

1.2.1 Structure Prediction

- This mini-lecture and tour will be about protein structure prediction
- At the end of this mini-lecture and tour, you should be able to
 - explain, in a general sense, the principles behind protein structure prediction
 - perform structure prediction using the I-TASSER web server

Structure Prediction Principles

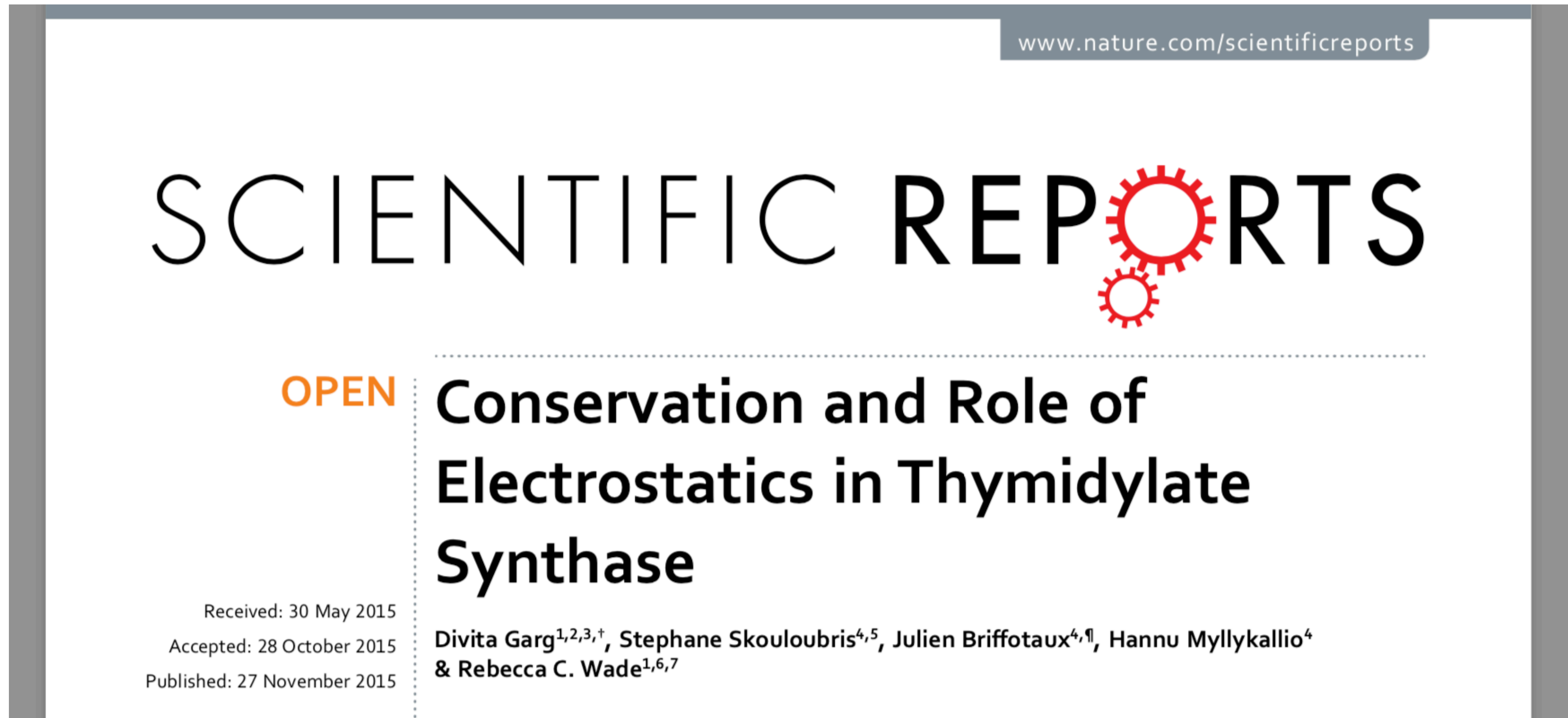
- **Homology modeling** is most useful for closer homology. Go through <http://www.bioinfo.rpi.edu/bystrc/courses/biol4550/lecture7/assets/player/KeynoteDHTMLPlayer.html#0> up to slide 11
- **Threading** is more useful for distant homologues. Watch <https://www.jove.com/video/3259/a-protocol-for-computer-based-protein-structure-function> up through minute 1
- Differences between approaches not completely distinct

Choosing Structure Prediction Software

- There are many software tools for protein structure prediction (see https://en.wikipedia.org/wiki/List_of_protein_structure_prediction_software)
- How should you decide which to use?
 - Ease of use
 - Web server - easier for sporadic use
 - Downloadable and scriptable - easier for large-scale applications
 - Accuracy
- The “Critical Assessment of protein Structure Prediction” (CASP) experiments are *blinded* tests of the ability to predict structure from sequence. (see <http://www.predictioncenter.org/index.cgi>)
- “I-TASSER (as 'Zhang-Server') was ranked as the No 1 server for protein structure prediction in recent community-wide CASP7, CASP8, CASP9, CASP10, CASP11, CASP12, and CASP13 experiments.”

Demonstration

- I will describe how to reproduce key results from this paper:



- Garg *et. al.* created a homology model from a minimal organism W.g.b. To reproduce their result, I needed to find the sequence.
- UniProt (<https://www.uniprot.org>) is a comprehensive biological sequence database
- A search for “thymidylate synthase” yields 50,099 results, including 566 that have been manually reviewed.
- The two results for “thymidylate synthase wigglesworthia” is much more manageable. Select the result that has been reviewed.

UniProt

UniProtKB thymidylate synthase

BLAST Align Retrieve/ID mapping Peptide search

UniProtKB results

UniProtKB consists of two sections:

- Reviewed (Swiss-Prot) - Manually annotated**
Records with information extracted from literature and curator-evaluated computational analysis.
- Unreviewed (TrEMBL) - Computationally analyzed**
Records that await full manual annotation.

The UniProt Knowledgebase (UniProtKB) is the central hub for the collection of functional information on proteins, with accurate, consistent and rich annotation. In addition to capturing the core data mandatory for each UniProtKB entry (mainly, the amino acid sequence, protein name or description, taxonomic data and citation information), as much annotation information as possible is added.

Help UniProtKB help video Other tutorials and videos Downloads

Filter by:

Reviewed (566) Swiss-Prot

Unreviewed (49,533) TrEMBL

Popular organisms

Mouse (14)

A. thaliana (12)

Human (8)

Rice (8)

B. subtilis (5)

Other organisms

Search terms

Filter "thymidylate" as:

BLAST Align Download Add to basket Columns

Quote terms: "thymidylate synthase"
Did you mean to search for thymidylate synthetase

Entry	Entry name	Protein names	Gene names	Organism	Length
<input type="checkbox"/> Q7W1A9	TYSY_BORPA	Thymidylate synthase	thyA BPP0787	Bordetella parapertussis (strain 12822 / ATCC BAA-587 / NCTC 13253)	323
<input type="checkbox"/> C5BMA3	TYSY_TERTT	Thymidylate synthase	thyA TERTU_0365	Teredinibacter turnerae (strain ATCC 39867 / T7901)	277
<input type="checkbox"/> Q8EV81	TYSY_MYCPE	Thymidylate synthase	thyA MYPE6870	Mycoplasma penetrans (strain HF-2)	289
<input type="checkbox"/> A3PYA8	TYSY_MYCSJ	Thymidylate synthase	thyA Mjls_2099	Mycobacterium sp. (strain JLS)	266
<input type="checkbox"/> POC0M5	TYSY_STAES	Thymidylate synthase	thyA SE_1120	Staphylococcus epidermidis (strain ATCC 12228)	318
<input type="checkbox"/> Q6FZ91	TYSY_BARQU	Thymidylate synthase	thyA BQ08630	Bartonella quintana (strain Toulouse) (Rochalimaea quintana)	264
<input type="checkbox"/> A9IW82	TYSY_BART1	Thymidylate synthase	thyA BT_1568	Bartonella tribocorum (strain CIP 105476 / IBS 506)	264
<input type="checkbox"/> Q6MID2	TYSY_BDEBA	Thymidylate synthase	thyA Bd3230	Bdellovibrio bacteriovorus (strain ATCC 15356 / DSM 50701 / NCIB 9529 / HD100)	264

UniProt

UniProtKB thymidylate synthase wigglesworthia

BLAST Align Retrieve/ID mapping Peptide search

UniProtKB results

UniProtKB consists of two sections:

- Reviewed (Swiss-Prot) - Manually annotated**
Records with information extracted from literature and curator-evaluated computational analysis.
- Unreviewed (TrEMBL) - Computationally analyzed**
Records that await full manual annotation.

The UniProt Knowledgebase (UniProtKB) is the central hub for the collection of functional information on proteins, with accurate, consistent and rich annotation. In addition to capturing the core data mandatory for each UniProtKB entry (mainly, the amino acid sequence, protein name or description, taxonomic data and citation information), as much annotation information as possible is added.

Help UniProtKB help video Other tutorials and videos Downloads

Filter by:

Reviewed (1) Swiss-Prot

Unreviewed (1) TrEMBL

Popular organisms

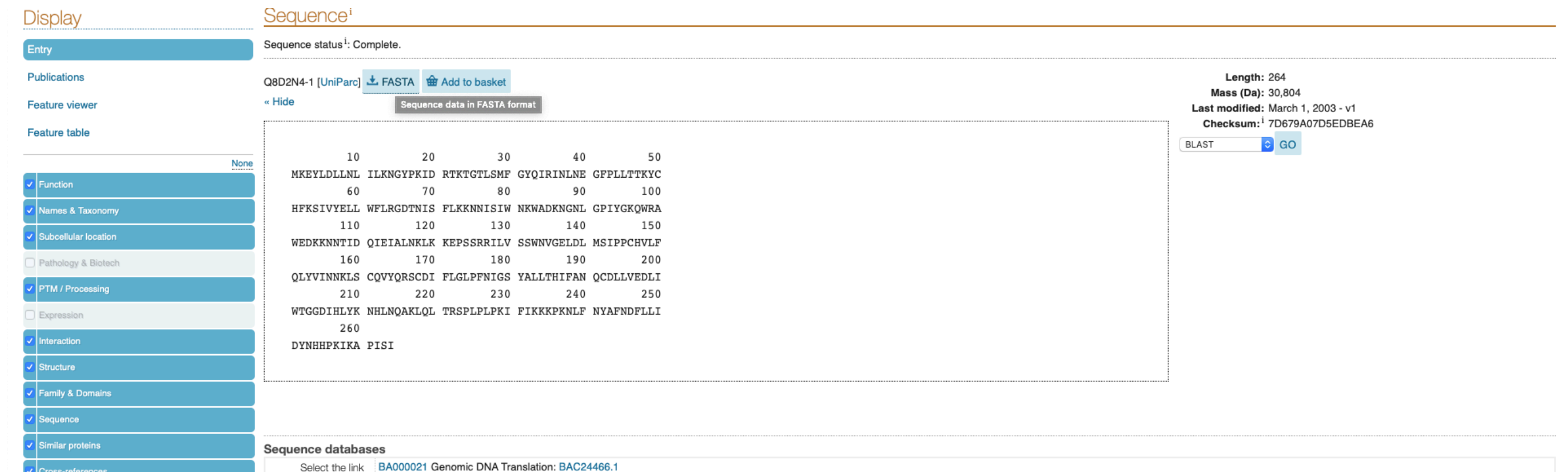
WIGBR (1)

BLAST Align Download Add to basket Columns

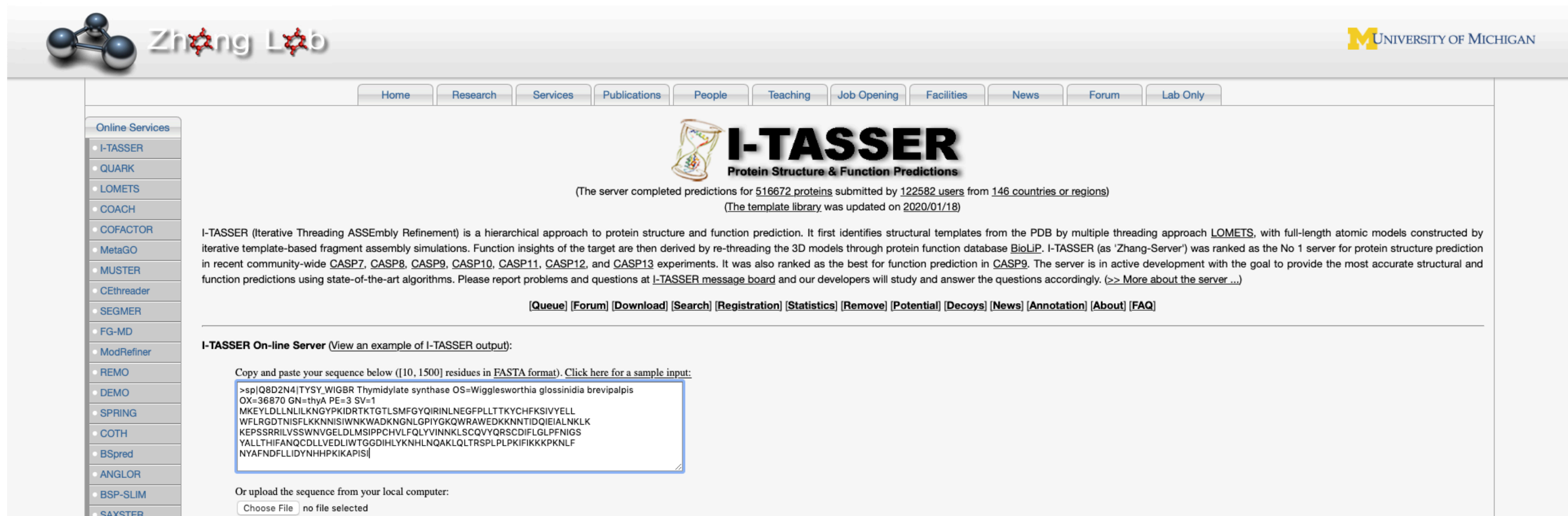
Quote terms: "thymidylate synthase"

Entry	Entry name	Protein names	Gene names	Organism	Length
<input type="checkbox"/> Q8D2N4	TYSY_WIGBR	Thymidylate synthase	thyA WIGBR3200	Wigglesworthia glossinidia brevipalpis	264
<input type="checkbox"/> H6Q4Z7	H6Q4Z7_WIGGL	Thymidylate synthase	thyA WIGMOR_0450	Wigglesworthia glossinidia endosymbiont of Glossina morsitans morsitans (Yale colony)	264

- To get the actual amino acid sequence, click on “Sequence” or scroll down and then click on FASTA. FASTA is a simple format for amino acid sequences based on one-letter codes.
- To run I-TASSER, you can visit the server interface at <https://zhanglab.ccmb.med.umich.edu/I-TASSER/>, register, paste the sequence into the appropriate field, and press “Run I-TASSER”.
- *But* let’s spare their servers and avoid the wait by not all submitting the same job.



The screenshot shows the UniProt entry for Q8D2N4-1. The left sidebar contains a 'Feature table' with various categories like Function, Name & Taxonomy, Subcellular location, etc. The main content area displays the 'Sequence' in FASTA format, with a 'FASTA' button and an 'Add to basket' button. The sequence is shown in a grid format, with columns for residue numbers (10, 20, 30, 40, 50) and the corresponding amino acid codes. The sequence is: MKEYLDLLNL ILKNGYPKID RTKTGTLSPF GYQIRINLNE GFPLLTYYC 60 70 80 90 100 HFKSIVYELL WFLRGDTNIS FLKNNISIW NKWADKNGNL GPIYQKQWRA 110 120 130 140 150 WEDKKNNTID QIEIALNKLK KEPSSRRILV SSWNVGELD MSIPPCHVLF 160 170 180 190 200 QLYVINNKLS CQVYQSCDI FLGLPFNIGS YALLTHIFAN QCDLLVEDLI 210 220 230 240 250 WTGGDIHLVK NHLNQAKLQL TRSPLPLPKI PIKKPKNLF NYAFNDPLLI 260 DYNHPKIKAI PISI. The right sidebar shows metadata: Length: 264, Mass (Da): 30,804, Last modified: March 1, 2003 - v1, Checksum: 7D679A07D5EDBEA6. There is a 'BLAST' button and a 'GO' button.



The screenshot shows the I-TASSER Protein Structure & Function Predictions server interface. The top navigation bar includes links for Home, Research, Services, Publications, People, Teaching, Job Opening, Facilities, News, Forum, and Lab Only. The main content area features the I-TASSER logo and a description of the server's capabilities. It states that the server completed predictions for 516672 proteins submitted by 122582 users from 146 countries or regions. The server is described as a hierarchical approach to protein structure and function prediction, using multiple threading approaches and full-length atomic models. A sidebar on the left lists various online services, including I-TASSER, QUARK, LOMETS, COACH, COFACTOR, MetaGO, MUSTER, CETHREADER, SEGMENT, FG-MD, ModRefiner, REMO, DEMO, SPRING, COTH, BSpreD, ANGLOR, BSP-SLIM, and SAXSTER. The main content area also includes a 'Queue' button and a 'Download' button. A text box for pasting the sequence is shown, with a sample input sequence: >sp|Q8D2N4|TYSY.WIGBR Thymidylate synthase OS=Wigglesworthia glossinidia brevipalpis OX=36870 GN=thYA PE=3 SV=1 MKEYLDLLNLKNGYPKIDRTKTGTLSPFQYQIRINLNEGFPPLTYYCHFKSIVYELL WFLRGDTNISFLKNNISIWKNWADKNGNLGPIYQKQWRAWEDKKNNTIDQIEIALNKLK KEPSSRRILVSSNVGELDMSIPPCHVLFQLYVINNKLSQVYQSCDIFLGLPFNIGS YALLTHIFANQCDLLVEDLIWTGGDIHLVKNHLNQAKLQLTRSPLPLPKIPIKKPKNLF NYAFNDPLLI DYNHPKIKAI PISI. Below the text box is a button to 'Choose File' and a note that no file was selected.

- I've submitted the sequence before. The results are archived at https://ccbatit.github.io/modelingworkshop/S516679_results/

I-TASSER results for job id S516679

(Click on [S516679_results.tar.bz2](#) to download the tarball file including all modeling results listed on this page. Click on [Annotation of I-TASSER Output](#) to read the instructions for how to interpret the results on this page. Model results are kept on the server for 60 days, there is no way to retrieve the modeling data older than 2 months)

Submitted Sequence in FASTA format

```
>protein
MKEYLDLLNLILKNGYPKIDRTKTGTLSPMGYQIRINLNEGFPLLTTRYCHFKSIYVELL
WFLRGDPTNISFLKKNNISIWNKWADKNGNLGPYIGKWRAWEDKKNNITDQIEIALNKLK
KEPSSRRILVSSWNVGELDLMSIPPCHVLFPQLFYVINNKLSCQVYQRSCTIFGLPFNIGS
YALLTHIFANQCDLLVEDLIWTGGDIHLYKNHLNAKLQLTRSPFLPKIPIKPKPNLF
NYAFNDFLIDYNHHPKIKAPISI
```

Predicted Secondary Structure

	20	40	60	80	100	120	140	160	180	200
Sequence	MKEYLDLLNLILKNGYPKIDRTKTGTLSPMGYQIRINLNEGFPLLTTRYCHFKSIYVELLWFLRGDPTNISFLKKNNISIWNKWADKNGNLGPYIGKWRAWEDKKNNITDQIEIALNKLKKEPSSRRILVSSWNVGELDLMSIPPCHVLFPQLFYVINNKLSCQVYQRSCTIFGLPFNIGSYALLTHIFANQCDLLVEDLIWTGGDIHLYKNHI									
Prediction	C#####CCCCCCCCCCCCSSSSCCSSSSSCCCCCCCCCCCCCC#####C####CCC#####CCCCCCCCCCCCSSSSSCCC#####CCCCCCCCSSSSSSCCSSSSSSSS#####C###SSSSSSSSSSSHH#									
Conf. Score	9789999999989087428999746998661589726778743445643067789999999618865999987286632330232676676677556453464999834099999998509997328998868445114569983168899974997899999786777777661599999999999986980148999988789858896									

H:Helix; S:Strand; C:Coil

Predicted Solvent Accessibility

	20	40	60	80	100	120	140	160	180	200
Sequence	MKEYLDLLNLILKNGYPKIDRTKTGTLSPMGYQIRINLNEGFPLLTTRYCHFKSIYVELLWFLRGDPTNISFLKKNNISIWNKWADKNGNLGPYIGKWRAWEDKKNNITDQIEIALNKLKKEPSSRRILVSSWNVGELDLMSIPPCHVLFPQLFYVINNKLSCQVYQRSCTIFGLPFNIGSYALLTHIFANQCDLLVEDLIWTGGDIHLYKNHI									
Prediction	7520151043007414637311110211032031204740000002302030002000000212210430374603001410464231020013001316374420000340630373251000000002172067000000000001035230100000010220110101000000000011060401100000000003201									

Values range from 0 (buried residue) to 9 (highly exposed residue)

(I-TASSER modeling starts from the structure templates identified by LOMETS from the PDB library. LOMETS is a meta-server threading approach containing multiple threading programs, where each threading program can generate tens of thousands of template alignments. I-TASSER only uses the templates of the highest significance in the threading alignments, the significance of which are measured by the Z-score, i.e. the difference between the raw and average scores in the unit of standard deviation. The templates in this section are the 10 best templates selected from the LOMETS threading programs. Usually, one template of the highest Z-score is selected from each threading program, where the threading programs are sorted by the average performance in the large-scale benchmark test experiments.)

(a) All the residues are colored in black; however, those residues in template which are identical to the residue in the query sequence are highlighted in color. Coloring scheme is based on the property of amino acids, where polar are brightly coloured while non-polar residues are colored in dark shade. ([more about the colors used](#))

(b) Rank of templates represents the top ten threading templates used by I-TASSER.

(c) Ident1 is the percentage sequence identity of the templates in the threading aligned region with the query sequence.

(d) Ident2 is the percentage sequence identity of the whole template chains with query sequence.

(e) Cov represents the coverage of the threading alignment and is equal to the number of aligned residues divided by the length of query protein.

(f) Norm. Z-score is the normalized Z-score of the threading alignments. Alignment with a Normalized Z-score >1 mean a good alignment and vice versa.

(g) Download Align. provides the 3D structure of the aligned regions of the threading templates.

(h) The top 10 alignments reported above (in order of their ranking) are from the following threading programs:
1: MUSTER 2: FFAS-3D 3: SPARKS-X 4: HHSEARCH2 5: HHSEARCH 6: Neff-PPAS 7: HHSEARCH 8: pGenTHREADER 9: PROSPECT2 10: PRC

(I-TASSER modeling starts from the structure templates identified by LOMETS from the PDB library. LOMETS is a meta-server threading approach containing multiple threading programs, the significance of which are measured by the Z-score, i.e. the difference between the raw and average scores in the unit of standard deviation. The templates in this section are sorted by the average performance in the large-scale benchmark test experiments.)

(a) All the residues are colored in black; however, those residues in template which are identical to the residue in the query sequence are highlighted in color. Coloring scheme is based on the property of amino acids, where polar are brightly coloured while non-polar residues are colored in dark shade. ([more about the colors used](#))

(b) Rank of templates represents the top ten threading templates used by I-TASSER.

(c) Ident1 is the percentage sequence identity of the templates in the threading aligned region with the query sequence.

(d) Ident2 is the percentage sequence identity of the whole template chains with query sequence.

(e) Cov represents the coverage of the threading alignment and is equal to the number of aligned residues divided by the length of query protein.

(f) Norm. Z-score is the normalized Z-score of the threading alignments. Alignment with a Normalized Z-score >1 mean a good alignment and vice versa.

(g) Download Align. provides the 3D structure of the aligned regions of the threading templates.

(h) The top 10 alignments reported above (in order of their ranking) are from the following threading programs:

1: MUSTER 2: FFAS-3D 3: SPARKS-X 4: HHSEARCH2 5: HHSEARCH I 6: Neff-PPAS 7: HHSEARCH 8: pGenTHREADER 9: PROSPECT2 10: PRC

- Here are the Top 10 threading templates. Based on the sequence identity, coverage, and Normalized Z-score, is thymidylate synthase from W.g.b. an easy or hard target?
- The sequence identity is not very high but in a reasonable range for homology modeling. The coverage is very high and Normalized Z-score is also high, so this not a hard target for I-TASSER.

[illegible]

Top 10 threading templates used by I-TASSER

(I-TASSER modeling starts from the structure templates identified by LOMETS from the PDB library. LOMETS is a meta-server threading approach containing multiple threading alignments, the significance of which are measured by the Z-score, i.e. the difference between the raw and average scores in the unit of standard deviation. The templates in this section are sorted by the average performance in the large-scale benchmark test experiments.)

[illegible]

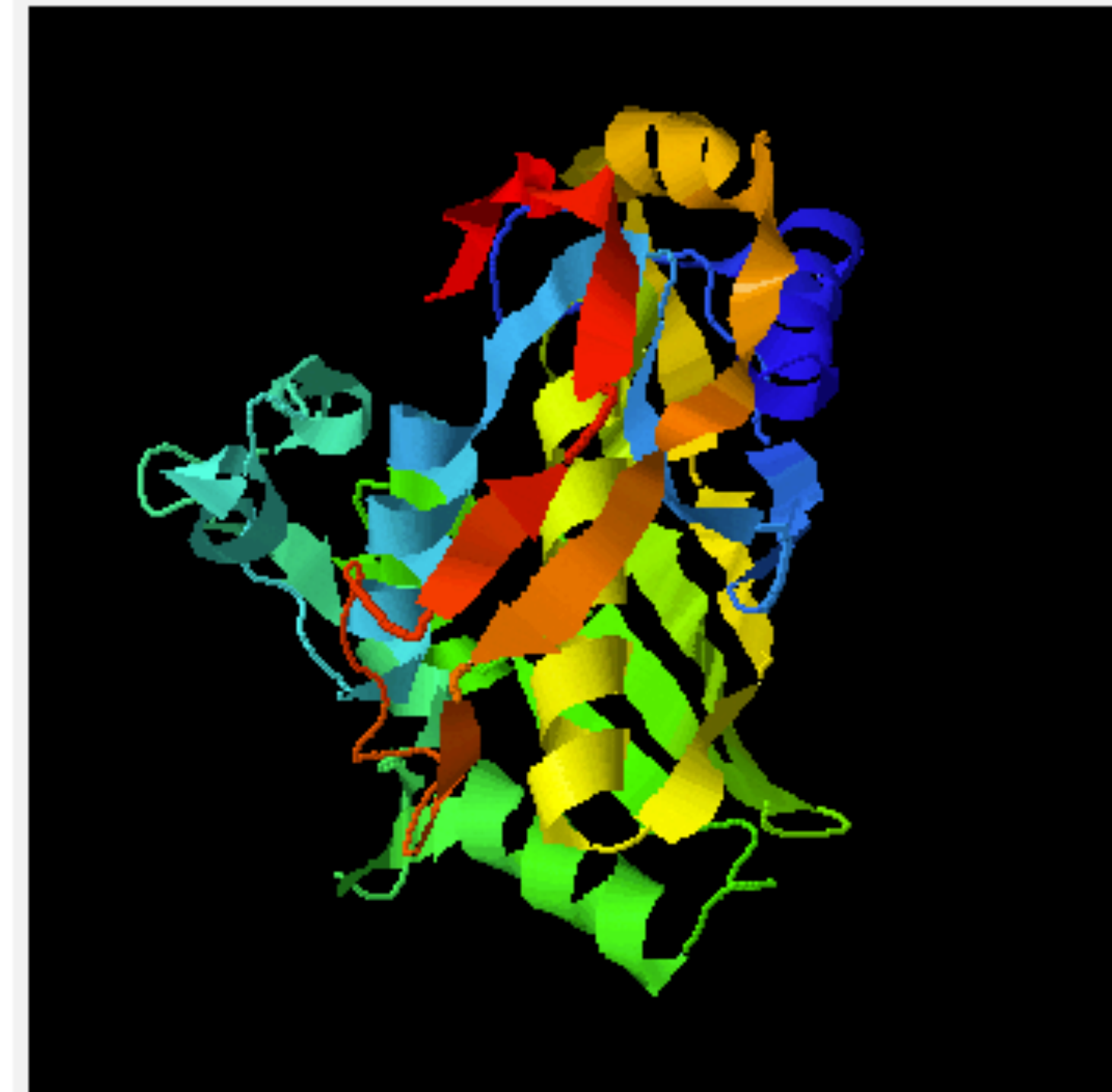
- The confidence of each model is quantitatively measured by C-score that is calculated based on the significance of threading template alignments and the convergence parameters of the structure assembly simulations. C-score is typically in the range of $[-5, 2]$, where a C-score of a higher value signifies a model with a higher confidence and vice-versa. TM-score and RMSD are estimated based on C-score and protein length...
- Is I-TASSER confident about its final model?

Top 5 final models predicted by I-TASSER

(For each target, I-TASSER simulations generate a large ensemble of structural quantitatively measured by C-score that is calculated based on the significance on C-score and protein length following the correlation observed between these rank models as seen in our benchmark tests. If the I-TASSER simulations conv

- [More about C-score](#)
- [Local structure accuracy profile of the top five models](#)

(By right-click on the images, you can export image file or change the



☐ Spin On/Off

- [Download Model 1](#)
- C-score=1.91 ([Read more about C-score](#))
- Estimated TM-score = 0.99 ± 0.04
- Estimated RMSD = $2.2 \pm 1.7 \text{ \AA}$

- The confidence of each model is quantitatively measured by C-score that is calculated based on the significance of threading template alignments and the convergence parameters of the structure assembly simulations. C-score is typically in the range of $[-5, 2]$, where a C-score of a higher value signifies a model with a higher confidence and vice-versa. TM-score and RMSD are estimated based on C-score and protein length...
- Is I-TASSER confident about its final model? Yes.

Top 5 final models predicted by I-TASSER

(For each target, I-TASSER simulations generate a large ensemble of structural quantitatively measured by C-score that is calculated based on the significance on C-score and protein length following the correlation observed between these rank models as seen in our benchmark tests. If the I-TASSER simulations conv

- [More about C-score](#)
- [Local structure accuracy profile of the top five models](#)

(By right-click on the images, you can export image file or change the



☐ Spin On/Off

- [Download Model 1](#)
- C-score=1.91 ([Read more about C-score](#))
- Estimated TM-score = 0.99 ± 0.04
- Estimated RMSD = $2.2 \pm 1.7 \text{ \AA}$

Discuss

- How can you be confident in a result from homology modeling/threading?
- Can homology modeling/threading be used to
 - predict the effect of a mutation
 - of a contact with a ligand in a binding site?
 - on a large-scale conformational change?
 - predict the effect of buffer conditions?

References

- Garg, D.; Skouloubris, S.; Briffotiaux, J.; Myllykallio, H.; Wade, R. C. Conservation and Role of Electrostatics in Thymidylate Synthase. Sci Rep 2015, 5 (1), 17356. <https://doi.org/10.1038/srep17356>, adapted under the CC BY 4.0 license.