EMBnet.journal
Bioinformatics in Action

Bouzin et al. (2019) EMBnet.journal 24, e925
http://dx.doi.org/10.14806/ej.24.0.925

Technical Notes

PyFuncover: full proteome search for a specific function using BLAST and PFAM

Yoan Bouzin[1], Benjamin Thomas Viart[1], María Moriel-Carretero[2], Sofia Kossida[1] ✉

[1]IMGT®, IGH, Univ Montpellier, CNRS, Montpellier, France
[2]CRBM, Univ Montpellier, CNRS, Montpellier, France
Competing interests: YB none; BTV none; MMC none; SK none

Abstract

Python Function uncover (PyFuncover) is a new bioinformatic tool able to search proteins with a specific function in a full proteome. The pipeline coded in python uses BLAST alignment and the sequences from a PFAM family as the search seed. We tested PyFuncover using the fatty acid-binding family (FABP) Lipocalin_7 from PFAM (version 32.0, September 2018) against the Homo sapiens NCBI proteome. After applying the scoring function in all the BLAST results, the data were classified and submitted to a GO-TERM analysis using bioDBnet. Analyses showed that all families of FABPs were ranked within the top scores. Included within this category were also families able to bind to hydrophobic molecules similar to fatty acids such as the retinol acid transporter and the cellular retinoic acid-binding protein.
Availability: PyFuncover source code is freely available at **https://github.com/Tuisto59/PyFuncover/** under the GPL licence.

Introduction

High-throughput technologies produce massive amount of data and bioinformatics approaches help predict and annotate protein function using increasingly complex and precise methods. One example is the NCBI annotation pipeline (Thibaud-Nissen *et al.*, 2016). The human genome sequence was released in 2003 but the annotation of the human proteome in January of 2018 (GRCh38.p12) still contains 2,404 uncharacterised proteins (out of 113,620). Protein families for which the relationship between sequence and function is more complex pose the most significant challenges. The enzymes are particularly tricky because only a small part of the protein is responsible for its function. Moreover, specific binding motifs for which knowledge is still partial and poorly annotated add up to this category.

In 2011 a tool called Ada-BLAST was published and used to predict a fatty acid-binding motif in the human protein BRCA1 (Hedgepeth *et al.*, 2015) and the horse Oxy-myoglobin (Patterson *et al.*, 2011), revealing in those already well-known proteins a new property. Today, this tool is no longer available. Inspired by the methodology explained in (Hong *et al.*, 2009; Patterson *et al.*, 2011; Dae Ko *et al.*, 2011; Hedgepeth *et al.*, 2015; Chintapalli *et al.*, 2015), we created PyFuncover.

PyFuncover is a pipeline able to rank each protein from a proteome according to a specific Protein FAMily (PFAM) (El-Gebali *et al.*, 2019). As a proof of concept, we used this tool to find proteins with putative fatty acid-binding property in the human proteome. We used as a seed the Lipocalin_7 domain family (PF14651[1]).

Workflow

To study a specific activity and to identify other proteins with potentially similar function, the first step is to recover a large set of protein sequences using as a seed the protein annotated with the desired function. Each chosen sequence will make ten iterations (PSI-BLAST accepts a list of multiple sequences, but only the first sequences are used) (see Figure 1, blue box).

Specific family sequences can be downloaded from the PFAM database as a multiple sequences alignment (MSA) from NCBI[2] or UniProt[3] using various formats. Each sequence has a header containing the protein accession, followed by a slash and the domain boundary. The accession of all PSI-BLAST reports is compiled, and each PFAM accession is checked if it is included in the PSI-BLAST results. Other sequences can be from a close family to the chosen one or belong to the same PFAM family. Gaps from the MSA are removed (see Figure 1, green boxes), and a BLAST database is made (see Figure 1, black box).

A whole proteome dataset can be downloaded or any set of proteins in FASTA format (see Figure 1, orange

Article history
Received: 04 February 2019
Accepted: 02 March 2019
Published: 25 April 2019

[1]https://pfam.xfam.org/family/Lipocalin_7
[2]https://www.ncbi.nlm.nih.gov/
[3]https://www.uniprot.org/

Bouzin et al. (2019) EMBnet.journal 24, e925
http://dx.doi.org/10.14806/ej.24.0.925

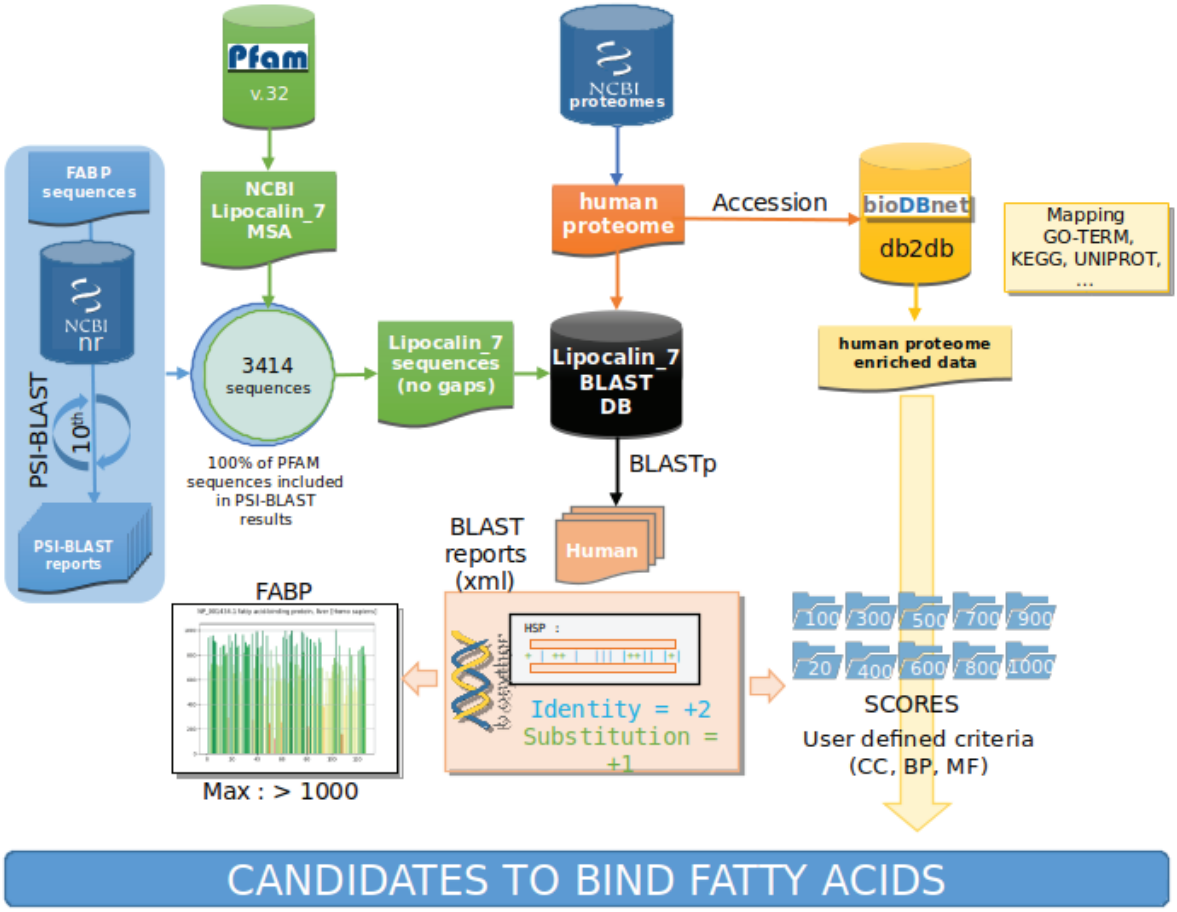

Figure 1. Workflow of PyFuncover.

box). For each sequence, a BLASTp is performed against the PFAM BLAST DB (see Figure 1, black box). For each protein (subject) that matches our sequence (query), BLAST produces alignments, called High Scoring Pairs (HSPs). A score of two, for all the identities, and a score of one, for all positive substitutions, is computed for each alignment. Accession numbers from NCBI are used to retrieve data from different databases (GO-Terms, UniProt, KEGG, PDB, BioCyc, Ensembl, GenBank…) (Ashburner *et al.*, 2000; Berman *et al.*, 2000; Clark *et al.* , 2016; Kanehisa et al., 2019; Karp et al., 2017; UniProt Consortium, 2018; Zerbino *et al.*, 2018) using the cross-reference database web application BioDBnet (db2db) (Mudunuri *et al.*, 2009) and compiled into a biologist-friendly table. This makes the results easy to open and parse using a spreadsheet software such as Excel.

Proof of concept

To test PyFuncover, we used a selection of human Fatty Acid-Binding Proteins (FABPs) (Table 1). The FABPs are part of the lipocalin_7 family (PF14651). The accession numbers of the 3414 sequences from the MSA of NCBI were compared with all the PSI-BLAST results. All the sequences were included into the PSI-BLAST results, and MSA were used to make the BLAST database. Using CDD-Search (Marchler-Bauer *et al.*, 2017), we checked the accessions of the PSI-BLAST reports. The accessions corresponded to the PFAM Lipocalin_7 or to the lipocalins 4 and 5 as expected since all three are members of the Calycin superfamily. The human proteome was downloaded to perform a BLASTp against the database made from the MSA. The XML reports were parsed using BioPython (Cock *et al.*, 2009).

Each amino-acid of each protein obtains a score. Scores can be represented as a barplot for visual analysis (Figure 2). Proteins were split into ten folders (from 100 up to 1000) based on its highest scored amino acid (Figure 2). For the FABPs input set, the highest score was 1052 for FABP7 (isoform X4, NP_001305971). Human

Table 1. List of the FABP used for the PSI-BLAST run.

FABP	UNIPROT Accession
FABP1	P07148
FABP2	P12104
FABP3 (FABP11)	P05413
FABP4	P15090
FABP5	Q01469
FABP6	P51161
FABP7	O15540
FABP8 (PMP2)	P02689
FABP9	Q0Z7S8
FABP12	A6NFH5

Figure 2. Score histogram per amino acid along the sequence of the FABP1 (isoform X1) of Homo sapiens. Colour range stands from less in red to high in green.

proteomes accession numbers were crossed with the GO-TERM database, using BioDBnet (see Figure 1, yellow box).

Considering the proteins with a score above 900 (arbitrarily chosen), we found members of all the nine FABPs families (Table 2). Above this threshold, we also found five (Cellular) Retinol-Binding Proteins (CRBPs) and two (Cellular) Retinoic Acid-Binding Proteins (CRABPs). This is remarkable, because FABPs, CRBPs and CRABPs are all three subfamilies of the intracellular Lipid-Binding Proteins (iLBPs) family. Moreover both retinol and retinoic acid display a partially similar structure to that of fatty acids (Smathers and Petersen, 2011). As expected the FABP1 family is ranked first using highest mean amino-acid score reaching 735 (Figure 3).

Conclusions

The dataset with a score above 900 contains the top one per cent of the input or 1,530 proteins. This number dramatically exceeds that described above as a proof of concept. This tool aims at helping biologists investigate their favourite set of proteins with a simple sequence-function scoring method. PyFuncover output table combines protein identification, score and several useful databases cross-references for handy investigation. Additionally, while we used it here to detect putative fatty acids-binding motifs, PyFuncover can be tailored to search other functional features matching the user's wishes.

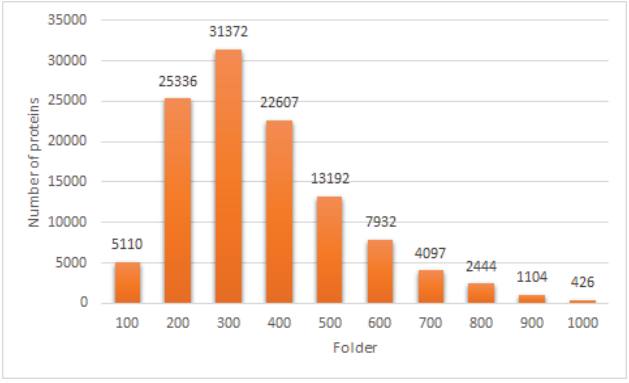

Figure 3. Number of protein in each folder.

Table 2. Proteins implicated in the binding of fatty acids and related hydrophobic molecules from Homo sapiens found in the 900 and 1000 folders.

Protein	Score 1000	Score 900
FABPs	FABP1, 3, 7, 8, 12	FABPS2, 3, 4, 5, 6, 7, 9
RBPs	RBP1, 5, 7	RPB2, 5
CRABPs		CRABP1, 2

Key Points

- PyFuncover is a new bioinformatic tool to search proteins with a specific function in a full proteome.
- Using the Lipocalin 7 family as input we observed in the top-ranked proteins all families of FABPs as well as families able to bind to hydrophobic molecules similar to fatty acids.
- PyFuncover output table combines protein identification, score and several useful databases cross-references for handy investigation.
- This tool aims at helping biologists investigate their favorite set of proteins with a simple sequence-function scoring method.

Acknowledgements

This work was supported by Merck Sharp and Dohme Avenir (GnoSTic) to S. Kossida and by the ATIP-Avenir program to M. Moriel-Carretero.

References

1. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. **25** (1), 25–29. http://dx.doi.org/10.1038/75556
2. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, *et al.* (2000) The Protein Data Bank. Nucleic Acids Res. **28** (1), 235–242. http://dx.doi.org/10.1093/nar/28.1.235
3. Chintapalli SV, Bhardwaj G, Patel R, Shah N, Patterson RL, *et al.* (2015) Molecular dynamic simulations reveal the structural determinants of Fatty Acid binding to oxy-myoglobin. PLoS One **10** (6), e0128496. http://dx.doi.org/10.1371/journal.pone.0128496
4. Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, and Sayers EW (2016) GenBank. Nucleic Acids Res. **44** (D1), D67–72. http://dx.doi.org/10.1093/nar/gkw1070
5. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, *et al.* (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics **25** (11), 1422–1423. http://dx.doi.org/10.1093/bioinformatics/btp163
6. El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, *et al.* (2019) The Pfam protein families database in 2019. Nucleic Acids Res. **47** (D1), D427–D432. http://dx.doi.org/10.1093/nar/gky995
7. Hedgepeth SC, Garcia MI, Wagner LE 2nd, Rodriguez AM, Chintapalli SV, *et al.* (2015) The BRCA1 tumor suppressor binds to inositol 1,4,5-trisphosphate receptors to stimulate apoptotic

calcium release. J. Biol. Chem. **290** (11), 7304–7313. http://dx.doi.org/10.1074/jbc.M114.611186

8. Hong Y, Chalkia D, Ko KD, Bhardwaj G, Chang GS, *et al.* (2009) Phylogenetic Profiles Reveal Structural and Functional Determinants of Lipid-binding. J. Proteomics Bioinform. **2**, 139–149. http://dx.doi.org/10.4172/jpb.1000071
9. Kanehisa M, Sato Y, Furumichi M, Morishima K, and Tanabe M (2019) New approach for understanding genome variations in KEGG. Nucleic Acids Res. **47** (D1), D590–D595. http://dx.doi.org/10.1093/nar/gky962
10. Karp PD, Billington R, Caspi R, Fulcher CA, Latendresse M, *et al.* (2017) The BioCyc collection of microbial genomes and metabolic pathways. Brief. Bioinform. http://dx.doi.org/10.1093/bib/bbx085
11. Kyung Dae Ko, Chunmei Liu, Rwebangira MR, Burge L, and Southerland W (2011) The development of a proteomic analyzing pipeline to identify proteins with multiple RRMs and predict their domain boundaries. In: 2011 IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW). IEEE, pp. 374–381 http://dx.doi.org/10.1109/BIBMW.2011.6112401
12. Marchler-Bauer A, Bo Y, Han L, He J, Lanczycki CJ, *et al.* (2017) CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. Nucleic Acids Res. **45** (D1), D200–D203. http://dx.doi.org/10.1093/nar/gkw1129
13. Mudunuri U, Che A, Yi M, and Stephens RM (2009) bioDBnet: the biological database network. Bioinformatics **25** (4), 555–556. http://dx.doi.org/10.1093/bioinformatics/btn654
14. Patterson RL, Hong Y, Chintapalli SV, Bhardwaj G, Zhang Z, *et al.* (2011) Adaptive-BLAST: A User-defined Platform for the Study of Proteins. J. Integr. OMICS **1** (1) http://dx.doi.org/10.5584/jiomics.v1i1.33
15. Smathers RL and Petersen DR (2011) The human fatty acid-binding protein family: Evolutionary divergences and functions. Hum. Genomics **5** (3), 170. http://dx.doi.org/10.1186/1479-7364-5-3-170
16. Thibaud-Nissen F, DiCuccio M, Hlavina W, Kimchi A, Kitts PA, *et al.* (2016) P8008 The NCBI Eukaryotic Genome Annotation Pipeline. J. Anim. Sci. **94** (suppl_4), 184–184.
17. UniProt Consortium T (2018) UniProt: the universal protein knowledgebase. Nucleic Acids Res. **46** (5), 2699. http://dx.doi.org/10.1093/nar/gkv1189
18. Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, *et al.* (2018) Ensembl 2018. Nucleic Acids Res. **46** (D1), D754–D761. http://dx.doi.org/10.1093/nar/gkx1098