HHS Public Access

Author manuscript

J Mol Biol. Author manuscript; available in PMC 2021 September 09.

Published in final edited form as:

J Mol Biol. 2021 May 28; 433(11): 166828. doi:10.1016/j.jmb.2021.166828.

ModFlex: towards function focused protein modeling

Mayya Sedova[#a], **Lukasz Jaroszewski**[#a], **Mallika Iyer**[#b], **Zhanwen Li**[a], **Adam Godzik**[a]

[a]University of California Riverside School of Medicine, Biosciences Division, Riverside, CA

[b]Graduate School of Biomedical Sciences, Sanford Burnham Prebys Medical Discovery Institute, La Jolla, CA

[#] These authors contributed equally to this work.

Abstract

There is a wide, and continuously widening, gap between the number of proteins known only by their amino acid sequence versus those structurally characterized by direct experiment. To close this gap, we mostly rely on homology-based inference and modeling to reason about the structures of the uncharacterized proteins by using structures of homologous proteins as templates. With the rapidly growing size of the Protein Data Bank, there are often multiple choices of templates, including multiple sets of coordinates from the same protein. The substantial conformational differences observed between different experimental structures of the same protein often reflect function related structural flexibility. Thus, depending on the questions being asked, using distant homologs, or coordinate sets with lower resolution but solved in the appropriate functional form, as templates may be more informative. The ModFlex server (https://modflex.org/) addresses this seldom mentioned gap in the standard homology modeling approach by providing the user with an interface with multiple options and tools to select the most relevant template and explore the range of structural diversity in the available templates. ModFlex is closely integrated with a range of other programs and servers developed in our group for the analysis and visualization of protein structural flexibility and divergence.

Keywords

Homology-based modeling; Structure modeling; Structure prediction; Structural flexibility; Functional substates; Modeling flexibility

Introduction

There is a wide gap between the number of proteins known only by their amino acid sequence and the proteins with experimentally solved structures; a gap that is becoming exponentially wider with the rapid advances of next generation sequencing technologies. Homology-based inference is the most commonly used method to bridge this gap, and allows us to reason about the structures, and even the functions, of otherwise uncharacterized proteins. Homology modeling [1], sometimes referred to as comparative

Corresponding Author: adam.godzik@medsch.ucr.edu .

modeling [2] to highlight its pragmatic approach to the issue of evolutionary relations between proteins, is a popular method of predicting structures of novel proteins and starts with identification of a homolog with a known structure to be used as a template. The selection of the appropriate template, followed by its alignment to the target, is the most critical step in the homology-based structure modeling approach. However, many proteins have multiple coordinate sets available in the Protein Data Bank (PDB) [3], representing structures solved in different functional states or in randomly different experimental conditions [4, 5]. This multiplicity of coordinate sets is reflective of the fact that proteins are highly flexible, existing in a multitude of different conformational states that form a set called a conformational ensemble [6, 7]. This flexibility is intrinsic to the function of many proteins and the relationship between structure, its flexibility and protein function is an actively studied field [8–13].

Our group and others have leveraged the multiplicity of coordinate sets in the PDB to create tools for studying protein flexibility and conformational diversity, such as PDBFlex [14], CoDNaS [15] and PCDB [16]. Multiple studies have shown that flexibility is intrinsic to protein function [8, 10, 11] and features critical for the function of the protein are often present only in some conformational states. Despite this, most homology-based structure modeling programs automatically offer users a single “best” template [17, 18], which is usually the template with the highest sequence identity to the target and/or the best resolution. However, depending on the goals of the modeling project, a different template with lower sequence identity and/or lower resolution could be more useful, if it represents the appropriate conformational and functional state of the protein. Therefore, we propose that the template selection process should involve the evaluation of multiple possible templates in different conformational and functional states of the protein. Here, we present a newly developed tool, the ModFlex server (https://modflex.org), that addresses this seldom mentioned problem in structure modeling by providing users with an interface via which the most appropriate template can be selected, or the diversity of the available templates can at least be explored.

The ModFlex server uses information from the previously developed PDBFlex database [14]. On the PDBFlex server, each protein with multiple coordinate sets deposited in the PDB is represented by a cluster that is further subdivided into subclusters representing individual conformational states (groups of conformations within a preselected similarity threshold). Upon submission of a query sequence to ModFlex, clusters representing homologous proteins from the PDBFlex server are identified using BLAST [19] and representatives of all their subclusters are imported into ModFlex and presented as possible templates for modeling. The ModFlex server is further integrated with other servers developed by our group, namely FATCAT [20, 21] and POSA [22], that can be used to analyze and visualize the flexibility and structural divergence between different possible templates. Additionally, each subcluster is annotated as to its ligand binding/(hetero)complex formation status, thereby allowing the user to make the most informed decision regarding the optimal choice of template.

We believe that this tool will be useful for the structural biology community by providing users with a range of possible template structures grouped by their sequence and structural similarity, thus allowing them to explore their range of structural and functional diversity.

Results

The ModFlex server output provides a sortable list of close (i.e. detectable by BLAST [19] with a 0.05 e-value significance threshold) structurally characterized homologs of a target protein. The candidates for modeling templates are grouped into clusters containing distinct structures of the same protein and, within each cluster, they are grouped into subclusters representing the structural diversity of the corresponding protein. The lists of clusters and subclusters are provided by the PDBFlex database [14]. The clustering data (*Results table*– see Figure 1) allows better selection of a modeling template as it informs the user about the structural variability among available templates at different levels of sequence similarity to the original query (input sequence). By selecting the templates' subclusters and putting them into the *Analysis cart* the user can perform a more detailed analysis of the conformational diversity in the list of pre-selected templates. The structures representing subclusters (i.e. candidates for modeling templates) that were added to the *Analysis cart* can be compared to each other in terms of CαRMSD, contact map overlap, alignment length and sequence identity, with results displayed in the form of a *Comparison matrix* (Figure 1). Additionally, the information about ligands bound to each structure and the information on whether a structure is part of a (hetero)complex is shown in the *Results table* and is also indicated by colors at the edge of the *Comparison matrix*. More detailed information, such as the name of the protein, can be displayed in pop-up windows by clicking on a specific PDB code. The *Comparison matrix* can be rearranged by manually selecting and removing the modeling template candidates or automatically, according to their ligand status or (hetero)complex presence. The main purpose of the *Analysis cart* and the associated tools is to allow users to select the optimal template by answering the following questions: a) do conformational changes in a protein family of interest correlate with a ligand, cofactor binding or some other experimental parameter? b) does CαRMSD accurately describe structural variability between the considered templates or do differences in the alignment lengths confound information provided by CαRMSD? c) how far in the sequence similarity space does one need to search in order to collect templates which provide sufficient structural (and biochemical) variability for a given modeling problem?

The ModFlex Server: User Interface

The ModFlex server input page is a simple form accepting a sequence in the fasta format. It also provides links to previous queries from the same user in the browser's history and automatically detects if the newly submitted query matches one of the previous ones. The *Search results* page is divided into three sections (Figure 1): a) *Query information* part which provides information about the query sequence which was used to start the search b) *Analysis cart* where the user-selected subset of templates are compared and analyzed in more detail including three-dimensional rendering and animation describing conformational differences c) a sortable *Results table* which shows the output of the BLAST search against

the PDBFlex database with hits organized according to the clusters and subclusters defined in PDBFlex.

The suggested way of using ModFlex is to run a query with a sequence of interest and then inspect the *Results table* and evaluate the distribution of templates in terms of sequence similarity to the query and their ligand/(hetero)complex status. If there are any templates with high sequence similarity, the user can assess whether they provide sufficient breadth of possibilities i.e. conformations with and without ligands, being part of (hetero)complexes or monomers etc. If there are no close templates or they do not cover any/all ligand and (hetero)complex possibilities, how far in terms of sequence identity does one need to search in order to collect templates that are sufficiently diverse, conformationally and biochemically (ligands, (hetero)complexes)? Once these questions have been considered the user can select candidate modeling templates to be included in the *Analysis cart* (modifying the initial default selection provided by the ModFlex server) and analyze their structural and sequence variability in more detail. Specifically, the user can check a) the overall range of structural variability – whether there are significant structural differences between the collected templates b) whether these differences are correlated with ligand and (hetero)complex status (it often requires learning about specific types of ligands present and checking if protein engineering was applied to the templates). If there are no clear insights from the analysis step, the user can change the contents of the *Analysis cart* and repeat the analysis.

Once the analysis of collected templates is complete, users can obtain 3D models of interest by clicking the “Modeling” icons displayed next to the templates of choice. The choice of the templates and the alignments can also be exported to other modeling programs or servers. In the following sections we illustrate several possible user scenarios by providing examples of actual modeling problems.

Proof-of-concept example - alternative conformations can be predicted based on the corresponding conformations of a homolog

As suggested previously, users may want to model a specific conformation for a protein of interest and would therefore need to explore the range of conformational diversity in the available templates. Here we describe a proof-of-concept example that shows how ModFlex can be successfully used in such a scenario:

In this scenario a user wishes to model the ligand-bound conformation of the human lactotransferrin protein and submits the sequence of its N-terminal domain (SwissProt entry P02788 residues 1-362) to ModFlex.

The search results show a number of possible template clusters, with varying numbers of subclusters. After sorting this list in decreasing order of sequence identity, one cluster (2pmsA) has 100% sequence identity to the query, with two subclusters that represent the bound (1h43A) and unbound (1l5tB) conformations (Figure 2A). But if the bound conformation was not known, one could proceed down the list to find a different cluster containing ligand-bound and apo subclusters. The next template has 63.66% sequence identity to the query and contains two subclusters - one representing the ligand-bound

structure (1jqfA) and one representing the apo conformation (1bp5A). By adding these three subclusters: template identical to the query but without the ligand (1l5tB), and more distant templates with and without ligands (1jqfA and 1bp5A, respectively) to the *Analysis cart* and analyzing their pairwise CαRMSDs, one can see that the structure of the query protein with the ligand (1l5tB) is very similar to the distant template with the ligand (1bp5A) and that the unliganded structure of the distant template represents a different conformation (Figure 2B), this can be also seen by aligning the structures in POSA (Figure 2C). Here we already know the answer, and indeed, one can show that the model created using a more distant template with the correct functional form (1jqfA) is more accurate than the one obtained using a template with higher sequence identity but the wrong functional state (Figure 2D).

Applications - predictions of alternative conformations for structurally uncharacterized proteins.

Besides the proof-of-concept example demonstrating accurate modeling of the already known ligand-bound structure, we tested applications of ModFlex to modeling of the currently unknown protein structures. Here we show three examples where we see strong evidence that apo and ligand-bound conformations of a protein can be modeled with a high level of confidence, and one example illustrating a very common situation where such a prediction is difficult.

1. Calpains [23] are calcium-dependent cysteine proteases in which conformational changes induced by binding of Ca^{2+} ions are required for activation (forming of a functional catalytic site). Here we test the modeling of mouse calpain small subunit 2 (NCBI Protein database ID: NP_081388.1). The BLAST search implemented in ModFlex returns thirteen PDBFlex clusters. After selecting clusters with over 50% sequence identity to the query, adding their structural subclusters to the *Analysis cart* and recalculating the *Comparison* matrix one can see a clear separation of Ca^{2+}-bound and apo structures by CαRMSD (Figure 3A). This suggests that one can build distinct, reliable models of these conformations of the mouse calpain and, probably, many other mammalian calpains.

2. Single-stranded DNA binding proteins play important roles in the biology of bacterial genomes [24]. Here we tested if it is possible to model nucleic acid-bound and apo forms of bacterial exodeoxyribonuclease I from *Klebsiella pneumoniae* (NCBI Protein database ID: WP_048289581.1). The ModFlex search returns only one sequence-based cluster containing exodeoxyribonuclease I from *Escherichia coli* K12 with 86% sequence identity to the query protein. The cluster comprises six structural similarity-based subclusters – five of them contain structures of exodeoxyribonuclease complexed with nucleic acids and one represents the structure of the apo form. The *Comparison matrix* calculated for all these subclusters shows separation of nucleotide-bound and apo conformations (Figure 3B) indicating that they can be confidently predicted for the test query from *Klebsiella* (and for many other bacterial homologs). However, the fact that the apo form is represented by only one structure diminishes confidence in the models of the apo form.

3. Proteins from the S100 family present in vertebrates are involved in diverse functions including regulation of phosphorylation, cytoskeleton dynamics, Ca^{2+} homeostasis, cell proliferation and differentiation, regulation of transcription and enzymatic functions [25]. They all share EF-hand Ca^{2+}-binding motifs and undergo conformational changes upon Ca^{2+}-binding. To test the possibility of modeling these conformational changes based on the known structures of S100 proteins we performed a ModFlex search with the sequence of S100-A16 protein from mouse (NCBI Protein database ID: NP_001343534.1) that lacks experimental structural characterization. Among the six templates with the lowest BLAST E-values, four represent Ca^{2+}-bound forms and two represent apo conformations. The bound and apo sets form two well-separated groups in the CαRMSD all-to-all comparison (Figure 3C) despite having quite diverse sequences. This suggests that the bound and apo conformations of the mouse S100-A16 protein and, most likely, other members of the S100 family can be reliably modelled based on existing experimental structures.

4. Beta-lactamases are a large family of enzymes which provide resistance to β-lactam antibiotics [26] and new, more potent variants of this enzyme are responsible for the recent emergence of multi-drug resistant strains of many pathogens (super-bugs) [27]. The attempt to model peptide-bound and apo conformations of NDM beta-lactamase 1 from *Acinetobacter baumannii* (NCBI Protein database ID: BBA83870.1) yields multiple templates in apo and ligand bound forms. Ligands include Mg^{2+}, Zn^{2+} ions, ampicillin and others. The CαRMSD all-to-all comparison indicates significant structural differences between the collected templates but these structures do not seem to form separate groups in the CαRMSD matrix when they are divided according to the presence of any of the ligands. This is quite a common scenario where there is no straightforward, easily observable relationship between ligands and protein conformations. While in such cases ligand-bound and apo forms cannot be confidently predicted, we assume that users with extensive knowledge of a specific protein family can still select optimal templates and obtain accurate models of the unknown structure of interest, by studying the known templates in more detail.

Modeling on templates with intermediate conformations between two experimental structures.

ModFlex provides a novel “experimental” function of building models based on templates which are conformational intermediates between any two experimentally characterized structural templates found for a given query. The intermediate templates are obtained using the morphing algorithm originally implemented in the FATCAT server and models are built on these structural intermediates using MODELLER [28]. This function could be useful in situations when ligand-bound and apo conformations are known and separated by some conformational distance (as indicated by CαRMSD) and, at the same time, some unknown structure of interest is expected to adopt an intermediate conformation (for example, the structure of an enzyme bound to an inhibitor or a transition-state analog). As shown

previously by Weiss and Levitt [29], interpolation algorithms and, specifically, the method developed by our group for the FATCAT server (and used here) can, in some cases, generate accurate models of such intermediate forms. However, morphing interpolations produce a series of models along the conformational trajectory and it is very difficult to decide which particular intermediate (i.e. model from the morphing trajectory) could be a good approximation of some unknown structure of interest. Consequently, this function may only have some limited applications - for example it can be used to perform extensive “sampling” of possible conformations without making a specific prediction. The models produced by such sampling could be used in applications of the molecular replacement method or X-ray crystallography [30, 31] to structurally flexible proteins in cases where models based on individual experimentally characterized templates do not provide successful phasing of crystallographic data.

Discussion

In this manuscript we describe the ModFlex server, which provides an interactive interface to support the process of choosing an optimal template for homology/comparative structure modeling. If a user is interested in modeling a specific functional form of a protein, the automated choice of the most similar template may not necessarily be optimal. The classical example of function driven protein conformational changes involves ligand binding. Others may include changes incurred by forming protein complexes or undergoing post-translational modifications. ModFlex offers a solution to this problem in cases where conformational changes are clearly correlated with the presence of specific ligands or other biochemical features, but it may also help users to identify templates that cover a broader range of potential structural variability for a given query in cases where such a correlation cannot be easily detected.

The accuracy of static homology-based models mostly depends on the template selection and accuracy of the alignment. There are many methods which allow assessment of the quality of a model based on the agreement of the model’s structural features with the distribution of these features in experimentally characterized native protein structures (see Table 1). We suggest using these methods in situations where similarity between the query sequence and the modeling templates is low and there are substantial differences between alternative homology-based models to choose from. These methods may also help in assessment of how realistic the models obtained by the interpolation between two template structures are (as the geometry of these models may be significantly distorted).

Arguably, the most important question related to flexibility-focused protein modeling is how the conservation of protein flexibility patterns changes with decreasing sequence identity between the query sequence and the template. This issue deserves a separate publication that is now in preparation. Our preliminary results indicate that the extent of similarity in the flexibility patterns of homologous proteins pairs drops significantly when the sequence identity of the pair is below 50%. Sequence similarity at this level can be still detected by the standard BLAST algorithm and this was the reason why we decided not to use more sensitive homology detection methods (since in more remote homologs, structural flexibility patterns of the template may not be predictive of flexibility patterns in the query protein).

There is evidence that proteins belonging to the same fold often undergo similar structural changes [32] even without similarity at the sequence level. Therefore, in the future, we may systematically identify folds where this is the case and, for these proteins, collect even very distant modeling templates.

Materials and Methods

Back-end functions

The back-end functions of the ModFlex server are based on the PDBFlex [14] database. In the first step, the back-end of the ModFlex server receives a query sequence and uses it to start a BLAST [33] search against the sequences of protein chains from the PDB database clustered at 95% sequence which are retrieved from the PDBFlex database. In the next step, for each statistically significant BLAST hit, all structural similarity based subclusters are pulled from the PDBFlex database. Protein structures representing these subclusters are made available to the user as possible modeling templates for the submitted query. The information about presence of ligands and information on whether template structures were determined as heterocomplexes are also pulled from PDBFlex. PDBFlex obtains information about ligands from the BioLiP database [34] and information about heterocomplexes is based on the mapping of PDB IDs to Uniprot IDs retrieved from the SIFTS [35] database.

Other functions performed by the back-end include:

- **a.** All-to-all sequence alignment followed by structure superposition, to provide CaRMSD and contact map overlap for a user-selected subset of modeling templates. The alignment is calculated with BLAST and CaRMSD and contact map overlap are calculated with in-house programs.
- **b.** Interpolation (morphing) between a user-selected pair of templates. The morphing trajectory is calculated using a program developed by our group and originally implemented on the FATCAT server [20, 21].
- **c.** Homology-based structure modeling of a query protein based on a single user-selected template or based on a selected morphing trajectory (series of templates) is performed with SCWRL [36] or MODELLER [28].

Front-end interface and functions

The front-end of the ModFlex website is written in JavaScript using the Vue.js (https://vuejs.org/) framework. Interactive comparison matrix display was developed using D3.js [37] visualization. Morphing trajectories and animations illustrating differences between a selected pair of templates are displayed in 3D using an interactive 3DMol.js library [38].

Information about the query sequence, results, user's template selections and modelling status are saved as session storage by the user's browser to enable quick access to previously run queries. These saved queries could be accessed or deleted by the user from the home page of the server. The front-end communicates with the back-end over Hypertext Transfer Protocol Secure (HTTPS).

Acknowledgments

This research was funded in part by the NIH NIGMS R35 GM118187 and NIH NIAID contract HHSN272201700060C

References

[1]. Schwede T, Kopp J, Guex N, Peitsch MC. SWISS-MODEL: An automated protein homology-modeling server. Nucleic Acids Res. 2003;31:3381–5. [PubMed: 12824332]

[2]. Marti-Renom MA, Stuart AC, Fiser A, Sanchez R, Melo F, Sali A. Comparative protein structure modeling of genes and genomes. Annu Rev Biophys Biomol Struct. 2000;29:291–325. [PubMed: 10940251]

[3]. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al.The Protein Data Bank. Nucleic Acids Res. 2000;28:235–42. [PubMed: 10592235]

[4]. Burra PV, Zhang Y, Godzik A, Stec B. Global distribution of conformational states derived from redundant models in the PDB points to non-uniqueness of the protein structure. Proc Natl Acad Sci U S A. 2009;106:10505–10. [PubMed: 19553204]

[5]. Kosloff M, Kolodny R. Sequence-similar, structure-dissimilar protein pairs in the PDB. Proteins. 2008;71:891–902. [PubMed: 18004789]

[6]. Cooper AThermodynamic fluctuations in protein molecules. Proc Natl Acad Sci U S A. 1976;73:2740–1. [PubMed: 1066687]

[7]. Frauenfelder H, Sligar SG, Wolynes PG. The energy landscapes and motions of proteins. Science. 1991;254:1598–603. [PubMed: 1749933]

[8]. Henzler-Wildman KA, Lei M, Thai V, Kerns SJ, Karplus M, Kern D. A hierarchy of timescales in protein dynamics is linked to enzyme catalysis. Nature. 2007;450:913–6. [PubMed: 18026087]

[9]. Bahar I, Lezon TR, Yang LW, Eyal E. Global dynamics of proteins: bridging between structure and function. Annu Rev Biophys. 2010;39:23–42. [PubMed: 20192781]

[10]. Boehr DD, McElheny D, Dyson HJ, Wright PE. The dynamic energy landscape of dihydrofolate reductase catalysis. Science. 2006;313:1638–42. [PubMed: 16973882]

[11]. Eisenmesser EZ, Millet O, Labeikovsky W, Korzhnev DM, Wolf-Watz M, Bosco DA, et al.Intrinsic dynamics of an enzyme underlies catalysis. Nature. 2005;438:117–21. [PubMed: 16267559]

[12]. Ramanathan A, Agarwal PK. Evolutionarily conserved linkage between enzyme fold, flexibility, and catalysis. PLoS Biol. 2011;9:e1001193. [PubMed: 22087074]

[13]. Monzon AM, Zea DJ, Marino-Buslje C, Parisi G. Homology modeling in a dynamical world. Protein Sci. 2017;26:2195–206. [PubMed: 28815769]

[14]. Hrabe T, Li Z, Sedova M, Rotkiewicz P, Jaroszewski L, Godzik A. PDBFlex: exploring flexibility in protein structures. Nucleic Acids Res. 2016;44:D423–8. [PubMed: 26615193]

[15]. Monzon AM, Rohr CO, Fornasari MS, Parisi G. CoDNaS 2.0: a comprehensive database of protein conformational diversity in the native state. Database (Oxford). 2016;2016.

[16]. Juritz EI, Alberti SF, Parisi GD. PCDB: a database of protein conformational diversity. Nucleic Acids Res. 2011;39:D475–9. [PubMed: 21097895]

[17]. Roy A, Kucukural A, Zhang Y. I-TASSER: a unified platform for automated protein structure and function prediction. Nat Protoc. 2010;5:725–38. [PubMed: 20360767]

[18]. Webb B, Sali A. Comparative Protein Structure Modeling Using MODELLER. Curr Protoc Bioinformatics. 2016;54:5 6 1–5 6 37. [PubMed: 27322406]

[19]. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, et al.Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25:3389–402. [PubMed: 9254694]

[20]. Ye Y, Godzik A. FATCAT: a web server for flexible structure comparison and structure similarity searching. Nucleic Acids Res. 2004;32:W582–5. [PubMed: 15215455]

[21]. Li Z, Jaroszewski L, Iyer M, Sedova M, Godzik A. FATCAT 2.0: towards a better understanding of the structural diversity of proteins. Nucleic Acids Res. 2020;48:W60–W4. [PubMed: 32469061]

[22]. Li Z, Natarajan P, Ye Y, Hrabe T, Godzik A. POSA: a user-driven, interactive multiple protein structure alignment server. Nucleic Acids Res. 2014;42:W240–5. [PubMed: 24838569]

[23]. Suzuki K, Hata S, Kawabata Y, Sorimachi H. Structure, activation, and biology of calpain. Diabetes. 2004;53 Suppl 1:S12–8. [PubMed: 14749260]

[24]. Lu D, Keck JL. Structural basis of Escherichia coli single-stranded DNA-binding protein stimulation of exonuclease I. Proc Natl Acad Sci U S A. 2008;105:9169–74. [PubMed: 18591666]

[25]. Donato RS100: a multigenic family of calcium-modulated proteins of the EF-hand type with intracellular and extracellular functional roles. Int J Biochem Cell Biol. 2001;33:637–68. [PubMed: 11390274]

[26]. Livermore DM. beta-Lactamases in laboratory and clinical resistance. Clin Microbiol Rev. 1995;8:557–84. [PubMed: 8665470]

[27]. Guo Y, Wang J, Niu G, Shui W, Sun Y, Zhou H, et al.A structural view of the antibiotic degradation enzyme NDM-1 from a superbug. Protein Cell. 2011;2:384–94. [PubMed: 21637961]

[28]. Webb B, Sali A. Protein structure modeling with MODELLER. Methods Mol Biol. 2014;1137:1–15. [PubMed: 24573470]

[29]. Weiss DR, Levitt M. Can morphing methods predict intermediate structures?J Mol Biol. 2009;385:665–74. [PubMed: 18996395]

[30]. Rossmann MG. Molecular replacement--historical background. Acta Crystallogr D Biol Crystallogr. 2001;57:1360–6. [PubMed: 11567146]

[31]. Schwarzenbacher R, Godzik A, Grzechnik SK, Jaroszewski L. The importance of alignment accuracy for molecular replacement. Acta Crystallogr D Biol Crystallogr. 2004;60:1229–36. [PubMed: 15213384]

[32]. Keskin O, Jernigan RL, Bahar I. Proteins with similar architecture exhibit similar large-scale dynamic behavior. Biophys J. 2000;78:2093–106. [PubMed: 10733987]

[33]. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215:403–10. [PubMed: 2231712]

[34]. Yang J, Roy A, Zhang Y. BioLiP: a semi-manually curated database for biologically relevant ligand-protein interactions. Nucleic Acids Res. 2013;41:D1096–103. [PubMed: 23087378]

[35]. Dana JM, Gutmanas A, Tyagi N, Qi G, O'Donovan C, Martin 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 Res. 2019;47:D482–D9. [PubMed: 30445541]

[36]. Wang Q, Canutescu AA, Dunbrack RL Jr. SCWRL and MolIDE: computer programs for side-chain conformation prediction and homology modeling. Nat Protoc. 2008;3:1832–47. [PubMed: 18989261]

[37]. Bostock M, Ogievetsky V, Heer J. D(3): Data-Driven Documents. IEEE Trans Vis Comput Graph. 2011;17:2301–9. [PubMed: 22034350]

[38]. Rego N, Koes D. 3Dmol.js: molecular visualization with WebGL. Bioinformatics. 2015;31:1322–4. [PubMed: 25505090]

[39]. Cheng J, Choe MH, Elofsson A, Han KS, Hou J, Maghrabi AHA, et al.Estimation of model accuracy in CASP13. Proteins. 2019;87:1361–77. [PubMed: 31265154]

[40]. Benkert P, Biasini M, Schwede T. Toward the estimation of the absolute quality of individual protein structure models. Bioinformatics. 2011;27:343–50. [PubMed: 21134891]

[41]. Uziela K, Menendez Hurtado D, Shu N, Wallner B, Elofsson A. ProQ3D: improved model quality assessments using deep learning. Bioinformatics. 2017;33:1578–80. [PubMed: 28052925]

[42]. Maghrabi AHA, McGuffin LJ. Estimating the Quality of 3D Protein Models Using the ModFOLD7 Server. Methods Mol Biol. 2020;2165:69–81. [PubMed: 32621219]

[43]. Olechnovic K, Venclovas C. VoroMQA: Assessment of protein structure quality using interatomic contact areas. Proteins. 2017;85:1131–45. [PubMed: 28263393]

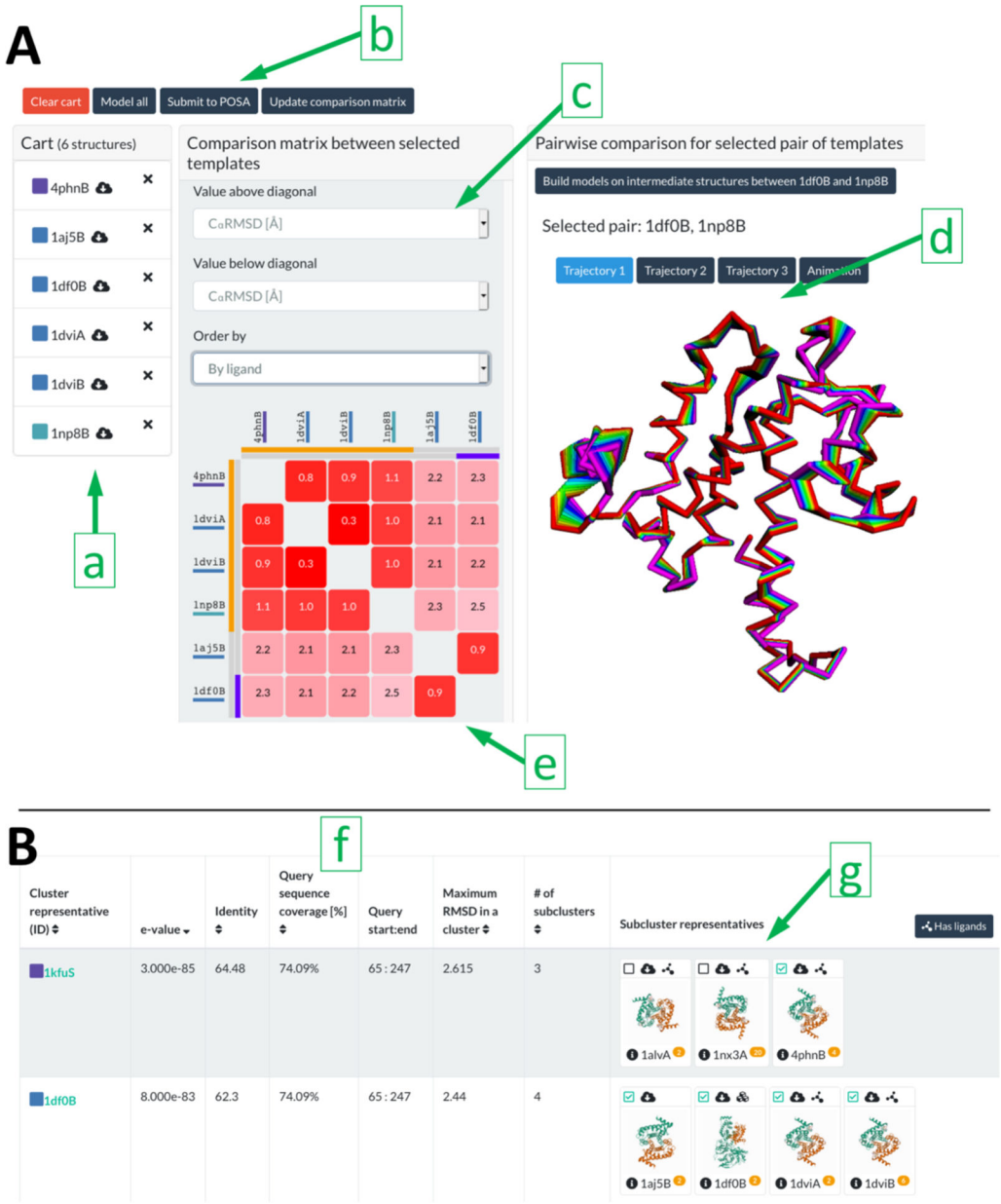

Figure 1.

A) Analysis cart section of the ModFlex interface allows in-depth structural analysis of a selected subset of modeling templates B) Results table shows results of the BLAST search performed on sequences representing PDBFlex clusters.

a) the list of structural templates currently included in the Analysis cart

b) buttons providing access to internal functions and external tools

c) selection fields controlling contents and ordering of the comparison matrix

d) 3D comparison panel; buttons allow selection of specific graphical representations of differences for a selected pair of structures

e) comparison matrix (by clicking on the table's cells users can select pairs of structures which will be compared in the 3D comparison panel on the right)

f) parameters extracted from the BLAST output (each BLAST hit corresponds to one PDBFlex cluster represented by a single sequence)
g) structural similarity-based subclusters shown for each PDBFlex cluster

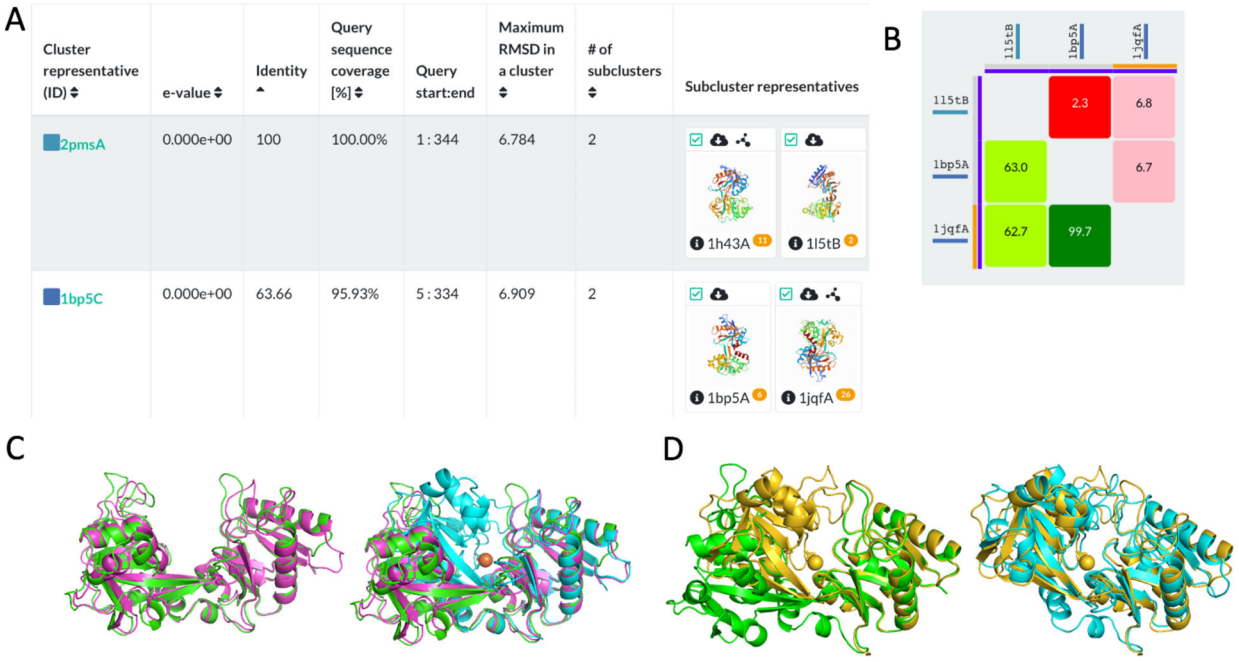

Figure 2.

Proof-of-Concept example - ModFlex results for human lactotransferrin whose ligand-bound and unbound structures are known (1h43A and 1l5tB PDB accessions and chain IDs for the bound and unbound structures, respectively). A) Results list showing top two templates based on sequence identity. B) Comparison matrix for selected templates showing CαRMSD above the axis and sequence ID below the axis (1h43A was excluded from comparison matrix). Chains 1l5tB and 1bp5A (apo conformations of lactotransferrin and transferrin) are very similar as indicated by the low CαRMSD between them. Chain 1jqfA (ligand-bound conformation of transferrin) has two times higher CαRMSD distance to the other two structures. Assuming that the ligand-bound conformation (1h43A) is unknown one would have a choice between using 1l5tB apo-lactotransferrin template (100% identical to the query) or ligand-bound human transferrin (only 63.66% sequence identity to the query) C) POSA alignment of 1bp5A (magenta), 1l5tB (green) and 1jqfA (cyan), showing the conformational similarity between 1bp5A and 1l5tB. D) POSA alignment of models created using the templates 1l5tB (green) and 1jqfA (cyan), and the “true” target structure, 1h43A (gold). This shows that 1jqfA is the optimal template if the user wants to create a model of the liganded conformation, even though its sequence identity to the query is lower than 1l5tB.

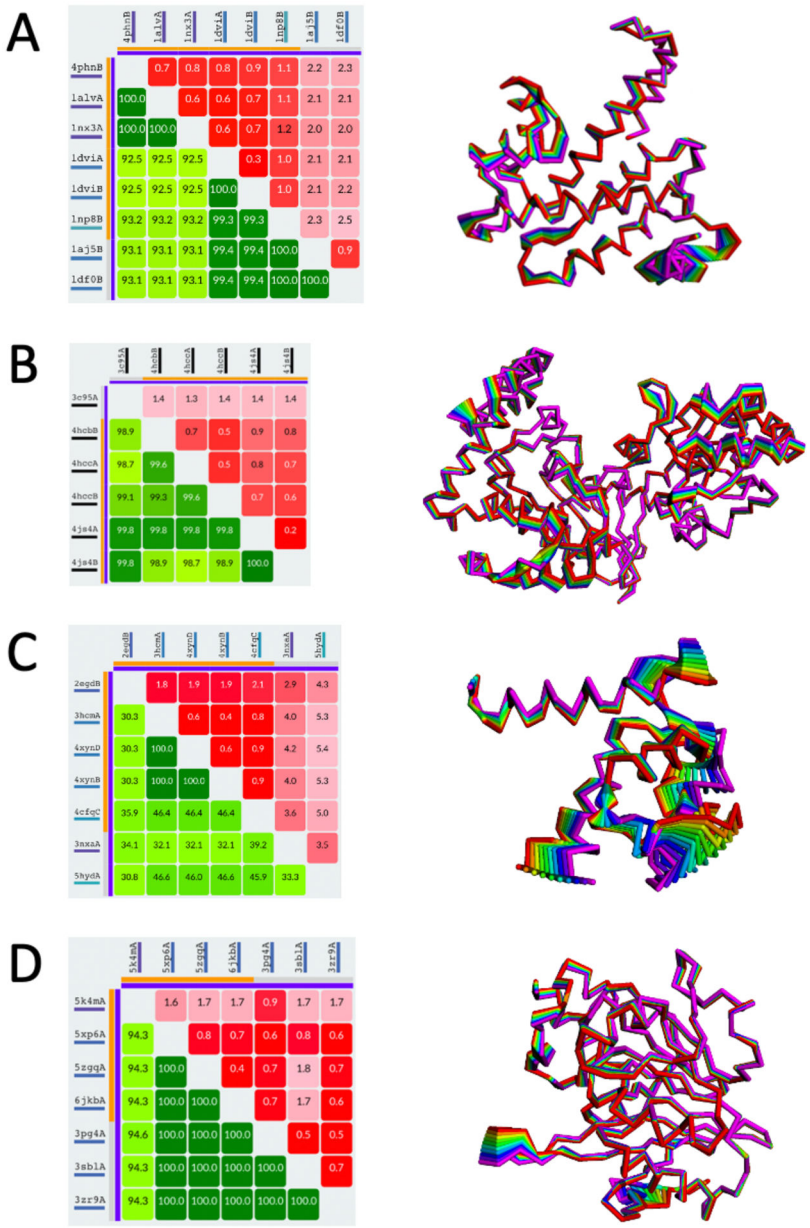

Figure 3.

Examples of comparison matrices showing CαRMSD above the axis and sequence ID below the axis (left column) and conformational transition trajectories between selected template structures (right column), where ligand-bound and ligand-free structures form two distinct clusters separated by larger conformational distance, thus allowing the user to predict (with some confidence) the ligand-bound and unbound conformations of their homologs without experimentally characterized structures. Metallo-beta-lactamases (D) represent a counterexample where no such clear trend can be observed.

A. Top modeling templates found for calpain small subunit 2 from *Mus musculus*, NCBI Protein database ID: NP_081388.1.

B. Top modeling templates found for exodeoxyribonuclease I from *Klebsiella pneumoniae*, NCBI Protein database ID: WP_048289581.1.

C. Top modeling templates found for protein S100-A16 from *Mus musculus*, NCBI Protein database ID: NP_001343534.1.

D. Top modeling templates found for NDM 1 from *Acinetobacter baumannii*, NCBI Protein database ID: BBA83870.1.

Table.

The selected popular webservers for evaluation of homology models of proteins. For in-depth discussion and comparison of these and other methods, please see description of the CASP13 experiment results [39].

Method	Server's URL	Reference
QMEAN	https://swissmodel.expasy.org/qmean/	[40]
ProQ3D	https://proq3.bioinfo.se/	[41]
ModFOLD8	https://www.reading.ac.uk/bioinf/ModFOLD/ModFOLD8_form.html	[42]
VoroMQA	http://bioinformatics.ibt.lt/wtsam/voromqa	[43]