BIOINFORMATICS ORIGINAL PAPER

Vol. 31 no. 1 2015, pages 33–39
doi:10.1093/bioinformatics/btu615

Genome analysis

Advance Access publication September 17, 2014

Making novel proteins from pseudogenes

P. R. Shidhi[1], Prashanth Suravajhala[2,3,4], Aysha Nayeema[5], Achuthsankar S. Nair[1], Shailja Singh[6] and Pawan K. Dhar[6,7,*]

[1]Department of Computational Biology and Bioinformatics, University of Kerala, Kariyavattom, Trivandrum- 695 581, India, [2]Bioinformatics.Org, 28 Pope Street, Hudson, MA 01749, USA, [3]Bioclues.org, India, [4]Bioclues.org, Denmark, [5]National College, University of Kerala, Trivandrum- 695 009, [6]Department of Life Sciences, School of Natural Sciences, Shiv Nadar University, Dadri, Uttar Pradesh- 201 314, and [7]Centre for Systems and Synthetic Biology, University of Kerala, Kariyavattom, Trivandrum- 695 581, India

Associate Editor: Burkhard Rost

ABSTRACT

Motivation: Recently, we made synthetic proteins from non-coding DNA of *Escherichia coli*. Encouraged by this, we asked: can we artificially express pseudogenes into novel and functional proteins? What kind of structures would be generated? Would these proteins be stable? How would the organism respond to the artificial reactivation of pseudogenes?

Results: To answer these questions, we studied 16 full-length protein equivalents of pseudogenes. The sequence-based predictions indicated interesting molecular and cellular functional roles for pseudogene-derived proteins. Most of the proteins were predicted to be involved in the amino acid biosynthesis, energy metabolism, purines and pyrimidine biosynthesis, central intermediary metabolism, transport and binding. Interestingly, many of the pseudogene-derived proteins were predicted to be enzymes. Furthermore, proteins showed strong evidence of stable tertiary structures. The prediction scores for structure, function and stability were found to be favorable in most of the cases.

Impact: To our best knowledge, this is the first such report that predicts the possibility of making functional and stable proteins from pseudogenes. In future, it would be interesting to experimentally synthesize and validate these predictions.

Contact: pawan.dhar@snu.edu.in

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on May 22, 2014; revised on August 28, 2014; accepted on September 12, 2014

1 INTRODUCTION

The term 'pseudogene' has been derived from the term 'pseudo' meaning false. These genes are also known as 'genomic fossils' (Lafontaine and Dujon, 2010). The first pseudogene was reported in 5S DNA of *Xenopus laevis* (Jacq *et al.*, 1977). Pseudogenes are obsolete stretches of DNA sequences that lack protein-coding potential owing to the presence of the frame shift mutation and premature stop codons even though they resemble functional genes (Mighell *et al.*, 2000). They are considered dysfunctional relatives of ancestral functional genes that might have lost function during evolution (Balakirev and Ayala, 2003). Pseudogenes have been reported in plants (Loguercio and Wilkins, 1998), bacteria (Ochman and Davalos, 2006), yeast (Harrison *et al.*, 2002), insects (Ramos-Onsins and Aguadé, 1998), nematodes (Harrison *et al.*, 2001) and mammals (Zhang and Gerstein, 2004).

Based on their origins, pseudogenes have been categorized into (i) processed pseudogenes—formed by retrotransposition of mRNA and have paralogs in the same genome (Li *et al.*, 2013); (ii) duplicated pseudogenes—sometimes called unprocessed pseudogenes arise because of the duplication of functional genes that later on acquire mutation and finally become non-functional; and (iii) unitary or disabled pseudogenes—thought to originate through disruptive mutation in the functional protein-coding genes (Mighell *et al.*, 2000). As new duplicated genes, they could serve as a source of genomic innovations, resulting in novel functions (Presgraves, 2005). Unprocessed and duplicated pseudogenes have intron–exon structures, whereas processed pseudogenes have exonic region only (Nishioka *et al.*, 1980). The long protein-coding genes tend to produce non-processed pseudogenes, whereas short protein-coding genes tend to produce processed pseudogenes (Goncalves, 2000).

Currently, the origin, evolution and function of pseudogenes are incompletely understood. The biological role of pseudogenes were first reported nearly 15 years ago (Korneev *et al.*, 1999) in the form of regulating neuronal nitric oxide synthase gene expression. Recent studies have indicated more functional roles for pseudogenes (Li *et al.*, 2013; Pink *et al.*, 2011; Poliseno et al., 2010). The relationship between pseudogenes and long non-coding RNAs (lncRNAs) is beginning to be understood. Antisense RNA derived from PTEN pseudogene has been found to regulate the transcription and mRNA stability of PTEN tumor suppressor gene (Johnsson *et al.*, 2013). Pseudogene-derived non-coding RNAs amplified the expression level of their parent gene and functioning as endogenous RNAs with the PTEN pseudogene. Further, pseudogene-derived small RNAs have been found to play a role in regional chromatin repression (Guo *et al.*, 2014).

Recent evidences indicate involvement of pseudogenes in regulating the growth of organism (Li *et al.*, 2013) by acting as miRNA decoy (Marques *et al.*, 2012) encoding short peptides or proteins (Bertrand *et al.*, 2002; Kandouz *et al.*, 2004). Studies show that siRNAs derived from pseudogenes of

*To whom correspondence should be addressed.

African *Trypanosoma brucei* suppress the gene expression through RNA interference (Wen *et al*., 2011).

Given their mechanisms of origins, development of these sequences over evolutionary scale of complexity and their potential functional roles, pseudogenes make a strong case for understanding fundamental biology and generating novel applications.

In this bioinformatics study on pseudogenes, our aim was to predict profile of proteins that can be made on demand. This thought has origins in our previous study (Dhar *et al*., 2009) where the feasibility of experimentally making novel and functional proteins from non-coding DNA was demonstrated.

Saccharomyces cerevisiae is among the most well-studied organism where pseudogenes have been identified and analyzed (Lafontaine and Dujon, 2010). It is also one of the most precisely sequenced and annotated eukaryotic genome (Brachat *et al*., 2003). Due to this reason, *S.cerevisiae* was considered in the present study.

2 METHODS

The Saccharomyces Genome Database version of *Saccharomyces cerevisiae* S288C was used in the present study. A total of 20 pseudogene sequences were retrieved using the yeast mine tool and computationally translated into protein sequences using Transeq tool of European Bioinformatics Institute (EBI) (Alvarez-Pérez *et al*., 2013; Goujon *et al*., 2010; Hoefman *et al*., 2014; Rice *et al*., 2000). From this dataset, 16 full-length pseudogenes were found that translated into complete protein sequences without any intervening stop codons. 4 pseudogenes with intervening stop codons was excluded. Thus, in this work, only 16 sequences were considered for detailed study.

2.1 Sequence-based functional prediction

The functional relatives of pseudogene-derived proteins were identified using the BLAST analysis (Altschul *et al*., 1990). The function of pseudogene-derived protein and its relatives was studied using ProtFun tool (Jensen *et al*., 2002, 2003). Protein localization of these sequences was studied using the WoLF PSORT server (Horton *et al*., 2007). STRING database (Franceschini *et al*., 2013) was used to predict physical and functional association network of proteins. The physiochemical properties of pseudogene sequences [molecular weight, theoretical pI, aliphatic index (Ikai, 1980) and GRAVY (Kyte and Doolittle, 1982)] were predicted using the Expasy Protparam server (Gasteiger, 2003), and mRNA secondary structure or folding patterns of pseudogenes were predicted using Mfold server (Zuker, 2003). The 3D structures of pseudogenes were predicted using I-TASSER (Zhang, 2008). Of the 16 sequences considered for the study, five sequences that displayed functional features were finally selected for stability prediction.

2.2 Stability of proteins

To compute the number of stabilization centers, pseudogene sequences were evaluated using SCide (Dosztanyi *et al*., 2003). Sequences showing evidence of stabilization centers were considered for calculating the total energy, including bonds, angles and torsion, improper, non-bonded and electrostatic constrains. While the total energy of the molecule was calculated using GROMAS69 force field implemented in Swiss pdb viewer (Guex and Peitsch, 1997), the cation–π interaction energies were calculated using the CaPTURE program (Gallivan and Dougherty, 1999). The non-covalent interactions such as hydrogen bonds, hydrophobic interactions, disulphide bridges and salt bridges (Baker and Hubbard, 1984; Berman, 1993; Creighton, 2005; Dill, 1990; Horovitz *et al*., 1990; Lins and Brasseur, 1995; Pace *et al*., 1996) were computed using WHAT IF (Vriend, 1990) and PIC Web server (Tina *et al*., 2007). The RASMOL (Sayle, 1995) molecular visualization software was used to visualize the interactions wherein non-canonical interactions, i.e. C–H....π, C–H....O and N–H....π interactions were computed using HBAT program (Tiwari and Panigrahi, 2007). These intermolecular interactions calculated by HBAT were visualized using RasMol, implemented as an add-on program in HBAT. The instability index was calculated based on a weight value using the Expasy Protparam server.

3 RESULTS

3.1 Sequence-based function prediction

3.1.1 Predicting the function of the pseudogene Pseudogene-encoded proteins were evaluated for their functions using ProtFun tool leading to the following functions: amino acid biosynthesis (25%), energy metabolism (19%), central intermediary metabolism (13%), purines and pyrimidines biosynthesis (13%), cell envelope (6%), regulatory functions (6%), fatty acid metabolism (6%), translation (6%), transport and binding (6%) (Table 1).

3.1.2 Identifying functional relatives Of the 16 pseudogene proteins, 8 of them (50%) showed the same function as their immediate relatives (Table 1). Pseudogene-derived proteins were found to map to central intermediary metabolism, energy metabolism, amino acid biosynthesis, purine and pyrimidine synthesis, transport and binding.

3.1.3 Predicting subcellular localization Localization of proteins is an indication of their probable role in the cell. The WoLF PSORT predicted most of the proteins to be localized to cytosol (31%), cytosol and nucleus (25%), nucleus (19%), mitochondria (13%), plasma membrane (6%) and extracellular membrane (6%) (Table 1).

3.1.4 Predicting protein-protein interactions Certain pseudogene-derived proteins were found to have interacting partners showing functions like hexose transporter (sugar transporter), L-serine/threonine dehydratases, dehydrogenase and serine hydrolase (Supplementary Fig. S1). Approximately one-third (31.25%) of the proteins were found to have interacting partners with either uncharacterized or hypothetical proteins. 12.5% of the pseudogene-derived protein sequences did not show any interacting partners.

As an example to highlight the importance of interaction information, the protein–protein network of EKA-9 (Fig. 1) showed interaction with serine dehydratases signature proteins. Pseudogene-derived proteins are shown to interact with various proteins that are experimentally validated further involving various biochemical pathways. From the pathway analysis, we observe that some of these pseudogenes interact with hexose transporter (HXT) families, which are linked to several unknown physiological functions further showing strong similarity with cell cycle mediators apart from its peers. These are involved in pathways specific to the cell cycle and glucose, whereas proteins shown interacting with pseudogenes are involved in glycine, serine and threonine metabolism, cysteine and methionine metabolism, biosynthesis of amino acid and citrate cycle pathways.

Table 1. Functional summary of pseudogenes and their relatives

Pseudogene	Function prediction (Profun)	Structural prediction (I-TASSER C-score)	Subcellular localization WoLF PSORT	Pseudogene relatives (with Seq ID)	Function prediction of Pseudogene relatives (ProtFun)	Sequence identity (%)	Query coverage (%)
EKA-1	Cell envelope	-2.83	Extracellular	Flocculin protein (BAN67853.1)	Cell envelope	66	95
EKA-2	Regulatory functions	-2.62	Cytosol and/or the nucleus	Flo1 Fusion protein (AAY57869.1)	Cell envelope	94	100
EKA-3	Fatty acid metabolism	-4.00	Mitochondria	TKP5 (XP_00164217 4.1)	Central intermediary metabolism	34	95
EKA-4	Central intermediary metabolism	-2.51	Cytosol	TKP5 (XP_00164256 1.1)	Central intermediary metabolism	63	100
EKA-5	Energy metabolism	-1.00	Cytosol and/or nucleus	K7_11200p (GAA23143.1)	Energy metabolism	84	42
EKA-6	Amino acid biosynthesis	-1.50	Mitochondria	K7_11200p (GAA23143.1)	Energy metabolism	84	100
EKA-7	Purines and pyrimidine biosynthesis	-3.70	Nucleus	HXT11 (EGA60048.1)	Purines and pyrimidine biosynthesis	98	100
EKA-8	Transport and binding	-0.12	Plasma membrane	HXT9 (CAY80568.2)	Transport and binding	98	100
EKA-9	Translation	0.58	Cytosol	Cha1 (EGA63144.1)	Energy metabolism	57	94
EKA-10	Amino acid biosynthesis	0.62	Cytosol	K7_03400 (GAA23953.1)	Amino acid biosynthesis	99	100
EKA-11	Purines and pyrimidine biosynthesis	-2.82	Cytosol and/or nucleus	Cos8 (EGA80150.1)	Translation	94	100
EKA-12	Energy metabolism	-2.45	Cytosol and/or nucleus	Cos8 (NP_011815.1)	Translation	97	100
EKA-13	Amino acid biosynthesis	-1.06	Nucleus	Cdc25 (EIW08991.1)	Purines and pyrimidine biosynthesis	96	88
EKA-14	Amino acid biosynthesis	-0.28	Cytosol	Formate dehydrogenase (CDH13920.1)	Amino acid biosynthesis	94	100
EKA-15	Energy metabolism	0.59	Cytosol	Fdh2 (EWG83443.1)	Energy metabolism	99	96
EKA-16	Central intermediary metabolism	-0.08	Nucleus	TY4 B gag pol fusion protein (P47024.3)	Translation	94	99

3.1.5 Predicting physiochemical properties Pseudogene proteins were studied for physicochemical properties (molecular weight, theoretical isoelectric point, aliphatic index and hydropathicity). Molecular weight of proteins was found to range from 7.42 KDa to 127.18 KDa. The isoelectric point (pI) value of proteins was found to vary from 4.67 to 9.54 (pI <7 show acidic nature whereas pI >7 indicate basic nature), aliphatic index value ranging from 64.03 to 102.57 (relatively higher value shows greater stability), hydropathicity value (return of GRAVY score) ranging from –0.734 to 0.443 indicates better interaction with water, as value ranges are low (Table 2).

3.1.6 Tertiary structure prediction The tertiary structure of 16 pseudogene-encoded protein sequences was predicted using I-TASSER server. The I-TASSER confidence score indicates the quality of predicted models based on *ab initio* and threading algorithm. Structures were predicted with C-score varying from –4 to 0.62 (optimal range –5 to 2) while considering the paradigm greater the C-score, greater is the possibility of a good tertiary structure (Supplementary Table S1).

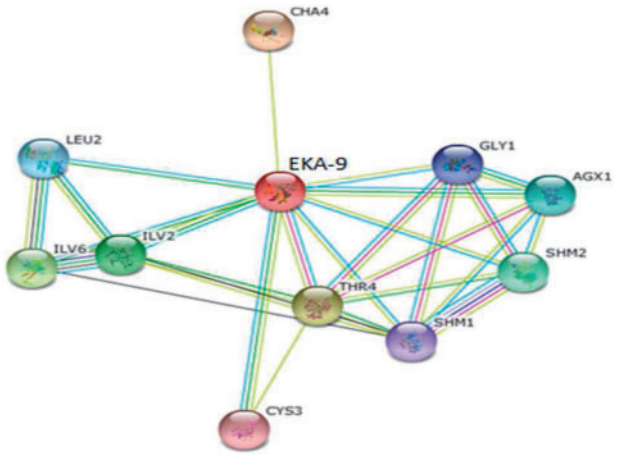

Fig. 1. Protein-protein interaction network of EKA-9

3.2 Predicting stability

3.2.1 Stabilization centers, cation–π, non-covalent and non-canonical interactions Based on the results obtained from the function prediction, five pseudogenes showing top-most function hits were considered for stability prediction. All predicted protein sequences showed low negative value for the total energy indicating that pseudogene-derived proteins would be probably *stable*, if expressed (Table 3). Approximately 40% of the proteins exhibited stabilization centers ranging from 40 to 100 with rest of the proteins equally distributed (20% each) across the rest of the stabilization centers (Table 3).

Studies on protein stability among the five sequences considered revealed remarkable observations with one showing >20 cation–π interactions, three showing the presence of <5 cation–π interactions and one not showing any cation–π interaction (Table 3). EKA-16 showed highest number of non-covalent interactions (1018) and non-canonical interactions (218) followed by EKA-8 and EKA-15. Table 4 describes non-covalent interactions, and Table 5 describes non-canonical interactions.

The MFOLD results (Table 3) showed EKA-13 (–73.1 kcal/mol) and EKA-9 (–95.2 kcal/mol) with higher ΔG, whereas EKA-16 (–740 kcal/mol) and EKA-15 (–201.6 kcal/mol) showed relatively lower ΔG.

3.2.2 Predicting instability index Sequences showed instability index ranging from 29.92 to 51.4 (Table 3) indicating that most of the proteins are likely to be stable, if expressed.

3.3 Correlation of stability parameters

A nearly consistent trend for all stability parameters was found for all the pseudogene-encoded proteins (Fig. 2). From these data, we infer that the total energy of the folded structures is in the favorable range (Fig. 3). Further, non-canonical

Table 2. Physicochemical properties of pseudogenes

Pseudogene	Molecular weight (Da)	Theoretical pI	Aliphatic index	GRAVY
EKA 1	7609.6	8.79	64.03	–0.164
EKA 2	27 755	4.67	75.31	–0.109
EKA 3	16 049.3	9.0	80.07	–0.139
EKA 4	34 507.4	6.75	89.9	–0.223
EKA 5	12 434.3	8.42	89.14	–0.102
EKA 6	8 646.4	9.54	100	0.14
EKA 7	11 638	6.54	74.31	–0.22
EKA 8	50 830.2	8.71	95.78	0.443
EKA 9	13 942.9	9.48	70.71	–0.411
EKA 10	22 812	4.7	102.57	0.101
EKA 11	27 454	8.21	85.13	–0.04
EKA 12	7 427.7	5.25	102.13	–0.12
EKA 13	11 554.1	9.33	72.82	–0.399
EKA 14	15 936	4.89	101.45	–0.159
EKA 15	26 487.3	9.3	87.58	–0.429
EKA 16	127 188.7	8.29	80.14	–0.734

Table 3. Representing total energy stabilization centers, cation–π interactions and ΔG

Pseudogene	Stabilization centers	Instability index	Totalenergy kcal/mol	Cation – π interaction	Cation – π interaction energykcal/mol	ΔG kcal/mol
EKA-8	47	34.6	–7686.523	2	–1.06	–412
EKA-9	12	30.54	–2619.806	1	6.5	–95.2
EKA-13	34	51.4	–1807.638	2	–5.1	–73.1
EKA-15	42	29.92	–6485.247	NA	NA	–201.6
EKA-16	192	41.36	–24 481.600	21	–54.71	–740

Table 4. Non-covalent interactions

Pseudogene	Hydrogen bond	Di – sulphide bridge	Saltbridge	Hydrophobic interaction	Total number of interactions
EKA-8	332	9	88	338	429
EKA-9	76	1	23	57	100
EKA-13	37	3	20	62	60
EKA-15	139	2	62	149	203
EKA-16	564	21	433	387	1018

Table 5. Non-canonical interactions

Pseudogene	C – H . . . π interactions	C – H . . . O interactions	N – H . . . π interactions	Total number of non-canonical interactions
EKA-8	18	83	15	116
EKA-9	1	14	13	28
EKA-13	NA	14	6	20
EKA-15	13	18	19	50
EKA-16	25	132	61	218

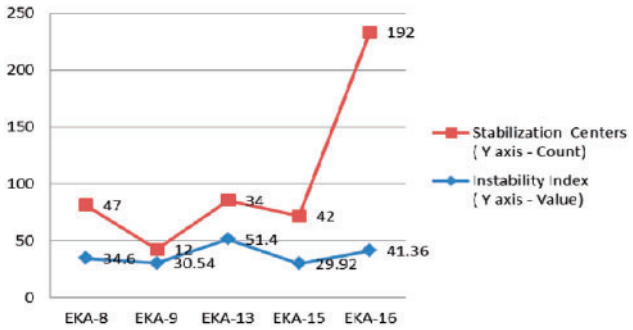

Fig. 2. Trend for Instability index and Stabilization Centers

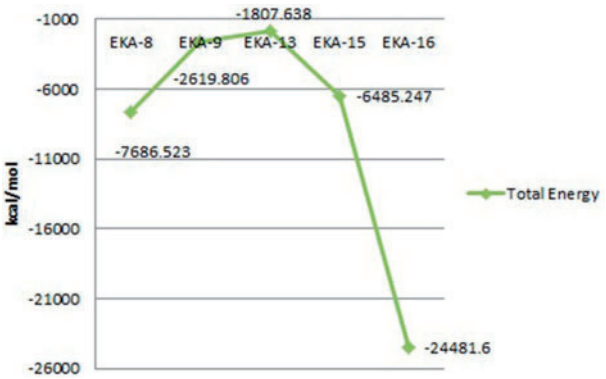

Fig. 3. Trend for Total Energy

interaction and non-covalent interaction trends indicate stability of the structures predicted (Figs 4 and 5).

4 DISCUSSION

The present study is an extension of previous work (Dhar *et al.*, 2009) where *Escherichia coli* non-coding DNA sequences were artificially expressed into functional proteins. This gave rise to a new question—what would happen if we artificially expressed pseudogene sequences? Would they make stable and functional proteins? How would their structure look like? What kind of molecules would they interact with? Given their increasingly complex role at both genetic and epigenetic levels (Guo *et al.*,

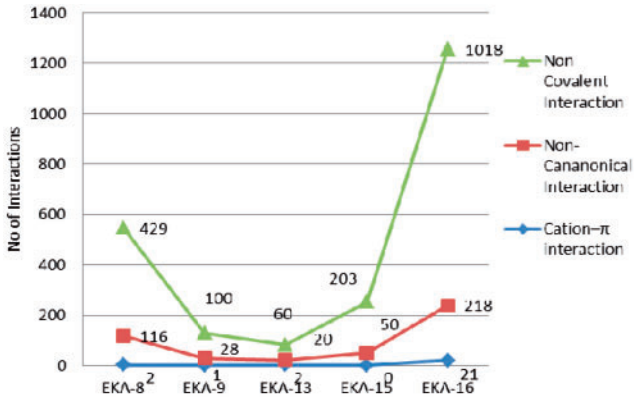

Fig. 4. Trend for non-canonical interactions, non-covalent interaction and cation–π interactions

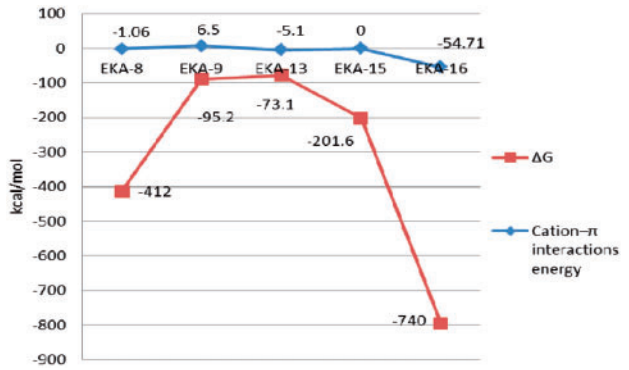

Fig. 5. Cation–π interaction energy and ΔG

2014), this study attempts to present a novel way of understanding pseudogene biology by artificially expressing candidate sequences that show most promising leads.

Of the 20 pseudogenes that were computationally translated to protein sequences, 16 sequences gave full-length open reading frames (ORF) without any stop codons and were considered in this study. To understand potential property of pseudogene proteins, sequence-based and structure-based prediction studies were performed using the best computational tools available.

Before this step, pseudogene sequences were sent for RNA structure prediction based on the reasoning that if pseudogene mRNA assumes a rigid secondary structure, it would be difficult to artificially synthesize proteins. Reports suggest that 'highly expressed genes' may not show stable mRNA secondary structure, whereas 'low expressed genes' may show highly stable mRNA secondary structure (Drummond *et al.*, 2005; Mukund *et al.*,1999). Thus, higher the value of ΔG, lower the possibility of forming a stable mRNA secondary structure. All pseudogene proteins exhibited ΔG ranging from –740 kcal/mol to –73.1 kcal/mol (Table 3) indicating that pseudogenes may have been possibly low-expressing genes in the past when they were in an active state. It would be interesting to see how they behave when artificially expressed using both a weak and a strong promoter.

To strengthen pseudogene predictions for experimental validation, it is important to address the reliability of function predictions. Functions of selected pseudogenes and its relatives were predicted using sequence information, as tools have been developed that reliably predict the function from sequence data (Jensen *et al.* 2002, 2003). It was found that many pseudogene-encoded proteins (EKA-1, EKA-4, EKA-5, EKA-7, EKA-8, EKA-10, EKA-14 and EKA-15) had function similar to that of their relatives (Table 1). Furthermore, several pseudogene proteins were predicted to play a role in amino acid biosynthesis and energy metabolism. We found that three pseudogenes (EKA-8, EKA-9 and EKA-15) from the set of 16 pseudogenes showed similar functions based on both protein–protein interaction network and Gene Ontology predictions.

Majority of pseudogene proteins were found to localize to cytosol. Interestingly, pseudogene-encoded proteins (EKA-5, EKA-12 and EKA-15) that show up in the energy metabolism category also localize to cytosol (cytoplasm), thus strengthening the belief in predictions. Proteins with regulatory functions (EKA-2) were also localized to cytosol (cytoplasm). Interestingly, EKA-7 and EKA-11 showing up under purines and pyrimidines biosynthesis functional category were found to be localized to the nucleus subcellular compartment (Table 1).

It was encouraging to observe that predicted pseudogene proteins showed *high aliphatic index* values and *lower instability index* indicating greater stability, if expressed. The low hydrophobicity score of predicted proteins indicate their polar nature.

Tertiary structure of potential pseudogene proteins was determined by using I-TASSER because of the wide acceptance of this tool in the community. The quality of model is estimated based on the C-score. The convergence parameters of the structure assembly simulations and threading template alignments are used for calculating the C-score. Typically, C-score value is between –5 and 2, where higher value of C-score indicates a model with a high confidence and vice versa. The C-score value for all pseudogene proteins was found to be in the range of –4 to 0.58, indicating a *strong* foldability of the predicted proteins

Structural stability of proteins is an important indicator of their potential function (Ramanathan *et al.*, 2011). To further understand the strength of structural predictions, we studied proteins using stability parameters like stabilization centers, total energy, cation–π interaction energies, non-covalent, non-canonical interaction and instability index, and encouraging evidence of protein structure stability was found (Tables 3–5). Further, the total energy of the proteins was calculated using GROMACS force field—wherein the lower the energy, the higher the possibility of stable configuration. The total energy of all the proteins individually was found to be negative (Table 3) indicating that all the proteins are likely to exhibit a stable structure, if expressed. Further, 5 of the 16 proteins exhibited several stabilization centers. Among all the proteins, EKA-8 and EKA-16 were found to have the lowest energy and highest numbers of stabilization centers (Table 3).

We also studied non-covalent interactions like hydrogen bonds, hydrophobic interactions, disulphide bridges, salt bridges and cation–π interactions in these proteins. The data obtained under all these categories give a strong indication of stable foldability of proteins. Among all the proteins, EKA-16 and EKA-8 show presence of higher non-covalent interaction indicating better stability (Table 4). These two proteins also showed the highest number of non-canonical interactions suggesting higher structural stability of proteins thus lending support to the 3D

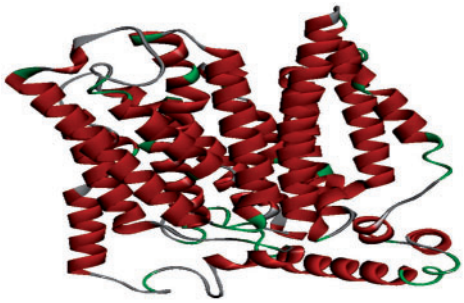

Fig. 6. Structure of EKA-8 with alpha helices and coils

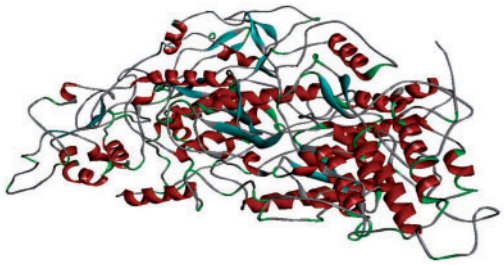

Fig. 7. Structure of EKA-16 with alpha helices, beta-pleated sheets and coils

structure stability profiles (Ramanathan *et al.*, 2011; Umezawa and Nishio, 1998).

Finally, to validate the strength of stability predictions, we performed tests that examine instability of these proteins. The instability index indicates whether a protein would be unstable *in vivo* (Guruprasad *et al.*, 1990)—the instability index of <40 is considered as a good evidence of stability (Ramanathan *et al.*, 2011). Interestingly, EKA-15 sequence exhibited the lowest instability followed by EKA-9, EKA-8 and EKA-16 (Table 3).

Overall, this study suggests that EKA-8 (Fig. 6) and EKA-16 (Fig. 7) are the two most promising pseudogenes for artificial expression into proteins. Experiments have been started to validate these predictions. It would be interesting to see how cell responds to deliberate expression of sequences that nature decided to switch off. Given the context dependency and emergent properties arising from protein interactions (Banerji, 2013), it would be interesting to see the experimental outcome of artificial pseudogene expression.

5 CONCLUSION

This work explores the possibility of making stable and functional proteins from pseudogenes. A comprehensive multi-parametric study, based on sequence and structural evidences identifies two pseudogenes (EKA 8 and EKA 16) as the most promising candidates for the future artificial protein synthesis and functional studies.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the facilities provided by Department of Computational Biology and Bioinformatics, University of Kerala.

Funding: Funded by State Inter-University Centre of Excellence in Bioinformatics and MHRD Centre of Excellence in Ayurinformatics.

Conflict of interest: none declared.

REFERENCES

Altschul,S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Alvarez-Pérez,S. *et al.* (2013) Multilocus sequence analysis of nectar pseudomonads reveals high genetic diversity and contrasting recombination patterns. *PLoS One*, **8**, e75797.

Baker,E.N. and Hubbard,R.E. (1984) Hydrogen bonding in globular proteins. *Prog. Biophys. Mol. Biol.*, **44,** 97–179.

Balakirev,E.S. and Ayala,F.J. (2003) Pseudogenes: are they "junk" or functional DNA? *Annu. Rev. Genet.*, **37,** 123–151.

Banerji,A. (2013) An attempt to construct a (general) mathematical framework to model biological "context-dependence". *Syst. Synth. Biol.*, **7,** 221–227.

Berman,H.M. (1993) Hydrogen bonding in biological structures. G.A. Jeffrey and W. Saenger. *Biophys. J.*, **64**, 1976.

Bertrand,N. *et al.* (2002) Proneural genes and the specification of neural cell types. *Nat. Rev. Neurosci.*, **3,** 517–530.

Brachat,S. *et al.* (2003) Reinvestigation of the Saccharomyces cerevisiae genome annotation by comparison to the genome of a related fungus: Ashbya gossypii. *Genome Biol.*, **4**, R45.

Creighton,T.E. (2005) Proteins: Structures and Molecular Properties. *Nucleic Acids Res.*, **31**, 3345–3348.

Dhar,P.K. *et al.* (2009) Synthesizing non-natural parts from natural genomic template. *J. Biol. Eng.*, **3**, 2.

Dill,K.A. (1990) Dominant forces in protein folding. *Biochemistry*, **29,** 7133–7155.

Dosztanyi,Z. *et al.* (2003) SCide: identification of stabilization centers in proteins. *Bioinformatics*, **19,** 899–900.

Drummond,D.A. *et al.* (2005) Why highly expressed proteins evolve slowly. *Proc. Natl Acad. Sci. U. S. A.*, **102**, 14338–14343.

Franceschini,A. *et al.* (2013) STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res.*, **41**, D808–D815.

Gallivan,J.P. and Dougherty,D.A. (1999) Cation-pi interactions in structural biology. *Proc. Natl Acad. Sci. USA*, **96**, 9459–9464.

Gasteiger,E. (2003) ExPASy: the proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res.*, **31**, 3784–3788.

Goncalves,I. (2000) Nature and Structure of Human Genes that Generate Retropseudogenes. *Genome Res.*, **10**, 672–678.

Goujon,M. *et al.* (2010) A new bioinformatics analysis tools framework at EMBL-EBI. *Nucleic Acids Res.*, **38**, W695–W699.

Guex,N. and Peitsch,M.C. (1997) SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis*, **18**, 2714–2723.

Guo,X. *et al.* (2014) Characterization of human pseudogene-derived non-coding RNAs for functional potential. *PLoS One*, **9**, e93972.

Guruprasad,K. *et al.* (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. *Protein Eng.*, **4**, 155–161.

Harrison,P. *et al.* (2002) A small reservoir of disabled ORFs in the yeast genome and its implications for the dynamics of proteome evolution. *J. Mol. Biol.*, **316**, 409–419.

Harrison,P.M. *et al.* (2001) Digging for dead genes: an analysis of the characteristics of the pseudogene population in the Caenorhabditis elegans genome. *Nucleic Acids Res.*, **29**, 818–830.

Hoefman,S. *et al.* (2014) Niche differentiation in nitrogen metabolism among methanotrophs within an operational taxonomic unit. *BMC Microbiol.*, **14**, 83.

Horovitz,A. *et al.* (1990) Strength and co-operativity of contributions of surface salt bridges to protein stability. *J. Mol. Biol.*, **216**, 1031–1044.

Horton,P. *et al.* (2007) WoLF PSORT: protein localization predictor. *Nucleic Acids Res.*, **35**, W585–W587.

Ikai,A. (1980) Thermostability and aliphatic index of globular proteins. *J. Biochem.*, **88**, 1895–1898.

Jacq,C. *et al.* (1977) A pseudogene structure in 5S DNA of Xenopus laevis. *Cell*, **12**, 109–120.

Jensen,L.J. *et al.* (2003) Prediction of human protein function according to gene ontology categories. *Bioinformatics*, **19**, 635–642.

Jensen,L.J. *et al.* (2002) Prediction of human protein function from post-translational modifications and localization features. *J. Mol. Biol.*, **319**, 1257–1265.

Johnsson,P. *et al.* (2013) A pseudogene long noncoding RNA network PTEN transcription and translation in human cells. *Nat. Struct. Mol. Biol.*, **20**, 440–446.

Kandouz,M. *et al.* (2004) Connexin43 pseudogene is expressed in tumor cells and inhibits growth. *Oncogene*, **23**, 4763–4770.

Korneev,S.A. *et al.* (1999) Neuronal Expression of Neural Nitric Oxide Synthase (nNOS) protein is suppressed by an antisense RNA transcribed from an NOS pseudogene. *J. Neurosci.*, **19**, 7711–7720.

Kyte,J. and Doolittle,R.F. (1982) A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.*, **157**, 105–132.

Lafontaine,I. and Dujon,B. (2010) Origin and fate of pseudogenes in *Hemiascomycetes*: a comparative analysis. *BMC Genomics*, **11**, 260.

Li,W. *et al.* (2013) Pseudogenes: pseudo or real functional elements? *J. Genet. Genomics*, **40**, 171–177.

Lins,L. and Brasseur,R. (1995) The hydrophobic effect in protein folding. *FASEB J.*, **9**, 535–540.

Loguercio,L.L. and Wilkins,T.A. (1998) Structural analysis of a hmg-coA-reductase pseudogene: insights into evolutionary processes affecting the hmgr gene family in allotetraploid cotton (Gossypium hirsutum L.). *Curr. Genet.*, **34**, 241–249.

Marques,A.C. *et al.* (2012) Evidence for conserved post-transcriptional roles of unitary pseudogenes and for frequent bifunctionality of mRNAs. *Genome Biol.*, **13**, R102.

Mighell,A.J. *et al.* (2000) Vertebrate pseudogenes. *FEBS Lett.*, **468**, 109–114.

Mukund,M.A. *et al.* (1999) Effect of mRNA secondary structure in the regulation of gene expression: unfolding of stable loop causes the expression of Taq polymerase in *E. coli*. *Curr. Sci.*, **76**, 1486–1490.

Nishioka,Y. *et al.* (1980) Unusual alpha-globin-like gene that has cleanly lost both globin intervening sequences. *Proc. Natl Acad. Sci. USA*, **77**, 2806–2809.

Ochman,H. and Davalos,L.M. (2006) The nature and dynamics of bacterial genomes. *Science*, **311**, 1730–1733.

Pace,C.N. *et al.* (1996) Forces contributing to the conformational stability of proteins. *FASEB J.*, **10**, 75–83.

Pink,R.C. *et al.* (2011) Pseudogenes: pseudo-functional or key regulators in health and disease? *RNA*, **17**, 792–798.

Poliseno,L. *et al.* (2010) A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. *Nature*, **465**, 1033–1038.

Presgraves,D.C. (2005) Evolutionary genomics: new genes for new jobs. *Curr. Biol.*, **15**, R52–R53.

Ramanathan,K. *et al.* (2011) Predicting therapeutic template by evaluating the structural stability of anti-cancer peptides—a computational approach. *Int. J. Pept. Res. Ther.*, **17**, 31–38.

Ramos-Onsins,S. and Aguadé,M. (1998) Molecular evolution of the Cecropin multigene family in Drosophila. functional genes vs. pseudogenes. *Genetics*, **150**, 157–171.

Rice,P. *et al.* (2000) EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet.*, **16**, 276–277.

Sayle,R. (1995) RASMOL: biomolecular graphics for all. *Trends Biochem. Sci.*, **20**, 374–376.

Tina,K.G. *et al.* (2007) PIC: protein interactions calculator. *Nucleic Acids Res.*, **35**, W473–W476.

Tiwari,A. and Panigrahi,S.K. (2007) HBAT: a complete package for analysing strong and weak hydrogen bonds in macromolecular crystal structures. *In Silico Biol.*, **7**, 651–661.

Umezawa,Y. and Nishio,M. (1998) CH/pi interactions in the crystal structure of class I MHC antigens and their complexes with peptides. *Bioorg. Med. Chem.*, **6**, 2507–2515.

Vriend,G. (1990) WHAT IF: a molecular modeling and drug design program. *J. Mol. Graph.*, **8**, 52–56, 29.

Wen,Y.-Z. *et al.* (2011) Pseudogene-derived small interference RNAs regulate gene expression in African *Trypanosome brucei*. *Proc. Natl Acad. Sci. USA*, **108**, 8345–8350.

Zhang,Y. (2008) I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics*, **9**, 40.

Zhang,Z. and Gerstein,M. (2004) Large-scale analysis of pseudogenes in the human genome. *Curr. Opin. Genet. Dev.*, **14**, 328–335.

Zuker,M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.*, **31**, 3406–3415.