Pakistan J. Zool., vol. 49(4), pp 1261-1265, 2017. DOI: http://dx.doi.org/10.17582/journal.pjz/2017.49.4.1261.1265

In Silico Analyses of the Pseudogenes of *Helicobacter pylori*

Neenish Rana, Nosheen Ehsan, Awais Ihsan and Farrukh Jamil*
Department of Biosciences, COMSATS Institute of Information Technology, Sahiwal, Punjab, Pakistan

ABSTRACT

Pseudogenes were previously regarded as molecular fossils, non-functional by-products of genome evolution. However, it has been indicated by several lines of evidences that some pseudogenes are active. Using current data of NCBI we have retrieved 65 pseudogenes from the genome sequence of human pathogenic bacteria *Helicobacter pylori* (*H. pylori*) strain 26695. Computational analysis of the genome showed 6 transcriptionally active pseudogenes that can produce stable mRNA secondary structure compared to their functional parents. Moreover it was observed that their putative protein products will be thermodynamically stable. The sequence-based predictions suggested that the pseudogenes-derived proteins may involve in different biological functions like translation, energy metabolism, amino acid metabolism and transport and binding.

Article Information
Received 12 April 2016
Revised 24 August 2016
Accepted 30 January 2017
Available online 28 June 2017

Authors' Contribution
FJ designed the study. NR and NE performed the experiments. AI and FJ collected and analyzed the data. FJ supervised and prepared the manuscript.

Key words
Pseudogenes, *Helicobacter*, Protein structure, mRNA stability, Protein models.

INTRODUCTION

Pseudogenes are genomic loci having sequence homology with other functional genes but they are biologically inactive due to certain aberrations in their sequences like deletions/insertions frameshift mutations and premature stop codons (Welch *et al*., 2015). Therefore, they were referred as genomic fossil or inert genes (Pink *et al*., 2015). However, recent studies have challenged this concept and proposed several different functions for different pseudogenes of unicellular and multicellular organism (Dhar *et al*., 2009; Balakirev and Ayala, 2003; Tariq *et al*., 2016). For example, at the RNA level they can compete with other gene's RNA by interacting with RNA binding proteins (Zheng and Gerstein, 2007), and as proteins they may affect parent or other unrelated enzymes. Therefore they may consequently affect vital metabolic pathways (Zou *et al*., 2009).

Pseudogenes have been categorized into three classes: processed, duplicated and unitary pseudogenes (Rouchka and Cha, 2009). Processed pseudogenes lack introns and they are generated from reverse transcription of their mRNA or integration into silent regions of the genome (Milligan and Lipovich, 2015); whereas duplicated pseudogenes are inactive due to certain disabling features in their regulatory regions such as unfaithful gene duplication, premature stop codons, frame shift mutations or removal of promoter region (Pink *et al*., 2015); while unitary pseudogenes exist with absence of functional counterparts (Zhang *et al*., 2010).

Helicobacter pylori is a human pathogen that exists in the gastric mucosa of human stomach, and it plays a vital role in causing gastric cancer and gastrointestinal disorders (Kusters *et al*., 2006). Its genome sequence (NCBI accession no; GCA_000307795) host 65 pseudogenes out of 1561 total predicted genes. It might be possible that they may produce stable proteins as a study has shown that non coding part of *E. coli* produced stable proteins (Dhar *et al*., 2009). In this frame work, computational analyses of *H. pylori*'s pseudogenes is a step toward understanding possible functions of their derived proteins.

METHODS

The genome of *H. pylori* (GCA_000307795) was analyzed and 65 pseudogenes sequences were retrieved from NCBI database and computationally translated into protein by using Transeq tool of European Bioinformatics Institute (EBI) (Goujon *et al*., 2010; Hoefman *et al*., 2014). Twenty one pseudogene-derived proteins showed significant homology with other functionally active proteins and these were considered for further analyses (*i.e.* predicting sequence based function prediction).

Sequence based function prediction

By using the basic local alignment search tool (BLAST)

* Corresponding author: farrukhccb@gmail.com
0030-9923/2017/0004-1261 $ 9.00/0

functional parents of the pseudogenes were identified (Altschul *et al*., 1990). Strength of the pseudogenes and their relative's promoters was calculated by using BPROM program and expressed in linear discrimination function (LDF) value (Solovyev and Salamov, 2011). Messenger RNA (mRNA) stability was predicted by using RNA fold web server (Zuker and Stiegler, 1981) on the basis of minimum free energy (MFE). ProtFun tool (Jensen *et al*., 2002, 2003) was used for predicting the possible functions of the pseudogene and sub-cellular localization of these proteins was studied by using ProtCompB program; while the physiochemical properties: molecular weights, theoretical isoelectric points, aliphatic index (Ikai, 1980) and hydropathicity (GRAVY) (Kyte and Doolittle, 1982) were predicted by ExPASy ProtParam server (Gasteiger *et al*., 2003). Tertiary structures of pseudogenes encoded proteins and their functional relative proteins were predicted using SWISS-MODEL and I-TASSER server (Zhang, 2008; Biasini *et al*., 2014).

Stability of pseudogene-derived proteins

GROMAS69 force field implemented in Swiss PDB viewer (Guex and Peitsch, 1997) to calculate total energy of the predicted model based on non-bonded and electrostatic constrains. Total cation-π interactions and their energies were calculated using CaPTURE Program (Gallivan and Dougherty, 1999). By using Expasy ProtParam server instability index was calculated.

RESULTS AND DISCUSSION

This study was designed to understand possible roles of the pseudogenes of *H. pylori* by using different computational analysis of the artificially transcribed and translated products of the genes. Analyses of the upstream sequence of the 21 pseudogenes and their functional parents showed that 4 pseudogenes (HP0052, HP0205, HP0502, and HP1522) have a stronger promoter region while 7 pseudogenes (HP0039, HP0041, HP0343, HP0482, HP0505, HP0548, HP0744 and HP0915) host weaker promoters sequence than those of their functional parents (Supplementary Table S1). Analyses showed that 6 pseudogenes (HP0143, HP0369, HP0432, HP0481, HP0548 and HP0679) have 100% sequence identity with 100% query coverage to known proteins of other *H. pylori* strains and their promoters also show similar strengths (Supplementary Table S1). It appears that these genes might be active and they may be poorly annotated.

The expression of the pseudogenes was evaluated on the basis of free energy values (MFE) of the secondary structures of their mRNAs. It has been proposed that highly expressed genes pose less stable mRNA secondary structure, while low expressed genes show more stable secondary structure (Mukund *et al*., 1999; Drummond *et al*., 2005). MFE values of the pseudogenes range from -563.5 to -18.5 kcal/mol and it correspond well to the MFE of their functional parents (-794 to -33.50 kcal/mol) (Supplementary Table S1). Analyses of the data showed that MFE values of the five pseudogenes (HP0143, HP0369, HP0432, HP0481 and HP0679) are same as those of their parent mRNA's. It seems that these 5 genes might be active under certain specific conditions or may transcribe to regulate other parent genes of the organism. There are only three pseudogenes (HP0094, HP0619 and HP0744) that produced more stable mRNA compared to their functional parents (Supplementary Table S1).

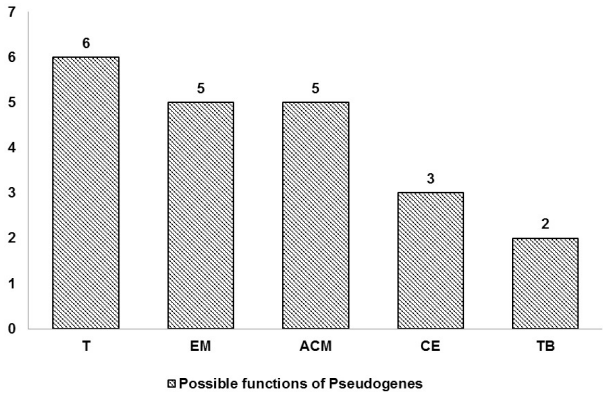

Fig. 1. Function prediction of pseudogene-derived proteins. 6 proteins were predicted to be involved in translation (T), 5 in energy metabolism (EM), 5 in amino acid metabolism (ACM), 3 in cell envelope (CE) while, the functions of 2 proteins predicted in transport and binding (TB).

Table I.- Summary of the stability parameters of selected pseudogenes.

Sequence ID	Stability centers	Instability index	Total energy (kcal/mol)
HP0039 (899692)	0	30.19	-398.3550
HP0041 (899153)	0	37.96	-75.8959
HP0052 (899240)	0	38.07	-249.5048
HP0254 (899058)	0	-2.93	-161.6401
HP0369 (900281)	0	13.82	-151.9675
HP0482 (899253)	22	39.59	-215.8401
HP0505 (899261)	4	11.59	-38.3692

Tertiary structures of the 21 pseudogene-encoded proteins were predicted and their putative functions were obtained from ProtFun tool. Most of the proteins were

Table II.- Physiochemical properties and sub-cellular localization of pseudogenes. This table shows the molecular mass, pI, aliphatic index and sub-cellular localization of the pseudogenes.

Sequence ID	Molecular mass (KDa)	Theoratical pI	Aliphatic Index	GRAVY	Sub-cellular localization
HP0039 (899692)	10.1977	5.82	93.33	-0.017	Inner Membrane
HP0041 (899153)	13.8909	9.27	80.16	-0.639	Periplasm
HP0052 (899240)	41.0625	8.03	79.54	-0.42	Outer Membrane
HP0254 (899058)	4.4624	10.18	77.11	-0.542	Outer Membrane
HP0369 (900281)	6.8228	7.73	63.79	-0.702	Cytoplasm
HP0482 (899253)	19.3029	7.68	88.29	-0.372	Outer Membrane
HP0505 (899261)	5.2963	10.38	129.18	0.347	Outer Membrane

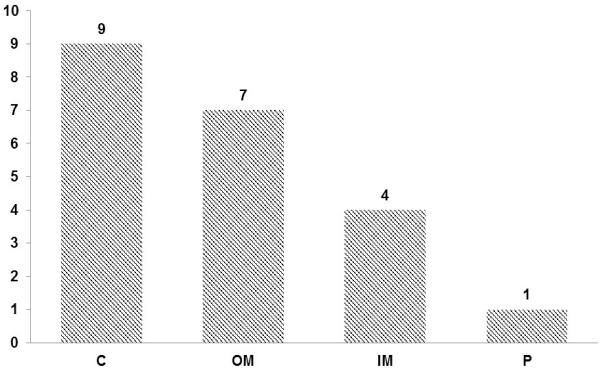

Fig. 2. Sub-cellular localization of pseudogene-derived proteins. 9 proteins showed high potential to be localize in cytoplasm (C), 7 proteins appear to reside in outer membrane (OM), 4 proteins localized to inner membrane (IM) while localization of remaining 1 predicted in periplasm (P).

found to be involved in translation (6 enzymes/proteins), energy metabolism (5), amino acid metabolism (5), transport and binding (2), while 3 proteins showed their potential to be a part of cell envelope (Fig. 1). Analyses showed that 9 proteins have potential to be localized in cytoplasm, 7 in outer membrane, 4 in inner membrane and 1 will be in periplasm (Fig. 2).

Out of 21 only 7 proteins form stable tertiary structures (Supplementary Table S2) as the overall energy of the protein was -38 to -398 kcal/mole and instability index was found to be less than 40 (Table I). This suggested *in vivo* stability of these proteins as stability index below 40 is considered as a good evidence of stability and it shows that proteins will be stable *in vivo* (Guruprasad *et al*., 1990). The physiochemical parameters like molecular masses of these stable proteins were determined that range from 4.46 to 41.06 KDa, suggesting the presence of different size proteins (Table II). The isoelectric point (pI) values of the proteins vary from 5.82 to 10.38 that indicated acidic nature of only one protein (HP0039m) and basic nature of 6 proteins (HP0041, HP0052, HP0254, HP0369, HP0482, HP0505), and the aliphatic index value of the proteins ranges from 63.79 to 129.18 (higher the value, should greater the stability of protein). The hydropathicity value (GRAVY score) showed that six proteins are hydrophobic in nature while one is hydrophilic (Table II). These stable proteins may be significant for the microorganism and may be expressed under specific conditions.

CONCLUSION

In conclusion, our study identifies 6 pseudogenes in *H. pylori* that appears to be active genes as they are 100% identical to other functional genes in other strains of *H. pylori*. Overall, we have identified 7 pseudogenes that may produce stable proteins, however further studies are required to explore exact role of these proteins.

ACKNOWLEDGEMENT

We gratefully acknowledge Higher Education Commission (HEC) of Pakistan for grants to establish Bioinformatics research laboratory at COMSATS, Sahiwal.

Supplementary material

There is supplementary material associated with this article. Access the material online at: http://dx.doi.org/10.17582/journal.pjz/2017.49.4.1261.1265

Statement of conflict of interest

Authors have declared no conflict of interest.

REFERENCES

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and

Lipman, D.J., 1990. Basic local alignment search tool. *J. mol. Biol.*, **215**: 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Balakirev, E.S. and Ayala, F.J., 2003. Pseudogenes: Are they "junk" or functional DNA? *Annu. Rev. Genet.,* **37**:123–151. https://doi.org/10.1146/annurev.genet.37.040103.103949

Biasini, M., Bienert, S., Waterhouse, A., Arnold, K., Studer, G., Schmidt, T., Kiefer, F., Cassarino, T.G., Bertoni, M., Bordoli, L. and Schwede, T., 2014. SWISS-MODEL: Modelling protein tertiary and quaternary structure using evolutionary information. *Nucl. Acids Res.*, **42**: W252–258. https://doi.org/10.1093/nar/gku340

Dhar, P.K., Thwin, C.S., Tun, K., Tsumoto, Y., Maurer-Stroh, S., Eisenhaber, F. and Surana, U., 2009. Synthesizing non-natural parts from natural genomic template. *J. biol. Engin.*, **3**: 2.

Drummond, D.A., Bloom, J.D., Adami, C., Wilke, C.O. and Arnold, F.H., 2005. Why highly expressed proteins evolve slowly. *Proc. natl. Acad. Sci. U.S.A.*, **102**: 14338–14343. https://doi.org/10.1073/pnas.0504070102

Gallivan, J.P. and Dougherty, D.A., 1999. Cation-pi interactions in structural biology. *Proc. natl. Acad. Sci. U.S.A.*, **96**: 9459–9464. https://doi.org/10.1073/pnas.96.17.9459

Gasteiger, E., Gattiker, A., Hoogland, C., Ivanyi, I., Appel, R.D. and Bairoch, A., 2003. ExPASy: The proteomics server for in-depth protein knowledge and analysis. *Nucl. Acids Res.*, **31**: 3784–3788. https://doi.org/10.1093/nar/gkg563

Goujon, M., Mcwilliam, H., Li, W., Valentin, F., Squizzato, S., Paern, J. and Lopez, R., 2010. A new bioinformatics analysis tools framework at EMBL–EBI. *Nucl. Acids Res.*, **38**: W695–W699. https://doi.org/10.1093/nar/gkq313

Guex, N. and Peitsch, M.C., 1997. SWISS-MODEL and the Swiss-Pdb Viewer: an environment for comparative protein modeling. *Electrophoresis*, **18**: 2714–2723. https://doi.org/10.1002/elps.1150181505

Guruprasad, K., Reddy, B.V. and Pandit, M.W., 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 Engin.*, **4**: 155–161. https://doi.org/10.1093/protein/4.2.155

Hoefman, S., van-der-Ha, D., Boon, N., van Damme, P., de-Vos, P. and Heylen, K., 2014. Niche differentiation in nitrogen metabolism among methanotrophs within an operational taxonomic unit. *BMC Microbiol.*, **14**: 83. https://doi.org/10.1186/1471-2180-14-83

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

Jensen, L.J., Gupta, R., Blom, N., Devos, D., Tamames, J., Kesmir, C., Nielsen, H., Staerfeldt, H.H., Rapacki, K., Workman, C., Andersen, C.A., Knudsen, S., Krogh, A., Valencia, A. and Brunak, S., 2002. Prediction of human protein function from post-translational modifications and localization features. *J. mol. Biol.*, **319**: 1257–1265. https://doi.org/10.1016/S0022-2836(02)00379-0

Jensen, L.J., Gupta, R., Staerfeldt, H.H. and Brunak, S., 2003. Prediction of human protein function according to gene ontology categories. *Bioinformatics*, **19**: 635–642. https://doi.org/10.1093/bioinformatics/btg036

Kusters, J. G., van Vliet, A.H.M. and Kuipers, E.J., 2006. Pathogenesis of *Helicobacter pylori* Infection. *Clin. Microbial. Rev.*, **19**: 449-490. https://doi.org/10.1128/CMR.00054-05

Kyte, J. and Doolittle, R.F., 1982. A simple method for displaying the hydropathic character of a protein. *J. mol. Biol.*, **157**: 105–132. https://doi.org/10.1016/0022-2836(82)90515-0

Milligan, M.J. and Lipovich, L., 2015. Pseudogene-derived lncRNAs: Emerging regulators of gene expression. *Front. Genet.*, **5**: 476. https://doi.org/10.3389/fgene.2014.00476

Mukund, M.A., Bannerjee, T., Ghosh, I. and Datta, S., 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.

Pink, R.C., Wicks, K., Caley, D.P., Punch, E.K., Jacobs, L. and Carter, D.R.F., 2015. Pseudogenes: Pseudo-functional or key regulators in health and disease? *RNA*, **17**: 792–798. https://doi.org/10.1261/rna.2658311

Rouchka, E.C. and Cha, I.E., 2009. Current trends in pseudogene detection and characterization. *Curr. Bioinform.*, **4**: 112–119. https://doi.org/10.2174/157489309788184792

Solovyev, V. and Salamov, A., 2011. Automatic annotation of microbial genomes and metagenomic sequences. In: *Metagenomics and its applications in agriculture, biomedicine and environmental studies* (ed. R.W. Li), Nova Science Publishers, pp. 61-78.

Tariq, F., Khalid, Q., Sehgal, S.A., Mannan S. and Jamil, F., 2016. Possible roles of the pseudogenes of *Salmonella* Typhimurium. *Pakistan J. Zool.,* **48**:

1805–1810.

Welch, J.D., Baran-Gale, J., Perou, C.M., Sethupathy, P. and Prins, J.F., 2015. Pseudogenes transcribed in breast invasive carcinoma show subtype-specific expression and ceRNA potential. *BMC Genom.*, **16**: 113. https://doi.org/10.1186/s12864-015-1227-8

Zhang, Y., 2008. I-TASSER server for protein 3D structure prediction. *BMC Bioinform.*, **9**: 40. https://doi.org/10.1186/1471-2105-9-40

Zhang, Z.D., Frankish, A., Hunt, T., Harrow, J. and Gerstein, M.B., 2010. Identification and analysis of unitary pseudogenes: historic and contemporary gene losses in humans and other primates. *Genom. Biol.,* **11**: R26. https://doi.org/10.1186/gb-2010-11-3-r26

Zuker, M. and Stiegler, P., 1981. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucl. Acids Res.*, **9**: 133–148. https://doi.org/10.1093/nar/9.1.133

Zheng, D. and Gerstein, M.B., 2007. The ambiguous boundary between genes and pseudogenes: the dead rise up, or do they? *Trends Genet.,* **23**: 219–224. https://doi.org/10.1016/j.tig.2007.03.003

Zou, C., Lehti-Shiu, M.D., Thibaud-Nissen, F., Prakash, T., Buell, C.R. and Shiu, S.H., 2009. Evolutionary and expression signatures of pseudogenes in *Arabidopsis* and rice. *Pl. Physiol.*, **151**: 3–15. https://doi.org/10.1104/pp.109.140632