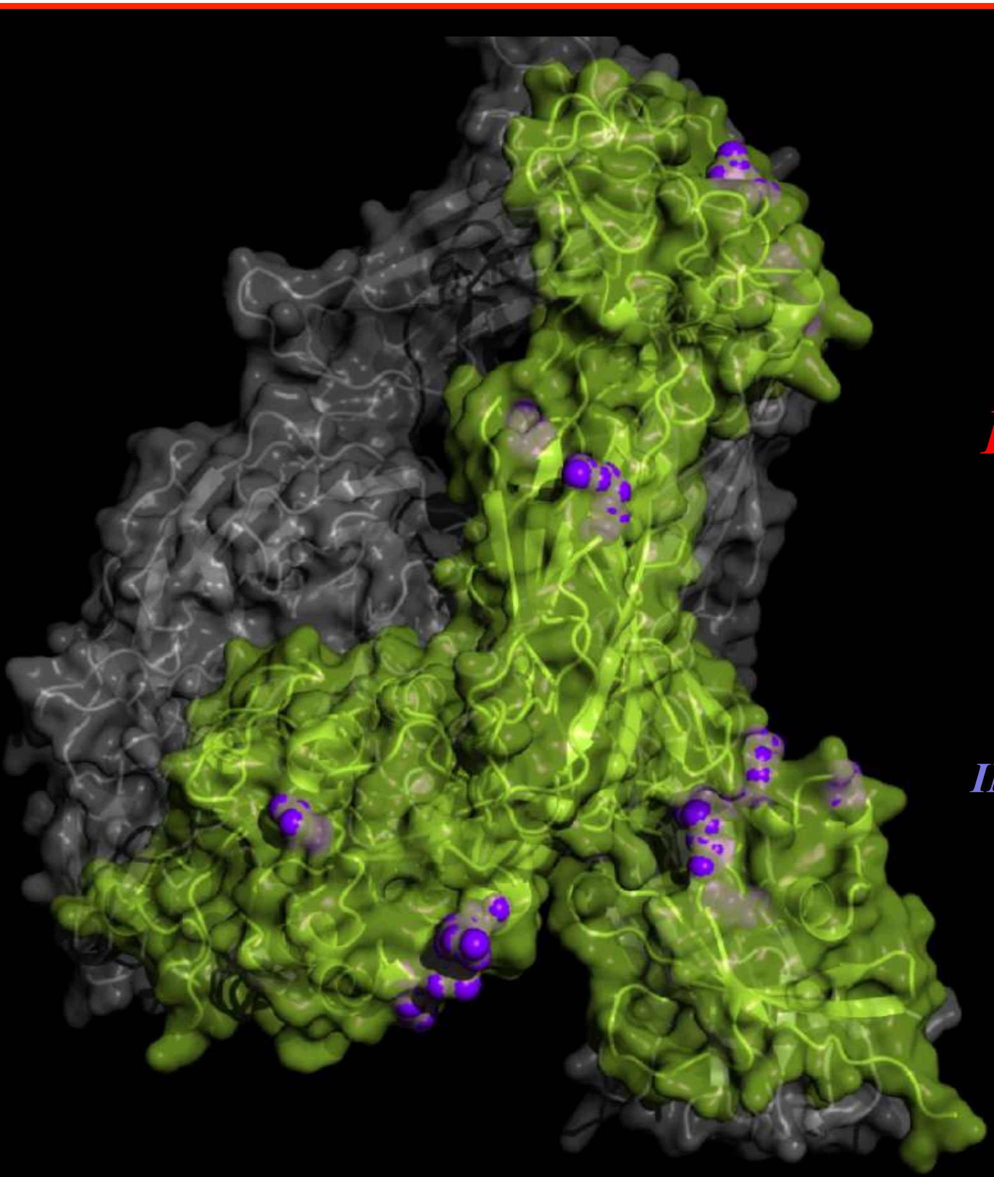


AlphaFold, the Artificial Intelligence approach (Nobel Prize 2024): a real (r)evolution or not?

Alexandre G. de Brevern

*INSERM UMR_S 1134, DSIMB Bioinformatics team,
Université Paris Cité,
Université de la Réunion, Paris, FRANCE.*

Conflicts of Interest

- None
- My comments are my own (and not those of INSERM, universities, etc.)

DSIMB: a bioinformatics team

My team

Associated to Université de la Réunion

ARCTIC OCEAN
Greenland
Svalbard (Nor.)
ARCTIC OCEAN
United States of America
Canada
Iceland
Norway
Sweden
Finland
Russia
United Kingdom
Ireland
Denmark
Neth.
Estonia
Latvia
Belarus
Pol.
Ger.
Ukraine
Kazakhstan
Mongolia
France
Italy
Rom.
Georgia
Bulgaria
Azer.
Uzb.
Kyrgyzstan
Tajikistan
United States of America
ATLANTIC OCEAN
Portugal
Spain
Greece
Turkey
Turkm.
N. Korea
S. Korea
Japan
China
Tunisia
Lebanon
Syria
Iraq
Iran
Afg.
Pakistan
Nepal
Bhutan
Morocco
Israel
Jordan
Kuwait
Algeria
Libya
Egypt
Saudi Arabia
U.A.E.
PACIFIC
Mexico
Western Sahara
India
Laos
Taiwan
Oman
Burma
PACIFIC OCEAN
Mauritania
Mali
Niger
Chad
Eritrea
Yemen
Bangl.
Thailand
Vietnam
Philippines
Guatemala
Honduras
Nicaragua
Senegal
Sudan
Djibouti
Costa Rica
Panama
Ven.
Guyana
French Guiana
Guinea-Bissau
Sierra Leone
Liberia
Cote d' Ivoire
Ghana
Nigeria
Cam.
C.A.R.
Ethiopia
Cambodia
Colombia
Togo
Somalia
Sri Lanka
Malaysia
Ecuador
Suriname
Gabon
Dem. Rep. of Congo
Uganda
Kenya
INDIAN OCEAN
Papua New Guinea
Indonesia
Rep. of the Congo
Tanzania
Peru
Brazil
Angola
Zambia
Malawi
Bolivia
Namibia
Zimbabwe
Madagascar
Mozambique
Paraguay
Botswana
Chile
ATLANTIC OCEAN
Swaziland
Lesotho
Australia
South Africa
Argentina
Uruguay
OCEAN

© 2007 Geology.com

PARADIGM

Paradigm

- FRANCIS CRICK (1970, Nature)

The central dogma of molecular biology deals with the detailed residue-by-residue transfer of sequential information. It states that such information cannot be transferred from protein to either protein or nucleic acid.

NATURE VOL. 227 AUGUST 8 1970

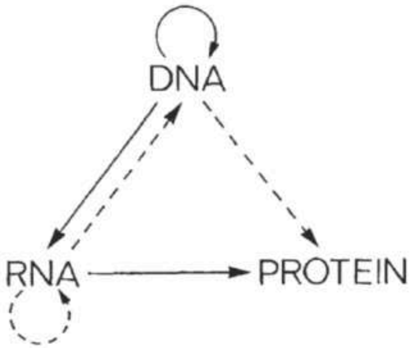
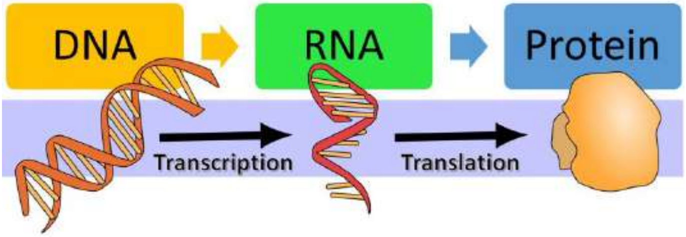


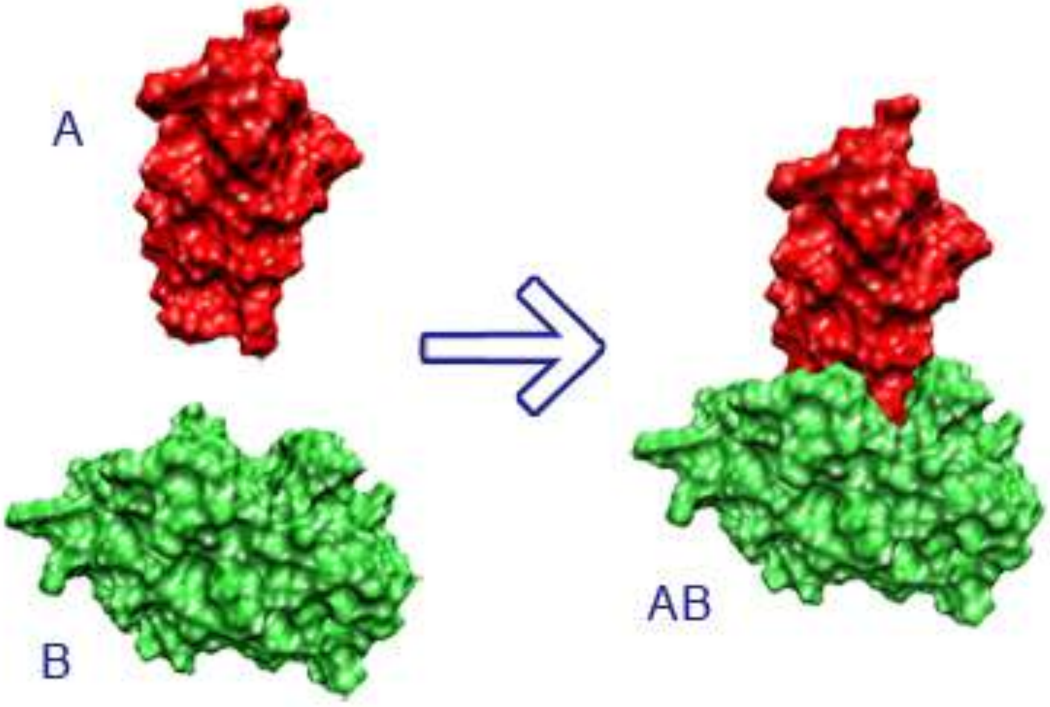
Fig. 3. A tentative classification for the present day. Solid arrows show general transfers; dotted arrows show special transfers. Again, the absent arrows are the undetected transfers specified by the central dogma.



INTEREST OF PROTEIN 3D STRUCTURES

Interest of protein 3D structures

- Because protein function(s) is at atomic scale



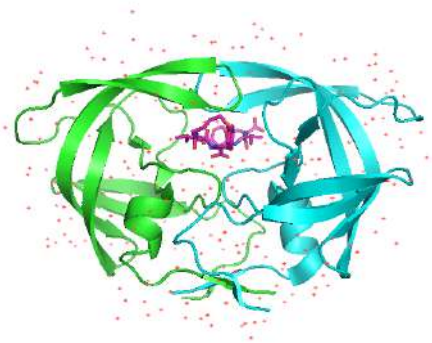
Inserm
Institut national
de la santé et de la recherche médicale

Interest of protein 3D structures

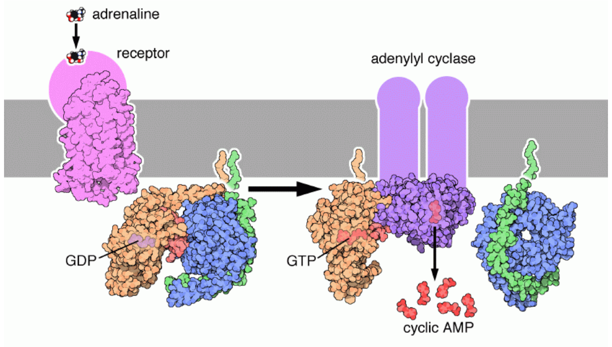
SIMB
Bioinformatique & Interactions

- Understanding the function(s) and more ...

Enzymes

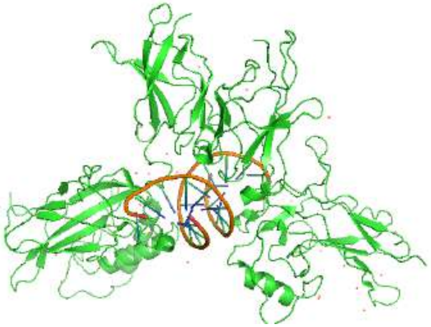


Receptors

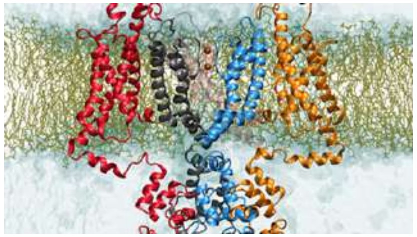


Drug design

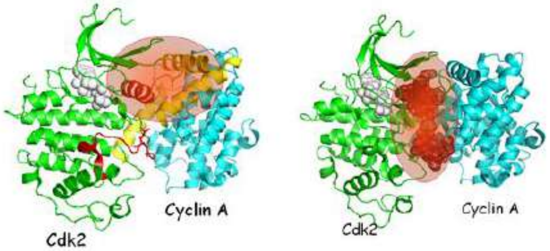
Transcription Factors



Transport

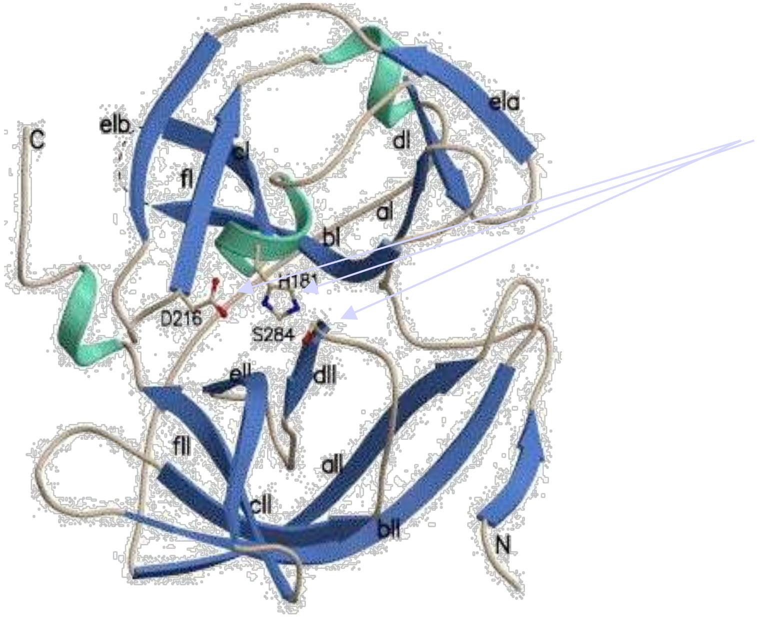


Protein-Protein Interaction

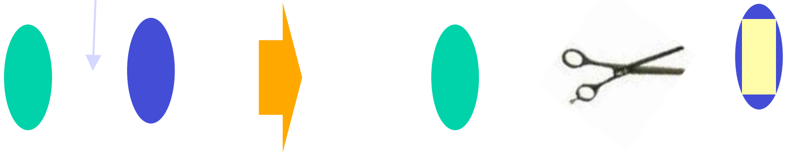


Interest of protein 3D structures

- Understand enzymatic mecanisms



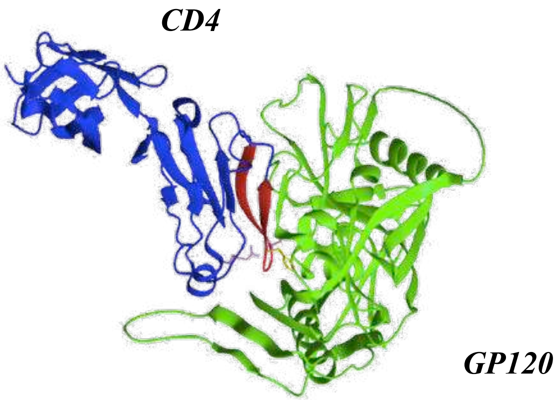
Catalytic triad

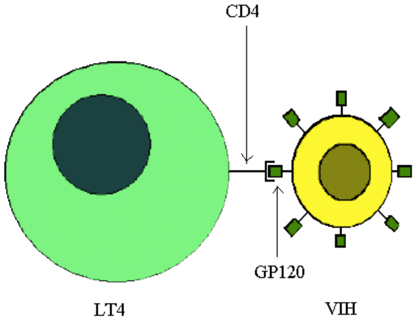


Protease serine

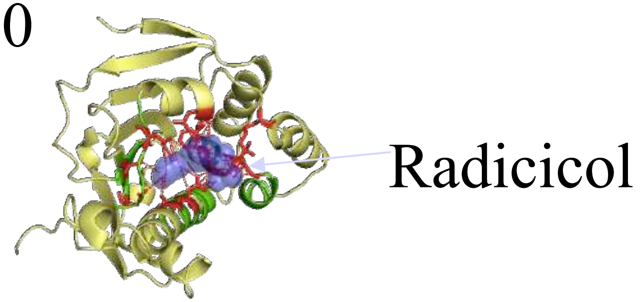
Interest of protein 3D structures

- Understand protein-ligand interactions





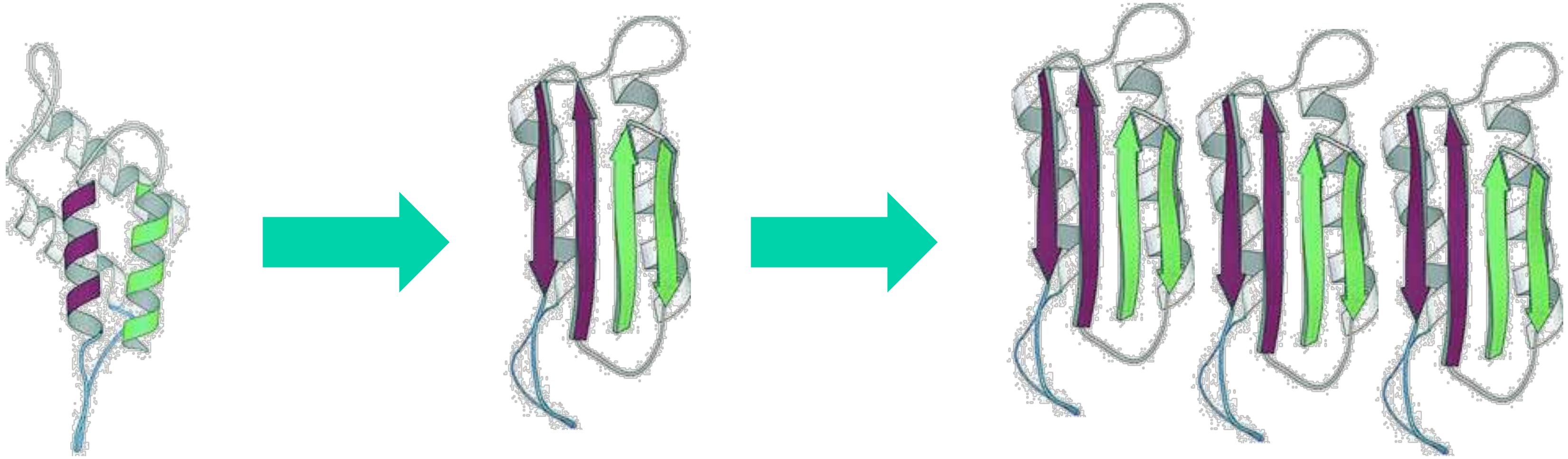
HSP90



i.e., chaperon proteins, inhibitors …

Interest of protein 3D structures

- Understand diseases



Agregation

prion (Creutzfeld-Jacob)

i.e. Alzheimer disease, Parkinson…

Snoopy’s question

You are right Alex, but a lot of proteins have not avalaible 3D structures

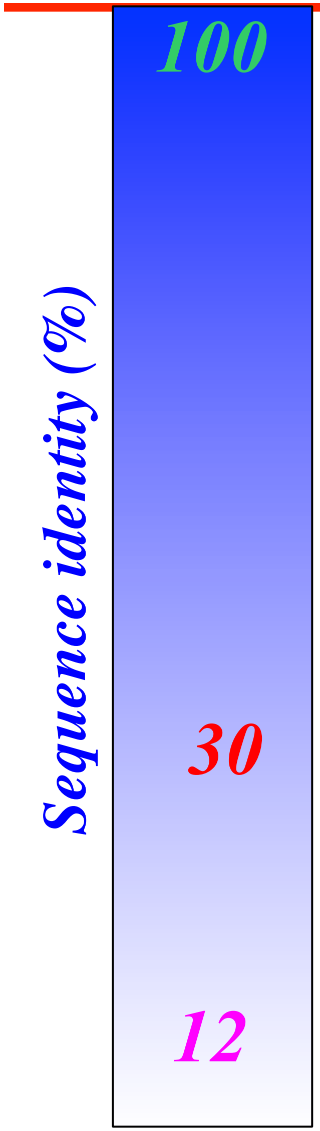
3D MODELLING

3D modelling

- However, then number of protein 3D structures is largely lower than the number of avalaible protein sequences…

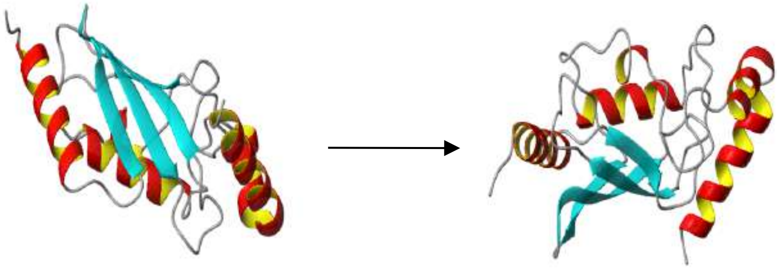
- So we use, since 40 years, different approaches to build from the sequences pertinent structural models.

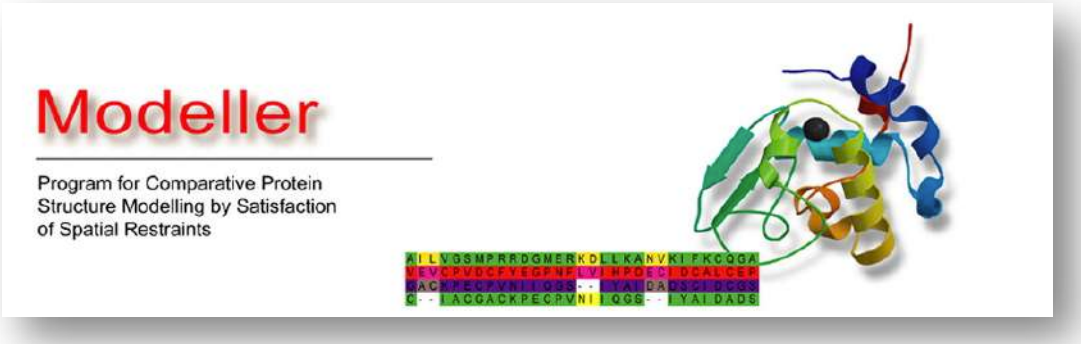
3D modelling



Homology modelling

```
ATPLGLPTHVVVAGLNPHTRESD
ATPLGIPTHVPPAGLNPHTRESD
!!!!! !!!!  !!!!!!!!!!!
```



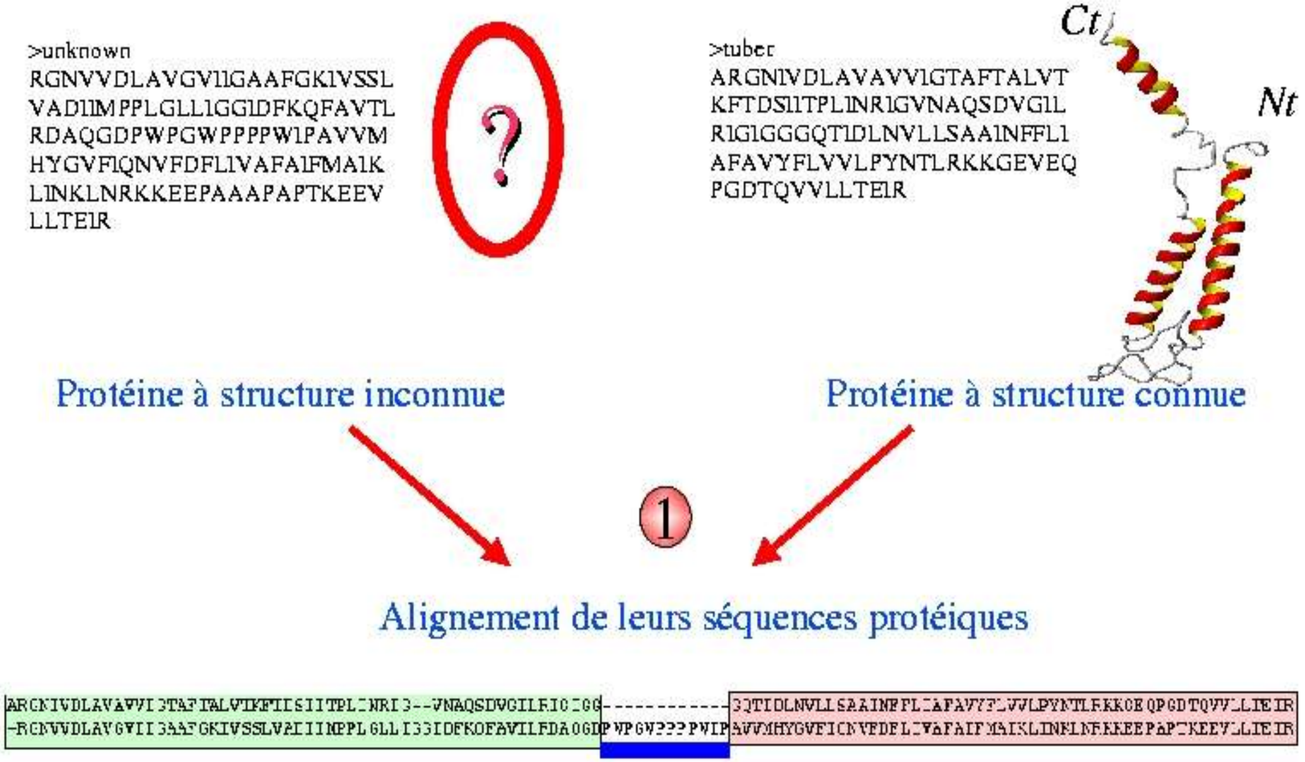


Modeller

Program for Comparative Protein Structure Modelling by Satisfaction of Spatial Restraints

AILVGSMPRRDGMERKDLLKANVKI FKCQGA
C--IACGACKPECPVNI IQGS--IYAIDADS

Comparative modelling



```
>unknown
RGNVVDLAVGVIIGAAFGKIVSSL
VADIIMPPLGLLIGGIDFKQFAVTL
RDAQGDPWPGWPPPPWIPAVVM
HYGVFIQNVFDFLIVAFAIFMAIK
LINKLNRKKEEPAAAPAPTKEEV
LLTEIR
```

Protéine à structure inconnue

```
>tuber
ARGNIVDLAVAVVIGTAFTALVT
KFTDSIITPLINRIGVNAQSDVGIL
RIGIGGGQTIDLNVLLSAAINFFLI
AFAVYFLVVLPYNTLRKKGEVEQ
PGDTQVVLLTEIR
```

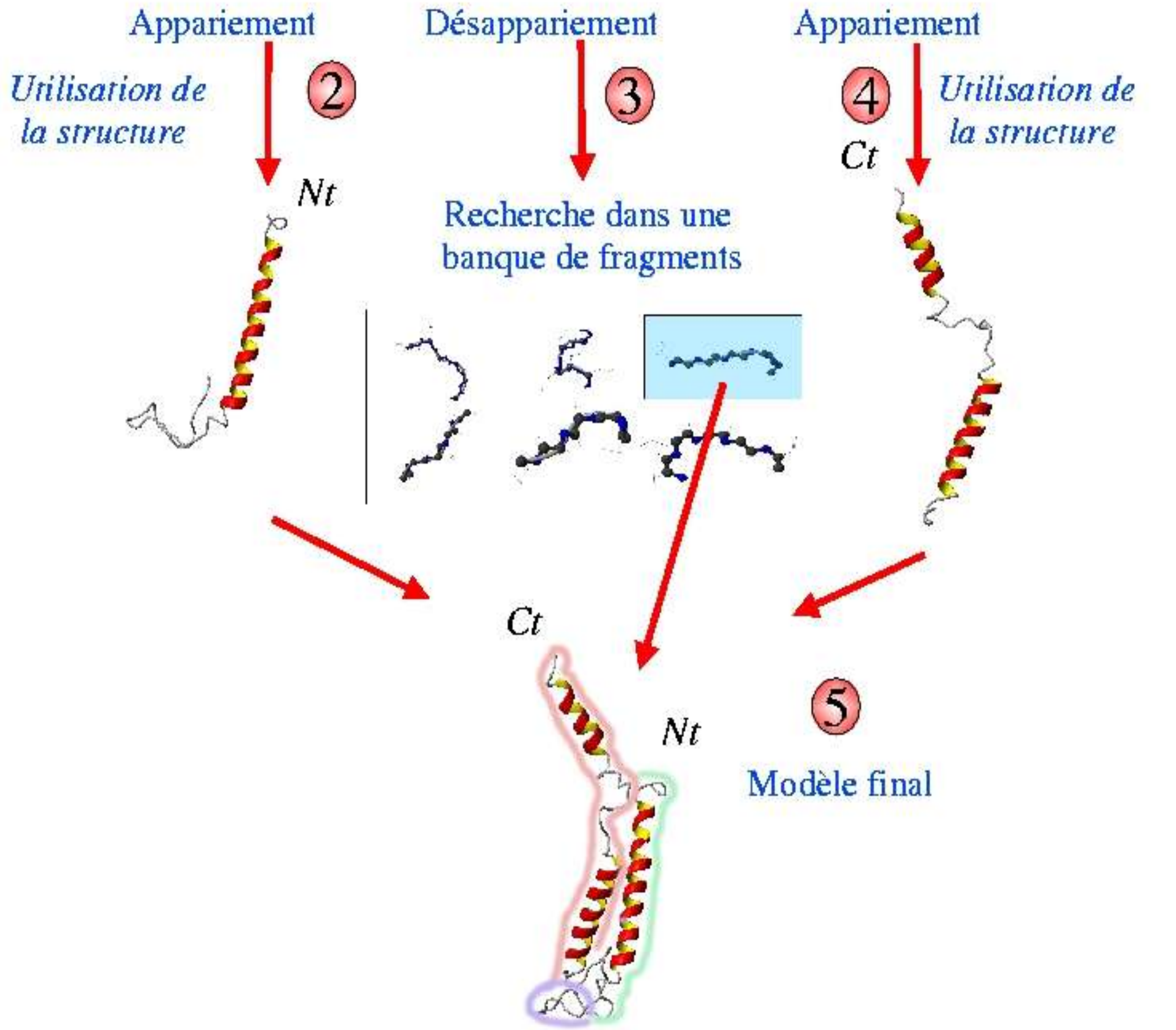
Protéine à structure connue

Alignement de leurs séquences protéiques

```
ARGNIVDLAVAVVIGTAFTALVTKFTDSIITPLINRIG--VNAQSDVGILRIGIGG------------GQTIDLNVLLSAAINFFLIAFAVYFLVVLPYNTLRKKGEQPGDTQVVLLTEIR
-RGNVVDLAVGVIIGAAFGKIVSSLVADIIMPPLGLLIGGIDFKQFAVTLRDAQGDPWPGWPPPPWIPAVVMHYGVFIQNVFDFLIVAFAIFMAIKLINKLNRKKEEPAPTKEEVLLTEIR
```

Alignement de leurs séquences protéiques

```
ARGNIVDLAVAVVIGTAFIALVTKFIISIITPLINRIG--VNAQSDVGILRICIGG------------GQTIDLNVLLSAAINFFLIAFAVYFLVVLPYNTLRKKGEQPGDTQVVLLIEIR
-RGNVVDLAVGVIIGAAFGKIVSSLVADIIMPPLGLLIGGIDFKQFAVILFDAOGDPVPGV???PVIPAVVMHYGVFICNVFDFLIVAFAIFMAIKLINKLNRKKEEPAPTKEEVLLIEIR
```



Appariement

Désappariement

Appariement

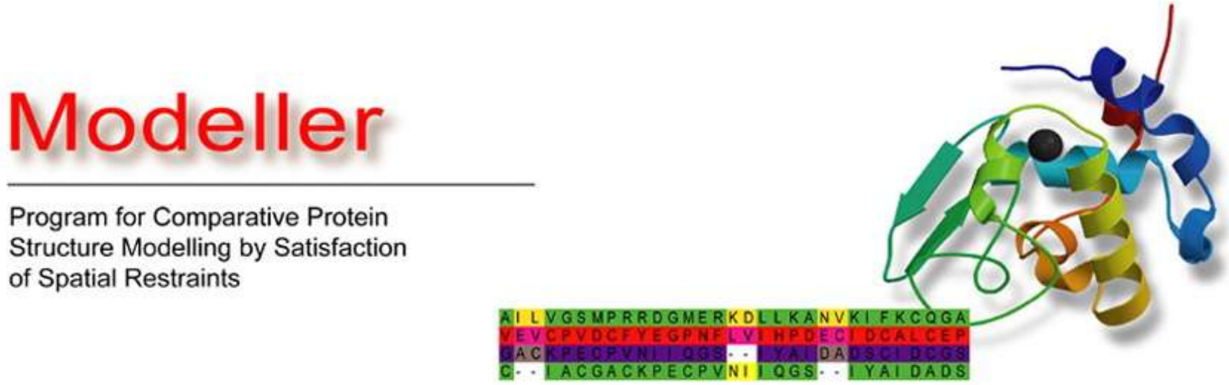
2
Utilisation de la structure
Nt
3
Recherche dans une banque de fragments
4
Ct
Utilisation de la structure
Ct
Nt
5

Modèle final

https://salilab.org/modeller/

Key: MODELIRANJE as noted on http://www.cbs.dtu.dk/~blicher/Courses/Homology_modelling_tutorial.pdf

HOW TO PLAY WITH MODELLER



Modeller

Program for Comparative Protein Structure Modelling by Satisfaction of Spatial Restraints

AILVGSMPRRDGMERKDLLKANVKIFKCQGA
VEVCPVDCFYEGPNFLVIHPDECIDCALCEP
GACKPECPVNIIQGS--IYAIDADSCIDCGS
C--IACGACKPECPVNIIQGS--IYAIDADS

Comparative modelling

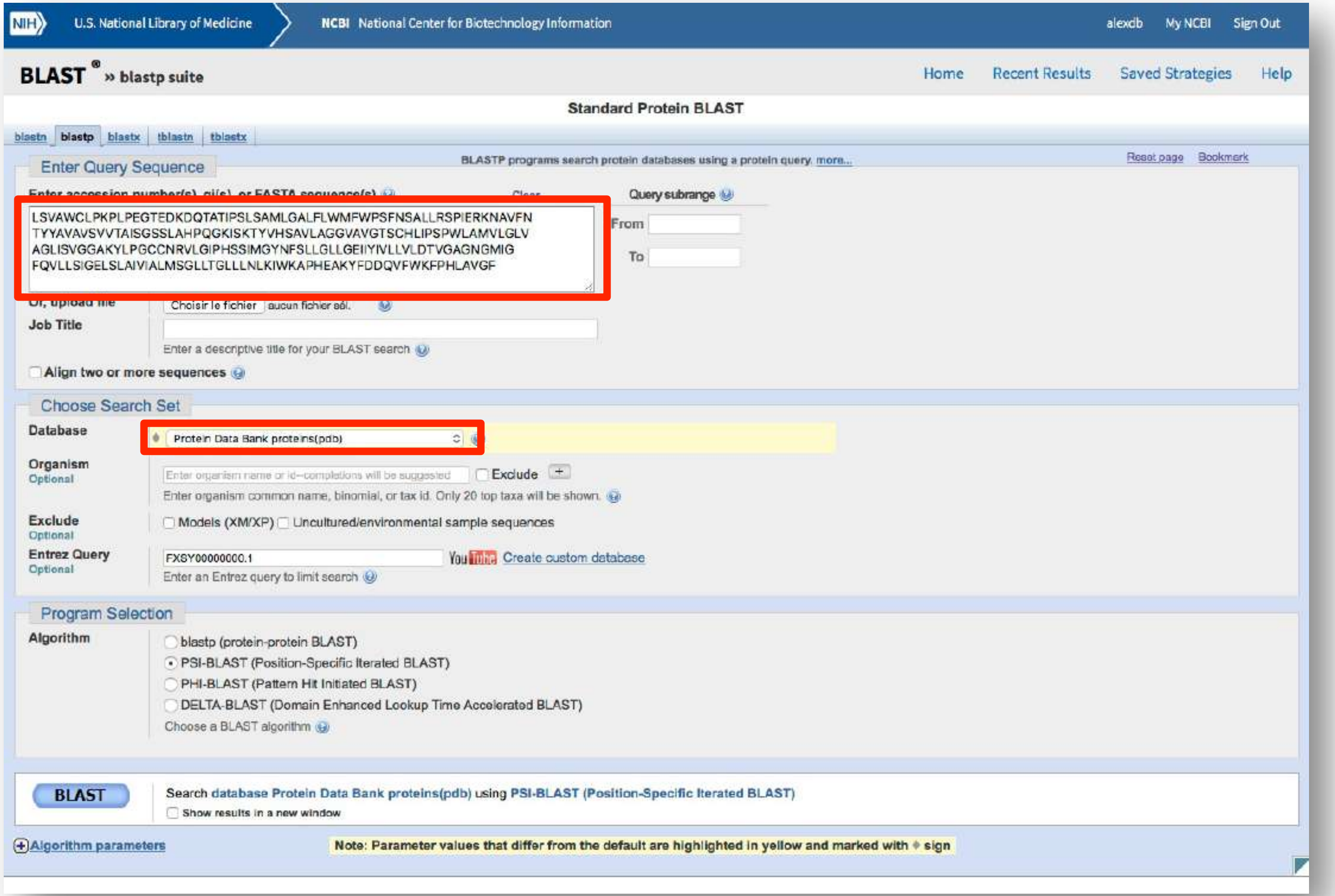
1. You need a sequence.

RhD protein ➔ UniProtKB - Q02161 (RHD_HUMAN)
http://www.uniprot.org/uniprot/Q02161

```
>sp|Q02161|RHD_HUMAN Blood group Rh(D) polypeptide OS=Homo sapiens GN=RHD PE=1 SV=3
MSSKYPRSVRRCLPLWALTLEAALILLFYFFTHYDASLEDQKGLVASYQVGQDLTVMAAI
GLGFLTSSFRRHSWSSVAFNLFMLALGVQWAILLDGFLSQFPSGKVVITLFSIRLATMSA
LSVLISVDAVLGKVNLAQLVVMVLVEVTALGNLRMVISNIFNTDYHMNMMHIYVFAAYFG
LSVAWCLPKPLPEGTEDKDQTATIPSLSAMLGALFLWMFWPSFNSALLRSPIERKNAVFN
TYYAVAVSVVTAISGSSLAHPQGKISKTYVHSAVLAGGVAVGTSCHLIPSPWLAMVLGLV
AGLISVGGAKYLPGCCNRVLGIPHSSIMGYNFSLLGLLGEIIYIVLLVLDTVGAGNGMIG
FQVLLSIGELSLAIVIALMSGLLTGLLLNLKIWKAPHEAKYFDDQVFWKFPHLAVGF
```

Comparative modelling

2. You need a sequence not too far away (with a structure).



Standard Protein BLAST

BLASTP programs search protein databases using a protein query. more...

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s)

```
LSVAWCLPKPLPEGTEDKDQTATIPSLSAMLGALFLWMFWPSFNSALLRSPIERKNAVFN
TYYAVAVSVVTAISGSSLAHPQGKISKTYVHSAVLAGGVAVGTSCHLIPSPWLAMVLGLV
AGLISVGGAKYLPGCCNRVLGIPHSSIMGYNFSLLGLLGEIIYIVLLVLDTVGAGNGMIG
FQVLLSIGELSLAIVIALMSGLLTGLLLNLKIWKAPHEAKYFDDQVFWKFPHLAVGF
```

Query subrange
From
To

Or, upload file: Choisir le fichier aucun fichier sél.

Job Title:

Enter a descriptive title for your BLAST search

☐ Align two or more sequences

Choose Search Set

Database: Protein Data Bank proteins(pdb)

Organism (Optional): Enter organism name or id--completions will be suggested ☐ Exclude +

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Exclude (Optional): ☐ Models (XM/XP) ☐ Uncultured/environmental sample sequences

Entrez Query (Optional): FXSY00000000.1

Enter an Entrez query to limit search

Program Selection

Algorithm:
○ blastp (protein-protein BLAST)
● PSI-BLAST (Position-Specific Iterated BLAST)
○ PHI-BLAST (Pattern Hit Initiated BLAST)
○ DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

Choose a BLAST algorithm

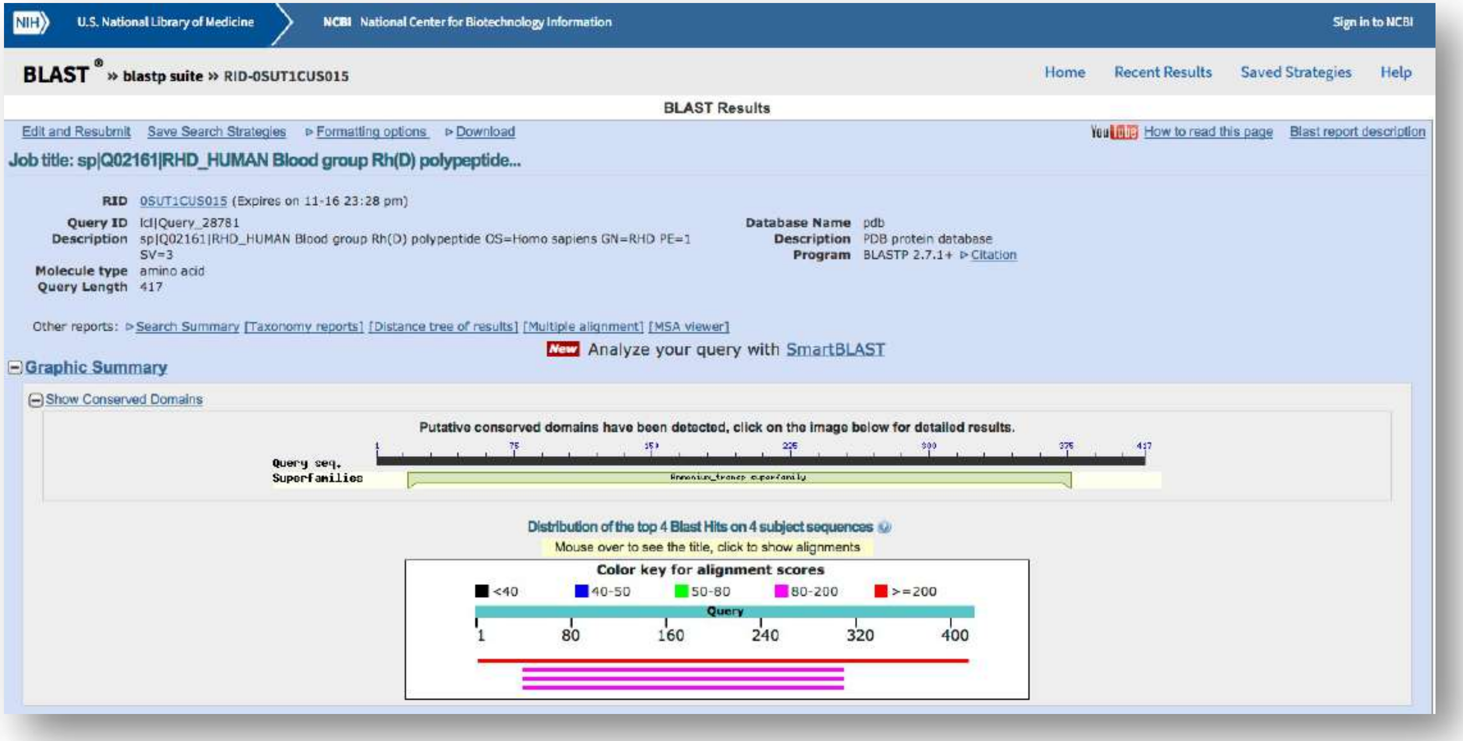
Search database Protein Data Bank proteins(pdb) using PSI-BLAST (Position-Specific Iterated BLAST)

☐ Show results in a new window

Note: Parameter values that differ from the default are highlighted in yellow and marked with ♦ sign

Comparative modelling

3. Analysis of the results



NIH
U.S. National Library of Medicine
NCBI National Center for Biotechnology Information
Sign in to NCBI
BLAST® » blastp suite » RID-0SUT1CUS015
Home
Recent Results
Saved Strategies
Help
Edit and Resubmit
Save Search Strategies
▷ Formatting options
▷ Download
YouTube How to read this page
Blast report description

BLAST Results

Job title: sp|Q02161|RHD_HUMAN Blood group Rh(D) polypeptide...

RID	0SUT1CUS015 (Expires on 11-16 23:28 pm)		
Query ID	lcl|Query_28781	Database Name	pdb
Description	sp|Q02161|RHD_HUMAN Blood group Rh(D) polypeptide OS=Homo sapiens GN=RHD PE=1 SV=3	Description	PDB protein database
Molecule type	amino acid	Program	BLASTP 2.7.1+ ▷ Citation
Query Length	417		

Other reports: ▷ Search Summary [Taxonomy reports] [Distance tree of results] [Multiple alignment] [MSA viewer]

New Analyze your query with SmartBLAST

⊟ Graphic Summary

⊟ Show Conserved Domains

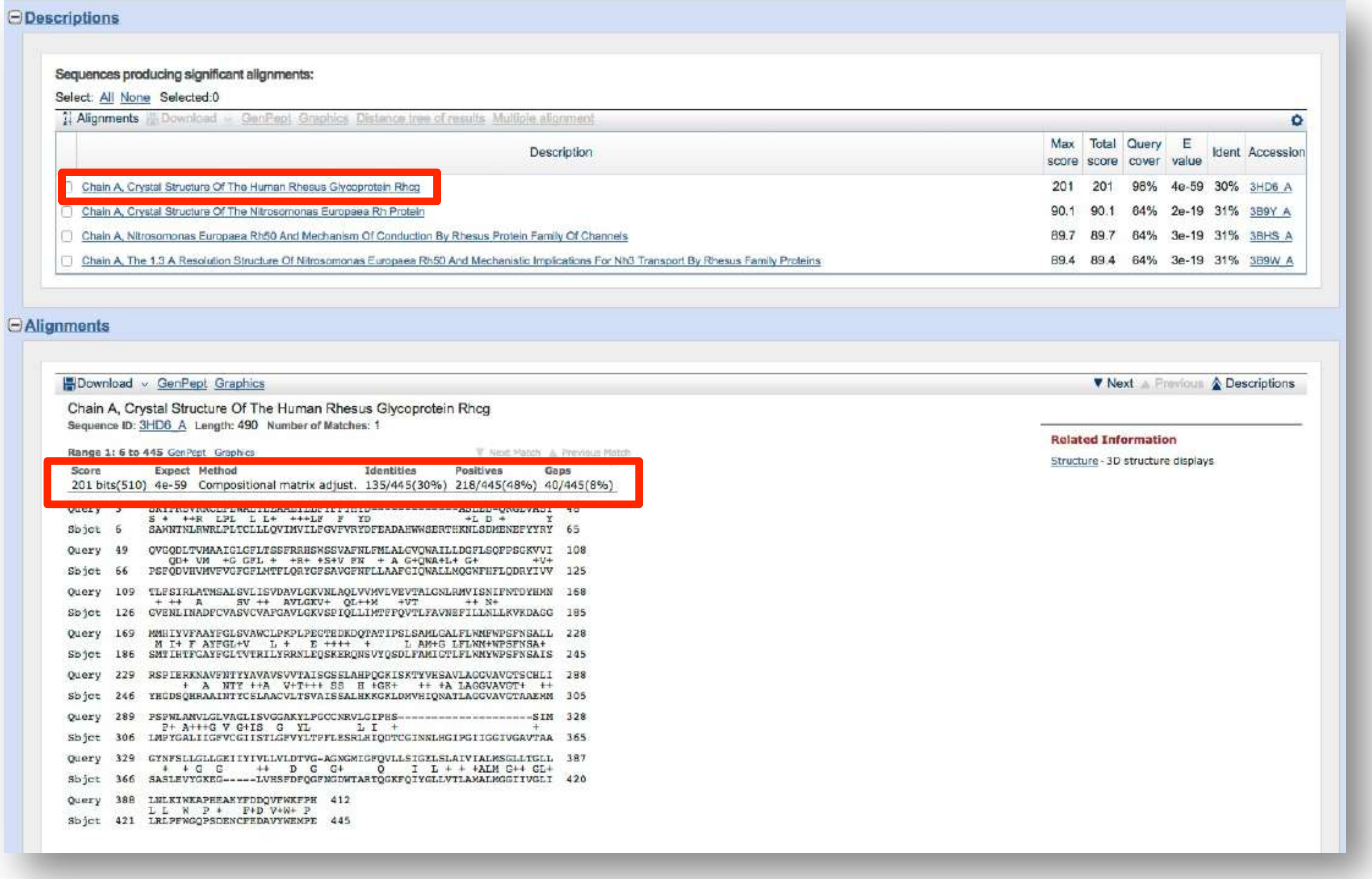
Putative conserved domains have been detected, click on the image below for detailed results.

Distribution of the top 4 Blast Hits on 4 subject sequences

Mouse over to see the title, click to show alignments

Comparative modelling

3. Analysis of the results



Descriptions

Sequences producing significant alignments:

Select: All None Selected:0

Alignments Download GenPept Graphics Distance tree of results Multiple alignment

	Description	Max score	Total score	Query cover	E value	Ident	Accession
□	Chain A, Crystal Structure Of The Human Rhesus Glycoprotein Rhcg	201	201	98%	4e-59	30%	3HD6_A
□	Chain A, Crystal Structure Of The Nitrosomonas Europaea Rh Protein	90.1	90.1	64%	2e-19	31%	3B9Y_A
□	Chain A, Nitrosomonas Europaea Rh50 And Mechanism Of Conduction By Rhesus Protein Family Of Channels	89.7	89.7	64%	3e-19	31%	3BHS_A
□	Chain A, The 1.3 A Resolution Structure Of Nitrosomonas Europaea Rh50 And Mechanistic Implications For Nh3 Transport By Rhesus Family Proteins	89.4	89.4	64%	3e-19	31%	3B9W_A

Alignments

Download GenPept Graphics ▼ Next ▲ Previous ▲ Descriptions

Chain A, Crystal Structure Of The Human Rhesus Glycoprotein Rhcg

Sequence ID: 3HD6_A Length: 490 Number of Matches: 1

Related Information

Structure - 3D structure displays

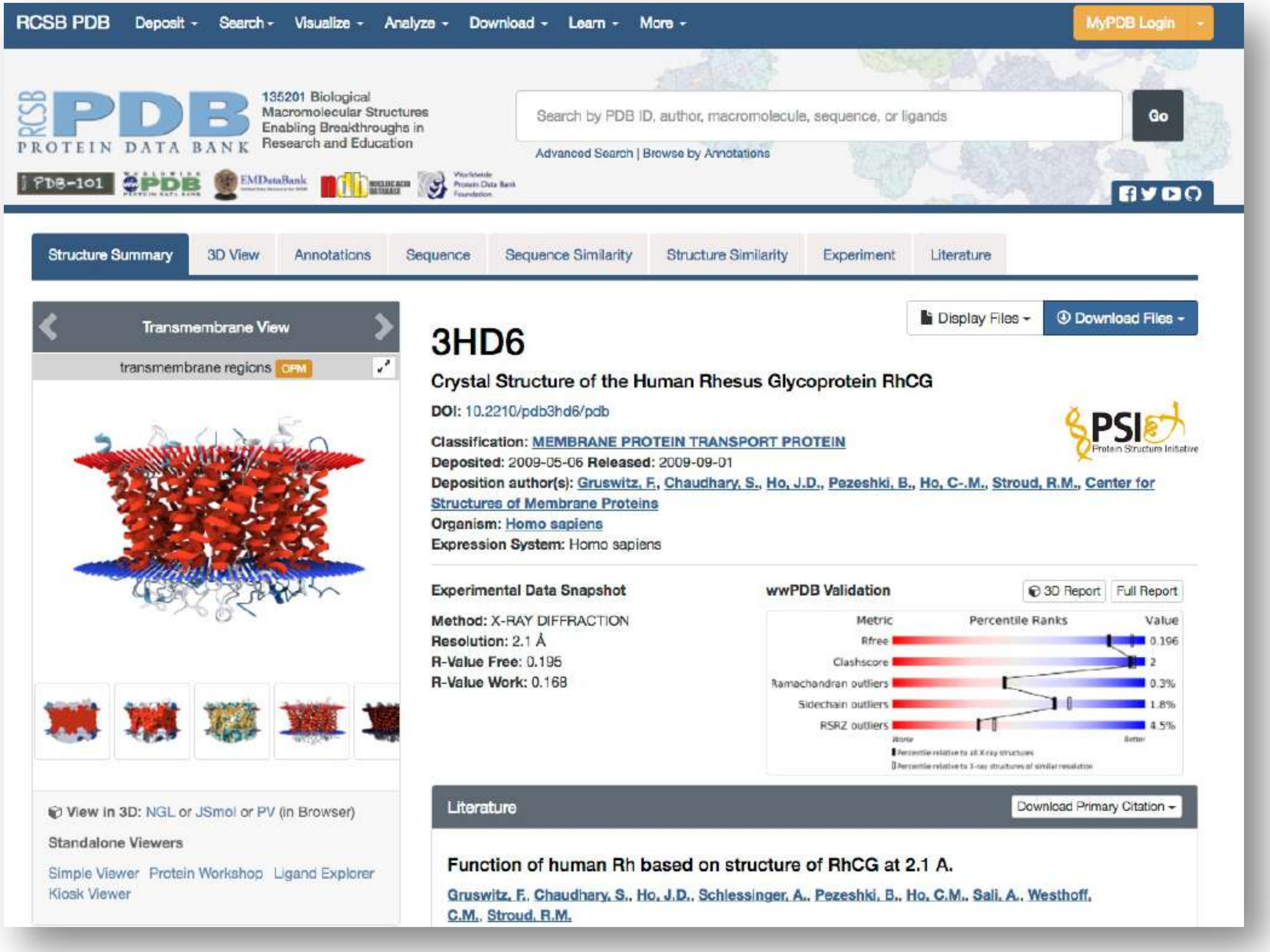
Range 1: 6 to 445 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
201 bits(510)	4e-59	Compositional matrix adjust.	135/445(30%)	218/445(48%)	40/445(8%)

```
Query  5    SKIFRSVRKNCLPIWALTLEAAEILLFITFIHYD---------------ASLED-QNGLVASY  48
            S +    ++R  LPL   L L+   +++LF   F   YD                +L D +       Y
Sbjct  6    SAWNTNLRWRLPLTCLLLQVIMVILFGVFVRYDFEADAHWWSERTHKNLSDMENEFYYRY  65

Query  49   QVGQDLTVMAAIGLGFLTSSFRRHSWSSVAFNLFMLALGVQWAILLDGFLSQFPSGKVVI  108
               QD+ VM  +G GFL +  +R+ +S+V FN  + A G+QWA+L+ G+         +V+
Sbjct  66   PSFQDVHVMVFVGFGFLMTFLQRYGFSAVGFNFLLAAFGIQWALLMQGWFHFLQDRYIVV  125

Query  109  TLFSIRLATMSALSVLISVDAVLGKVNLAQLVVMVLVEVTALGNLRMVISNIFNTDYHMN  168
             + ++  A      SV ++  AVLGKV+  QL++M   +VT          ++ N+
Sbjct  126  GVENLINADFCVASVCVAFGAVLGKVSPIQLLIMTFFQVTLFAVNEFILLNLLKVKDAGG  185

Query  169  MMHIYVFAAYFGLSVAWCLPKPLPEGTEDKDQTATIPSLSAMLGALFLWMFWPSFNSALL  228
             M I+ F AYFGL+V    L +    E ++++   +        L AM+G LFLWM+WPSFNSA+
Sbjct  186  SMIIHTFGAYFGLTVTRILYRRNLEQSKERQNSVYQSDLFAMIGTLFLWMYWPSFNSAIS  245

Query  229  RSPIERKNAVFNTYYAVAVSVVTAISGSSLAHPQGKISKTYVHSAVLAGGVAVGTSCHLI  288
                 +   A  NTY ++A   V+T+++ SS   H +GK+    ++ +A LAGGVAVGT+  ++
Sbjct  246  YHGDSQHRAAINTYCSLAACVLTSVAISSALHKKGKLDMVHIQNATLAGGVAVGTAAEMM  305

Query  289  PSPWLAMVLGLVAGLISVGGAKYLPGCCNRVLGIPHS--------------------SIM  328
              P+ A+++G V G+IS  G  YL          L I  +                     +
Sbjct  306  LMPYGALIIGFVCGIISTLGFVYLTPFLESRLHIQDTCGINNLHGIPGIIGGIVGAVTAA  365

Query  329  GYNFSLLGLLGEIIYIVLLVLDTVG-AGNGMIGFQVLLSIGELSLAIVIALMSGLLTGLL  387
              +  + G  G      ++    D  G  G+      Q     I  L + + +ALM G++ GL+
Sbjct  366  SASLEVYGKEG-----LVHSFDFQGFNGDWTARTQGKFQIYGLLVTLAMALMGGIIVGLI  420

Query  388  LNLKIWKAPHEAKYFDDQVFWKFPH  412
            L L   W   P +      F+D  V+W+  P
Sbjct  421  LRLPFWGQPSDENCFEDAVYWEMPE  445
```

Comparative modelling

4. Selection of the structural template



RCSB PDB Deposit ▾ Search ▾ Visualize ▾ Analyze ▾ Download ▾ Learn ▾ More ▾ MyPDB Login

RCSB PDB PROTEIN DATA BANK

135201 Biological Macromolecular Structures Enabling Breakthroughs in Research and Education

Search by PDB ID, author, macromolecule, sequence, or ligands Go

Advanced Search | Browse by Annotations

PDB-101 · Worldwide PDB · EMDataBank · Nucleic Acid Database · Worldwide Protein Data Bank Foundation

Structure Summary | 3D View | Annotations | Sequence | Sequence Similarity | Structure Similarity | Experiment | Literature

Transmembrane View

transmembrane regions OPM

Display Files ▾ Download Files ▾

3HD6

Crystal Structure of the Human Rhesus Glycoprotein RhCG

DOI: 10.2210/pdb3hd6/pdb

PSI Protein Structure Initiative

Classification: MEMBRANE PROTEIN TRANSPORT PROTEIN

Deposited: 2009-05-06 **Released:** 2009-09-01

Deposition author(s): Gruswitz, F., Chaudhary, S., Ho, J.D., Pezeshki, B., Ho, C.-M., Stroud, R.M., Center for Structures of Membrane Proteins

Organism: Homo sapiens

Expression System: Homo sapiens

Experimental Data Snapshot

Method: X-RAY DIFFRACTION

Resolution: 2.1 Å

R-Value Free: 0.195

R-Value Work: 0.168

wwPDB Validation 3D Report Full Report

Metric	Value
Rfree	0.196
Clashscore	2
Ramachandran outliers	0.3%
Sidechain outliers	1.8%
RSRZ outliers	4.5%

Percentile Ranks
Worse
Better
Percentile relative to all X-ray structures
Percentile relative to X-ray structures of similar resolution

View in 3D: NGL or JSmol or PV (in Browser)

Standalone Viewers

Simple Viewer Protein Workshop Ligand Explorer Kiosk Viewer

Literature Download Primary Citation ▾

Function of human Rh based on structure of RhCG at 2.1 A.

Gruswitz, F., Chaudhary, S., Ho, J.D., Schlessinger, A., Pezeshki, B., Ho, C.M., Sali, A., Westhoff, C.M., Stroud, R.M.

Comparative modelling

4. Selection of the structural template: now the sequence

```
>3HD6:A|PDBID|CHAIN|SEQUENCE
GPSSPSAWNTNLRWRLPLTCLLLQVIMVILFGVFVRYDFEADAHWWSERTHKNLSDMENEFYYRYPSFQDVHVMVFVGFG
FLMTFLQRYGFSAVGFNFLLAAFGIQWALLMQGWFHFLQDRYIVVGVENLINADFCVASVCVAFGAVLGKVSPIQLLIMT
FFQVTLFAVNEFILLNLLKVKDAGGSMTIHTFGAYFGLTVTRILYRRNLEQSKERQNSVYQSDLFAMIGTLFLWMYWPSF
NSAISYHGDSQHRAAINTYCSLAACVLTSVAISSALHKKGKLDMVHIQNATLAGGVAVGTAAEMMLMPYGALIIGFVCGI
ISTLGFVYLTPFLESRLHIQDTCGINNLHGIPGIIGGIVGAVTAASASLEVYGKEGLVHSFDFQGFNGDWTARTQGKFQI
YGLLVTLAMALMGGIIVGLILRLPFWGQPSDENCFEDAVYWEMPEGNSTVYIPEDPTFKPSGPSVPSVPMVSPLPMASSV
PLVPGGLVPR
```

Comparative modelling

5. A new alignment:

```
>3HD6:A|PDBID|CHAIN|SEQUENCE
GPSSPSAWNTNLRWRLPLTCLLLQVIMVILFGVFVRYDFEADAHWWSERTHKNLSDMENEFYYRYPSFQDVHVMVFVGFG
FLMTFLQRYGFSAVGFNFLLAAFGIQWALLMQGWFHFLQDRYIVVGVENLINADFCVASVCVAFGAVLGKVSPIQLLIMT
FFQVTLFAVNEFILLNLLKVKDAGGSMTIHTFGAYFGLTVTRILYRRNLEQSKERQNSVYQSDLFAMIGTLFLWMYWPSF
NSAISYHGDSQHRAAINTYCSLAACVLTSVAISSALHKKGKLDMVHIQNATLAGGVAVGTAAEMMLMPYGALIIGFVCGI
ISTLGFVYLTPFLESRLHIQDTCGINNLHGIPGIIGGIVGAVTAASASLEVYGKEGLVHSFDFQGFNGDWTARTQGKFQI
YGLLVTLAMALMGGIIVGLILRLPFWGQPSDENCFEDAVYWEMPEGNSTVYIPEDPTFKPSGPSVPSVPMVSPLPMASSV
PLVPGGLVPR
```

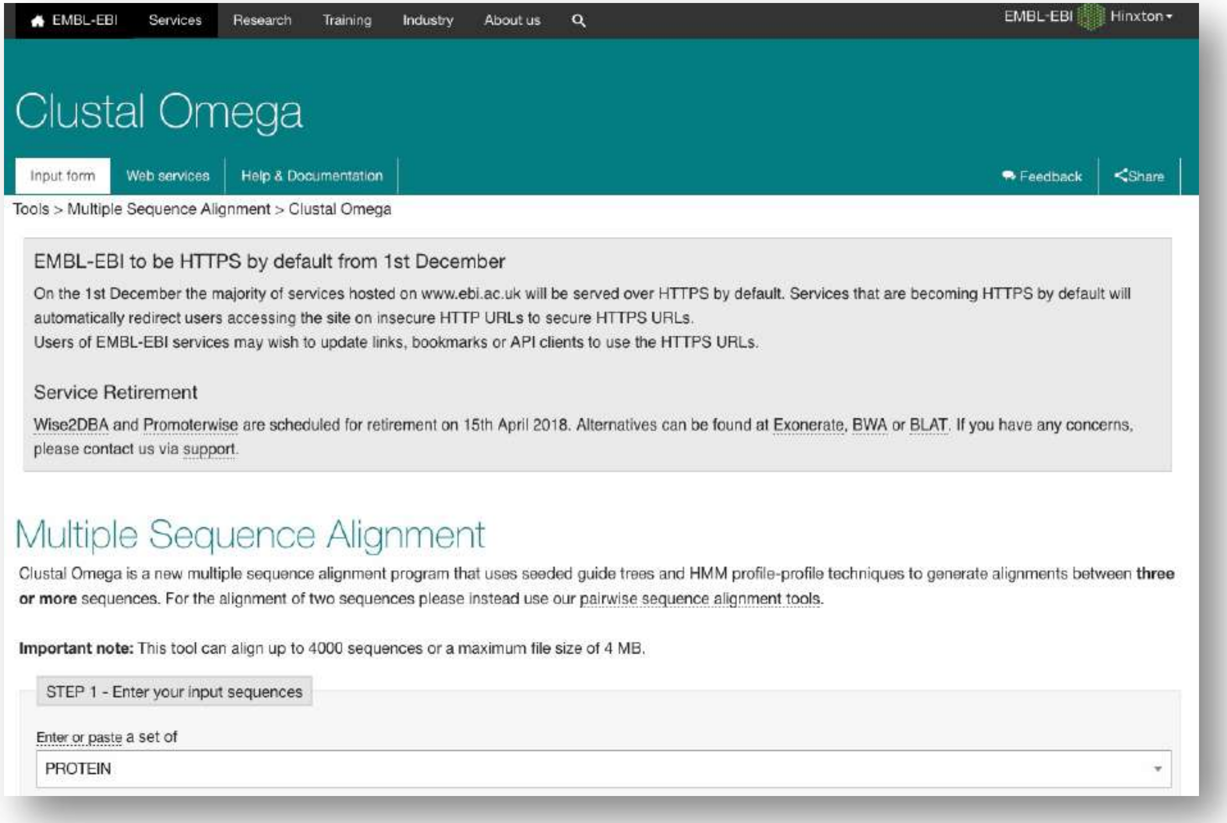
+

```
>sp|Q02161|RHD_HUMAN Blood group Rh(D) polypeptide OS=Homo sapiens GN=RHD PE=1 SV=3
MSSKYPRSVRRCLPLWALTLEAALILLFYFFTHYDASLEDQKGLVASYQVGQDLTVMAAI
GLGFLTSSFRRHSWSSVAFNLFMLALGVQWAILLDGFLSQFPSGKVVITLFSIRLATMSA
LSVLISVDAVLGKVNLAQLVVMVLVEVTALGNLRMVISNIFNTDYHMNMMHIYVFAAYFG
LSVAWCLPKPLPEGTEDKDQTATIPSLSAMLGALFLWMFWPSFNSALLRSPIERKNAVFN
TYYAVAVSVVTAISGSSLAHPQGKISKTYVHSAVLAGGVAVGTSCHLIPSPWLAMVLGLV
AGLISVGGAKYLPGCCNRVLGIPHSSIMGYNFSLLGLLGEIIYIVLLVLDTVGAGNGMIG
FQVLLSIGELSLAIVIALMSGLLTGLLLNLKIWKAPHEAKYFDDQVFWKFPHLAVGF
```

Comparative modelling

5. A new alignment:

https://www.ebi.ac.uk/Tools/msa/clustalo/



Clustal Omega

Tools > Multiple Sequence Alignment > Clustal Omega

EMBL-EBI to be HTTPS by default from 1st December

On the 1st December the majority of services hosted on www.ebi.ac.uk will be served over HTTPS by default. Services that are becoming HTTPS by default will automatically redirect users accessing the site on insecure HTTP URLs to secure HTTPS URLs.
Users of EMBL-EBI services may wish to update links, bookmarks or API clients to use the HTTPS URLs.

Service Retirement

Wise2DBA and Promoterwise are scheduled for retirement on 15th April 2018. Alternatives can be found at Exonerate, BWA or BLAT. If you have any concerns, please contact us via support.

Multiple Sequence Alignment

Clustal Omega is a new multiple sequence alignment program that uses seeded guide trees and HMM profile-profile techniques to generate alignments between **three or more** sequences. For the alignment of two sequences please instead use our pairwise sequence alignment tools.

Important note: This tool can align up to 4000 sequences or a maximum file size of 4 MB.

STEP 1 - Enter your input sequences

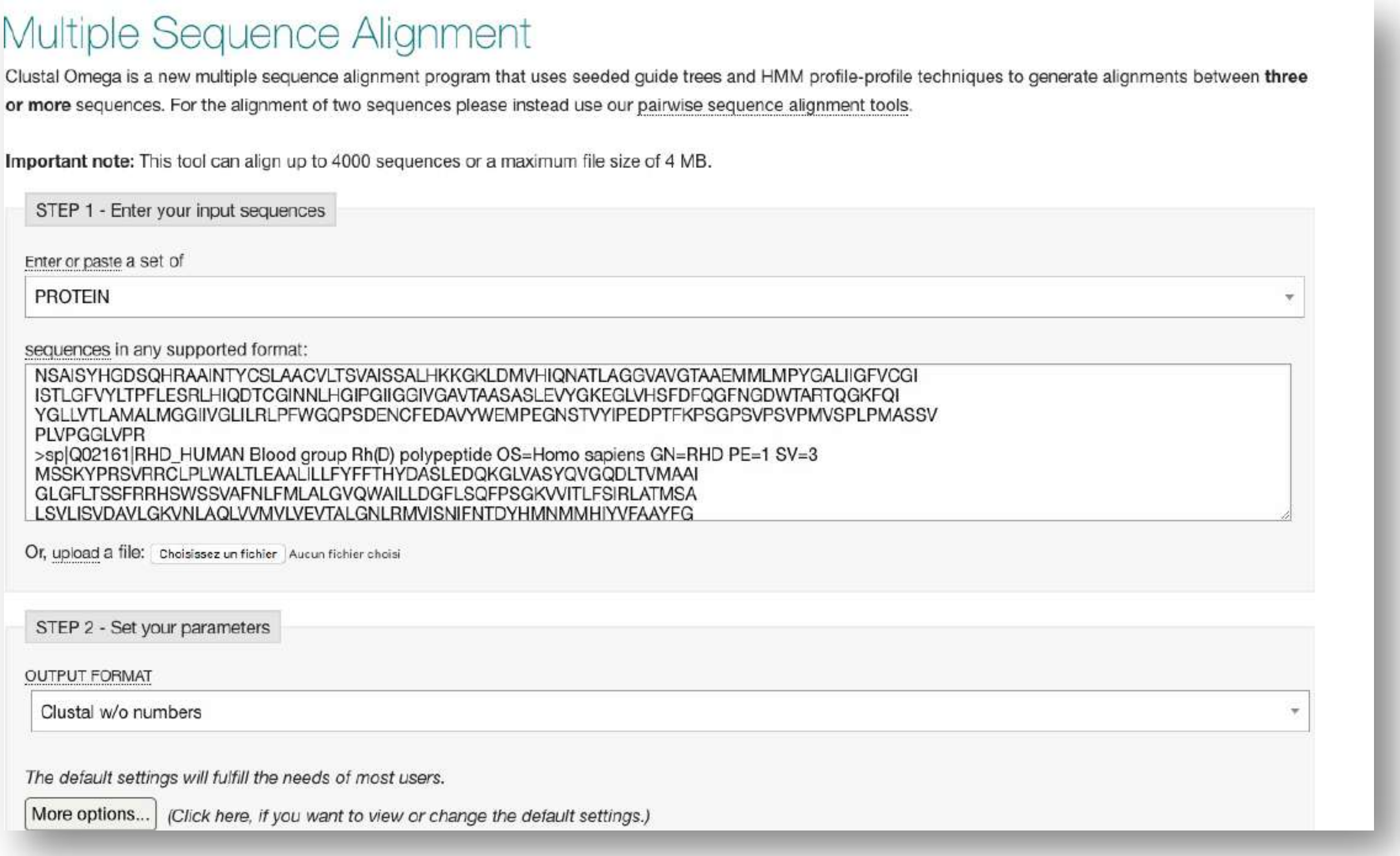
Enter or paste a set of

PROTEIN

Comparative modelling

5. A new alignment:

https://www.ebi.ac.uk/Tools/msa/clustalo/



Multiple Sequence Alignment

Clustal Omega is a new multiple sequence alignment program that uses seeded guide trees and HMM profile-profile techniques to generate alignments between **three or more** sequences. For the alignment of two sequences please instead use our pairwise sequence alignment tools.

Important note: This tool can align up to 4000 sequences or a maximum file size of 4 MB.

STEP 1 - Enter your input sequences

Enter or paste a set of

PROTEIN

sequences in any supported format:

```
NSAISYHGDSQHRAAINTYCSLAACVLTSVAISSALHKKGKLDMVHIQNATLAGGVAVGTAAEMMLMPYGALIIGFVCGI
ISTLGFVYLTPFLESRLHIQDTCGINNLHGIPGIIGGIVGAVTAASASLEVYGKEGLVHSFDFQGFNGDWTARTQGKFQI
YGLLVTLAMALMGGIIVGLILRLPFWGQPSDENCFEDAVYWEMPEGNSTVYIPEDPTFKPSGPSVPSVPMVSPLPMASSV
PLVPGGLVPR
>sp|Q02161|RHD_HUMAN Blood group Rh(D) polypeptide OS=Homo sapiens GN=RHD PE=1 SV=3
MSSKYPRSVRRCLPLWALTLEAALILLFYFFTHYDASLEDQKGLVASYQVGQDLTVMAAI
GLGFLTSSFRRHSWSSVAFNLFMLALGVQWAILLDGFLSQFPSGKVVITLFSIRLATMSA
LSVLISVDAVLGKVNLAQLVVMVLVEVTALGNLRMVISNIFNTDYHMNMMHIYVFAAYFG
```

Or, upload a file:

STEP 2 - Set your parameters

OUTPUT FORMAT

Clustal w/o numbers

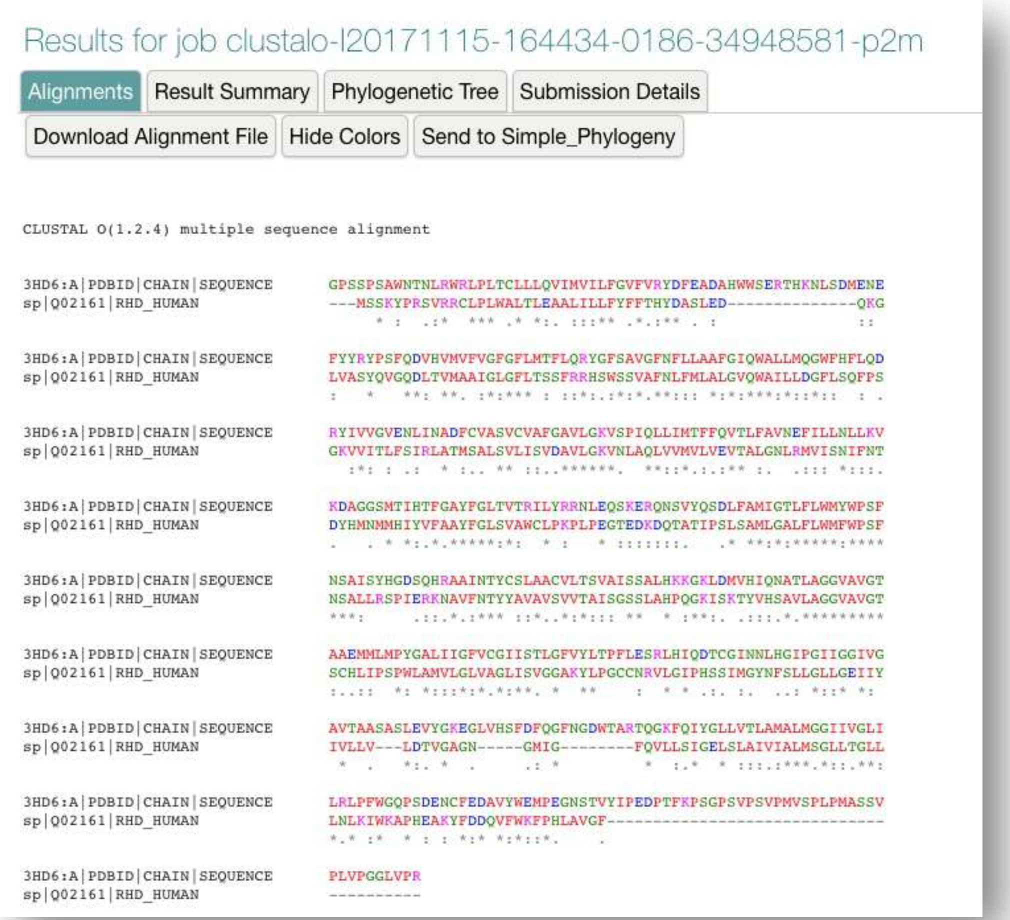
The default settings will fulfill the needs of most users.

(Click here, if you want to view or change the default settings.)

Comparative modelling

5. A new alignment: the results

https://www.ebi.ac.uk/Tools/msa/clustalo/



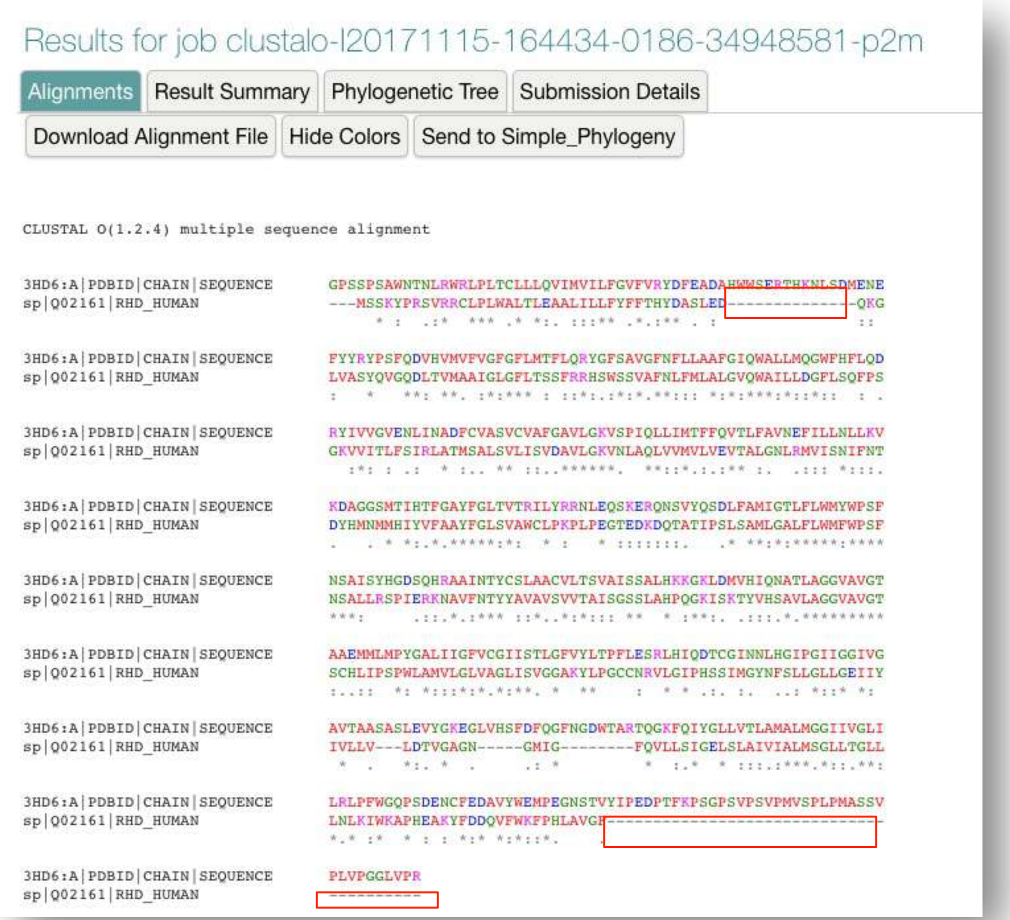
Results for job clustalo-I20171115-164434-0186-34948581-p2m

```
CLUSTAL O(1.2.4) multiple sequence alignment


3HD6:A|PDBID|CHAIN|SEQUENCE      GPSSPSAWNTNLRWRLPLTCLLLQVIMVILFGVFVRYDFEADAHWWSERTHKNLSDMENE
sp|Q02161|RHD_HUMAN              ---MSSKYPRSVRRCLPLWALTLEAALILLFYFFTHYDASLED--------------QKG
                                      * :  .:*  *** .* *:. :::** .*.:** . :              ::

3HD6:A|PDBID|CHAIN|SEQUENCE      FYYRYPSFQDVHVMVFVGFGFLMTFLQRYGFSAVGFNFLLAAFGIQWALLMQGWFHFLQD
sp|Q02161|RHD_HUMAN              LVASYQVGQDLTVMAAIGLGFLTSSFRRHSWSSVAFNLFMLALGVQWAILLDGFLSQFPS
                                 :   *   **: **. :*:*** : ::*:.:*:*.**::: *:*:***:*::*::  : .

3HD6:A|PDBID|CHAIN|SEQUENCE      RYIVVGVENLINADFCVASVCVAFGAVLGKVSPIQLLIMTFFQVTLFAVNEFILLNLLKV
sp|Q02161|RHD_HUMAN              GKVVITLFSIRLATMSALSVLISVDAVLGKVNLAQLVVMVLVEVTALGNLRMVISNIFNT
                                   :*: : .:  * :.. ** ::..******.  **::*.:.:** :.  .::: *:::.

3HD6:A|PDBID|CHAIN|SEQUENCE      KDAGGSMTIHTFGAYFGLTVTRILYRRNLEQSKERQNSVYQSDLFAMIGTLFLWMYWPSF
sp|Q02161|RHD_HUMAN              DYHMNMMHIYVFAAYFGLSVAWCLPKPLPEGTEDKDQTATIPSLSAMLGALFLWMFWPSF
                                 .   . * *:.*.*****:*:  * :   * :::::::.   .* **:*:*****:****

3HD6:A|PDBID|CHAIN|SEQUENCE      NSAISYHGDSQHRAAINTYCSLAACVLTSVAISSALHKKGKLDMVHIQNATLAGGVAVGT
sp|Q02161|RHD_HUMAN              NSALLRSPIERKNAVFNTYYAVAVSVVTAISGSSLAHPQGKISKTYVHSAVLAGGVAVGT
                                 ***:     .::.*.:*** ::*..*:*::: **  * :**:. .:::.*.*********

3HD6:A|PDBID|CHAIN|SEQUENCE      AAEMMLMPYGALIIGFVCGIISTLGFVYLTPFLESRLHIQDTCGINNLHGIPGIIGGIVG
sp|Q02161|RHD_HUMAN              SCHLIPSPWLAMVLGLVAGLISVGGAKYLPGCCNRVLGIPHSSIMGYNFSLLGLLGEIIY
                                 :..::  *: *:::*:*.*:**. *  **    :  * * .:. :.  ..: *::* *:

3HD6:A|PDBID|CHAIN|SEQUENCE      AVTAASASLEVYGKEGLVHSFDFQGFNGDWTARTQGKFQIYGLLVTLAMALMGGIIVGLI
sp|Q02161|RHD_HUMAN              IVLLV---LDTVGAGN-----GMIG--------FQVLLSIGELSLAIVIALMSGLLTGLL
                                  *  .   *:. *  .     .: *        *  :.*  * :::.:***.*::.**:

3HD6:A|PDBID|CHAIN|SEQUENCE      LRLPFWGQPSDENCFEDAVYWEMPEGNSTVYIPEDPTFKPSGPSVPSVPMVSPLPMASSV
sp|Q02161|RHD_HUMAN              LNLKIWKAPHEAKYFDDQVFWKFPHLAVGF------------------------------
                                 *.* :*  * : : *:* *:*::*.    .

3HD6:A|PDBID|CHAIN|SEQUENCE      PLVPGGLVPR
sp|Q02161|RHD_HUMAN              ----------
```

Comparative modelling

5. A new alignment: the results

https://www.ebi.ac.uk/Tools/msa/clustalo/



Results for job clustalo-I20171115-164434-0186-34948581-p2m

```
CLUSTAL O(1.2.4) multiple sequence alignment

3HD6:A|PDBID|CHAIN|SEQUENCE      GPSSPSAWNTNLRWRLPLTCLLLQVIMVILFGVFVRYDFEADAHWWSERTHKNLSDMENE
sp|Q02161|RHD_HUMAN              ---MSSKYPRSVRRCLPLWALTLEAALILLFYFFTHYDASLED--------------QKG
                                      * :  .:*  *** .* *:. :::** .*.:** . :               ::

3HD6:A|PDBID|CHAIN|SEQUENCE      FYYRYPSFQDVHVMVFVGFGFLMTFLQRYGFSAVGFNFLLAAFGIQWALLMQGWFHFLQD
sp|Q02161|RHD_HUMAN              LVASYQVGQDLTVMAAIGLGFLTSSFRRHSWSSVAFNLFMLALGVQWAILLDGFLSQFPS
                                 :   *   **: **. :*:*** : ::*:.:*:*.**::: *:*:***:*:.*::  : .

3HD6:A|PDBID|CHAIN|SEQUENCE      RYIVVGVENLINADFCVASVCVAFGAVLGKVSPIQLLIMTFFQVTLFAVNEFILLNLLKV
sp|Q02161|RHD_HUMAN              GKVVITLFSIRLATMSALSVLISVDAVLGKVNLAQLVVMVLVEVTALGNLRMVISNIFNT
                                   :*: : .:  * :.. ** ::..******.  **::*.:.:** :.  .::: *:::.

3HD6:A|PDBID|CHAIN|SEQUENCE      KDAGGSMTIHTFGAYFGLTVTRILYRRNLEQSKERQNSVYQSDLFAMIGTLFLWMYWPSF
sp|Q02161|RHD_HUMAN              DYHMNMMHIYVFAAYFGLSVAWCLPKPLPEGTEDKDQTATIPSLSAMLGALFLWMFWPSF
                                 .   . * *:.*.*****:*:  * :   * :::::::.   .* **:*:*****:****

3HD6:A|PDBID|CHAIN|SEQUENCE      NSAISYHGDSQHRAAINTYCSLAACVLTSVAISSALHKKGKLDMVHIQNATLAGGVAVGT
sp|Q02161|RHD_HUMAN              NSALLRSPIERKNAVFNTYYAVAVSVVTAISGSSLAHPQGKISKTYVHSAVLAGGVAVGT
                                 ***:     .::.*.:*** ::*..*:*::: **  * :**:. .:::.*.*********

3HD6:A|PDBID|CHAIN|SEQUENCE      AAEMMLMPYGALIIGFVCGIISTLGFVYLTPFLESRLHIQDTCGINNLHGIPGIIGGIVG
sp|Q02161|RHD_HUMAN              SCHLIPSPWLAMVLGLVAGLISVGGAKYLPGCCNRVLGIPHSSIMGYNFSLLGLLGEIIY
                                 :..::  *: *:::*:*.*:**. *  **    :  * * .:. :.  ..: *::* *:

3HD6:A|PDBID|CHAIN|SEQUENCE      AVTAASASLEVYGKEGLVHSFDFQGFNGDWTARTQGKFQIYGLLVTLAMALMGGIIVGLI
sp|Q02161|RHD_HUMAN              IVLLV---LDTVGAGN-----GMIG--------FQVLLSIGELSLAIVIALMSGLLTGLL
                                  *  .   *:. *  .     .: *         *  :.*  * :::.:***.*::.**:

3HD6:A|PDBID|CHAIN|SEQUENCE      LRLPFWGQPSDENCFEDAVYWEMPEGNSTVYIPEDPTFKPSGPSVPSVPMVSPLPMASSV
sp|Q02161|RHD_HUMAN              LNLKIWKAPHEAKYFDDQVFWKFPHLAVGF------------------------------
                                 *.* :*  * : : *:* *:*::*.    .

3HD6:A|PDBID|CHAIN|SEQUENCE      PLVPGGLVPR
sp|Q02161|RHD_HUMAN              ----------
```

Comparative modelling

6. Modeller

a. the script

```
#!/usr/bin/env python
# Homology modeling by the automodel class
from modeller import *                       # Load standard Modeller classes
from modeller.automodel import *             # Load the automodel class
    # Redefine the special_patches routine to include the additional disulfides
    # (this routine is empty by default):
log.verbose()                                # request verbose output
env = environ()                              # create a new MODELLER environment to build this
model in

a = automodel(env,
              alnfile = 'one.ali',          # alignment filename
              knowns   = '3HD6',            # codes of the templates
              sequence = 'PROTEINE-RHD')     # code of the target
a.starting_model= 1                          # index of the first model
a.ending_model = 5                           # index of the last model
                                             # (determines how many models to calculate)
a.make()                                     # do the actual homology modeling
```

Comparative modelling

6. Modeller

b. the real alignement for Modeler

syntaxe

Alignment in PIR format

```
>P1;3HD6
structureX:3HD6:1    :A:443  :A:: : :
-----SAWNTNLRWRLPLTCLLLQVIMVILFGVFVRYDFE----------
--------NEFYYRYPSFQDVHVMVFVGFGFLMTFLQRYGFSAVGFNFLL
AAFGIQWALLMQGWFHFLQDRYIVVGVENLINADFCVASVCVAFGAVLGK
VSPIQLLIMTFFQVTLFAVNEFILLNLLKVKDAGGSMTIHTFGAYFGLTV
TRILYRRNLEQSKERQNSVYQSDLFAMIGTLFLWMYWPSFNSAISYHGDS
QHRAAINTYCSLAACVLTSVAISSALHKKGKLDMVHIQNATLAGGVAVGT
AAEMMLMPYGALIIGFVCGIISTLGFVYLTPFLESRLHIQDTCGINNLHG
IPGIIGGIVGAVTAAS----------------------DWTARTQGKFQI
YGLLVTLAMALMGGIIVGLILRLPFWGQPSDENCFEDAVYWEMPEGNS--
----------------------------------------
*
```

syntaxe

Alignment in PIR format

```
>P1;PROTEINE-RHD
sequence:PROTEINE-RHD:   1 : :  417 : : :: :
---MSSKYPRSVRRCLPLWALTLEAALILLFYFFTHYDASLED-------
-------QKGLVASYQVGQDLTVMAAIGLGFLTSSFRRHSWSSVAFNLFM
LALGVQWAILLDGFLSQFPSGKVVITLFSIRLATMSALSVLISVDAVLGK
VNLAQLVVMVLVEVTALGNLRMVISNIFNTDYHMNMMHIYVFAAYFGLSV
AWCLPKPLPEGTEDKDQTATIPSLSAMLGALFLWMFWPSFNSALLRSPIE
RKNAVFNTYYAVAVSVVTAISGSSLAHPQGKISKTYVHSAVLAGGVAVGT
SCHLIPSPWLAMVLGLVAGLISVGGAKYLPGCCNRVLGIPHSSIMGYNFS
LLGLLGEIIYIVLLVLDTVG----------------AGNGMIGFQVLLSI
GELSLAIVIALMSGLLTGLLLNLKIWKAPHEAKYFDDQVFWKFPHLAVGF
----------------------------------------
*
```

Comparative modelling

6. Modeller

   c. the template

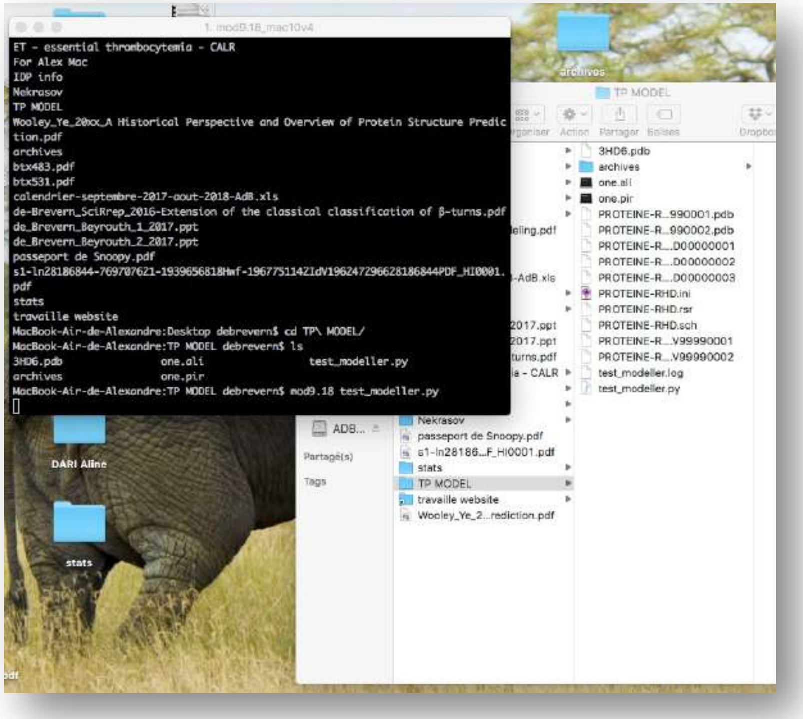
`The PDB …`

Comparative modelling

6. Modeller

   d. now the work

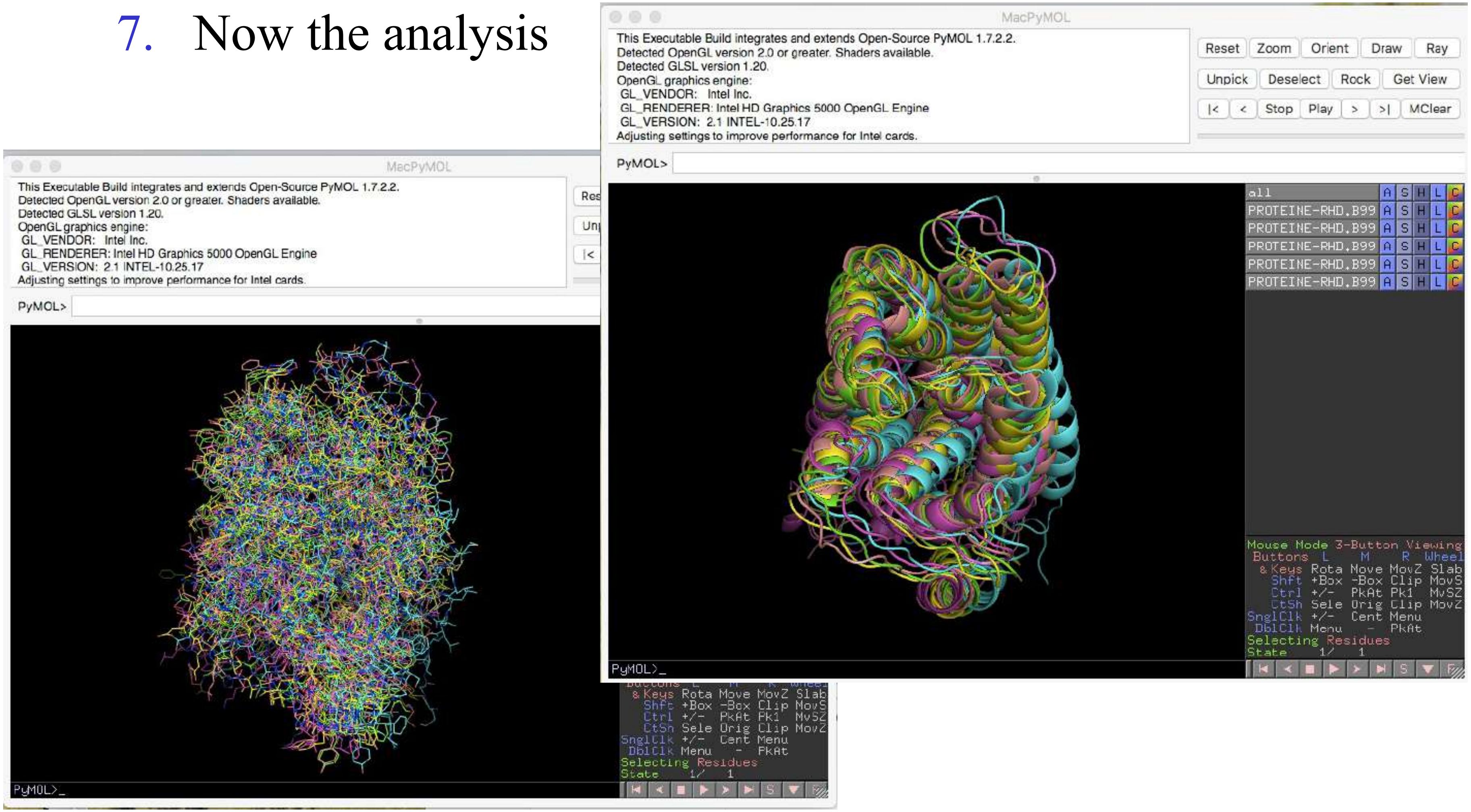
```
> mod9.25 test_modeller.py
```



```
ET - essential thrombocytemia - CALR
For Alex Mac
IDP info
Nekrasov
TP MODEL
Wooley_Ye_20xx_A Historical Perspective and Overview of Protein Structure Predic
tion.pdf
archives
btx483.pdf
btx531.pdf
calendrier-septembre-2017-aout-2018-AdB.xls
de-Brevern_SciRrep_2016-Extension of the classical classification of β-turns.pdf
de_Brevern_Beyrouth_1_2017.ppt
de_Brevern_Beyrouth_2_2017.ppt
passeport de Snoopy.pdf
s1-ln28186844-769707621-1939656818Hwf-1967751142IdV196247296628186844PDF_HI0001.
pdf
stats
travaille website
MacBook-Air-de-Alexandre:Desktop debrevern$ cd TP\ MODEL/
MacBook-Air-de-Alexandre:TP MODEL debrevern$ ls
3HD6.pdb                one.ali                 test_modeller.py
archives                one.pir
MacBook-Air-de-Alexandre:TP MODEL debrevern$ mod9.18 test_modeller.py
▯
```

Comparative modelling

7. Now the analysis

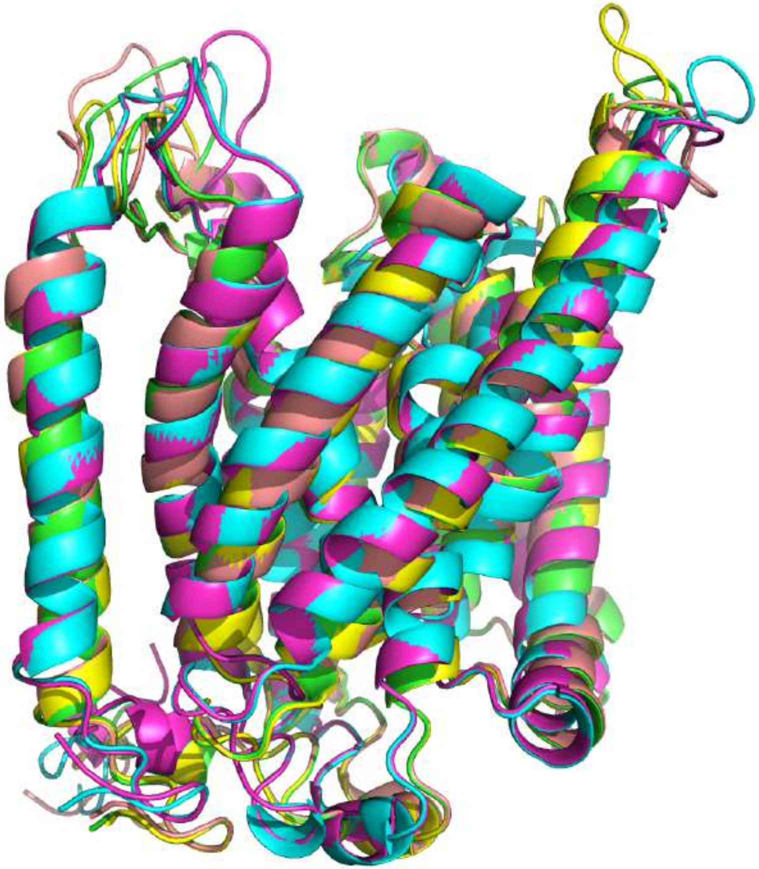


```
This Executable Build integrates and extends Open-Source PyMOL 1.7.2.2.
Detected OpenGL version 2.0 or greater. Shaders available.
Detected GLSL version 1.20.
OpenGL graphics engine:
 GL_VENDOR:   Intel Inc.
 GL_RENDERER: Intel HD Graphics 5000 OpenGL Engine
 GL_VERSION:  2.1 INTEL-10.25.17
Adjusting settings to improve performance for Intel cards.
```

```
This Executable Build integrates and extends Open-Source PyMOL 1.7.2.2.
Detected OpenGL version 2.0 or greater. Shaders available.
Detected GLSL version 1.20.
OpenGL graphics engine:
 GL_VENDOR:   Intel Inc.
 GL_RENDERER: Intel HD Graphics 5000 OpenGL Engine
 GL_VERSION:  2.1 INTEL-10.25.17
Adjusting settings to improve performance for Intel cards.
```

Comparative modelling

7. Now the analysis

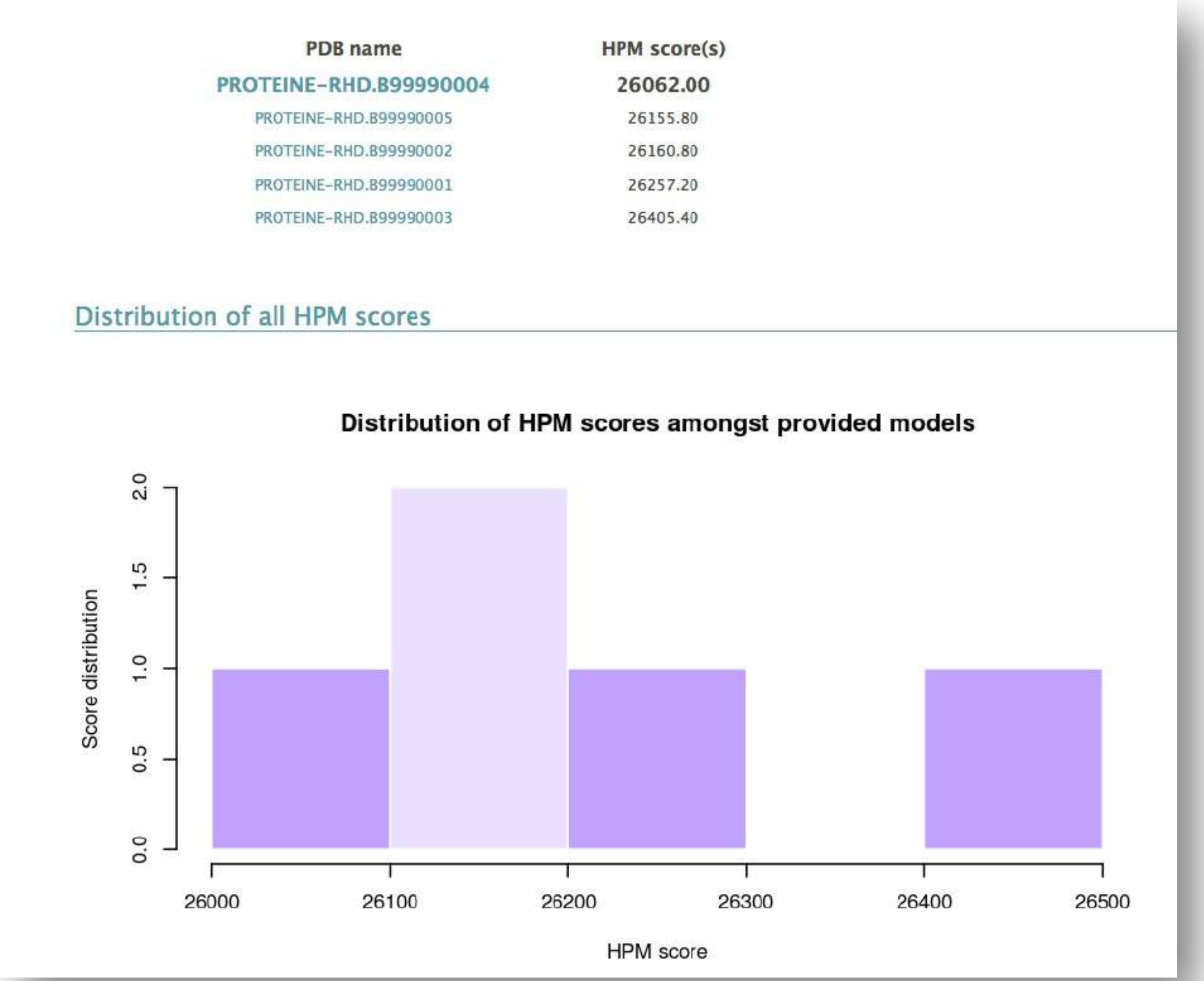


Comparative modelling

- Need a specific assessment

Comparative modelling

7. Now the analysis

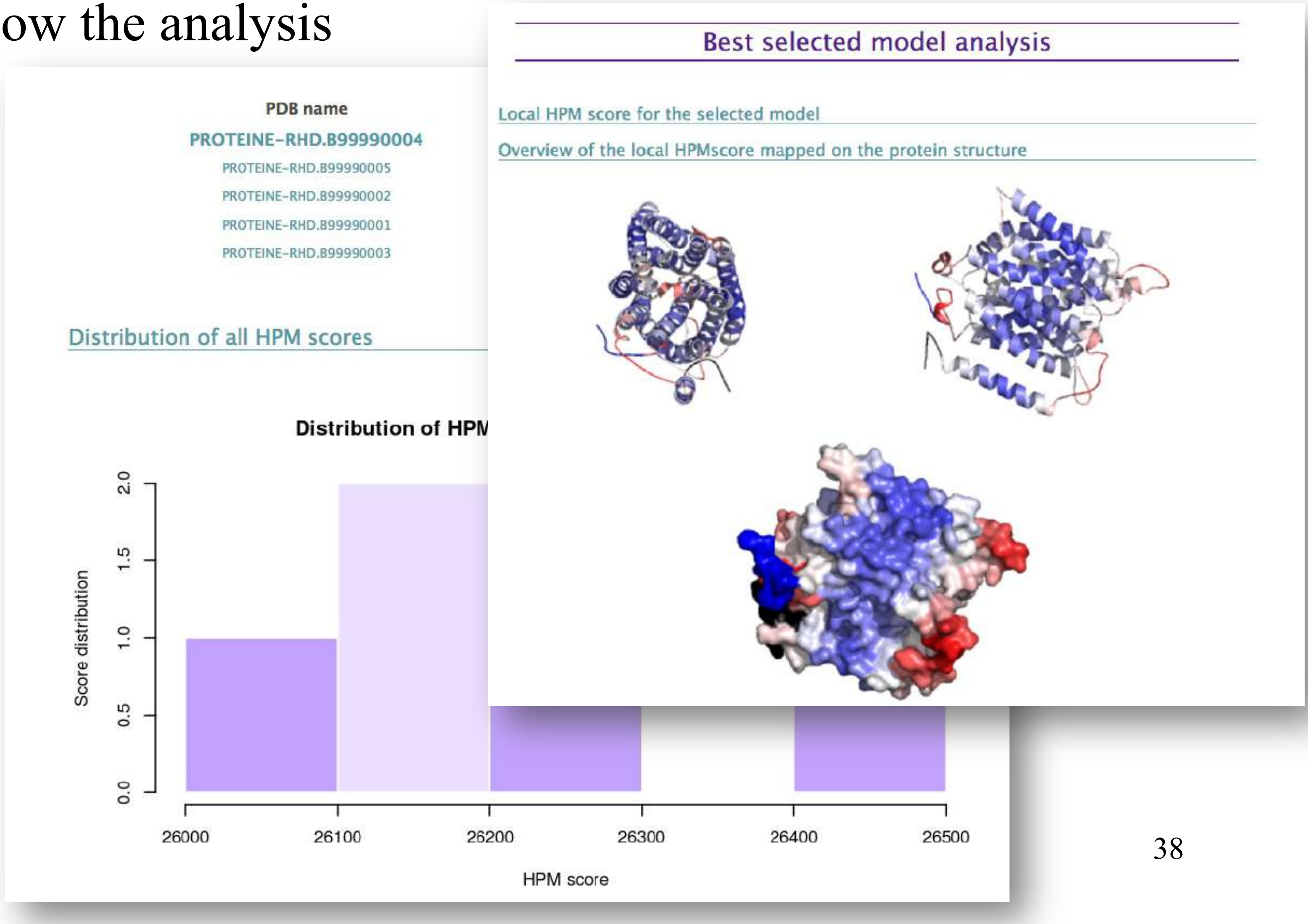


PDB name	HPM score(s)
PROTEINE-RHD.B99990004	**26062.00**
PROTEINE-RHD.B99990005	26155.80
PROTEINE-RHD.B99990002	26160.80
PROTEINE-RHD.B99990001	26257.20
PROTEINE-RHD.B99990003	26405.40

Distribution of all HPM scores

Comparative modelling

7. Now the analysis



PDB name

PROTEINE-RHD.B99990004
PROTEINE-RHD.B99990005
PROTEINE-RHD.B99990002
PROTEINE-RHD.B99990001
PROTEINE-RHD.B99990003

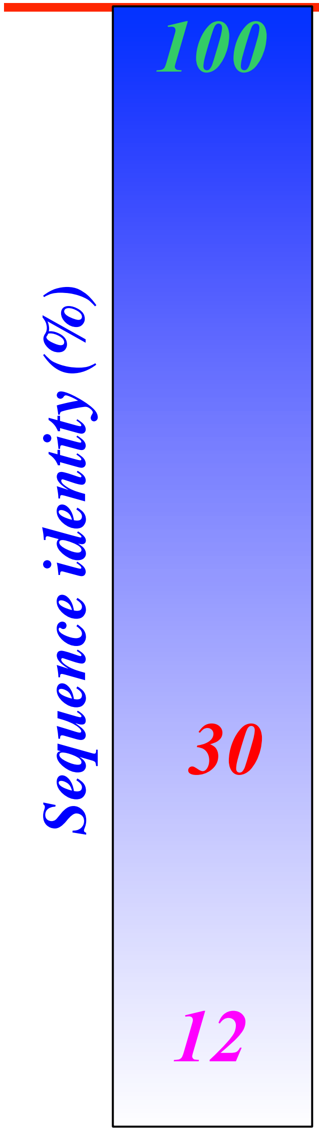
Distribution of all HPM scores

Best selected model analysis

Local HPM score for the selected model

Overview of the local HPMscore mapped on the protein structure

3D modelling

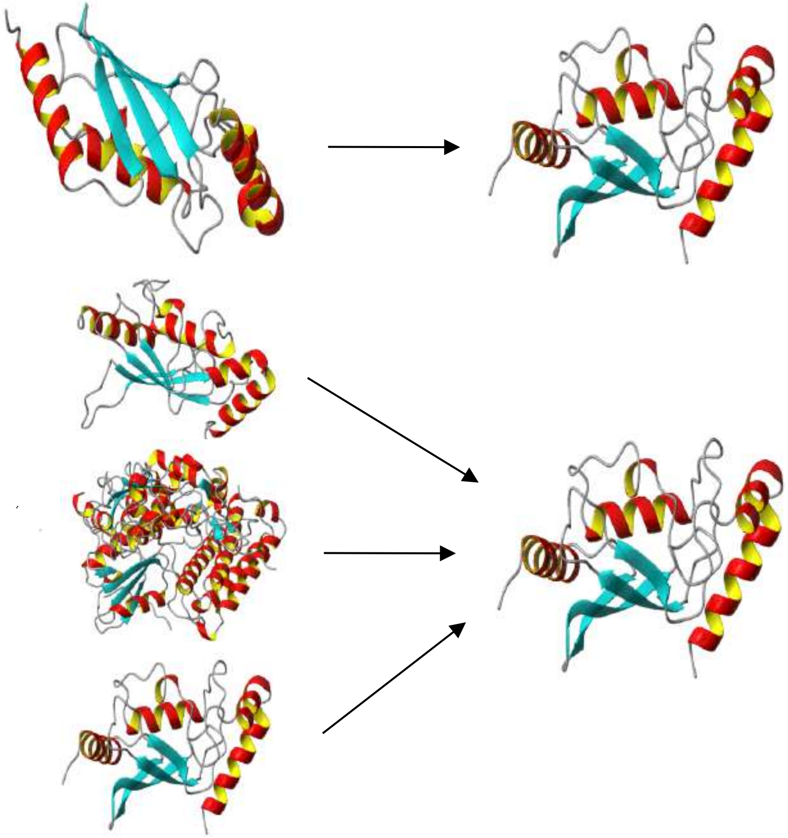


Homology modelling

```
ATPLGLPTHVVVAGLNPHTRESD
ATPLGIPTHVPPAGLNPHTRESD
!!!!! !!!!  !!!!!!!!!!!
```

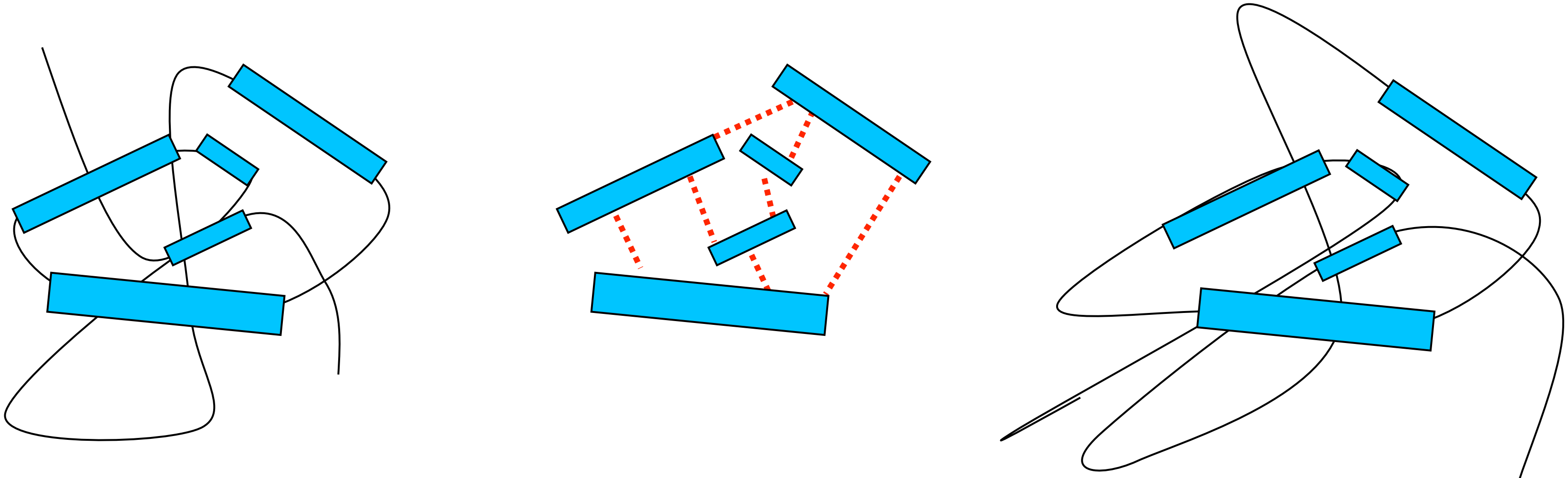
Threading

```
ETPLGLPTHVVVEGLNPHTRESD
IRVLGIPTHVPPIGLNPHTRIID
   !! !!!!   !!!!!!!  !
```



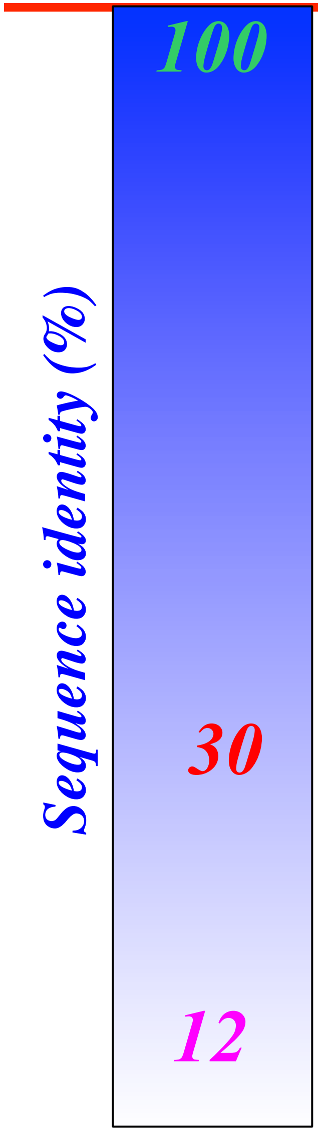
Threading

- The main idea



- Searching for structural similarity => notion of protein core

3D modelling



Homology modelling

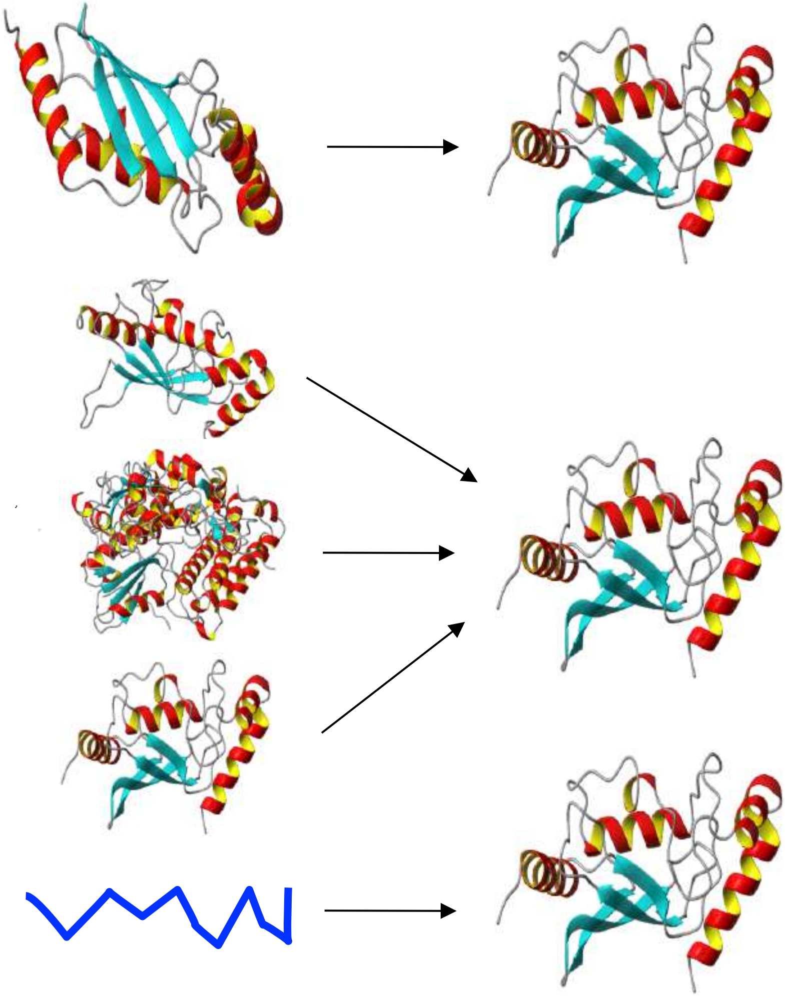
```
ATPLGLPTHVVVAGLNPHTRESD
ATPLGIPTHVPPAGLNPHTRESD
!!!!! !!!!   !!!!!!!!!!!
```

Threading

```
ETPLGLPTHVVVEGLNPHTRESD
IRVLGIPTHVPPIGLNPHTRIID
   !! !!!!    !!!!!!!  !
```

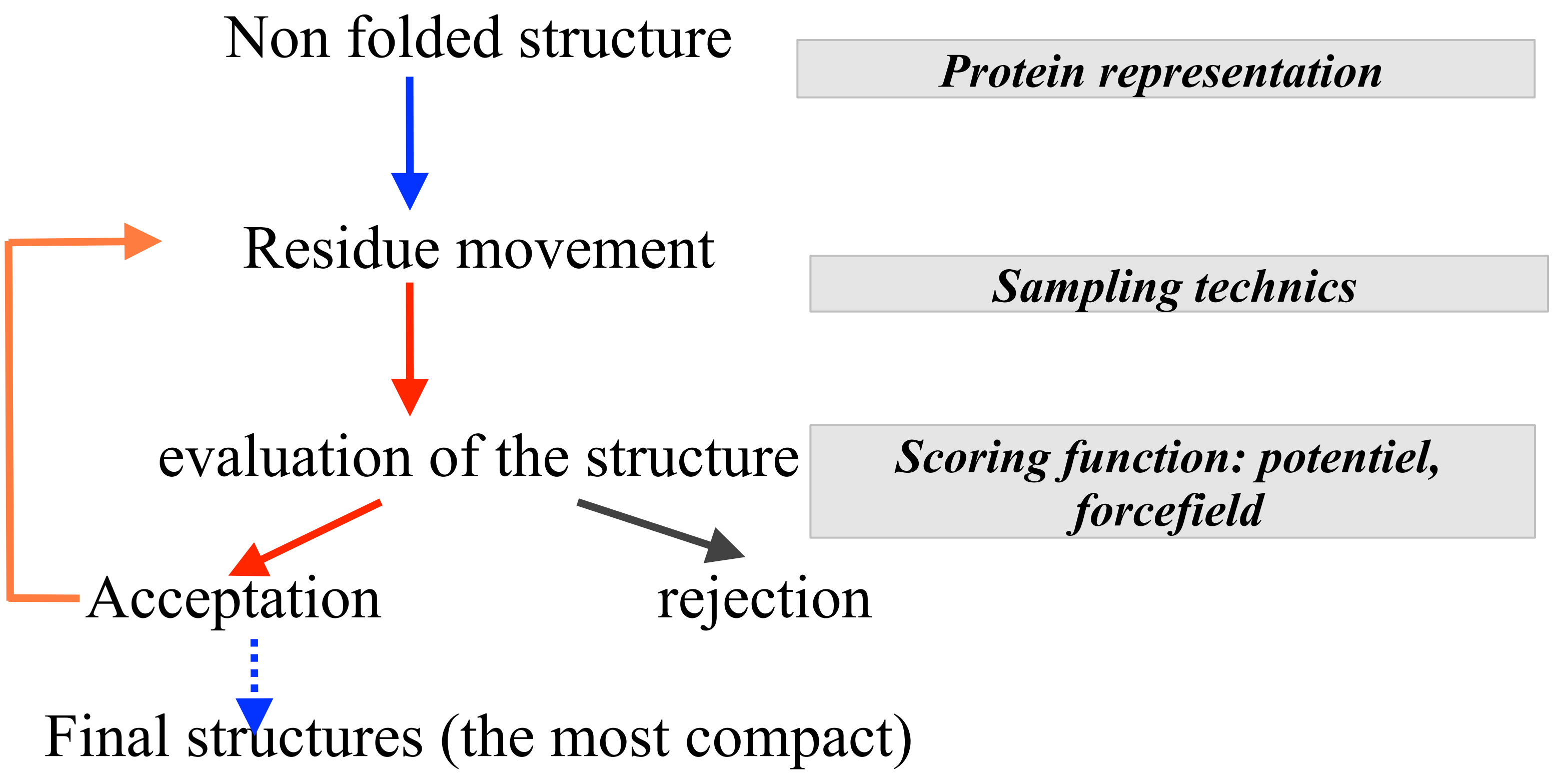
ab initio

```
ETPLGLPPHVVVEGLNPPPRESD
IRVLGIPVHVPPIGPNVVVRIID
   !! ! !!   ! !   !  !
```



ab initio

- **Principle:** the native structure corresponds to a global minima (in terms of energy)



Non folded structure

Protein representation

Residue movement

Sampling technics

evaluation of the structure

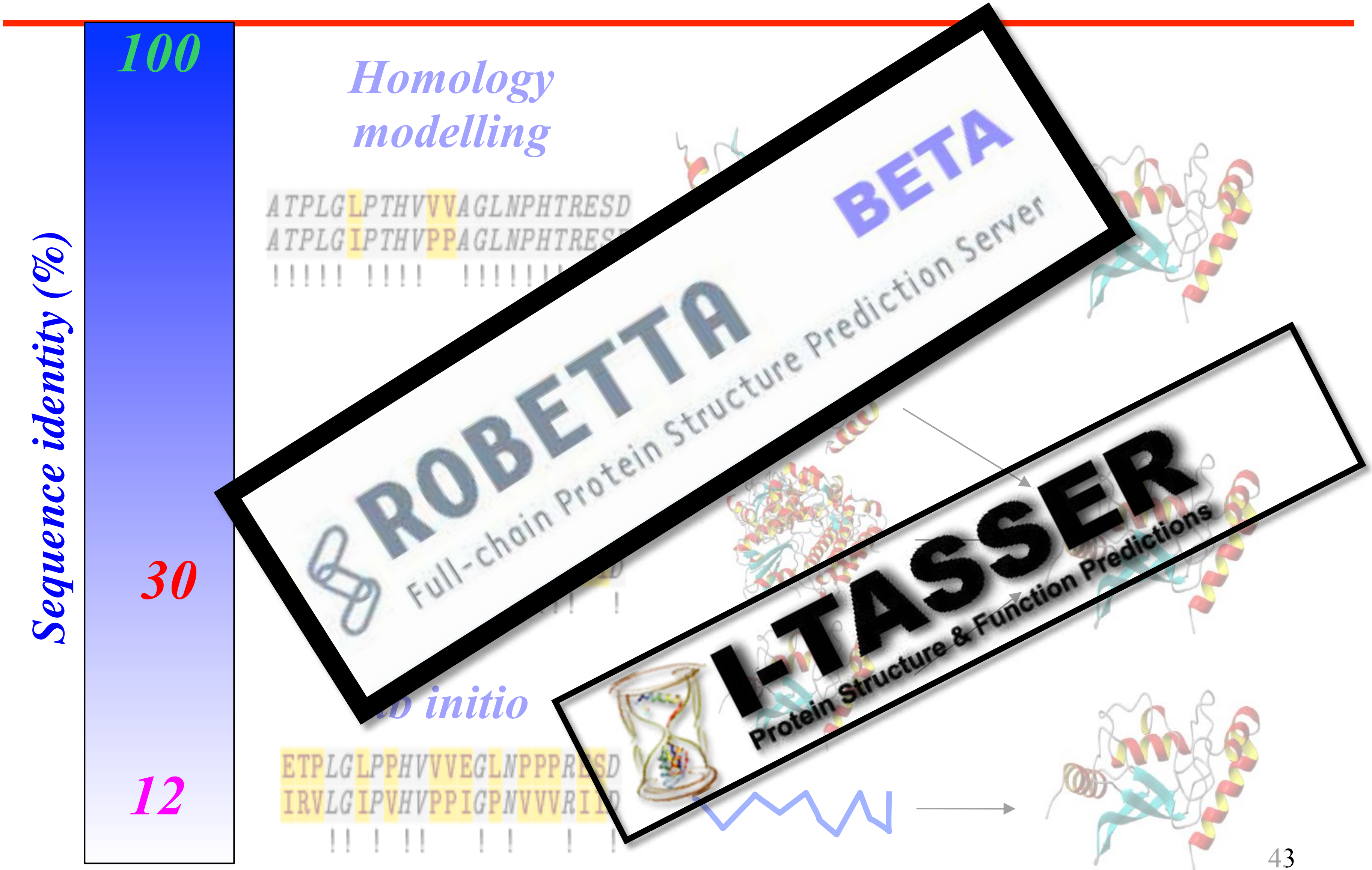
Scoring function: potentiel, forcefield

Acceptation

rejection

Final structures (the most compact)

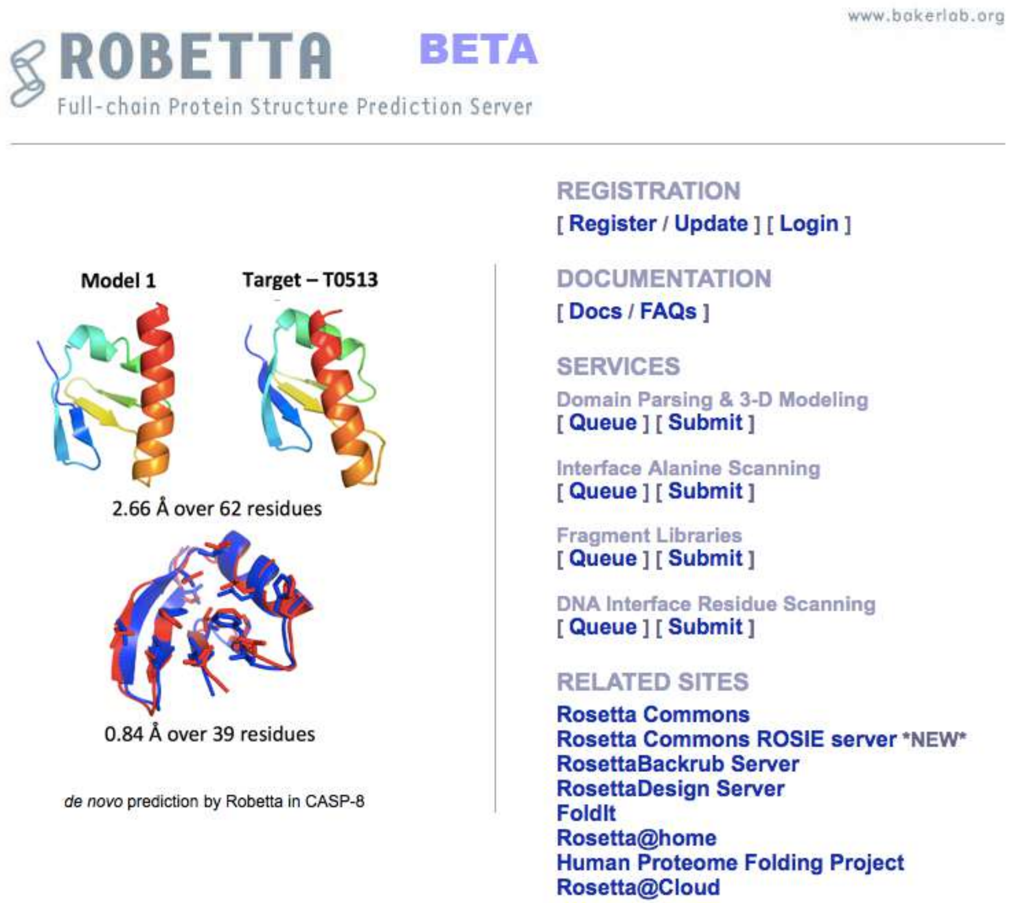
3D modelling



de novo

- Robetta

http://robetta.bakerlab.org



www.bakerlab.org

ROBETTA BETA
Full-chain Protein Structure Prediction Server

Model 1 Target – T0513

2.66 Å over 62 residues

0.84 Å over 39 residues

de novo prediction by Robetta in CASP-8

REGISTRATION
[Register / Update] [Login]

DOCUMENTATION
[Docs / FAQs]

SERVICES
Domain Parsing & 3-D Modeling
[Queue] [Submit]

Interface Alanine Scanning
[Queue] [Submit]

Fragment Libraries
[Queue] [Submit]

DNA Interface Residue Scanning
[Queue] [Submit]

RELATED SITES
Rosetta Commons
Rosetta Commons ROSIE server *NEW*
RosettaBackrub Server
RosettaDesign Server
FoldIt
Rosetta@home
Human Proteome Folding Project
Rosetta@Cloud

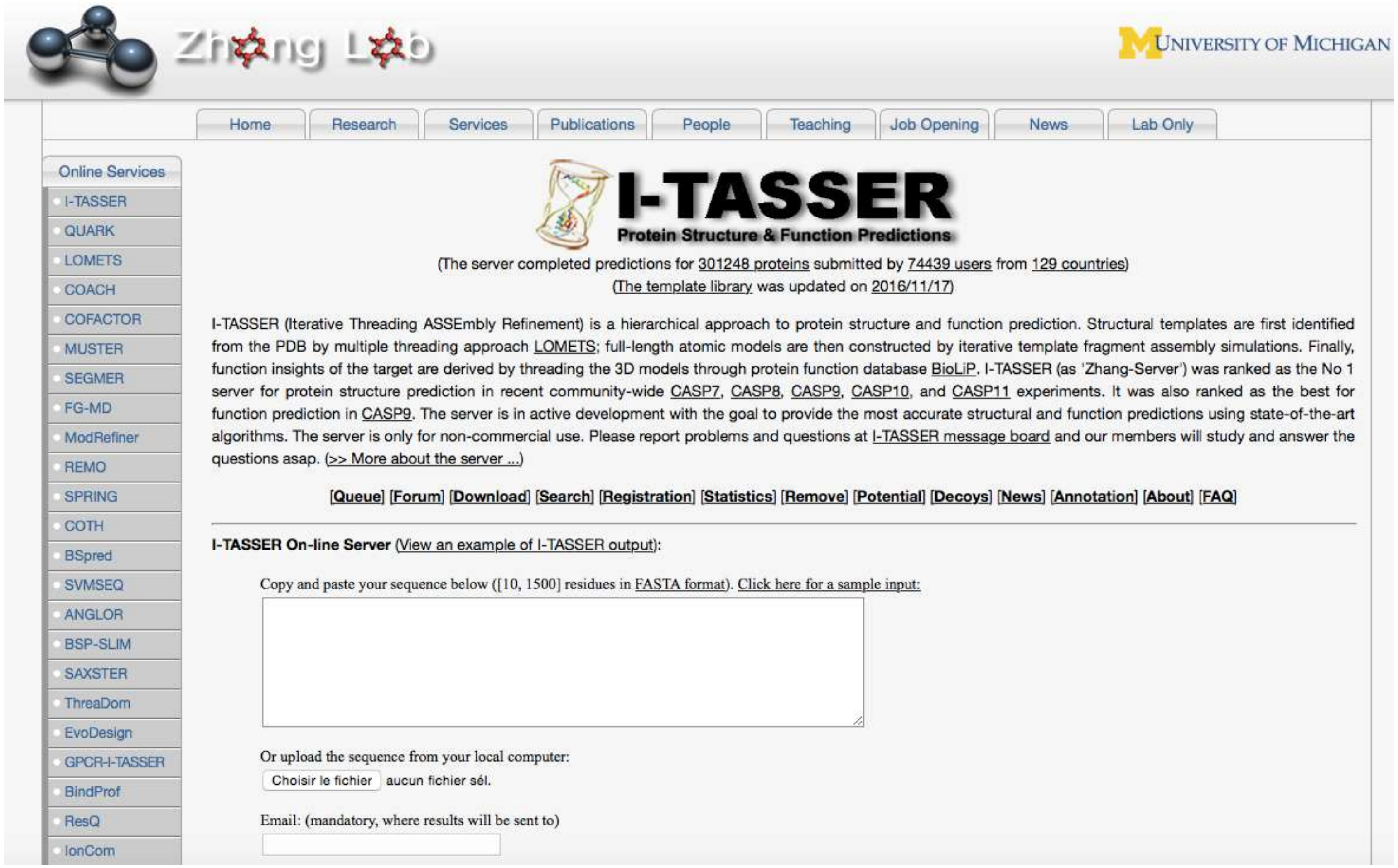
Inserm
Institut national
de la santé et de la recherche médicale

de novo

Bioinformatique & Interactions

- I-Tasser

http://zhanglab.ccmb.med.umich.edu/I-TASSER/



(The server completed predictions for 301248 proteins submitted by 74439 users from 129 countries)
(The template library was updated on 2016/11/17)

I-TASSER (Iterative Threading ASSEmbly Refinement) is a hierarchical approach to protein structure and function prediction. Structural templates are first identified from the PDB by multiple threading approach LOMETS; full-length atomic models are then constructed by iterative template fragment assembly simulations. Finally, function insights of the target are derived by threading the 3D models through protein function database BioLiP. I-TASSER (as 'Zhang-Server') was ranked as the No 1 server for protein structure prediction in recent community-wide CASP7, CASP8, CASP9, CASP10, and CASP11 experiments. It was also ranked as the best for function prediction in CASP9. The server is in active development with the goal to provide the most accurate structural and function predictions using state-of-the-art algorithms. The server is only for non-commercial use. Please report problems and questions at I-TASSER message board and our members will study and answer the questions asap. (>> More about the server ...)

[Queue] [Forum] [Download] [Search] [Registration] [Statistics] [Remove] [Potential] [Decoys] [News] [Annotation] [About] [FAQ]

I-TASSER On-line Server (View an example of I-TASSER output):

Copy and paste your sequence below ([10, 1500] residues in FASTA format). Click here for a sample input:

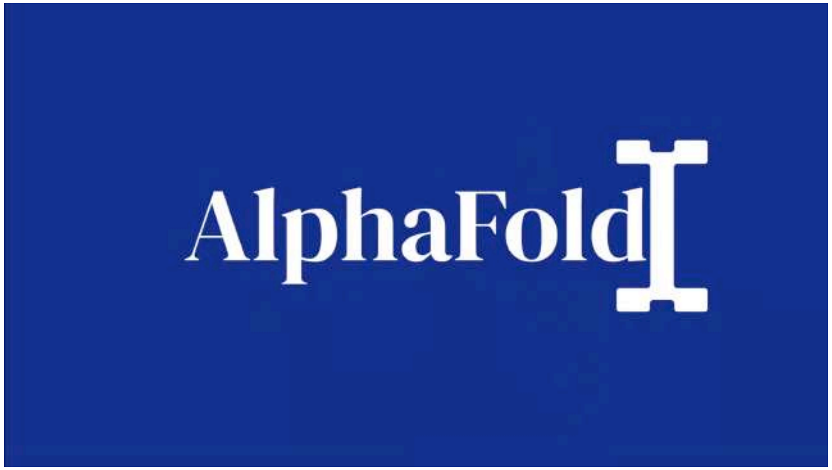
Or upload the sequence from your local computer:

aucun fichier sél.

Email: (mandatory, where results will be sent to)

Snoopy’s question

Are you sure you have not forgotten something ?



ALPHAFOLD2

The (r)evolution …

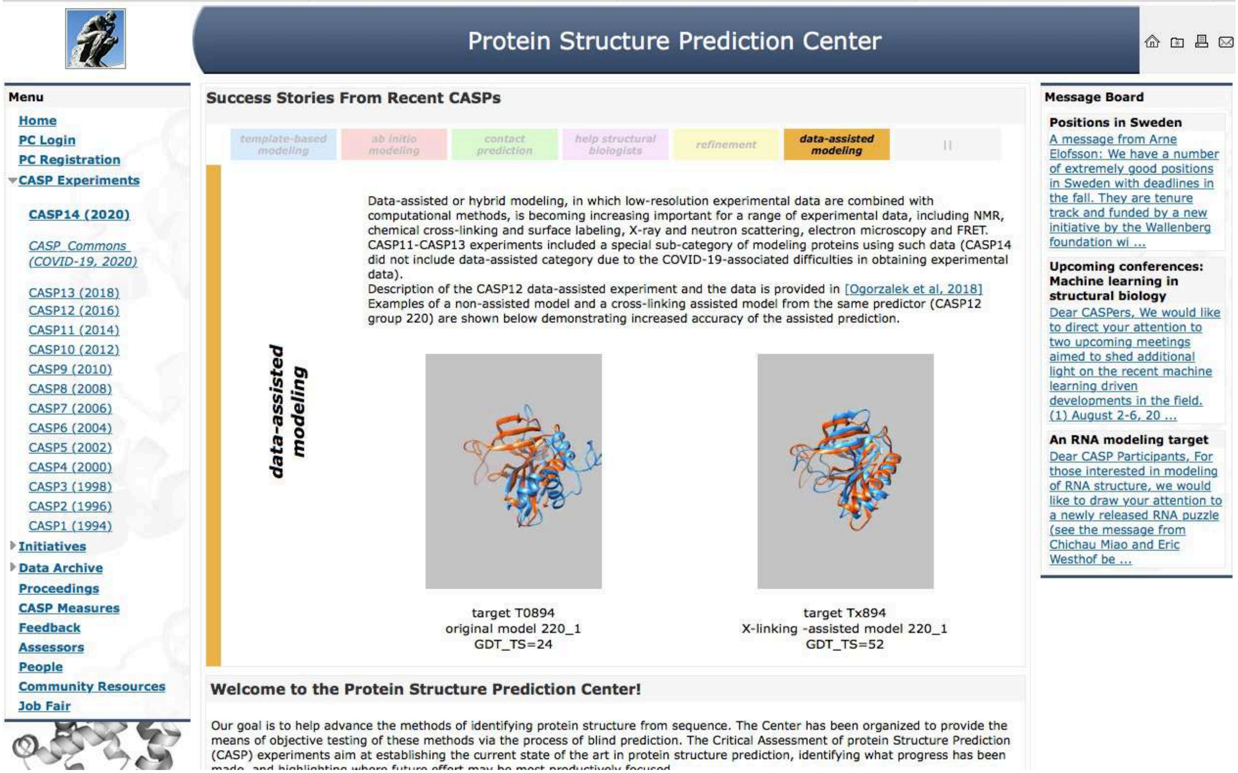
AlphaFold2

- What is CASP?

AlphaFold2

➢ What is CASP?

Critical Assessment of Structural Prediction



Success Stories From Recent CASPs

Data-assisted or hybrid modeling, in which low-resolution experimental data are combined with computational methods, is becoming increasing important for a range of experimental data, including NMR, chemical cross-linking and surface labeling, X-ray and neutron scattering, electron microscopy and FRET. CASP11-CASP13 experiments included a special sub-category of modeling proteins using such data (CASP14 did not include data-assisted category due to the COVID-19-associated difficulties in obtaining experimental data).
Description of the CASP12 data-assisted experiment and the data is provided in [Ogorzalek et al, 2018]
Examples of a non-assisted model and a cross-linking assisted model from the same predictor (CASP12 group 220) are shown below demonstrating increased accuracy of the assisted prediction.

target T0894
original model 220_1
GDT_TS=24

target Tx894
X-linking -assisted model 220_1
GDT_TS=52

Welcome to the Protein Structure Prediction Center!

Our goal is to help advance the methods of identifying protein structure from sequence. The Center has been organized to provide the means of objective testing of these methods via the process of blind prediction. The Critical Assessment of protein Structure Prediction (CASP) experiments aim at establishing the current state of the art in protein structure prediction, identifying what progress has been made, and highlighting where future effort may be most productively focused.

Message Board

Positions in Sweden
A message from Arne Elofsson: We have a number of extremely good positions in Sweden with deadlines in the fall. They are tenure track and funded by a new initiative by the Wallenberg foundation wi ...

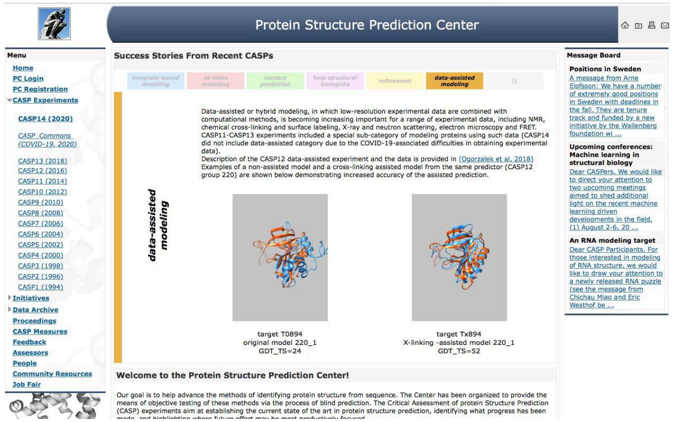
Upcoming conferences: Machine learning in structural biology
Dear CASPers, We would like to direct your attention to two upcoming meetings aimed to shed additional light on the recent machine learning driven developments in the field. (1) August 2-6, 20 ...

An RNA modeling target
Dear CASP Participants, For those interested in modeling of RNA structure, we would like to draw your attention to a newly released RNA puzzle (see the message from Chichau Miao and Eric Westhof be ...

AlphaFold2

- What is CASP?

Critical Assessment of Structural Prediction



Success Stories From Recent CASPs

Data-assisted or hybrid modeling, in which low-resolution experimental data are combined with computational methods, is becoming increasing important for a range of experimental data, including NMR, chemical cross-linking and surface labeling, X-ray and neutron scattering, electron microscopy and FRET. CASP11-CASP13 experiments included a special sub-category of modeling proteins using such data (CASP14 did not include data-assisted category due to the COVID-19-associated difficulties in obtaining experimental data).
Description of the CASP12 data-assisted experiment and the data is provided in [Ogorzalek et al, 2018]
Examples of a non-assisted model and a cross-linking assisted model from the same predictor (CASP12 group 220) are shown below demonstrating increased accuracy of the assisted prediction.

target T0894
original model 220_1
GDT_TS=24

target Tx894
X-linking -assisted model 220_1
GDT_TS=52

Welcome to the Protein Structure Prediction Center!

Our goal is to help advance the methods of identifying protein structure from sequence. The Center has been organized to provide the means of objective testing of these methods via the process of blind prediction. The Critical Assessment of protein Structure Prediction (CASP) experiments aim at establishing the current state of the art in protein structure prediction, identifying what progress has been made, and highlighting where future effort may be most productively focused.

Message Board

Positions in Sweden
A message from Arne Elofsson: We have a number of extremely good positions in Sweden with deadlines in the fall. They are tenure track and funded by a new initiative by the Wallenberg foundation wi ...

Upcoming conferences: Machine learning in structural biology
Dear CASPers, We would like to direct your attention to two upcoming meetings aimed to shed additional light on the recent machine learning driven developments in the field. (1) August 2-6, 20 ...

An RNA modeling target
Dear CASP Participants, For those interested in modeling of RNA structure, we would like to draw your attention to a newly released RNA puzzle (see the message from Chichau Miao and Eric Westhof be ...

=> Goal: blind proposition of structural models, i.e. evaluation of the different methodologies.

AlphaFold2

- How CASP had evolved?

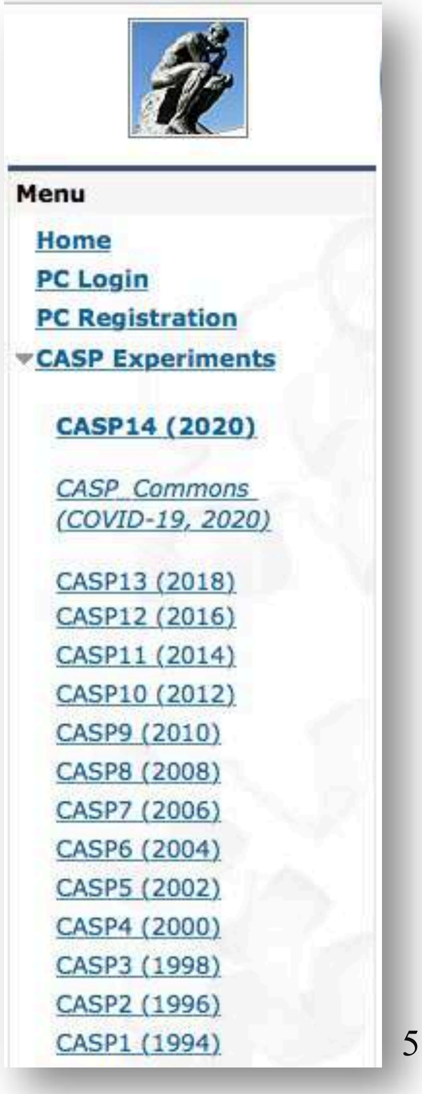
Very crude:

(i) Threading with comparative modelling

(ii) Threading

(iii) *de novo*

(iv) Improvements of *de novo*



Menu

Home

PC Login

PC Registration

▾CASP Experiments

CASP14 (2020)

CASP Commons (COVID-19, 2020)

CASP13 (2018)

CASP12 (2016)

CASP11 (2014)

CASP10 (2012)

CASP9 (2010)

CASP8 (2008)

CASP7 (2006)

CASP6 (2004)

CASP5 (2002)

CASP4 (2000)

CASP3 (1998)

CASP2 (1996)

CASP1 (1994)

AlphaFold2

- How CASP had evolved?

Very crude:

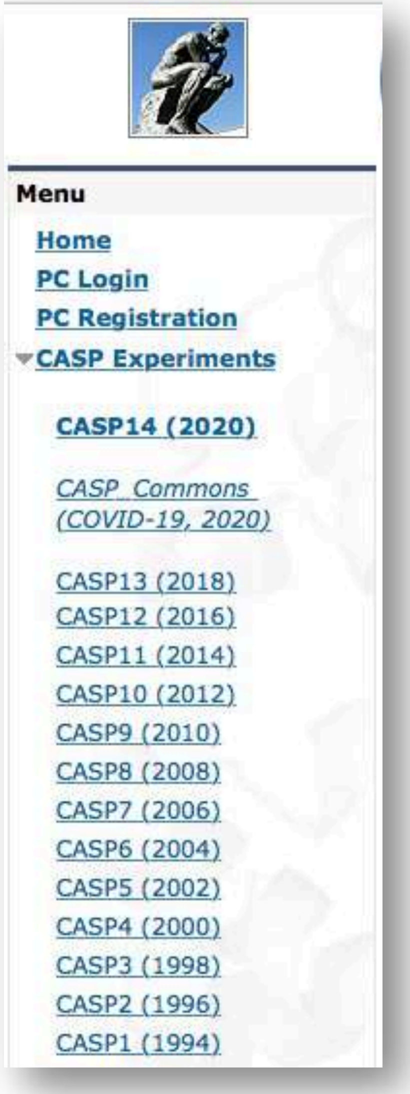
(i) Threading with comparative modelling

(ii) Threading

(iii) *de novo*

(iv) Improvements of *de novo*

(v) AlphaFold (2018), v2 (2020)



Menu

- Home
- PC Login
- PC Registration
- CASP Experiments
  - CASP14 (2020)
  - *CASP Commons (COVID-19, 2020)*
  - CASP13 (2018)
  - CASP12 (2016)
  - CASP11 (2014)
  - CASP10 (2012)
  - CASP9 (2010)
  - CASP8 (2008)
  - CASP7 (2006)
  - CASP6 (2004)
  - CASP5 (2002)
  - CASP4 (2000)
  - CASP3 (1998)
  - CASP2 (1996)
  - CASP1 (1994)

AlphaFold2

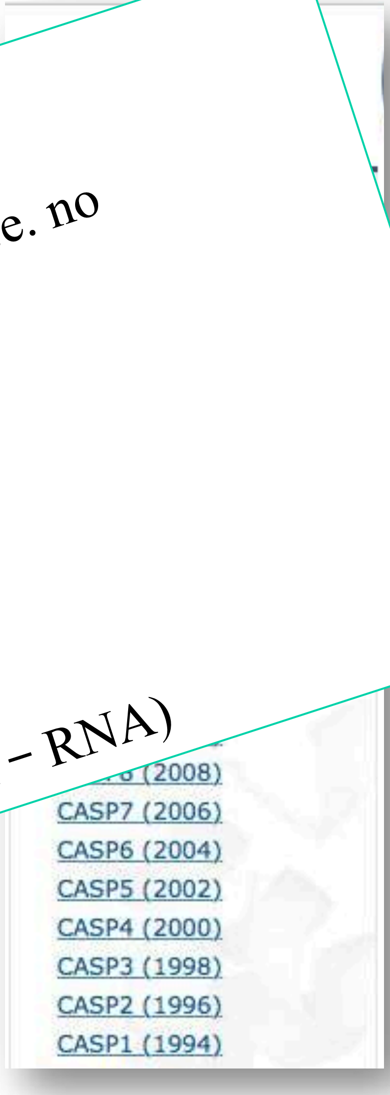
- How CASP had evolved?

Very crude:

(i) Thre

Systematic bias:
Limited number of avalaible structures
Not all types of structures (type of fold, type of protein, i.e. no transmembrane protein)
How to evaluate (RMSD, GDT_TS, …)
Human supervised help …
A lot of money for some labs..

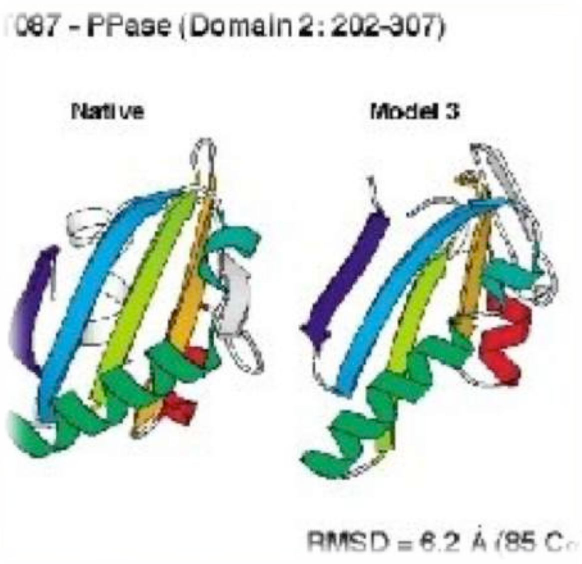
Evolution:
Question of disorder
Question of complexes (protein – protein, protein – RNA)


(2008)
CASP7 (2006)
CASP6 (2004)
CASP5 (2002)
CASP4 (2000)
CASP3 (1998)
CASP2 (1996)
CASP1 (1994)

AlphaFold2

- How CASP had evolved?

1994-2002 : David Baker, add improvements …
but still difficult when it is difficult



AlphaFold2

- How CASP had evolved?

1994-2002 : David Baker, add improvements …
but still difficult when it is difficult

2002-2010: add more and more constraints,
to test (a lot of computational filters)

Rosetta (Baker) & I-Tasser (Zhang)

AlphaFold2

- How CASP had evolved?

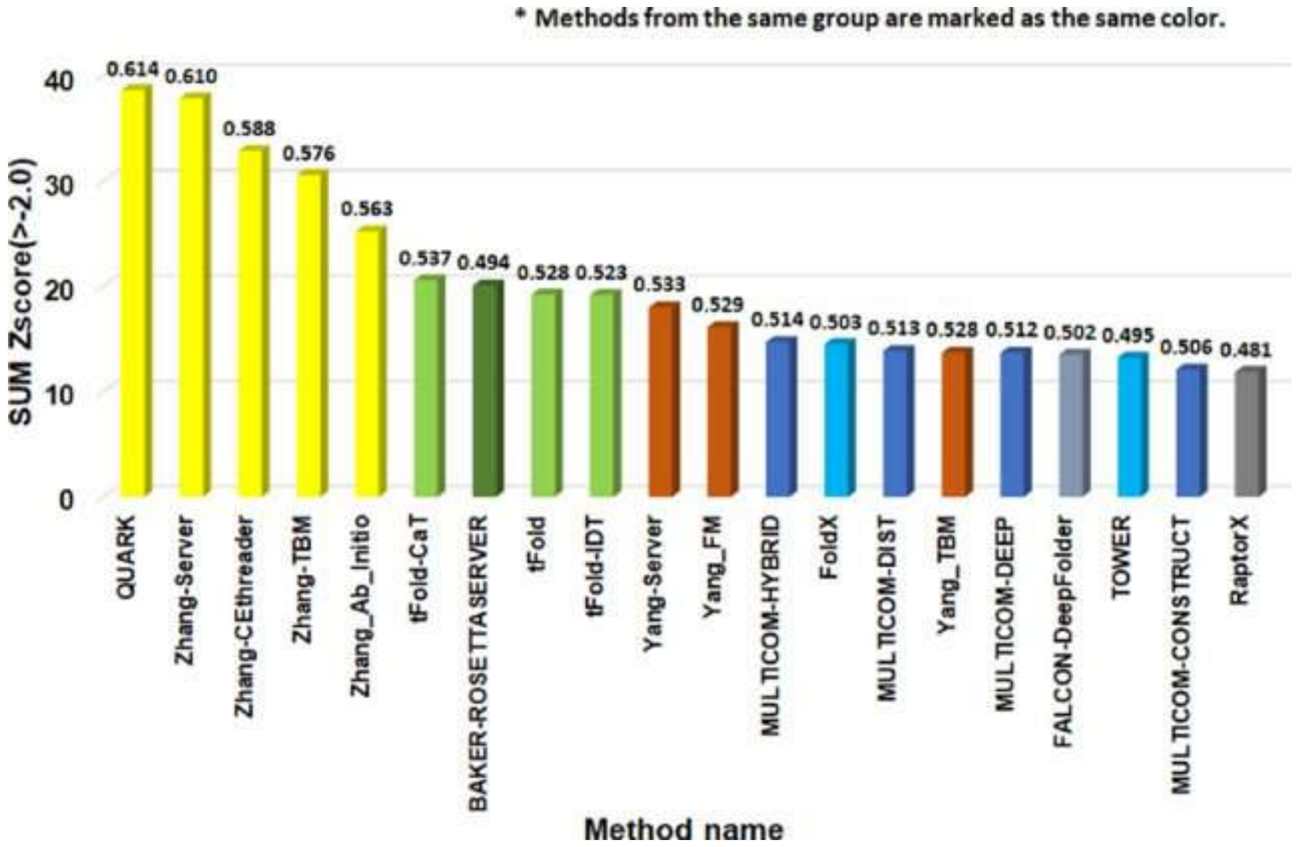
1994-2002 : David Baker, add improvements …
but still difficult when it is difficult

2002-2010: add more and more constraints,
to test (a lot of computational filters)

2012-2016: slight improvements

AlphaFold2

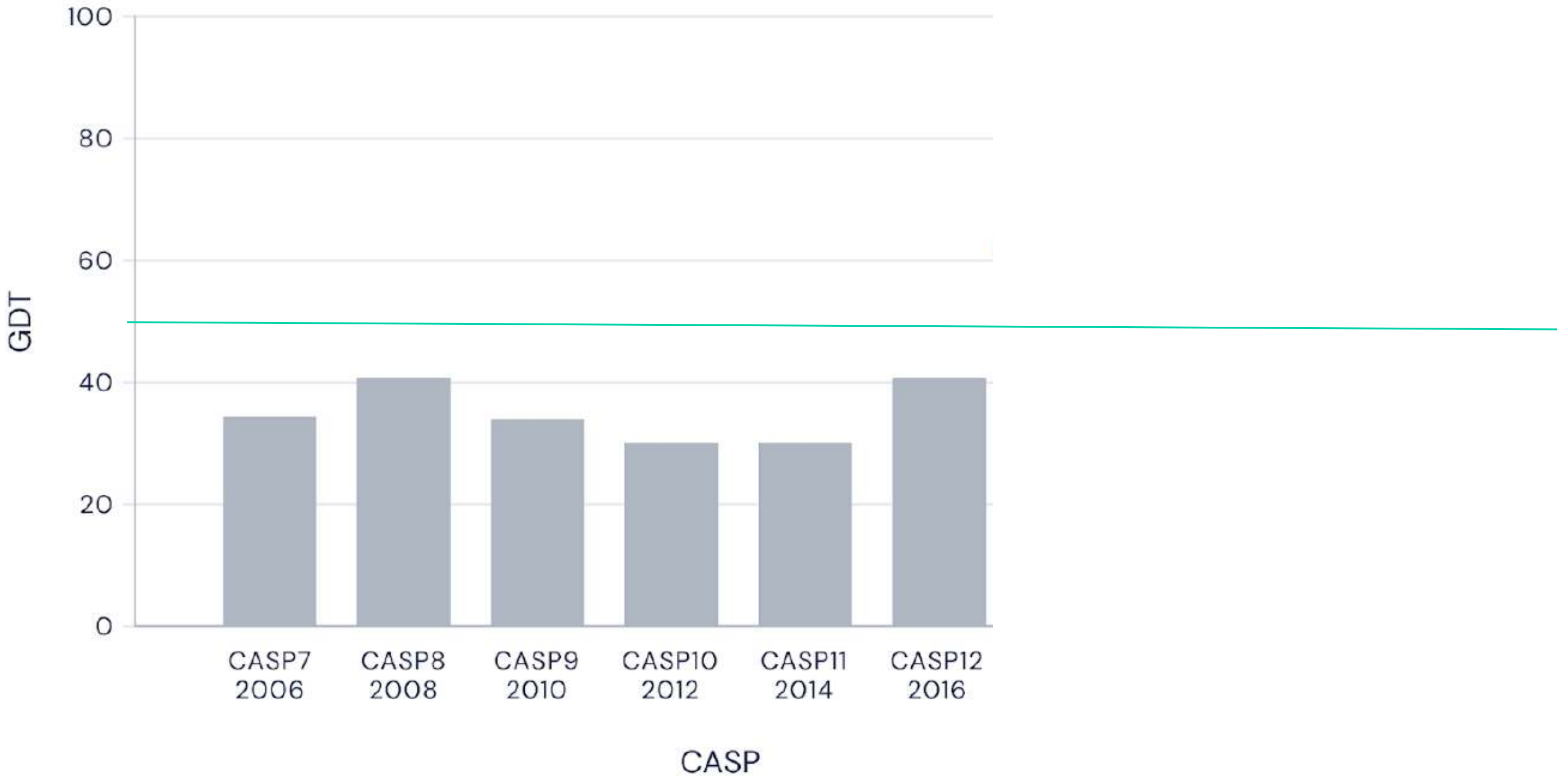
- 2016 (*on specific folds, with specific criteria*)



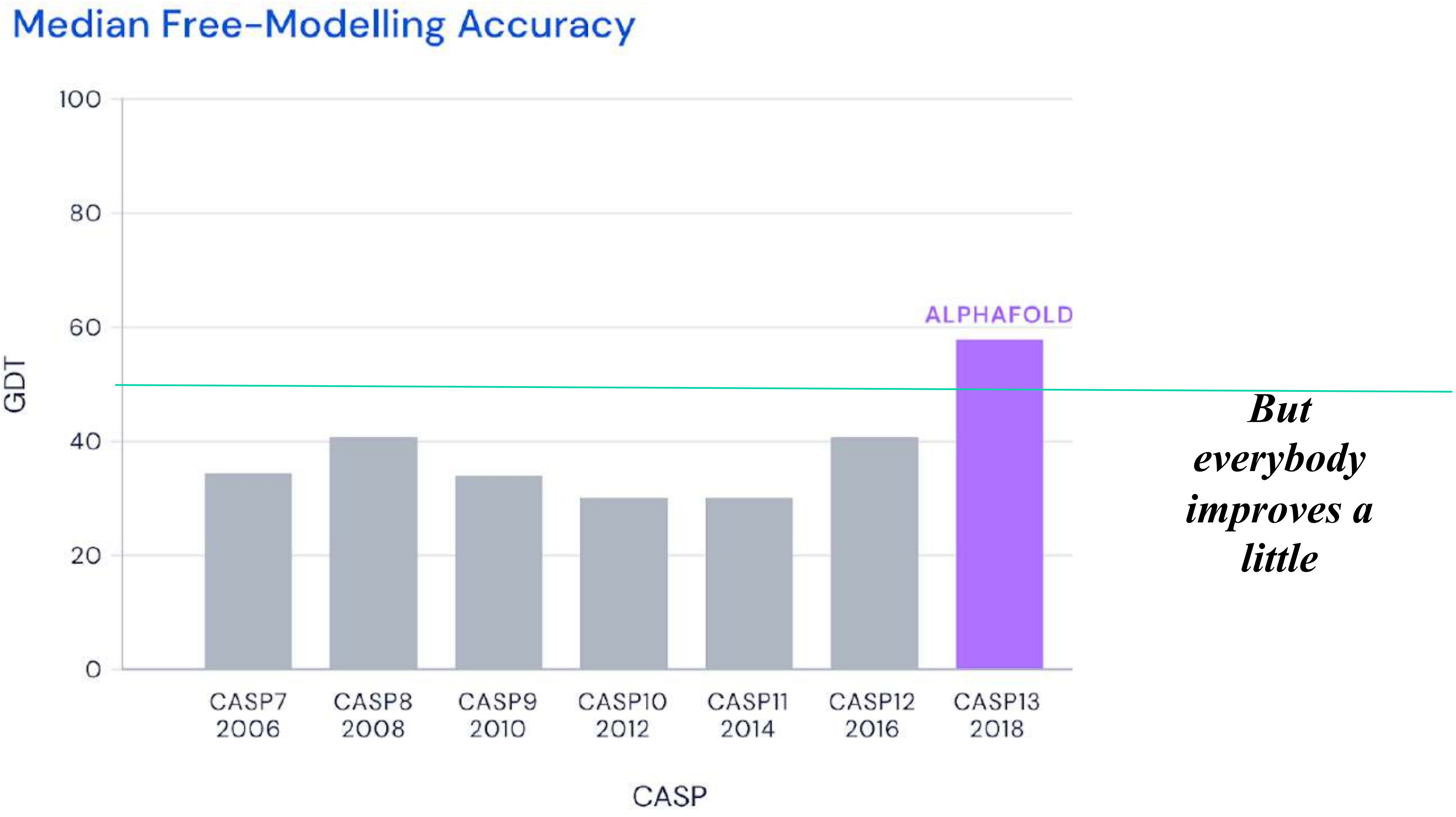
* Methods from the same group are marked as the same color.

AlphaFold2

Median Free-Modelling Accuracy

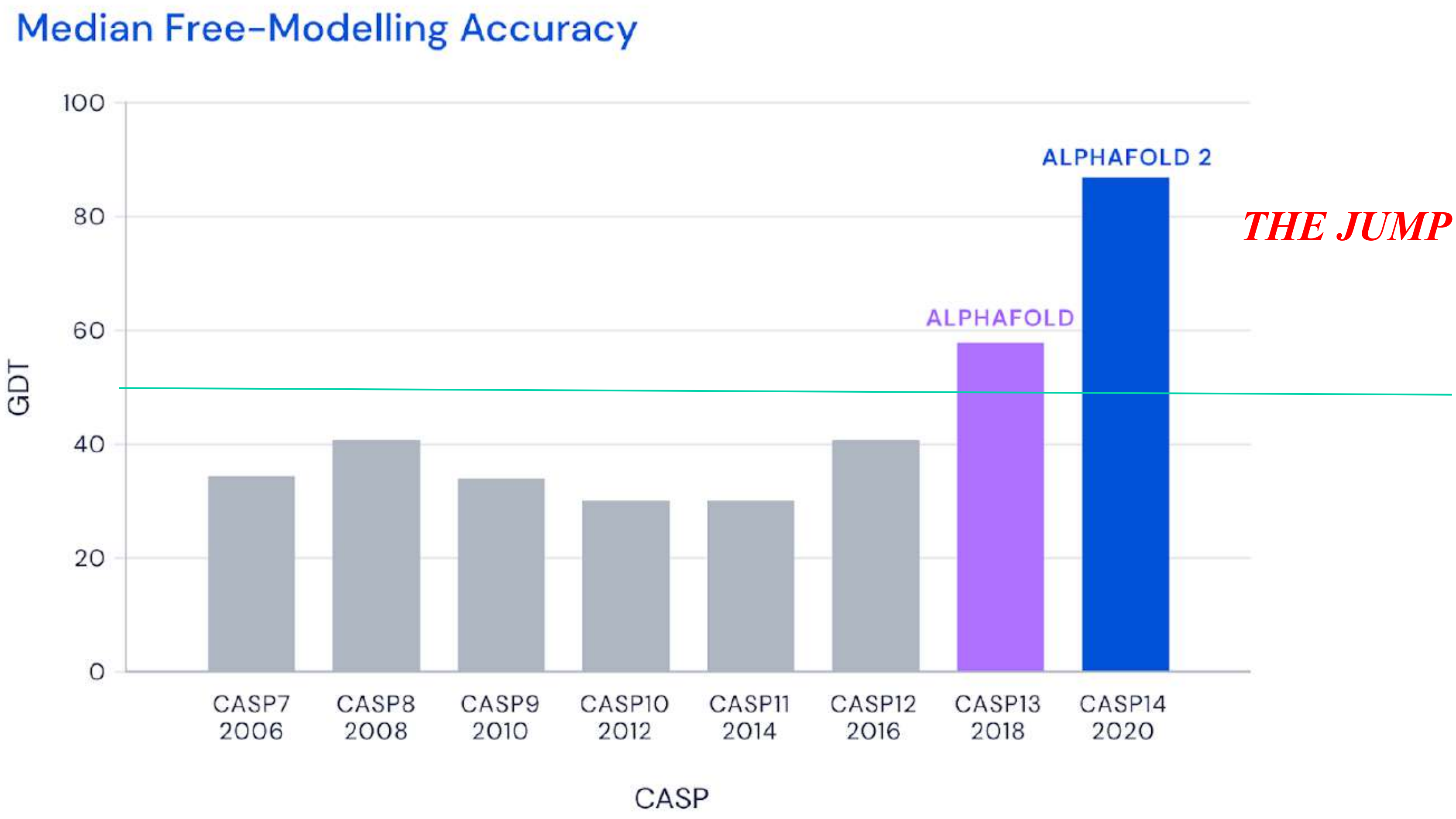


AlphaFold2



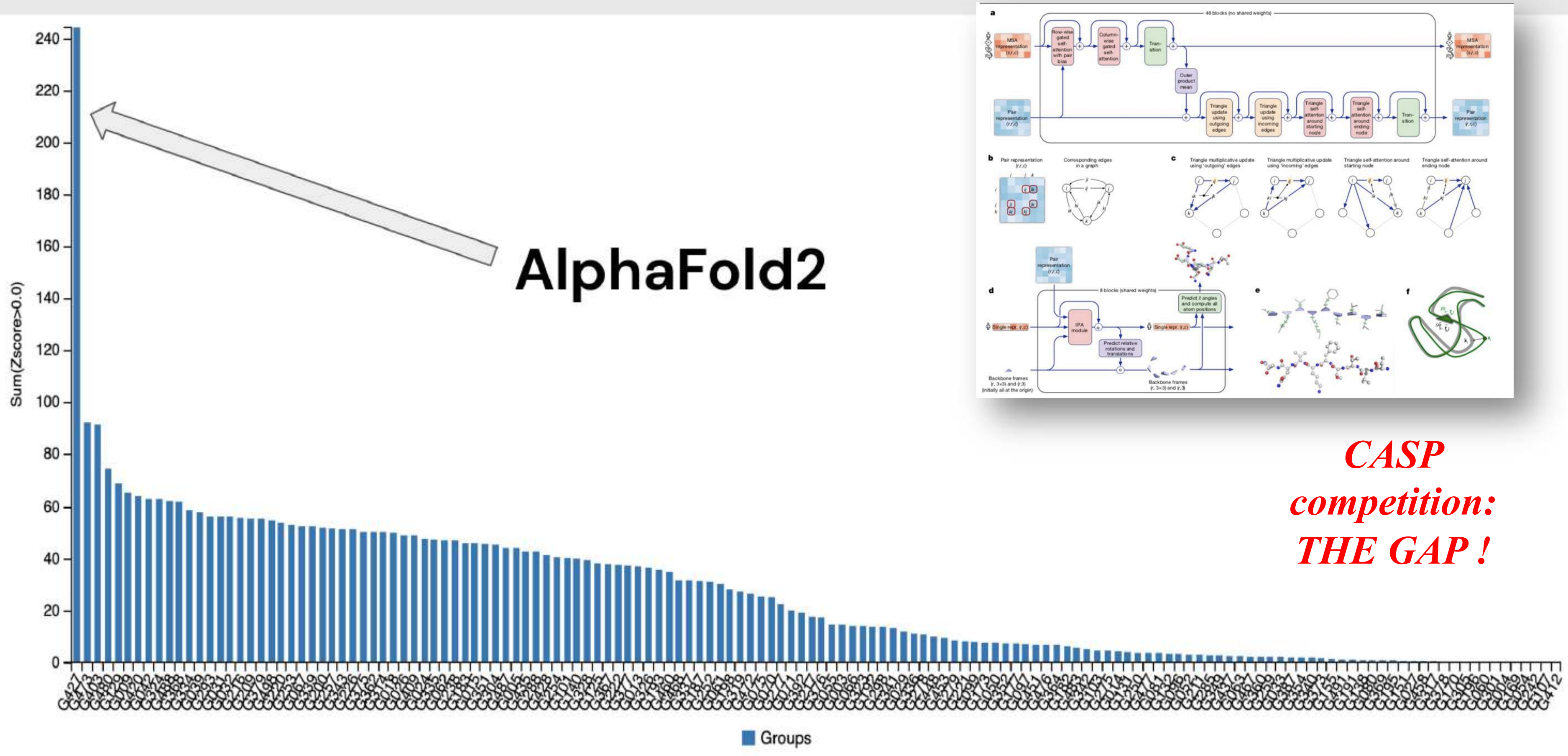
But everybody improves a little

AlphaFold2



THE JUMP

AlphaFold2



CASP competition: THE GAP !

#	GR code	GR name	Domains Count	SUM Zscore (>-2.0)	Rank SUM Zscore (>-2.0)	AVG Zscore (>-2.0)	Rank AVG Zscore (>-2.0)	SUM Zscore (>0.0)	Rank SUM Zscore (>0.0)	AVG Zscore (>0.0)	Rank AVG Zscore (>0.0)
1	427	AlphaFold2	92	244.0217	1	2.6524	1	244.0217	1	2.6524	1
2	473	BAKER	92	90.8241	2	0.9872	2	92.1241	2	1.0013	

AlphaFold2

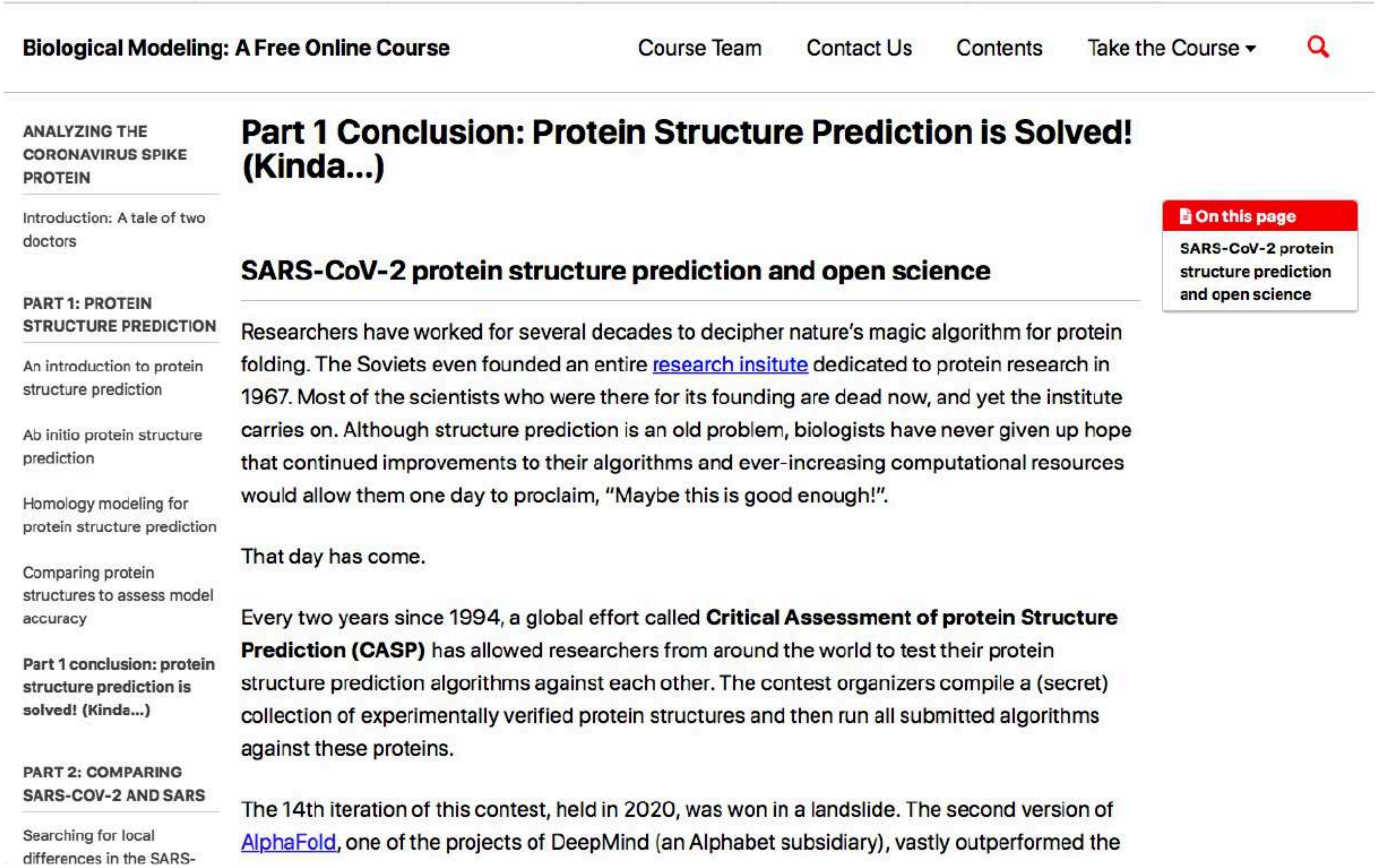
- In all papers !!

- Specialized and not

  Figaro, le Monde, ….

AlphaFold2

- In all papers !!


Biological Modeling: A Free Online Course
Course Team
Contact Us
Contents
Take the Course ▾
ANALYZING THE CORONAVIRUS SPIKE PROTEIN
Introduction: A tale of two doctors
PART 1: PROTEIN STRUCTURE PREDICTION
An introduction to protein structure prediction
Ab initio protein structure prediction
Homology modeling for protein structure prediction
Comparing protein structures to assess model accuracy
Part 1 conclusion: protein structure prediction is solved! (Kinda...)
PART 2: COMPARING SARS-COV-2 AND SARS
Searching for local differences in the SARS-
On this page
SARS-CoV-2 protein structure prediction and open science

Part 1 Conclusion: Protein Structure Prediction is Solved! (Kinda...)

SARS-CoV-2 protein structure prediction and open science

Researchers have worked for several decades to decipher nature's magic algorithm for protein folding. The Soviets even founded an entire research insitute dedicated to protein research in 1967. Most of the scientists who were there for its founding are dead now, and yet the institute carries on. Although structure prediction is an old problem, biologists have never given up hope that continued improvements to their algorithms and ever-increasing computational resources would allow them one day to proclaim, "Maybe this is good enough!".

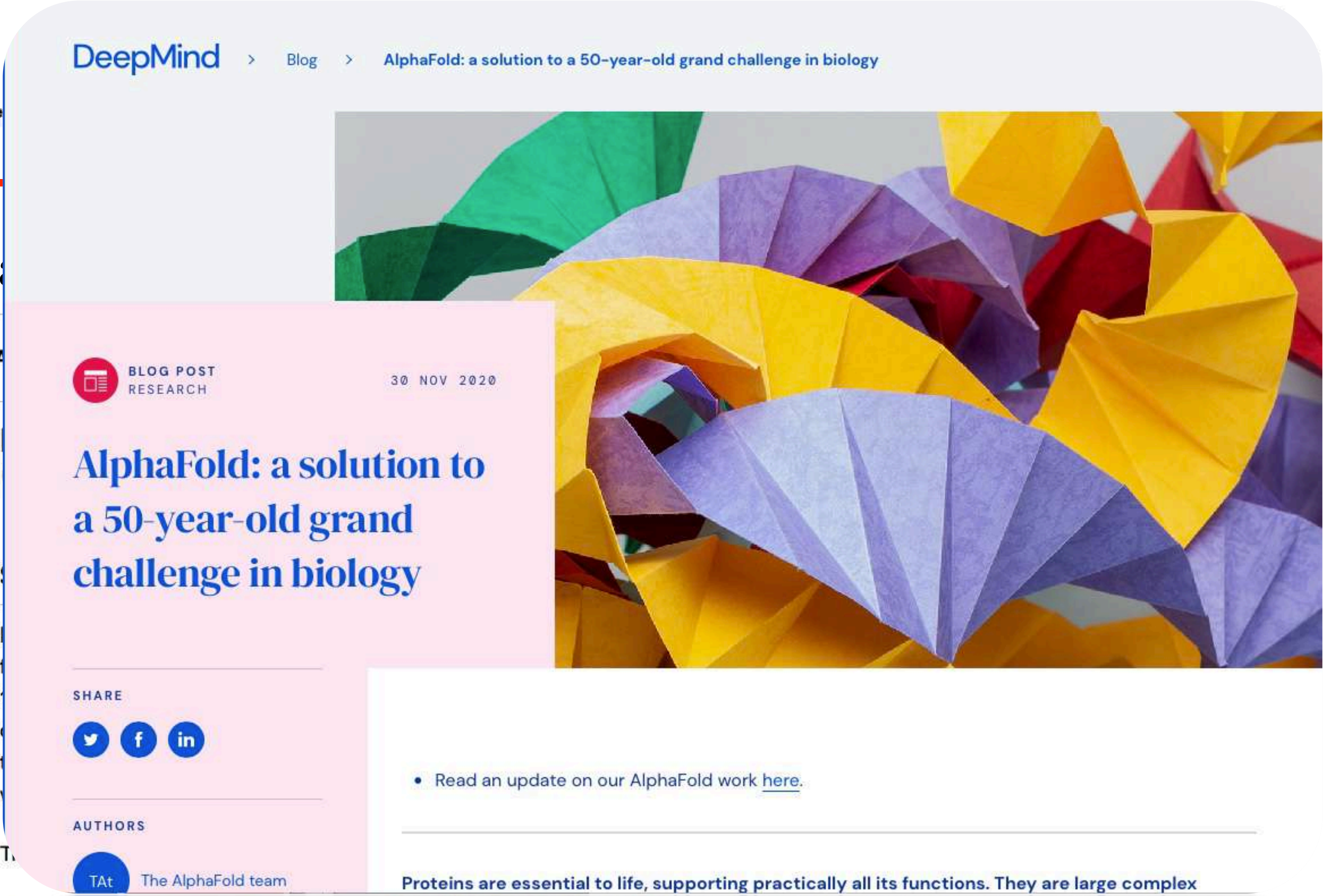
That day has come.

Every two years since 1994, a global effort called **Critical Assessment of protein Structure Prediction (CASP)** has allowed researchers from around the world to test their protein structure prediction algorithms against each other. The contest organizers compile a (secret) collection of experimentally verified protein structures and then run all submitted algorithms against these proteins.

The 14th iteration of this contest, held in 2020, was won in a landslide. The second version of AlphaFold, one of the projects of DeepMind (an Alphabet subsidiary), vastly outperformed the

- In a

Biological Modeling: A


DeepMind > Blog > AlphaFold: a solution to a 50-year-old grand challenge in biology

BLOG POST
RESEARCH

30 NOV 2020

AlphaFold: a solution to a 50-year-old grand challenge in biology

SHARE

AUTHORS

TAt The AlphaFold team

- Read an update on our AlphaFold work here.

Proteins are essential to life, supporting practically all its functions. They are large complex

T

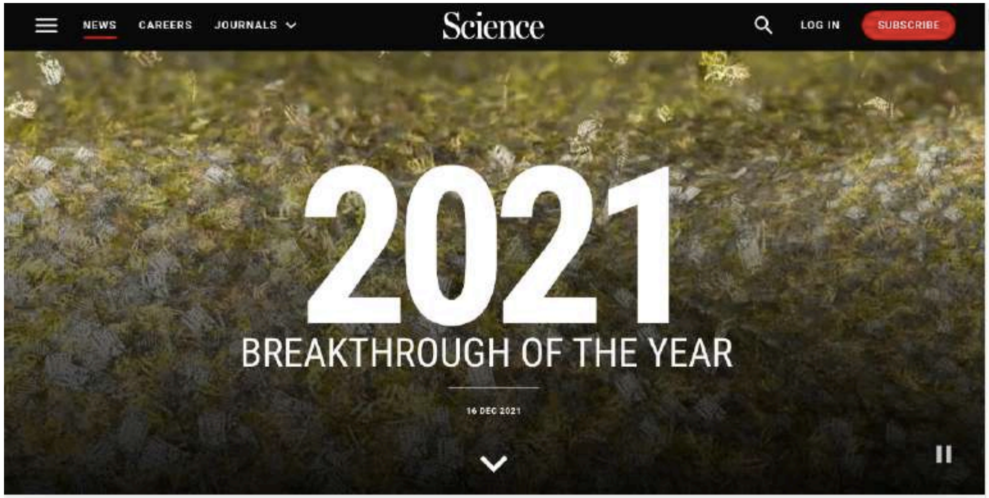
Every two years since 1994, a global effort called **Critical Assessment of protein Structure Prediction (CASP)** has allowed researchers from around the world to test their protein structure prediction algorithms against each other. The contest organizers compile a (secret) collection of experimentally verified protein structures and then run all submitted algorithms against these proteins.

The 14th iteration of this contest, held in 2020, was won in a landslide. The second version of AlphaFold, one of the projects of DeepMind (an Alphabet subsidiary), vastly outperformed the

AlphaFold2

- In all papers !! ➔ *Nature* 2021 (now > 30.000 citations)

Breakthrough of the year *Science* 2021

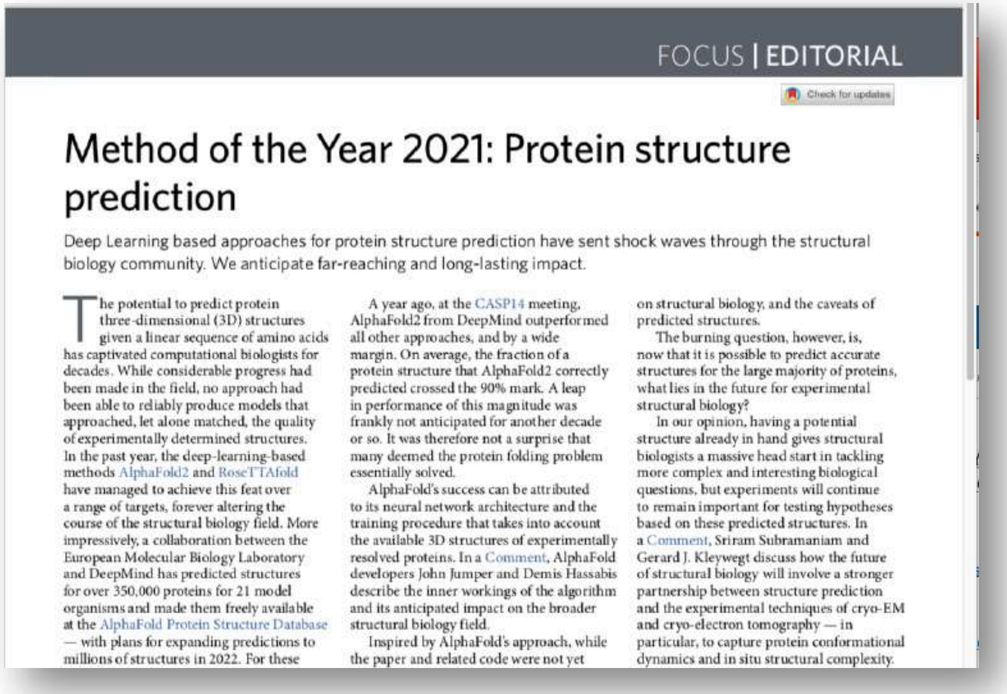


AlphaFold2

- In all papers !! ➔ *Nature* 2021 (now > 30.000 citations)

Breakthrough of the year *Science* 2021

Method of the year *Nature Methods* 2021



FOCUS | EDITORIAL

Check for updates

Method of the Year 2021: Protein structure prediction

Deep Learning based approaches for protein structure prediction have sent shock waves through the structural biology community. We anticipate far-reaching and long-lasting impact.

The potential to predict protein three-dimensional (3D) structures given a linear sequence of amino acids has captivated computational biologists for decades. While considerable progress had been made in the field, no approach had been able to reliably produce models that approached, let alone matched, the quality of experimentally determined structures. In the past year, the deep-learning-based methods AlphaFold2 and RoseTTAfold have managed to achieve this feat over a range of targets, forever altering the course of the structural biology field. More impressively, a collaboration between the European Molecular Biology Laboratory and DeepMind has predicted structures for over 350,000 proteins for 21 model organisms and made them freely available at the AlphaFold Protein Structure Database — with plans for expanding predictions to millions of structures in 2022. For these

A year ago, at the CASP14 meeting, AlphaFold2 from DeepMind outperformed all other approaches, and by a wide margin. On average, the fraction of a protein structure that AlphaFold2 correctly predicted crossed the 90% mark. A leap in performance of this magnitude was frankly not anticipated for another decade or so. It was therefore not a surprise that many deemed the protein folding problem essentially solved.

AlphaFold's success can be attributed to its neural network architecture and the training procedure that takes into account the available 3D structures of experimentally resolved proteins. In a Comment, AlphaFold developers John Jumper and Demis Hassabis describe the inner workings of the algorithm and its anticipated impact on the broader structural biology field.

Inspired by AlphaFold's approach, while the paper and related code were not yet

on structural biology, and the caveats of predicted structures.

The burning question, however, is, now that it is possible to predict accurate structures for the large majority of proteins, what lies in the future for experimental structural biology?

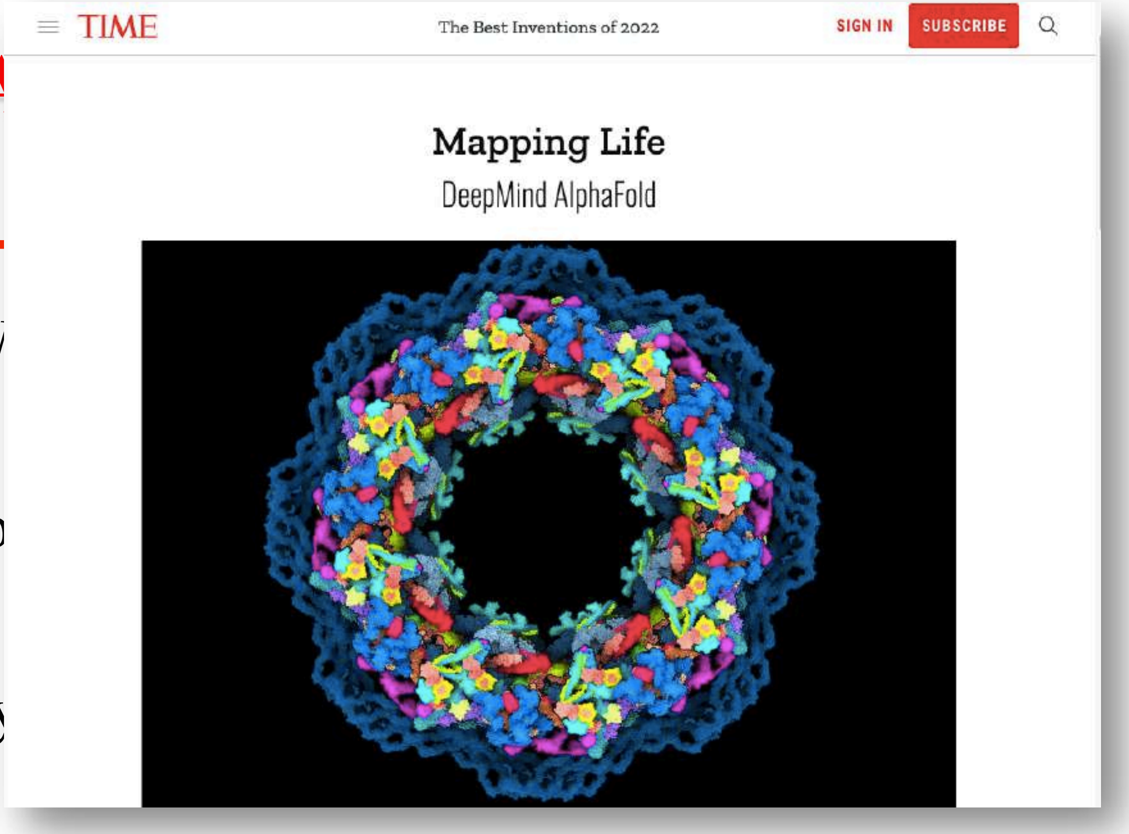
In our opinion, having a potential structure already in hand gives structural biologists a massive head start in tackling more complex and interesting biological questions, but experiments will continue to remain important for testing hypotheses based on these predicted structures. In a Comment, Sriram Subramaniam and Gerard J. Kleywegt discuss how the future of structural biology will involve a stronger partnership between structure prediction and the experimental techniques of cryo-EM and cryo-electron tomography — in particular, to capture protein conformational dynamics and in situ structural complexity.

A

- In all papers !! ➔ *N*

Breakthrough o

Method of the y



Mapping Life

DeepMind AlphaFold

Best invention of 2022 (*Life*)

AlphaFold2

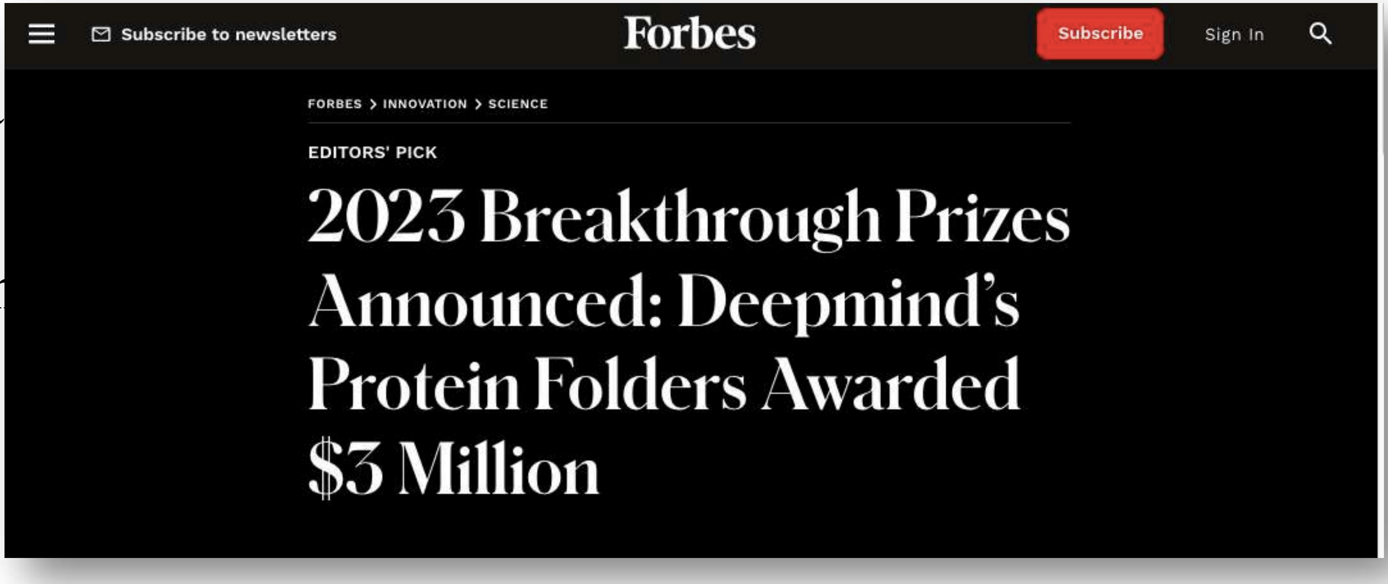
- In all papers !! ➔ *Nature* 2021 (now > 30.000 citations)

Brea

Meth

Best

Prices ⋯.



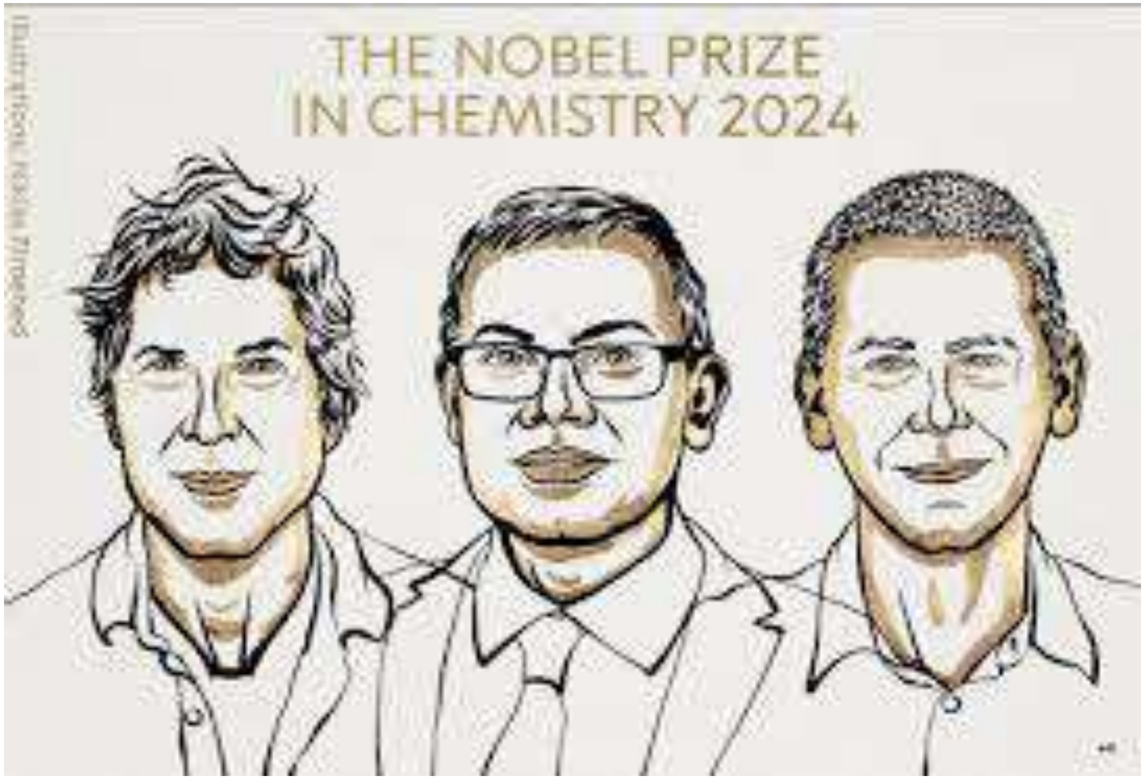
FORBES > INNOVATION > SCIENCE

EDITORS' PICK

2023 Breakthrough Prizes Announced: Deepmind's Protein Folders Awarded $3 Million

AlphaFold2

- And now Nobel prize 2024 (Demis Hassabis & John Jumper)



David Baker ***Demis Hassabis*** ***John Jumper***

AlphaFold2

- So is it real?
- Why?
- How?

AlphaFold2

1. What is behind

AlphaFold2

1. What is behind

Google

>50 engineers (at least) x >5 years

Deep Learning approaches (as Facebook, DeepMind..)

$$$ for excellent bioinformatics specialists

Google’s GPU power (impressive)

translation: heavy, heavy, very heavy

AlphaFold2

2. mechanisms

AlphaFold2

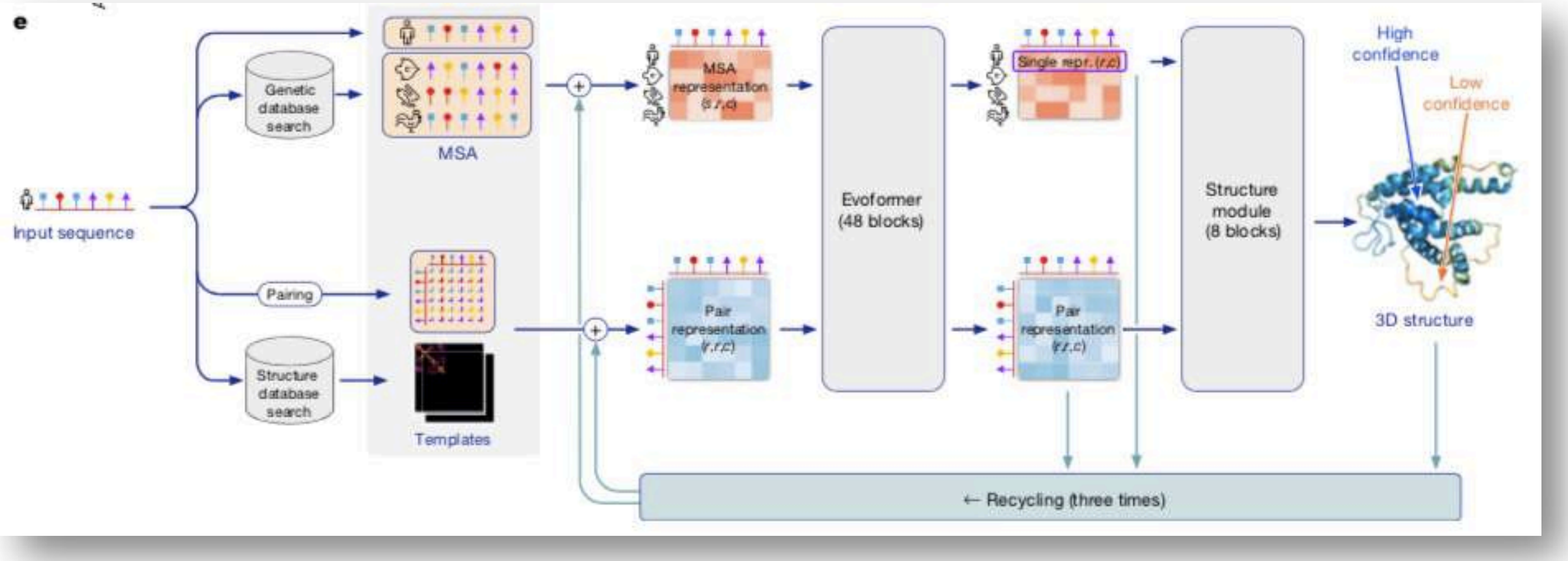
2. mechanisms

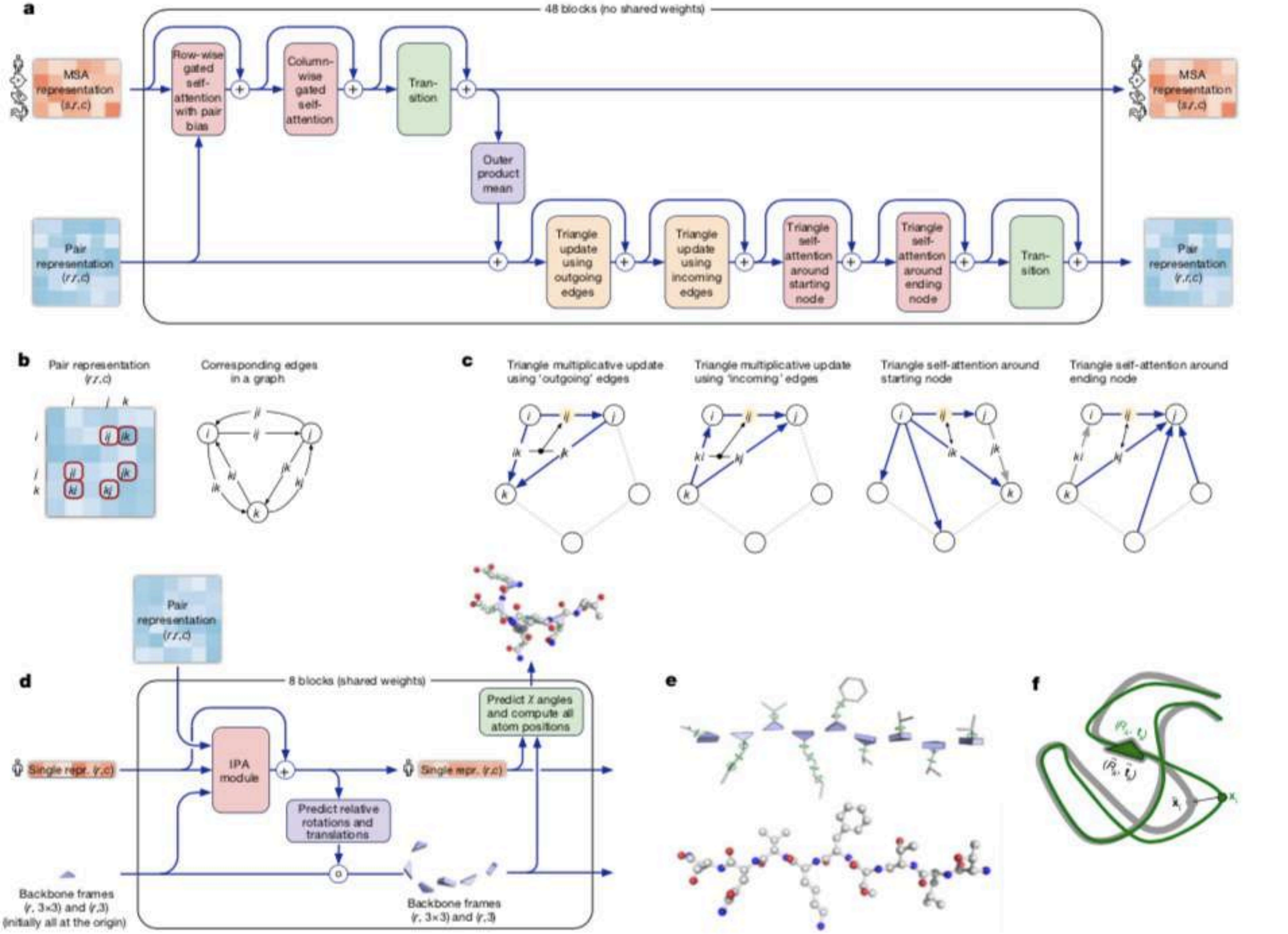
AF1 → CNN

AF2 → LLM

AlphaFold2

3. AlphaFold2


e
Input sequence
Genetic database search
MSA
Pairing
Structure database search
Templates
MSA representation (s,r,c)
Pair representation (r,r,c)
Evoformer (48 blocks)
Single repr. (r,c)
Pair representation (r,r,c)
Structure module (8 blocks)
High confidence
Low confidence
3D structure
← Recycling (three times)


a
48 blocks (no shared weights)
MSA representation (s,r,c)
Row-wise gated self-attention with pair bias
Column-wise gated self-attention
Tran-sition
MSA representation (s,r,c)
Outer product mean
Pair representation (r,r,c)
Triangle update using outgoing edges
Triangle update using incoming edges
Triangle self-attention around starting node
Triangle self-attention around ending node
Tran-sition
Pair representation (r,r,c)
b
Pair representation (r,r,c)
i
j
k
ij
ik
ji
jk
ki
kj
Corresponding edges in a graph
c
Triangle multiplicative update using 'outgoing' edges
Triangle multiplicative update using 'incoming' edges
Triangle self-attention around starting node
Triangle self-attention around ending node
d
Pair representation (r,r,c)
8 blocks (shared weights)
Single repr. (r,c)
IPA module
Predict relative rotations and translations
Single repr. (r,c)
Predict χ angles and compute all atom positions
Backbone frames (r, 3×3) and (r,3) (initially all at the origin)
Backbone frames (r, 3×3) and (r,3)
e
f
(R_k, t_k)
$(\tilde{R}_k, \tilde{t}_k)$
$\tilde{x}_i$
x_i

AlphaFold2

4. The questions

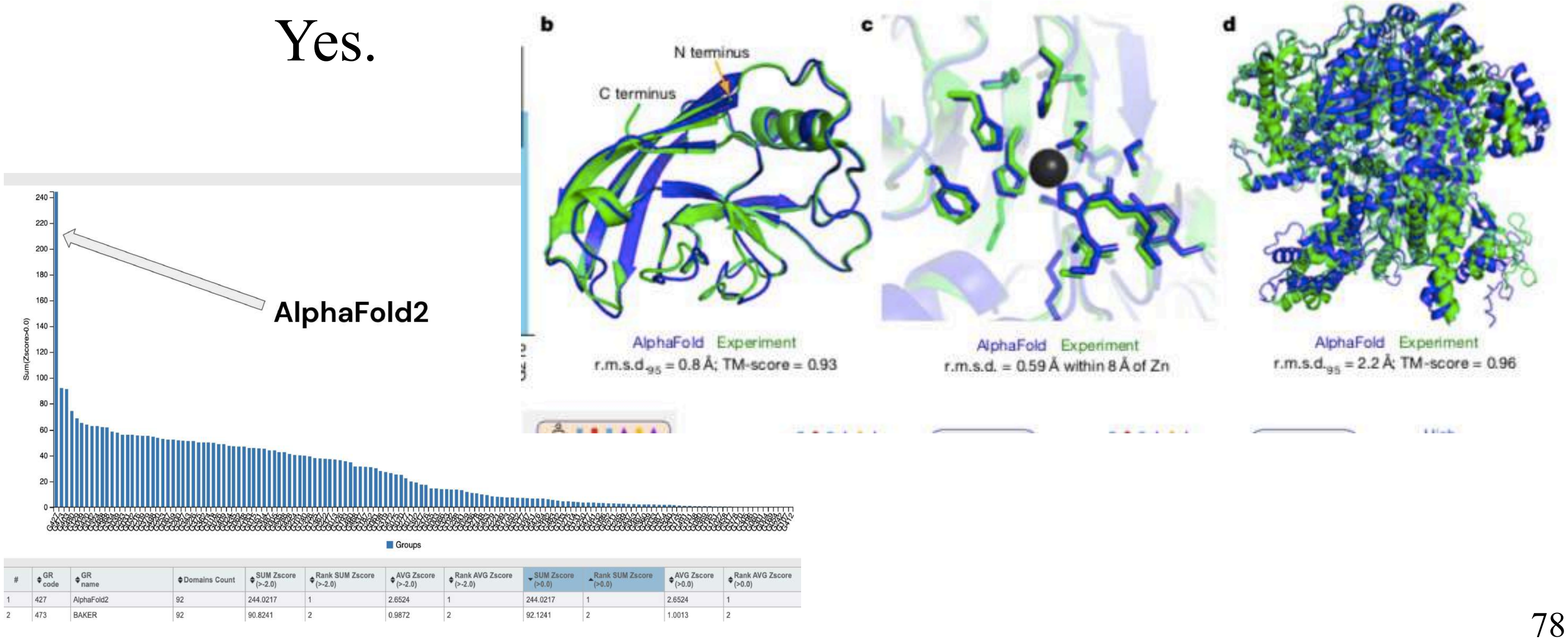
- Is it so good?
- Is the protein folding problem resolved?
- Is there some limitations?

AlphaFold2

4. The questions

➤ Is it so good?

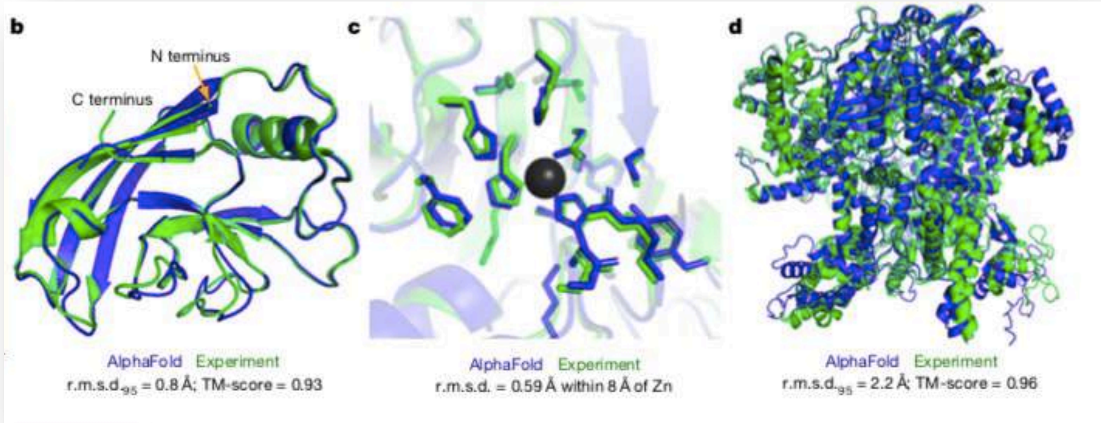
Yes.



#	GR code	GR name	Domains Count	SUM Zscore (>-2.0)	Rank SUM Zscore (>-2.0)	AVG Zscore (>-2.0)	Rank AVG Zscore (>-2.0)	SUM Zscore (>0.0)	Rank SUM Zscore (>0.0)	AVG Zscore (>0.0)	Rank AVG Zscore (>0.0)
1	427	AlphaFold2	92	244.0217	1	2.6524	1	244.0217	1	2.6524	1
2	473	BAKER	92	90.8241	2	0.9872	2	92.1241	2	1.0013	2

AlphaFold2

4. The questions


b
N terminus
C terminus
AlphaFold Experiment
r.m.s.d.$_{95}$ = 0.8 Å; TM-score = 0.93
c
AlphaFold Experiment
r.m.s.d. = 0.59 Å within 8 Å of Zn
d
AlphaFold Experiment
r.m.s.d.$_{95}$ = 2.2 Å; TM-score = 0.96

- Is it so good?

Yes.

They used Multiple Sequence Alignments

(they tested more than anyone before)

They are expending the local protein fold space

They have incorporated all types of SOA approaches

They have computational power never seen before

Inserm
Institut national
de la santé et de la recherche médicale

AlphaFold2

SIMB
Bioinformatique & Interactions

4. The questions

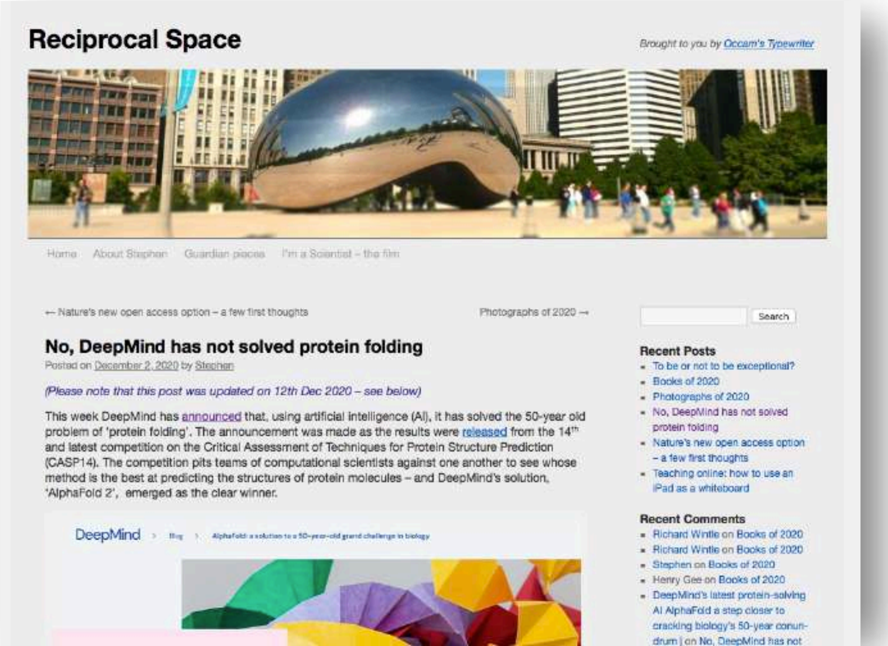
- Is the protein folding problem resolved?

AlphaFold2

4. The questions

- Is the protein folding problem resolved?

No. Protein folding is not protein fold…



Reciprocal Space

Brought to you by Occam's Typewriter

← Nature's new open access option – a few first thoughts

Photographs of 2020 →

No, DeepMind has not solved protein folding

Posted on December 2, 2020 by Stephen

(Please note that this post was updated on 12th Dec 2020 – see below)

This week DeepMind has announced that, using artificial intelligence (AI), it has solved the 50-year old problem of 'protein folding'. The announcement was made as the results were released from the 14^{th} and latest competition on the Critical Assessment of Techniques for Protein Structure Prediction (CASP14). The competition pits teams of computational scientists against one another to see whose method is the best at predicting the structures of protein molecules – and DeepMind's solution, 'AlphaFold 2', emerged as the clear winner.

DeepMind > Blog > AlphaFold: a solution to a 50-year-old grand challenge in biology
Search

Recent Posts

- To be or not to be exceptional?
- Books of 2020
- Photographs of 2020
- No, DeepMind has not solved protein folding
- Nature's new open access option – a few first thoughts
- Teaching online: how to use an iPad as a whiteboard

Recent Comments

- Richard Wintle on Books of 2020
- Richard Wintle on Books of 2020
- Stephen on Books of 2020
- Henry Gee on Books of 2020
- DeepMind's latest protein-solving AI AlphaFold a step closer to cracking biology's 50-year conundrum | on No, DeepMind has not

AlphaFold2

4. The questions

- Is there some limitations?

It is a strange questions as now
(i) you can use it at home
(ii) there is a database of already done model

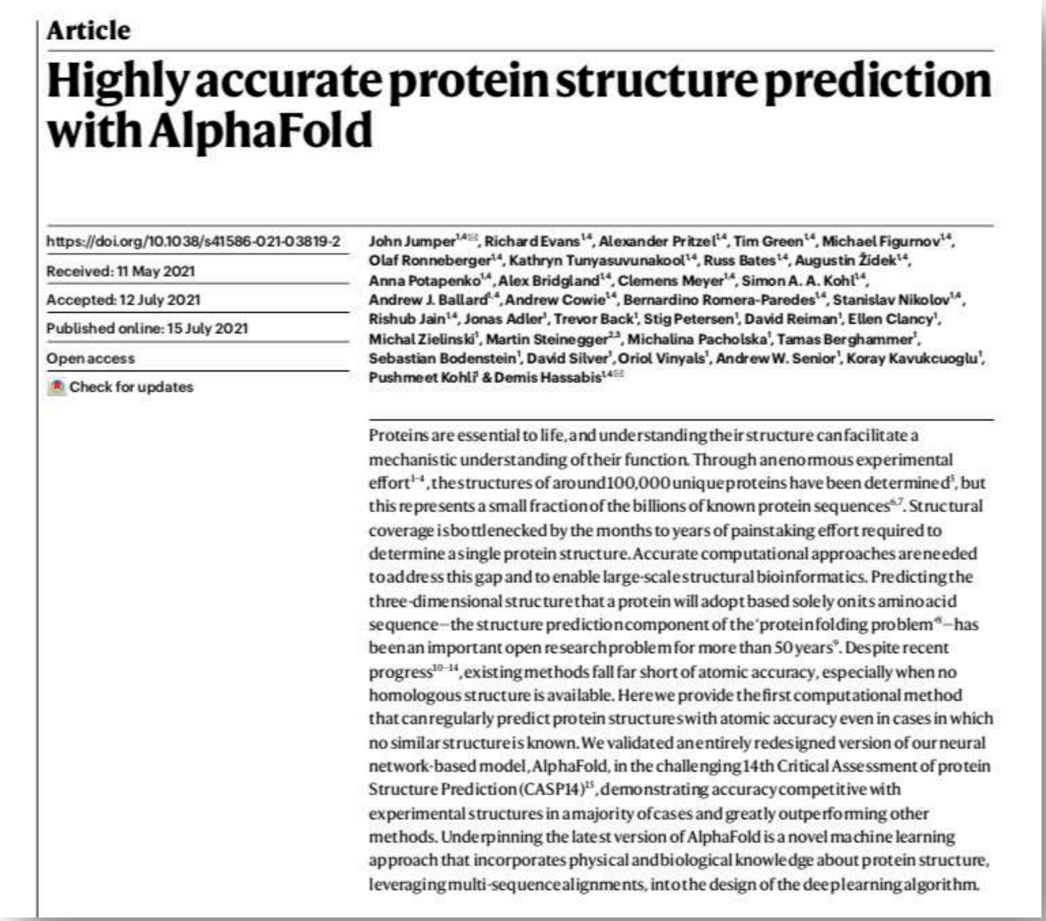
AlphaFold2

- Is there some limitations?

  (i) you can use it at home

Algorithm is published and entirely avalaible (was not the case for v1)

Jumper, J et al. (2021) *Nature*, 596(7873):583-589.



Article

Highly accurate protein structure prediction with AlphaFold

https://doi.org/10.1038/s41586-021-03819-2

Received: 11 May 2021

Accepted: 12 July 2021

Published online: 15 July 2021

Open access

Check for updates

John Jumper[1,4,✉], Richard Evans[1,4], Alexander Pritzel[1,4], Tim Green[1,4], Michael Figurnov[1,4], Olaf Ronneberger[1,4], Kathryn Tunyasuvunakool[1,4], Russ Bates[1,4], Augustin Žídek[1,4], Anna Potapenko[1,4], Alex Bridgland[1,4], Clemens Meyer[1,4], Simon A. A. Kohl[1,4], Andrew J. Ballard[1,4], Andrew Cowie[1,4], Bernardino Romera-Paredes[1,4], Stanislav Nikolov[1,4], Rishub Jain[1,4], Jonas Adler[1], Trevor Back[1], Stig Petersen[1], David Reiman[1], Ellen Clancy[1], Michal Zielinski[1], Martin Steinegger[2,3], Michalina Pacholska[1], Tamas Berghammer[1], Sebastian Bodenstein[1], David Silver[1], Oriol Vinyals[1], Andrew W. Senior[1], Koray Kavukcuoglu[1], Pushmeet Kohli[1] & Demis Hassabis[1,4,✉]

Proteins are essential to life, and understanding their structure can facilitate a mechanistic understanding of their function. Through an enormous experimental effort[1–4], the structures of around 100,000 unique proteins have been determined[5], but this represents a small fraction of the billions of known protein sequences[6,7]. Structural coverage is bottlenecked by the months to years of painstaking effort required to determine a single protein structure. Accurate computational approaches are needed to address this gap and to enable large-scale structural bioinformatics. Predicting the three-dimensional structure that a protein will adopt based solely on its amino acid sequence—the structure prediction component of the 'protein folding problem'[8]—has been an important open research problem for more than 50 years[9]. Despite recent progress[10–14], existing methods fall far short of atomic accuracy, especially when no homologous structure is available. Here we provide the first computational method that can regularly predict protein structures with atomic accuracy even in cases in which no similar structure is known. We validated an entirely redesigned version of our neural network-based model, AlphaFold, in the challenging 14th Critical Assessment of protein Structure Prediction (CASP14)[15], demonstrating accuracy competitive with experimental structures in a majority of cases and greatly outperforming other methods. Underpinning the latest version of AlphaFold is a novel machine learning approach that incorporates physical and biological knowledge about protein structure, leveraging multi-sequence alignments, into the design of the deep learning algorithm.

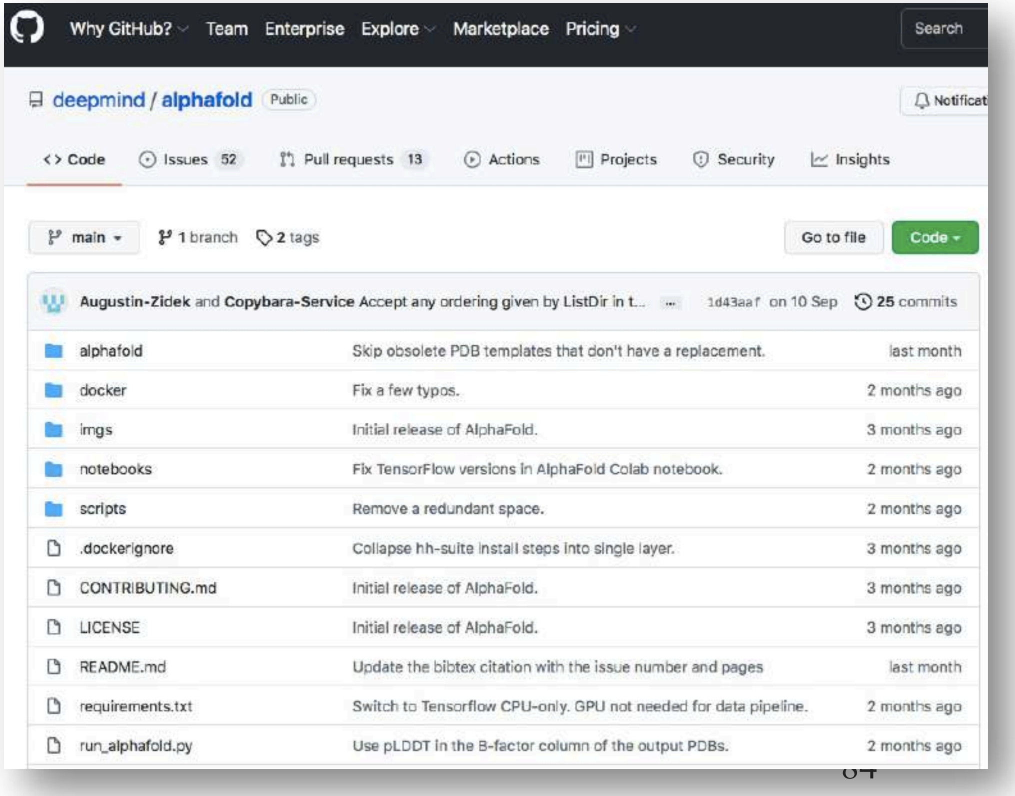
AlphaFold2

- Is there some limitations?

  (i) you can use it at home

Algorithm is published and entirely avalaible (was not the case for v1)

https://github.com/deepmind/alphafold



Augustin-Zidek and Copybara-Service Accept any ordering given by ListDir in t... 1d43aaf on 10 Sep 25 commits

alphafold	Skip obsolete PDB templates that don't have a replacement.	last month
docker	Fix a few typos.	2 months ago
imgs	Initial release of AlphaFold.	3 months ago
notebooks	Fix TensorFlow versions in AlphaFold Colab notebook.	2 months ago
scripts	Remove a redundant space.	2 months ago
.dockerignore	Collapse hh-suite install steps into single layer.	3 months ago
CONTRIBUTING.md	Initial release of AlphaFold.	3 months ago
LICENSE	Initial release of AlphaFold.	3 months ago
README.md	Update the bibtex citation with the issue number and pages	last month
requirements.txt	Switch to Tensorflow CPU-only. GPU not needed for data pipeline.	2 months ago
run_alphafold.py	Use pLDDT in the B-factor column of the output PDBs.	2 months ago

Inserm
Institut national
de la santé et de la recherche médicale

AlphaFold2

SiMB
Bioinformatique & Interactions

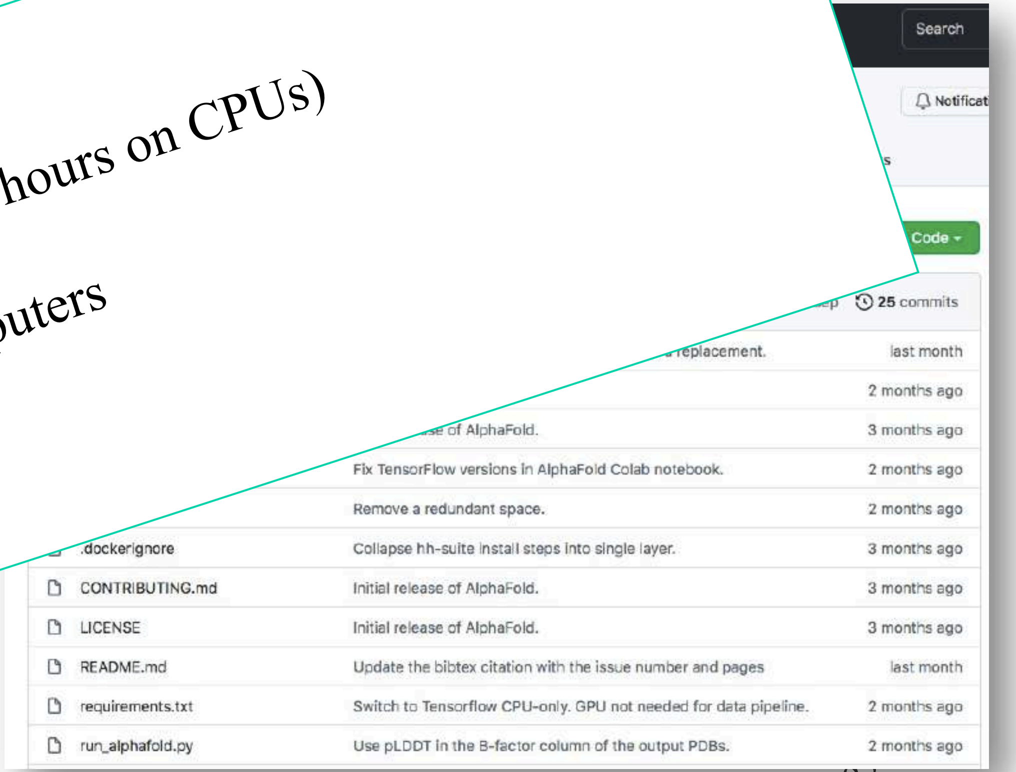
- Is there some limitations?

(i) you can use it at home

Algorithm is published
entirely availa
the c

Obligatory:
2To on HD
A lot of GPUs (10 mn GPUs = = 5 hours on CPUs)
A lot of memories
Translation: not everybody computers

Needed:
A specialist to install it



	replacement.	last month
		2 months ago
	se of AlphaFold.	3 months ago
	Fix TensorFlow versions in AlphaFold Colab notebook.	2 months ago
	Remove a redundant space.	2 months ago
.dockerignore	Collapse hh-suite install steps into single layer.	3 months ago
CONTRIBUTING.md	Initial release of AlphaFold.	3 months ago
LICENSE	Initial release of AlphaFold.	3 months ago
README.md	Update the bibtex citation with the issue number and pages	last month
requirements.txt	Switch to Tensorflow CPU-only. GPU not needed for data pipeline.	2 months ago
run_alphafold.py	Use pLDDT in the B-factor column of the output PDBs.	2 months ago

AlphaFold2

➢ Is there some limitations?

(i) you can use it at home

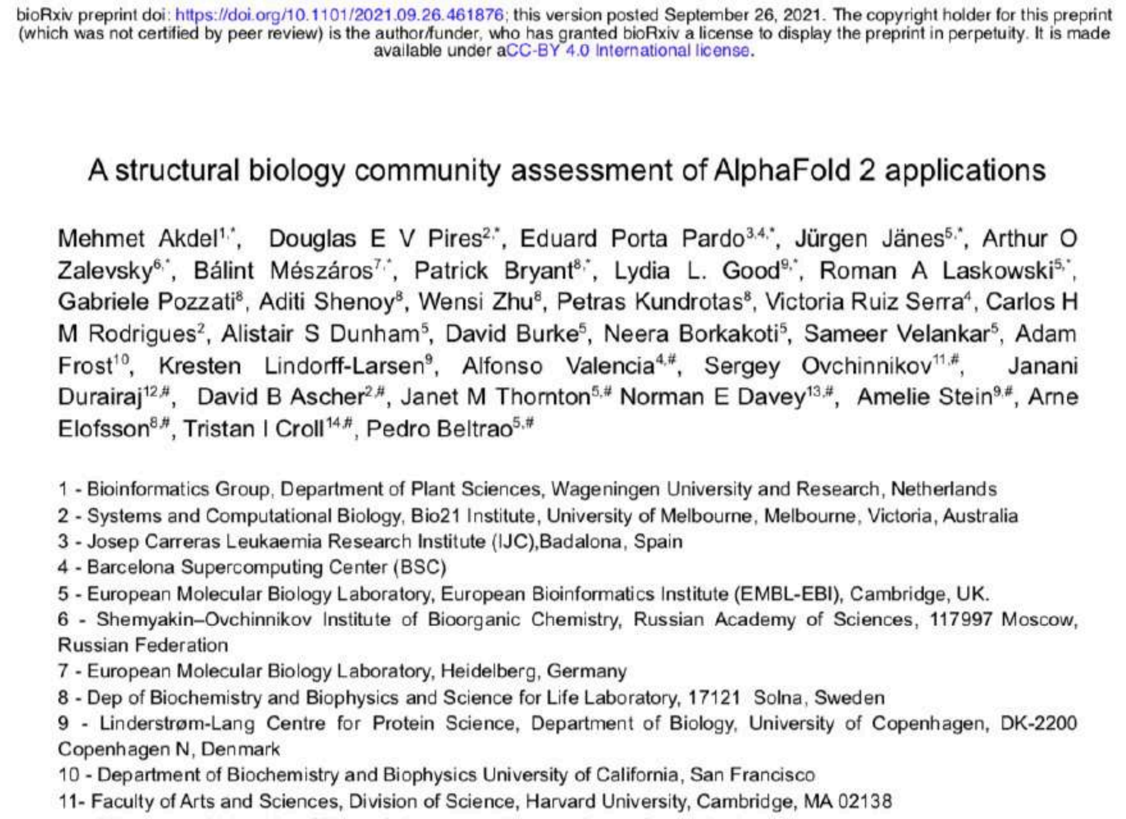
So people have used it.

Recent results from a big consortium

“For 11 proteomes, an average of 25% additional residues can be confidently modelled when compared to homology modelling”

➔ Automatic homology modelling ...

Akdel et al (2021) *bioRxiv* => (2022) *Nat Struct Biol*



A structural biology community assessment of AlphaFold 2 applications

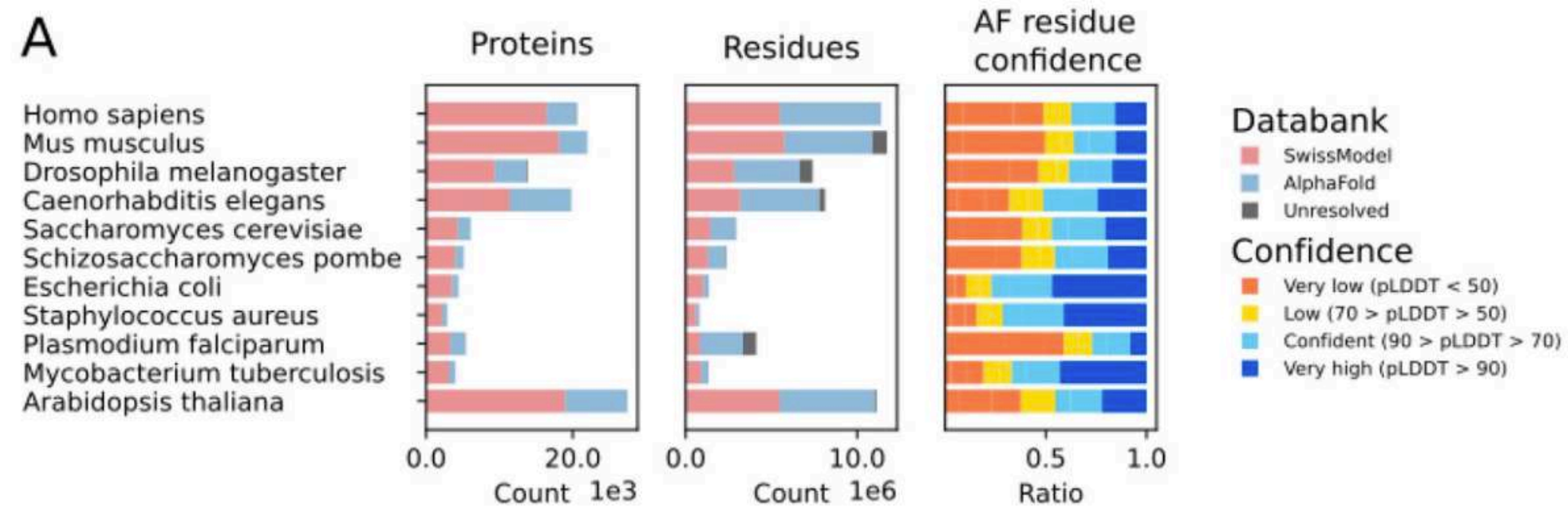
Mehmet Akdel[1,*], Douglas E V Pires[2,*], Eduard Porta Pardo[3,4,*], Jürgen Jänes[5,*], Arthur O Zalevsky[6,*], Bálint Mészáros[7,*], Patrick Bryant[8,*], Lydia L. Good[9,*], Roman A Laskowski[5,*], Gabriele Pozzati[8], Aditi Shenoy[8], Wensi Zhu[8], Petras Kundrotas[8], Victoria Ruiz Serra[4], Carlos H M Rodrigues[2], Alistair S Dunham[5], David Burke[5], Neera Borkakoti[5], Sameer Velankar[5], Adam Frost[10], Kresten Lindorff-Larsen[9], Alfonso Valencia[4,#], Sergey Ovchinnikov[11,#], Janani Durairaj[12,#], David B Ascher[2,#], Janet M Thornton[5,#] Norman E Davey[13,#], Amelie Stein[9,#], Arne Elofsson[8,#], Tristan I Croll[14,#], Pedro Beltrao[5,#]

1 - Bioinformatics Group, Department of Plant Sciences, Wageningen University and Research, Netherlands
2 - Systems and Computational Biology, Bio21 Institute, University of Melbourne, Melbourne, Victoria, Australia
3 - Josep Carreras Leukaemia Research Institute (IJC),Badalona, Spain
4 - Barcelona Supercomputing Center (BSC)
5 - European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Cambridge, UK.
6 - Shemyakin–Ovchinnikov Institute of Bioorganic Chemistry, Russian Academy of Sciences, 117997 Moscow, Russian Federation
7 - European Molecular Biology Laboratory, Heidelberg, Germany
8 - Dep of Biochemistry and Biophysics and Science for Life Laboratory, 17121 Solna, Sweden
9 - Linderstrøm-Lang Centre for Protein Science, Department of Biology, University of Copenhagen, DK-2200 Copenhagen N, Denmark
10 - Department of Biochemistry and Biophysics University of California, San Francisco
11- Faculty of Arts and Sciences, Division of Science, Harvard University, Cambridge, MA 02138

AlphaFold2

- Is there some limitations?

  (i) you can use it at home



modelling

➔ Automatic homology modelling ...

Akdel et al (2021) *bioRxiv* => (2022) *Nat Struct Biol*

4 - Barcelona Supercomputing Center (BSC)
5 - European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Cambridge, UK.
6 - Shemyakin–Ovchinnikov Institute of Bioorganic Chemistry, Russian Academy of Sciences, 117997 Moscow, Russian Federation
7 - European Molecular Biology Laboratory, Heidelberg, Germany
8 - Dep of Biochemistry and Biophysics and Science for Life Laboratory, 17121 Solna, Sweden
9 - Linderstrøm-Lang Centre for Protein Science, Department of Biology, University of Copenhagen, DK-2200 Copenhagen N, Denmark
10 - Department of Biochemistry and Biophysics University of California, San Francisco
11- Faculty of Arts and Sciences, Division of Science, Harvard University, Cambridge, MA 02138

AlphaFold2

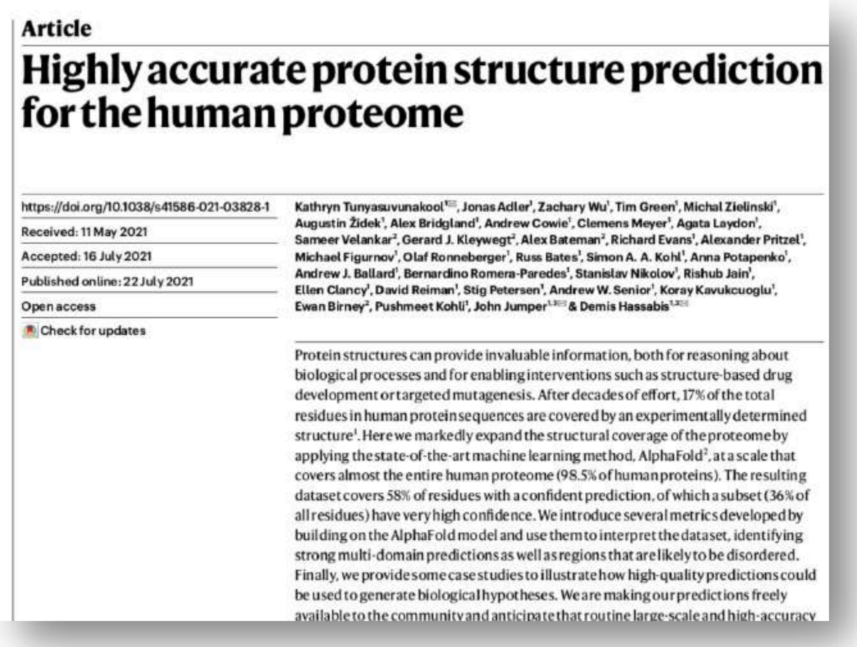
- Is there some limitations?

  (ii) there is a database of already done model

EBI: https://www.alphafold.ebi.ac.uk

AlphaFold2, at a scale that covers .. 98.5% of human proteins. The resulting dataset covers 58% of residues with a confident prediction, of which a subset (36% of all residues) have very high confidence.
➔ 36% for drug design

Tunyasuvunakool K, et al (2021), *Nature*. 596(7873):590-596.



Article

Highly accurate protein structure prediction for the human proteome

https://doi.org/10.1038/s41586-021-03828-1
Received: 11 May 2021
Accepted: 16 July 2021
Published online: 22 July 2021
Open access
Check for updates

Kathryn Tunyasuvunakool[1], Jonas Adler[1], Zachary Wu[1], Tim Green[1], Michal Zielinski[1], Augustin Žídek[1], Alex Bridgland[1], Andrew Cowie[1], Clemens Meyer[1], Agata Laydon[1], Sameer Velankar[2], Gerard J. Kleywegt[2], Alex Bateman[2], Richard Evans[1], Alexander Pritzel[1], Michael Figurnov[1], Olaf Ronneberger[1], Russ Bates[1], Simon A. A. Kohl[1], Anna Potapenko[1], Andrew J. Ballard[1], Bernardino Romera-Paredes[1], Stanislav Nikolov[1], Rishub Jain[1], Ellen Clancy[1], David Reiman[1], Stig Petersen[1], Andrew W. Senior[1], Koray Kavukcuoglu[1], Ewan Birney[2], Pushmeet Kohli[1], John Jumper[1,3] & Demis Hassabis[1,3]

Protein structures can provide invaluable information, both for reasoning about biological processes and for enabling interventions such as structure-based drug development or targeted mutagenesis. After decades of effort, 17% of the total residues in human protein sequences are covered by an experimentally determined structure[1]. Here we markedly expand the structural coverage of the proteome by applying the state-of-the-art machine learning method, AlphaFold[2], at a scale that covers almost the entire human proteome (98.5% of human proteins). The resulting dataset covers 58% of residues with a confident prediction, of which a subset (36% of all residues) have very high confidence. We introduce several metrics developed by building on the AlphaFold model and use them to interpret the dataset, identifying strong multi-domain predictions as well as regions that are likely to be disordered. Finally, we provide some case studies to illustrate how high-quality predictions could be used to generate biological hypotheses. We are making our predictions freely available to the community and anticipate that routine large-scale and high-accuracy

AlphaFold2

- Is there some limitations?

  (ii) there is a database of already done model

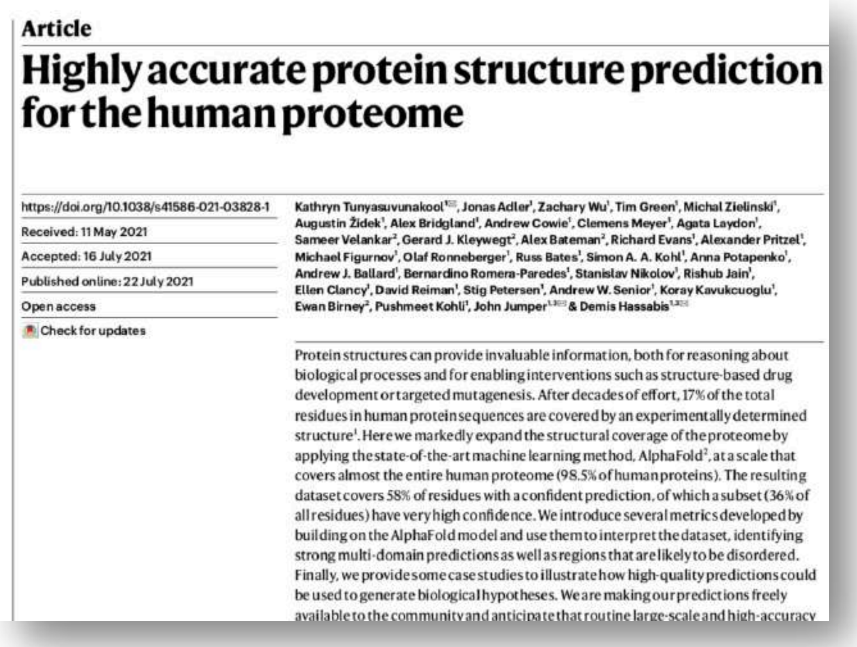
EBI: https://www.alphafold.ebi.ac.uk

AlphaFold2, at a scale that covers .. 98.5% of human proteins. The resulting dataset covers 58% of residues with a confident prediction, of which a subset (36% of all residues) have very high confidence.

➔ 36% for drug design

➔ 42% question about fold

Tunyasuvunakool K, et al (2021), *Nature*. 596(7873):590-596.



Article

Highly accurate protein structure prediction for the human proteome

https://doi.org/10.1038/s41586-021-03828-1

Received: 11 May 2021

Accepted: 16 July 2021

Published online: 22 July 2021

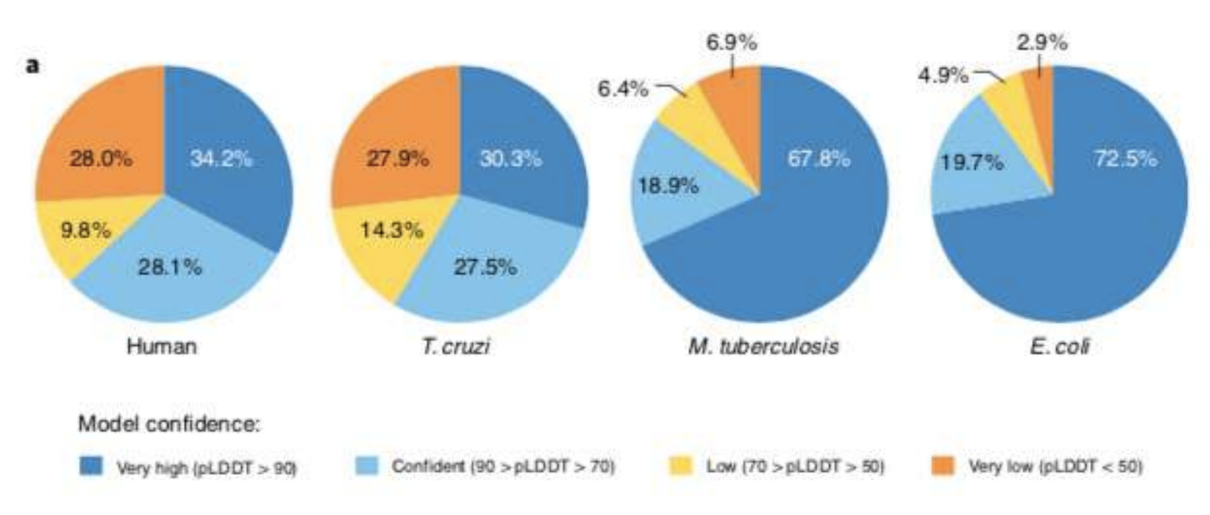
Open access

Check for updates

Kathryn Tunyasuvunakool[1]✉, Jonas Adler[1], Zachary Wu[1], Tim Green[1], Michal Zielinski[1], Augustin Žídek[1], Alex Bridgland[1], Andrew Cowie[1], Clemens Meyer[1], Agata Laydon[1], Sameer Velankar[2], Gerard J. Kleywegt[2], Alex Bateman[2], Richard Evans[1], Alexander Pritzel[1], Michael Figurnov[1], Olaf Ronneberger[1], Russ Bates[1], Simon A. A. Kohl[1], Anna Potapenko[1], Andrew J. Ballard[1], Bernardino Romera-Paredes[1], Stanislav Nikolov[1], Rishub Jain[1], Ellen Clancy[1], David Reiman[1], Stig Petersen[1], Andrew W. Senior[1], Koray Kavukcuoglu[1], Ewan Birney[2], Pushmeet Kohli[1], John Jumper[1,3]✉ & Demis Hassabis[1,3]✉

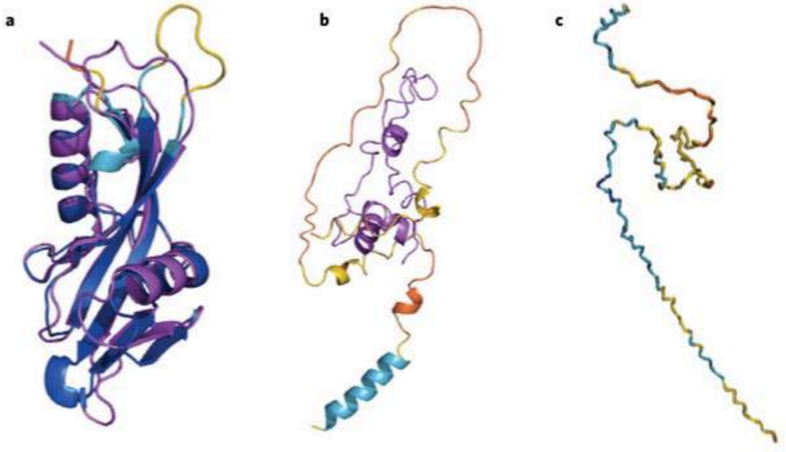
Protein structures can provide invaluable information, both for reasoning about biological processes and for enabling interventions such as structure-based drug development or targeted mutagenesis. After decades of effort, 17% of the total residues in human protein sequences are covered by an experimentally determined structure[1]. Here we markedly expand the structural coverage of the proteome by applying the state-of-the-art machine learning method, AlphaFold[2], at a scale that covers almost the entire human proteome (98.5% of human proteins). The resulting dataset covers 58% of residues with a confident prediction, of which a subset (36% of all residues) have very high confidence. We introduce several metrics developed by building on the AlphaFold model and use them to interpret the dataset, identifying strong multi-domain predictions as well as regions that are likely to be disordered. Finally, we provide some case studies to illustrate how high-quality predictions could be used to generate biological hypotheses. We are making our predictions freely available to the community and anticipate that routine large-scale and high-accuracy

AlphaFold2



- *Protein structures predicted using artificial intelligence will aid medical research, but the greatest benefit will come if clinical data can be similarly used to better understand human disease.*

Janet M. Thornton, Roman A. Laskowski and Neera Borkakoti. (2021) *Nat Med*. 27:1666-1671.



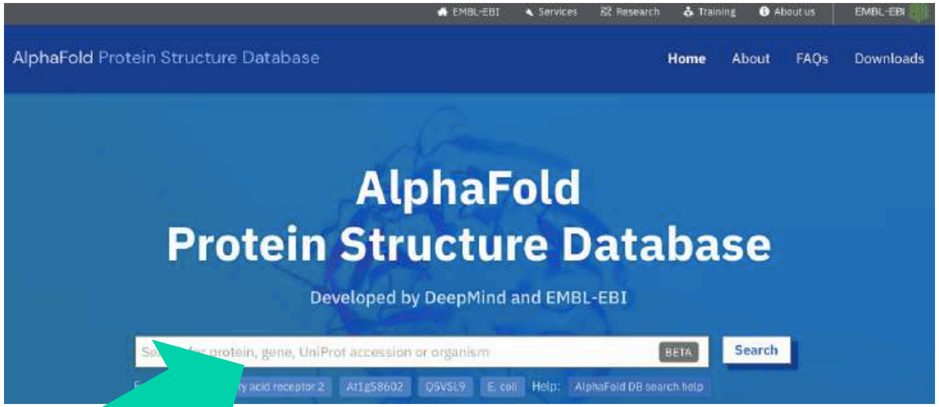
The good, the bad and the uglý˘

AlphaFold2

- Is there some limitations?

(ii) there is a database of already done model

EBI: https://www.alphafold.ebi.ac.uk



AlphaFold
Protein Structure Database

Developed by DeepMind and EMBL-EBI

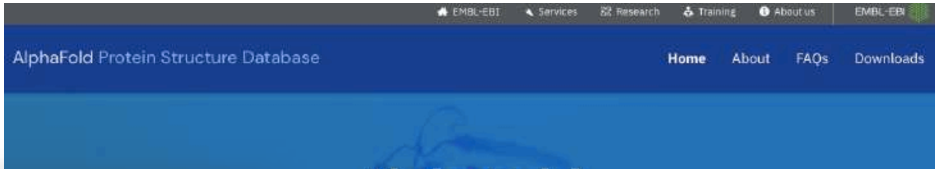
So you ask your favourite protein

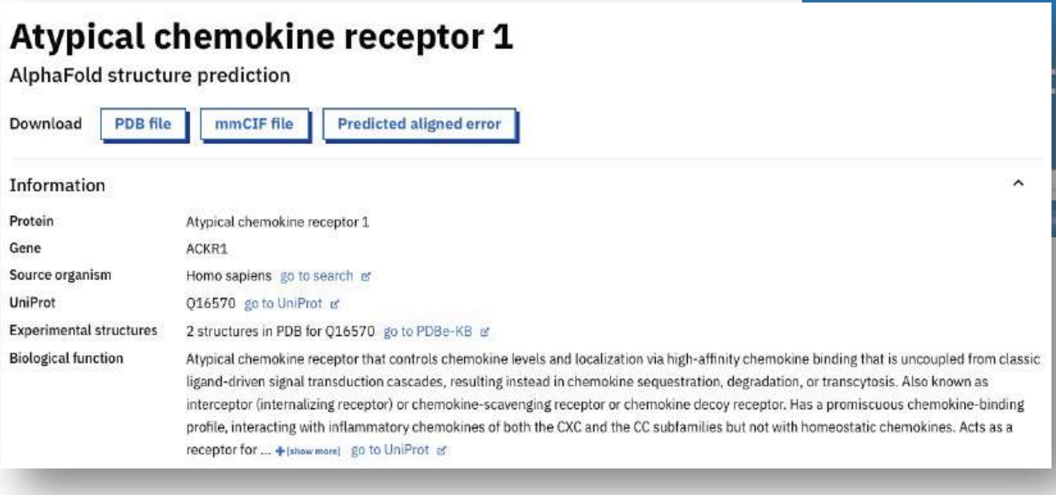
AlphaFold2

- ➤ Is there some limitations?

  (ii) there is a database of already done model

EBI: https://www.alphafold.ebi.ac.uk





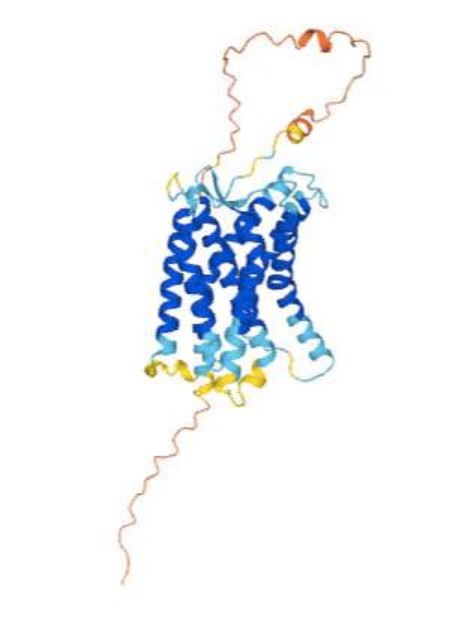
Atypical chemokine receptor 1

AlphaFold structure prediction

Download PDB file mmCIF file Predicted aligned error

Information

Protein	Atypical chemokine receptor 1
Gene	ACKR1
Source organism	Homo sapiens go to search
UniProt	Q16570 go to UniProt
Experimental structures	2 structures in PDB for Q16570 go to PDBe-KB
Biological function	Atypical chemokine receptor that controls chemokine levels and localization via high-affinity chemokine binding that is uncoupled from classic ligand-driven signal transduction cascades, resulting instead in chemokine sequestration, degradation, or transcytosis. Also known as interceptor (internalizing receptor) or chemokine-scavenging receptor or chemokine decoy receptor. Has a promiscuous chemokine-binding profile, interacting with inflammatory chemokines of both the CXC and the CC subfamilies but not with homeostatic chemokines. Acts as a receptor for ... +[show more] go to UniProt



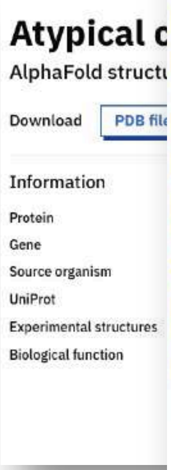
Yes, it is a transmembrane one···

And i do not like the final model...

AlphaFold2

- Is there some limitations?

EBI: https://v



Atypical c

AlphaFold structu

Download PDB fil

Information

Protein

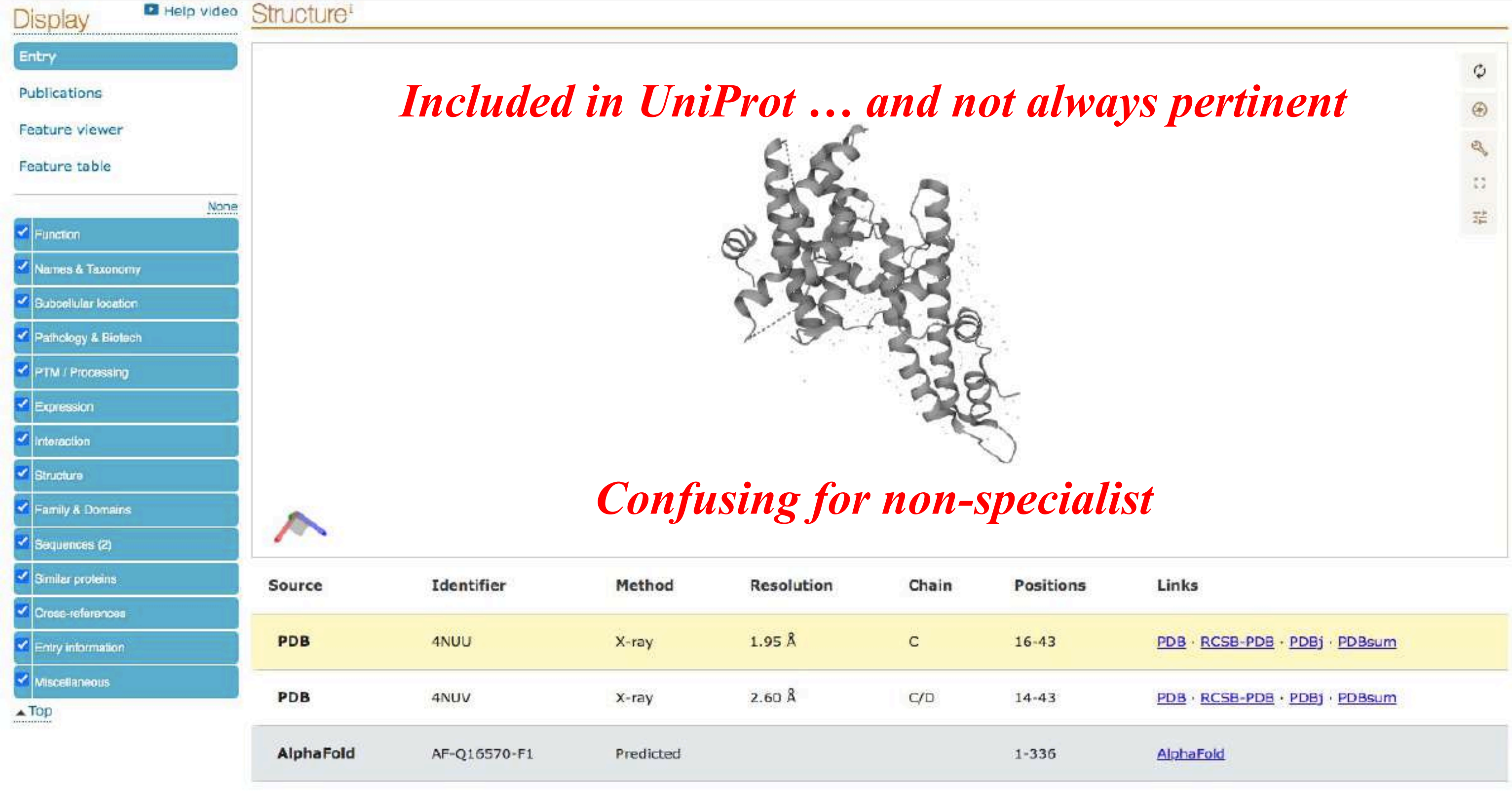
Gene

Source organism

UniProt

Experimental structures

Biological function



Included in UniProt … and not always pertinent

Confusing for non-specialist

Source	Identifier	Method	Resolution	Chain	Positions	Links
PDB	4NUU	X-ray	1.95 Å	C	16-43	PDB · RCSB-PDB · PDBj · PDBsum
PDB	4NUV	X-ray	2.60 Å	C/D	14-43	PDB · RCSB-PDB · PDBj · PDBsum
AlphaFold	AF-Q16570-F1	Predicted			1-336	AlphaFold

Yes, it is a transmembrane one…

And i do not like the final model...

AlphaFold2

- Is there some limitations?

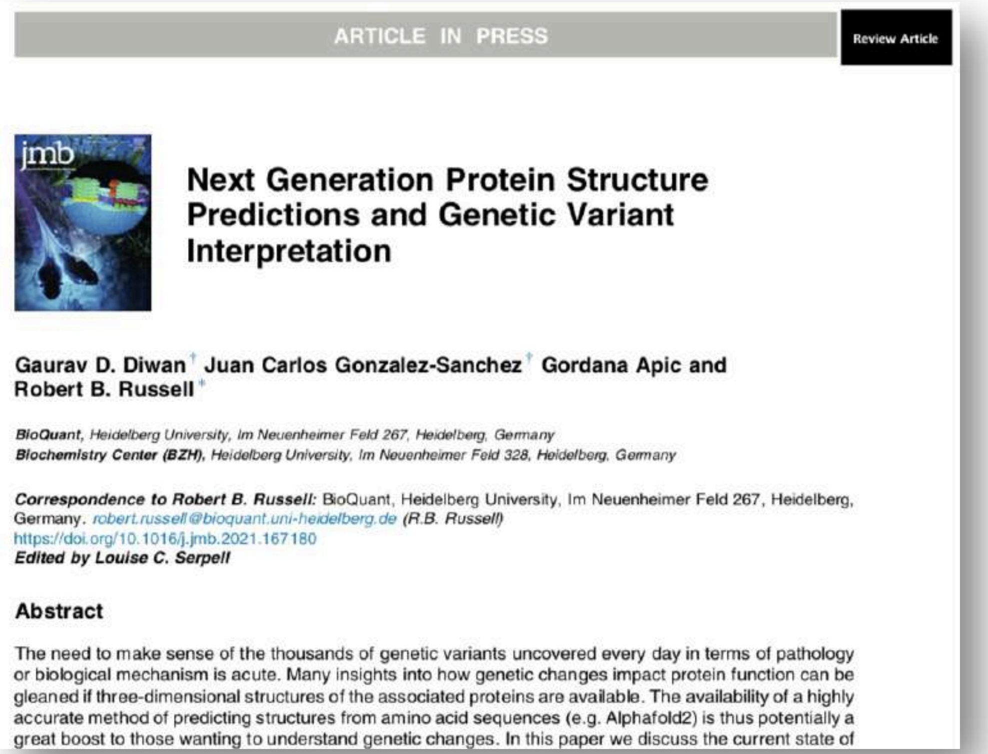
SNPs == pathologies

AlphaFold2

➢ Is there some limitations?

An accurate prediction of topology can certainly help these efforts, but what is really needed is a means to study precise side-chain orientations, interactions with non-protein molecules and the dynamics of the system. Not to mention, one typically makes use of a host of other non-structural information, such as evolutionary conservation, sequence annotation data and, of course, the vast and growing scientific literature.

Diwan GD, Gonzalez-Sanchez JC, Apic G, Russell RB. (2021), *J Mol Biol*. 4:167180.



ARTICLE IN PRESS

Review Article

jmb

Next Generation Protein Structure Predictions and Genetic Variant Interpretation

Gaurav D. Diwan[†] **Juan Carlos Gonzalez-Sanchez**[†] **Gordana Apic and Robert B. Russell**[*]

***BioQuant**, Heidelberg University, Im Neuenheimer Feld 267, Heidelberg, Germany*
***Biochemistry Center (BZH)**, Heidelberg University, Im Neuenheimer Feld 328, Heidelberg, Germany*

Correspondence to Robert B. Russell: BioQuant, Heidelberg University, Im Neuenheimer Feld 267, Heidelberg, Germany. *robert.russell@bioquant.uni-heidelberg.de (R.B. Russell)*

https://doi.org/10.1016/j.jmb.2021.167180
Edited by Louise C. Serpell

Abstract

The need to make sense of the thousands of genetic variants uncovered every day in terms of pathology or biological mechanism is acute. Many insights into how genetic changes impact protein function can be gleaned if three-dimensional structures of the associated proteins are available. The availability of a highly accurate method of predicting structures from amino acid sequences (e.g. Alphafold2) is thus potentially a great boost to those wanting to understand genetic changes. In this paper we discuss the current state of

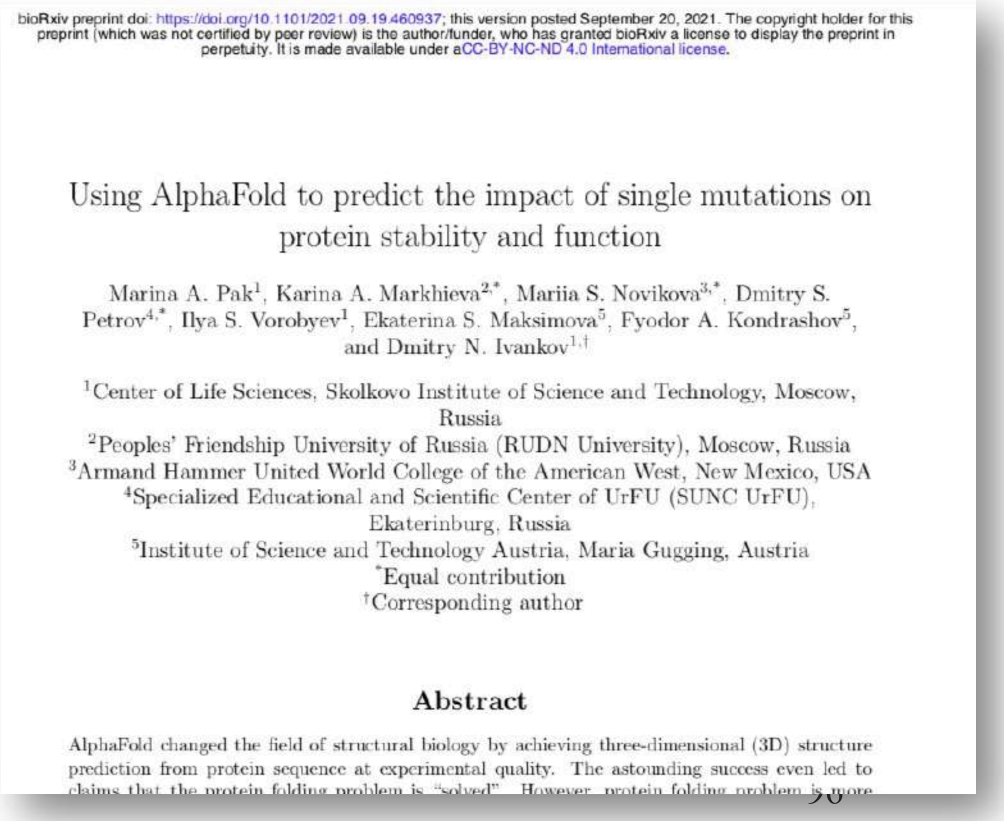
AlphaFold2

- Is there some limitations?

Evaluation of variants ?

We found a very weak or no correlation between AlphaFold output metrics and change of protein stability or fluorescence. Our results imply that AlphaFold cannot be immediately applied to other problems or applications in protein folding.

Pak et al (2021), BioRxiv
https://www.biorxiv.org/lookup/doi/10.1101/2021.09.19.460937



Using AlphaFold to predict the impact of single mutations on protein stability and function

Marina A. Pak[1], Karina A. Markhieva[2,*], Mariia S. Novikova[3,*], Dmitry S. Petrov[4,*], Ilya S. Vorobyev[1], Ekaterina S. Maksimova[5], Fyodor A. Kondrashov[5], and Dmitry N. Ivankov[1,†]

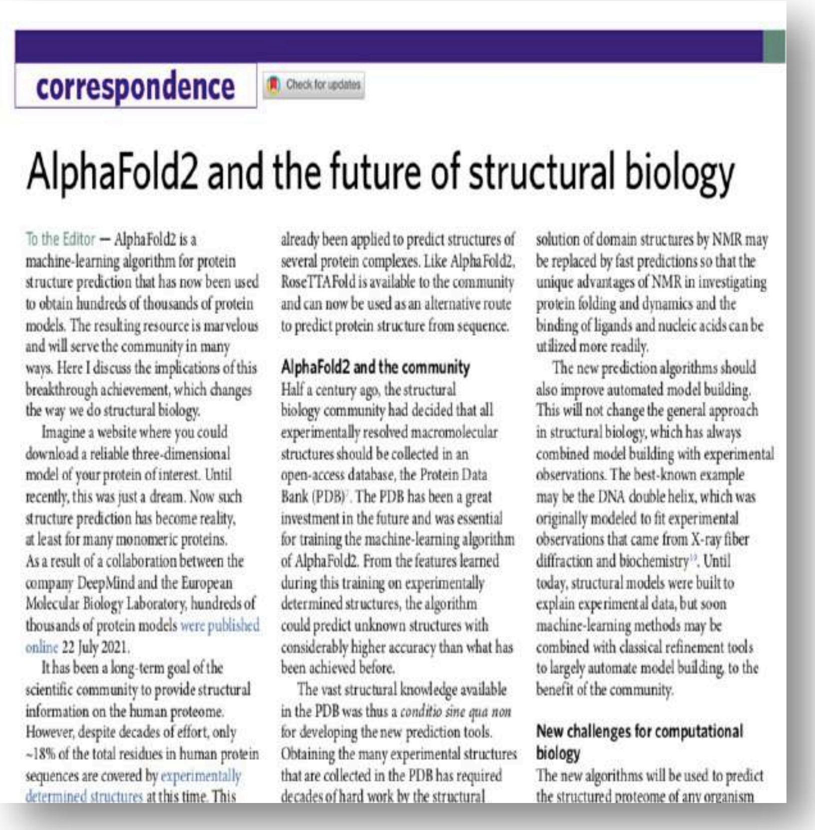
[1]Center of Life Sciences, Skolkovo Institute of Science and Technology, Moscow, Russia
[2]Peoples' Friendship University of Russia (RUDN University), Moscow, Russia
[3]Armand Hammer United World College of the American West, New Mexico, USA
[4]Specialized Educational and Scientific Center of UrFU (SUNC UrFU), Ekaterinburg, Russia
[5]Institute of Science and Technology Austria, Maria Gugging, Austria
*Equal contribution
†Corresponding author

Abstract

AlphaFold changed the field of structural biology by achieving three-dimensional (3D) structure prediction from protein sequence at experimental quality. The astounding success even led to claims that the protein folding problem is "solved". However, protein folding problem is more

AlphaFold2

- The new prediction algorithms do not solve the protein folding problem in the sense that they do not reveal how a sequence encodes three-dimensional structure.
- However, they do solve the problem in practical terms, as they can reliably predict structure from sequence, ***at least in many cases.***
- ***Although only time will tell***, this advance is expected to represent a breakthrough in structural biology that is comparable to previous major advances,



correspondence

Check for updates

AlphaFold2 and the future of structural biology

To the Editor — AlphaFold2 is a machine-learning algorithm for protein structure prediction that has now been used to obtain hundreds of thousands of protein models. The resulting resource is marvelous and will serve the community in many ways. Here I discuss the implications of this breakthrough achievement, which changes the way we do structural biology.

Imagine a website where you could download a reliable three-dimensional model of your protein of interest. Until recently, this was just a dream. Now such structure prediction has become reality, at least for many monomeric proteins. As a result of a collaboration between the company DeepMind and the European Molecular Biology Laboratory, hundreds of thousands of protein models were published online 22 July 2021.

It has been a long-term goal of the scientific community to provide structural information on the human proteome. However, despite decades of effort, only ~18% of the total residues in human protein sequences are covered by experimentally determined structures at this time. This

already been applied to predict structures of several protein complexes. Like AlphaFold2, RoseTTAFold is available to the community and can now be used as an alternative route to predict protein structure from sequence.

AlphaFold2 and the community

Half a century ago, the structural biology community had decided that all experimentally resolved macromolecular structures should be collected in an open-access database, the Protein Data Bank (PDB)[1]. The PDB has been a great investment in the future and was essential for training the machine-learning algorithm of AlphaFold2. From the features learned during this training on experimentally determined structures, the algorithm could predict unknown structures with considerably higher accuracy than what has been achieved before.

The vast structural knowledge available in the PDB was thus a *conditio sine qua non* for developing the new prediction tools. Obtaining the many experimental structures that are collected in the PDB has required decades of hard work by the structural

solution of domain structures by NMR may be replaced by fast predictions so that the unique advantages of NMR in investigating protein folding and dynamics and the binding of ligands and nucleic acids can be utilized more readily.

The new prediction algorithms should also improve automated model building. This will not change the general approach in structural biology, which has always combined model building with experimental observations. The best-known example may be the DNA double helix, which was originally modeled to fit experimental observations that came from X-ray fiber diffraction and biochemistry[2,3]. Until today, structural models were built to explain experimental data, but soon machine-learning methods may be combined with classical refinement tools to largely automate model building, to the benefit of the community.

New challenges for computational biology

The new algorithms will be used to predict the structured proteome of any organism

Cramer P. (2021) *Nat Struct Mol Biol*. 28(9):704-705.

AlphaFold2

Conclusions on AF2

- Yes, it is excellent

AlphaFold2

Conclusions on AF2

- Yes, it is excellent
- No, it is not perfect and a lot of works are still needed.

AlphaFold2

Conclusions on AF2

- Yes, it is excellent
- No, it is not perfect and a lot of works are still needed.
- So, an excellent new tool, with results that must be evaluated (*as always*)

An analysis

Biochimie 207 (2023) 11–19

Contents lists available at ScienceDirect

Biochimie

journal homepage: www.elsevier.com/locate/biochi

An agnostic analysis of the human AlphaFold2 proteome using local protein conformations

Alexandre G. de Brevern[*]

Université Paris Cité and Université des Antilles and Université de la Réunion, INSERM UMR_S 1134, BIGR, DSIMB Bioinformatics team, F-75014, Paris, France

ARTICLE INFO

Article history:
Received 26 April 2022
Received in revised form
14 October 2022
Accepted 17 November 2022
Available online 20 November 2022

Keywords:
Secondary structure
Helix
Sheet; turn
polyproline II
Structural alphabet; protein structure
Deep learning

ABSTRACT

Knowledge of the 3D structure of proteins is a valuable asset for understanding their precise biological mechanisms. However, the cost of production of 3D structures and experimental difficulties limit their obtaining. The proposal of 3D structural models is consequently an appealing alternative. The release of the AlphaFold Deep Learning approach has revolutionized the field. The recent near-complete human proteome proposal makes it possible to analyse large amounts of data and evaluate the results of the approach in greater depth. The 3D human proteome was thus analysed in light of the classic secondary structures, and many less-used protein local conformations (PolyProline II helices, type of γ-turns, of β-turns and of β-bulges, curvature of the helices, and a structural alphabet). Without questioning the global quality of the approach, this analysis highlights certain local conformations, which maybe poorly predicted and they could therefore be better addressed.

1. Introduction

Proteins are essential constituents of cells, one of the major macromolecules of life. Composed of 20 amino acids, proteins ensure a variety of essential biological functions. The dogma of molecular biology emphases the link between the nucleic sequence and the protein sequence to arrive at protein structures and their functions. Nowadays, access to these sequences has become particularly cheap [1]. Databases contain millions of protein sequences, while the three-dimensional structures of proteins are much more difficult to obtain experimentally [2].

Hence for more than 30 years, different computational approaches have been implemented to propose three-dimensional (3D) structural models of proteins from their amino acid sequence [3]. The classic categories of these approaches include homology or comparative modelling, threading, *ab initio*, *de novo* approaches, and meta-servers; these last combined several approaches. It is possible to notice the best-known tools such as Modeller [4], the golden standard of homology/comparative modelling and *de novo* most recent approaches Rosetta and I-TASSER [5–7]. These two last have won numerous Critical Assessment for Protein Structure Prediction (CASP) meetings [8,9].

Arrived at the CASP13 (2018), the company DeepMind presented its new Deep Learning approach named AlphaFold [10]. It won the Free Modelling category, i.e. the prediction of novel protein folds [11], whereas template-Based Modelling category, i.e. protein folds already found in the Protein Data Bank, was won by Zhang's group [12]. Two years later, AlphaFold version 2 obtained particularly remarkable at CASP14 (2020) [13,14]; some models were within the uncertainties of the experimental resolution, an impressive result. This improvement combined the delicate use of evolution, contacts within proteins, and large GPU computing power that allowed the implementation of a particularly complex and elegant architecture [15,16].

AlphaFold 2 was a hot topic for 2020 and 2021 [17–20], leading to a revolution in protein structural model building [21,22] and opening potential new opportunities, e.g. new drug design researches [23,24]. Three points can be noticed (i) the code can be downloaded freely on GitHub (https://github.com/deepmind/alphafold) and is really useable [25], (ii) different online notebooks for non-specialists are easy to use (e.g. https://colab.research.google.com/github/sokrypton/Colabfold/blob/main/batch/AlphaFold2_batch.ipynb) [26], and (iii) EBI provides structural model databases [27]. Indeed, model building is expensive in

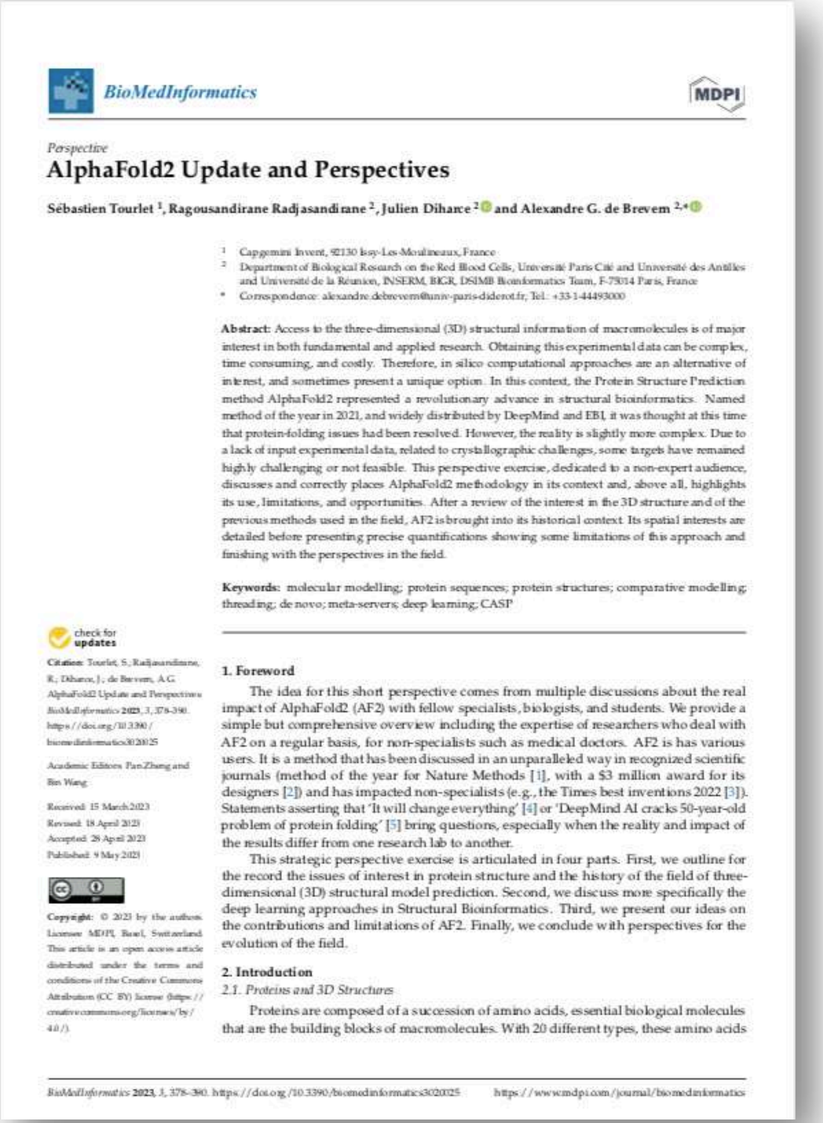
[*] INSERM UMR_S 1134, DSIMB Bioinformatics team, Paris Cité, 8 rue Maria Helena Vieira da Silva, 75014, Paris, France.
E-mail address: alexandre.debrevern@u-paris.fr.

https://doi.org/10.1016/j.biochi.2022.11.009
0300-9084/© 2022 Elsevier B.V. and Société Française de Biochimie et Biologie Moléculaire (SFBBM). All rights reserved.

Not all local conformations are properly predicted !

de Brevern A.G. An agnostic analysis of the human AlphaFold2 proteome using local protein conformations. *Biochimie* (2023) **207**:11-19.

Perspectives



BioMedInformatics

Perspective

AlphaFold2 Update and Perspectives

Sébastien Tourlet [1], Ragousandirane Radjasandirane [2], Julien Diharce [2] and Alexandre G. de Brevern [2,*]

[1] Capgemini Invent, 92130 Issy-Les-Moulineaux, France
[2] Department of Biological Research on the Red Blood Cells, Université Paris Cité and Université des Antilles and Université de la Réunion, INSERM, BIGR, DSIMB Bioinformatics Team, F-75014 Paris, France
* Correspondence: alexandre.debrevern@univ-paris-diderot.fr; Tel.: +33-1-44493000

Abstract: Access to the three-dimensional (3D) structural information of macromolecules is of major interest in both fundamental and applied research. Obtaining this experimental data can be complex, time consuming, and costly. Therefore, in silico computational approaches are an alternative of interest, and sometimes present a unique option. In this context, the Protein Structure Prediction method AlphaFold2 represented a revolutionary advance in structural bioinformatics. Named method of the year in 2021, and widely distributed by DeepMind and EBI, it was thought at this time that protein-folding issues had been resolved. However, the reality is slightly more complex. Due to a lack of input experimental data, related to crystallographic challenges, some targets have remained highly challenging or not feasible. This perspective exercise, dedicated to a non-expert audience, discusses and correctly places AlphaFold2 methodology in its context and, above all, highlights its use, limitations, and opportunities. After a review of the interest in the 3D structure and of the previous methods used in the field, AF2 is brought into its historical context. Its spatial interests are detailed before presenting precise quantifications showing some limitations of this approach and finishing with the perspectives in the field.

Keywords: molecular modelling; protein sequences; protein structures; comparative modelling; threading; de novo; meta-servers; deep learning; CASP

check for updates

Citation: Tourlet, S.; Radjasandirane, R.; Diharce, J.; de Brevern, A.G. AlphaFold2 Update and Perspectives. *BioMedInformatics* **2023**, *3*, 378–390. https://doi.org/10.3390/biomedinformatics3020025

Academic Editors: Pan Zhang and Bin Wang

Received: 15 March 2023
Revised: 18 April 2023
Accepted: 28 April 2023
Published: 9 May 2023

1. Foreword

The idea for this short perspective comes from multiple discussions about the real impact of AlphaFold2 (AF2) with fellow specialists, biologists, and students. We provide a simple but comprehensive overview including the expertise of researchers who deal with AF2 on a regular basis, for non-specialists such as medical doctors. AF2 is has various users. It is a method that has been discussed in an unparalleled way in recognized scientific journals (method of the year for Nature Methods [1], with a $3 million award for its designers [2]) and has impacted non-specialists (e.g., the Times best inventions 2022 [3]). Statements asserting that 'It will change everything' [4] or 'DeepMind AI cracks 50-year-old problem of protein folding' [5] bring questions, especially when the reality and impact of the results differ from one research lab to another.

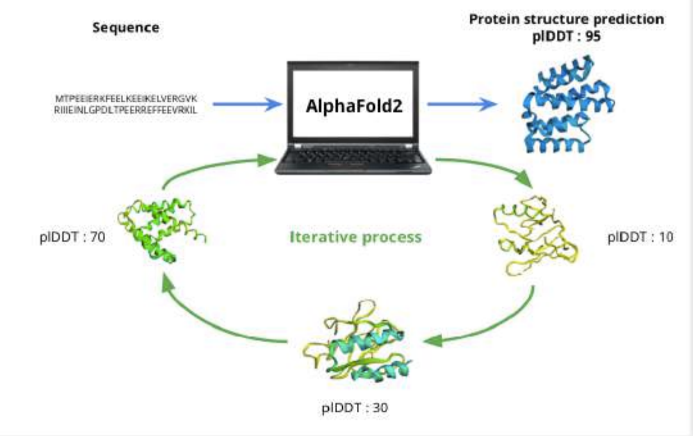
This strategic perspective exercise is articulated in four parts. First, we outline for the record the issues of interest in protein structure and the history of the field of three-dimensional (3D) structural model prediction. Second, we discuss more specifically the deep learning approaches in Structural Bioinformatics. Third, we present our ideas on the contributions and limitations of AF2. Finally, we conclude with perspectives for the evolution of the field.

2. Introduction

2.1. Proteins and 3D Structures

Proteins are composed of a succession of amino acids, essential biological molecules that are the building blocks of macromolecules. With 20 different types, these amino acids

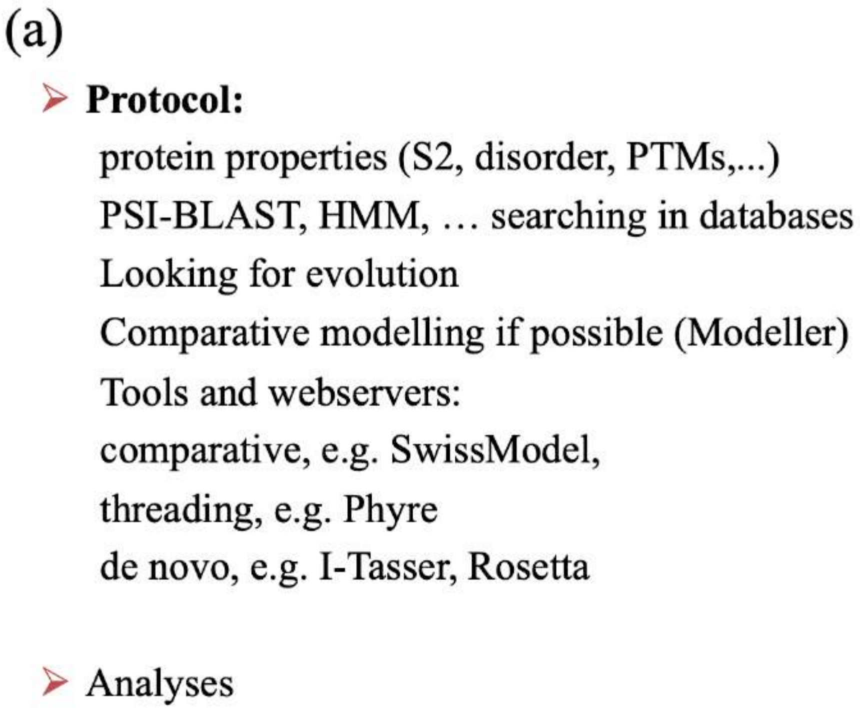
Analyses of the impact of AlphaFold2 on the daily life of a Structural Bioinformatics lab.


Sequence
Protein structure prediction
pIDDT : 95
MTPEEIERKFEELKEEIKELVERGVK
RIIIEINLGPDLTPEERREFFEEVRKIL
AlphaFold2
pIDDT : 70
Iterative process
pIDDT : 10
pIDDT : 30

Tourlet S., Radjasandirane R., Diharce J., de Brevern A.G. AlphaFold2 Update and Perspectives. *BioMedInformatics* (2023) **3**(2), 378-390.

AlphaFold2

What I was doing before AlphaFold2



(a)

- **Protocol:**

  protein properties (S2, disorder, PTMs,...)

  PSI-BLAST, HMM, … searching in databases

  Looking for evolution

  Comparative modelling if possible (Modeller)

  Tools and webservers:

  comparative, e.g. SwissModel,

  threading, e.g. Phyre

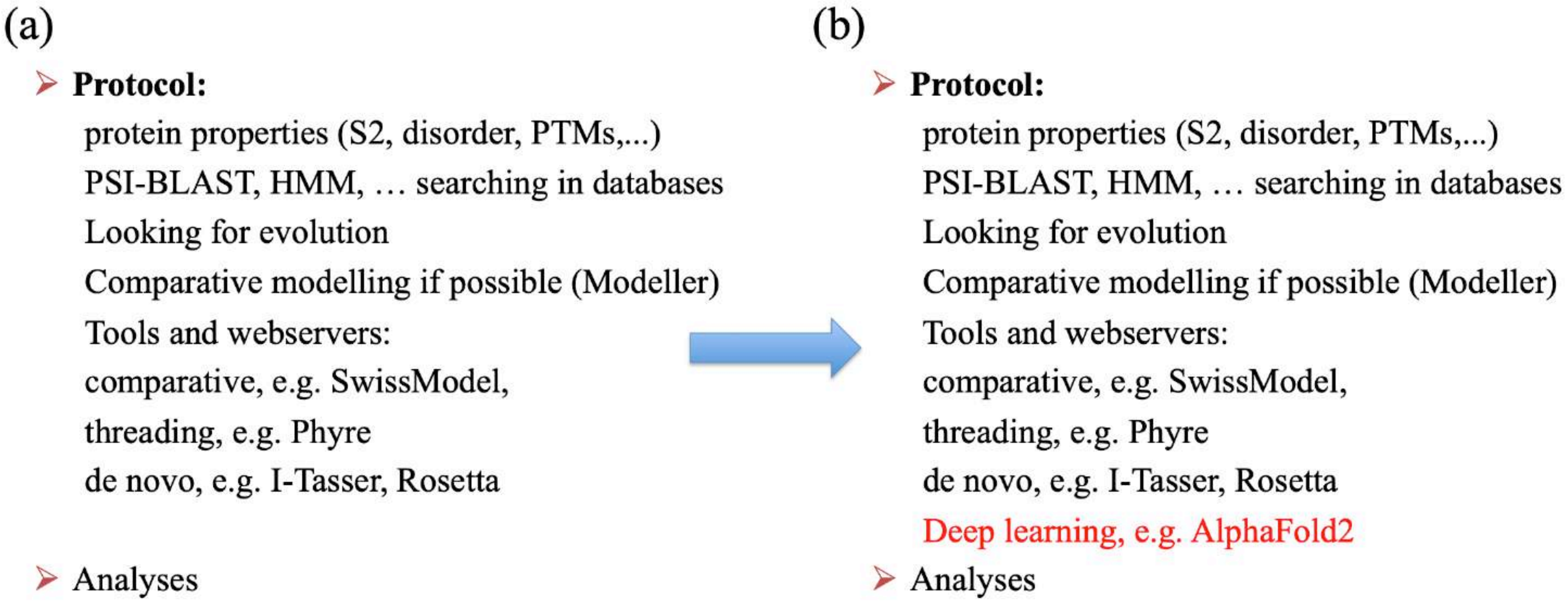
  de novo, e.g. I-Tasser, Rosetta

- Analyses

Tourlet S., Radjasandirane R., Diharce J., de Brevern A.G. AlphaFold2 Update and Perspectives. *BioMedInformatics* (2023) **3**(2), 378-390.

AlphaFold2

What I was doing before AlphaFold2 *What I am doing now*



(a)

- **Protocol:**
  protein properties (S2, disorder, PTMs,...)
  PSI-BLAST, HMM, … searching in databases
  Looking for evolution
  Comparative modelling if possible (Modeller)
  Tools and webservers:
  comparative, e.g. SwissModel,
  threading, e.g. Phyre
  de novo, e.g. I-Tasser, Rosetta
- Analyses

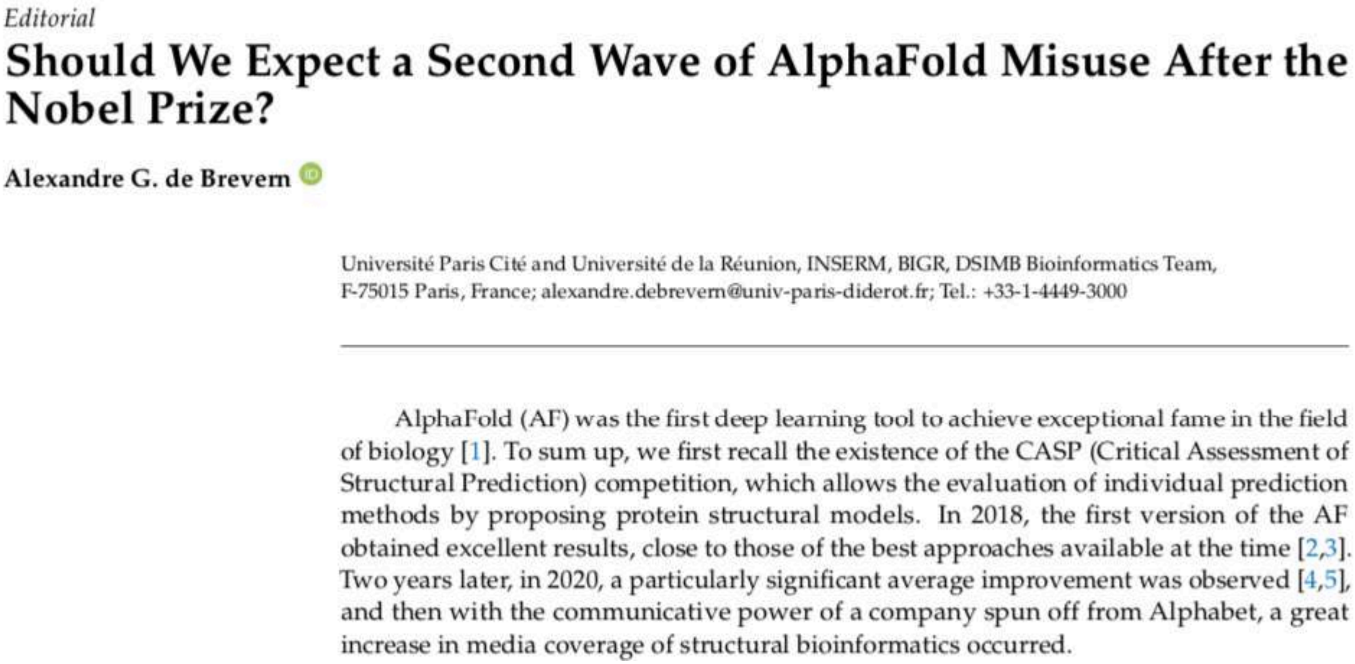
(b)

- **Protocol:**
  protein properties (S2, disorder, PTMs,...)
  PSI-BLAST, HMM, … searching in databases
  Looking for evolution
  Comparative modelling if possible (Modeller)
  Tools and webservers:
  comparative, e.g. SwissModel,
  threading, e.g. Phyre
  de novo, e.g. I-Tasser, Rosetta
  Deep learning, e.g. AlphaFold2
- Analyses

Tourlet S., Radjasandirane R., Diharce J., de Brevern A.G. AlphaFold2 Update and Perspectives. *BioMedInformatics* (2023) **3**(2), 378-390.

Last paper

- *Editorial :* **Should We Expect a Second Wave of AlphaFold Misuse After the Nobel Prize?**



Editorial

Should We Expect a Second Wave of AlphaFold Misuse After the Nobel Prize?

Alexandre G. de Brevern

Université Paris Cité and Université de la Réunion, INSERM, BIGR, DSIMB Bioinformatics Team,
F-75015 Paris, France; alexandre.debrevern@univ-paris-diderot.fr; Tel.: +33-1-4449-3000

AlphaFold (AF) was the first deep learning tool to achieve exceptional fame in the field of biology [1]. To sum up, we first recall the existence of the CASP (Critical Assessment of Structural Prediction) competition, which allows the evaluation of individual prediction methods by proposing protein structural models. In 2018, the first version of the AF obtained excellent results, close to those of the best approaches available at the time [2,3]. Two years later, in 2020, a particularly significant average improvement was observed [4,5], and then with the communicative power of a company spun off from Alphabet, a great increase in media coverage of structural bioinformatics occurred.

CONCLUSIONS

AlphaFold2

- It seems, but it is not so easy to do a good structural model.
- Link with experiments can be very complicated
- Analysis of initial structural data is essential
- Good knowledge of appropriate tools is important
- It takes a lot of time, needs to be properly *think*.

THANK YOU

Snoopy’s question

Is really everything perfect ?

Question & Design

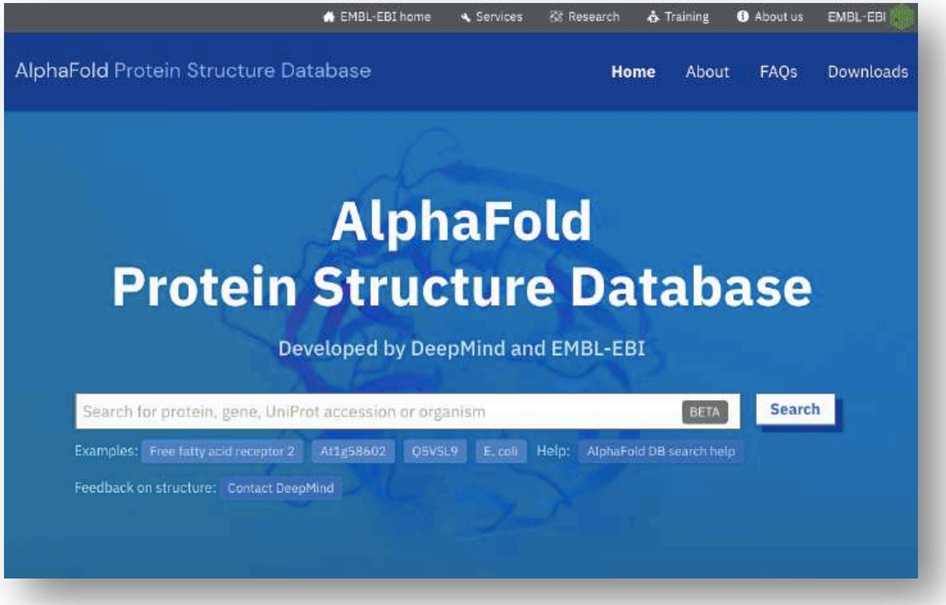
Question: Can we evaluate at a local protein level the general quality of AlphaFold2?

Question & Design

Question: Can we evaluate at a local protein level the general quality of AlphaFold2?

Design:

Dataset: AlphaFold2 human proteome structural model provided by EBI.

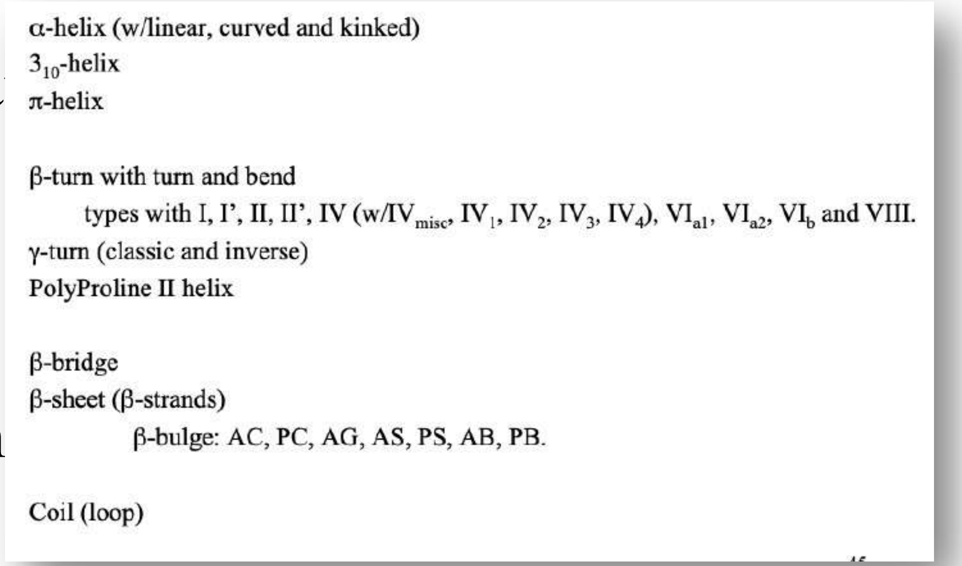


AlphaFold
Protein Structure Database

Developed by DeepMind and EMBL-EBI

Question & Design

Question: Can we evalua
quality of AlphaFold2?



α-helix (w/linear, curved and kinked)
3_{10}-helix
π-helix

β-turn with turn and bend
types with I, I', II, II', IV (w/IV_{misc}, IV_1, IV_2, IV_3, IV_4), VI_{a1}, VI_{a2}, VI_b and VIII.
γ-turn (classic and inverse)
PolyProline II helix

β-bridge
β-sheet (β-strands)
β-bulge: AC, PC, AG, AS, PS, AB, PB.

Coil (loop)

Design:

Dataset: AlphaFold2 hum
provided by EBI.

Methods: Assignment of local protein conformations
(DSSP, ProMotif, SEGNO, HELANAL, Protein Blocks)

Question & Design

Question: Can we evaluate at a local protein level the general quality of AlphaFold2?

Design:

Dataset: AlphaFold2 human proteome structural model provided by EBI.

Methods: Assignment of local protein conformations

(DSSP, ProMotif, SEGNO, HELANAL, Protein Blocks)

pLDDT (confidence index)

Question & Design

Question: Can we evaluate at a local protein level the general quality of AlphaFold2?

Design:

Dataset: AlphaFold2 human proteome structural model provided by EBI.

Methods: Assignment of local protein conformations

(DSSP, ProMotif, SEGNO, HELANAL, Protein Blocks)

pLDDT (confidence index)

Z-score to analyse over- and under-representation

(two PDB non-redundant structural datasets were also used for comparison)

Results

- AF2 human proteome: 23.511 structural models (representing 98.5% of human proteome)

Results

- AF2 human proteome: 23.511 structural models (representing 98.5% of human proteome)

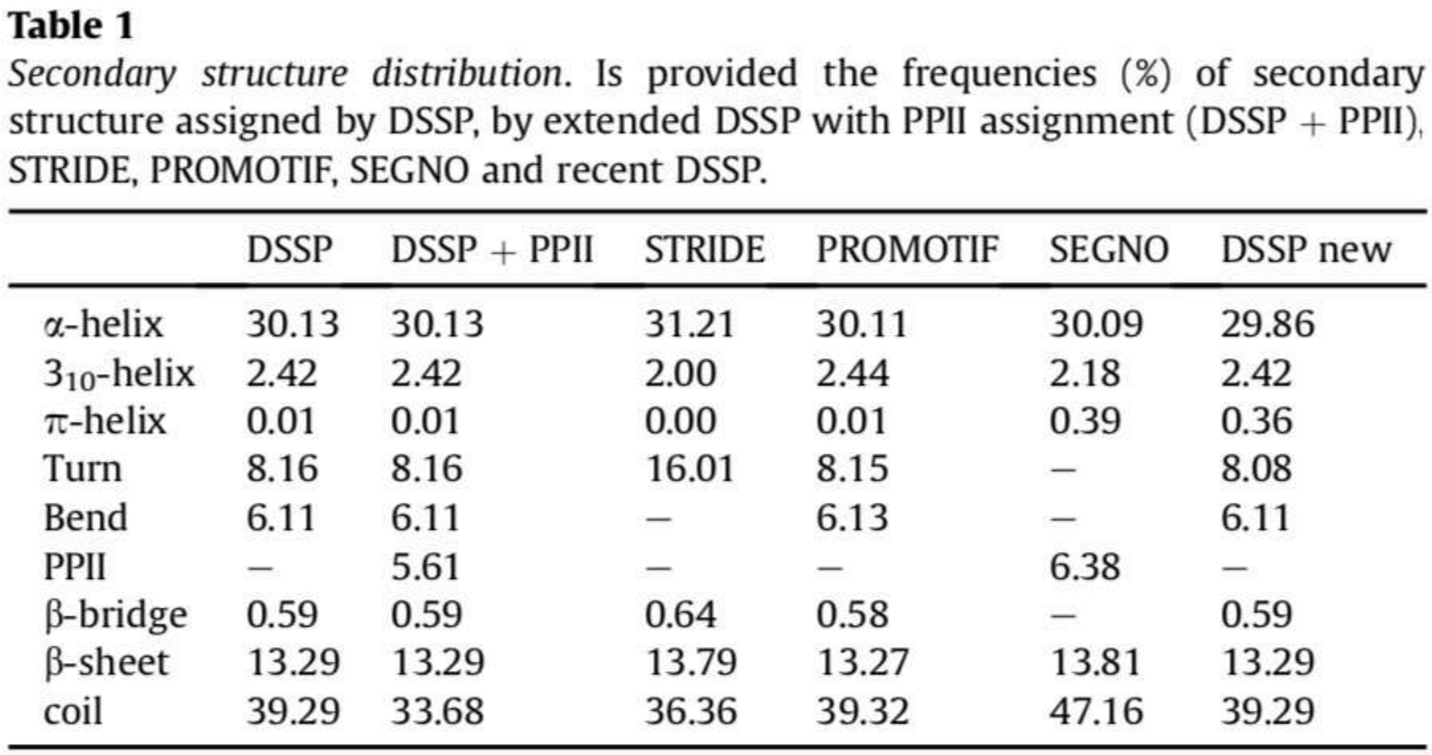


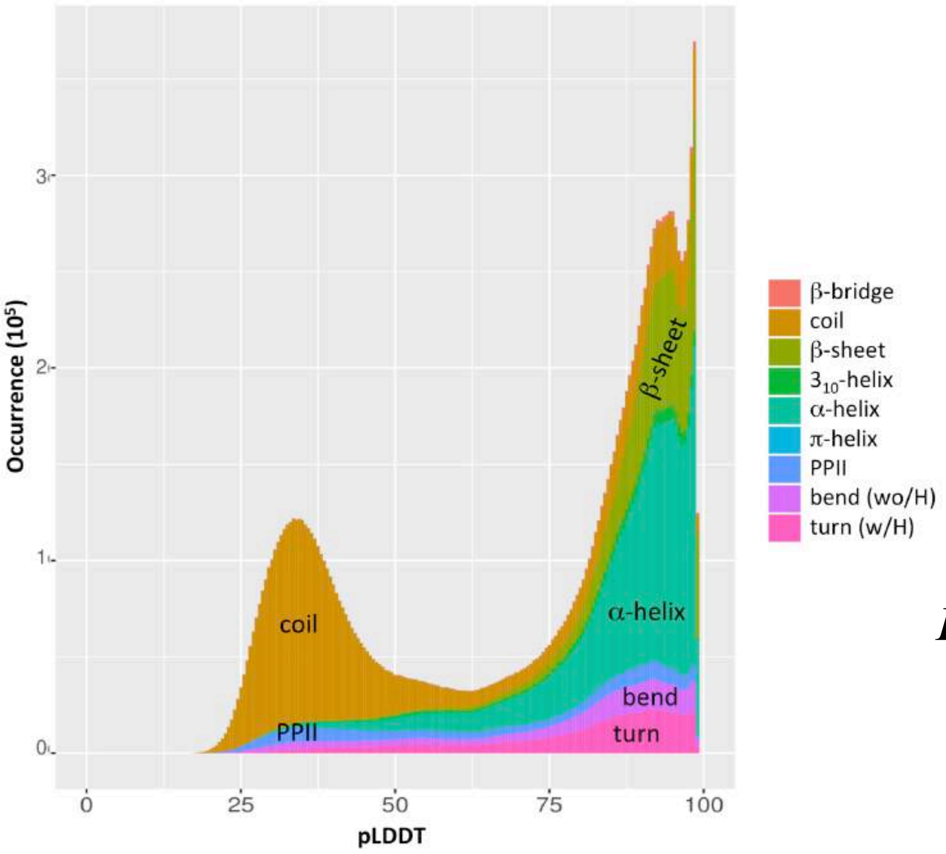
Table 1

Secondary structure distribution. Is provided the frequencies (%) of secondary structure assigned by DSSP, by extended DSSP with PPII assignment (DSSP + PPII), STRIDE, PROMOTIF, SEGNO and recent DSSP.

	DSSP	DSSP + PPII	STRIDE	PROMOTIF	SEGNO	DSSP new
α-helix	30.13	30.13	31.21	30.11	30.09	29.86
3_{10}-helix	2.42	2.42	2.00	2.44	2.18	2.42
π-helix	0.01	0.01	0.00	0.01	0.39	0.36
Turn	8.16	8.16	16.01	8.15	–	8.08
Bend	6.11	6.11	–	6.13	–	6.11
PPII	–	5.61	–	–	6.38	–
β-bridge	0.59	0.59	0.64	0.58	–	0.59
β-sheet	13.29	13.29	13.79	13.27	13.81	13.29
coil	39.29	33.68	36.36	39.32	47.16	39.29

Results

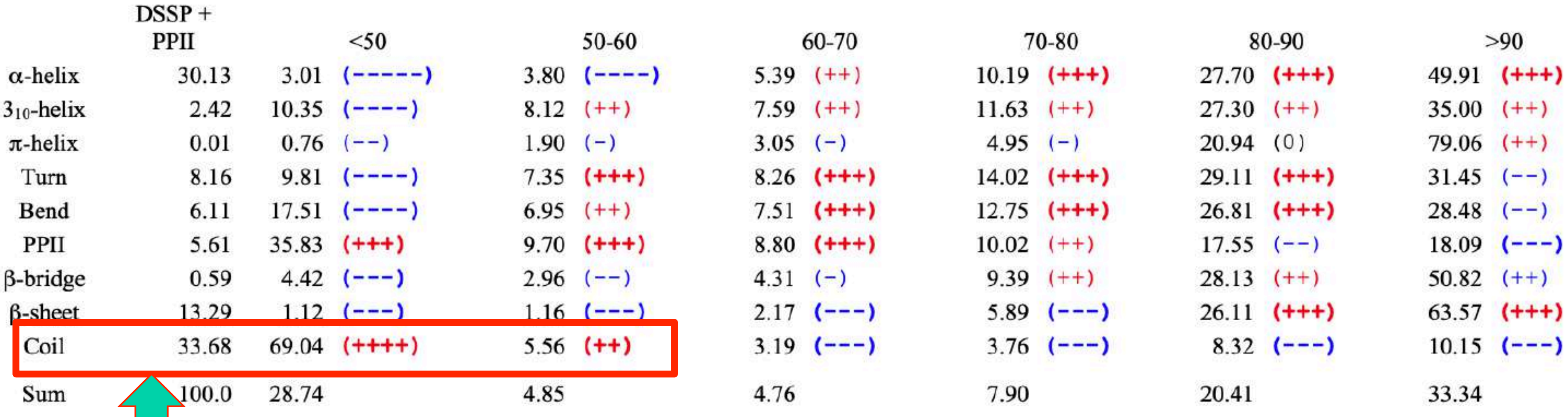
- AF2 human proteome: 23.511 structural models (representing 98.5% of human proteome)



DSSP (8-states) +PPII

Results

- AF2 human proteome: 23.511 structural models (representing 98.5% of human proteome)

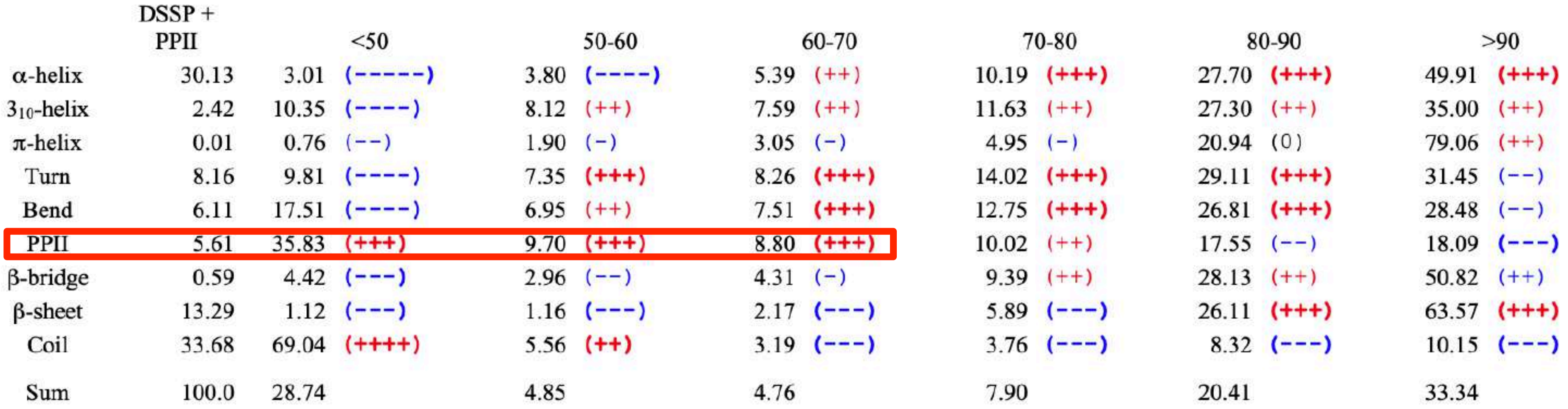


	DSSP + PPII	<50		50-60		60-70		70-80		80-90		>90	
α-helix	30.13	3.01	(-----)	3.80	(----)	5.39	(++)	10.19	(+++)	27.70	(+++)	49.91	(+++)
3_{10}-helix	2.42	10.35	(----)	8.12	(++)	7.59	(++)	11.63	(++)	27.30	(++)	35.00	(++)
π-helix	0.01	0.76	(--)	1.90	(-)	3.05	(-)	4.95	(-)	20.94	(0)	79.06	(++)
Turn	8.16	9.81	(----)	7.35	(+++)	8.26	(+++)	14.02	(+++)	29.11	(+++)	31.45	(--)
Bend	6.11	17.51	(----)	6.95	(++)	7.51	(+++)	12.75	(+++)	26.81	(+++)	28.48	(--)
PPII	5.61	35.83	(+++)	9.70	(+++)	8.80	(+++)	10.02	(++)	17.55	(--)	18.09	(---)
β-bridge	0.59	4.42	(---)	2.96	(--)	4.31	(-)	9.39	(++)	28.13	(++)	50.82	(++)
β-sheet	13.29	1.12	(---)	1.16	(---)	2.17	(---)	5.89	(---)	26.11	(+++)	63.57	(+++)
Coil	33.68	69.04	(++++)	5.56	(++)	3.19	(---)	3.76	(---)	8.32	(---)	10.15	(---)
Sum	100.0	28.74		4.85		4.76		7.90		20.41		33.34	

It is expected (IDRs, ..)

Results

- AF2 human proteome: 23.511 structural models (representing 98.5% of human proteome)

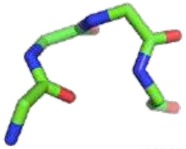


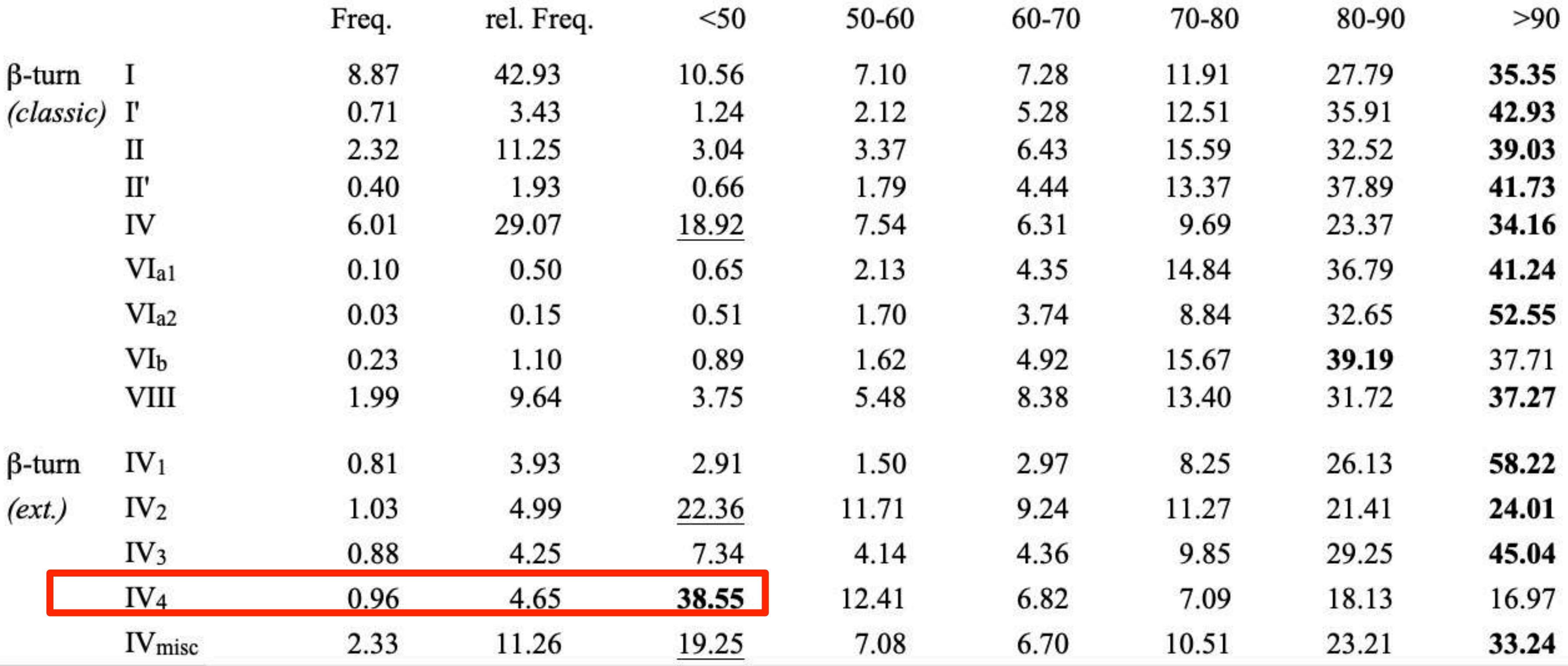
	DSSP + PPII	<50		50-60		60-70		70-80		80-90		>90	
α-helix	30.13	3.01	(-----)	3.80	(----)	5.39	(++)	10.19	(+++)	27.70	(+++)	49.91	(+++)
3_{10}-helix	2.42	10.35	(----)	8.12	(++)	7.59	(++)	11.63	(++)	27.30	(++)	35.00	(++)
π-helix	0.01	0.76	(--)	1.90	(-)	3.05	(-)	4.95	(-)	20.94	(0)	79.06	(++)
Turn	8.16	9.81	(----)	7.35	(+++)	8.26	(+++)	14.02	(+++)	29.11	(+++)	31.45	(--)
Bend	6.11	17.51	(----)	6.95	(++)	7.51	(+++)	12.75	(+++)	26.81	(+++)	28.48	(--)
PPII	5.61	35.83	(+++)	9.70	(+++)	8.80	(+++)	10.02	(++)	17.55	(--)	18.09	(---)
β-bridge	0.59	4.42	(---)	2.96	(--)	4.31	(-)	9.39	(++)	28.13	(++)	50.82	(++)
β-sheet	13.29	1.12	(---)	1.16	(---)	2.17	(---)	5.89	(---)	26.11	(+++)	63.57	(+++)
Coil	33.68	69.04	(++++)	5.56	(++)	3.19	(---)	3.76	(---)	8.32	(---)	10.15	(---)
Sum	100.0	28.74		4.85		4.76		7.90		20.41		33.34	

- PolyProline II helices are found often associated with low confidence index.

Results

β-turns:





		Freq.	rel. Freq.	<50	50-60	60-70	70-80	80-90	>90
β-turn	I	8.87	42.93	10.56	7.10	7.28	11.91	27.79	**35.35**
(classic)	I'	0.71	3.43	1.24	2.12	5.28	12.51	35.91	**42.93**
	II	2.32	11.25	3.04	3.37	6.43	15.59	32.52	**39.03**
	II'	0.40	1.93	0.66	1.79	4.44	13.37	37.89	**41.73**
	IV	6.01	29.07	18.92	7.54	6.31	9.69	23.37	**34.16**
	VI$_{a1}$	0.10	0.50	0.65	2.13	4.35	14.84	36.79	**41.24**
	VI$_{a2}$	0.03	0.15	0.51	1.70	3.74	8.84	32.65	**52.55**
	VI$_{b}$	0.23	1.10	0.89	1.62	4.92	15.67	**39.19**	37.71
	VIII	1.99	9.64	3.75	5.48	8.38	13.40	31.72	**37.27**
β-turn	IV$_1$	0.81	3.93	2.91	1.50	2.97	8.25	26.13	**58.22**
(ext.)	IV$_2$	1.03	4.99	22.36	11.71	9.24	11.27	21.41	**24.01**
	IV$_3$	0.88	4.25	7.34	4.14	4.36	9.85	29.25	**45.04**
	IV$_4$	0.96	4.65	**38.55**	12.41	6.82	7.09	18.13	16.97
	IV$_{misc}$	2.33	11.26	19.25	7.08	6.70	10.51	23.21	**33.24**

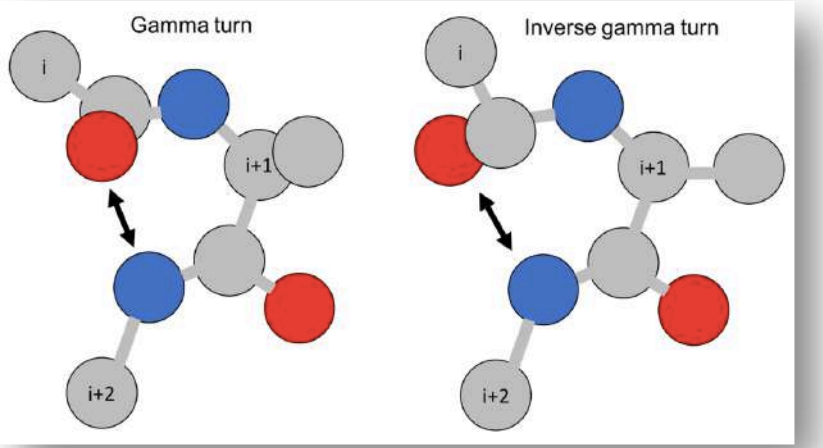
- A small issue with β-turn type IV$_4$ (frequency 0.96% of β-turns), near all maximum frequency are with pLDDT > 90.

Results

γ-turns:

		Freq.	rel. Freq.	<50	50-60	60-70	70-80	80-90	>90
γ-turn	*classic*	0.09	1.43	17.66	8.31	8.34	11.39	23.73	**29.18**
	Inverse	6.20	98.57	**54.98**	14.28	8.07	5.45	7.91	9.28

- A big issue with inverse γ-turn (frequency 98.6% of γ-turns), with 55% with pLDDT < 50.

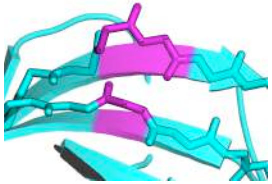


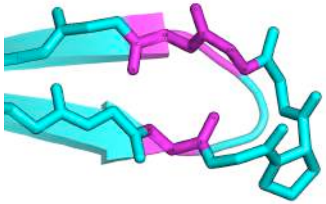
Results

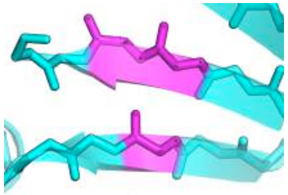
β-bulges:

		Freq.	rel. Freq.	<50	50-60	60-70	70-80	80-90	>90
β-bulge	AG1	0.97	41.66	2.54	1.26	3.77	11.33	34.20	**46.91**
	AC	1.08	46.22	2.33	0.74	1.81	5.08	26.31	**63.73**
	PC	0.04	1.56	1.79	0.53	1.05	2.86	13.44	**80.33**
	AW	0.13	5.45	2.46	0.89	2.23	8.90	36.45	**49.06**
	PW	0.01	0.30	3.70	3.51	2.53	4.19	16.96	**69.10**
	AB	0.01	0.37	4.01	1.49	2.36	12.74	32.47	**46.93**
	PB	0.02	0.67	20.73	8.59	6.54	9.53	19.27	**35.30**
	AS	0.08	3.46	1.73	1.08	2.69	7.79	28.46	**58.24**
	PS	0.01	0.31	0.73	0.18	0.55	2.29	9.72	**85.88**

- No systematic problem for β-bulge.





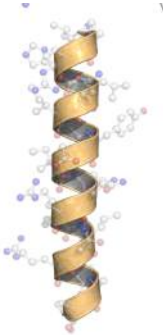


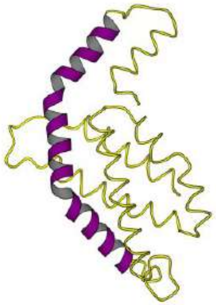
Results

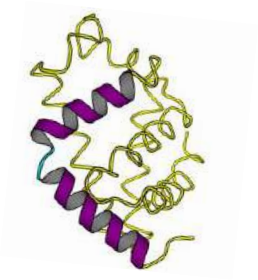
helix geometry:

		Freq.	rel. Freq.	<50	50-60	60-70	70-80	80-90	>90
Helix	*linear*	1.54	8.96	2.72	3.21	4.41	8.29	24.86	**56.49**
	curved	11.01	64.00	2.81	3.72	4.97	8.88	24.75	**54.85**
	kinked	4.65	27.04	2.52	3.72	5.82	11.66	30.28	**45.99**

- No systematic problem for helix geometry.





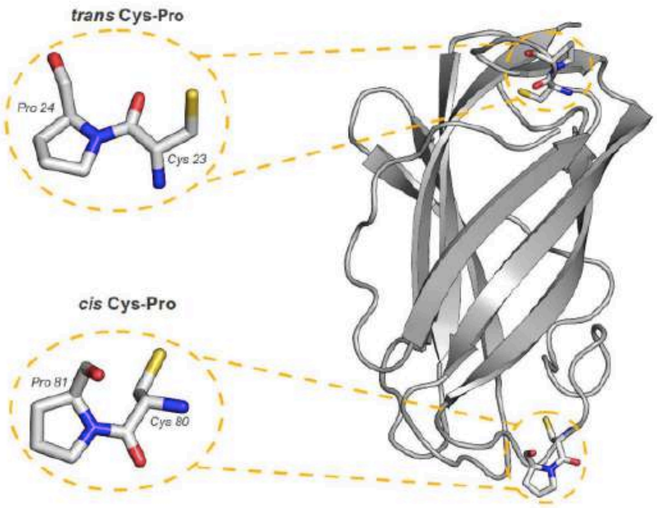


Results

Omega angles:

		Freq.	rel. Freq.	<50	50-60	60-70	70-80	80-90	>90
cis ω	All residue	4.75	--	**94.81**	2.22	0.40	0.58	1.00	0.98
cis ω	Proline	0.24	3.80	**86.62**	8.91	1.20	0.86	1.28	1.10

- A systematic problem for cis ω angle (0°) for Proline and every type of residues.



Craveur P., Joseph A.P., Poulain P., Rebehmed J., de Brevern A.G. Cis-trans isomerization of omega dihedrals in Proteins. *Amino Acids* (2013) **45**(2):279-89.

Results

Human proteome analysed by DSSP (+PPII) and the other approaches.

- PolyProline II helices are found often associated with low confidence index.

- Some less classical local protein conformations are found with low confidence index, i.e. γ-turns and cis ω angles.

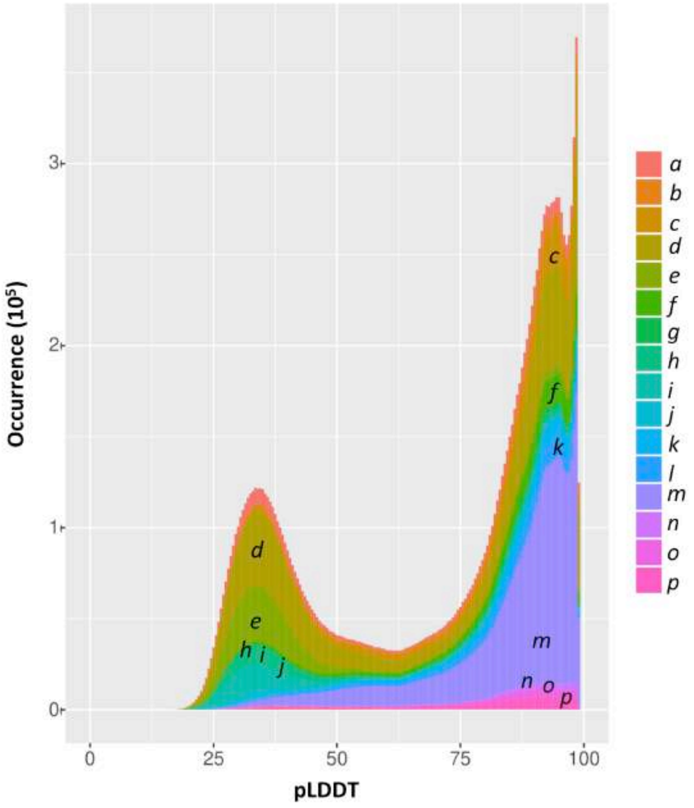
  55% of inverse γ-turns have pLDDT <50

  39% of type IV_4 β-turns have pLDDT <50

  94% of cis ω angles have pLDDT <50

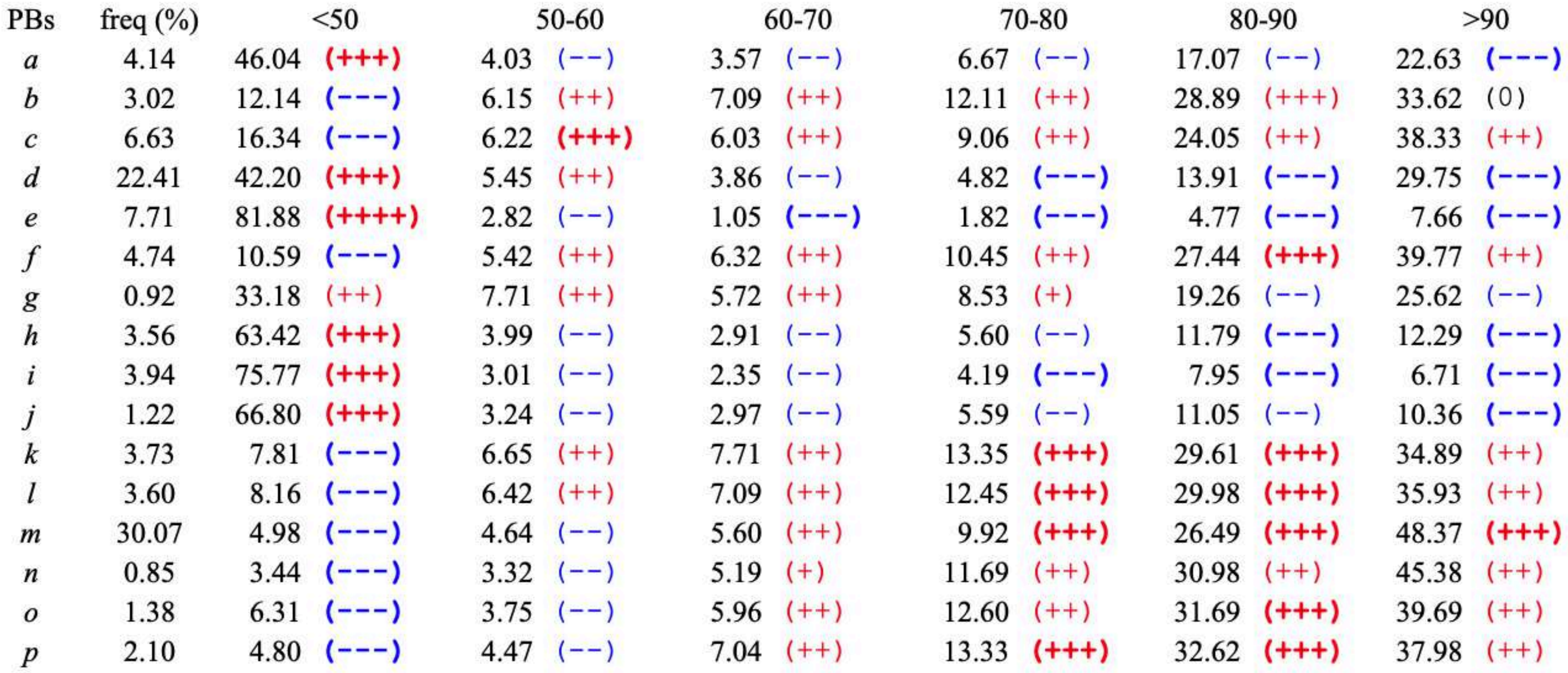
Additional results

- Analysis was also done with Protein Blocks (a series of 16 small local protein conformations of 5 residues, de Brevern et al, *Proteins*, 2000).



Additional results

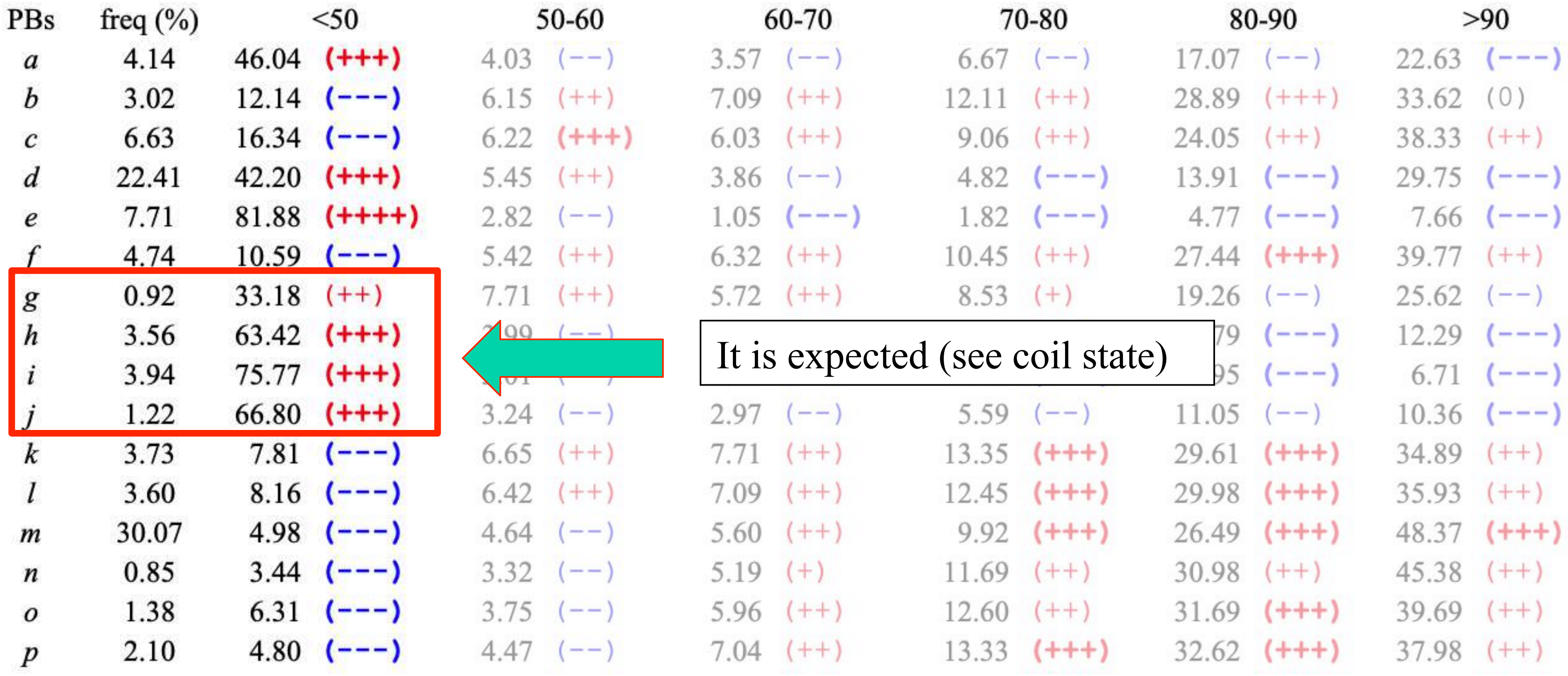
- Analysis was also done with Protein Blocks (a series of 16 small local protein conformations of 5 residues, de Brevern et al, *Proteins*, 2000).



PBs	freq (%)	<50		50-60		60-70		70-80		80-90		>90	
a	4.14	46.04	(+++)	4.03	(−−)	3.57	(−−)	6.67	(−−)	17.07	(−−)	22.63	(−−−)
b	3.02	12.14	(−−−)	6.15	(++)	7.09	(++)	12.11	(++)	28.89	(+++)	33.62	(0)
c	6.63	16.34	(−−−)	6.22	(+++)	6.03	(++)	9.06	(++)	24.05	(++)	38.33	(++)
d	22.41	42.20	(+++)	5.45	(++)	3.86	(−−)	4.82	(−−−)	13.91	(−−−)	29.75	(−−−)
e	7.71	81.88	(++++)	2.82	(−−)	1.05	(−−−)	1.82	(−−−)	4.77	(−−−)	7.66	(−−−)
f	4.74	10.59	(−−−)	5.42	(++)	6.32	(++)	10.45	(++)	27.44	(+++)	39.77	(++)
g	0.92	33.18	(++)	7.71	(++)	5.72	(++)	8.53	(+)	19.26	(−−)	25.62	(−−)
h	3.56	63.42	(+++)	3.99	(−−)	2.91	(−−)	5.60	(−−)	11.79	(−−−)	12.29	(−−−)
i	3.94	75.77	(+++)	3.01	(−−)	2.35	(−−)	4.19	(−−−)	7.95	(−−−)	6.71	(−−−)
j	1.22	66.80	(+++)	3.24	(−−)	2.97	(−−)	5.59	(−−)	11.05	(−−)	10.36	(−−−)
k	3.73	7.81	(−−−)	6.65	(++)	7.71	(++)	13.35	(+++)	29.61	(+++)	34.89	(++)
l	3.60	8.16	(−−−)	6.42	(++)	7.09	(++)	12.45	(+++)	29.98	(+++)	35.93	(++)
m	30.07	4.98	(−−−)	4.64	(−−)	5.60	(++)	9.92	(+++)	26.49	(+++)	48.37	(+++)
n	0.85	3.44	(−−−)	3.32	(−−)	5.19	(+)	11.69	(++)	30.98	(++)	45.38	(++)
o	1.38	6.31	(−−−)	3.75	(−−)	5.96	(++)	12.60	(++)	31.69	(+++)	39.69	(++)
p	2.10	4.80	(−−−)	4.47	(−−)	7.04	(++)	13.33	(+++)	32.62	(+++)	37.98	(++)

Additional results

- Analysis was also done with Protein Blocks (a series of 16 small local protein conformations of 5 residues, de Brevern et al, *Proteins*, 2000).

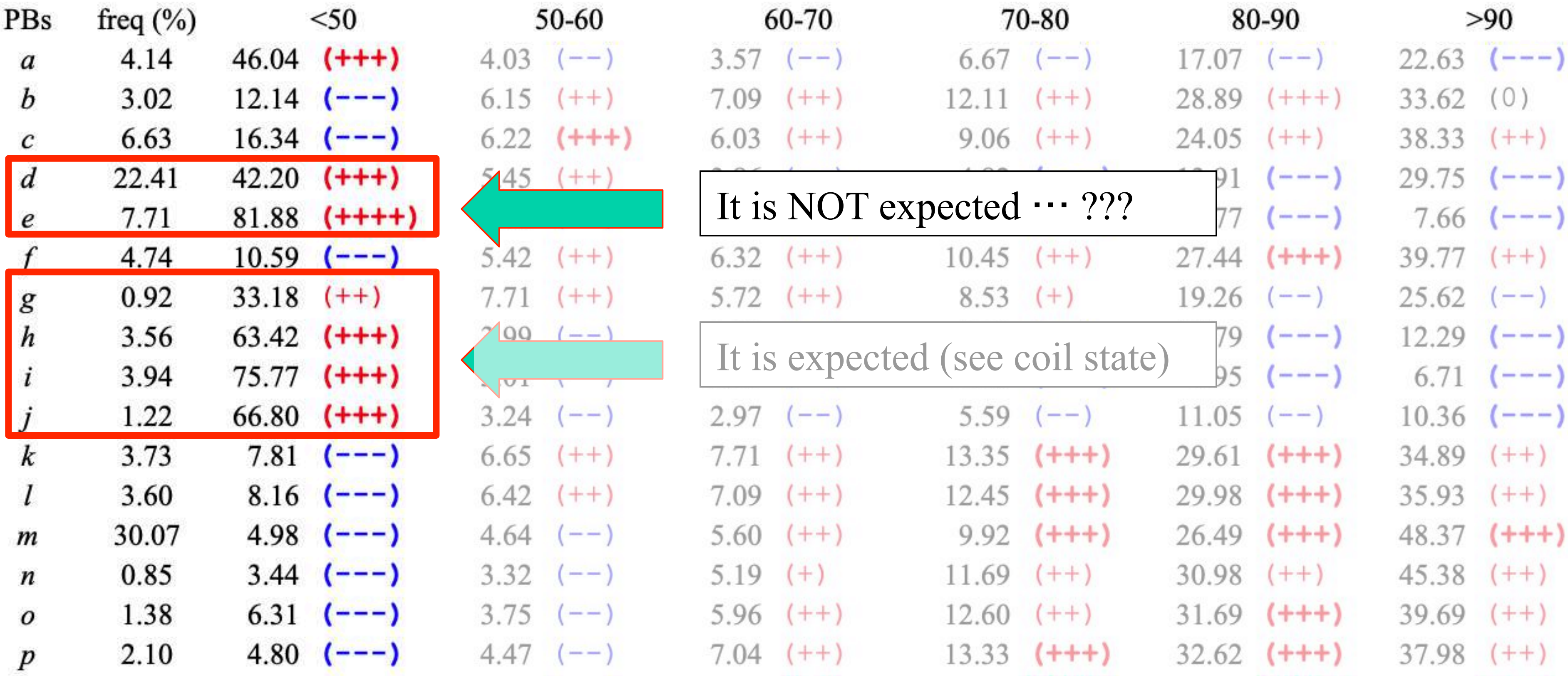


PBs	freq (%)	<50	50-60	60-70	70-80	80-90	>90
a	4.14	46.04 (+++)	4.03 (--)	3.57 (--)	6.67 (--)	17.07 (--)	22.63 (---)
b	3.02	12.14 (---)	6.15 (++)	7.09 (++)	12.11 (++)	28.89 (+++)	33.62 (0)
c	6.63	16.34 (---)	6.22 (+++)	6.03 (++)	9.06 (++)	24.05 (++)	38.33 (++)
d	22.41	42.20 (+++)	5.45 (++)	3.86 (--)	4.82 (---)	13.91 (---)	29.75 (---)
e	7.71	81.88 (++++)	2.82 (--)	1.05 (---)	1.82 (---)	4.77 (---)	7.66 (---)
f	4.74	10.59 (---)	5.42 (++)	6.32 (++)	10.45 (++)	27.44 (+++)	39.77 (++)
g	0.92	33.18 (++)	7.71 (++)	5.72 (++)	8.53 (+)	19.26 (--)	25.62 (--)
h	3.56	63.42 (+++)	[illegible]99 (--)	[illegible]	[illegible]	[illegible]79 (---)	12.29 (---)
i	3.94	75.77 (+++)	[illegible]	[illegible]	[illegible]	[illegible]95 (---)	6.71 (---)
j	1.22	66.80 (+++)	3.24 (--)	2.97 (--)	5.59 (--)	11.05 (--)	10.36 (---)
k	3.73	7.81 (---)	6.65 (++)	7.71 (++)	13.35 (+++)	29.61 (+++)	34.89 (++)
l	3.60	8.16 (---)	6.42 (++)	7.09 (++)	12.45 (+++)	29.98 (+++)	35.93 (++)
m	30.07	4.98 (---)	4.64 (--)	5.60 (++)	9.92 (+++)	26.49 (+++)	48.37 (+++)
n	0.85	3.44 (---)	3.32 (--)	5.19 (+)	11.69 (++)	30.98 (++)	45.38 (++)
o	1.38	6.31 (---)	3.75 (--)	5.96 (++)	12.60 (++)	31.69 (+++)	39.69 (++)
p	2.10	4.80 (---)	4.47 (--)	7.04 (++)	13.33 (+++)	32.62 (+++)	37.98 (++)

It is expected (see coil state)

Additional results

- Analysis was also done with Protein Blocks (a series of 16 small local protein conformations of 5 residues, de Brevern et al, *Proteins*, 2000).



PBs	freq (%)	<50		50-60		60-70		70-80		80-90		>90	
a	4.14	46.04	(+++)	4.03	(--)	3.57	(--)	6.67	(--)	17.07	(--)	22.63	(---)
b	3.02	12.14	(---)	6.15	(++)	7.09	(++)	12.11	(++)	28.89	(+++)	33.62	(0)
c	6.63	16.34	(---)	6.22	(+++)	6.03	(++)	9.06	(++)	24.05	(++)	38.33	(++)
d	22.41	42.20	(+++)	5.45	(++)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	(---)	29.75	(---)
e	7.71	81.88	(++++)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	(---)	7.66	(---)
f	4.74	10.59	(---)	5.42	(++)	6.32	(++)	10.45	(++)	27.44	(+++)	39.77	(++)
g	0.92	33.18	(++)	7.71	(++)	5.72	(++)	8.53	(+)	19.26	(--)	25.62	(--)
h	3.56	63.42	(+++)	[illegible]	(--)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	(---)	12.29	(---)
i	3.94	75.77	(+++)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	(---)	6.71	(---)
j	1.22	66.80	(+++)	3.24	(--)	2.97	(--)	5.59	(--)	11.05	(--)	10.36	(---)
k	3.73	7.81	(---)	6.65	(++)	7.71	(++)	13.35	(+++)	29.61	(+++)	34.89	(++)
l	3.60	8.16	(---)	6.42	(++)	7.09	(++)	12.45	(+++)	29.98	(+++)	35.93	(++)
m	30.07	4.98	(---)	4.64	(--)	5.60	(++)	9.92	(+++)	26.49	(+++)	48.37	(+++)
n	0.85	3.44	(---)	3.32	(--)	5.19	(+)	11.69	(++)	30.98	(++)	45.38	(++)
o	1.38	6.31	(---)	3.75	(--)	5.96	(++)	12.60	(++)	31.69	(+++)	39.69	(++)
p	2.10	4.80	(---)	4.47	(--)	7.04	(++)	13.33	(+++)	32.62	(+++)	37.98	(++)

It is NOT expected ··· ???

It is expected (see coil state)

Additional results

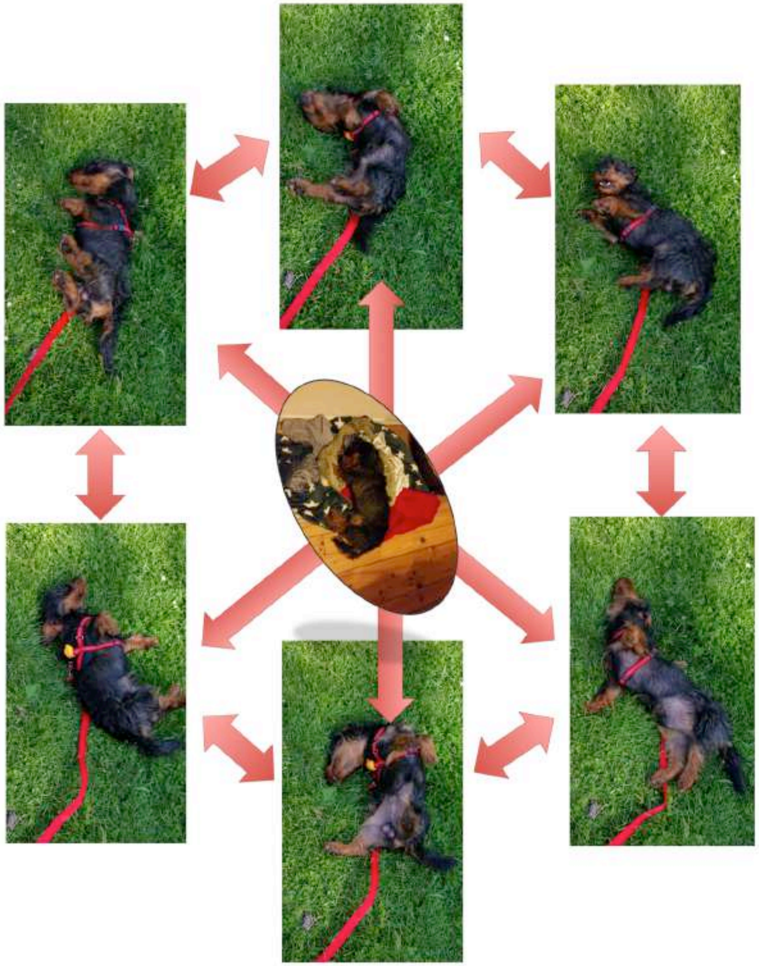
- Analysis was also done with Protein Blocks (a series of 16 small local protein conformations of 5 residues, de Brevern et al, *Proteins*, 2000).

- Over-representation in low confidence region of Protein Blocks *a*, *d* and *e* (geometrically N-cap, central and C-cap part of a β-strand).

- However, the frequency of β-sheets is lower than expected in this dataset.

- Wouldn't we have unfinished β-sheets but with well-prepared β-strands (the prediction of β-sheets is always the most difficult).

THANK YOU



A dachshund-analogy to illustrate the analysis of protein dynamics at the light of protein local backbone conformation taken from Narwani et al, *JBSD*, 2020.