BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Hypothesis

Volume 9(15)

Exploration of freely available web-interfaces for comparative homology modelling of microbial proteins

Vijay Nema[1]* & Sudhir Kumar Pal[2]

[1]Microbiology and Clinical Pathology, 73, G-Block, MIDC, Bhosari, Pune, Maharashtra-411 026, India; [2]Dr. DY Patil University, Plot no.-50, Sector-15, CBD Belapur, Navi Mumbai-400614, India; Vijay Nema – Email: vnema@nariindia.org; Tel: +91-2733 1200, Fax-+91-2027121071; *Corresponding author

Received August 05, 2013; Accepted August 07, 2013; Published August 28, 2013

Abstract:
Aim: This study was conducted to find the best suited freely available software for modelling of proteins by taking a few sample proteins. The proteins used were small to big in size with available crystal structures for the purpose of benchmarking. Key players like Phyre2, Swiss-Model, CPHmodels-3.0, Homer, $(PS)^2$, $(PS)^2$-V2, Modweb were used for the comparison and model generation. **Results:** Benchmarking process was done for four proteins, Icl, InhA, and KatG of *Mycobacterium tuberculosis* and RpoB of *Thermus Thermophilus* to get the most suited software. Parameters compared during analysis gave relatively better values for Phyre2 and Swiss-Model. **Conclusion:** This comparative study gave the information that Phyre2 and Swiss-Model make good models of small and large proteins as compared to other screened software. Other software was also good but is often not very efficient in providing full-length and properly folded structure.

Keywords: *Mycobacterium tuberculosis*, Proteins, Homology modelling, Freeware, Benchmarking.

Background:
To rationally develop new anti-infective agents, it is essential to study the genetics and physiology of microbes. The main cause of hindrance in the designing of new drugs rationally, is the lack of structural information of some very well known targets of drugs for example RpoB protein in *M. tuberculosis*.

The literature has good amount of data available about the *in-vitro* identification of mutations in various microbial genes and the data have been precisely correlated with the phenotypic expression of resistance. However, unavailability of crystal structures of some proteins makes it difficult to predict and understand the exact impact of the mutation on structural changes and binding of drugs or other inhibitors. The only way to understand its three dimensional structure and the related properties to be used in understanding protein-ligand binding is to model the protein with *in-silico* approaches using amino acid sequences as starting point. There are numerous tools available for this purpose. Some very good tools are too costly and for most of the scholars affordability is always a concern. Online or freely available tools generated by academia have been a great help for such researchers. This study aims to explore and compare a few freely available on-line tools. As per the reviewed literature, software chosen for the comparison are top in the list of available free software for homology modelling of protein structure. The information about the best software would allow resource limited settings to still move forward with their studies while saving cost, time and effort. In order to fulfil the aim of this study, structure generation of proteins with known crystal structures was attempted and it was tried to find the best online homology modelling server.

In this study, the proteins chosen for benchmarking experiment were different in nature. One was Icl protein of *M. tuberculosis*

and other was the RpoB protein of *Thermus Thermophilus*. Since Icl is a small protein of *M. tuberculosis* and RpoB is a long protein of different microorganism, a range of proteins was covered in this experiment. The other two proteins InhA and KatG from *M. tuberculosis* with sizes smaller and larger than the tested proteins were taken in order to confirm and validate the results obtained.

Software:

Phyre2

Protein Homology/analogy Recognition engine 2 (PHYRE2) is a free online homology modelling server **[1, 2].** Phyre2 uses the alignment of hidden Markov models via HHsearch to significantly improve accuracy of alignment and detection rate. Phyre2 also incorporates a new *ab-initio* folding simulation called Poing to model those regions of proteins in question which have no detectable homology to known structures **[3].** Poing is also used to combine multiple templates. Distance constraints from individual models are treated as linear elastic springs. Poing then synthesises entire protein in the presence of these springs and at the same time models unconstrained regions using its physics simulation **[1].**

Swiss-Model

It is a web-based integrated service dedicated to protein structure homology modelling. A personal working environment is provided for each user where several modelling projects can be carried out in parallel. Protein sequence and structure databases necessary for modelling are accessible from the workspace and are updated in regular intervals. Tools for template selection, model building and structure quality evaluation can be invoked from within the workspace **[4].**

CPHmodels-3.0

CPHmodels-3.0 is a web-server predicting 3D-structure of protein by use of single template homology modelling. The server employs a hybrid of the scoring functions of CPHmodels-2.0 and a novel remote homology-modelling algorithm. The web server is available at http: //www.cbs.dtu.dk/services /CPHmodels/ **[5].**

ModWeb

ModWeb is a server for automated comparative protein structure modelling (http://salilab.org/modweb). It accepts one or many sequences in the FASTA format and calculates models for them based on the best available template structures from the Protein Data Bank. The structural templates used to build models in ModPipe consist of a set of non-redundant chains extracted from structures in the PDB **[6].** Sequence-structure matches are established using multiple variations of sequence-sequence, profile-sequence, sequence-profile and profile-profile alignment methods. Significant alignments (E-value better than 1.0) covering at least 30 amino acid residues are selected for modelling. Models are built for each one of the sequence-structure matches using comparative modelling by satisfaction of spatial restraints as implemented in Modeller **[7].** The resulting models are evaluated using several model assessment schemes like MTALL (Training set is based on the template structure), MSALL (Training set is based on similar secondary structure), RMSD (Predicted RMSD), etc and the best scoring models are returned to the user **[8].**

(PS)²

(PS)² is an automated homology modelling server **[9].** The method uses an effective consensus strategy by combining PSI-BLAST **[10],** IMPALA **[11]**, and T-Coffee **[12]** in both template selection and target-template alignment. MODELLER **[13],** the modelling package, is used to built the final three dimensional structure. The PROCHECK program, after generating a predicted model with no other refinements, was used to evaluate the quality of this model based on the G-factor. Finally, the predicted model is displayed by Chime and automatically sent to users **[9].** The main drawback of PS2 server is that it cannot process the sequence more than 800 amino acid.

(PS)²-V2

(PS)²-v2 is the advanced version of (PS)², an automatic homology modelling server **[9].** The method uses a new substitution matrix called S2A2. This is a 60x60 substitution matrix based on secondary structure propensities of 20 amino acids (aa). It is a handy tool for the detection of remote homologous and target-template alignment. The final 3-D structure is built using the modelling package MODELLER. After generated a model, the programs ProQ and ProQres are used to evaluate the quality of this model based on a number of structural features predicts the quality of a protein model and to find correct models in contrast to other methods which are optimized to find native structures **[14].** Finally, the predicted model was displayed using software for molecular visualization (AstexViewer) and automatically sent to users.

Homer

Homer (HOmology ModellER) is a comparative modelling server for protein structure prediction. It builds a model structure from an alignment (in FASTA format) and a single template structure (PDB format). Homer performs loop modelling and side-chain optimization on request. The program utilizes FRST to generate the model and a per-residue energy profile. Scoring functions are widely used in the final step of model selection in protein structure prediction. A novel combination of four knowledge-based potentials recognizing different features of native protein structures is introduced and tested. The pairwise, solvation, hydrogen bond, and torsion angle potentials contain largely orthogonal information. Of these, the torsion angle potential is found to show the strongest correlation with model quality. Combining these features with a linear weighting function, it was possible to construct a robust energy function capable of discriminating native-like structures on several benchmarking sets.

Methodology:

Common parameters for comparison

The Expectation value or Expect value (E value) represents the number of different alignments with low or better scores that is expected to occur in a database search by chance. The lower the Expectation value, the more significant the score and the alignment are. Phyre2 does not provide E-value.

Sequence identity and *template selection* are other important parameters to begin with. Sequence identity is the extent to which two sequences have the same residues at the same positions when aligned together. Sequence coverage shows how much sequence is covered for generating the model. Template selection and the resolution of template are very important.

Templates having finer resolution (generally <2.0Å) are treated to be good templates. Better template results in better alignment and finally better three dimensional structures.
Other important parameters which were included for comparison were *RMS deviation* of the modelled structure with the template used and second one was the *energy of the structure* obtained calculated by Swiss PDB Viewer **[4].**

Icl modelling and benchmarking
The Icl protein of *M. tuberculosis* [Uniprot-Id P0A5H3] was selected for testing the software followed by the benchmarking process. Isocitrate lyase (Icl) has 429aa and a good quality crystal structure [PDB-ID 1F8MA] having a resolution of 1.8Å. Models of Icl were generated using all the above mentioned software. No further processing of modelled structures was done. The models generated by different software were compared among themselves with respect to certain parameters as provided in the software (default parameters). For benchmarking, the original structure with PDB code 1F8M was taken and all the structures obtained through modelling were superimposed on it.

RpoB benchmarking
In-silico structures of RpoB protein of *T. Thermophilus* [(PDB-ID 1IW7C, 2.60Å) & 1119aa] was modelled through the same procedure as like Icl.

Validation of the method by applying them on other two proteins
To validate the results obtained from Icl and RpoB benchmarking, two new proteins of *M. tuberculosis* were chosen whose crystal structures are available. First is Enoyl-[acyl-carrier-protein] reductase [NADH] InhA protein [Uniprot-Id P0A5Y6] [PDB-ID 3OEW, 2.20 Å] and second is Catalase-peroxidase KatG protein [Uniprot-Id Q08129**]** [PDB-ID 2CCA, 2.00 Å]**.** Crystal structures of these proteins are available with RCSB-IDs as of 3OEW and 1SJ2 respectively. InhA protein is 269aa long and KatG is 740aa long.

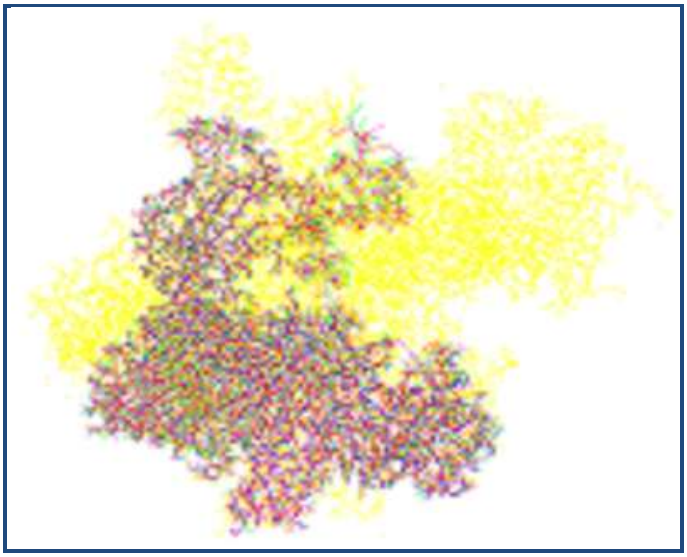

Figure 1: Superimposed figure of all the models of Icl protein of *M. tuberculosis* generated by different software phyre2 in orange (brick colour), HOMER in blue, ModWeb in red, CPHmodels in green, PS2 in grey, ps2-v2 in pink, swiss-model in cyan. Crystal structure is in yellow.

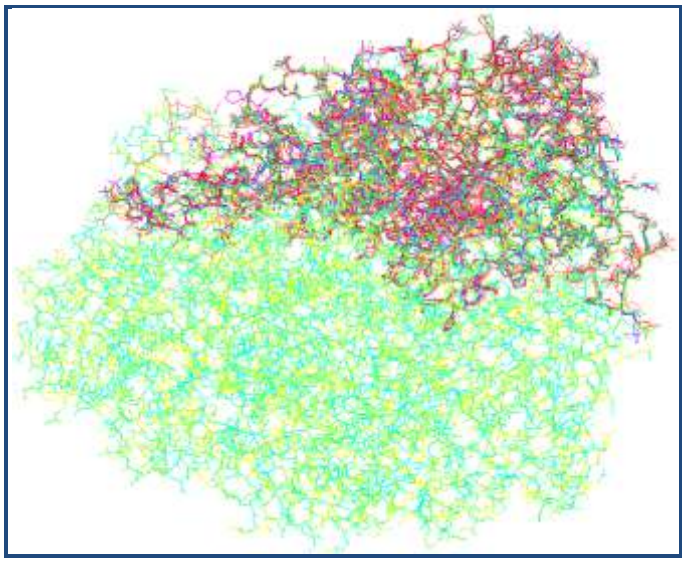

Figure 2: Superimposed figure of all the models of RpoB protein of *Thermus thermophilus* generated by different software phyre2 in grey, HOMER in blue, ModWeb in pink, CPHmodels in red, ps2-v2 in green. Crystal structure is in yellow.

Results:
Modelling and benchmarking
Icl: In Icl modelling 1F8M, 1F8I, 1F61 templates were selected for model generation. All these were structures of Icl protein of *M. tuberculosis* with or without ligand. This showed that all the software work well in opting the right and best template for model generation. The structures were modelled by Phyre2, Swiss-Model, CPHmodels-3.0, Homer, ModWeb, $(PS)^2$, and $(PS)^2$-V2. Different software picked different templates based on various selection methods. Some software like Swiss-Model take one letter amino acid code as the query sequence and other software like CPHmodels-3.0 take protein sequence in FASTA format. The detailed results are provided in **Table 1 (see supplementary material)** and the models generated with different software were superimposed to see the structural similarity **(Figure 1).** Structures generated by $(PS)^2$ and $(PS)^2$-V2 are absolutely same with an RMS deviation of 0.0Å. All the figures of generated models by individual software are presented in the supplementary files. The sequence of protein of Icl is 429aa long. The model generated by Phyre2 and Swiss-Model is of 429aa and the rest software make 427aa long structure. Homer generated 425aa long model. The templates picked by all software were same as like the crystal structure and hence the results of modelling obviously explain the benchmarking success.

RpoB: RpoB protein of *T. Thermophilus* is modelled using the same software as used for Icl. **Table 1** shows the results obtained. For RpoB benchmarking process, different software selected different templates for homology modelling. 1YNJ (Taq RNA polymerase) and 1SMY (Thermus thermophilus RNA polymerase) were selected by Phyre2, 2CW0M (Crystal structure of *T. Thermophilus* RNA polymerase holoenzyme at 3.3Å resolution) used by Swiss-Model, 1IW7C (Crystal structure of the RNA polymerase holoenzyme from *T. Thermophilus* at 2.6Å resolution) by CPHmodels-3.0, and 2A6HC (Crystal structure of the *T. thermophilus* RNA polymerase holoenzyme in complex with antibiotic sterptolydigin) by ModWeb. Homer

software covered 1117aa out of 1119 for model generation. All other software covered the whole sequence in generating the model. All the generated structures were superimposed to the crystal structures of the protein and also to each other **(Figure 2)** and RMS values were recorded **(Table 1).** Structure generated by Swiss-Model using template 2CW0M showed RMS deviation of 0Å.

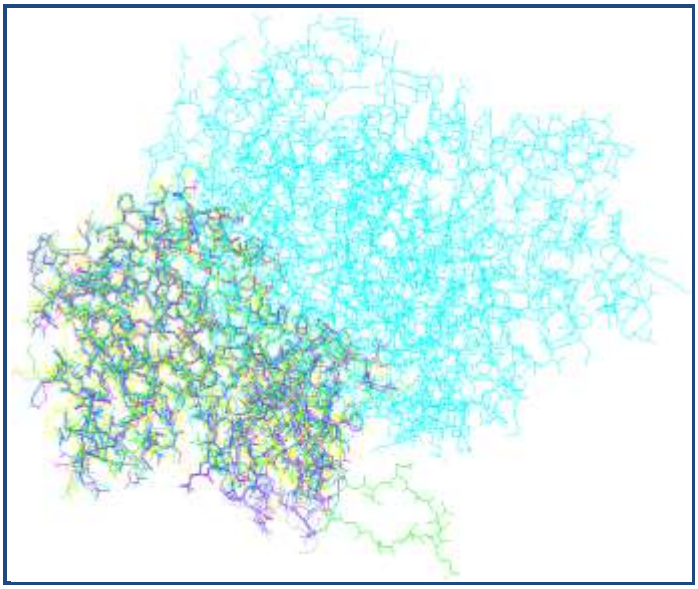

Figure 3: Superimposed figure of all the models of InhA protein of *M. tuberculosis* generated by different software phyre2 in blue, Homer in yellow, ModWeb in green, CPHmodels in grey, $(PS)^2$ in pink, $(PS)^2$-v2 in red, swiss-model in cyan. Crystal structure is in brick red.

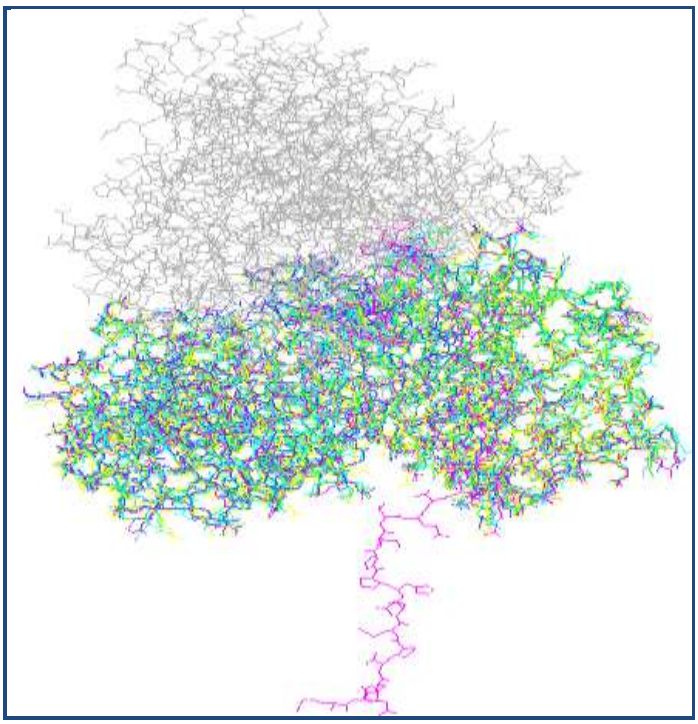

Figure 4: Superimposed figure of all the models of KatG protein of *M. tuberculosis* generated by different software phyre2 in blue, Homer in yellow, ModWeb in green, CPHmodels in cyan, $(PS)^2$ in pink, $(PS)^2$-v2 in red, swiss-model in grey. Crystal structure is in brick red.

InhA: Results of InhA benchmarking were as like the results of Icl and RpoB benchmarking **(Table 1).** In the model generation of inhA (269aa), every software used different template to generate the final three dimensional model. Phyre2 covered the whole sequence. Other software left some residues while generating the model. The E-value obtained by Swiss-Model, CPHmodels-3.0, PS2, and PS2-V2 is a non zero value. Superimposition of models generated and crystal structure provided the information regarding the difference between the obtained structures from different software **(Figure 3).** The figure revealed that the model generated by ModWeb had a loop protruding outside the superimposed core structure whereas all other software made similar structure alike crystal structure.

KatG: Modelling of KatG protein of *M. tuberculosis* having sequence length of 740aa, showed varied results **(Table 1)**. Except Phyre2 and Swiss –model none other software covered the whole sequence for model generation. E-value of this model given by Swiss-Model is 0.00 e^{-1}. Other software except Phyre2 gave 0 as the E-value. . All the generated structures were superimposed to each other and differences are recorded **(Figure 4).** The figure revealed that the model generated by $(PS)^2$ had a long loop protruding outside the superimposed core structure whereas all other software made similar structure as the crystal structure.

Discussion:

Proteins with available crystal structures helped finding the best software to be used for modelling of proteins with crystal structure. The default parameters of all the software were chosen so as to compare them without user-biasness. However, the option for using advanced parameters as per requirement and skills is always available. The main idea behind the benchmarking process with the help of four different proteins was to judge the software quality to a next level. This step was to decipher whether these software work well with short as well as long sequences.

From Icl benchmarking process it was deciphered that all the software work well for shorter sequences and provide enough fair results. In Icl benchmarking, Phyre2 and Swiss-Model covered the whole sequence whereas other software left some residues either because of the unavailability of the correct residue profile or to make a smaller but better conformation having good resolution. Although Phyre2 picked up the right template and covered the whole sequence, there is a reportable amount of deviation in the RMS value between *in-silico* structure and the template used. It was observed that when crystal structure of a protein sequence is present, Swiss-Model could pick it up and made the structure accordingly with low RMS deviation. In case of non-availability of crystal structure and hence no proper template, software (Swiss-Model) made model of higher RMS deviation. This is because the template selected is of different kind and the alignment is not expected to be the same.

To confirm if this software also provide good result for long chain protein sequences, same methodology of model generation was applied on the RpoB protein of *T. Thermophilus* with Uniprot-ID 1IW7 using chain C. In case of Icl modelling templates selected were more or less same. However, in case of

RpoB modelling there was a variety in template selection. The reason of this was that sequence-structure alignment might not be good. On performing a protein-protein blast on the basis of non redundant or ref-sequence database, 4GZZ (Crystal structures of bacterial RNA Polymerase paused elongation complexes) at 4.29Å resolution was found as the best hit.

The decision of choosing the best online software is also validated with the use of two more proteins one is small protein InhA (269aa) and second is a relatively long chain protein, KatG (740aa). Both the proteins are of *M. tuberculosis* having crystal structure information in PDB database. **Table 1** also shows the results of InhA and KatG protein modelling. For InhA, all software gave better results as we got in the case of Icl but in the case of KatG same problem of incomplete sequence coverage is encountered by almost all software as we got in RpoB protein. As the length of protein increases, it becomes difficult for automated software to find the right and best suited template. The alignment also becomes tedious to perform. This problem was taken care by Phyre2 that uses multiple templates to cover the whole sequence and search better folding patterns to make the complete three dimensional structure of the protein. This results in increased RMS deviation but overall quality of structure is highly reliable. Modelled structure can be used for further analysis after post modelling modifications.

Disclosure statement:
No competing financial interests exist.

References:

[1] Kelley LA & Sternberg MJ, *Nat Protoc.* 2009 **4:** 363 [PMID: 19247286]
[2] Soding J, *Bioinformatics.* 2005 **21**: 951 [PMID: 15531603]
[3] Malik IA *et al. J Pak Med Assoc.* 1993 **43**: 118 [PMID: 8411614]
[4] Guex N & Peitsch MC, *Electrophoresis.* 1997 **18**: 2714 [PMID: 9504803]
[5] Nielsen M *et al. Nucleic Acids Res.* 2010 **38**: W576 [PMID: 20542909]
[6] Eswar N *et al. Nucleic Acids Res.* 2003 **31**: 3375 [PMID: 12824331]
[7] Sali A & Blundell TL, *J Mol Biol.* 1993 **234**: 779 [PMID: 8254673]
[8] Eramian D *et al. Protein Sci.* 2008 **17**: 1881 [PMID: 18832340]
[9] Chen CC *et al. Nucleic Acids Res.* 2006 **34**: W152 [PMID: 16844981]
[10] Schaffer AA *et al. Nucleic Acids Res.* 2001 **29**: 2994 [PMID: 11452024]
[11] Schaffer AA *et al. Bioinformatics.* 1999 **15**: 1000 [PMID: 10745990]
[12] Notredame C *et al. J Mol Biol.* 2000 **302**: 205 [PMID: 10964570]
[13] Marti-Renom MA *et al. Annu Rev Biophys Biomol Struct.* 2000 **29**: 291 [PMID: 10940251]
[14] Wallner B & Elofsson A, *Protein Sci.* 2006 **15**: 900 [PMID: 16522791]

Edited by P Kangueane
Citation: **Nema & Pal,** Bioinformation 9(15): 796-801 (2013)

BIOINFORMATION

open access

Supplementary material:

Table 1: The values of the evaluated parameters as provided by different software upon modelling of Icl, RpoB, InhA and KatG

		Parameters									
Software	**Protein**	**Query length (aa)**	**Model length (aa)**	**Template selected**	**E-value**	**Sequence identity (%)**	**Sequence coverage (%)**	**Resolution of template selected (Å)**	**RMS Deviation of modelled Vs target structure (Å)**	**Energy calculated by SPDBV (KJ/mol)**	**Method**
PHYRE2	Icl	429	429	1F61A	-	100	100	2	4.48	-7968.9	HMM via Hhsearch
	RpoB	1119	1119	1YNJC & 1SMYC	-	100	100	3.20 & 2.70	2.78 & 2.37	-12830.9	HMM via Hhsearch
	InhA	269	269	2H7MA	-	100	100	1.62	0.24	-2476.3	HMM via Hhsearch
	KatG	740	740	2CCAA	-	100	100	2	2.23	-6123.4	HMM via Hhsearch
SWISS-MODEL	Icl	429	429	1F8MB	0	100	100	1.8	0.08	-21726.6	QMEAN
	RpoB	1119	1119	2CW0M	0	100	100	3.3	0	-	QMEAN
	InhA	269	2-269	2H7IA	8.67 e-147	100	99.62	1.62	0.07	-11015.3	QMEAN
	KatG	740	740	2CCAB	0.00 e-1	100	100	2	0.34		QMEAN
CPHmodels-3.0	Icl	429	427	1F8MA	0	100	99.5	1.8	0.5	-19588.8	PDB Blast
	RpoB	1119	1119	1IW7C	0	100	100	2.6	1.12	-49351.4	PDB Blast
	InhA	269	268	1BVRA	1 e-155	100	99.62	2.8	0.62	-9123.54	PDB Blast
	KatG	740	717	1SJ2A	0	100	96.5	2.41	0.58	-27451.7	PDB Blast
ModWeb	Icl	429	427	1F8MA	0	100		1.8	0.32	-8601.2	MTALL
	RpoB	1119	1119	2A6HC	0	100	100	2.4	0.74	-2416.4	MTALL
	InhA	269	2-269	2B35B	0	100	99.62	2.3	2.41	-1629.7	MTALL
	KatG	740	29-743	2CCAA	0	100	96.48	2		-11099.4	MTALL
PS2	Icl	429	427	1F8MA	-	100	100		0.29	-8236.3	Composition based Stats
	RpoB	Sequence length of more than 800aa are not taken up.									
	InhA	269	268	2H7IA	4 e-74	100	99.62	1.62	0.28	-292.0	Composition based Stats
	KatG	740	715	2CCAA	0	100	96.62	2	2.44	-9201.3	Composition based Stats
PS2-V2	Icl	429	427	1F8MA	-	100	99.53	1.8	0.29	-8236.3	S2A2 substitution matrix
	RpoB	1119	1119	2A6HC	0	100	100	2.4	1.16	-4700.5	S2A2 substitution matrix
	InhA	269	268	2H7IA	1.8 e-23	100	99.62	1.62	0.28	-292.0	S2A2 substitution matrix
	KatG	740	715	2CCAA	0	100	96.62	2	2.44	-9201.37	S2A2 substitution matrix
HOMER	Icl	429	425	1F8IA	-	99.8		2.25	0	-18053.6	ALIGN2RAW, SCWRL
	RpoB	1119	1117	2A68C	-	100	99.82	2.5	0.8	-23782.6	ALIGN2RAW, SCWRL
	InhA	269	265	1QSGG	n/a	100	98.51	1.75	0.98	8149476	ALIGN2RAW, SCWRL
	KatG	740	2-711	2FXHA	-	67.5	95.81	1.9	0.51	exponential	ALIGN2RAW, SCWRL