ISSN 2303-4521
Periodicals of Engineering and Natural Sciences *Original Research*
Vol. 10, No. 1, February 2022, pp.481-490

Correlation of model quality between predicted proteins and their templates

Muhamed Adilović[1*], Altijana Hromić-Jahjefendić[1]
[1] Genetics and Bioengineering, International University of Sarajevo

ABSTRACT

Protein structure prediction is an important process that carries a lot of benefits for various areas of science and industry. Template modeling is the most reliable and most popular method, depending on the solved structures from the Protein Data Bank. An important part of it is template selection, using different methods, which is a challenging task that requires special attention because the proper selection of protein template can lead to a more accurate protein prediction. This study focuses on the relationships between predicted proteins, taken from the Swiss-model repository, and their templates, on a larger scale. Features of predicted proteins are taken into account, including protein length, sequence identity, and sequence coverage. Quality assessment scores are compared and analyzed between the predicted proteins and their templates. Overall, quality assessment scores of predicted proteins show a moderate positive correlation to the sequence identity with the templates. Moreover, based on our data, the level of template quality is noticeably correlated with the predicted protein structuers, because templates with higher quality scores will, on average, also allow for the modeling of predicted proteins with higher quality scores.

Keywords: protein structure prediction; protein quality assessment; template-prediction correlation; Protein Data Bank; Swiss-model

Corresponding Author:

Muhamed Adilović
Genetics and Bioengineering
International University of Sarajevo
Hrasnička Cesta 15, Ilidža
E-mail: madilovic@ius.edu.ba

1. Introduction

Proteins are one of the main components of living organisms and the main workers in cells [1]. Understanding their mechanisms of action is helpful for the understanding of many biological pathways and treatment of various diseases [1], [2]. Since protein function is closely related to its structure, focusing on structural biology is the key component in the overall study of proteins [3].

There are several methods of determining protein structure, but the two most important ones are the physical determination of structure through various methods, and prediction of protein structure through different algorithms. Physically, the structure is determined using different techniques, among which the most popular ones are X-ray crystallography, nuclear magnetic resonance (NMR), and cryogenic electron microscopy (cryo-EM) [4]. These types of structures are generally made publicly available through Protein Data Bank (PDB) – an online repository of solved protein structures [5].

When it comes to protein structure prediction, it also has different methods, with the main ones being template-based prediction and *ab initio*. The template-based uses uploaded protein structures available at the PDB, finds the most similar ones mostly by comparing the sequence of amino acids (although different tools have different methods), and builds/predicts the novel protein using the available template [6]. *Ab initio*, on the other hand, tries to independently predict the structure of a novel protein (although it can also use parts of the existing

templates) [7]. Template-based prediction is inherently more accurate, however, the accuracy of *ab initio* is improving, and there are also situations where it is more useful due to specific constraints, e.g. [8], [9].
In general, *in silico* study of protein structures is very popular since it can find potential targets for later research while using a fraction of the resources [10], which is why there has been a lot of development in computational biology and bioinformatics, especially in the area of structural biology, e.g. using machine learning (ML) in order to study various aspects of proteins including their structure [11]–[13], structural quality [14, p.], [15]–[17], or classification [18], [19].
Validation of protein's 3D structure is another important aspect, and there are different quality assessment (QA) tools developed for this purpose.
The issue is that, with the prediction of proteins' structure using template-based methods, structural properties of templates, including the quality, might be transferred onto the predicted structures. This is an important aspect to be considered since the selection of the best template for the prediction of the protein is a challenging task, with novel methods of template selection showing an improvement in the quality of the prediction of proteins [20], which indicates that there is a potential for further optimization of this process.
The aim of this research is to assess whether there is a "transfer of characteristics", namely the quality level, from the templates to predicted protein structures, and to what degree. In order to do this, the aim is also to assess the correlation of predicted proteins' QA levels with their structural features and, if there is a connection, to group them into corresponding subgroups for the proper study and analysis.

2. Materials and methods

This study contains two databases – a database of template structures collected from the PDB, and a database of predicted proteins from Swiss-model (S-M) [21]. Both databases have been cross-referenced and filtered so that the analyses in this study have been done only on those predicted proteins which contain templates from the first database, and vice versa – on the templates which have predicted proteins in the second database.

2.1. Sample collection and retrieval of the information

The first database contains template proteins from the PDB solved with X-ray crystallography. From the total of 35,710 present, 6656 are used at least once in the second database containing predicted proteins. Additional details for the first database as well as QA results and descriptive statistics are found in the previous study done only on the protein templates [22].
The second database containing the predicted protein structures has been taken from the S-M repository available online [21]. It contained more than 400,000 proteins during the collection process, which have been cross-referenced with the first database, so the final database containing predicted protein structures contains 49,000 proteins all of which have been predicted based on the templates that are in the first database.
The online repository also contains two important parameters: percentage of sequence used from the template, and percentage of similarity to the template sequence which have been taken for further usage and analysis. Additionally, proteins in the online repository are grouped according to the organism from which the sequences have been taken, which has also been adapted to the database.
The final list of protein features used in the second database is available in table 1. Moreover, all features from the first database (template database) can also be cross-referenced to the second database (database of predicted proteins).

Table 1. Primary Experimental Information Retrieved from the PDB [23]

Criteria	**Description**
Template	PDB protein used for the prediction/homology modeling
Organism	Organism from which the protein sequence originates
Residue Count	The total number of residues in the protein model
Sequence Identity	Percentage of similarity between the template and the protein
Sequence Coverage	Percentage of residues from the template used during modeling

2.2. Quality assessment

The list of different methods used for the Quality assessment is in Table 2. These are all the same QA tools used for the assessment of database containing template proteins from the PDB. The main difference is that predicted proteins do not have R value, since it is the measurement obtained experimentally during the determination of the protein structure.

Table 2. Quality assessment methods performed on proteins

Criteria	**Measurement**
Ramachandran	Percentage of outliers in an atom model [24]
Energy	Total energy of a model normalized for the residue count [25]
Verify3D	Percentage of residues above 0.2 threshold [26], [27]
PROCHECK	Percentage of satisfactory evaluations [28]
ERRAT	Percentage of nonrandom distribution of atoms [29]
PROVE	Percentage of buried outlier protein atoms [30]
QMEAN	Feature scale 0-1 [31]
DOPE	Real numbers [32]
VoroMQA	Real numbers [33]

2.3. Correlation between the first and the second database

The features from the second database (organism of origin, protein length, sequence identity, and sequence coverage) have been analyzed against the QA scores using Pearson correlation and based on them, predicted proteins were adequately divided and compared. The comparison is done by taking into account the QA scores of template proteins and predicted proteins.
The following correlation was made between the two databases: predicted proteins were first divided based on the sequence identity to the template into 20 distinct groups – 5 points in sequence identity is taken as a cut-off value, and since the identity goes from 0 to 100, the end result is 20 groups.
From each of the group, proteins were filtered based on the number of predicted proteins having the same template - if there are at least 30 proteins with the same template - the mean of their quality scores is taken into account, otherwise, the sample is deemed too small and is excluded.
Cross-comparison has then been performed with t-test - comparing the means of each of the protein samples among themselves, and analyzing the difference in the means of quality scores of predicted proteins with the difference in the quality scores of their templates, taking into consideration statistically significant differences.

3. Results

3.1. Descriptive statistics

The total number of proteins in the 2nd database is 48523. The total number of templates from the 1st database used in the prediction of the proteins from the 2nd database is 6656. This means that most of the PDB proteins are not used as a template, others are used more than once, and some are very commonly used as templates. E.g. 4UXV is the most common template – 1387 proteins are predicted from it. It is "Cytoplasmic domain of bacterial cell division protein EzrA". 5XG2 is the second most common template – 1007 proteins are predicted from it. It is "Crystal structure of a coiled-coil segment (residues 345-468 and 694-814) of Pyrococcus yayanosii Smc". Out of 6656 templates, 4392 (~66%) have been used for the prediction of only up to 3 proteins (inclusive).
Table 3 includes the summary of descriptive statistics for protein features.

Table 3. Descriptive statistics of protein features

	Length	Sequence Identity	Sequence Coverage
Mean	287.60	35.11	73.27
Std. Error	0.82	0.11	0.12
Median	248.00	27.89	80.66
Mode	127.00	100.00	100.00
Std. Dev.	179.54	24.68	26.32
Range	1534	97.31	179.85
Minimum	16	2.69	2.08
Maximum	1550	100.00	181.93

Figures 1-3 show a visual representation of the distribution of protein characteristics across the sample collected.

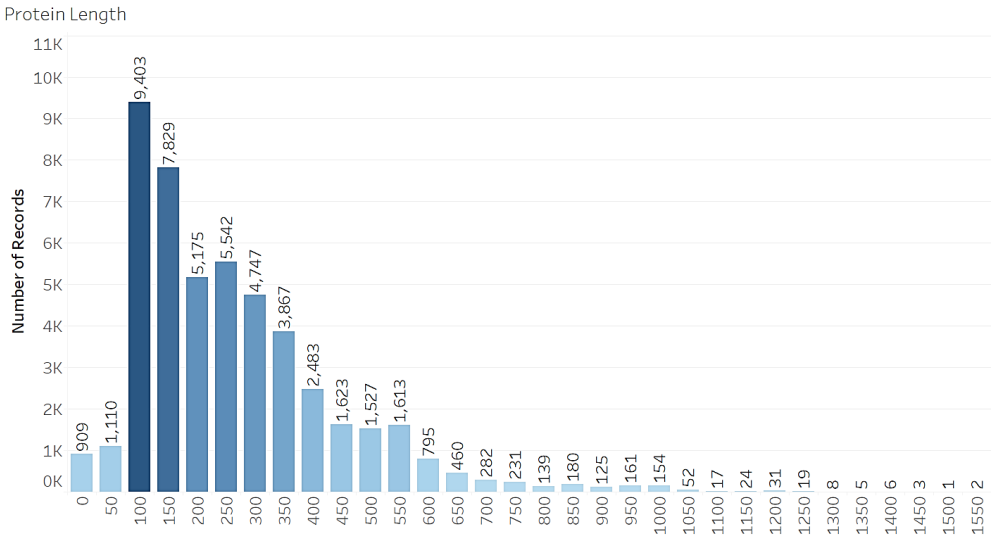

Figure 1. distribution of proteins according to their length

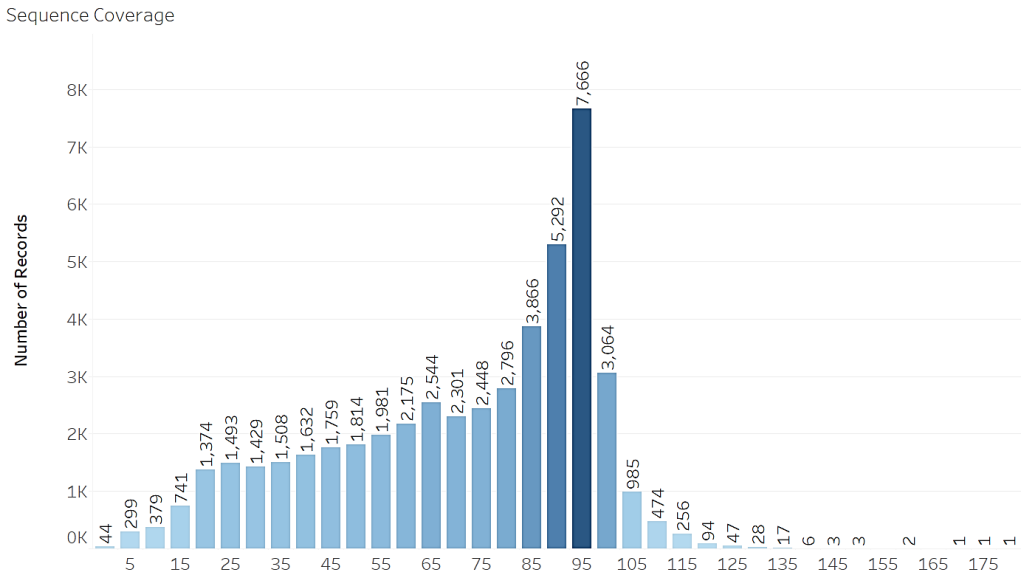

Figure 2. distribution of proteins according to the sequence coverage from the template

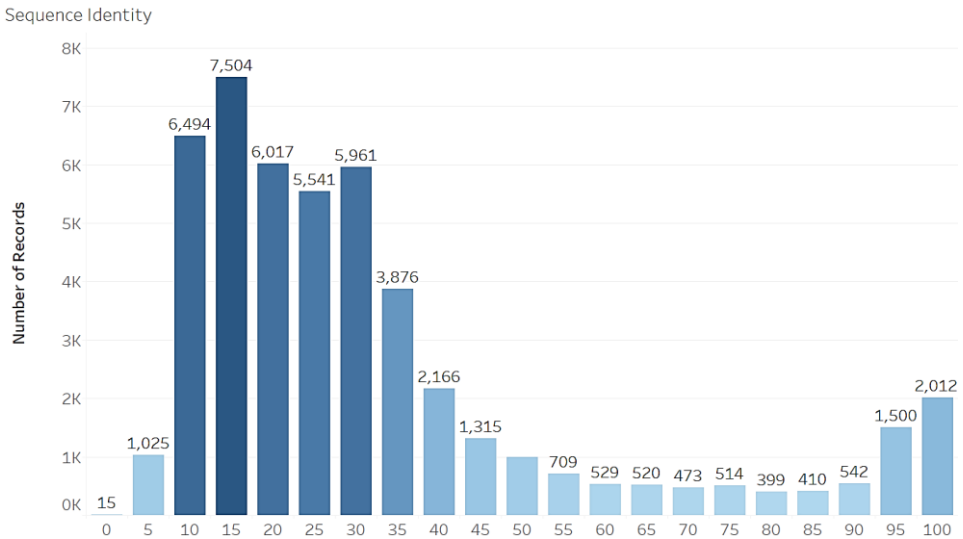

Figure 3. distribution of proteins according to the sequence identity with the template

3.2. Quality assessment

Table 4 shows descriptive statistics of QA scores for all of the proteins in the dataset. Energy, Prove, and Dope have the inverted scale, with lower scores indicating better quality, while other QA tools have the positive scale.

Table 4. Descriptive statistics of QA scores

	Energy	Ramach.	Verify3D	Prove	Errat	Procheck	Qmean	Dope	VoroMQA
Mean	-2.10	98.15	64.87	5.62	86.77	55.62	66.03	-0.67	35.22
Std. Error	0.00	0.01	0.13	0.01	0.05	0.09	0.03	0.00	0.06
Median	-2.13	98.41	73.33	5.70	89.03	50.00	65.00	-0.67	37.40
Mode	-1.85	100.00	0.00	5.70	100.00	44.00	65.05	-0.73	45.98
Std. Dev.	0.32	1.65	28.27	1.83	10.36	18.78	7.27	0.78	12.13
Range	9.90	45.00	100.00	25.70	100.00	89.00	98.97	8.80	61.73
Minimum	-10.36	55.00	0.00	0.00	0.00	11.00	1.03	-4.31	1.53
Maximum	-0.46	100.00	100.00	25.70	100.00	100.00	100.00	4.49	63.26

3.3. Correlation between the first and the second database

Figure 4 shows correlation coefficients between each of the three main protein features (length, sequence identity, and sequence coverage) and QA scores. The last column shows the average of all scores. The values shown are absolute since some QA tools have inverted scales. Due to sequence identity showing the highest mean correlation across all QA scores, it has been chosen as a reference value to divide proteins into groups for proper comparison, removing possible influence on the results.

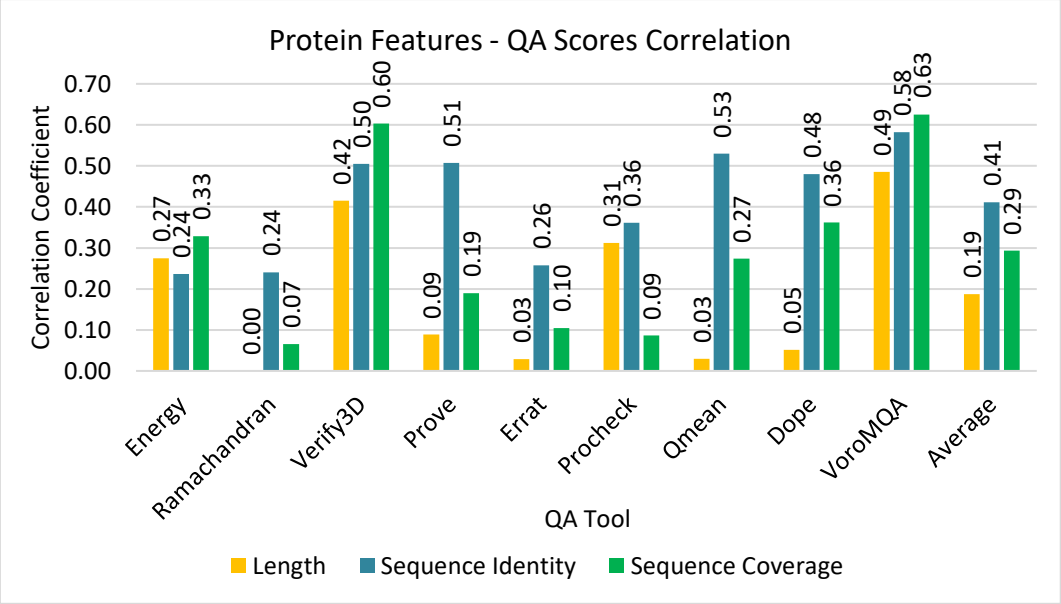

Figure 4. Correlation between protein features and QA scores

Figure 5 (a and b) shows, in details, the correlation between sequence identity and different QA methods. Quality assessments for each protein characteristic have been merged into two groups/figures, based on the similarity of the output score, and with the intention to simplify the presentation of the data. The first part of the figure (a) contains the results from the following quality assessments: Dope, Energy, and Prove, while the second part of figure (b) contains the results from the following quality assessments: Ramachandran (allowed percentage), Errat, Procheck, Qmean, Verify3D, and VoroMQA.

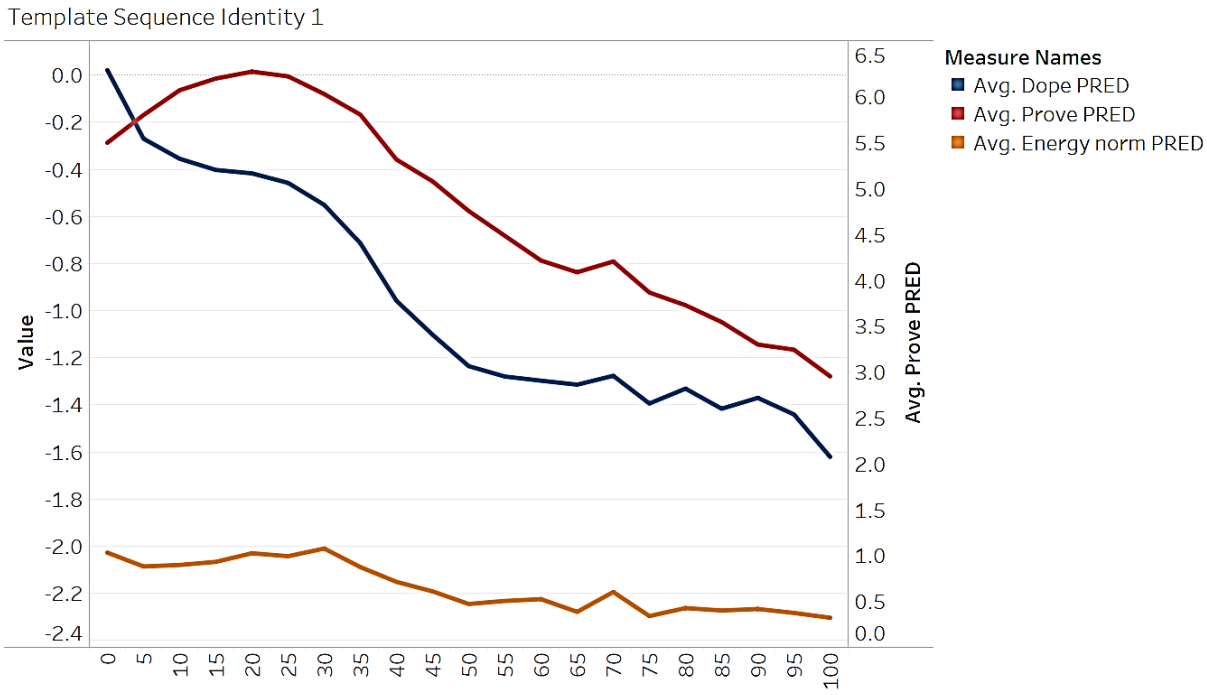

Figure 5a. Quality assessment scores based on the sequence identity between the predicted protein and the template

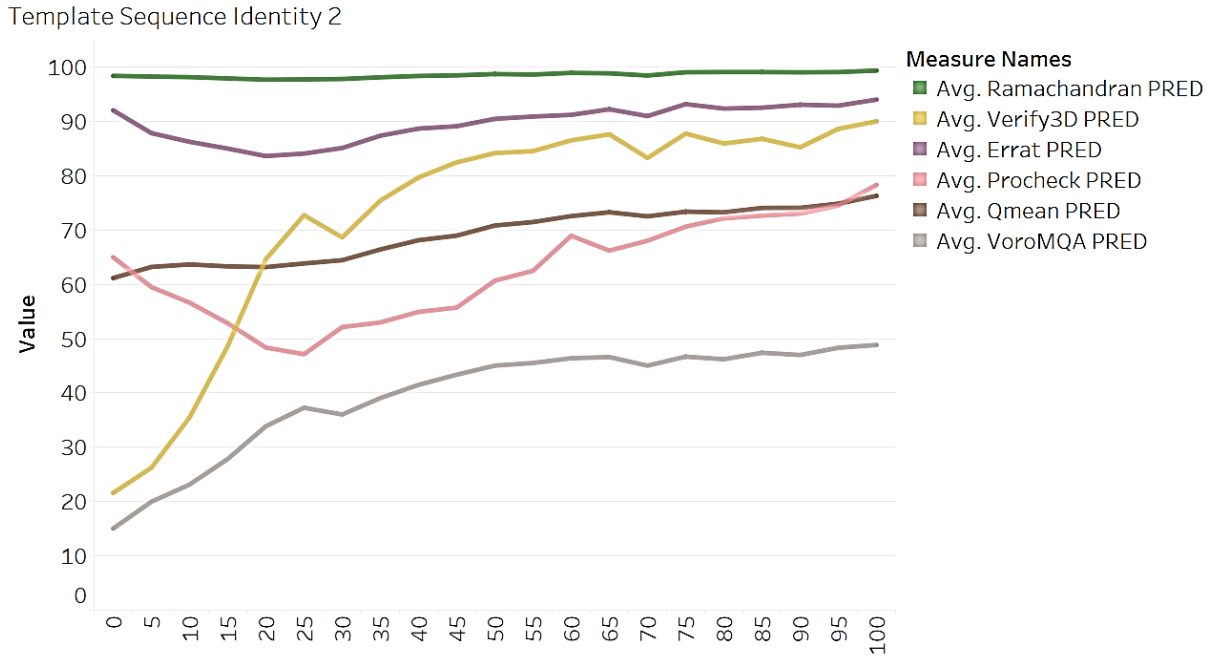

Figure 5b. Quality assessment scores based on the sequence identity between the predicted protein and the template

After organizing the predicted proteins into 20 groups based on the sequence identity with the template (since that feature of predicted proteins has been shown to have a consistent correlation with the quality scores of predicted proteins), they were filtered so that there are at least 30 proteins with the same template per subgroup. Then, a total of 498 cross-comparisons have been made and the end result is that the difference in quality scores between the predicted proteins is consistent with the difference in quality scores between the templates 82% of the time. This means that, if two protein templates are compared, when one template has higher QA scores than the other, e.g., it can also be expected that the proteins predicted from that template will also have higher QA scores, than the proteins predicted from the template with the lower QA scores. This relation is illustrated in Figure 6, which does not contain actual data but simply a representation of the relationship. T1 (blue) represents

the QA scores of template 1, with QA scores of its predicted proteins shown below it, while the template 2 (T2) with its predicted proteins is shown in yellow. Dashed lines represent the average QA scores of predicted proteins. Note that the templates have been positioned higher on the Y-axis for the ease of representation – this does not necessarily indicate that the templates have higher QA scores than the predicted proteins, however, the analysis does show that template proteins have higher QA scores, on average, when compared to the predicted proteins (results shown in the discussion part).

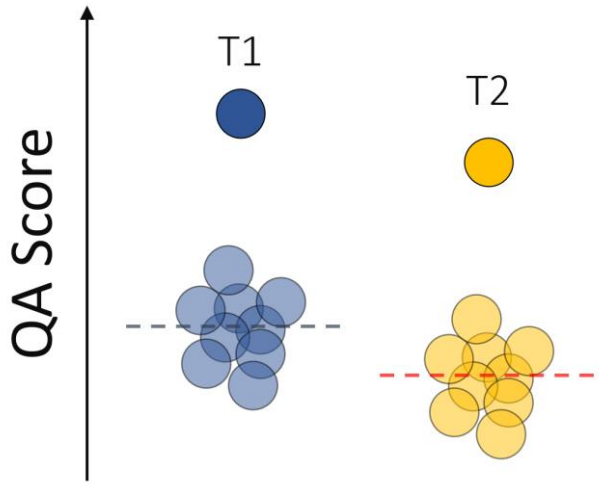

Figure 6. The relation of the QA scores between templates and predicted proteins

4. Discussion

When the relationship between the templates and the predicted proteins is analyzed, it is interesting to see the large disparity in the number of proteins being used as templates – out of 35710 experimentally determined proteins from the sample (first database), only ~18% (6656) have been used as templates for the prediction of 48523 proteins from the second database. Two of them have been used as a template for the prediction of more than 1000 other proteins each, while two-thirds of them (~66%) have been used as a template only 1, 2, or 3 times.

The sequence identity of predicted proteins to their templates is low on average (35.11%), which is also visible from figure 3, but it is still reasonable since it has been shown that proteins with sequence identity as low as 20% can actually be homologous in terms of structure and function [34].

Sequence coverage, on the other hand, is larger, with mean value of 73.27, indicating that S-M tends to include a bigger portion of the protein template during the prediction. This is also visible from figure 2, showing that the largest proportion of predicted proteins from the database have between 95% and 100% sequence coverage. This is expected due to the fact that some amino acids have similar physiochemical properties and even though they might differ between the template and the prediction, the end effect on the structure might be similar [1]. Most of the proteins fall into the shorter group, visible from figure 1, with an average length of 287.60, but this is expected since the first database, containing templates, contains only monomeric proteins.

Regarding the QA, it is interesting to note that the quality assessment software could be divided into two categories with respect to the quality scores – those which give similar quality score to templates and predicted proteins, and those which give significantly lower scores, on average. The results for the templates are shown in the separate study [22], however they are briefly mentioned here where necessary.

Negative energy and Ramachandran assessment give very similar average scores to templates and predicted proteins (Energy standardized: -2.08 vs -2.10: Ramachandran: 99.53 vs 98.15). This shows that modeling of the proteins is generally performed in such a way that the protein model is optimized in terms of packing and *phi/psi* angles. It is interesting to note here that Ramachandran scores seem to be very good, most of the time, which would put Ramachandran as the least reliable method for the determination of structural quality, on its own.

Other tools show large differences between the mean QA scores of templates and predicted proteins (Verify3D: 83.47 vs 64.87; Prove: 1.48 vs 5.62; Errat: 92.84 vs 86.77; Procheck: 81.33 vs 55.62; Qmean: 88.14 vs 66.03; DOPE: -1.9 vs -0.67; VoroMQA: 51.21 vs 35.22), as expected since experimental structure determination is a

more reliable method. Moreover, QA tools are trained on the solved structures, which introduces a bias into the scoring process.

Regarding the features of predicted proteins: sequence length and coverage of template sequences do not show a consistent correlation with the quality scores, and their average correlation is very low and low, respectively. However, it is visible that certain QA tools might be "susceptible" to the length of the protein, with Verify3D and VoroMQA showing a moderate correlation between the QA scores and protein lengths, which is consistent with the previous study done only on the protein templates [22].

The characteristic of predicted proteins which showed a consistent correlation with the quality scores is sequence identity to the template –a consistent increase in the quality of predicted proteins that have higher sequence identity to the template is visible, compared to those with lower sequence identity. This is the reason why the predicted proteins have been divided according to the sequence identity in the last part of the analysis. The average Pearson correlation coefficient is moderately positive – 0.41, and these results are partially expected. The higher the sequence identity, the closer the two proteins are structurally (template and predicted protein), which would, as a consequence, result in their QA scores being more similar as well. This is consistent with other studies which have shown that it is possible to predict protein model QA scores from the sequence alignment – a step necessary for the prediction [35].

Finally, this study shows that the quality scores of predicted proteins are generally consistent with the quality scores of their templates in more than 80% of the cases. Comparing predicted proteins based on two templates, average quality scores of predicted proteins are higher if the corresponding template has a higher quality score, and vice versa (with the adjustment to the sequence identity). Although correlation doesn't necessarily mean causation, relatively high number (82%) indicates that there is "a transfer of property" between the templates and predicted proteins, when it comes to the QA scores, even with the adjustment for the possible sequence identity bias by making the comparison only between the proteins of similar sequence identities.

5. Conclusion

Analyzing the relationships between predicted proteins (from S-M repository) and their templates (from the PDB) on a larger scale, certain trends are visible. Sequence identity can play an important role on the QA scores of predicted proteins, with most of the QA results showing moderate positive correlation to it. Sequence coverage and protein length do not show the same level of correlation, although it is moderate in some instances, indicating that certain QA tools can be biased towards the protein length, with longer proteins having better QA scores. Correlating the QA scores between the predicted proteins and their templates, a significant link can be noticed between the predicted proteins having higher QA scores on average. This occurs if the template they are predicted from also has a higher QA score, when compared to the predicted proteins and templates of lower QA scores. This is an important aspect that should be taken into consideration during the protein prediction process and template selection. Further analysis of QA scores on a local level might give additional insights into the trends of QA tools when scoring protein 3D structures.

References

[1] "Introduction to Proteins: Structure, Function, and Motion, Second Edition," *CRC Press*. https://www.crcpress.com/Introduction-to-Proteins-Structure-Function-and-Motion-Second-Edition/Kessel-Ben-Tal/p/book/9781498747172 (accessed Oct. 02, 2019).

[2] R. A. Chica, "Protein Engineering in the 21st Century," *Protein Sci. Publ. Protein Soc.*, vol. 24, no. 4, pp. 431–433, Apr. 2015, doi: 10.1002/pro.2656.

[3] C. A. Orengo, A. E. Todd, and J. M. Thornton, "From protein structure to function," *Curr. Opin. Struct. Biol.*, vol. 9, no. 3, pp. 374–382, Jun. 1999, doi: 10.1016/S0959-440X(99)80051-7.

[4] "Comparison of Crystallography, NMR and EM - Creative Biostructure." https://www.creative-biostructure.com/comparison-of-crystallography-nmr-and-em_6.htm (accessed Oct. 30, 2019).

[5] R. P. D. Bank, "RCSB PDB: Homepage." https://www.rcsb.org/ (accessed Oct. 02, 2019).

[6] A. Fiser, "Template-based protein structure modeling," *Methods Mol. Biol. Clifton NJ*, vol. 673, pp. 73–94, 2010, doi: 10.1007/978-1-60761-842-3_6.

[7] J. Lee, P. L. Freddolino, and Y. Zhang, "Ab Initio Protein Structure Prediction," in *From Protein Structure to Function with Bioinformatics*, D. J. Rigden, Ed. Dordrecht: Springer Netherlands, 2017, pp. 3–35. doi: 10.1007/978-94-024-1069-3_1.

[8] S. Vangaveti, T. Vreven, Y. Zhang, and Z. Weng, "Integrating ab initio and template-based algorithms for protein–protein complex structure prediction," *Bioinformatics*, doi: 10.1093/bioinformatics/btz623.

[9] S. Abeln, J. Heringa, and K. A. Feenstra, "Strategies for protein structure model generation," 2017.
[10] Y. Zhang, "Protein Structure Prediction: Is It Useful?," *Curr. Opin. Struct. Biol.*, vol. 19, no. 2, pp. 145–155, Apr. 2009, doi: 10.1016/j.sbi.2009.02.005.
[11] J. Cheng, A. N. Tegge, and P. Baldi, "Machine Learning Methods for Protein Structure Prediction," *IEEE Rev. Biomed. Eng.*, vol. 1, pp. 41–49, 2008, doi: 10.1109/RBME.2008.2008239.
[12] M. Gao, H. Zhou, and J. Skolnick, "DESTINI: A deep-learning approach to contact-driven protein structure prediction," *Sci. Rep.*, vol. 9, no. 1, pp. 1–13, Mar. 2019, doi: 10.1038/s41598-019-40314-1.
[13] S. Wang, J. Peng, J. Ma, and J. Xu, "Protein Secondary Structure Prediction Using Deep Convolutional Neural Fields," *Sci. Rep.*, vol. 6, p. 18962, Jan. 2016, doi: 10.1038/srep18962.
[14] S. P. Nguyen, Y. Shang, and D. Xu, "DL-PRO: A novel deep learning method for protein model quality assessment," in *2014 International Joint Conference on Neural Networks (IJCNN)*, Jul. 2014, pp. 2071–2078. doi: 10.1109/IJCNN.2014.6889891.
[15] R. Cao, B. Adhikari, D. Bhattacharya, M. Sun, J. Hou, and J. Cheng, "QAcon: single model quality assessment using protein structural and contact information with machine learning techniques," *Bioinformatics*, vol. 33, no. 4, pp. 586–588, Feb. 2017, doi: 10.1093/bioinformatics/btw694.
[16] K. Uziela, D. Menéndez Hurtado, N. Shu, B. Wallner, and A. Elofsson, "ProQ3D: improved model quality assessments using deep learning," *Bioinformatics*, vol. 33, no. 10, pp. 1578–1580, May 2017, doi: 10.1093/bioinformatics/btw819.
[17] R. Cao, Z. Wang, Y. Wang, and J. Cheng, "SMOQ: a tool for predicting the absolute residue-specific quality of a single protein model with support vector machines," *BMC Bioinformatics*, vol. 15, no. 1, p. 120, Apr. 2014, doi: 10.1186/1471-2105-15-120.
[18] C. L. P. Gupta, A. Bihari, and S. Tripathi, "Protein Classification using Machine Learning and Statistical Techniques: A Comparative Analysis," *ArXiv190106152 Cs Q-Bio Stat*, Jan. 2019, Accessed: Oct. 02, 2019. [Online]. Available: http://arxiv.org/abs/1901.06152
[19] A. Dalkiran, A. S. Rifaioglu, M. J. Martin, R. Cetin-Atalay, V. Atalay, and T. Doğan, "ECPred: a tool for the prediction of the enzymatic functions of protein sequences based on the EC nomenclature," *BMC Bioinformatics*, vol. 19, no. 1, p. 334, Sep. 2018, doi: 10.1186/s12859-018-2368-y.
[20] A. Runthala and S. Chowdhury, "Refined template selection and combination algorithm significantly improves template-based modeling accuracy," *J. Bioinform. Comput. Biol.*, vol. 17, no. 02, p. 1950006, Nov. 2018, doi: 10.1142/S0219720019500069.
[21] S. Bienert *et al.*, "The SWISS-MODEL Repository-new features and functionality," *Nucleic Acids Res.*, vol. 45, no. D1, pp. D313–D319, 04 2017, doi: 10.1093/nar/gkw1132.
[22] M. Adilović and A. Hromić-Jahjefendić, "Feature Importance in the Quality of Protein Templates," *Period. Eng. Nat. Sci. PEN*, vol. 9, no. 2, Art. no. 2, Apr. 2021, doi: 10.21533/pen.v9i2.1830.
[23] "PDB101: Learn: Guide to Understanding PDB Data: Introduction," *RCSB: PDB-101*. http://pdb101.rcsb.org/learn/guide-to-understanding-pdb-data/introduction (accessed Oct. 02, 2019).
[24] G. J. Kleywegt and T. A. Jones, "Phi/psi-chology: Ramachandran revisited," *Struct. Lond. Engl. 1993*, vol. 4, no. 12, pp. 1395–1400, Dec. 1996, doi: 10.1016/s0969-2126(96)00147-5.
[25] H. Zhou and Y. Zhou, "Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction," *Protein Sci.*, vol. 11, no. 11, pp. 2714–2726, 2002, doi: 10.1110/ps.0217002.
[26] R. Lüthy, J. U. Bowie, and D. Eisenberg, "Assessment of protein models with three-dimensional profiles," *Nature*, vol. 356, no. 6364, pp. 83–85, Mar. 1992, doi: 10.1038/356083a0.
[27] J. U. Bowie, R. Lüthy, and D. Eisenberg, "A method to identify protein sequences that fold into a known three-dimensional structure," *Science*, vol. 253, no. 5016, pp. 164–170, Jul. 1991, doi: 10.1126/science.1853201.
[28] R. A. Laskowski, M. W. MacArthur, D. S. Moss, and J. M. Thornton, "PROCHECK: a program to check the stereochemical quality of protein structures," *J. Appl. Crystallogr.*, vol. 26, no. 2, Art. no. 2, Apr. 1993, doi: 10.1107/S0021889892009944.
[29] C. Colovos and T. O. Yeates, "Verification of protein structures: patterns of nonbonded atomic interactions," *Protein Sci. Publ. Protein Soc.*, vol. 2, no. 9, pp. 1511–1519, Sep. 1993, doi: 10.1002/pro.5560020916.
[30] J. Pontius, J. Richelle, and S. J. Wodak, "Deviations from standard atomic volumes as a quality measure for protein crystal structures," *J. Mol. Biol.*, vol. 264, no. 1, pp. 121–136, Nov. 1996, doi: 10.1006/jmbi.1996.0628.

[31] P. Benkert, M. Biasini, and T. Schwede, "Toward the estimation of the absolute quality of individual protein structure models," *Bioinforma. Oxf. Engl.*, vol. 27, no. 3, pp. 343–350, Feb. 2011, doi: 10.1093/bioinformatics/btq662.
[32] M. Shen and A. Sali, "Statistical potential for assessment and prediction of protein structures," *Protein Sci. Publ. Protein Soc.*, vol. 15, no. 11, pp. 2507–2524, Nov. 2006, doi: 10.1110/ps.062416606.
[33] K. Olechnovič and Č. Venclovas, "VoroMQA: Assessment of protein structure quality using interatomic contact areas," *Proteins Struct. Funct. Bioinforma.*, vol. 85, no. 6, pp. 1131–1145, 2017, doi: 10.1002/prot.25278.
[34] W. R. Pearson, "An Introduction to Sequence Similarity ('Homology') Searching," *Curr. Protoc. Bioinforma. Ed. Board Andreas Baxevanis Al*, vol. 0 3, Jun. 2013, doi: 10.1002/0471250953.bi0301s42.
[35] X. Deng, J. Li, and J. Cheng, "Predicting Protein Model Quality from Sequence Alignments by Support Vector Machines," *J. Proteomics Bioinform.*, vol. Suppl 9, Nov. 2013, doi: 10.4172/jpb.S9-001.