Syst Synth Biol (2014) 8:169–171
DOI 10.1007/s11693-014-9148-4

COMMENTARY

Identifying pseudogenes from hypothetical proteins for making synthetic proteins

P. R. Shidhi · Achuthsankar S. Nair · Prashanth Suravajhala

Received: 4 April 2014 / Revised: 10 April 2014 / Accepted: 11 April 2014 / Published online: 20 April 2014

Abstract Nature selected certain regions of the genome for encoding proteins. Most of the sequences were used to encode only RNA. What happened to the remaining sections of the genome? It is possible that some sequences were retired and retained as non-functional entities called pseudogenes. Though several evolutionary prospects with functional endpoints exist, we looked at the possibility of hypothetical proteins correlating with the emergence of pseudogenes and potential of such genes to make novel synthetic molecules. In this commentary, we consider two key aspects: (1) does any correlation exist between hypothetical proteins and pseudogenes and (2)—can we make novel and functional proteins from pseudogenes?

Keywords Pseudogenes · Hypothetical proteins · Synthetic proteins

A hypothetical protein (HP) is a protein whose existence is predicted but the proof of its expression remains uncertain. Interestingly, many HPs in the recent past have been known to be expressed in vivo. With various methods known to identify components in cell membrane, the functional significance of large number of proteins, especially those that (a) do not have function, (b) are not expressed, (c) are unique or common among genomes, merit detailed study (Sivashankari and Shanmughavel 2006). Majority of genomic regions encoding hypothetical proteins are non-characterized thereby making orphan genes, interesting candidates for study (Dujon 1996). Most of these regions are predicted by computational methods due to non-similarity to known proteins or EST (Expressed Sequence Tag) sequences. The undiscovered regions or proteins whose functions are not known could be of great interest because some of these regions might contribute to the development of human disease (Bianchi et al. 1999). The post annotation experience indicates that many proteins turn out to be obsolete end products, thus giving rise to the formation of pseudogenes. Currently, a broad spectrum of computational biology tools can be utilized to check whether a functional molecule can be made from sequences considered non-coding (Dhar et al. 2009). Furthermore, some hypothetical pseudogenes or spurious open reading frames that are translated could be their protein product equivalents with important regulatory functions.

Classification scoring systems to find *bona fidelity*

While manual annotation is followed by computational predictions, deciphering function of candidate pseudogenes in exploring the putative pathways is important. Our previous studies were aimed to find whether or not any structural templates for mitochondrial proteins harbor domains that were significant in their expression (Unpublished). Although community is engaged in carrying out protein interaction studies, experiments are time consuming and a consensus on this topic is still missing. Computational evidence exists to show that hypothetical proteins could have functional homologs (Galperin and Koonin

P. R. Shidhi · A. S. Nair
Department of Computational Biology and Bioinformatics,
University of Kerala, Kariyavattom Campus,
Thiruvananthapuram 695581, Kerala, India

P. Suravajhala (✉)
Bioclues.Org, Hyderabad, India
e-mail: prash@bioclues.org

P. Suravajhala
Bioinformatics.Org, 28 Pope Street, Hudson, MA 01749, USA

2004). Furthermore, confidence in the form of reliability scores could help understand, if the classifiers employed for predicting functions using bioinformatics methods, are indeed correct. Although classifiers are based on the manual annotation and predictions, scoring patterns of pseudogenes and hypothetical proteins matching with non-coding regions and their putative synthetic products is a challenge. To fulfill this aim, we believe there is a need for developing a computational approach in the form of a feature selection that is in consensus with the experiments for checking whether or not any non-coding regions of these pseudogenes are artificially made in the laboratory. Due to contextual annotation, several works apart from feature selection in the recent past have caught interest. Apart from genomic context analysis which provides functional clues about hypothetical proteins, a wide variety of methods, viz. K-nearest neighbors (Lan et al. 2013), minimum-redundancy maximum-relevance (Wang et al. 2013) have been employed in the recent-past.

Can we make synthetic products from hypothetical pseudogenes?

Can a pseudogene, whose existence has been predicted but no experimental evidence exists, be 'synthetically expressed' in an organism? Recent experimental evidence suggests that novel proteins can be made from sequences historically considered junk (Dhar et al. 2009). Thus, it would be interesting to artificially make synthetic proteins from hypothetical proteins that show strong correlation with pseudogenes, which will allow us to understand their potential function. Our recent (unpublished observations) on 28 synthetic protein sequences from GenBank show that these proteins can be functionally related to mitochondria and also strengthen the premise of making functional proteins with unknown domains whose structures are not deciphered. Based on this, we reason that proteins containing Domain of Unknown Function or pseudogene like features can be considered concomitant homologues to some synthetic proteins. For example, Fanconi Anemia (FA) genes considered 'hypothetical' were shown to be related to breast cancer genes even though some of their counterparts have been linked to pseudogenesis (Knies et al. 2012). There was nothing in common with any other genes until when breast cancer gene BRCA2 was shown to be similar with Fanconi Anemia gene FANCD1 (Fortugno 2007). As a result, annotations using functional and comparative genomic studies have shown a great potential for genes where multiple independent pseudogenes may result in genomic instability, thus rendering significant disadvantage to the survival of the organism.

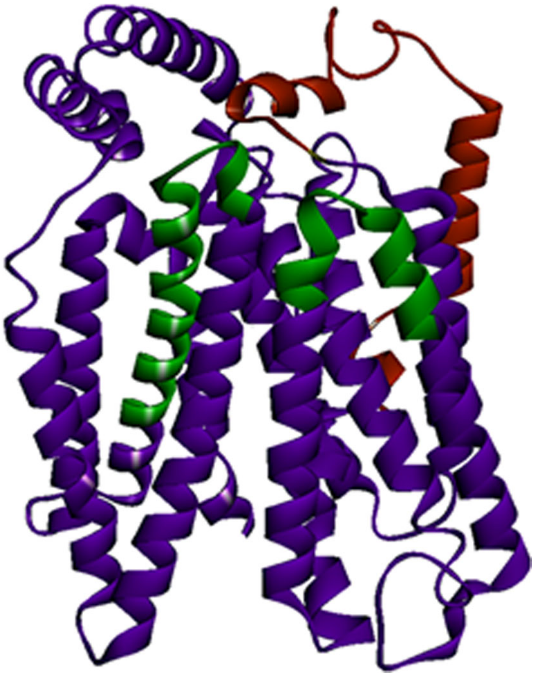

Fig. 1 An example of Pseudogene derived from a protein having functional features. Here, we have taken an example of a Major Facilitator Superfamily (MFS) profile which is shown in *violet colour* and Sugar transport proteins signature 1 and 2 shown in *green colour*. The structure was visualized using Discovery Studio

An example with respect to the functional entity could be attributed to bacterial small MutS region (Smr) like proteins that are linked to DNA repair. These proteins are similar to the C terminal regions of MutS1 proteins which are *bona fide* ATPase domains. In principle, when the Smr proteins were 'hypothetical', there were three prepositions of these proteins, viz. Smr domains alone, the MutS proteins with Smr domains at the C terminal region and the MutS proteins with C-terminal ATPase domains which are quite similar to the Smr domains present in MutS1 proteins with the latter linked to pseudogeny. Can there be new derivatives of Smrs' synthesized in the laboratory with specific sequence inserts of the hypothetical genes-turned-pseudogenes?

To conceptualize the possibility of making a functional protein from pseudogenes, we computationally predicted the structure of a pseudogene derived protein, viz. Major Facilitator Superfamily (Fig. 1). Major Facilitator Superfamily (MFS) proteins are quite common across 17 distinct families within the MFS. Whether or not these proteins corroborate gene duplication events to generate conserved motifs, transmembrane, hypothetical proteins turning out to be pseudogenes is unclear. When the 3D structure of MFS was predicted using I-TASSER, the C-score for the structure was found to be −0.12, which is within the typical range (−5 to 2) for a reliable prediction whereas (negative) total energy (−7,686.523 kcal/mol) for the structure suggests that the structure is a stable one (Zhang 2008).

We argue in support of a need for using sequence and structural homologs to find a good correlation between hypothetical proteins and pseudogenes leading to synthesis of novel proteins. With recent studies on human genome deciphered from nearly hundred novel proteins through pseudogenesis that are linked to cancer, there is a need to construct visual representations of 'pseudogenecity and synthetome' on a cloud of protein interaction networks (Branca et al. 2014). We believe these will help check *bona fidelity* of identifying pseudogenes from hypothetical proteins and address an important question: Is there a need for assessment of protein annotation data to identify the extent of pseudogenicity?

References

Bianchi MM, Sartori G, Vandenbol M, Kaniak A, Uccelletti D, Mazzoni C, Di Rago JP, Carignani G, Slonimski PP, Frontali L (1999) How to bring orphan genes into functional. Yeast 15(6): 513–526

Branca R, Orre L, Johansson H, Granholm V, Huss M, Pérez-Bercoff Å et al (2014) HiRIEF LC-MS enables deep proteome coverage and unbiased proteogenomics. Nat Methods 11(1):59–62

Dhar PK, Thwin CS, Tun K, Tsumoto Y, Maurer-Stroh S, Eisenhaber F, Surana U (2009) Synthesizing non-natural parts from natural genomic template. J Biol Eng 3(3):2

Dujon B (1996) The yeast genome project: what did we learn? Trends Genet 12:263–270

Fortugno LP (2007) Frontiers in Breast Cancer Research. Nova Publishers, New York, p 195

Galperin MY, Koonin EV (2004) Conserved hypothetical' proteins: prioritization of targets for experimental study. Nucleic Acids Res 32(18):5452–5463

Knies K, Schuster B, Ameziane N, Rooimans M, Bettecken T, de Winter J, Schindler D (2012) Genotyping of fanconi anemia patients by whole exome sequencing: advantages and challenges. PLoS ONE 7(12):e52648

Lan L, Djuric N, Guo Y, Vucetic S (2013) MS-kNN: protein function prediction by integrating multiple data sources. BMC Bioinformatics 14(3):S8

Sivashankari S, Shanmughavel P (2006) Functional annotation of hypothetical proteins—A review. Bioinformation 1(8):335–338

Wang J, Zhang D, Li J (2013) PREAL: prediction of allergenic protein by maximum relevance minimum redundancy (mRMR) feature selection. BMC Syst Biol 7(5):S9

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