Vol.64: e21200118, 2021
https://doi.org/10.1590/1678-4324-2021200118
ISSN 1678-4324 Online Edition

Article - Engineering, Technology and Techniques

Reduced Distance Matrix to Verify the Similarity between Protein Structures

Otaviano Martins Monteiro[1*]
https://orcid.org/0000-0002-7736-0392

Sandro Renato Dias[2]
https://orcid.org/0000-0001-5288-5929

Thiago de Souza Rodrigues[2]
https://orcid.org/0000-0003-0041-8303

[1] Federal Center for Technological Education of Minas Gerais (CEFET-MG), Postgraduate Program in Mathematical and Computational Modeling (PPGMMC), Belo Horizonte, Minas Gerais, Brazil; [2] Federal Center for Technological Education of Minas Gerais (CEFET-MG), Department of Computing (DECOM), Belo Horizonte, Minas Gerais, Brazil.

Editor-in-Chief: Paulo Vitor Farago
Associate Editor: Andressa Novatski

Received: 2020.02.27; Accepted: 2021.05.27

*Correspondence: otavianomartins@hotmail.com (O.M.M.).

HIGHLIGHTS

- Development of a reduced distance matrix, approximately 70% less than [1].
- The reduced distance matrix has a shorter processing time than the algorithms [1-3].
- Several clusters were constructed with similar protein interactions.
- The reduced distance matrix obtained satisfactory accuracy.

Abstract: This paper focuses on developing a reduced distance matrix to improve the computational performance during the protein interactions clustering. This proposed matrix considers as centroids two alpha carbon atoms from a protein structure and stores the distances between these centroids and the other atoms from this same structure. Each row in this matrix represents a database record and each column is a distance value. Through this build matrix, clusters were performed using K-Means Clustering. The precision and performance of this presented technique were compared with aCSM, RID and another distance matrix methodology that considers the distances between all atoms from each protein structure. The results were satisfactory. The reduced distance matrix obtained a high precision and the best computational performance.

Keywords: protein interactions; reduced distance matrix; clustering.

INTRODUCTION

Disulfide bond interactions are covalent bonds formed by the interaction between sulfur atoms from cysteines, which after oxidations became cystines. These bonds are responsible for maintaining the conformational stability (three-dimensional structure) of a protein by linking distant parts of a polypeptide chain or different chains [4,5]. The Figure 1 illustrates a disulfide bond interaction.

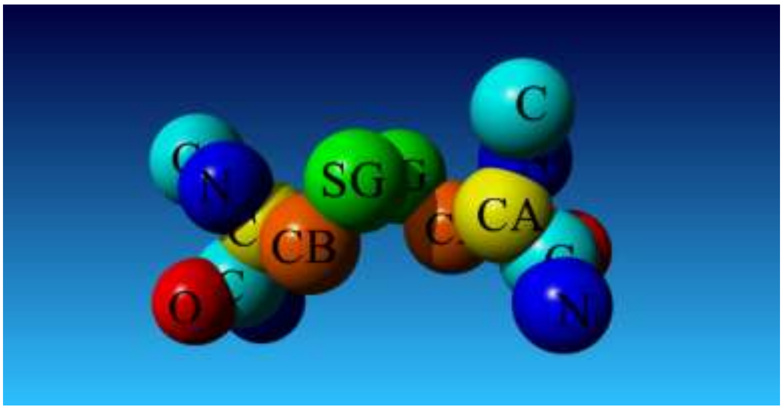


Figure 1. Disulfide bond formed between two cysteine molecules.

The Figure 1 illustrates the construction of the disulfide bond by linking two cysteines, through the Sulfur Gamma atoms (SG) of each cysteine, illustrated in green. After this union, these molecules are called cystines. This link stabilizes the main chain from each side, formed by the atoms Nitrogen (N) in dark blue, Carbon Alpha (CA) in yellow, Carbon (C) in light blue and Oxygen (O) in red. In this figure, each interacting side also contains an atom from the residue before the main chain (C) in light blue, an atom from the residue after the main chain (N) in dark blue and a beta carbon atom (CB) in orange.

These interactions are important for protein stability and conformation. Disulfide bonds, along with hydrogen bonds, ionic pairs, and van der Waals interactions, define the shape and keep the protein structure stable [5-7].

Disulfide bonds are also important in folding the polypeptide chain into specific secondary and tertiary structures. Additionally, their groupings with other polypeptides form the quaternary structures [5,8].

There are studies that prove the increased stability of a protein by introducing disulfide bonds in distinct regions of the same protein [5, 9, 10]. The study of two variants of the Subtilisin BPN' enzyme, shows that protein stability was increased by 1,000 times, from 10 mutation points, including the addition of a new disulfide bond [11]. In the literature we can also find studies that prove the protein stability decrease when removing disulfide bonds [12].

Several proteins structures, including disulfide bonds, were studied by various researchers and inserted in biological databases, such as the Protein Data Bank (PDB). This database stores through text files information about three-dimensional proteins structure. This information includes: atomic coordinates, functions, interactions and others important protein characteristics [13]. Many software, such as LSQKAB [14,15], work through PDB text files.

The LSQKAB is based on Kabsch algorithm [16], which calculates the optimal rotation matrix between two protein structures, minimizing the Root-Mean-Square Deviation (RMSD). However, using LSQKAB to calculate the similarity across all records in a database is time consuming [5].

One alternative to do this task in a viable computer time is using the RID tool. The first step is to choose one record to be the reference. Then, the LSQKAB is used to overlap all other records with the reference, generating a delta file with the result atomic distances for each overlapping file. Finally, a score is calculated for each file generated, to optimize the search [2,5]. This technique gets satisfactory but not optimal results when working with disulfide bonds.

There are techniques that work with protein structures by generating signatures for biological graphs. Some examples are Cutoff Scanning Matrix (CSM) [17] and one of its variations, atomic Cutoff Scanning (aCSM) [3]. These techniques generate signatures for biological graphs based on atoms distance pattern from a protein structure. The generated signatures are used in classification tasks [17].

In this paper, we studied the disulfide bonds interactions clustering, extracted from PDB. The goal in this work is to form groups with disulfide bonds records that have low Root-Mean-Square values when compared with each other. This study is important, because finding interactions of disulfide bonds with low values of distance differences, is possible for a specialist in the field to analyze and propose mutations between them, which can bring improvements to the structures of these macromolecules.

Protein studies benefit several segments, such as health, in the development of drugs and vaccines. In the industry, improving digestive enzymes that are used in various processes. In the environment, changing the enzymes for the degradation of contaminants, among other countless areas [5].

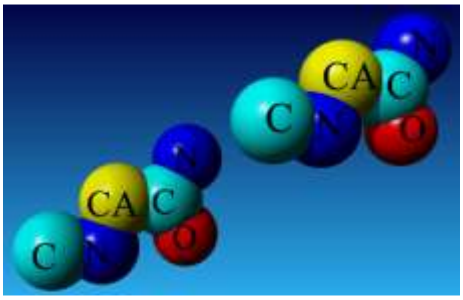
Inspired by one CSM steps, we developed methodologies based on matrix distances. First, we developed a complete distance matrix, that uses the Euclidean distance calculation between all atoms in an interaction file, to build a distance matrix [1]. In order to improve the computational performance, we developed a reduced distance matrix using two alpha carbon atoms from each interaction as centroids to reduce the amount of distances to be calculated. We use LSQKAB to compare the accuracy of clusters formed by the methodologies based on distance matrix, aCSM and RID. We also evaluate the processing time. In this way, the complete distance matrix obtained the best precision and satisfactory run time. The reduced distance matrix obtained the best processing time and satisfactory precision.

MATERIAL AND METHODS

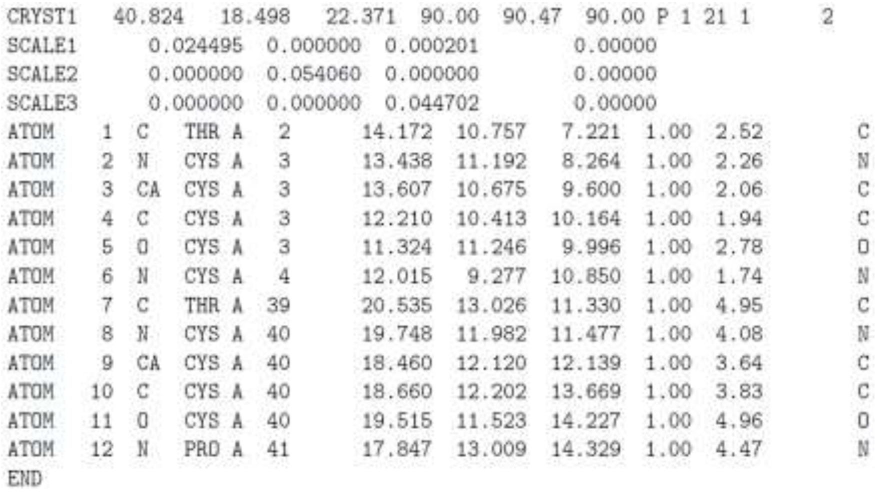
Disulfide Bond Interactions Data Type

The disulfide bonds evaluated in this work were extracted from the PDB in text file format by the RID tool. The information contained in these text files were obtained by x-ray diffraction, for the reason of obtaining a more precise coordinate for each atom. This choice was more viable than working with Nuclear Magnetic Resonance Spectroscopy (NMR), for the reason that NMR reports several coordinates of probabilities for each atom. This excess of information could hinder the process of building the distance matrices.

The RID tool obtained for each interacting side from each disulfide bond, the main chain, an atom from the residue before the main chain and an atom from the residue after the main chain. The RID tool ignored some atoms such as beta carbons, gamma sulfur and the side chain atoms. These ignored atoms were not inserted into the text files, called interaction files. Because the focus of this database is on maintaining the main chain, to conserve the protein's function in a mutation event. Consequently, these ignored atoms are not necessary for the focus of this database. The Figure 2 (b) shows the structure of “1ejg_CYS-3-A_CYS-40-A_mc6.pdb”, an interaction file. The Figure 2 (a) shows the three-dimensional view of this file.



(a)



```
CRYST1   40.824   18.498   22.371  90.00  90.47  90.00 P 1 21 1      2
SCALE1      0.024495  0.000000  0.000201        0.00000
SCALE2      0.000000  0.054060  0.000000        0.00000
SCALE3      0.000000  0.000000  0.044702        0.00000
ATOM      1  C   THR A   2      14.172  10.757   7.221  1.00  2.52           C
ATOM      2  N   CYS A   3      13.438  11.192   8.264  1.00  2.26           N
ATOM      3  CA  CYS A   3      13.607  10.675   9.600  1.00  2.06           C
ATOM      4  C   CYS A   3      12.210  10.413  10.164  1.00  1.94           C
ATOM      5  O   CYS A   3      11.324  11.246   9.996  1.00  2.78           O
ATOM      6  N   CYS A   4      12.015   9.277  10.850  1.00  1.74           N
ATOM      7  C   THR A  39      20.535  13.026  11.330  1.00  4.95           C
ATOM      8  N   CYS A  40      19.748  11.982  11.477  1.00  4.08           N
ATOM      9  CA  CYS A  40      18.460  12.120  12.139  1.00  3.64           C
ATOM     10  C   CYS A  40      18.660  12.202  13.669  1.00  3.83           C
ATOM     11  O   CYS A  40      19.515  11.523  14.227  1.00  4.96           O
ATOM     12  N   PRO A  41      17.847  13.009  14.329  1.00  4.47           N
END
```

(b)

Figure 2. This figure illustrates the structure of an interaction file, through two panels. They are respectively: (a) Three-dimensional visualization of this structure, by [18]; (b) Information contained in the interaction file.

The Figure 2 (a) illustrates the two interacting sides that form the interaction file. The leftmost Carbon (C) is an atom from the residue before the main chain of the left side. The next atoms from this side form this main chain. They are: Nitrogen (N), Alpha Carbon (CA), Carbon (C) and Oxygen (O). The last Nitrogen (N) from the left side is an atom from the residue after the main chain. The right side from the interacting file also contains 6 atoms, which follow the same pattern as the left side. In the Figure 2 (b), the first line refers to the crystallographic cell used to extract various information from the atomic structure, being identified by CRYST1. The lines 2, 3, and 4, indicated by SCALEn show the operators for transforming orthogonal coordinates into crystallographic cell coordinates. Below these first 4 lines, the coordinate section starts. This section is one of the most important in the file. Through information contained in [13,19], it was possible to construct the table 1 to explain in detail the items in this section.

Table 1. Coordinate section [13,19].

Columns	Data type	Field	Definition
1 – 6	Record name	ATOM	
7 – 11	Integer	Serial	Atom serial number
12 – 16	Atom	Name	Atom name
17	Character	AltLoc	Alternate location indication
18 – 20	Residue name	ResName	Residue name
22	Character	ChainID	Chain identify
23 – 26	Integer	ResSec	Residue sequence number
27	AChar	iCode	Code insertion of residues
31 – 38	Float (8,3)	X	Orthogonal coordinate for X
39 – 46	Float (6,2)	Y	Orthogonal coordinate for Y
47 – 54	Float (6,2)	Z	Orthogonal coordinate for Z
55 – 60	Float (6,2)	Occupancy	Occupancy
61 – 66	Float (6,2)	TempFactor	Temperature factor
77 – 78	LString (2)	Element	Element symbol right- justified
79 – 80	LString (2)	Charge	Charge on the Atom

As indicated by the Table 1, the first string in the coordinate section from the Figure 2 (b) refers to the record name, contained in that row. The second string indicates the atom's serial number. The third string shows the atom name, where carbon (C), nitrogen (N), alpha carbon (CA) and oxygen (O). The atoms 1 through 6 refer to the left side from the interacting file. The atoms 7 to 12 refer to the right side. The fourth string refers to the Alternate location indication. The next four strings show the residue name, chain identify, residue sequence number and code insertion of residues respectively. The strings 9, 10 and 11 indicates the atom coordinates in Å for the x, y and z axes. The next 4 strings respectively refer to the probability of the atom being at that location, temperature factor, right-aligned element symbol, and charge. The file "1ejg_CYS-3-A_CYS-40-A_mc6.pdb", illustrated in the Figure 2 B does not contains "alternate location indication", "code insertion", and "charge".

Experiments

To verify the efficiency of the reduced distance matrix, the experiments were performed with 16,383 disulfide bond interaction files. This is the same database used in the complete distance matrix experiments [1]. The results obtained were compared with the complete distance matrix, aCSM and RID.

We decided to compare the results with the complete distance matrix [1], because the software presented in this paper originated from this technique. The purpose of this comparison is to verify how much the reduced distance matrix has improved performance time and check if occurred loss in accuracy.

The aCSM was chosen because this software is one CSM version that considers all atoms from the protein structure to calculate the pattern of distances. The different versions from the CSM are considered the state of the art in generating biological graph patterns.

The RID tool was selected because it is a software specialized in working with protein interactions.

Conducting the Experiments

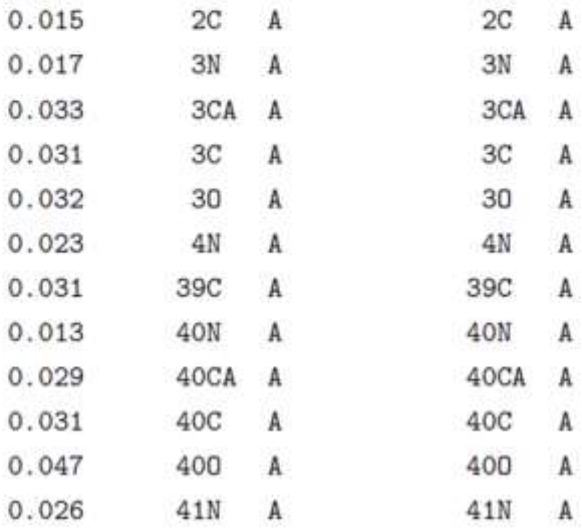
All methodologies were evaluated by clustering through K-Means Clustering, defining 500, 750 and 1,000 groups. There are other well-known clustering techniques, such as DBSCAN that groups the database records according to their density, another example is Hierarchical Clustering, which builds a hierarchy of clusters according the similarity of records, among other techniques. But in this article, we chose K-Means Clustering because this technique has a satisfactory processing time, works with all database records without exclusions and uses Euclidean distance as metric. The distance matrices described in this article, aCSM and RID also use Euclidean distance in at least one of its stages.

K-Means clustering is an algorithm that divides data into k groups, based on the distance of the centroid from each cluster. In this case, centroid is the point at which the sum of distances of all elements from this cluster is the smallest possible. K-means uses an iterative algorithm that minimizes the sum of distances of each item to the centroid of its group. This clustering algorithm moves objects between clusters, until the sum of distances reaches the lowest possible value [20].

The accuracy of the formed clusters was verified using the LSQKAB overlaps results. Atomic overlaps were performed between all records that were in the same group. Each file was compared twice to all other files in his cluster. In the first comparison, one of these records remained "fixed" and the other was fit to the same. In the second comparison, the opposite occurred between these files. A delta file was generated for each comparison.

Each delta file generated contains the information of the distance differences between each overlapping atom pair. The tolerance value set was up to 0.5 Å for each overlapping amino acid pair. This guarantees a maximum RMS value of 0.5 Å and consequently a maximum RMSD of up to 0.5 Å. This value ensures strong similarity between the compared files, resulting in a satisfactory overlap between these records [21]. The RMSD is the most common way to evaluate similarities between protein structures [22].

The delta file structure is represented by Figure 3. Through the file “1ejg_CYS-3-A_CYS-40-A_mc6.pdb-1cbn_CYS-3-A_CYS-40-A_mc6.pdb.deltas ”, one of the generated files.



```
0.015        2C   A         2C   A
0.017        3N   A         3N   A
0.033        3CA  A         3CA  A
0.031        3C   A         3C   A
0.032        3O   A         3O   A
0.023        4N   A         4N   A
0.031       39C   A        39C   A
0.013       40N   A        40N   A
0.029       40CA  A        40CA  A
0.031       40C   A        40C   A
0.047       40O   A        40O   A
0.026       41N   A        41N   A
```

Figure 3. Delta file “1ejg_CYS-3-A_CYS-40-A_mc6.pdb-1cbn_CYS-3-A_CYS-40-A_mc6.pdb.deltas” [23].

The delta file name is given by joining the names of the records that formed it, separated by “-”. In this example, the delta file was formed by the fixed file “1ejg_CYS-3-A_CYS-40-A_mc6.pdb”, and the record “1cbn_CYS-3-A_CYS-40-A_mc6.pdb”. The first column refers to the differences value in atomic distances between these files. The next two columns refer to the first file information and the last two refer to the second file. This information is protein residue number, atom and chain identification [5].

The most important information for each delta file is its first column. It represents the difference in distances between atoms in each record relative to the same atoms in the other compared structure [23].

In this work was evaluated the accuracy and the time spent by each algorithm for constructing records and performing clustering.

Parameters Used to Adapt Techniques to Experiments

The objective of this subsection is to indicate the chosen parameters values and the necessary adjustments made in the evaluated algorithms, aiming at the accomplishment of the experiments.

The development of the reduced distance matrix is detailed in the development subsection. The methodology that uses the complete distance matrix did not require parameter adaptations for the experiments. The aCSM and RID techniques required adaptations.

The aCSM algorithm generates the distance pattern from a protein according to its frequency of atoms at certain cutoff distances. To obtain this frequency of atoms is necessary to pass as parameters the minimum and maximum distances to be accounted in the frequency and the increment step. To obtain these parameters, we first calculated the largest Euclidean distance between atoms in the same database record. A value of 11.22 Å was found. This value was considered as the maximum distance parameter. According to experiments made by [17], was defined 151 columns to represent the frequency of atoms in each record. We obtained the step value from the division between the maximum distance value and the number of columns.

That way, the step value was 0.074 Å. The minimum distance started with 0, incrementing with the step value. After running aCSM with these parameters, the distance patterns of all records were obtained based on the frequency of their respective atoms. The distance patterns obtained were used to perform the grouping with k-means clustering.

The first step in the RID technique is to choose a database file to be used as a reference. The criterion considered is the resolution value contained in the PDB file. Thus, the interaction file chosen was the "1ejg_CYS-3-A_CYS-40-A_mc6.pdb". After this choice, the LSQKAB software was used to perform atomic overlaps of all other 16,382 interaction files against the chosen reference. As a result, 16,382 delta files were generated. These values were imported by Matlab and grouped by the k-means Clustering.

Complete Distance Matrix

Inspired by one CSM stage, a methodology based on distance matrix was developed. This methodology allows the comparison and clustering of interactions extracted from the PDB, through the atomic distance similarities, in an efficient way. However, the distance matrix cited was constructed considering the distance differences between all atoms from each interaction [1]. The flowchart of this methodology is illustrated by the Figure 4.

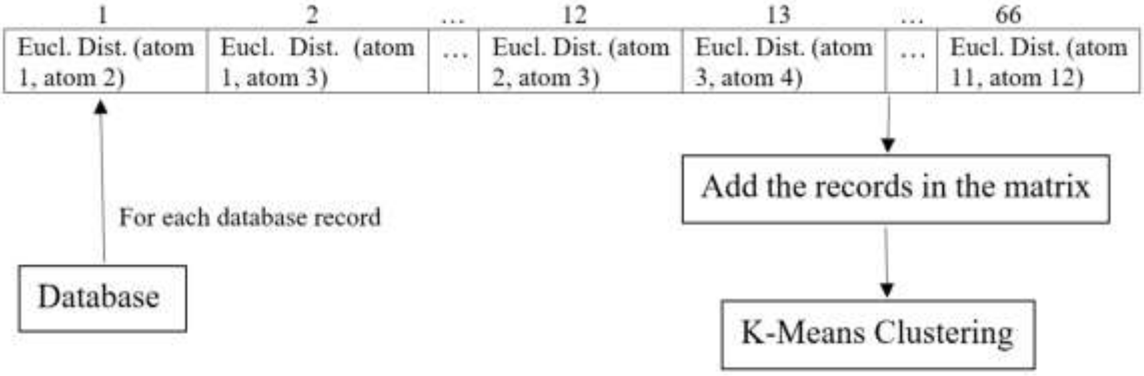

1
2
...
12
13
...
66
Eucl. Dist. (atom 1, atom 2)
Eucl. Dist. (atom 1, atom 3)
...
Eucl. Dist. (atom 2, atom 3)
Eucl. Dist. (atom 3, atom 4)
...
Eucl. Dist. (atom 11, atom 12)
For each database record
Database
Add the records in the matrix
K-Means Clustering

Figure 4. Complete distance matrix flowchart [1].

According to Figure 4, for each interaction file contained in the database, the Euclidean distance was calculated between all 12 atoms from each record, considering the x, y and z coordinates. The results of each record were inserted into a vector of 66 positions. The first position refers to the Euclidian distance between x, y and z coordinates of atom 1 (leftmost Carbon) to atom 2 (Nitrogen belonging to the main chain from the left side). The vector started at position 1, because Matlab starts the vectors from this position. The position 2 from the vector refers to Euclidean distance between atoms 1 (leftmost Carbon) and 3 (Alpha Carbon from the left side). And so on, until the position 66 refers to the calculation between atoms 11 and 12 (Oxygen and Nitrogen on the right side). Subsequently, these vectors were joined in a matrix. Due to the fact that the studied files are the same size, the constructed matrix had 16,383 rows and 66 columns. Each row represents a record and each column indicates a Euclidean distance value [1]. The table 2 explains in more detail the 66 positions.

Table 2. Definitions of the positions from the vector from complete distance matrix.

Positions	Definition
1 – 11	Euclidian distance between the leftmost Carbon atom and the other 11 atoms
12 – 21	Euclidean distance between the Nitrogen and the other atoms, except the leftmost Carbon
22 – 30	Euclidean distance between the left CA and the other atoms, except the first two atoms
31 – 38	Euclidean distance between the Carbon and the other atoms, except the first three atoms
39 – 45	Euclidean distance between the Oxygen and the other atoms, except the first four atoms
46 – 51	Euclidean distance between the Nitrogen and the other atoms, except the first five atoms
52 – 56	Euclidean distance between the first Carbon in the right side and other atoms on the right
57 – 60	Euclidean distance between the Nitrogen in the right side and the other 4 remaining atoms
61 – 63	Euclidean distance between the right CA and the other 3 remaining atoms
64 – 65	Euclidean distance between the Carbon and the others 2 remaining atoms
66	Euclidean distance between the last 2 atoms (Nitrogen and Oxygen)

According Table 2, positions 1 through 11 refer to the Euclidean distance calculation between the leftmost Carbon and the other 11 atoms contained in the interaction file. Positions 12 to 21 refer to the calculation between the Nitrogen belonging to the main chain on the left side and the other atoms, except the leftmost Carbon. That is, the Euclidean distance between Nitrogen and the leftmost carbon is not performed in this time because this calculation has already been done in position 1. Following this idea, the next positions of the vector refer to the Euclidean distance between the other atoms in the interaction file, without repeating the calculation between the same atoms.

Development of the Reduced Distance Matrix

The reduced distance matrix was inspired by the complete distance matrix. While the complete distance matrix depends on the calculation of the Euclidean distance between all atoms in the same interaction file, the reduced distance matrix depends only on the Euclidean distance calculations between the centroids (alpha carbon atoms) in relation to the other atoms. The Figure 5 illustrates the positions of Alpha Carbon atoms (CA) on both sides of the interacting files.

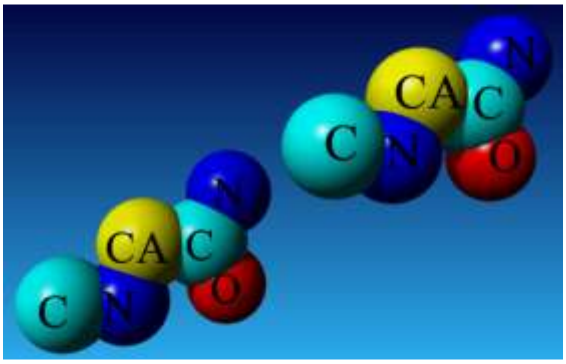


Figure 5. The alpha carbon atoms, in yellow color, were chosen to be the centroids.

The alpha carbon atoms were chosen to be the centroids because they are in the center of the main chain. Additionally, they are attached to the side chain.

The function of the reduced distance matrix consists of calculating the Euclidean distance for the x, y and z coordinates between the first centroid atom and the other 10 non-centroid atoms. Then, the Euclidean distance between the second centroid atom and the other 10 non-centroid atoms is calculated. Thus, 20 values were obtained to represent each interaction file. Then a matrix with 16,383 rows and 20 columns was constructed. Each row represents a record and each column indicates a distance value. Thus, the reduced distance matrix works with less than one third of the quantity of columns used in the complete distance matrix, which uses 66 columns to represent each interaction file. The flow chart of the reduced distance matrix is illustrated by Figure 6.

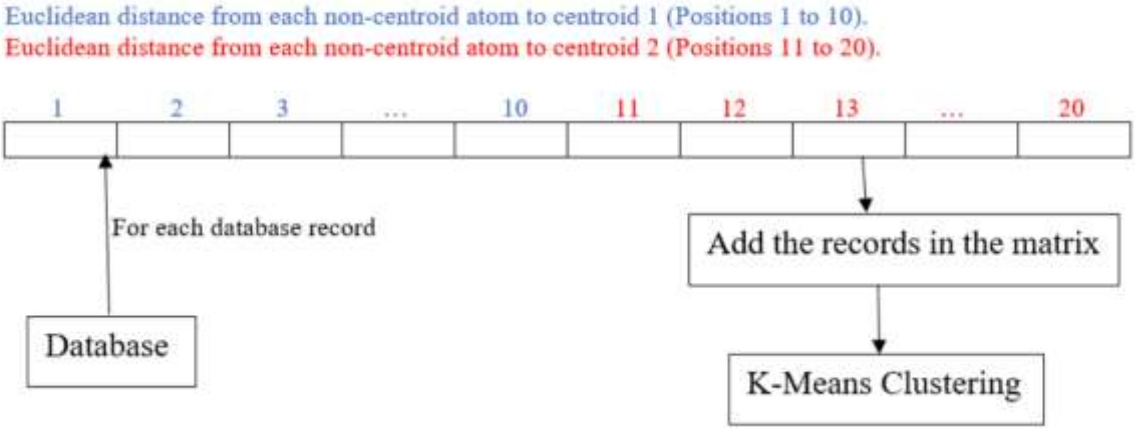
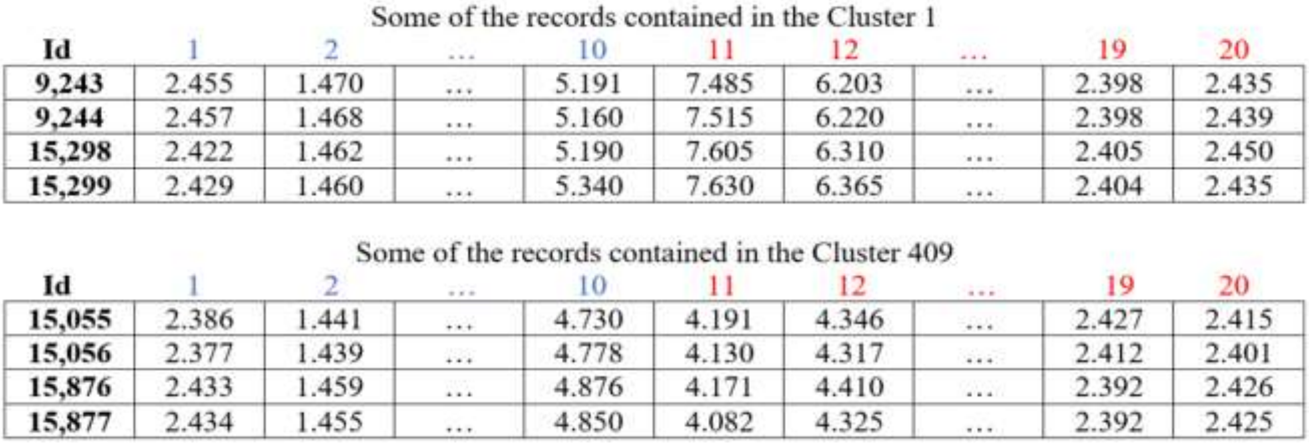

Euclidean distance from each non-centroid atom to centroid 1 (Positions 1 to 10).
Euclidean distance from each non-centroid atom to centroid 2 (Positions 11 to 20).
1
2
3
...
10
11
12
13
...
20
For each database record
Add the records in the matrix
Database
K-Means Clustering

Figure 6. Reduced distance matrix flow chart.

According to Figure 6, each interaction file contained in the database was represented by a 20 positions vector. The first position from the vector refers to the calculation of the Euclidean distance between the coordinates x, y e z from the first centroid atom (alpha carbon on the left side of the interacting file) and the first non-centroid atom (leftmost Carbon). The positions 2 to 10 refer to the calculation of the Euclidean distance between the first centroid atom and the other 10 non-centroid atoms. The positions 10 to 20 refer to the calculation of the Euclidean distance between the second centroid atom (alpha carbon on the right side of the interacting file) and the other 10 non-centroid atoms contained in the same record. Later, these vectors were joined in a matrix, which was grouped by the K-means Clustering technique, in the Matlab software [20]. The Figure 7 illustrates the values contained in some positions of the reduced distance matrix, for the clusters 1 and 409.



Some of the records contained in the Cluster 1

Id	1	2	...	10	11	12	...	19	20
9,243	2.455	1.470	...	5.191	7.485	6.203	...	2.398	2.435
9,244	2.457	1.468	...	5.160	7.515	6.220	...	2.398	2.439
15,298	2.422	1.462	...	5.190	7.605	6.310	...	2.405	2.450
15,299	2.429	1.460	...	5.340	7.630	6.365	...	2.404	2.435

Some of the records contained in the Cluster 409

Id	1	2	...	10	11	12	...	19	20
15,055	2.386	1.441	...	4.730	4.191	4.346	...	2.427	2.415
15,056	2.377	1.439	...	4.778	4.130	4.317	...	2.412	2.401
15,876	2.433	1.459	...	4.876	4.171	4.410	...	2.392	2.426
15,877	2.434	1.455	...	4.850	4.082	4.325	...	2.392	2.425

Figure 7. Examples of some records contained in cluster 1 and 409.

In the Figure 7, the starting positions values (1 and 2), and the ending positions values (19 and 20) were slightly different between the records from one cluster to the records from the other cluster. On the other hand, the central positions (10, 11 and 12) were very different for records from distinct clusters. The differences across the 20 positions of each record contributed to the K-Means in the group formations.

RESULTS AND DISCUSSION

The Table 3 indicates the number of satisfactory overlaps obtained by the reduced distance matrix and the complete distance matrix for experiments considering 500, 750 and 1,000 clusters. These values refer to the average after 31 executions.

Table 3. Number of records in clusters according to satisfactory overlap rates.

Technique	Reduced Distance Matrix			Complete Distance Matrix		
Quantity of clusters	**500**	**750**	**1,000**	**500**	**750**	**1,000**
0% to 10%	486	338	318	442	345	281
10% to 20%	1,727	1,249	968	1,473	1,069	763
20% to 30%	1,432	1,303	1,105	1,421	1,033	962
30% to 40%	1,319	1,171	1,162	1,114	1,299	1,084
40% to 50%	1,225	1,041	947	1,051	912	890
50% to 60%	1,341	1,166	1,009	1,389	1,023	904
60% to 70%	1,368	1,202	1,110	1,311	1,017	1,070
70% to 80%	1,880	1,821	1,579	1,058	1,319	969
80% to 90%	2,215	2,199	2,088	2,209	1,582	1,382
90% to 100%	3,390	4,893	6,097	4,915	6,784	8,078

According to the table 3, the reduced distance matrix formed several clusters, with many records, which had satisfactory overlap rates over than 90%. The best results were with the 1,000 clusters experiments, which obtained 6,097 records in this range of hits. In this same scenario, the complete distance matrix got even better results, averaging 8,078 records. On the other hand, the reduced distance matrix obtained more records in clusters that obtained satisfactory overlap rates between 80% and 90%, 70% and 80% and still 60% and 70%, when compared to the complete distance matrix.

For experiments with 500 and 750 clusters, although both methodologies had lower percentages of satisfactory overlaps, the behavior of the results was similar. The complete distance matrix had more records that were in clusters with a satisfactory overlap rate greater than 90%. The reduced distance matrix had more records that remained in clusters that had satisfactory overlap rates between 60% until 70%, 70% and 80%, and 80% until 90%, when compared to the complete distance matrix.

The Figure 8 illustrates the number of records in clusters with different ranges of satisfactory overlaps rates. In this experiment, we considered the reduced distance matrix, the complete distance matrix, beyond the results obtained by RID and aCSM for this same database. The results of these last two algorithms have been described in [1]. This figure refers to the results with 1,000 clusters, considering the average of 31 runs.

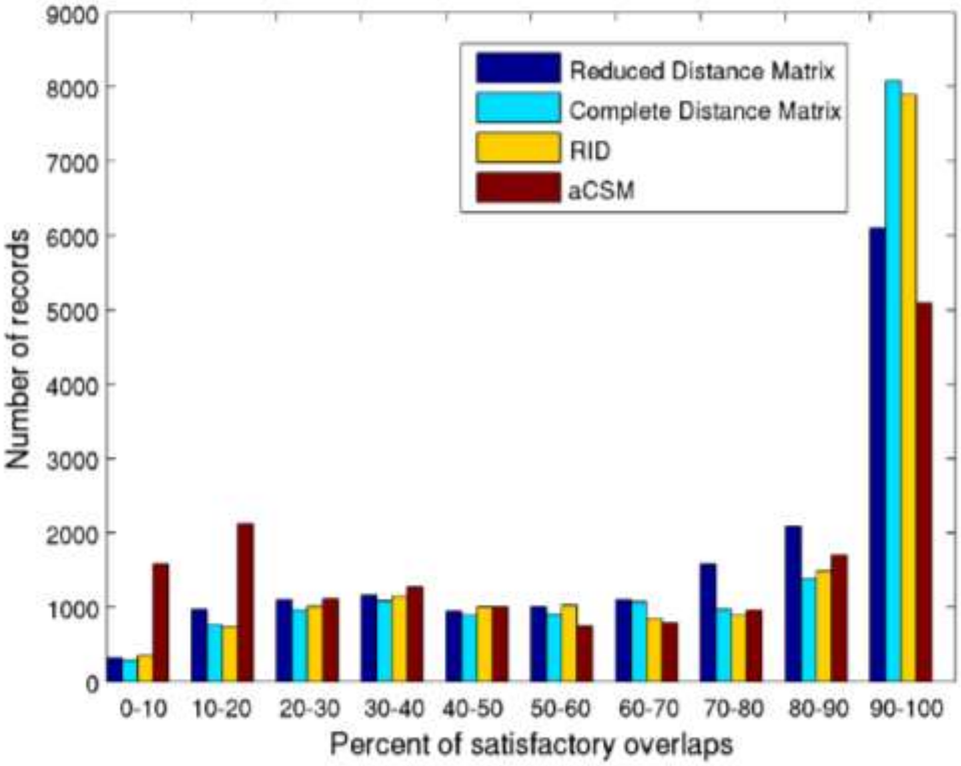


Figure 8. Number of records per cluster according to satisfactory overlap rate.

According to Figure 8, the number of records that the reduced distance matrix obtained in clusters with more than 90% of satisfactory overlaps was higher than the aCSM algorithm, but this amount was less than the complete distance matrix and RID. However, the reduced distance matrix obtained more records than the others three algorithms in clusters with the following percentage ranges 80% until 90%, 70% until 80% and 60% until 70%.

Analyzing the number of records in clusters with low satisfactory overlap rates, we found that the reduced distance matrix obtained fewer records than aCSM in the two worst ranges (0% to 10% and 10% to 20%), obtaining quantities close to the complete distance matrix and RID.

The results for the experiments defining the maximum number of clusters in 500 and 750 groups, followed similar behaviors although all methodologies obtained lower percentages of satisfactory overlaps.

The Figure 9 refers to a boxplot with the total average of satisfactory overlaps. This boxplot evaluated the reduced distance matrix, complete distance matrix, aCSM and RID. The displayed values refer to the average after 31 runs for each methodology, considering 1,000 clusters.

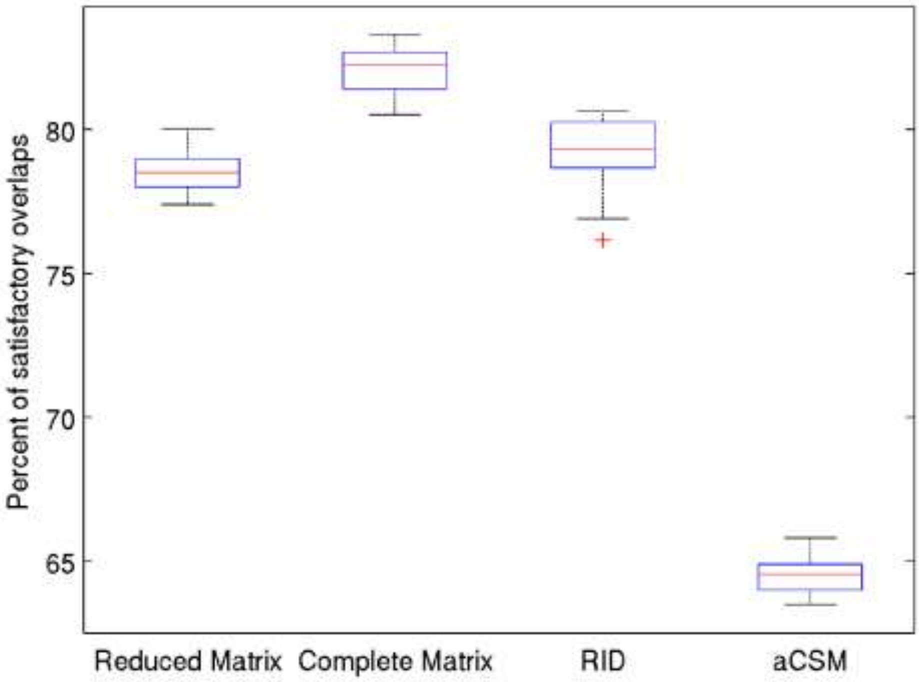


Figure 9. Boxplot comparing the amount of satisfactory overlap of the four techniques.

According to the boxplot, the results of the reduced distance matrix were satisfactory, being considerably higher than the aCSM. The presented technique also obtained a precision close to the RID and the complete distance matrix.

The reduced distance matrix presented a smaller variation of results than the RID. All results of the presented methodology were between 77% and 80% of satisfactory overlaps. The RID algorithm had some executions that got more than 80.5%, but on the other hand, some executions had values close to 76.5%, for example, the outlier below the last quartile in the Figure 10. The medians of the reduced distance matrix and the RID had close values (central red lines).

The best results of the reduced distance matrix were very close to the worst results of the complete distance matrix. This difference was less than 0.5%. On the other hand, the best results of the complete distance matrix were next to 83,30% of satisfactory overlaps. This value is superior than the best results from all other compared techniques.

The accuracy results obtained by the reduced distance matrix were satisfactory. Even using a matrix 70% smaller than the complete distance matrix, the average of satisfactory overlaps of both methodologies were close. A smaller distance matrix takes less time to be processed by clustering algorithms. Obtaining a very short distance matrix and still with a satisfactory percentage of overlaps accepted is a satisfactory result.

After the boxplot, can be used the analysis of variance (ANOVA) to compare the distribution of the 4 groups. ANOVA indicates whether there are statistical differences between the results obtained by each technique. This analysis can be used when the residuals comply with the following assumptions: normality, independence, and equality [24]. In statistics, the residuals are the differences between the obtained and expected values. The residues indicate the natural variation of the data [25]. The verification of the normality assumption is illustrated in Figure 10.

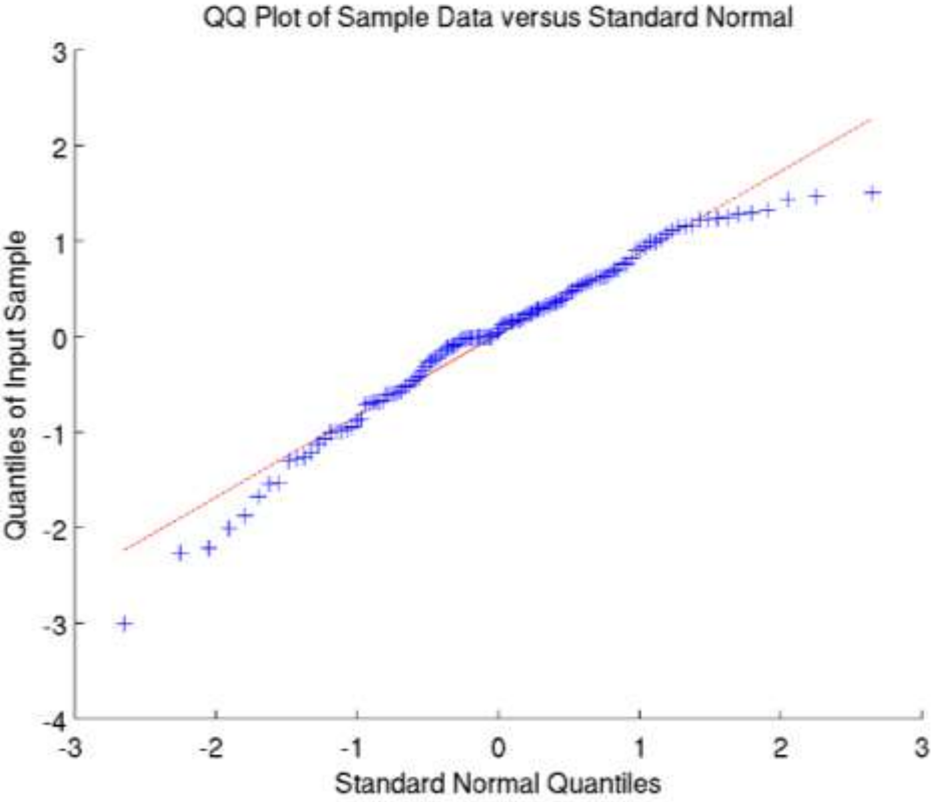


Figure 10. Assumption of normality.

The validation of normality can be obtained by the graph of normal probability of the residues, in the Figure 10. In this figure, each residue is compared with its expected value (considering that they follow a normal distribution). In other words, the residues (blue points in the graph) should be close to the expected values (red line). Small variations, as in Figure 10 are tolerable. The Figure 11 shows the validation of the independence assumption.

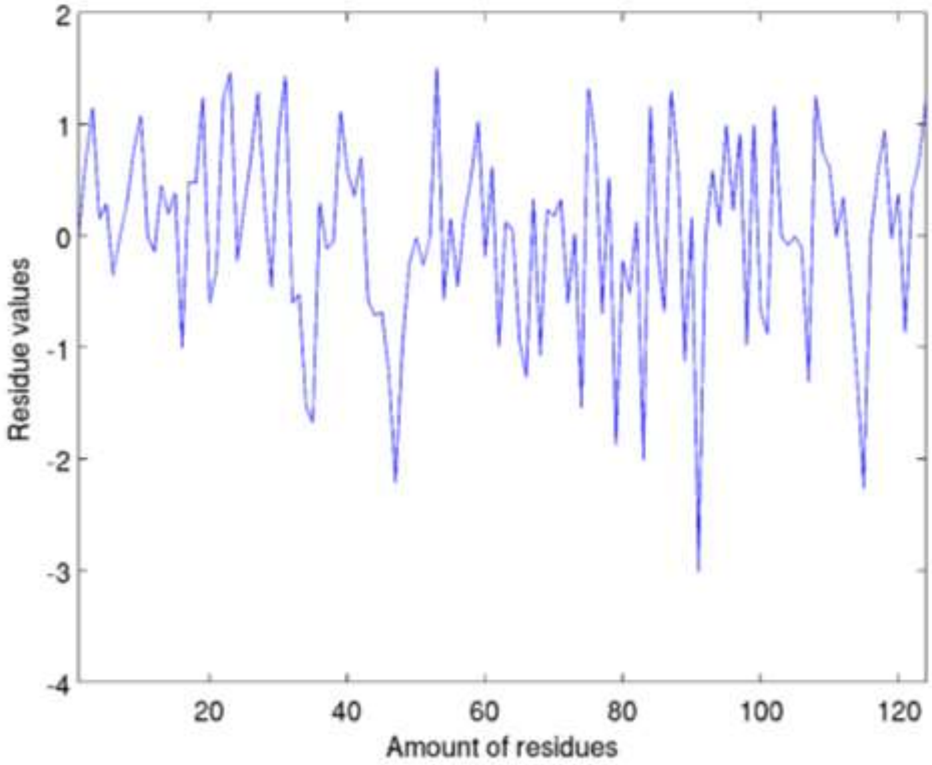


Figure 11. Assumption of independence.

The independence premise must ensure that one observation does not influence the other. In other words, the residues cannot be predictable. Thus, the results contained in Figure 11 indicate that this assumption is complied, because the residue values do not have a defined pattern, making it impossible to predict the next residual value.

The last premise to be assessed is the equality assumption. It is illustrated in the Figure 12.

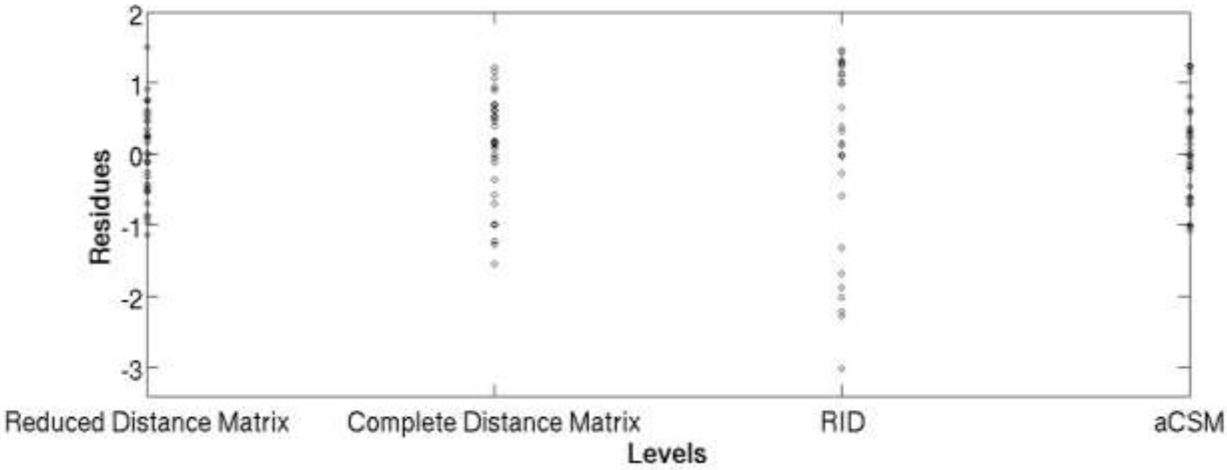


Figure 12. Assumption of equality.

To comply with the equality assumption, each group have to contain the residues equally distributed. That is, this assumption is fulfilled when each group has the same amount of residues above and below the value 0, on the y-axis of Figure 12. Small variations are tolerable. According the Figure 12, residues too complies the equality assumption. In this way, all three premises are fulfilled, enabling the use of ANOVA.

We use the One-way ANOVA from Matlab [26]. The p-value obtained was 2,44463^{-107}, a very small amount, indicating that at least one of the evaluated techniques has different accuracy from the others. Through the results from ANOVA is possible to perform the Tukey test to verify with greater precision which of the evaluated techniques obtained results statistically different from the others [26].

The Figure 13 refers to the Tukey test to verify which group have a considerable statistical difference between the total average of accepted overlaps for the reduced distance matrix in relation to the three other techniques. This analysis was made using the results obtained from the experiments of 1,000 clusters.

In this analysis, the dashed lines refer to the results of the reduced distance matrix. If these lines touch the results of another technique, there is no significant difference between these two techniques. The confidence level considered was 99.

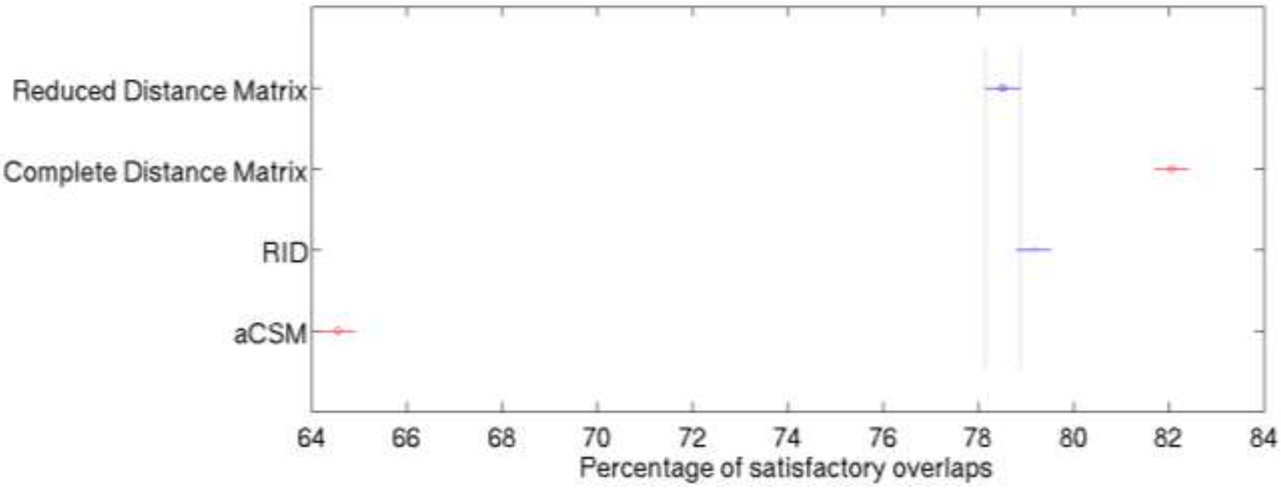


Figure 13. Statistical analysis comparing the results of the reduced distance matrix with the other three techniques evaluated, in the 1,000 clusters experiments.

According to the Tukey's test, illustrated by the Figure 13, the reduced distance matrix obtained better precision results than the aCSM. This difference was significantly considerable because aCSM obtained values between 64% and 66%, while the results of the reduced distance matrix were between 78% and 80% of satisfactory overlaps. When comparing the results of the methodology presented with the RID, the dashed lines indicate no significant statistical difference for both techniques, according to the defined parameters. That is, the average of satisfactory overlaps of the reduced distance matrix was close to the RID. The complete distance matrix presented statistical superiority in relation to all other compared techniques.

Time Consuming

The methodology based on complete distance matrix had the shortest total time to perform the experiments, when compared with aCSM and RID [1]. In this section, we compared the time taken by the reduce distance matrix and the complete distance matrix to construct the data structure and group it by K-means clustering. The results obtained are in Table 4.

Table 4. Time consuming.

Technique	Reduced Distance Matrix			Complete Distance Matrix		
Quantity of clusters	**500**	**750**	**1,000**	**500**	**750**	**1,000**
Build the matrix	00:02	00:02	00:02	00:10	00:10	00:10
Clustering	00:20	00:29	00:49	00:40	01:09	01:23
Total time	00:22	00:31	00:51	00:50	01:19	01:33

This analysis of the execution time highlights the main contribution of the reduced distance matrix. The Table 4 contains the time consumption in seconds of the reduced distance matrix and complete distance matrix for constructing data structures and grouping, considering 500, 750, and 1000 clusters. According this table, the reduced distance matrix was almost 5 times faster in the step of building the matrix to be grouped, when compared to the time taken to build the complete distance matrix. This considerable difference occurred due to the reduction of the amount of calculations required to construct the matrix.

The time to group the reduced distance matrix by k-means clustering was also considerably shorter in the 500, 750 and 1000 cluster experiments. This is because the reduced distance matrix works with less than one third of the number of columns in the complete distance matrix, so it was less difficult for the k-means to form groups.

Thus, the total time spent to do the whole clustering process was many shorter by the methodology presented in this article.

Complexity of Algorithm

The reduced distance matrix has the computational complexity to transform an interaction file in a vector form into:

$$O(n\ c - c\ c), \tag{1}$$

Where “n” is the number of atoms contained in the interaction file and “c” is the amount of defined centroids. This result was better than the complete distance matrix, which has the complexity of:

$$O((n^2 - n) / 2), \tag{2}$$

to perform this same procedure.

As a result, the reduced distance matrix has the least computational complexity of all the compared techniques. Because according to [23], the complete distance matrix presented a lower computational complexity when compared to aCSM and RID. Thus, the reduced distance matrix is the fastest algorithm among the techniques compared in this article.

CONCLUSION

The results obtained in this work indicate that it is possible through the reduced distance matrix, to form several groups of interaction files that result in good atomic overlaps when compared to the other records in the same cluster. The precision of the reduced distance matrix was higher than aCSM and there was no statistical difference when comparing the results with the RID. On the other hand, the complete distance matrix obtained a higher precision when compared to the other techniques.

The major contribution from the reduced distance matrix is her performance. Its execution time is considerably faster than the other compared algorithms, because this technique has a lower computational complexity. The biggest highlight of the performance time refers to the construction of the data, in which the presented technique was almost 5 times faster than the second placed one.

Thus, the technique presented in this paper justifies being a good option to compare and group protein interaction files.

Acknowledgments: we are grateful to the Federal Center for Technological Education of Minas Gerais (CEFET-MG) for the support offered and for the doctoral scholarship used at the beginning of this research. We are also grateful to the Coordination of Superior Level Staff Improvement (CAPES) for the current doctoral scholarship.

REFERENCES

1. Monteiro OM, Dias SR, Rodrigues TS. [Development of a Methodology Based on a Distance Matrix for Verifying Protein Similarities]. Proceeding Series of the Brazilian Society of Computational and Applied Mathematics. 2020; 7(1): 0103691-7. doi: 10.5540/03.2020.007.01.0369.
2. Dias SR, Garrat RC, Nagem RAP. Proposition of Site-Directed Mutagenesis of Proteins Based on a Residue-Residue Interaction Database. The Brazilian Society for Biochemistry and Molecular Biology [Internet]. 2011 Apr 30. [cited 2019 Nov 29]; 3(1). Available from: http://www.sbbq.org.br/reuniao/cdrom/ra2011/resumos/R8102.pdf
3. Pires DEV, Minardi RCM, Santos MA, Silveira CH, Campos FF, Meira W. aCSM: noise-free graph-based signatures to large-scale receptor-based ligand prediction. Bioinformatics. 2013 Apr 01;29(7):855-61. doi: 10.1093/bioinformatics/btt058. PubMed PMID: 23396119.
4. Hunter L. Molecular biology for computer scientists. In: Hunter L, editor. Artificial intelligence and Molecular Biology. California: American Association for Artificial Intelligence; 1993. p. 1-46.
5. Dias SR. Residue Interaction Database – [Proposition of site-directed mutations based on interactions observed in proteins of known three-dimensional structure] [PhD Thesis]. Universidade Federal de Minas Gerais; 2012.
6. Gromiha MM, Selvaraj S. Inter-residue interactions in protein folding and stability. Progress in Biophysics and Molecular Biology. 2004 Oct; 86(02): 235-277. doi: 10.1016/j.pbiomolbio.2003.09.003. PubMed PMID: 15288760.
7. Pace CN, Fu H, Fryar KL, Landua J, Trevino SR, Shirley BA, et al. Contribution of Hydrophobic Interactions to Protein Stability. J Mol Biol. 2011 Mar 03. 408(3): 514-528. doi: https://doi.org/10.1016/j.jmb.2011.02.053. PubMed PMID: 21377472; PubMed Central PMCID: PMC3086625.
8. Lehninger AL, Nelson DL, Cox MM. [Principles of Biochemistry]. 4th ed. Sarvier; 2007.
9. Pantoliano MW, Whitlow M, Wood JF, Hardman KD, Rollence ML, Bryan PN. Large increases in general stability for Subtilisin BPN' through incremental changes in the free energy of unfolding. Biochemistry. 1989 Sep 5. 28(18):7205-13. doi: 10.1021/bi00444a012. PubMed PMID: 2684274.
10. Pantoliano MW, Whitlow M, Wood JF, Rollence ML, Finzel BC, Gilliland GL, et al. The engineering of binding affinity at metal ion binding sites for the stabilization of proteins: Subtilisin as a test case. Biochemistry. 1988 Nov 01. 27(22):8311-7. doi: 10.1021/bi00422a004. PubMed PMID: 3072018.
11. Almog O, Gallagher DT, Ladner JE, Strausberg S, Alexander P, Bryan P, et al. Structural basis of thermostability analysis of stabilizing mutations in Subtilisin BPN'. J Biol Chem. 2002 May 13. 277(30):27553-8. doi: 10.1074/jbc.M111777200. PubMed PMID: 12011071.
12. Sakaguchi M, Takezawa M, Nakazawa R, Nozawa K, Kusakawa T, Nagasawa T, et al. Role of Disulphide Bonds in a Thermophilic Serine Protease Aqualysin I from Thermus aquaticus YT-1. J. Biochem. 2008 Jan 23. 143(5):625-32. doi: 10.1093/jb/mvn007. PubMed PMID: 18216068.
13. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic Acids Res. 2000 Jan 01;28(1):235-42. doi: 10.1093/nar/28.1.235. PubMed PMID: 10592235; PubMed Central PMCID: PMC102472.
14. Winn MD, Ballard CC, Cowtan KD, Dodson EJ, Emsley P, Evans PR, McNicholas SJ, et al. Overview of the CCP4 suite and current developments. Acta Crystallographica. 2011 Apr; 67(4), 235-242. doi: 10.1107/S0907444910045749. PubMed PMID: 21460441; PubMed Central PMCID: PMC3069738.
15. CCP4 [Internet]. LSQKAB (CCP4: Supported Program). C2011 [cited 2019 Dec 02]. Available from: http://www.ccp4.ac.uk/html/lsqkab.html
16. Kabsch W. A solution for the best rotation to relate two sets of vectors. Acta Crystallographica. 1976 Sep; 32(5):922-3. doi:10.1107/S0567739476001873.
17. Pires DEV, Minardi RCM, Santos MA, Silveira CH, Santoro MM, Meira W. Cutoff Scanning Matrix (CSM): structural classification and function prediction. BMC genomics, BioMed Central. 2011 Dec 22. 12(S4): S12. PubMed PMID: 22369665. PubMed Central: PMC3287581.

18. Yasara. Yasara View Download. Version 19.9.17 [software]. 2019 [cited 2019 Dec 04]. Available from: http://www.yasara.org/viewdl.htm.
19. PDB [Internet]. Coordinate Section; c2010 [cited 2019 Nov 07]. Available from: http://www.wwpdb.org/documentation/file-format-content/format33/sect9.html
20. MathWorks [Internet]. k-means clustering. C2006 [cited 2020 Jul 13]. Available from: https://www.mathworks.com/help/stats/kmeans.html
21. Tsai CS. An Introduction to Computational Biochemistry. 1st ed. John Wiley & Sons: New York; 2002. doi: 10.1002/0471223840.
22. Kufareva I, Abagyan R. Methods of protein structures comparison. Methods in Molecular Biology. 2012 Jan 11; 857(1): 231-57. doi: 10.1007/978-1-61779-588-6_10 .PubMed PMID: 22323224. PubMed Central: PMC4321859.
23. Monteiro OM. [Development of a Methodology Based on a Distance Matrix for Verifying Protein Similarities]. [dissertation]. Centro Federal de Educação Tecnológica de Minas Gerais; 2017.
24. Campelo F. [Internet]. Design and Analysis of Experiments 06 Simple Comparisons; c2015 [cited 2020 Oct 18]. Available from: https://github.com/fcampelo/Design-and-Analysis-of-Experiments/blob/master/06-SimpleComparisons/Chapter06.pdf
25. Rossi A. [Internet]. Regression Diagnosis; c2019 [cited 2020 Oct 18]. Available from: https://lamfo-unb.github.io/2019/04/13/Diagnostico-em-Regressao/.
26. MathWorks [Internet]. Anova1. C2006 [cited 2020 Oct 19]. Available from: https://www.mathworks.com/help/stats/anova1.html.