Young Scientist- Tomorrow's Science Begins Today
Vol. 5, Issue 1 – 2021
ISSN: 2581-4737

# Comparative Analysis among Six Proteins OCT4, SOX2, KLF4, C-MYC, NANOG, and LIN28 using Bio Info Tools

**Purnima[1], Saurabh Mishra[2]**
[1]*Department of Biotechnology, Mewar University, Gangrar, Chittorgarh, Rajasthan, India.*
[2]*Department of Biotechnology, Mewar University, Chittorgarh, Rajasthan, India.*

## Abstract

Now, it is possible to treat various fatal diseases using iPSCs. Take damaged nerve tissue as an example; in that case, medical doctors might remove healthy skin cells from the patient. These skin cells are reprogrammed into the healthy nerve cell. The developed nerve cells would be transferred into the affected part of the body. This approach is helpful in the treatment and healing of diseases such as Parkinson's disease.

Research is going on to find molecules that can either replace or catalyze the functionality of transcription factors and other regulators in iPS cell induction. While different aspects and other properties of stem cells are a mystery for researchers, reprogramming factors provide a new path for the research of therapeutic agents. With the exponential growth in the generation of iPSC in many ways, there is an urgent requirement for a cost-effective, animal-free alternative.

The analytical parameters of the reprogramming factors and the interaction studies are not well known. The question we raised here is whether any relationship exists between all the parameters and functionality of the selected transcription factors. So, this research paper is centralized on solving the given problem. Modern high-throughput molecular technologies can sorta set of gene products simultaneously. The present research paper used Bio Info Tools to discuss the detailed comparative analysis among six proteins "OCT4, SOX2, KLF4, C-MYC, NANOG, and LIN28".

**Keywords:** Stem cells, Induced Pluripotent Stem Cells (iPSCs),OCT4, SOX2, KLF4, C-MYC, NANOG, and LIN28.

## Introduction

Stem cells can divide into differentkinds of cells with specific functions and produce many similar cells through mitosis in multicellular organisms. Mammals have two kinds of stem cells: adult stem cells, which are typically present in all tissues, and embryonic stem cells, which are located in the inner cell mass of the blastocyst.

Adult stem cells may be isolated from various organs of children and adults, including bone marrow, heart, liver, and other digestive organs. Blood and tissues from the umbilical cord are rich in pluripotent adult stem cells. The most recent method for treating heart disease and liver cirrhosis involves somatic stem cells.

In 2012, Shinya Yamanaka received the Physiology/Medicine Nobel Prize for discovering induced pluripotent stem cells. He identified a new method to modify mature cells into specialized cells with the potential of pluripotency. These reprogrammed cells behaved as ES cells and could produce varieties of cells required for the proper functioning of the body. The generated reprogrammed cells using different regulators are"induced pluripotent stem cells (iPSCs)."

## Induced Pluripotent Stem Cells (iPSCs)

Creating induced pluripotent stem cells offered a novel method for transforming adult Somatic cells into pluripotent stem cells. This could be possible by introducing a group of regulators involved in the transcription of genes and the desired expression. Reprogramming was introduced in 2006 by Yamanaka's group and changed the prospectus of researchers towards the working of cells.

This research develops a platform for molecular biologists to study various diseases and their treatments. Differentiated somatic cells of the body are reversed back to the cell with pluripotent potential by reprogramming. In the past, a series of experiments were conducted to successfully produce a clone of organisms through the transfer of nuclear material into somatic cells.

In iPSC research, transcription factors are added to somatic body cells to generate pluripotent stem cells. Although the resulting reprogrammed cells are not exactly like natural stem cells, they do have the potential to generate new pluripotent cells. The selected transcription factors made it possible to derive induced pluripotent stem cells.

These transcription factors are biochemically proteins and regulate the process genetically to produce embryonic stem cells like pluripotent cells. At Kyoto University, Shinya Yamanaka and his colleagues conducted the first experiment using induced pluripotent stem cells. The transcription factors "Oct3/4, Sox2, c-Myc, and Klf4" were employed to reprogram the murine fibroblast cells into pluripotent cells.

Similar combinations of transcription factors were used in a previous attempt to create iPSCs in human fibroblasts. Junying Yu did the researchwith James Thomson and his team at the University of Wisconsin-Madison using the transcription factors "Nanog, Lin28, Oct4, and Sox2".They got successful in producing reprogrammed cells from the skin of a human.

This research opens a new door to generating pluripotent cells without the risk of issues associated with human embryos. This approach is useful in treating specific patients according to their cell types.

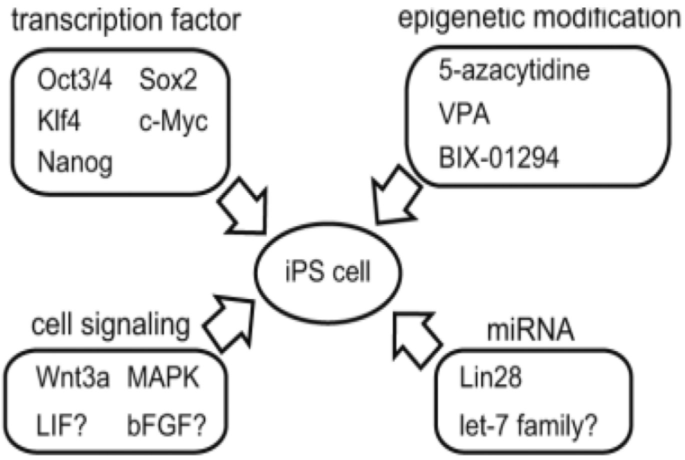

**Figure 1: Factors influence iPS cell induction**

Direct reprogramming was conducted by various teams who used an overlapping combination of "Oct4, Sox2, Nanog, Klf4, c-Myc, and Lin28", suggesting that Oct4 and Sox2 are the master regulators. At the same time, the other four transcription factors are co-regulators of reprogramming the cell.

## Review of Literature

During the developmental stage, stem cells can differentiate into various cell types. In the body of any living organism, they repair the damaged system through the capacity of pluripotency without affecting the other healthy cells or tissues. On dividing stem cells, the newly generated cells may behave the same as stem cells or differentiate into a specific type of cell [1, 2, and 3].

These are artificially synthesized in the laboratory and used to produce different cell types. Naturally, pluripotency is the characteristic of stem cells, but now- a day's any immature somatic cell in our body can be converted into a pluripotent cell through reprogramming. The Broad Stem Cell Research Center at the University of California, Los Angeles team produced induced pluripotent stem cells for the first time in 2007. The BSCRC team that created iPSC included Drs. April Pyle, William Lowry, Amander Clark, and Kathrin Plath [5].

In 2006, the Kyoto University team led by Shinya Yamanaka produced induced pluripotent stem cells from mice for the first time. After one year, James Thomson and his team atthe University of Wisconsin-Madison conducted the same experiment successfully with human cells. In 2006, Yamanaka selected a group of protein regulators in his experiment and proved that these were able to produce pluripotency in somatic cells. These developed resulting cells were termed as iPSCs. From a wide variety of regulators expressed in ES cells, Yamanaka chose a set of 24 transcription factors [6-9].

In subsequent investigations, genes that encode the chosen transcription factors were introduced into mouse skin fibroblasts. A combination of Myc, Sox2, Klf4, and Oct3/4 transcription factors was discovered to be sufficient to transform mouse embryonic fibroblasts into pluripotent stem cells. The resulting colonies exhibited a striking similarity to ES cells [10]. In 2007, Yamanaka and James Thomson's team was credited with creating the first human-induced pluripotent stem cells. Among the selected transcription factors, Oct4 & Sox2 are common by all research groups, except Yamanaka used Myc& Klf4, whereas another group introduced Nanog & Lin28.

The abovementioned six transcriptions were used in different sets to produce iPS cells and considered reprogramming factors [11]. The produced iPS cells are genetically similar to embryonic stem cells, as both share the same structure and characteristics. This invention provides a major tool in designing individual genetically specific medicine and research associated with stem cells. In addition, iPS cells offer an advanced method to diagnose the pathogenic cause of human disease and their study for drug discovery and toxicity screening [12].

## Research Methodology

Researchers have retrieved sequence data of the human protein LIN28 & other factors from the "National Center for Biotechnology Information (NCBI)" server. They found different homologous forms of the same protein; homolog B was selected to analyze further. The researcher has chosen the standard FASTA format to access the amino acid sequences of reprogramming factors by the data retrieval tool of NCBI.

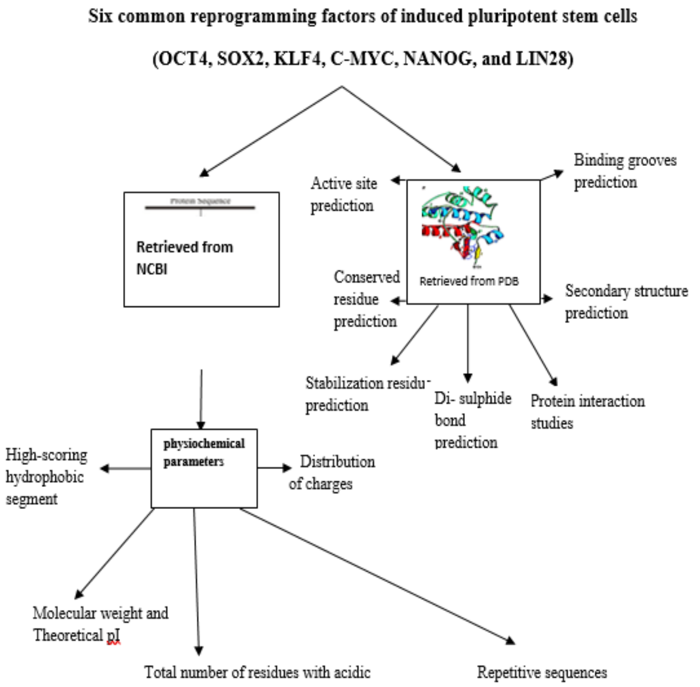
Six common reprogramming factors of induced pluripotent stem cells
(OCT4, SOX2, KLF4, C-MYC, NANOG, and LIN28)
Retrieved from NCBI
Active site prediction
Binding grooves prediction
Retrieved from PDB
Conserved residue prediction
Secondary structure prediction
Stabilization residue prediction
Di- sulphide bond prediction
Protein interaction studies
physiochemical parameters
High-scoring hydrophobic segment
Distribution of charges
Molecular weight and Theoretical pI
Total number of residues with acidic
Repetitive sequences

**Figure 2: Six common reprogramming factors of induced pluripotent stem cells ("OCT4, SOX2, KLF4, C-MYC, NANOG, and LIN28")**

The researcher used PROTPARAM to calculate physiochemical parameters like Theoretical isoelectric point, Total number of atoms, Total number of acidic (Aspartate & Glutamate) and essential amino acids (Arginine, Lysine & Histidine) atomic composition, Instability point, and Grand average of hydropathicity [2]. One can enter the amino acid sequence in the input box or access parameters through the UniProt database accession number.

The SAPS tool was used to analyze the number & type of amino acids, the nonpolar nature of residues, and repetitive segments within the protein sequence.

To explore the secondary structure aspects, ‘‘SCRATCH protein predictor’’ can be used to analyze the protein structure using different computational methods.There is a module Dipro to find the locations and number of disulphide bonds in the given sequence.

The hydrophobicity of any protein depends upon the number and nature of amino acids. Three-dimensional protein structure stabilizes by specific folds produced due to the inner core structure of hydrophobic residues. ProtScale is used to find the hydrophobicity of reprogramming proteins in the proteomic analysis [4].ProtScale offers a variety of algorithms to calculate hydrophobic regions; we have selected Kyte& Doolittle method in our work. The output will be produced as an X-Y plot that interprets hydrophobicity through high signals.

PDB is the repository of three-dimensional structures of protein modeledby Nuclear magnetic resonance technique and x-ray crystallography. The secondary structure comprises alpha sheets, beta strands, and others.We can access PDB's structural, sequence, literature, and other protein annotations. PDBsum database is a brief description of the PDB. We accessed the secondary structure through PDB by entering the code in the search box of the PDBsum home page [13, 14].

The SCide server was used to find stabilization centers, and the SRide program was used to find the stabilizing residues. The server can predict residues responsible for stabilizing the three-dimensional structure of proteins. These Stabilizing residues can be predicted by calculating the hydrophobic nature of side chains and defining a score based on the conserved residues. In the program default cutoff value of the conservation,the score is taken as C6. These residues are also helpful in collecting information about the family of the given sequences.

ConSurf tool is used to find conserved amino acid residue patterns. The server calculates the score by creating position specific matrix. Based onthe score, predicted conserved residues were used for evolutionary studies. This analysis was based on the amino acid sequence and the protein's three-dimensional structure.

The PyMOL has been used to present the binding sites on the surface of reprogramming factors of iPSCs. Protein structure is imported into the tool in “.pdb” files to predict binding sites. The uploaded structure was accessed through the surface parameter of the popup menu S (show).This structure may be processed for docking any ligand at the active site of given proteins to study molecule-ligand interactions that would be helpful in drug design.

If the tertiary structure of a given protein is not present in PDB, then protein modeling is required to study structural topology. We have used a homology modeling approach to predict protein structure. The primary requirement of this method is selectinga template (known protein structure) for the target (query) sequence. Modeller 9.11 program is used to predict three-dimensional structures of the reprogramming factors.

## Results and Discussion

First, the Amino acid sequences were retrieved in the FASTA format through the NCBI Entrez tool.

Amino acid squencence analysis of selected reprogramming factors was performed through the ProtParam tool, and the results are shown in Table 1

**Table 1: Physiochemical parameters of reprogramming factors**

| **Reprogramming Factors** | **No. of amino acids** | **Molecular weight (Dalton)** | **Theoretical pI** | **Total No. of (Asp + Glu)** | **Total No. of (Arg + Lys)** | **The instability index (II)** | **Aliphatic index** | **Grand average of hydropathicity (GRAVY)** |
|---|---|---|---|---|---|---|---|---|
| LIN28 | 250 | 27083.6 | 9.15 | 28 | 38 | 79.67 | 45.64 | -0.890 |
| OCT4 | 360 | 38570.61 | 5.69 | 38 | 33 | 53.24 | 66.61 | -0.435 |
| SOX2 | 317 | 34309.82 | 9.74 | 21 | 34 | 58.73 | 48.71 | -0.742 |
| C Myc | 439 | 48804.08 | 5.33 | 64 | 51 | 92.23 | 66.42 | -0.772 |
| NANOG isoform 1 | 305 | 34619.57 | 6.32 | 25 | 24 | 66.12 | 49.87 | -0.844 |
| NANOG isoform 2 | 289 | 32837.63 | 5.79 | 25 | 23 | 66.37 | 50.59 | -0.841 |
| KLF4 | 513 | 54670.54 | 8.69 | 43 | 50 | 67.27 | 56.74 | -0.565 |

**Charge Distribution Prediction:** Amino acid sequence-based analysis was performed through the ProtScale tool. The composition of proteins in terms of Carbon, Hydrogen, Nitrogen, Oxygen, and Sulphurwas calculated. A few sets of amino acids are repeated in the sequences and create complexity; such repeats are called Tandem and Periodic repeats.

Table 2: Atomic composition and repeats of the reprogramming factors

| Reprogramming Factors | Atomic Formula | Total no. of atoms | Atomic composition | Tandem and periodic repeats |
|---|---|---|---|---|
| OCT4 | $C_{1718}H_{2657}N_{469}O_{517}S_{13}$ | 5374 | Carbon 1718<br>Hydrogen 2657<br>Nitrogen 469<br>Oxygen 517<br>Sulfur 13 | Aligned matching blocks:<br>[14-17] PPGG<br>[41-44] PPGG<br>[22-25] PGGP<br>[42-45] PGGP<br>[95- 99] PEGEA<br>[340-344]PEGEA |
| LIN28 | $C_{1161}H_{1859}N_{353}O_{363}S_{16}$ | 3752 | Carbon 1161<br>Hydrogen1859<br>Nitrogen 353<br>Oxygen 363<br>Sulfur 16 | Aligned matching blocks:<br>[ 120- 123] QKRK<br>[ 245- 248] QKRK<br>[ 207- 210] PQEA<br>[ 222- 225] PQEA |
| SOX2 | $C_{1467}H_{2321}N_{443}O_{457}S_{26}$ | 4714 | Carbon 1467<br>Hydrogen 2321<br>Nitrogen 443<br>Oxygen 457<br>Sulfur 26 | Aligned matching blocks:<br>[ 22- 25] GGNS<br>[ 135- 138] GGNS<br>[ 16- 23] noossss<br>[ 24- 31] noossss<br>n= NQ;0=ST;s=AG |
| C Myc | $C_{2107}H_{3330}N_{604}O_{702}S_{14}$ | 6757 | Carbon 2107<br>Hydrogen 3330<br>Nitrogen 604<br>Oxygen 702<br>Sulfur 14 | Aligned matching blocks:<br>[ 159- 162] SGSP<br>[ 279- 282] SGSP<br>[ 237- 240] PLVL<br>[ 294- 297] PLVL |
| NANOG isoform 1 | $C_{1503}H_{2303}N_{415}O_{485}S_{21}$ | 4727 | Carbon 1503<br>Hydrogen 2303<br>Nitrogen 415<br>Oxygen 485<br>Sulfur 21 | Aligned matching blocks:<br>[35- 42] YPSLQMSS<br>[174- 180] YPSL_YSS<br>[ 196- 200] WSNQT<br>[ 206- 210] WSNQT<br>[ 263- 266] LEAA<br>[ 267- 270] LEAA |
| NANOG | $C_{1421}H_{2187}N_{395}O_{460}S_{21}$ | 4484 | Carbon 1421<br>Hydrogen 2187 | [ 180- 184] WSNQT<br>[ 190- 194] WSNQT |

| | | | | |
|---|---|---|---|---|
| isoform 2 | | | Nitrogen 395<br>Oxygen 460<br>Sulfur 21 | [ 247- 250] LEAA<br>[ 251- 254] LEAA |
| KLF4 | $C_{2399}H_{3707}N_{703}O_{719}S_{24}$ | 7552 | Carbon 2399<br>Hydrogen 3707<br>Nitrogen 703<br>Oxygen 719<br>Sulfur 24 | There are no high scoring hydrophobic/ transmembrane segments. |

## Disulphide bond topology

Disulphide bonds are the type of covalent bond between two sulfur-containing amino acid Cys that provides more stability to the protein. Intramolecular disulphide bonds within polypeptide chains are the major factor responsible for stabilizing any protein. Intermolecular disulphides between polypeptides provide stability to the quaternary structure of the protein.

**Table 3: Positions of di-sulphide bonds in reprogramming factors**

| **Reprogramming Factors (UniProtKB )** | **Total no. of cysteine** | **Predicted number of bonds** | **Disulphide bonds by decreasing the order of probability** | |
|---|---|---|---|---|
| | | | Cysteine position 1 | Cysteine position 2 |
| **Oct 4 (Q01860)** | 9 | 3 | 252<br>63<br>185 | 279<br>70<br>198 |
| **Sox2** (P48431) | The sequence has LESS THAN TWO cysteines and, therefore, cannot form disulfide bonds | | | |
| **Klf4** (O43474) | 9 | 4 | 243<br>7<br>19<br>185 | 251<br>13<br>29<br>27 |
| **Cmyc** (P01106) | 10 | 4 | 25<br>117<br>300<br>171 | 70<br>133<br>342<br>188 |
| **Nanog** (Q9H9S0) | 9 | 4 | 7<br>227<br>19<br>169 | 13<br>235<br>29<br>211 |
| **LIN28** | 9 | 4 | 13<br>164<br>117<br>152 | 44<br>174<br>139<br>161 |

In our study, work has been centralized on the structural organization of polypeptide chains of reprogramming factors, so intramolecular disulphide bonds are being considered. Table 3 shows that OCT4 contains less and SOX2 does not contain any disulphide bond; respectively, these are more reactive and involved in various pathways as a regulator to control the expression of stem cells.

**Hydrophobic Segment:** Hydrophobic amino acids (such as glycine, alanine, valine, leucine, phenylalanine, tryptophan, and methionine) play an indispensable role in the folding of proteins. The hydrophobic effect of proteins maintains their stability, insertion into the nonpolar medium, and folding. The energy needed to maintain folds within the proteins comes from the hydrophobic core region that contains side chains.

We used Kyote&Doolite algorithm to get the hydrophobicity of the selected proteins. The output is in the form of a graph and doesnot produce any statistical score. The graph X-axis shows no. of amino acids, and Y axis presents hydropathy. When interpreting the results, we can only consider strong signals. Figures 3 to 8 represent hydropathy plots of the reprogramming factors.

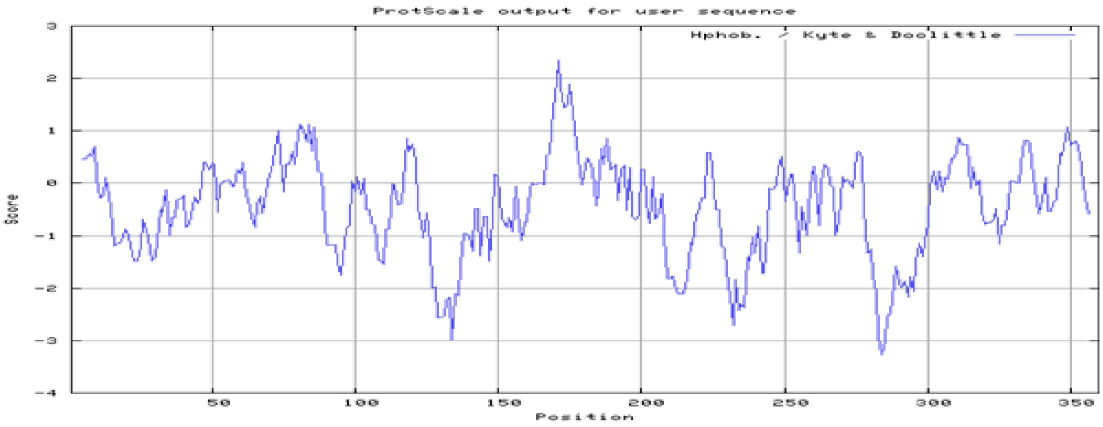

**Figure 3:Hydrophobicity plot of protein LIN28**

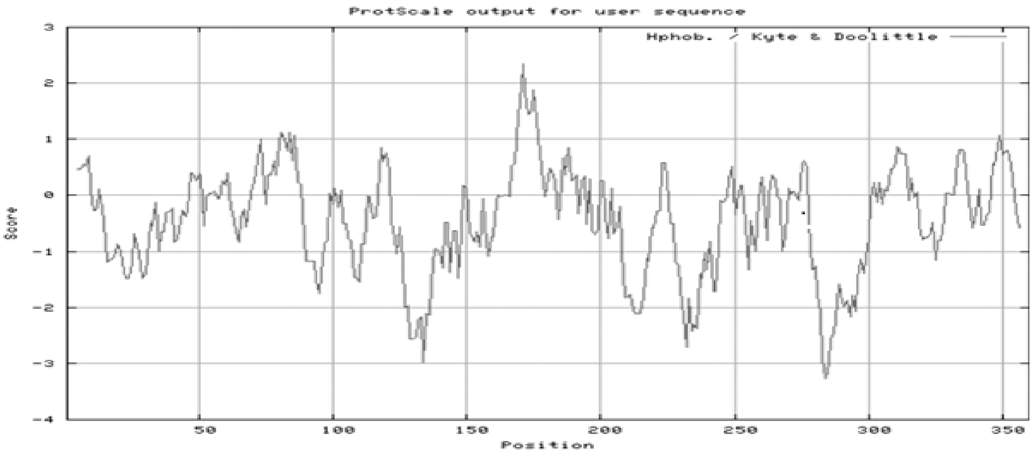

**Figure 4: Hydrophobicity plot of OCT4 protein sequence**

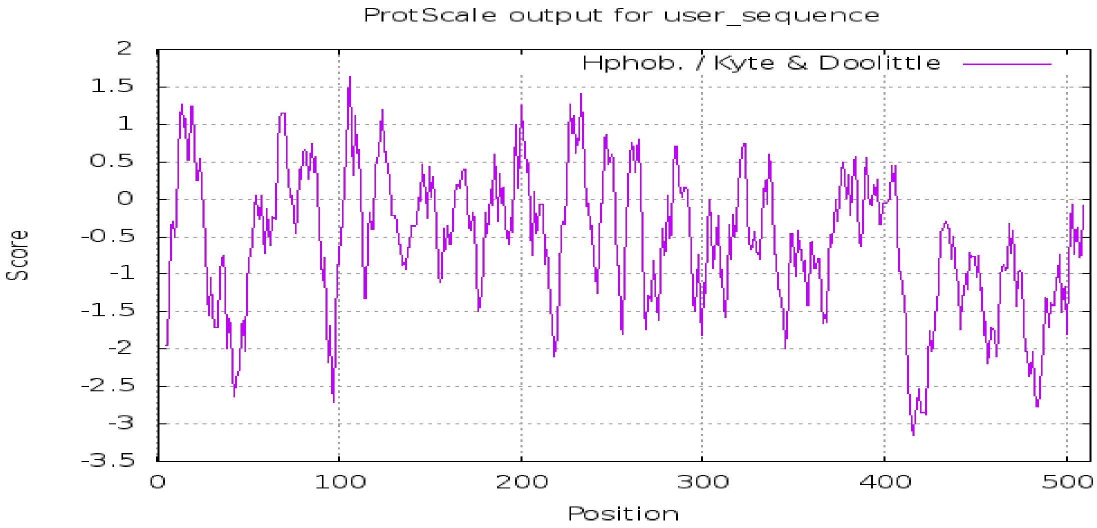

Figure 5: Hydrophobicity plot of the protein KLF4

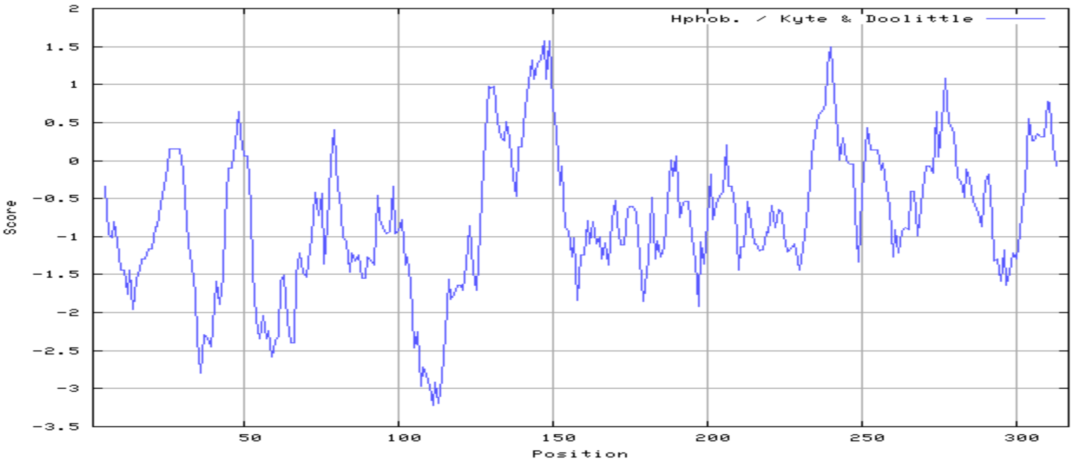

Figure 6: Hydrophobicity plot of SOX2 protein sequence

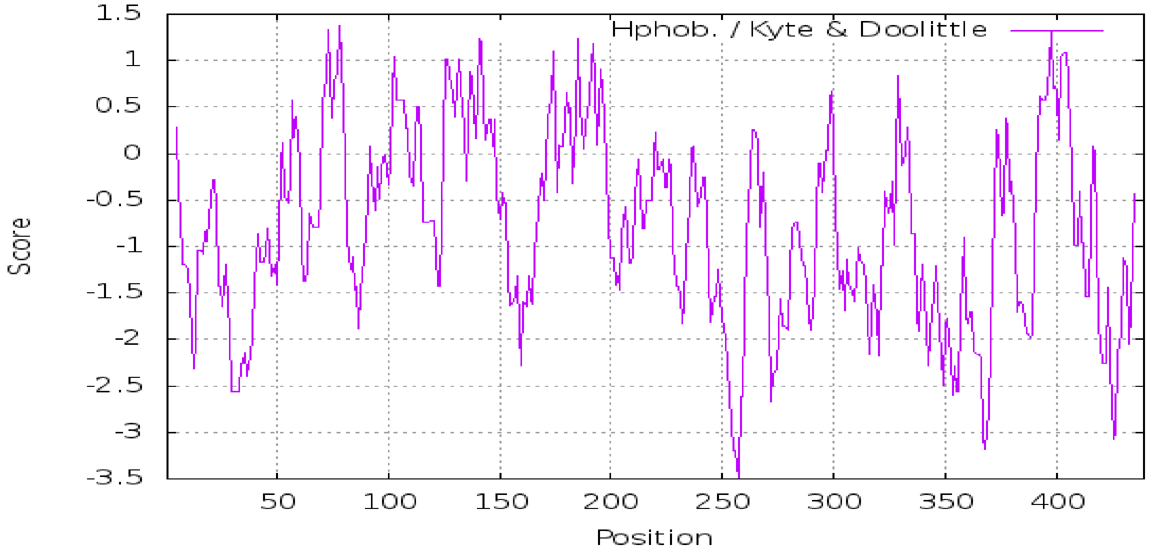

Figure 7:Hydrophobicity plot of the protein C-MYC

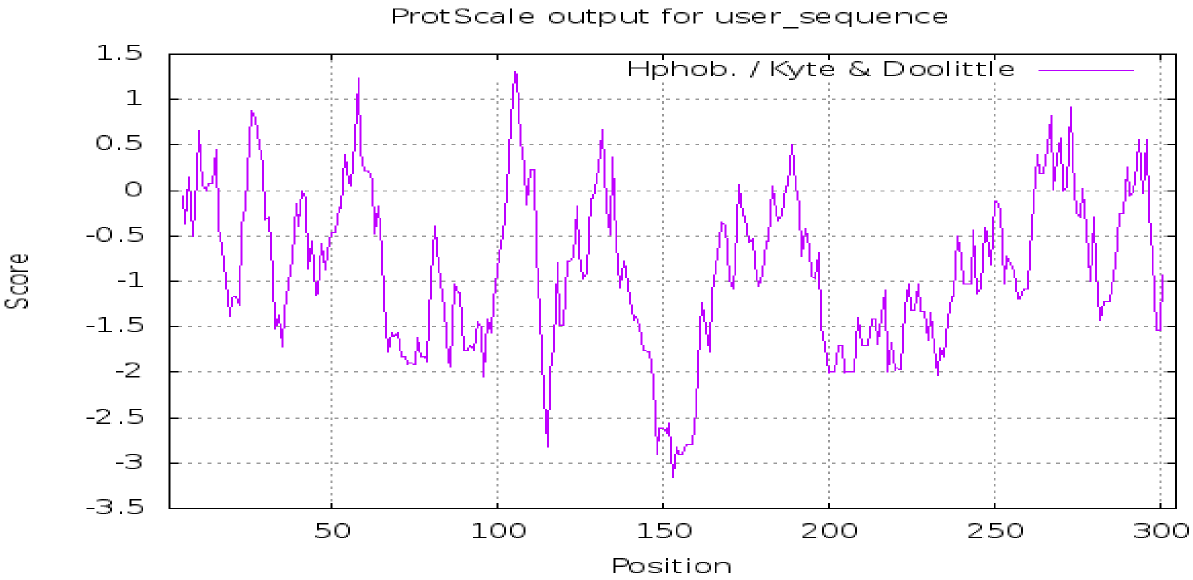

**Figure 8: Hydrophobicity plot NANOG protein sequence**

We observed the plots and found that LIN28 and OCT4 show strong signals at 160-180 amino acids.KLF and SOX2 show more hydrophilicity. CMYC has various hydrophobic regions upto the length. NANOG shows hydrophobic regions at 50-60 and 100-120 amino acids.

This analysis proves that KLF4 and SOX2 proteins show more variations and the least stability in respect of three-dimensional structures.

## Conclusion

In our work, we have used advanced and computational programs to study genomics and proteomics of reprogramming factors associated with induced pluripotency of somatic cells. Major research efforts include protein structure prediction, gene annotations, protein-protein interactions, protein structural analysis, and evolutionary relationships among six master regulators of stem cell development and differentiation.

Our study provides detailed information on selected reprogramming factors which will help designdrugs for specific patient diseases. Adult bone marrow transplantation into the damaged part of the body is the most common practice to replace the injured part. This study might help in controlling embryonic stem cells of pancreatic islets to secrete insulin in diabetic patients. Various types of cancer can be treated through injections of stem cells.

Apart from wide applications of stem cells, technical challenges exist, such as the availability of pure stem cell lines without mutations, delivery to a specific site into the body, rejection by the body, and cell proliferation leading to tumors. Ethical issues are also associated with stem cell research, as embryonic stem cells are derived from extra blastocysts developing embryos. The use of embryos for research purposes is not acceptable to society. We have to think that "Is an embryo an organism?"

To overcome these situations, iPSCs are the most promising tool for recovering from various diseases. We can establish iPSC lines for patient-specific diseases by understanding the behavior of tissues to their interacting genes. Human iPSCs have the potential to develop all three primary germ layers and act as stem cell markers.

## References

1. Hans. R, "The Potential of Stem Cells: An Inventory," in Nikolaus Knoepffler; Dagmar Schipanski; Stefan Lorenz Sorgner. Humanbiotechnology as Social Challenge. Ashgate Publishing, 2007, pp. 28.
2. S. Mitalipov and D. Wolf, "Totipotency, pluripotency and nuclear reprogramming," in Adv. Biochem. Eng. Biotechnol, 2009, pp. 185-199.
3. F. Ulloa-Montoya, C.M .Verfaillie, W.S. Hu, "Culture systems for pluripotent stem cells," in J BiosciBioeng, 2005, pp.12-27.
4. M.J.Evans and M.H. Kaufman, "Establishment in culture of pluripotential cells from mouse embryos," in Nature, July 1981, pp. 154-6.
5. 'stem cell researchers reprogram human skin cells into cells with the same properties as embryonic stem cells, Feb. 11, 2008. Available at https://stemcell.ucla.edu/induced-pluripotent-stem-cells.
6. K.Takahashi and S.Yamanaka," Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors," in Cell, Vol 126, 2006, pp. 663-676.
7. H. Zaehres and H.R Scholer, "Induction of pluripotency: from mouse to human," in Cell, Vol 131, 2007, pp. 834-835.
8. G. Amabile and A. Meissner, "Induced pluripotent stem cells: current progress and potential for regenerative medicine," in Trends Mol Med, Vol 15, 2009, pp.59-68.
9. D. Huangfu, K. Osafune, R. Maehr, W. Guo, A. Eijkelenboom, S.Chen, et al." Induction of pluripotent stem cells from primary human fibroblasts with only Oct4 and Sox2," in Nat Biotechnol, Vol 26, 2008, pp. 1269-1275.
10. J. Yu, M.A Vodyanik, K. Smuga-Otto, J. Antosiewicz-Bourget, J.L. Frane, S. Tian S, et al., "Induced pluripotent stem cell lines derived from human somatic cells," in Science, Vol 318, 2007, pp. 1917-1920.
11. M.Nakagawa, M. Koyanagi, K. Tanabe, K. Takahashi, T. Ichisaka, T. Aoi, et al. "Generation of induced pluripotent stem cells without Myc from mouse and human fibroblasts," in NatBiotechnol, Vol 26, 2008, pp. 101-106.
12. R.S. Deshmukh, K.A. Kovács, A. Dinnyés, "Drug Discovery Models and Toxicity Testing Using Embryonic and Induced Pluripotent Stem-Cell-Derived Cardiac and Neuronal Cells," in Stem Cells Int, Vol 2012, May 2012, pp. 379569.
13. Laskowski, R. A. (2009). PDBsum new things. In Nucleic Acids Research, Vol 37, D355-D359.
14. Hutchinson, E. G., & Thornton, J. M. (1990). HERA: a program to draw schematic diagrams of protein secondary structures. Proteins, Vol 8, 203-212.