. Moreover, refinement of the predicted model by the Molecular dynamics simulations can prove to be major breakthrough in adjustments to side chain stereochemistry, backbone conformation and model correction (Zhao et al., 2013; Nygaard et al., 2013). In another distinctive view, functional and evolutionary similarities between the target and template should be given more attention apart from sequence similarity to avoid the false positives.

In conclusion, complementing the expensive and time consuming wet lab set up, the *in silico* modeling methods in the coming years with their ability to predict reliable 3D conformation of proteins close to their native structures and overruling the limited shortcomings by inheriting novel biological concepts will definitely support and coordinate the critical structural biological studies at the forefront.

ACKNOWLEDGEMENT

This work was carried out with the support of Department of Science and Technology, New Delhi, India as INSPIRE Fellowship (IF120575).

REFERENCES

- Aloy P, Russell RB. Structural systems biology: modelling protein interactions. *Nat Rev Mol Cell Biol* 2006;7:188–97.
- Altschul S, Madden T, Schaffer A, Zhang J, Zhang Z, Miller W et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997;25:3389–402.
- MacCallum JL, Perez A, Schnieders MJ, Hua L, Jacobson MP, Dill KA. Assessment of protein structure refinement in CASP9. *Proteins* 2011;79:74–90.
- Nygaard R, Zou Y, Dror RO, Mildorf TJ, Arlow DH, Manglik A et al. The dynamic process of $\beta(2)$ - adrenergic receptor activation. *Cell* 2013;152:532–42.
- Sali A, Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol* 1993; 23:779–815.
- Schmidt A, Lamzin VS. Veni, vidi, vici - atomic resolution unravelling the mysteries of protein function. *Curr Opin Struct Biol* 2002;12:698–703.
- Simons KT, Strauss C, Baker D. Prospects for ab initio protein structural genomics. *J Mol Biol* 2001; 306:1191–9.
- Wass MN, David A, Sternberg M.J. Challenges for the prediction of macromolecular interactions. *Curr Opin Struct Biol* 2011;21:382–90.
- Wu S, Zhang Y. LOMETS: A local Meta - threading - server for protein structure prediction. *Nucleic Acids Res* 2007;35:3375–82.
- Wu S, Skolnick J, Zhang Y. Ab initio modeling of small proteins by iterative TASSER simulations. *BMC Biology* 2007;5:17.
- Zhao G, Perilla JR, Yufenyuy EL, Meng X, Chen B, Ning J et al. Mature HIV-1 capsid structure by cryo-electron microscopy and all-atom molecular dynamics. *Nature* 2013;497:643–6.