Protein structure prediction using homology modeling [Edit]

From Learn Bioinformatics course • 13 min read

What are proteins?

Proteins are large biomolecules which are responsible for performing most of the functions within an organisms cells, including responding to stimuli, acting as catalysts for other reactions, transporting molecules from one place to another and performing cell signaling. Just like DNA sequences, protein sequences are *strings of molecules* but unlike DNA sequences, there are *20* different molecules called *amino-acids* that make up protein sequences.

Protein structure

Every *1D protein sequence* string folds into ***3D structures***. These 3D protein structures are determine how a protein responds to various environments and which other molecules it interacts with, and hence is critical in the ability of the protein to perform its functions. The 3D structure of protein is described by providing the *coordinates (x-y-z)* of every *atom* in the protein, in *3D space*.

Determining protein structure

Protein structures can be determined using experimental procedures like *X-ray crystallography* and *Nuclear Magnetic Resonance (NMR)*. However, these techniques are slow and cumbersome, and cannot be applied to all the proteins. Therefore, *high-throughput* computational methods are used to *predict* 3D structures of proteins from sequences.

Homology Modeling

One of most popular computational methods for protein structure prediction is *Homology Modeling*. Homology modeling leverages the property of *evolutionary conservation* of from the same common ancestor (*homology*) tend to have similar 3D structures. In homology modeling, this property of conservation of protein structure is used to predict structures of newly discovered protein sequences whose structures cannot be resolved using traditional experimental methods.

The main idea is that protein sequence with unknown structure is searched against the sequence database of proteins where structures of all proteins are experimentally known and the unknown structure is *modeled* from the *evolutionarily* closest or best matching protein from the database.

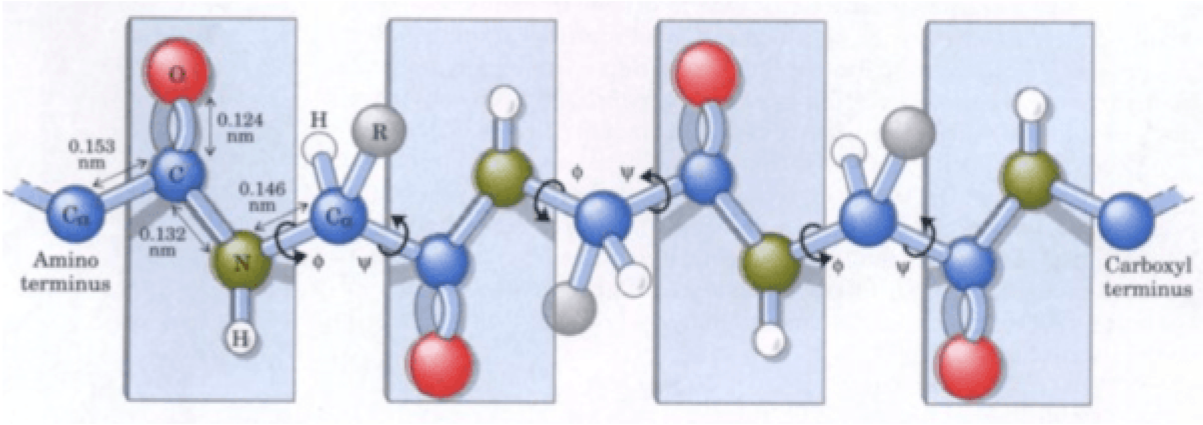
In this article, we describe the approach and methodology of homology modeling, i.e. how it works. We also describe how to use the SWISS-MODEL tool for performing homology modeling.

Detailed description of Homology Modeling method

In this section, we will provide an overview of the steps involved in homology modeling. Note that a number of these steps are active areas of research.

An mentioned previously, homology modeling starts with knowledge of the structure of a number of proteins and their sequences which have been determined by experimental methods. The method uses this previous knowledge to predict the structure of proteins for which we know the sequence, but don't yet know the 3D structure.

For predicting the structure of the protein, we'll first predict the co-ordinates of N, C_a, C_b (backbone) and then the co-ordinates of the R-group (side-chain) of each amino acid.



Step 1: Template recognition and initial alignment

First we find the evolutionarily closest proteins to the target (protein for which we want the predict the structure). This is achieved using database search algorithms like BLAST (Basic Local Alignment Search Tool) which perform sequence alignment of the target sequence against the database of proteins sequences. The PDB (Protein Data Bank) is one such database. The best matching protein sequence from the database, to our target is assumed to be the evolutionarily closest and its structure will be used as a **template** to the model the structure of the target. The database search tool also gives an alignment, i.e. information of which regions of the target match which regions of the template.

Step 2: Alignment correction

The initial alignment between the target and template obtained during the database search may not be optimum in certain difficult regions of the alignments. For example, the initial alignment may violate some rules of amino-acid substitutions like substituting hydrophilic residue with hydrophobic residue in the core of the protein.

Given that we have already found an initial template, we can now use more expensive alignment algorithms to find a better alignment. For example, we can use multiple sequence alignment algorithms for this step. Multiple sequence alignment are useful for identifying regions that are highly divergent, and hence better detecting the appropriate locations for insertions and deletions.

Step 3: Backbone generation

After the target-template alignment is optimized, the protein backbone structure (N-C_a-C_b) for the target is generated. This is achieved by simply copying the coordinates of the template backbone to the target based on the alignment. That is to say, the coordinates of an atom in the target protein is said to be the same as the coordinates of the corresponding atom in the template protein, as suggested by the alignment from the previous step. Obviously, this process is highly dependent on the accuracy of the template structure, and any errors in our initial database will cause errors in our prediction.

The backbone step does not handle the two types of mismatches present in the alignment, insertions and deletions. Incorporating these mismatches into the backbone is the most difficult part in homology modeling.

The secondary structure of the protein consists of helices, strands and loops. Since conformational changes implied by insertions and deletions can't happen in helices and strands, they must happen in the loops.

There are two main approaches to model loops: knowledge based and energy based. The former approach searches for conformations of loops that have similar sequences and endpoints as the target in the database of known structures. The latter models the loop conformation in an ab initio manner by predicting the loop structures with lowest structural energies using force field functions and molecular dynamics. These methods provide fairly accurate results for short loops with up to 5-8 residues.

Step 5: Side-chain modeling

Modeling the side chains involves predicting the value of C_a-C_b torsion angle for each R-group attached to the backbone. The conformation of side-chains in the structures, also called the rotamers, depends on the values on this torsion angle. The side-chain are generally modeled in a knowledge based manner using rotamer libraries which contain preferred conformations for all 20 R-groups under various chemical neighborhoods.

Step 6: Model optimization

Now that all the aspects of protein structure are modeled for the target, it is time make fine changes in the structure to reduce the overall energy. This is achieved in an iterative manner. In each iteration, the backbone conformations and rotamer conformations are changed alternatively to lower the overall energy of the predicted structures.

Model optimization can also be performed by running a molecular dynamics simulation, which starts with the current predicted structure, and makes small changes to the structure based on the simulation, i.e. we simulate what would happen to each of the atoms of the protein under the forces surrounding it on a femtosecond (10^{-15}) timescale.

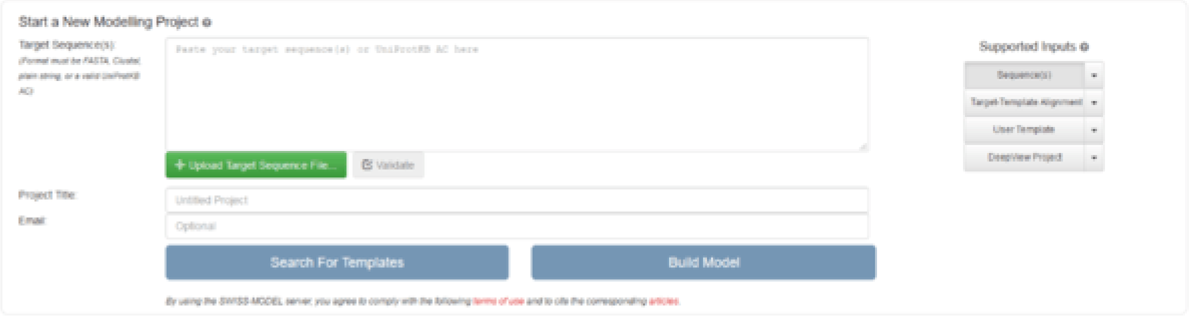
The final step is to check the predicted structure for errors. Errors are introduced in the predicted protein structures due to low alignment between target and template or due to errors in template structures. Checks are performed on the predicted structure to see if all the bond lengths, bond angles and torsion angles fall in characteristic ranges found from experimentally determined protein structures. Energy checks are also performed which see if the different types of structure based energies like Van der Waals and Electrostatic are at expected levels.

Homology modeling using SWISS-MODEL

We will take detailed look at homology modeling procedure by predicting structure for protein *Ornithine carbamoyltransferase* (UniProtKB accession: P96134) present in bacteria *Thermus thermophilus* using the SWISS-MODEL tool.

Target-template recognition

First step is to search the target sequence against the database of sequences with known protein structures. Paste accession number into the window and hit the "Search For Templates" button



Start a New Modelling Project

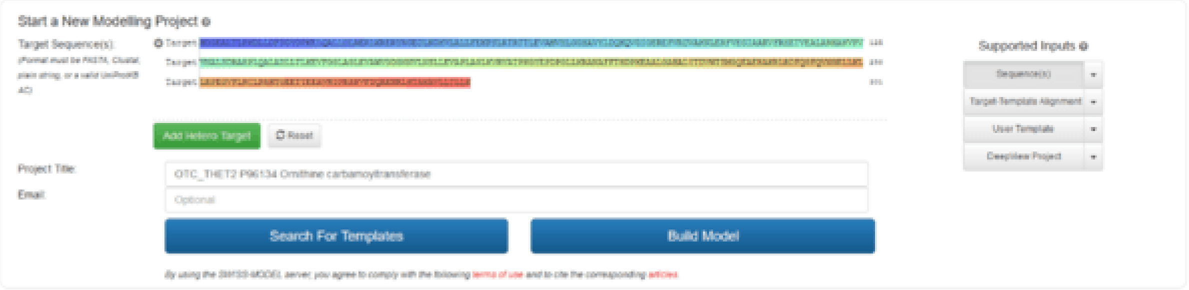
Target Sequence(s): (Format must be FASTA, Clustal, plain string, or a valid UniProtKB AC)

Project Title: Untitled Project

Email: Optional

By using the SWISS-MODEL server, you agree to comply with the following terms of use and to cite the corresponding articles.

Paste your target sequence(s) or UniProtKB AC here
+ Upload Target Sequence File...
Validate
Search For Templates
Build Model
Supported Inputs
Sequence(s)
Target-Template Alignment
User Template
DeepView Project



Start a New Modelling Project

Target Sequence(s): (Format must be FASTA, Clustal, plain string, or a valid UniProtKB AC)

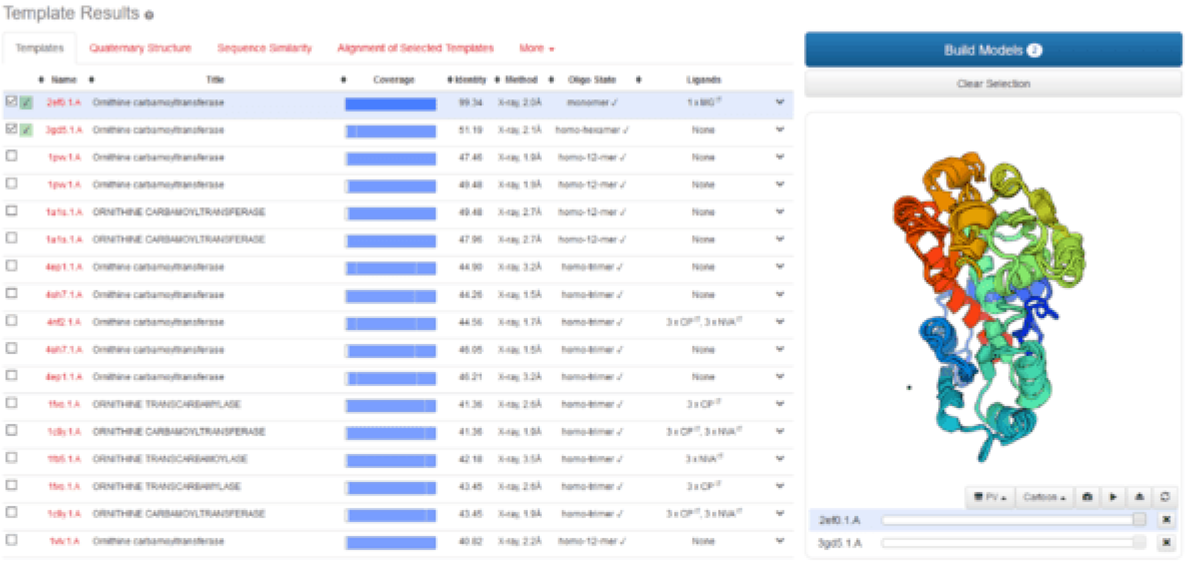
Project Title: OTC_THET2 P96134 Ornithine carbamoyltransferase

Email: Optional

By using the SWISS-MODEL server, you agree to comply with the following terms of use and to cite the corresponding articles.

how well their sequences align with the target protein sequence. The first selected structure template is the best matching (99% identity). The second template matches the target sequence with 51% identity. The superimposed protein structures of the two top matching templates can be seen in the window on the right.

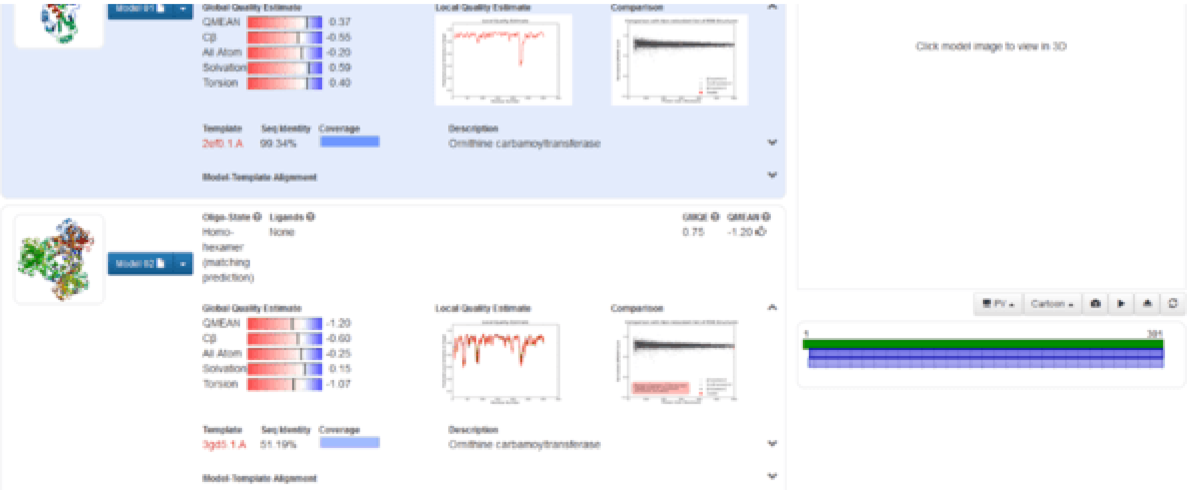
We will use the top two results to build or predict two structures for the target sequence and then select the best predicted structure.



Template Results

	Name	Title	Coverage	Identity	Method	Oligo State	Ligands	
☑ ✓	2ef0.1.A	Ornithine carbamoyltransferase		99.34	X-ray, 2.0Å	monomer ✓	1 x BIG	⌄
☑ ✓	3gd5.1.A	Ornithine carbamoyltransferase		51.19	X-ray, 2.1Å	homo-hexamer ✓	None	⌄
□	1pvv.1.A	Ornithine carbamoyltransferase		47.46	X-ray, 1.9Å	homo-12-mer ✓	None	⌄
□	1pvv.1.A	Ornithine carbamoyltransferase		49.48	X-ray, 1.9Å	homo-12-mer ✓	None	⌄
□	1a1s.1.A	ORNITHINE CARBAMOYLTRANSFERASE		49.48	X-ray, 2.7Å	homo-12-mer ✓	None	⌄
□	1a1s.1.A	ORNITHINE CARBAMOYLTRANSFERASE		47.96	X-ray, 2.7Å	homo-12-mer ✓	None	⌄
□	4ep1.1.A	Ornithine carbamoyltransferase		44.90	X-ray, 3.2Å	homo-trimer ✓	None	⌄
□	4oh7.1.A	Ornithine carbamoyltransferase		44.26	X-ray, 1.5Å	homo-trimer ✓	None	⌄
□	4nf2.1.A	Ornithine carbamoyltransferase		44.56	X-ray, 1.7Å	homo-trimer ✓	3 x CP, 3 x NVA	⌄
□	4oh7.1.A	Ornithine carbamoyltransferase		46.05	X-ray, 1.5Å	homo-trimer ✓	None	⌄
□	4ep1.1.A	Ornithine carbamoyltransferase		46.21	X-ray, 3.2Å	homo-trimer ✓	None	⌄
□	1fvo.1.A	ORNITHINE TRANSCARBAMYLASE		41.36	X-ray, 2.6Å	homo-trimer ✓	3 x CP	⌄
□	1c9y.1.A	ORNITHINE CARBAMOYLTRANSFERASE		41.36	X-ray, 1.9Å	homo-trimer ✓	3 x CP, 3 x NVA	⌄
□	1fb5.1.A	ORNITHINE TRANSCARBAMOYLASE		42.18	X-ray, 3.5Å	homo-trimer ✓	3 x NVA	⌄
□	1fvo.1.A	ORNITHINE TRANSCARBAMYLASE		43.45	X-ray, 2.6Å	homo-trimer ✓	3 x CP	⌄
□	1c9y.1.A	ORNITHINE CARBAMOYLTRANSFERASE		43.45	X-ray, 1.9Å	homo-trimer ✓	3 x CP, 3 x NVA	⌄
□	1vlv.1.A	Ornithine carbamoyltransferase		40.82	X-ray, 2.2Å	homo-12-mer ✓	None	⌄

The modeling results can be seen below. The two predicted structures are ranked according to the quality of their models. Next, we will assess the quality of both the predicted structures to see which is the best.



Global Quality Estimate

QMEAN	0.37
Cβ	-0.55
All Atom	-0.20
Solvation	0.59
Torsion	0.40

Template	Seq Identity	Coverage	Description
2ef0.1.A	99.34%		Ornithine carbamoyltransferase

Oligo-State: Homo-hexamer (matching prediction)
Ligands: None
GMQE: 0.75
QMEAN: -1.20

Global Quality Estimate

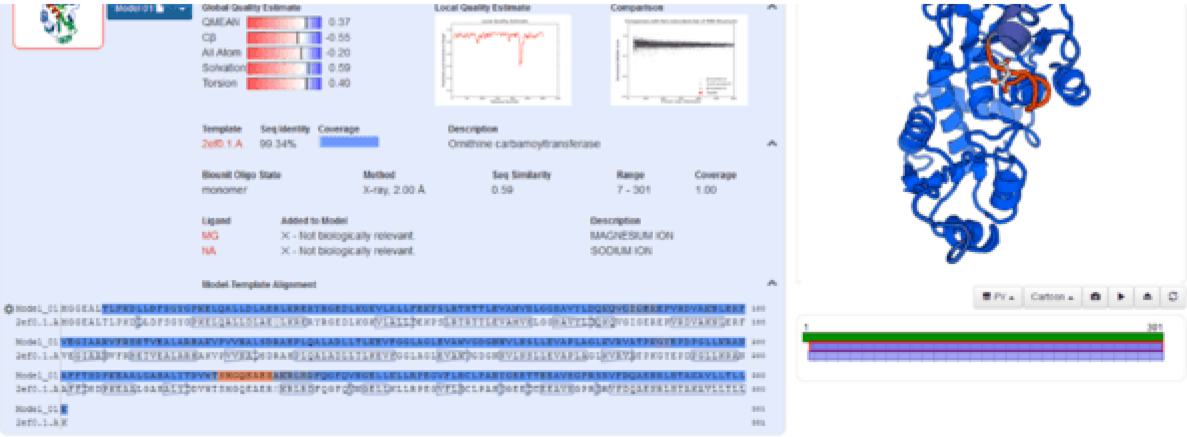
QMEAN	-1.20
Cβ	-0.60
All Atom	-0.25
Solvation	0.15
Torsion	-1.07

Template	Seq Identity	Coverage	Description
3gd5.1.A	51.19%		Ornithine carbamoyltransferase

The QMEAN is one of the primary measures used to assess the model quality. QMEAN is a composite scoring function based on different geometrical properties of protein structures and provide both global (*i.e.*for the entire structure) and local (*i.e.* per residue) absolute quality estimates.

QMEAN consists of four individual terms. The four individual terms of the global QMEAN quality scores are also listed. The white area in the bar-plots (numerical values close to zero) indicates that the property is similar to what is observed in experimental structures. Positive values indicate that the model scores higher than experimental structures on average, negative numbers indicate that the model scores lower than experimental structures on average.

For the first model (built using 2ef0.1.A as a template) the QMEAN terms mostly fall within the white region.



Model 01

Global Quality Estimate

QMEAN	0.37
Cβ	-0.55
All Atom	-0.20
Solvation	0.59
Torsion	0.40

Local Quality Estimate

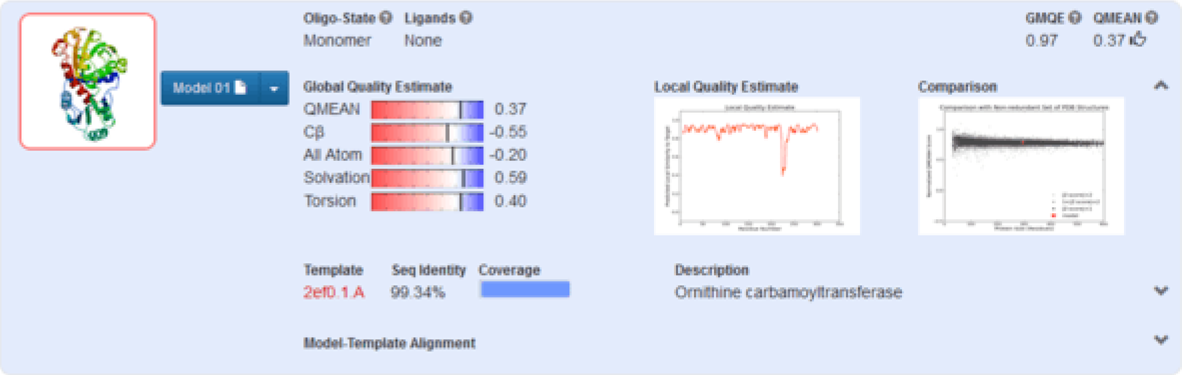
Comparison

Template	Seq Identity	Coverage	Description
2ef0.1.A	99.34%		Ornithine carbamoyltransferase

Biounit Oligo State	Method	Seq Similarity	Range	Coverage
monomer	X-ray, 2.00 Å	0.59	7 - 301	1.00

Ligand	Added to Model	Description
MG	✕ - Not biologically relevant	MAGNESIUM ION
NA	✕ - Not biologically relevant	SODIUM ION

Model-Template Alignment



Oligo-State	Ligands
Monomer	None

GMQE	QMEAN
0.97	0.37

Model 01

Global Quality Estimate

QMEAN	0.37
Cβ	-0.55
All Atom	-0.20
Solvation	0.59
Torsion	0.40

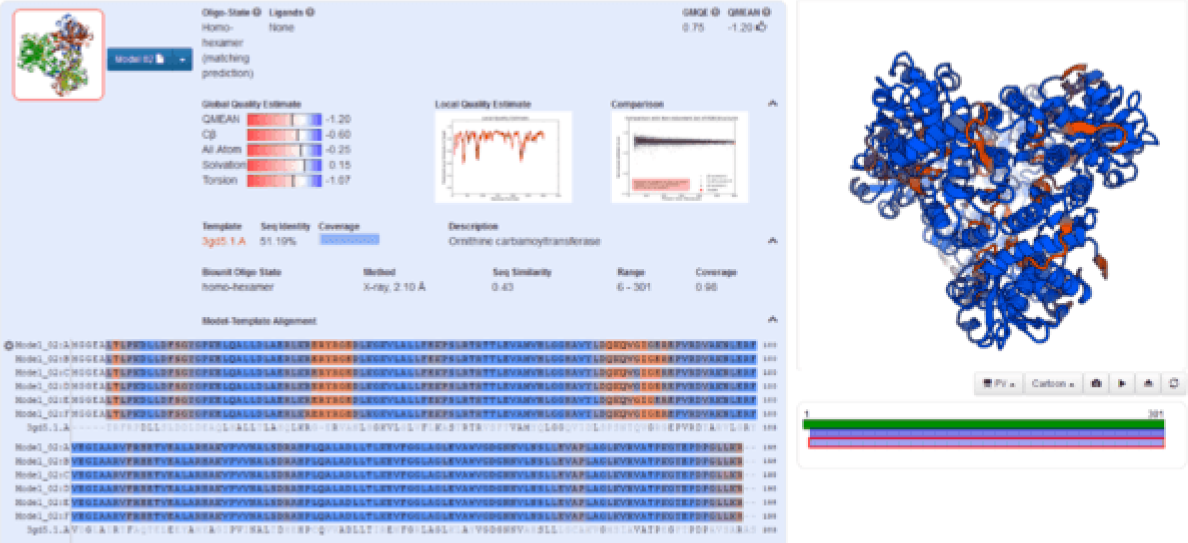
Local Quality Estimate

Comparison

Template	Seq Identity	Coverage	Description
2ef0.1.A	99.34%		Ornithine carbamoyltransferase

Model-Template Alignment

However, for the second model (built using 3gd5.1.A as a template) most QMEAN terms significantly differ from the optimal



Oligo-State	Ligands
Homo-hexamer (matching prediction)	None

GMQE	QMEAN
0.75	-1.20

Model 02

Global Quality Estimate

QMEAN	-1.20
Cβ	-0.60
All Atom	-0.25
Solvation	0.15
Torsion	-1.07

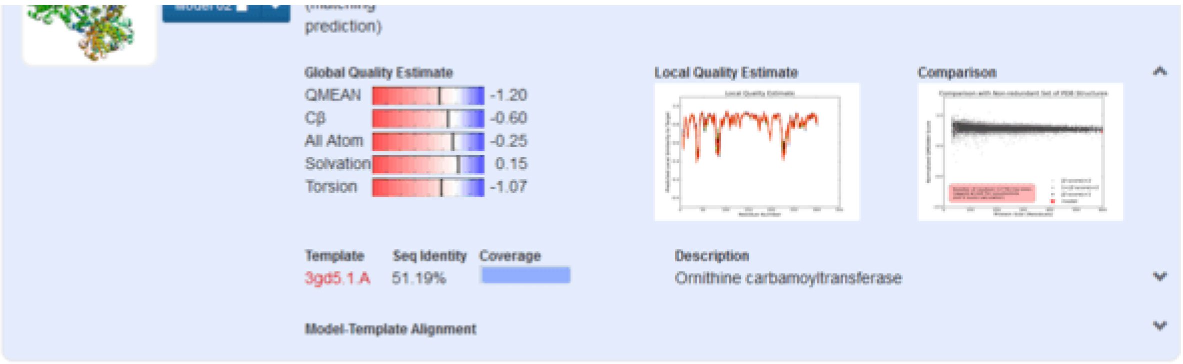
Local Quality Estimate

Comparison

Template	Seq Identity	Coverage	Description
3gd5.1.A	51.19%		Ornithine carbamoyltransferase

Biounit Oligo State	Method	Seq Similarity	Range	Coverage
homo-hexamer	X-ray, 2.10 Å	0.43	6 - 301	0.98

Model-Template Alignment



Model 02

(matching prediction)

Global Quality Estimate

QMEAN	-1.20
Cβ	-0.60
All Atom	-0.25
Solvation	0.15
Torsion	-1.07

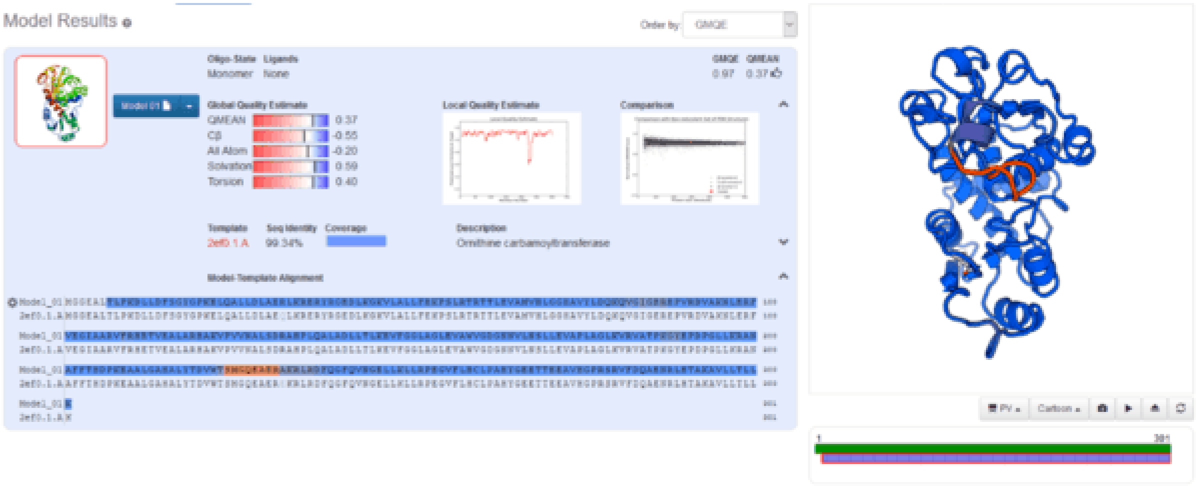
Local Quality Estimate

Comparison

Template	Seq Identity	Coverage	Description
3gd5.1.A	51.19%		Ornithine carbamoyltransferase

Model-Template Alignment

Therefore, the structure predicted by the template 2ef0.1.A is the most optimal model and can be used as the *predicted structure* for our target sequence.



Model Results

Order by: GMQE

Model 01

Oligo-State	Ligands
Monomer	None

GMQE	QMEAN
0.97	0.37

Global Quality Estimate

QMEAN	0.37
Cβ	-0.55
All Atom	-0.20
Solvation	0.59
Torsion	0.40

Local Quality Estimate

Comparison

Template	Seq Identity	Coverage	Description
2ef0.1.A	99.34%		Ornithine carbamoyltransferase

Model-Template Alignment

```
Model_01 [illegible]
2ef0.1.A [illegible]
Model_01 [illegible]
2ef0.1.A [illegible]
Model_01 [illegible]
2ef0.1.A [illegible]
Model_01 [illegible]
2ef0.1.A [illegible]
```

Resources

- SWISS-MODEL
- MODELLER

References

- "Homology Modeling" by Elmar Krieger, Sander B. Nabuurs, and Gert Vriend

Protein structure Swiss-model Homology modeling Protein structure prediction Protein

Category: Biology • Last updated: 2 years ago egories

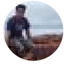
MARK COMPLETED

★★★★★ Leave a Review

‹ PREVIOUS

Detecting Mutations with Read Mapping and Suffix Trees

ABOUT THE CONTRIBUTORS



Akshay Yadav

Masters in Bioinformatics and 6 years of work experience in Bioinformatics & Computational Biology

Keshav Dhandhania

BSc, MSc 2014 @ MIT (AI, Deep Learning). Former Competitive Programmer.

Add a comment

Back to top

POPULAR PATHS & COURSES

Machine Learning

Deep Learning

Natural Language Processing

Big Data

Data Science

Bioinformatics

Algorithms

egories

NEW PATHS & COURSES

Web Development

UX & UI Design

Startups

Product Management

Cryptocurrencies

Finance

College Admissions

ABOUT US

Our Mission

How to Contribute

Help and FAQ

CommonLounge Meta

Team

Privacy, Terms & Refunds

GET IN TOUCH

hello@commonlounge.com

Facebook

Twitter

San Francisco, CA, USA

+1 844 318 7406