Writing with amino acids: designing the folding and binding of model proteins
Coluzza, I.

Citation for published version (APA):
Coluzza, I. (2005). Writing with amino acids: designing the folding and binding of model proteins

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
6 Refoldable proteins and substrate interaction

Loosely speaking, folding is the process by which proteins explore their conformational space and find those conformations that have the lowest free energy. Many proteins can fold into more than one structure. The relative stability of these structures can often be changed by external agents such as the absorption of light, the binding to a substrate or the hydrolysis of a fuel molecule such as ATP.

In this chapter we explore the refolding of model proteins that is caused by binding to a substrate. To this end, we used the approach described in Chapter 2 to design protein-like lattice heteropolymers that have different stable conformations depending on whether they are bound to a substrate or free in solution. We considered three different systems that differ in the size of the interacting units. For small substrates, we observe that an increase in temperature induces the protein to unbind from the substrate, yet remain in its native state. For larger protein-substrate interaction sites, the bound conformation remains folded while the unbound conformation is denatured. We also considered the case where the unbound protein does not have a well-defined native structure. We found that, in that case, binding to a specific substrate could induce folding in the disordered protein.

6.1 Introduction

Proteins can change their conformation depending on their physical or chemical environment. A common example of such a structural change is unfolding, which happens when the protein is heated or exposed to denaturing substances. Denaturing is not a particularly subtle conformational change as the resulting state of the protein is disordered. However, in many cases proteins respond to an external stimulus by changing their spatial arrangement from one specific conformation to another. Driven conformational changes are thought to be at the root of many of the tasks that proteins perform. This is the reason for why proteins are often compared to nano-machines. Examples of proteins that perform a task by changing their conformation are motor proteins that generate forces or transport materials from one part of a cell to another.

The motor action is usually induced by the binding and subsequent hydrolysis of a “fuel” molecule, such as ATP. The energy that is released during the burning of this fuel is used to induce a large-scale rearrangement of the protein backbone. This, in its turn, causes the protein to move with respect to the substrate to which it is bound. Whilst this general picture of motor action is generally believed to be correct, the details of the motor action are, at present, largely
6 Refoldable proteins and substrate interaction

unknown.

Several models have been proposed to account for the mechanism of molecular-motor action. We will not review these models but rather describe some common features. The motor action is usually assumed to involve changes in the tertiary structure of the protein that is represented as a set of springs and links that connect larger structural units. When the APT is hydrolyzed, some of the physical links are cut and the energy stored in the springs force the protein to move to a new equilibrium conformation. In a number of cases, this picture could be supported by structural data derived from x-ray crystallography. In these experiments, the proteins are "frozen" during different stages of their power cycle. The resulting protein conformations are thereupon crystallized and studied by x-ray crystallography. In this way, conformational changes in the protein could be studied in detail. Clearly, this approach yields insight in the sequence of allosteric transitions that are involved in the motor action of proteins. However, necessarily, the technique neglects the contribution of thermal fluctuations and focuses on those parts of the molecule that are highly ordered. Much less is known about the evolution of disordered domains in the protein or about the role that fluctuations play during the power cycle of a motor protein. That fluctuations can be very important follows, for instance, from a recent study by Hawkins and McLeish [45] who proposed a coarse-grained model for E-coli lac and trp repressor. The binding of this repressor to its substrate can be strongly influenced by the binding of a second molecule to the repressor: even though this second binding hardly changes the conformation of the repressor, it does affect the thermal fluctuations in the molecule and this, in its turn, was shown to have a large effect on the binding of the repressor to its substrate.

In the present chapter, we use simulations of a simple lattice model to explore the effect of substrate binding on the structure and fluctuations of model proteins. Our approach is as follows: we consider pairs of proteins and substrates. The substrate are designed such that the binding will induce a conformational change in the protein. For simplicity, we assume that the interactions between the monomeric units of the protein and the substrate are similar to the intramolecular interactions between amino acids that belong to the same protein. In the context of our lattice model, this means that the amino-acid-substrate interactions are determined by the same interaction matrix as the intra-molecular interactions of the amino acids belonging to the same protein.

6.2 Methods

The model system that we consider is similar to the one discussed in the previous chapter, namely a protein confined in a cubic box in the presence of an immobile substrate. The conformational energy is defined by Eq. 5.2.

We start by designing simultaneously the sequence of amino acids to be compatible with both the initial (unbound) state (A) and the final bound state (B). Moreover, we impose that the most stable structure in the bound state is not the same as in the unbound state. As before, the design stage involves a Monte Carlo sampling over amino-acid sequences. The acceptance of a sequence-changing trial move is determined by three criteria: the first two are Metropolis-like rules that ensure that sequence changes that greatly increase the energy of either state
Figure 6.1: Spatial arrangement of the chain in the structures used to test the model
Figure 6.2: Spatial arrangements of the chain used to study the folding upon binding of a random domain.
6 Refoldable proteins and substrate interaction

A or state B (or both) are rejected. The third acceptance criterion is related to the structural heterogeneity of the amino-acid sequences. This rule (Eq. 2.14) was described the preceding chapters.

Once a suitable sequence has been generated, we focus on the equilibrium properties of the system, and calculate the free energy as function of the number of native contacts (Eq. 3.2), and of the number of contacts with the substrate $Q_s$. In this way, we measure how the increase in the number of contact with the substrate correlates with the change in the native structure of the chain.

In the case where we studied protein folding induced by the substrate, we had to generate sequences that were in the coil state when unbound. We found that if we increased the configurational design "temperature" to a value that would generate random coil states in solution, then we would also find disordered bound states, yet when we chose a lower configurational temperature, the proteins in solution would typically fold into some compact state. To resolve this problem, we devised a scheme to control the randomness of the protein sequence. We randomly selected a certain percentage of amino acids that could evolve without constraints. In practice the design algorithm would randomly change their identity when selected for a mutation move, regardless of the Metropolis acceptance criteria. These amino acids would typically be irrelevant for folding. We found that, in this way, it was possible to design proteins that would be in the random coil state when unbound, yet in an ordered state when bound.

6.3 Results

To study the influence of a substrate on the equilibrium properties of our model protein we considered three different conformational changes induced by substrates of different sizes. In Fig. 6.1 we show the target structures between which the transitions occur: $1 \leftrightarrow 2$, $3 \leftrightarrow 4$, and $5 \leftrightarrow 6$. In addition we studied the case of a protein that does not fold into a native state when it is in solution. However, upon binding the molecule assumes a designed target structure (see figure 6.2). This model could be relevant for the understanding of the role that random domains play in protein binding. The same design procedure was applied in every case. We therefore limit our explanation to design of the sequence that, upon binding, undergoes a conformational change from structure 1 (Fig. 6.1a) to structure 2 (Fig. 6.1b). Following the procedure explained in chapter 2 we optimize the conformational energy of the chain in both structure 1 (see Fig. 6.1a) and 2 (see Fig. 6.1b).

After eight independent simulations (typical length per run: $10^9$ MC trial moves), we collect those sequences for which structure A is most stable in solution whilst structure B is most stable when bound. In all cases, we found at least one sequence that satisfied these constraints. In Tab. 6.2 we show the sequences that we selected to study the conformational changes shown in figure 6.1.

The design procedure is slightly different in the case where the molecule is in the disordered state in solution and folds upon binding (Fig. 6.2). Following the procedure explained in the Methods section above, we designed the protein in the bound state with different percentages of random amino acids ranging from 0% to 60%. The results are a group of sequences D0-D60 shown in Tab. 6.3.
### Table 6.2: Sequences generated for the test structures (Fig. 6.1).
Each letter represents a different amino acid [3]. The letters in bold are the amino acids of the substrate

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Table 6.2: Sequences generated for the test structures (Fig. 6.1). Each letter represents a different amino acid [3]. The letters in bold are the amino acids of the substrate</th>
</tr>
</thead>
<tbody>
<tr>
<td>RHFSYTRRMDDRCWVCDACVMCT</td>
<td>Sequence A</td>
</tr>
<tr>
<td>PHWLEYNKENPKIMEQRKWGEDP</td>
<td></td>
</tr>
<tr>
<td>KFAEQNKM MSQ</td>
<td></td>
</tr>
<tr>
<td>LEASPSKIREGYPGRTRDFYWCKDLEC</td>
<td>Sequence B</td>
</tr>
<tr>
<td>MNCKILECNWCKIRECMHFRDPDF YWCKQVECM- NCKV VATGQHQHI</td>
<td></td>
</tr>
<tr>
<td>PRDGLWGRDQPRDFMIFRDYMKDCLW CKEWNKEMICRENNKDLWCK ENMKECMICK- EWFKDCWCKEKNKEMIC</td>
<td></td>
</tr>
<tr>
<td>CREM PRQFMIGHQHHHPGLVTSTYAVVAAVT SYYPSQAHAVGSTQ</td>
<td>Sequence C</td>
</tr>
</tbody>
</table>

### Table 6.3: Sequences generated for the study of binding random domain proteins (Fig.6.2).
Each letter represents a different amino acid [3]. The letters in bold are the amino acids of the substrate

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Table 6.3: Sequences generated for the study of binding random domain proteins (Fig.6.2). Each letter represents a different amino acid [3]. The letters in bold are the amino acids of the substrate</th>
</tr>
</thead>
<tbody>
<tr>
<td>MKCREWLKDREIMK DCEWNRFREPLKD HQITVMFPPWQCYCSTSAYGDVVIYNSNQFAGTH</td>
<td>D0</td>
</tr>
<tr>
<td>MKCREWLKDREIMK DCEWNRFREPLKD HQITVMFPPWQCYCSTSAYGDVVIYNSNQFAGTH</td>
<td>D6</td>
</tr>
<tr>
<td>MKCREWLKDREIMK DCEWNRFREPLKD HQITVMFPPWQCYCSTSAYGDVVIYNSNQFAGTH</td>
<td>D12</td>
</tr>
<tr>
<td>CRNPECFKQWEGCK MRECIEDWELGKM PDVAFHHSCQNTNHYAWQGVFILTLDRYHM</td>
<td>D18</td>
</tr>
<tr>
<td>KMIPWECMNDWCKM RLWERMIEWDPR NY-CFCKEADFVILYNYSTQHGRRQSTVAAKTT</td>
<td>D24</td>
</tr>
<tr>
<td>GECPELRWWRFRE MCKDPEVKQFNMM DIYIK- TATHCADCQDSVPQSLNYHQLYGIW</td>
<td>D30</td>
</tr>
<tr>
<td>GECPELRWWRHIRE MCKDPEHKQFNNA DMIISIFNTKSCYDLSFNGAGPQAQHTTVVW</td>
<td>D36</td>
</tr>
<tr>
<td>GGCPELRWWRFKE MCKDPNHNCEFNM DMSIK- TATTACQDYIWNFLGKQSQHYHYVP</td>
<td>D42</td>
</tr>
<tr>
<td>GGCPELRWWMQFRR WHDDPNEVEHFNI KMSIK- TATGCWCYRHLNFAMYQAYKSKVL</td>
<td>D48</td>
</tr>
<tr>
<td>GGCPELRWWMQFRR WHDDPNEVEHFNI KMSIK- TATGCWCYRHLNFAMYQAYKSKVL</td>
<td>D54</td>
</tr>
<tr>
<td>GGCPELRWDMHFRE WCKDPNWKHFNIV RMI- AIKGSPTCQKVALDFSQEAYNMVM</td>
<td>D60</td>
</tr>
</tbody>
</table>
6 Refoldable proteins and substrate interaction

6.3.1 Free energy calculations

Having designed the sequences for the structures depicted in figures 6.1 and 6.2, we verified that the generated sequences do indeed fold into the respective target structures in both the bound and (where applicable) the unbound states. In Fig. 6.3.a-6.5.a, we plot the free energy of sequences A,B,C respectively, as a function of the number of native contacts \( Q \) (Eq. 3.2) at a temperature of \( T = 0.1 \). In each plot we distinguish between conformations that do and do not touch the substrate. A common feature of the three proteins is that they fold into different target structures, depending on their binding state. For example, in Fig. 6.3, the lowest free-energy conformation in the bound state corresponds to structure 1 (\( Q = 18 \)), whilst in the unbound state structure 2 (\( Q = -12 \)) has the lower free energy. The same applies for sequences B and C that were designed for the transitions 3 ↔ 4 and 5 ↔ 6 respectively. This result demonstrates it is feasible to design model proteins that undergo conformational changes upon binding to a substrate. In addition, in Fig. 6.3 the barrier to go from conformation 2 (\( Q = -12 \)) to conformation 1 (\( Q = 6 \)) is higher for the unbound conformation than for the bound conformations. This suggests that binding is likely to precede refolding. Additional evidence that this may be the case comes from the shape of the free energy surface \( F(Q, Q_s) \) that is a function of the number of native contacts \( Q \) and of the number of contacts with the substrate \( Q_s \) (see Figs. 6.7.a-6.9.a). For instance, Fig. 6.7.a shows that when the protein has an order parameter close to that of the unbound native state (conformation 2), the bound states with high values of \( Q_s \) are not favorable. Rather, the free-energy landscape for refolding is fairly flat except in the vicinity of the target state (1). This suggests that the protein is first weakly absorbed on the substrate. From then on refolding and increased binding\(^1\) to the substrate occur together.

Let us next compare the behavior of the different substrate sizes and consider the effect of temperature. To test the thermal stability of the different conformational changes, we increase the temperature until we reach a regime where the native state of the free protein is in equilibrium with the native bound conformation. For small substrates (1 ↔ 2 and 3 ↔ 4) the temperature increase will favor the unbound conformation, without denaturing the protein. The situation is different for the transition 5 ↔ 6, where the size of the substrate is larger. In this case, there is a temperature region where the bound state still folds into the designed structure, whilst the unbound state is denatured. Hence in this case the substrate acts to increase the thermal stability of a particular protein conformation.

Note that an induced conformational change such as the one from 1 ↔ 2 could act as a form conformation-mediated signal transmission. The interaction of a protein with a small molecule or small binding site, induces a substantial rearrangement of the chain, which changes the nature of the exposed surface of the protein. It would interesting to study the nature of this signal transmission more extensively, such a study would fall outside the scope of this thesis.

Let us finally consider the case of a protein coil that folds into a compact state when binds to a substrate. In Fig. 6.6.a-b we plot the free energy of, respectively, the free and bound states of sequences D, as function of the number of native contacts \( Q \) (Eq. 3.2). Not surprisingly,

\(^1\)Note that the definition of the binding free energy here is not yet system size independent. In order to correct for this we could perform the same simulation with periodic boundary conditions or compute the concentration of protein in the bulk.
6 Refoldable proteins and substrate interaction

Figure 6.3: Plots of the free energy $F(Q)$ of sequence A as a function of the number of native contacts $Q$ (Eq. 3.2), at $T = 0.10$ (a) and at temperature $T = 0.30$ (b). States that touch the substrate (squares) have been plotted separately from those that do not (circles). The curve corresponding to the touching states is longer, because in the definition of the order parameter we take into account also the native contacts with the substrate. All data were obtained with a combined parallel tempering and umbrella sampling simulation.

Figure 6.4: Plots of the free energy $F(Q)$ of sequence B as a function of the number of native contacts $Q$ (Eq. 3.2), at $T = 0.10$ (a) and at temperature $T = 0.30$ (b). States that touch the substrate (squares) have been plotted separately from those that do not (circles). The curve corresponding to the touching states is longer, because in the definition of the order parameter we take into account also the native contacts with the substrate. All data were obtained with a combined parallel tempering and umbrella sampling simulation.
Figure 6.5: Plots of the free energy $F(Q)$ of sequence C as a function of the number of native contacts $Q$ (Eq. 3.2), at $T = 0.10$ (a) and at temperature $T = 0.50$ (b). States that touch the substrate (squares) have been plotted separately from those that do not (circles). The curve corresponding to the touching states is longer, because in the definition of the order parameter we take into account also the native contacts with the substrate. All data were obtained with a combined parallel tempering and umbrella sampling simulation.

Figure 6.6: Plots of the free energy $F(Q)$ of sequences D0-D60 (0-60% of random amino acids) as a function of the number of native contacts $Q$ (Eq. 3.2), at $T = 0.10$. States that touch the substrate are plotted separately (b) from those that do not (a). The gap between the curves does not have a physical meaning, and has been introduced only to separate the free energies of sequences that folds in the bulk from the one that do not. Note that all proteins fold upon binding.
above a certain threshold of "randomness" (30%) the unbound chain does not have a compact native state. However we found that even in the case where 60% of the amino acids were chosen at random, the substrate-bound state would still fold into the native state. Of course, these numbers should be treated with caution as the quantitative results of the simulations are expected to be model dependent. The main message of Fig. 6.6.a-b is that a disordered protein (or, for that matter, a disordered protein domain) can be involved in selective binding and, in the process, fold into an ordered conformation. Although we have not tested directly whether protein protein can be mediated by an apparently random domain, the present results strongly suggest that this is the case.

To summarize, we have used a lattice model to describe the conformational changes in protein systems. This particular phenomenon is often triggered by the interaction of the protein with an external agent, that we model as substrate fixed in the simulation box. The first result was the successful design of a system with two equilibrium conformations, one for the bound state and the other one for the free state. We then computed the free energy of the proteins at different temperatures – distinguishing between touching and non-touching conformations. The behavior at low temperature was mainly characterized by a strong preference for the bound states 1, 3 and 5. When the temperature of the system was increased, the free energy of the unbound states decreased because of the increased importance of translational entropy. However the response of the system, was not simply a gradual denaturation of the proteins. Rather, it depended on the number of interactions with the substrate. At higher temperatures the inter-molecular interactions of small substrates (1 ↔ 2 and 3 ↔ 4) were not strong enough to compensate for the increase of the translational-entropy term, however the intra-molecular bonds were still stable and they could keep the protein in the native state 1 and 3 respectively. This picture starts to change when we consider a larger substrate. In this case the strength of the intra-molecular and inter-molecular interactions were comparable. As consequence, the translational-entropy term was never strong enough to favor the unbound states, before the protein was unfolded.
Figure 6.7: Plots of the free energy landscape $F(Q, Q_s)$ of sequence A as a function of the number of native contacts $Q$ (Eq. 3.2) and of the number of contacts with the substrate $Q_s$, at $T = 0.10(a)$ and at temperature $T = 0.30(b)$. All data were obtained with a combined parallel tempering and umbrella sampling simulation.
Figure 6.8: Color Online) Plots of the free energy landscape $F(Q,Q_s)$ of sequence B as a function of the number of native contacts $Q$ (Eq. 3.2) and of the number of contacts with the substrate $Q_s$, at $T = 0.10$(a) and at temperature $T = 0.30$(b).
Figure 6.9: Color Online) Plots of the free energy landscape $F(Q, Q_s)$ of sequence C as a function of the number of native contacts $Q$ (Eq. 3.2) and of the number of contacts with the substrate $Q_s$, at $T = 0.10$(a) and at temperature $T = 0.50$(b).