University *of Ljubljana*
Faculty *of Mathematics and Physics*

SEMINAR

PROTEIN SELF-ASSEMBLY

Urška Jelerčič

Advisors:
Prof. Dr. Roman Jerala
Prof. Dr. Rudolf Podgornik

Ljubljana, 9.12.2009

Abstract

Self-assembly is one of the driving mechanisms in many different systems. In this paper, I concentrate on biological systems, starting with work already done on DNA and then moving to the proteins. First I describe the problem of protein folding and introduce the statistical approach of modeling the process. The folding is then further explained in the means of energy landscapes as well as the kinetics and convergent evolution. In the second part of the paper I present the ideas and results gathered during preparations for iGEM competition 2009. I discuss the use of coiled coils as building blocks and theoretically and experimentally show structures that can be self-assembled.

Contents

1 INTRODUCTION

Self-assembly is a process by which molecules spontaneously assemble into some structure or molecular machine under the appropriate conditions and adopt a defined arrangement in space. The fundamental advantage over mechanically directed assembly is that it requires no tools to move and orient components. Selective binding between matching surfaces is uniquely defined with attractive forces between components that prevail random connections [1].

Biological systems are almost entirely driven by self-assembly. This is an essential process for many of the key activities of living cells. These include the formation of protein (protein folding) and nucleic acid complexes, formation of plasma membrane, flagella, cytoskeleton assembly as well as pathological conditions such as the formation of prions and viral particles, and many others.

Protein self-assembly in this paper refers to organization of certain protein structural motifs into defined 2D or 3D structures. Since this is the field Slovenian iGEM team researched for this year's competition [2], I will be presenting the assembly of coiled-coil motifs into planar networks and regular polyhedra. Given that the self-assembly reduces the entropy, we may reasonably ask ourselves whether such processes can even be possible without some external assistance. Since we know that different forms of self-regulation (such as all types of crystallization) occur in nature, it follows that there must be another thermodynamic quantity that determines the equilibrium. We will see that this quantity is Helmholz energy and the statistical aproach of describing the assembly will be presented.

Paper will be divided into two parts - first briefly mentioning the rich field of DNA self-assembly [3][4], where researchers manufactured practically any desired shape by using only system of complementary DNA strands. Furthermore I will describe why we tried to apply the knowledge and experience of DNA assembly to proteins [5] although proteins are technically much more complex. Since the vast majority of setbacks connected to using proteins is hidden in the problem of protein folding, next section addresses the topic in more detail, especially from physicists point of view. Protein folding is also a self organized process that can help us better understand other types of self-assembly. The rest of the paper describes the use of particular secondary motifs - coiled coils for manufacturing the structures (therefore we can call coiled coils building blocks), shows which structures can be theoretically made using simple combinations of these motifs and presents experimental results.

2 DNA

Recently, DNA manipulation achieved spectacular results [3][4] relying only on the base-pairing properties of DNA duplex and a huge amount of talent and imagination. Almost any type of 2D and 3D structure could be assembled relying only on synthetic DNA and self-assembly. Designed nanostructures, composed of intertwined complementary segments form extraordinary patterns and nano-objects ranging from tetrahedron, octahedron, cube, buckyball, dodecahedron to astonishing two dimensional lattices and shaping different patterns by DNA origami. The design and preparation of such assemblies and conditions, under which these constructs form, can be relatively simple due to limited and easily predictable nucleotide interactions.

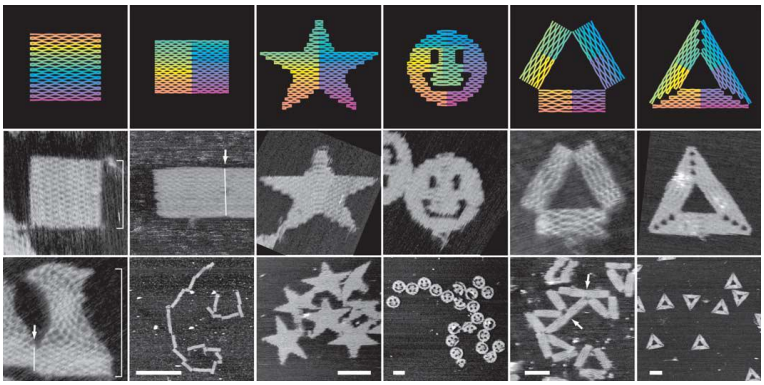

Figure 1: *DNA origami* [4]

DNA has been thus tamed to form complex and well-defined designed structures but in nature proteins (and not nucleic acids) are actually used as embodiments of function and as scaffolds for structures.

3 DNA Vs. PROTEINS

DNA self-assembly has been well researched and nowadays practically any structure or a pattern can be somehow prepared. However, in nature nucleic acids do not represent the main building units but proteins do. Let us see what are the differences between DNA and proteins as building blocks.

Polypeptides are composed of 20 different amino acids [6] in comparison to 4 nucleotides in DNA. Those aminoacids confer to polypeptides a range of different properties, from charge, hydrophobicity to different interaction properties and chemical reactivity. Additionally polypeptide backbone allows polypeptides to assume much larger conformational variability or on the other hand restrict it more than nucleic acids. Also, in nature most of the dynamic structural assemblies as well as cellular nanomachines are made of polypeptides. Proteins can perform a wide range of functions. They can bind proteins, nucleic acids, metal ions and an enormous range of other atomically precise structures, both biological and non-biological. Protein molecules can represent effective scaffold structures: they can have the strength and stiffness like those of epoxies, polycarbonates and other engineered polymers that are also used in nanotechnology applications. On the other hand they can be extremely elastic and their absorption of energy makes them superior to many organic or inorganic materials such as steel or Kevlar.

When comparing the proteins and DNA, it is clear that nucleic acids surpass polypeptides only in two categories. Namely nucleic acids can by directly synthesized by chemical synthesis or easily replicated either in vitro (PCR) or by cell replication machinery, while for polypeptides the chemical synthesis is currently limited. Polypeptides larger than 50 residues are mainly produced by the transcription/translation machinery of cells and their isolation may be demanding. The second argument in favor of nucleic acids is that Watson-Crick base pairing is relatively easy to design to encode folding of nucleic acids in a DNA origami-like fashion. Prediction of tertiary structures (e.g. protein folding problem) by polypeptides is currently still a challenging problem and there are only a handful of proteins designed "from scratch".

4 PROTEIN FOLDING

4.1 The problem

The function of the protein is mainly determined by its structure, which can be described on four different levels: primary, secondary, tertiary and quaternary [6]. Primary structure corresponds to the sequence of aminoacids, secondary shows larger building units (for example α-helix and β-sheet), tertiary connects secondary motifs in one polipeptide chain while quaternary presents the whole threedimensional functional protein (many polipeptide chains). The protein folding theory concentrates on how to predict tertiary or quaternary structure based on the information from primary, because it was unambiguously proven in the late 50s[1] that native conformation depends solely on the sequence of aminoacids.

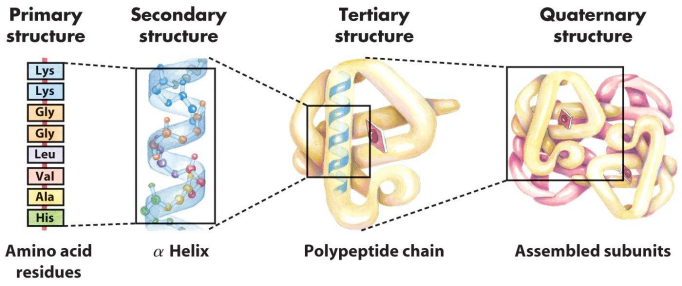

Figure 2: *Levels of protein structure* [7]

Proteins are vast sistems (consisted of several thousand aminoacids), irregular, unsymmetrical and very flexible which allows them to change conformation rapidly [8]. They can take many different forms but only one is fully functional - called *native*. Actually, even most native states of proteins are flexible and are comprised not of only one conformation but of a set of closely related structures. This flexibility is essential if they need to perform any biological function. These and many others are all the reasons why protein structure prediction remains challenging.

Today we can follow two qualitatively different ways to unveil the connection between sequence and structure:

[1] Christian B. Anfinsen won Nobel Prize in Chemistry in 1972 for this discovery

- Knowledge-based: Based on the already measured data: homology modeling, fold recognition and de novo prediction
- Ab initio methods: Use computer simulation of the real physical process without using any empirical information.

It is clear that knowledge-based methods show obvious practical advantages and superior results when compared to pure ab initio approaches, since proteins are too complex to model easily. But if we want to know the real details of protein folding as it happens we must resort to ab initio strategies which give us a wider range of applicability and grater ability of generalization. Although ab initio methods are slowly improving, the main approach used nowadays in practice is still combined knowledge-based.

4.2 Energy functions and statistical approach

There are first a few assumptions we have to consider. There are many details in real folding that complicate its description: many cellular processes are involved in converging to native structure; some proteins have been shown to fold cotranslationally and many of them are known to be assisted by molecular chaperones; some non-peptide molecules may be covalently attached to the protein chain or some cofactor or ion may be needed to reach the native structure; some residues may be post-translationally changed into side chains that are not included in the standard twenty,... We will ignore all these factors and assume that what we learn about the mechanism of folding of small, fast-folding proteins in vitro will apply to their folding in vivo and, to a large extent, to the folding of individual domains in larger proteins. This simplified problem is called *restricted protein folding problem* and is one of the most common forms of folding problem.

Now we can ask ourselves: How does the protein fold so fast into its functional native structure? We will search the answer by using statistical mechanics and thermodynamics [8][9][10][11] and will assume also that all the macroscopic parameters (T, N_w=number of water molecules,...) remain constant. Our considered sistem will be one protein unit surrounded by N_w water molecules.

As we have already learned; the sistem that obeys classical mechanics can be completely described with Euclidean coordinates and momenta. We will use the following denotation for the atoms that belong to the protein: x^μ and π_μ, with $\mu = 1, ..., N$) and those belonging to the water molecules: X^m and Π_m, with $m = N+1, ..., N+N_w$). The whole set of microscopic states shall be called phase space and denoted by $\Gamma \times \Gamma_w$, explicitly indicating that it is formed as the direct product of the protein phase space Γ and the water molecules phase space Γ_w.

Hamiltonian (or energy) function in this case has the form:

$$H(x^\mu, X^m, \pi_\mu, \Pi_m) = \sum_\mu \frac{\pi_\mu^2}{2M_\mu} + \sum_m \frac{\Pi_m^2}{2M_m} + V(x^\mu, X^m) \tag{1}$$

where M_μ and M_m denote the atomic masses and $V(x^\mu, X^m)$ is the potential energy. We will assume that the equilibrium is attained at the temperature T and does not change. Thus we can define the partition function of canonical ensemble:

$$Z = \frac{1}{h^{N+N_w} N_w!} \int_{\Gamma \times \Gamma_w} \exp[-\beta H(x^\mu, X^m, \pi_\mu, \Pi_m)] dx^\mu dX^m d\pi_\mu d\Pi_m \tag{2}$$

where h is Planck's constant and $N_w!$ accounts for the quantum indistinguishability of the N_w water molecules. It should be noted again that we assume N_w to be constant and as such represents just a change of reference in the Helmholtz free energy.

In order to reduce computational demands, water coordinates and momenta are customarily averaged out. The integration over the water momenta Π_m in equation (2) yields a T-dependent factor that shall be dropped because again it only changes reference. On the other hand, the integration over the water coordinates X^m is not so trivial, and, except in the case of very simple potentials, it can only be performed formally. To do this, we define the *potential of mean force* or *effective potential energy* by:

$$W(x^\mu; T) \equiv -k_B T \ln \left(\int \exp(-\beta V(x^\mu, X^m)) dX^m \right) \tag{3}$$

and associated *effective Hamiltonian*:

$$H_{eff}(x^\mu, \pi_\mu; T) = \sum_\mu \frac{\pi_\mu^2}{2M_\mu} + W(x^\mu; T) \tag{4}$$

In this manner we can rewrite partition function:

$$Z = \int_{\Gamma} \exp[-\beta H_{eff}(x^{\mu}, \pi_{\mu}, T)] dx^{\mu} d\pi_{\mu} \tag{5}$$

It is common in literature to integrate over protein momenta π_{μ} because this choice largely simplifies the discussion about the mechanisms of protein folding. However, performing this average brings up a number of difficulties mostly due to the fact that the probability density in the x^{μ}-space is not invariant under canonical transformation. Bearing that in mind, we integrate nevertheless and define new form of partition function that will be used in the following derivations:

$$Z = \int_{\Omega} \exp[-\beta W(x^{\mu}, T)] dx^{\mu} \tag{6}$$

where Ω now denotes the positions part of the protein phase space Γ.

The probability density function in the protein conformational space Ω is defined as:

$$p(x^{\mu}) = \frac{\exp[-\beta W(x^{\mu}, T)]}{Z} \tag{7}$$

We can see that $W(x^{\mu})$ completely determines the conformational preferences of the polypeptide chain as a function of each single point of Ω. Usually we do not describe system point by point but define states that are neither single points of Ω nor the whole set, but finite subsets $\Omega_i \subset \Omega$ comprising many different conformations that are related in some sense.
Partition function of a certain state Ω_i is now:

$$Z_i \equiv \int_{\Omega_i} \exp[-\beta W(x^{\mu}, T)] dx^{\mu} \tag{8}$$

We can also define microscopic probability density function, which tells us the probability of system being in this particular state:

$$p_i(x^{\mu}) \equiv p(x^{\mu} \Big| x^{\mu} \in \Omega_i) = \frac{\exp[-\beta W(x^{\mu})]}{Z_i} \tag{9}$$

and the entropy of Ω_i:

$$S_i \equiv -k_B \int_{\Omega_i} p_i(x^{\mu}) \ln p_i(x^{\mu}) dx^{\mu} \tag{10}$$

4.3 Energy landscapes

An average-length polypeptide chain is very large (N protein atoms). But the size of conformational space grows exponentially on N and is therefore astronomical [8]. This is not just practical problem but has been for several years also theoretical. The so called *Levinthal paradox*[2] says that, if, in the course of folding, a protein is required to sample all possible conformations and the conformation of a given residue is independent of the conformations of the rest (which is false), then the protein will never fold to its native structure. The hipothesis was very general[3], but the difference between the time that a small protein (≈100 AA) would need to fold randomly (≈ 10^{10} years) and the time the process really takes (several ms) is absurdly vast. Protein folding obviously cannot be a completely random trial-and-error process (i.e. a random walk in conformational space).

The paradox would exist if the energy function ($W(x^{\mu})$) would be so called *golf-course* energy landscape, but Lavinthal quickly proposed the solution: *ant-trail* landscape. This view, which is typically referred to as the old view of folding, is largely influenced by the situation in simple chemical reactions, where the barriers surrounding the minimum energy paths that connect the different local minima are very steep compared to $k_B T$, and the dynamical trajectories are, consequently, well defined. In protein folding, however, due to the fact that the principal driving forces are much weaker than those relevant for chemical reactions and comparable to $k_B T$ (e.g. rotations of the individual side chain), short-lived transient interactions may form randomly among different residues in the chain and the system describes stochastic trajectories that are never the same. Hence, since the native state may be reached in many ways, it is unlikely that a single minimum energy path dominates over the rest of them.

[2]Stated in 1969 in the talk: '*How to fold graciously*'

[3]For example, one could equally well apply it to the formation of crystals, and conclude that crystallization can never occur!

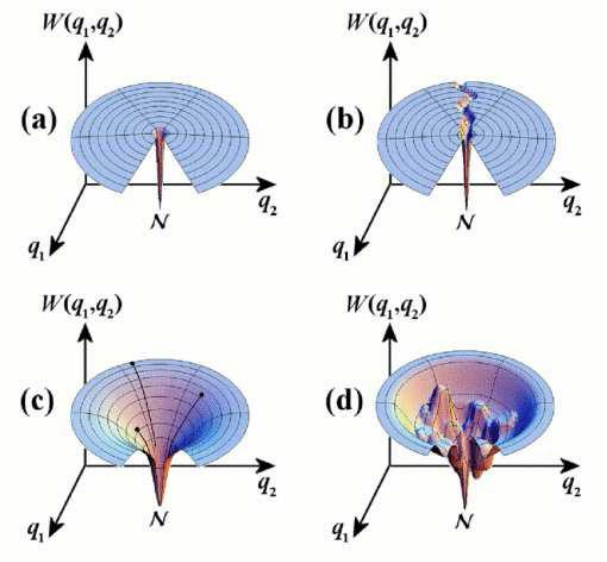

Figure 3: *Energy landscapes: a)golf-course, b)ant-trail, c)energy funnel, d)realistic rugged funnel* [8]

The new view of folding has been proposed in the late 80s and is based on the better known statistics of spin glasses. According to it, when a large number of identical proteins is introduced in a test tube a conformational equilibrium is attained between the native ensemble of states N and the ensemble made up of the rest of conformations (the unfolded state U). At the microscopic level each single molecule is following a partially stochastic trajectory determined by the intrinsic energetics of the system (given by $W(x^\mu)$). All trajectories are different, some aiming towards the native state and some towards the unfolded state, but, if we focus on a single molecule at an arbitrary time, the probability that it is wandering in the native basin is very high (typically more than 99%). The process of folding is therefore greatly accelerated. But how (physically) can this fast fold be guaranteed? It turns out that the energy landscape has to be funneled [8][12][13][14] towards the native state so that any microscopic trajectory has more probability to evolve in the native direction than in the opposite one at every point of the conformational space.

Blindly taking into account the characteristics of proteins, it is clear that such systems present a large degree of *frustration*[4]. That means that there is not a single conformation of the chain which optimizes all the interactions at the same time. This fact leads to more realistic rugged[5] energy landscape with low-energy states and high barriers, although ruggednes must be relatively small in order to avoid getting trapped in deep local minima during the course of folding.

But there is a practically important detail: is the native state the global minimum of the effective potential energy $W(x^\mu)$ (*thermodynamically controlled* folding process) or it is just the lowest-lying kinetically-accessible local minimum (*kinetically controlled* process)? Today it is widely accepted that the thermodynamic hypothesis is fulfilled most of the times. Theoretically, the difference is minor, but it greatly changes the computaional approach. In the case of global minimum, the prediction of the native state may be tackled both dynamically and by simple minimization of the function $W(x^\mu)$, whereas, if the thermodynamic hypothesis is broken, the native structure may still be found performing molecular dynamics simulations, but minimization procedures could be misleading (because the absence of the kinetic information). Currently the theoretical algorithms include a number of strong assumptions but are still incapable of generalized folding prediction.

[4]The paradigm for a frustrated system is the spin glass, a magnetic system in which spins are randomly arrayed in a dilute alloy. The interactions between spins are equally often, at random ferromagnetic (the spins want to point in the same direction) and antiferromagnetic (the spins want to point in opposite directions). These two conflicting local tendencies can not be satisfied completely in any arrangement of spin orientations. Thus, the system is said to be frustrated

[5]First funnel models assumed smooth landscapes which exhibit cooperative phase transitions that are determined by the temperature. At high temperatures, the large number of high energy structures predominate, but as the temperature of the system is lowered, the system will occupy the lower energy states. Dynamically, below a transition temperature, such systems will fall into a funnel of low energy states and may remain trapped there. In these processe (e.g. crystallization), once a large enough nucleus of low energy structure is formed, the rest of the low energy structure forms rapidly. Systems with rough energy landscapes also exhibit effective phase transitions. When the temperature of such a system is lowered, it tends to occupy the lower energy states and at a transition temperature will become trapped in one of them. Generally, these transitions are accompanied by a considerable slowing of the motion as the system tries to exit over the high energy barriers(phenomenon known as glass transition in super-cooled liquids).

4.4 Kinetics

The other part of folding information is provide by kinetic theory [10]. Protein molecules are random coiled (denatured) in nonaqueus medium. Polar residues seek to form hydrogen bonds and therefore create nonpermanent α-helices and β-sheets. Since these bonds form momentarily and can switch from one molecule to another ($\tau \approx 10^{-12}s$) they are considered random fluctuations. When the pH or temperature is changed, so that the medium becomes effectively aqueus, the protein chain begins to fold and permanent bonds in α-helices and β-sheets grow with time. The helical content (a measure of secondary structure) first overshoots and then decays to the native value. The experiments in which researchers measure excess heat capacity of the soultion when proteins are added show two peaks: peak at higher temperature marks transition from denatured state to an intermediate state known as *molten globule*, and the second peak marks the transition[6] to the native state.

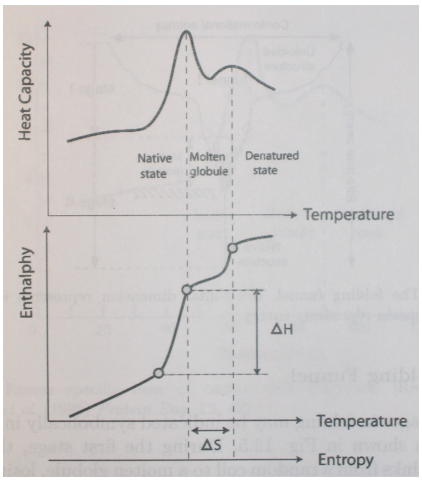

Figure 4: *Measuring excess heat capacity during folding and plotting enthalphy* $H(S)_p$*:* $H = \int C_p dT$ *against temperature* [10]

Folding seems to go trough two stages: a fast and slow one. The fast stage lasts the order of 10^{-3}, during which the denatured state becomes a molten globule. In the next stage, the molten globule slowly evolves into the native state. The shape of the energy funnel can be now a bit altered (accounting for different time constants of two stages):

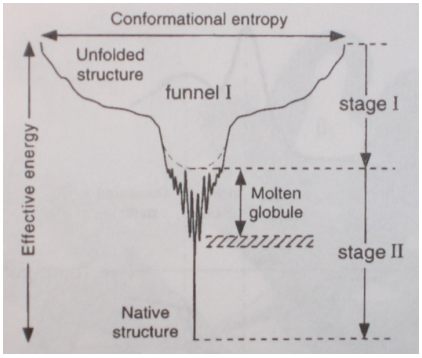

Figure 5: *The folding funnel* [10]

At this point let us discuss the entropy 'paradox' that might occur. Folding into arranged structure (from totally unfolded, chaotic state) lowers the entropy, which would seem to conflict with second law of thermodynamics[7], stating that the entropy of an isolated system which is not in equilibrium will tend to increase over time, approaching a maximum value at equilibrium. How can proteins fold then? It turns out that in nonisolated systems entropy is not the main contributor to stability, but Helmholz free

[6] If the protein molecule had been infinetely large, we might imagine that these peaks would become delta functions, signifying first-order phase transitions. As they stand, we call them pseudo phase transitions.

[7] The same concern applies to self-assembly of proteins into organized structures

energy is (the minimum):

$$F \equiv U - TS \tag{11}$$

The term TS represents the part of energy residing in random thermal motion, which is negligible in long polypeptide chains compared to hydrophobic and electrostatic interactions. Also, we have to remember that our system comprises also of many water molecules that form water net around the protein. When protein folds and creates bonds within itself, the bonds with water break and change, resulting in incrasing the entropy of water.

4.4.1 Convergent evolution

There is also another empirical fact complicating the folding description - proteins of very different amino acid sequences share the same folded structure (a phenomenon common in dissipative processes). This property is known as *convergent evolution* and means that molten globule state already has a certain degree of universality. Let us consider another analogy: the turbulence. As protein folding is thought to be a stochastic process with dissipation, the same applies to the turbulent flow. In the steady state known as fully developed homogenous turbulence, the energy spectrum $E(k)$ obeys Kolmogorov's law (for k in the *inertial range*[8]):

$$E(k) \sim \epsilon^{2/3} k^{-5/3} \tag{12}$$

where ϵ is the rate of energy dissipation. Simmilar power-law was empiricaly established for proteins:

$$\widetilde{g}(k) \sim k^{-5/3} \tag{13}$$

in denatured (beginning) state, where $\widetilde{g}$ is the correlation function of the protein.

Also, a turbulent fluid can be modeled by a tangle of vortex lines. A vortex either terminates on an external wall or ends on itself to form a ring. Kelvin's theorem in hydrodynamics states that, in the absence of viscosity, vortex lines do not cross. Thus the motion of the vortex line can be modeled by SAW[9], as already shown for polypeptide chains.

Obviously there are certain similarities which give us the idea to use something well known from turbulence theory and apply it to processes in polymers that we do not entirely understand. A helping tool in explaining the principle of convergent evolution in proteins is *energy cascade*. Let us first review the mechanisms in turbulence: Energy input occurs at large scales, trough the creation of vortex rings. They are unstable because of 'vortex streching' (i.e. the core of the vortex spontaneously contracts to a smaller radius and increases in lenght. Eventually the vortex ring breaks up into smaller rings. The ring ceases to exist as such, when its radius bedcomes comparable to the core radius. Its energy is the dissipated as heat and, most importantly, the final distribution of the rings is independent of initial conditions, for it depends only on the nature of vortex instability.

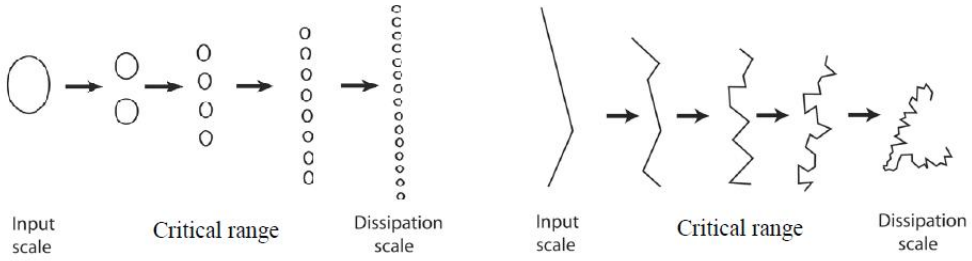

Figure 6: *Energy cascade in turbulence (left) and proteins (right)* [10]

How can we apply this to proteins? The protein molecule folds in aqueus solution because of the intaractions between protein molecules and water net. Water nets have vibrational frequences $\sim 10GHz$. This lies in the same range as those of the low vibrational modes of protein molecule. Therefore, there is resonant transfer of energy and additional energy exchange due to random impacts. The resonant transfer involves shape vibrations and therefore occurs at the largest lenght scales. It is then transferred to intermediate lenght scales trough nonlinear couplings of the vibrational modes. There is thus little dissipation, until energy is further dispersed to smaller lenghts, where it dissipates as heat. The manner

[8] inertial range a.k.a. universal range: $\eta^{-1} \gg k \gg \xi^{-1}$, where η is a small lenght scale at which dissipation takes place and ξ is correlation lenght

[9] self-avoiding walk algorithms

of energy transfer is independent of initial conditions, therefore, after a few steps in the cascade, all the starting information is lost. This approach also implies importance of minimizing the cascade time in order to increase the probability of succesful fold, but there is yet no mathematical model that would combine this condition with minimization of local free energy.

5 COILED-COILS

Athough we saw that there are generally great problems with protein structure algorithms, some special interactions which can be predicted do exist. Among the best understood types of protein-protein interactions are *coiled-coils* where the stability of interacting helices can be obtained with relatively good accuracy using knowledge-based methods. These are also the types of secondary motifs we used for our project.

Coiled-coils are protein structural motifs [15][16][17][18] where α-helices wrap around each other to form an intertwined superhelix. Left-handed helices require 3.5 residues for each turn, therefore every seven residues (heptad) makes two helical turns, which span approximately 1 nm. Those residues are designated $(abcdefg)_n$ in one helix and $(a'b'c'd'e'f'g')_n$ in the other. Residues at positions a and d are usually occupied by nonpolar core residues found at the interface of two helices and are essential for the oligomerization (formation of a hydrophobic core). Residue at positions e and g are solvent exposed polar residues at the edge of hydrophobic core that can interact with positions e' and g' on the neighboring helix through electrostatic interactions. Residues b, c and f are typically hydrophilic and exposed to the solvent and may be available to introduce additional functional properties into the coiled-coil.

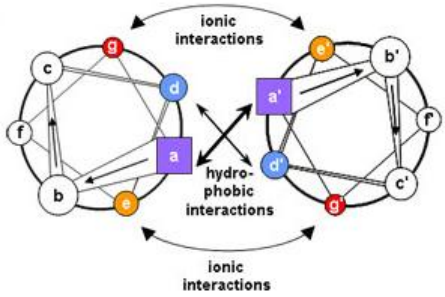

Figure 7: *Structure of parallel coiled-coil* [5]

Coiled coils can assemble either in a parallel or antiparallel orientation, with respect to the orientation of polypeptide backbone. They can form either heterodimers or homodimers and all of those combinations have already been designed and verified experimentally for pairs of coiled-coil forming segments. Stability of coiled-coil is determined by the interactions between heptads, orientation and length of the segment.

5.1 Orthogonal coiled coils - de novo design

One of the main challenges in connection with the design [20][20][21][22][23] of the pairs is to ensure exclusivity of the pairing. Our selected heterodimeric coiled-coil-forming segments (compared to two associated DNA strands) had to satisfy the following requirements:

- any coiled-coil-forming segment must form coiled-coil with exactly one segment,
- any other combination is forbidden/significantly disfavored.

In this way selected system of coiled-coil forming segments is called *orthogonal*. Additionally we have to ensure that selected proteins are not too long (in order to avoid bending as well as decreased energetic difference between desired and undesired pairs) but still have approximately the same length to form symmetrical structures if not otherwise required by some special design. Several other characteristics (parallel or antiparallel orientation, homodimers, heterodimers) also have to be considered.

The design starts by choosing the most appropriate amino acids for the every position in sequence. Generally, the rule states that hydrophobic residues at positions a or d and opposite charge at positions e and g on the equivalent positions between the two helices stabilize, whereas burial of polar residues (Asn) at positions a or d and the same charge at positions e and g destabilize the structure. In this simplified

model we consider a limited number of variable residues only at positions that significantly affect the stability (a, d, e, g) and neglecting the effect on other positions.

Accordingly to recent work by Hegeman and coworkers [21], there are many factors that affect stability: hydrophobic burial, propensity, solubility, electrostatic interaction of flanking residues and others but the most sucessful coiled-coils prediction algorithm considers only three main factors: core, electrostatic and propensity.

How the algorithm works? A rudimentary packing score is assigned to all dimmers to distinguish cores which make large contributions to stability from those which do not by scoring hydrophobic pairings highest. Core interactions are summed over the overlapping part of the sequences, taking into account only combinations at aa' and dd'. The weights characterizing each amino acid (AA) combination were derived from number of experimental measures. Electric parameters are based on opposing charge pairings and place energetic penalties on similar charge pairings with $g_i e'_{i+1}$ end $e_{i+1} g'_i$ treated same for simplicity. These interactions are summed similarly as core - the only difference is that a g_i residue forms a columbic interaction with an e'_{i+1} residue of the next heptad on the opposite helix. Electrostatic weights were related to free energy contributions based on data from double mutant analysis. Propensity represents normalized sum over all AA in the sequence and states the relative contribution of each AA to the wholesome stability of the helix. It informs upon the frequency or preference with which a given residue occurs in particular conformation. The least square fit[10] is then used to combine all these parameters and stability can be rationalized. T_m is calculated for each possible register combination to find the most stable one (the highest T_m), including its orientation (parallel or antiparallel).

5.2 Identification of a set of orthogonal coiled-coil pairs

We have also designed an algorithm [5] which maximizes the temperature difference between the least stable desired pair and most stable undesired pair in the selected set (previously calculated using protein prediction algorithm). As a result we found a system of 8 orthogonal designed peptides: P1, P2, P3, P4, P5, P6, P7 and P8, which form four pairs: P1+P2, P3+P4, P5+P6 and P7+P8, where the temperature difference between the least stable desired pair and most stable undesired pair was predicted to be more than 60°C.

	PARALLEL							
	P1	P2	P3	P4	P5	P6	P7	P8
P1	33	100	29	27	31	32	30	29
P2		-6	10	7	11	12	11	9
P3			10	93	19	20	19	17
P4				5	17	18	17	15
P5					13	101	-15	-16
P6						16	-13	-15
P7							12	96
P8								9

	ANTIPARALLEL							
	P1	P2	P3	P4	P5	P6	P7	P8
P1	-62	5	-30	-33	-28	-27	-29	-30
P2		-100	-49	-52	-47	-46	-48	-49
P3			1	-87	-40	-38	-41	-42
P4				-3	-42	-41	-43	-44
P5					-81	7	-39	-40
P6						-78	-37	-38
P7							3	-84
P8								1

Figure 8: *Pair interactions (degrees Celsius) for parallel (favoured) and antiparallel (unfavoured) orientation* [5]

5.3 Coiled-coils self-assembly

The assembly of two halfs of coiled-coil is very similar to protein folding process. System again tends to form energetically most stable structure, although entropy seems to decrease (the entropy part here is even smaller then in folding of a single protein, because number of interactions is bigger). Entropically the same happens when water freezes to ice or when some other crystalization process occurs.

In our case we used coiled-coils linked into polypeptide chains. Generally we used combination of three halves (or *elements*) of coiled-coil motif in one chain (named *building block*) and different combinations of their pair halves in other. This way we can uniquely determine specifical binding of each chain and ensure a high probability of formation of predicted structures.

Our orthogonal system of pairs also theoretically enables temperature regulation (similarly as with DNA) - the high temperatures (compared to room temperature) guarantee strong interaction also at severe conditions, whereas undesired pairs with very low melting temperatures cannot form pairs at room

[10] Tested on 59^2 bZIP interactions and was able to correctly identify 92% of noninteractions and 92% of strong interactions.

temperature. In this way we could dictate the order of forming pairs just by increasing or decreasing the temperature.

5.4 Topology

In order to select the most appropriate building blocks [5] for experiments we listed all possible combinations of chains with three elements and searched for all the structures they can make. Let us see the most interesting ones:

5.4.1 Topology of two-dimensional lattice made of single type of polypeptide chain

[Legend: a and a' - parallel heterodimer and its pair; A and A' - antiparallel heterodimer and its pair; A - antiparallel homodimer (pairs with itself); a - parallel homodimer (pairs with itself); arrows in figures represent coiled-coil segments]

Two-dimensional lattice can form through many different ways as well it can be assembled from different primitive cells. Let us first consider only options using only one type of polypeptide chain comprising three coiled-coil segments, which is the simplest type of building block. It turns out that trigonal and hexagonal lattices are geometricaly easy obtainable as seen in the examples below, where we use three segments of different antiparallel/parallel homodimer/heterodimer combinations.

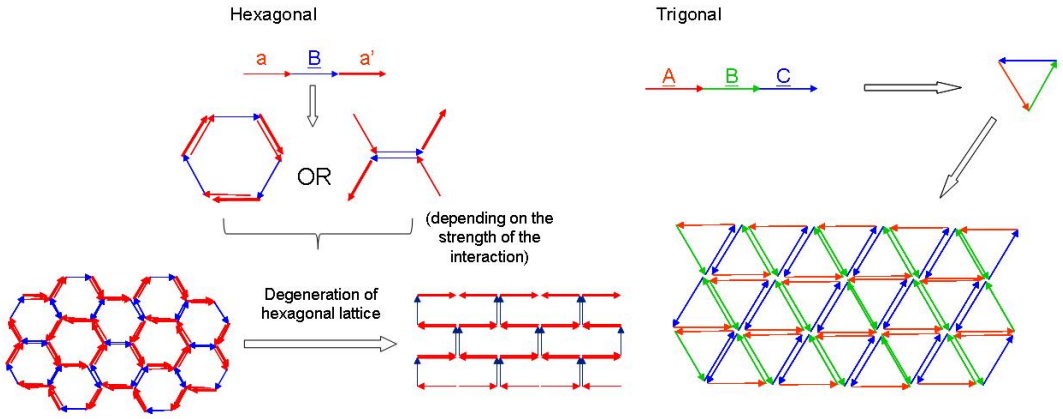

Figure 9: *Schematic assembly of trigonal and hexagonal lattice: a pairs parallelly with a', A(B, C) antiparallelly with A(B, C)* [5]

5.4.2 Creating three-dimensional polygons from a single type of polypeptide chain

Note: again we restrict our discussion only to cases with one-type-three-element chains.

The simplest polyhedron that we can form is tetrahedron or triangular pyramid. Tetrahedron is formed from four polypeptide chains.

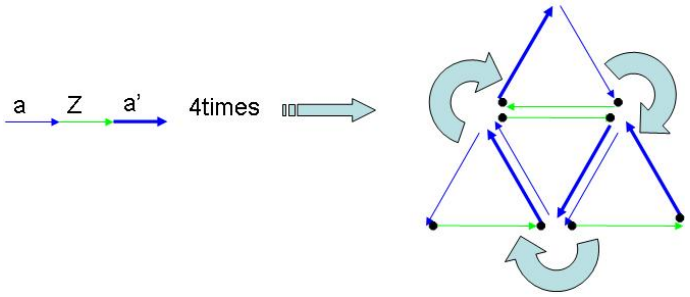

Figure 10: *Assembly of tetrahedron: a pairs parallelly with a', Z antiparallelly with itself* [5]

The next structure that can be obtained is n-sided prism, where n depends on the polypeptide concentration, temperature and length of the linker (among other parameters). Some types of chains (combinations of coiled-coil forming segments) even pre-define whether n can be even or odd (i.e. a-b-a' forms prisms with even number of sides, contrary to a-B-a' where there are no limitation on the accessible

number of sides). This assembly is possible to obtain from different ways - below is presented example a-b-a', which can form a box (parallelepiped) as the smallest 3D polygon of this type of polypeptide chain:

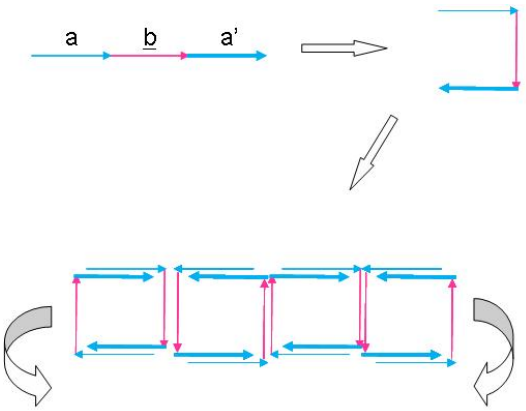

Figure 11: *Assembly of prism (cube): a pairs parallelly with a', b parallelly with itself* [5]

5.4.3 Extension to self-assemblies made of several different polypeptide chains

Here is where real diversity begins. Let us look at only two examples:

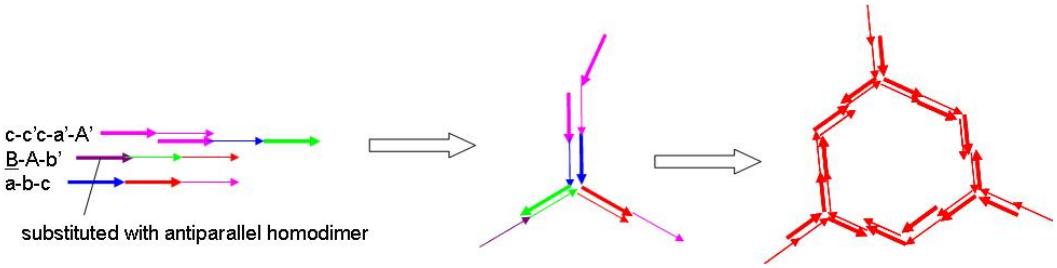

Figure 12: *Example A* [5]

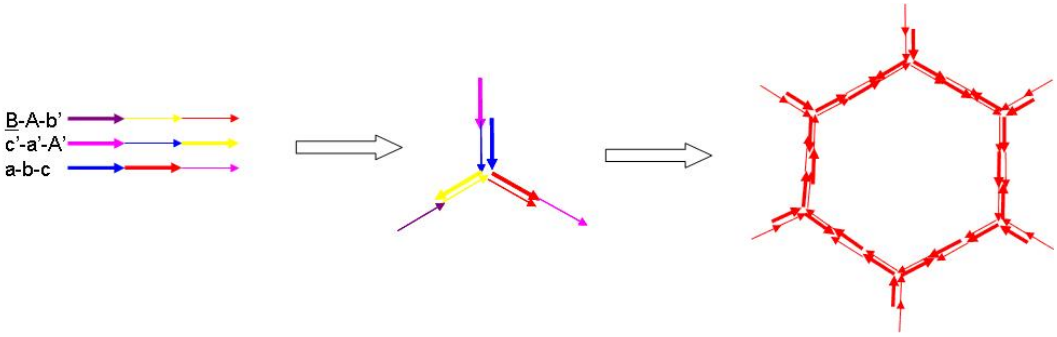

Figure 13: *Example B* [5]

Both and many more derive from the same three-armed motif but can assemble into topologically very different structures. If we design a rigid linker (or no linker at all) this manner we can expect creation of the following planar structures:

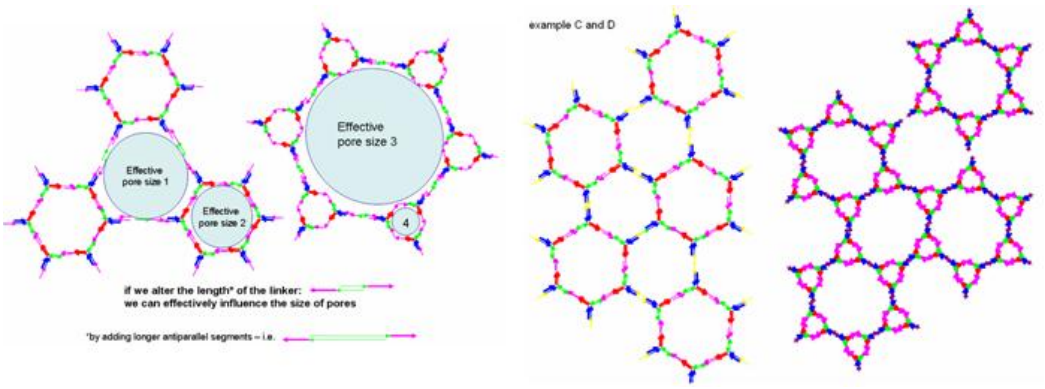

Figure 14: *Assembly of planar nano-networks* [5]

In order to allow the extension into the third dimension we need the linkers between elements to allow flexibility (as already shown for DNA polyhedra). In this way we can form theoretically any regular polyhedron with three-armed motifs in the vertexes. Below you can see the formation of a cube, formed from the elements in the example B:

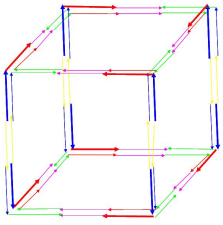

Figure 15: *Assembly of a cube (or any regular polyhedra - depending on concentration)* [5]

6 EXPERIMENTAL RESULTS

We have seen that certain building block can form 2D structures as well as 3D, that is why we selected one of the simplest combinations that could theoretically form a cube as well as cover the plane in a polyhedral (hexagonal) lattice (named *construct K2*) [5]. This combination comprised two parallel coiled-coil heterodimers (designed P1 and P2) and one parallel homodimer (Gcn4-p1(I-L)). Between each coiled-coil-forming domain we introduced a dipeptide linker Gly-Ser, to allow the limited flexibility of coiled-coil segments. Additionally the polypeptide construct included a hexahistidine peptide tag, to facilitate purification, detection and attachment site of additional functions, such as metal or fluorophore binding, to the assembled product.

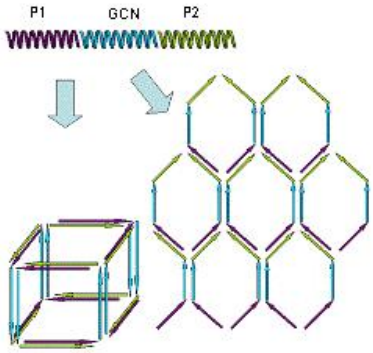

Figure 16: *K2 construct and possible assemblies* [5]

All the required proteins were sucessfully produced and isolated (oligomers analysed using SDS-PAGE

and Western Blot[11]). Circular Dichroism (CD)[12] spectra of individual P1 or P2 peptides show that each of them is disordered in solution while their mixture (P1+P2) shows a high level of α-helical content and therefore proves the generation of coiled coil.

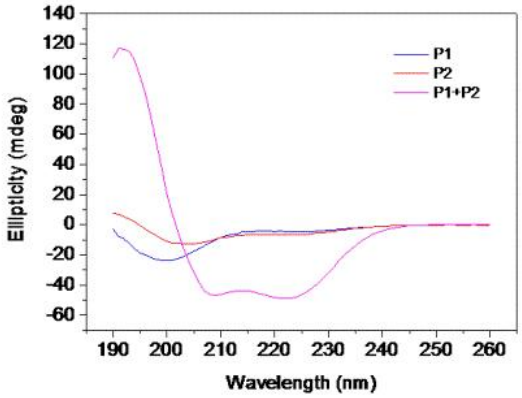

Figure 17: *CD spectra of separate P1 and P2 proteins and their combination, which forms coiled-coil* [5]

We analyzed the secondary structure of a polypeptide K2 (assembled by process of slow chemical annealing[13] we invented) under the native conditions. CD spectra showed strong helical signal, confirming that the coiled-coil interactions occur also in the context of a longer polypeptide and in the presence of linker sequences.

Concentration of the polypeptide represents an important factor in determining the type of self-assembled structures, either polygons, which require the assembly of eight chains or lattice, which requires over hundreds of molecules. At low polypeptide concentrations formation of polygons should be favored since they should form oligomers that are big enough to form a closed structure, satisfying all coiled-coil-forming potentials.

DLS (dynamic light scattering)[14] results showed that solution contained two populations of K2 aggregates and both were present at appreciable amount. Hydrodynamic radius (RH) of one population was 8 nm, and for the other 88 nm, indicating the presence of initial steps of larger assemblies in addition to small aggregates, compatible with the expected box. Note that the scattering power of larger aggregates dominates the correlation curve and the observable presence of smaller aggregates indicates that their number is significant in the solution.

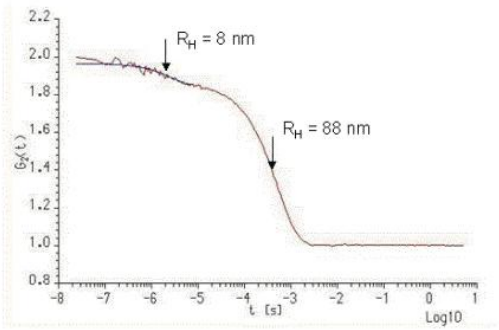

Figure 18: *DLS plot* [5]

[11] SDS-PAGE (sodium dodecyl sulfate polyacrylamide gel electrophoresis) is a widely used technique which separates proteins according to their molecular weight. Western blot is an analitical method used for detection of specific proteins.

[12] A type of spectroscopy based on the differential absorption of left- and right-handed circularly polarized light. The far-UV CD spectrum of proteins can reveal important characteristics of their secondary structure (each of the secondary motifs - e.g. α-helix, β-sheet,... has a distinguishable spectrum)

[13] More at [5]

[14] A technique which can be used to determine the size distribution profile of polymers in solution. When the light hits particles the light scatters in all directions and one observes a time-dependent fluctuation in the scattering intensity. The dynamic information of the particles is derived from an autocorrelation of the intensity trace recorded during the experiment. After complex data analysis the hydrodynamic radius (RH) is obtained.

Self-assembled structures of K2 annealed at different concentrations were analyzed by AFM (atomic force microscopy) operating in acoustic alternative current mode and TEM (transmission electron microscopy). TEM was used to analyze the structure of self-assembled K2 at $5\mu g/ml$ and $0.5\mu g/ml$. We can see the presence of a polygonal lattice with edges measuring below 10 nm at high concentrations and a cube at low. At low protein concentrations small aggregates with dimensions below 10 nm were also observed on AFM, which is consistent with the expected size of the self-assembled box.

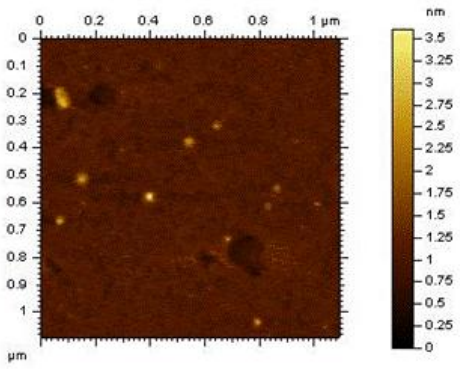

Figure 19: *AFM image: K2 concentration* $0.5\mu g/ml$ [5]

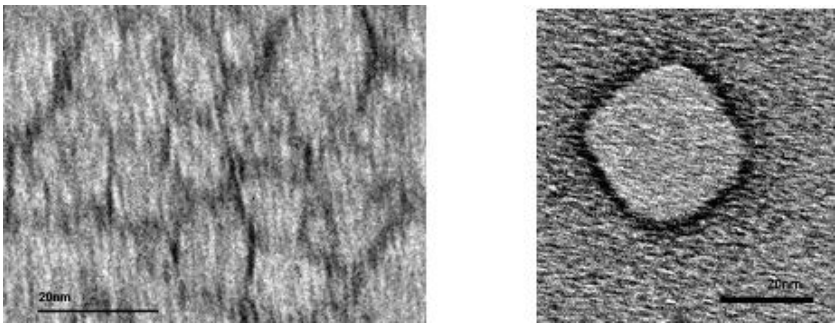

Figure 20: *TEM image: K2 concentration* $5\mu g/ml$ *(left) and* $0.5\mu g/ml$ *(right)* [5]

This shows that we can indeed form topologically different structures made of polypeptides. K2 thus represents the first example of self-assembled polygon composed of coiled-coil segments. There is clearly room for improvement, as the lattice contains a lot of defects, which could probably be solved by optimizing the polypeptide concentration, refolding conditions or modifications of the linker.

7 CONCLUSIONS

The biological self-assembly related topics are relatively new and still developing research area, interesting not only for its complexity but also for practical applications self-assembled structures enable. They can be used as scaffolds for attachment of different biological molecules, they can formate biosensors, artificial enzymes and on the other hand be biomineralized to conduct electricity or used as bio-nanorobots. Relatively simple manufacture of such polypeptide nanomaterials is yet another argument for future application in industrial processes.

To demonstrate above mentioned ability to use such coiled-coil assemblies in practice we manufactured an ultrafiltration membrane [5] which successfully filtered viruses from the solution. The design of polypeptide lattice has been carefully chosen to allow changing the size of pores in order to be able to adjust to the size and type of particles we want to filter. Moreover, we have shown that assembly and disassembly can be regulated by adding small molecules, further demonstrating the power of coiled-coil units to build nanomaterials with programmable properties.

A lot of work still has to be done, primarily on the field of tertiary structure prediction, but spectacular results from DNA self-assembly and recent work on polypeptides encourage further research in this area to the point, where our imagination and creativity will be the sole limits.

8 References

1. Sweeney B., Zhang T., Schwartz R. 2008. *Exploring the parameter space of complex self-assembly through virus capsid models.* Biophysical Journal, 94, 3: 772–783
2. http://2009.igem.org (1.12.2009)
3. He Y., Ye T., Su M., Zhang C., Ribbe A.E., Jiang W., Mao C. 2008. *Hierarchical self-assembly of DNA into symmetric supramolecular polyhedra.* Nature, 452: 198-202
4. Rothemund P.W.K., 2006. *Folding DNA to create nanoscale shapes and patterns.* Nature, 440.
5. http://2009.igem.org/Team:Slovenia (1.12.2009)
6. Jelerčič U. 2008. FMF. *Proteini.* Seminar 3. letnik
7. Nelson D. L., Cox M. M. 2005. *Lehninger Principles of Biochemistry.* Freeman, 4th ed.
8. Pablo Echenique. 2008. *Introduction to protein folding for physicists.* http://arxiv.org/abs/0705.1845 (1.12.2009)
9. Lazaridisa T., Karplus M. 2003. *Thermodynamics of protein folding: a microscopic view.* Biophysical Chemistry, 100 367–395
10. Huang K. 2005. *Statistical Physics and Protein Folding.* World Scientific.
11. Kuščer I., Žumer S. 2006. *Toplota.* DMFA - založništvo
12. Sutto L., Tiana G., Broglia R.A. 2005. *Hierarchy of events in protein folding: beyond the Go model* at http://arxiv.org/ftp/q-bio/papers/0601/0601044.pdf (1.12.2009)
13. Bryngelson J. D., Onuchic J. N., Socci N. D., Wolynes P. G. 1994. *Funnels, Pathways and the Energy Landscape of Protein Folding: A Synthesis* at http://arxiv.org/abs/chem-ph/9411008v1 (1.12.2009)
14. Onuchic J. N., Wolynes P. G. 2004. *Theory of protein folding.* Current Opinion in Structural Biology, 14:70–75
15. Grigoryen G., Keating A.E. 2008. *Structural specificity in coiled-coil interactions.* Current Opinion in Structural Biology, 18: 477-483
16. Parry A.D.D., Fraser R.D.B., Squire J.M. 2008. *Fifty years of coiled-cois and alpha-helical bundles: A close relationship between sequence and structure.* Journal of Structural Biology
17. Brown J.H., Cohen C., Parry D.A.D. 1996. *Heptad Breaks in alpha-Helical Coiled Coils: Stutters and Stammers.* PROTEINS: Structure, Function and Genetics, 26: 134-145
18. Phillips G.N.Jr. 1992. *What Is the Pitch of the alpha-Helical Coiled Coil.* PROTEINS: Structure, Function and Genetics, 14: 425-429
19. Bromley E.H., Sessions R.B., Thomson A.R., Woolfson D.N. 2009. *Designed alpha-helical tectons for constructing multicomponent synthetic biological systems.* Journal of the American Chemical Society, 131(3): 928-93
20. Fong J.H, Keating A.E., Singh M. 2004. *Predicting specificity in bZIP coiled-coil protein interactions.* Genome Biology, 5: R11
21. Hagemann U.B., Mason J.M., Müller K.M., Arndt K.M. 2008. *Selectional and mutational scope of peptides sequestering the Jun-Fos coiled-coil domain.* Journal of Molecular Biology, 381:73-88
22. Mason J.M., Müller K.M., Arndt K.M. 2007. *Considerations in the design and optimization of coiled-coil structures.* Methods in Molecular Biology, 352: 35-70
23. Mason J.M., Müller K.M., Arndt K.M. 2006. *Semirational design of Jun-Fos coiled coils with increased affinity: Universal implications for leucine zipper prediction and design.* PNAS, 103, 24: 8989-8994