JOURNAL OF EPIGENETICS, VOL. 01, ISS. 02 (SPRING 2019), 31-39

JOURNAL OF EPIGENETICS

Journal homepage: http://jep.usb.ac.ir/

Document Type: Original Article

Does Protein Similarity of Pluripotency Factors Mean Their Gene Ontology Semantic Similarity?

Reza Majidzadeh Heravi [a,*], Mohsen Qayoomian [b], Morteza Hashemi Attar [a]

[a] *Department of Animal Science, Faculty of agriculture, Ferdowsi university of Mashhad, Mashhad, Iran; Email: rmajidzadeh@um.ac.ir, mh_544@yahoo.com*

[b] *Department of Biology, Faculty of Science, Ferdowsi university of Mashhad, Mashhad, Iran; Email: m.qayoomian@gmail.com*

***** *Corresponding author at: Email Address: rmajidzadeh@um.ac.ir*

ARTICLE INFO

Article history
Received 07 September 2019
Accepted 28 March 2020
Available online 28 March 2020
DOI: 10.22111/jep.2020.31556.1016

KEYWORDS:
Gene ontology, Phylogeny, Pluripotency, Protein, Similarity.

A B S T R A C T

Recognition and prediction of biological function of proteins based on amino acid sequences is a simple method employed in so many software and operators. However, the sequence similarity does not always imply to similarity of biological function. The aim of this study was to determine the semantic similarity of gene ontology (GO) of six pluripotency factors, Oct4, Sox2, C-Myc, Klf-4, Lin28 and Nanog in six species and evaluate their conformity with their protein sequence similarity and phylogenetic distance. C-myc factor exhibited a significant correlation between phylogenetic distance and protein similarity. The other factors like Sox2, Klf-4 and Lin-28 showed the correct changes of phylogenetic distance and protein similarity, but Nanog and Oct4 factors did not display a correct correlation between two indices because, the increase of protein similarity was not followed with the decrease of phylogenetic distance. Following the study, the protein or nucleotide similarity was assumed as dependent variable and GO similarity in three categories of biological process (BP), molecular function (MF) and, cell component (CC) were expected as the independent variables. With this assumption, regression analysis was accomplished to determine the best model for protein and nucleotide similarity estimation. The protein or nucleotide similarity also displayed a significant regression with GO similarity for C-myc factor and category of BP and CC were selected to estimate protein or nucleotide similarity by model, but a significant regression was not observed for other pluripotency factors for estimation of protein or nucleotide similarity. It means that except of C-myc, GO similarity of other studied pluripotency factors didn't reflect the protein or nucleotide similarity. It is suggested that related data for five pluripotency factors, including Oct-4, Sox2, Klf4, Lin28 and Nanog in the six studied species should be reviewed.

* Corresponding author: Tel.: +985138805747.
E-mail address: rmajidzadeh@um.ac.ir

1. Introduction

Stem cells are defined as precursor cells that have the capacity to self-renew and to generate multiple mature cell types (Simara et al., 2013). The potency of stem cells can range from totipotent, which are able to give rise to all of the cells in an organism, including extra embryonic tissues, (e.g. zygote) to unipotent, which are only able to differentiate into one type of cell (e.g. spermatogonia). Pluripotent stem cells exhibited the potential of self-renew by dividing and developing into the three primary germ cells (Simara et al., 2013). Due to their tremendous potential for therapeutic use, research on deriving, expanding and manipulating human pluripotent stem cells, including embryonic stem cells (hESCs) and the related induced pluripotent stem cells (hiPSCs), has grown exponentially.

The pluripotent genes that play main role in reprogramming of somatic cells, are Klf4, C-Myc, Sox2 and Oct4 (pou51) (Takahashi et al., 2007). The products of these genes are pluripotent factors that are not only markers of pluripotent cells, but also actually important for maintaining the pluripotent state (Takahashi et al., 2007). Recent success in reprogramming of somatic cells mostly emphasizes on important and essential role of pluripotency factors. In these studies, cooperation of Klf4, Oct4, Sox2 and C-Myc factors reprogramed mouse embryonic cells and mature fibroblast cells into a pluripotent state (Takahashi et al., 2007; Wernig et al., 2007). These are the same four factors that have proven to be able to reprogram human skin fibroblasts (Takahashi et al., 2007; Park et al., 2008). While another study showed that Nanog, Sox2, Oct4 and Lin28 factors are enough to induce pluripotency from human somatic cells (Yu et al., 2007).

Gene ontology (GO) project started to organize and characterize a large amount of biological data aiming the simple task of calculating a rational relationship and any possible linkage among them (Ashburner et al., 2000; Consortium, 2004). GO was known as a standard and reliable tool for interpreting gene products in different databases. In recent years, GO terms were used in description of different protein functions in databases like NOPdb which is about nuclear protein or SCCOPPI which is about relationship of domain in a protein molecule or MolMovDB which is about macromolecule motility (Echols et al., 2003; Leung et al., 2006; Winter et al., 2006). In addition, the GO terms were successfully applied in large scale evaluation of protein like Swiss-Prot, TrEMBL and InterPro.

Based on sequence similarity method, the function of a protein can be estimated based on the similarity of its sequence to those have been functionally known in the database. In this method, it is assumed that sequence similarity is equal to the biological function similarity. Since this hypothesis corresponds to reality and BLAST method is a simple and popular way, it is widely used by scientists. Although BLAST of sequences does not always imply to similarity of biological function between interested proteins (Šali, 1999; Gerlt and Babbitt, 2000). Zhong-Hui et al. (2006) showed protein pairs into a GO group exhibited more sequence similarity than protein pairs selected out of a GO group or randomized ones (Duan et al., 2006). They also declared that similarity of sequences can be used as a key index for prediction of protein function (Duan et al., 2006).

Numerous studies have been conducted to determine the biological function of a protein based on the similarity of its sequence to known proteins, which showed good reliability in many cases (Šali, 1999; Gerlt and Babbitt, 2000). Determination of biological functional similarity was also calculable based on GO terms after organizing the GO terms, so in most cases the results were consistent with the facts, although there were conflicting cases too (depending on the type of protein studied) (Duan et al., 2006). Among the conducted studies, no relationships between the GO similarity and the sequence similarity of proteins were defined. The aim of this study was to determine the GO term similarity of six pluripotent factors of Oct4, Sox2, C-Myc, Klf4, Lin28 and Nanog in six animal species, then the similarity indices obtained in GO were compared with the index of sequence similarity and the phylogenetic distance. Finally, it was evaluated whether the GO term similarity follows the sequence similarity that assumed as standard in this research.

2. Materials and Methods

2.1. Animal Species and Protein

Six pluripotent factors were studied in 5 animal species and human (Homo sapiens). Animal species were included cow (Bos taurus), sheep (Ovis oris), mouse (Mus musculus), rat (Rattus norvegicus) and pig (Sus scrufa). Six pluripotent factors including C-

Myc, Sox2, Oct4, Klf-4 and Nanog that were considered in each organism. Their protein sequences were searched and extracted from Swiss-Prot database. The confirmed sequences were selected from mentioned database and used in this study. The Accession number of each protein was shown in Table 1.

2.2. Phylogenetic Relationship between Pluripotent Factors in 6 Species

To know evolution distance based on protein sequences, the phylogenic relationship was determined for each pluripotent factor for each species using MEGA6 software (Tamura et al., 2011). The results were shown as tree diagram. Phylogenic distance was also determined as pair comparison for each factor. Drawing of tree diagram has been done by use of the method UPGMA. Estimation of pairwise distance was calculated based on Poisson model.

2.3. Semantic Similarity of GO between Pluripotent Factors

The semantic similarity of GO was determined between pluripotent factors using web ProteInOn software (Faria et al., 2007). This software is used for searching and comparing of GO terms in protein molecules and it implements the several scale of semantic similarities for calculation of protein and similarities. To characterize of proteins, this software is also able to combine the data of protein-protein reactions with GO terms.

In order to estimating GO similarities between pluripotency factors, the pairwise comparison was performed among six species. For this purpose, at first step, the search method was determined that in this study was "*Calculate the semantic similarity of protein*". For the second step, GO and the measurement method was selected that was Rensik. Finally, for third step, the accession number of pluripotent factors (Table1) of six species was inserted in protein field. The experimental data were used for all factors, except for Klf4 that due to incomplete experimental data, electronic annotation data was not avoidable.

2.4. Protein Similarity of Pluripotency Factors

The Sequence similarity of proteins was determined using the alignment of protein sequences of studied factors. Therefore, a web multiple alignment software was used on National Center Biotechnology Information (NCBI) website. The GenInfo Identifier (GI) number for each protein inserted in the query field and the paired comparisons of proteins were calculated as percentage of similarity.

2.5. Nucleotide Similarity of Pluripotency Factors

In order to match the protein sequence with the nucleotide sequence for each factor, the sequence of each protein was translated into nucleotide sequence using CLC Genomics Workbench 6.0 (CLC bio, Cambridge, MA, USA). Several translations were acquired in the program output that aligned in NCBI web site by BLAST software. The translation that showed high similarity score to the sequence of involved gene, was used to determine nucleotide similarity for the factors studied in the six organisms. The nucleotide similarity was calculated by BLASTN, a web software in NCBI website.

2.6. Statistical Analysis

The estimated similarity by alignment of protein or nucleotide sequences, was considered as the dependent factor and GO semantic similarity in three groups (BP, CC and MF) was considered as the independent factors. Robust regression analysis was used to estimate the strongest and weakest observations and their effect on the estimating regression equation.

Table 1- The Accession Number of Studied Proteins (from Swiss-prot Database).

	Sox2	C-myc	Klf4	Oct4	Nanog	Lin28
Homo sapiens	P48431	P01106	O43474	Q01860	Q9H950	Q9H9Z2
Mus Musculus	P48432	P01108	Q60793	P20263	Q80Z64	Q8K3Y3
Rattus Norvegicus	D4A543	P09416	Q923V7	Q6MG27	A8QWW8	D3ZZA6
Bos Taurus	A2VDX8	Q2HJ27	A7YWE2	O97552	Q4JM65	E1BHM3
Ovis Oris	P54231	Q28566	C7ENG0	C5IX17	C5IX19	H6WP37
Sus Scrufa	B1Q0D1	Q29031	Q52JJ4	Q9TSV5	Q1W1Y4	B1PXG0

All data were analyzed by SAS 9.4 software (SASInstitute and 2004) using Robustreg and Reg procedures. In the Reg procedure, Akaike information criterion (AIC) determined the quality of each model relative to other defined models. Correlation between estimated similarities determined by Pearson correlation and the correlation over 0.6 considered as high and the correlation between 0.2 to 0.6 as medium and under 0.2 considered as weak (Balaji and Srinivasan, 2007).

DOI: 10.22111/jep.2020.31556.1016

3. Results

3.1. Phylogenetic Relationship and GO Similarity of Pluripotency Factors in Six Species

3.1.1. C-myc

The Phylogenetic relationship of proteins was determined based on the number of different amino acids in the comparable sequence of proteins in different species. Evaluation of C-myc protein cleared a close phylogenetic distance between *Rattus norvegicus* and *Mus musculus* and these two species shows high distance with human and other animals (Table 2).

Table2- The Phylogeny Distance of C-myc Factor in Human and Five Studied Animal (Amino Acid Difference /site).

	Organism	1	2	3	4	5
1	*Bos taurus*					
2	*Ovis aries*	0.008				
3	*Sus scrofa*	0.036	0.036			
4	*Homo sapiens*	0.087	0.087	0.061		
5	*Mus musculus*	0.734	0.725	0.742	0.759	
6	*Rattus norvegicus*	0.717	0.709	0.742	0.759	0.020

Table 3 to 5 shows the result of GO semantic similarities in three categories of MF, BP and CC for C-myc factor. In all three categories of Gene Ontology, *R. norvegicus* C-myc exhibited the lowest similarity to other organisms. That was consistent with the phylogenetic distance result but *M. musculus* C-myc has more similarity in compare with rat with other animals in three categories which was not consistent with phylogenetic distance. *H. sapiens* C-myc showed the least similarity to other organisms in three categories of GO (Table 3, 4, 5). Phylogenetic distances displayed very low correlation with the GO similarities and were not statistically significant.

Table 3- GO Semantic Similarity of c-myc on Molecular Function in Human and Five Studied Animal (%).

	Animal	1	2	3	4	5
1	*Bos Taurus*					
2	*Ovis aries*	100				
3	*Sus scrofa*	100	100			
4	*Homo sapiens*	89	89	89		
5	*Mus musculus*	100	100	100	89	
6	*Rattus norvegicus*	80.3	80.3	80.3	69.3	80.3

3.1.2. Sox2

Sox2 phylogenetic distance in species of cattle (*B.taurus*) and sheep (*O.oris*) was zero and in the species of rat (*R.norvegicus*), mouse (*M.musculus*) and pig (*S.scurfa*) was low and housed in one branch. Human (*H. Sapience*) in this category was far from other species.

GO semantic similarity was evaluated for Sox2 in three categories of GO and the results were not consistent with phylogeny results except to *S.scurfa* and *R.norvegicus* that showed 100% similarity in the CC and BP categories.

Table 4- GO Semantic Similarity of c-myc on Cellular Component in Human and Five Studied Animal (%).

	Animal	1	2	3	4	5
1	*Bos Taurus*					
2	*Ovis aries*	100				
3	*Sus scrofa*	100	100			
4	*Homo sapiens*	45.5	45.5	45.5		
5	*Mus musculus*	88.1	88.1	88.1	40	
6	*Rattus norvegicus*	46.5	46.5	46.5	67.7	41.1

3.1.3. Nanog

Paired comparison through the phylogenetic distance for Nanog showed the least distance between *S. Scurfa* and *H. sapiens* whereas the most phylogenetic distance was found between *B.taurus* and *O. Oris*. It was noticeable that the protein sequence of *H. sapiens* and *S.scurfa* were aligned 100% in protein BLAST (with 75% identify) but in the alignment of protein sequence of Nanog in *B.tarus* and *O.oris* the cover was 55% (with 85% identify).

Table 5- GO Semantic Similarity of c-myc on Biological Process in Human and Five Studied Animal (%).

	Animal	1	2	3	4	5
1	*Bos Taurus*					
2	*Ovis aries*	100				
3	*Sus scrofa*	100	100			
4	*Homo sapiens*	71.4	71.4	71.4		
5	*Mus musculus*	81.2	81.2	81.2	70.3	
6	*Rattus norvegicus*	69.7	69.7	69.7	77.3	69.7

GO similarity was the highest in the CC and BP category between *O.oris* and *B.taurus* as well as between *O.oris* and *R.norvegicus*. In the MF category, in addition to mentioned cases, *H. sapiens* showed 100% similarity to *R.norvegicus* and *B.taurus*.

3.1.4. Oct4

The closest phylogenetic relationship was observed between *H.sapiens*, *B.taurus* and *S.scurfa* for Oct4 factor. *R.norvegicus* and *M. musculus* were also close together. The most distance for phylogenic relationship belonged to *B.taurus* and *O.oris*. Go semantic similarity was complete (100%) in the three categories of Go between *B.turus* and *S.scurfa* but *H.sapiens* showed low semantic similarity to *B.taurus* and *S.scurfa* in all three categories which did not correspond to phylogenetic relationship results. *M.mousculus* and *R.norvegicus* displayed the complete GO semantic similarity only in the CC category that was consistent with phylogenetic relationship results.

3.1.5. Lin28

Two groups were observed for Lin28 in the evaluation of phylogenic relationship. The first group included of *H.sapeins*, *R.norvegicus*, *M.mousculus* and *O.oris* and the second group was *B.taurus* and *S.scurfa*. Go semantic similarity in the BP category were also divided into two groups as *B.taurus* and *O.oris* in same group and *H.sapiens*, *R.norvegicus*, *M.mousculus* and *S.scurfa* in second group. The difference between phylogenetic and GO semantic similarity evaluation was the replacement of *O.oris* with *S.scurfa*. There is no consistency in the result of phylogenic relationship and BP category with the MF category. Go semantic similarity in the CC category was vague because of incomplete protein sequence in the database.

3.1.6. Klf4

This pluripotency factor showed a close phylogenic distance equal to 0.004 between *B.taurous* and *O.oris*. There was not any phylogenic relationship between the other species. GO semantic similarity exhibited the same results for Klf4 in the three categories of BP, MF and CC. Semantic similarity between *S.scurfa* and *B.taurus* or *M.mausculus* and *R.norvegicus* were consistent with the phylogenetic results. GO data in the BP category for sheep were incomplete for Klf4 factor.

3.2. Conformity of Protein Sequence Similarity of Pluripotency Factors with Their GO Similarity

Regression analysis was conducted to determine if the semantic similarity of GO could predict the protein sequence similarity of pluripotency factors. In addition to regression analysis, Pearson correlation was measured between protein sequence similarity and GO similarity. Through the pluripotency factors, only c-myc showed a significant regression ($p<0.05$) and Klf-4 and Oct-4 tended to be significant ($p<0.1$). The model that used the semantic similarity of BP and CC categories, was the best model to predict the protein sequence similarity of C-myc (Table 6). Five out of 6 the pluripotency factors used the ingredient of BP similarity in their models, it seems that this ingredient was more effective than other ingredient to estimate the protein similarity. Best models that were selected based on Mallows' Cp were shown in Table 6. The correlation rate of protein sequence similarities with the BP and CC categories was 60.7% and 36.6% subsequently for C-myc factor. The correlation rate of protein sequence similarity with the similarity of BP category was significant ($p<0.05$). A medium correlation between the protein or nucleotide similarity and GO component was shown in Table 7.

Table 6- The Best Model for Prediction of Protein Sequence Similarity based on GO Similarity in Six Pluripotency Factors.

Factors		Coefficient of parameter				R^2	Probability
	n*	BP(b_0)	MF(b_1)	CC(b_2)	Intercept		
Sox2	1	---	0.0170	---	96.3793	0.0366	0.4947
C-myc	2	0.4523	---	-0.1514	67.2777	0.6033	0.0039
Klf-4	2	0.7079	-0.7830	---	101.0441	0.5299	0.0712
Oct-4	3	0.1411	-0.2600	0.1106	92.1165	0.4909	0.0519
Nanog	1	0.0488	---	---	67.1985	0.0186	0.6276
Lin-28	1	0.0149	---	---	94.2186	0.0136	0.6794

*the number of variable in the model

Table 7- Pearson Correlation Coefficient between Protein or Nucleotide Sequence Similarity and the Similarity of GO Categories for C-myc Factor.

	Nucleotide	Protein	GOMF	GOBP	GOCC
Nucleotide	1.000				
Protein	0.844**	1.000			
GOMF	0.339ns	0.310 ns	1.000		
GOBP	0.574*	0.607*	0.680**	1.000	
GOCC	0.284 ns	0.365 ns	0.729**	0.918**	1.000

ns: non-significant
*, **: significant at 0.05 and 0.01 respectively

DOI: 10.22111/jep.2020.31556.1016

3.3. Conformity of GO Similarity of Pluripotency Factors with Their Nucleotide Sequence Similarity

The evaluation of conformity of GO semantic similarity with nucleotide sequence similarity displayed a significant regression for the pluripotency factor of C-myc and Klf4 ($p<0.01$). The prediction model for the nucleotide sequence similarity of C-myc, same as protein similarity model, included two variables of BP and CC. The best model was one variable model with the CC gradient for Klf4 factor (Table 8). GO semantic similarity showed a significant correlation equal to 82.3% with nucleotide sequence similarity for Klf4 factor (Table 9).

Table 8- The Best Model for Prediction of Nucleotide Sequence Similarity based on GO Similarity in Six Pluripotency Factors.

Factors	n*	Regression parameters				R^2	Probability
		BP(b_0)	MF(b_1)	CC(b_2)	Intercept		
Sox2	1	-0.0523	---	---	97.7449	0.1075	0.2328
C-myc	2	0.6541	---	-0.2402	54.1951	0.7044	0.0007
Klf-4	1	---	---	0.0941	86.0708	0.6936	0.0028
Oct-4	3	0.0944	0.1314	0.0840	87.1085	0.3754	0.1451
Nanog	1	-0.0787	---	---	87.09263	0.2061	0.2196
Lin-28	1	0.2765	---	---	87.94960	0.0404	0.4727

* The number of variable in the model

Table 9-Pearson Correlation Coefficient between Protein or Nucleotide Sequence Similarity and Similarity of GO Categories for Klf-4 Factor.

	Nucleotide	Protein	GOMF	GOBP	GOCC
Nucleotide	1.000				
Protein	0.596^{ns}	1.000			
GOMF	0.566^{ns}	0.500^{ns}	1.000		
GOBP	0.651*	0.610^{ns}	0.972**	1.000	
GOCC	0.832**	0.409^{ns}	0.593^{ns}	0.640*	1.000

ns: non-significant

*, **: significant at 0.05 and 0.01 respectively

4. Discussion

This study was conducted to evaluate the rate of Gene Ontology semantic similarity with phylogenetic distance or sequence similarity in the six pluripotency factors. The pairwise similarity of phylogenic distances was compared with the pairwise similarities of the protein sequences for each pluripotency factor, and a negative correlation was found for the factors of C-myc, Sox2, Klf-4 and Lin-28, indicating the correct changes of the two similarity indices. Nanog and Oct4 did not show a correct correlation between phylogenetic distance and protein similarity (Table 10). Failure to match the result of protein similarity with phylogenic distance can be lead to the difference of calculation methods. Phylogenetic distance between two sequences was calculated based on different amino acids in total protein sequence while homology calculation between two protein sequences is a statistical method. Two proteins may 100% aligned through the BLAST function but their similarity were calculated in low rate, for instance Nanog similarity was identified 75% in *S.scurfa* and *H.sapiens* while there is an alignment equal to 100% for them. The identification of protein similarity is a useful estimation to analyze the evaluation distance but it should be considered that there is not a linear relationship and, a little change in similarity percentage of proteins creates a large phylogenic distance between them (Pearson, 2013). In addition, based on a rule of thumb, a protein similarity higher than 30% were statistically significant and both proteins were similar (Pearson, 2013).

Correlation of phylogenetic and GO similarity did not show a clear trend in pluripotency factors and correct trend was only found for C-myc in the three categories of GO. The correlation rate for this factor was weak in CC category and medium in the MF and BP categories.

To estimate the best regression model for prediction of protein similarity using GO semantic similarity was able to find a category of GO that show high conformity with protein similarity. Therefore, in this study a suitable model for pluripotency factors were determined. Among the prediction models for protein or nucleotide similarity, those of C-myc exhibited a significant regression model with high correlation for the parameters. Considering the high and medium correlation of the used ingredients in regression model of protein and nucleotide for C-myc and the significant regression model along with correct change of GO similarity with phylogenetic distance, it can be resulted that the GO similarity of C-myc may be able to explain the protein and nucleotide similarities. The correlation of GO similarity and protein sequence similarity was cleared and this

correlation was significant in the MF category on *H.sapiens* database in Swiss-Prot (Lord et al., 2003). In this study, the BP category of GO showed high correlation with the protein and nucleotide sequences and this ingredient was used in all prediction models. It is suggested that the BP category probably exhibits an accurate relation between the protein sequences and GO terms in the pluripotency factors. With exception of C-myc factor, there was no significant correlation between the GO terms and the sequence of protein and nucleotide of other pluripotent factors. Since the compilation of GO information, many approaches were developed for the annotation of GO terms based on sequence (Schug et al., 2002; Vinayagam et al., 2004). Some of this approaches include the several web depended tools for the GO annotation. These tools often search the similar proteins into GO map in the database proteins like Genebank and Swiss-Prot that result a biological concept for an unknown protein (Hennig et al., 2003; Zehetner, 2003). It is noteworthy that the expressed results based on the sequence similarity and GO definitions by protein interpretation systems do not mean a high percentage of confidence in the prediction of protein function (Duan et al., 2006). On the other hand, if the hypothesis of protein sequence and molecular function relationship is correct (Duan et al., 2006), it was probably better to revise the GO information of pluripotency factors or to add the new information to database. In this study, the use of BP category in the regression model of 5 out of 6 factors suggested this category covers the most accurate ingredients for prediction of protein similarity.

Table 10- Correlation of Phylogeny Indexes of Pluripotency Factors with Protein Similarity and GO Semantic Similarity

Factors	Protein similarity	MF	BP	CC
C-myc	-0.825**	-0.273^{ns}	-0.395 ^{ns}	-0.114 ^{ns}
Sox2	-0.114 ^{ns}	0.302 ^{ns}	-0.322 ^{ns}	0.048 ^{ns}
Klf-4	-0.322 ^{ns}	0.174 ^{ns}	nd	0.120 ^{ns}
Oct-4	0.373^{ns}	-0.300^{ns}	0.130^{ns}	0.126^{ns}
Nanog	0.096 ^{ns}	0.098 ^{ns}	-0.279 ^{ns}	-0.062 ^{ns}
Lin-28	-0.548*	0.190^{ns}	0.068^{ns}	nd

ns: non-significant

*,**: significant at 0.05 and 0.01 respectively

The prediction model of nucleotide sequence for Klf-4 factor was a model with CC ingredient alone and a significant regression. While that ingredient did not involve in protein prediction model of Klf-4, other ingredients (BP and MF category) were effective. It seems that the reason should been searched in the correlation of GO and nucleotide or protein similarity. The nucleotide similarity showed high correlation to the similarity of CC ingredient for Klf-4, while protein similarity correlated to the BP and MF categories. These conflictions were probably related to the medium correlation of protein and nucleotide sequence similarity that was not statistically significant. Therefore, it is suggested that the protein and nucleotide sequences of Klf-4 factor should be revised in six studied species in this manuscript.

In this study, the phylogenetic distance, the protein and nucleotide similarity of six pluripotency factors were compared with their GO semantic similarity in human and five animals. Through the studied factors, the phylogenetic distance of C-myc followed GO semantic similarity and short phylogenetic distance showed high semantic similarity. This following was statistically significant and logic for the protein and nucleotide similarity in C-myc evaluation but there was no the reasonable relationship for others pluripotency factors. All pluripotency factors, exception of C-myc, showed no significant regression for estimation of protein or nucleotide similarity. It's suggested that there is no a significant correlation between protein or nucleotide similarity with GO similarity. The results of this study recommended that the sequence data and GO data of five pluripotency factors of Oct-4, Sox2, Klf4, Lin28 and Nanog were revised for six studied species.

Acknowledgements

The authors gratefully acknowledge the financial support the Excellence Centre for Animal Sciences, and Faculty of Agriculture, Ferdowsi University of Mashhad (FUM) 2/27833; 10/07/2013 and Dr. Banayan Aval for his English editing service.

References

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., 2000. Gene Ontology: tool for the unification of biology. *Nature genetics 25, 25.*

Balaji, S., Srinivasan, N., 2007. Comparison of sequence-based and structure-based phylogenetic trees of homologous proteins: Inferences on protein evolution. *Journal of biosciences 32, 83-96.*

Consortium, G.O., 2004. The Gene Ontology (GO) database and informatics resource. *Nucleic acids research 32, D258-D261.*

Duan, Z.-H., Hughes, B., Reichel, L., Perez, D.M., Shi, T., 2006. The relationship between protein sequences and their gene ontology functions. *BMC bioinformatics 7, S11.*

DOI: 10.22111/jep.2020.31556.1016

Echols, N., Milburn, D., Gerstein, M., 2003. MolMovDB: analysis and visualization of conformational change and structural flexibility. *Nucleic Acids Research 31, 478-482.*

Faria, D., Pesquita, C., Couto, F.M., Falc˜ao, A.e.O., 2007. ProteInOn: A Web Tool for Protein Semantic Similarity, pp. 1-11.

Gerlt, J.A., Babbitt, P.C., 2000. Can sequence determine function? Genome Biology 1, reviews0005. 0001.

Hennig, S., Groth, D., Lehrach, H., 2003. Automated Gene Ontology annotation for anonymous sequence data. *Nucleic Acids Research 31, 3712-3715.*

Leung, A.K.L., Trinkle-Mulcahy, L., Lam, Y.W., Andersen, J.S., Mann, M., Lamond, A.I., 2006. NOPdb: nucleolar proteome database. *Nucleic Acids Research 34, D218-D220.*

Lord, P.W., Stevens, R.D., Brass, A., Goble, C.A., 2003. Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics 19, 1275-1283.*

Park, I.-H., Zhao, R., West, J.A., Yabuuchi, A., Huo, H., Ince, T.A., Lerou, P.H., Lensch, M.W., Daley, G.Q., 2008. Reprogramming of human somatic cells to pluripotency with defined factors. *Nature 451, 141.*

Pearson, W.R., 2013. An introduction to sequence similarity ("homology") searching. *Current Protocols in Bioinformatics 42, 3.1. 1-3.1. 8.*

Šali, A., 1999. Genomics: Functional links between proteins. *Nature 402, 23.*

SAS Institute, , 2004. SAS/STAT User's Guide: Statistics. Version 9.2 Edition. SAS Inst. Inc., Cary, NC.

Schug, J., Diskin, S., Mazzarelli, J., Brunk, B.P., Stoeckert, C.J., 2002. Predicting gene ontology functions from ProDom and CDD protein domains. *Genome Research 12, 648-655.*

Simara, P., Motl, J.A., Kaufman, D.S., 2013. Pluripotent stem cells and gene therapy. *Translational Research 161, 284-292.*

Takahashi, K., Tanabe, K., Ohnuki, M., Narita, M., Ichisaka, T., Tomoda, K., Yamanaka, S., 2007. Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell 131, 861-872.*

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., Kumar, S., 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol 28, 2731-2739.*

Vinayagam, A., König, R., Moormann, J., Schubert, F., Eils, R., Glatting, K.-H., Suhai, S., 2004. Applying support vector machines for gene ontology based gene function prediction. *BMC Bioinformatics 5, 116.*

Wernig, M., Meissner, A., Foreman, R., Brambrink, T., Ku, M., Hochedlinger, K., Bernstein, B.E., Jaenisch, R., 2007. In vitro reprogramming of fibroblasts into a pluripotent ES-cell-like state. *Nature 448, 318.*

Winter, C., Henschel, A., Kim, W.K., Schroeder, M., 2006. SCOPPI: a structural classification of protein–protein interfaces. *Nucleic Acids Research 34, D310-D314.*

Yu, J., Vodyanik, M.A., Smuga-Otto, K., Antosiewicz-Bourget, J., Frane, J.L., Tian, S., Nie, J., Jonsdottir, G.A., Ruotti, V., Stewart, R., 2007. Induced pluripotent stem cell lines derived from human somatic cells. *Science 318, 1917-1920.*

Zehetner, G., 2003. OntoBlast function: From sequence similarities directly to potential functional annotations by ontology terms. *Nucleic Acids Research 31, 3799-3803.*

JOURNAL OF EPIGENETICS, VOL. 01, ISS. 02 (SPRING 2019), 31-39 DOI: 10.22111/jep.2020.31556.1016

JOURNAL OF EPIGENETICS

Journal homepage: http://jep.usb.ac.ir/

آیا شباهت پروتئینی فاکتور های پرتوانی به مفهوم شباهت معنایی ژن آنتولوژی آنهاست؟

رضا مجیدزاده هروی[1]*، محسن قیومیان[2]، مرتضی هاشمی عطار[3]

چکیده

شناخت اعمال بیولوژیکی پروتئینها بر اساس ترتیب اسید آمینه‌ها یک روش ساده‌ای است که در خیلی از نرم افزارها و همچنین توسط پژوهشگران بکار گرفته می شود. اگرچه شباهت ترتیب اسید آمینه‌ای پروتئینها بر شباهت عمل زیستی آنها دلالت ندارد. هدف از این مطالعه تعیین شباهت ژن آنتولوژی شش فاکتور پرتوانی Oct4، Sox2، C-Myc، Klf-4، Lin28، و Nanog در شش گونه و ارزیابی تطابق آنها با شباهت سکانس پروتئینی و فاصله فیلوژنی بود. فاکتور پرتوانی C-myc همبستگی معنی‌داری را بین فاصله فیلوژنی و شباهت پروتئینی ارائه داد. دیگر فاکتورها مانند Sox2، Klf-4 و Lin28 تغییرات صحیح رابطه فیلوژنی و شباهت پروتئینی را نشان دادند اما Oct4 و Nanog ارتباط صحیحی را بین دو شاخص ارائه ندادند چون با افزایش شباهت پروتئینی رابطه فیلوژنی کمتر نشد. در ادامه آزمایش شباهت پروتئینی و یا نوکلوئیدی بعنوان متغیر مستقل و شباهت ژن آنتولوژی در سه شاخه مرحله بیولوژیک (BP)، عمل مولکولی (MF) و جزء سلولی (CC) بعنوان متغیر وابسته فرض شد. با این فرض آنالیز رگرسیون بهترین مدل را برای تخمین شباهت پروتئینی و نوکلئوتیدی ارائه داد. رگرسیون شباهت پروتئینی یا نوکلئوتیدی با شباهت آنتولوژی برای فاکتور C-myc معنی‌دار بود و شباهت شاخه های BP و CC برای تخمین شباهت پروتئینی و یا نوکلئوتیدی توسط مدل انتخاب شدند، اما برای سایر فاکتور های پرتوانی رگرسیون معنی داری برای تخمین شباهت پروتئینی یا نوکلئوتیدی مشاهده نشد. این نتایج نشان می دهد که برای فاکتورهای پرتوانی مورد مطالعه به غیر از C-myc از روی شباهت آنتولوژی نمی‌توان شباهت سکانس پروتئینی یا نوکلئوتیدی را نتیجه‌گیری کرد.

واژگان کلیدی: ژن آنتولوژی، فیلوژنی، پرتوانی، شباهت، پروتئین

[1]- استادیار، گروه علوم دامی، دانشگاه فردوسی مشهد (نویسنده مسئول) rmajidzadeh@um.ac.ir

[2]- کارشناس ارشد، گروه زیست شناسی، دانشگاه فردوسی مشهد M.Qayoomian@gmail.com

[3] - مربی، گروه علوم دامی، دانشگاه فردوسی مشهد mh_544@yahoo.com

* Corresponding author: Tel.: +985138805747.
E-mail address: rmajidzadeh@um.ac.ir