BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Excel@FIT 2020
http://excel.fit.vutbr.cz

EnzymeMiner: Web Server for Automated Mining of Soluble Enzymes

Simeon Borko*

Abstract
Millions of protein sequences are being discovered at an incredible pace, representing an inexhaustible source of biocatalysts. Despite genomic databases growing exponentially, classical biochemical characterization techniques are time-demanding, cost-ineffective and low-throughput. Therefore, computational methods are being developed to explore the unmapped sequence space efficiently. Selection of putative enzymes for biochemical characterization based on rational and robust analysis of all available sequences remains an unsolved problem. To address this challenge, I have developed EnzymeMiner – a web server for automated screening and annotation of enzymes that enables selection of hits for wet-lab experiments. EnzymeMiner prioritizes sequences that are more likely to preserve the catalytic activity and are expressible in a soluble form in heterologous host organism *Escherichia coli*. EnzymeMiner reduces the time devoted to data gathering, multi-step analysis, sequence prioritization and selection from days to hours. EnzymeMiner is a universal tool applicable to any enzyme family that provides an interactive and easy-to-use web interface freely available at https://loschmidt.chemi.muni.cz/enzymeminer/.

Keywords: enzyme mining — novel biocatalysts — web server

Supplementary Material: *N/A*

*xborko02@stud.fit.vutbr.cz, *Faculty of Information Technology, Brno University of Technology*

1. Introduction

There are currently more than 259 million unique protein sequences in the NCBI nr [1] database (release 2020-02-10). Despite their enormous promise for biological and biotechnological discovery, experimental characterization has been performed on only a small fraction of the available sequences. Currently, there are about 560,000 protein sequences reliably curated in the UniProtKB/Swiss-Prot [2] database (release 2020_01).

The low ratio of characterized to uncharacterized sequences reflects the sharp contrast in low-throughput biochemical techniques versus high-throughput next-generation sequencing technology. Although more efficient biochemical techniques employing miniaturization and automation have been developed [3, 4, 5], the most widely used experimental methods do not provide sufficient capacity for biochemical characterization of proteins spanning the ever-increasing sequence space. Therefore, computational methods are currently the only way to explore the immense protein diversity available among the millions of uncharacterized sequence entries.

Two different computational strategies are generally used for exploration of the unknown sequence space. The first strategy takes a novel uncharacterized sequence as input and predicts functional annotations.

The method involves annotating the unknown input sequences by predicting protein domains [6], Enzyme Commission (EC) number [7] or Gene Ontology terms that are a subject of the initiative named the Critical Assessment of Functional Annotation [8]. These methods are often universal and applicable to any protein sequence. However, they often lack specificity as the automatic annotation rules or statistical models need to be substantially general. A significant advantage of these methods is their seamless integration into available databases. Submission of a query sequence to a database is sufficient, with no need for running computation- and memory-intensive bioinformatics pipelines locally. A model example of this approach is the automatic annotation workflow of the UniProtKB/TrEMBL database [2].

The second strategy takes a well-known characterized sequence as an input and applies a computational workflow, typically based on a homology search, to identify novel uncharacterized entries in genomic databases that are related to the input query sequence [5, 9]. The homology search is often followed by a filtration step, which checks the essential sequence properties, e.g., domain structure or presence of essential residues. The main advantage of these methods is the higher specificity of the analysis. A disadvantage is that it may be complicated to apply the developed workflow to protein families other than those for which it was designed. Moreover, these workflows typically require running complex bioinformatics pipelines and are usually not available through a web interface.

The fundamental unsolved problem is how to deal with the overwhelming number of sequence entries identified by these methods and select a small number of relevant hits for in-depth experimental characterization. For example, a database search for members of the haloalkane dehalogenase model family using the UniProt web interface yields 3,598 sequences (UniProtKB release 2020_01). It is impossible to rationally select several tens of targets for experimental testing without additional bioinformatics analyses to help prioritize such a large pool of sequences.

To address the challenge of exploring the unmapped enzyme sequence space and rational selection of attractive targets, I have developed the EnzymeMiner web server. Enzymes are specific proteins that serve as biocatalysts for chemical reactions. They have broad application in both industry and academy, mostly in biosynthetic pathways for ecological production of chemical compounds. EnzymeMiner identifies novel enzyme family members, comprehensively annotates the targets and facilitates efficient prioritization and selection of representative hits for experimental characterization. To the best of my knowledge, there is currently no other tool available that allows such a comprehensive analysis in a single easy-to-run integrated workflow on the web.

2. Method

EnzymeMiner implements a three-step workflow: (i) homology search, (ii) essential residue based filtering, and (iii) hits annotation (Figure 1). To execute these tasks, the server requires two different types of input information: (i) query sequences, and (ii) essential residue templates. The query sequences serve as seeds for the initial homology search. The essential residue templates, defined as pairs of a protein sequence and a set of essential residues in that sequence, allow the server to prioritize hits that are more likely to display the enzyme function. Therefore, the essential residues may be the catalytic and ligand- or cofactor-binding residues that are indispensable for proper catalytic function. Each essential residue is defined by its name, position and a set of allowed amino acids for that position.

In the first homology search step, a query sequence is used as a query for a PSI-BLAST [10] two-iteration search in the NCBI nr database [1]. If more than one query sequence is provided, a search is conducted for each sequence separately. Besides a minimum E-value threshold 10-20, the PSI-BLAST hits must share a minimum of 25% global sequence identity with at least one of the query sequences. Artificial protein sequences, i.e., sequences described by the term artificial, synthetic construct, vector, vaccinia virus, plasmid, halotag or replicon, are removed. EnzymeMiner sorts the PSI-BLAST hits by E-value and passes a maximum of 10,000 best hits to the next steps in the workflow. The default parameters for the homology search step, as well as the other steps, can be modified using advanced options in the web server.

In the second essential residue based filtering step, the homology search hits are filtered using the essential residue templates. First, the hits are divided into template clusters. Each cluster contains all hits matching essential residues of a particular template. Essential residues are checked using global pairwise alignment with the template calculated by USEARCH [11]. When multiple essential residue templates match, the hit is assigned to the template with the highest global sequence identity. Second, for each cluster, an initial multiple sequence alignment (MSA) is constructed using Clustal Omega [12]. The MSA is used to revalidate the essential residues of identified hits

by checking the corresponding column in the MSA. Sequences not matching essential residues of the template are removed from the cluster. Third, the MSA is constructed again for each template cluster and the essential residues are checked for the last time. The final set of identified sequences reported by EnzymeMiner contains all sequences left in the template clusters. In the third annotation step, the identified sequences are annotated using several databases and predictors: (i) transmembrane regions are predicted by TMHMM [13], (ii) Pfam domains are predicted by InterProScan [14], (iii) source organism annotation is extracted from the NCBI Taxonomy [15] and the NCBI BioProject database [16], (iv) protein solubility is predicted by the in-house tool SoluProt for prediction of soluble protein expression in *Escherichia coli*, and (v) sequence identities to queries, hits or other optional sequences are calculated by USEARCH [11].

The sequence space of the identified hits is visualized using representative sequence similarity networks (SSNs) generated at various clustering thresholds using MMseqs2 [17] and Cytoscape [18]. SSNs provide a clean visual approach to identify clusters of highly similar sequences and rapidly spot sequence outliers. SSNs proved to facilitate identification of previously unexplored sequence and function space [19]. The SSN generation method used in EnzymeMiner is inspired by the EFI-EST tool [20]. The minimum alignment score to include an edge between two representative sequences in an SSN is 40.

3. Validation

The EnzymeMiner workflow has been experimentally validated using the model enzymes haloalkane dehalogenases [5]. The sequence-based search identified 658 putative dehalogenases. The subsequent analysis prioritized and selected 20 candidate genes for exploration of their protein structural and functional diversity. The selected enzymes originated from genetically unrelated Bacteria, Eukarya and, for the first time, also Archaea and showed novel catalytic properties and stabilities. The workflow helped to identify novel and very interesting haloalkane dehalogenases, including (i) the most catalytically efficient, (ii) the most thermostable enzyme showing a melting temperature of 71 °C, (iii) three different cold-adapted enzymes active at near to 0 °C, (iv) highly enantioselective enzymes, (v) enzymes with a wide range of optimal operational temperature from 20–70 °C and an unusually broad pH range from 5.7–10, and (vi) biocatalysts degrading the warfare chemical yperite or different environmental pollutants. The sequence mining, annotation, and visualization steps from the workflow published by Vanacek and co-workers [5] were fully automated in the EnzymeMiner web server. The successful use case for the haloalkane dehalogenase family is described in the form of an easy-to-follow tutorial available on the EnzymeMiner web page. Additional extensive validation of the fully automated version of EnzymeMiner, experimentally testing the properties of another 50 genes of the haloalkane dehalogenases, is currently ongoing in the Loschmidt laboratories.

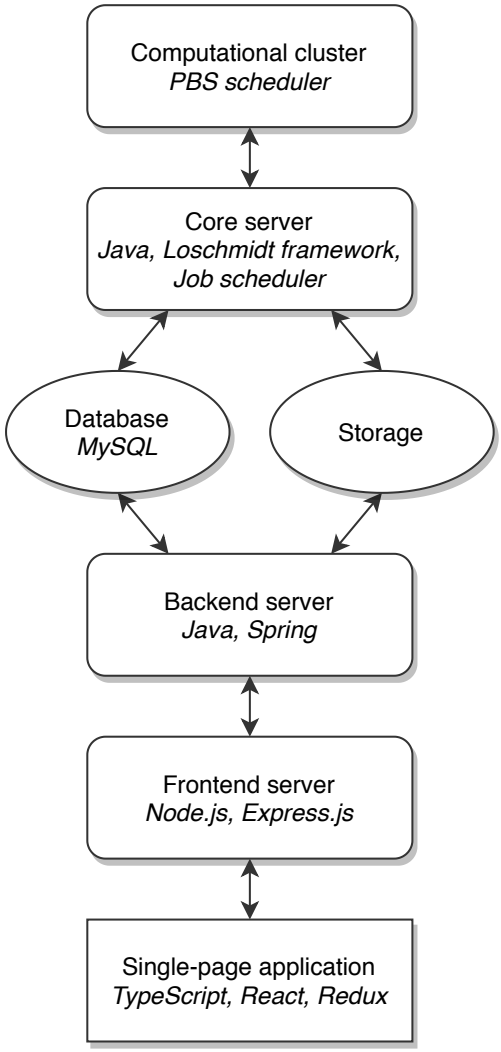

Figure 1. Multiple-server architecture of the EnzymeMiner web server. Only the *Frontend* server is publicly accessible.

4. Implementation

As many steps of the workflow are computationally demanding, a special architecture of the web server is needed. EnzymeMiner has a layered multiple-server architecture, where each server has a specific purpose and responsibility (Figure 1).

The *Core* server implements the main EnzymeMiner workflow using an in-house Java framework for

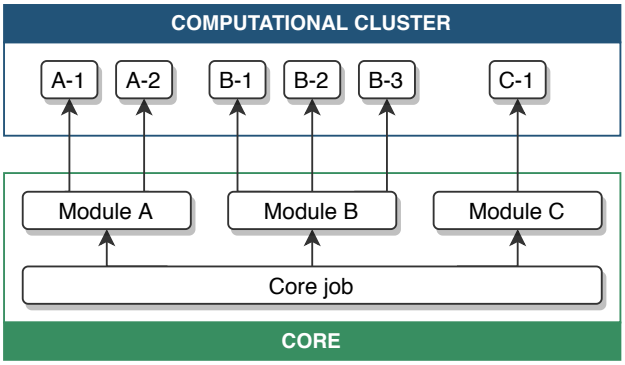

Figure 2. EnzymeMiner computing units. There are three types of computing units: (i) core jobs, (ii) core modules and (iii) cluster jobs.

bioinformatic applications from Loschmidt Laboratories. The *Core* reads jobs submitted to the *Database* and corresponding input files from the *Storage*. For each job, the *Core* instantiates and schedules twelve core modules that implement the three-step workflow. Independent modules are executed in parallel to speed up the calculation. To compute extremely demanding tasks, core modules submit cluster jobs to the *Computational cluster*. The cluster uses the Portable Batch System (PBS) to maximize parallelism and optimize usage of the available computing resources. To summarize, three types of computing units (Figure 2) are used in EnzymeMiner: (i) user-submitted core job, (ii) core modules scheduled by the *Core* server, and (iii) cluster jobs scheduled by the *Computational cluster*.

The *Backend* server implements a REST API for the single-page application. It implements endpoints to (i) submit new jobs to the *Database* and save input files to the *Storage*, (ii) download the job status and job results, and (iii) fetch server usage statistics.

The *Frontend* serves the static files of the single-page application. It is the only server in the architecture that is publicly accessible. The *Frontend* server also serves as a proxy to the *Backend* server.

The *Single-page application* provides the graphical user interface for the EnzymeMiner workflow. It is running in the user's browser and it is implemented in TypeScript using React framework for HTML rendering and Redux library for state management. The application uses Bootstrap and Material-UI components. Webpack builds the JavaScript and CSS files for the browser.

5. Conclusions and outlook

The EnzymeMiner web server identifies putative members of enzyme families and facilitates their prioritization and well-informed selection for experimental characterization to reveal novel biocatalysts. The identified enzymes might have broad application in both industry and academy, mostly in biosynthetic pathways for ecological production of chemical compounds. Such a task is difficult using the web interfaces of the available protein databases, e.g., UniProtKB/TrEMBL and NCBI Protein, since additional analyses are often required. The output of EnzymeMiner is an interactive selection table containing the annotated sequences divided into sheets based on various criteria. The table helps to select a diverse set of sequences for experimental characterization. Two key annotations are (i) the predicted solubility score, which can be used to prioritize the identified sequences and increase the chance of finding enzymes with soluble protein expression, and (ii) the sequence identity to query sequences complemented with an interactive SSN displayed directly on the web, which can be used to find diverse sequences. EnzymeMiner is a universal tool applicable to any enzyme family. It reduces the time needed for data gathering, multi-step analysis and sequence prioritization from days to hours. All the EnzymeMiner features are implemented directly on the web server and no external tools are required. The fact that EnzymeMiner is a web application brings both advantages and disadvantages. User is able to run the pipeline using a user-friendly web interface without the need to install software and download databases. On the other hand, user cannot modify the pipeline and the speed of the computation depends on the current load of the computing cluster.

In the next EnzymeMiner version, I plan two major improvements. First, I will implement automated tertiary structure prediction based on homology modeling and threading for all identified sequences. The structural predictions will allow subsequent analysis of active site pockets/cavities and access tunnels that will help to identify additional attractive targets for experimental characterization. Second, I will implement automated periodical mining. When enabled, EnzymeMiner will rerun the analysis periodically and inform the user about novel sequences found since the last search.

EnzymeMiner web server was submitted to highly impacted *Nucleic Acids Research* journal in March 2020.

6. Acknowledgement

I thank Ing. Jiří Hon for consultations, guidance, support and valuable advice. I also thank the Loschmidt laboratories for their scientific supervision and the participants of the 1st Hands-on Computational Enzyme Design Course for giving valuable feedback on the En-

zymeMiner user interface. Their comments inspired me to make the sequence similarity network visualization much more interactive. Computational resources were supplied by the project "e-Infrastruktura CZ" (e-INFRA LM2018140) and ELIXIR (LM2015047) provided within the program Projects of Large Research, Development and Innovations Infrastructures.

References

[1] Eric W. Sayers, Richa Agarwala, Evan E. Bolton, et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 47(D1):D23–D28, January 2019.

[2] The UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research*, 47(D1):D506–D515, January 2019.

[3] Pierre-Yves Colin, Balint Kintses, Fabrice Gielen, et al. Ultrahigh-throughput discovery of promiscuous enzymes by picodroplet functional metagenomics. *Nature Communications*, 6(1):1–12, December 2015.

[4] Thomas Beneyton, Stéphane Thomas, Andrew D. Griffiths, et al. Droplet-based microfluidic high-throughput screening of heterologous enzymes secreted by the yeast Yarrowia lipolytica. *Microbial Cell Factories*, 16(1):18, January 2017.

[5] Pavel Vanacek, Eva Sebestova, Petra Babkova, et al. Exploration of Enzyme Diversity by Integrating Bioinformatics with Expression Analysis and Biochemical Characterization. *ACS Catalysis*, 8(3):2402–2412, March 2018.

[6] Sara El-Gebali, Jaina Mistry, Alex Bateman, et al. The Pfam protein families database in 2019. *Nucleic Acids Research*, 47(D1):D427–D432, January 2019.

[7] Yu Li, Sheng Wang, Ramzan Umarov, et al. DEEPre: sequence-based enzyme EC number prediction by deep learning. *Bioinformatics*, 34(5):760–769, March 2018.

[8] Naihui Zhou, Yuxiang Jiang, Timothy R. Bergquist, et al. The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. *Genome Biology*, 20(1):244, November 2019.

[9] Wai Shun Mak, Stephen Tran, Ryan Marcheschi, et al. Integrative genomic mining for enzyme function to enable engineering of a non-natural biosynthetic pathway. *Nature Communications*, 6(1):1–10, November 2015.

[10] S. Altschul. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402, September 1997.

[11] Robert C. Edgar. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, 26(19):2460–2461, October 2010.

[12] Fabian Sievers, Andreas Wilm, David Dineen, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology*, 7(1):539, January 2011.

[13] Anders Krogh, Björn Larsson, Gunnar von Heijne, and Erik L.L Sonnhammer. Predicting transmembrane protein topology with a hidden markov model: application to complete genomes11Edited by F. Cohen. *Journal of Molecular Biology*, 305(3):567–580, January 2001.

[14] E. Quevillon, V. Silventoinen, S. Pillai, et al. InterProScan: protein domains identifier. *Nucleic Acids Research*, 33(Web Server):W116–W120, July 2005.

[15] S. Federhen. The NCBI Taxonomy database. *Nucleic Acids Research*, 40(D1):D136–D143, January 2012.

[16] T. Barrett, K. Clark, R. Gevorgyan, et al. BioProject and BioSample databases at NCBI: facilitating capture and organization of metadata. *Nucleic Acids Research*, 40(D1):D57–D63, January 2012.

[17] Martin Steinegger and Johannes Söding. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*, 35(11):1026–1028, November 2017.

[18] P. Shannon. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13(11):2498–2504, November 2003.

[19] Janine N. Copp, Eyal Akiva, Patricia C. Babbitt, and Nobuhiko Tokuriki. Revealing Unexplored Sequence-Function Space Using Sequence Similarity Networks. *Biochemistry*, 57(31):4651–4662, August 2018.

[20] John A. Gerlt, Jason T. Bouvier, Daniel B. Davidson, et al. Enzyme Function Initiative-Enzyme Similarity Tool (EFI-EST): A web tool for generating protein sequence similarity networks. *Biochimica et Biophysica Acta (BBA) - Proteins*

and Proteomics, 1854(8):1019–1037, August 2015.