Developing and analyzing idealized models for molecular recognition

Hans Behringer [a,1], Thorsten Bogner [a], Alexey Polotsky [b], Andreas Degenhard [a] and Friederike Schmid [a]

[a] *Fakultät für Physik , Universität Bielefeld, D–33615 Bielefeld, Germany*

[b] *Sérvice de Physique de l'Etat Condensé CEA, 91191 Gif-sur-Yvette cedex, France*

Abstract

We study equilibrium aspects of molecular recognition of two biomolecules using idealized model systems and methods from statistical physics. Starting from the basic experimental findings we demonstrate exemplarily how an idealized coarse-grained model for the investigation of molecular recognition of two biomolecules can be developed. In addition we provide details regarding two model systems for the recognition of a flexible and a rigid biomolecule respectively, the latter taking into account conformational changes. We focus particularly on the interplay and influence of the correlations of the residue distributions of the biomolecules on the recognition process.

Key words: Biophysics, molecular recognition, model reduction, coarse-graining

[1] e-mail: behringe@Physik.Uni-Bielefeld.DE

1 Introduction

Biological systems such as the immune system are characterized by a complicated and complex interplay of biological processes. A crucial ingredient for the immune system to work efficiently is the ability of antibodies to specifically recognize corresponding antigens (Alberts et al., 1994; Kleanthous, 2000). In general molecular recognition can be viewed as the ability of a certain biomolecule, also referred to as the *recognition agent* or *recognition molecule* in this article, to interact preferentially with a particular target molecule although a vast variety of different but structurally similar rival molecules are present. Recognition processes are governed by the interplay of noncovalent interactions of comparable strengths such as ionic binding, the van der Waals interaction, the formation of hydrogen bonds and hydrophobicity. The noncovalent interactions between the residues of the biomolecules lead to the formation of a complex where the two biomolecules form a mutual interface consisting of one or more patches on their surfaces. In addition long-range electrostatic interactions are believed to pre-orientate the molecules so that the probability of a contact of the interface patches upon a collision of the molecules is increased (e.g. Janin, 2000; Wodak and Janin, 2003). The simultaneous presence of different types of interactions and the fact that the associated energy scales do not separate leads to a complicated interplay among them. Therefore a detailed description of recognition processes poses a difficult and involved problem. An understanding of the principles of molecular recognition processes is not only important from a scientific point of view but also for biotechnological and biomedical applications. The knowledge of these principles is a necessary input for the design of synthetic heteropolymers with molecular recognition ability so that they can interact with a biological environment, i.e. biomolecules, cells and tissues, in a programmable way (see e.g. the review by Peppas and Huang (2002)).

Modern computer facilities make it possible to study the molecular recognition process between two molecules on a single molecule level (e.g. Halperin et al., 2002; Brooijmans and Kuntz, 2003). However, it is not yet possible to incorporate the heterogeneity of the environment on a single-molecule level into such studies. An important question that arises in the study of molecular recognition processes is the phenomenon of specificity which is basically the fact that recognizing biomolecules bind to each other although a huge amount of competing rival molecules are present. In an aqueous environment noncovalent bonds are typically of the order of 1-2 kcal/mole and are therefore only slightly stronger than the thermal energy $k_{\mathrm{B}}T_{\mathrm{room}} \simeq 0.62$ kcal/mole at physiological conditions. The specificity of biomolecular recognition is thus only achieved if a large number of functional groups of the two molecules to recognize each other precisely match and thus a sufficient number of corresponding noncovalent bonds can be formed. This principle is often called complementarity in the

literature (Pauling and Delbrück, 1940). Specificity is thus a genuinely cooperative effect. The problem of specificity of molecular recognition processes can be tackled using methods from statistical physics. Characteristic properties of the heterogeneity of the biological environment, which cannot yet be taken into account on a single molecule level, can be incorporated into the analysis on a statistical basis. In addition statistical methods allow the identification of the relevant degrees of freedom that influence the recognition processes. In this context the study of idealized models can provide insight into the general principles of specificity in recognition processes.

The investigation of molecular recognition using methods from statistical physics can therefore serve as an example to illustrate model reduction for biophysical processes. For the construction of such reduced, idealized models the relevant degrees of freedom have to be identified. In an ideal situation this is done by starting with the microscopic model that contains all information about the system and then applying approximations that are suitable for the particular context to be described. One then arrives at a reduced model which contains only the relevant degrees of freedom which are sufficient to describe the behaviour of the system. The approximations applied to the microscopic model are usually justified by experimental observations. This approach is nicely illustrated by the theory of magnetism. Starting from a complete microscopic quantum mechanical description of a solid one arrives at the Heisenberg model which contains only those degrees of freedom any more that are relevant for the explanation of magnetic phenomena (e.g. Ashcroft and Mermin, 2001). The full quantum mechanical model of the solid contains for example degrees of freedom which are related to the vibrations of atoms. However these lattice vibrations, which are important for the propagation of sound, do not influence the magnetic properties of the solid. Therefore they are neglected during the reduction process. On the other hand experimental investigations of magnetic solids have revealed the importance of local magnetic moments, which are related to the spins of the atoms, for magnetic phenomena. Therefore these degrees of freedom and their interactions are kept. In most biological systems, however, a detailed microscopic model does not exist and therefore one usually starts directly form experimental observations to identify the relevant information for constructing a model to describe the behaviour of the system.

In this article we review how coarse-grained models for molecular recognition processes can be developed and summarize results we obtained by employing these models for the investigation of the principle mechanisms underlying molecular recognition (Polotsky et al., 2004a,b; Bogner et al., 2004). In the next section we will particularly demonstrate how these models are developed starting from the findings of experimental investigations of recognition processes. The models are then analyzed using methods from statistical physics. In subsection 2.3 we briefly discuss other coarse-grained models for molecular

recognition investigated in the literature. The last section 3 summarizes our results and gives a fairly detailed perspective of possible extensions of the discussed models. Such extensions are motivated by questions arising currently in biomolecular and biotechnological sciences.

2 Coarse-grained models for investigating molecular recognition

In recent years the structural properties of protein-protein complexes formed during the recognition process have been clarified by many studies (Janin and Chothia, 1990; Jones and Thornton, 1996; Lo Conte et al., 1999; Jones and Thornton, 2000; Janin, 2000; Chakrabarti and Janin, 2002; Wodak and Janin, 2003). Although different complexes show rather different properties one fundamental ingredient in molecular recognition can be identified. The 20 amino acids appearing in natural proteins can be classified with respect to their degree of hydrophobicity where two types of residues are distinguished, namely polar and hydrophobic ones. The studies of the protein-protein complexes revealed that the residues at the interface between the two proteins are more hydrophobic than those of the rest of the protein surface which contains in addition a considerable fraction of polar residues. From investigations of the interior of proteins it is also known that it contains mostly hydrophobic residues. Therefore, the hydrophobicity is expected to be the most dominant ingredient for both protein-protein recognition and also protein folding which leads to the tertiary structure of the protein.

In this section we present how idealized models for the investigation of molecular recognition can be developed and investigated. Aspects concerning the development of idealized models will be particularly stressed in subsection 2.1. In the proposed modelling approaches the structure of the recognition molecule is described as a heteropolymer chain consisting of hydrophobic (H) and polar (P) residues. The degree of hydrophobicity is therefore represented in a coarse-grained way by only two distinct values. Analogously the interface patch on the surface of the target molecule is modelled by a heterogeneous surface pattern, where each residue is again either hydrophobic or polar. In addition the target molecule is approximated by a flat surface structure, i.e. no geometrical contributions related to the curvature of the interface are taken into account at this point. In the two following subsections we consider models of molecular recognition between two rigid proteins on the one hand (section 2.1) and a rigid and a flexible biomolecule on the other hand (section 2.2). Most protein-protein recognition processes involve proteins that do not change their conformations during the association process and thus remain rigid. Nevertheless there are notable examples where at least one participating biomolecule is flexible and therefore may undergo a conformational change upon association (e.g. Peppas and Huang, 2002; Wodak and Janin, 2003). An important

example of recognition processes where one biomolecule is flexible is the recognition of a flexible DNA molecule by a rigid protein (e.g. Chakraborty, 2001; Bruinsma, 2002).

2.1 Recognition between two rigid biomolecules

In this subsection an idealized model is developed that can be used as a starting point to analyze the principles that govern the recognition process of two rigid proteins which do not change their conformation when they interact with each other. In order to investigate specific adsorption, one has to analyze the association of the target with different recognition molecules. The recognition agent, the biomolecule that recognizes the target, is therefore modelled as a heteropolymer whose structure is not yet specified. It is in addition surrounded by solvent molecules (S) which are also polar molecules. The tertiary structure of the protein is determined by requiring the chain to be compactly folded which means that the total number of solvent contacts of the residues is minimized (Li et al., 1996). In this sense the recognition agent is also a rigid molecule. A finite number of possible compact and not identical structures is presented to the target molecule thereby mimicking the heterogeneity of the environment encountered by the target molecule in a biological situation. As different recognition agents are considered one has to take the energy contributions from the interior of the recognizing protein into account. Apart from the internal energy the interaction energy at the interface between the proteins and the energy arising from the contacts with solvent molecules contribute.

In a first approximation we model the various contributions to the total energy in the following way. For simplicity the recognizing molecule is considered on a regular lattice so that each lattice site is occupied by a residue. Although various lattices will serve as possible simplifications of the problem here we consider only square lattices for the discussion of the model. The actual calculations are then carried out on either two-dimensional square lattices or on three-dimensional cubic lattices. The recognition molecule is surrounded by solvent molecules occupying sites at the boarder of the employed lattice and is in contact with the target protein also presented on a lattice whose sites are occupied by H- and P-residues. The interactions can then be expressed in terms of a spin model with the total energy given by

$$E_{\mathrm{tot}} = \sum_{\langle i,j \rangle} \sum_{\langle \alpha,\beta \rangle} \tau_i^{\alpha} \tau_j^{\beta} E_{\alpha,\beta}. \tag{1}$$

The variables τ_i^{α} describe the distribution of residues on the lattice and therefore the structure of the protein that is to recognize the target biomolecule. The parameters $E_{\alpha,\beta}$ model the interactions between two residues of type α

and β, where α and β are either of type polar (P), hydrophobic (H) or solvent (S). The first sum runs over neighbouring sites i and j on the lattice so that only residues close to each other can contribute to the total energy, the second sum is a sum over the different types of residues and molecules, i.e. $\alpha, \beta \in \{\mathrm{H}, \mathrm{P}, \mathrm{S}\}$. Note the sum over lattice sites contains the contacts of the residues with the solvent molecules and with the residues of the target biomolecule on the interface. Therefore this sum comprises the energy contributions arising from both intermolecular contacts within the recognition agent and intramolecular contacts at the interface with surface residues of the target and with solvent molecules. The spin variables τ_i^α describe whether site i is occupied by a residue/molecule of type α. Thus $\tau_i^\alpha = 1$ if site i is occupied by a residue/molecule of type α and zero otherwise. Therefore the variables τ_i^α model the structure of the biomolecules and determine the degree of chemical heterogeneity in the recognition molecule. The parameters $E_{\alpha,\beta}$ specify the energy contribution due to a contact between a residue/molecule of type α and a residue/molecule of type β. Contacts between residues of the same type, i.e. HH or PP contacts, lead to favourable energy contributions. Contrarily, contacts between residues of different types (i.e. HP contacts) are accompanied by unfavourable energy contributions. The four types of contacts are schematically illustrated in figure 1. In addition contacts between solvent

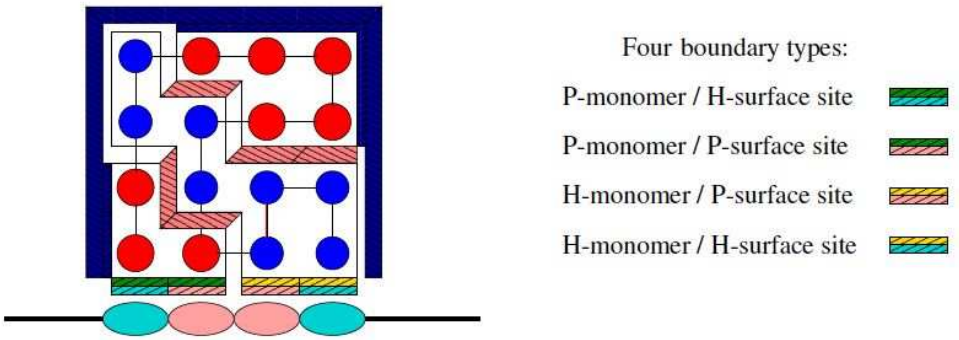

Fig. 1. A compactly folded chain attached to a target molecule and surrounded by solvent molecules (dark blue area) at the other three sides. The different types of contacts that occur are graphically illustrated.

molecules indicated by the outer dark blue area in figure 1 and hydrophobic residues also lead to an unfavourable energy, whereas contacts with polar ones to a favourable energy contribution. The overall energetics of the model can then be summarized by the relations $E_{\mathrm{HH}} \approx E_{\mathrm{PP}} \approx E_{\mathrm{PS}} < E_{\mathrm{HP}} \approx E_{\mathrm{HS}}$, where the particular values of the interaction parameters still have to be specified. Regarding the energetic preference based on the different energy contributions the hydrophobic residues tend to be buried away from polar ones and the polar solvent molecules. This models the hydrophobic effect that is assumed to be dominant in protein-protein recognition.

A fixed sequence of residues of the recognition molecule leads to a variety of

different conformations i.e. compactly folded structures of this molecule. Following the definition of the energy model in (1), the energy for each conformation can be computed, for both situations, away from the structured surface of the target molecule and attached to the surface. In both cases an energy spectrum is obtained including the computed energy values for the possible conformations. The adsorption energy is defined for a particular recognition molecule and a selected target molecule as the energy difference between the ground state energy of the polymer free in the solvent $E_{0,\text{free}}$ and the ground state energy of the adsorbed polymer $E_{0,\text{ads}}$, denoted as

$$E_{\text{ads}} = E_{0,\text{ads}} - E_{0,\text{free}}. \tag{2}$$

In the framework of the proposed model (1), the process of adsorption is studied with respect to a varying E_{ads}, i.e. a stronger adsorption and thus the mutual recognition of two biomolecules corresponds to a larger E_{ads}.

Regarding a given sequence of residues different conformations, i.e. compactly folded structures, exist. For all of these conformations and for all of the possible surface patterns at the interface patch of the target molecule one can now compute the associated energy spectra. In case a unique surface pattern with highest adsorption energy exists the chosen sequence is called *selective* with respect to this unique surface pattern. This in turn means that the two agents, the compactly folded recognition molecule and the target presenting the surface structure at the interface, recognize each other. Note that the pattern recognized as specific may only be part of the whole contact interface between the biomolecules. In general the complete recognition interface can be far more complex structured. However, in our modelling we have assumed a limited size of possible interaction regions located within a larger structure. This is schematically illustrated in figure 2.

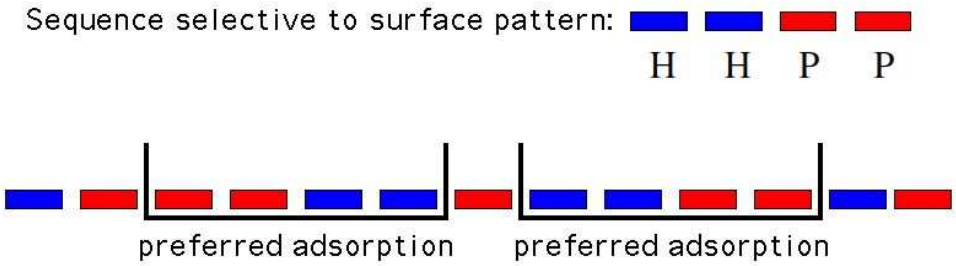

Fig. 2. Example of a surface pattern as recognized by a selective molecule. The pattern may only be one part of a much more complex target molecule.

One of the questions arising in the context of specific adsorption is whether the model allows to detect general rules underlying the recognition process. To accomplish this task statistical features from a set of sequences each selec-

tive for a particular surface pattern can be computed (Bogner et al., 2004). In this approach the sequence of residues is treated as a random vector which is then Fourier transformed. Thus, if common features within a set of sequences do exist, the Fourier components will show a strong correlation. Within a statistical analysis such a correlation manifests in the covariance matrix. Diagonalizing this matrix yields the eigenvalues which are a measure for the squared variances of these components. Low variances correspond to characteristic components regarding the set of residue sequences. Figure 3 shows the eigenvector corresponding to the lowest eigenvalue for six different surface patterns identified as selective for a presented recognition molecule. The shown eigenvectors are calculated for a three-dimensional $3 \times 3 \times 3$ system. As a result we obtain that the highest frequency components are identified

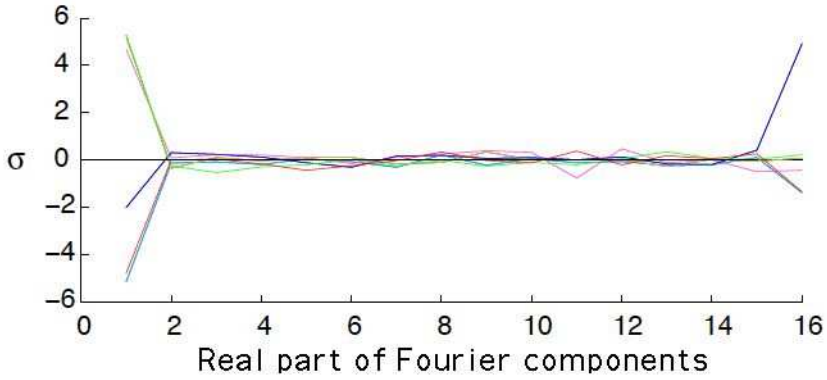

Fig. 3. Coordinates in Fourier space of the smallest variance eigenvector for six different surface patterns involved in selective recognition of different recognition molecules for the $3 \times 3 \times 3$ system. The frequency components are ordered from highest frequency (left) to lowest frequency (right), the latter corresponding to the highest indexing number in the figure.

as those with the largest contribution and accordingly dominate the recognition process. Therefore, a universal feature that governs molecular recognition within our model is identified as the local, small-scale structure of the modelled biomolecules. This general behaviour is found in both (small) two-dimensional and three-dimensional systems (Bogner et al., 2004). In addition, for some examined systems also the lowest frequency components, shown in figure 3 by the highest indexing number, add a major contribution. According to this observation the mean hydrophobicity in combination with the small-scale structure governs a specific recognition process.

2.2 Recognition between a flexible and a rigid biomolecule

As already mentioned above some recognition processes involve at least one flexible biomolecule so that conformational changes can occur. An important example is protein-DNA recognition. Recognition in this context is the specific

adsorption of the flexible biomolecule onto the chemically patterned surface structure of the rigid biomolecule (Sarai and Kono, 2005). An investigation of such processes will be an essential contribution for understanding cell biology and functioning. Regarding the fixed structure of the target molecule, this is then fully characterized by the chemical heterogeneity on its surface. However, to explore the contribution from chemical heterogeneity in full, the flexible recognition agents need to be enabled to adjust their conformations such that they predominantly attach to the most attractive surface sites. This is accomplished by relaxing the constraint of a compactly folded recognition agent when computing the energy of a conformation.

Recent simulation studies with respect to flexible recognition molecules at solid planar surfaces suggest a strong relationship between strongly correlated structures in the flexible biomolecule and the rigid surfaces (e.g. summarized by Chakraborty, 2001). In addition it was found that, upon increasing the strength of the interactions, the adsorption transition of heteropolymers on heterogeneous surfaces is followed by a second sharp transition, where the recognizing molecules freeze into conformations that match the surface patterns. In nature such a process where adsorption is followed by a freezing transition might be involved in the protein-DNA recognition, where the protein slides along the DNA molecule before finding its specific docking site (e.g. Bruinsma, 2002). However, regarding the immune system as another biological system based on recognition principles, i.e. the interaction between antigens and antibodies, it is not only important that a molecule recognizes a particular surface, since it is crucial that it does not adsorb to other molecular structures (Janeway et al., 1999). As a first approach to investigate such mechanisms the importance of cluster size matching between the surface pattern and the flexible recognition molecule can be investigated. To accomplish this task, one has to calculate the shift of the adsorption transition as a function of correlation lengths on the two interacting partners. Although simulations investigating the matching of cluster sizes were performed for particularly selected heterogeneous patterns, these do not allow to examine the dependence of recognition processes on the degree of heterogeneity (Semler and Genzer, 2003).

Again we use a coarse-grained model where coarse-graining is applied again on two levels, namely the description of the structure of the involved biomolecules and the interaction energy between them. Here we resort to a well known model from polymer physics to describe the flexible biomolecule that can undergo conformational changes during the recognition process as the constraint of being compactly folded is dropped (e.g. Doi and Edwards, 1986). The molecule is a chain of length N measured in residue units, each of equal size a. The conformation of the flexible biomolecule in space is specified by the vector $\mathbf{r}(n) = \{\mathbf{x}(n), z(n)\}$ with $n \in [0, N]$ parameterizing the curve representing the chain-like heteropolymer. In principle this vector can have continuous or discrete entries so that one works again on a lattice as has been done in

subsection 2.1. The type of the residue at position n in the chain is specified by the variable $\xi(n)$ which can be either continuous or discrete and is again related to its degree of hydrophobicity. The position of the residues on the surface to which the flexible polymer can attach is labelled by the two-dimensional vector $\mathbf{x}$, the type of the residue is specified by the variable $\sigma(\mathbf{x})$. The strength and the sign of the interaction between a monomer in the recognition molecule and a surface site is then determined by the product $\xi(n)\sigma(\mathbf{x})$. In particular, a positive contribution defines attraction and a negative one defines repulsion.

In what follows we formulate the model in terms of continuous variables. The sequence of monomers $\xi(n)$ for the flexible biomolecule is then assumed to be Gaussian distributed with mean value $\xi_0 = \langle \xi(n) \rangle$. The effective energy of the system consisting of the flexible molecule and the surface of the target is then given by

$$E_{\mathrm{tot}} \;=\; \frac{3}{2\beta a^2} \int_0^N dn \left(\frac{\partial \mathbf{r}}{\partial n} \right)^2 \;+\; \int_0^N dn \, V\left[z(n) \right] \cdot \xi(n) \cdot \sigma\left[\mathbf{x}(n) \right] . \tag{3}$$

The first term accounts for the flexibility of the biomolecule (e.g. Doi and Edwards, 1986). The parameter β is inverse proportional to the temperature $T = 1/(k_{\mathrm{B}}\beta)$ where k_{B} is the Boltzmann constant. The second term represents the energy contribution due to the interaction of the residues of the flexible biomolecule chain and the residues of the surface pattern. This contribution is basically determined by the residue-surface potential V which is taken to be attractive and short-range so that only residues that are in contact with each other contribute to the energy. The product $\xi\sigma$ models the hydrophobic effect so that hydrophobic residues (associated with negative values of ξ and σ) are buried away from polar ones (associated with positive values of ξ and σ).

Using this model for the energy of the two interacting biomolecules we investigated the influence of the correlation lengths of the residues of the two interacting biomolecules. The correlation function for the residues on the flexible recognition molecule (R) chain is taken to be

$$c(n_1, n_2) = \langle (\xi(n_1) - \xi_0)(\xi(n_2) - \xi_0) \rangle = \Delta_{\mathrm{R}}^2 \exp\left(-\Gamma_{\mathrm{R}} |n_1 - n_2| \right) . \tag{4}$$

The parameter Γ_{R} corresponds to the inverse contour correlation length on the flexible biomolecule and Δ_{R} gives the variance or correlation strength of the single-residue distribution. Note that specifying the correlation function and the mean ξ_0 then determines the Gaussian distribution function. Similarly the corresponding correlation function for the residues on the surface of the target molecule (T) is characterized by the parameters Δ_{T} and Γ_{T}. Following the introduced notation, the family of parameters Δ and Γ are a measure for

the chemical heterogeneity of the interacting partners. A possible experimental realization of predefined heterogeneity is referred to as *colouring*. Here a protein-like sequence is generated from an originally homogeneous chain containing only hydrophobic monomer units coloured with hydrophilic residues (Semler and Genzer, 2006).

The strength Δ and correlation length $1/\Gamma$ are then employed as parameters to describe the matching of the correlations of the flexible recognition molecule (R) and of the surface pattern of the target molecule (T). Within a first approximation it is further assumed that the chain and surface distributions should be neutral on average. i.e. the corresponding mean values are taken as $\langle \xi \rangle = \langle \sigma \rangle = 0$. Figure 4 schematically illustrates the extended model allowing for flexible biomolecular conformations of one of the two involved molecules.

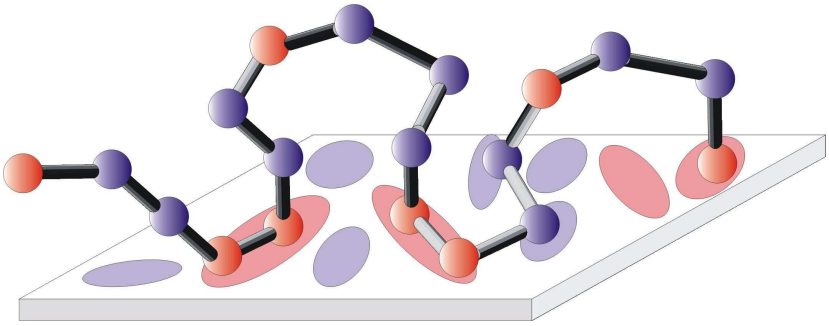

Fig. 4. Schematic illustration of the adsorption of a polymer chain onto a chemically structured surface. The molecule is free to adopted the most appropriate conformation.

Employing the model (3) it is possible to compute the adsorption transition as

$$\frac{\Delta_{\mathrm{T}}^2 \cdot \Delta_{\mathrm{R}}^2}{\Gamma_{\mathrm{T}} + \sqrt{6\Gamma_{\mathrm{R}}}} = \text{constant} \tag{5}$$

specifying the region of adsorption and desorption in the parameter space of variances Δ and inverse correlation lengths Γ (Polotsky et al., 2004a,b). Tuning selected parameters the equation (5) therefore describes the adjustment of the remaining parameters at the transition point. In addition, the term in (5) dominates the computation of the mean affinity between residues in the chain and sites of the target when compared with a homogeneous monomer distribution (Polotsky et al., 2004b). Therefore it is possible to deduce some dominating features governing the interplay between the employed cluster sizes: Incorporation of chemical heterogeneity at constant average load favours adsorption. In particular, utilizing correlation functions for describing the characteristics of the heterogeneity, it is found that increasing the correlation length and strength of the correlations the adsorption transition is shifted towards lower affinities.

In this subsection we briefly discuss other coarse-grained approaches, which are based on concepts used for the statistical description of the physics of glassy systems, to the investigation of statistical properties of molecular recognition processes.

In the approach proposed by Janin (1996) the density of states $M(E)$ for the binding modes between two rigid proteins is considered as a function of the binding energy E. Apart from the native complex, the lowest energy binding state with $E = 0$, there exists a variety of non-native, energetically less favourable states. These alternative binding modes are, for example, related to different relative orientations of the two proteins. The specificity of molecular recognition is basically related in this approach to the existence of an energy gap G between the energy of the native state and the one of the lowest lying non-native states. The density of states of the protein-protein complex is calculated numerically within a docking analysis (for details, in particular concerning the geometric representation of the proteins, see e. g. (Bernauer et al., 2005)), where the binding energy is assumed in a coarse-grained way to be proportional to the area of the interface between the two proteins, an assumption that seems to be well justified (e.g. Janin, 2000; Wodak and Janin, 2003). It is found that the low energy part of the numerically evaluated density of states can be well described by the density of states of the random energy model, which is one of the simplest models to describe the behaviour of glassy systems (e.g. Parisi, 2003). This observation allows the definition of two characteristic temperatures. The so-called specificity transition temperature T_{S} is given by the inverse slope of the tangent on the entropy $S(E) = \ln M(E)$ that passes through the energy $E = 0$ of the native state. At this temperature the system is found to be in the native binding mode with probability $1/2$ and below T_{S} the native binding mode becomes abundant. The slope of the tangent on the entropy curve at the lowest lying non-native states at $E = G$ defines the so-called glass transition temperature T_{g}. For temperatures below T_{g} only non-native states whose energy is close to $E = G$ can compete with the native state any more and the system can be trapped in a low-lying non-native state hinting at a glassy phase of the system. Within the random energy model approximation, i. e. a Gaussian density of states, the specificity tarnsition temperature T_{S} is determined by the ratio of the energy gap G and the width ΔE of the density of states. In this approach the recognition of two molecules A and B is described with respect to the spectrum of their association modes A-B irrespective of a possible presence of rival molecules C which might compete to form complexes A-C or B-C. In (Janin, 1996), however, it has been argued how rival molecules can be incorporated into this approach.

Wang and Verkhivker (2003) proposed an approach that is in spirit similar

to the one by Janin (1996), however, formulated on a more microscopic level. They also consider the binding modes of two rigid proteins. For a set of residues in multibody contact across the interface between the two proteins they assign an energy contribution ε which is assumed to be a random variable due to the sequence and interaction heterogeneity on the surface of the proteins. Again they consider the native binding mode and the non-native ones. They introduce in addition an overlap parameter Q which measures how close a non-native mode a is to the native one n. The overlap Q basically counts the number of residue contacts in the binding state a that also appear in the native mode n. Under certain assumptions (in particular a Gaussian distributed random variable ε and a restriction to pair contacts between residues) they calculate the average number of non-native binding states a that have an energy E and a fixed overlap Q between a and the native mode n. It turns out that for each value of Q the system can be modelled independently by a random energy model. For each Q they therefore get a glass transition temperature T_{g} below which the system is trapped in a low energy state of the subset of states characterised by Q. From the free energy of their system Wang and Verkhivker obtain the phase diagram of the binding between the two proteins. They find three phases, namely a native and non-native binding phase and a glass phase already hinted at in the Janin approach. Wang and Verkhivker (2003) then relate the specificity of the biomolecular binding to the fact that the binding temperature (specificity temperature in the nomenclature of Janin (1996)) should by higher that the glass temperature to avoid a nondiscrimination of the native binding mode with states trapped in the glass phase. As in the work of Janin the specificity then turns out to be affected by the ratio of the gap between the native binding state and the lowest non-native mode and the width of the density of states.

3 Perspectives and summary

Before summarizing the results let us discuss how the present approach for investigating molecular recognition can be extended to incorporate further ingredients. Here we focus on extensions of the model for the molecular recognition of two rigid proteins.

In the studies of protein-protein complexes the structural properties of the so-called recognition site, the contact interface between the biomolecules, have been addressed extensively. (Janin and Chothia, 1990; Jones and Thornton, 1996; Lo Conte et al., 1999; Chakrabarti and Janin, 2002). In most associations of two proteins the molecules are basically rigid although minor rearrangements of the amino acid side chains do occur. This observation of minor rearrangements of the amino acid side chains can be incorporated into the idealized model for rigid protein recognition studied above in section 2.1. At

the interface between the two proteins the residues have been modelled to be in contact and therefore contribute to the energy of the system. In order to incorporate the minor rearrangements of the side chains of the residues the following modified energy contribution of the interactions between residues at the interface is proposed (Behringer et al., 2006):

$$E_{\text{interface}} = \sum_{k} \sum_{\langle \alpha,\beta \rangle} \tau^{\alpha}_{\text{T},k} \tau^{\beta}_{\text{R},k} E_{\alpha,\beta}(\sigma_k). \tag{6}$$

The first sum takes into account all residues k of the interface and the variables $\tau_{\text{T},k}$ and $\tau_{\text{R},k}$ characterize the type of the residues of the target protein (T) and the recognition agent (R), respectively, at the interface site k. Note that apart form the interface energy $E_{\text{interface}}$ the total energy then comprises the further contributions from the residue interactions in the interior of the recognizing protein and the energy contributions from the contacts of the residues with solvent molecules. These contributions are modelled as in section 2.1. The additional variable σ_k in the general interface energy term (6) can take on different discrete values and takes the quality of the contact of the two residues at position k into account. On a very coarse-grained level one may, for example, distinguish only between good and bad contacts. In the case of a good contact the interaction between the residues $\tau_{\text{T},k}$ and $\tau_{\text{R},k}$ leads to a large favourable energy contribution whereas for a bad contact one has only a small contribution. A good contact may imply for example that the distance between the two residues is small, a steric hindrance on the other hand may result in a large distance and consequently one has a bad contact. For residues with a polar moment a good contact may be established if the moments are appropriately aligned to each other. The variable σ therefore models effects that are related to energy contributions stemming form the minor rearrangements of the side-chains of the amino acids when a complex is formed. This extension taking more details of the interface into account then allows to study the complementarity of the proteins at the interface from two perspectives. On the one hand one can ask whether hydrophobic residues are indeed buried away from polar ones at the interface leading to a complementarity related to the composition of the proteins at the interface. On the other hand one can now consider the complementarity of the shape of the proteins at the interface. Again one can now study the influence of the small-scale structure of the biomolecules on molecular recognition in this extended model approach.

The specificity of molecular recognition is closely related to the heterogeneity of the environment of the two recognizing molecules which has the consequence that different molecules compete for binding with the target molecule. In the approach considered above the heterogeneity has been mimicked by allowing different structures of the recognition agent where each possible structure was taken into account with equal probability. In reality however the different molecules will appear with different frequencies in the plasma of a cell, for

example. These frequencies have been optimized by natural evolution over a long period of time.

In the model system for molecular recognition this can be incorporated by introducing a first design or learn step before the actual analysis of the association is carried out. In this preceding design step the target molecule is fixed and the recognition agents are designed so that an ensemble of recognition agents is created. In this ensemble those biomolecules that are well optimized with respect to the target in the sense that their complementarity with the target is large appear with an increased probability. Let τ_{T} denote the structure of the fixed target molecule and τ_{R} the various possible structures of the recognizing molecule. Then this design step leads to a probability distribution $P(\tau_{\mathrm{R}}|\tau_{\mathrm{T}})$ for the structures τ_{R} given the fixed structure τ_{T} of the target. The design has to be performed according to some specific model that mimics features of natural evolution processes. To illustrate this step a bit further in a rather unbiological context let us assume that the design is just done by thermal fluctuations. Then the structure τ_{R} has a high probability to be present in the ensemble if it leads to a favourable interaction energy with the fixed target τ_{T}, a structure which has an unfavourable interaction energy with the target has a low probability. The distribution $P(\tau_{\mathrm{R}}|\tau_{\mathrm{T}})$ is then basically the Boltzmann distribution $P(\tau_{\mathrm{R}}|\tau_{\mathrm{T}}) \sim \exp(-\beta E_{\mathrm{RT}})$ with E_{RT} being the interaction energy between the target and the recognizing molecule. In general the probability distribution $P(\tau_{\mathrm{R}}|\tau_{\mathrm{T}})$ will depend on further parameters which describe the conditions under which this first learn step has been carried out. In this way statistical characteristics of biomolecules that have emerged during evolution can be incorporated into the analysis of molecular recognition.

In conclusion, we presented idealized model systems which allow the study of the principle mechanisms governing molecular recognition processes from a statistical point of view. We considered in detail how an idealized model for the recognition of two rigid biomolecules can be developed starting from the basic findings of experimental investigations of protein-protein complexes. The analysis of the model showed that the local correlations on the surface of the biomolecules seem to be an important feature in molecular recognition processes. In addition we considered a modified model where one of the two molecules participating in the recognition process is allowed to be flexible. It can therefore adjust its conformation with respect to a rigid surface to achieve a high complementarity with the surface pattern during the adsorption process. In this context we studied how the matching of correlation lengths of the residue distribution on the two molecules affect the adsorption process and therefore influence the recognition procedure. However, the number of favourable contacts will be balanced by the loop entropy of the flexible chain. In order to calculate conformational characteristics of the adsorbed chain we aim to compute the averaged size of loops and adsorbed segments (Polotsky et al., 2006).

References

Alberts, B., Bray, D., Lewis, L., Raf, M., Roberts, K., Watson, J., 1994. *Molecular Biology of the Cell*. Garland Publishing, Inc., New York.

Ashcroft, N. W., Mermin, N. D., 2001. *Solid State Physics*. Brooks/Coole Thomson Learning, Singapore.

Behringer, H., Degenhard, A., Schmid, F. 2006. Coarse-grained lattice model for molecular recognition. Phys. Rev. Lett. 97, 128101.

Bernauer, J., Poupon, A., Azé, J., Janin, J. 2005. A docking analysis of the statistical physics of protein-protein recognition. Phys. Biol. 2, S17-S23.

Bogner, T., Degenhard, A., Schmid, F. 2004. Molecular recognition in a lattice model: An enumeration study. Phys. Rev. Lett. 93, 268108.

Brooijmans, N., Kuntz, I. D., 2003. Molecular recognition and docking algorithms. Annu. Rev. Biomol. Struct. 32, 335-373.

Bruinsma, R. F., 2002. Physics of protein-DNA interaction. Physica A 313, 211-237.

Chakrabarti, P., Janin, J., 2002. Dissecting protein-protein recognition sites. Proteins: Struct., Funct., Genet. 47, 334-343.

Chakraborty, A. K., 2001. Disordered heteropolymers: Models for biomimetic polymers and polymers with frustrating quenched disorder. Phys. Rep. 342, 1-61.

Doi, M., Edwards, S. F., 1986. *The Theory of Polymer Dynamics*. Oxford University Press, Oxford.

Halperin, I., Ma, B., Wolfson, H., Nussinov, R., 2002. Principles of docking: An overview of search algorithms and a guide to scoring functions. Proteins 47, 409-443.

Janeway C. A., Travers P., Walport M., Capra J. D., 1999. *Immuno Biology*. Elsevier Science Ltd/Garland Publishing, London.

Janin, J., 1996. Quantifying biological specificity: The statistical mechanics of molecular recognition. Proteins: Struct., Funct., Genet. 25, 438-445.

Janin, J., 2000. Kinetics and thermodynamics of protein-protein interactions. In (Kleanthous, 2000, Ch. 1).

Janin, J., Chothia, C., 1990. The structure of protein-protein recognition sites. J. Biol. Chem. 265, 16027-16030.

Jones, S., Thornton, J. M., 2000. Analysis and classification of protein-protein interactions form a structural perspective. In (Kleanthous, 2000, Ch. 2).

Jones, S., Thornton, J. M., 1996. Principles of protein-protein interactions. Proc. Natl. Acad. Sci. USA, 93, 13-20.

Kleanthous, C., ed., 2000. *Protein-Protein Recognition*. Oxford University Press, Oxford.

Li H., Helling R., Tang C., Wingreen N., 1996. Emergence of Preferred Structures in a Simple Model of Protein Folding. Science 273, 666-669.

Lo Conte, L., Chothia, C., Janin, J., 1999. The atomic structure of protein-protein recognition sites. J. Mol. Biol. 285, 2177-2198.

Parisi, G., Glasses, replicas and all that, in Barrat, J.-L., Feigelman, M., Kur-

chan, J., Dalibard, J. (eds.) 2003. *Slow Relaxations and Nonequilibrium Dynamics in Condensed Matter*. Springer, Berlin.

Pauling L., Delbrück, M., 1940. The nature of the intermolecular forces operative in biological processes. Science 92, 77-79.

Peppas, N. A., Huang, Y., 2002. Polymers and gels as molecular recognition agents. Pharmaceutical research 19, 578-587.

Polotsky, A., Degenhard, A., Schmid, F., 2004a. Influence of sequence correlations on the adsorption of random copolymers onto homogeneous planar surfaces. J. Chem. Phys. 120, 6246-6256.

Polotsky, A., Degenhard, A., Schmid, F., 2004b. Polymer adsorption onto random planar surfaces: Interplay of polymer and surface correlation. J. Chem. Phys. 121, 4853-4864.

Polotsky, A., Degenhard, A., Schmid, F., 2006. Lattice Model of Random Heteropolymer Adsorption: Generating Functions' Approach and the Morita Approximation. To be published

Sarai, A., Kono, H., 2005. Protein-DNA Recognition Patterns and Predictions. Annu. Rev. Biophys. Biomol. Struct. 34, 379-398.

Semler, J. J., Genzer, J., 2003. Monte Carlo simulations of copolymer adsorption at planar chemically patterned surfaces: Effect of surface domain sizes. J. Chem. Phys. 119, 5274-5280.

Semler, J. J., Genzer, J., 2006. Design of random copolymers with statistically controlled monomer sequence distributions via Monte Carlo simulations. J. Chem. Phys. 125, 014902.

Wang, J., Verkhivker, G. M., 2003. Energy landscape theory, funnels, specificity, and optimal criterion of biomolecular binding. Phys. Rev. Lett. 90, 188101.

Wodak, S. J., Janin, J., 2003. Structural basis of macromolecular recognition. Adv. Prot. Chem. 61, 9-73.