CONTINUOUS HETEROGENEITY ANALYSIS OF BIOLOGICAL MACROMOLECULES

Abstract— The study of the flexibility and the different structural conformations of macromolecules is of great importance to understand the function and behaviour of this molecules in both, the biological environments and the industrial applications. Thanks to the development of techniques such as Electron Microscopy it is possible to visualize these macromolecules, which has to be previously analyzed through different computational methods to perform the final study of the structure in three-dimensions. In most of the cases, these computational methods are semiautomatic, so they need the intervention of an experienced user to achieve the correct results. Consequently, there is a great interest in the development of automatic and simple computational algorithms capable of analyzing the structural flexibility of molecules independently of the user. Following the previous idea, the objective of this master thesis is to optimize and improve a new method designed to study molecular flexibility in a fully automatic manner in order to simplify its execution, improve its performance, increment the amount of information extracted from the data and to open new possible approaches for the application of the software.

Keywords— Electron Microscopy (EM), Spherical Harmonics (SPH), Zernike Polynomials (ZP), Multidimensional Scaling (MS), Root Mean Square Distance (RMSD).

INTRODUCTION

The study of the three-dimensional structure of macromolecules is essential to understand properly how the machinery of life works. The function of proteins and other complexes does not only depend on the chemistry behind the molecule (which determines the type molecular interactions that can be established or the final composition of a molecule among other properties) but it is also dependant on the final conformation adopted by the macromolecule (for example, structural changes are responsible of the specificity of the interactions with a given target, leading to completely different responses and outcomes).

In order to extract the information required to reconstruct the three-dimensional structure of a macromolecule, several techniques have been developed including X-Ray crystallography [1] Nuclear Magnetic Resonance [2] and Electron Microscopy [3]. Electron Microscopy relies on the acquisition of several micrographs from a sample in the form of a movie. Afterwards, each molecule (commonly referred as particle) in the micrographs will be extracted, isolated and processed to reconstruct the final three-dimensional structure of the sample. Since is not possible to control the position of the macromolecule in the sample, the isolated images will represent different projections from different projection angles, which can be used to determine the three-dimensional shape of the particle after an angular assignation [4].

However, single particle analysis suffers from different error sources that decrease the quality of the structures reconstructed. For example, single particle analysis assumes that the projections obtained from the micrographs correspond to the same specimen in a specific conformation. This as-

sumption is not always satisfied, as the sample may have contaminants or undesired specimens coming from an imperfect purification process. Moreover, molecules do not remain still in the sample, but they are subjected to a continuous motion due to thermal fluctuations and conformational changes.

Since biological macromolecules are intrinsically flexible structures, new reconstruction algorithms try to classify the particles in order to recover several conformations of the molecule under study (known as discrete heterogeneity). Nevertheless, it is not possible to guarantee that the number of images acquired are enough to reconstruct all the possible conformations with enough resolution (making impossible to analyze the continuous heterogeneity of molecules based on the reconstructed volumes).

In order to overcome the challenges coming from the flexibility of biological molecules, different computational algorithms have been introduced in order to analyze this continuous heterogeneity of molecules. One of the main algorithms used to perform this modeling is Normal Mode Analysis (NMA) [5]. Although NMA is a powerful tool, it is semiautomatic meaning that it requires the intervention of the user in order to return the desired result. This will bias the analysis towards the knowledge of the user, making the results prone to suffer from errors and decreasing their repeatability and reproducibility. The limitations present in semiautomatic methods can only be overcome by developing new algorithms capable of performing continuous heterogeneity analysis on their own, leading to more reliable results.

During the next sections, a mathematical basis capable of analyzing the continuous heterogeneity of molecules automatically will be described and optimized through the development of different methods in order to simplify its execution while improving its performance.

METHODS

This section starts with the mathematical description of the basis that will be used to study the continuous heterogeneity of biological molecules[1]. After the description, the different methods and algorithms that were developed during the master thesis to improve this new tool will be introduced.

Spherical Harmonics and Zernike Polynomials

In mathematics, a basis is usually defined as a set of components (such as vectors or functions) of a given space that are able to span any component belonging to that space. One simple example is the Hilbert space and its basis, that allow to express the spatial position of any point in a N-dimensional space.

Another interesting property of a basis (which is essential for the method proposed along the thesis) is the possibility of moving from a given basis to a completely different one. A representative example of the usefulness of this property is the Fourier Transform, which is commonly used when dealing with waves. The Fourier Transform moves the components of a wave from the so called *Real Space* to the *Fourier Space* or *Frequency Space*. In this space, a wave is represented by a series of delta functions whose height represents the strength of a sine wave of that specific frequency in the composition of the initial wave. The possibility of jumping from one basis to another is extremely useful for the analysis of the data: in many cases, studying a given piece of information may be complex in the *Data Space*, but it might be simpler in the *Alternative Space* defined by a new basis.

Following the previous idea, it is possible to define a new basis [6] to move the structural information of a macromolecular complex resolved by EM to another "space" where the conformational changes are more easily studied. This basis can be defined as a functional space declared over the sphere composed by two set of functions: the Spherical Harmonics (1) and the Zernike Polynomials (2).

$$Y_l^m(\theta,\varphi) = Ne^{im\varphi}P_l^m(cos\theta) \tag{1}$$

[1]The description provided is just for the ease of the reader. For a more exhaustive explanation of the basis, the interested reader is referred to [6]

$$R_l^n(\rho) = \sum_{k=0}^{\frac{l-n}{2}} \frac{(-1)^k (l-k)!}{k!\left(\frac{l+n}{2}-k\right)!\left(\frac{l-n}{2}-k\right)!} \rho^{l-2k} \quad (2)$$

By joining these two components, the basis can be defined as:

$$Z_l^{n,m}(x_r, y_r, z_r) = R_l^n(r) Y_l^m(x_r, y_r, z_r) \quad (3)$$

As explained before, the goal of (3) is to study the structural changes suffered by a given macromolecule. This problem can be understood as finding the deformation that drives one conformation of a molecule towards a different structure. The solution of the previous problem can be seen as:

$$\min_{c_{l,m,n}} ||V_1(\vec{r}) - V_2(\vec{g}(\vec{r}))||^2 \quad (4)$$

Where V_1 and V_2 represent two different conformations of a molecule known beforehand and $g(r)$ is a deformation defined as:

$$\vec{g}(\vec{r}) = \vec{r} + \sum c_{l,n,m} Z_l^{n,m}(\vec{r})$$

Being $c_{l,n,m}$ the deformation coefficients needed to minimize (4).

Volume normalization

The evaluation of the volume at $\vec{g}(\vec{r})$ proposed in (4) cannot be directly performed in general. The volumes reconstructed after processing the micrographs acquired by the EM cannot be formed by an infinite set of points due to the storing limitations of computers. Instead, the volumes are discretized into finite units know as voxels. In fact, the size of this voxels will determine the size of the smallest detail that can be seen in the reconstructed volume (i.e. the voxel size determines the resolution of the volumes). Since the deformation obtained is, in general, a number of arbitrarily large precision (whose lower and upper bounds are again restricted by the computer memory), the position defined by $g(\vec{r})$ might not lie within an exact voxel but in any position within the voxel volume. In order to approximately evaluate the volume at $g(\vec{r})$, it is needed to interpolate the value at this point (in this way, the accuracy of the evaluation will increase). However, interpolation algorithms strictly depend on the values of the voxels that are used to performed the calculations.

As stated by (4), the minimization of the cost function depends on the difference on the voxel value of V_1 at position $\vec{r}$, and the interpolated value of V_2 at position $\vec{g}(\vec{r})$. If the histograms of the two volumes are not properly normalized, the minimization algorithm will not be able to appropriately determine the deformation coefficients that best define the deformation desired.

To that end, a new normalization procedure is proposed. The main objective of this normalization is to align the histograms of the volumes V_1 and V_2 in such a way that the middle region of the histograms is as superimposed as possible.

$$\hat{V} = \frac{V - \tilde{V}_{bg}}{P_{99}}$$

Where P_{99} is the percentile 99^{th} percentile of the voxel values belonging to the foreground (the volume V) and $\tilde{V}_{bg}$ is the median value of the volume background (both, foreground and background can be extracted after applying a mask to the whole volume).

For the normalization to be effective, it is needed to constrain the maximum and minimum values of the normalized volumes as the quotient may result in large absolute values for the maxima and minima of the new histograms. In order to avoid this undesired effect of the normalization, large positive and negative value are constrained to:

$$constrain = \pm \frac{z(0.999)}{z(0.99)} = \pm 1.3284$$

Where $z(x)$ represents the inverse of the normal distribution evaluated at the value x (which represents the percentile P_x).

Multi-resolution analysis

Interpolation introduces another important source of error coming from the resolution of the maps. The mathematical basis introduced in equation (3) is able to compute deformations associated to low and high frequency movements depending

on the degree of the polynomials involved: the higher the degree the basis reaches, the higher the frequency associated to the movements.

However, deformation is analyzed from lower to higher frequency movements. This implies that overlapping of the volumes at the first iterations is crucial to interpolate appropriately the voxels when low frequencies are being analyzed. If the volumes provided to the software show a large spatial resolution, conformational changes appearing in high frequency regions (e.g. small alpha helixes and beta sheets present in the structure of the samples) will prevent the basis from computing low frequency deformations due to interpolation and overlapping limitations.

In order to prevent errors driven by the resolution of the volumes, it is possible to work with several low-pass filtered versions of the input maps to perform a multi-resolution analysis. Following the previous idea, the cost function shown in equation (4) will become a sum over the different filtered pairs:

$$\min_{c_{l,m,n}} \sum_{f} ||V_1^f(\vec{r}) - V_2^f(\vec{g}(\vec{r}))||^2$$

Multi-resolution analysis is a powerful tool to compute the deformation among maps with varying resolution or with fine details as it will be explained during the next sections.

Deformation penalization

One of the effects that have been observed when applying the basis to deform a given volume is the presence of regions that are moved in an excessive fashion. These errors mainly appear when the degree of the polynomials involved in the calculation is large, meaning that the software has more degrees of freedom to modify the structure of the volumes under study.

The deformation phenomenon is a combination of different effects appearing due to the prepossessing steps described before (volume normalization and multi-resolution analysis) plus the effect of the cost function to be minimized during the deformation procedure.

The new normalization method introduced relies on the implementation of percentiles to compute the new voxel values. However, the presence of percentiles may be problematic if the number of zero values present in the foreground is large compared to the number of voxels belonging to the actual volume. As an example, it can be considered that the mask used is circular with a radius equal to half the box size containing the volume. If the box is large, the number of zero values (background) associated to the mask will be much larger than those of the macromolecular complex contained in the box. When this happens, the percentile will be progressively shifted towards the zero value. As a result, most of the voxels belonging to the macromolecular complex will be clipped leading to a volume close to a binary mask (the volume is mostly formed by zeros and clipping values). Since all the voxels within the volume are clipped, the software can freely deformed in any direction within the macromolecular complex (as the correlation with all the surrounding voxels will be high) leading to excessive deformations.

The discussion involving the multi-resolution analysis follows the same reasoning as the one described before. Due to the application of filters, the voxel values will become closer to each other allowing the software to deform in any direction without introducing a cost difference.

The normalization errors can be solved by applying masks that are limited to the volume region where the macromolecular complex is found. For this software, entropy thresholding was implemented to generate mask that overcome the issue of including to many background voxels in the calculation of the percentiles. In the case of the filter, narrow Gaussians are preferred to prevent the undesired effect of having similar voxel values.

Apart from improving the prepossessing steps, it is also possible to introduce restrictions to the cost function to be minimized during the deformation process. An extra term was added that accounts for the deformation and the mass difference between the deformed and the original volumes. In this way, the cost function will increase its value

when excessive deformations are introduced, reducing even further this effect. The importance of this penalization compared to the deformation cost is controlled by a regularization parameter. This allows the user to adjust the behaviour of the software to the needs of the experiment.

Multidimensional scaling

Equation (4) shows the procedure to be followed to determine the deformation that has to be applied to a given volume to determine how it moves towards a new structure. Nonetheless, the basis proposed in (3) is not only intended to be applied to a single pair of conformations.

Considering the case where the processing of the micrographs acquired by the EM rendered an arbitrary number of volumes N (which represent different conformations of the same specimen). For this new case, it is possible to apply the basis previously introduced to all possible pair combinations of the N volumes. After computing the deformation coefficients and the deformed maps, a distance measurement can be established among all the different deformed volumes with respect to their reference pair. Following the previous procedure, a distance matrix will be obtained. This distance matrix shows how close are the structure of two volumes in terms of the deformation extracted with the basis. However, a direct representation of the distance matrix is not possible, as the coordinates that can be extracted from the distances will lie on a *N-dimensional* space.

In order to find a representation of this coordinates in a lower dimensional space, an approximation has to be used. The challenge is to find a set of coordinates in one, two or three dimensions whose distances are as close as possible to the *N-dimensional* matrix defined by the basis algorithm. The solution to this problem requires a tool commonly known as *Multidimensional Scaling* [7]. As stated before, this tool finds a set of points whose distance are as close as possible to an input matrix of arbitrary dimensions. Although there are several formulations of the algorithm, here it is provided the classical formulation for the convenience of the reader [8]. Given a distance matrix $D \in \mathbb{R}^N$ which has been centered by a centering matrix J, the set of points in dimension m that best describe the contents in matrix D are:

$$P = E_m \Delta_m^{1/2} \tag{5}$$

Being E_m and Δ_m the first m eigenvectors and eigenvalues associated to matrix D. The matrix P containing the coordinates of each point is commonly referred as an *structure map* of the macromolecular complex.

The previous classical definition of the *Multidimensional Scaling* problem is convenient as it can be implemented easily in a computer. However, the mathematical formulation of the tool described before does not contemplate the possibility of combining several distance matrices computed using different metrics. This is an useful application, as different distance metrics will provide with a different result that can be combined to create a consensus that maximizes the information extracted from the *Multidimensional Scaling*.

For the case proposed in this master thesis, the comparison of the N different volumes using, for example, deformation distance (the amount of deformation in pixels suffered by a given volume when deformed towards a different structure) and correlation distance (correlation between the deformed volume and it reference conformation converted to distance) matrices may result in two completely different set of points P when equation (5) is applied. In fact, since the two matrices will generally report a different perspective of the problem, both will provide with valuable information when analyzing the structural transitions hidden in the data. Since each metric provides with a different (but valid) point of view, the goal is to find a consensus among the matrices to have the best possible solution to the problem.

In order to find the best combination of the resulting matrices obtained by (5), a new method based on entropy minimization of the point matrices P is proposed. Starting with a series of coordinates matrices P_i obtained after applying classical *Multidimensional Scaling* to a given dis-

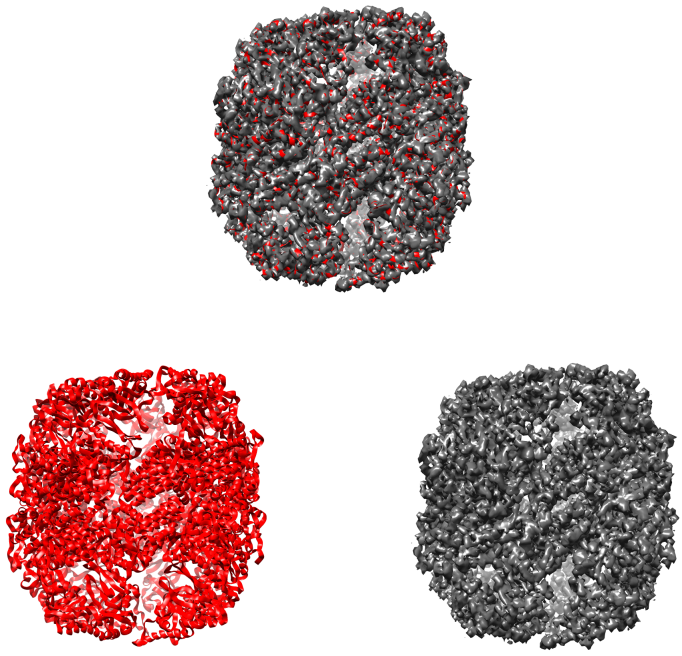

Fig. 1: Simulated EM map from PDB file. The top image shows the alignment between the simulated map (gray) and the original PDB file (red).

tance matrix D_i, the objective is to find the coefficients α_i such that:

$$C = \sum_i \alpha_i P_i$$

In the previous linear combination, the coefficients α_i must also verify that:

$$\sum_i \alpha_i = 1$$

Since all the coefficients add up to one, the final weights can be regarded as the "importance" of the information stored in the set of points P_i when included in the consensus. The criterion followed to determine the best combination of weights for each matrix was to find the combination that minimizes the entropy of the matrix C. This can be done by iterating along the different possible values of α_i, convolving the resulting matrix C with a Gaussian kernel and analyzing the Shannon entropy of the final images for each of the combinations of α_i.

However, it may remain unclear why to choose a minimum of entropy instead of a maximum. Since entropy can be considered as a measure of the amount of information in an image [9] (meaning that an image with maximum entropy has maximum information), then using a minimum of entropy as a criterion will result in a set of points C with minimum information. Although the previous statement is true in general, it might not be appropriate for the calculation of the structure maps.

Let's consider a set N volumes (that represent a series of different conformations defining a transition between the first and the N^{th} volume). The expected result to be obtained after applying classical *Multidimensional Scaling* to the output distance matrix D is a reproduction of the trajectory that the first volume follows to become the N^{th} volume. Having a trajectory of points means that we have a highly order structure with low entropy (otherwise, the points would be spread around space). If instead several distance matrices D_i are used, a similar trajectory should be recovered as it best describes the structural relations established among the volumes analyzed. This means that the criterion will be to search a minimum of entropy to increase the probability of recovering a trajectory if present.

Since the samples imaged in EM are generally composed by a single specimen, the different possible structures that will be recovered after the 3D reconstruction will represent different conformations of the same macromolecule. Thanks to the previous fact, it is possible to safely apply a minimum-of-entropy criterion as it is desirable to determine if the conformations recovered define a "transitional conformational change" or if they represent isolated conformations instead.

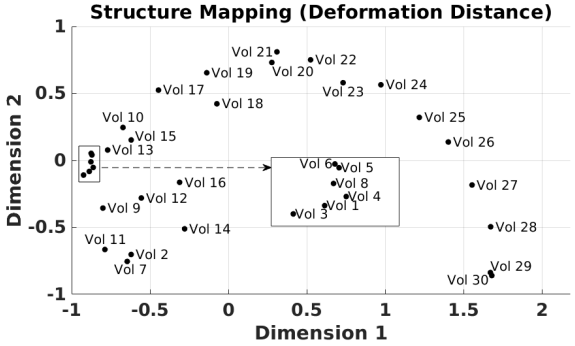

Fig. 2: Deformation structure mapping obtained after running the software with a simulated trajectory of 30 PDBs.

RESULTS AND DISCUSSION

In this section, the results obtained after applying the methods described previously to different data will be described and discussed. The test performed were run with simulated and experimental data (including both, EM maps and PDB files) in order to assess the the accuracy of the conformational changes computed by the algorithm under different conditions that are representative of real EM studies.

Conformational changes between simulated EM maps

As explained before, one direct application of the basis introduced in equation (3) is to compare two different conformations of a macromolecule. However, in many cases the data that can be accessed through databases or simulated data appears in the form of a PDB file. A PDB file contains information about the position of the atoms, commonly derived from an electron density map. Since the basis is only intended to be applied to electron density maps, it will be needed to perform a conversion of the atomic coordinates to electron density. An example is provided in **Figure 1**.

The simulation of electron density maps from a PDB file is, however, introducing some complications in the analysis of the data. The main challenge is to overcome the interpolation errors arising from non-overlapping regions in the volumes to be compared. As explained in the previous section, filtering the input maps is necessary to improve the overlapping, allowing the software to compute more accurately the deformation due to low frequency movements.

The data analyzed was composed by a set of 30 different simulated PDB files that define an open-close trajectory of a GroEL complex. The results obtained after computing the structure mapping trough spherical harmonics is show in **Figure 2**, **Figure 3** and **Figure 4**.

Figure 2 shows the structure mapping associated to the deformation distance metric. This metric does not take into account the shapes of the volumes that were compared, and it mostly provides information about the low frequency movements that the different volumes have suffered. As it can be seen from the image, the right branch is defining a trajectory as expected, but the left branch shows a disordered cloud of points. The presence of this cloud is better understood when comparing **Figure 2** and **Figure 3**. In this region, the differences between the volumes are small and have mostly a high frequency contribution. Since the degree of the basis used was not high enough, it was not possible to compute the deformation due to the high frequency regions, leading to the unstructured area shown in **Figure 2**.

However, **Figure 3** has a well resolved trajectory even in the region of high frequency differences. This change in the result appears due to the application of a different metric to compute the structure mapping: the correlation distance. Thanks to correlation, the information stored in the shape of the different volumes is taken into account when com-

puting the dissimilarity measure among the different possible volume pairs. As a result, areas defined mainly by high frequencies can be differentiated leading to a better structure mapping. For this specific case, correlation distance provide us with a well defined trajectory as it takes into account both the low frequency information (coming from the basis) and the high frequency information (mainly coming from the correlation measure).

Figure 4 shows the consensus of minimum entropy found between the two previous structure maps. As expected, most of the weight is given to the correlation metric, as it already includes low and high frequency information. According to the consensus and to the structure maps, for well defined trajectories with enough samples/volumes, correlation distance is preferred as it will provide with information that is not present when working with the deformation distance. In addition, correlation metric will also play an essential role when the spatial resolution of the maps to be analyzed is large.

Random conformational changes in simulated EM maps

In may cases, the different experimental EM maps reconstructed form the images of the sample do not represent a conformational trajectory. Instead, it is possible that the different specimens found after analysing the data through a 3D classification correspond to a random set of different

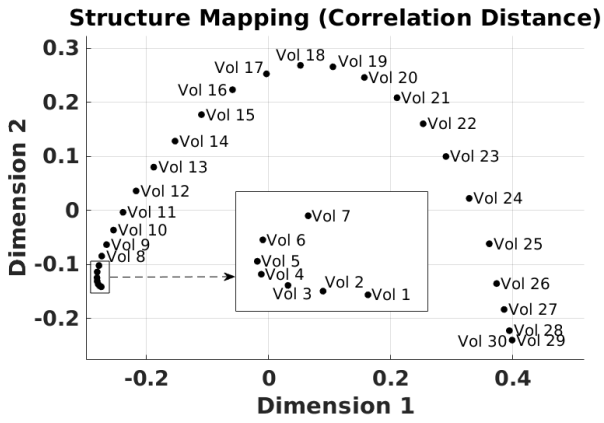

Fig. 3: Correlation structure mapping obtained after running the software with a simulated trajectory of 30 PDBs.

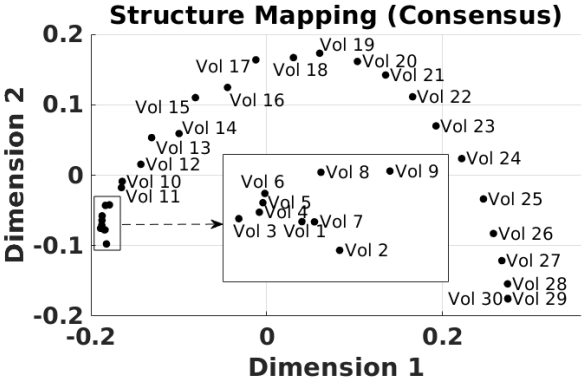

Fig. 4: Consensus structure mapping obtained after running the software with a simulated trajectory of 30 PDBs.

states of the sample that cannot be structurally related. This may happen when the energy landscape of the sample has several viable conformations that minimize the energy of the macromolecule and the structures reconstructed represent random minima spread along the energy landscape.

The study of a random walk simulated through molecular dynamics is, according to the previous idea, of great interest to determine the behaviour of the software under the aforementioned conditions. The test performed consisted on the analysis of 23 different PDB files obtained after simulating a random walk followed by a GroEL complex through molecular dynamics. The consensus obtained after applying the entropy method to the deformation and correlation metrics is shown in **Figure 5**.

When simulating a random walk through molecular dynamics, two cases may appear depending on the parameters used to create the different random conformations during the simulation. If the sampling of the conformational changes suffered by the initial macromolecular state is slow, the resulting structure map should look like a Gaussian around the initial conformation. However, if the sampling rate is increased, the conformations obtained will resemble a random trajectory, as the sampling is able to capture intermediate states that share structural information with their predecessors.

According to **Figure 5**, the random walk simulated corresponds to the second case. The resulting structure mapping (obtained after applying

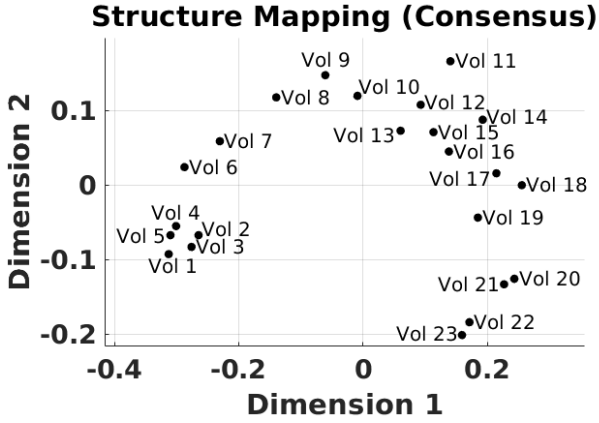

Fig. 5: Consensus structure mapping obtained after running the software with a simulated random walk of 23 PDBs.

the entropy consensus to the deformation and correlation metrics) shows a trajectory between the initial and the final conformations of the random walk, although the intermediate steps are randomly arrange when compared to **Figure 4**. The results shows that the software is able to capture correctly the information of a trajectory even if it is disturbed by random noise. This result is a better indicator of how the software will behave in the real and non-ideal world where the movement suffered by macromolecular is also affected by thermal noise.

Despite the usefulness of analyzing random trajectories, the study of randomly distributed Gaussian conformation is also of great value, as it shows the effects appearing on the structure mapping when the structures to be fed into the software do not have a close relation. This should be performed as a future study in order to assess more deeply the efficiency and performance of the entropy consensus and the optimized basis.

Compatibility of deformation and PDB spaces

Although the software described and improved along this master thesis is intended to be applied to EM maps, it is interesting to study the compatibility of the space defined by the deformation coefficients and the space defined by the atomic positions in a PDB file. If the spaces are compatible, the deformation computed with the EM maps could be directly applied to the PDB files to obtain a deformed version of the original PDB. This introduces a new range of possible studies like energy minimization or docking of the deformed structures to get new information about the conformational changes computed by the mathematical basis.

The first step is to determine how the deformation should be applied to the PDB files. The software defines the deformation computed from the original volume as:

$$V_R(\vec{r}) = V_I(\vec{r} + \vec{g}(\vec{r}))$$

Being $g(r)$ the deformation computed through the coefficients and the basis. The previous equality holds if the degree of the polynomials goes to infinity. Otherwise, the result would be just an approximation of V_R up to a high frequency conformational change (i.e. the specific features of V_I are kept untouched during the deformation process).

The previous equation is referred to a coordinate system whose origin is placed on V_R. However, when the deformation is applied to a PDB file, the reference volume is lost so is the coordinate system. It is possible to rewrite the previous equation in terms of the coordinate system defined by V_I by applying a change of variable of the form:

$$\vec{r}' = \vec{r} + \vec{g}(\vec{r})$$

And the previous equation becomes:

$$V_R(\vec{r}' - \vec{g}(\vec{r})) = V_I(\vec{r}')$$

This implies that the deformation to be applied to the PDB files to check if the spaces are compatible is inverted. In fact, this means that, due to the change in the coordinate system, a compression in a volume should translate into an expansion in the PDB if the sign is not taken into account.

Table 1: RMSD (in Å) measured from the original and deformed PDB

	C2	C8	C17	C24
C0 Original	1.202	5.044	7.918	8.536
C0 Deformed	0.853	2.752	5.038	5.468

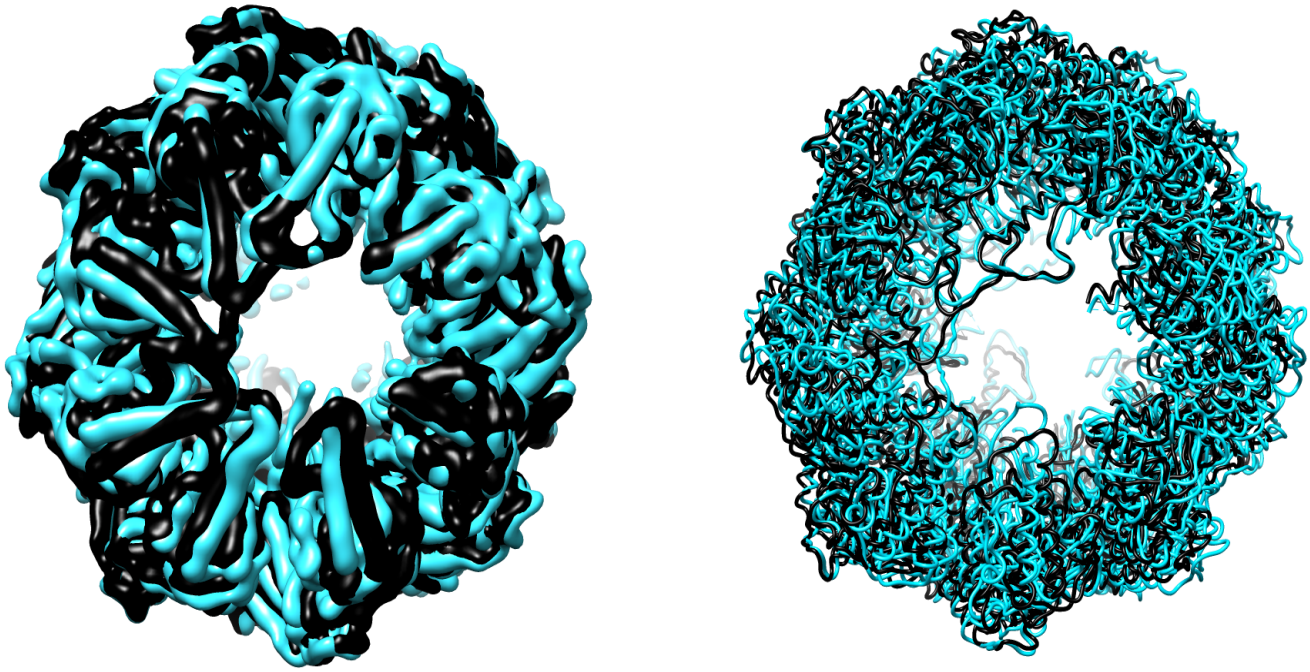

Fig. 6: Deformed EM maps (left) and corresponding PDB files (right). The image shows the deformed (black) and reference (blue) maps and PDB.

Figure 6 shows the deformation applied to a volume and its correspondent PDB file with Chimera. The procedure followed was identical to the way volumes are deformed: for each atomic position, the mathematical basis value is computed to find the deformation $g(r)$ through the deformation coefficients. Then, the value of $g(r)$ is inverted and use to move the atomic coordinates to the new location. This last step involved an interpolation when working with volumes as a voxel value value was needed. When working with PDB files, there is no need to interpolate as we are working with coordinates, so the deformation displacement can be directly applied.

As it can be seen from the image, the deformation applied to the PDB reproduces the structural change defined by the maps. Moreover, the RMDS between the deformed and the reference PDB is decreased as shown in **Table 1**. This result implies that the spaces defined by the deformation coefficients and the atomic coordinates in a PDB file are compatible, so the deformations computed thorough the EM maps can be directly applied to the PDB files.

Comparison of simulated and experimental EM maps

All the results presented during the previous sections were obtained after applying the mathematical basis previously described and studied to different sets of simulated data. Thanks to this data, it is possible to control the conditions of the tests, allowing to better refine and understand the mode of operation of the software. However, it is difficult to assess how accurately our simulations can reproduce real EM maps due to the large amount of factors that can affect the final reconstruction of a real sample. The lack of a way to validate how well simulated EM maps reproduce reality makes difficult to assess quantitatively the performance of the simulation algorithms. We propose the usage of the optimized mathematical basis to give a measurement of this validation.

One of the main challenges that appear when comparing simulated and experimental EM maps are the differences in the structural information stored in the volumes. Due to construction, simulated EM maps have intrinsically much larger res-

olution compared to experimental maps. Due to the resolution differences, interpolation limitations will start playing an important role, introducing errors that will prevent the software from comparing properly the data. As discussed before, this undesired effect can be minimize either by applying a multi-resolution analysis or by prepossessing the maps to give them approximately the same spatial resolution. It is important to mention that, even after this prepossessing step, the structural differences among the experimental and simulated maps might be large due to reconstruction errors.

The data used for the test was a set of ten volumes representing different conformations of the rabbit ryanodine receptor. Five of the maps of the set were obtained experimentally while the other five were simulated from the PDB files extracted from the five previous experimental maps. This implies that, by construction, it is expected to observe five groups of points in the structure mapping representing a experimental-simulated map pair.

The results obtained are presented in **Figure 7**. For this application, deformation distance was found to be the best choice. The main reason are the structural differences present between the simulated and experimental maps that will drive the correlation distance to a "class grouping" (i.e. the structural map will show two well differentiated groups representing the experimental and the simulated maps, making impossible to analyze the similarities present between the two classes).

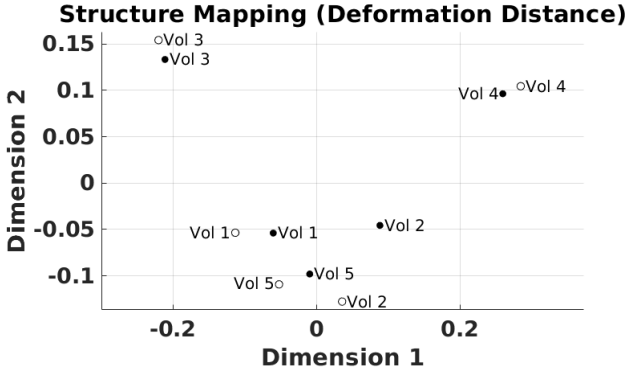

Fig. 7: Deformation structure mapping obtained after applying the mathematical basis to a set of experimental-simulated map pairs.

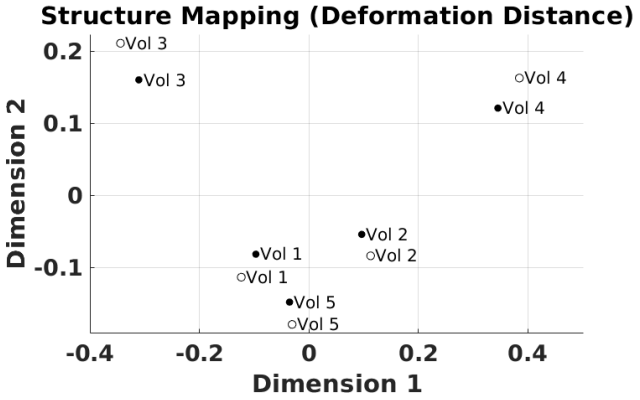

Fig. 8: Structure mapping obtained after performing a submatrix analysis to a set of experimental-simulated map pairs.

Since deformation distance is a metric that does not take into account structural features, it will provide more reliable and informative results. Most of the pairs can be properly identified in the structure mapping obtained, although there is a pair that was not identified correctly. This effect appears when the conformational changes between two different structures are small, reducing also the value of the deformation distance. Ideally, the smallest deformation distance should be found when comparing the experimental-simulated pair. Nevertheless, this is not true if the distance of two conformations is small, as there will be other pairs showing short deformation distances as they are structurally similar.

In order to overcome this limitation, it is possible to analyze the distance matrix by parts. In this way, the first five rows of the 10×10 distance matrix will represent the comparison of the simulated EM maps with the experimental maps and with themselves. By analyzing this two groups independently, it is possible to asses if the relationships among the volumes are similar when the structure mapping of the two groups are compared. This result is shown in **Figure 8**. As it can be seen from the structure mapping, the independent analysis of the submatrices reveals all the pairs avoiding the errors present in **Figure 7**.

According to the results obtained, it can be concluded that the basis is a good method to validate our simulated data. In this way, it is possible to

measure how well our simulations represent a real scenario and to correct the maps in case the dissimilarities are large enough.

CONCLUSIONS

Macromolecular complexes are not static entities that remain in a given conformation when performing their functions. Instead, their structure is highly dynamic, as it usually changes from one conformation to another depending on the conditions of their environment. Moreover, the three-dimensional structure of macromolecular complexes is strictly related to is function, as it exposes different interaction regions that may remain hidden otherwise.

The analysis of conformational changes is, as described before, essential to understand properly the macromolecules under study. There are, however, several limitations when working with algorithms to described the continuous heterogeneity of molecules, being some drawbacks the lack of full automation of the process and the difficulty of executing the software properly for a given experiment. During this master thesis, a new software to study macromolecular flexibility has been described and optimized to performed this tasks more easily, simplifying the process and improving the results obtained.

The software has shown its performance when applied to a range of different data simulating a variety of scenarios that might be found when studying the flexibility of macromolecular complexes in EM. In all the cases, the expected result is obtained and a consensus among different metrics was computed to maximize the amount of information that can be extracted from the samples. In this way, it is possible to determine if the reconstructed maps coming from the images acquired are structurally related (i.e. they define a conformational trajectory when the structure map is computed) or not (the structure map shows a random layout of points).

In addition, the PDB and deformation spaces has been shown to be compatible. This has important implications, as it opens new fronts to study how the deformation computed affects to the atomic positions. The deformations computed may lead to changes in the interactions found in the macromolecule analyzed that can be studied through energy minimization algorithms and docking.

Lastly, the software were also proven to work correctly when faced with a dataset containing simulated and experimental data. Thanks to this comparison, it is possible to determine how accurately our simulations can reproduce real data reconstructed from the images acquired with the EM.

REFERENCES

[1] J. Drenth and J. Mesters, *Principles of protein X-ray crystallography: Third edition*, vol. 9780387333342. Springer New York, 2007.

[2] J. Jonas, "High-resolution nuclear magnetic resonance studies of proteins," 3 2002.

[3] M. J. Ellis and H. Hebert, "Structure analysis of soluble proteins using electron crystallography," 7 2001.

[4] S. H. Scheres, R. Núñez-Ramírez, C. O. Sorzano, J. M. Carazo, and R. Marabini, "Image processing for electron microscopy single-particle analysis using XMIPP," *Nature Protocols*, vol. 3, pp. 977–990, 5 2008.

[5] C. O. Sanchez Sorzano, A. L. Alvarez-Cabrera, M. Kazemi, J. M. Carazo, and S. Jonić, "StructMap: Elastic distance analysis of electron microscopy maps for studying conformational changes," *Biophysical Journal*, vol. 110, pp. 1753–1765, 4 2016.

[6] D. Herreros Calero Supervisor, C. Óscar Sánchez Sorzano Tutor, and J. Ripoll Lorenzo Leganés, ""Image Processing Algorithms for Electron Microscopy"," tech. rep., 3 2018.

[7] J. Kruskal and M. Wish, *Multidimensional Scaling*. SAGE Publications, Inc., 7 2011.

[8] W. K. Härdle, L. Simar, W. K. Härdle, and L. Simar, "Multidimensional Scaling," in *Applied Multivariate Statistical Analysis*, pp. 397–412, Springer Berlin Heidelberg, 2012.

[9] R. M. Gray, "Entropy," in *Entropy and Information Theory*, pp. 61–95, Boston, MA: Springer US, 2011.