

EMERALD-UI: an interactive web application to unveil novel protein biology hidden in the alternative alignment space

Andrei Preoteasa¹, Andreas Grigorjew^{1,2}, Alexandru I. Tomescu^{1,*}, Hajk-Georg Drost^{3,4,*}

¹Department of Computer Science, University of Helsinki, Helsinki, 00560, Finland

²LAMSADE, CNRS UMR7243, Université Paris Dauphine-PSL, Place du Maréchal de Lattre de Tassigny, Paris, 75775, France

³Computational Biology Group, Max Planck Institute for Biology Tuebingen, Tuebingen, Germany

⁴Digital Biology Group, Faculty of Life Sciences, University of Dundee, Dundee, United Kingdom

*Corresponding authors. Hajk-Georg Drost, Computational Biology Group, Max Planck Institute for Biology Tuebingen, Tuebingen, Germany; Digital Biology Group, Faculty of Life Sciences, University of Dundee, Dundee, United Kingdom. E-mail: hdrost001@dundee.ac.uk; Alexandru I. Tomescu, Department of Computer Science, University of Helsinki, Helsinki, 00560, Finland. E-mail: alexandru.tomescu@helsinki.fi
Associate Editor: Arne Elofsson

Abstract

Summary: Life over the past four billion years has been shaped by proteins and their capacity to assemble into three-dimensional conformations. Protein sequence alignments have been the enabling technology for exploring the evolution and functional adaptation of proteins across the tree of life. Recent advancements in scaling the prediction of three-dimensional protein structures from primary sequence alone, revealed that different modes of conservation and function operate on the sequence and structure level. This difference in protein conservation patterns and their underlying functional change that could emerge in suboptimal alignment configurations is often ignored in optimal protein alignment approaches. We introduce EMERALD-UI, an open-source interactive web application which is designed to reveal unexplored biology by visualising stable structural conformations or protein regions hidden in the alternative alignment space.

Availability: EMERALD-UI is available at <https://algbio.github.io/emerald-ui/>. The source code of the version described in this manuscript is available at <https://github.com/algbio/emerald-ui> and archived at Software Heritage: swh: 1: dir: 8b5a70160396d5e9a2e6d015c3b6f1426176d9a4.

1 Introduction

When aligning two protein sequences, a user-defined (and often heuristic) substitution matrix determines which alignment configuration is deemed most optimal by a respective alignment algorithm (Zahid *et al.* 2015, Kaminski *et al.* 2022, Llinares-López *et al.* 2023, Petti *et al.* 2023). For two divergent protein sequences, for example, this could mean that thousands of alternative alignment configurations with biological relevance may exist, but with lower total substitution-matrix score (Sierk *et al.* 2010, Chen and Kihara, 2011). The core assumption of this optimal alignment approach is that due to its well-established heuristic insight, the optimal score is proportional to a biologically meaningful alignment configuration, especially when dealing with similar proteins with greater than 90% sequence identity (Naor and Brutlag 1994). Although this assumption seems reasonable, when extrapolated to millions of pairwise comparisons, this approach can introduce a systematic optimal alignment bias, where alternative alignment configurations would have been

more reasonable when biologically assessed (Grigorjew *et al.* 2023). We recently introduced the open-source command line tool EMERALD which allows users to systematically explore the suboptimal-alignment space and sample alignment-safe positions from the theoretically true alignment configuration (Grigorjew *et al.* 2023). Using EMERALD, users can extract individual alignment-configuration positions which are stable across all or a proportion of suboptimal alignments. However, this tool lacks a comprehensive user interface and analytics capacity to reveal alternative biology or protein structural information hidden in alternative alignment configurations.

EMERALD-UI solves this problem by providing an interactive web application to visualise the suboptimal-alignment space between two protein sequences and map alignment-stable positions onto the three-dimensional structure on the respective protein folds. We envision that users will be able to trial how alternative (suboptimal) alignment configurations will translate into differences in protein structural or protein-protein interaction space. Especially, when deployed to proteins involved in

protein complexes, we foresee new biological and functional insights emerging from new (and previously unrecognised) configurations of protein complexes that will arise from the systematic exploration of the alternative-alignment space. Particularly when dealing with the comparison of highly divergent proteins, the alternative-alignment space will reveal an evolutionary repertoire of structural conformations which can also be investigated in the context of their role in biological pathways, kinetic efficiency (e.g. enzymes) or drug target predictions.

2 EMERALD: a standalone command line tool

In Grigorjew *et al.* (2023), we introduced the notion of *alignment-safety* for pairwise protein sequence alignments: EMERALD explicitly calculates the space of alternative (suboptimal) alignments. From this space, a set of *safety windows* can be extracted as contiguous alignment intervals that are shared across many (or all) alternative alignments.

In detail, EMERALD considers a standard directed acyclic graph (DAG) representing a dynamic-programming alignment (using Needleman–Wunsch alignment with affine gaps) (Gotoh 1982). To also represent sub-optimal alignments, this DAG incorporates all edges contained in an alignment of cost at most Δ from the optimal one. In this graph, EMERALD computes alignment intervals (safety windows) that are common to an α proportion of all alignments in this DAG ($\alpha = 1$ means common to *all* alignments). These safety windows are projected back onto the coordinates in the two protein sequences.

In a large-scale experiment using the SwissProt database, we showed that especially for divergent sequence pairs, safety windows can retain a high fraction of structurally stable protein residues while covering only a small fraction of the sequence (Grigorjew *et al.* 2023). This result supports the idea that exploring suboptimal-alignment space reveals structurally significant conserved regions that a single optimal alignment would otherwise miss.

In practice, the two EMERALD parameters Δ and α impact alignment-safety computations as follows:

- Increasing Δ enlarges the suboptimal-alignment space. Moreover, as Δ increases, safety windows typically become shorter and/or fewer (and overall “safe coverage” tends to drop), because more alternative alignments create more disagreement about which residue pairings are robust.
- Increasing α makes the safety criterion stricter, meaning that safety-windows tend to become shorter or fewer, but result in higher-confidence. Decreasing α on the other hand makes the results more permissive, yielding longer or more windows, but less stable.

3 EMERALD-UI: an interactive web-application

The new EMERALD-UI web application provides an interactive and responsive user interface for generating and exploring the suboptimal protein alignment space when comparing two

protein sequences. Moreover, while EMERALD as a standalone command line tool requires user installation, EMERALD-UI is running natively in any modern web-browser environment without manual installation requirements.

The UI takes two user-defined protein sequences as input and internally runs EMERALD directly in the browser environment to ultimately display the alternative alignment results as well as their associated protein structures with safety windows projected onto the structure model. All results generated with EMERALD-UI can be exported to the user machine. EMERALD-UI also provides direct URL links to downstream databases such as UniProt (Ahmad *et al.* 2025), AlphaFold (Varadi *et al.* 2022), FoldSeek (van Kempen *et al.* 2024), KEGG (Kanehisa and Goto 2000, Kanehisa 2019, Kanehisa *et al.* 2025), LIGYSIS (Utgés *et al.* 2025) for further inspection of how alternative alignment configurations translate into differences in structural space (Fig. 1).

The application supports multiple input methods. The user can either upload their own protein sequence files in FASTA-format, manually input protein sequences by copy-pasting, upload PDB or CDF files, or search the Uniprot database directly from EMERALD-UI.

With the protein pair defined, the user can then fine tune EMERALD parameters. The α slider changes the safety window highlighting threshold. The Δ slider changes the suboptimal alignment depth. The input also includes advanced parameters which can be used to change the EMERALD alignment substitution matrix, gap cost and gap-opening cost. We support BLOSUM 45, 50, 62, 80, 90, PAM 30, 70, 250 and basic IDENTITY. With these advanced parameters the user can explore how the alternative-alignment space and corresponding structural changes change when different evolutionary scenarios (different cost-matrices) are employed.

For each case, the generated alignment is visualised in a custom graph drawing. An alignment (match or mismatch) of characters is visualised by a diagonal line between them, whilst an indel is seen as either a vertical or horizontal line. The alignment safety windows are highlighted in green within the graph. To illustrate alternative suboptimal alignment configurations, the graph highlights these alternative configurations as different alignment paths along the graph. The “traditional” optimal alignment is visualised in blue. The graph is highly interactive so that users can zoom in and out of the graph to select their preferred path and alternative alignment configuration. Hovering over any section of the alignment graph shows the characters at that position, their positional indices and fraction of how many (sub)optimal alignment paths go through that position. Further analysis can be done through the side panel which lets the user highlight and view information about safety windows and non-safety-windows. The side panel is also designed to encourage users to select a specific alignment path and thereby explore specific subsections of the suboptimal-alignment space. The graph can be exported directly as a PNG, JPEG or SVG file and used in publications. The side panel includes a settings tab where users can toggle different features in the graph on and off depending on which features they wish to visualise in the exported image.

Finally, EMERALD-UI connects any alternative alignment assessment to downstream databases for further exploration of how the alternative-alignment space can fuel novel discovery in

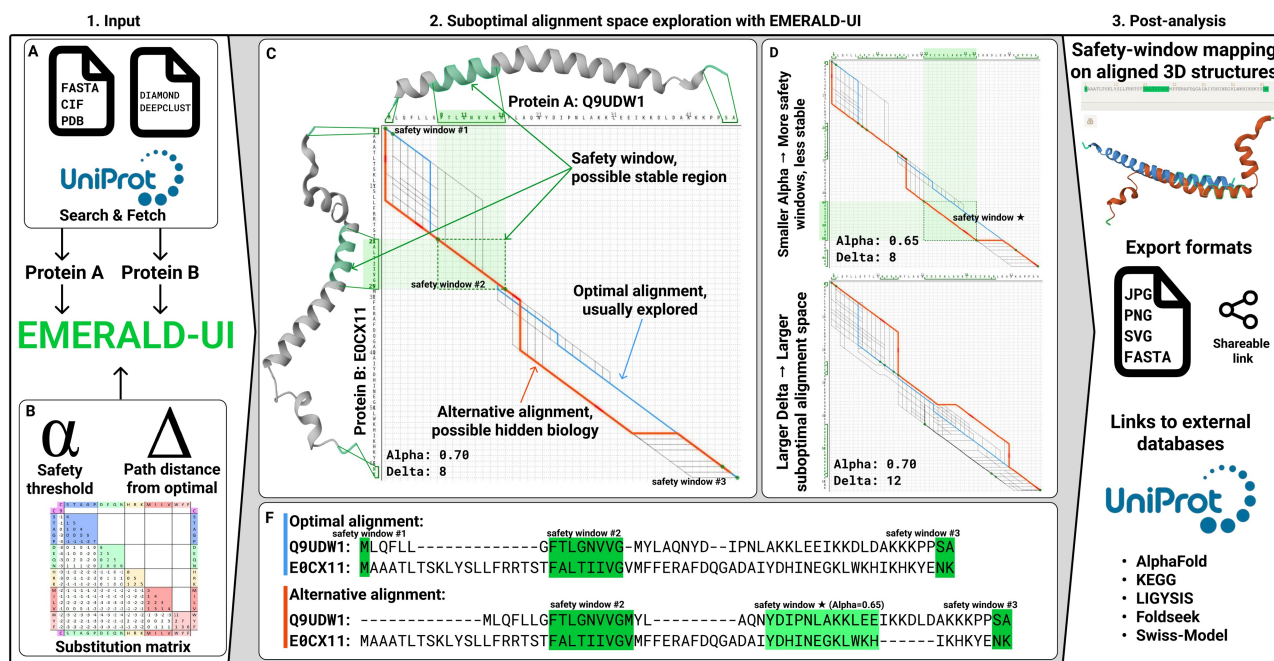


Figure 1 Schematic overview of the EMERALD-UI workflow. (1) Users can import their protein sequences as FASTA, CIF, PDB, UniProt and DIAMOND DeepClust clusters for comparisons across the alternative-alignment space (panel A). (2) Using the protein sequences specified by the user, EMERALD-UI then computes the pairwise (suboptimal) alignments using EMERALD either by relying on the default settings or the users adjusted (or even advanced) settings (panel B) and visualises the corresponding alignment graph which includes all alternative alignment configurations of the alternative-alignment space (panel C). When lowering the alpha parameter new safety windows across the alternative-alignment space are revealed (panel D, top). This is captured by paths diverting from the optimal alignment path (blue path). When increasing the delta parameter on the other hand, more alternative alignment paths are revealed (panel D, bottom). Users can then select an alternative alignment path (orange path) for further structural and downstream assessments (panels C, D, F). (3) Finally, users can explore where the safety windows map on the three-dimensional structure, export the analysis and forward these proteins to downstream protein annotation databases for further exploration. Alignment available at: <https://algbio.github.io/emerald-ui/?seqA=E0CX11&seqB=Q9UDW1>.

the structural and biological pathway space. For example, when using proteins from UniProt, EMERALD-UI projects the generated safety windows directly onto the structure of the protein through Mol* (industry standard for 3D visualization) [Sehnal et al. \(2021\)](#) and links the respective proteins to KEGG and LIGYSIS for downstream pathway and ligandability assessments.

4 Code highlights and software architecture

EMERALD-UI has a simple and lightweight software architecture, which enables real-time and no-delay execution without local installation or server deployment. The application uses EMERALD natively in the browser through Web Assembly (WASM). The raw output of EMERALD is compiled in real time and then displayed as a graph. The graph rendering and deployment logic was written ground-up with the D3.js JavaScript library, meaning that exploring the graph (zooming and panning) is fast and smooth even with large inputs. The graph is interactive and lets the user explore and highlight specific sections of the alignment. Each safety window can be highlighted and viewed separately in a side panel (see documentation on EMERALD-UI webpage). In addition to the optimal alignment being highlighted in the graph (in blue) it is also shown as a traditional pairwise alignment configuration below the alignment graph. If the user selects a custom path in the

graph (in orange), the corresponding alternative alignment is also shown as a pairwise alignment in the same panel as the optimal pairwise alignment.

EMERALD-UI also supports exporting different data in different formats. If the user has selected sections in the graph, i.e. a safety window, a non-safety window or a full alternative alignment, they can be exported as raw text directly from the application. The graph itself can be exported as an image file. The alignment can also be directly shared as a URL, which when opened, takes the user directly to the exact same protein alignment.

The protein retrieval tool embedded in EMERALD-UI pulls the latest protein data through the UniProt API. The search box supports searching proteins by scientific names, accession codes or ID codes and even natural language. The search results are compiled into a list table with the accession code, protein name, organism and length. The user is also given the option to preview the protein, navigate to the UniProt page of the protein and further explore the corresponding 3D structure of the protein through Mol*.

Acknowledgments

We thank Sebastian Schmidt for discussions on EMERALD-UI. We also thank all members of the DrostLab and Algorithmic Bioinformatics Lab for their feedback on user-experience and overall UI features.

Author contributions

Andrei Preoteasa (Formal analysis [Equal], Investigation [Equal], Methodology [Equal], Resources [Equal], Software [Lead], Visualization [Lead], Writing—review & editing [Supporting]), Andreas Grigorjew (Formal analysis [Supporting], Methodology [Equal], Software [Supporting]), and Alexandru Tomescu (Conceptualization [Equal], Data curation [Equal], Formal analysis [Equal], Funding acquisition [Equal], Investigation [Equal], Methodology [Equal], Project administration [Equal], Resources [Equal], Software [Supporting], Supervision [Equal], Validation [Equal], Visualization [Supporting], Writing—original draft [Equal], Writing—review & editing [Equal]), and Hajk-Georg Drost (Conceptualization [Lead], Data curation [Equal], Formal analysis [Equal], Funding acquisition [Equal], Investigation [Equal], Methodology [Equal], Project administration [Equal], Resources [Equal], Software [Equal], Supervision [Equal], Validation [Equal], Visualization [Equal], Writing—original draft [Lead], Writing—review & editing [Equal])

Conflicts of interest

None declared.

Funding

This work was supported by the Max Planck Society and a Royal Society Wolfson Fellowship [RSWF\R1\241004 awarded to H.G.D.]. H.G.D. also thanks Detlef Weigel for generous sponsorship and the BMBF-funded de. NBI Cloud within the German Network for Bioinformatics Infrastructure (de.NBI) [031A532B, 031A533A, 031A533B, 031A534A, 031A535A, 031A537A, 031A537B, 031A537C, 031A537D, 031A538A] for computational support. This work was partially funded by the Research Council of Finland [grants No. 346968, 358744] and by the European Union [ERC, SCALEBIO, 101169716, awarded to A.I.T.]. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them. Finally, A.G. is partially supported by the Agence nationale de la recherche (ANR) project ANR-21-CE48-0022 [S-EX-AP-PE-AL], awarded to Michael Lampis.

Data availability

EMERALD-UI is available at <https://algbio.github.io/emerald-ui/>. The source code of the version described in this manuscript is available at <https://github.com/algbio/emerald-ui> and archived at Software Heritage: swlh: 1: dir: 8b5a70160396d5e9a2e6d015c3b6f1426176d9a4.

References

Ahmad S, Jose da Costa Gonzales L, Bowler-Barnett E *et al.* The UniProt website API: facilitating programmatic access to

- protein knowledge. *Nucleic Acids Res* 2025;**53**:W547–W553. <https://doi.org/10.1093/nar/gkaf394>
- Chen H, Kihara D. Effect of using suboptimal alignments in template-based protein structure prediction. *Proteins Struct Funct Bioinform* 2011;**79**:315–34.
- Gotoh O. An improved algorithm for matching biological sequences. *J Mol Biol* 1982;**162**:705–8.
- Grigorjew A, Gynter A, Dias FHC *et al.* Sensitive inference of alignment-safe intervals from biodiverse protein sequence clusters using EMERALD. *Genome Biol* 2023;**24**:168. <https://doi.org/10.1186/s13059-023-03008-6>.
- Kaminski K, Ludwiczak J, Alva V *et al.* pLM-BLAST—distant homology detection based on direct comparison of sequence representations from protein language models, *bioRxiv*, pages 2022–11, 2022.
- Kanehisa M. Toward understanding the origin and evolution of cellular organisms. *Protein Sci* 2019;**28**:1947–51. <https://doi.org/10.1002/pro.3715>.
- Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 2000;**28**:27–30. <https://doi.org/10.1093/nar/28.1.27>
- Kanehisa M, Furumichi M, Sato Y *et al.* KEGG: biological systems database as a model of the real world. *Nucleic Acids Res* 2025; **53**:D672–7. <https://doi.org/10.1093/nar/gkae909>
- Llinares-López F, Berthet Q, Blondel M *et al.* Deep embedding and alignment of protein sequences. *Nat Methods* 2023; **20**:104–11.
- Naor D, Brutlag DL. On near-optimal alignments of biological sequences. *J Comput Biol* 1994;**1**:349–66.
- Petti S, Bhattacharya N, Rao R *et al.* End-to-end learning of multiple sequence alignments with differentiable Smith–Waterman. *Bioinformatics* 2023;**39**:btac724.
- Sehna D, Bittrich S, Deshpande M *et al.* Mol* Viewer: modern web app for 3D visualization and analysis of large biomolecular structures. *Nucleic Acids Res* 2021;**49**:W431–7. <https://doi.org/10.1093/nar/gkab314>
- Sierk ML, Smoot ME, Bass EJ *et al.* Improving pairwise sequence alignment accuracy using near-optimal protein sequence alignments. *BMC Bioinformatics* 2010;**11**:146–15.
- Utgés JS, MacGowan SA, Barton GJ. LIGYSIS-web: a resource for the analysis of protein-ligand binding sites. *Nucleic Acids Res* 2025;**53**:W351–60. <https://doi.org/10.1093/nar/gkaf411>
- van Kempen M, Kim SS, Tumescheit C *et al.* Fast and accurate protein structure search with foldseek. *Nat Biotechnol* 2024; **42**:243–6. <https://doi.org/10.1038/s41587-023-01773-0>
- Varadi M, Anyango S, Deshpande M *et al.* AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res* 2022;**50**:D439–44. <https://doi.org/10.1093/nar/gkab1061>
- Zahid SK, Hasan L, Khan AA *et al.* A novel structure of the Smith–Waterman Algorithm for efficient sequence alignment. In: *2015 Third International Conference on Digital Information, Networking, and Wireless Communications (DINWC)*, Moscow, Russia, p. 6–9. IEEE, 2015. <https://ieeexplore.ieee.org/document/7054208>