ColabFold v1.5.5: AlphaFold2 using MMseqs2

Easy to use protein structure and complex prediction using AlphaFold2 and Alphafold2-multimer. Sequence alignments/templates are generated through MMseqs2 and HHsearch. For more details, see bottom of the notebook, checkout the ColabFold GitHub and read our manuscript. Old versions: v1.4, v1.5.1, v1.5.2, v1.5.3-patch

Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. ColabFold: Making protein folding accessible to all. *Nature Methods*, 2022

Input protein sequence(s), then hit `Runtime -> Run all`

query_sequence: " LEEKKVCQGTSNKLTQLGTFEDHFLSLQRMFNNCEVVLGNLEITYVQ "

- Use `:` to specify inter-protein chainbreaks for **modeling complexes** (supports homo- and hetro-oligomers). For example **PI...SK:PI...SK** for a homodimer

jobname: " test "

num_relax: 0

- specify how many of the top ranked structures to relax using amber

template_mode: none

- `none` = no template information is used. `pdb100` = detect templates in pdb100 (see notes). `custom` - upload and search own templates (PDB or mmCIF format, see notes)

Show code

```
jobname test_479e5
sequence LEEKKVCQGTSNKLTQLGTFEDHFLSLQRMFNNCEVVLGNLEITYVQRNYDLSFLKTIQEVAGYVLIA
length 771
```

Install dependencies

Show code

```
installing colabfold...
CPU times: user 156 ms, sys: 24.6 ms, total: 180 ms
Wall time: 44.3 s
```

MSA options (custom MSA upload, single sequence, pairing mode)

msa_mode: mmseqs2_uniref_env

pair_mode: unpaired_paired

- "unpaired_paired" = pair sequences from same species + unpaired MSA, "unpaired" = seperate MSA for each chain, "paired" - only use paired sequences.

Show code

Advanced settings

model_type: auto

- if `auto` selected, will use `alphafold2_ptm` for monomer prediction and `alphafold2_multimer_v3` for complex prediction. Any of the mode_types can be used (regardless if input is monomer or complex).

num_recycles: 3

- if `auto` selected, will use `num_recycles=20` if `model_type=alphafold2_multimer_v3`, else `num_recycles=3` .

recycle_early_stop_tolerance: auto

- if `auto` selected, will use `tol=0.5` if `model_type=alphafold2_multimer_v3` else `tol=0.0`.

relax_max_iterations: 200

- max amber relax iterations, `0` = unlimited (AlphaFold2 default, can take very long)

pairing_strategy: greedy

- `greedy` = pair any taxonomically matching subsets, `complete` = all sequences have to match in one line.

Sample settings

- enable dropouts and increase number of seeds to sample predictions from uncertainty of the model.
- decrease `max_msa` to increase uncertainity

max_msa: auto

num_seeds: 1

use_dropout: ☐

Save settings

save_all: ☑

save_recycles: ☐

save_to_google_drive: ☑

- if the save_to_google_drive option was selected, the result zip will be uploaded to your Google Drive

dpi: 200

- set dpi for image resolution

Don't forget to hit `Runtime -> Run all` after updating the form.

Show code

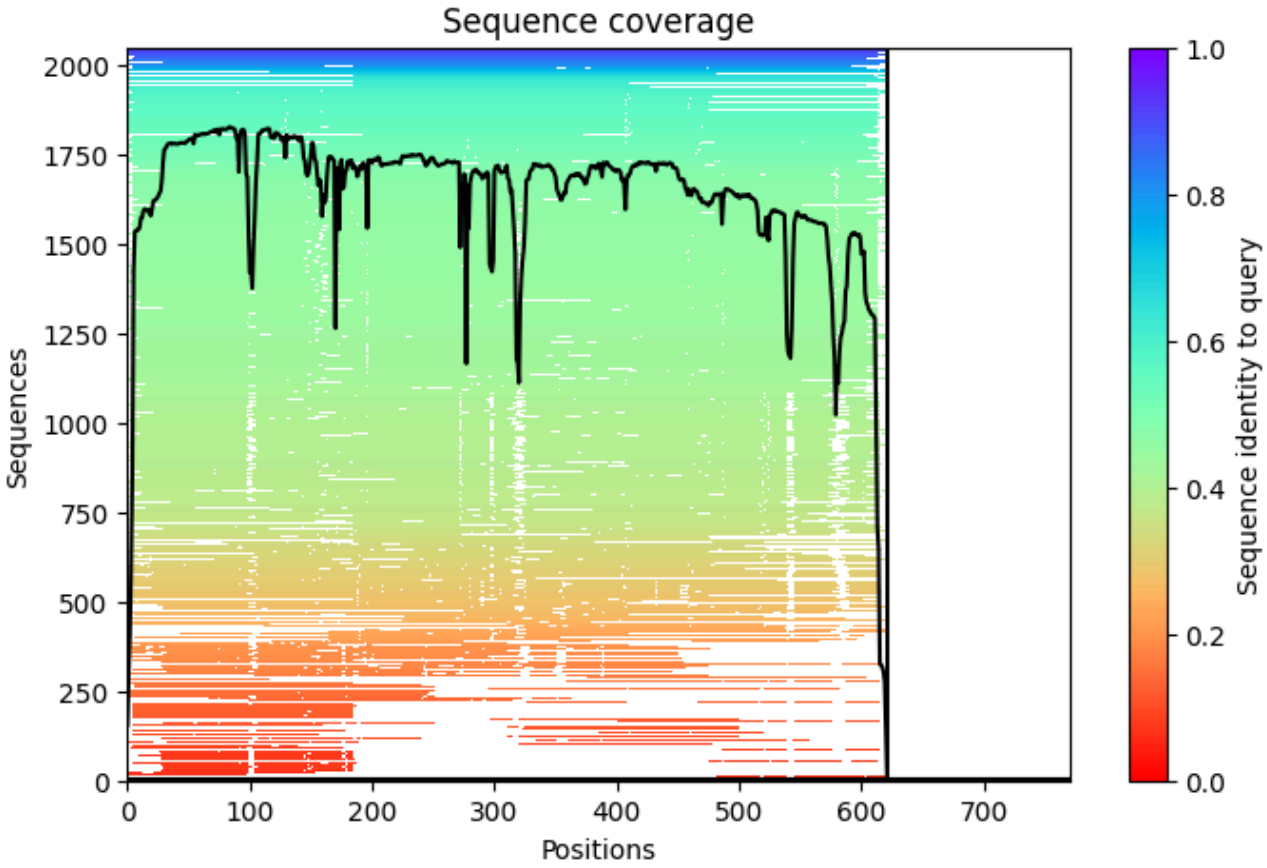
```
You are logged into Google Drive and are good to go!
```

Run Prediction

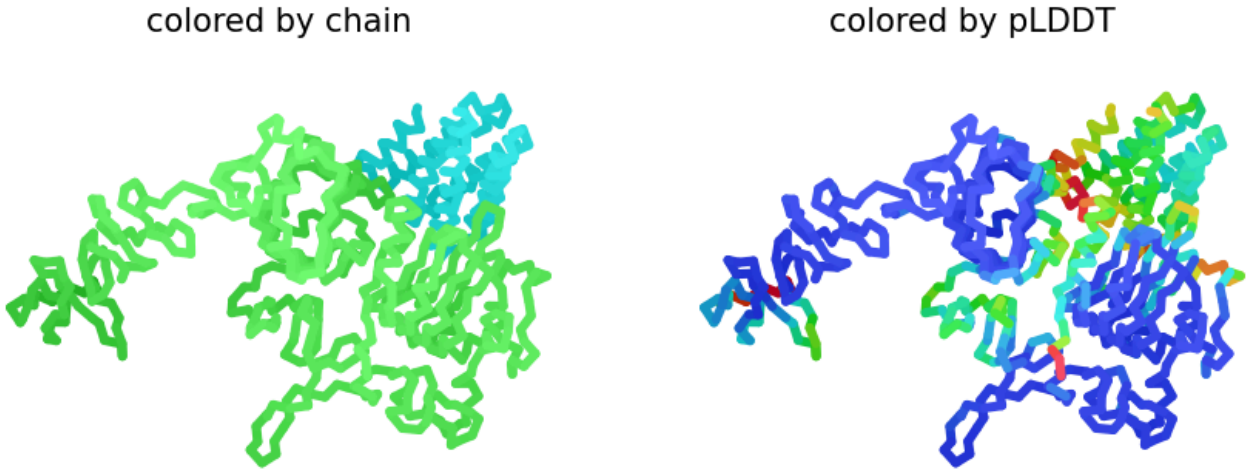
display_images: ☑

Show code

```
Downloading alphafold2_multimer_v3 weights to .: 100%|██████████| 3.82G/3.82G
2024-08-15 11:42:54,555 Running on GPU
2024-08-15 11:42:54,886 Found 5 citations for tools or databases
2024-08-15 11:42:54,887 Query 1/1: test_479e5 (length 771)
COMPLETE: 100%|██████████| 300/300 [elapsed: 00:01 remaining: 00:00]
COMPLETE: 100%|██████████| 300/300 [elapsed: 00:00 remaining: 00:00]
```



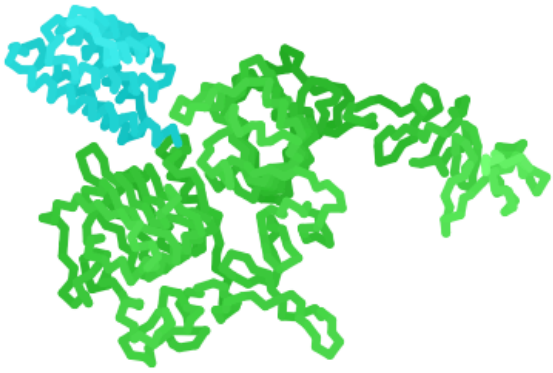

```
2024-08-15 11:42:58,722 Setting max_seq=508, max_extra_seq=1542
2024-08-15 11:46:06,129 alphafold2_multimer_v3_model_1_seed_000 recycle=0 pLD
2024-08-15 11:48:35,592 alphafold2_multimer_v3_model_1_seed_000 recycle=1 pLD
2024-08-15 11:51:04,513 alphafold2_multimer_v3_model_1_seed_000 recycle=2 pLD
2024-08-15 11:53:33,608 alphafold2_multimer_v3_model_1_seed_000 recycle=3 pLD
2024-08-15 11:53:33,610 alphafold2_multimer_v3_model_1_seed_000 took 628.4s (
```



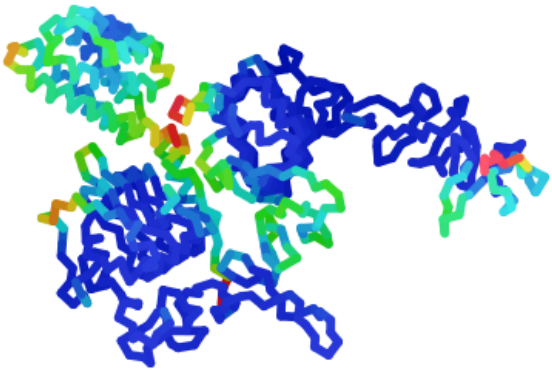

```
2024-08-15 11:56:07,662 alphafold2_multimer_v3_model_2_seed_000 recycle=0 pLD
2024-08-15 11:58:36,612 alphafold2_multimer_v3_model_2_seed_000 recycle=1 pLD
2024-08-15 12:01:05,494 alphafold2_multimer_v3_model_2_seed_000 recycle=2 pLD
2024-08-15 12:03:34,440 alphafold2_multimer_v3_model_2_seed_000 recycle=3 pLD
```

```
2024-08-15 12:03:34,441 alphafold2_multimer_v3_model_2_seed_000 took 595.6s (
```

colored by chain

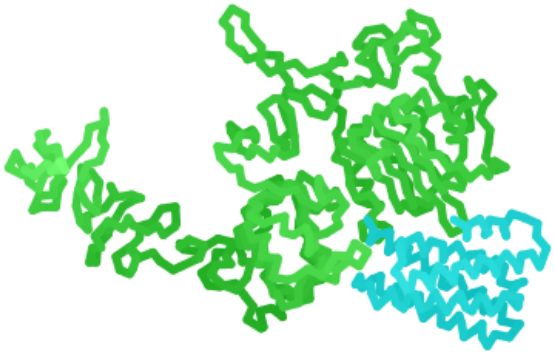


colored by pLDDT

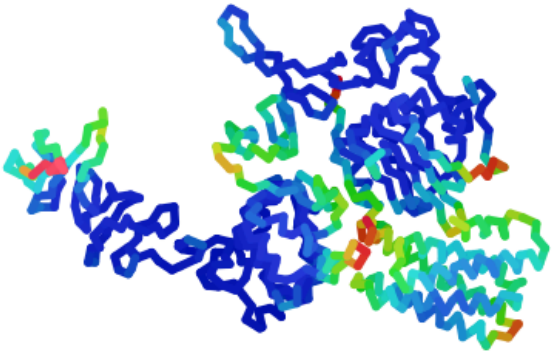


```
2024-08-15 12:06:08,350 alphafold2_multimer_v3_model_3_seed_000 recycle=0 pLD
2024-08-15 12:08:37,393 alphafold2_multimer_v3_model_3_seed_000 recycle=1 pLD
2024-08-15 12:11:06,255 alphafold2_multimer_v3_model_3_seed_000 recycle=2 pLD
2024-08-15 12:13:35,311 alphafold2_multimer_v3_model_3_seed_000 recycle=3 pLD
2024-08-15 12:13:35,313 alphafold2_multimer_v3_model_3_seed_000 took 595.6s (
```

colored by chain

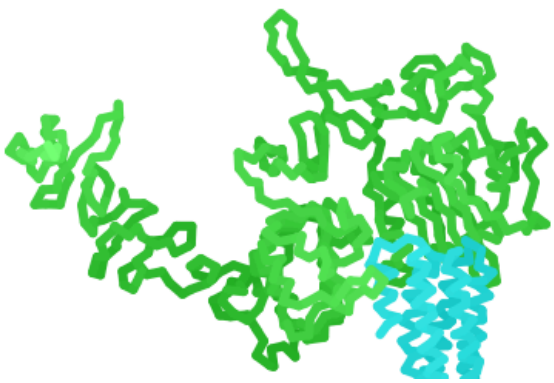


colored by pLDDT

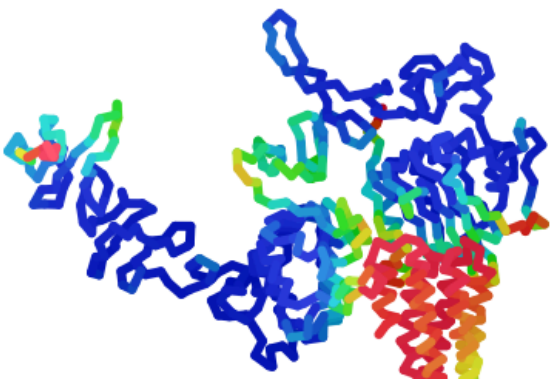


```
2024-08-15 12:16:06,116 alphafold2_multimer_v3_model_4_seed_000 recycle=0 pLD
2024-08-15 12:18:35,790 alphafold2_multimer_v3_model_4_seed_000 recycle=1 pLD
2024-08-15 12:21:04,573 alphafold2_multimer_v3_model_4_seed_000 recycle=2 pLD
2024-08-15 12:23:33,373 alphafold2_multimer_v3_model_4_seed_000 recycle=3 pLD
2024-08-15 12:23:33,374 alphafold2_multimer_v3_model_4_seed_000 took 595.9s (
```

colored by chain

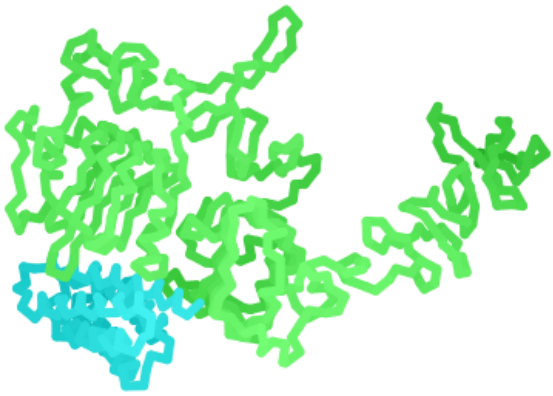


colored by pLDDT

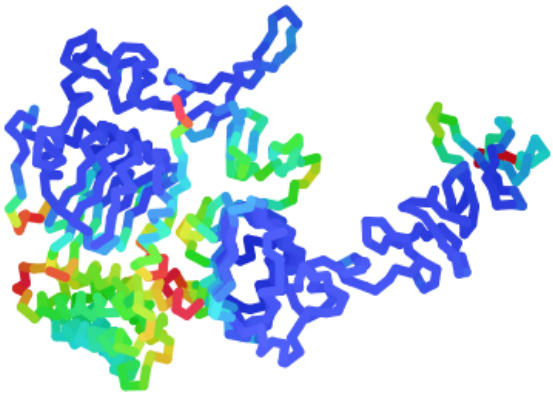


```
2024-08-15 12:26:06,209 alphafold2_multimer_v3_model_5_seed_000 recycle=0 pLD
2024-08-15 12:28:35,202 alphafold2_multimer_v3_model_5_seed_000 recycle=1 pLD
2024-08-15 12:31:04,097 alphafold2_multimer_v3_model_5_seed_000 recycle=2 pLD
2024-08-15 12:33:33,724 alphafold2_multimer_v3_model_5_seed_000 recycle=3 pLD
2024-08-15 12:33:33,726 alphafold2_multimer_v3_model_5_seed_000 took 596.2s (
```

colored by chain



colored by pLDDT



```
2024-08-15 12:33:37,873 reranking models by 'multimer' metric
2024-08-15 12:33:37,874 rank_001_alphafold2_multimer_v3_model_4_seed_000 pLDD
2024-08-15 12:33:37,875 rank_002_alphafold2_multimer_v3_model_3_seed_000 pLDD
2024-08-15 12:33:37,876 rank_003_alphafold2_multimer_v3_model_5_seed_000 pLDD
2024-08-15 12:33:37,876 rank_004_alphafold2_multimer_v3_model_1_seed_000 pLDD
2024-08-15 12:33:37,877 rank_005_alphafold2_multimer_v3_model_2_seed_000 pLDD
2024-08-15 12:33:40,384 Done
0
```

Display 3D structure

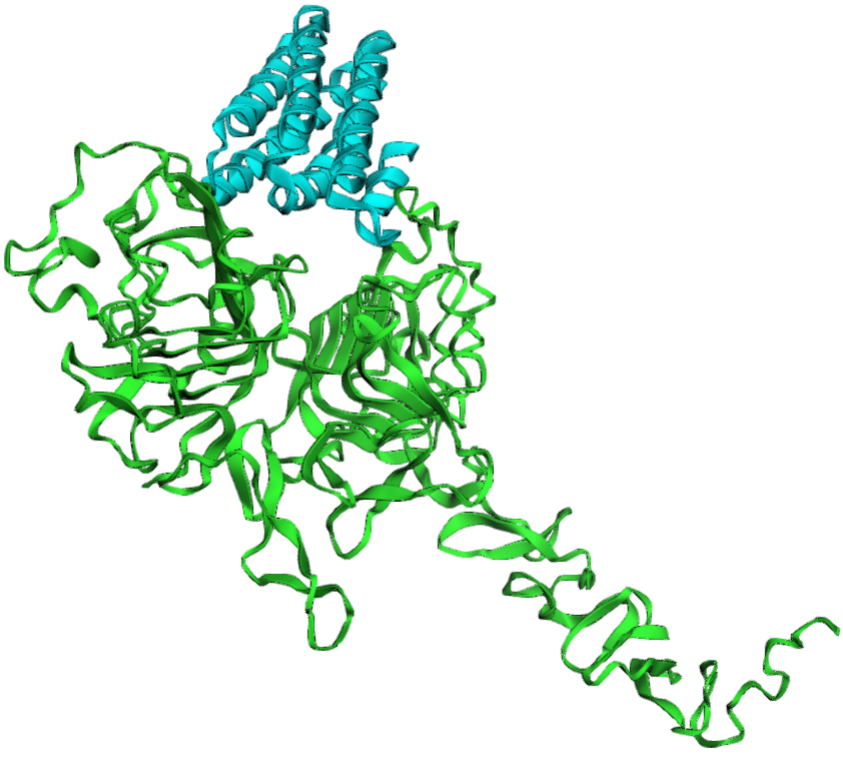
rank_num: 1

color: chain

show_sidechains:

show_mainchains:

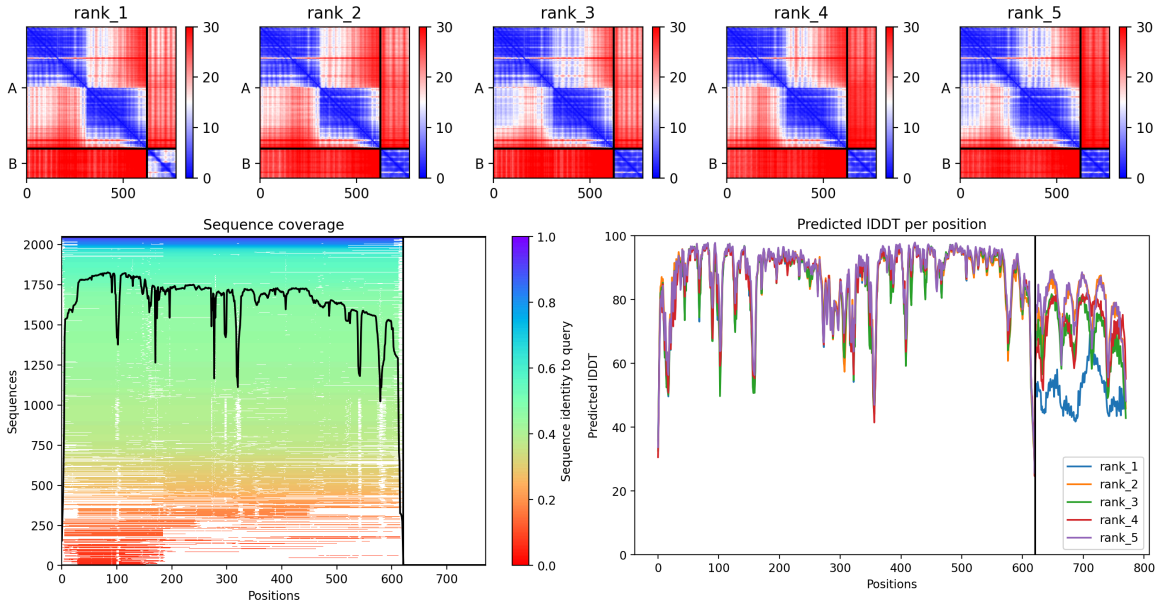
Show code



Plots

Show code

Plots for test_479e5



Package and download results

If you are having issues downloading the result archive, try disabling your adblocker and run this cell again. If that fails click on the little folder icon to the left, navigate to file: `jobname.result.zip`, right-click and select "Download" (see screenshot).

Show code

```
Uploaded test_479e5.result.zip to Google Drive with ID 1JixtcdsXKJeHfJiNJcv_X
Downloading "test_479e5.result.zip":
```

Instructions

Quick start

1. Paste your protein sequence(s) in the input field.
2. Press "Runtime" -> "Run all".

3. The pipeline consists of 5 steps. The currently running step is indicated by a circle with a stop sign next to it.

Result zip file contents

1. PDB formatted structures sorted by avg. pLDDT and complexes are sorted by pTMscore. (unrelaxed and relaxed if `use_amber` is enabled).
2. Plots of the model quality.
3. Plots of the MSA coverage.
4. Parameter log file.
5. A3M formatted input MSA.
6. A `predicted_aligned_error_v1.json` using AlphaFold-DB's format and a `scores.json` for each model which contains an array (list of lists) for PAE, a list with the average pLDDT and the pTMscore.
7. BibTeX file with citations for all used tools and databases.

At the end of the job a download modal box will pop up with a `jobname.result.zip` file. Additionally, if the `save_to_google_drive` option was selected, the `jobname.result.zip` will be uploaded to your Google Drive.

MSA generation for complexes

For the complex prediction we use unpaired and paired MSAs. Unpaired MSA is generated the same way as for the protein structures prediction by searching the UniRef100 and environmental sequences three iterations each.

The paired MSA is generated by searching the UniRef100 database and pairing the best hits sharing the same NCBI taxonomic identifier (=species or sub-species). We only pair sequences if all of the query sequences are present for the respective taxonomic identifier.

Using a custom MSA as input

To predict the structure with a custom MSA (A3M formatted): (1) Change the `msa_mode`: to "custom", (2) Wait for an upload box to appear at the end of the "MSA options ..." box. Upload your A3M. The first fasta entry of the A3M must be the query sequence without gaps.

It is also possilbe to proide custom MSAs for complex predictions. Read more about the format here.

As an alternative for MSA generation the HHblits Toolkit server can be used. After submitting your query, click "Query Template MSA" -> "Download Full A3M". Download the A3M file and upload it in this notebook.

PDB100

As of 23/06/08, we have transitioned from using the PDB70 to a 100% clustered PDB, the PDB100. The construction methodology of PDB100 differs from that of PDB70.