Bioinformatics Solutions Inc.

RAPTOR 3.0
User Manual

For use with the RAPTOR protein structure prediction software

www.BioinformaticsSolutions.com

Index

Introduction……………………………………………………………………………2
Installation…………………………………………………………………………..4
Quick Tour……………………………………………………………………………5
Menu System…………………………………………………………………………..5
Navigation Panel and Output Panel……………………………………………………7
Using RAPTOR………………………………………………………………………..9

Introduction

Homology Modeling

Suppose you know the amino acid sequence of a target protein and you want to know its three-dimensional (3D) structure, yet to be solved experimentally by X-ray crystallography or NMR. An underlying premise for homology modeling is that a set of proteins are homologous, their 3D structures are more conserved than their sequences. The homology modeling method constructs the three-dimensional structure for a target sequence by using the homologous proteins of the target.

General Procedures to Create Homologous Models

- Homologue selection: Identify one or several homologous proteins from the structure database (i.e. PDB). Some computer tools such as PSI-BLAST can be used for this action.
- Sequence alignment: Build a multiple sequence alignment among the target sequence and the selected homologous sequences.
- Core determination: Identify the most conserved segments (cores) and variable segments (loops) in the multiple sequence alignment.
- Core modeling: Predict coordinates of core residues of the target sequence from those of the known structure(s).
- Loop modeling: predict conformations for the loops in the target sequence.
- Side chain packing: construct the side chain coordinates.
- Refinement and Evaluation: The quality of predicted structure can be measured by using some software.

Does Homology Modeling Always Work?

Given a target sequence, if there are no homologous proteins found from the structure database, you cannot use homology modeling in this case. In practice, when the sequence identity in the alignment is below 25%, the homology is insignificant and you can not expect to obtain a good homologous model from homology modeling.

Fold Recognition (Protein Threading)

Fold recognition is based on the observation that the number of distinct structures was not growing as fast as the PDB as a whole and 90% of the new structures submitted to PDB in the past several years have similar structure folds to some structures in PDB. Currently, there are more than 1000 folds.

Protein threading predicts protein structures by using statistical knowledge of the relationship between the structure and the sequence. The prediction is made by "threading" each amino acid of the target sequence to a position in the template structure; evaluation is performed with respect to how well the target fits the template. After the best-fit template is selected, the model is built on the alignment with the chosen template.

Fold Recognition involves the following procedures:

Preparation

- The construction of a structure template database: Select protein structures from the PDB as structural templates.
- The design of a scoring function: Design a good scoring function to measure the fitness between target sequences and template.
    - A good scoring function should consider:
        - mutation potential
        - environment fitness potential
        - pair-wise potential
        - secondary structure compatibilities
        - gap penalties.

      The quality of the scoring function is closely related to the prediction accuracy.

Given a Target Sequence

- Threading alignment: Align the target sequence with each structure template by optimizing the designed scoring function. If there are 'N' structure template in the database, after this step, there will be 'N' alignments.
- Ranking alignment: All the obtained alignments are ranked by using various measuring methods and the best alignment is identified.
- Build the structural model from the selected alignment as homology modeling does, i.e. core determination, core modeling, loop modeling, side-chain packing.

<u>Fold recognition is most effective for hard targets that homology modeling cannot handle</u>. In practice, when the sequence identify is below 25%, in many cases, fold recognition can give reasonably good prediction.

What is RAPTOR?

RAPTOR (RApid Protein Threading predictOR) is a protein threading software package developed by Dr. Jinbo Xu and Dr. Ming Li. It applies novel Linear Programming techniques to the protein threading problem and has achieved great success. RAPTOR has been consistently ranked in the top tier in recent CASP's. In CASP5, RAPTOR paper was voted as the "most innovative paper" by peers in the research community.

Installation

Files required:

RAPTOR1.tar.gz	Executable and Template Library
Install_script1.sh	Install Script 1
RAPTOR2.tar.gz	RefSeq Database used by PHI-BLAST
Install_script2.sh	Install Script 2

How to Install

Copy all the files to a temporary direction and enter that directory.
You may need to run "chmod u+x *.sh" to make the two script files executable.

Then run Install_script1.sh and *Install_script2.sh* respectively.
This will install RAPTOR in RAPTOR/ under your home directory.

If you do not have *RAPTOR2.tar.gz*
Install *RAPTOR1.tar.gz* first.

Then you can download NR database by yourself from ftp://ftp.ncbi.nih.gov/blast/db/

Here are instructions:
Download *nr.00.tar.gz* and *nr.01.tar.gz* to a directory
Uncompress them in that directory and you will obtain a bunch of files whose names start with "nr.00." or "nr.01.".

Move those files to data/nr/ under RAPTOR/

Registration

Run RAPTOR server which is in bin/. A registration window will pop up and a key is required for the registration.

Organization of directories and important files

RAPTOR	
bin/	Binaries
data/	
fssp/	Template fssp Files
PSM/	Template PSM Files
parameters/	
fssp.list	Template List
RAPTOR.conf	Configuration File of RAPTOR
GuiProperties.conf	Configuration File of the GUI
Ip-files/	Parameter Files used in IP
nocore-files/	Parameter Files used in NoCore
nocore2-files/	Parameters files used in NPCore
pdb	Template PDB Files
WEIGHTS/	Parameter Files used by Support Vector Machine

Quick Tour

To test RAPTOR, you can load a test sequence and run it with RAPTOR. To do that, you click "File" in the menu and select "Load Sequence/XML File". In the file browser, you can go to data/seq/ and load one test sequence into the work space. After that, you will see an icon on the left panel and the content of the sequence will be displayed in a window on the right.

Then you can select "Run" in the menu and select "Run Selected" from the dropdown menu. A configuration panel will pop up. The only option that you may need to change is the path of the database used by PHI-BLAST depending on how you install the database. Click "Advanced" tab and find the "Database for PHI-BLAST". If you have installed NR database, the path should be [home directory]/RAPTOR/data/nr/nr. If you have installed RefSeq database, the path should be [home directory]/RAPTOR/data/nr/RefSeq. [home directory] is the path of your home directory.

Click "Run" and RAPTOR will start to run. It will take about one hour to run one sequence depending on the sequence length. After the sequence is finished, a tabbed window will appear on the right. You will find PSP matrix obtained by PHI-BLAST, predicted secondary structure, ranking list of templates and all the alignments.

Menu System

File
File->Load
You can load a sequence file (.seq) or a output file (.xml).

File->Close Selected
You can close the output windows for the selected sequence

File->Close All
Close the windows for all the sequences in the workspace

File->Delete Output
Delete the XML file for the selected sequence

File->Exit
Exit the GUI

Edit
Edit->RAPTOR Config
This will pop up a configuration panel where you can set up the configuration of RAPTOR

Run
Run->Run Selected
This will pop up the configuration panel and after you press "Run" the sequence will be run

Run->Run All
This will pop up the configuration panel and after you press "Run" all the sequences in the work space will be run

Window
You can select different window from the drop down menu

Help
You can launch a browser to read the manual or visit BSI website.

Configuration Panel (Basic)

Threading Method
There are three threading methods available in RAPTOR: NoCore, NPCore and IP. You can select to run one, two or all of them in a run.

3D Modeling
You can let RAPTOR call Modeller automatically after doing threading. Select the check box and locate the Modeller program in the file browser. If you prefer to do 3D modeling with ICM PRO, RAPTOR can also output ICM Pro input files. You just select the check box and specify an output path.

Output Path
This is the directory in which RAPTOR will be run and all the output files will be stored in it.

Output Files
You need to specify how many templates are saved in the templates. If you save too many templates in the XML file, the file will take up too much disk space.

Keep raw files
You can select to keep or remove RAPTOR raw output files.

Advanced
Template Settings

List Path
The path of the template list. It is a text file which stores the names of all the templates in the template library.

FSSP Path
The directory where all the .fssp files are stored

PSM Path
The directory where all the .psm files are stored

PDB Path
The directory where all the trimmed .pdb files are stored.

Database for PHI-BLAST
If you use NR database, it should be [nr path]\nr
If you use RefSeq database, it should be [refseq path]\refseq_protein
Example: if all the NR files are in /home/usr/RAPTOR/data/NR,
then this field should be like: /home/usr/RAPTOR/data/NR/nr

If the RefSeq files are in /home/usr/RAPTOR/data/RefSeq/, then this field
Should be like /home/usr/RAPTOR/data/RefSeq/refseq_protein

PDB File Viewer
This is the view that will be called automatically in RAPTOR. A RasMol viewer comes with RAPTOR. If you have your own viewer, make sure it can be called to display a pdb file from command line like this: yourviewer xxxx.pdb

Template Ranking Method
RAPTOR supports two template ranking methods:
Support Vector Machine (SVM) and Z-score. Normally, you should use SVM.
For very long or short sequences, you can use Z-score for possible better result.

Navigation Panel and Output Window

Navigation Panel

The left hand side is the navigation panel. Each Sequence is represented by an icon (icon picture inserted). After running RAPTOR, the RAPTOR output is represented by (a raptor head icon, insert picture). You can browser different sequences and their outputs by clicking different icons in the navigation panel.

Output Window

PHI-BLAST Profile

The output window is composed of a set of tab windows. The first tab window is PHI-BLAST profile. It is a 20 row matrix, each row corresponding to some amino acid. The column width is the length of the query sequence. Thus each residue in a query sequence has a 20-element vector with it. Each element represents the occurring frequency of certain amino acid at that position in the multiple sequence alignment obtained from PHI-BLAST output.

The frequency is from 0 to 1. To make it easier for you to read the profile, the frequency is divided into 10 segments. Each segment will be represented by a color. In this way, the matrix can be represented by a rectangle in the window which is composed of many small square cells. The color of cell is determined by the occurring frequency. You can easily find out the conserved residues and non-conserved residues by differentiating colors.

Secondary Structure

Different colors are used to represent helices, beta sheets, loops (add color in html).
Some acronyms

AA	amino acid
PHD	PsiPred predicted secondary structure.
E	Beta Strand
H	Helices.
Space	Loops
Rel	Confidence of predicted secondary structure type
PrE	Chance of being beta strand
PrH	Chance of being helix
PrL	Cchance of being loop

Rank by Score

Top Window
Each method is represented by a folder icon (add a picture). If you double click it, the templates will be displayed, ranked by their E-values. The smaller the E-value, the better. Also displayed are other scored used internally.

Table fields:
tName: template name
eValue: E value
tLen: template length
sLen: target length
Score:alignment score
mScore: mutation score
fScore: environmental fitness score
gScore: gap score
ssScore: secondary structure score
pScore: pairwise score
cScore: contact capacity score
SVMout: score output by the Support Vector Machine

Bottom Window
If you click a template, its alignment will be displayed in a drop down window. The color of the template is consistent with its actual secondary structure and the color of the target is consistent with its predicted secondary structure.

If you click the template name, a browser will be launched and connected to the PDB website.
If you click “View 3D structure with RasMol”, a RasMol window will pop up and the structure will be displayed.
If you click “Functional Annotation” tab, a window will drop down and show the functional information extracted from the template pdb file.

Alignments

The left side of the toolbar allows you to select some session(s) and specify how many templates you want to display. The right side of the tool bar allows you to compare any two alignments. To specify an alignment, you can use method name and its rank.

Error

This window displays the errors that occurred during the run.

Using RAPTOR

Input file and Output file

RAPTOR accept FASTA format sequence file as input. Here is an example of FASTA format sequence:

```
>2acy(len=98)
AEGDTLISVDYEIFGKVQGVFFRKYTQAEGKKLGLVGWVQNTDQGTVQGQLQGPASKV
RHMQEWLETKGSPKSHIDRASFHNEKVIVKLDYTDFQIVK
```

The default suffix for sequence file is ".seq". If the file you loaded does not have right suffix, ".seq" will be appended to the file name.

The output of RAPOR is stored in XML file. You can load an XML file saved by RAPTOR and display its content.

All the raw files of RAPTOR are stored in a directory whose name is the sequence name in the output directory. Suppose the sequence name is XXXX.
Here is the structure of directory XXXX

```
XXXX
        PSP    PHI-BLAST output files
        SS      PHI-PRED output files
        [method name]
                MODEL          alignment files  .pir file
                OUT            ranking files   .scoreRank file
                <Modeller Output>  modelleroutput .pdb file
                <ICM Pro Input>    ICM Pro input files
```

The structure of output directory:

```
PSP     PHI-BLAST output      file
SS      secondary structure prediction output files
[method name] temporarily store threading output
XXXX
```

Where [method name] can be NoCore, NPCore, or IP. Directories embraced by <> are only generated when the corresponding checkbox is selected and the path is specified in the configuration panel

PHI-BLAST Database

In RAPTOR, PHI-BLAST is used to generate position specific matrix (sequence profile) of a target sequence. By default, PHI-BLAST uses NR database, but the size of NR database is very large (1 G after compression). So an alternative database is RefSeq, which is a curated non-redundant sequence database of genomes, transcripts and proteins maintained by NCBI. RefSeq is much smaller, about half size of NR. We conducted a comparison of the two. The profiles obtained from them are almost the same. So you can always use RefSeq to replace NR.

NR database can be downloaded from ftp://ftp.ncbi.nih.gov/blast/db/nr.00.tar.gz & ftp://ftp.ncbi.nih.gov/blast/db/nr.01.tar.gz

RefSeq can be downloaded from ftp://ftp.ncbi.nih.gov/blast/db/refseq_protein.tar.gz. After uncompressing, you can obtain a bunch of index files. You need to put them in some directory and specify the path in the configuration panel (add a hyperlink here).

Threading Methods

Dynamic Programming vs. Integer Programming

RAPTOR has three threading methods available: NoCore, NPCore, and IP. NoCore and NPCore both use dynamic programming to optimize the scoring function. IP uses integer programming to optimize the scoring function. The difference is that if a scoring function considers pair-wise contact, dynamic program can only find a local optimum solution while integer programming can find the global optimal solution. Most of other threading servers are based on dynamic programming and RAPTOR's integer programming is unique.

NoCore vs. NPCore

NoCore and NPCore are both based on dynamic programming. The difference is that in NPCore, the template and target are first divided into cores before doing threading. A core is a conserved segment of a protein. NoCore and NPCore are very effective for easy targets.

Running One Sequence with Different Methods

IP's running time is longer than NoCore and NPCore. Thus, given a target sequence, you can run NoCore first. If the prediction is not good, try NPCore. If both cannot give good predictions, you can try IP. This will save you much time. Of course, you can also run more than one methods at one time. RAPTOR can keep up to three methods' output in the XML file. When you run NPCore after running NoCore, the output will be automatically inserted into the XML file. If you run NoCore for the second time with different configuration, the old result in the XML file will be overwritten by the new result.

The fist step of RAPTOR is to run PHI-BLAST. If you already run NoCore, then when you run NPCore, this step will be skipped,as the PHI-BLAST is stored in PSP/ under the output directory. If the program finds those files, PHI-BLAST will be skipped. This will save running time.

Judging Prediction Quality from Alignment

First, you can compare the actual secondary structure of the template with the predicted secondary structure of the query sequence. As the accuracy of secondary structure is around 80%, this is an important measure of the prediction quality. Then you can look at the gaps in the alignment. *The fewer the gaps, the better the prediction quality. The shorter the gaps, the better the prediction quality.* Ending gaps normally can be ignored. Sometimes, the ending gaps may be very long. This means the program can only give good prediction for part of the query sequence. What if the ending gaps are too long? In many cases, for long sequences, they may have more than one domain. Thus the ending gaps may be very long. You can cut them into domains first and run each domain with RAPTOR.

Using Modeller

If you are an academic user, you can download Modeller for free from http://www.salilab.org/modeller/download_installation.html
And you need to register at http://www.salilab.org/modeller/registration.html to get a license key in order to install Modeller.

Customizing Templates

RAPTOR/data/parameters/fssp.list stores the names f all the templates in the template library. If you are interested in a specific template, you can save its name in another file and specify the path in the configuration panel.

You can also create your own template library. You need a pdb file and generate PSM and fssp file from it. Then put PSM file in RAPTOR/data/PSM and fssp file in RAPTOR/data/fssp

Using RasMol

The default viewer for pdb files is RasMol. The default display mode is cartoon. The structure is colored according to the secondary structure. You can rotate the structure by using the left key of the mouse. To move the structure, press the right mouse key and drag. To shrink or enlarge the display, press "shift" key, press the right mouse key and drag. For a full reference of RasMol, you can visit http://www.umass.edu/microbio/rasmol/

RAPTOR Reference List

Feng Jiao, Jinbo Xu, Libo Yu, Dale Schuurmans. Protein Fold Recognition Using Gradient Boost Algorithm. Accepted by CSB 2006.

Jinbo Xu. Protein Fold Recognition by Predicted Alignment Accuracy. ACM/IEEE Transactions on Computational Biology and Bioinformatics, 2(2):157-165. 2005.

Jinbo Xu, Ming Li, Dongsup Kim, Ying Xu. RAPTOR: optimal protein threading by linear programming. Journal of Bioinformatics and Computational Biology 1:1(2003) 95-117.

Jinbo Xu and Ming Li. Assessment of RAPTOR's linear programming approach in CAFASP3. Proteins: Structure, Function, and Genetics, 53(S6): 579--584, Oct. 2003. Invited paper for CASP5, voted by peers as the "most innovative method in CASP5".

BSi

Bioinformatics Solutions Inc.

Technical Support

Email: raptor@bioinfor.com
Phone: 1-519-8858288 ext. 16