The Search for New Proteins: From Data Mining to Expression

Cruz Tetlalmatzi Sofía
Advisor Daniel Cox

August 13th 2015

The development of biotechnology lets us engineer proteins to use in practical applications by taking advantage of the Central Dogma of Molecular Biology. β-solenoids have proven to have application potential with current works on fibril formation, nanoparticle deposits, 2D arrays and supercapacitors. Applications make it necessary to expand the known catalogue of useful proteins. To satisfy this purpose we have followed the whole process from data mining, through simulation, onto initial synthesis and characterization. We ended with two square β-solenoids with application potential, one found in data mining with initial unknowns structure, $2BM6_P$, and one with known structure but with no previous experimental testing, 3DU1.

Introduction

The Central Dogma of molecular biology states that the general flow of information in a cell goes from DNA to RNA to proteins. This is, genes store information in DNA, which is transcribed into RNA, as a messenger from the nucleus to the rest of the cell, and then translated into the amino acids who then fold to form a protein.

By differences in physical and chemical composition, proteins then serve a range of functions that cover nearly every aspect the cell must tend to. Living systems use 20 amino acids as the building bricks of the whole range of proteins. Every amino acid has the same basic chemistry: there is a central carbon whose four attachments are a hydrogen (H), a carboxyl group (COOH), an amino group (NH_2) and a variable radical, side chain, (R), Figure 1. The C-terminus and N-terminus happen to possess the ability to form a peptide bond, forming a backbone. Thus, amino acids can polymerize to form peptides or proteins.

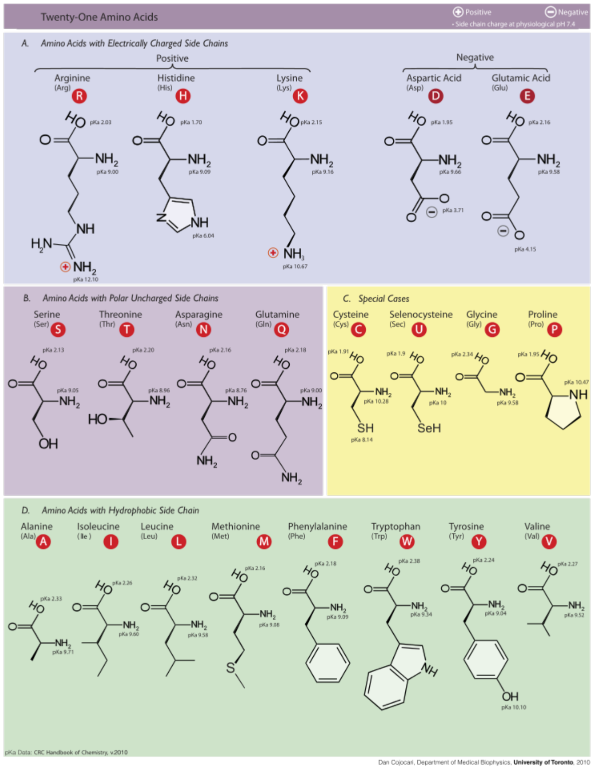
As the chain gets longer, the backbone chain begins to interact with itself forming hydrogen bonds, bending. The way it folds depends on the radicals, on how they bend by interacting with each other and with the environment. The two basic ways to fold are in a spiral by coiling, α-helix, or a plane by rows ,β-sheet, Figure 2.

A β-solenoid is a combination of the two. The chain is coiled but the periods are lengthy enough to form sheets on the sides; the final geometry is a prism with number of sides depending on the sequence.

Taking advantage of the Central Dogma, biotechnologies have developed applications for engineered proteins. β-solenoids are interesting in this context, for they have shown a relative easiness to polymerize, forming fibrils of micron length. The fibrils are the first step to applications, for they can serve as templates to deposit to grow nanoparticles; there are also approaches to organize the fibrils in rows and stacking them to form a 2-dimensional lattice. [1]

As an example, an approach with direct everyday application is looking to use proteins in the model of a supercapacitor. Supercapacitors look to have big surface area to electrodes with small distance between mobile ions and electrodes to enable high values for charge and energy storage. Using graphene as a conductor, previous attempts have managed a big surface area but struggled in minimizing distance. By using β-solenoids to keep the graphene sheets in place, shorter distances are obtained, Figure 3, boosting the capacitor's storage. It is hoped that low cost, high energy density supercapacitors can be realized that will be more competitive with batteries. Moreover, these supercapacitors will be manufactured in a more sustainable manner, offer higher safety than batteries, and contain no toxic components.

With such practical applications, expanding the β-solenoid library is important. In this paper it is shown how the search for new useful proteins is done, from the very beginning of finding a new interesting proteins,



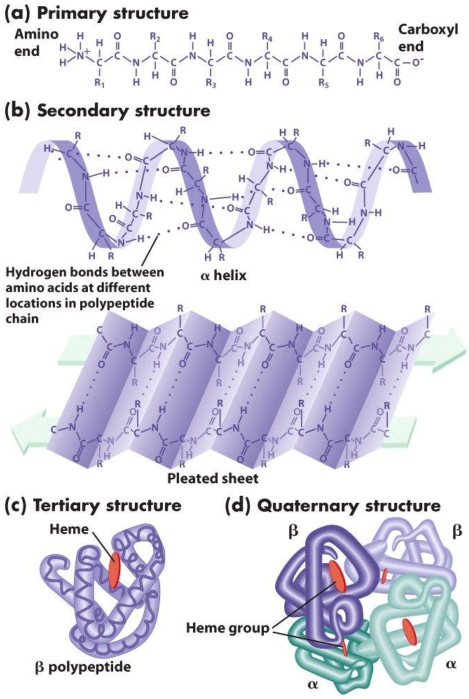


Figure 2: Basic protein structures. [3]

Figure 1: 20 amino acids used in biological systems. [2]

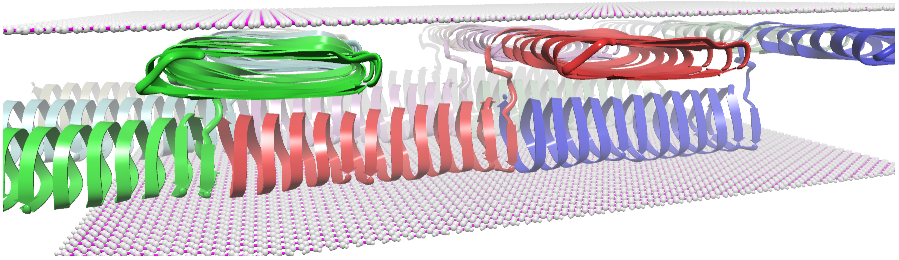


Figure 3: β-solenoids between graphene sheets give a supercapacitor both small distances between mobile ions and electrodes, and big surface area to electrodes.

working our way through all the way to testing its viability to be handled in the lab, seeking the most basic form of protein manipulation, fibril polymerization.

Taking it one step at a time

0.1 Data Mining

Protein research is extremely common nowadays and proteins are roughly as numerous as genes, meaning there is at least one protein per function per species, and that is a lot. Millions of proteins have been discovered, analysed and their sequence identity uploaded to the science community at different molecular biology databases. Relatively, only a few of these have an official structure in PDB (Protein Data Base) [4], an open source repository for atomic resolution structures for proteins when they are discovered . Initially, all the proteins worked with for application by our group are included in PDB. In order to expand our knowledge of protein structures for basic and applied research, first we have to find them.

Mining for proteins and genes has been a major issue for molecular biologists. Database computational methods have been successfully implemented, mostly taking advantage of both the genetic and the proteic codes. The most common searcher, BLAST (Basic Local Alignment Search Tool) , uses local alignment to find homologous sequences to an initial given sequence. Local alignment means the searcher breaks the input into short words (3 amino acids long), then for each triplet it finds sequences that match it, at the matching spot it expands to adjacent words on both directions and stops expanding when the algorithm detects a low enough score for similarity. [5]

We are specially interested in this, given that we look to expand our set of proteins with a β-solenoid structure. As the structure is defined by the sequence, we looked for sequences that shared enough homology to the worked proteins to maintain the β-solenoid, but different enough for it to be worth while working with.

The search conducted by BLAST went through three filters first, items with less than 50% likelihood were discarded, then only those with a distinct different function from the original and each other survived, of the remaining, the most different one from the original structure was chosen to be worked with. These are shown in Table 1.

PDB ID	Characteristics	Dif. Function Homologous	Chosen Protein
2BM6	Antibiotic Resistace Mycobacterium	6	Peptidase Cyanobacteria ($2BM6_P$)
1THJ	Anhydrase Methanosarcina	8	Sulfate Cystobacter ($1THJ_S$)
1EZG	Antifreeze Beetle	5	Collagen-like Oyster ($1EZG_C$)
3ULT	Antifreeze Ryegrass	11	Zonadhesin Mallard ($3ULT_Z$)
1M8N	Antifreeze Budworn	0	–
1HV9	GlmU E . coli	0	–
1HM0	Amide Streptococcus	0	–
2PNE	Antifreeze Snowflea	0	–

Table 1: Sequences homologous to proteins with known structure

Notice how for half the proteins searched for are unique to our best knowledge. This is, no one has discovered any protein that is functionally distinct from the original one.

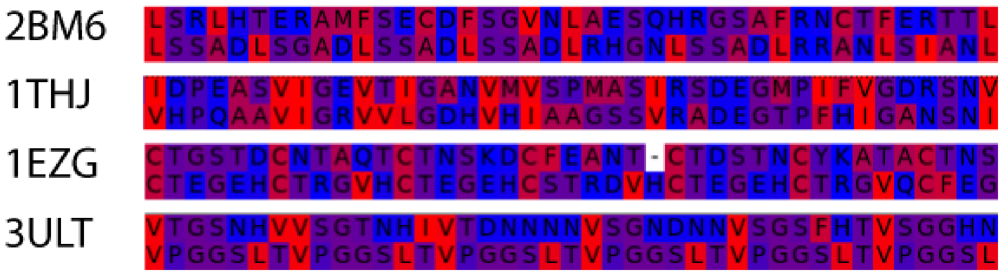
0.2 Structure

A considerable amount of official protein structures have been discovered by experimental methods, namely rigorous X-ray crystallography or nuclear magnetic resonance techniques as can be seen in every PDB validation report. Nonetheless, the quantity of functionally different proteins exceeds the experimental capacity of structurally characterizing every single one. Thus, efforts have been made to figure out structures with less dependance on experimental methods. Here is where molecular dynamics simulation comes in.

The solution to this new problem is far from trivial. The folding of the amino acid chain is very sensitive to the orientation amino acids are given due to their mutual interactions, i.e. every atom counts. In average an amino acid is formed by 19 atoms, and proteins are formed formed by chains of a couple of hundreds to a few tens of thousands amino acids, often coupled in groups. This makes it statistically impossible to determine the native structure from a given random sequence of amino acids.

Instead of trying to figure out the whole structure from scratch we have taken to our advantage the homology of our sequences. As we can see in Figure 4, β-solenoids preserve a quasi-cyclic sequence bounded by hydrophobic amino acids (red). The periodicity of these amino acids is preserved in the homologous sequence. Thus, using JACKAL: A Protein Modeling Package, the amino acids in between the periodic hydrophobes have been mutated one period at a time. For each step, JACKAL will output a few hundred random conformations along the energy space and work its way to the native structure through energy minimization as seen in Figure 5.

By doing this, we restrict the unknown structure puzzle to a few amino acids at a time, whilst making sure the outcome preserves hydrophobes facing inward. For, it is known that the common medium in which proteins are suspended is water and, having a hydrophobic group facing inwards is energetically more favourable than otherwise. The resulting structure we can see in Figure 6.



2BM6
LSRLHTERAMFSECDFSGVNLAESQHRGSAFRNCTFERTTL
LSSADLSGADLSSADLSSADLRHGNLSSADLRRANLSIANL

1THJ
IDPEASVIGEVTIGANVMVSPMASIRSDEGMPIFVGDRSNV
VHPQAAVIGRVVLGDHVHIAAGSSVRADEGTPFHIGANSNI

1EZG
CTGSTDCNTAQTCTNSKDCFEANT-CTDSTNCYKATACTNS
CTEGEHCTRGVHCTEGEHCSTRDVHCTEGEHCTRGVQCFEG

3ULT
VTGSNHVVSGTNHIVTDNNNNVSGNDNNVSGSFHTVSGGHN
VPGGSLTVPGGSLTVPGGSLTVPGGSLTVPGGSLTVPGGSL

Figure 4: Sequence snippet of original protein (above) and homologous protein (below) with amino acids marked by hydrophobicity (polar in blue, hydrophobes in red).

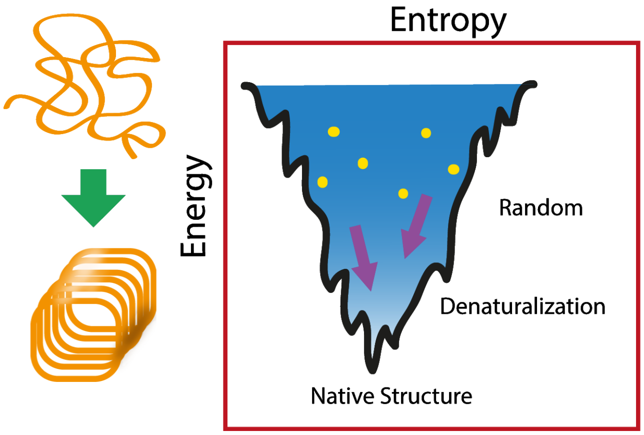


Figure 5: For every mutated segment, JACKAL will start with several random configurations of the new amino acids and find its way to the native structure through energy minimization.

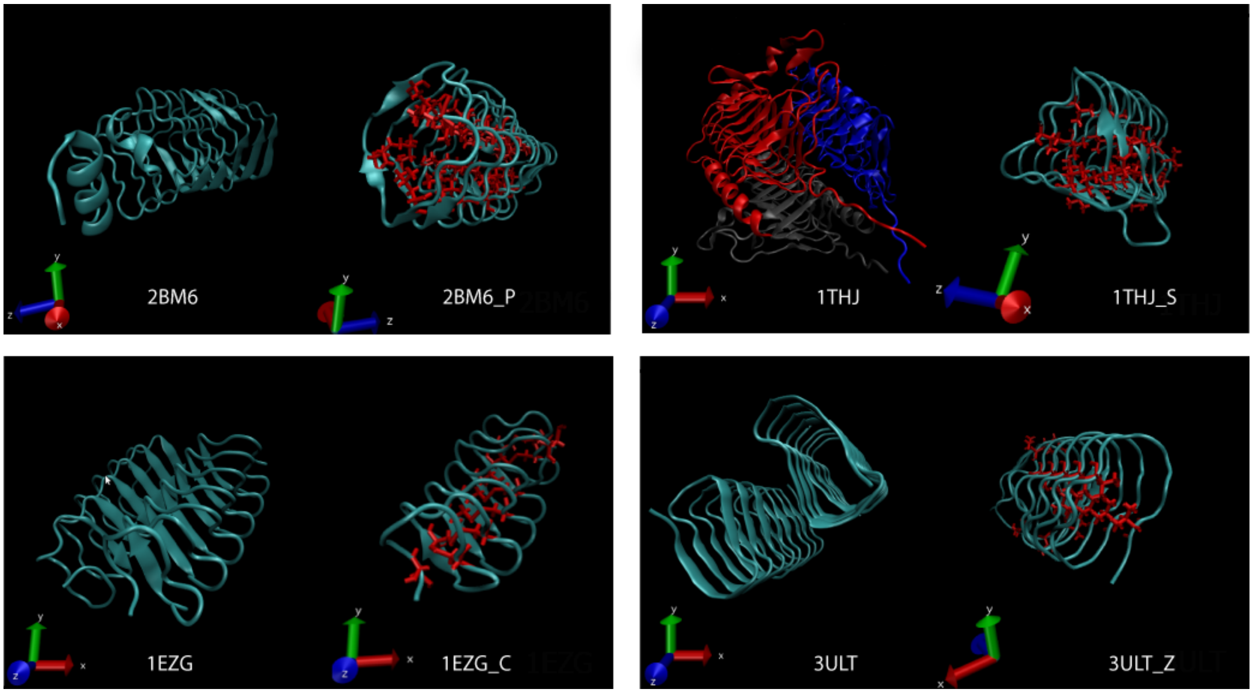


Figure 6: Structure of original protein (left) and of homologous protein (right). Structures in ribbon representation, diagram traces backbone and excludes bonds. Hydrophobic amino acids have been highlighted in red on the homologous protein.

0.3 Stability

Once having a proposed structure, we must ensure its stability. Meaning that it must be tested if under strains of its usual environment the hypothetical structure preserves its conformation without unfolding. This is translated to simulating how the forces exerted between atoms in the structure and the solvent (water) interact. So far there has only been a classical approach so in terms of mathematics, we have to solve:

$$F = ma$$

For the following potential: [6]

$$\begin{aligned} V = & \\ & \sum_{\text{bonds}} k_b \left(b - b_0\right)^2 + \sum_{\text{angles}} k_\theta \left(\theta - \theta_0\right)^2 + \sum_{\text{dihedrals}} k_\phi \left[1 + \cos\left(n\phi - \delta\right)\right] \\ & + \sum_{\text{impropers}} k_\omega \left(\omega - \omega_0\right)^2 + \sum_{\text{Urey-Bradley}} k_u \left(u - u_0\right)^2 \\ & + \sum_{\text{nonbonded}} \epsilon \left[\left(\frac{R_{\text{min}ij}}{r_{ij}}\right)^{12} - \left(\frac{R_{\text{min}_{ij}}}{r_{ij}}\right)^{6}\right] + \frac{q_i q_j}{\epsilon r_{ij}} \end{aligned} \tag{1}$$

For every atom in the system.

Multiple forces are considered, the set of these is called a *force field* and are represented in Equation 1 . The forces taken into account are the following: The first term for bond force between atoms; second, the angle force formed between three atoms; third, the dihedral angle force between four atoms; fourth, the impropers' out of plane bending; fifth, the Urey-Bradley component of bending due to interaction with atoms two bond apart. These first four of these terms are shown in Figure 7.The last two terms account for nonbonded (3 bonds apart or more) interactions, which are the Van der Waals energy modeled with standard Lennard-Jones potential , Figure 8, and finally the Coulomb electrostatic potential. The k_i are constants determined to favour the dynamics for a given molecule, different force fields have different k_i.

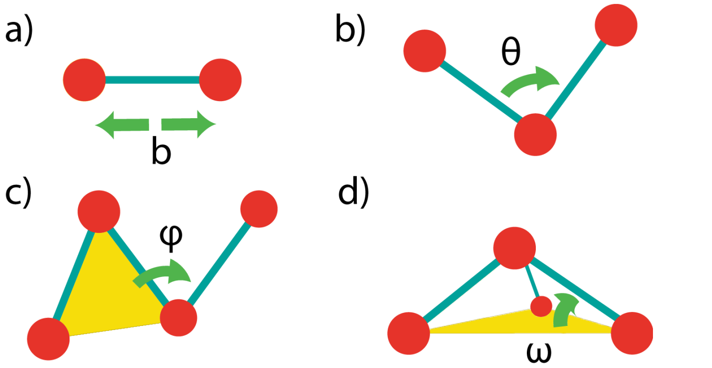


Figure 7: Forces taken into account in molecular dynamics simulation. a) bonds, b) angle, c) dihedrals, d) impropers. The yellow space between three atoms indicates the plane they form.

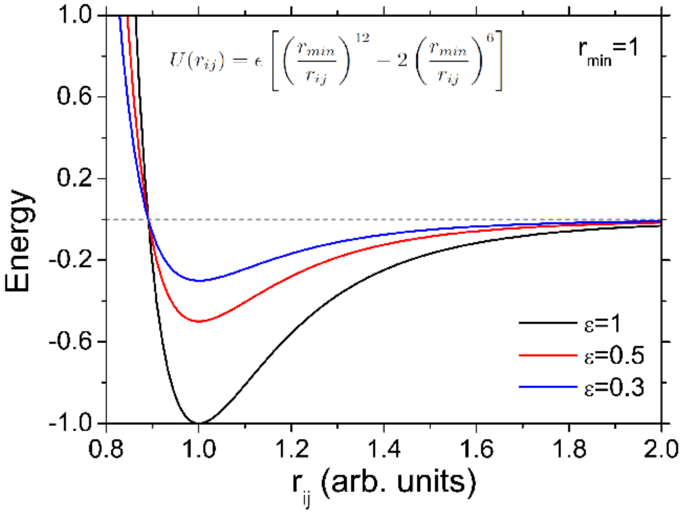


Figure 8: Standard Lennard-Jones potential for Van der Waals energy. [7]

The simulation is conducted using GROMACS (GROningen MAchine for Chemical Simulations). We will simulate the proteins floating in water for enough time to consider it stable (10 ns by convention). First we must choose the force field to implement. As this depends on the type of protein to be simulated, it has been tested that β-solenoids are better simulated under CHARMM27 force field.

Then the environment must be created, this is a cubic box filled with water and with added ions,usually Na^+ or Cl^-, to neutralize the charge of the whole system. Afterwards, the molecule's structure must be relaxed, for this we have minimized the energy of the protein to be less than 1000 KJ/mol/nm. The reduction of energy can be seen in Figure 9.The simulation of the structure will occur at room temperature and the molecules of the structure must be excited to match this temperature. Thus, there is a gradual linear heating from 0K to 300K. It must be noted that the temperature is only determined by the velocity of the molecules and no phase transitions or non classical phenomena for low temperatures are considered. [8]

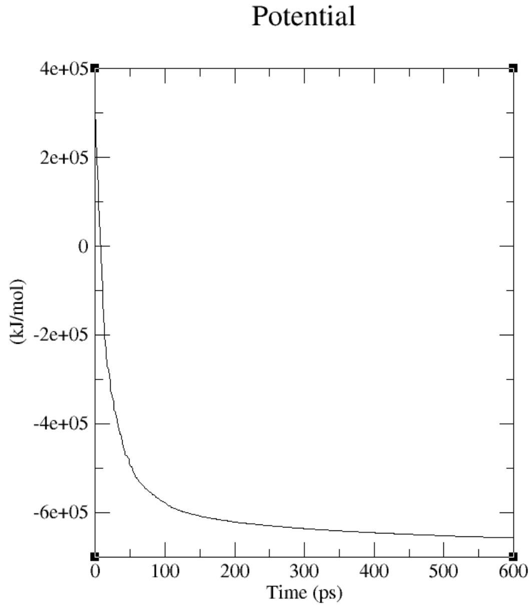


Figure 9: Potential reduction due to energy minimization, relaxation of the structure.

Last, the solvent must be relaxed around the solute to match the initial condition and subsequent interactions of the structure during the simulation. For this equilibration, two canonical ensembles must be employed, isochoric-isothermal (NVT; fixed number of particles N, volume V and temperature T) and isobaric-isothermal (NPT; fixed number of particles N, pressure and temperature T). In other words, we look to first stabilize temperature and then pressure [9].

Temperature is a purely statistical measure so stabilization is done with a thermostat. This is a method to expand the system beyond its boundaries to add or remove energy, approximation the canonical ensemble. The energy fluctuation then makes N-particle system have an average constant temperature, rather than a fixed temperature overall. To introduce the extra energy, the Nosé-Hoover model used a Lagrangian with additional, artificial coordinates and velocities. We use the Nose-Hoover thermostat given it has an advantage of getting the temperature in there reversibly.

$$L_{NH} = \sum^{N} \frac{m_i}{2} s^{22}_{\ i} - \phi(\mathbf{q}) + \frac{Q}{s}\dot{s}^2 - gkTlns \tag{2}$$

$$\mathbf{q}_i' = \mathbf{q}_i \ , \ \mathbf{p}_i' = \mathbf{p}_i/s \ , \ dt' = \frac{dt}{s} \tag{3}$$

The Nose-Hoover Lagrangian, Equation 2, is written for a N-particle system, with m_i masses, Q the effective mass and potential energy ϕ. Where the primed variables from Equation 3 are the real variables for position, momenta and time, respectively, the unprimed variables are the virtual coordinates and s an additional degree of freedom acting as an external system on the simulated system. [10]

Now our system is ready to run the simulation. After running the simulation, only one test is needed to tell whether the structure is stable or not. The Root Mean Square Deviation (RMSD) of atomic positions of the backbone chain indicates how much the structure was modified from the initial condition, omitting translation and rotation.

$$RMSD(\mathbf{x}_i(t), \mathbf{x}_i^0) = \sqrt{\frac{1}{N}\sum_{i=1}^{N} \mid U\{\mathbf{x}_i(t) - \mathbf{x}_C(t)\} - \{\mathbf{x}_i^0(t) - \mathbf{x}_C^0(t)\} \mid^2} \tag{4}$$

RMSD is calculated using Equation 4 , where U is the optimal transformation $U^{\{\mathbf{x}_i(t)\}\rightarrow\{\mathbf{x}_i^0\}}$ via rotation and translation that superimposes the $X(t)$ structure with X^0 structure and, where x_C and x_C^0 are their respective centers of geometry [11]. If the RMSD is in average constant, then the structure is stable.

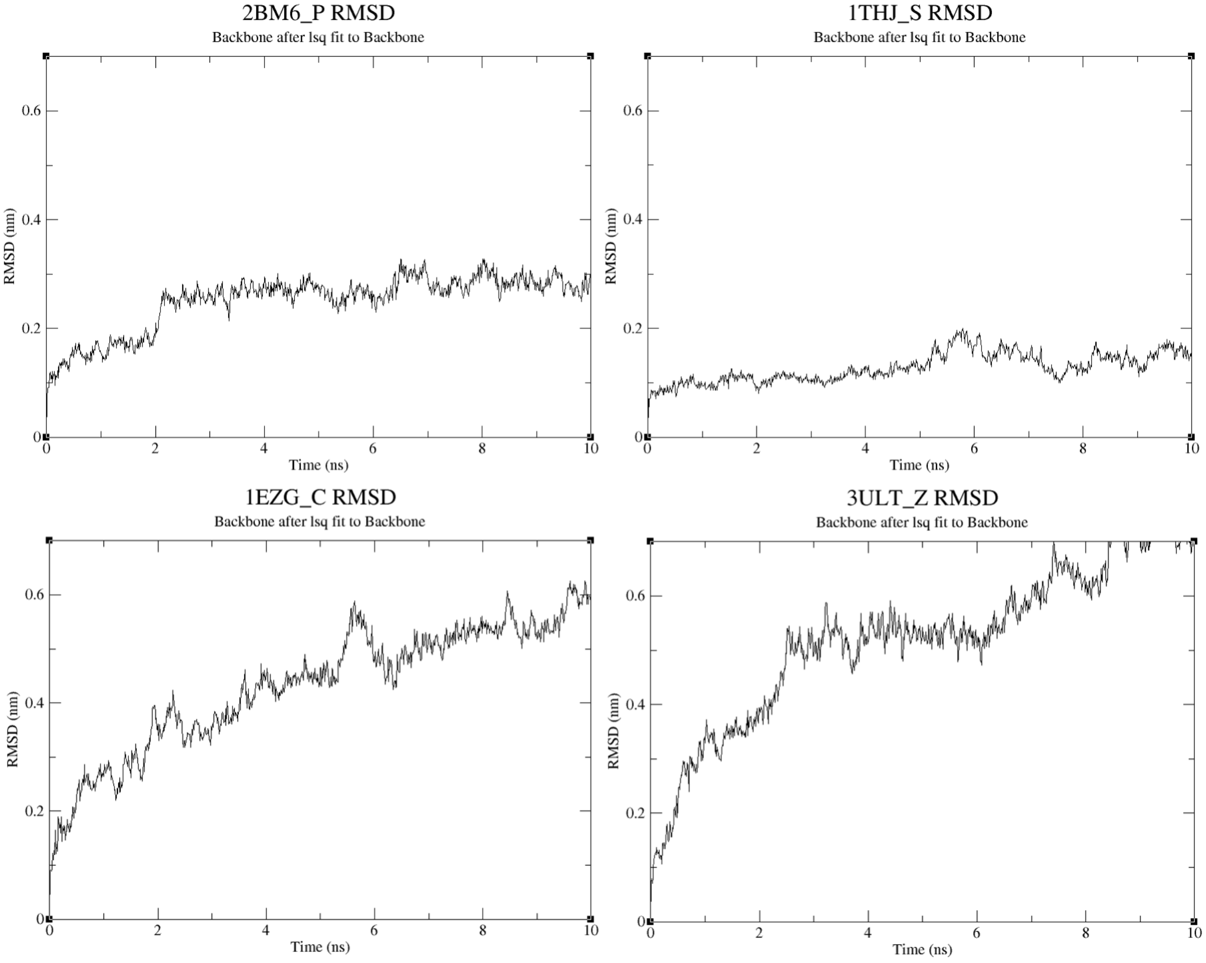


Figure 10: Root Mean Square Deviation (RMSD) of all homologous proteins. $2BM6_P$ and $1THJ_S$ achieved a stable plateau, whereas $1EZG_C$ and $3ULT_Z$ didn't.

It's easy to see in Figure 10 that both the homologous proteins to 2BM6 and 1THJ reached a plateau, meaning that after modifying their initial geometry they settled to a stable conformation. Whereas, the ones for 1EZG and 3ULT present a slope that exceeds the conventional measure of stable structure, of 0.5 nm, and is nowhere near stable. Looking at the structure through the simulation, and as we can appreciate in Figure 11, it was noticed that $3ULT_Z$ uncoiled as if being a spring pulled at beginning and end, as this protein's sequence is strictly periodic, this undoubtedly indicates that this β-solenoid geometry does not correspond to this sequence's native structure. On the other hand, $1EZG_C$ protein uncoiled around a specific point, dividing the solenoid in two. After pinpointing where the break happened, an irregularity in the sequence was detected. This corresponds to the amino acids in the 4th cycle seen in Figure 4; this period is 7 amino acids long instead of 6.

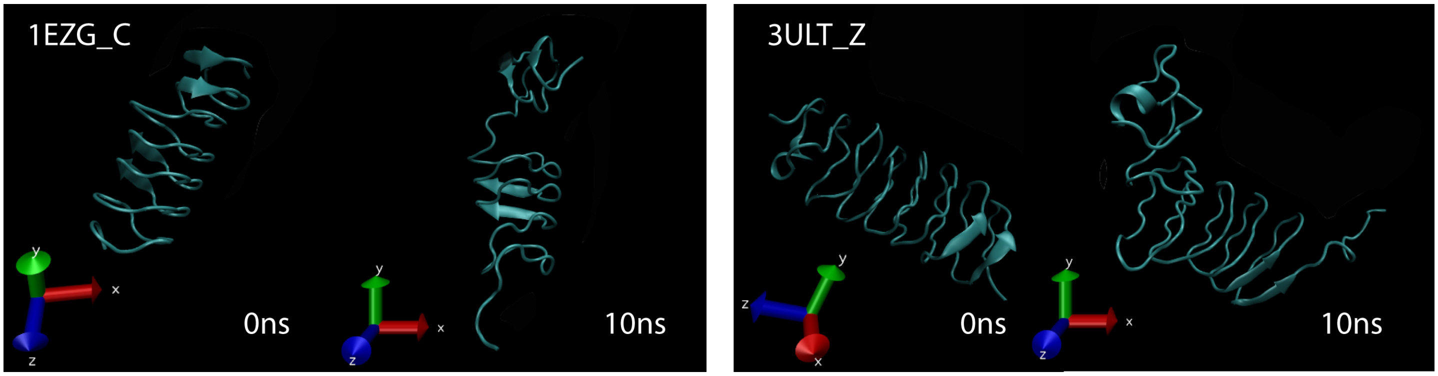


Figure 11: Uncoiled structures of $1EZG_C$ and $3ULT_Z$ before simulation and after 10 ns simulation.

To test the hypothesis that without the change in period length the structure would be stable, another simulation was conducted changing this cycle by eliminating the extra serine (S).

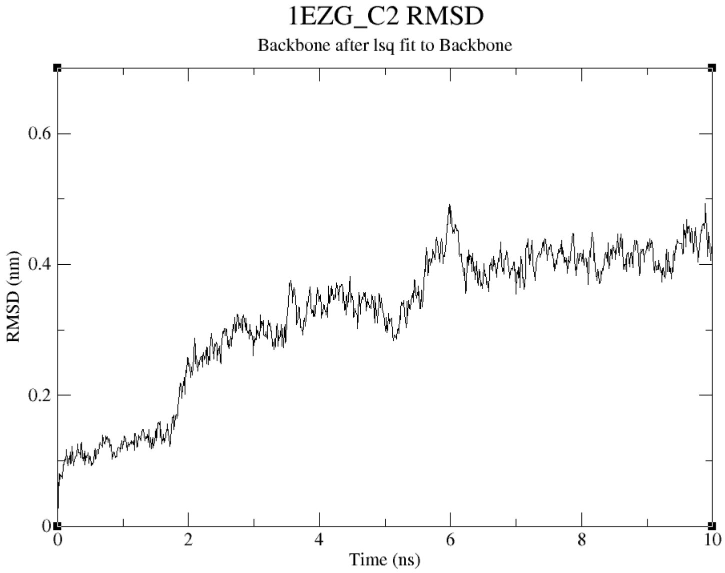


Figure 12: RMSD for $1EZG_{C2}$; $1EZG_C$ with the deleted serine (S).

As we can now observe in Figure 12, the structure obtains a stable plateau though, the multiple step behaviour might lead to further uncoiling. By analyzing the simulation movie it is noted again a fault in the same region as before. Both this and the previous behaviour indicate another kind of coiling correspond to that region and is very probably linked to the function of the protein.

Having some structures which passed the stability test, before working with them in the lab, the potential proteins must be biochemically compatible to the bacterias that are going to express them and the structure must be double checked.

Of the three candidates, two had to be discarded by biochemical reasons. 1THJ has proved to be difficult to work with in past laboratory attempts, not folding correctly and in consequence, not forming polymers. 1EZG is hard to work with as bacteria are not prone to handle proteins held together by disulphide bonds cysteines (C) use, due to the internal reductant composition of the cell. $2BM6_P$ turned out to be a reliable and interesting structure because its homologous is bacterian, as opposed to eukaryotic like 1EZG, which represents an advantage for expression. Also, 2BM6 resembles a protein previously worked with inconclusively, 3DU1 Figure 16. 3DU1 is a differential regulation protein in Nostoc Cyanobacteria. Both proteins have a four sided β-solenoid geometry and are bounded by Leucines 15.

So, $2BM6_P$'s structure must be double checked before proceeding. The structure is revised by going one step further in how realistic the simulation is. This is achieved by assigning a Maxwell-Boltzmann distribution of initial velocities, corresponding to 300 K (Figure 13), for a finite number of seeds to every atom in the system and then running the simulation. The simulation ran for 5 different random seed numbers between 100000-199999 and, all these RMSD results backed the conclusion that the structure is in fact stable, as can be seen in Figure 14.

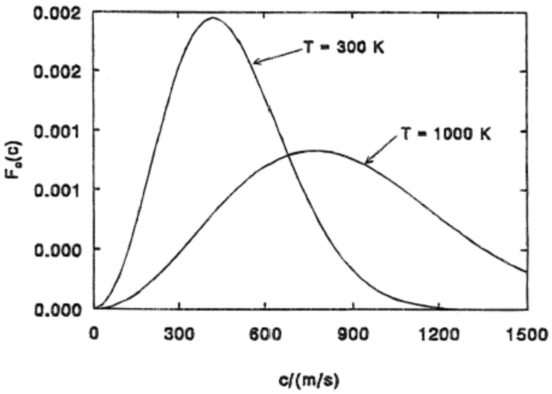


Figure 13: Maxwell-Boltzmann distribution for 300 K and 1000 K. [?]

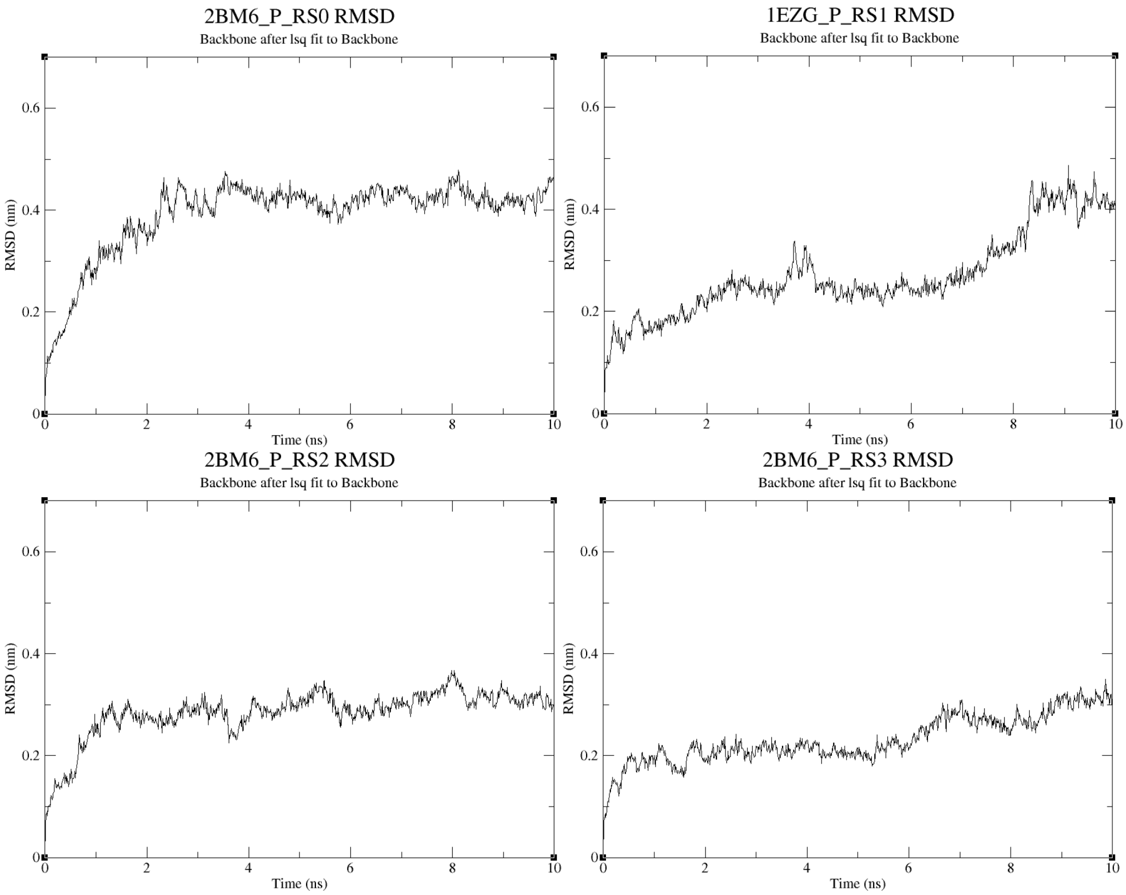


Figure 14: RMSD simulation for 4 different random seed numbers for $2BM6_P$ structure

Before taking it to the lab, there must be a few adjustments to the structure. The natural protein isn't long enough and we are seeking to form fibril polymers. Therefore the N- and C-terminus must have added amino acids to help binding, (QLS) and (KVNVL) respectively. To adjust the length, the modified $2BM6_P$ sequence was doubled and merged using the recently added aminos acids as glue. The glue sequence of amino acids is designed to for salt bond in the interface between solenoids, making the attachment more stable.

In order to complete the previously undone work, it was decided to conduct a parallel work between 3DU1 and $2BM6_P$. With this consideration, a molecular dynamics simulation and regular RMSD test were done for 3DU1. The results indicate that the structure is stable as can be seen in Figure 17

L A A D L S G A S L Q G A D L S Y A N L E S A I L R K A N L Q G A D L T G A I L

Figure 15: 3DU1 sequence

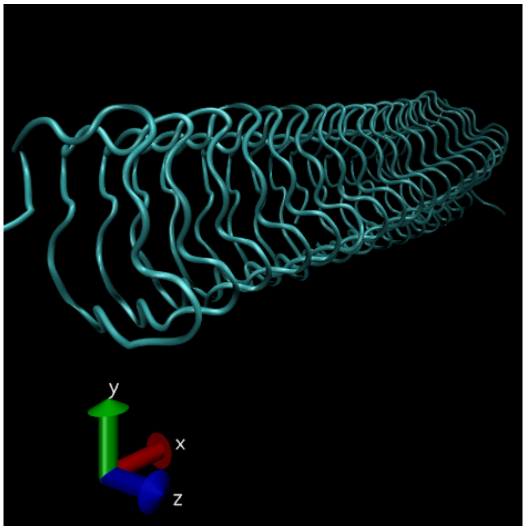


Figure 16: (3DU1 structure

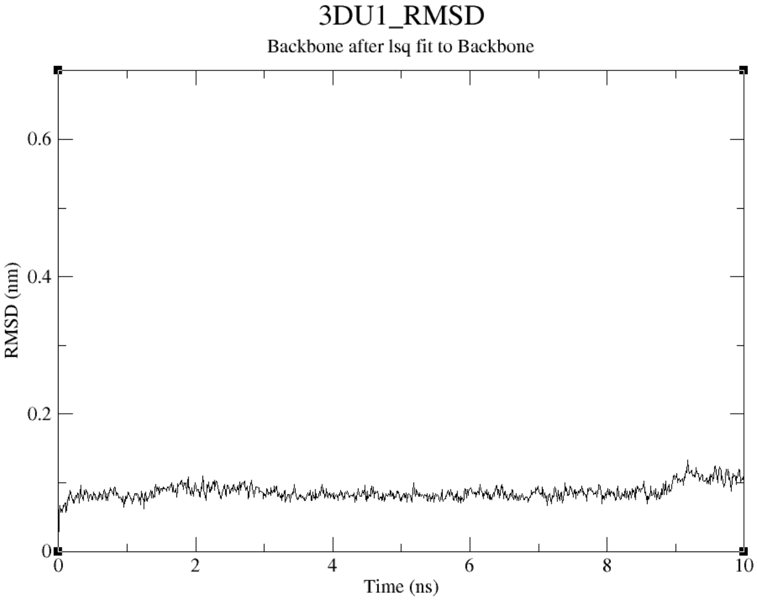


Figure 17: RMSD for 3DU1.

0.4 Expression

After the computational phase has been completed, the next step is expression. For this it is necessary to order the gene from a biotechnology company, in this case Life Technologies. The protein sequence must be reverse-translated to DNA sequence taking in account codon optimization for a more viable gene. A histidine tail is added at the very beginning to facilitate purification and a short assembly sequence before and after the protein sequence. The assembly sequence is simply a couple of nucleotides long strands overlapped between the plasmid and the protein gene that function as a complimentary key to place the plasmid in place.

The gene then has to be inserted to the plasmid matching the assembly sequence corresponding to the order, Figure 18. The plasmid in question has first been opened by enzymes at the assembly loci, restriction digestion, then, by means of a thermal cycle the protein gene has been inserted in place. However, after this procedure there are many plasmids with an incorrect assembly. To discard the faulty plasmids, they must be transformed (inserted into) a strain of E. coli specialized for doing so, o, Strain A. The final expression strain includes two antibiotic resistence genes, so if a given cell was inserted with a plasmid containing both resistance genes, the chances of it having the complete gene are very high. By adding these antibiotics to the medium they grow in, all other cells are discarded.

Afterwards, two colonies from Strain A with good plasmids are chosen, the plasmids extracted from them, sent to sequence to verify they haven't suffered from significant mutation and then inserted to a strain specialized for expression, Strain B. Once the plasmid is inserted to Strain B, a colony is chosen, a pre induced sample is collected, an expression inducer is added to it and samples are collected every hour for four hours and overnight Figure 19. These samples are separated by a buffer between soluble and insoluble. Finally they run in a gel to determine if the protein gene was expressed at all and if so, how long it takes for the E. coli to express it.

It was found that $2BM6_P$, row between 28-36 kDa ,had a good expression at three hours and 3DU1, row between 17-28 kDa, at the overnight sample, both for insoluble 20. This means the protein aggregated inside

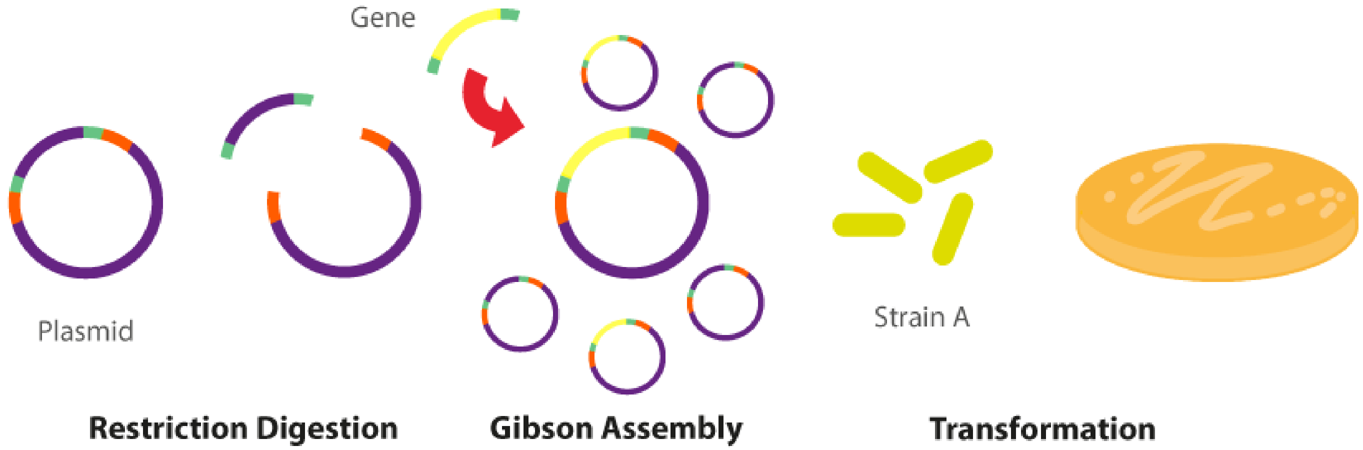


Figure 18: Plasmid insertion.

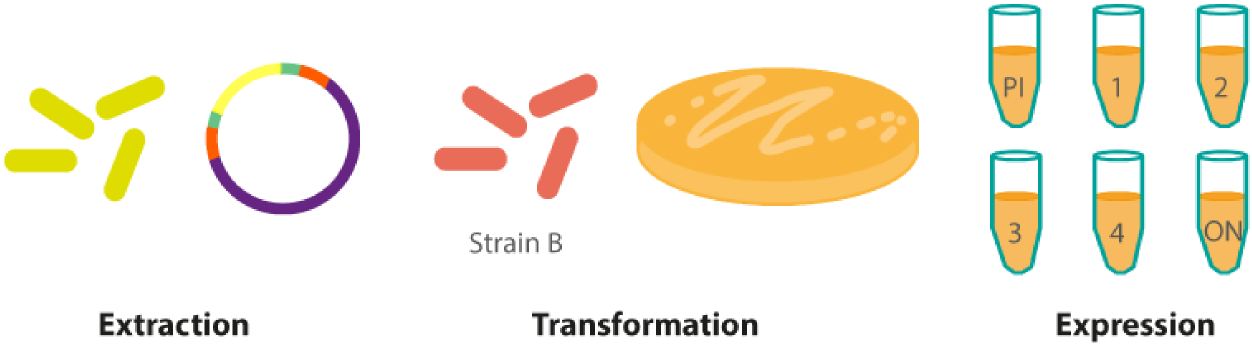


Figure 19: Protein expression

the bacteria, forming inclusion bodies, making them insoluble.

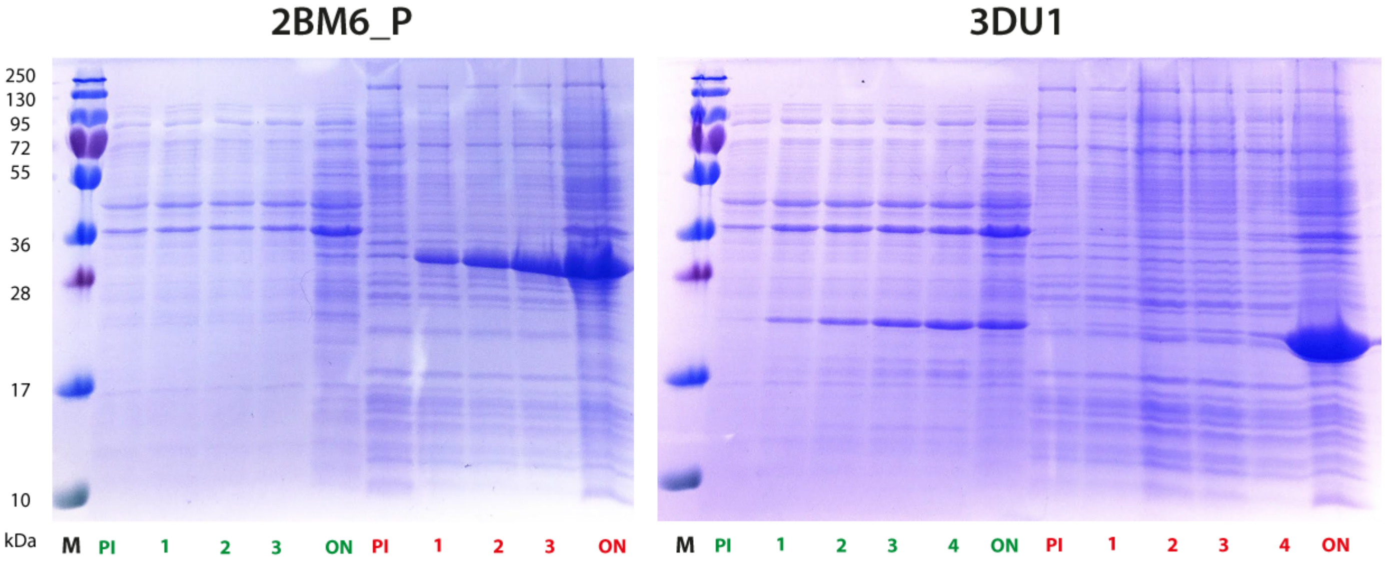


Figure 20: Protein expression gels for each protein, for every sample taken : (PI) pre induce, 1-4 hours, (ON) overnight. Molecular weight markers (M) at the far left of each gel. Good expression is known from really blue coloring, $2BM6$ between 28-36 kDa, 3DU1 between 17-28 kDa.

The line of work from what has been achieved thus far extends to large scale expression of the proteins, their purification, and finally culminating with the growth of fibrils. The chemical complexity and specific to gene and protein behaviour increase the difficulty of subsequent steps; as one goes from theoretical research, to experimental, to applications, this paper ends with the certainty that our worked proteins are well expressed.

Conclusion

The work done here, in an effort to expand our catalogue of usable proteins, after discarding numerous proteins, has left us with two square shaped β-solenoids, $2BM6_P$ and 3DU1, that have almost proven to have application potential, with folding and polymerization tests left to secure this.

We have seen how statistical mechanics deeply intervenes in the development of the computational tools and in deciding the parameters to run a simulation. Our results prove how this classical approach works well, although one can't resist wondering how would a non-classical approach modify current models.

This problem, though simple to pose, is incredibly interdisciplinary involving the collaborative effort of physics, chemistry and biology in both theoretical and experimental ways. From this experience we appreciate the outlines of doing research on a field dedicated to biotechnology, how huge amounts of effort behind the tools used for this paper make it relatively simple to work with and, how this is allowing us to actually used the scientific knowledge for practical applications.

Acknowledgment

We acknowledge useful discussions with everyone at the Cox Group and Toney Lab, specially guidance from Krishna, Amanda, Zeyu and of course, Daniel Cox and Michael Toney themselves. We thank the Department of Physics of UC Davis, the Physics REU Program for providing students this opportunity and Rena Zieve, for making the Davis REU so much more than a research program.

References

[1] Peralta M.D.R., *et.al.* (2015). *Engineering Amyloid Fibrils from β-Solenoid Proteins for Biomaterials Applications.* ACS Nano. 2015 Jan 27;9(1):449-63. doi: 10.1021/nn5056089.

[2] Wikipedia, the free encyclopedia. (2015). *Amino acid.* Consulted July, 2015 from: $https://en.wikipedia.org/wiki/Amino_acid$

[3] Wayne W. LaMorte. Boston University School of Public Health (2014). *Nucleic Acids.* Consulted July, 2015 from: $http://sphweb.bumc.bu.edu/otlt/MPH-Modules/PH/PH709_BasicCellBiology/PH709_BasicCellBiology26.html$

[4] RCSB Protein Data Bank. (2015). *An Information Portal to 111241 Biological Macromolecular Structures.* Consulted July, 2015 from: $http://www.rcsb.org/pdb$

[5] Facultad de Ingienería, Universidad de la República Uruguay. (2009). *BLAST (Basic Local Alignment Search Tool).* Consulted July, 2015 from: $http://www.rcsb.org/pdb$

[6] Theoretical and Computational Biophysics Group, university of Illinois. (2007). *The CHARMM Force Field.* Consulted July, 2015 from: $http://www.ks.uiuc.edu/Training/Tutorials/science/forcefield-tutorial/forcefield-html/node5.html$

[7] MBN Explorer. (2014). *LENNARD-JONES POTENTIAL.* Consulted July, 2015 from: $http://www.mbnexplorer.com/users-guide/4-energy-and-force-calculation/41-pairwise-potentials/414-lennard-jones-potential$

[8] Justin A. Lemkul, Ph.D. (2015) Department of Pharmaceutical Sciences University of Maryland, Baltimore *GROMACS Tutorial: Lysozyme in Water*. Consulted July, 2015 from: *http : //www.bevanlab.biochem.vt.edu/Pages/Personal/justin/gmx − tutorials/lysozyme/*

[9] Wei Cai. Stanford University. (2011). *Handout 9. NPT and Grand Canonical Ensembles*. Consulted July, 2015 from: $http : //micro.stanford.edu/\ caiwei/me334/Chap9_{N}PT_{G}rand_{C}anonical_{E}nsemble_{v}04.pdf$

[10] Yanxiang zhao. George Washington Univerity. (2014). *Brief introduction to the thermostats*. Consulted July, 2015 from: *https : //home.gwu.edu/ yxzhao/ResearchNotes/ResearchNote007Thermostat.pdf*

[11] Theoretical and Computational Biophysics Group, university of Illinois. (2007). *Component rmsd: root mean square displacement (RMSD) with respect to a reference structure.*. Consulted July, 2015 from: *http : //www.ks.uiuc.edu/Research/namd/2.9/ug/node55.htmlSECTION000132215000000000000*

[12] Thermopedia. (2011). *MAXWELL-BOLTZMANN DISTRIBUTION*. Consulted July, 2015 from: *http : //www.thermopedia.com/content/942/*

[13] Schroeder Daniel V. (1999). *An Introduction to Thermal Physics*. 1st Ed, Pearson Prentice Hall.