Accurate predictions of enzymatic biochemistry as an enabler for generation of de-novo sequences

Anonymous authors

Abstract

Terpene synthases (TPSs) generate the scaffolds of the largest class of natural products, including several first-line medicines. The amount of available TPS protein sequences is increasing exponentially, but computational characterization of their function remains an unsolved challenge. We assembled a curated dataset of one thousand characterized TPS reactions and developed a method to devise highly accurate machine-learning models for functional annotation in a low-data regime. Our models significantly outperform existing methods for TPS detection and substrate prediction. By applying the models to large protein sequence databases, we discovered seven TPS enzymes previously undetected by state-of-the-art computational tools and experimentally confirmed their activity. Furthermore, we discovered a new TPS structural domain and distinct subtypes of previously known domains. Our work demonstrates the potential of machine learning to speed up the discovery and characterization of novel TPSs. Furthermore, in-silico functional annotations provide the ML community with a large dataset of pseudo-labeled exemplary TPS sequences. The accurate models for TPS detection and substrate prediction can serve as oracles to check the presence of desired biochemical activity in the generated sequences. We envision the published dataset of exemplary TPS sequences and the accurate TPS-annotation models to boost the generation of de-novo enzymatic TPS sequences.

Introduction

Terpene synthases (TPSs) are ubiquitous enzymes that produce the hydrocarbon scaffolds for the largest and the most diverse class of natural products called terpenoids. The most scents a human has ever experienced are terpenoids (Caputi and Aprea 2011). Terpenoids include widely used flavors, fragrances, and first-line medicines. E.g., Paclitaxel is the first-line anticancer medicine with billion-dollar peak annual sales (Gallego-Jara et al. 2020). Yet, these fantastic small molecules, terpenes and terpenoids, are too complex to be efficiently synthesized industrially (Quílez del Moral, Pérez, and Barrero 2020) and are typically extracted from plants, which is resource-intensive. For instance, each patient requires 3-10 Pacific yew trees for the mentioned anticancer treatment Paclitaxel (Kathiravan et al. 2012).

A more sustainable and efficient production of terpenoids can be achieved via synthetic biology (Zhang and Hong 2020). However, the discovery of TPSs responsible for the biosynthesis of these invaluable natural products is not trivial. The amount of available protein sequences is increasing exponentially. Preprocessing the results of high-throughput DNA sequencing by detecting the most likely candidates for TPS function can significantly accelerate the progress in bioprospecting, natural product biosynthesis, and synthetic biology. The main challenge when working with TPSs is the small amount of characterized data. It makes applying state-of-the-art ML models to TPSs challenging due to the curse of dimensionality.

Results

We assembled a curated dataset of 2028 characterized TPS reactions and developed highly accurate machine-learning oracles for functional TPS annotation of protein sequences. On top of it, we created a model for TPS detection and substrate[1] classification. The devised in-silico oracle for functional TPS annotation significantly outperforms existing methods, bumping the mean average precision of substrate classification from 0.69 to 0.89, see Fig. 1a. Also, we developed a novel computational procedure to segment AlphaFold2-generated structures into TPS structural domains. Taking the structural information into account improves the classification of substrates for the TPS hits from the fast PLM-based detector. As it is depicted in Fig. 1b, our overall performance is significantly better compared to PSI-BLAST (Altschul et al. 1997), Profile Hidden Markov Models (Eddy 1998), the recent deep-learning model Foldseek (Barrio-Hernandez et al. 2023) for protein search in the structure space of AlphaFold2 predictions (Barrio-Hernandez et al. 2023), and a state-of-the-art EC number predictor CLEAN (Yu et al. 2023).

[1] Substrate is a molecule the enzyme acts upon.

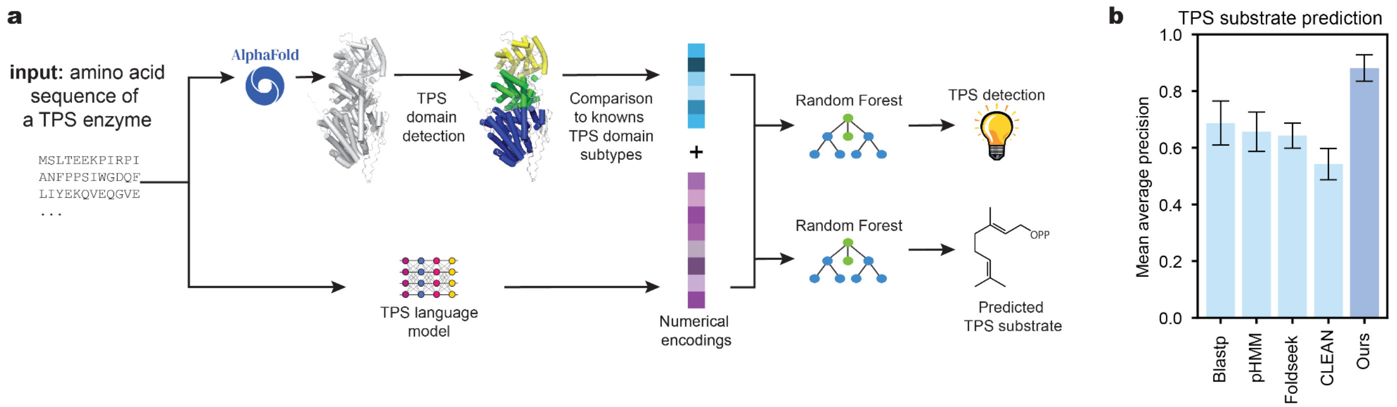

Fig 1. Our predictive pipeline outperforms existing methods on TPS detection and substrate classification. **a,** Overview of our predictive pipeline. **b**, Evaluation of TPS substrate classification performance.

By applying the developed models to large protein sequence databases, we discovered and experimentally confirmed the predicted activity of seven TPS enzymes previously undetected by state-of-the-art bioinformatic tools. The selected TPS hits by the oracle were undetected by traditional approaches like Pfam (Bateman et al. 2004), SUPFAM (Pandit et al. 2004), or InterProScan (Jones et al. 2014). Some experimentally confirmed hits came from organisms not typical for TPS activity, like archaea or viruses.

We published the dataset of 1192 characterized TPS synthases. Furthermore, with these TPS sequences of wet lab characterization, we published roughly two hundred thousand putative TPS sequences characterized by our oracles. Finally, we made the trained oracle models publicly available.

Methods

We fine-tuned a self-supervised TPS language model (PLM) using a masked language modeling objective on top of roughly 200k putative TPS sequences mined with standard tools like Pfam (Bateman et al. 2004). On top of embeddings from the PLM, we created Random Forest models for TPS detection and substrate classification, see the bottom branch of Fig. 1a.

Segmentation of AlphaFold2-generated TPS structures was accomplished via structural alignment of known TPS-specific structural domains on top of the AlphaFold2 structure. Pairwise comparisons of the segmented domains were also performed via sequence-independent alignment. We used PyMOL (DeLano and Bromberg 2004) and TM-align (Y. Zhang and Skolnick 2005) for alignment. The TM-score of pairwise alignment was then used as a precomputed metric for K-medoids (Park and Jun 2009) and HDBSCAN (McInnes, Healy, and Astels 2017) clustering algorithms. The resulting clusters defined subtypes of the established TPS domains.

The experimental confirmation of the predicted activity was achieved by heterologous expression in budding yeast.

Discussion

We demonstrated the potential of ML to speed up the functional characterization of TPS, with confirmation of the models' efficiency via wet lab experiments. Furthermore, we used the models to mine a large dataset of in-silico annotated TPS sequences.

Our next goal is to use the published TPS data and the oracles to generate novel TPS sequences acting upon the substrate(s) of interest. In other words, the aim is to develop generative ML models producing functional de-novo amino acid sequences constrained on a substrate molecule(s). There are several optimization criteria in the generation process. The higher the confidence of our oracles that a novel sequence encodes an active TPS, the better. The higher the confidence of the TPS classifier that the de-novo TPS accepts the required substrate, the better. Finally, novel sequences are preferred among generated ones that pass tests by our oracles with comparable confidence. The sequences with the lowest similarity to known TPSs are more likely to produce new products, expanding the chemical space of known small molecules.

References

Altschul, S. F., T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. 1997. “Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs.” *Nucleic Acids Research* 25 (17): 3389–3402.

Barrio-Hernandez, Inigo, Jingi Yeo, Jürgen Jänes, Milot Mirdita, Cameron L. M. Gilchrist, Tanita Wein, Mihaly Varadi, Sameer Velankar, Pedro Beltrao, and Martin Steinegger. 2023. “Clustering Predicted Structures at the Scale of the Known Protein Universe.” *Nature*, September. https://doi.org/10.1038/s41586-023-06510-w.

Bateman, Alex, Lachlan Coin, Richard Durbin, Robert D. Finn, Volker Hollich, Sam Griffiths-Jones, Ajay Khanna, et al. 2004. “The Pfam Protein Families Database.” *Nucleic Acids Research* 32 (Database issue): D138–41.

Caputi, Lorenzo, and Eugenio Aprea. 2011. “Use of Terpenoids as Natural Flavouring Compounds in Food Industry.” *Recent Patents on Food, Nutrition & Agriculture* 3 (1): 9–16.

Eddy, S. R. 1998. “Profile Hidden Markov Models.” *Bioinformatics* 14 (9): 755–63.

Gallego-Jara, Julia, Gema Lozano-Terol, Rosa Alba Sola-Martínez, Manuel Cánovas-Díaz, and Teresa de Diego Puente. 2020. “A Compressive Review about Taxol®: History and Future Challenges.” *Molecules* 25 (24). https://doi.org/10.3390/molecules25245986.

Jones, Philip, David Binns, Hsin-Yu Chang, Matthew Fraser, Weizhong Li, Craig McAnulla, Hamish McWilliam, et al. 2014. “InterProScan 5: Genome-Scale Protein Function Classification.” *Bioinformatics* 30 (9): 1236–40.

Kathiravan, G., Sripathi M. Sureban, Harsha N. Sree, V. Bhuvaneshwari, and Evelin Kramony. 2012. “Isolation of Anticancer Drug TAXOL from Pestalotiopsis Breviseta with Apoptosis and B-Cell Lymphoma Protein Docking Studies.” *Journal of Basic and Clinical Physiology and Pharmacology* 4 (1): 14–19.

McInnes, Leland, John Healy, and Steve Astels. 2017. “Hdbscan: Hierarchical Density Based Clustering.” *Journal of Open Source Software* 2 (11): 205.

Pandit, Shashi B., Rana Bhadra, V. S. Gowri, S. Balaji, B. Anand, and N. Srinivasan. 2004. “SUPFAM: A Database of Sequence Superfamilies of Protein Domains.” *BMC Bioinformatics* 5 (March): 28.

Park, Hae-Sang, and Chi-Hyuck Jun. 2009. “A Simple and Fast Algorithm for K-Medoids Clustering.” *Expert Systems with Applications* 36 (2, Part 2): 3336–41.

Quílez del Moral, José Francisco, Álvaro Pérez, and Alejandro F. Barrero. 2020. “Chemical Synthesis of Terpenoids with Participation of Cyclizations plus Rearrangements of Carbocations: A Current Overview.” *Phytochemistry Reviews* 19 (3): 559–76.

Yu, Tianhao, Haiyang Cui, Jianan Canal Li, Yunan Luo, Guangde Jiang, and Huimin Zhao. 2023. “Enzyme Function Prediction Using Contrastive Learning.” *Science* 379 (6639): 1358–63.

Zhang, Caizhe, and Kui Hong. 2020. “Production of Terpenoids by Synthetic Biology Approaches.” *Frontiers in Bioengineering and Biotechnology* 8 (April): 347.