Making novel proteins from pseudogenes

P. R. Shidhi¹, Prashanth Suravajhala²,³,⁴, Aysha Nayeema⁵, Achuthsankar S. Nair¹, Shailja Singh⁶ and Pawan K. Dhar⁶,⁷,*

¹Department of Computational Biology and Bioinformatics, University of Kerala, Kariyavattom, Trivandrum- 695 581, India, ²Bioinformatics.Org, 28 Pope Street, Hudson, MA 01749, USA, ³Bioclues.org, India, ⁴Bioclues.org, Denmark, ⁵National College, University of Kerala, Trivandrum- 695 009, ⁶Department of Life Sciences, School of Natural Sciences, Shiv Nadar University, Dadri, Uttar Pradesh- 201 314, and ⁷Centre 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 protein-coding potential. They are considered as functional 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 (Brachet 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 Transseq tool of European Bioinformatics Institute (EBI) (Alvarez-Perez 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 physicochemical 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 (Doustanjii 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 amount 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.

34
The tertiary structure of 16 pseudogene-encoded protein sequences was predicted using the I-TASSER server. The I-TASSER confidence score indicates structural prediction (I-TASSER score) ranging from 4.67 to 9.54 (pI to 127.18 KDa). The isoelectric point (pI) value of proteins was found to range 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, while 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 above.

### 3.2 Predicting stability

#### 3.2.1 Stabilization centers, cation–/C25 interactions

Studied on protein stability among the five sequences considering revealed remarkable observations with one showing >20 cation–/C25 interactions, three showing the presence of C5 cat-

#### 3.2.2 Predicting instability index

The MFOLD results (Table 3) showed EKA-13 (–73.1 kcal/mol) and EKA-16 (–74.6 kcal/mol) showed relatively lower AG.

### 3.3 Correlation of stability parameters

#### 3.3.1 Predicting physicochemical properties

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 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, while 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 above.

### Table 1. Functional summary of pseudogenes and their relatives

<table>
<thead>
<tr>
<th>EKA-16</th>
<th>EKA-15</th>
<th>EKA-14</th>
<th>EKA-13</th>
<th>EKA-12</th>
<th>EKA-11</th>
<th>EKA-10</th>
<th>EKA-9</th>
<th>EKA-8</th>
<th>EKA-7</th>
<th>EKA-6</th>
<th>EKA-5</th>
<th>EKA-4</th>
<th>EKA-3</th>
<th>EKA-2</th>
<th>EKA-1</th>
<th>Pseudogene</th>
</tr>
</thead>
<tbody>
<tr>
<td>Function</td>
<td>Function</td>
<td>Function</td>
<td>Function</td>
<td>Function</td>
<td>Function</td>
<td>Function</td>
<td>Function</td>
<td>Function</td>
<td>Function</td>
<td>Function</td>
<td>Function</td>
<td>Function</td>
<td>Function</td>
<td>Function</td>
<td>Function</td>
<td>Function</td>
</tr>
<tr>
<td>4.1</td>
<td>4.2</td>
<td>4.3</td>
<td>4.4</td>
<td>4.5</td>
<td>4.6</td>
<td>4.7</td>
<td>4.8</td>
<td>4.9</td>
<td>5.0</td>
<td>5.1</td>
<td>5.2</td>
<td>5.3</td>
<td>5.4</td>
<td>5.5</td>
<td>5.6</td>
<td>5.7</td>
</tr>
<tr>
<td>WoLF PSORT</td>
<td>WoLF PSORT</td>
<td>WoLF PSORT</td>
<td>WoLF PSORT</td>
<td>WoLF PSORT</td>
<td>WoLF PSORT</td>
<td>WoLF PSORT</td>
<td>WoLF PSORT</td>
<td>WoLF PSORT</td>
<td>WoLF PSORT</td>
<td>WoLF PSORT</td>
<td>WoLF PSORT</td>
<td>WoLF PSORT</td>
<td>WoLF PSORT</td>
<td>WoLF PSORT</td>
<td>WoLF PSORT</td>
<td>WoLF PSORT</td>
</tr>
<tr>
<td>Subcellular localization</td>
<td>Subcellular localization</td>
<td>Subcellular localization</td>
<td>Subcellular localization</td>
<td>Subcellular localization</td>
<td>Subcellular localization</td>
<td>Subcellular localization</td>
<td>Subcellular localization</td>
<td>Subcellular localization</td>
<td>Subcellular localization</td>
<td>Subcellular localization</td>
<td>Subcellular localization</td>
<td>Subcellular localization</td>
<td>Subcellular localization</td>
<td>Subcellular localization</td>
<td>Subcellular localization</td>
<td>Subcellular localization</td>
</tr>
<tr>
<td>EKA-16</td>
<td>EKA-15</td>
<td>EKA-14</td>
<td>EKA-13</td>
<td>EKA-12</td>
<td>EKA-11</td>
<td>EKA-10</td>
<td>EKA-9</td>
<td>EKA-8</td>
<td>EKA-7</td>
<td>EKA-6</td>
<td>EKA-5</td>
<td>EKA-4</td>
<td>EKA-3</td>
<td>EKA-2</td>
<td>EKA-1</td>
<td>Pseudogene</td>
</tr>
<tr>
<td>Function</td>
<td>Function</td>
<td>Function</td>
<td>Function</td>
<td>Function</td>
<td>Function</td>
<td>Function</td>
<td>Function</td>
<td>Function</td>
<td>Function</td>
<td>Function</td>
<td>Function</td>
<td>Function</td>
<td>Function</td>
<td>Function</td>
<td>Function</td>
<td>Function</td>
</tr>
<tr>
<td>4.1</td>
<td>4.2</td>
<td>4.3</td>
<td>4.4</td>
<td>4.5</td>
<td>4.6</td>
<td>4.7</td>
<td>4.8</td>
<td>4.9</td>
<td>5.0</td>
<td>5.1</td>
<td>5.2</td>
<td>5.3</td>
<td>5.4</td>
<td>5.5</td>
<td>5.6</td>
<td>5.7</td>
</tr>
</tbody>
</table>
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.,...
Making novel proteins from pseudogenes

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 hydropobicity 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...
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


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

![Fig. 6](image_url)

![Fig. 7](image_url)
Making novel proteins from pseudogenes


