QUALITY-MODEL CLUSTERING TOOL: A MODULE FOR CLUSTERING PROTEIN MODELS BASED ON QUALITY ATTRIBUTES

Abstract: The process of protein modeling usually involves the production of a variety of structures requiring efficient tools for structure model comparison attempting to choose the best three-dimensional (3D) structure. This paper introduces an alternative method for clustering 3D protein models that, instead of using attributes related to structural alignment to group the data, use quality-attributes of those models to represent and cluster them. This method stands out by removing the need to define a priori a base model for structural alignments. Even so, it is possible to present the most representative structure in each cluster, which is useful for docking or molecular dynamics studies. All the results were statistically analyzed and compared with decisions made by professionals to validate the proposed algorithm. The experiment simulated a usual protein comparative modeling process for different CATH classifications. The calculated variance levels after the dimensional reduction validate the workflow for different protein chain sizes. All the molecular descriptors for the input files are calculated by MHOline 2.0, an online scientific workflow for studies on bioinformatics and computational biology, available for free on www.mholline2.lncc.br, or hand made using specific programs (e.g., MODELLER, PROCHECK) and adjusting the data to the template specified in this document. The Quality-Model Clustering Tool (QMC) and the data set used in this work are available for download on the git repository (github.com/ruanmedina/Quality-Model-Clustering)

Keywords: Protein modeling. Clustering methods. Quality attributes. Bioinformatics.

1M.Sc. student in Computational Modelling, Federal University of Juiz de Fora - Juiz de Fora, MG, Brazil, ruan.medina@engenharia.ufjf.br.
2M.Sc. in Computational Modeling, Federal University of Juiz de Fora - Juiz de Fora, MG, Brazil, arturossi10@gmail.com.
3Ph.D. in Computational Modeling, Federal University of Juiz de Fora - Juiz de Fora, MG, Brazil, priscila.capriles@ufjf.edu.br.
**Resumo:** O processo de modelagem computacional de proteínas geralmente envolve a produção de uma variedade de conformações, o que requer ferramentas eficientes de comparação de modelos estruturais para a seleção das melhores estruturas tridimensionais (3D). Este artigo apresenta um método alternativo para agrupar modelos 3D de proteínas que, ao vez de usar atributos relacionados ao alinhamento estrutural para agrupar os dados, utiliza de atributos de qualidade desses modelos para representá-los e agrupá-los. Esse método se destaca ao eliminar a necessidade de definir *a priori* um template para alinhamentos estruturais. Mesmo assim, é possível apresentar a estrutura mais representativa em cada grupo formado, o que é útil para estudos de docking e dinâmica molecular. Todos os resultados foram analisados estatisticamente e comparados com as decisões tomadas por profissionais para validar o algoritmo proposto. O experimento simulou um processo usual de modelagem comparativa de proteínas para diferentes classificações de CATH. Os níveis de variância calculados após a redução dimensional proposta validam o workflow para diferentes tamanhos de cadeias de proteínas. Todos os descritores moleculares utilizados nos arquivos de entrada podem ser calculados pelo MHOLline 2.0, um workflow científico on-line para estudos em bioinformática e biologia computacional, disponível gratuitamente em www.mholline2.lncc.br ou calculados de forma individual usando programas específicos (por exemplo, MODELLER, PROCHECK) e ajustando os dados para o modelo especificado neste documento. A ferramenta Quality-Model Clustering (QMC) e o conjunto de dados usados neste trabalho estão disponíveis para download no repositório *git* (github.com/ruanmedina/Quality-Model-Clustering).

**Palavras-chave:** Modelagem de proteínas. Métodos de agrupamento. Atributos de qualidade. Bioinformática.
1 INTRODUCTION

The knowledge of the three-dimensional (3D) structure of proteins is essential to study diseases such as parasitoses, viruses, and cancer (LEE; FREDDOLINO; ZHANG, 2017). Nowadays, the 3D protein structure prediction (PSP) are often guided by computational experiments in many kinds of research, since experimental PSP remains costly and time-consuming (VERLI, 2014). Comparative modeling is an example of a computational method for PSP (ESWAR et al., 2006). This method constructs 3D models using known structures of macromolecules as template. These structures are obtained from databases (e.g., PDB - Protein Data Bank). Comparative modeling is highly dependent on the quality of templates and the evolutionary relationship between them and the modeled protein (CAPRILES et al., 2010). Additionally, the process demands the production of a large number of conformations and requires several refinements and validation steps, making the final decision of the best models difficult to make (VERLI, 2014).

A way to automate the 3D modeling process has been presented by the scientific workflow MHOLline (CAPRILES et al., 2010). Starting with an input *fasta* file, it runs a BLASTp alignment against the PDB database, and it associates the output into four groups (G0-G3) according to the modeling viability. Proteins classified in the G2 group, based on the expected quality of models, are modeled using the MODELLER program (ŠALI; BLUNDELL, 1993). MHOLline implements other programs such as PROCHECK (LASKOWSKI et al., 1993) and Molprobity (CHEN et al., 2010) to validate the stereochemistry of 3D protein models, SIGNALP (PETERSEN et al., 2011) to identify signal peptide regions, PSIPRED (MCGUFFIN; BRYSON; JONES, 2000) to predict the secondary structure of proteins, and TMHMM (KROGH et al., 2001) and HMMTOP to identify transmembrane regions (TUSNÁDY; SIMON, 2001).

Once the comparative modeling technique is a stochastic method that generates a sample of 3D structures, the decision-making about the best geometry is a challenge. Structure-based clustering techniques, such as RMSD-based techniques, is commonly applied to reduce the data into a subset of models (SIEW et al., 2000; JAMROZ; KOLINSKI, 2013; WU et al., 2017). These tools require a reference structure for 3D superposition against models. However, considering the user desires to know which model is the best, selecting the reference one can also be a challenge.

In this paper, we present a method to clustering 3D protein models based on the analysis of the quality of the generated models, using a set of attributes describing the stereochemistry and energy of these structures. The proposed method uses clustering algorithms to identify groups of lower average energy and cluster similar data into proper groups. Hence, it supplies
the user with a set of automatically refined structures. Initially, we developed the algorithm to incorporate the MHOLline workflow, but the method also can be used as a standalone system.

2 MATERIAL AND METHODS

2.1 Implementation and Operation

The system considers a given ensemble of conformations of a protein to evaluate and compares the quality of these models to search for preferential conformation(s). Figure 1 shows the operating pipeline developed in this work (in Python3). In the first step, we define the representation of each protein model based on quality attributes such as molpdf, DOPE score, DOPE-HR score, Normalized DOPE score, and ‘Allowed’ and ‘Outlier’ residues. The first four are results from evaluation functions of the comparative modeling software MODELLER, and the last two from the analysis generated by the Molprobity program. These attributes are often already available when doing comparative modeling, so choosing them for this initial analysis exclude the demand for running additional programs. The text follows explaining each attribute (more details about the first four functions are available in the MODELLER manual):

- **molpdf**: Is the MODELLER’s objective function \( F(R) \), which minimizes the energy of the atoms interactions and geometric features restraints.

- **DOPE score**: The Discrete Optimized Protein Energy score is a statistical potential optimized for model assessment based on the standard MODELLER energy function. It is often used to select the best structure from a collection of models built by MODELLER.

- **DOPEHR score**: Attribute very similar to DOPE score, but obtained at a higher resolution (using a bin size of 0.125Å rather than 0.5Å).

- **Normalized DOPE score**: The normalized DOPE score derives from the statistics of raw DOPE scores. This is a Z-score which positive values represent a ‘bad’ model and scores lower than -1 or so represents a ‘native-like’ structure.

- **Allowed residues**: Percentage of residues on allowed regions from stereochemical analysis using the Molprobity program.

- **Outlier residues**: The number of residues in not allowed regions stereochemical analysis using the Molprobity program.

Following the definition of data representation, the software automatically normalizes the attributes according to user-defined criteria. Two normalization methods have
been incorporated: MinMaxScaler, fits the attribute values between 0 and 1, and StandardScaler, fits in a Gaussian distribution (mean around 0 and standard deviation of 1). Then, the data are submitted to a dimensionality reduction via Principal Component Analysis (PCA) algorithm, reducing the number of attributes to two, without losing the essence of the data set (JOLLIFFE; CADIMA, 2016).

In the next step, the parameters of the clustering are established. We implemented six (6) different clustering methods, using the Scikit-learn Python module (PEDREGOSA et al., 2011): K-means (LLOYD, 1982), Ward (MURTAGH; LEGENDRE, 2014), Spectral Clustering (LUXBURG, 2007), Affinity Propagation (DUECK, 2009), Mean Shift (WU; YANG, 2007) and DBSCAN (ESTER et al., 1996). The first three algorithms require the estimated number of clusters ($k$), so the workflow uses the Silhouette Coefficient evaluation and a variation of the Elbow Method. Both algorithms are iteratively tested from $k_{min} = 1$ to the user-entered $k_{max}$ (default = 10). Next, a brief explanation of each method:

- **Silhouette Coefficient**: The silhouette analyzes the relation between intra-cluster distance and the nearest-cluster distance, *i.e.*, how close each point in a cluster is to the points in its cluster (value of +1) when compared with neighboring clusters (value of −1). To evaluate the clustering of the samples as a whole, we use the mean
Silhouette Coefficient over all samples (ROUSSEEUW, 1987).

- **Elbow Method**: The Elbow Method analyzes the variation of the mean of the Sum of Squared Errors (SSE) of the clusters with increasing $k$. As the SSE related to $2 \leq k \leq k_{max}$ is always descending (more groups necessarily decrease the mean intra-cluster distance), mathematically, we search for the $k$ related to the more significant drop of error. Nevertheless, this method is not well suited to treat samples with two or fewer clusters. In these cases, one must use the Silhouette Coefficient method (KODINARIYA; MAKWANA, 2013).

Once having the required parameters, the system fits the data for each method using suitable settings. The parameters used for each method were omitted, but may be accessed through the git repository. Here, we present some main characteristics of the tested methods:

- **Affinity Propagation**: It imports the concept of ‘Message Passing’ between data points. Each of the data points is a possible representative one for a new cluster. To discover the number of clusters and to form the groups of data, the method uses a similarity function to calculate and update a similarity matrix in an iterative process (DUECK, 2009).

- **K-means**: It is a Partitioning Method that usually requires the expected number of clusters. The basic idea is to find clustered data that minimize a particular error criterion of the distance between each instance to the centroid of the cluster allocated at the moment. In every iteration, the method recalculates the centroid and reallocates the data points (LLOYD, 1982).

- **Ward**: It is an Agglomerative Hierarchical Clustering Method that requires the number of desired clusters. This method considers every instances a cluster at the beginning of the interactions and tends to merge the pair of groups with the minimum inter-cluster distance (MURTAGH; LEGENDRE, 2014).

- **Spectral Clustering**: The method uses the information derived from eigenvalues - or spectrum - of the similarity matrix of the data. Then, a standard clustering method - such as K-means - is employed to cluster in lower-dimensional space. All these operations make the method very computationally demanding (LUXBURG, 2007).

- **DBSCAN**: It is a Density-based Method that works with the concept of ‘Dense Regions’ of data. It characterizes a cluster as a dense region surrounded by a non-dense one. The algorithm searches for clusters by searching the neighborhood of each object in the
database and checks whether it contains more than the minimum number of objects to form a cluster. Otherwise, it considers the point of data as outliers (ESTER et al., 1996).

- **Mean Shift**: Built upon the concept of Kernel Density Estimation (KDE), it estimates the clusters based on a density function in a range from one arbitrary point. The mode of the density within the range is estimated, and become the next center to evaluate. The process converges when the mode and center are the same (WU; YANG, 2007).

At the end of the clustering step, the algorithm returns the medoid structure for each cluster. For those clustering methods that do not work with a medoid structure, the system calculates the estimated medoid from the centroids of each cluster. The pipeline output files, by default, are presented in the “./Clusters_Data” directory and summarized in the “analysis.out” file.

### 2.2 Computational Experiment and Validation

To test the efficacy of the algorithm, we used a benchmark of 137 different protein sequences. Each pair of protein sequences with an associated template for comparative modeling was divided into five (5) categories (Class A to E) based on the length of the amino acid sequence, as shown in Table 1. The benchmark protein set is a subset of the presented in the work of (ZHANG; SKOLNICK, 2005) that describes the development of TM-align and uses a set of 200 nonhomologous PDB proteins for testing structure alignment algorithms (ZHANG; SKOLNICK, 2005). To the benchmark applied in this paper, we added some proteins aiming to fill the gaps in the desired dataset, considering the plurality of classification and protein length. We removed some proteins with dubious CATH/SCOP classification. For this first analysis, we preferred to apply the method to well-known proteins presented in consolidated and straightforward benchmarks to facilitate the analysis process. The entire data set can be accessed through the git repository (github.com/ruanmedina/Quality-Model-Clustering). The summary of the data is presented in Table 1.

We aim to simulate a real comparative modeling process. First, we extract the sequences from the pdb files using a Python3 script with BioPython (COCK et al., 2009), followed by the search of potential templates using BLAST (ALTSCHUL et al., 1990) in MHOLine 2.0 workflow. The template selection criteria applied for each protein in the dataset followed the suggestion of its BLAST results, but disregarding the candidates with high identity (close to 100%) and low identity (less than 25%). Once having the protein-template pairs, we generated 50 models for each protein using a Python3 script for MODELLER (ESWAR et al., 2006), and the addiction of BioPython library to manipulate data and download the chosen template
DATASET SUMMARY

| Template coverage | Class A | Class B | Class C | Class D | Class E |
|-------------------|---------|---------|---------|---------|---------|
| < 50%             | 17      | 39      | 27      | 5       | 4       |
| ≥ 50%             | 6       | 27      | 12      | 4       | 3       |

| CATH classification | Mainly Alpha | Mainly Beta | Alpha Beta | Few SS<sup>a</sup> |
|---------------------|--------------|-------------|------------|-------------------|
|                     | 10           | 19          | 6          | 0                 |
|                     | 5            | 3           | 8          | 3                 |
|                     | 5            | 44          | 25         | 6                 |
|                     | 3            | 0           | 0          | 0                 |

| Number of different pairs of protein-template | 23 | 66 | 39 | 9<sup>c</sup> | 7<sup>d</sup> |

Table 1: Summary of the benchmark dataset. <sup>a</sup> SS: Secondary Structure. <sup>b</sup> Visually determined. <sup>c</sup> 4 proteins were used twice with different templates. <sup>d</sup> 3 proteins were used twice with different templates.

structure files. Then, we submitted all generated models to Molprobity (CHEN et al., 2010) to assess the stereochemical quality of the models. Lastly, the method presented in this work analyzes and clusters the protein models, and then exhibits the final results.

### 3 RESULTS AND DISCUSSION

Using the methodology described and the implemented QMC, the user may automate part of the refinement step, such as in protein modeling and pattern evaluation of quality attributes. This method allows the user to perform conformation clustering without knowing a reference structure - often needed in clustering methods based on geometric properties - nor having to calculate all-versus-all alignment between the generated pdb models.

In this section, we show an example of usage, preliminary studies about the attributes used to represent the conformations, and variance reviews when reducing the data dimensionality.

All tests were performed on a Kubuntu 18.04 64 bits with 8Gb of RAM and an Intel® Core™ CPU (8 cores) i7-7700 3.6GHz.

#### 3.1 Usage and Base Case Scenario

We have already described some of the QMC applicability and how flexible it can be by leaving the user free to choose the best parameters for the situation, but also suggesting default ones for general studies. However, what kind of information may a researcher get
from using this tool? Figure 2 shows an overview of a general modeling process for the protein 1F4H, which is a hydrolase of Escherichia coli (strain K12) organism. Suppose that doing a BLAST search, the most similar structure found was the protein 3IAP, which is also a hydrolase of the same organism with a 99% of identity and a very low e-value. The methodology described follows with a comparative modeling step using MODELLER, generating $N_{\text{conf}}$ (default $N_{\text{conf}} = 50$) slightly different structures. MolProbity evaluates each conformation. By the end of this step, the quality attributes needed to represent each structure have already been computed.

Having an input table of attributes, the QMC normalizes and reduces the dimensionality of the data into two transformed attributes capable of representing the data as much as it is possible in two-dimensional (2D) space. When doing that, the researcher can access which of the original attributes carry the most significant variances for the data, as shown in Figure 2. Also, the method presents the percentage of the original variance represented in two-dimensional space.

Subsequently, QMC fits the transformed two-dimensional data in different clustering methods. In Figure 2, we showed the Mean Shift results as an example. The optimized clusters are presented to the user, who may analyze the dispersion of the reduced quality attributes of the generated models. Each cluster is defined as a representative structure placed next to the centroid of the group. The representative structures may be further analyzed as a form to get sufficiently different geometric conformations, e.g., different lengths of sheets, or different possible positions for coils, as shown in Figure 2. This example shows that it is possible to get different structural conformations even though there is a high level of identity between the sequence to be modeled and the chosen template.

3.2 Variance and Attributes Analysis

Since QMC uses PCA for component analysis, it is necessary to investigate if the reduction in a two-dimensional space is sufficient to represent the essence of the data. It is crucial to have the same tendency expressed when considering all the quality attributes with the original variance.

First, results from preliminary tests (not presented) over the correlation of the attributes used so far, was the base to choose the new dimensionality order (i.e., two-dimensional). The six (6) quality attributes may be separated into three classes: (a) molpdf; (b) DOPE score, DOPE-HR score, and Normalized DOPE score; and (c) ‘Allowed’ and ‘Outlier’ residues. Attributes of different classes presented little correlation between each other, but, frequently, high correlation with attributes in the same class. With this set of attributes, there is no reason to use dimensions higher than three. Furthermore, it was not possible to assign an attribute of each class for
Figure 2: An overview of a general comparative modeling process for the protein 1F4H showing the kind of analysis and information researchers may get from clustering based on quality attributes.

representing the models and remove the dimensionality reduction phase, because the higher levels of variation fluctuate among the attributes of the same class for different proteins.

So, why not reduce the dimensionality to three? We investigated the contribution of each of the attributes to understand how it was possible to reach the levels of the information we keep reaching in PCA reductions. Figure 3 shows the frequency of each attribute at the top-two list of most significant attributes (that carries the most variance) for each of the tested proteins in this work. To select the most significant attributes, we considered the highest components of the PCA eigenvectors related to each component of the new space. The results show that frequently an attribute of class (b) is taken as the most significant for the first coordinate of the PCA, and a MolProbity attribute - class (c) - is taken for the second one.

A special mention should be made about the molpdf attribute in class (a). While it is not chosen frequently, in cases where there is no variation between the MolProbity calculated attributes for the models, i.e., when all the residues are in allowed regions of torsion in Ramachandran, it appears to play an important role. In such cases, using a 3D reduction would result in choosing at least two attributes from the same class, which does not improve our representation. This over-dimension happens because - in these cases of none outliers residues...
- the class (c) is no longer one of the most representative and becomes a set of attributes without no information. Besides, working in two dimensions reduces the computational complexity of calculations in the clustering methods. It also facilitates the visualization of formed groups in 2D scatters plots that can be understood by any tool user.

![Figure 3: Frequency of each attribute at the top-two most significant values for each tested protein. The different colors show whether the biggest part of the variance relative to a given attribute was placed in the first (Coor1) or, second (Coor2) coordinate of PCA transformed data.](image)

We still have to investigate how much variance we are capturing from the data. Moreover, it is interesting to know whether the two transformed coordinates are enough to represent proteins as their sizes grow. Figure 4 shows how is the percentage of the original variance represented in two-dimensional space created by PCA for each of the defined classes with different protein range sizes. We estimated the percentage of the original variance by the sum of the PCA eigenvalues related to each component of the new space. The results show that, for all the classes, the represented variance is considered high. The observed total mean - (0.86 ± 0.044)% - was sufficient relevant with little standard deviation. Outlier data only occurred in the bigger classes (B and C), but always as over-representation situations. Most of these outliers derived from proteins in which the set of models does not show outliers residues in the Ramachandran analysis.

Moreover, we applied a Kruskal-Wallis H-test to the data. The test presented a strongly non-representative $p – value = 0.455$ indicating that there is no evidence in this set of data for believing the classes medians are significantly different from each other. We choose to use a non-parametric test due to the non-Gaussian distribution detected.

Importantly, in Figure 4, the length of error bars tends to decrease with increasing chain
size of proteins. The lower chain proteins appear to have a homogeneous distribution of their variance by the quality attributes used so far. Also, for these protein classes, it is more common the occurrence of the situation mentioned previously, in which class \( c \) loses all significance, making the representation of the models difficult. This difficulty indicates the future need to study increasing the number of quality attributes used as input of the algorithm or an adaptation of the order of dimensionality reduction for these cases.

4 CONCLUSION

In this work, we presented an alternative method and a tool for clustering protein models based on quality attributes instead of geometric properties. It was possible to verify the capability of the chosen attributes to describe the data and the ability to capture sufficient variance of data, even in two-dimensional reduced space obtained by PCA method. The clustered data reduces the analysis processes by selecting a small subset of representative structures. The proposed method has proven to be useful mainly when the researcher does not know the reference structure to apply geometric evaluations, which usually occurs during protein modeling processes.
Further works should validate the usefulness of the technique using more benchmark proteins, such as I-TASSER SPICKER Set-II, CASP11, or construct a database with new benchmarks. It is also necessary to compare the efficiency for results clustered by quality versus geometry, to identify the pros and cons of using each of the approaches in different situations.

ACKNOWLEDGMENTS

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brazil (CAPES) – Finance Code 1834993, Fundação de Amparo à Pesquisa do Estado de Minas Gerais – (FAPEMIG) – Finance Code 11651/2017 and Universidade Federal de Juiz de Fora – (UFJF) – Grupo de Educação Tutorial da Engenharia Computacional.

REFERENCES

ALTSCHUL, S. F. et al. Basic local alignment search tool. Journal of molecular biology, Elsevier, v. 215, n. 3, p. 403–410, 1990.

CAPRILES, P. et al. Structural modelling and comparative analysis of homologous, analogous and specific proteins from trypanosoma cruzi versus homo sapiens: putative drug targets for chagas disease treatment. BMC Genomics, p. 11:610, 2010.

CHEN, V. et al. Molprobity: all-atom structure validation for macromolecular crystallography. BMC bioinformatics, Acta Crystallographica, D66, p. 12–21, 2010.

COCK, P. J. A. et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics, v. 25, n. 11, p. 1422–1423, 03 2009. ISSN 1367-4803. Disponível em: <https://doi.org/10.1093/bioinformatics/btp163>.

DUECK, D. Affinity propagation: clustering data by passing messages. [S.l.]: Citeseer, 2009.

ESTER, M. et al. A density-based algorithm for discovering clusters in large spatial databases with noise. v. 96, n. 34, p. 226–231, 1996.

ESWAR, N. et al. Comparative protein structure modeling using modeller. Current Protocols in Bioinformatics, v. 15, n. 1, p. 5.6.1–5.6.30, 2006. Disponível em: <https://currentprotocols.onlinelibrary.wiley.com/doi/abs/10.1002/0471250953.bi0506s15>.

JAMROZ, M.; KOLINSKI, A. Clusco: clustering and comparison of protein models. BMC bioinformatics, BioMed Central, v. 14, n. 1, p. 62, 2013.

JOLLIFFE, I. T.; CADIMA, J. Principal component analysis: a review and recent developments. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, The Royal Society Publishing, v. 374, n. 2065, p. 20150202, 2016.

KODINARIYA, T. M.; MAKWANA, P. R. Review on determining number of cluster in k-means clustering. International Journal, v. 1, n. 6, p. 90–95, 2013.
KROGH, A. et al. Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. *Journal of Molecular Biology*, v. 305, n. 3, p. 567–580, 2001.

LASKOWSKI, R. et al. Procheck - a program to check the stereochemical quality of protein structures. *J. App. Cryst.*, v. 26, p. 283–291, 1993.

LEE, J.; FREDDOLINO, P. L.; ZHANG, Y. Ab initio protein structure prediction. In: *From protein structure to function with bioinformatics*. [S.l.]: Springer, 2017. p. 3–35.

LLOYD, S. Least squares quantization in pcm. *IEEE transactions on information theory*, IEEE, v. 28, n. 2, p. 129–137, 1982.

LUXBURG, U. V. Tutorial on spectral clustering. *Statistics and computing*, Springer, v. 17, n. 4, p. 395–416, 2007.

MCGUFFIN, L.; BRYSON, K.; JONES, D. The psipred protein structure prediction server. *Bioinformatics*, v. 4, p. 401–405, 2000.

MURTAGH, F.; LEGENDRE, P. Ward’s hierarchical agglomerative clustering method: which algorithms implement ward’s criterion? *Journal of classification*, Springer, v. 31, n. 3, p. 274–295, 2014.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011.

PETERSEN, T. et al. Signalp 4.0: discriminating signal peptides from transmembrane regions. *Nature Methods*, v. 8, p. 785–786, 2011.

ROUSSEEUW, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, Elsevier, v. 20, p. 53–65, 1987.

ŠALI, A.; BLUNDELL, T. L. Comparative protein modelling by satisfaction of spatial restraints. *Journal of molecular biology*, Elsevier, v. 234, n. 3, p. 779–815, 1993.

SIEW, N. et al. Maxsub: an automated measure for the assessment of protein structure prediction quality. *Bioinformatics*, Oxford University Press, v. 16, n. 9, p. 776–785, 2000.

TUSNáDY, G.; SIMON, I. The hmmmto transmembrane topology prediction server. *Bioinformatics*, v. 17, p. 849–850, 2001.

VERLI, H. Bioinformática: da biologia à flexibilidade molecular. Sociedade Brasileira de Bioquímica e Biologia Molecular, 2014.

WU, H. et al. Identify high-quality protein structural models by enhanced-means. *BioMed research international*, Hindawi, v. 2017, 2017.

WU, K.-L.; YANG, M.-S. Mean shift-based clustering. *Pattern Recognition*, Elsevier, v. 40, n. 11, p. 3035–3052, 2007.
ZHANG, Y.; SKOLNICK, J. Tm-align: a protein structure alignment algorithm based on the tm-score. *Nucleic acids research*, Oxford University Press, v. 33, n. 7, p. 2302–2309, 2005.