International Journal of Computer Information Systems and Industrial Management Applications (IJCISIM)
ISSN: 2150-7988 Vol.1 (2009), pp.321-328
http://www.mirlabs.org/ijcisim

STPred Server: Protein Structure Prediction Server

Sayantan Ghosh, Ketan Pandey, Febin J. Prabhudass

Department of Bioinformatics, School of Biotechnology, Chemical and Biomedical Engineering
VIT University, Vellore, 632014, India

sayantan7@gmail.com, evilsindicate@gmail.com and febinprabhudass@vit.ac.in

Abstract

A plethora of online structure prediction for proteins servers can be found over the internet. Increasing complexities during structure prediction may be attributed to the quotidian rise of these servers. One of the key limitations of these servers is the lack of rapid, robust modeling, paltry results and the inability to give reproducible methods for detecting, matching and analyzing protein structures. The most commonly used approaches involve getting an unknown sequence (and a template, in some servers) and then aligning the sequence with its own pre-defined parameters, then obtaining a predicted secondary structure. We introduce a server, using Java Server Pages architecture upon Apache Tomcat, for protein structure prediction wherein the user is provided maximum control over the parameters and variables, which define the relationship and homology of the unknown sequence with known sequence databases. Added to that, the package uses Profile Hidden Markov Models for template selection and Python programs of Modeller for structure prediction based on the selected templates. It coherently implies the user doing his modeling studies as if on (more reliable) standalone software, and that too, without the hassles of any coding, writing tedious scripts or fallacious guess work resulting in protracted homology models. There is an auxiliary module to the server to determine the composition of the user's unknown protein structure and correct mistakes, if any. All the steps in course are fully transparent so as to give full independence of changing the variables as and when, suited to the user, to get perfect results.

Keywords
Structure Prediction, Clustal-W, HMMER, Modeller, DOPE Score.

1. Introduction

Protein structure prediction [10] is a set of techniques in Bioinformatics that aim to predict the local secondary structures of proteins and RNA sequences based only on knowledge of their primary structure - amino acid or nucleotide sequence, respectively. For proteins, a prediction consists of assigning regions of the amino acid sequence [11,12] as alpha helices, beta strands (often noted as "extended" conformations), or turns [11]. Specialized algorithms [10] have been developed for the detection of specific well-defined patterns such as trans-membrane helices and coiled coils in proteins, or canonical microRNA structures in RNA [10,11].

The best modern methods of structure prediction in proteins reach about 80 per cent accuracy allowing the use of the predictions in fold recognition and ab-initio protein structure prediction, classification of structural motifs, and refinement of sequence alignments. The accuracy of current protein structure prediction methods is assessed in weekly benchmarks such as EsyPred3D and Bioinfo Bank Meta Server.

There are three modules in the STPred server. These are divided into **Structure Composition**, **Template Selection** and **Structure Prediction**. All these modules are essential to determine the most probable model for the secondary structure of a given unknown protein sequence as we shall see in the proceeding pages.

2. Tools and Techniques

2.1 Clustal W

Clustal W helps perform multiple sequence alignment for the unknown sequences given as a flat file format. In addition, it also creates Phylogenetic trees which can be used for evolutionary studies [4]. The command line options are divided into 2 groups:

- Data: This includes the file name of the sequence and the output file for multiple sequence alignment.
- Verb: This subgroup includes options for creating trees in “newick” format [4,37,40]

clustal –infile=filename.ali outfile.ali [37]

2.2 HMMER

Hmmbuild

If we have a multiple sequence alignment of a protein domain or protein sequence family. To use HMMER [12,13,31] to search for additional remote homologues of the family, we want to first build a profile HMM from the alignment. The following command builds a profile HMM from the alignment of 50 globin sequences in globins50.msf:

> *hmmbuild globin.hmm globins50.msf* [12]

Hmmalign

Another use of profile HMMs is to create multiple sequence alignments of large numbers of sequences. A profile HMM can be build of a “seed” alignment of a small number of representative sequences, and this profile HMM can be used to efficiently align any number of additional sequences. For example, to align the 630 globin sequences in **globins630.fa** to our globin model **globin.hmm**, and create a new alignment file called **globins630.ali**, we’d do:

> hmmalign -o globins630.ali globin.hmm globins630.fa [12]

Hmmsearch

As an example of searching for new homologues using a profile HMM, we’ll use the globin model to search for globin domains

> hmmsearch globin.hmm Artemia.fa [12]

2.3 MODELLER

Align2d

This command aligns a block of sequences (second block) with a block of structures (first block). It is the same as the alignment.align() [3] command except that a variable gap opening penalty is used. This gap penalty depends on the 3D structure of all sequences [19,25,35,36] in block 1. The variable gap penalty can favor gaps in exposed regions, avoid gaps within secondary structure elements, favor gaps in curved parts of the main chain [22,23] , and minimize the distance between the two positions spanning a gap. The alignment.align2d () [3] command is preferred for aligning a sequence with structure(s) in comparative modeling because it tends to place gaps in a better structural context.

```
aln = alignment(env)
aln.append_model(mdl,align_codes='1vhbA',atom_files
='1vhb.pdb')
aln.append(file='c:/STPred_serverV2/BAHG123.ali',
align_codes='BAHG123');aln.align2d()
```

Code: Alignment.py [3]

Automodel

If we do not have an initial alignment between the templates and target sequence, MODELLER can derive one for we, fully automatically. All MODELLER requires is a PIR file containing the target sequence and the template PDB codes (their sequences are not required -- just use a single '*' character -- as MODELLER will read these from the PDBs). Use the automodel class as per usual, but call the automodel.auto_align() [3] method before automodel.make()[3].

```
a=automodel(env,alnfile='c:/STPred_serverV2/BAHG12
3-1vhbA.ali',knowns='1vhbA', sequence='BAHG123',
assess_methods=(assess.DOPE, assess.GA341))
a.make()
```

Code: Automodel.py [3]

Model Evaluation

If several models are calculated for the same target, the "best" model can be selected in several ways. For example, we could pick the model with the lowest value of the MODELLER objective function or the DOPE assessment score, or with the highest GA341 assessment score [12,13] , all of which are reported in the log file, above. (The objective function, molpdf, is always calculated, and is also reported in a REMARK in each generated PDB file. The DOPE and GA341 scores, or any other assessment scores, are only calculated if we list them in assess_methods.) The molpdf and DOPE scores are not 'absolute' measures, in the sense that they can only be used to rank models calculated from the same alignment. Other scores are transferable. For example GA341 scores always range from 0.0 (worst) to 1.0 (native-like); however GA341 is not as good as DOPE at distinguishing 'good' models from 'bad' models.

```
# read model file
mdl = complete_pdb(env, 'TvLDH.B99990002.pdb')
# Assess with DOPE:
s = selection(mdl)    # all atom selection
s.assess_dope(output='ENERGY_PROFILE
NO_REPORT', file='TvLDH.profile',
normalize_profile=True,smoothing_window=15)
```

Code: Evaluate.py [3]

3. Flow Chart:

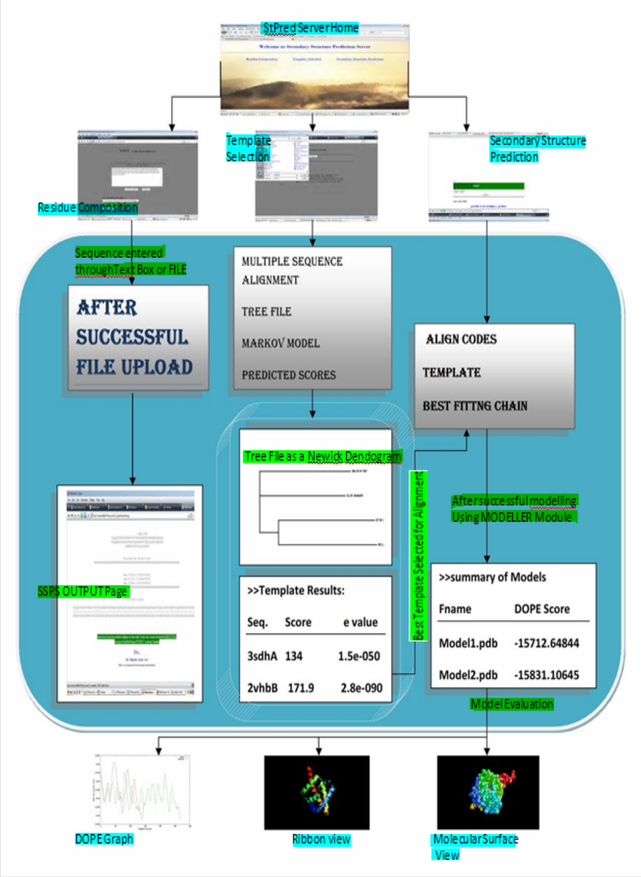


Fig 1:Flowchart

Our server is organised into parallel and serial hierarchical arrangements. This gives much flexibility to the server. The modules have been arranged in a crossed-linked parallel hierarchy that attributes to its multi-tasking abilities and enables the user to work with multiple sequences at a given instance of time. The working of the server is given as follows:

1. The unknown sequence can be checked for validity using the Residue Composition module. This module provides information about the percentage of various components of the protein along with their total number and molecular weight.

2. The actual procedure of calculating models starts at the Template Selection Module. This module takes in the unknown sequence as input and gives out the following outputs:

i. Multiple sequence alignment of the unknown sequence with its known templates.

ii. Tree file depicting phylogeny.

iii. Scores and e-values of different templates.

After proper analysis of the MSA, tree files, scores and evalues, a suitable template is selected. In case of weak homology, upto 3 templates can be chosen for modelling. These templates and the unknown sequence are fed to the Secondary Structure Prediction module.

3. Once the unknown sequence is uploaded, the server performs align2d. This gives the alignment of the unknown sequence with the structure of the PDB template. Once the server performs align2d, model calculation is done. The server allows up to 5 models to be generated from the template. The **final output** gives the details of the models that have been calculated along with their molpdf and DOPE scores.

4. Implementation

4.1 Residue Composition

The input for this module can either be a raw sequence which contains only the amino acid sequence for the protein or the user can upload a file containing the protein in any standard format found in major protein sequence data formats viz. FASTA, PIR, CLUSTAL etc. [33,37]

After entering the sequence into the server, the user is forwarded to the results page. This page contains vital information relevant to the given protein sequence. The results present the user with a preliminary analysis of the protein with which to proceed to the next level of a more accurate determination of the possible secondary structure.

The results obtained give the user the length of the protein sequence i.e. how many amino acids are present in the protein and also the percentage composition of each class of secondary structure (namely alpha- helix, beta- sheet and coil). [11,26] The percentage composition gives an idea of the overall structural composition of the protein and the same can be verified by further analysis on the STPred server.

4.2 Template Selection

This module is crucial for obtaining pdb templates to the input sequence. The general idea is that protein sequences which are similar in amino acid content also share common structural properties [5,20]. These templates are obtained from a standard protein database [29,32] which consists of a large number of protein sequences already analyzed by experts and the user can determine the functions and physiochemical properties of his/her input sequence.

The sequences should be in a single file. The following formats are recognized and accepted by the server: NBRF-PIR, EMBL-SWISSPROT, Pearson (FASTA), Clustal (.aln), GCG-NSF (Pileup), GCGS-RSF and GDE flat file [29,33]. This method is more accurate since there is more data available for the server to work with and only those templates which fulfill all conditions are shown in the results page. Since there are multiple sequences, the server does a Multiple Sequence Analysis (MSA) of the input sequence.

After this, a Newick tree [Fig.2] is calculated [34,38] which diagrammatically shows the relative similarity among the sequences given by the user. The code is given as follows:

```
my $tree = parse( -format => 'newick', -string =>
$string )->first;
my $treedrawer = Bio::Phylo::Treedrawer->new(
     -shape   => 'RECT', # histogram curvy,diag
     -mode    => 'PHYLO', # cladogram
     -format => 'SVG' );
```

Code: Treedrawer.pl [34,38]

Subsequently, a HMM Profile is calculated [12,13]. It is much more accurate for multiple sequence file than in the case of a single sequence input.

Once the HMM Profile is calculated for either input methods, the user has to choose any one of the following databases against which to search for suitable templates [14,31] . The databases available in the server are

- Human Protein Family: This database contains templates for the human protein family. In case the query sequence is a human protein sequence.
- Globins Family: It is a Globin sequence database to be used if the query is a globin protein.
- RCSB PDB Concise Database: It contains the PDB Database templates of all currently confirmed crystallographic models.
- RCSB PDB Extensive Database: It contains the entire PDB database. This database may contain duplicate entries and also is slower to process than the other databases.

The output will show the MSA file, the Newick tree file along with the Markov Model for the aligned sequences and predicted Scores and E-values for templates. To select the most accurate template, take note of the Scores and E-values. Sequences which have a **high score** and a corresponding **low E-value** score are more suitable for secondary structure analysis available in the server [2,3,7].

4.3 Structure Prediction

In this module, the aim is to find the most probable secondary structure for the query sequence based on the templates which were found from the Template Selection module. The module is based on certain Python scripts which are a part of Modeller. The final structure is displayed in 3D so that all the chains and bonds along with the active site for the protein [10,16,18] can be viewed clearly.

The input to be provided to the server is the query protein sequence along with the templates found earlier. The user can choose to give either a single sequence as template or up to 3 sequences to serve as template upon which to model the predicted structure. The input sequence has to be in PIR format (*.ali) [3] and the templates can be in PDB format.

To ensure proper modelling the user can try to upload the templates in decreasing order of homology. This can easily be done by checking the Scores for the templates.

The server then creates an **alignment profile** for the templates and the query sequence.

```
aln = alignment(env)
aln.append_model(mdl,align_codes='1vhbA',
atom_files='1vhb.pdb')
aln.append(file='c:/STPred_serverV2/BAHG123.ali',
align_codes='BAHG123')
aln.align2d()
```

Code: Alignment.py [3]

Based on this profile, probable models are created which are theoretical models. There are individual values assigned to each model such as **'molpdf'** scores, **'DOPE'** scores and **'GA341'** scores. To identify the most probable model for the query sequence, these values are taken into account. For example, a model with a high molpdf score and a negative DOPE score along with a GA341 score closest to 1 will be taken as the most probable model for the protein.

The user can also evaluate any of the models against the template to check for energy deviations against the selected template. The final results will show the user the most probable structure for the protein along with as many as 4 more models which can also be taken as possible models but with a lower accuracy. The DOPE energy graph shows the alignment of the most probable model against the selected templates. This can show DOPE per-residue score against alignment positions in the templates. The DOPE score signifies the error level in the predicted models.

5. Results

5.1 Residue Composition

The output obtained from residues composition studies is divided into 3 parts.

- **The number of amino acids** – An accurate calculation of the number of amino acids in the unknown sequence.
- **Percentage of helices, coils and sheets –** This gives the prediction of % composition of helices, coils and sheets based o preliminary studies.
- **Molecular weight –** The molecular weight is calculated to understand the structure of the unknown protein completely.

5.2 Template Selection

In template selection, multiple sequence alignment is performed with the unknown sequence, which gives the following results:

-Tree File – This is obtained in NEWICK [2] format. The output is a phylogenetic dendogram. As shown in the figure, for example, our unknown sequence, named MELN HUMAN is most related only to ZFB4 HUMAN. The other two templates are far away in phylogeny. This gives us a sure shot 'good' template, thus, eliminating weak templates from the list, which in turn helps in calculating more accurate model in the final stage.

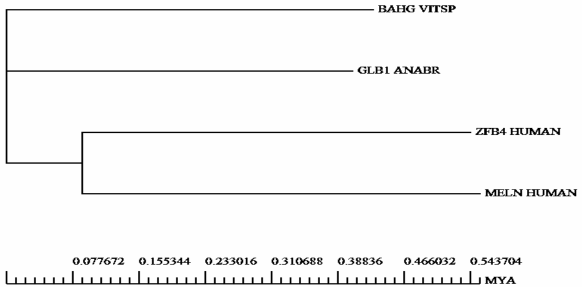
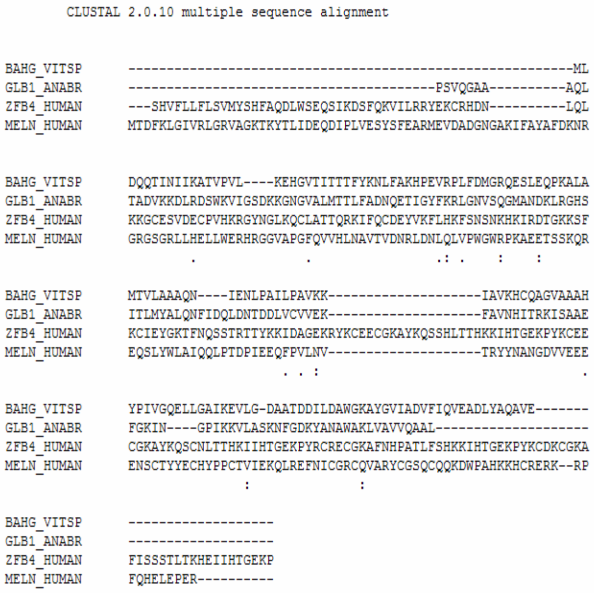
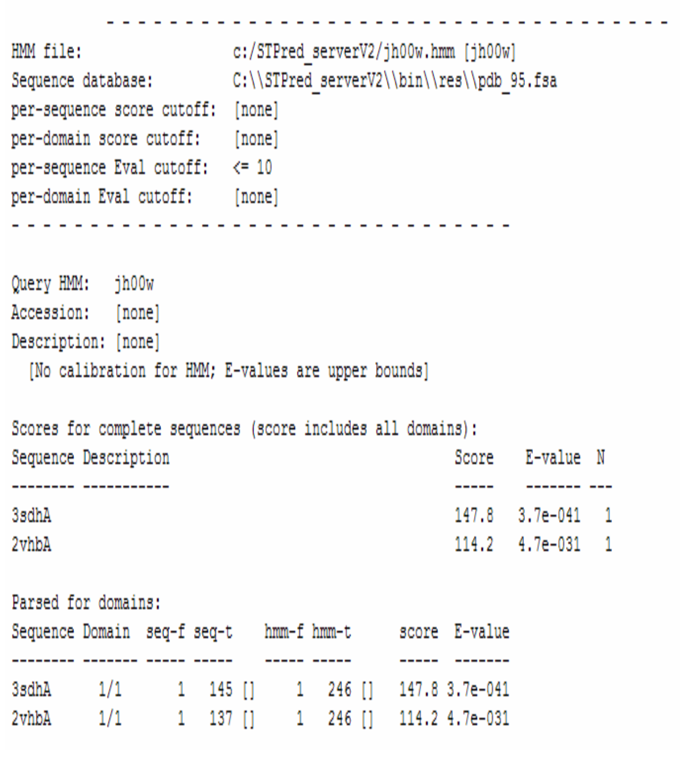


Fig 2: Phylogenetic Dendogram for the input file



```
CLUSTAL 2.0.10 multiple sequence alignment


BAHG_VITSP      ----------------------------------------------------------ML
GLB1_ANABR      ----------------------------------------PSVQGAA----------AQL
ZFB4_HUMAN      ---SHVFLLFLSVMYSHFAQDLWSEQSIKDSFQKVILRRYEKCRHDN----------LQL
MELN_HUMAN      MTDFKLGIVRLGRVAGKTKYTLIDEQDIPLVESYSFEARMEVDADGNGAKIFAYAFDKNR


BAHG_VITSP      DQQTINIIKATVPVL----KEHGVTITTTFYKNLFAKHPEVRPLFDMGRQESLEQPKALA
GLB1_ANABR      TADVKKDLRDSWKVIGSDKKGNGVALMTTLFADNQETIGYFKRLGNVSQGMANDKLRGHS
ZFB4_HUMAN      KKGCESVDECPVHKRGYNGLKQCLATTQRKIFQCDEYVKFLHKFSNSNKHKIRDTGKKSF
MELN_HUMAN      GRGSGRLLHELLWERHRGGVAPGFQVVHLNAVTVDNRLDNLQLVPWGWRPKAEETSSKQR
                        .              .                .: .    :    :

BAHG_VITSP      MTVLAAAQN----IENLPAILPAVKK--------------------IAVKHCQAGVAAAH
GLB1_ANABR      ITLMYALQNFIDQLDNTDDLVCVVEK--------------------FAVNHITRKISAAE
ZFB4_HUMAN      KCIEYGKTFNQSSTRTTYKKIDAGEKRYKCEECGKAYKQSSHLTTHKKIHTGEKPYKCEE
MELN_HUMAN      EQSLYWLAIQQLPTDPIEEQFPVLNV--------------------TRYYNANGDVVEEE
                                    . . :                                  .

BAHG_VITSP      YPIVGQELLGAIKEVLG-DAATDDILDAWGKAYGVIADVFIQVEADLYAQAVE-------
GLB1_ANABR      FGKIN----GPIKKVLASKNFGDKYANAWAKLVAVVQAAL--------------------
ZFB4_HUMAN      CGKAYKQSCNLTTHKIIHTGEKPYRCRECGKAFNHPATLFSHKKIHTGEKPYKCDKCGKA
MELN_HUMAN      ENSCTYYECHYPPCTVIEKQLREFNICGRCQVARYCGSQCQQKDWPAHKKHCRERK--RP
                               :              :

BAHG_VITSP      -------------------
GLB1_ANABR      -------------------
ZFB4_HUMAN      FISSSTLTKHEIIHTGEKP
MELN_HUMAN      FQHELEPER----------
```

Fig3.1: Multiple Sequence Alignment and HMM Search results



```
              - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
HMM file:                  c:/STPred_serverV2/jh00w.hmm [jh00w]
Sequence database:         C:\\STPred_serverV2\\bin\\res\\pdb_95.fsa
per-sequence score cutoff: [none]
per-domain score cutoff:   [none]
per-sequence Eval cutoff:  <= 10
per-domain Eval cutoff:    [none]
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

Query HMM:   jh00w
Accession:   [none]
Description: [none]
  [No calibration for HMM; E-values are upper bounds]

Scores for complete sequences (score includes all domains):
Sequence Description                                   Score    E-value  N
-------- -----------                                   -----    ------- ---
3sdhA                                                  147.8   3.7e-041   1
2vhbA                                                  114.2   4.7e-031   1

Parsed for domains:
Sequence Domain  seq-f seq-t    hmm-f hmm-t      score  E-value
-------- ------- ----- -----    ----- -----      -----  -------
3sdhA      1/1       1   145 []     1   246 []   147.8 3.7e-041
2vhbA      1/1       1   137 []     1   246 []   114.2 4.7e-031
```

Fig3.2: Multiple Sequence Alignment and HMM Search results

-Multiple sequence alignment of the unknown sequences

This gives the multiple sequence alignment of the input file in CLUSTAL format [Fig.3.1, 3.2]. The matches are represented by the corresponding residues. The mismatches are assigned a dash.

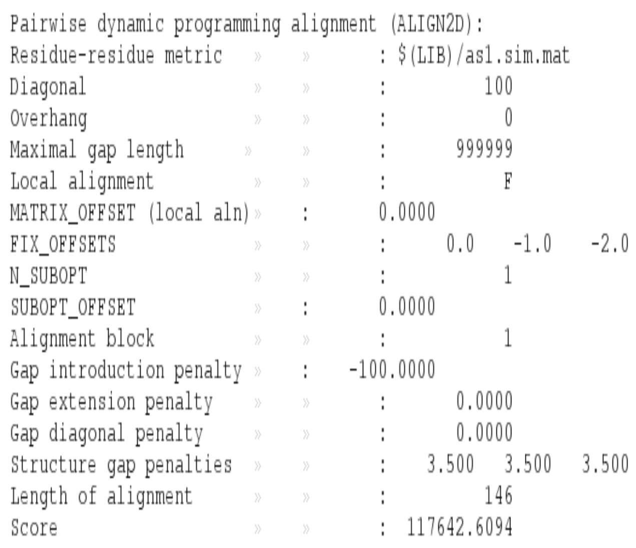
-Predicted scores and E-value

This gives the percentile error or mismatch between the sequences along with the scores for the alignment. A high score and a low E-value denotes that a proper match has been obtained. It also shows the line-wise alignment and a histogram showing the matches with the template.

5.3 Structure Prediction

Once the unknown sequence is uploaded, the server performs align2d. This gives the alignment of the unknown sequence with the structure of the PDB template [15]. Several details are provided regarding **Overhang**, **Gap Penalties** and **Score** [Fig.4]. Once the server performs align2d, model calculation is done. The number of models needed is specified by the user. The server allows up to 5 models to be generated from the template. The output gives the details of the models that have been calculated along with their molpdf and DOPE scores.

Now, the appropriate model is selected by the user based on the **molpdf** and **DOPE** scores for evaluation.



```
Pairwise dynamic programming alignment (ALIGN2D):
Residue-residue metric     »    »       : $(LIB)/as1.sim.mat
Diagonal                   »    »       :         100
Overhang                   »    »       :           0
Maximal gap length         »    »       :      999999
Local alignment            »    »       :           F
MATRIX_OFFSET (local aln)» :       0.0000
FIX_OFFSETS                »    »       :      0.0    -1.0    -2.0
N_SUBOPT                   »    »       :           1
SUBOPT_OFFSET              »    :       0.0000
Alignment block            »    »       :           1
Gap introduction penalty   »    :    -100.0000
Gap extension penalty      »    »       :      0.0000
Gap diagonal penalty       »    »       :      0.0000
Structure gap penalties    »    »       :    3.500   3.500   3.500
Length of alignment        »    »       :         146
Score                      »    »       :  117642.6094
```

Fig 4: Output from the Align2d program

The DOPE profile graph is traced based on the template and the model selected. As shown in the graph in fig 5, the template DOPE values are represented by the green line and that of the model are represented by the red line. The areas which coincide represent a perfect alignment and the areas which don’t coincide show an error in the alignment. If the active site is found in those areas of the graph that coincide, the model is considered to be accurate and no further changes need to be made. In case the active site falls in the regions that don’t coincide, it is advised to change the template by choosing the next nearest template and follow the same steps again.

Sample Final Output Structures:

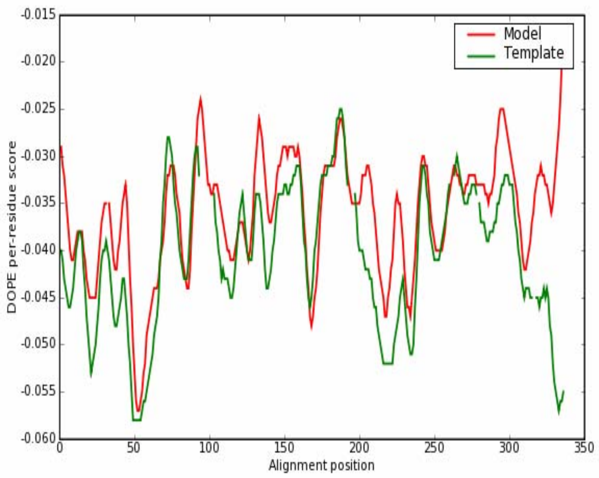


Fig 5: DOPE Per Residue Graph

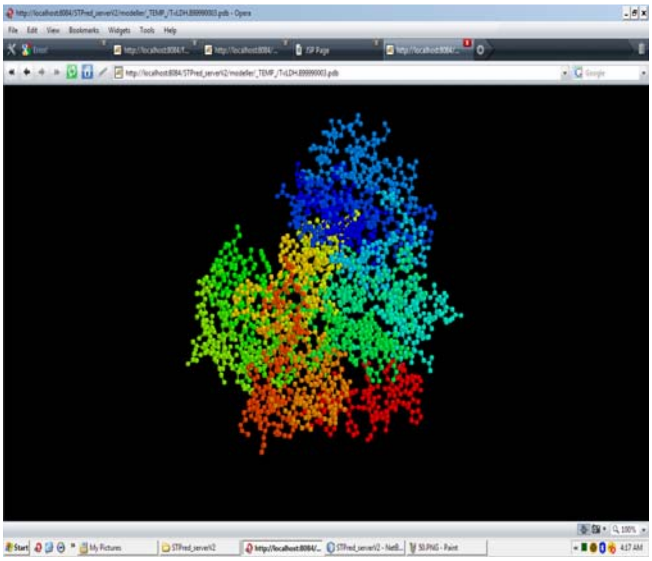

http://localhost:8084/STPred_serverV2/modeller/_TEMP_/T-LDH.B99990003.pdb - Opera
File Edit View Bookmarks Widgets Tools Help
JSP Page
http://localhost:8084/STPred_serverV2/modeller/_TEMP_/T-LDH.B99990003.pdb
Google
Start

Fig 6: The predicted structure from STPred Server.

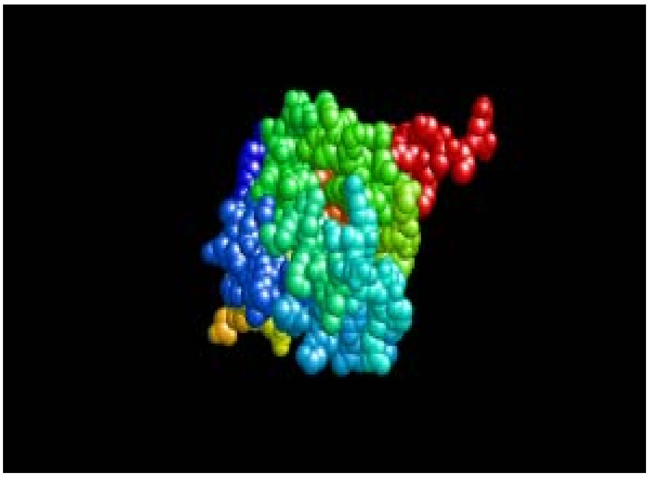


Fig 7: BAHG modelled with 1VHB as template. Molecular Surface View

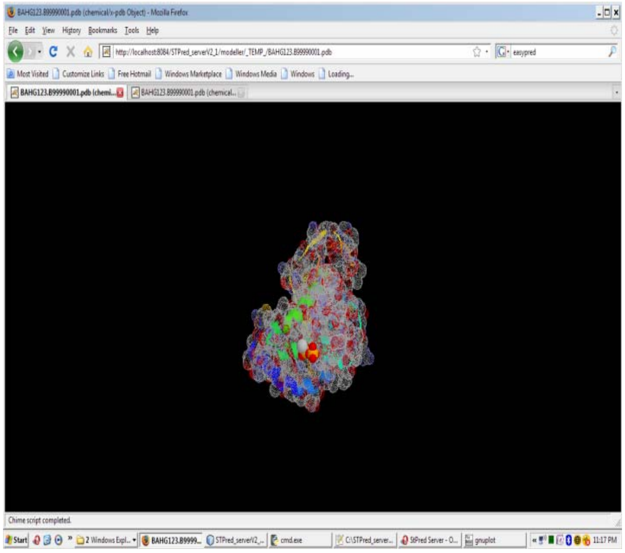

BAHG123.B99990001.pdb (chemical/x-pdb Object) - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://localhost:8084/STPred_serverV2_1/modeller/_TEMP_/BAHG123.B99990001.pdb
easypred
Most Visited
Customize Links
Free Hotmail
Windows Marketplace
Windows Media
Windows
Loading...
BAHG123.B99990001.pdb (chemi...
BAHG123.B99990001.pdb (chemical...
Chime script completed.
Start

Fig 8: BAHG modelled with 1VHB as template... Note the hetero molecule at the bottom

6. Discussion

The STPpred server ensures that the user is provided with accurate, fast and apt results. In addition to accuracy our server is faster than most other preexisting standalone servers. This increased speed can be attributed to the use of JAVA Server Pages (JSP) technology [8,9] which creates separate **threads** for a single **process** coming as a request instead of creating several processes for a single request. All this attributes to the high performance of the server. Added to this, use of Hidden Markov Models in the program algorithm [12,13] increases the accuracy of the templates selected and Modeller programs [3] produce accurate and detailed output structures. The STPred server has been benchmarked against various pre-existing online servers and software. We have compared our server results with the results provided by those servers in order to continually improve the STPred server.

Secondary structure content prediction (SSCP)

This server [15,16,17] uses neural networks to perform the alignment. Alignments are obtained by combining, weighting and screening the results of several multiple alignment programs. This final 3D structure is built using MODELLER [3].

On the other hand, our server uses hidden markov models profiles (HMM profiles) [5,12,14] for template selection and subsequently uses the templates to predict the structure. Similar to the EsyPred3D server, the final structure is built using MODELLER.

Template Selection

This is a novel idea introduced in the STPred server [1]. This is the first server to provide the user with the option of selecting an appropriate template depending on the scores and E-values. This facility was previously available to the user; however, the procedure was tedious. Our server aids the user by providing this option within the serve itself. This will save time as well as effort.

EsyPred3D Server

This server uses neural networks to perform the alignment. Alignments are obtained by combining, weighting and screening the results of several multiple alignment programs. This final 3D structure is built using MODELLER [3]. On the other hand, our server uses hidden markov models profiles (HMM profiles) [13] for template selection and subsequently uses the templates to predict the structure [17,21]. Similar to the EsyPred3D server[30], the final structure is built using MODELLER.

Structure Prediction Meta Server

The Structure Prediction Meta Server [6] provides access to various fold recognition, function prediction and local structure prediction methods. The Server takes the amino acid sequence of the query protein, the reference name for the prediction job, and the E-mail address as input. The Meta Server accepts only sequences, which have not been submitted before. In case of duplicate sequences the second user will be notified with a link to the previous submission. Sequences longer than 800 amino acids are not accepted by some services. Each server has its own process queuing system managed by the Meta Server. All results of fold recognition servers are translated into uniform formats. The information extracted from the raw output of the servers includes the PDB codes of the hits, the alignments and the similarity (reliability) scores specific for every server. The secondary structure assignments for all hits are taken from the mapped FSSP. Underscored amino acids indicate the first residue after an insertion in the template sequence. The Meta server provides translation of the alignments in standard formats like FASTA [29], PDB or CASP. The Meta Server is coupled to consensus servers.

In contrast, our server is a standalone server taking in query sequences as flat files and calculating the 3D PDB structures using its own algorithms and MODELLER [3] module, however, the server is faster in comparison to most other servers and nearly as fast as the Meta server.

7. Future Enhancements

Utmost care has been taken to provide accurate results along with maintaining the speed of the server. However, there’s still room for improvement.

The residue composition algorithm can be improved and made more accurate by using a larger test dataset. Currently the server gives individual class of secondary structures. This system can be converted into a windowed system, which will enable our server to predict more complex structures with higher accuracy.

For the Template Selection module, predetermined HMM profiles for a large neural network test dataset can be incorporated into the server. It can also be connected to the BLAST server for getting a larger set of templates from the NCBI database. This would further lead to increase in both accuracy and automation of the server.

Currently, the Secondary Structure Prediction module takes 5 to 15 minutes for calculating the structure. This can be improved upon in the future by using the Java packages [27,28] directly instead of redirection through Python functions. At present, the server gives a maximum of 5 predicted structures. This can be increased along with the ability to give separate structures with hetero-atoms.

8. References

[1] Altschul S.F., Carroll R.J., Lipmann D.J.: weights for data related by a tree. J Mol Biol. (pp. 207:647-53), 1989.

[2] Andreas D.Baxevanis, B.F.Franccis: Bioinformatics: A practical guide to the Analysis of Genes and Proteins Protein Structure Prediction and Analysis (pp.223-51) Predictive Methods using protein Sequences (pp.197-220) Using Perl to facilitate biological Analysis (pp.481-495), 2005.

[3] Andrej Šali : Modeller Tutorial MODELLER, a protein structure modeling program. [available at: http://salilab.org/modeller/], 1989.

[4] Apache Batik SVG Toolkit Batik is a Java based toolkit for applications which handle images in the Scalable Vector Graphics (SVG) format for viewing, generation or manipulation. [available at: http://xmlgraphics.apache.org/batik/], 1999.

[5] Bernstein F.C., Koetzle T.F., Williams G.J.B., Meyer E.F., Brice M.D. et al: The Protein Data Bank. J Mol Biol. (pp.112,535-42)

[6] the prediction servers and 3D-Jury: Ginalski K, Elofsson A, Fischer D, Rychlewski L. "3D-Jury: a simple approach to improve protein structure predictions." Bioinformatics. 19(8):1015-8; 2003.

[7] Bowie J.U., Luthy R., Eisenberg D.: A method to identify protein sequences that fold into a known 3-D Structure Science (pp. 164-170), 1991.

[8] Casey Kochmer , Geert Van Damme: Professional JSP: Second Edition JSP and XML (Chap. 12) Debugging JSP and Servlets (Chap.19), 2001.

[9] David M. Geary: Designing scalable and extensible Web Applications Advanced JavaServer Pages (chap. 5), 2001

[10] David W. Mount: Bioinformatics: Sequence and Genome Analysis. Multiple Sequence Alignment (pp. 169-221) Sequence database searching for similar sequences (pp. 230-275) Introduction to Probability and Statistical Analysis of Sequence Alignments (pp. 122 129) Protein Classification and Structure prediction (pp.411-490) Bioinformatics Programming using Perl and Perl modules (pp. 549-595), 2005.

[11] Dayhoff M.O, Baker W.C., Hunt L.T. Establishing homologies in protein sequences Comparison of commonly used methods. J Mol Evol. (pp. 21:112-125) , 1983.

[12] Durbin R., Eddy S., Krogh A, and Mitchison G : Biological Sequence Analysis – Probabilistic Models of Proteins and Nucleic Acids. Cambridge, UK: Cambridge University Press., 1998.

[13] Eddy S. : HMMER: Profile Hidden Markov Models for biological Sequence Analysis. An implementation of profile HMM methods for sensitive database searches using multiple sequence alignments as query. [available at: http://hmmer.wustl.edu/], 2001.

[14] Efron B. Bootstrapping methods. Another look at the jackknife. Ann. Stat. (pp. 1- 26), 1979.

[15] Eisenhaber F., Imperiale F., Argos P., Froemmel C.: Prediction of Secondary Structural Content of Proteins from Their Amino Acid Composition Alone. New Analytic Vector Decomposition

Methods Proteins: Struct., Funct., Design, N2,157-168, 1996.

[16] Eisenhaber F., Froemmel C., Argos P.: Prediction of Secondary Structural Content of Proteins from Their Amino Acid Composition Alone. The Paradox with Secondary Structural Class Proteins: Struct., Funct., Design, N2, 169-179, 1996.

[17] Eisenhaber F., Persson B., Argos P.: Prediction of Protein Structure. Recognition of Primary, Secondary, and Tertiary Structural Features from Amino Acid Sequence Critical Reviews in Biochemistry & Molecular Biology, N1, 1-94, 1995.

[18] Fischer D., Eisenberg D.: Protein fold recognition using sequence derived predictions Protein Science (p.5, pp. 947-55), 1996.

[19] Frank Eisenhaber and Federica Imperiale: Secondary Structural Content of Proteins from their Amino Acid composition [available at: http://coot.embl.de/SSCP//sscp_seq.html], 1995.

[20] Gasteiger E., Hoogland C., Gattiker A., Duvaud S., Wilkins M.R., Appel R.D., Bairoch A.: Protein Identification and Analysis Tools on the ExPASy Server, 2005.

[21] Gibart J.F., MAdej T., Bryant S.H.: Surprising similarities in Structure composition Current Opinion Structural Biology (p.6, pp. 377- 85), 1996.

[22] H. Nakashima, K. Nishikawa and T. Ooi: The folding type of a protein is relevant to the amino acid composition. Biochem.99(1986), pp. 153–162, 1986.

[23] Hagen J.B.: the origin of Bioinformatics Nat Rev. Genet I. (pp. 231-36), 2000.

[24] Harshawardhan P. Bal : Bioinformatics, Principles and Applications Web based Sequence Analysis: HMMER. pp:88-106, 2005.

[25] Hirst J.D. and Sternberg. M.J.E.: Prediction of structural and functional features of proteins. Structure prediction of proteins and nucleic acids by artificial neural networks. Biochemistry. (pp.31:7211-18), 1992.

[26] John M. Walker (ed): The Proteomics Protocols Handbook, Humana Press. (pp. 571-607), 2005.

[27] Kenneth Litwak: Pure Java 2 Exceptions (pp.189-191) Java.io (pp.541-580) Java.util (pp. 701- 720), 1999.

[28] Kito D. Mann: Java Server Faces in Action JSF fundamentals (pp. 38-45) The Request Processing Lifecycle (pp. 57- 69), 2004.

[29] Ladunga I., Weise B.A., smith R.F.: FASTA:SWAP AND FASTA:PAT Pattern database searches using combinations of aligned amino acids and a novel scoring theory. J Mol. Biol. (259:840-854), 1996.

[30] Lambert C, Leonard N, De Bolle X, Depiereux E. ESyPred3D: Prediction of proteins 3D structures. Bioinformatics.18(9):1250-1256; 2002.

[31] Martin Bond and Debbie Law: Basic principles of JSPs Tomcat Kick Start (Chap. 5), 2003.

[32] Nixon K.C. Carpenter J.M. On outgroups Cladistics (pp. 413-426), 1993.

[33] Ross C., Williams B. and Kay R.F.: Phylogenetic Analysis of anthropoid relationships Journal of Human Evolution. (pp: 35:221-306), 1998.

[34] Rutger Vos, Aki Mimoto, Klaas Hartmann, Jason Caravas: Bio-Phyo Module for perl The base class for the Bio::Phylo package for Phylogenetic analysis using object-oriented perl5. [available at: http://search.cpan.org/~rvosa/Bio-Phylo/], 2005.

[35] Schwede T., Kopp J., Guex N.,Peitsch M. C.: Swiss Model: An Automated protein homology modeling Server. Nucl. Acids Res. (31,3381-5), 2003.

[36] Susan Costantini, Giovanni Colonna and Angelo M. Gacchiano: Amino acid propensities for secondary structures are influenced by the protein structural class., 2006.

[37] Thompson JD, Gibson TJ :CLUSTALW Nucleic Acids Research.Pp:22:4673- 4680.,1994

[38] Thomas Williams, Colin Kelley: Gnuplot An Interactive Plotting Program to visualize Mathematical functions and data. [available at: http://www.gnuplot.info/], 2007.

[39] Waterman M.S., Smith T.F., Beyer W.A. Some biological sequence matrices. Adv. Maths (pp. 20:367-87), 1976.

[40] Weston P.H.: Methods for rooting cladistic trees Models in Phylogeny Reconstruction (pp. 125-155), 1994.