



The consequences of translational and rotational entropy lost by small molecules on binding to proteins

Christopher W. Murray* & Marcel L. Verdonk

Astex Technology Ltd, 250 Cambridge Science Park, Milton Road, Cambridge, CB4 0WE, UK

Received 12 June 2002; accepted in final form 11 December 2002

Key words: rigid-body entropy, fragment binding, structure-based design, empirical scoring functions.

Summary

When a small molecule binds to a protein, it loses a significant amount of rigid body translational and rotational entropy. Estimates of the associated energy barrier vary widely in the literature yet accurate estimates are important in the interpretation of results from fragment-based drug discovery techniques. This paper describes an analysis that allows the estimation of the rigid body entropy barrier from the increase in binding affinities that results when two fragments of known affinity and known binding mode are joined together. The paper reviews the relatively rare number of examples where good quality data is available. From the analysis of this data, we estimate that the barrier to binding, due to the loss of rigid-body entropy, is 15–20 kJ/mol, i.e. around 3 orders of magnitude in affinity at 298 K. This large barrier explains why it is comparatively rare to observe multiple fragments binding to non-overlapping adjacent sites in enzymes. The barrier is also consistent with medicinal chemistry experience where small changes in the critical binding regions of ligands are often poorly tolerated by enzymes.

Introduction

Drug discovery approaches based on high throughput chemistry and high throughput screening have not delivered major increases in the efficiency of drug discovery [1]. The focus on screening drug-like compounds rather than lead-like compounds could be one reason for the mixed success of the high throughput paradigm [2, 3]. It has been observed that the lead optimisation process often increases molecular weight. An alternative strategy to lead discovery that has been gaining momentum recently is based on very small molecules (or fragments) [4–10]. The idea is to start with a fragment which will probably possess weak affinity for the target, and only introduce extra molecular weight when there are concomitant increases in affinity. This should avoid generating leads with high molecular weight and relatively poor activity, because these kinds of leads have historically proved difficult to optimise.

A fragment-based drug discovery approach would be expected to work particularly well when two fragments bind to different but adjacent sites in an enzyme, a phenomenon that we shall refer to as multiple site binding. Figure 1 illustrates how big increases in affinity should be anticipated when two fragments that exhibit multiple site binding are joined together in an ideal fashion. Page and Jencks were the first to point out that the expected affinity of the joined molecule should be greater than the sum of the affinities of the two fragments [11–13]. The reason is that a fragment loses a significant amount of rigid body rotational and translational entropy when it forms a complex. In other words, a small molecule has to overcome a significant barrier to binding arising from its loss of rigid body entropy. Page and Jencks assumed that the barrier was the same for molecules A, B and C in Figure 1 because theoretically one expects the rigid body entropy of a free molecule in solution to have only a weak dependence on molecular weight (i.e., a logarithmic dependence [11–13]). It follows that the sum of the measured affinities for fragment A and B will include two unfavourable rigid body entropy bar-

*Corresponding author. Email: c.murray@astex-technology.com, Tel: +44 (1223) 226 228, Fax: +44 (1223) 226 201.

riers whereas the actual measured affinity for molecule C will include only one unfavourable term. Therefore, the affinity of molecule C should be substantially greater than the sum of the two measured affinities for fragments A and B when the two molecules are joined in an ideal fashion.

This paper is concerned with examining the energetics associated with linking fragments. The rigid body entropy loss will be calculated using data on multiple site binding where there is assay information on the fragments and joined molecule. An advantage of this work over Page and Jencks' original analysis [13] is that more relevant experimental data is available and that we can restrict our analysis to situations where there is structural information on the binding mode of the molecules. The latter restriction is important because the analysis assumes that fragments are joined optimally, i.e., no additional favourable interactions or unfavourable strain energies are introduced by the linker, and this approximation is easier to assess when binding modes are available.

The present paper shows that the energy associated with the rigid body entropy loss is significant and we will go on to discuss what consequences this has for fragment-based drug discovery. In particular we will consider the following questions:

- why multiple site binding is a relatively rare event
- why the optimisation of *validated* low potency fragments is tractable
- why some areas of molecules in a lead series exhibit hyper-sensitive SAR
- why hits from high throughput screening and high throughput chemistry are often difficult to optimise.

Experimental examples of single-site and multiple-site binding

This section reviews relevant experimental information on fragment binding to proteins. It draws the distinction between three types of observed fragment binding events:

- a) Single site binding. All the observed fragments overlap a common region in the enzyme.
- b) Proximal multiple site binding. Fragments are seen to bind to two or more spatially distinct regions in the enzyme, but the binding regions are proximal in the sense that the fragments from each site could in principle be joined together without generating non-drug-like molecules.

- c) Distal multiple site binding. Fragments bind to two or more spatially distinct regions in the enzyme, and the binding regions are distal in the sense that when fragments from each site are joined together, very large, non-drug-like molecules are formed.

a) Examples of single site binding


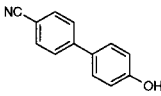
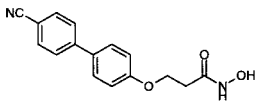
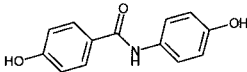
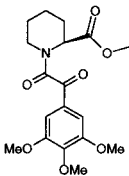
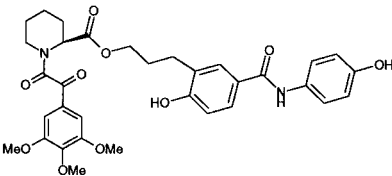
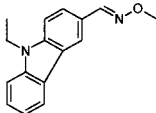
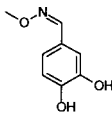
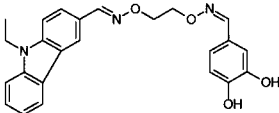

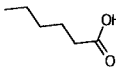
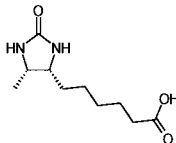
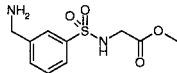
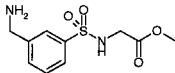
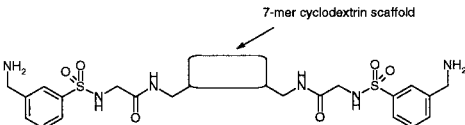
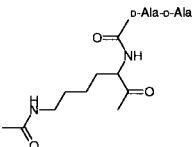
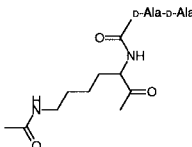
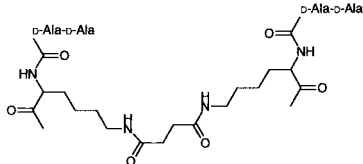
From the literature, one can infer situations where single site binding has been observed in fragment-based drug discovery methods. The examples from the literature include Urokinase [7, 14], DNA gyrase [6], the DNA binding domain of human papiloma virus E2 [15] (there were two binding sites observed but they are too spatially distant to fit the definition of a proximal multiple binding site adopted here and only one of the sites led to inhibition), Lck SH2 domain [16], Erm methyltransferase [17] and Trypsin [10].

b) Examples of proximal multiple site binding

The first four rows of Table 1 show examples of proximal multiple site binding that we are aware of and consider interesting in the context of this paper. As discussed above, we restrict ourselves to examples where there is affinity data on both the fragments and a joined molecule involving the fragments, and where there is some structural information on the fragments and/or the joined molecule. A convincing example of proximal multiple site binding is that for stromelysin (MMP3) [18] (example 1). Acetohydroxamic acid and a simple biphenyl derivative were identified using NMR, and subsequently joined together by a flexible alkoxy linker to give a potent inhibitor. Another SAR by NMR example is given for FK506 binding protein [4] (example 2). Here a pipercolinic acid derivative is attached via a flexible linker to a benzanilide fragment to give a potent binder. A non-structure based approach has been used to design inhibitors of tyrosine kinase c-Src [8] (example 3). Here monomers with greater than 70% activity at 500 μ M were joined 'randomly' with a flexible linker to yield a 64 nm inhibitor. The affinities of biotin and its derivatives for the protein avidin (example 4) have been extensively studied [19]. Structural data is available on a close analogue of 'Molecule C' [23] but not on the fragments.

We have been unable to locate other useful examples of proximal multiple site binding. There are a couple of other kinds of examples worth mentioning though. In a mutated form of protein tyrosine phosphatase 1B, phenylphosphate is observed to bind in

Table 1. Examples of multiple-site binding for stromelysin (1) [18], FK506 binding protein (2) [4], tyrosine kinase (3) [8], avidin (4) [19], tryptase (5) [20] and vancomycin (6) [21, 22].

Example	Fragment A	Fragment B	Molecule C
(1)	 $K_i = 17000 \mu\text{M}$ $m = 75$	 $K_i = 20 \mu\text{M}$ $m = 195$	 $K_i = 0.025 \mu\text{M}$ $m = 282$
(2)	 $K_i = 100 \mu\text{M}$ $m = 229$	 $K_i = 2 \mu\text{M}$ $m = 365$	 $K_i = 0.049 \mu\text{M}$ $m = 620$
(3)	 $K_i = 40 \mu\text{M}$ $m = 252$	 $K_i = 41 \mu\text{M}$ $m = 167$	 $K_i = 0.064 \mu\text{M}$ $m = 417$
(4)	 $K_i = 34 \mu\text{M}$ $m = 100$	 $K_i = 260 \mu\text{M}$ $m = 116$	 $K_i = 0.00000041 \mu\text{M}$ $m = 214$
(5)	 $K_i = 17 \mu\text{M}$ $m = 258$	 $K_i = 17 \mu\text{M}$ $m = 258$	 $K_i = 0.0006 \mu\text{M}$ $m = 1590$
(6)	 $K_i = 4.8 \mu\text{M}$ $m = 372$	 $K_i = 4.8 \mu\text{M}$ $m = 372$	 $K_i = 0.0011 \mu\text{M}$ $m = 742$

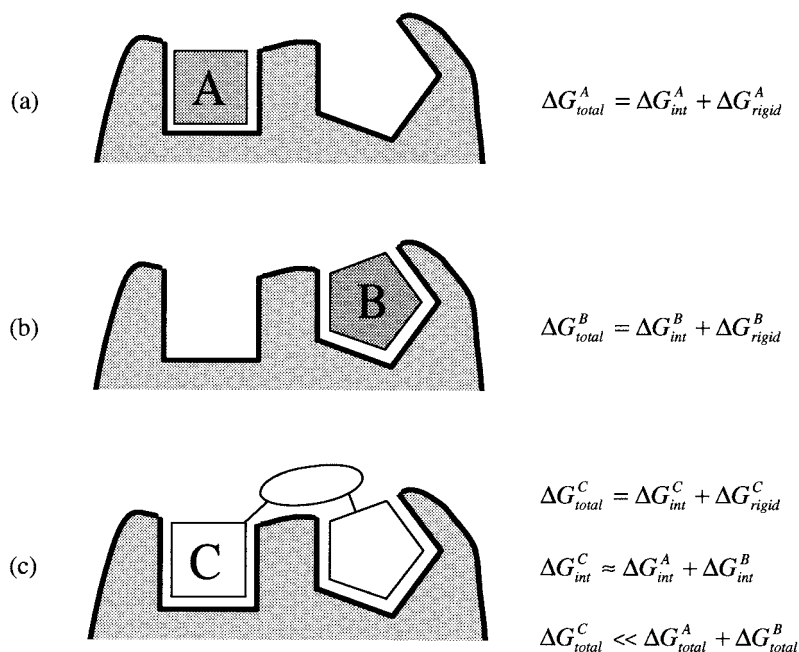


Figure 1. Schematic illustrating the energetics of joining together fragments that exhibit multiple site binding. (a) fragment A binds to the enzyme and the free energy can be decomposed into an intrinsic affinity for the enzyme, ΔG_{int}^A , and an unfavourable term, ΔG_{rigid}^A , representing the rigid body entropy loss; (b) fragment B binds to a different site on the enzyme; and (c) molecule C binds with affinity greater than the sum of affinities for fragments A and B when the fragments are joined together ideally.

two sites that are proximal to each other in the active site [24]. There may be other examples of this kind in the crystallographic literature that we have been unable to locate. Proximal multiple site binding of this type is not helpful to our analysis because it is not usually possible to obtain binding affinities for each site separately. It is also common to see salt molecules bound to enzymes at the same time as ligands. An example is Stroud's work [25] on thymidylate synthase where fragments of dUMP were observed to co-bind with a phosphate ion. Similarly, in purine nucleoside phosphorylase, base analogues are often seen to co-bind with a phosphate or sulphate ion [26]. In the examples of salts binding, the concentration of salt is usually very high and there rarely is good affinity data for the salt binding. Additionally, many of the assumptions used to calculate free energies of binding will be inappropriate for high concentrations of salt. For these reasons, we have not considered the co-binding of a salt ion as multiple site binding.

c) Examples of distal multiple site binding

Although distal multiple site binding is not the main emphasis of this work there are a few examples in the literature where there is data on the binding affinities

of the individual ligands and the combined molecules and where there is some structural information on the binding mode of the ligands. Two of these examples will be used in our analysis.

Trypsin is a trypsin-like serine protease that exists as a tetramer with four active sites. The measured affinities for trypsin of simple molecules such as benzylamine will be the average of the affinities of each of the four sites. One successful design strategy (example 5 in Table 1) has been to join together two benzylamines in such a way that the joined molecule can span two active sites and the benzylamines are near optimally positioned in the S1 pockets of two monomers [20, 27]. Schaschke et al. have joined two benzylamine derivatives using a large, cyclic, semi-rigid scaffold and have demonstrated that 'Molecule C' has the expected binding mode with the enzyme (although the structural data was not good enough to be completely certain of the binding mode of the benzylamine moieties).

Whitesides has worked extensively on the binding properties of multivalent vancomycin derivatives with multivalent Ace-Lys-D-Ala-D-Ala derivatives (example 6 in Table 1) [28]. Here, we consider only the dimeric analogues [21, 22, 29] because this allows

us to use the same, simple analysis approach as for the other examples considered here. In the complex of 'Fragment A' with vancomycin, Rao et al. state that it can be assumed that the lysine sidechain contributes little to affinity and remains freely rotating but that it is frozen in complexes of dimers [29].

Theoretical analysis of fragment binding

a) Estimating the loss of rotational and translational entropy

Consider the binding of a small rigid fragment to an enzyme pocket. In solution, the fragment has a considerable amount of translational and rigid body rotational entropy associated with its free movement through the solution and with its tumbling motion, and much of this rigid body entropy is lost on specific binding to the larger protein. It is difficult to estimate the magnitude of this loss of rigid body entropy because the ligand forms new low frequency vibrational modes on formation of the complex and these modes can retain significant amounts of entropy. This entropy compensation will be variable and may depend on the ligand and protein.

A starting point for the analysis of the size of entropy loss is a consideration of the rigid body entropy associated with a small molecule in the gas phase [30, 11], which can be estimated by the Sackur-Tetrode equation:

$$S_{trans} = R \ln \left[\left(\frac{2\pi mkT}{h^2} \right)^{3/2} (ve^{5/2}) \right] \quad (1)$$

where R is the ideal gas constant ($8.31451 \text{ J} \cdot \text{mol}^{-1} \cdot \text{K}^{-1}$), k is Boltzmann's constant and h is Planck's constant. m is the molecular weight and v is the volume open to a molecule. T is the temperature of the system. The general form of the Sackur-Tetrode equation is a reasonable approximation for the solution phase provided that the volume open to the molecule in solution is changed appropriately from its gas phase value [31–34].

The rigid body rotational entropy can be estimated from the classical statistical mechanical expression:

$$S_{rot} = R \ln \left[\frac{\pi^{1/2}}{\sigma} \left(\frac{8\pi^2 I_A kT}{h^2} \right)^{1/2} \left(\frac{8\pi^2 I_B kT}{h^2} \right)^{1/2} \left(\frac{8\pi^2 I_C kT}{h^2} \right)^{1/2} \right] \quad (2)$$

where I_A , I_B and I_C are the three principle moments of inertia and σ is a symmetry number. Gas phase expressions for rotational entropy are a good approximation to entropies in the solution phase – experimental values of rotational entropies in solution are typically within 2% of values estimated from the gas phase [33].

It is generally recognised that, when a ligand binds to a protein binding site, it loses a large part of its rigid-body entropy. As a first approximation, this loss of entropy is often assumed to be a constant. But equations (1) and (2) clearly show that both the translational and rotational entropy of a molecule in solution are dependent on its molecular weight. In fact, the overall rigid-body entropy (i.e. the sum of translational and rotational entropy) can be approximated with:

$$S_{rigid} \approx A + B \cdot R \ln m \quad (3)$$

Where A and B are not functions of the molecular weight, but do depend on the shape of the molecule. An analysis of the 3D geometries, generated using Corina [35], of 1500 'random' compounds with molecular weights varying between 20 and 800, shows that, under standard conditions, $A \approx 57 \text{ J} \cdot \text{mol}^{-1} \cdot \text{K}^{-1}$ and $B \approx 5$. This means that, for molecular weights varying between 20 and 800, the rigid-body entropy varies between 182 and 335 $\text{J} \cdot \text{mol}^{-1} \cdot \text{K}^{-1}$, which corresponds to a free energy difference of 46 kJ/mol at 298 K. Hence, it would appear that it is a poor approximation to assume that the loss of rigid-body entropy upon binding to a protein is a constant.

Intramolecular reactions are often used as model systems for enzyme-catalysed reactions [36, 37]. Because in intramolecular reactions the reactants are covalently bonded to each other, these reactions often proceed much faster than analogous intermolecular reactions. The same holds for enzyme-catalysed reactions; the enzyme holds the reactants close together in the enzyme-substrate complex, hence accelerating the reaction. The entropy loss in a bimolecular reaction is about 150 $\text{J} \cdot \text{mol}^{-1} \cdot \text{K}^{-1}$ [36], mainly due to the loss of rigid-body entropy; this loss of entropy does not take place for intramolecular reactions. Page and Jencks assume that this entropy loss does not take place for enzyme-catalysed reactions either, and conclude that, therefore, as a result of the binding of the substrates to the enzyme, rate enhancements of about 8 orders of magnitude are to be expected [36]. Such rate enhancements, and even higher, are observed frequently for enzyme-catalysed reactions.

Although the approach used by Page and Jencks gives good agreement with experiment for enzyme-catalysed reactions, we do not believe it is useful for our analysis. In bimolecular reactions, one of the reactants loses all of its rigid-body entropy. This may be a reasonable approximation for many enzyme-substrate complexes, where the enzyme often forms special interactions, e.g. (semi-) covalent bonds, with the substrate, but we believe (non-substrate) ligands retain a considerable amount of rigid-body entropy when bound to a protein binding site. Also, enzyme-substrate complexes often stabilise transition states, leading to additional rate enhancements. Such effects make it difficult to compare rate enhancements in enzyme-catalysed reactions with rigid-body entropy losses of a ligand on binding to a protein.

It is difficult to come up with a satisfactory method for estimating the rigid-body entropy retained by the ligand, when bound (non-covalently) to a protein. Searle and Williams give estimates of between 8.8 and 44.8 kJ/mol (at 298 K) for the total entropy loss after taking into account the compensation due to the introduction of new vibrational modes in the complex [38]. The higher end of the range is associated with complexes that have very large enthalpic contributions (i.e., strong polar interactions) and are therefore held more tightly by the enzyme. The lower end of the range is associated with complexes that have very weak enthalpic interactions (i.e., only hydrophobic interactions) and therefore are more loosely held by the enzyme.

Finkelstein and Janin [31] describe a model that accounts for the rigid-body movements of a ligand when bound to a protein binding site. Their model assumes that the translational entropy of a ligand, bound to a protein binding site is still given by the Sackur-Tetrode equation, but that the volume accessible to the ligand is vastly reduced by the interactions it forms with the protein. The translational entropy loss of the ligand upon binding to the protein can then be estimated as:

$$\Delta S_{trans} \approx 3R \ln \left(\delta x / v^{1/3} \right) \quad (4)$$

Where δx represents the average r.m.s. amplitude (in Å) in principal directions of the ligand, when bound to the protein. Finkelstein and Janin derive a similar expression for the loss of rotational entropy, using equation (2) as a starting point:

$$\Delta S_{rot} \approx 3R \ln \left(\delta \alpha / 2\pi^{2/3} \right) \quad (5)$$

Where $\delta \alpha$ represents the average r.m.s. amplitude (in radians) for rotations about the principal directions of the ligand, when bound to the protein. Interestingly, in this model, the loss of translational and rotational entropy is independent of molecular weight. In another, more detailed, theoretical analysis, Gilson and colleagues [39] come to the same conclusion.

In conclusion, the theoretical estimates are crude and unsatisfactory but there is universal agreement that there is a substantial free energy barrier to binding arising from the loss in rigid body entropy. In this study, we will follow the models presented by Finkelstein [31] and Gilson [39], and assume that the loss of rigid-body entropy is independent of molecular weight. Hence, we will assume that ΔG_{rigid} (see Figure 1) is a constant and estimate its value from the examples of multiple-site binding in Table 1. The assumption of a constant is not ideal because, following Searle and Williams, one expects it to be larger for ligands forming strong polar interactions (small δx and $\delta \alpha$ values in Finkelstein's model) compared with ligands forming mainly lipophilic interactions (larger δx and $\delta \alpha$ values).

b) The effect of losses in rigid body entropy for model systems

Two important conclusions can be drawn from the above analysis:

- 1) As the rigid-body entropy of a compound in solution is proportional to the logarithm of its molecular weight (equation (3)), the increase in entropy for the protein-fragment complex over and above the entropy of the unliganded protein is negligible. For example, the free energy change associated with the change in rigid-body entropy in going from a molecular weight of 30,000 (e.g., a typical unliganded protein) to a molecular weight of 30,300 (protein-ligand complex) is less than 0.1 kJ/mol.
- 2) The losses in entropy for small fragments are similar to those for larger molecules. This means that one expects a large increase in affinity when two weak binding fragments are joined together – the affinity could be larger than that expected by summing the isolated affinities of the two fragments.

These points were illustrated by the schematic in Figure 1. Here we try to analyse the situation in Figure 1 in more detail. The binding affinity for fragment A can be written as:

$$\Delta G_{total}^A = \Delta G_{int}^A + \Delta G_{rigid} \quad (6)$$

where ΔG_{rigid} is the free energy associated with the loss of rigid body entropy on binding to the enzyme; and ΔG_{int}^A contains other free energy terms that contribute to binding and includes favourable enthalpic and entropic interactions (such as hydrogen bonds, lipophilic interactions, and the entropy gained by the expulsion of water molecules from the binding site) and unfavourable terms (such as entropy losses associated with freezing rotatable bonds or enthalpy costs associated with fragment A not being bound in its lowest energy state). We will refer to ΔG_{int}^A as the *intrinsic* binding affinity associated with fragment A.

Similarly the binding affinity for fragment B can be written as:

$$\Delta G_{total}^B = \Delta G_{int}^B + \Delta G_{rigid} \quad (7)$$

Imagine joining these two molecules together to form a molecule C and writing the expression for binding affinity of molecule C as:

$$\begin{aligned} \Delta G_{total}^C = & \Delta G_{int}^A + \Delta G_{int}^B + \Delta G_{rigid} + \Delta G_{rot}^C \\ & + \Delta G_{strain}^C + \Delta G_{binding}^{A-B} \end{aligned} \quad (8)$$

The decomposition in equation (8) begins with the intrinsic binding affinities of fragments A and B, and the loss of rigid body entropy for molecule C. The next term, ΔG_{rot}^C takes account of unfavourable entropic terms as a result of introducing new rotatable bonds when molecule A is connected to B. ΔG_{strain}^C takes account of any free energy loss that might have occurred as a result of either not presenting fragments A and B optimally to their respective pockets or as a result of strain introduced through the linker region not being in its lowest energy conformation. Schematics representing both these possibilities are given in Figure 2. The final term, $\Delta G_{binding}^{A-B}$ accounts for any new direct favourable or unfavourable interactions with the enzyme that the linker might form. It also includes indirect interactions facilitated by the linker. For example the act of joining the two parts together may perturb the solvation energies of A and B or, more importantly, the presence of fragment A in molecule C will increase the intrinsic binding affinity of fragment B by forming part of the pocket where B binds. This latter effect is illustrated schematically in Figure 3.

It is useful to replace the first two terms of equation (8) using equations (6) and (7), i.e.:

$$\begin{aligned} \Delta G_{total}^C = & \Delta G_{total}^A + \Delta G_{total}^B - \Delta G_{rigid} \\ & + \Delta G_{rot}^C + \Delta G_{strain}^C + \Delta G_{binding}^{A-B} \end{aligned} \quad (9)$$

The section on multiple site binding gives a number of experimental examples in which the affinity

of two fragments and the subsequent joined molecule are known, and where there is structural information on the molecular binding. This is the information required to estimate ΔG_{rigid} using equation (9) providing that approximations for the final three terms can be derived.

Much work has been done on estimating the entropic penalty associated with introducing rotatable bonds. Searle and Williams estimate the entropy loss is between 1.6 and 3.6 kJ/mol per rotatable bond frozen from an examination of the entropies of fusion in homologous series [38]. Page and Jencks estimate the entropy loss per rotatable bond frozen to be around 4.5 kJ/mol, based on an analysis of entropy changes in cyclization reactions [11]. Empirical scoring functions often use a single value per rotatable bond to estimate the entropic penalty - the following values have been adopted: 1.4 kJ/mol [40], 1.0 kJ/mol [41], 1.2 kJ/mol [42], 2.8 kJ/mol [43] and 2.9 kJ/mol [44]. Here, an approach based on the values given by Mammen et al. is adopted for the estimation of ΔG_{rot}^C [45]. They present an entropic model based on probabilities derived from torsional energy maps of force fields. This gives entropies that vary according to the rotatable bond that is fixed - for example, freezing an sp^3 - sp^3 rotatable bond between two carbons gives a free energy loss of 2.2 kJ/mol at 298 K. We have assumed that, apart from rotations of CH_3 and NH_3^+ groups, all rotatable bonds in the ligands are frozen upon binding to the protein. An exception was made in the case of the vancomycin example, where we followed Rao and colleagues [21, 22, 29], i.e. we have assumed that the bonds in the lysine side chain are not frozen upon binding of the monomer, but that these bonds are all frozen in the complex of the dimer.

One of the main difficulties in applying equation (9) comes from estimating the strain energy, ΔG_{strain}^C (see Figure 2). Most successful examples of joining fragments use flexible linkers and this maximises the chance that: (i) the fragments A and B will be properly presented when they are joined together in molecule C and (ii) that the linker region can adopt a low energy conformation. The only practical way of using equation (9) is to assume that the strain energy is zero, but to realize that calculated values of ΔG_{rigid} could be significant underestimates if there is strain energy [13]. Our restriction to examples where there is structural information on the ligands is primarily driven by the need to justify the assumption that both the strain energy and the final term of equation (9) are nearly zero.

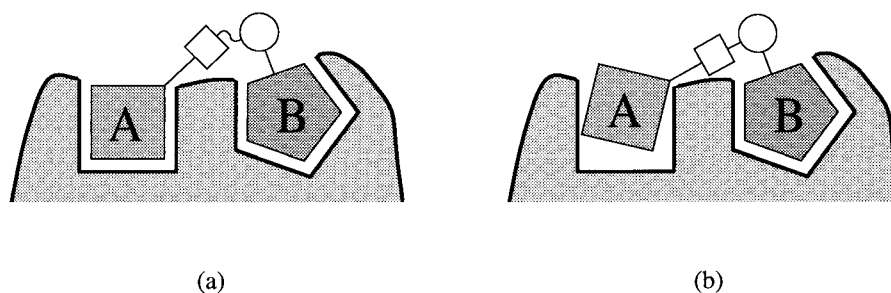


Figure 2. Schematic representing two situations where the strain energy, ΔG^C_{strain} is greater than zero. (a) shows a highly strained linker region but perfectly presented fragments and (b) shows an unstrained linker region but fragment A is poorly presented to its preferred pocket. Both configurations will be relevant to the free energy of binding for this molecule.

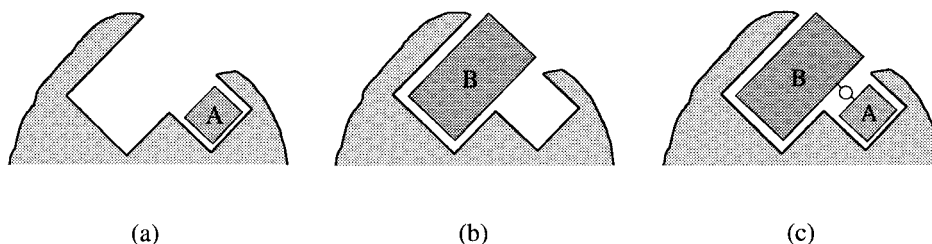


Figure 3. Schematic representing one situation where $\Delta G^{A-B}_{binding}$ would be expected to be favourable. (a) shows fragment A bound to its pocket, (b) shows fragment B bound to its pocket but notice how B's binding mode creates a cavity in the region where A binds and (c) shows how the cavity is removed when B is joined to A. The presence of B increases the affinity of A for its pocket. We believe this effect is important for example 4 in Table 1 (avidin). The effect is less important for example 1 in Table 1 (stromelysin) because the affinity of fragment B is measured in the presence of fragment A.

The final term of equation (9) represents additional interactions between the joined molecule and the enzyme (or between the joined molecule and the solvent). These interactions are additional in the sense that they are not represented accurately by the sum of the intrinsic affinities of fragments A and B, and they may arise from a number of sources, as discussed above. In the analysis that follows it will be assumed that $\Delta G^{A-B}_{binding}$ can be neglected. This assumption will be more exact for some examples than for others. Consider the schematic in Figure 3 for two real examples. In the case of stromelysin (example 1 in Table 1), the activity of the biphenyl molecule was actually determined in the presence of the hydroxamic acid so that the mechanism in Figure 3 should not lead to significant errors in our analysis. In the case of avidin (example 4 in Table 1), it seems likely from inspection of the crystal structure of biotin in avidin [23] that when the hexanoic acid (4b) binds, it would create a less solvent exposed pocket where the ureido fragment binds. The creation of this lipophilic cavity on binding of the hexanoic acid would probably be unfavourable in the absence of the ureido fragment (4a). As a re-

sult, the $\Delta G^{A-B}_{binding}$ term will be relatively large and negative.

c) Estimates of ΔG_{rigid} using data from multiple site binding

Under the assumption that the final two terms in equation (9) are zero, it is possible to re-write equation (9) as:

$$\Delta G_{rigid} = \Delta G^A_{total} + \Delta G^B_{total} - \Delta G^C_{total} + \Delta G^C_{rot} \quad (10)$$

Table 2 gives the values of ΔG_{rigid} that can be estimated from equation (10) using the appropriate examples of multiple site binding given earlier. Example 2 (FKBP) was omitted from the analysis because there was good reason to believe that ΔG^C_{strain} was quite large for this system. Shuker et al. indicate that 'Fragment A' binds in a displaced orientation (1 to 2 Å) when it is contained in the larger 'Molecule C' as compared to its preferred binding mode as an isolated fragment [4]. This suggests that the two fragments have not been joined optimally. The relatively modest increase in potency on joining these two tight binding fragments supports this hypothesis (see Table 1).

Example 3 (a tyrosine kinase) is also excluded, because there is no structural information on the binding mode of this molecule or the original fragments. It is possible that the increase in activity may be due to a general increase in lipophilicity or that one or both of the fragments are not properly presented in the joined molecule. In fact, if examples 2 and 3 are included in the analysis, the calculated values for ΔG_{rigid} come out with non-physical values (i.e., they are negative).

The values obtained for ΔG_{rigid} range from 7 to 29 kJ/mol, corresponding to 1.5 to 5 orders of magnitude in binding affinity. There are two reasons why we obtain this range of ΔG_{rigid} values:

- (1) The assumption that ΔG_{rigid} is a constant is not ideal because the entropy associated with the newly formed vibrational modes in the complex will depend on the types of interactions formed between protein and ligand (see above); in other words, there genuinely is a range of ΔG_{rigid} values, depending on the system.
- (2) The approximations in equation (10) may have led to under or overestimates of ΔG_{rigid} . In the avidin case, for example, the assumption that $\Delta G_{binding}^{A-B}$ is zero is probably not true because the binding affinity of hexanoic acid (4b) is probably much smaller in the absence of 4a, than when it is contained in the molecule 4c. The reasons for this were discussed in detail in the last section, and it could explain why the ΔG_{rigid} obtained for this example is significantly higher than those of the other examples. In the tryptase example, we have indicated that only 4 rotatable bonds are introduced upon linking fragments A and B, however, the cyclodextrin linker is very large and not completely rigid. Additionally, the structural information on the binding of 5c is not of sufficient quality to be sure that the benzylamine moieties are properly presented to their respective pockets in the joined molecule [20]. Both these mechanisms could explain why ΔG_{rigid} appears to be underestimated in the tryptase example.

In conclusion, ΔG_{rigid} varies with the system, but the actual range is probably smaller than that seen in Table 2, because the range in Table 2 is also caused by the approximations in equation (10). On the whole, we think it is best to adopt 15–20 kJ/mol, i.e. about 3 orders of magnitude in the affinity, as the best estimate of the rigid body rotational and translational barrier to binding, but to realise that it may vary, depending on the types of interactions formed in the complex.

Discussion

a) Multiple site binding is a rare event

In this work we have combined a crude treatment of rigid body rotational and translational entropy with experimental data on multiple site binding to derive an estimate for the entropic barrier to binding. It differs from previous approaches because the energetics associated with linking fragments is dealt with in greater detail, and because a wider range of multiple site binding examples from the literature is used. The resulting estimate is a free energy barrier to binding of about 3 orders of magnitude. This is a large barrier but it is not inconsistent with the observation that proximal multiple site binding is a comparatively rare event. There is little evidence in the literature that fragments of measurable activity can be routinely seen binding to different proximal regions of enzymes. This is because only exceptionally tight binding fragments can be expected to overcome the substantial entropic barrier. It is important to realise that this is not just a ‘sensitivity’ issue (although, clearly, the sensitivity of the biophysical technique is important for successful observation of fragment binding). If there is not enough affinity available in a pocket to overcome the entropic barrier to binding then isolated fragments will simply not bind to the enzyme. Fejzo et al. have suggested that the druggability of an enzyme can be related to its ability to bind fragments: enzymes that do not have pockets deep enough to allow fragments to bind are unlikely to be easy drug targets [5].

b) The activity of drug molecules is very unevenly distributed through the molecule

In the development of SAR on lead molecules, it has been observed that often, one portion of the molecule is highly sensitive to small changes in molecular structure [46]. So sensitive, in fact, that sub-optimal changes in this portion of the lead molecule frequently eradicate activity. Smaller versions of the lead molecule are inactive if they do not contain what has been called the ‘molecular anchor’ or ‘minimal recognition motif’ [46]. We suggest that the molecular anchor is a fragment that will bind to the enzyme with low affinity. Here we examine a model system containing a molecular anchor to see if our analysis is consistent with these observations. In this analysis, we have assumed that $\Delta G_{rigid} = 15 - 20$ kJ/mol.

Consider a fragment A that binds to an enzyme with an activity of 100 μ M (–22.8 kJ/mol). According

Table 2. Estimates of ΔG_{rigid} from equation (10) for the multiple site binding examples shown in Table 1. Δn_{rot}^C is the number of extra rotatable bonds in molecule C that are frozen upon binding to the protein, compared to fragments A and B. ΔG_{rot}^C was calculated as outlined in the text, using Mammen et al. [45].

Example	$\Delta G_{\text{total}}^A$ (kJ/mol)	$\Delta G_{\text{total}}^B$ (kJ/mol)	$\Delta G_{\text{total}}^C$ (kJ/mol)	ΔG_{rot}^C (kJ/mol)	Δn_{rot}^C	ΔG_{rigid} (kJ/mol)
Stromelysin (Ex. 1)	−10.1	−26.8	−43.4	+7.6	3	14.1
Avidin (Ex. 4)	−25.5	−20.5	−70.7	+4.2	2	28.9
Tryptase (Ex. 5)	−27.2	−27.2	−52.6	+8.5	4	6.7
Vancomycin (Ex. 6)	−30.3	−30.3	−51.1	+30.9	13	21.4

to equation (6), fragment A is actually contributing −42.8 to −37.8 kJ/mol of favourable interactions (depending on which value we use for ΔG_{rigid}). Now consider a much larger, potent drug molecule, C, that contains fragment A and has a potency of 3 nM (−48.6 kJ/mol). The drug molecule, C, must overcome a similar rigid-body entropic barrier to binding and forms −68.6 to −63.6 kJ/mol of favourable interactions. Note that the majority of favourable interactions are provided by fragment A, despite of the fact that the affinity of molecule C is 33,000 times more than the affinity of fragment A. What happens if fragment A is removed altogether from molecule C? The new molecule will form −20.8 to −30.8 kJ/mol of favourable interactions to offset against the 15–20 kJ/mol barrier to binding and this leads to an affinity of 2 mM to 1 M. This would be considered inactive in any drug program that already had molecules in the nanomolar range. It follows that molecule C would exhibit hypersensitive SAR when changes are made to fragment A.

We believe that this kind of analysis is consistent with much of medicinal chemistry experience on drug design projects. It suggests that activity in drugs is not evenly distributed through the molecule and this has important implications in the best strategies for drug discovery. In support of this, Kuntz et al. have tried to fit a set of ligands with ‘maximal binding affinities’ against the number of atoms or non-hydrogen atoms and have shown that the plots were severely non-linear [47].

c) optimisation of validated low potency fragments is tractable

Fragment based discovery approaches are currently being pursued by a number of workers [4–10]. The analysis presented here supports the potential utility

of such approaches. It shows that the identification of low molecular weight fragments with millimolar affinity is suggestive of a fragment with very high intrinsic affinity for the enzyme (i.e., ΔG_{int}^A from equation (6) is large and negative). Picking up the extra intrinsic affinity required to reach the nanomolar range may not be as challenging as it first appears. Also, early concentration on key SAR portions of the molecule will avoid having to optimise molecules in which the key SAR regions of the enzyme are poorly accessed.

d) hits from high throughput screening are difficult to optimise

Hann et al. have examined a binary model representing ligand binding to an enzyme [3]. Potential interactions on the enzyme are represented as a linear succession of +’s or −’s. Ligands of differing lengths are also represented by a string of +’s and −’s. In the model, the ligands must form an exact complementary match with some linear sequence of the enzyme. The analysis shows the difficulty in finding an exact match as the ligand increases in size because there is a combinatorial explosion in the number of possibilities.

Here we highlight the need for improved models that take into account the uneven distribution of activity in good lead molecules. Such a model must assume a non-even distribution of potential affinities available via binding to adjacent enzyme pockets. Although it may not be possible to derive a model that is complex enough to capture all the key facets of drug binding yet simple enough to be of some utility, it is still interesting to consider what the essential features of such a model might be.

Consider the situation shown in Figure 4(a) where an enzyme contains 4 regions, *a*, *b*, *c* and *d* with each region capable of achieving an intrinsic affinity of 6, 2, 2 and 2 orders of magnitude respectively when

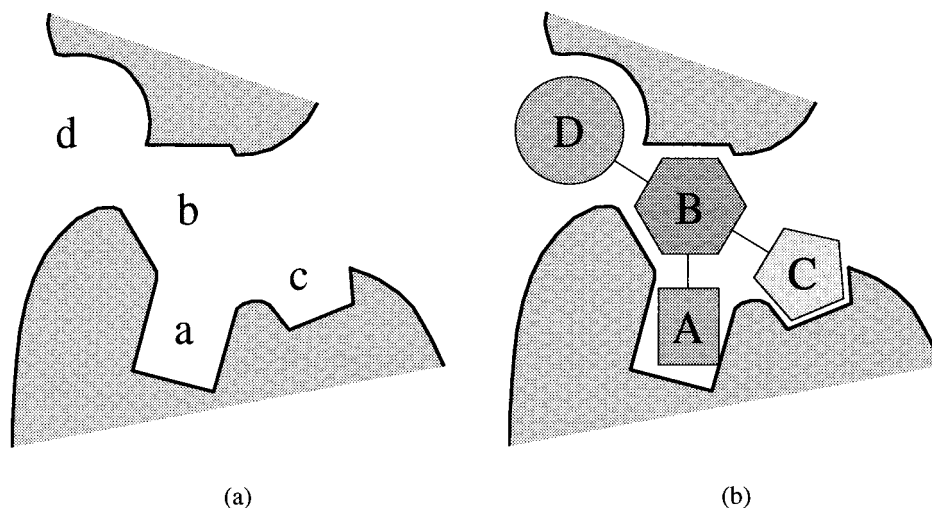


Figure 4. Schematic representing a compound ABCD interacting with an enzyme with four pockets a, b, c and d. (a) shows the pockets a, b, c and d which can yield intrinsic affinities of 6, 2, 2 and 2 orders of magnitude when the perfect fragment is optimally presented, and (b) shows a poor quality screening hit in which the A moiety is poorly presented to pocket A but interactions of the rest of the molecule with the other pockets prevent easy optimisation by varying one component of the molecule at a time.

a perfect moiety is bound (one order of magnitude is equivalent to 5.7 kJ/mol). Assume that drug-sized molecules, ABCD, are used to probe the enzyme with moiety A probing region a; moiety B probing region b; and so on. From the analysis in this paper, the barrier to binding of ABCD can be assumed to be about 3 orders of magnitude.

The first point to note is that following the arguments above, one expects moiety A to be much more sensitive to small structural changes than moieties B, C and D. The probability of finding a correct A will therefore be much lower than the probability of finding a correct B. Not only is a near perfect A moiety required but there will also be a **very** severe dependency on the context in which A is presented to its pocket. The presentation of A will be dependent on the remainder of the molecule, BCD (see Figure 4(b)). Trying to include this dependency in a simple model is challenging yet it lies at the heart of the problem because if drug-sized hits from HTS typically contain a sub-optimal *but well presented* A moiety then one might expect such leads to be easy to optimise. This could be achieved by simply replacing the sub-optimal A fragment with an optimal choice without excessively increasing molecular weight. In contrast, it is known that drug-like leads from HTS are often difficult to optimise [2].

We suggest that moderately active ABCD molecules that have arisen from random screening are most likely to contain both sub-optimal *and* poorly pre-

sented A moieties. We also postulate that it is the poor presentation of the A fragment that makes such hits difficult to optimise (see Figure 4(b)). For these hits, there will be no simple optimisation path to more active compounds using the usual stepwise changes that characterise medicinal chemistry. Not only would A need to be changed to A', but it will also be necessary to change B in order to correctly present A'. As Figure 4 shows, changes in B are also likely to affect the binding characteristics of moieties C and D. Successful optimisation of this hit would therefore require more than one part of the molecule to be changed *at the same time* in order to see a substantial increase in affinity. It would therefore be difficult to obtain a high quality lead from this hit, although of course one would expect to obtain moderate increases in affinity through addition of molecular weight to some portion of the molecule. (For example, in Figure 4(b), adding lipophilicity to moiety D would be expected to increase affinity moderately but such a change might take one outside drug-like space and eventually lead to pharmacokinetic problems.) We believe combinatorial libraries arranged around a rigid scaffold are particularly susceptible to poor presentation of the key binding moiety and this may explain the poor hit rate observed with diversity-based combinatorial libraries compared with historical compound collections.

e) Implications for empirical scoring functions

An empirical scoring function typically consists of a linear combination of terms that reflect the strength of binding of a ligand to an enzyme or receptor given the geometry of interaction [40–43, 48, 49]. For example, the Chemscore function [48] could be written as:

$$\Delta G_{total} = \Delta G_o + \Delta G_{hbond} + \Delta G_{metal} \\ + \Delta G_{lipo} + \Delta G_{flex}$$

These terms are, respectively, a constant term, a weighted count on hydrogen bonds (and their quality), a weighted count on contacts with metals, an estimate of the lipophilic contact energy and a term to punish ligand flexibility. Functions are used that reflect the expected physical form for each of these terms and the terms are weighted with respect to each other using a regression technique. The regression method chooses weights that best reproduce the experimental binding affinities of ligands for which the experimental binding mode is known.

The constant, ΔG_o , might be expected to represent the entropic barrier to binding but in most regressions it is the term with least significance (in a statistical sense) [40, 48]. The reasons are probably connected with the absence of terms to punish bad interactions and the lack of data on these bad interactions in the training sets; the training sets consist of x-ray structures of protein-ligand complexes, and as rule do not contain bad interactions. The size and sign of ΔG_o is important in establishing the magnitude of the other terms in the equation and will affect the performance of the equation when comparing molecules of different sizes in virtual screening applications. Here we have estimated the entropic barrier to binding to be 15–20 kJ/mol and there may be some value in using this value in future empirical functions. It may not do better on typical training sets but it might be expected to be more predictive when attempting to step outside the training set in virtual screening applications. Jain [42] employed a functional form that is logarithmic in molecular weight and, after optimisation of the weights, the resulting term is $5.94 \ln m$. We believe, however, that, following Finkelstein's [31] and Gilson's [39] work, the rigid-body entropy barrier to binding is independent of molecular weight, and hence ΔG_o should be too.

Conclusions

This paper has used the concept of multiple site binding in which two fragments bind to different sites in an enzyme. In some examples, the fragments have been joined together to give potent inhibitors and the activities and binding modes of the fragments and joined molecules are known. The paper has outlined a novel analysis to allow the calculation of the rigid body entropic barrier to binding, using the examples on multiple-site binding. This barrier to binding is estimated to be 15–20 kJ/mol (about 3 orders of magnitude in affinity). This estimate is considerably lower than that reported by Page and Jencks (about 45 kJ/mol or 8 orders of magnitude), but still represents a large barrier to binding. This large barrier was used to rationalise experience with fragment based drug discovery approaches. The use of a more realistic entropic barrier to binding may also be important in developing more predictive empirical scoring functions for use in virtual screening.

Acknowledgements

The authors would like to acknowledge Robin Carr, Harren Jhoti, Mike Hartshorn and Richard Taylor for useful discussions concerning this work. The authors would also like to thank the referees of this paper for providing many useful and constructive suggestions.

References

1. Hird, N., *Drug Disc. Today*, 5 (2000) 307.
2. Teague, S.J., Davis, A.M., Leeson, P.D. and Oprea, T., *Angew. Chem. Int. Ed.*, 38 (1999) 3743.
3. Hann, M.M., Leach, A.R. and Harper, G., *J. Chem. Inf. Comput. Sci.* 41 (2001) 856.
4. Shuker, S.B., Hajduk, P.J., Meadows, R.P. and Fesik, S.W. *Science*, 274 (1996) 1531.
5. Fejzo, J., Lepre, C.A., Peng, J.W., Bemis, G.W., Ajay, Murcko, M.A. and Moore, J.M., *Chem. Biol.*, 6 (1999) 755.
6. Boehm, H.-J., Boehringer, M., Bur, D., Gmuender, H., Huber, W., Klaus, W., Kostrewa, D., Kuehne, H., Luebbbers, T., Meunier-Keller, N. and Mueller, F., *J. Med. Chem.*, 43 (2000) 2664.
7. Nienaber, V.L., Richardson, P.L., Klighofer, V., Bouska, J.J., Giranda, V.L. and Greer, J., *Nature Biotech.*, 18 (2000) 1105.
8. Maly, D.C., Choong, I.C. and Ellman, J.A. *Proc. Natl. Acad. Sci. USA*, 97 (2000) 2419.
9. Erlanson, D.A., Braisted, A.C., Raphael, D.R., Randal, M., Stroud, R.M., Gordon, E.M. and Wells, J.A., *Proc. Natl. Acad. Sci. USA*, 97 (2000) 9367.
10. Blundell, T.L., Jhoti, H. and Abell, C., *Nature Rev.*, 11 (2002) 45.

11. Page, M.I. and Jencks, W.P. *Proc. Natl. Acad. Sci. USA*, 68 (1971) 1678.
12. Page, M.I. *Chem. Soc. Rev* 1973, 295
13. Jencks, W.P. *Proc. Natl. Acad. Sci. USA*, 78 (1981) 4046.
14. Hajduk, P.J., Boyd, S., Nettlesheim, D., Nienaber, V., Severin, J., Smith, R., Davidson, D., Rockway, T. and Fesik, S.W., *J. Med. Chem.*, 43 (2000) 3862.
15. Hajduk, P.J., et al., *J. Med. Chem.*, 40 (2000) 3144.
16. Hajduk, P.J., Zhou, M.-M. and Fesik, S.W., *Bioorg. Med. Chem. Lett.*, 16 (2000) 2403.
17. Hajduk, P.J., Dinges, J., Schkeryantz, J.M., Janowick, D., Kaminski, M., Tufano, M., Augeri, D.J., Petros, A., Nienaber, V., Zhong, P., Hammond, R., Coen, M., Beutel, B., Katz, L. and Fesik, S.W., *J. Med. Chem.*, 42 (1999) 3852.
18. Hajduk, P.J., Sheppard, G., Nettlesheim, D.G., Olejniczak, E.T., Shuker, S.B., Meadows, R.P., Steinman, D.H., Carrera, G.M., Marcotte, P.A., Severin, J., Walter, K., Smith, H., Gubbins, E., Simmer, R., Holzman, T.F., Morgan, D.W., Davidsen, S.K. and Fesik, S.W. *J. Am. Chem. Soc.*, 119 (1997) 5818.
19. Green, N.M., *Adv. Protein Chem.* 29 (1975) 85.
20. Schaschke, N., Matschiner, G., Zettl, F., Marquardt, U., Bergner, A., Bode, W., Sommerhoff, C.P. and Moroder, L. *Chem. Biol.*, 8 (2001) 313.
21. Rao, J. and Whitesides, G.M., *J. Am. Chem. Soc.* 119 (1997) 10286.
22. Rao, J., Lahiri, J., Isaacs, L., Weis, R.M. and Whitesides, G.M., *Science* 280 (1998) 708.
23. Pugliese, L., Coda, A., Malcovati, M. and Bolnesi, M., *J. Mol. Biol.*, 231 (1993) 698.
24. Puius, Y.A., et al., *Proc., Natl., Acad. Sci. USA*, 94 (1997) 13420.
25. Stout, T.J., Sage, C.R. and Stroud R.M., *Structure*, 6 (1998) 839.
26. Mao, C. et al., *Biochemistry*, 37 (1998) 7135.
27. Rice, K.D., Gangloff, A.R., Kuo, E.Y.-L., Dener, J.M., Wang, V.R., Lum, R., Newcomb, W.S., Havel, C., Putnam, D., Cregar, L., Wong, M. and Warne, R.L., *Bioorg. Med. Chem. Lett.*, 10 (2000) 2357.
28. Mammen, M., Choi, S.-K. and Whitesides, G.W. *Angew. Chem. Int. Ed.*, 37 (1998) 2754.
29. Rao, J., Lahiri, J., Weis, R.M., Whitesides, G.M., *J. Am. Chem. Soc.* 122 (2000) 2698.
30. McQuarrie, D.A. *Statistical Mechanics*, D.A., Harper and Row, New York, 1976.
31. Finkelstein, A.V. and Janin, J., *Protein Engineering* 3 (1989), 1-3.
32. Wertz, D.H., *J. Am. Chem. Soc.* 102 (1980) 5316.
33. Mammen, M., Shakhnovich, E.I., Deutch, J.M. and Whitesides G.M., *J. Org. Chem.* 63 (1998) 3821.
34. Murphy, K.P., Xie, D., Thompson, K.S., Amzel, M. and Freire, E., *Proteins: Struct. Func. Gen.* 18 (1994) 63.
35. Sadowski, J. and Gasteiger, J., *Chem. Rev.* 93 (1993) 2567.
36. Page, M.I. *Angew Chemie Int. Ed.* 16 (1977) 449.
37. Kirby, A.J., *Adv. Phys. Org. Chem.* 1980, 17, 225.
38. Searle, M.S. and Williams, D.H., *J. Am. Chem. Soc.* 114 (1992) 10690.
39. Gilson, M.K., Given, J.A., Bush, B.L., McCammon, A., *Biophysical Journal* 72 (1997), 1047-1069.
40. Böhm, H.-J., *J. Comput.-Aided Mol. Design*, 8 (1994) 243.
41. Böhm, H.-J., *J. Comput.-Aided Mol. Design* 12 (1998) 309.
42. Jain, A.J., *J. Comput.-Aided Mol. Design* 10 (1996) 10, 427.
43. Head, R.D., Smythe, M.L., Oprea, T.I., Waller, C.L., Green, S.M. and Marshall, G.R., *J. Am. Chem. Soc.*, 118 (1996) 3959.
44. Andrews, P.R., Craik, D.J. and Martin, J.L., *J. Med. Chem.* 27 (1984) 1648.
45. Mammen, M., Shakhnovich, E.I. and Whitesides G.M., *J. Org. Chem.* 63 (1998) 3168.
46. Rejto, P.A. and Verkhiver, G.M., *Proc. Natl. Acad. Sci.* 93 (1996) 8945.
47. Kuntz, I.D., Chen, K., Sharp, K.A. and Kollman, P.A., *Proc. Natl. Acad. Sci.* 96 (1999) 9997.
48. Eldridge, M.D., Murray, C.W., Auton, T.R., Paolini, G.V. and Mee, R.P., *J. Comput.-Aided Mol. Design*, 11 (1997) 425.
49. Wang, R., Liu, L., Lai, L. and Tang, Y., *J. Mol. Model.* 4 (1998) 379.