Acta Spartae Vol. 2 No. 1 2016
Pages 1–4

Predicting the Function of Hypothetical Protein PANDA_003700 Partial Using Computational Analysis Methods

Cameron Bixby and Padmanabhan Mahadevan[1]
Department of Biology, University of Tampa, Tampa, FL 33606, [1]Faculty Advisor

ABSTRACT

Legions of hypothetical proteins are currently awaiting accurate characterization of their functions, but due to the influx of sequence data, the rate of analysis is not able to keep up with the rate of discovery of hypothetical proteins. However, various computational methods such as, Pfam, BLAST, and Swiss Homology Modeling are helping researchers predict the function of hypothetical proteins. Instead of using only experimental methods which are time consuming and difficult, computational methods are helping pave the way for faster hypothetical protein analysis. In this study, a series of computational tests were performed in order to predict the function of hypothetical protein PANDA_003700 partial (EFB18608.1). The predicted function of the hypothetical protein was found to be that of an mRNA turnover protein 4 which is involved in ribosomal assembly.

1 INTRODUCTION

The majority of the gene products produced after an organism is sequenced are proteins whose function is not known, called hypothetical proteins (HPs). Proteins that are predicted from nucleic acid sequences and proteins with unknown functions are considered hypothetical proteins (Lubec et al., 2005). Therefore as large amounts of hypothetical proteins are discovered from genomic sequencing, they will continue to enter the spotlight of many studies in the Bioinformatics and Genomics field. About half of the proteins in most genomes are candidates for HPs (Lubec et al., 2005). Therefore, determining the function of the HPs is very important when trying to complete the genomic and proteomic information of a sequenced organism. HPs are observed across a variety of phylogenetic lineages but their functions are not characterized (Galperin & Koonin, 2004).Therefore, the challenge to characterize the function of HPs using experimental and computational methods has become more important in Genomic studies.

Typically, the work dedicated to discovering the functions of HPs can be separated into two parts: prediction of protein function through its sequence and prediction of the 3-D structure. In terms of predicting the function of a HP through its sequence, researchers will use computational methods in order to compare their HP against functional proteins in hope of high sequence similarities. In finding the similarities between sequences, researchers can infer the function of the protein, explore protein families and evolutionary relationships (Lubec et al., 2005). The most common tool in calculating sequence similarity is the Basic Local Alignment Search Tool (BLAST) which has a version that can blast a protein query against a database of proteins. Exploration of various protein families to see if the HP shares any common evolutionary origin is another route taken by researchers in order to gather more information on their HP. Protein families are sets of protein regions which share a significant degree of sequence similarity (Punta et al., 2012). Therefore using databases like Pfam can display various relationships between a HP and other functional proteins. It is also important to mention that protein domains are also considered another area in which a researcher can use the HP sequence to discover its domains. Protein domains are viewed as the basic components of proteins and from this it helps determine the functional characterization (Veretnik at al., 2004).

The sequence of a hypothetical protein can provide a lot of insight in terms of the prediction of the protein structure itself which can then further help determine the function of the HP. This ties in with the goal of structural genomics which is to create a complete inventory of protein folds/structures that can help predict functions for all proteins (Mittl & Grütter, 2001). One way to determine the 3-D structure of a protein is by either x-ray crystallography or NMR, then the structure can be compared against other structures in a protein database (Zarembinski et al., 1998). However, those experimental methods are usually difficult, complex, and time consuming. Therefore, with limited experimental models of proteins to compare with a HP, homology modeling has become a reliable way to determine the 3-D structure by using a HPs amino acid sequence. It is important to mention that homology searches are more accurate if the sequence similarity between the HP and the homolog of another known protein is greater than 30%. Overall, stronger the sequence similarity of a HP to other functional proteins, the likelihood of predicting its structure and function increases tremendously.

Computational tools have allowed researchers to generate more information about HPs in sequenced genomes across various organisms such as, mammals. Hypothetical proteins constitute a large portion of mammalian proteomes (Lubec et al., 2005) which can possibly reveal inferred evolutionary relationships between other mammals. Therefore, a hypothetical protein was chosen at random from the sequenced genome of the Giant Panda (*Ailuropoda melanoleuca*). The HP chosen was PANDA_003700 partial and in an effort to gain insight into the process of determining the function of this HP, various computational analysis methods were implemented in this study. It was hypothesized that the function of the HP from the Giant Panda could be determined with the use of various genomic computational analysis methods. Therefore, the results from this study could provide new insight to the function of hypothetical protein PANDA_003700 partial from the *Ailuropoda melanoleuca* sequenced genome.

2 MATERIALS AND METHODS

A search for "*Ailuropoda melanoleuca* hypothetical protein" in the online GenBank database was conducted to generate a list of random hypothetical proteins for that sequenced organism. A random

Protein accession number and organism name	Percent id	E value
EFB18608.1_ hyp. Prot. PANDA_003700 [*Ailuropoda melanoleuca*]	100	2.59E–159
XP_002915673.1_ Mrt4 [*Ailuropoda melanoleuca*]	98.122	1.57E–157
XP_012416701.1_ Mrt4 [*Odobenus rosmarus divergens*]	98.122	7.19E–157
XP_006732520.1_ Mrt4 [*Leptonychotes weddellii*]	97.653	1.12E–156
XP_008692886.1_ Mrt4 [*Ursus maritimus*]	97.653	1.41E–156
XP_544532.2_ Mrt4 [*Canis lupus familiaris*]	96.262	4.78E–154
XP_007083826.1_ Mrt4 [*Panthera tigris altaica*]	96.244	8.03E–154
XP_014937071.1_ Mrt4 [*Acinonyx jubatus*]	95.775	6.11E–153
XP_003989662.1_ Mrt4 [*Felis catus*]	95.305	2.78E–152
XP_012512390.1_ Mrt4 [*Propithecus coquereli*]	93.897	4.82E–152
XP_002750411.1_ Mrt4 [*Callithrix jacchus*]	93.868	2.85E–150
XP_004603443.1_ Mrt4 [*Sorex araneus*]	91.628	4.71E–148
XP_012003767.1_ Mrt4 [*Ovis aries musimon*]	92.019	1.16E–147
XP_002716054.1_ Mrt4 [*Oryctolagus cuniculus*]	91.08	1.52E–146
XP_010343819.1_ Mrt4 [*Saimiri boliviensis boliviensi*]	91.038	2.18E–146
XP_008820052.1_ Mrt4 [*Nannospalax galili*]	91.08	3.49E–146
EHH49593.1_ hypo. Prot. EGM [*Macaca fascicularis*]	90.141	3.73E–145
XP_006068171.1_ Mrt 4 [*Bubalus bubalis*]	94.527	2.93E–144

Table 1. The Top 17 BLAST hits observed for the hypothetical protein sequence (EFB18608.1).

hypothetical protein with an amino acid length between 100–300 amino acids was then selected (Benson et al., 2007). The *Ailuropoda melanoleuca* hypothetical protein (EFB18608.1) selected from the GenBank search was blasted using a protein blast. The search set selected for the protein blast was Non-redundant protein sequences (nr) (McGinnis & Madden, 2004). Using the fasta sequence from the selected *Ailuropoda melanoleuca* hypothetical protein selected from Genbank, a search against PROSITE signatures was conducted (Hulo et al., 2004).

Using the HPs accession number EFB18608.1, it was searched in the UniProt database for a possible Entry ID and description (Apweiler et al., 2004). The fasta sequence from the *Ailuropoda melanoleuca* hypothetical protein selected from Genbank was used in a Pfam sequence search (Bateman et al., 2002). A homology model was built using the *Ailuropoda melanoleuca* hypothetical protein in Swiss Model (Schwede et al., 2003).

The selected *Ailuropoda melanoleuca* hypothetical protein (EFB18608.1) sequence from GenBank was uploaded onto the Phyre server using the Normal modeling mode (Kelley & Sternberg, 2009). Using the PDB file generated from the Swiss Model server, it was uploaded onto the Profunc server for further analysis (Laskowski, Watson & Thornton, 2005)

All the fasta sequences from each of the blast hits within an e–140 parameter were selected and aligned on the MAFFT server. Under the Uppercase/Lowercase section on MAFFT, the parameter chosen was 'same as input'. The other parameters were left in their default format. The fasta formatted file was then converted into MEGA. The new .meg file was then analyzed by creating a Neighbor joining tree. A Bootstrap method was selected and the number of bootstrap values chosen was 500 (Kumar et al., 2008).

3 RESULTS

Only the top 17 blast hits are displayed in Table 1. All of the top blast hits had an e value of 144 or higher and sequence identity of 90% or higher. This suggests that the matches were very similar based on these values and could be considered significant in regards to the PANDA_003700 partial hypothetical protein.

No hits that corresponded with any of the Prosite signatures were observed for the hypothetical protein PANDA_003700 partial. The Pfam result found a protein family match to the Ribosomal protein L10 family for the hypothetical protein. This is a significant match according to the e value of 4.7e–1

After the hypothetical protein had been entered into Uniprot, an entry ID was found (D2H270). Looking further into the entry ID, it listed the hypothetical protein as part of the Ribosomal protein L10 family. The subcellular location was the ribosome, and the proteins amino acid length was 213. This corresponds with the Pfam result which also found a match for the Ribosomal L10 family.

The Swiss Model server produced a protein model of the PANDA hypothetical protein using the template information from the website. The coverage was 91% and the sequence similarity was 42%. The model that was produced was listed as an mRNA turnover protein 4. The results suggest that the model quality is reasonable based on various metric values such as, solvation and torsion (Figure 1). The greater the values are towards the blue coloring, the better the quality of the model.

Q-score	Z-score	Name
0.764	18.6	Crystal Structure of Mrt4
0.299	5.8	N- terminal fragment of ribosomal protein L10
0.232	4.5	pre-translocation 70s tRNA structure
0.232	4.5	Pre-translocation 70s tRNA
0.215	3.7	70s ribosome

Table 2. Fold Results generated by Profunc of the model created by Swiss Model.

Confidence	ID %	Template Information
100	42	mRNA turnover protein 4
100	17	acidic ribosomal protein p0 homology
100	34	60s ribosomal protein I28
100	21	50s ribosomal protein I14e
100	24	60s ribosomal protein I18a

Table 3. The top 5 templates observed from the Phyre Results for hypothetical protein PANDA_003700 partial.

The results from Profunc revealed that the fold search for the best match to be the Mrt4 protein crystal structure (Table 2). This again agrees with the BLAST results.

The confidence and coverage percentages obtained from the Phyre server for the hypothetical protein suggested that the fold of the hypothetical protein is very similar to that of Mrt4 protein. This suggests that the function of the hypothetical protein is likely to be similar to that of the Mrt4 protein. The top 5 models were selected based on the raw alignment quality and aligned residues. The number one template had a confidence of 100% (Table 3) and an identification percentage of 42. The other four templates had lower identification percentages, but their confidence levels were still 100% (Table 3). Due to the low identification percentages of models #2–4, model #1 was the chosen model of focus.

Figure 1. The bar chart showing model quality metrics produced by the Swiss Model server.

A phylogenetic tree (not shown) was analyzed to help compare the hypothetical protein against the top BLAST hits in order to help create an inferred evolutionary relationship. The overall robustness of the tree was fairly low, with only 43% of all the nodes having a bootstrap value over 50%. This means that more than half were not considered robust. Some of the internal nodes had bootstrap values up to 99%, such as the node between the *Equus przewalskii* (XP 008540202.1) and *Equus callibus* (XP 001501832.2).

4 DISCUSSION

The results of the subsequent computational tests and analyses supported the hypothesis that the function for the hypothetical protein PANDA_003700 could be predicted. This study also helped suggest an inferred evolutionary relationship between other organisms and the Ailuropoda melanoleuca HP. After analyzing the protein BLAST results, the top 17 hits were selected and it was observed that majority of those hits were predicted sequences of an mRNA turnover protein 4 (Mrt4 protein) (Table 1). The results observed in the protein blast were also very accurate considering that the lowest e-value was 7e-144 and the lowest sequence identity was 90% out of all the 17 selected hits (Table 1). The Higher the e-value and sequence identity observed, the more confident the result that was generated from BLAST. As for the function of an Mrt4 protein, it serves as a component in the ribosomal assembly machinery (Rodríguez-Mateos et al., 2009). It has been found that the Mrt4 protein have shown extensive similarity to the ribosomal P0 protein, therefore it is classified as a P0-like protein since it has influence on assembly of the pre-60S particle (Michalec et al., 2010). This allows further interpretation that the HP under study is involved in ribosomal assembly machinery based on sequence similarity from the protein BLAST results.

The sequence of the HP PANDA_003700 partial was analyzed through Pfam to see if any protein families were observed to better predict the function of the HP. There was one hit which was the Ribosomal L10 family, or more specifically a Ribosomal protein L10. The Ribosomal protein L10 gene encodes a ribosomal protein that is part of the 60s subunit. The e-value was 4.7e–19 which is considered significant, therefore it can be inferred once again that the HP being studied participates in ribosomal assembly machinery. The HP sequence was also uploaded to Prosite, but no hits were observed. Although no hits were observed for the Prosite database, Prosite also contains protein families similar to Pfam. Therefore, in this study it was more useful to use Pfam since the Ribosomal L10 protein family was observed as an accurate hit.

Despite the dubious results from the Prosite database, the HP PANDA_003700, partial did have an active identification in UniProt, which D2H270. Looking further into its entry ID on UniProt, the listed description was ribosome biogenesis. This further suggests, that the predicted HPs function has a role in ribosomal assembly machinery, and more specifically, it being an Mrt4 protein. The structure of the HP was also predicted in order to develop a better understanding about the function of the protein

The structure of the Mrt4 protein was produced by the Phyre results generated a PDB molecule called an mRNA turnover protein 4 with a confidence level of 100% and had a coverage of 91%. Based on this information, it can be predicted that this is a possible structure of the HP due to the 100% confidence level and high coverage value. The first fold template represented shows the PDB molecule is likely an Mrt4 protein, but the identity percentage was below 50% (Table 3). As for the other 4 templates represented in Table 3, their PDB molecules were all ribosomal proteins, but their identity percentage was 35% or lower. Although the percentages of the domains were not above 50%, that does not mean that all

the fold templates were useless in building the secondary structure of the HP. Many proteins will have a similar fold even if they are distantly related. However, the first and third templates were the most useful due to them having an identity percentage higher than 30% with the HP. Therefore, the use of them as templates to predict the 3-D structure for the HP PANDA_003700 partial was acceptable. It is also important to mention the Phyre server detects twice as many remote homologies as standard sequence-profile searching (Kelley & Sternberg, 2009), so the server will typically contain higher amounts of lower percentage templates that will be used to predict the secondary structure. Regardless, the top 5 templates used to build the secondary structure of the HP were all observed to be involved in ribosomal assembly as ribosomal proteins which further supports the HPs involvement in the ribosomal machinery as an Mrt4 protein.

In addition to the Phyre testing, Swiss-Homology modeling was used to generate a 3-D molecule using the sequence of the HP PANDA_003700 partial which supported its possible identity as an Mrt4 protein. Similar to the Phyre server, the Swiss-model had a 91% coverage and the title was an Mrt4 protein. However, the sequence similarity was below 50% which suggests there were a decent amount of errors in producing the 3-D model. Therefore, the comparative structure could not be compared to a 3-D structure produced from experimental methods such as, X-ray crystallography. However, the Swiss-models overall quality (Figure 1) can be used to help infer the 3-D structure of the HP PANDA_003700 partial and determine the molecular basis of the HP. The Profunc results also helped back up the model produced from the Swiss-Homology modeling. In terms of fold matches, the highest fold matches were all participants in ribosomal structure or machinery (Table 2). The top match being a crystal structure of an Mrt4 protein.

Overall, the majority of the results from this study suggest that the function of the HP PANDA_003700 partial, is that of an Mrt4 protein which is involved in ribosomal assembly machinery. It also suggests that the same Mrt4 protein found in the *Ailuropoda melanoleuca* has been observed in other eukaryotic organisms such as *Ursus maritimus* (XP 008692886.1) and *Odobenus rosmarus* (XP 012416701.1) (Table 1). However, further experimentation should be conducted in order to prove with 100% confidence that the HP analyzed was indeed an mRNA turnover protein 4. Therefore, more experiments using computational or experimental methods should be conducted in order to further support the results reported in this study.

REFERENCES

Apweiler, R., Bairoch, A., Wu, C.H., et al. 2004, Nucleic Acids Res, 32, D115–9

Bateman, A., Birney, E., Cerruti, L., et al. 2002, Nucleic Acids Res, 30(1), 276–80

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., et al. 2007, Nucleic Acids Res, 35, D21–5

Galperin, M.Y. & Koonin, E.V. 2004, Nucleic Acids Res, 32(18), 5452–5463

Hulo, N., Sigrist, C.J.A., Le Saux, V., et al. 2004, Nucleic Acids Res, 32, D134–7

Kelley, L.A. & Sternberg, M.J.E. 2009, Nat Prot, 4(3), 363 71

Kumar, S., Nei, M., Dudley, J., et al. 2008, Brief. Bioinform, 9(4), 299–306

Laskowski, R.A., Watson, J.D., & Thornton, J.M. 2005, Nucleic Acids Res, 33, W89–93

Lubec, G., Afjehi-Sadat, L., Yang, J., et al. 2005, Prog Neurobiology, 77(12), 90–127

McGinnis, S. & Madden, T.L. 2004, Nucleic Acids Res, 32, W20–5

Michalec, B., Krokowski, D., Grela, P., et al. 2010, Int J Biochem Cell Bio, 42(5), 736–48

Mittl, P.R.E. & Grütter, M.G. 2001, Curr Opin Chem Bio, 5(4), 402–8

Punta, M., Coggill, P.C., Eberhardt, R.Y., et al. 2012, Nucleic Acids Res, 40 (D1): D290–D301

Rodríguez-Mateos, M., Abia, D., Garca-Gómez, J.J., et al. 2009, Nucleic Acids Res, 37(11), 3514–3521

Schwede, T., Kopp, J., Guex, N., et al. 2003, Nucleic Acids Res, 31(13), 3381–5

Veretnik, S., Bourne, P.E., Alexandrov, N.N., et al. 2004, J Mol Biol, 339(3)647–78

Zarembinski, T.I., Hung, L.W., Mueller-Dieckmann H.-J., et al. 1998, Proc Nat Acad Sci U.S., 95(26), 15189–15193