Numerical modelling of frustrated fibrous self assembly of Ure2p prion.

Clara Delahousse, under the supervison of Dr. Martin Lenz[1]

[1]**Laboratoire de Phyique Theoriques et de Modeles Statistiques, Soft Biophysics Group**

Abstract

The goal of this study is to determine the structure of Ure2p yeast prion, a misfolded form of Ure2p protein responsible for a stress-triggered phenotype in yeast, after its self assembly into fibrous aggregates. Using biological data which provides in particular information on residue interactions, this work aims at finding the conformation of the prion and compare it with the one of a native protein, which doesn't form fibrils, so as to investigate the mechanisms leading to pathological fiber formation. As self assembly happens as a way to minimize interaction energy between elements, this work replicates the self assembly principle and develops a conjugate gradient method based algorithm which computes and minimizes the energy of one protein inside of a fiber. The minimized parameters correspond to the smallest energetic cost and could therefore be biologically relevent and may characterize the real fibers shapes. This report describes the prion monomer conformation as well as the prion dimer conformation and the different types of fibers obtained. Then the deformation energy associated to different fiber conformations is computed, which highlights the role of geometrical frustration in fiber formation.

Introduction

Self assembly is the process by which elements gather so as to minimize their overall energy. When the assembling elements are ill-fitting, the energy minimization may require them to undergo geometrical frustration: the assembling elements are deformed in the process. Self assembly is extremely useful in biology, as many supra-molecular structures are formed in this fashion, such as microtubules or actin filaments for instance. However, frustrated self assembly can also be detrimental for living organisms. Actually, some ill-fitting proteins can gather and form deleterious fibers involved in neurodegenerative diseases like Alzheimer's and Parkinson's, or in prion diseases. This study focusses on the Ure2p protein which is a yeast prion known to self assemble into fibers, but whose deformed configuration once inside a fiber has yet to be fully characterized. There are several hypothesis regarding Ure2p self assembly. In any case the native or non prion protein doesn't assemble into fibrils, only the prion is assembly competent. Therefore the aim of this work is first to determine the prion conformation. Once in its prion form, the protein could either form fibrils directly, from prion monomers, or first form prion dimers which would then assemble into a fiber. Besides, two different kinds of fibers have been observed depending on the temperature, with one of them being an amyloid fiber with a core of cross beta sheets. The hypothesis proposed by this work is that depending on the temperature, the protein can fold into two distinct assembly competent forms, leading to the formation of two different kinds of fibrous aggregates. This report proposes a conjugate gradient method based algorithm which uses crosslinking data to define intra-molecular and inter-molecular interactions between residues and minimizes the energy associated to these interactions for a protein inside a fiber. Knowing that all proteins in a fiber have the same neighbors, the goal is to test possible fiber configurations respecting this criteria. Through this process we hope to determine possible deformed protein configurations after fibrous self assembly, and help investigate the origin of the formation of protein fibers. In the following report, Section 1 presents the biological data available and how it is physically modelled as well as the parametrization used to test all relevant fiber conformations, and describes the algorithm used. Section 2 focusses on the description of the assembly competent proteins, dimers and fibers obtained, and the caracterization of the different fibres. Lastly Section 3 highlights the role of geometrical frustration in fiber formation, and classifies the fibers obtained into two categories : amyloid and non amyloid fibers.

1 Methods

This sections aims at describing the numerical modelling of the self assembly process of a biological prion through first describing the biological data available and how it can be integrated in a model in section 1.1, then highlighting the interest of fiber parameterization in section 1.2, and finally describing the algorithmic process set up in section 1.3

1.1 Modelling the biological data

In order to make use of the biological data at hand in a physical model, it is necessary to first describe the physical relevance of this data in 1.1.1 and then simplify it to create a model of protein aggregation in 1.1.2.

1.1.1 Biological data

The following paragraphs present the data gathered by Dr. Ronald Melki's team at the Laboratoire d'enzymologie et de biochimie structurale de Gif-sur-Yvette, which allows for a thorough description of the protein and gives information on the protein conformation inside a fiber.

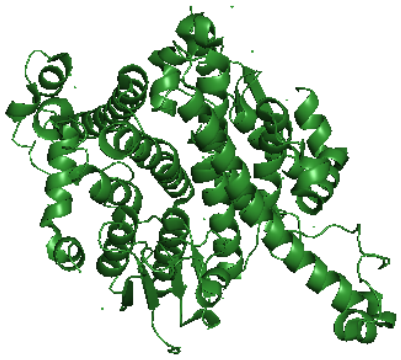

Figure 1. Ure2p C-terminal dimer visualized with Pymol.

Prions are proteins folded into a transmissible pathological conformation which leads to them self assembling and forming deleterious aggregates, often disrupting the affected organism. Ure2p is a yeast prion that propagates a stress-triggered phenotype in baker's yeast. In its native conformation Ure2p is involved in a pathway regulating he use of nitrogen sources, which is lost in cells containing the prion form. Ure2p is composed of 2 distinct domains [8]:

- The N-terminal domain, ranging from amino acid 1 to 95 which is poorly structured and flexible
- T C-terminal domain, ranging from amino acid 96 to 354, which contains many alpha helices and is much more rigid. (fig 1)

The terms N-terminal and C-terminal refer to both ends of a protein. Since to form a peptide the amine group of an amino acid binds to the carboxylic group of the next one, at one end an amino group remains unbound (N-terminal) and at the other end it is a carboxylic group (C-terminal). The amino acids are the numbered starting from the N-terminus. The structure of the rigid C-terminal domain has been solved and is available for visualization thanks to a data base which especially contains all the coordinates of the atoms forming the protein

in Å, and which residue each atom belongs to [2]. This information is accessible through the Pymol software [6] The native or non-prion protein is active and soluble and forms dimers through the interaction of 2 C-terminal domains. The N terminal domain of Ure2p is known as the prion forming domain because it is required in the change of conformation that leads to the prion structure [5]. When such configuration change has happened the protein is said to be assembly competent. There is little information about this conformational change, other than that the N-terminal domain is believed to wrap around the C-terminal domain. Experiments have also shown that there is a crystalline structure inside some fibers which is consistent with the conservation of the integrity of the C-terminal domain. [1]. In addition, native Ure2p forms dimers that conserve the crystalline structure of the C-terminal of monomeric Ure2p. Therefore two hypothesis for fiber formation are explored in this work: fibers could either be formed through self assembly of monomeric Ure2p, or through self assembly of dimeric Ure2p (figure 2). The following paragraph describes experimental data giving insights on the formation of such dimers and fibers.

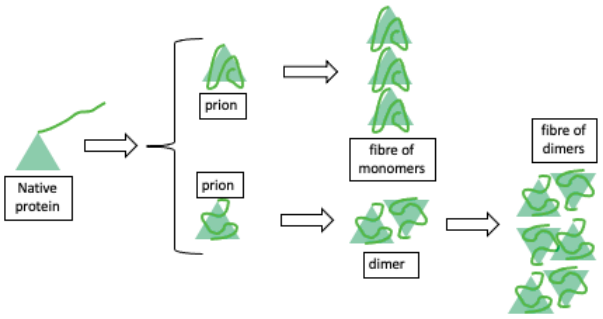

Figure 2. **Illustration of the two fiber formation processes considered.**

Several experiments have been carried on by the team of Dr. Ronald Melki so as to get information on Ure2p's structure inside a fiber. Cross linking and oxidation experiments provide data on the distance between residues in the fiber which is summarized in figure 3. Besides, RMN experiments show that all the dimers in the fiber are in the same environment, meaning that they all have the same interactions with the same neighbors. See appendix for more details on the experiments.

1.1.2 Modelling

This section describes the spring and spheres model chosen to represent the protein aggregates. In this model each residue is represented by a hard sphere of fixed radius. Complex physical interactions fix the distance between residues in reality, and these are modelled as an harmonic deformation cost.

As developed before, the protein is constituted of two domains which each have a different conformation and rigidity characteristics and require different modelling

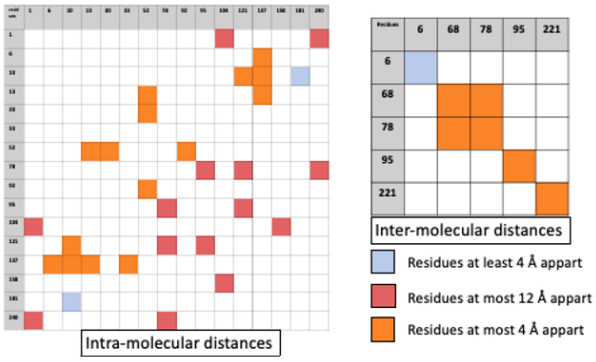

Figure 3. **Crosslinking distances data. The rows and columns refer to the residues that have been marked and each case is colored depending on how the couple of residues associated interact.**

strategies. However in both cases only the residues involved in the crosslinking and oxidation experiments are modelled, as they are the only ones for which distance constraints can be specified. The N-terminal domain which is highly flexible is modelled as a gaussian chain of amino acids, linked by spring core interactions which rigidity is $\frac{1}{K}\ N.m^{-1}$, with K the number of amino acids in the chain. Per this high flexibility, the structure of the N-terminal has not been solved and there is little information on the lengths between its residues, which have to be approximated. Here the relaxed length between amino acids is chosen as 1 Å which is the length of a covalent bond. As the coordinates of the residues of the C-terminal are well documented, they are extracted from the biological data base to build the model. These residues are also linked through spring core interactions which spring constant is set to 10 $N.m^{-1}$ since the C-terminal is more rigid than the N-terminal, but no further order of magnitude is known. The number of interactions between residues un the C-terminal is $i = 6*(r-1)-3*r$ with r the number of residues, which leads to an isostatic structure : a structure for which all the degrees of freedom are constrained only once, see figure 4 .

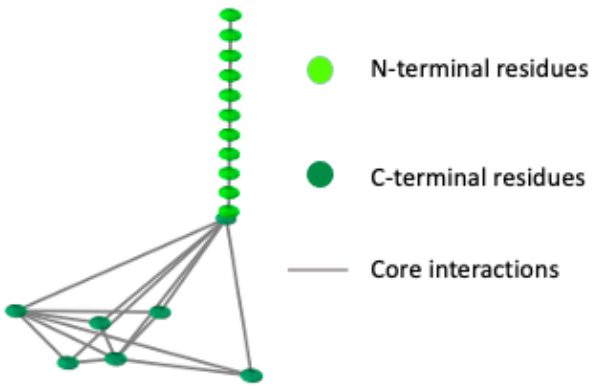

Figure 4. **Model native protein.**

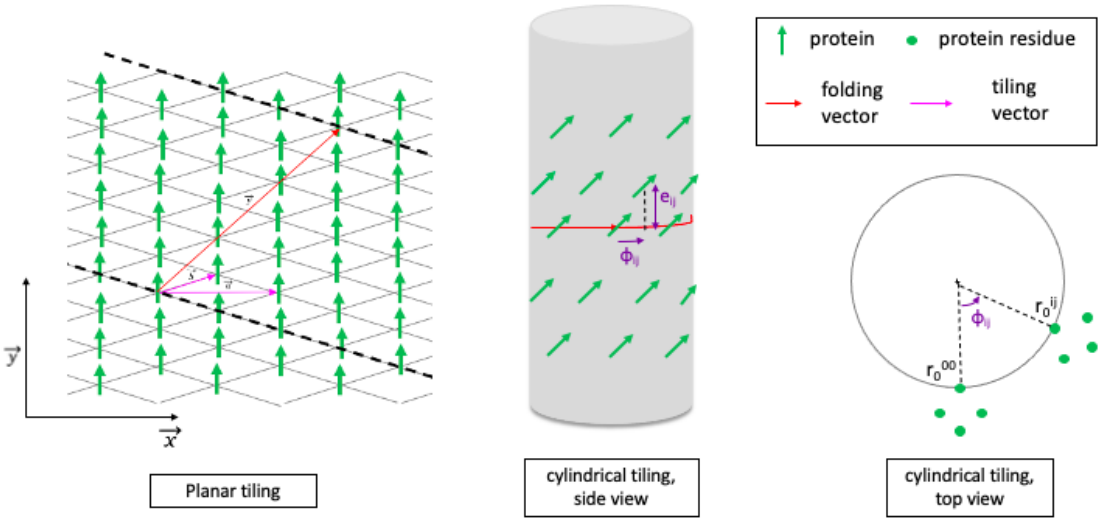

Figure 5. Description of the folding process of a parallelogrammatic lattice described by vectors $\vec{a}$ and $\vec{b}$ onto a cylinder, with the folding vector $\vec{p}$. First panel represents the tilling on a plane before folding. Panels two and three represent two views of the tilled cylinder after the folding. In panel three the protein is represented by its residues with residue zero on the surface of the cylinder.

Experiments provide information on the distances between residues in the prion's configuration. These distance constraints are modelled as non physical inter-residue interactions, which only enforce the distance constraints. The crosslinking data is modelled as an attractive interaction when the residues are known to be closer than a given threshold and as a repulsive interaction when they are known to be further away than a given threshold. The attractive interactions correspond to a low energetic cost when the interacting residues are at a distance bellow a threshold d_0^a, and a harmonic energy barrier when the distance is above the threshold. Conversely for repulsive interactions the energy is high when the residues are closer than the threshold. This threshold corresponds to the size of the crosslinker used in the experiments (see appendix 3.1.2). All in all the three kinds of interactions involved are defined as:

$$E_{spring} = \frac{1}{2} * k_{spring} * (d - d_0^s)^2 \tag{1}$$

$$E_{attra} = \begin{cases} \frac{1}{2} * k_{attra} * (d - d_0^a)^2 & \text{if } d > d_0^a \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

$$E_{rep} = \begin{cases} \frac{1}{2} * k_{rep} * (d - d_0^r)^2 & \text{if } d < d_0^r \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

where d_0^s, d_0^a, d_0^r are the cutoff distances for the spring, attractive and repulsive interactions respectively, and d is the distance between the two interacting residues. Experimental data indicates which residues are in interaction, however there is no way to know which dimer or monomer two interacting residues belong to. Therefore all possible interaction partners need to be tested by the algorithm. Lastly, the steric repulsion between the C-terminal parts of two proteins are modelled as repulsive cut-off interactions between each residue of the C-terminal of protein zero and the residues of the C-terminal of its interaction partner. The cut-off distance is chosen as twice the radius of a spherical residue : $E_{ste} = E_{rep}$ with $d_0^r = 2r_r esidue$. This ensures that the residues cannot interpenetrate.

1.2 Fiber parameterization

This section describes analytically the tilling method used in the minimization algorithm to parameterize all fiber configurations respecting the identical environment biological constraint.

Considering a general form for a fiber boils down to tilling a cylinder. To obtain such a tilled cylinder, let us consider a tilled plane which will be folded onto itself to obtain a cylinder. Here this tilling and folding process must account for the biological constraint of all proteins having the same neighbor configuration. The first step is to identify which tilled planes respect the constraint, and the see in which cases the constraint holds when folding the plane. In mathematics, a plane covered by the regular repetition of a motive is called a wallpaper. There are seventeen wallpaper types depending on the symmetries conserved, which form the wall paper group. Only two of these satisfy the identical neighbors constraint in the case of the cylinder (see appendix and figure 6) :

- The p1 wallpaper which consists in a regular parallelogrammatic lattice
- The p2 wallpaper which is also based on a simple parallelogrammatic lattice whose motive is repeated with a 180 degrees rotation symmetry around each node.

Such wall papers are described by lattice vectors defined as $\vec{b}$ defined as $\vec{a} = a_1 * \vec{x}$ and $\vec{b} = b_1 * \vec{x} + b_2 * \vec{y}$ in **??**. The folding process can be characterized by a vector $\vec{p}$ called folding vector which links

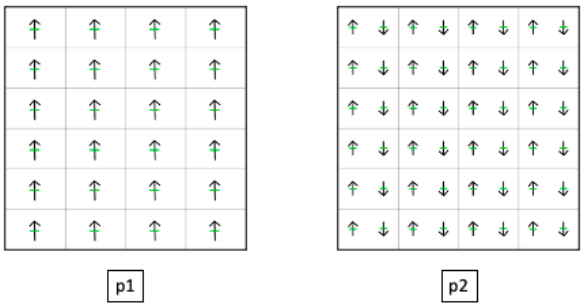

Figure 6. p1 and p2 wallpaper groups on rectangular lattices and with an arrow as a base motive.

two nodes that coincide once the plane is folded. The folding vector is a linear combination of the lattice vectors: $\overrightarrow{p} = m * \overrightarrow{a} + n * \overrightarrow{b}$, with m and n integers. THe folding vextor fixes the radius of the tilled cylinder to $R = \frac{1}{2\pi}\sqrt{(ma_1 + nb_1)^2 + (nb_2)^2}$. Once the cylinder is defined, fiber parameterization provides a way to generate the position of any residue of any protein in the fiber from a minimal set of parameters. The choice of a set of linearly independent parameters is crucial in the following when it comes to the method of conjugate gradient. The planar tilling is described by the coordinates of the lattice vectors, and the coordinated of one protein around the tiling's node, which defines the motive. The protein used to define the motive is called protein zero in the following. However, this set of parameters is not linearly independent in the case of the tilled cylinder, as the fixed radius adds a constraint: the center of mass of the protein must be on the surface of the cylinder. This constraint is equivalent to fixing the position of one of the residues of the protein to $(R, 0, 0)$. Let us fix the position of the first residue of protein zero, residue zero (see figure 5) so that $(x_1, y_1, z_1, ...x_{n-1}, y_{n-1}, z_{n-1}, a_1, b_1, b_2)$ is the set of linearly independent parameters describing the problem from now on. Considering protein ij as the protein at a distance $i + \overrightarrow{a} + j * \overrightarrow{b}$ from protein zero, the coordinates of any residue in this protein is generated thanks to a rotation of angle ϕ^{ij} around the cylinder's axis and a translation of e^{ij} along the cylinder's axis, with ϕ^{ij} and e^{ij} determined as follows : if $(\overrightarrow{e_r}, \overrightarrow{e_\theta}, \overrightarrow{e_z})$ is the referential of the cylinder, then $\overrightarrow{e_r}$ vector is the one comming out of the unfolded plane, the $\overrightarrow{e_\theta}$ vector is in the direction of the folding vector $\overrightarrow{p}$ and the $\overrightarrow{e_z}$ vector is orthogonal to $\overrightarrow{p}$ in the plane. Is it necessary to project the $\overrightarrow{a}$ and $\overrightarrow{b}$ vectors in the $(\overrightarrow{e_r}, \overrightarrow{e_\theta}, \overrightarrow{e_x})$ base. The calculations give out :

$$\overrightarrow{e_\theta} = \frac{(na_1 + mb_1)\overrightarrow{x} + mb_2\overrightarrow{y}}{\sqrt{(na_1 + mb_1)^2 + (mb_2)^2}}$$
$$\overrightarrow{e_z} = \frac{-mb_2\overrightarrow{x} + (na_1 + mb_1)\overrightarrow{y}}{\sqrt{(na_1 + mb_1)^2 + (mb_2)^2}}$$

and therefore :

$$\phi^{ij} = 2\pi\frac{(na_1 + mb_1)(ia_1 + jb_1) + jmb_2^2}{(na_1 + mb_1)^2 + (mb_2)^2} \quad (4)$$
$$e^{ij} = \frac{b_2 a_1(jn - im)}{\sqrt{(na_1 + mb_1)^2 + (mb_2)^2}} \quad (5)$$

The relations 4 to 5 are used in the energy derivation (see appendix 3.5) and to generate the coordinates of any protein in the final conformation knowing only the minimization parameters, and therefore represent fibers or dimers.

1.3 Description of the algorithm

Now that the parameterization of the problem is complete this section focusses on its numerical implementation.

The algorithm implemented minimizes the total energy of protein zero. This energy encompasses the spring interaction energy characterizing the deformability of a protein, attractive or repulsive interactions between residues of the same protein or in different proteins and steric repulsion between C-terminal residues: $E_{total} = E_{spring} + E_{attra} + E_{rep} + E_{ste}$ (see equations 1 to 3)In order to increase the convergence speed of the conjugate gradient method it is necessary to provide a function computing the derivative of the energy with respect to the minimization parameters. These derivatives are computed in the appendix.

The analytical parameterization highlights all the parameters that must be explored by the algorithm so as to test all the possible fiber configurations and hope to get a global minimum of the energy. They can be divided in two groups :

- The discrete parameters, which correspond to the symmetry group p1 or p2, the folding vector coordinates m and n and the so called interaction partners for each interaction involving a residue of protein zero and one in another protein. These take a finite number of values and are not to be optimized by the algorithm, but rather the minimization has to be run for each set of these parameters.
- The continuous parameters, which are the linearly independent $(x_1, y_1, z_1, ...x_{n-1}, y_{n-1}, z_{n-1}, a_1, b_1, b_2)$ and which are optimized by the conjugate gradient algorithm.

This fixes the codes architecture as follows:

- Definition of all the interactions between residues as modelled from the biological data.
- Initialization of the continuous parameters: proposition of a set of coordinates for the tilling vectors $\overrightarrow{a}$ and $\overrightarrow{b}$, and initialization of the coordinates of the model protein zero as developped in section 1.1.2.
- For a given symmetry group, a given set of coordinates for the folding vector, and a given set of interaction partners, optimization of the continuous

parameters in a method of conjugate gradient fashion through the "$scipy.optimize.fmincg$" python function. Store the final energy and the corresponding minimized parameters, as well as the discrete parameters.

- Repeat the last step for all the combinations of discrete parameters.
- Search for the minimum energy across all sets of discrete parameters. The corresponding parameters describe the most likely fiber configuration and deformed protein configuration.

In the end the minimization step will be carried out using the "$scipy.optimize.fmin_{cg}$" python function which uses a non linear conjugate gradient algorithm, see supplementary materials and [7]. Its inputs are a one dimensional array containing the function to be minimized, the parameters to be optimized, and the derivative of the function to be minimized. It returns the minimized parameters and the minimum value found for the function. However, as the implementation of the gradient function necessary to use the Method of conjugate gradient is still a work in progress, the following results are obtained using a minimization method which doesn't require the analytical computation of the gradient function: the Nelder Mead method (see appendix), implemented in the "$scipy.optimize.minimize$" function of python.

2 Results

The approach described above gives out numerical results presented in this section. In order to observe fiber formation, the native protein must first become assembly competent (see 2.1) and then either form fibers directly or form dimers first (seen 2.2 and 2.3). The same process is followed in the algorithm, and several fiber configurations are tested so as to explore the role of geometrical frustration in fiber formation in section 2.4. The simulations are still running and the results here are a preliminary sample of fiber configurations obtained.

2.1 Assembly competent protein

Let us first describe the configuration of a protein inside a dimer or a fiber of monomers. The biological data described in section 1.1.1 specifies both intra and inter-molecular distances between residues. Depending on if the fiber is formed by monomeric or dimeric Ure2p, this data doesn't have the same interpretation.

- If the fiber is formed of monomeric Ure2p, then the intra-molecular distances designate distances between the residues of the same monomer and inter-molecular distances correspond to distances between the residues of protein zero and the other marked residue in protein zero's interaction partner.
- If the fiber is formed of dimeric Ure2p, then the intra-molecular distances correspond to distances between one residue of the first monomer in the dimer, and another residue in the second monomer of the dimer. Inter-molecular distances correspond to distances between one residue in a dimer, and one residue in another dimer. This adds one level of complexity because it is not possible to biologically distinguish the monomers of a dimer, and therefore the residues of monomer one could be interacting with the ones of either monomer one or two in a different dimer, and the same goes for residues of monomer two. The results presented in the following are obtained with first monomers and secons monomers interacting with first and second monomers respectively, but the other possibilities must be taken into acount as well.

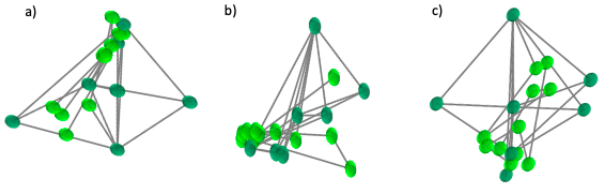

Figure 7. **Three different assembly competent proteins: prion forms of Ure2p inside a fiber. In light green : N-terminal residues, in dark green : C-terminal residues. The numbering corresponds to the one in figure 8**

Figures 7 and 9 respectively represent monomers inside the fibers of figure 8 and the dimers formed using the interactions described above. These figures show that in all the cases presented the crystalline structure of the C-terminal of Ure2p is deformed upon dimer or fiber formation, although the pyramid like shape is conserved. The flexible N-terminal is deformed and wraps around the C-terminal in different fashions. The quantitative deformation of each part of the protein is discussed in section 2.4, but this qualitative observation is consistent with the biological expectation to find crystalline structures in the fibers [1]. Furthermore, assembly competent protein a) differs from the others in that the N-terminal folds in a different direction with respect to the pyramid like C-terminal axis.

2.2 Formation of dimers

This section then focusses on the formation of dimers of assembly competent proteins

The dimer formed with the precursor Ure2p is presented above 9. The comparison with figure 1 in figure 9 highlights the role of the N-terminal part of the protein in dimer and fiber formation, as the full protein dimer has a different conformation than the C-terminal dimer. This dimer is then used as a tiling motive (instead of a monomer) on order to form fibers of dimeric Ure2p. However once again the pyramid-like form of the C-terminal is conserved in the assembly competent monomer and in the dimer.

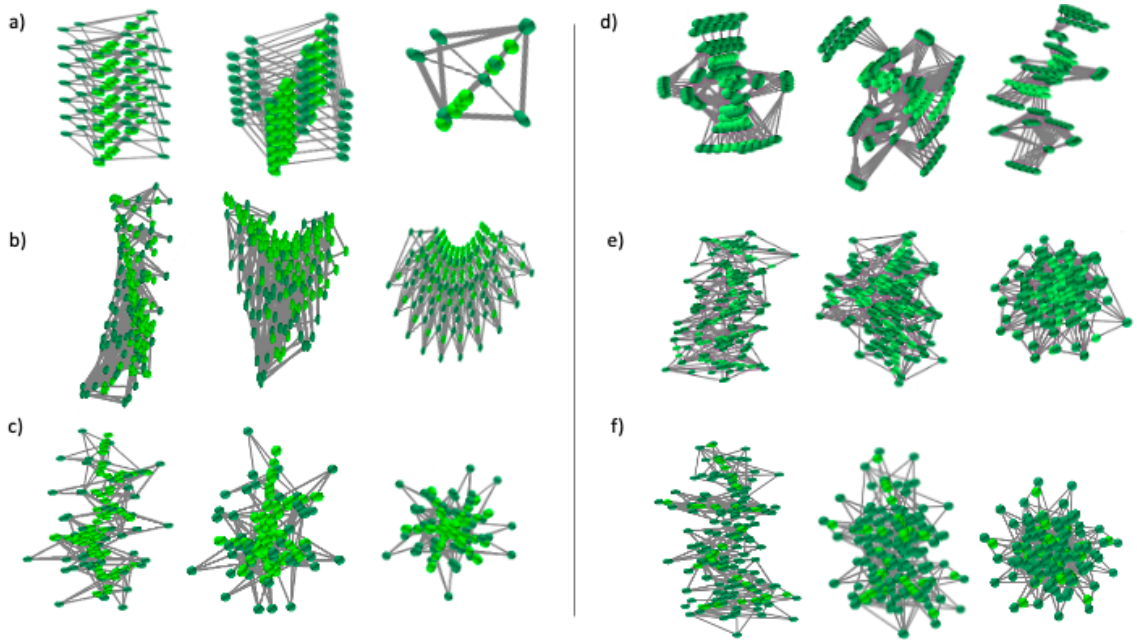

Figure 8. 6 different types of fibers obtained through simulation, each see under three different angles. Each fiber is obtained by optimizing the parameters $(x_1, y_1, z_1, ...x_{n-1}, y_{n-1}, z_{n-1}, a_1, b_1, b_2)$ for different folding vector coordinates m and n, for the wall paper group p_1 and allowing either the closest or the two closest proteins to be interaction partners. In the first panel fibers a),b) and c) are three fibers formed from Ure2p monomers. Specifically, for fiber a) $m = -1$ or 1, $n = 0$, interactions allowed with the closest neighbor. For fiber b) $m = -1$ or 1, $n = 0$, interactions allowed with the two closest neighbors. For fiber c) $m = -1$ and $n = -1$ or $m = 1$ and $n = 1$, interactions allowed with the closest neighbor. In the second panel fibers d),e) and f) are three fibers formed from Ure2p dimers. Specifically, for fiber a) $m = -1$ or 1, $n = 0$, interactions allowed with the closest neighbor. For fiber b) $m = -1$ or 1, $n = 0$, interactions allowed with the two closest neighbors. For fiber c) $m = -1$ and $n = -1$ or $m = 1$ and $n = 1$, interactions allowed with the closest neighbor.

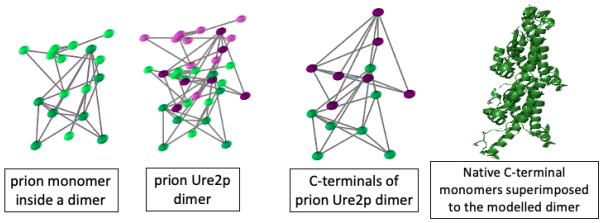

Figure 9. Characteristics of a dimer of Ure2p prions. In light green and light purple : N-terminal residues of protein 1 and 2 respectively, In dark green and dark purple : C-terminal residues of protein 1 and 2 respectively. In the last panel the shape of the C-terminal of a native protein has been extracted thanks to the Pymol software, and superimposed to the C-terminals of the model dimer.

2.3 Formation of fibers

The aim of this section is to describe different types of fibers that can be obtained from monomeric or dimeric Ure2p.

Let us first focus on monomeric Ure2p fibers. Figure 8 represents four fibers obtained. These fibers can be divided into two categories. In fiber a, the N-terminal residues of a protein are aligned in the centre of the fiber and are organized in lines perpendicular to the axis of the fiber. This structure is a reminder of the beta sheet structures that are present in amyloid fibers. The amyloid fiber form is consistent with the litterature [3]. The existence of several forms of proteins is consistent as well since other forms of proteins have been observed [3]. When it comes to fibers formed with prion dimers, two categories of fibers seem to emerge as well with fiber d) of figure 8 being structured differently than the rest. In this fiber the N-terminal seem to align like in fiber a), however this time the beta sheet like chains formed by the N-terminal residue are oriented along the fiber's axis which is not consistent with an amyloid form. In addition, fibers e) and f), along with fiber c) have a star shaped structure, with N-terminal residues organized in several directions.

Figure 10 shows the total energy of a molecule (protein or dimer) inside a (monomeric or dimeric respectively) fiber. This final energy takes into account the energies linked to unsatisfied distance constraints or deformation of the N and C terminals. Overall the final energies observed are lower for proteins in dimeric fibers than for proteins in the monomeric fibers case, even when adding the deformation needed to get a dimer from two monomers. This seems to indicate that forming dimers and then fibers has a lower energetic cost, but further data must be gathered to validate this hypothesis. No clear dependency of the total energy of a protein inside a fiber on the fiber configuration (m, n and interaction partners dependency) arises from this data,

however further simulations must be caried on.

2.4 Geometrical frustration

This section highlights the role of geometrical frustration in the formation of Ure2p fibers.

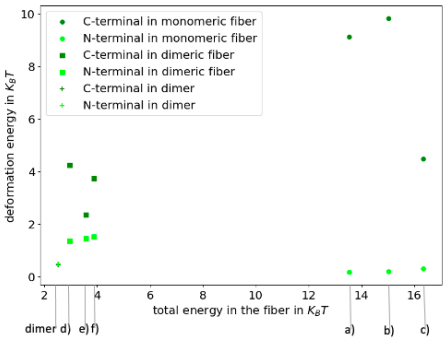

Figure 10. Energy deformation for the C and N protein terminals as a function of the post minimisation energy of the protein in its fiber for fibers of monomers.

Geometrical frustration happens in a constrained system when it is more energetically beneficial for the elements of the system to deform than not to respect the constraints [4]. In the case of this study, if the protein is always forced to deform to accommodate the constraints that come with being in a fiber, then geometrical frustration plays a role in fiber formation. Here geometrical frustration is quantified by the deformation energy, which is the energy necessary to deform the springs linking the residues of the protein. The deformation energy depends on the spring stiffness and is different for the N-terminal and the C-terminal parts of Ure2p. Figure 10 represents the deformation energy of the C and N-terminals of Ure2p for the six configurations presented in figure 8. In all cases the deformation energy of the C-terminal part is higher than the one for the N-terminal part which is highly flexible, emphasizing the role of geometrical frustration as well.

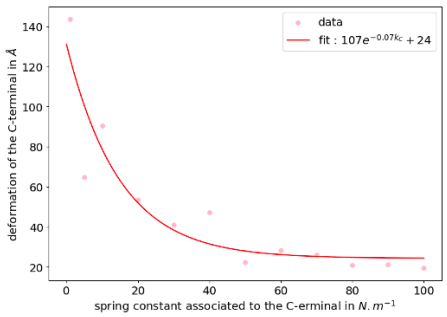

Figure 11. Total deformation of the C-terminal as a function of its rigidity for a given set of discrete paramters and a constant rigidity of the N-terminal.

Figure 11 is obtained by optimizing the energy of a protein for a given fiber conformation, with a constant N-terminal spring constant but for different C-terminal spring constants, and computing for each simulation the total deformation of the C-terminal, meaning the sum of the differences in the distances between residues involved in core interactions before and after optimization. This deformation converges toward a non-zero value, showing that geometrical frustration of the C-terminal happens during fiber formation even for high rigidities. The data is also fitted with an exponential function. The exponent obtained is smaller than one in absolute value, which indicates that the deformation of the C-terminal decreases slowly with its rigidity.

Conclusions

This work presents a method numerically replicating self assembly of Ure2p protein. The algorithm developed is a general one and can be used to test all fiber conformations respecting the identical surroundings constraints for the self assembling elements. It can be used on any protein provided enough data is gathered to impose distance constraints on the residues. When applied to the Ure2p protein, this method helps investigate the formation of fibers of dimeric or monomeric Ure2p.

The optimization method used to minimize the energy of a protein inside a fiber is a Nelder Mead method for now and will eventually be a conjugate gradient method. In any case, the optimization could converge toward a local minimum for the energy rather than the global one. Given the energy dependence in both the coordinates of protein zero and the tilling parameters, it is not possible to compute the complete energy landscape and ensure that the global minimum is reached. However, each simulation is run three times with different initial parameters (a_1, b_1, b_2) each time. In all cases tested all three simulations reach the same minimum which is consistent with it being a global minimum.

The results presented above are coherent with the biological literature [3] since they hint toward the formation of amyloid-like fibers to be energetically beneficial for the Ure2p prion, in the case of monomer fibers. Besides, simulations also show the existence of different possible kinds of fibers as there are small final energy differences between fibers in both the monomeric and dimeric case. Simulations also seem to indicate that proteins in dimeric fibers have a lower total energy and need to deform less to respect the constraints imposed by the biological data. Finally, these results highlight the importance of geometrical frustration in the formation of such fibers, since in all cases geometrical deformation of the rigid C-terminal is observed and since that some deformation happens even for highly rigid structures. The geometrical frustration of the C-terminal observed may seem inconsistent with the biological data stating that there are crystalline forms of the protein in fibers [1]. However, the data doesn't specifiy that the crystalline form observed in fibers is the exact same as the one observed in native protein, and it is possible that the C-terminal, although deformed conserves a crystalline conformation as the pyramid-like shape seems to be conserved.

The upcoming work will consist on improving the algorithm's performances by finishing implementing the gradient function necessary to use a conjugate gradient method. Besides, only few fiber conformations have been tested yet and it is necessary to carry on more simulations to understand better the energetics of Ure2p fiber formation.

References

[1] Carole Gardiennet Yannick Sourigues Christian Wasmer Birgit Habenstein Anne Schütz Beat H. Meier Ronald Melki Anja Böckmann Antoine Loquet, Luc Bousset. Prion fibrils of ure2p assembled under physiological conditions contain highly ordered, natively folded modules. *Journal of Molecular Biology*, 394, 2009.

[2] Belrhali H. Janin J. Melki R. Morera S. Bousset, L. Crustal structure of the globular region of the prion protein ure2p from yeart saccharomyces cervisiae. *Structure*, 9, 2001.

[3] Doucet J Melki R. Bousset L, Briki F. The native-like conformation of ure2p in fibrils assembled under physiologically relevant conditions switches to an amyloid-like conformation upon heat-treatment of the fibrils. *J Struct Biol.*, 2003.

[4] Witten T. Lenz, M. Geometrical frustration yields fibre formation in self-assembly. *Nature Physics*, 13, 2017.

[5] Paulette Decottignies Steven Dubois Pierre Le Marechal Luc Bousset, Virginie Redeker and Ronald Melki. Structural characterization of the fibrillar form of the yeast saccharomyces cerevisiae prion ure2p. *Biochemistry*, 43, 2004.

[6] LLC Schrödinger and Warren DeLano. Pymol.

[7] Jonathan Richard Shewchuk. An introduction to the conjugate gradient method without the agonizing pain. *School of Computer Science Carnegie Mellon University*, 1994.

[8] Komar A.A. Walter S. Buchner J. Cullin C. Melki R. Thual C., Bousset L. Stability, folding, dimerization and assembly properties of the yeast prion ure2p. *Biochemistry*, 40, 2001.

3 Appendix

3.1 Github link to the code used for the simulations

"https://github.com/claradelahousse/Ure2p_fibers.git"

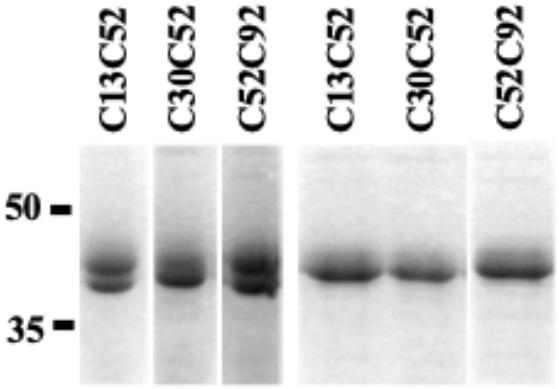

Figure 12. oxidation data obtained by Ronald Melki's group.

3.2 Experiments

3.2.1 Cysteine oxidation experiments

Cysteines are amino acids that can bind together when oxidized. This binding happens when the cysteines are close enough together, which defines a cut-off distance for cysteine-cysteine interactions. In the experiments carried on, some cysteine couples in the Ure2p protein are oxidized. Then the solution of Ure2p proteins passes through a gel, which separates the components of the solution in terms of molecular weight. The results are presented in FIGURE. If one column presents only one line it means that only one conformation of the Ure2p protein is present in the gel. There are several oxidized cysteine couples for which the gel present several lines 12, which means that there are conformation of the Ure2p protein in which this cysteine couple binds together, and some where it doesn't. These observations are coherent with the hypothesis of several soluble protein conformations, the native form and the prion form for instance, as well as the existence of several fiber conformations and therefore several types of polymerized Uer2p.

3.2.2 Crosslinking experiments

A crosslinking experiment is based on the following protocol 13:

- Some residues in the proteins of interest are marked.
- Crosslinkers are introduced in the solution containing the proteins. They will bind to one marked residue at each of their extremities.
- The proteins are denatured and parceled into small pieces.
- The pieces marked by crosslinkers stay together and can be identified

If 2 marked residues are bound by a cross linker, this means that the distance between them in the proteins of interest is bellow the size of the crosslinker. Similarly, if they aren't bound, it is above. The size of the crosslinker serves as a cut-off for the distance between two marked

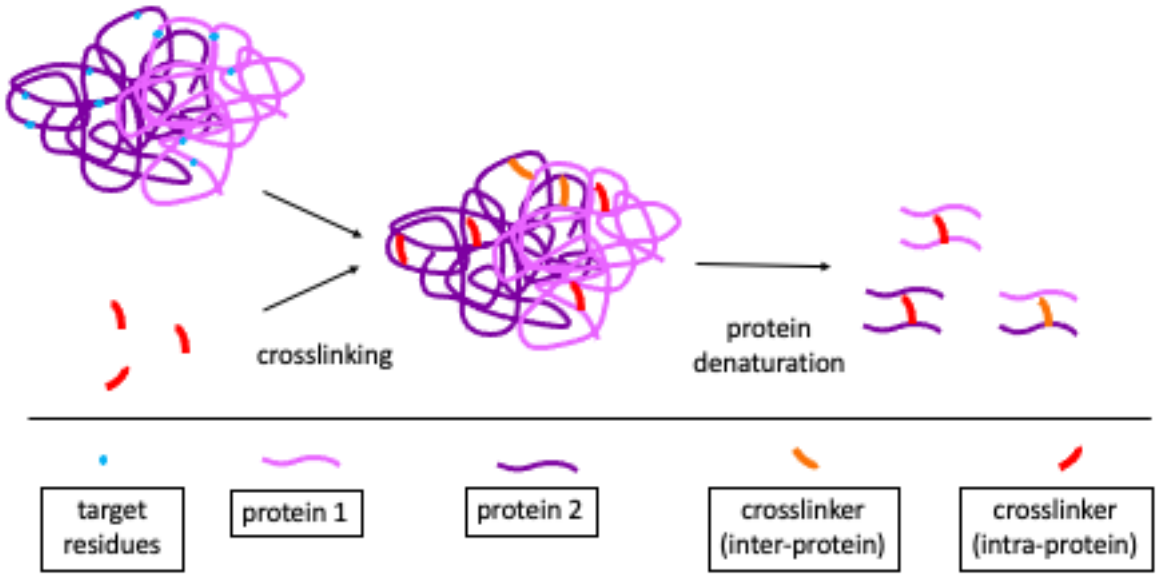

Figure 13. **Crosslinking experiment protocol.**

residues. Crosslinking can be used to refine a cysteine oxidation experiment by providing a clearer cut-ff distance for two residues known to be interacting.

3.3 The Method of Conjugate Gradient

Let us describe the conjugate gradient minimization [7]. This method is effective on systems of the form $A\vec{x} = \vec{b}$, so it can be applied to quadratic forms such as $f(\vec{x}) = \frac{1}{2}\vec{x}^T A\vec{x} - \vec{b}^T x$ which is minimized by the solution of $A\vec{x} = \vec{b}$.

Actually, when one wants to minimize a function f of a vector $\vec{x} = x^i{}_{i\in[1,d]}$ in a space of dimension $d > 1$, one way is to start from a vector $\vec{x}_0$, and take several steps $(\vec{x1}, \vec{x_2}, ...)$ untill $f(\vec{x_n})$ is close enough to 0. At each step k, the error $\vec{e_k} = \vec{x_k} - \vec{x}$ indicaes how far the current sep remains from he solution, and the residue $\vec{r_k} = \vec{b} - A\vec{x_k}$ how far from the correct value of $\vec{b}$. In this case the convergence speed depends on a wise choice of the direction and size of each step. In the Method of Steepest Descent, each step is taken in the direction where the slope is the steepest, meaning the direction x^j for which $\frac{\partial f}{\partial x^j}$ is the smallest. This so called direction of steepest descent can be assimilated to the residue at step k : $r_k = -f'(\vec{x_k}$ The point where the steepest slope at step k reaches zero i.e the point where the steepest slope's direction is orthogonal to the gradient fixes the size of step k.

The Method of Conjugate Directions is a refinement of the Method of Steepest Descent, in which a set of A-orthogonal directions of minimization $(d_i)_{i\in[1,n]}$ is chosen in order to avoid the algorithm taking several stems in the same direction. The A-orthogonality of two vectors $\vec{d_1}$ and $\vec{d_2}$ is defined as: $\vec{d_1}^T A\vec{d_2} = 0$. The condition of A-orthogonality instead of orthogonality ensures that the size of a step can be computed like in the Method of Steepest Descent, by finding the point where the steepest slope reaches 0, and warrants a convergence in n steps as well. To perform this method, it is necessary to find a set of A-orthogonal directions. The Method of Conjugate Gradient is based on a clever choice of such vectors.

Performing the Method of Conjugate Gradient boils down to performing the Conjugate Directions one with search directions constructed to be orthogonal to the residuals. By definition of the step size, at each step the residual is orthogonal to the previous search directions, until the residual is zero and the problem is solved. Residual k is also orthogonal to all the previous residuals, which means that to build the following search direction, the only vector that matters is the search direction of the previous step: there is no need to store the previous search directions which increases available memory.

3.4 The Nelder Mead Method

The implementation of the Gradient function of the conjugate gradient method is a work in progress. In the meantime to obtain fiber formation results another slower energy minimization method is used. The Nelder Mead method is chosen here because it doesn't rely on the use of gradient. Instead, for a problem of dimension N this method tests N+1 values of the function, and orders them based on how close they are to the target value. This collection of values is called a simplex and can be assimilated to a polygon of N+1 sides, that will deform as the optimization is carried on. At each iteration the worst point x_N is moved through a reflection with respect to the center of mass of the simplex.

- If this new position is better than the second worst, but not better that the best, meaning if the value of the function is closer to the target than in the one at point $x_{N-!}$ but not at x_0, then the point is integrated to the simplex.

- If the new position is better than the best, then an

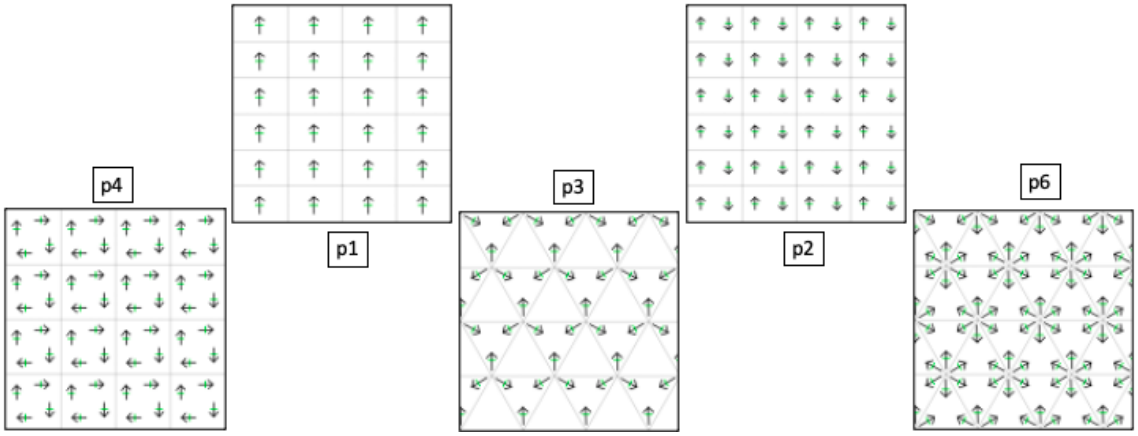

Figure 14. **Symetry groups respecting the identical neighbors constraint on a plane.**

expansion is performed : another step (of predefined length) is taken in the same direction. If the expansion position is better than the reflected one, then it replaces the worst point, else the reflected one replaces the worst point.

- If the new position is worse than the worst point, a contraction is performed : the worst point is replaced by a point in the same direction as the reflection but closer to the centroid.

This process goes on until the function reaches a value close enough to the target, with a predefined tolerance.

3.5 Relevant Wall paper groups

The main experimental result about a protein's surroundings is that all the proteins in a fiber have the same immediate environment : they all have the same neighbor configuration. There can be several configuration that respects this constraint. As a matter of fact, forming a fiber can be seen as folding a 2D lattice around a cylinder. In this case one way of listing all the possible configuration is to understand which 2D lattices still respect the identical environment constraint once folded onto a cylinder.

First let us focus on the case of a 2D lattice tiled with proteins. All proteins being in the same environment implies that the tiling must be invariant by translation from one protein to another. Therefore this lattice can be considers as a mathematical wallpaper, which is an object covering a plane so that the drawing is unchanged under certain isometries REF. Here these unchanged isometries must coincide with the ones used to tile the plane, i.e to go from one protein to the next. Wallpapers are classified into 17 groups depending on the transformation group they leave invariant. There are 4 times of planar isometries:

- translation: the pattern remains identical when shifted from one or several units
- rotation: the patter remains identical when rotated around a point
- reflection: the pattern remains identical when flipped across an axis
- glide reflection: the pattern remains identical when flipped then shifted.

Among those only the translations and rotations are compatible with there being only fully identical proteins. This reduces the number of possible planar symmetry groups to 5 (figure 14:

- p1: a parallelogrammatic lattice with two translation axes
- p2: a parallelogrammatic lattice with two translation axes and four second degree (180 degrees) rotations
- p4: parallelogrammatic lattice with two translation axes, four second degree (180 degrees) rotations and one fourth degree (90 degrees) rotation.
- p3: a triangular lattice with seven third degrees (120 degrees) rotations
- p6: a triangular lattice with six second degrees (180 degrees) rotations, six third degree (120 degrees) rotations and one sixth degree (60 degrees) rotation.

Next, it is important to consider which of these 4 groups still respect the identical environment constraint when folded onto a cylinder, and under which conditions. The folding of the plane results in the distances being stretched in all directions except the one of the axis of the cylinder. If the distances between proteins are stretched unevenly, then the identical environment constraint which was respected on the 2D plane does not hold anymore. Actually, this issue arises for triangular lattices where only some proteins are stretched. It also arises in the p4 symmetry group as the proteins are aligned in two orthogonal directions. Lastly, the constraint holds in the p1 and p2 symmetry groups (figure 6).

3.6 Analytical model

The convergence speed of the "$scipy.optimise.fmin_cg$" funciton increases when provided with an analytical expression of the derivative of the minimisation function. These derivatives are computed as follow :

- Derivatives with respect to protein zero coordinates

$$\frac{\partial E_h}{\partial x_a^{00}} = k\frac{d-d_0}{d}((x_a^{00}-x_b^{ij})-(x_a^{00}-x_b^{ij})\frac{\partial x_b^{ij}}{\partial x_a^{00}}-(y_a^{00}-y_b^{ij})\frac{\partial y_b^{ij}}{\partial x_a^{00}})$$
$$\frac{\partial E_h}{\partial y_a^{00}} = k\frac{d-d_0}{d}((y_a^{00}-x_b^{ij})-(y_a^{00}-y_b^{ij})\frac{\partial y_b^{ij}}{\partial y_a^{00}}-(x_a^{00}-x_b^{ij})\frac{\partial x_b^{ij}}{\partial y_a^{00}})$$
$$\frac{\partial E_h}{\partial z_a^{00}} = k\frac{d-d_0}{d}((z_a^{00}-z_b^{ij})-(z_a^{00}-z_b^{ij})\frac{\partial z_b^{ij}}{\partial z_a^{00}})$$

- Derivatives with respect to tilling vectors coordinates :

$$\frac{\partial E_h}{\partial a_1} = k\frac{d-d_0}{d}((x_a^{00}-x_b^{ij})\frac{\partial x_a^{00}}{\partial a_1}-(x_a^{00}-x_b^{ij})\frac{\partial x_b^{ij}}{\partial a_1}+(y_a^{00}-y_b^{ij})\frac{\partial y_a^{00}}{\partial a_1}-(y_a^{00}-x_b^{ij})\frac{\partial y_b^{ij}}{\partial a_1}+(z_a^{00}-z_b^{ij})\frac{\partial z_a^{00}}{\partial a_1}-(z_a^{00}-z_b^{ij})\frac{\partial z_b^{ij}}{\partial a_1})$$
$$\frac{\partial E_h}{\partial b_1} = k\frac{d-d_0}{d}((x_a^{00}-x_b^{ij})\frac{\partial x_a^{00}}{\partial b_1}-(x_a^{00}-x_b^{ij})\frac{\partial x_b^{ij}}{\partial b_1}+(y_a^{00}-y_b^{ij})\frac{\partial y_a^{00}}{\partial b_1}-(y_a^{00}-x_b^{ij})\frac{\partial y_b^{ij}}{\partial b_1}+(z_a^{00}-z_b^{ij})\frac{\partial z_a^{00}}{\partial b_1}-(z_a^{00}-z_b^{ij})\frac{\partial z_b^{ij}}{\partial b_1})$$
$$\frac{\partial E_h}{\partial b_2} = k\frac{d-d_0}{d}((x_a^{00}-x_b^{ij})\frac{\partial x_a^{00}}{\partial b_2}-(x_a^{00}-x_b^{ij})\frac{\partial x_b^{ij}}{\partial b_2}+(y_a^{00}-y_b^{ij})\frac{\partial y_a^{00}}{\partial b_2}-(y_a^{00}-x_b^{ij})\frac{\partial y_b^{ij}}{\partial b_2}+(z_a^{00}-z_b^{ij})\frac{\partial z_a^{00}}{\partial b_2}-(z_a^{00}-z_b^{ij})\frac{\partial z_b^{ij}}{\partial b_2})$$

With E_h corresponding to any harmonic energy with a spring constant k ans assuming that residue b of protein ij is in interaction with residue a of protein zero, and with :

$$\frac{\partial x_0^{00}}{\partial a_1} = \frac{n(a_1n+b_1m)}{(\sqrt{(a_1n+b_1m)^2+b_2^2m^2})}$$
$$\frac{\partial x_0^{00}}{\partial b_1} = \frac{m(a_1n+b_1m)}{(\sqrt{(a_1n+b_1m)^2+b_2^2m^2})}$$
$$\frac{\partial x_0^{00}}{\partial b_2} = \frac{m^2b_2}{(\sqrt{(a_1n+b_1m)^2+b_2^2m^2})}$$
$$\frac{\partial \phi^{ij}}{\partial a_1} = 2\pi * \frac{(nb_1j+ia_1)+i(a_1n+b_1m)}{(a_1n+b_1m)^2+b_2^2m^2} - \frac{2n(a_1n+b_1m)(b_2^2jm+(b_1j+ia_1)(a_1n+b_1m))}{((a_1n+b_1m)^2+b_2^2m^2)^2}$$
$$\frac{\partial \phi^{ij}}{\partial b_1} = 2\pi * \frac{((j(a_1n+b_1m)+m(b_1j+ia_1))}{((a_1n+b_1m)^2+b_2^2m^2)} - \frac{2m(a_1n+b_1m)(b_2^2jm+(b_1j+ia_1)(a_1n+b_1m))}{(a_1n+b_1m)^2+b_2^2m^2)^2}$$
$$\frac{\partial \phi^{ij}}{\partial b_2} = 2\pi * \frac{2a_1b_2m(jn-im)(a_1n+b_1m)}{((a_1n+b_1m)^2+b_2^2m^2)^2}$$
$$\frac{\partial z^{ij}}{\partial a_1} = \frac{b_2m(jn-im)(a_1b_1n+m(b_1^2+b_2^2))}{((a_1n+b_1m)^2+b_2^2m^2)^{\frac{3}{2}}}$$
$$\frac{\partial z^{ij}}{\partial b_1} = \frac{-(nb_2ja_1-mb_2ia_1)m(na_1+mb_1)}{((a_1n+b_1m)^2+b_2^2m^2)^{\frac{3}{2}}}$$
$$\frac{\partial z^{ij}}{\partial b_2} = \frac{a_1(jn-im)(a_1n+b_1m)^2}{((a_1n+b_1m)^2+b_2^2m^2)^{\frac{3}{2}}}$$