Structome-TM: Complementing dataset assembly for structural phylogenetics by addressing size-based biases

Ashar J. Malik[1,2,3,*] and David B. Ascher[1,2,3,*]

[1]School of Chemistry and Molecular Biosciences, The University of Queensland, Brisbane, Australia

[2]Australian Centre for Ecogenomics, The University of Queensland, Brisbane, Australia

[3]Computational Biology and Clinical Informatics, Baker Heart and Diabetes Institute, Melbourne, Victoria, Australia

[*]Correspondence to Ashar J. Malik: ashar.malik@uq.edu.au, David B. Ascher: d.ascher@uq.edu.au

Abstract

The increasing availability of protein structures presents new opportunities for detecting homologous proteins. While global alignment methods are widely used and generally effective, they often struggle to identify related structures obscured by differences in protein size and domain architecture. To address this limitation, we introduce Structome-TM, a web server for protein structure comparison and phylogenetic analysis using TM-score as a metric. By emphasizing local structural alignments, Structome-TM improves the detection of homologous proteins that global methods might overlook. With a user-friendly interface and efficient pairwise comparisons, Structome-TM expands the structural neighbourhood landscape and facilitates phylogenetic reconstruction.

Availability and implementation

The Structome-TM web server is available at https://biosig.lab.uq.edu.au/structome_tm/

Introduction

The large-scale availability of protein structure data has revolutionized the exploration of deep phylogenetic relationships [1, 2]. This process typically involves two stages: (1) gathering structures to define the taxa space and (2) comparing these structures to infer evolutionary histories. Structome-Q (previously Structome [2]) is a pre-calculated resource that efficiently addresses both stages. It rapidly identifies structurally similar proteins to a query structure using the Q-score metric provided by GESAMT [3] and performs pairwise structure comparisons to construct a neighbour-joining tree.

The Q-score metric is defined as:

$$Q = \frac{N_{\text{align}}^2}{\left(1 + \left(\frac{\text{rmsd}}{R_0}\right)^2\right) N_1 N_2}$$

where N_{align} is the number of aligned residue pairs, rmsd is the root-mean-square deviation of the alignment, R_0 is a scaling parameter (set to 3.0 Å), and N_1 and N_2 are the amino acid counts of the two structures. By normalizing over the product of the structure lengths ($N_1 N_2$), Q-score attempts to quantify global structural similarity. While highly effective in many scenarios, this approach inherently scores poorly for structures that share only localized regions of similarity, such as proteins with partial domain matches. As a result, these proteins may rank low and be excluded from further analyses, potentially overlooking biologically significant homologues.

In contrast, the TM-score [4], defined as:

$$\text{TM-score} = \frac{1}{L_{\text{target}}} \sum_{i=1}^{N} \frac{1}{1 + \left(\frac{d_i}{d_0(L_{\text{target}})}\right)^2}$$

where L_{target} is the length of the target protein, N is the number of aligned amino acid, d_i is the distance between aligned amino acid pairs, and

$$d_0(L_{\text{target}}) = 1.24\sqrt[3]{L_{\text{target}} - 15} - 1.8,$$

adopts a different approach. Implementation of this metric through Foldseek [5] normalizes based on alignment length rather than overall structure size, thus prioritising local similarity. This makes it particularly advantageous for identifying meaningful alignments where only partial structural matches exist, such as multi-domain proteins sharing one domain with the query. This distinction allow TM-score to highlight potential relationships that might be underestimated by Q-score.

As a stand-alone resource, Structome-TM introduces an alternative methodology, particularly suited for cases where local structural alignments offer critical insights. Together, Structome-Q and Structome-TM provide complementary tools for structural phylogenetics, enabling a more nuanced exploration of protein relationships.

Methods

Structome-TM follows the same architecture as Structome-Q (previously Structome [2], now available at https://biosig.lab.uq.edu.au/structome_q/). In summary, the RCSB PDB (www.rcsb.org) hosts a large number of protein structures many of which are short peptides or redundant. To reduce the size of this dataset, filtering and clustering is employed. Proteins longer than 50 amino acids are clustered at 90% sequence identity using usearch [6]. Each cluster is represented in Structome-TM by its centroid, which is assumed to be a reasonable proxy for all cluster members when determining structural neighbourhoods. Only centroids are pairwise compared, thereby making the calculations computationally tractable. When a query structure is submitted, unless it is itself a cluster centroid, the results correspond to its cluster's centroid. For the remainder of this work, "query structure" refers to the centroid for brevity. Structome-TM was last updated in December 2024, ensuring the inclusion of the most recent data. Pairwise comparisons between centroids are performed using Foldseek, and the resulting TM-scores are reported when Structome-TM is queried. In addition to TM-score, information collated from several resources, namely SCOP [7], CATH [8], ECOD [9] and the NCBI taxonomy database [10, 11], is also listed.

The web resource is designed using Flask, and makes use of HTML, JavaScript and AJAX for the front-end. It utilizes Mol* [12] for visualizing protein structures. The Biopython library generates neighbour-joining trees, which are rendered interactively

using the D3 JavaScript library.

Usage

Structome-TM presents results in an interactive and user-friendly manner. Results are displayed as tables where rows correspond to comparisons between the query structure and centroids in the Structome-TM database with a TM-score > 0.1. Clicking a row reveals the list of member proteins for the respective cluster and shows the query structure superimposed on the selected centroid. A TM-score distribution plot summarizes the TM-scores listed in the hits table, providing an overview of alignment quality. Checkboxes allow centroids to be selected for tree generation. Once the tree generation module is triggered, a job is submitted, and upon successful completion, a neighbour-joining tree is rendered. The tree uses "1 - TM-score" as the distance metric and is displayed in a simplified format, with an option for downloading for further annotation. An additional section lists the taxa represented in the tree, and clicking on a taxon on the tree reveals detailed information about it.

Multi-domain target ranking

The comparison of flavohemoglobin from *Saccharomyces cerevisiae* with hemoglobin from *Anser indicus* (see Fig. 1) highlights how the complementary approach employed by Structome-TM can capture meaningful alignments that might be missed when global similarity is emphasized. Flavohemoglobin comprises a globin domain fused with an FAD-binding domain [13], resulting in a noticeable size mismatch between the two proteins. Structome-Q assigns a low Q-score of "0.213" due to this size difference and its emphasis on global alignment. This approach down-weights the relatively good structural alignment, producing a low score which may not be easily discernable from random matches (background noise), thereby increasing the likelihood of missing this potential result. In contrast, Structome-TM achieves a TM-score of "0.803" by emphasizing local alignment quality. This complementary approach helps build a more comprehensive structural neighbourhood, ensuring that hits comprising other domains are detected and can be included in downstream analysis.

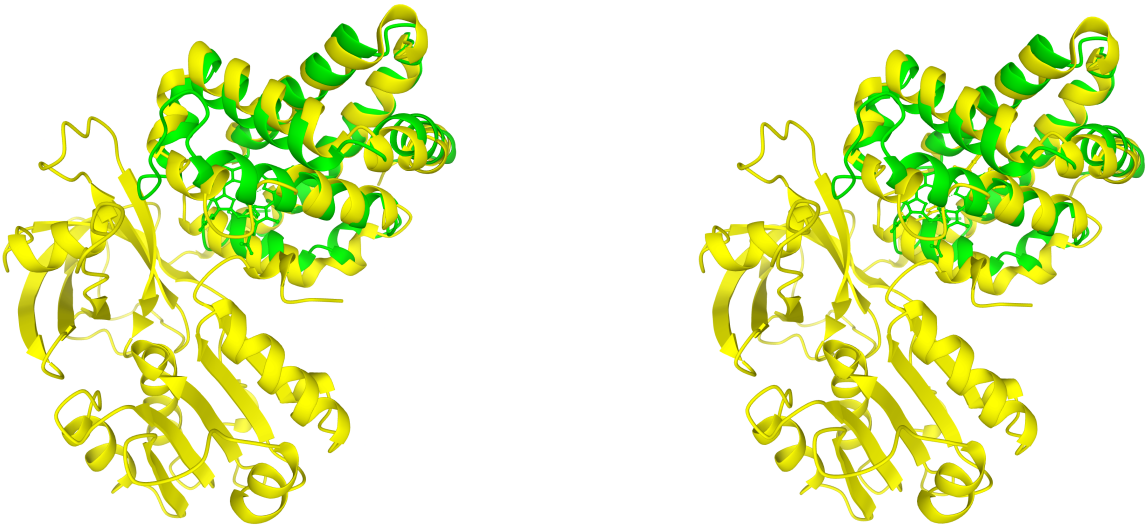

Figure 1: Protein structure alignment between hemoglobin from *Anser indicus* (RCSB PDB accession 1a4f, chain A, 141 amino acids, shown in green) and flavohemoglobin from *Saccharomyces cerevisiae* (RCSB PDB accession 4g1b, chain A, 399 amino acids, shown in yellow). The left panel illustrates the alignment from Structome-Q using the Q-score metric, while the right panel shows the alignment from Structome-TM using TM-score. The Q-score for this alignment is 0.213, whereas the TM-score is 0.803. The lower similarity score from Q-score reflects its global normalization accounting for overall protein sizes. A score of 0.213 indicates that this hit will probably be lost in the background noise.

Outlook

Future developments for Structome-TM aim to enhance its usability and functionality. Key enhancements will include the ability for users to compare custom structures, beyond the centroid proteins currently supported along with language processing capabilities to offer meaningful insights through generative text summarization of results.

Together, Structome-TM and Structome-Q serve as complementary tools, enriching the process of constructing a comprehensive structural neighbourhood. By offering an alternative metric, Structome-TM opens new possibilities for phylogenetic tree reconstruction and facilitates deeper evolutionary insights.

Funding

DBA is supported by the investigator grant from the National Health and Medical Research Council (NHMRC) of Australia [GNT1174405] and the Victorian government Operational Infrastructure Support program.

References

[1] Malik, A. J., Poole, A. M., and Allison, J. R. "Structural phylogenetics with confidence". *Molecular biology and evolution* 37.9 (2020), pp. 2711–2726.

[2] Malik, A. J. et al. "Structome: a tool for the rapid assembly of datasets for structural phylogenetics". *Bioinformatics Advances* 3.1 (2023), vbad134.

[3] Krissinel, E. "Enhanced fold recognition using efficient short fragment clustering". *Journal of molecular biochemistry* 1.2 (2012), p. 76.

[4] Zhang, Y. and Skolnick, J. "TM-align: a protein structure alignment algorithm based on the TM-score". *Nucleic acids research* 33.7 (2005), pp. 2302–2309.

[5] Kempen, M. van et al. "Foldseek: fast and accurate protein structure search". *Biorxiv* (2022), pp. 2022–02.

[6] Edgar, R. C. “Search and clustering orders of magnitude faster than BLAST”. *Bioinformatics* 26.19 (2010), pp. 2460–2461.

[7] Fox, N. K., Brenner, S. E., and Chandonia, J.-M. “SCOPe: Structural Classification of Proteins—extended, integrating SCOP and ASTRAL data and classification of new structures”. *Nucleic acids research* 42.D1 (2014), pp. D304–D309.

[8] Sillitoe, I. et al. “CATH: increased structural coverage of functional space”. *Nucleic acids research* 49.D1 (2021), pp. D266–D273.

[9] Schaeffer, R. D. et al. “ECOD: new developments in the evolutionary classification of domains”. *Nucleic acids research* 45.D1 (2017), pp. D296–D302.

[10] Dana, J. M. et al. “SIFTS: updated Structure Integration with Function, Taxonomy and Sequences resource allows 40-fold increase in coverage of structure-based annotations for proteins”. *Nucleic acids research* 47.D1 (2019), pp. D482–D489.

[11] Schoch, C. L. et al. “NCBI Taxonomy: a comprehensive update on curation, resources and tools”. *Database* 2020 (2020), baaa062.

[12] Sehnal, D. et al. “Mol* Viewer: modern web app for 3D visualization and analysis of large biomolecular structures”. *Nucleic acids research* 49.W1 (2021), W431–W437.

[13] Schuster, C. D. et al. “Globin phylogeny, evolution and function, the newest update”. *Proteins: Structure, Function, and Bioinformatics* 92.6 (2024), pp. 720–734.