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

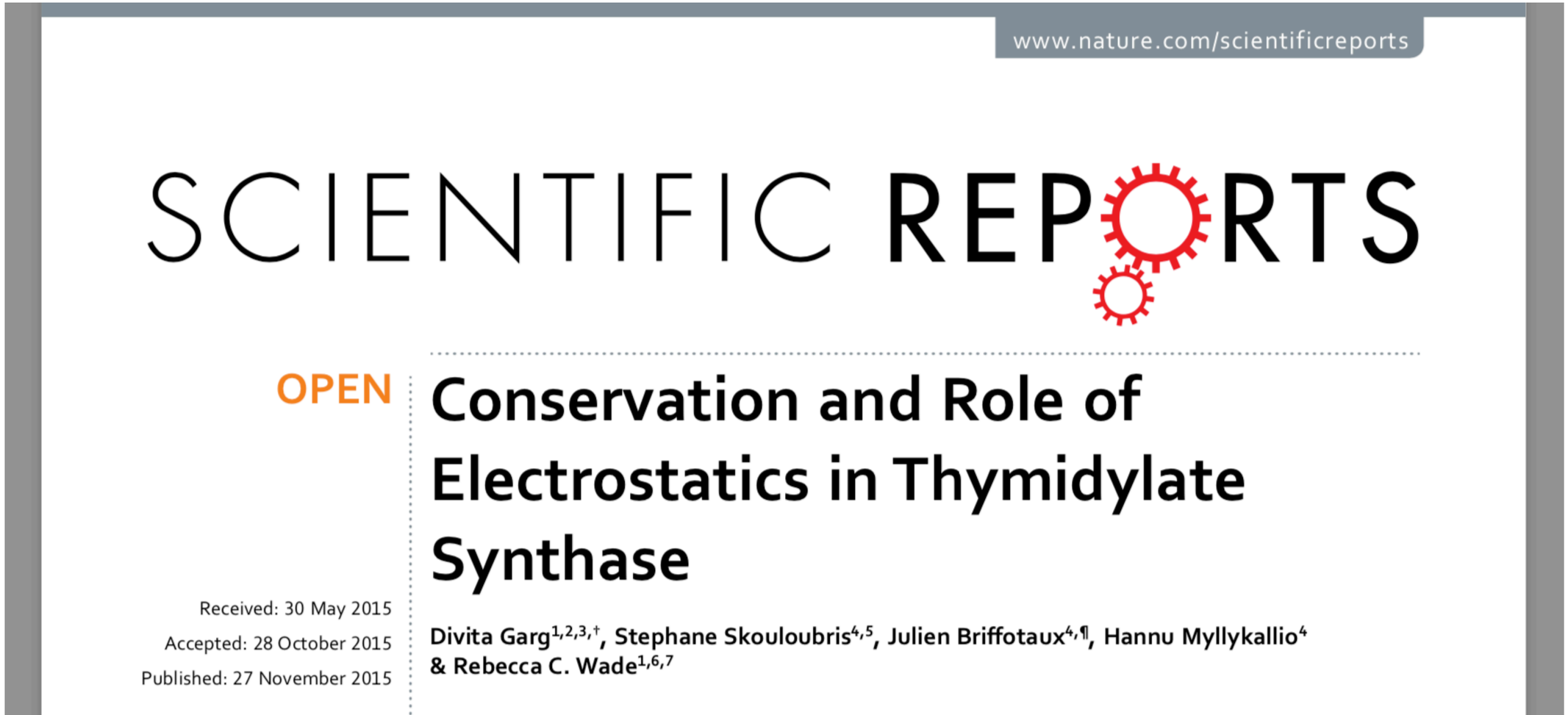
- **Homology modeling** is most useful for closer homology. Go through http://www.bioinfo.rpi.edu/bystrc/courses/biol4550/lecture7/assets/player/KeynoteDHTMLPlayer.html#0 up to slide 11
- **Threading** is more useful for distant homologues. Watch https://www.jove.com/video/3259/a-protocol-for-computer-based-protein-structure-function 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:



www.nature.com/scientificreports

SCIENTIFIC REPORTS

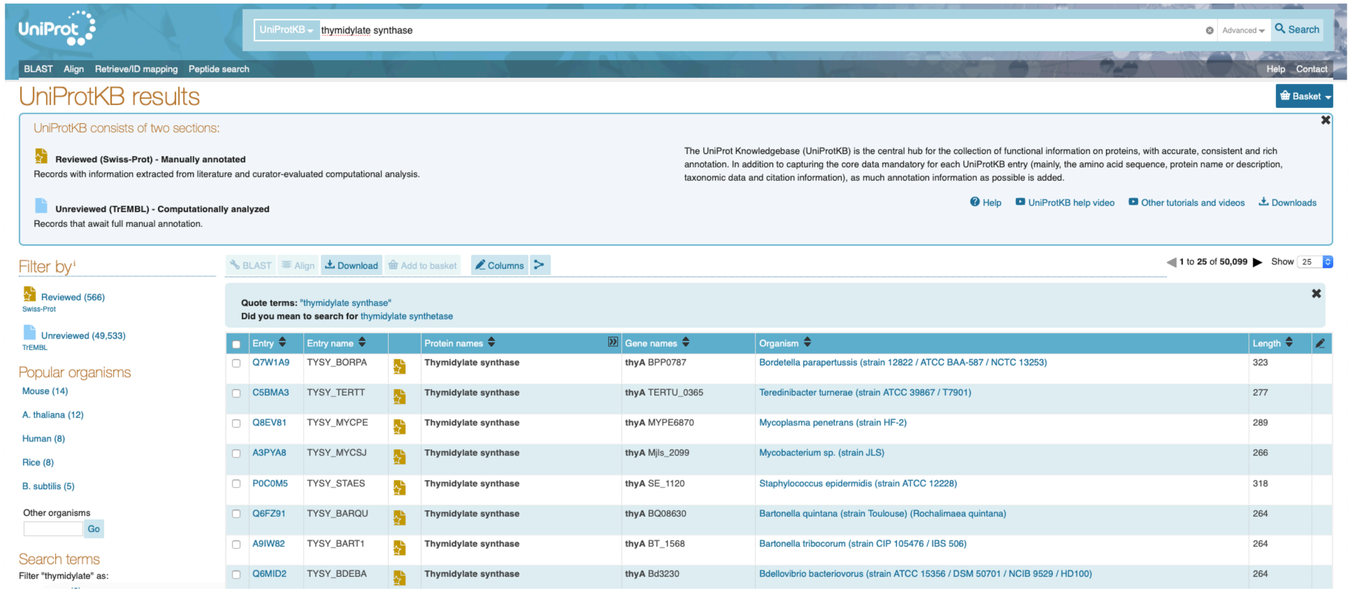
OPEN

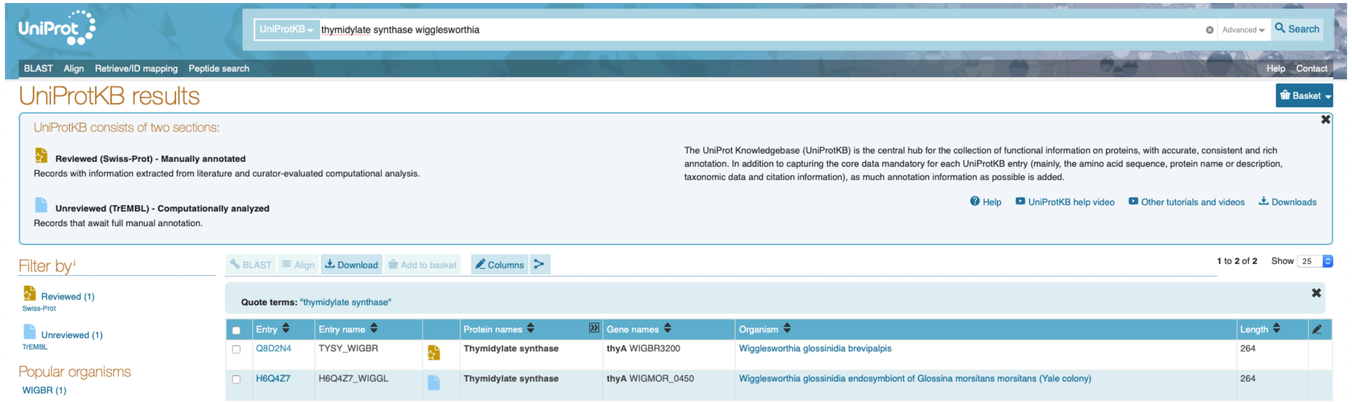
Conservation and Role of Electrostatics in Thymidylate Synthase

Received: 30 May 2015
Accepted: 28 October 2015
Published: 27 November 2015

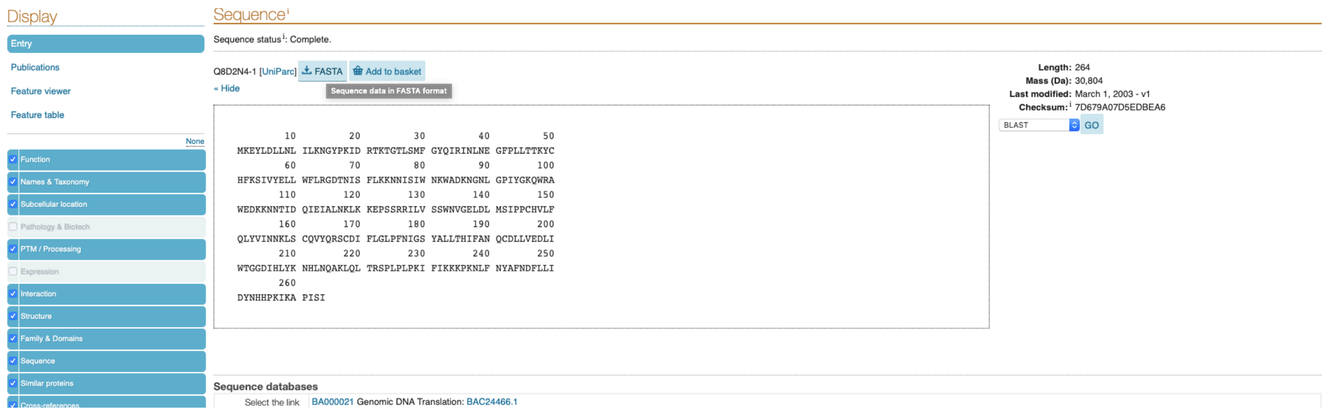
Divita Garg[1,2,3,†], Stephane Skouloubris[4,5], Julien Briffotaux[4,¶], Hannu Myllykallio[4] & Rebecca C. Wade[1,6,7]

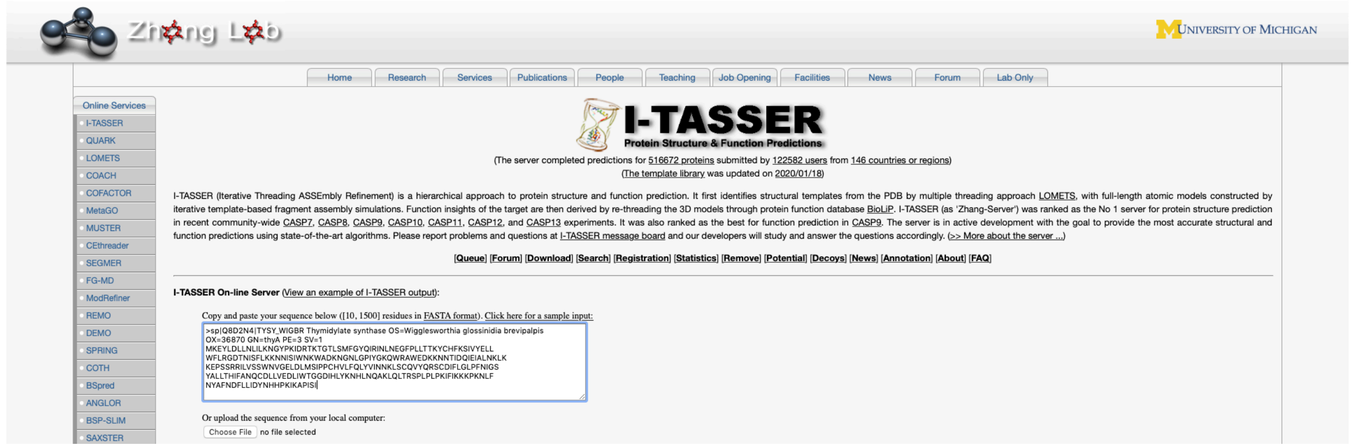
- Garg *et. al.* created a homology moodel from a minimal organism W.g.b. To reproduce their result, I needed to find the sequence.
- UniProt (https://www.uniprot.org) is a comprehensive biological sequence database
- A search for “thymidylate synthase” yields 50,099 results, including 566 that have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



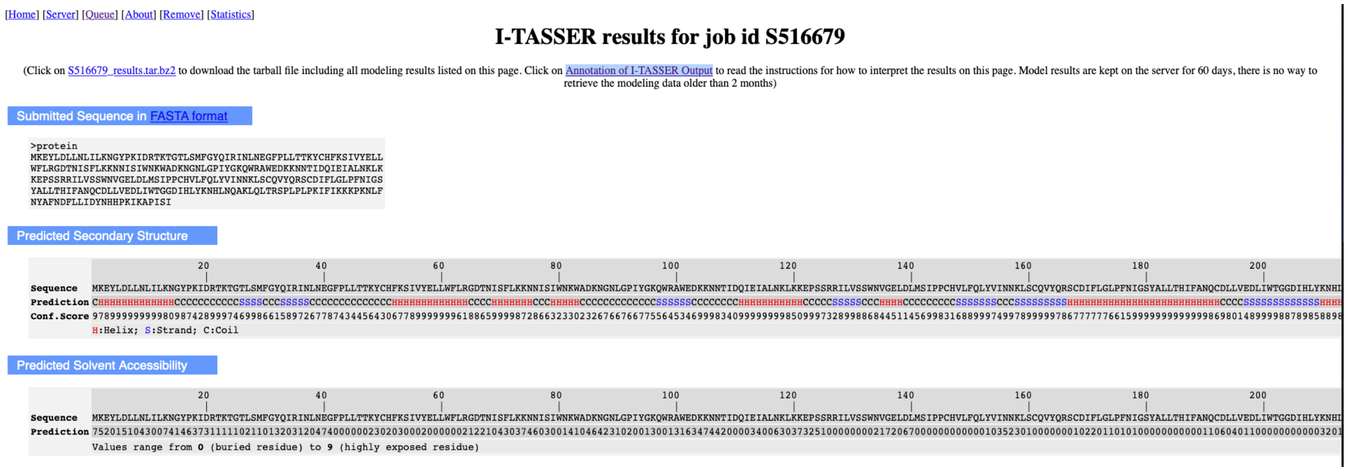


- To get the actual amino acid sequence, click on “Sequence” or scroll down and then click on FASTA. FASTA is a simple format for amino acid sequences based on one-letter codes.
- To run I-TASSER, you can visit the server interface at https://zhanglab.ccmb.med.umich.edu/I-TASSER/, register, paste the 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.





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



[Home] [Server] [Queue] [About] [Remove] [Statistics]

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
MKEYLDLLNLILKNGYPKIDRTKTGTLSMFGYQIRINLNEGFPLLTTKYCHFKSIVYELL
WFLRGDTNISFLKKNNISIWNKWADKNGNLGPIYGKQWRAWEDKKNNTIDQIEIALNKLK
KEPSSRRILVSSWNVGELDLMSIPPCHVLFQLYVINNKLSCQVYQRSCDIFLGLPFNIGS
YALLTHIFANQCDLLVEDLIWTGGDIHLYKNHLNQAKLQLTRSPLPLPKIFIKKKPKNLF
NYAFNDPLLIDYNHHPKIKAPISI
```

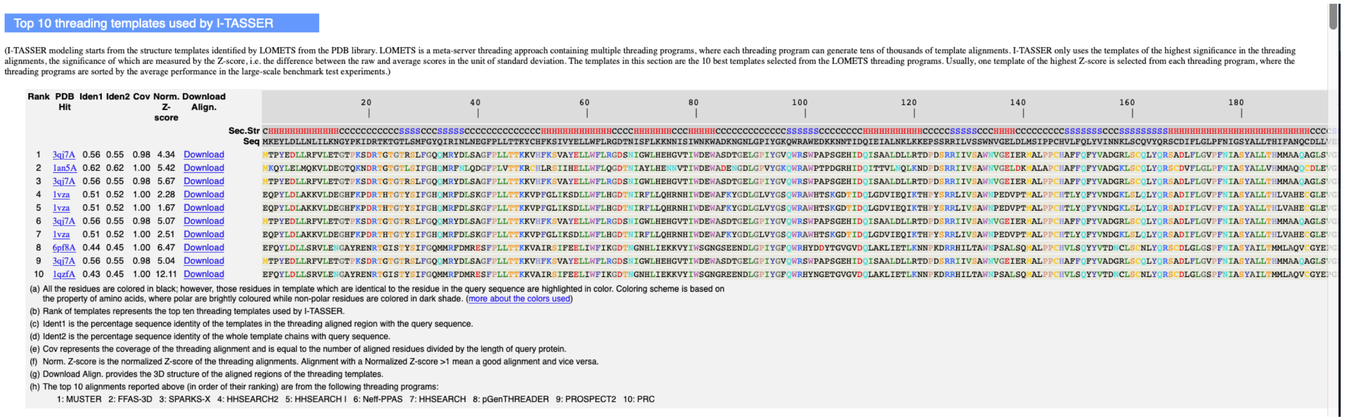
Predicted Secondary Structure

```
                    20        40        60        80       100       120       140       160       180       200
                     |         |         |         |         |         |         |         |         |         |
Sequence   MKEYLDLLNLILKNGYPKIDRTKTGTLSMFGYQIRINLNEGFPLLTTKYCHFKSIVYELLWFLRGDTNISFLKKNNISIWNKWADKNGNLGPIYGKQWRAWEDKKNNTIDQIEIALNKLKKEPSSRRILVSSWNVGELDLMSIPPCHVLFQLYVINNKLSCQVYQRSCDIFLGLPFNIGSYALLTHIFANQCDLLVEDLIWTGGDIHLYKNHI
Prediction CHHHHHHHHHHHCCCCCCCCCCSSSSCCCSSSSSCCCCCCCCCCCCCHHHHHHHHHHHCCCCHHHHHHCCCHHHHHCCCCCCCCCCCCCSSSSSSCCCCCCCCHHHHHHHHHHCCCCCSSSSSCCCHHHHCCCCCCCCCSSSSSSCCCSSSSSSSSHHHHHHHHHHHHHHHHHHHHHHHHCCCCSSSSSSSSSSSSSSHHHH
Conf.Score 9789999999998098742899974699866158972677874344564306778999999961886599998728663233023267667667755645346999834099999998509997328998868445114569983168899974997899999786777777661599999999999986980148999988789858898
           H:Helix; S:Strand; C:Coil
```

Predicted Solvent Accessibility

```
                    20        40        60        80       100       120       140       160       180       200
                     |         |         |         |         |         |         |         |         |         |
Sequence   MKEYLDLLNLILKNGYPKIDRTKTGTLSMFGYQIRINLNEGFPLLTTKYCHFKSIVYELLWFLRGDTNISFLKKNNISIWNKWADKNGNLGPIYGKQWRAWEDKKNNTIDQIEIALNKLKKEPSSRRILVSSWNVGELDLMSIPPCHVLFQLYVINNKLSCQVYQRSCDIFLGLPFNIGSYALLTHIFANQCDLLVEDLIWTGGDIHLYKNHI
Prediction 752015104300741463731111102110132031204740000002302030002000000212210430374603001410464231020013001316347442000034006303732510000000021720670000000000010352301000000010220110101000000000011060401100000000003201
           Values range from 0 (buried residue) to 9 (highly exposed residue)
```

- 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?

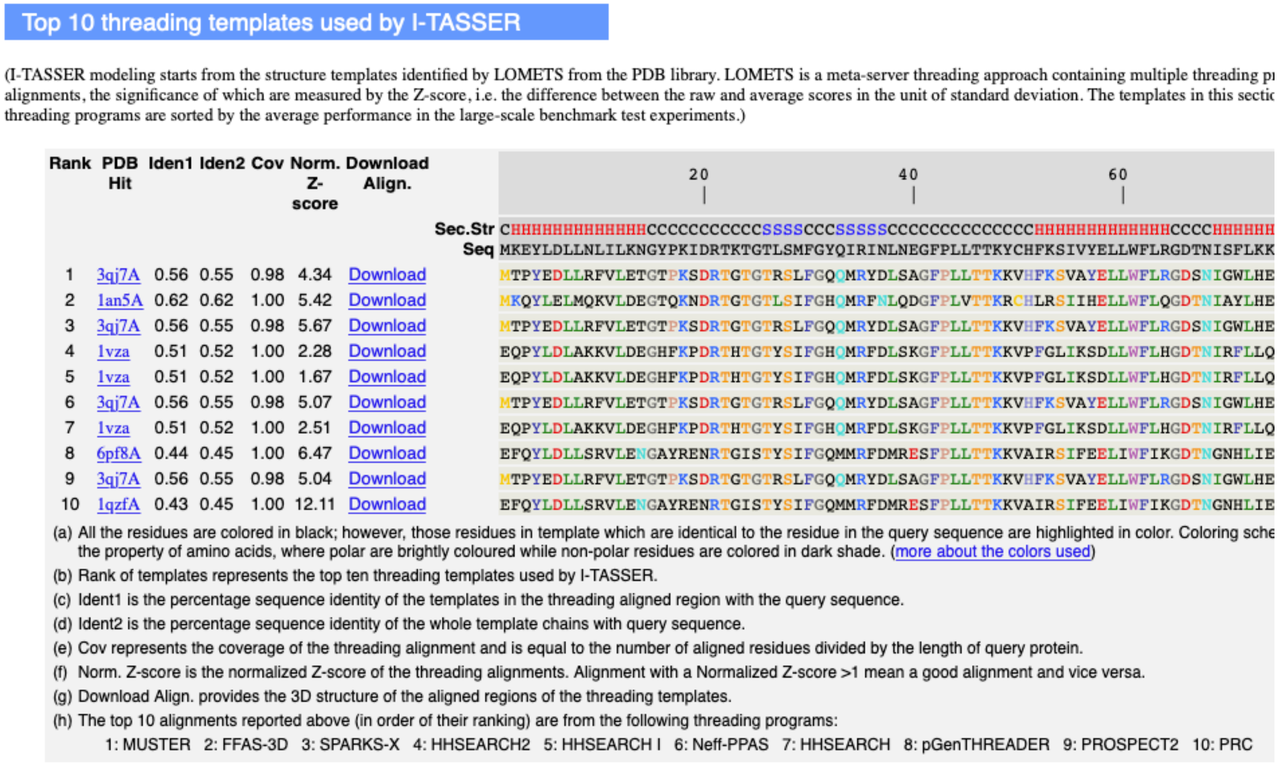


Top 10 threading templates used by I-TASSER

(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 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.)

Rank	PDB Hit	Iden1	Iden2	Cov	Norm. Z-score	Download Align.
1	3qj7A	0.56	0.55	0.98	4.34	Download
2	1an5A	0.62	0.62	1.00	5.42	Download
3	3qj7A	0.56	0.55	0.98	5.67	Download
4	1vza	0.51	0.52	1.00	2.28	Download
5	1vza	0.51	0.52	1.00	1.67	Download
6	3qj7A	0.56	0.55	0.98	5.07	Download
7	1vza	0.51	0.52	1.00	2.51	Download
8	6pf8A	0.44	0.45	1.00	6.47	Download
9	3qj7A	0.56	0.55	0.98	5.04	Download
10	1qzfA	0.43	0.45	1.00	12.11	Download

(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



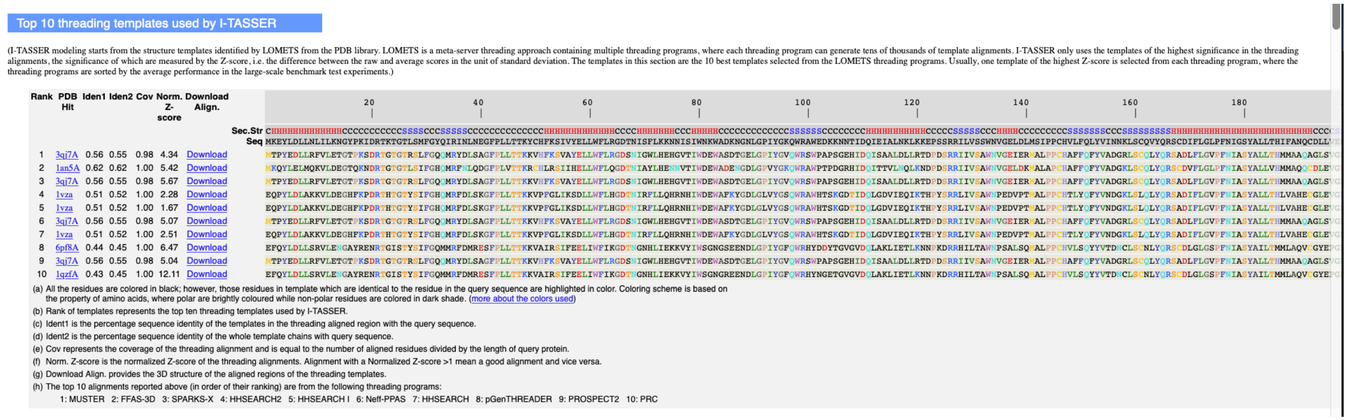
Top 10 threading templates used by I-TASSER

(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 pr 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 sectic threading programs are sorted by the average performance in the large-scale benchmark test experiments.)

Rank	PDB Hit	Iden1	Iden2	Cov	Norm. Z-score	Download Align.
1	3qj7A	0.56	0.55	0.98	4.34	Download
2	1an5A	0.62	0.62	1.00	5.42	Download
3	3qj7A	0.56	0.55	0.98	5.67	Download
4	1vza	0.51	0.52	1.00	2.28	Download
5	1vza	0.51	0.52	1.00	1.67	Download
6	3qj7A	0.56	0.55	0.98	5.07	Download
7	1vza	0.51	0.52	1.00	2.51	Download
8	6pf8A	0.44	0.45	1.00	6.47	Download
9	3qj7A	0.56	0.55	0.98	5.04	Download
10	1qzfA	0.43	0.45	1.00	12.11	Download

(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 sche 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.

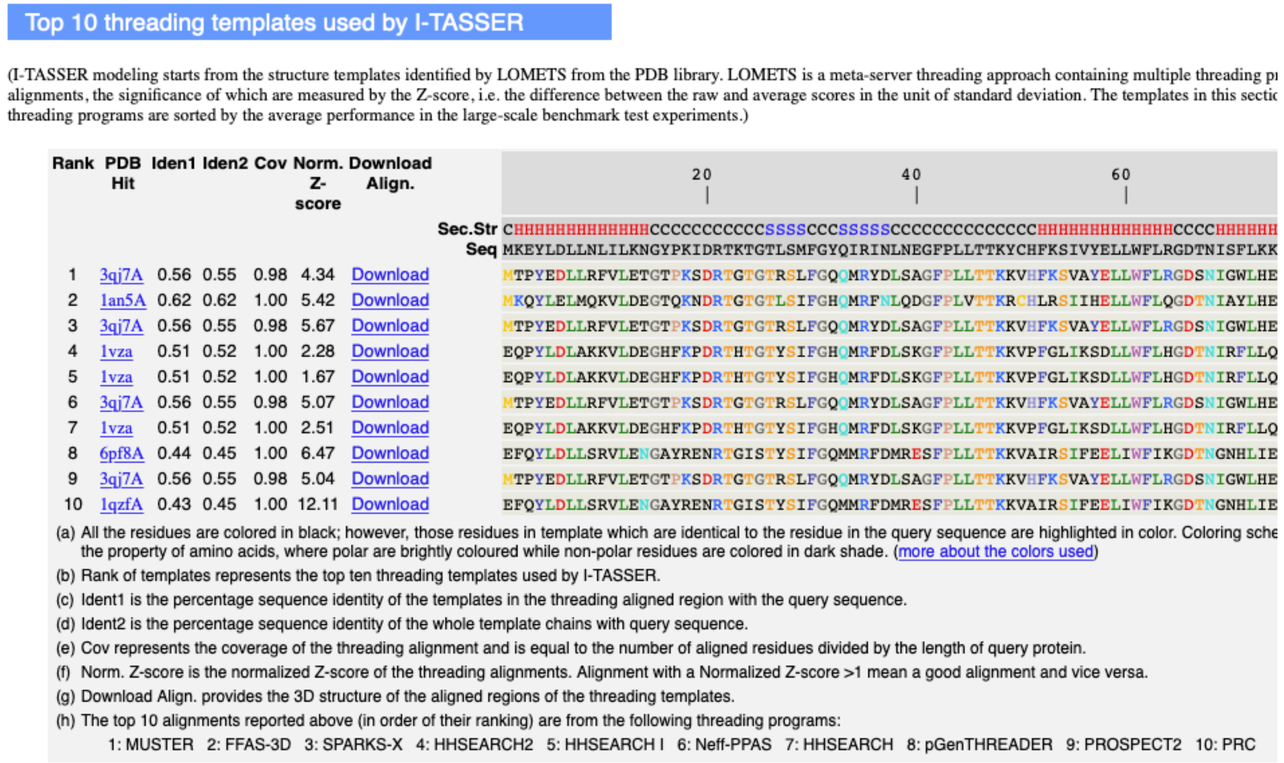


Top 10 threading templates used by I-TASSER

(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 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.)

Rank	PDB Hit	Iden1	Iden2	Cov	Norm. Z-score	Download Align.
1	3qj7A	0.56	0.55	0.98	4.34	Download
2	1an5A	0.62	0.62	1.00	5.42	Download
3	3qj7A	0.56	0.55	0.98	5.67	Download
4	1vza	0.51	0.52	1.00	2.28	Download
5	1vza	0.51	0.52	1.00	1.67	Download
6	3qj7A	0.56	0.55	0.98	5.07	Download
7	1vza	0.51	0.52	1.00	2.51	Download
8	6pf8A	0.44	0.45	1.00	6.47	Download
9	3qj7A	0.56	0.55	0.98	5.04	Download
10	1qzfA	0.43	0.45	1.00	12.11	Download

(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



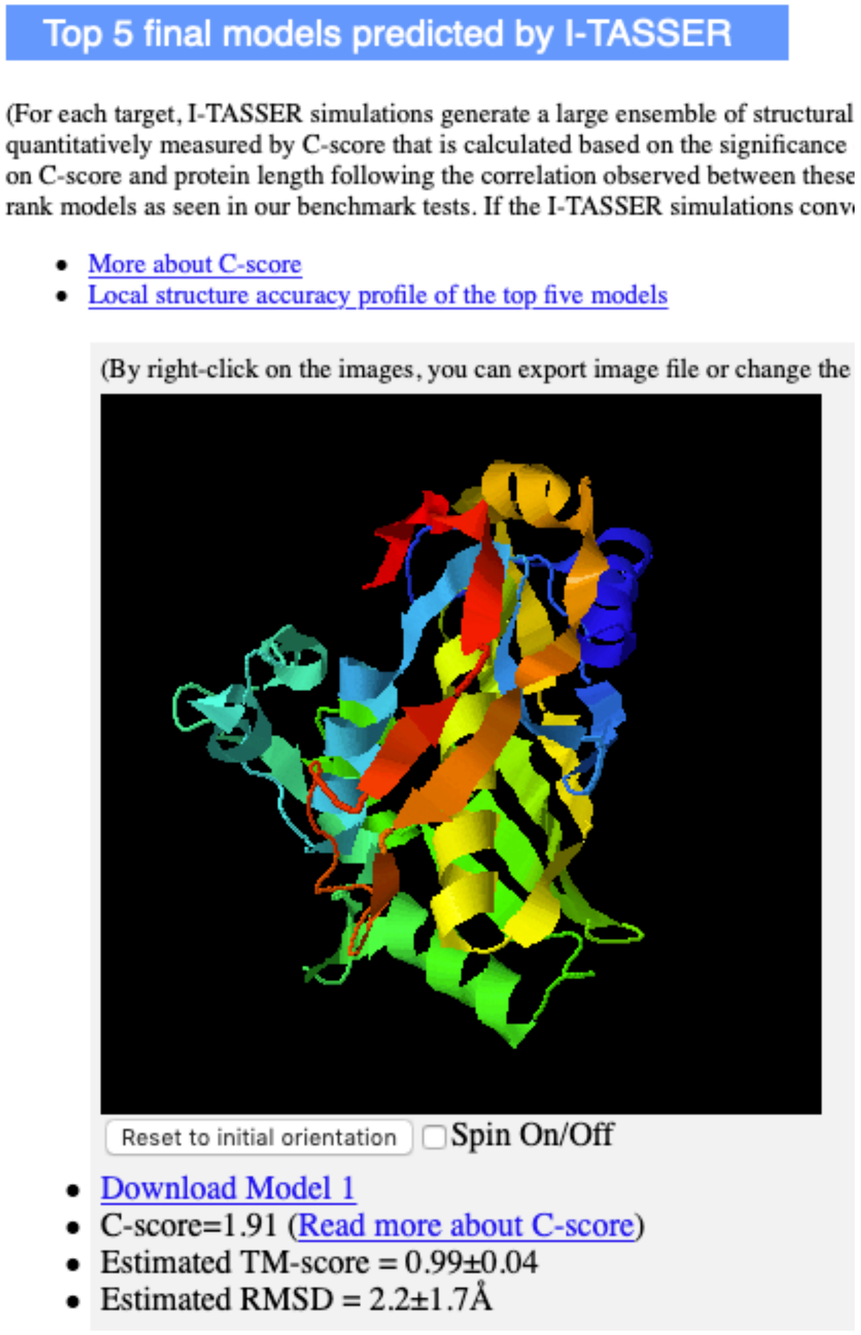
Top 10 threading templates used by I-TASSER

(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 pr 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 sectic threading programs are sorted by the average performance in the large-scale benchmark test experiments.)

Rank	PDB Hit	Iden1	Iden2	Cov	Norm. Z-score	Download Align.
1	3qj7A	0.56	0.55	0.98	4.34	Download
2	1an5A	0.62	0.62	1.00	5.42	Download
3	3qj7A	0.56	0.55	0.98	5.67	Download
4	1vza	0.51	0.52	1.00	2.28	Download
5	1vza	0.51	0.52	1.00	1.67	Download
6	3qj7A	0.56	0.55	0.98	5.07	Download
7	1vza	0.51	0.52	1.00	2.51	Download
8	6pf8A	0.44	0.45	1.00	6.47	Download
9	3qj7A	0.56	0.55	0.98	5.04	Download
10	1qzfA	0.43	0.45	1.00	12.11	Download

(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 sche 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

- 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?



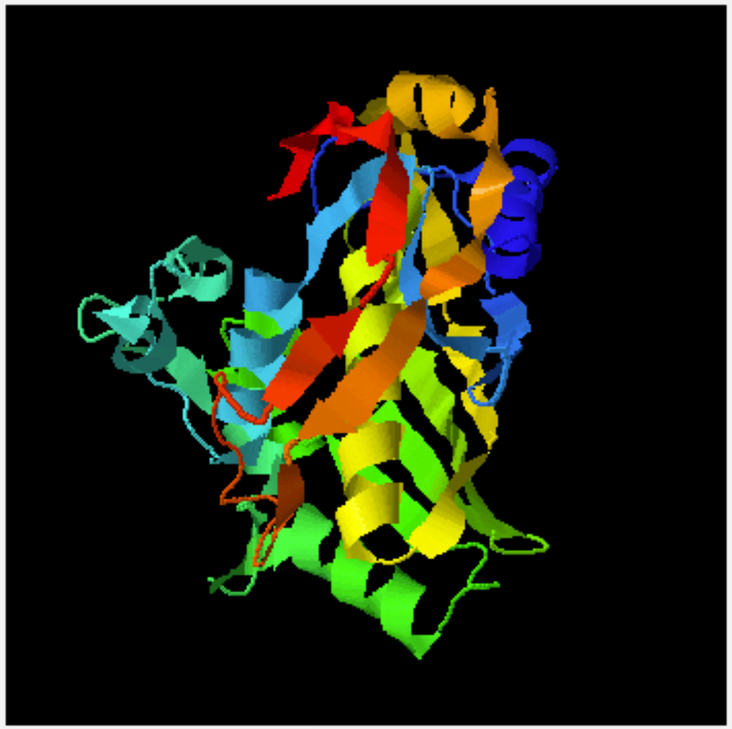
- The confidence of each model is quantitatively measured by C-score that is calculated based on the significance of threading template alignments and the convergence parameters of the structure assembly simulations. C-score is typically in the range of [-5, 2], where a C-score of a higher value signifies 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



Reset to initial orientation ☐ 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±1.7Å

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.; Briffotaux, 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.