International Journal of Scientific Research in Biological Sciences **Research Paper**
Vol.5, Issue.2, pp. 1-6, April (2018) **E-ISSN:** 2347-7520

3D structure prediction of OCT 4, an important Reprogramming Facto of induced Pluripotent Stem Cells (iPSCs)

P.Chauhan[1*], N. Wal[2]

[1]Ph.D Scholar, Department of Microbiology, Mewar University, Gangrar, Chittorgarh, Rajasthan, India
[2]Associate Professor, Department of Microbiology, Mewar University, Chittorgarh, Rajasthan, India

[*]*Corresponding Author: bi.poornima@gmail.com, Tel.: 9045063529*

Available online at: www.isroset.org

Received: 06/Mar/2018, Revised: 15/Mar/2018, Accepted: 12/Apr/2018, Online: 30/Apr/ 2018

***Abstract*—** Oct 4 is one of the transcription factors among six reprogramming factors (OCT4, SOX2, KLF4, C-MYC, NANOG, and LIN28) selected by Takahashi and Yamanaka to induce somatic cells into pluripotent stem cells (iPSCs).Stem cell research is used in treatment of a number of diseases including genetic disorders. Several questions regarding reprogramming factors of stem cells are remaining unanswerable due to limited experimental availability and ehilical issues. Proteomic analysis of OCT 4 is still remaining unpredicted as protein structure is not available in PDB. The aim of this study was prediction of the tertiary structure of OCT4 protein using homology modeling approach through MODELLER program. Quality and reliability assessments were performed on predicted model and found the model reliable.

***Keywords*—**: iPSCs, therapeutic targets, homology modeling, template, reprogramming factors, OCT 4

I. INTRODUCTION

Stem cells have the capacity to divide mitotically to produce specialized cells and more stem cells. Embryonic stem cells and adult stem cells are generally found in humans [1].
**

In 2006, Shinya Yamanaka made a ground breaking discovery that would win him the Nobel Prize in Physiology or Medicine. They found a new method to reprogrammed adult specialized cells into stem cells. These reprogrammed stem cells had the capacity to make various types of cells in the body and are named as induced pluripotent stem cells, or iPS cells [2]. Combinations of six reprogramming factors Oct 4, Sox2, c-Myc, LIN28 and Klf4 were used to develop pluripotency in mouse and human fibroblast cells [3, 4, 5].

Oct-4 (octamer-binding transcription factor 4) also known as POU5F1 (POU domain, class 5, transcription factor 1) is a protein that in humans is encoded by the POU5F1 gene [6].To analyze detailed mechanism and interaction of Oct 4 with other transcription factor, 3D structure of the protein must be studied [7]. In this paper, we have focused on structural prediction of human OCT4 protein with the help of bioinformatics tools

The aim of protein modelling is to predict the 3D structure of a protein from its primary structure with an accuracy that is comparable to the best results achieved experimentally. RCSB PDB database have detailed structural information of various proteins of different organisms. Basically X-ray crystallography or nuclear magnetic resonance (NMR) techniques are used to determine the protein structure, which are expensive, time consuming and complex process. Therefore computational approaches are used to determine the 3D structure from a protein sequence.

Homology modelling algorithm is the most popular that is based on alignment result of the template and target sequence. Template is a known protein structure, searched by alignment tool against PDB. If the idenditity of the template with target sequence is greater than 30% then it can act as template [8.9].

Our work focuses on predicting the 3D structure of OCT4 using homology modeling approach through MODELLER 9.11 program, validating and analyzing their active sites. Stem cell research has indispensible role in treating cancer, spinal cord injuries, and muscle damage, genetic disorders and a number of other diseases.

II. RELATED WORK

iPSCs were first produced in 2006 from mouse cells and in 2007 from human cells in a series of experiments by Shinya Yamanaka's team at Kyoto University, Japan, and by James Thomson's team at the University of Wisconsin-Madison. In 2006, Yamanaka proved that introduction of a small set of transcription factors into a differentiated cell was sufficient to revert the cell to a pluripotent state. The

resulting cells were called induced pluripotent stem cells (iPSCs). Yamanaka selected a set of 24 transcription factors among a large number of transcription factors that were expressed in ES cell [10-13]. In further experiment all 24 genes encoding these transcription factors were introduced in skin fibroblast of mouse. The generated colonies shown remarkable resemblance to ES cells and a combination of only four transcription factors (Myc, Oct3/4,Sox2 and Klf4) were found sufficient to convert mouse embryonic fibroblasts to pluripotent stem cells[14].

In 2007, Yamanaka'sand James Thomson's laboratories were the first to produce human iPS cells. Yamanaka used the four factor (Myc, Oct4, Sox2 and Klf4) whereas another group identified partially overlapping combination of reprogramming factors such as Oct4, Sox2, Nanog, and Lin-28. Therefore, there are six common reprogramming factors (OCT4, SOX2, KLF4, C-MYC, NANOG, and LIN28) which are widely used for generating the iPS cells [15]. These iPS cells are morphologically and phenotypically similar to embryonic stem (ES) cells and thus offer exciting possibilities in stem cell research and regenerative medicine**.** Moreover, iPS cells are useful tools for studying the pathogenesis of human disease**,** for drug discovery and toxicity screening [16, 17].

Stem cell driven regenerative systems are highly complex and dynamic, consisting of large numbers of different cells expressing many molecules controlling their fates. Therefore, mathematical models and computational tools are necessary - both to aid the interpretation of experimental data and to simulate the behavior of stem cell systems based on hypothetical assumptions. This problem can be resolved by using bioinformatics tools and algorithms.

III. METHODOLOGY

Methodology

Protein retrieval and sequence analysis:

The primary sequence of the **POU domain, class 5, transcription factor 1** (Accession No.Q01860) of Homo sapiens was analyzed from the EXPASY public domain protein database and National Center for Biotechnology Information (NP_002692) [18].The OCT 4 protein sequence was retrieved in FASTA format and used for further analysis.

Model building

BLASTP search with default parameters against the Protein data bank (PDB) was used to find the best suitable templates for homology modeling [19]. Multiple sequence alignment was performed between selected template and target sequences using TCOFFEE tool [20].MODELLER program [21] predicted three dimensional structure of OCT4 by using backbone of the selected template.

Model evaluation

Structural Validation of the tertiary structure of OCT4 protein was done by ProSA-web [22, 23] Z-scores and Procheck Ramachandran plot [24]. Among five model predicted by MODELLER energy minimization was performed by GROMOS96 force field.

Calculation of Highly Conserved Amino Acids

The conservation patterns of COMT using ConSurf server [25, 26] has been developed. The conservation scores at each amino acid position were calculated using the same web server. This server can calculate the evolutionary conservation of amino acid positions in proteins using an empirical Bayesian inference, starting from protein structure and sequence.

Generation of Surface Cavity

The PyMOL [27] has been used to for generation of surface cavity as well as identification of binding grooves of OCT4. ".pdb" files were used to generate the surface structure and the cavities of the given protein.

IV. RESULTS AND DISCUSSION

Homology modeling approach works on the conserved residues among target and templtes. Template selection was performed by BLASTp similarity searching program against PDB database. Templates are selected on the basis of identity, E value and other parameters. Table 1 shows the parameters of best five templates that producing significant alignment. PDB ID 3L1P_A was selected as a template to predict tertiary structure of OCT4 protein.

Table 1.

Summary of Sequences producing significant alignment				
Accession	***Identity***	***E value***	***Query coverage***	***Total score***
3L1P_A	88%	6e-96	***42%***	***282***
2XSD_C	66%	1e-59	***39%***	***190***
1CQT_A	***57%***	***4e-51***	***39%***	***168***
1OCT_C	***57%***	***5e-51***	***39%***	***167***
1O4X_A	***56%***	***1e-50***	***39%***	***167***

To find conserve residues between target and selected template Multiple sequence alignment was performed using TCOFFEE server .Alignment result predicted that region of **134-290** amino acids of the target sequence shown conserve residue [Fig 1]. Homology modeling approach used this sequence alignment and structure of template protein for predicting backbone of the OCT4 protein by MODELLER.

Figure 1. Multiple sequence alignment between target and selected template

Table 2.

Summary of successfully produced models by MODELLER		
Filename	***Molpdf***	***Dope score***
POU5F.B99990001.pdb	1262.44336	-17633.31445
POU5F.B99990001.pdb	219.14990	17987.83984
POU5F.B99990001.pdb	1299.38696	17494.85352
POU5F.B99990001.pdb	1245.15808	17991.05664
POU5F.B99990001.pdb	1265.86719	18090.18555

On the basis of DOPE (**D**iscrete **O**ptimized **P**rotein **E**nergy) score POU5F.B99990001 was selected the best one. DOPE is a statistical parameter that accounts for the finite and spherical shape of the native structures. It is used to assess the quality of predicted protein structure [28]. The models returning the minimum molpdfs can be chosen as best probable structures and can be further used for evaluating with the DOPE score. The PyMOL Molecular Graphics System was used to analyse and visualization of the predicted protein.

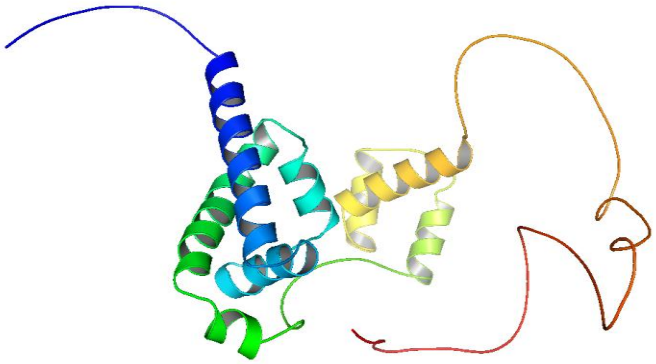
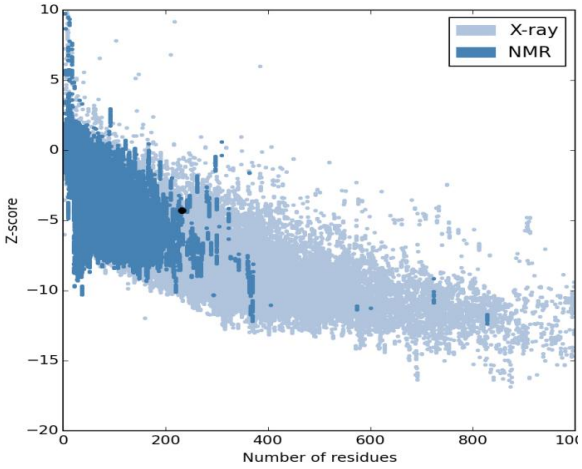


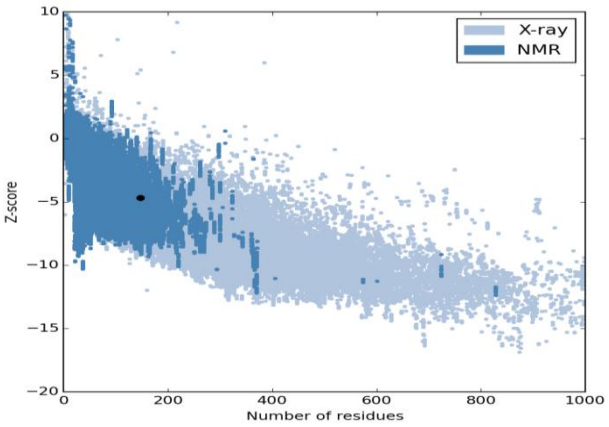
Figure 2. Predicted 3D structure of OCT4 PROTEIN from amino acid 126 to 290

Model evaluation

The Z-score is indicative of overall model quality and is used to check whether the input structure is within the range of scores typically found for native proteins of similar size. PROSAweb was used to find the Z-score of template and query. Z-score of query protein was -4.29 and Z-score of template was 4.68. Procheck checks the stereochemical quality of a protein structure by analyzing residue-by-residue geometry and overall structure geometry. This tool was used to determine the Ramachandran plot to assure the quality of the model. The result of the Ramachandran plot showed 93.4% of residues in favorable region representing that it is a reliable and good quality model (Fig. 7). A model having more than 90% residues in favorable region is considered as good quality model.



(a)



(b)

Figure. 6 Z-score of (a) query and (b) template protein using PROSA web

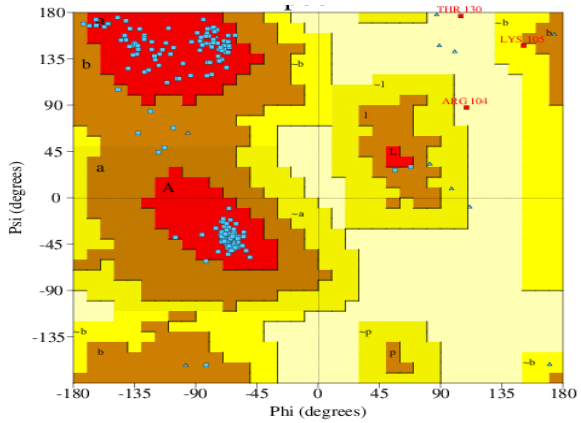


Figure 7.Ramachandran Plot of predicted 3D structure of OCT 4 protein structure

Table 3.

	Ramachandran Plot statistics predicted 3D structure of OCT 4 protein structure	
	No. of residues	**Percentage**
Most favoured regions [A,B,L]	185	***93.4***
Additional allowed regions [a,b,l,p	10	***5.1***
Generously allowed regions [~a,~b,~l,~p]	1	***0.5***
Disallowed regions [XX]	2	***1.0***
Non-glycine and non-proline residues	198	***100***

On analyzing the predicted three dimensional model of the OCT4 revealed that loop **291-360** is disordered and does not appear in the PDB structure. Most probably the long active site loop is flexible in the absence of a ligand and could not be seen in the diffraction map [29]. Therefore, we used loop modeling to increase the accuracy of the models. This Knowledge based approach searches the PDB for known loops with endpoints that match the residues between which the loop has to be inserted and simply copies the loop conformation. Table 4 shows the summary of selected templates.

Table 4.

ALIGN PDB CODES	Summary of selected templates for loop modelling			
	Length	**chains**	**Organim**	**UniProtKB ID**
4HC4A	323 residues	2	*E. coli*	P62623
3QK5A	587 residues	2	*(Rat)*	P97612
3DLSA	335 residues	6	*Homo sapiens*	Q96RG2
3L1PA	155 residues	4	*Mus musculus*	P20263

Figure 8. Multiple sequence alignment of target (region 258 to 360 amino acids) and multiple templates

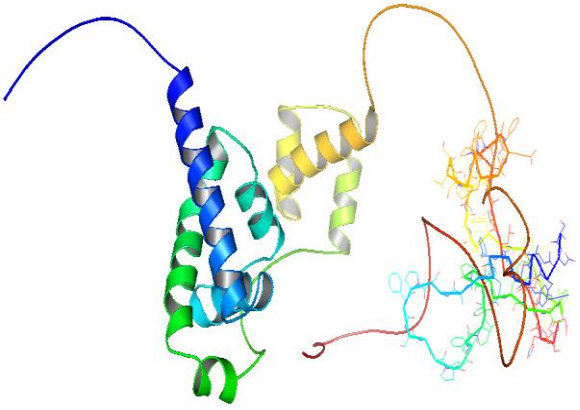


Figure 9. Superimposed structure of modeled OCT 4 protein and modeled loop region

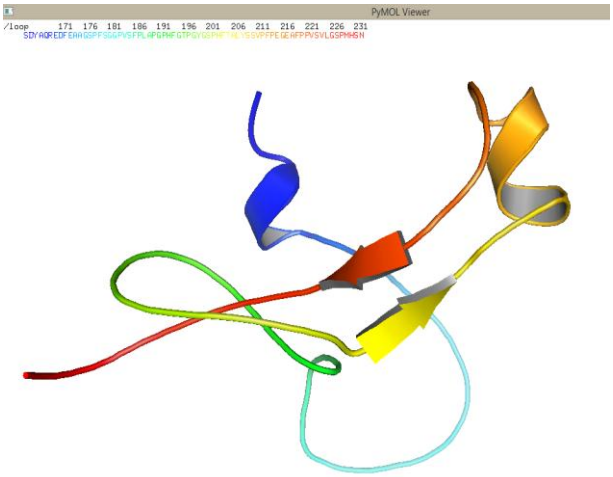


Figure 10. Modelled loop region from 291 to 360

V. CONCLUSION AND FUTURE SCOPE

Our analytical and interaction studies provide a rational molecular platform for initiating *in silico* drug design studies to the designing of new modulators of reprogramming factors of iPS, to guide site-directed mutagenesis. On the other hand, provide OCT4 molecule a raw data with lots of information regarding molecular interactions, for further protein engineering.

Various researches have been conducted on comparative analysis of ES cells and iPS cells that supported resemblance of both type of cells. analytical studied are still remain unpredicted that may be a great help to the future researchers to understand the mechanism(s) as well as path- way of nuclear reprogramming process.

ACKNOWLEDGMENT

I thanks to Chiranjib Chakraborty for suggesting me such a dynamic topic and guide me at each step of my work

REFERENCES

[1]H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, et al, "*The protein data bank*", Nucleic Acids Research, Vol. 28, pp. 235–242, 2000.

[2] M.D. Bethesda, "Stem Cell Information", National Institutes of Health. 2017.

[3] V. Brendel, P. Bucher, I. Nourbakhsh, B. E. Blaisdell, S. Karlin, "Methods and algorithms for statistical analysis of protein sequences". Proceedings of the National Academy of Sciences of the United States of America, Vol. 89, pp. 2002–2006, 1992.

[4] W. L. DeLano, The PyMOL molecular graphics System, 2002.

[5] N. Eswar, M.A. Marti-Renom, B. Webb, M.S. Madhusudhan, D. Eramian, et al. "*Comparative Protein Structure Modeling with MODELLER*", Current Protocols in Bioinformatics, Vol .15, pp. 5.6.1-5.6.30, 2006.

[6] A. Fiser, R.K. Do, A. Sali, 2000. "*Modeling of loops in protein structures*", Protein Science. Vol. 9. pp. 1753-1773, 2000.

[7] E. Gasteiger, C. Hoogland, A. Gattiker, S. Duvaud, M.R.Wilkins, A. Bairoch, 2005."*Protein Identification and Analysis Tools on the ExPASy Server;*(In) John M. Walker (ed)", The Proteomics Protocols Handbook, Humana Press. pp. 571-607, 2005.

[8] K. Guruprasad, B.V.P. Reddy, M.W. Pandit, 1990. "*Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting in vivo stability of a protein from its primary sequence*". Prot. Eng, Vol. 4, pp. 155-164, 1990.

[9] J. C. Hisstock, A.M. Lesk, "*Prediction of protein function from sequence and structure*". Q Rev Biophys, Vol. 36, pp. 307-340, 2003.

[10] K. Takahashi, et al. "*Induction of pluripotent stem cells from adult human fibroblasts by defined factors*", Cell , Vol. 131, pp. 861–72, 2007.

[11] K. Takahashi, S. Yamanaka, "*Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors*", Cell Vol. 126, pp. 663–76, 2006.

[12] H. Zaehres, H .R. Scholer, "*Induction of pluripotency: from mouse to human".* Cell, Vol. 131, pp. 834–5. 2007.

[13] M. Tada, Y. Takahama, K. Abe, N. Namatsuji, T. Tada, "*Nuclear reprogramming of somatics cells by in vitro hybridization with ES cells*", Current biology, Vol. 11, pp. 1553-1558. 2001.

[14] J. Yu, M.A. Vodyanik, K. Smuga-Otto,et al, "*Induced pluripotent stem cell lines derived from human somatic cells*", Science, Vol. 5858, pp.1917–20, 2007.

[15] K. Okita, T. Ichisaka, S. Yamanaka, "*Generation of germline--competent induced pluripotent stem cells*". Nature, Vol. 448, pp.313-317, 2007.

[16] N. Yokoo, et al, "*The effects of cardiovactive drugs on cardiomyocytes derived from human induced pluripotent stem cells*". Biochem Biophys Res Commun , Vol. 387, pp.482-88, 2009.

[17] Q. Lian, et al, "*Future perspective of induced pluripotent stem cells for diagnosis, drug screening and treatment of human diseases*", Thromb Haemost, Vol. 104, pp.39-44, 2010.

[18] E. A. Kimbrel, R. Lanza, "Pluripotent stem cells: The last 10 years", *Regenerative Medicine*, Vol. 11, pp. 831–847, 2016.

[19] Notredame, Higgins, Heringa," *T-Coffee: A novel method for multiple sequence alignments",* JMB, Vol. 302, pp. 205-217, 2000.
[20] B.G. Rost, B.G. Yachdav, J. Liu,"*The PredictProtein Server*", Nucleic Acids Research, 32(Web Server issue), W 321-W326.2004.
[21] E. W. Sayers, T. Barrett, D.A. Benson, E. Bolton, S.H. Bryant, et al. "*Database resources of the national center for bio- technology information*", Nucleic Acids Research, Vol. 39, pp. 38–51, 2011.
[22]M.Y. Shen, A. Sali, "*Statistical potential for assessment and prediction of protein structures*", 2006.
[23] M.J. Sippl, "*Recognition of Errors in Three-Dimensional Structures of Proteins*", Proteins, Vol. 17, pp. 355-362, 1993.
[24] M. Stadtfeld, K. Hochedlinger, "Induced *pluripotency: history, mechanisms, and applications*", Genes Dev, Vol. 4, pp. 2239–2263, 2010.
[25] H. Ashkenazy, E. Erez, E. Martz, T. Pupko, N. Ben-Tal, Nucleic Acids Research, Vol. 38, W529– W533. 2010.
[26] F. Glaser, T. Pupko, I. Paz, R. E. Bell, D. Bechor-Shental, E. Martz, N. Ben-Tal, Bioinformatics, Vol. 19, pp. 163–164, 2003.
[27] W.L. DeLano, "The PyMOL molecular graphics System", 2002.
[28] M.Y.Shen and A. Sali, "*Statistical Potential For Assessment and prediction of Protein Structures*", Protein Sci, Vol. 15, pp. 2507–2524.2006.
[29] M. Wiederstein, J. Sippl, "*ProSA-web: Interactive Web Service for the Recognition of Errors in Three-dimensional Structures of Proteins*", Nucleic Acids Research, Vol. 35, W407-W410. 200

AUTHORS PROFILE

Mrs. P. Chauhan pursed B.Sc., M.Sc. and M.Phil. Bioinformatics. She did her projects work from IGIB (Delhi), SVBP Univ of Agric. & Tech (Meerut) and research scholar of Mewar University (Rajasthan). She is currently working as PGT- Biology in Delhi Public School (Muzaffarnagar). She has published various research papers in reputed international journals, presented posters in different conferences. Her main research work focuses on protein modelling, docking, proteomic and Genomic analysis through Bioinformatics algorithms. She has 13 years of teaching experience.

