Selecting notable CASP_Commons/COVID-19 models based on QA (EMA) results

dr. Kliment Olechnovič

Life Sciences Center, Vilnius University

2020-12-04

What we wanted:

▶ to highlight what models were consistently favoured by the QA (EMA) methods

What we did:

▶ calculated consensus similarity scores for models that were ranked highly by QA methods.

Idea

all models,
global consensus
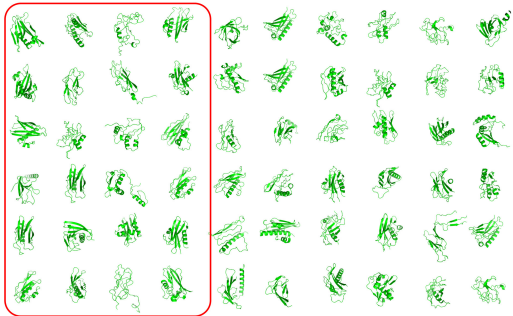

single QA-based ranking,
top N consensus

multiple QA-based rankings,
combined top N consensus

(that is what we did
for various N)

Below is an algorithm we used for every target:

► Calculate similarities (CAD-score or lDDT) of every model to others.
► Remove duplicated models based on the CAD-score threshold of 0.8.
► For every N from 1 to 50 do:
    1. Select top N models from every available QA ranking.
    2. Bundle all the selected models together.
        ► If a model was selected by more than one QA method, it is included multiple times: this way "popular" models gain more weight.
    3. Calculate a consensus similarity value (average CAD-score or lDDT of all pairwise comparisons) for every model inside the bundle.
► Print results as a table, highlight models with highest consensus scores for every N.

Used QA predictions

The set of QA methods was pruned to remove repeating variations of methods, the number of QA methods was reduced from 30 to the following 23:

QA035	angleQA	QA037	Elofsson
QA041	VoroMQA-dark	QA081	SBROD
QA092	QMEANDisCo	QA110	Seder2020
QA114	Ornate	QA131	BAKER
QA179	DAVIS-QAconsensus	QA182	Bhattacharya
QA204	ModFOLD8_rank	QA210	Kiharalab
QA227	MESHI	QA236	Seder2020hard
QA273	Takeda-Shitaka-Lab	QA274	ProQ3D
QA307	qa3dcnn_prof	QA363	GCMQA
QA373	Yang-Server	QA389	MULTICOM-CONSTRUCT
QA438	Destini	QA452	VoroCNN
QA490	bioinsilico_sbi		

The following choices:

- ▶ use a non-redundant or full set of QA methods.
- ▶ use CAD-score or lDDT
- ▶ remove or keep duplicated models

had very small effect on the final results.

Output: full tables

All result tables: https://bit.ly/CASP-COVID-QA.

Top part of a full table for target C1901:

The table is sorted by 'cs_top1'. Highest column values are highlighted. Table can be resorted by clicking on a column header.

model	cs_top1	cs_top2	cs_top3	cs_top4	cs_top5	cs_top6	cs_top7	cs_top8	cs_top9	cs_top10	cs_top15	cs_top20	cs_top25	cs_top30	cs_top35	cs_top40	cs_top45	cs_top50
C1901TS156_1	0.720	0.705	0.680	0.650	0.605	0.576	0.551	0.536	0.522	0.512	0.478	0.458	0.443	0.435	0.429	0.424	0.418	0.412
C1901TS156_2	0.659	0.683	0.674	0.659	0.612	0.584	0.559	0.546	0.531	0.521	0.485	0.466	0.451	0.443	0.437	0.432	0.426	0.420
C1901TS156_5	0.641	0.684	0.674	0.647	0.599	0.570	0.545	0.531	0.516	0.507	0.473	0.453	0.438	0.430	0.424	0.419	0.413	0.407
C1901TS401_1	0.578	0.582	0.581	0.577	0.535	0.513	0.490	0.477	0.464	0.455	0.423	0.407	0.393	0.386	0.381	0.377	0.372	0.366
C1901TS044_1	0.402	0.398	0.395	0.395	0.401	0.405	0.409	0.409	0.415	0.420	0.425	0.432	0.428	0.430	0.423	0.416	0.410	0.403
C1901TS215_1	0.342	0.359	0.354	0.361	0.371	0.377	0.382	0.385	0.385	0.388	0.388	0.380	0.373	0.373	0.371	0.365	0.360	0.355
C1901TS210_3	0.324	0.327	0.326	0.325	0.326	0.326	0.327	0.328	0.332	0.332	0.338	0.341	0.341	0.340	0.339	0.337	0.335	0.336
C1901TS299_3	0.316	0.318	0.319	0.318	0.319	0.319	0.319	0.319	0.319	0.319	0.319	0.319	0.320	0.320	0.320	0.321	0.321	0.321
C1901TS213_2	0	0.392	0.392	0.393	0.400	0.405	0.408	0.408	0.412	0.419	0.425	0.435	0.429	0.429	0.422	0.415	0.408	0.402
C1901TS414_1	0	0.374	0.375	0.374	0.378	0.380	0.382	0.385	0.386	0.388	0.389	0.389	0.391	0.397	0.399	0.398	0.397	0.396
C1901TS210_1	0	0.368	0.368	0.366	0.374	0.373	0.372	0.373	0.372	0.372	0.374	0.376	0.376	0.379	0.380	0.378	0.375	0.373
C1901TS210_4	0	0.320	0.324	0.322	0.328	0.332	0.332	0.335	0.333	0.332	0.338	0.337	0.334	0.333	0.330	0.329	0.329	0.328
C1901TS131_3	0	0	0.448	0.447	0.457	0.466	0.473	0.487	0.500	0.499	0.491	0.481	0.473	0.467	0.463	0.458	0.452	0.447
C1901TS213_1	0	0	0.395	0.395	0.403	0.407	0.409	0.410	0.414	0.420	0.426	0.434	0.429	0.429	0.423	0.415	0.409	0.403
C1901TS152_5	0	0	0.378	0.375	0.382	0.382	0.381	0.382	0.381	0.382	0.385	0.384	0.385	0.386	0.389	0.390	0.387	0.383
C1901TS215_2	0	0	0.349	0.351	0.359	0.369	0.376	0.389	0.394	0.397	0.390	0.384	0.380	0.378	0.378	0.373	0.367	0.362
C1901TS299_2	0	0	0.343	0.341	0.341	0.341	0.341	0.341	0.340	0.341	0.340	0.339	0.340	0.342	0.343	0.343	0.343	0.343
C1901TS309_1	0	0	0.341	0.340	0.341	0.342	0.343	0.345	0.345	0.345	0.346	0.346	0.345	0.345	0.348	0.347	0.348	0.349
C1901TS405_5	0	0	0.333	0.330	0.331	0.331	0.336	0.336	0.335	0.333	0.346	0.344	0.340	0.338	0.335	0.333	0.334	0.334
C1901TS131_4	0	0	0	0.444	0.460	0.466	0.471	0.486	0.497	0.497	0.486	0.477	0.470	0.464	0.460	0.455	0.449	0.444
C1901TS369_1	0	0	0	0.390	0.396	0.402	0.407	0.407	0.406	0.410	0.408	0.410	0.415	0.415	0.417	0.414	0.411	0.407

All result tables: https://bit.ly/CASP-COVID-QA.

Some tables of "notable" models, i.e. models with the highest consensus score for at least a single N:

C1901

model	cs_top1	cs_top2	cs_top3	cs_top4	cs_top5	cs_top6	cs_top7	cs_top8	cs_top9	cs_top10	cs_top15	cs_top20	cs_top25	cs_top30	cs_top35	cs_top40	cs_top45	cs_top50
C1901TS156_1	0.720	0.705	0.680	0.650	0.605	0.576	0.551	0.536	0.522	0.512	0.478	0.458	0.443	0.435	0.429	0.424	0.418	0.412
C1901TS156_2	0.659	0.683	0.674	0.659	0.612	0.584	0.559	0.546	0.531	0.521	0.485	0.466	0.451	0.443	0.437	0.432	0.426	0.420
C1901TS131_1	0	0	0.448	0.447	0.457	0.466	0.473	0.487	0.500	0.499	0.491	0.481	0.473	0.467	0.463	0.458	0.452	0.447

C1902

model	cs_top1	cs_top2	cs_top3	cs_top4	cs_top5	cs_top6	cs_top7	cs_top8	cs_top9	cs_top10	cs_top15	cs_top20	cs_top25	cs_top30	cs_top35	cs_top40	cs_top45	cs_top50
C1902TS156_1	0.606	0.535	0.518	0.521	0.511	0.505	0.509	0.513	0.513	0.511	0.500	0.497	0.491	0.485	0.482	0.477	0.472	0.464
C1902TS438_2	0.513	0.498	0.500	0.504	0.513	0.517	0.518	0.523	0.529	0.527	0.530	0.530	0.526	0.524	0.522	0.517	0.511	0.502

C1903

model	cs_top1	cs_top2	cs_top3	cs_top4	cs_top5	cs_top6	cs_top7	cs_top8	cs_top9	cs_top10	cs_top15	cs_top20	cs_top25	cs_top30	cs_top35	cs_top40	cs_top45	cs_top50
C1903TS156_2	0.448	0.464	0.453	0.472	0.472	0.468	0.462	0.468	0.465	0.464	0.458	0.456	0.454	0.454	0.453	0.453	0.450	0.449
C1903TS369_5	0.431	0.452	0.453	0.453	0.446	0.444	0.443	0.442	0.440	0.438	0.439	0.437	0.441	0.445	0.447	0.448	0.447	0.447
C1903TS414_4	0	0	0	0	0	0	0	0	0	0	0.451	0.456	0.459	0.464	0.466	0.469	0.469	0.469
C1903TS414_5	0	0	0	0	0	0	0	0	0	0	0	0	0.461	0.467	0.470	0.471	0.471	0.472

C1904

model	cs_top1	cs_top2	cs_top3	cs_top4	cs_top5	cs_top6	cs_top7	cs_top8	cs_top9	cs_top10	cs_top15	cs_top20	cs_top25	cs_top30	cs_top35	cs_top40	cs_top45	cs_top50
C1904TS401_1	0.606	0.538	0.492	0.470	0.456	0.442	0.430	0.422	0.414	0.410	0.400	0.391	0.383	0.379	0.377	0.373	0.368	0.365
C1904TS156_1	0.529	0.547	0.540	0.541	0.540	0.529	0.516	0.512	0.504	0.500	0.492	0.486	0.479	0.475	0.472	0.470	0.465	0.461
C1904TS131_1	0	0	0	0.457	0.475	0.475	0.475	0.478	0.477	0.478	0.489	0.491	0.488	0.487	0.483	0.482	0.475	0.472

How to interpret consensus scores

To interpret the significance of consensus score values, take a look at the
histogram of all consensus CAD-score values. Basically, consensus scores
around 0.6 are relatively significant.
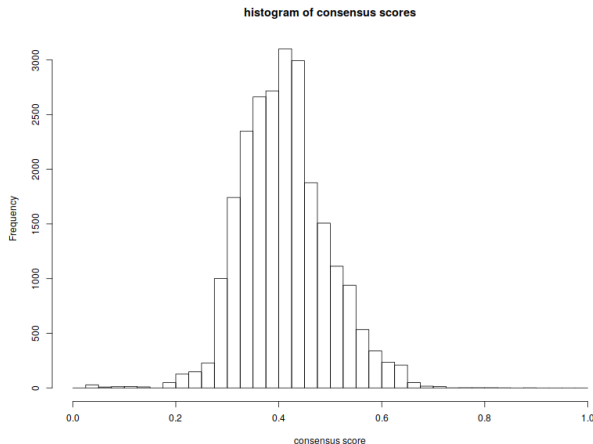
► Probably, a better strategy would be to just select AlphaFold models. Anyway, in many cases AlphaFold models were the best according to QA-based consensus.

► Membrane proteins and monomers out of context (e.g. out of obligatory complexes) are difficult for most of the QA methods to evaluate.

  ► A domain of C1905 was a part of an obligatory dimer in membrane, so QA-based consensus did not work well for C1905.

All result tables: https://bit.ly/CASP-COVID-QA.

All scripts and results are at
https://github.com/kliment-olechnovic/casp_commons_covid_analysis.

Thanks to
- Česlovas Venclovas
- Gabriel Studer
- Andriy Kryshtafovych
- all the predictors