Template-based protein structure modeling using the RaptorX web server

Morten Källberg[1–3], Haipeng Wang[1,3], Sheng Wang[1], Jian Peng[1], Zhiyong Wang[1], Hui Lu[2] & Jinbo Xu[1]

[1]Toyota Technological Institute at Chicago, Chicago, Illinois, USA. [2]Department of Bioengineering, University of Illinois at Chicago, Chicago, Illinois, USA. [3]These authors contributed equally to this work. Correspondence should be addressed to J.X. (jinboxu@gmail.com).

Published online 19 July 2012; doi:10.1038/nprot.2012.085

A key challenge of modern biology is to uncover the functional role of the protein entities that compose cellular proteomes. To this end, the availability of reliable three-dimensional atomic models of proteins is often crucial. This protocol presents a community-wide web-based method using RaptorX (http://raptorx.uchicago.edu/) for protein secondary structure prediction, template-based tertiary structure modeling, alignment quality assessment and sophisticated probabilistic alignment sampling. RaptorX distinguishes itself from other servers by the quality of the alignment between a target sequence and one or multiple distantly related template proteins (especially those with sparse sequence profiles) and by a novel nonlinear scoring function and a probabilistic-consistency algorithm. Consequently, RaptorX delivers high-quality structural models for many targets with only remote templates. At present, it takes RaptorX ~35 min to finish processing a sequence of 200 amino acids. Since its official release in August 2011, RaptorX has processed ~6,000 sequences submitted by ~1,600 users from around the world.

INTRODUCTION

Proteomes constitute the backbone of cellular function by carrying out the tasks encoded in the genes expressed by a given cell type. Recent decades have seen rapid growth in high-throughput procedures capable of identifying the proteomic profile of a cell in any state[1,2]. It does, however, remain challenging to efficiently classify the operational role of the individual protein entities identified in such procedures. Functional properties of a protein domain, such as enzymatic activity[3] or the ability to interact with other proteins[4], can often be derived from the approximate spatial arrangement of its amino acid chain in the folded state. Knowledge of the structure of a newly discovered protein is thus highly valuable in determining the role it plays in biological processes, and it can serve as an important stepping stone in generating hypotheses or suggesting experiments to further explore the protein's nature. Although the Protein Data Bank (PDB)[5] provides experimentally solved structural data for an increasing number of protein domains, solving protein structures remains costly, time consuming and, in certain instances, technically difficult. Consequently, the vast majority of protein sequences available in public databases do not have a solved structure at this point in time. More than ~10 million unique protein sequences have been deposited, whereas only ~70,000 have had their structures solved. To bridge this gap, a wide array of computational protocols for protein secondary and tertiary structure prediction from its amino acid sequence are continuously being developed.

Computational structure prediction methods can, in principle, be divided into two categories, template-based and template-free modeling, with some composite protocols combining aspects of both. Methods in the former group include comparative modeling methods[6], which, given a target sequence, identify evolutionarily related templates with solved structure by sequence or sequence-profile comparison (e.g., BLAST and HHpred[7]) and construct structure models based on the scaffold provided by these templates. Alternative methods build on the observation that known protein structures appear to comprise a limited set of stable folds. It is thus often found that evolutionarily distant or unrelated protein sequences share common structural elements, which is used by threading methods[8,9] such as MUSTER[10], SPARKS[11,12] and RAPTOR[13–15]. It has been demonstrated that, in some cases, incorporating structural information to match the query sequence to potential templates enables similarity in fold to be detected despite the lack of an explicit evolutionary relationship.

Template-based modeling (TBM) can generate useful approximate models for a large number of sequences with relative ease if close templates are available. Current methods do, however, become unreliable when there are no homologs with solved structures in PDB or when templates under consideration are distant homologs[16]. Template-free methods offer an alternative for modeling such difficult cases. Pure *ab initio* methods[17–19] aim at building a 3D model without using structure homologous information; the successful application of such methods is, however, limited to short target sequences (< 120 residues) at present. In addition, a number of semi-*ab initio* approaches exist that assemble short structural fragments or use statistical information to spatially restrain the building of a model structure. Finally, so-called composite methods, which combine subsets of the previously mentioned approaches, have been very successful in recent Critical Assessment of Protein Structure Prediction (CASP) competitions, most notably the TASSER methodology developed by Zhang[20].

Although all of the aforementioned methods have made key contributions to the field of structure prediction, it remains challenging to accurately predict the structure of a target sequence with a sparse sequence profile and no close homologs in the PDB. It has been estimated that 76% of the 4.2 million models deposited in MODBASE[21], a database repository for theoretical structure models, are built from remote homologs. Thus, any improvement in structure prediction methods addressing these cases will have a substantial effect on the utility of such theoretical models, as well as on our ability to assign functional properties on the basis of common fold patterns.

The RaptorX server

TBM crucially depends on the quality of the target-template alignment. Our previous program RAPTOR has been successful

in efficiently optimizing the general protein-threading scoring function, and it has been among the best structure prediction protocols available, as demonstrated at previous CASP evaluations[13]. RAPTOR and other state-of-the-art threading programs are, however, limited by a linear scoring function, which cannot accurately represent any correlation that may exist among the features used for assessing alignment quality (for instance, secondary structure and sequence profile are known to be correlated). Further, the application of structural information in the alignment process does not take into consideration the level of similarity between target and template. The use of structural information when modeling a target with a high-similarity template might introduce noise, whereas structural information becomes relatively more important when modeling a challenging target with sparse sequence profile.

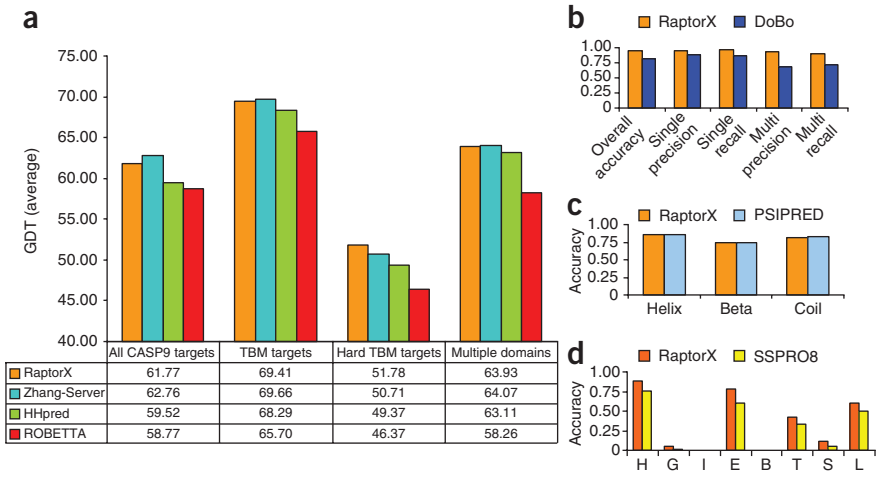
To better address cases in which no close template exists, we have studied and implemented a number of novel modeling strategies in our new software RaptorX[22], taking a completely different approach than that used in RAPTOR. First, a profile-entropy scoring method, taking into consideration the number of nonredundant homologs available for the target sequence and template structure, is used to assess the quality of information content in sequence profiles[23], thereby allowing us to optimize the modeling strategy specifically to the target. Second, we use conditional random fields (CRFs) to integrate a variety of biological signals in a nonlinear threading score function not previously used by any threading software[24]. Finally, we have implemented a multiple-template threading (MTT) procedure[25], enabling the use of multiple templates to model a single target sequence. Unlike other MTT methods, which mainly increase the alignment coverage, our MTT method can partially correct errors in pairwise alignments by exploiting intertemplate similarity and thus can improve the final model quality.

Results from the recently concluded CASP9 competition clearly indicated the value of the above-mentioned innovations. RaptorX was ranked second overall, slightly outperformed only by Zhang's servers[26], which combined results from ~10 individual homology modeling/threading programs and further conducted extensive postprocessing refinement of results from the individual methods. In addition, RaptorX generated the best alignments for the 50 most difficult TBM CASP9 targets[27], outperforming all other servers.

Aside from structure modeling, RaptorX can be used to obtain custom pairwise target-template alignments and to generate an arbitrary number ($<1{,}000$) of alternative pairwise alignments through probabilistic sampling, as well as to generate single-target to multiple-template alignments. Further, RaptorX also provides a conditional neural field (CNF)–based[28] prediction protocol for determining the three-state or eight-state secondary structure distribution for each residue in a target protein.

To supplement structure prediction, RaptorX also provides domain parsing of long protein sequences and disorder prediction to help users interpret secondary and tertiary structure prediction results. To help users gauge whether prediction results obtained from RaptorX will fit the purpose of their work, we have included an overview of the modeling accuracy one can expect from the individual modules in the RaptorX server in **Figure 1**. For each module, the performance of RaptorX is compared with that of competing methods. In the case of 3D structure prediction, the comparison is with I-TASSER (i.e., Zhang-Server)[20], Robetta[16] and HHpred[7] on ~110 CASP9 target sequences, with performance measured by the averaged global distance test (GDT) score in four target categories. We use the CASP9 targets to measure the performance of RaptorX because this makes it easy to compare RaptorX with other top servers blindly. For domain parsing, we compare RaptorX with DoBo[29] on both single-domain and multidomain CASP9 targets. Finally, the performance of secondary structure prediction is assessed by comparing the prediction accuracy of RaptorX in the helix, β-sheet and coil environments with that of PSIPRED[30]. We also compare the eight-state secondary structure prediction accuracy of RaptorX with that of SSPro8 (ref. 31), which to the best of our knowledge is the only publicly available server providing eight-state secondary structure prediction.

A comparison of the services offered by the RaptorX server with those available from servers based on alternative structure prediction protocols is given in **Table 1**. Servers are compared with respect to the following features: Is the prediction result from a single tool or consensus results from a collection of protocols (meta-server)? Does the server do domain parsing for a large target sequence?



	All CASP9 targets	TBM targets	Hard TBM targets	Multiple domains
RaptorX	61.77	69.41	51.78	63.93
Zhang-Server	62.76	69.66	50.71	64.07
HHpred	59.52	68.29	49.37	63.11
ROBETTA	58.77	65.70	46.37	58.26

Figure 1 | Performance assessment of core prediction modules in the RaptorX server. (**a**) Comparison of structure prediction performance by global distance test (GDT) score for RaptorX and three other publicly available protocols on the CASP9 targets. Performance is compared in four categories: All CASP9 targets, template-based modeling (TBM) targets, hard TBM targets and multidomain targets. (**b**) Performance comparison for domain parsing between RaptorX and DoBo. Metrics are given for overall performance, and performance on single-domain and multidomain CASP9 target proteins. Specifically, accuracy is the overall proportion of both single-domain and multidomain proteins identified correctly; single (multi) recall is the fraction of single-domain (multidomain) proteins that are predicted; single (multi) precision is the fraction of correctly predicted single-domain (multidomain) proteins among all the predictions. A multidomain protein is correctly predicted only if its domain boundaries are correctly identified. (**c**) Performance comparison between RaptorX and PSIPRED for secondary structure prediction. The accuracies achieved for three-state prediction (helix, sheet and coil) are compared. (**d**) Performance comparison between RaptorX and SSPRO8 for eight-state secondary structure prediction. The accuracies achieved for eight-state prediction for the classes H, G, I, E, B, T, S, L (using SSPRO8 nomenclature) are compared.

TABLE 1 | Comparison of RaptorX with several publicly available protein structure prediction servers.

Name	URL	Prediction options	M/S	DP	TM/FM	FA
RaptorX	http://raptorx.uchicago.edu/	Secondary, tertiary, alignment sampling, multiple-template threading	S	Yes	TM	Yes
I-TASSER[26]	http://zhanglab.ccmb.med.umich.edu/I-TASSER/	Tertiary	M	Yes	TM, FM	Yes
Phyre[50]	http://www.sbg.bio.ic.ac.uk/phyre2/	Secondary, tertiary	M	No	TM	Yes
HHpred[51]	http://toolkit.tuebingen.mpg.de/hhpred	Secondary, tertiary	S	No	TM	No
Robetta[52]	http://robetta.bakerlab.org/	Tertiary	S	Yes	TM, FM	No
GenThreader[30]	http://bioinf.cs.ucl.ac.uk/web_servers/	Secondary, tertiary, others	S	Yes	TM	No

DP, domain parsing; FA, functional annotation; TM, template-based modeling; FM, template-free modeling; M/S, meta-server or single server.

Is modeling based on homology detection or *ab initio*? And does the server provide biological function annotation? Two additional key features distinguished RaptorX from other structure prediction servers, namely the ability to do MTT (improving the overall alignment by using information from multiple templates) and the ability to do alignment sampling. The utility of these two features is discussed in detail below.

To the best of our knowledge, the RaptorX server is not biased toward any specific types of proteins. However, it does have some limitations mainly because of the insufficient coverage of several sequence and structure databases. The secondary structure prediction accuracy on average is slightly decreased if the query sequence does not have a sufficient number of sequence homologs in the NR (nonredundant) database. The domain prediction is limited by the coverage of the Pfam database, which currently covers ~75% of all the protein sequences[32]. The tertiary structure prediction is limited by the coverage of the template database. RaptorX cannot produce reliable models for a query sequence if it does not have even a template of low similarity in the PDB.

Figure 2 outlines the three modeling tasks users can accomplish using the RaptorX server, namely tertiary structure prediction, secondary structure prediction and custom alignment. Each task is decomposed into a number of timed conceptual stages, with the logical flow from one stage to the next indicated by the connecting arrows. In the following, we describe the basic concept of computation done in each of these stages while referring the reader to previous publications for more detailed accounts. As indicated in **Figure 2**, structure modeling is the last stage in the structure prediction workflow before the final result is returned to the user. Although the focus of this work is on the necessary steps before the construction of the 3D model of a target sequence, we recognize that this final step in itself constitutes an important and complex computational task. The RaptorX server deploys the software package MODELLER[33] to construct structural models from an

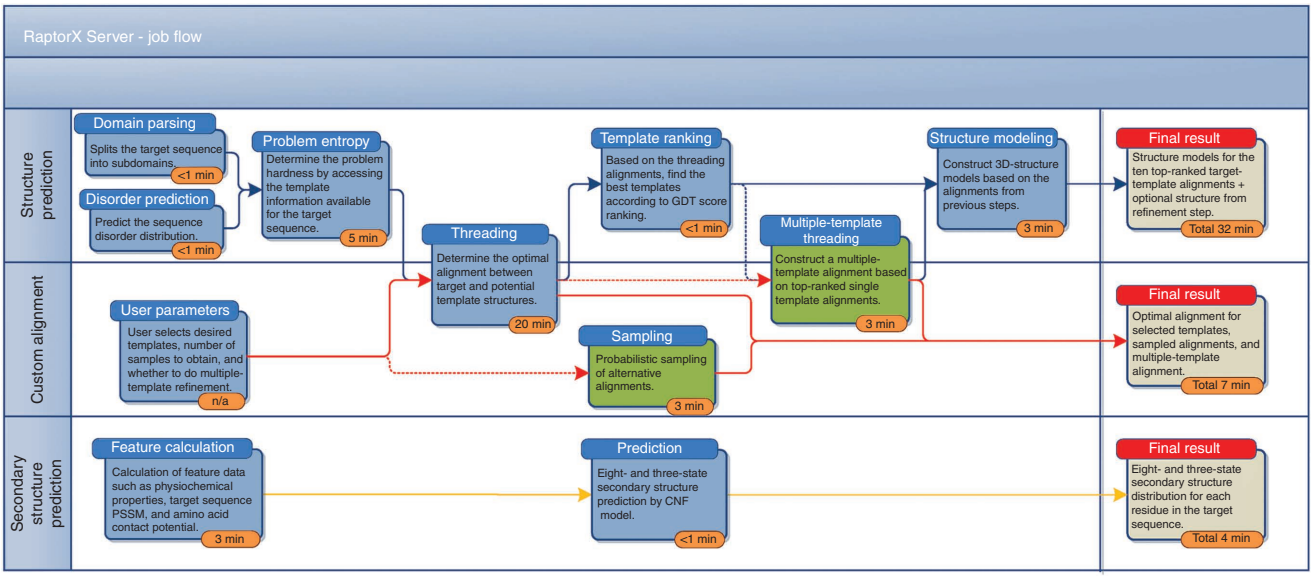


Figure 2 | Workflow used by the RaptorX server. Outline of the three modeling tasks users can accomplish using the RaptorX server, namely tertiary structure prediction, custom alignment and secondary structure prediction. For each stage, details of the computation and approximate completion time for a 250-residue target sequence are given (for threading; the indicated time is for a full template library scan). Blue boxes indicate mandatory stages, green boxes indicate optional stages and gray boxes indicate the resulting output. The blue, red and yellow directed paths indicate the flow for structure prediction, custom alignment jobs and secondary structure prediction, respectively. Dashed and solid paths indicate that the subsequent steps are optional and required, respectively. n/a, not applicable.

alignment between the best set of templates and the target sequence using the same procedure described in reference 23.

Applications of RaptorX

The secondary and tertiary structure models generated by RaptorX 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[34] and epitope prediction[35]. Another application is found in determining the binding topology of small ligand molecules to putative binding sites on the domain structure generated. Such molecular docking studies can be carried out using software packages such as AutoDock[36], and often have an important guiding role in rational drug design pipelines. A related application is so-called macromolecular docking, in which the quaternary structure formed by two or more single protein domains is determined using software packages such as DOCK[37]. The latter of these two applications is of particular interest in so-called protein-protein interaction studies[38,39].

In addition to studying the biophysics of potential molecular interaction, the protein structure model generated by RaptorX can also serve as input for more specialized function prediction protocols. For instance, a wide range of servers based on machine learning models tuned to identify key functional residues are available. One example is the recently published NAPS (a residue-level nucleic acid-binding prediction server), which, given a protein structure, can determine which residues may be DNA or RNA binding[40].

Further, RaptorX can be used for improving a multiple-sequence alignment of sequences without structure by using tools such as T-Coffee (specifically, M-Coffee)[41,42]. Consider the following scenario: we wish to construct a multiple-sequence alignment of sequences A, B and C (none of which have a solved structure). For each sequence A, B and C, RaptorX can be used to identify related template sequences. Suppose that some good templates (sequences with structure) are identified by RaptorX for A and B. Then the alignment of A and B to their respective top templates can be used by T-Coffee to construct a better multiple-sequence alignment for A, B and C. The better multiple-sequence alignments are achieved when the structure information from the top templates discovered by RaptorX is taken into consideration, as T-Coffee can often generate better alignments with structure data available to guide the process.

Experimental design

Nonlinear alignment scoring function. RaptorX uses a profile entropy–dependent scoring function for protein threading. The detection of good templates for a target protein with a sparse sequence profile, by the use of sequence profile information in the form of a hidden Markov model or a position-specific scoring matrix (PSSM), is often inadequate. To address this concern, our scoring method takes into consideration the sequence profile sparsity (i.e., the number of nonredundant homologs available for the sequence and template), as well as the complex correlation among various protein features. Given this information, we can weigh the relative importance placed on sequence and structure features in the threading step. For instance, a target sequence that only has a few sequence homologs will have a sequence profile with a low entropy score (i.e., sparse sequence profile). In this case, RaptorX will place more weight on structure information, whereas a target with a high entropy score will rely more heavily on sequence profile information in scoring alignments.

The protein threading step is done by constructing a CRF model for finding an optimal alignment. In this formulation, biological properties calculated for the input sequence s and template sequence t serve as so-called observations for predicting the state (match or gap) of each position in the resulting optimal alignment a. A CRF representation is particularly well suited for modeling this problem, as it can efficiently deal with a set of highly correlated input features for determining the optimal sequence of alignment states by using nonlinear scoring functions. This property ultimately stems from the fact that CRFs seek to optimize the conditional probability $P(a|s, t)$ (i.e., how likely is an alignment given the input) rather than the joint probability $P(a, s, t)$ that is sought to be optimized in generative models.

The nonlinearity in our scoring function is achieved by using a collection of regression trees to determine the log likelihood of each alignment state in the CRF model. Rather than explicitly trying to express all possible correlations among basic features (which would likely lead to a prohibitively large number of complex features), the regression tree is used for learning only the most important subset of correlations. Each regression tree consists of a set of mutually exclusive paths, each of which can be represented as a conjunction of rules on the input features. The criterion represented by a given path can be as simple as a cutoff on a single feature, such as '(mutation score < -50), then the log-likelihood of a match state is ln(0.9)', or a complex conjunction such as '($-50 <$ mutation score < -10) and (secondary structure score > 0.9) and (solvent accessibility score > 0.6), then the log-likelihood of a match state is ln(0.7)'.

By expressing the likelihood of different states at a given alignment position using regression trees, we can apply varying standards when aligning different regions of the target and template, in much the same way a PSSM provides different mutation potentials for the 20 amino acids at each sequence position. However, in contrast to PSSMs, regression trees can incorporate any type of protein feature, not just those based on sequence statistics. More details on the exact formulation of the described threading strategy can be found in reference 24.

Assessment of alignment quality. RaptorX predicts the quality of an alignment by using a neural network that estimates the similarity, measured by TMscore[43] (normalized by the target length), between the target and template and then by ranking all candidate templates according to the predicted quality. To this end, the following features are used: Sequence profile similarity, primary sequence similarity, statistical potential–based sequence similarity, secondary structure similarity, solvent accessibility similarity, contact capacity similarity and environmental fitness and the number of gap openings and gap positions.

Multiple-template threading. Given the steady increase in solved protein structures, it is probable that more than one good template for a given target is available, or that a set of templates provides better coverage of the target than is possible using just one template. On the basis of the optimal pairwise sequence-template alignments generated from a complete screening of a template library or from a custom alignment job, RaptorX offers the option to align a single target sequence to any number of its top templates by the use of MTT[25]. Although the increase in target coverage can improve structure modeling results in some instances, the key aspect of MTT is the ability to improve individual pairwise sequence-template alignments

by exploiting inter-template similarity. Such improvements are generally not possible using existing multiple-template methods that simply assemble pairwise alignments into a single multiple-protein alignment (using the target protein as a pivot), resulting in errors from the pairwise alignments persisting in the single-target to multiple-template alignment.

The ability to make this improvement is, in short, because of the use of a probabilistic-consistency transformation, the key idea of which is to generate a set of pairwise alignments that is as consistent as possible with each other. First, all possible pairwise alignments between target and template pairs are expressed as a probabilistic alignment matrix, with each possible alignment being associated with a probability. A binary alignment matrix, which can be thought of as a special probabilistic alignment matrix, is also generated between any two templates using structure alignment tools. The entries in all the matrices are then iteratively adjusted to achieve the maximum consistency among all the matrices simultaneously, thereby improving individual alignments by taking into account information from multiple target-template pairs. On the basis of this set of consistent probabilistic alignment matrices, a superior single-target to multiple-template alignment can be constructed. Such a multiple protein alignment not only has a better target coverage but also better alignment accuracy. More technical details of the described strategy are accounted for in reference 25.

Probabilistic sampling of alignments. In addition to inspecting the optimal alignment, especially in cases in which only remote templates are available, it can often be informative for users to obtain a number of alternative alignments. Alignment sampling allows the user to see how different subsequences of a target align with the biologically important areas of a template structure. Further, it gives the option of building a set of alternative structure models for the same target, and bases the decision of which is more suitable for a specific application on structure data rather than alignment data.

The probabilistic nature of our CRF threading method allows for sampling of any number of alternative alignments, as it defines the probability distribution over all possible alignments conditioned on the target and template sequence. The decision of which alignment to use for model building can then ultimately be guided by the user's choice of model quality assessment method (which has the option to incorporate much more information than the threading model's scoring function) or the user's own domain knowledge. To sample the alignment space, we use a forward-backward algorithm[44]. In the 'forward step,' a revised form of the Smith-Waterman algorithm is used to compute a $m \times n \times 3$ dynamic-programming table, G, with m and n being the length of the target and template protein, respectively, and three the number of alignment states. In this table, $G(i, j, h)$ denotes the probability sum of all the alignments with the constraint that sequence position i is aligned to template position j with state h. Once G is calculated, we can sample alternative alignments from C to N terminus in the 'backward step.'

Function annotation of structure models. Similarity in the fold of two proteins may indicate the existence of an evolutionary relationship, which in turn may imply a shared functional role. The Structural Classification of Proteins (SCOP) database provides a description of the structural and evolutionary relation of most proteins in the PDB[45,46]. Whenever a structure model is constructed, RaptorX provides a distribution statistic of the 'class', 'fold', 'super-family', 'family' and 'protein type' from some or all of ten top-ranked templates as identified in the SCOP database version 1.75, with each template contribution weighted by its predicted alignment quality (normalized among the ten structures). Only the templates with a predicted alignment quality of at least 85% of the highest predicted quality are used, as in most cases the predicted alignment quality error is less than 15%. The SCOP distribution of high-ranked templates, in addition to the 3D model of the target sequence, will give the user an initial feel for the nature of the protein being modeled and thus provide a starting point for further exploration of the structure in question.

Secondary structure prediction. The secondary structure prediction module is based on a CNF model[28] developed by Wang *et al*[47]. CNFs possess properties found in both neural networks and CRFs, obtaining nonlinear modeling capabilities in joining information from diverse protein features for a single residue from the former, and the ability to model the interdependence in secondary structure for adjacent residues from the latter. Further, CNF provides a probability distribution over the secondary structure classes, rather than simply returning a single class prediction. Returning a distribution makes it is possible for the user to take the uncertainty of class assignment into consideration when interpreting results, a feat not possible with a discrete class prediction model. Models for three- and eight-class prediction (see PROCEDURE) are available, both of which are learned from training data sets in which a residue with known secondary structure class is represented as a combination of position-specific and position-independent features.

Domain parsing. For each submitted sequence, RaptorX will first examine whether the target sequence consists of multiple domains by searching it against the Pfam database[48]. If at least one significant Pfam entry is identified (E value < 0.001), RaptorX will cut the sequence into domains and conduct tertiary structure prediction and functional annotation for each domain separately. This is done because domains in a multidomain protein are likely to have different functions; therefore, it is better to conduct function annotation for each domain independently. In addition, if the target has fewer than 500 amino acids, RaptorX will predict the 3D model for the entire sequence even if it was found to be a multidomain protein. On the other hand, if the target has more than 500 residues, no 3D model for the whole sequence is generated, as it is unlikely that a good template for the full sequence exists in the PDB.

Note that domain parsing only affects 3D structure prediction and functional annotation. Both secondary structure prediction and disorder prediction are directly applied to the whole target sequence.

Disorder prediction. For each submitted sequence, RaptorX conducts disorder prediction by running DISOPRED[49] and visualizes the prediction result using a method similar to that deployed in the secondary structure prediction module. In certain instances, inspecting the disorder prediction result can help users better evaluate the reliability of the tertiary structure prediction. If, for instance, a large segment of the sequence is predicted to be disordered with a high confidence score, the 3D structure prediction for this segment is very likely unreliable, which may affect the accuracy of other regions in the structure model. To obtain more reliable results, users are suggested to remove large disordered regions from the sequence and resubmit the remaining sequence segments to RaptorX.

MATERIALS

EQUIPMENT

Computer

- A personal computer connected to the Internet and a web browser with JavaScript enabled. RaptorX is compatible with three popular web browsers: Google Chrome, Firefox and Microsoft Internet Explorer

Data

- The amino acid sequence(s) of the protein(s) of interest should be 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 will be ignored and will not affect the prediction. To prevent a single sequence from occupying the server for a very long time, RaptorX takes a protein sequence with at most 2,000 amino acids

PROCEDURE

Submitting a job ● TIMING 10 min

1| Go to the RaptorX homepage at http://raptorx.uchicago.edu/.

2| Select 'New job' from the menu at the top of the page.

3| Use the tab menu to select between submitting an 'Alignment Job' and a 'Structure Prediction Job.'

4| In the 'Job Identification' section of the form, supply a job name (default is 'My job') and an e-mail address to be used for notification when the job has been completed. The e-mail provided here will also serve as the username by which the job account is identified on the server for accessing results at a later date. An error message will appear if no e-mail address is provided.

▲ **CRITICAL STEP** As RaptorX does not require a user to register before submitting a job, it is important to provide a correct e-mail address. Otherwise, you will not be able to retrieve the results of your job.

5| In the 'Sequences' section of the form, provide one or more sequences in FASTA format. The sequence(s) can either be supplied by copying and pasting into the text box or by uploading a flat text file containing the data.

▲ **CRITICAL STEP** For a given prediction or alignment job, the FASTA identifier is used to identify the individual sequence(s) when browsing through the job results; it is therefore important to provide a descriptive sequence name. Although the length of the sequence name is not limited, it is better not to use a very long sequence name.

6| This step differs depending on whether an alignment job (option A) or a structure prediction job (option B) is being submitted:

(A) Alignment job

(i) Indicate the structure(s) you wish the supplied sequence(s) from Step 5 to be aligned with. 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.

(ii) Under 'Alignment options', check the types of alignment you wish to generate. The options given are as follows: '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 CRF model; or 'Multiple template alignment', which returns a multiple protein alignment between the selected templates and the input target sequence.

? TROUBLESHOOTING

(B) Structure prediction job

(i) Specify the parameters in 'Job Settings.' Specifically, choose whether multiple-template modeling is to be used, and whether secondary, tertiary or both secondary and tertiary structure modeling is to be done.

(ii) Specify the prediction type in the drop-down menu (select between performing 'Structure prediction' and 'Secondary structure prediction,' or both) and whether to use MTT when multiple good templates are available for the target.

7| Press the 'Submit' button to queue the job on the server. Successful submission will redirect the user to a page of pending and finished jobs for the account used.

▲ **CRITICAL STEP** Upon submission, the data entered in the form will be validated and the user will be notified of any errors that need to be corrected in a box appearing at the top of the page. Please note that there is a limit to the number of pending jobs allowed for one user (as identified by their username and IP address used at submission) in order to maintain sufficient server capacity to serve all users. Specifically, each user can have no more than 20 sequences pending prediction at any point in time, and a single job can contain at most ten sequences. Further, the results of a job are only stored for 14 d after the job is completed.

Job monitoring and job availability
● TIMING 25–60 min

8| To track pending and finished 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 that on which it was initially created, the user will need to go to the server front page and supply the account e-mail in the login field on the right. This will generate an e-mail message to the address given with a hyperlink to the page containing the jobs for the account associated with that e-mail.
? TROUBLESHOOTING

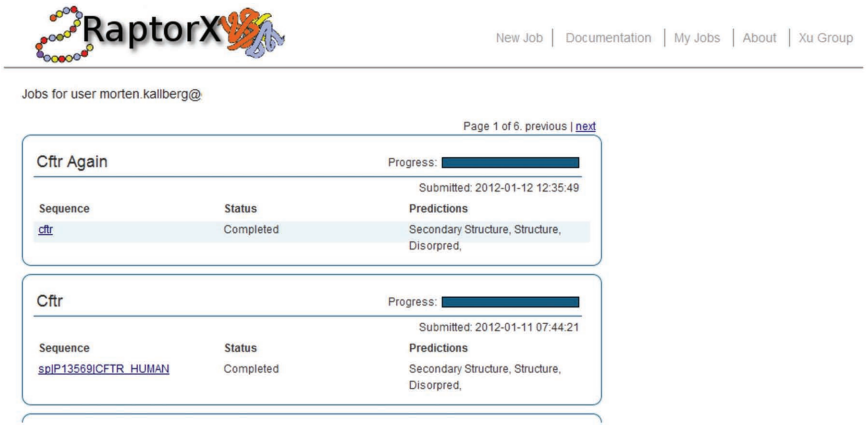


Figure 3 | Job-listing interface. Selecting 'My jobs' displays this job overview for the user's account, which gives the status of each prediction in the job along with overall information of the predictions being done for each sequence submitted.

9| Once you have logged in to the server, selecting 'My jobs' in the menu at the top of the page displays a job overview for the account (**Fig. 3**). Here the status of each prediction in the job is given along with overall information about the predictions being done for each sequence submitted. To track the job status in real time, simply refresh the page and the completion status of the prediction submitted for each sequence in a job will be updated. Clicking on a sequence name will take the user to the result page for this sequence.
? TROUBLESHOOTING

Viewing secondary structure predictions ● TIMING 5 min

10| Click on a secondary structure job in the overview to display a summary page similar to the one depicted in **Figure 4**.

11| Secondary structure prediction is provided in two modes, using both three-state and eight-state models. You can switch between the two modes using the blue tab menu (see label 1 in **Fig. 4**). The three-state model gives the distribution between the classes 'α-helix', 'extended strand in β-ladder' and 'loop/irregular'. In addition to these the eight-state model prediction classes are 'residue in isolated β-bridge', '3-helix (3/10 helix)', '5-helix (π-helix)', 'hydrogen bonded turn (3, 4 or 5 turn)', and 'bend'.
? TROUBLESHOOTING

12| For each residue, a figure depicting the distribution of secondary structure 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. 4**). 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. 4**).

13| The right-hand column provides information on the status of the prediction job (see label 3 in **Fig. 4**); to download the prediction results for the sequence, including the full class distribution for both models and the most likely secondary class sequence from the three-state model in PSIPRED-like format[30], click the link labeled 'Download' (see label 4 in **Fig. 4**).

Viewing tertiary structure and functional predictions ● TIMING 10 min

14| Click on a structure job in the job overview to obtain a job summary similar to the one depicted in **Figure 5**.

15| In a structure prediction job, a protein structure is built for each of the ten top-ranked alignments between the target sequence and the structures in the template library. The interface provides the rank of the currently selected alignment result (see label 1 in **Fig. 5**), with the highest-ranked model being selected as default (on the basis of the best template). Click the 'Selected alternative models' button to bring up a selection menu from which it is possible to switch between alternative models (see label 4 in **Fig. 5**). For each model, the PDB code of the template used and the estimated GDT score of the alignment is provided. If MTT is used, a model with the multiple templates will be available as well (see label 3 in **Fig. 5**).
? TROUBLESHOOTING

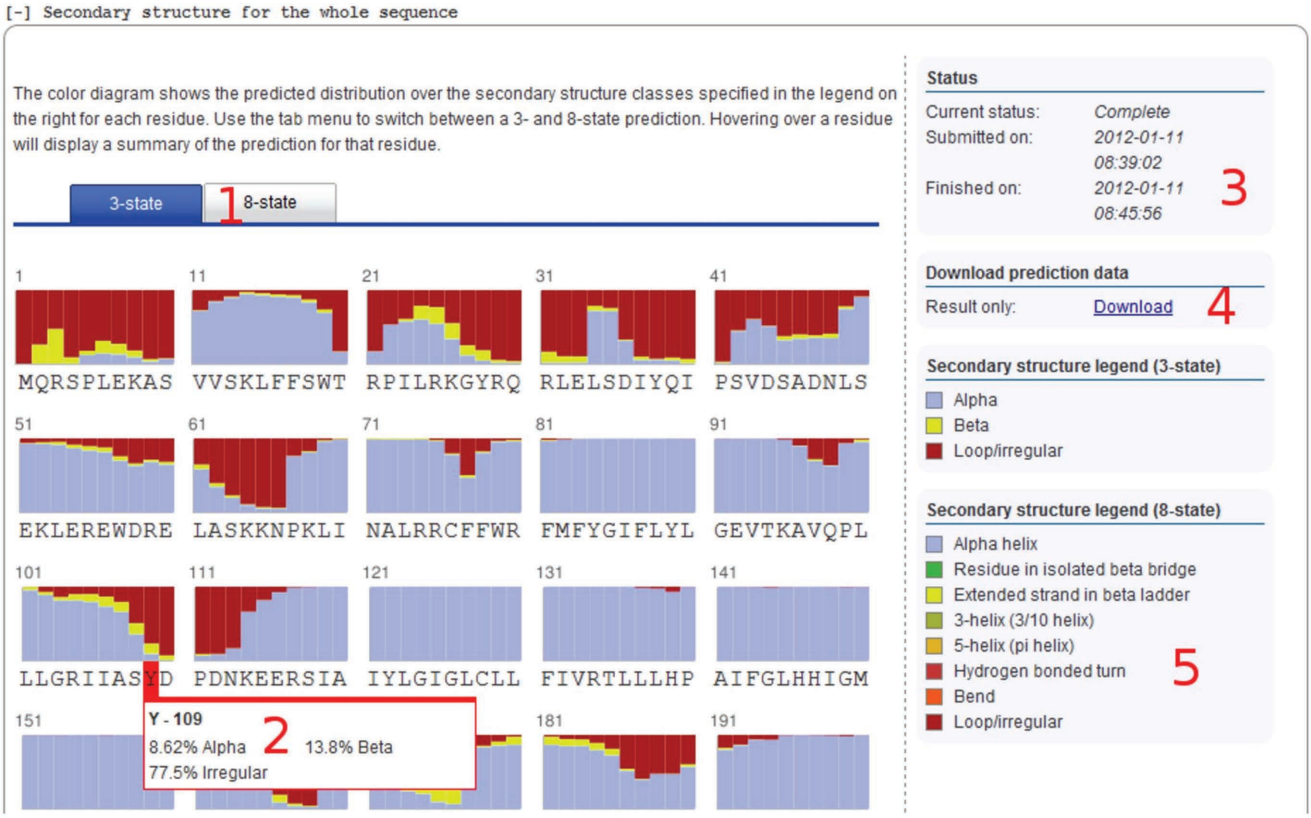


Figure 4 | Secondary structure result interface. The numbered labels indicate the location of the following screen features: (1) tabs for switching between the three-state and eight-state prediction; (2) hovering over a residue will give detailed statistics on the secondary-state distribution; (3) the status: a current running time of the job; (4) a download link for the prediction results; and (5) a color-code legend for secondary structure diagram.

16| Judge the quality of a selected structure model from the reported alignment score. The score falls between 0 and 100, with 100 indicating a perfect model (see label 2 in **Fig. 5**). As rule of thumb, a model scoring <50 can be considered highly likely to show the correct fold of the target sequence. For each model, the PDB identifier for the template structure used for the currently selected model and the specific polypeptide chain from the PDB file used for the model is displayed. Click the link to go to structures record at the PDB (http://www.pdb.org/; see label 3 in **Fig. 5**). Further, the complete SCOP (http://scop.berkeley.edu/) classification of the template for the currently selected model is given if available. Clicking the link will take you to the relevant record in the SCOP database (see label 5 in **Fig. 5**).

17| A Jmol structure viewer providing a visualization of the currently selected model is loaded underneath. Use the mouse to rotate and zoom on the structure. Right-clicking the model will bring up a menu of further options for changing the visualization (see label 6 in **Fig. 5**). To the right of the structure viewer, a menu for controlling the representation of the currently selected model is available. Here the user can zoom on the structure, switch between coloring modes and select a wire-frame display of the structure (see label 7 in **Fig. 5**).

18| The alignment of the target and template sequence 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 scheme used is as follows: red = hydrophobic, blue = acidic, magenta = basic, green = hydroxyl + amine. An asterisk ('*') under aligned residues signifies matching residues, whereas a colon (':') signifies that the aligned residues are in the same functional group. Hover over aligned residues to highlight the target residue in the Jmol viewer (see label 8 in **Fig. 5**).

19| The right-hand column provides information on the status of the prediction job (see label 9 in **Fig. 5**). Click on the links to download the prediction results, including the PDB files for the ten top-ranked with corresponding alignments, the set of alignments between the target sequence and all structures in the template library used, a list containing the complete ranking of all alignments in acceding order according to GDT score and a BLAST search result of the target sequence against the nonredundant PDB database (see label 10 in **Fig. 5**). Below the box with download links, a brief user guide for the Jmol viewer is given (see label 11 in **Fig. 5**).

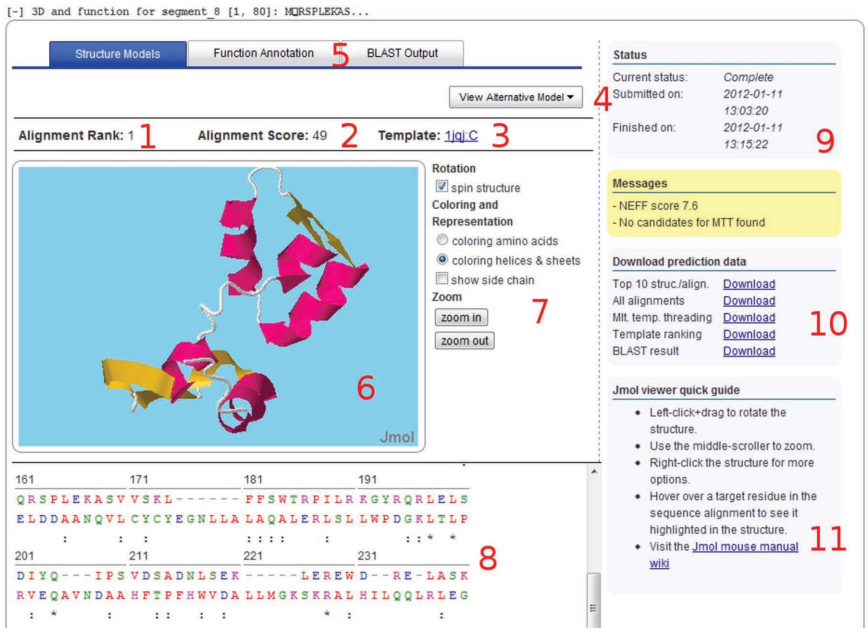


Figure 5 | Tertiary structure result interface. The numbered labels indicate the location of the following screen features: (1) the rank of currently selected model; (2) the quality score of the model; (3) the PDB IDs for the set templates used for modeling; (4) a drop-down menu for selecting alternative structure models; (5) tabs for switching between structure prediction, function annotation and BLAST output; (6) interactive viewer displaying the currently selected model structure; (7) menu for controlling the interactive viewer; (8) alignment used for structure modeling; (9) indication of the status: a current running time of the job; (10) download links for prediction results; and (11) a user guide for the interactive structure viewer.

Disorder prediction ● **TIMING 2 min**

20| If a structure prediction job has been submitted (Step 6B), a disorder prediction for the entire target sequence is also done. Graphics comparable to those described for secondary structure prediction are used to visualize the probability that a given residue is either in a disorder segment (marked in red) or nondisorder segment (marked in blue). Hover over the residue to display the exact probabilities (**Fig. 6**).

Domain parsing ● **TIMING 2 min**

21| If a structure prediction job has been submitted (Step 6B), RaptorX first uses a domain parsing procedure to explore whether the target sequence appears to consist of multiple domains or constitutes a single folding unit. If multiple domains are found, the domain parsing results will be available in table format outlining the span of each segment, the Pfam family it is predicted to belong to and a confidence measure (*E* value) for the domain assignment. View the table by clicking the '+' under 'Domain parsing' to view the table (**Fig. 6**).

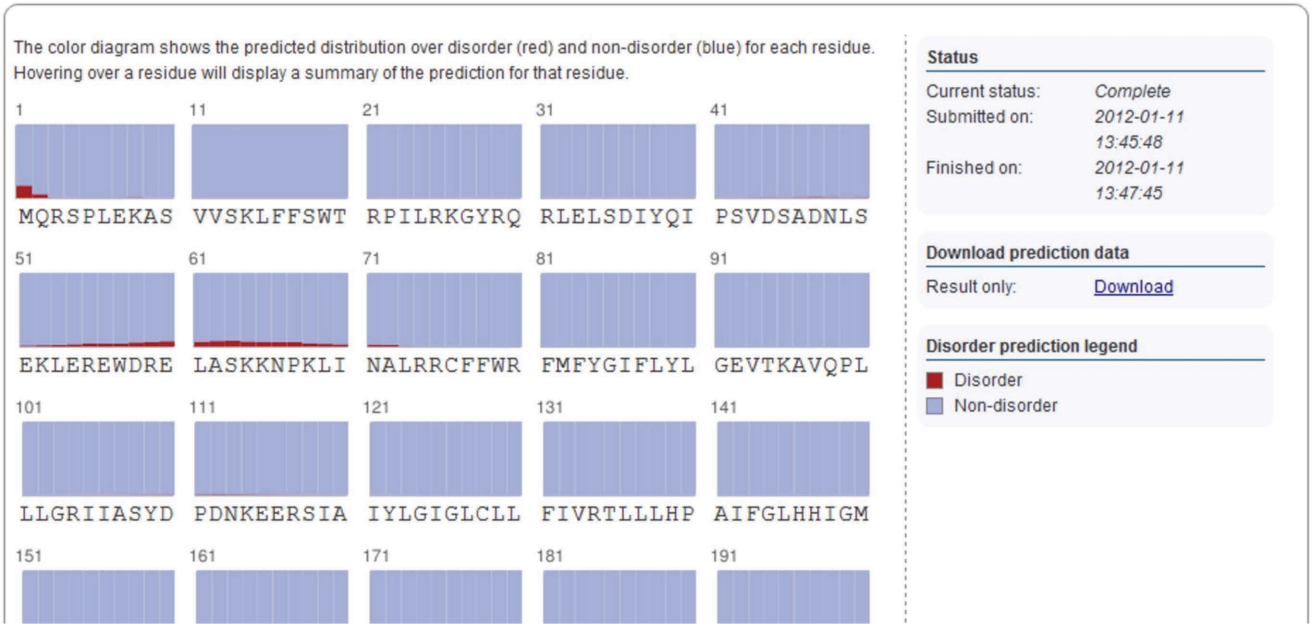


Figure 6 | Disorder prediction result display. Graphics comparable to those described for the secondary structure result interface (**Fig. 4**) are used to visualize the probability that a given residue is either in a disorder segment (marked in red) or nondisorder segment (marked in blue). Again, hovering over a residue will give detailed statistics on the disorder prediction, whereas the right-hand side shows the status of the job with a download link for the disorder prediction results and a color-code legend for the disorder prediction diagram.

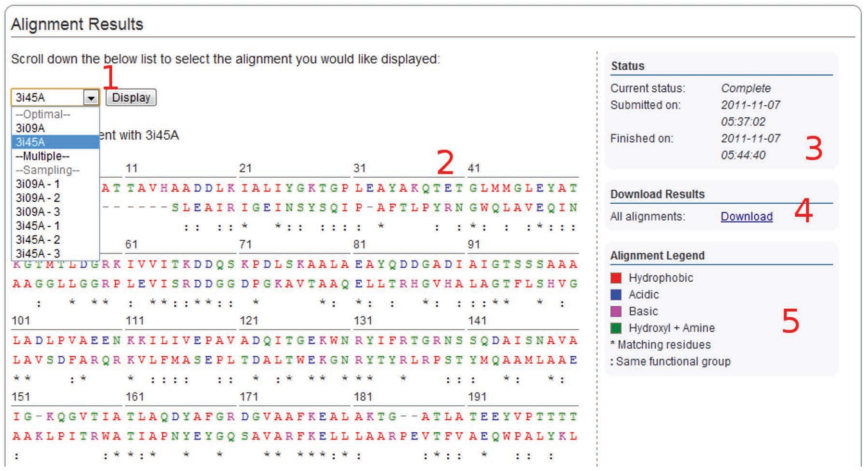


Figure 7 | Custom alignment result interface. The numbered labels indicate the location of the following screen features: (1) a drop-down menu for switching between alternative alignments; (2) the alignment between target sequence and template; (3) indication of the status: a current running time of the job; (4) a link for download of the prediction result; and (5) a legend indicating the alignment color coding.

Viewing custom alignment results

● **TIMING 5 min**

22| Click on an alignment job in the job overview to obtain a summary similar to the one depicted in **Figure 7**.

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. The alignment of the target and template sequences will be displayed after a selection is made and the 'Display' button is pressed. Each position in the alignment is color-coded according to the chemical nature of the residue. The scheme used is as follows: red = hydrophobic, blue = acidic, magenta = basic, green = hydroxyl + amine. An asterisk ('*') under aligned residues signifies matching residues, whereas a colon (':') signifies that the aligned residues are in the same functional group.

25| The right-hand column provides information on the status of the job. Click on the links to download the alignment results, including the set of alignments between the target sequence and all structures in the template library used.

? TROUBLESHOOTING

Troubleshooting advice can be found in **Table 2**.

TABLE 2 | Troubleshooting table.

Step	Problem	Possible reason	Solution
6A(ii)	I wish to create a custom alignment to structure template XXXX, but I cannot find it in when searching the drop-down menu	The template library used on the server is 'nonredundant,' meaning that several highly similar structures in the PDB are omitted and only one representative structure included in our library	Use the supplied list of equivalent structures to identify the structure in the library equivalent to your desired template
8	I submitted a few sequences to RaptorX a couple of days ago, but have never received any response from the server	Usually RaptorX can process at least one of your submitted sequences within 24 h even if it is overloaded. If this problem happens, RaptorX may be down for maintenance or you may have provided an incorrect email address	Click on the 'contact' menu at the bottom of the RaptorX web page and send a message to the system administrator
9	I do not see any results displayed in the result page	To improve the appearance of the results page, the prediction results are not expanded automatically for submitted sequences consisting of many domains	There should be at least four result entries in the results page, including secondary and tertiary structure prediction, domain parsing and disorder prediction. Clicking on any of them will display the relevant result

(continued)

TABLE 2 | Troubleshooting table (continued).

Step	Problem	Possible reason	Solution
11	The probability of observing the same secondary structure class for a given residue differs between the three- and eight-state models. For instance, residue 8 is in an α-helix with probability of 17% and 14% in the two models, respectively	The two models providing the distribution over the secondary structure classes are optimized using data sets with different possible states for each residue. Consequently, differences in the α-helix propensity between the two models may for instance be due to other types of helices being possible in the eight-state model	None; this is a potential consequence of the model
15	I chose to do MTT for my structure job, but do not see any MTT results in the drop-down menu	The construction of a better structure model from MTT is only done if our method predicts that a model better than the first-ranked single-template model can be obtained by joining information from multiple templates	If you still wish to construct a multiple-template alignment from a set of desired template structures, this can be accomplished through the custom alignment interface

● TIMING

Steps 1–7, submitting a job: 10 min
Steps 8 and 9, job monitoring and job availability: 25–60 min
Steps 10–13, viewing secondary structure predictions: 5 min
Steps 14–19, viewing tertiary structure and functional predictions: 10 min
Step 20, disorder prediction: 2 min
Step 21, domain parsing: 2 min
Steps 22–25, viewing custom alignment results: 5 min

Prediction of 3D structure, secondary structure and functional annotation of a small protein sequence (~300–400 residues) takes approximately 30–35 min; processing a medium-sized domain (~350–400 residues) will take 40–45 min, whereas for large domains (~800 residues) running times approaching 65 min should be expected (for a further breakdown of the time needed to complete different job types, see **Fig. 2**). The actual time between submission of a prediction job and the availability of the final result on the server does, however, also depend on the number of jobs currently queued on the server. RaptorX uses a fair-share job schedule policy to prevent users from holding up the whole server by submitting too many sequences in a short time. That is, whenever RaptorX finishes one sequence, RaptorX will proceed to the next user and conduct predictions for one of this user's sequences. Currently, the RaptorX server is deployed on a 24-CPU machine with 94 GB of available RAM. By using this framework, an average of 120 structure and secondary structure prediction jobs are completed in a 24-h period.

ANTICIPATED RESULTS

Once a job is completed, the user is notified by an e-mail message containing a link to the result page. For each sequence, the structure prediction result page contains the following: predicted secondary structure, disorder prediction, domain parsing, (if the submitted sequence contains multiple domains) up to 11 template-based 3D models and a simple functional annotation for each putative domain, and (if the submitted sequence is a single-domain protein or if it contains less than 500 amino acids) up to 11 template-based 3D models and a simple functional annotation for the whole sequence. **Figure 2** indicates the expected output for each of the three core modules. **Figures 3–8** show some example outputs.

[-] Domain parsing results for the whole sequence

Segment	Span	Source	Reference Parent	Parent Span	E-value	Annotations
segment_1	1251-1374	Pfam	PF00005	1-117	6.90e-17	ABC_tran: ABC transpo
segment_2	465-576	Pfam	PF00005	1-117	1.20e-16	ABC_tran: ABC transpo
segment_3	81-350	Pfam	PF00664	1-274	1.90e-15	ABC_membrane: ABC tra
segment_4	862-1143	Pfam	PF00664	1-271	3.10e-14	ABC_membrane: ABC tra
segment_5	547-618	Pfam	PF02463	136-209	8.80e-11	SMC_N: RecF/RecN/SMC
segment_6	1228-1416	Pfam	PF02463	16-210	2.10e-10	SMC_N: RecF/RecN/SMC
segment_7	446-621	Pfam	PF09818	240-412	1.30e-07	ABC_ATPase: Predicted
segment_8	1-80	Pfam	N/A			
segment_9	351-445	Pfam	N/A			
segment_10	622-861	Pfam	N/A			
segment_11	1144-1227	Pfam	N/A			
segment_12	1417-1480	Pfam	N/A			

Figure 8 | Domain parsing result display. If multiple domains are found, the domain parsing results outline the span of each segment, the Pfam family it is predicted to belong to, a confidence measure (*E* value) for the domain assignment and a possible functional annotation of domain region.

ACKNOWLEDGMENTS This work is supported by the US National Institutes of Health grants R01GM0897532, a US National Science Foundation grant DBI-0960390, a Microsoft PhD Research Fellowship, an FMC Educational Fund Fellowship and the Toyota Technical Institute at Chicago summer intern program. We are grateful to the University of Chicago Beagle team, TeraGrid and Canada's Shared Hierarchical Academic Research Computing Network (SHARCNet) for their support of computational resources.

AUTHOR CONTRIBUTIONS J.X. conceived and supervised the project. M.K. and H.W. designed and developed the web server. H.L. oversaw server development. J.P. developed the threading algorithm. S.W. designed the template database. Z.W. developed the protein secondary structure prediction algorithm. M.K. and J.X. wrote the paper.

COMPETING FINANCIAL INTERESTS The authors declare no competing financial interests.

Published online at http://www.nature.com/doifinder/10.1038/nprot.2012.085. Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Aebersold, R. & Mann, M. Mass spectrometry-based proteomics. *Nature* **422**, 198–207 (2003).
2. Källberg, M. & Lu, H. An improved machine learning protocol for the identification of correct Sequest search results. *BMC Bioinformatics* **11**, 591 (2010).
3. Bairoch, A. The ENZYME database in 2000. *Nucleic Acids Res* **28**, 304–305 (2000).
4. Hannum, G. *et al.* Genome-wide association data reveal a global map of genetic interactions among protein complexes. *PLoS Genet* **5**, e1000782 (2009).
5. Berman, H.M. *et al.* The Protein Data Bank. *Nucleic Acids Res* **28**, 235–242 (2000).
6. Martí-Renom, M.A. *et al.* Comparative protein structure modeling of genes and genomes. *Annu. Rev. Biophys. Biomol. Struct.* **29**, 291–325 (2000).
7. Soding, J. Protein homology detection by HMM-HMM comparison. *Bioinformatics* **21**, 951–960 (2005).
8. Bowie, J.U., Lüthy, R. & Eisenberg, D. A method to identify protein sequences that fold into a known three-dimensional structure. *Science* **253**, 164–170 (1991).
9. Jones, D.T., Taylor, W.R. & Thornton, J.M. A new approach to protein fold recognition. *Nature* **358**, 86–89 (1992).
10. Wu, S. & Zhang, Y. MUSTER: Improving protein sequence profile-profile alignments by using multiple sources of structure information. *Proteins* **72**, 547–556 (2008).
11. Zhang, C., Liu, S., Zhou, H. & Zhou, Y. An accurate, residue-level, pair potential of mean force for folding and binding based on the distance-scaled, ideal-gas reference state. *Protein Sci.* **13**, 400–411 (2004).
12. Zhang, W., Liu, S. & Zhou, Y. SP5: improving protein fold recognition by using torsion angle profiles and profile-based gap penalty model. *PLoS ONE* **3**, e2325 (2008).
13. Xu, J. & Li, M. Assessment of RAPTOR's linear programming approach in CAFASP3. *Proteins* **53**, 579–584 (2003).
14. Xu, J., Li, M., Kim, D. & Xu, Y. RAPTOR: optimal protein threading by linear programming. *J. Bioinform. Comput. Biol.* **1**, 95–117 (2003).
15. Xu, J., Li, M., Lin, G., Kim, D. & Xu, Y. Protein threading by linear programming. *Pac. Symp. Biocomput.* 264–275 (2003).
16. Baker, D. & Sali, A. Protein structure prediction and structural genomics. *Science* **294**, 93–96 (2001).
17. Liwo, A., Lee, J., Ripoll, D.R., Pillardy, J. & Scheraga, H.A. Protein structure prediction by global optimization of a potential energy function. *Proc. Natl. Acad. Sci. USA* **96**, 5482–5485 (1999).
18. Simons, K.T., Kooperberg, C., Huang, E. & Baker, D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J. Mol. Biol.* **268**, 209–225 (1997).
19. Wu, S., Skolnick, J. & Zhang, Y. *Ab initio* modeling of small proteins by iterative TASSER simulations. *BMC Biol.* **5**, 17 (2007).
20. Zhang, Y. I-TASSER: fully automated protein structure prediction in CASP8. *Proteins* **77**, 100–113 (2009).
21. Pieper, U. *et al.* MODBASE, a database of annotated comparative protein structure models and associated resources. *Nucleic Acids Res* **37**, D347–D354 (2009).
22. Peng, J. & Xu, J. RaptorX: exploiting structure information for protein alignment by statistical inference. *Proteins* **79**, 161–171 (2011).
23. Peng, J. & Xu, J. Low-homology protein threading. *Bioinformatics* **26**, i294–i300 (2010).
24. Peng, J. & Xu, J. Boosting Protein Threading Accuracy. *Lect. Notes Comput. Sci.* **5541**, 31–45 (2009).
25. Peng, J. & Xu, J. A multiple-template approach to protein threading. *Proteins* **79**, 1930–1939 (2011).
26. Roy, A., Kucukural, A. & Zhang, Y. I-TASSER: a unified platform for automated protein structure and function prediction. *Nat. Protoc.* **5**, 725–738 (2010).
27. Mariani, V., Kiefer, F., Schmidt, T., Haas, J. & Schwede, T. Assessment of template based protein structure predictions in CASP9. *Proteins* **79**, 37–58 (2011).
28. Peng, J., Bo, L. & Xu, J. Conditional neural fields. In *Advances in Neural Information Processing Systems 22* (eds. Bengio Y., Schuurmans D., Lafferty J., Williams C.K.I. and Culotta A.) 1419–1427 (Neural Information Processing Systems Foundation, 2009).
29. Eickholt, J., Deng, X. & Cheng, J. DoBo: Protein domain boundary prediction by integrating evolutionary signals and machine learning. *BMC Bioinformatics* **12**, 43 (2011).
30. Buchan, D.W. *et al.* Protein annotation and modelling servers at University College London. *Nucleic Acids Res* **38**, W563–W568 (2010).
31. Pollastri, G., Przybylski, D., Rost, B. & Baldi, P. Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. *Proteins* **47**, 228–235 (2002).
32. Punta, M. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **40**, D290–D301 (2012).
33. Fiser, A. & Sali, A. Modeller: generation and refinement of homology-based protein structure models. *Methods Enzymol.* **374**, 461–491 (2003).
34. Zhao, H., Yang, Y. & Zhou, Y. Highly accurate and high-resolution function prediction of RNA binding proteins by fold recognition and binding affinity prediction. *RNA Biol.* **8**, 988–996 (2011).
35. Kulkarni-Kale, U., Bhosle, S. & Kolaskar, A.S. CEP: a conformational epitope prediction server. *Nucleic Acids Res.* **33**, W168–W171 (2005).
36. Morris, G.M. *et al.* AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J. Comput. Chem.* **30**, 2785–2791 (2009).
37. Lorber, D.M. & Shoichet, B.K. Hierarchical docking of databases of multiple ligand conformations. *Curr. Top Med. Chem.* **5**, 739–749 (2005).
38. Singh, R., Park, D., Xu, J., Hosur, R. & Berger, B. Struct2Net: a web service to predict protein-protein interactions using a structure-based approach. *Nucleic Acids Res.* **38**, W508–W515 (2010).
39. Singh, R., Xu, J. & Berger, B. Struct2net: integrating structure into protein-protein interaction prediction. *Pac. Symp. Biocomput.* 403–414 (2006).
40. Carson, M.B., Langlois, R. & Lu, H. NAPS: a residue-level nucleic acid-binding prediction server. *Nucleic Acids Res.* **38**, W431–W435 (2010).
41. Wallace, I.M., O'Sullivan, O., Higgins, D.G. & Notredame, C. M-Coffee: combining multiple sequence alignment methods with T-Coffee. *Nucleic Acids Res.* **34**, 1692–1699 (2006).
42. Notredame, C., Higgins, D.G. & Heringa, J. T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* **302**, 205–217 (2000).
43. Zhang, Y. & Skolnick, J. Scoring function for automated assessment of protein structure template quality. *Proteins* **57**, 702–710 (2004).
44. Charniak, E. *Statistical Language Learning* (MIT Press, 1993).
45. Murzin, A.G., Brenner, S.E., Hubbard, T. & Chothia, C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536–540 (1995).
46. Andreeva, A. *et al.* Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Res.* **36**, D419–D425 (2008).
47. Wang, Z., Zhao, F., Peng, J. & Xu, J. Protein 8-class secondary structure prediction using conditional neural fields. *Proteomics* **11**, 3786–3792 (2011).
48. Finn, R.D. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **38**, D211–D222 (2010).
49. Ward, J.J., McGuffin, L.J., Bryson, K., Buxton, B.F. & Jones, D.T. The DISOPRED server for the prediction of protein disorder. *Bioinformatics* **20**, 2138–2139 (2004).
50. Kelley, L.A. & Sternberg, M.J.E. Protein structure prediction on the Web: a case study using the Phyre server. *Nat. Protoc.* **4**, 363–371 (2009).
51. Soding, J., Biegert, A. & Lupas, A.N. The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res.* **33**, W244–W248 (2005).
52. Kim, D.E., Chivian, D. & Baker, D. Protein structure prediction and analysis using the Robetta server. *Nucleic Acids Res.* **32**, W526–W531 (2004).