W60–W64 *Nucleic Acids Research, 2020, Vol. 48, Web Server issue* *Published online 29 May 2020*
doi: 10.1093/nar/gkaa443

FATCAT 2.0: towards a better understanding of the structural diversity of proteins

Zhanwen Li[1], Lukasz Jaroszewski[1], Mallika Iyer[2], Mayya Sedova[1] and Adam Godzik[1,*]

[1]Division of Biomedical Sciences, University of California Riverside School of Medicine, Riverside, CA 92521, USA and [2]Graduate School of Biomedical Sciences, Sanford Burnham Prebys Medical Discovery Institute, La Jolla, CA 92037, USA

Received March 17, 2020; Revised April 28, 2020; Editorial Decision May 11, 2020; Accepted May 19, 2020

ABSTRACT

FATCAT 2.0 server (http://fatcat.godziklab.org/), provides access to a flexible protein structure alignment algorithm developed in our group. In such an alignment, rotations and translations between elements in the structure are allowed to minimize the overall root mean square deviation (RMSD) between the compared structures. This allows to effectively compare protein structures even if they underwent structural rearrangements in different functional forms, different crystallization conditions or as a result of mutations. The major update for the server introduces a new graphical interface, much faster database searches and several new options for visualization of the structural differences between proteins

INTRODUCTION

The FATCAT server (http://fatcat.godziklab.org/) provides access to the flexible protein structure alignment program developed in our group (1). It is part of the protein structure analysis development environment that also includes the POSA (2) and PDBFlex servers (3). A java version of the FATCAT algorithm, jFATACT (4) is used as the default structure similarity search algorithm at the RCSB PDB portal (5).

Protein structure comparison has a long history (6) and many popular algorithms address this problem using different heuristics and concepts (7). However, most of the structure comparison programs treat proteins as rigid bodies despite the fact that the proteins are highly dynamic and flexible. The flexibility of protein structures is increasingly being appreciated as important to their function and many proteins are now being solved in different conformations reflecting different functional states. Flexible protein structure alignment algorithms address the problem of comparing structures in different conformational states by introducing special types of gaps ('twists') in the alignment that allow for translations and rotations of parts of the structure (1,8).

Another limitation of most of the existing structure comparison programs is that they typically focus on the question of evaluating the similarity/difference between two structures by a simple numerical score, with the root mean square difference (RMSD) being the most popular choice. Such a score is invaluable in classifying proteins into families or folds but is less useful for describing and visualizing the differences between otherwise similar structures. For instance, many structural biology studies focus on the description of changes brought about by substrate binding or complex formation and simple RMSD values do not capture the fact that despite sometimes large global structural differences, proteins undergoing such changes remain highly similar. For instance, the RMSD between two conformations of the same protein may be as high as the RMSD between two structures without any similarity.

The FATCAT server was originally described in the NAR webserver issue in 2004 (9), this first major upgrade is based on the same algorithm, but provides several new functionalities, such as a morphing movie and several interactive visualization options as well as new interface and significant speed-up in the database searches.

BASIC USE OF THE FATCAT SERVER

We developed the FATCAT server to address two challenges: aligning protein structures in different conformations and visualizing the differences between them in an intuitive way. The main server page provides a general introduction to the flexible protein structure alignment problem and gives users several choices in the menu at the top of the page. In particular, the user can choose from the following options: PAIRWISE ALIGNMENT, DATABASE SEARCH, HOMOLOGY SEARCH, REFERENCES, HELP, OTHER SERVERS, GODZIK LAB. In the following, we will discuss the first three options, as the others are self-explanatory and deal mostly with housekeeping features of the server.

*To whom correspondence should be addressed. Tel: +1 951 827 7276; Email: adam.godzik@medsch.ucr.edu

Pairwise structural alignment (PAIRWISE ALIGNMENT option)

For basic pairwise alignment, the input to the FATCAT server consists of the PDB IDs of two protein chains (users can also upload their own structures for comparison). In addition, the user can to choose between the default 'flexible' alignment option and the 'rigid' option, with the latter setting the flexibility gap penalty to infinity and thus running the standard, rigid body alignment. The output provides the 'flexible' or 'rigid' alignment of the two structures, together with several options for visualizing the structural differences between them. Since the original NAR manuscript (9) several new interactive visualization options were added to the server. In addition to the widely used display of the overlapped structures in an optimal superposition (see Figure 1, column A), the difference between structures can be visualized as a C-α displacement field (Figure 1, column B), series of intermediate structures (Figure 1, column C) or a difference distance—or—contact map (DDM or DCM, respectively)—see Figure 1D). FATCAT also provides a trajectory of one structure 'morphing' into the other visualized as an animation (see column E in an animated version of Figure 1 at: http://fatcat.godziklab.org/fatcathelp_files/FATCAT_fig_a.ppsx). To calculate the conformational path between the two compared structures, FATCAT uses a novel morphing algorithm (Rotkiewicz *et al.*, in preparation), which was independently evaluated and was found to correctly reproduce experimentally characterized intermediate structures on the trajectory between two conformations (10). In the current version of the server, all visualizations use 3dmol which provides high quality images and animations (11).

The examples shown in Figure 1 illustrate several types of structural differences and their FATCAT results and visualizations. The first row shows a comparison between an apo (12) and nitric oxide bound (13) structures of myoglobin. The structural differences are mostly limited to a s single loop, with RMSD of 2.52 Å and no flexibility gaps in the alignment. It is interesting to note that the differences in the loop closing the binding site are best visible in the cartoon superposition (column A) and the difference distance map (column D), but the displacement and intermediate structure visualization (columns B and C, respectively) show smaller changes across the entire structure. Second row shows the induced fit in *Escherichia coli* Isocitrate Dehydrogenase (14,15). There the two rigid subdomains move in respect to each other. This is easiest to see in the Difference Distance map, where large almost white blocks along the diagonal identify the almost rigid subdomains. The third example shows combined subdomain rearrangement and structural changes within one of the subdomains upon calcium binding in the rabbit Troponin C (16,17). The last example is used as a default test example on the FATCAT website.

Search by structural similarity (DATABASE SEARCH option)

Usually the first question asked by structural biologists after characterizing a novel protein structure is: 'What is the structure similar to?' Unless homology to structurally characterized proteins is evident from sequence similarity, answering this question is often a non-trivial task. The answer often provides valuable insights into the protein's function and its possible evolutionary relationships.

Publicly available structure comparison algorithms, the best known of which is the Dali server (18,19) provide an option for comparing a query structure to known protein structures stored in the PDB database. By providing the database search option on the FATCAT server we aim to provide an option for searching deeper in the structure space for similarities that could be missed with a rigid search. At the same time, FATCAT is based on different heuristics than Dali (or other similar programs) and often identifies novel similarities, not detected or not scored highly by other approaches.

The FATCAT server can be used to search the PDB (5) or SCOPe domain (20) databases, each clustered at 40% or 90% sequence similarity, for similar structures. RCSB PDB (5) uses jFATCAT (4) with the rigid option to prepare lists of similar structures for all PDB entries and makes them available from the 'structure similarity' PDB pages. As compared to the original server, the search algorithm has been optimized and most searches can be completed in less than 10 minutes. The output consists of a list of similar structures rank-ordered by the estimated statistical significance of the structural similarity, with result pages for all the similar structures provided in the same format as the pairwise comparison discussed above. The searches can also be carried out using a structure uploaded by the user.

Here, again, the main advantage of the FATCAT server is that in structure similarity searches it goes beyond rigid body similarity. Even in the case of single domain proteins, one internal 'hinge movement' may make structural similarity difficult to recognize visually and by standard rigid-body comparisons. The option of searching the PDB and SCOP databases by structural similarity was available in the original FATCAT server. However, in the new server the format of the output of this search is redesigned, additional visualizations are available and the individual pairwise result pages follow the new format described in the previous section. New interactive visualization of the distribution of *P*-value against other alignment features is now available to assist with the selection of the most biologically relevant structural similarity (see Figure 2A). This is not always a simple task since structure comparison algorithms balance minimizing RMSD while maximizing the length of the alignment. Similar to other servers, the FATCAT server provides a single score (here, the *P*-value) describing the significance of the hits but in some cases the most interesting similarities are not captured by the ranking of *P*-values.

The interactive chart can display scatter plots of any pair of the following variables: structure length (length), *P*-value (pvalue), alignment length (opt-len), RMSD (opt-rmsd), full alignment length (align-len), total length of gaps (gap) and sequence identity (seq-identity). Users can select a pair of these variables to be used as X and Y axes of the plot and the third variable to be used for coloring of plot points (rainbow color scale is displayed below the chart). By clicking a point on the chart user can select structural similarities between the query structure and structures from

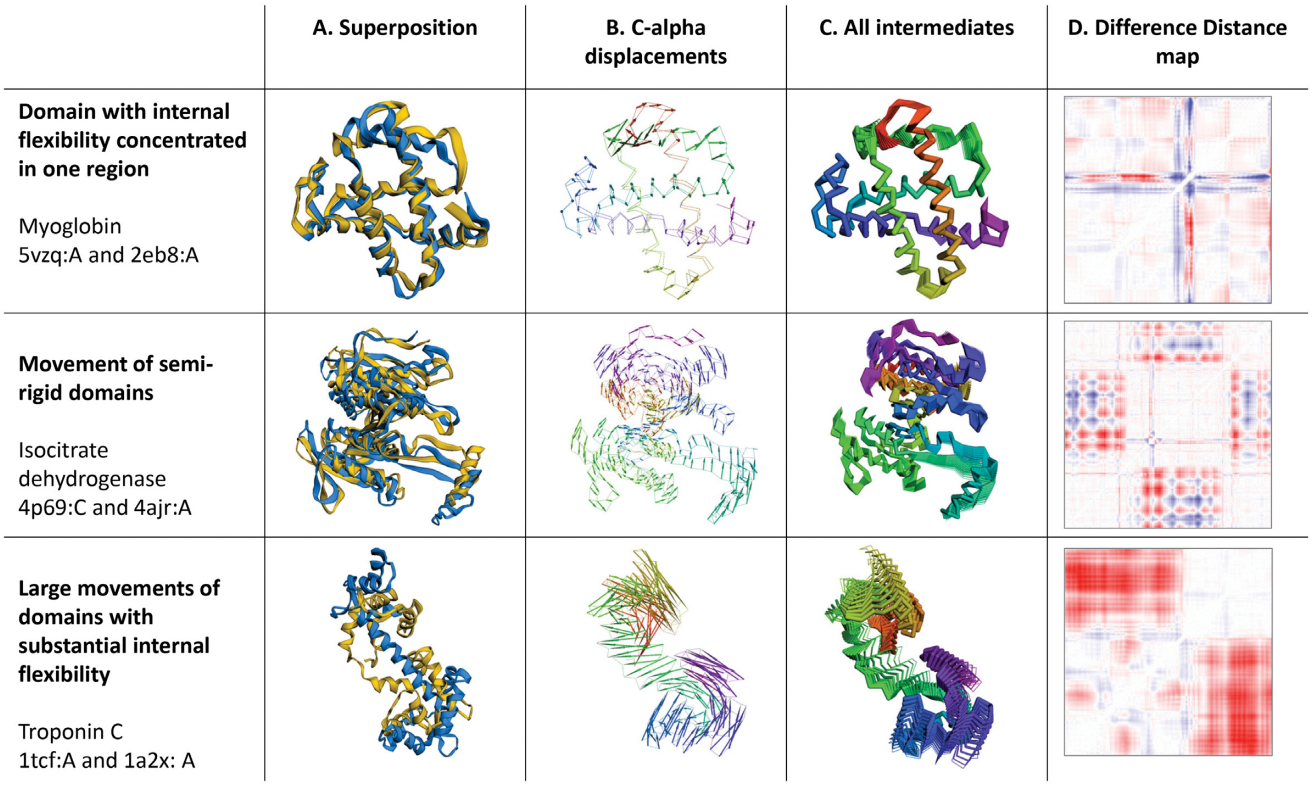

Figure 1. Different types of protein flexibility illustrated by FATCAT. Three pairs of protein structures with different character of conformational differences were compared using FATCAT. In contrast to traditional structural superposition (left-most column) graphically presented FATCAT results make it possible to quickly assess the character of conformational differences between two structures. For instance, C-alpha displacements reveal regions where rotational movements and DDMs allow identification of internally rigid domains (diagonally located squares without substantial changes in distances e.g. isocitrate dehydrogenase) and regions of high local flexibility (narrow 'stripes' of large conformational changes as seen in DDM of myoglobin)

the database. The selections made in the chart and in the list of 'hits' below the chart are connected. The list of similarities can be filtered and sorted by each of the output parameters. It is also possible to extract the selected list of hits, alignments, sequences or structures in the text format. At the same time each of the individual pairwise comparisons found during the search can be examined via the same interface that is used for pairwise structural comparison (see the section *Pairwise Alignment* and Figure 1).

In the example shown in Figure 2, the SCOPe domain database was searched with the structure of the rabbit troponin C (PDB 1tcf:A). As expected, the significant results all belong to the same SCOP fold a.39 (EF-hand like). As seen in the Figure 2 A, the alignment lengths have wide distribution with two regions corresponding to two-domain and single domain structures of EF-hand proteins. Coloring by sequence identity reveals the closest homologs which tend to have the most similar structures (and the lowest *P*-values), but it is not always the case (as *P*-values are based solely on structural similarity).

Structural comparison of homologous proteins (HOMOLOGY SEARCH option)

Finding proteins with structures similar to the structure of interest is not the only question which can be answered by structure comparison algorithms. Even if the protein or proteins of interest show strong sequence similarity (implying similarity of their structures), flexible structural comparison may reveal non-trivial conformational differences. Such differences between close homologs or even different structures of the same proteins are the result of natural protein flexibility. They are at least as interesting from the biological point of view as remote similarities between distant homologs, as they are often essential to protein function.

In the current version of the FATCAT server, there is an option to search for closely homologous proteins by *sequence* similarity. The output looks similar to that of the database search, with the results rank ordered by *P*-value and links to the pairwise alignment analysis pages (see the section *Pairwise Alignment* and Figure 1) provided in the output. The result lists from a classical structure similarity and homology search could be partly overlapping, but search by sequence similarity is two orders of magnitude faster, at the same time, limits the output to proteins whose homology can be easily recognized by a sequence search alone. In the HOMOLOGY SEARCH option, the only search database available is the whole PDB database with no preclustering.

CONCLUSIONS AND FUTURE DIRECTIONS

As the focus of structural biology changes to the analysis of the molecular details of protein function and most

Figure 2. Interactive output of structure similarity search. The SCOP database clustered at 40% sequence identity was queried by the structure of troponin C (PDB entry: 1tcf chain A). (**A**) The scatter plot of PDB 'hits' identified in the search. X-axis: FATCAT *P*-value in logarithmic scale. Y-axis: length of the alignment. (**B**) The top of the list of PDB 'hits' with options for searching, sorting and filtering. The selection of results can be made on the scatter plot or on the list of results.

proteins are now solved in multiple of functional states, tools for structure analysis has to provide new functionality. The updated FATCAT server is providing new visualizations to add in this goal. The server is continuously updated and some of the new features, still not described in this manuscript, may become available after this manuscript is published.

DATA AVAILABILITY

FATCAT server (http://fatcat.godziklab.org) is freely available to all users. Linux executables of the program are available for download from the server homepage. The source code is available by request and will be released as open source in the future.

ACKNOWLEDGEMENTS

We want to acknowledge the PDB team for providing API access to the PDB database that is used by our server, as well as to all individual crystallography groups that deposit their coordinates to the PDB. Many ideas implemented in the current version of the FATCAT server originated in the discussions with the author of the original version of the server, Dr Yuzhen Ye.

FUNDING

NIH NIGMS [R35 GM118187]. Funding for open access charge: UCR Institutional Funds.
Conflict of interest statement. None declared.

REFERENCES

1. Ye,Y. and Godzik,A. (2003) Flexible structure alignment by chaining aligned fragment pairs allowing twists. *Bioinformatics*, **19**(Suppl. 2), ii246–ii255.
2. Li,Z., Natarajan,P., Ye,Y., Hrabe,T. and Godzik,A. (2014) POSA: a user-driven, interactive multiple protein structure alignment server. *Nucleic Acids Res.*, **42**, W240–W245.
3. Hrabe,T., Li,Z., Sedova,M., Rotkiewicz,P., Jaroszewski,L. and Godzik,A. (2016) PDBFlex: exploring flexibility in protein structures. *Nucleic Acids Res.*, **44**, D423–D428.
4. Prlic,A., Bliven,S., Rose,P.W., Bluhm,W.F., Bizon,C., Godzik,A. and Bourne,P.E. (2010) Pre-calculated protein structure alignments at the RCSB PDB website. *Bioinformatics*, **26**, 2983–2985.
5. Burley,S.K., Berman,H.M., Bhikadiya,C., Bi,C., Chen,L., Di Costanzo,L., Christie,C., Dalenberg,K., Duarte,J.M., Dutta,S. *et al.* (2019) RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. *Nucleic Acids Res.*, **47**, D464–D474.
6. Perutz,M.F., Rossmann,M.G., Cullis,A.F., Muirhead,H., Will,G. and North,A.C. (1960) Structure of haemoglobin: a three-dimensional Fourier synthesis at 5.5-A. resolution, obtained by X-ray analysis. *Nature*, **185**, 416–422.
7. Kolodny,R., Koehl,P. and Levitt,M. (2005) Comprehensive evaluation of protein structure alignment methods: scoring by geometric measures. *J. Mol. Biol.*, **346**, 1173–1188.
8. Shatsky,M., Nussinov,R. and Wolfson,H.J. (2002) Flexible protein alignment and hinge detection. *Proteins*, **48**, 242–256.
9. Ye,Y. and Godzik,A. (2004) FATCAT: a web server for flexible structure comparison and structure similarity searching. *Nucleic Acids Res.*, **32**, W582–W585.
10. Weiss,D.R. and Levitt,M. (2009) Can morphing methods predict intermediate structures? *J. Mol. Biol.*, **385**, 665–674.
11. Rego,N. and Koes,D. (2015) 3Dmol.js: molecular visualization with WebGL. *Bioinformatics*, **31**, 1322–1324.
12. Abe,S., Ueno,T., Reddy,P.A., Okazaki,S., Hikage,T., Suzuki,A., Yamane,T., Nakajima,H. and Watanabe,Y. (2007) Design and structure analysis of artificial metalloproteins: selective coordination of His64 to copper complexes with square-planar structure in the apo-myoglobin scaffold. *Inorg. Chem.*, **46**, 5137–5139.
13. Wang,B., Shi,Y., Tejero,J., Powell,S.M., Thomas,L.M., Gladwin,M.T., Shiva,S., Zhang,Y. and Richter-Addo,G.B. (2018) Nitrosyl myoglobins and their nitrite Precursors: Crystal structural and quantum mechanics and molecular mechanics theoretical investigations of preferred Fe -NO ligand orientations in myoglobin distal pockets. *Biochemistry*, **57**, 4788–4802.
14. Goncalves,S., Miller,S.P., Carrondo,M.A., Dean,A.M. and Matias,P.M. (2012) Induced fit and the catalytic mechanism of isocitrate dehydrogenase. *Biochemistry*, **51**, 7098–7115.
15. Wang,S., Shen,Q., Chen,G., Zheng,J., Tan,H. and Jia,Z. (2014) The phosphatase mechanism of bifunctional kinase/phosphatase AceK. *Chem. Commun. (Camb.)*, **50**, 14117–14120.
16. Soman,J., Tao,T. and Phillips,G.N. Jr (1999) Conformational variation of calcium-bound troponin C. *Proteins*, **37**, 510–511.
17. Vassylyev,D.G., Takeda,S., Wakatsuki,S., Maeda,K. and Maeda,Y. (1998) Crystal structure of troponin C in complex with troponin I fragment at 2.3-A resolution. *Proc. Natl. Acad. Sci. U.S.A.*, **95**, 4847–4852.
18. Holm,L. and Sander,C. (1995) Dali: a network tool for protein structure comparison. *Trends Biochem. Sci.*, **20**, 478–480.
19. Holm,L. (2020) Using dali for protein structure comparison. *Methods Mol. Biol.*, **2112**, 29–42.
20. Fox,N.K., Brenner,S.E. and Chandonia,J.M. (2014) SCOPe: structural classification of proteins–extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Res.*, **42**, D304–D309.