

RaptorX server: A Resource for Template-Based Protein Structure Modeling

Morten Källberg, Gohar Margaryan, Sheng Wang,
Jianzhu Ma, and Jinbo Xu

Abstract

Assigning functional properties to a newly discovered protein is a key challenge in modern biology. To this end, computational modeling of the three-dimensional atomic arrangement of the amino acid chain is often crucial in determining the role of the protein in biological processes. We present a community-wide web-based protocol, RaptorX server (<http://raptorx.uchicago.edu>), for automated protein secondary structure prediction, template-based tertiary structure modeling, and probabilistic alignment sampling.

Given a target sequence, RaptorX server is able to detect even remotely related template sequences by means of a novel nonlinear context-specific alignment potential and probabilistic consistency algorithm. Using the protocol presented here it is thus possible to obtain high-quality structural models for many target protein sequences when only distantly related protein domains have experimentally solved structures. At present, RaptorX server can perform secondary and tertiary structure prediction of a 200 amino acid target sequence in approximately 30 min.

Key words Protein structure prediction, Homology modeling, Protein threading, Secondary structure prediction, Model quality assessment

1 Introduction

The advent of high-throughput procedures capable of identifying the entities making up cellular proteomes [1, 2] is one of the milestone accomplishments of recent decades. The availability of these high-dimensional datasets does, however, present us with the challenge of efficiently determining the functional role of the expressed protein entities. The biological activity of a protein domain, such as enzymatic catalysis [3] or signaling transduction [4], is often highly related to the three-dimensional arrangement of its amino acid chain. Structural models of newly discovered proteins are thus valuable in uncovering their biological function and can serve as an important stepping stone in generating hypotheses or suggesting

experiments to further explore their nature. While the Protein Data Bank (PDB) [5] provides experimentally determined structural data for a number of protein domains, the vast majority of protein sequences available in public databases currently do not have solved structures.

To this end template-based modeling methods can generate approximate models for a large number of sequences with relative ease if a closely related template domain sequence with solved structure is available. Current methods do, however, become unreliable when there are no homologs with solved structures in the PDB or when templates under consideration are distant homologs [6]. Template-based modeling is critically dependent on the quality of the target–template alignment. To better address cases where no close template exists, we studied and implemented a number of novel modeling strategies in our new software RaptorX server [7, 8]. RaptorX server takes into consideration the number of non-redundant homologs available for the target sequence and a template structure to assess the quality of information content in sequence profiles [9]. This allows us to optimize the modeling strategy specifically to the target. Second, RaptorX server uses conditional neural fields (CNF), a variant of conditional random fields (CRF), to integrate a variety of context-specific biological signals in a nonlinear probabilistic scoring function [10]. Finally, RaptorX server has also implemented a multiple-template threading (MTT) procedure [11], enabling the use of multiple templates to model a single-target sequence. Results from CASP9 and the recently concluded CASP10 competitions clearly demonstrate the value of the abovementioned innovations. RaptorX server ranked second being only outperformed by a server employing consensus analysis of results from multiple single methods and extensive post-threading refinement [12].

Aside from structure modeling, RaptorX server provides options for custom pairwise target–template alignments and single-target multiple-template alignments. Furthermore, RaptorX server utilizes a CNF [13]-based prediction protocol for determining the three-state secondary structure, eight-state secondary structure, and solvent accessibility distributions for each residue in the target sequence. RaptorX server also provides disorder prediction of an input protein sequence.

The secondary and tertiary structure models generated by RaptorX server can serve as starting points for further analysis in a number of diverse application areas. For example, the predicted 3D models can be used for binding site epitope prediction as well as in protein docking and protein–protein interaction studies [14, 15].

2 Materials

The following are necessary for the use of RaptorX server.

1. A personal computer connected to the Internet and a web browser with Java Script enabled: RaptorX server is compatible with three popular web browsers: Google Chrome, Firefox, and Internet Explorer. Nevertheless, the former two browsers may be slightly better than the third one in visualizing the prediction results.
2. The amino acid sequence(s) of the protein(s) of interest in FASTA format: The allowed characters in the sequence are the one-letter codes for the 20 standard amino acids. Spaces and line breaks in the sequence string are ignored and do not affect the prediction. To prevent a single sequence from occupying the server for a very long time, we currently limit the length of user-submitted sequences to 2,000 amino acids.

3 Methods

In this section we present two separate use cases of the RaptorX server. First, we cover the main use case of obtaining the secondary and tertiary structures of a target sequence. Second, we demonstrate the use of RaptorX server to generate alignments between the target sequence and user-specified template structures.

3.1 Modeling the Secondary and Tertiary Structures of a Target Sequence

1. In the web browser navigate to <http://raptorx.uchicago.edu>.
2. From the menu at the top of the page select “New job.”
3. Use the tab menu to choose between “Alignment Job” and “Structure Prediction Job.”
4. In the “Job Identification” section of the form provide a job name (defaults to “my job”) and an e-mail address that will be used for notification upon job completion. The e-mail given also serves as the username for accessing results at a later date. Since RaptorX server does not require any user registration, it is important that a correct e-mail address is provided.
5. In the “Sequences” box, provide one or more FASTA-formatted sequences. These can be supplied by copy and pasting into the text box or by uploading a flat-text file with the data. The FASTA identifier is used to identify the individual sequence(s) when browsing prediction results; therefore, we recommend using descriptive sequence names. In the “Job Settings” section, choose if multiple-template modeling is to be used (recommended) and if you wish to do secondary, tertiary, or both secondary and tertiary structure modeling.

6. Press the submit button to queue the prediction job. The data entered in the form will be validated, and the user will be notified of any errors that need correction in a box appearing at the top of the page. If the submission is successful the user will be redirected to an overview page displaying pending and completed prediction jobs. It should be noted that the number of pending jobs allowed for one user is limited to 20.
7. In order to track pending and completed prediction jobs the user needs to be logged in to the server. If the login from a previous session has expired or the account needs to be accessed from a different machine than the one used for the initial submission, the user can supply the account e-mail in the login field on the RaptorX server front page. An e-mail message will be sent to the address containing a hyperlink to the overview page.
8. Select “My jobs” in the menu at the top of the page to display the job overview for the account. Here, the status of each prediction in the job is given along with overall information on the predictions being done for each submitted sequence. To track the job status in real time simply refresh the page, and the completion status of the prediction for each submitted sequence in a job will be updated (*see Note 1*).
9. Click on the structure labeling link in the job overview page to bring up a summary page similar to the one depicted in Fig. 1.
10. Structure labeling prediction is provided in four modes. The available results include three-class and eight-class secondary structures, disorder prediction, as well as three-state solvent accessibility. You can switch between the modes using the blue tab menu (*see Label 1* in Fig. 1). The three-class secondary structure prediction gives the distribution between the classes alpha-helix, extended strand in beta ladder, and loop/irregular. In addition to these, the eight-class prediction classes include residue in isolated beta-bridge, 3-helix (3/10 helix), 5-helix (π -helix), hydrogen-bonded turn (3, 4, or 5 turn), and Bend (*see Note 2*). Disorder prediction classifies residues as disorder or non-disorder, while solvent accessibility classes are buried, medium, and exposed.
11. For each residue a figure depicting the distribution of structure labeling classes is given, indicating the relative likelihood of a given residue belonging to each of these classes. The legend for the color coding of the states can be found in the column on the right-hand side of the page (*see Label 5* in Fig. 1). Hover over a residue to display the exact probability distribution of secondary structure classes in a pop-up box next to the residue (*see Label 2* in Fig. 1).
12. The right-hand column provides information on the status of the prediction job (*see Label 3* in Fig. 1); to download the



Fig. 1 Example of a secondary structure prediction result

prediction results for the sequence, including the full class distributions of the four secondary structure predictions, click the link labeled “Download” (see Label 4 in Fig. 1).

13. Click on a 3D structure link in the job overview page to obtain a job summary similar to the one depicted in Fig. 2a, b (see Note 3).
14. In a structure prediction job, a protein structure is built for each of the (≤ 10) top-ranked alignments between the target and sequences from the template library. The rank of the candidate model is provided in the results overview (see Label 1 in Fig. 2a), with the highest ranked model being selected as the default. Clicking the “View alternative models” button will bring up a menu from which the user can switch between models (see Label 5 in Fig. 2a). For each model, the PDB code of the template along with the p -value and uGDT score of the alignment is given. If MTT is used, a model based on several templates will be available as well (see Label 4 in Fig. 2a) (see Note 4).
15. The quality of the model is given by p -value, uGDT, and global distance test (GDT) (see Labels 2 and 3 in Fig. 2a) of its


a

[-] 3D model(s) for domain 1 [80, 394]

View Alternative Model

Alignment Rank: 1 **1** P-value: 2.423e-15 **2** uGDT(GDT): 302.25(95.95) **3**

Templates: 2iaeC 2ie3C 3c5wC 1u32A 1s70A 1it6A 1wao1 2o8aA 1tcoA 1s95A **4**



Rotation
☒ spin structure
 Coloring and Representation
☐ coloring amino acids
☒ coloring helices & sheets
☐ show side chain
 Zoom
 zoom in **7**
 zoom out

6 Jmol

103 113 123 G

```

L S E S Q V K S L C E K A K E I L T K E S N V Q E V R C - - - P V T V C G D V
L S E S Q V K S L C E K A K E I L T K E S N V Q E V R C - - - P V T V C G D V
L S E S Q V K S L C E K A K E I L T K E S N V Q E V R C - - - P V T V C G D V
L S E S Q V K S L C E K A K E I L T K E S N V Q E V R C - - - P V T V C G D V
L Q E N E I R G L C L S R E I F L S Q P I L L E L E A - - - P L K I C G D I
M T E A E V R G L C I K S R E I F L S Q P I L L E L E A - - - P L K I C G D I
  
```

8

9 Status

Current status: Complete
 Submitted on: 2013-02-25 19:39:02
 Finished on: 2013-02-25 20:02:21

10 Download prediction data

[Download](#) domain structures for available models.
[Download](#) confidence scores for available models.
[Download](#) alignments/scores for available models.

11 Jmol viewer quick guide

- Left-click+drag to rotate the structure.
- Use the middle-scroller to zoom.
- Right-click the structure for more options.
- Hover over a target residue in the sequence alignment to see it highlighted in the structure.
- Visit the [Jmol mouse manual wiki](#)

Sequence box help

- The topmost row is the sequence you entered aligned to


b

[-] 3D model(s) for domain 1 [80, 394]

View Alternative Model

Alignment Rank: 5 P-value: 1.354e-13 uGDT(GDT): 267.25(84.84)

Template: 1s70A



Rotation
☒ spin structure
 Coloring and Representation
☐ coloring amino acids
☒ coloring helices & sheets
☐ show side chain
 Zoom
 zoom in
 zoom out

Jmol

102 112 122 132

```

Q L S E S Q V K S L C E K A K E I L T K E S N V Q E V R C F V T V C G D V H G Q
Q M T E A E V R G L C I K S R E I F L S Q P I L L E L E A P L K I C G D I H G Q
3 2 1 1 1 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0
  
```

142 152 162 172

```

F H D L M E L F R I G G K S P D T N Y L F M G D Y V D R G Y Y S V E T V T L L V
  
```

1

Status

Current status: Complete
 Submitted on: 2013-02-25 19:39:02
 Finished on: 2013-02-25 20:02:21

Download prediction data

[Download](#) domain structures for available models.
[Download](#) confidence scores for available models.
[Download](#) alignments/scores for available models.

Jmol viewer quick guide

- Left-click+drag to rotate the structure.
- Use the middle-scroller to zoom.
- Right-click the structure for more options.
- Hover over a target residue in the sequence alignment to see it highlighted in the structure.
- Visit the [Jmol mouse manual wiki](#)

Sequence box help

- The topmost row is the

Fig. 2 (a) Example of a tertiary structure prediction result with multiple templates. **(b)** Example of a tertiary structure prediction result with a single template and local quality score

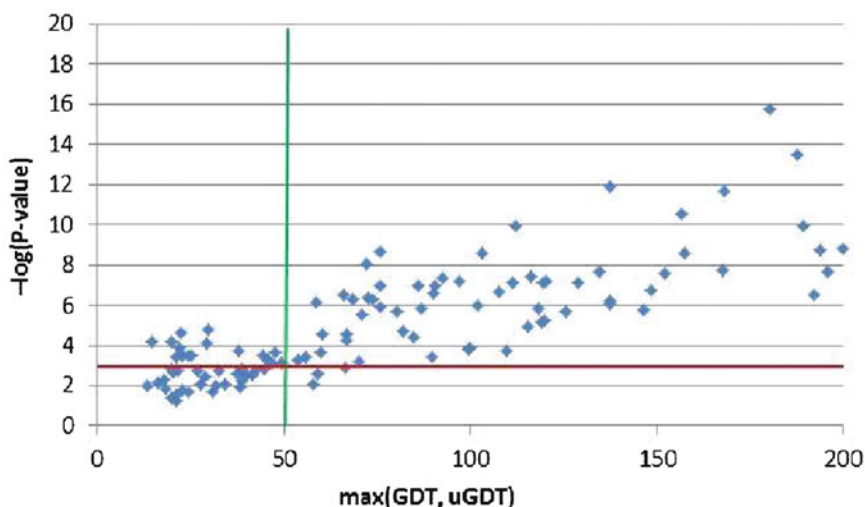


Fig. 3 The relationship between p -value and the model quality on the 123 CASP10 targets

alignment with the selected template. The uGDT is the unnormalized GDT score which is defined as $1 \times N(1) + 0.75 \times N(2) + 0.5 \times N(4) + 0.25 \times N(8)$, where $N(x)$ is the number of residues with the local RMSD smaller than x Å. GDT is uGDT normalized by a protein domain length. GDT measures the quality of a model by comparing it with the native structure and has a value ranging from 0 to 100, indicating the worst and the best quality, respectively. As shown in Fig. 3, the p -value is a reliable indicator of model quality. When the p -value is small (i.e., $<10^{-5}$), the models have a uGDT or a GDT greater than or equal to 50. Even in the case of a p -value smaller than 10^{-4} , only three models have both uGDT and GDT less than 50. That is, the prediction from our threading method is reliable when the p -value is less than 10^{-4} . For each model, the PDB identifier for the template structure and the specific polypeptide chain from the PDB file used to build the currently selected model are displayed. Click the link to go to the structure record in the PDB (<http://www.pdb.org>) (see Label 4 in Fig. 2a).

16. A graphic representation of the currently selected model is provided in the Jmol viewer. Use the mouse to rotate and zoom the structure. Right-clicking on the model will bring up a menu of further options for changing the visualization settings (see Label 6 in Fig. 2a). To the right of the structure viewer a menu for controlling the representation of the currently selected model is available (see Label 7 in Fig. 2a).
17. The alignment of the target and template sequences used for constructing the current model is displayed below the Jmol viewer. Each position in the alignment is color coded according

to the chemical nature of the residue. The color scheme used is the following: Red=Hydrophobic, Blue=Acidic, Magenta=Basic, and Green=Hydroxyl+Amine. RaptorX server also provides the predicted RMSD at each aligned position rounded to the nearest integer as indicators of reliability in the last row (*see Label 1* in Fig. 2b). Hover over the aligned residues to highlight the corresponding target residues in the Jmol viewer (*see Label 8* in Fig. 2a).

18. The column to the right provides an overview of the prediction job status (*see Label 9* in Fig. 2a). Click on the appropriate links to download the prediction results. Multiple download options are available: PDB files for the top-ranked models, the corresponding alignments with their local reliability scores, and the confidence scores such as the *p*-value, uGDT, and GDT mentioned above (*see Label 10* in Fig. 2a). Underneath the download links a third box with a brief user's guide for the Jmol viewer is given (*see Label 11* in Fig. 2a) followed by a guide on the sequence box.

3.2 Custom Template Alignment

19. Repeat **steps 1–5** from Subheading 3.1 above.
20. Indicate the structure(s) you wish the supplied sequence(s) from **step 5** to be aligned to. Enter the PDB ID in the text box, and select the desired structure from the drop-down menu that appears. Repeat to add additional structures to the list (*see Note 5*).
21. Under “Alignment options,” check the types of alignments you wish to generate. The options available are “Optimal pairwise alignment” which returns the best possible pairwise alignment between the target sequence and the selected templates; “Probabilistic sampling” which returns a user-specified number of alternative alignments sampled according to the alignment probability distribution generated by the CNF model; or “Multiple template alignment” which returns a multiple-protein alignment between the selected templates and the input target sequence.
22. Click on an alignment job in the job overview to obtain a summary similar to the one depicted in Fig. 4.
23. In an alignment job, in addition to the optimal alignments between the target sequence and the provided template structures, a set of sampled alternative alignments may also be generated. To generate a sample alignment, check the “Probabilistic sample” box and indicate the number of samples desired.
24. Click on the alignment drop-down selection box to bring up a selection menu from which it is possible to switch between alternative alignments (*see Label 1* in Fig. 4). The alignment of the target and template sequences will be displayed after a selection is made, and the “Display” button is pressed.

an alpha-helix with probability 17 % in the three-state model and 14 % in the eight-state model. As the two models give the distribution of secondary structure groups from two different class sets, the differences in the alpha-helix propensity between the two models could be due to other types of helices being possible in the eight-state model.

3. It should be noted that the prediction results are not expanded automatically when a results page is loaded. This is done to provide a better overview for the submitted sequence consisting of many domains. For any one submission there will be at least four entries in the result page including secondary and tertiary structure prediction, domain parsing, and disorder prediction. Clicking on any of them will display the relevant result.
4. Even if MTT is selected you may not see any MTT results in the drop-down menu. MTT is only deployed if our method predicts that a model based on several template structures is more accurate than the top-ranked single-template model. Should you still want to construct a multiple-template alignment, this can be accomplished through the custom alignment interface.
5. When looking up a template structure in the drop-down menu you may not always be able to find the desired PDB identifier. This is due to the template library used on the server being “non-redundant”; thus, several highly similar structures in the PDB are omitted, and only one representative structure is kept in the library. To resolve this problem, we supply a list of equivalent structures to identify the structure in the library equivalent to your desired template.

Acknowledgments

This work is supported by the National Institute of Health grant R01GM0897532, National Science Foundation DBI-0960390 and CAREER award, Alfred P. Sloan Fellowship, and TTIC summer intern program. The authors are grateful to the University of Chicago Beagle team, TeraGrid, and Canadian SHARCNet for their support of computational resources.

References

1. Aebersold R, Mann M (2003) Mass spectrometry-based proteomics. *Nature* 422(6928):198–207
2. Källberg M, Lu H (2010) An improved machine learning protocol for the identification of correct Sequest search results. *BMC Bioinformatics* 11:591
3. Bairoch A (2000) The ENZYME database in 2000. *Nucleic Acids Res* 28(1):304–305
4. Hannum G et al (2009) Genome-wide association data reveal a global map of genetic interactions among protein complexes. *PLoS Genet* 5(12):e1000782
5. Berman HM et al (2000) The protein data bank. *Nucleic Acids Res* 28(1):235–242
6. Baker D, Sali A (2001) Protein structure prediction and structural genomics. *Science* 294(5540):93–96

7. Peng J, Xu J (2011) RaptorX: exploiting structure information for protein alignment by statistical inference. *Proteins* 79(Suppl 10):161–171
8. Kallberg M et al (2012) Template-based protein structure modeling using the RaptorX web server. *Nat Protoc* 7(8):1511–1522
9. Peng J, Xu J (2010) Low-homology protein threading. *Bioinformatics* 26(12):i294–i300
10. Peng J, Xu J (2009) Boosting protein threading accuracy. *Lect Notes Comput Sci* 5541:31
11. Peng J, Xu J (2011) A multiple-template approach to protein threading. *Proteins* 79(6):1930–1939
12. Roy A, Kucukural A, Zhang Y (2010) I-TASSER: a unified platform for automated protein structure and function prediction. *Nat Protoc* 5(4):725–738
13. Peng J, Bo L, Xu J (2009) Conditional neural fields. In: Bengio Y, Schuurmans D, Lafferty J, Williams CKI, Culotta A (eds) *Advances in neural information processing systems*, vol 22. p 1419–1427
14. Singh R et al (2010) Struct2Net: a web service to predict protein–protein interactions using a structure-based approach. *Nucleic Acids Res* 38:W508–W515
15. Singh R, Xu J, Berger B (2006) Struct2net: integrating structure into protein–protein interaction prediction. *Pac Symp Biocomput* 2006:403–414

Protein Structure Prediction

Kihara, D. (Ed.)

2014, XI, 253 p. 64 illus., 48 illus. in color., Hardcover

ISBN: 978-1-4939-0365-8

A product of Humana Press