OULU 2007
A 493

ACTA UNIVERSITATIS OULUENSIS

Serena Donnini

COMPUTING FREE ENERGIES OF PROTEIN-LIGAND ASSOCIATION

FACULTY OF SCIENCE,
DEPARTMENT OF BIOCHEMISTRY,
BIOCENTER OULU,
UNIVERSITY OF OULU

ACTA UNIVERSITATIS OULUENSIS
A Scientiae Rerum Naturalium 493

SERENA DONNINI

COMPUTING FREE ENERGIES OF PROTEIN-LIGAND ASSOCIATION

Academic dissertation to be presented, with the assent of the Faculty of Science of the University of Oulu, for public defence in Raahensali (Auditorium L10), Linnanmaa, on October 19th, 2007, at 12 noon

OULUN YLIOPISTO, OULU 2007

Acta Univ. Oul. A 493, 2007

Supervised by
Doctor André Juffer

Reviewed by
Doctor Pak-Lee Chau
Professor Antti Poso

ISBN 978-951-42-8573-8 (Paperback)
ISBN 978-951-42-8574-5 (PDF)
http://herkules.oulu.fi/isbn9789514285745/
ISSN 0355-3191 (Printed)
ISSN 1796-220X (Online)
http://herkules.oulu.fi/issn03553191/

Cover design
Raimo Ahonen

OULU UNIVERSITY PRESS
OULU 2007

Donnini, Serena, Computing free energies of protein-ligand association
Faculty of Science, Department of Biochemistry, University of Oulu, P.O.Box 3000, FI-90014 University of Oulu, Finland, Biocenter Oulu, University of Oulu, P.O.Box 5000, FI-90014 University of Oulu, Finland
Acta Univ. Oul. A 493, 2007
Oulu, Finland

Abstract

Spontaneous changes in protein systems, such as the binding of a ligand to an enzyme or receptor, are characterized by a decrease of free energy. Despite the recent developments in computing power and methodology, it remains challenging to accurately estimate free energy changes. Major issues are still concerned with the accuracy of the underlying model to describe the protein system and how well the calculation in fact emulates the behaviour of the system.

This thesis is largely concerned with the quality of current free energy calculation methods as applied to protein-ligand systems. Several methodologies were employed to calculate Gibbs standard free energies of binding for a collection of protein-ligand complexes, for which experimental affinities were available. Calculations were performed using system description with different levels of accuracy and included a continuum approach, which considers the protein and the ligand at the atomic level but includes solvent as a polarizable continuum, and an all-atom approach that relies on molecular dynamics simulations.

In most such applications, the effects of ionic strength are neglected. However, the severity of this approximation, in particular when calculating free energies of charged ligands, is not very clear. The issue of incorporating ionic strength in free energy calculations by means of explicit ions was investigated in greater detail and considerable attention was given to the affinities of charged peptides in the presence of explicit counter-ions. A second common approximation is concerned with the description of ligands that exhibit multiple protonation states. Because most of current methods do not model changes in the acid dissociation constants of titrating groups upon binding, protonation equilibria of such ligands are not taken into account in free energy calculations. The implications of this approximation when predicting affinities were analysed.

Finally, when calculating free energies of binding, a correct description of the interactions between the protein and the ligand is of fundamental importance. However, active sites of enzymes, where strained conformations may hold a functional role, are not always accurately modelled by molecular mechanics force fields. The case of a strained planar proline in the active site of triosephosphate isomerase was investigated using an hybrid quantum mechanics/molecular mechanics method, which implies a higher level of accuracy.

Keywords: binding affinity, continuum methods, double decoupling method, free energy calculations, ionic strength, LIE, molecular dynamics, proline puckers, pK_a, QM/MM, thermodynamic integration, triosephosphate isomerase

Acknowledgements

This work was carried out in the Department of Biochemistry of the University of Oulu. I found this a very friendly and helpful environment where knowledge is freely exchanged on a daily basis. I want to warmly thank everybody for maintaining such an atmosphere. Thanks in particular to Kalervo Hiltunen, the head of the Department for maintaining such a positive working environment and to Anneli Kaattari, Virpi Hannus, Tuula Koret and all the staff for their help and availability with all kinds of issues. I would also like to thank the Biocenter that together with the Department have provided a frame for developing my own scientific character. In particular, I would like to thank Sinikka Eskelinen to whom I initially introduced myself and who promptly put me in contact with my current supervisor. Her dedication, as well as that of all the staff of the Biocenter and of the Department for maintaining such a favourable framework for scientific development is admirable. Thanks to Rik Wierenga and Tuomo Glumoff and all the members of the Structural Biology group with whom I more closely interacted and to Helmut and Anna, neighboring colleagues. Part of this work is the fruit of the collaboration with Rik Wierenga's group. I want to thank everybody for their valuable contributions and friendly attitude.

I am very grateful to André Juffer, my supervisor, who has provided in the first place the opportunity to carry out my Ph.D. studies. Thanks for his expertise and for giving me the opportunity to grow as a scientist. His door would always be open for discussions and by encouraging scientific confrontation and investigation he provided a very stimulating environment. Many thanks to everybody in the Biocomputing research group that I have seen growing quickly since the first day I arrived, for the everyday mutual support. In particular, Niko for always helping me in the lab and over distance with many practical issues, and to Outi, for sharing our mathematical interest. Thanks also to Marc for his beneficial contributions to the whole group and to Cesar for helping me with many chemistry issues.

The work presented in this thesis was partly carried out in Groningen, in the research group of Alan Mark. That time represents an important period of scientific growth for me, as well. I would like to thank Alan for his guidance and supervision and all the group members for providing a stimulating and friendly working environment. In particular, Alessandra Villa for fruitful discussions and supervision of my work. I would

also like to thank Luca Settimo, at the time in Turku, for our fruitful collaboration.

Thanks to Mari Ylianttila for sharing with me essential information concerning the preparation of the thesis and to Sally Ulich, for the careful proofreading of this thesis. The time spent in Oulu during my Ph.D. thesis means also much more to me. I wish to thank all my friends for the time spent together, in the cold and dark winter and interminable days of summer.

Finally, I wish to thank my family for their constant support and encouragement. My husband, in particular, with whom I share common scientific interests, has been a close collaborator and valuable advisor.

This work was financially supported by the Biocenter Oulu and the Finnish Ministry of Education, which I gratefully acknowledge.

List of original articles

This thesis is based on the following articles, which are referred to in the text by the Roman numerals:

I Donnini, S. & Juffer, A. H. (2004) Calculation of affinities of peptides for proteins. J Comp Chem 25: 393–411.

II Donnini, S., Villa, A., Groenhof G., Wierenga R. K., Mark A. E. & Juffer, A. H. Understanding a 1000-fold affinity difference: a computational study. Manuscript.

III Donnini, S., Mark A. E., Juffer, A. H. & Villa, A. (2005) Incorporating the effect of ionic strength in free energy calculations using explicit ions. J Comp Chem 26: 115–122.

IV Donnini, S., Groenhof G., Wierenga R. K. & Juffer, A. H. (2006) The planar conformation of a strained proline ring: a QM/MM study. Proteins: Struct Funct Bioinf 64: 700-710.

Abbreviations

3PP	3-phosphonopropionic acid
ENP	endopeptidase protein
Eq	equation
FEP	free energy perturbation
L. mexicana	*Leishmania mexicana mexicana*
LIE	linear interaction energy
MC	Monte Carlo
MD	molecular dynamics
MHC	major histocompatibility complex
MM	molecular mechanics
NMR	nuclear magnetic resonance
PDB	Protein Data Bank
PGA	2-phosphoglycolic acid
PMF	potential of mean force
QM	quantum mechanics
RMSD	root mean square deviation
SH2	Src-homology 2 domain
SH3	Src-homology 3 domain
T. brucei	*Trypanosoma brucei brucei*
TI	thermodynamic integration
TIM	triosephosphate isomerase

Contents

1 Introduction

Selective binding to other molecules is one of the most fundamental properties of proteins (Creighton 1993, Lodish *et al.* 1999). Most biological processes, are in fact regulated by protein-ligand association reactions. A deeper understanding of protein-ligand interaction would therefore greatly improve our knowledge of a wide range of biological processes. One possible route to quantitatively investigate such interactions in detail is to consider the thermodynamical aspects of the binding process by computational means. For this, one of the most important thermodynamic functions in biochemistry is the free energy. It can be thought of as a measure of the tendency of the system to evolve towards one state or another. If the free energy change associated with a given process is known in advance, one can predict in which direction the system could potentially evolve. In the context of protein-ligand associations, the change in free energy of the system will determine the extent of such interactions.

With advances in computational power, the calculation of free energy by means of theoretical methods has become a powerful and broadly used tool (Simonson *et al.* 2002). An important advantage of relying on theoretical approaches is that it is quite possible to consider certain aspects of the binding process that are difficult or even impossible to assess by experimental means. For instance, one can investigate the importance of individual contributions of groups of atoms to the energetics of binding and the impact of the dynamical properties of the protein on the binding process. In addition, structure-based drug design and redesign of enzymes would greatly benefit from computational methods and in particular from methodologies that can accurately predict the affinity of ligands for proteins (Norledge *et al.* 2001, Veselovsky & Ivanov 2003).

Binding, as any other molecular process, is the result of the interactions occurring between the different species involved. The underlying physical principles of any of these interactions are the same and they are based on the atomic properties of the molecules in the system. In order to calculate free energies for a particular process, the system has to be described in terms of such properties and interactions. However, the number of degrees of freedom that are consequently required to accurately describe the system can become prohibitively large. In such cases, it is then necessary to rely on reasonable approximations to be able to describe the process within the boundaries of

available computational resources. These approximations are usually concerned with the level of accuracy of the system description and the portion of conformational space that can be sampled, *i.e,* how well all the possible states of the system are represented by the model (van Gunsteren *et al.* 1993). Also, in the context of a free energy calculation of a biological system, the correct description of the physiological (macroscopic) conditions - such as the ionic strength - is far from trivial, rendering the accurate prediction of free energy a difficult and challenging objective (Reinhardt *et al.* 2001, Chipot & Pearlman 2002).

Different levels of description have been developed to model different types of process and to allow a compromise between the accuracy of the description and the performance of the calculation (McCammon 1998, Wang *et al.* 2001). These range from statistical mechanical approaches coupled with atomistic simulations (Beveridge & DiCapua 1989, King 1993) to methods based upon empirical descriptors such as the burial of the hydrophobic surface (Ooi *et al.* 1987, Janin 1995, Juffer *et al.* 1995). This makes it very difficult to make conclusive statements concerning the ability of a certain protein to bind specific ligands and it is often unclear which level of description is in fact required.

This thesis is largely concerned with methodologies that are available for the calculation of free energies of biomolecular systems, in particular free energies of binding. An important objective of this work was to consider the applicability of current free energy calculation methods to protein systems. A variety of computational methods based upon different approximations, including continuum electrostatics and molecular dynamics, were employed to predict - among other properties - the free energies of binding of protein-ligand complexes. As the work progressed, a number of issues emerged that lacked a clear interpretation and which were further investigated. These were concerned with the prediction of the affinities of *charged* ligands and ligands with titrating sites. A great deal of effort was also expended on the proper inclusion of ionic strength into the description and the correct interpretation of predicted affinities in relation to the experimental values. An important observation that emerged from the present work was the inability of current molecular mechanics force fields to describe strained conformations in the protein environment, which was found to be a major limiting factor in the proper modelling of active sites in proteins.

2 Review of the literature

2.1 Proteins: fundamental biological macromolecules

Proteins affect almost every property that characterizes a living organism. The expression of the information encoded in the DNA depends almost entirely on proteins, enzymes are proteins and proteins are crucial components of muscles and of collagen, only to mention but a few of their functions. In spite of this variety, proteins are a relatively homogeneous class of molecules. They are basically all composed of one or more polymers of amino acids. There are twenty amino acids in nature, each characterized by particular chemical properties. This is why different combinations of amino acids determine specific three-dimensional structures. It is this diversity in structure that is responsible for the wide range of functions that characterize proteins. The relationship between the functional properties of proteins and their spatial structure is therefore of key importance for the understanding of protein systems.

2.1.1 Determination of protein structures

The most widely used methods for the determination of protein structures are X-ray crystallography (Matthews 1977) and NMR spectroscopy (Wüthrich 1986). The majority of structures available from public databases such as the protein data bank (PDB) (Berman *et al.* 2000), are obtained using these methods. Structures obtained by NMR spectroscopy are usually not as detailed and accurate as those obtained by X-ray crystallography, but NMR has the advantage of using a protein in solution rather than as a crystal. Most protein crystals consist of about 50% solvent by volume, therefore protein molecules are in an aqueous environment. However, crystallographic conditions like high salt concentrations and possible contacts between single protein molecules in the crystal can alter their properties. In solution the protein is free of constraints of the crystal lattice and alternative conformations are usually detected by NMR spectroscopy. Also other dynamical properties can be determined, such as rates of rotation of certain groups of atoms in the structure.

In 2006, the PDB contained about 42,000 structures of which more than 35,000 were solved by X-ray crystallography. In the same year more than 5,000 structures

have been deposited in the PDB. Together with the number, the complexity of the structures has also increased, with several examples of large macromolecules. Interestingly, the number of unique protein folds (about 1,400, as defined by the Structural Classification of Proteins - SCOP - (Murzin *et al.* 1995) or by the Class, Architecture, Topology, and Homologous superfamily - CATH - (Orengo *et al.* 1997) databases) has not correspondingly increased in the last couple of years. This suggests that during evolution some folds have been selected and functional diversity has been obtained as a variation of these (Orengo & Thornton 2005).

2.1.2 Molecular interactions in proteins

Given the large size of polypeptide chains, knowledge of the covalent structure is not sufficient to fully characterize proteins. The three-dimensional structure is the result of many simultaneous interactions that take place among different parts of the molecule (Creighton 1993). These noncovalent interactions not only determine the structure of a protein but also mediate its function with the environment, that is with water, ions, other proteins, membranes, *etc.* These interactions are based on the same set of noncovalent forces, that, for convenience are usually classified as: van der Waals interactions, electrostatics interactions and hydrogen bonds. Because the environment of most proteins is water, its properties play a very important role in protein systems (Privalov & Gill 1988). Water is responsible for the hydrophobic effect. This interaction is often the most important characteristic of forces between molecules dissolved in water.

Van der Waals interactions (or dispersion forces) are weak attractive interactions between any two atoms in close proximity. Due to the movement of electrons around the nucleus, each atom behaves like an oscillating dipole. This leads to coupled, transient dipoles of opposite charge between close atoms, that mutually interact. Such attractive force is offset by the repulsion of the electrons of the atoms involved as they approach each other. Even though weak and close range, these interactions become significant when many atoms of a molecule are close to each other (Israelachvili 1973). In such cases steric complementarity can determine high specificity of interaction. Van der Waals interactions are usually considered to be independent of the orientation of the interacting molecules. This is, however, not always the case. For example, the polarizability of a C-H bond is almost twice as great along the bond as perpendicular to it.

In the context of molecular interactions it is rather common to address electrostatic

interactions in terms of interactions between point charges centred at the nuclei of the atoms. This is however only a reduction of the real picture that should be that of interacting electron clouds. For simplicity, and in order to introduce the next sections, in the following description we will adopt the viewpoint of point charges. The interaction between two charges is described by the Coulomb's law. The Coulomb energy varies inversely with only the first power of the distance between charges, and it is therefore effective also at large distances. Coulomb's law, however, applies to two point charges in a vacuum. In other homogeneous environments, this interaction is screened by the dielectric constant. Less homogeneous environments have to be treated explicitly, in particular, the interface between protein and solvent. Two charges on a protein of low dielectric constant do not interact through the shortest distance of low dielectric but through the high dielectric medium outside the protein. This effectively reduces the interaction between the charges. Furthermore, Coulomb's law is valid only for point charges and therefore applies at distances significantly greater than atomic dimensions.

Electrostatic interactions also involve molecules that do not have a net charge but have different electronegativities. Because of differences in electronegativity, most covalent bonds are in fact polarized and thus there is an effective (partial) charge on most atoms. Such separation of charge in a molecule determines its dipole moment. Dipoles interact with point charges, other dipoles and with other separation of charges such as quadrupole or octapole, in a way that is dependent on the relative orientation of the interacting groups. These interactions however, are not as effective as those between point charges, but diminish faster with increasing distance. Molecules with a permanent dipole moment are called polar molecules. Non-polar molecules can acquire a dipole moment in an electric field, as a consequence of the distortion the field causes in their electronic distributions. Such induced dipole moments are the effect of an external field and are therefore only temporary. Polar molecules' dipoles are also temporarily modified by an applied field. The tendency of the charge distribution of a molecule to be altered by an electric field is called polarizability, and it basically depends on how tight the electrons are held by the nuclei.

Hydrogen bonds are a special case of electrostatic interaction that occur when two electronegative atoms compete for the same hydrogen atom (Pimentel & McLellan 1971, Ippolito *et al.* 1990). Because of the small size but substantial charge of the hydrogen atom, it can interact strongly with an electronegative atom even though it is covalently bond to another. The electrostatic interaction between the dipole of the covalent bond and a partial negative charge is the main component of the hydrogen bond. Is

is rather difficult to establish the strength of a hydrogen bond, usually between 8-40 kJ mol^{-1}, because of the variety of hydrogen bonds, but also uncertainty in the estimates.

Interactions between non-polar molecules and water are not energetically very favourable because non-polar molecules cannot participate in hydrogen bonds with the polar water molecules. This causes interactions among non-polar groups themselves to be much more favourable than in other solvents. The preference of non-polar atoms for nonaqueous environments is known as the hydrophobic interaction. The nature of this hydrophobic effect has been the subject of endless controversy since it was first introduced in 1959 (Kauzman 1959). It seems now clear that it is not just an entropic effect as was postulated for many years, but has both entropic and enthalpic contributions which vary dramatically with temperature (Privalov & Gill 1989, Muller 1990). Conversely, electrostatic interactions between polar molecules in water are not particularly favourable because of comparable competing interactions with water.

Interactions between multipole charges and dipoles of the protein and between these and the solvent and eventually ions dissolved in it make interactions in a protein system very complex. In fact, how to analyze and calculate electrostatic interactions in proteins is still an open issue (Warshel & Russell 1984, Matthew 1985, Kollmann & Merz 1990, Sharp & Honig 1990).

2.1.3 Protein-ligand complexes

Noncovalent molecular interactions are responsible for the three-dimensional structure or fold of the protein and are essential in mediating the function of the protein in its environment. Normally, the interaction of a protein with the environment occurs through formation of a complex between the protein and a ligand, where the ligand can be another protein, a membrane or any other molecule a certain protein is interacting with. Three-dimensional structures of the protein-ligand complexes provide the most detailed information about the interactions of a protein with its ligand. Usually, there is steric complementarity at the interface between the interacting protein and ligand and the interface is as closely packed as the protein interior; polar groups are paired in hydrogen bonds and electrostatic charges are usually neutralized (Janin & Chothia 1990). This underlines how the nature of intermolecular protein-ligand interactions is the same as in proteins themselves.

Generally, the structure of a protein does not change substantially upon binding of a ligand. Small adjustments are important to maximize the number of favourable

interactions in the ligand bound and in the free protein structures, to permit rapid rates of association and dissociation. However, conformational changes upon binding can occur. Often these involve movements of flexible loops at the protein surface, as for instance in loop 6 of the enzyme triosephosphate isomerase (Wierenga *et al.* 1992). There are also examples of more substantial changes in the structure of the protein, that are usually of functional significance. For instance, many ligands bind between domains that move together to include the ligand (Lesk & Chothia 1984). In this way the ligand can associate and dissociate and the interactions between protein and ligand are maximized. In allosteric proteins, domain or subunit movement upon binding produce alterations of other sites on the protein that might bind or release other molecules (Leslie & Wonacott 1984, Perutz 1989).

The first model of a protein-ligand complex makes use of the well known analogy with a key fitting into a lock (Fisher 1894). Even though very explicative, this model does not account for the intrinsic dynamics of proteins. A subsequent induced fit model (Koshland 1958) takes into account conformational changes of the three dimensional structure of the ligand and the receptor as they adapt to each other during the binding process. However, in the context of enzymatic reactions, for which the enzyme-substrate complex formation is the first step, there has been recently growing evidence that protein dynamics could have a much more active role than previously expected. In particular, internal modes of the protein itself can in fact promote a reaction by decreasing the activation barrier (Wand 2001, Agarwal 2005). From this prospective even the induced fit model seem not to completely encompass the complexity of protein interactions.

2.2 Thermodynamics of binding

Most functions of proteins are mediated by the interaction of the protein with a certain ligand. The result of such interaction is the binding of the protein and the ligand and the formation of a protein-ligand complex. In this section a short overview of how to interpret and measure protein-ligand binding is given.

2.2.1 Free energy of binding

The affinity of a protein for a specific ligand determines whether a particular interaction is relevant under a given set of conditions and, therefore, if the protein-ligand complex

will form. The affinity is a measure of the overall free energy of the interactions occurring upon binding and quantifies the tendency of two molecules to bind together (Weber 1975, Atkins 1998). The more negative the free energy the stronger the binding occurs. The free energy of binding or affinity at constant temperature and pressure for the following reversible association reaction between a protein P and a ligand L:

$$P + L \rightleftharpoons PL \tag{1}$$

can be expressed as the difference in the free energy of the products and reactants:

$$\Delta G = \Delta G^{\ominus} + RTln\frac{[PL]}{[P][L]} - RTln\frac{[PL]^{\ominus}}{[P]^{\ominus}[L]^{\ominus}} \tag{2}$$

Here, $[P]$, $[L]$ and $[PL]$ are the concentrations of protein, ligand and protein-ligand complex, respectively. The symbol $^{\ominus}$ expresses standard quantities and $\Delta G^{\ominus}$ is the standard free energy of binding, which corresponds to the free energy of binding when the reactants and products are in their standard state. The standard state of a substance is its pure form at 1 bar pressure and 298.15 K. For a soluble substance 1 M concentration is usually considered. Note that the ratios in Eq (2) correspond to the association constant (K_a) for the binding reaction of P and L under a given set of conditions.

The affinity is a function of the concentration of the protein and the ligand. While most of the affinities are measured in simple dilute solutions, proteins often function in extremely concentrated solutions as in the cytosol. Therefore, rather than concentrations, Eq (2) should be expressed in terms of activities. The activity is equal to the product of the concentration and the activity coefficient, a dimensionless number measuring deviation from nonideality. Note that the presence of other molecules in the environment could effectively favour binding reactions, so that association reactions are more favourable in a concentrated solution that might be expected.

When the amount of the reactants and products of a certain process do not change anymore, equilibrium has been reached. At this point $\Delta G = 0$ and Eq (2) becomes:

$$\Delta G^{\ominus} = -RTln\frac{[PL]}{[P][L]} + RTln\frac{[PL]^{\ominus}}{[P]^{\ominus}[L]^{\ominus}} \tag{3}$$

The last term in Eq (3) effectively becomes zero and we have therefore a relation between the equilibrium association constant of a reaction and its standard free energy of binding.

The free energy is related to two other thermodynamic state functions, the enthalpy (H) and the entropy (S) by the following equation:

$$\Delta G = \Delta H - T\Delta S \tag{4}$$

In Eq (4) the change of free energy of a system that accompanies a certain process, is a function of the change in enthalpy and entropy of such system. This equation is a way in which the second law of thermodynamics can be expressed in terms of the properties of the system only, rather than those of the system and the surroundings. The second law of thermodynamics, that is obtained from Eq (4) dividing each term by T and changing the sign, states that the total entropy change of the universe (the environment plus the system) increases for a spontaneous process. As a consequence the ΔG of a spontaneous process must be negative. A positive and a negative ΔH characterize an endothermic and exothermic reactions, respectively. For an endothermic reaction to spontaneously occur, it must therefore be accompanied by a relatively greater increase in the entropy of the system.

2.2.2 Rates of binding and dissociation

In the previous section we have characterized the reaction in Eq (1) in terms of thermodynamic state functions. A state function refers to the properties of the initial or final states of a reaction. This is why ΔG indicates in which direction the reaction goes, *i.e.* which state has the lower free energy, but it does not tell anything about how fast a certain process occurs.

It is possible to express the association constant for the reaction in Eq (1) in terms of the rates of association, k_a, and dissociation, k_d, of the products and reactions, respectively:

$$K_a = \frac{k_a}{k_d} \tag{5}$$

k_a and k_d vary considerably depending on the size of the molecule and the type of reaction (Fersht 1977). Generally, the diffusion rate identifies the upper limit for the speed of a reaction. Most efficient enzymes are known to work at rates that approach the diffusion rate of the ligands into the active site (Albery & Knowles 1976). But there are exceptions to this. For example, favourable electrostatic interactions can guide charged ligands to the binding site and increase rate constants (Koppenol & Margoliash 1982).

2.2.3 Measurement of affinity

There are two fundamentally different classes of method for the measurement of affinities. One is based on the detection of the rate of formation of products or the rate

of depletion of reactants, whereas the other is based on the measurement of the heat exchange that accompanies a reaction.

Spectrophotometry, spectrofluorimetry, automatic titration, and the use of radioactively labeled substrates (Fersht 1977) are between the most common methods of the first class. These methods monitor the change in the absorbance, in the acidity or in the radioactive properties of the solution while the reaction takes place. Once the change in the concentration of substrates or products as a function of time is known, kinetic equations can be applied and the association constant can be derived. Depending on the type of the reactions different equations must be applied in order to correctly describe it.

Calorimetric analysis is a different method for the measurement of affinities that directly measures the energy exchange during a certain reaction (Hinz 1983). A calorimeter is a thermally insulated container where a reaction system can be performed and the energy exchange between the system and its environment measured. Also in this case, equations must be applied to correctly interpret the measured energy in terms of thermodynamic state functions of the system of interest.

2.2.4 Microscopic basis of macromolecular thermodynamics

Thermodynamical relations, like Eq (4), describe and connect different properties of a system but do not provide any explanation at the molecular level of the observable properties. A link between the bulk (observable) thermodynamic properties of matter and its molecular properties is provided by statistical thermodynamics. As suggested by the statistical approach of this thermodynamical description, observable properties can be described as average properties of the large number of molecules of which a system is composed.

In order to describe microscopically a system of N particles we would need to know the coordinates and velocities of each particle, *i.e* the configuration of the system. At any instant the system would move to a new configuration and because of collisions between particles the properties of each particle would change. In practice, such description is not feasible, because of the extremely high number of possible configurations. However, some of the configurations of the system are more probable than others, and the probability of finding the system in one of these is higher. For very large N values,

like in real thermodynamical systems, the most probable configuration dominates the others and only configurations very close to it will have a probability significantly different from zero. In this conditions, the properties of the most probable configuration represent the average properties of the system. (Hill 1986.)

The average number of particles in the most probable configurations can be calculated using the Boltzmann distribution. With the constraints that the total energy E of the system and the number of particles N are constant, the fraction of particles n_i/N in the configuration i with energy E_i is given by:

$$\frac{n_i}{N} = \frac{e^{-\beta E_i}}{\sum_i e^{-\beta E_i}} \tag{6}$$

where $\beta = \frac{1}{kT}$, with k the Boltzmann constant and T the temperature. Note that n_i decreases exponentially with an increase in energy. The denominator of Eq (6) is the molecular partition function, q:

$$q = \sum_i e^{-\beta E_i} \tag{7}$$

q gives an indication of the average number of states that are thermally accessible to a molecule at the temperature of the system.

From Eq (6), the total energy of a system, E, can be expressed as a function of its molecular properties:

$$E = \sum_i n_i E_i \tag{8}$$

$$E = \frac{N}{q} \sum_i E_i \, e^{-\beta E_i} \tag{9}$$

The molecular partition function q is defined for a system of independent molecules, *i.e.* no interactions occur between them and E is therefore constant. In order to treat systems of interacting particles, the concept of *ensemble* is introduced. An ensemble is a (infinitely large) collection of replications of the system. Even though the energy of the single members of such a collection can fluctuate because of interactions between its particles, the total energy of the ensemble remains constant, and therefore also the average energy E of each member will be constant. Eqs (8) and (9) can be rewritten as:

$$E = \frac{1}{N} \sum_i E_i \tag{10}$$

$$E = \frac{1}{Q} \sum_i E_i \, e^{-\beta E_i} \tag{11}$$

where E_i is now the energy of a single member of a collection of N replications, and Q is the canonical partition function. The relation between Q and q is: $Q = q^N$ for independent and distinguishable particles, and $Q = \frac{q^N}{N!}$ for independent and indistinguishable particles. The Boltzmann distribution provides here the most probable configurations of the ensemble, with n_i the number of members of the ensemble with energy E_i. (Hill 1986.)

The energy E corresponds to the value of the internal energy U of the system relative to its value at $T = 0$: $U = U(0) + E$. In the following, for simplicity, we assume that $U(0) = 0$, and therefore $U = E$. From thermodynamics, at constant volume V, a reversible change in the heat of the system, dq_{rev}, equals the change in internal energy, dU, and dq_{rev} is, in turn, related to the infinitesimal change in the entropy of the system, dS, at temperature T (Atkins 1998):

$$dU = dq_{rev} = TdS \tag{12}$$

Combining Eqs (11) and (12) it is possible to describe the entropy in terms of the microscopic properties of the system:

$$S = \frac{U}{T} + klnQ \tag{13}$$

In fact, once the partition function of a system is known all other properties of the system can be expressed as a function of it (Hill 1986). In this way statistical thermodynamics provides a molecular theory of equilibrium properties of macroscopic systems. In particular, from Eq (13) we can derive the expression for the (Helmholtz) free energy A:

$$A = -kTlnQ \tag{14}$$

Note that Q is the partition function of a system with N, V and T constant. To define analogously the Gibbs free energy, G, we need to consider a system for which N, P (pressure) and T are constant. However, when expressing free energy differences of processes for which the pressure-volume work, pV, is negligible, the two expressions become analogous. This is often the case of reactions in solution. We see in Eq (14) a different interpretation of the free energy: it is proportional to the logarithm of the average number of thermally accessible states.

2.3 Modelling of molecular systems

Molecular modelling is concerned with ways to mimic the behaviour of molecules and molecular systems (Leach 2001). Over the last years, the increase in computing power has extended the range of models that can be considered and systems that can be studied. This is why molecular modelling is almost invariably associated with computer modelling. Depending on the type and size of the system one is interested in, different models can be applied. These range from quantum mechanical to empirical force field models and employ different levels of accuracy in the system description. Hybrid models also exist where different parts of a same system are treated with a different level of description. This is because usually more accurate models are also computationally more expensive and limit the applicability to relatively small systems. Once a reasonable model for a system of interest has been found - usually a set of equations that describe the interactions occurring in a system - there is still need for some algorithms that reproduce the evolution in time or space of such a system based on the model. We refer to such algorithms as simulation methods.

2.3.1 From quantum mechanics to continuum models

The highest level of description of a molecular system is in principle given by the rules of relativistic quantum mechanics. All other levels of description are approximations to this first level. These approximations range from models that still consider fine atomic details to a more course macroscopic description. Even though a less accurate model might not be used to describe some properties of a system or some process, it gains applicability to other systems and phenomena. In the following sections some of the most common models are introduced.

Quantum mechanics models

Quantum mechanics (QM) models describe atoms as a function of electronic and nuclear positions. This is the main difference between QM and more course models where usually electrons are not explicitly taken into account. This makes it possible with QM models to obtain properties of a system that depend on the electronic distribution. In particular chemical reactions in which bonds are broken or formed and processes that involve multiple electronic states, such as photochemical conversion, can be simulated.

All matter has both particle and wave character. The extent of wave character depends on the mass and velocity of a particle. Because of the size and nature of electrons, the wave character must be considered to correctly describe their properties. All static and dynamic properties of an electron can then be derived from its wavefunction. The wavefunctions are the solutions of the Schrödinger equation. However, the Schrödinger equation cannot be solved exactly for atoms with more than one electron. Therefore, some approximations are made to describe molecules at the quantum mechanics level. One such approximation is to decouple the motion of the electrons from the motion of the nuclei. This Born-Oppenheimer approximation (Born & Oppenheimer 1927) allows one to solve the electronic and nuclei wavefunctions independently. The most common approaches to solve the Schrödinger equation in the context of this approximation are the Hartree-Fock, semi-empirical and density functional theory methods (Jensen 1999). These methods assume that electron-electron interaction can be described in an average way and the correlation between the motion of individual electrons is accounted for by expansion of the wavefunction in the Hartree-Fock method, use of empirical parameters in semi-empirical methods and use of exchange-correlation functionals in density functional theory.

Molecular mechanics models

When large numbers of particles are involved, QM models cannot be applied, simply because it would require prohibitive computational resources to describe such systems at the QM level of theory. A different approach can be considered, that is to ignore electronic motions and treat the atoms as a function of nuclear positions only. The electronic energy is then written as a parametric function of the nuclear coordinates, where the parameters are fitted to experimental data or higher level (QM) calculations data. In such methods, generally referred to as molecular mechanics (MM) or empirical force field (FF) methods, the molecules are, in fact, described by a ball and spring model. The balls represent the atoms and can have different sizes, whereas the springs are the bonds and can be more or less stiff. In MM methods, quantum aspects of nuclear motion are also neglected and the dynamics of the atoms is described by classical mechanics, *i.e.* Newton's second law, that is the equivalent of the Schrödinger equation in quantum mechanics. Because of these approximations, MM models cannot be used to model properties that depend on the electronic distribution of a molecule, such as chemical reactions, where the connectivity of the atoms change, or photochemical conversion.

The set of equations that describe the interactions of the particles in a system is usually called the force field. The total force field energy of a MM system, E_{MM}, is generally calculated as a sum of different terms. Each term can be thought of as the energy required to distort a molecule in a specific fashion:

$$E_{MM} = E_{str} + E_{bend} + E_{tors} + E_{vdw} + E_{el} \tag{15}$$

Here, E_{str} is the energy for stretching a bond, E_{bend} is the energy for bending an angle, E_{tors} is the energy for rotation around a bond or torsional energy, and E_{vdw} and E_{el} describe the interactions between non-bonded atoms. Non-bonded atoms are usually atoms on different molecules (*i.e.* not connected by bonds) and atoms that are separated by at least three bonds. In contrast with E_{vdw} and E_{el}, the first three terms in Eq (15) are often referred to as bonded terms. An additional term is sometimes added to account for correlation between the first three (non-bonded) terms.

Every term in Eq (15) can be expressed by a simple function of the positions of the atoms in the system, *i.e* of the coordinates of the nuclei. In most force fields, the bonds and angles functions are represented by simple harmonic functions, while torsions are described by periodic functions. The non-bonded interactions are generally described by a Lennard-Jones potential for the van der Waals interactions and by a Coulomb potential for the electrostatic interactions. One example of such functional forms is as follows:

$$\begin{aligned} E_{MM} = & \sum_{bonds} \frac{k_b}{2}\left(r - r_0\right)^2 + \sum_{angles} \frac{k_a}{2}\left(\theta - \theta_0\right)^2 + \\ & + \sum_{torsions} \frac{k_{tors}}{2}\left(1 + cos(n\phi - \phi_0)\right) + \\ & + \sum_{i=1}^{N} \sum_{j=i+1}^{N} \left(\left[\left(\frac{C_{ij}^A}{r_{ij}} \right)^{12} - \left(\frac{C_{ij}^B}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}} \right) \end{aligned} \tag{16}$$

where k is the force constant, r the bond distance, θ the angle, ϕ the torsion, C the Lennard-Jones parameter, q the atomic (partial) charge, ε_0 the vacuum permittivity and i, j indicate any two atoms. The equilibrium values are denoted by the subscript "0".

The non-bonded interactions, in the last sum of Eq (16), are described as pair-wise interactions. In reality, these interactions are more complex, *i.e.* a charged atom induces polarization of the surrounding atoms by perturbing their electronic clouds. Usually, the average effect of the charge-induced polarization is implicitly accounted for by the parameters of the Lennard-Jones potential and the magnitude of the atomic charges.

One of the basic assumptions of MM methods comes from the observation that molecules tend to be composed of units that are similar in different molecules. In this way parameters developed from data on small molecules can be applied to much larger molecules, such as polymers. For example, C-H bond lengths are roughly the same in all molecules. This transferability of structural features holds also for energetics features. For instance, heat of formation of linear alkanes is directly proportional to the number of CH_2 in the molecule, and heat of formation of longer alkanes can be estimated on the basis of additivity of CH_2 groups energies (Benzon 1976).

It should be noted that the numerical value of E_{MM} has no meaning by itself. The zero point of the energy in each term in Eq (16) has been chosen for convenience. The force field energy is sometimes referred to as the steric energy as in some sense it is the excess energy relative to a hypothetical molecule of non interacting fragments. Therefore, it is not possible to compare energies of structurally different molecules, unless the zero point of the energy scale is the same. However, most of the force fields are concerned with reproducing the geometries and possibly conformational relative energies, for which steric energy is sufficient. The application of force fields to the calculation of some of the properties of a system will be discussed in greater detail below.

Many different force fields have been developed. They differ mainly in the functional form of each term of Eq (15), inclusion or not of cross-terms and type of information that is being used for the derivation of the parameters. Some of the most common force fields for describing protein systems are AMBER (Wang *et al.* 2004), CHARMM (Brooks *et al.* 1983), GROMOS (van Gunsteren *et al.* 1996) and OPLS (Jorgensen & Tirado-Rives 1988, Jorgensen *et al.* 1996). The quality of a force field will be obviously affected by the form of the energy expression and the accuracy of the parameters. The functional forms generally offer a compromise between accuracy and computational efficiency and in general, if the force field is designed to treat large molecules, they will be kept very simple. Depending on how the parameters are derived, a certain force field will be more suitable for describing a particular system than another. However, even if a force field is well parametrized to reproduce a rotational energy profile for a small organic molecule, there is no guarantee that the relative energies of slightly larger molecules will be correctly reproduced. In particular, for large systems small inaccuracies in the functional form and in the parameters can have a significant influence on the energy surface described for that molecule. Therefore, no “best” force field exists, but each has advantages and disadvantages.

QM/MM models

In the QM/MM method, part of a the system is treated at the QM level while the rest is described with a MM force field (Warshel & Levitt 1976). This is why it is called a hybrid model and it is, in general, a very useful approach for systems that contain regions for which the MM description is not accurate enough. In a QM/MM model there are three types of interaction that have to be considered: interactions between the atoms in the MM subsystem, between the atoms in the QM subsystem and interactions between the two subsystems. While the first two are modelled in a straightforward manner by a QM and a MM model, respectively, the third type of interactions is more difficult to describe and many ways can be applied. In one approach, known as mechanical embedding of the QM subsystem, link hydrogen atoms are introduced at the bonds between the QM and MM subsystems. These atoms exist only in the QM calculations. Interactions between QM and MM subsystems are then modelled by force field terms. The ONIOM method (Maseras & Morokuma 1995, Svensson *et al.* 1996) is one alternative approach, in which the energy of the total system is obtained from the energy of the QM subsystem in vacuum plus the energy of the whole system calculated at the MM level, minus the energy of the QM subsystem calculated this time at the MM level.

Continuum models of solvent

Even though many processes occur in solvent, often the main interest is only in the behaviour of the solute. In such cases, for calculation purposes, it would be convenient to neglect the solvent, but its dielectric properties affect the behaviour of the solute to such an extent that the description of the system would not be realistic. For many applications, however, solvent can be approximated by a continuous medium with the dielectric properties of the solvent. This saves expensive calculation because individual solvent molecules are not explicitly taken into account. Many such continuum solvent models (Smith & Pettitt 1994) have been developed and can be used with both quantum mechanics and force field models. Generally, the short and long range effects of solvent are included in a solvation term that is added to the energy function of the rest of the system. This term, the solvation free energy, is the free energy for transferring a molecule from vacuum to solvent and can be written as:

$$\Delta G_{solvation} = \Delta G_{cavity} + \Delta G_{vdw} + \Delta G_{electrostatic} \tag{17}$$

Eq (17) is based on the thermodynamic cycle in Figure 1. Because free energy is a state function, $\Delta G_{solvation}$ can be calculated as the sum of the other terms within the cycle. In the next paragraphs each term in Eq (17) will be briefly discussed.

ΔG_{cavity} is the work to create a cavity in the solvent. This term depends on the entropy penalty for the reorganization of solvent molecules and the loss of solvent-solvent van der Waals interactions. ΔG_{vdw} accounts for the solute-solvent van der Waals interactions and it is usually a favourable term. ΔG_{cavity} and ΔG_{vdw} are generally short range effects and occur within the first solvation shell of the solute. Because the number of solvent molecules in the first solvation shell is approximately proportional to the surface area (A) of the solute, ΔG_{cavity} and ΔG_{vdw} are usually calculated as $\gamma \Delta A$, where γ is the proportionality constant. The sum of ΔG_{cavity} and ΔG_{vdw} is often indicated as ΔG_{np}, because it accounts for the non-polar contribution to the solvation free energy. The proportionality constant is estimated from experimentally determined free energies for the transfer of alkanes from vacuum to water (Chothia 1976). In fact, the non-polar term describes all contributions that are not explicitly described by the electrostatic term.

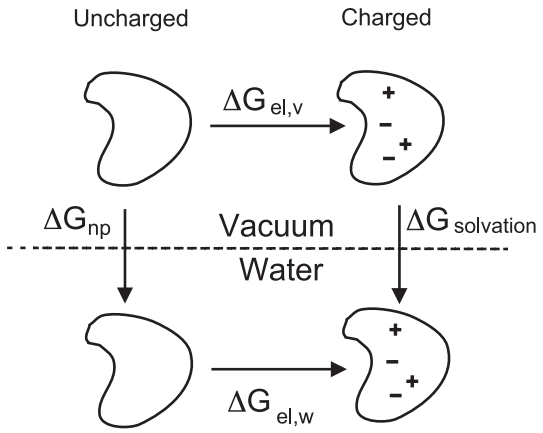

Fig 1. Thermodynamic cycle for the calculation of the solvation free energy ($\Delta G_{solvation}$) using a continuum approach. ΔG_{np} **=** ΔG_{cavity} **+** ΔG_{vdw}**;** $\Delta G_{electrostatic}$ **=** $\Delta G_{el,w}$ **-** $\Delta G_{el,v}$ **where** $\Delta G_{el,w}$ **and** $\Delta G_{el,v}$ **are the work to charge a molecule in water and vacuum, respectively. See text for reference.**

$\Delta G_{electrostatic}$, the last contribution in Eq (17), includes the effect of the solvent on the electrostatic interactions and can be thought as the difference in the work of charging a solute in water and in a vacuum (Figure 1). The electric charge distribution of the solute will polarize the medium, *i.e.* induce a dipole in the surrounding solvent, which in turn induces an electric field within the solute (reaction field) producing an electrostatic stabilization. This is a long range effect that has the consequence of producing an effective screening of charges. The use of a (macroscopic) dielectric constant greater than one for the protein is a way of accounting for this screening. The electrostatic component of the free energy of solvation was first derived by Born (1920) for a spherical charge and extended later by Onsanger (1936) to a dipole in a spherical cavity. This models can be incorporated into quantum mechanics using the self-consistent reaction field (SCRF) method and the polarizable continuum method (PCM) (Miertus *et al.* 1981). For larger systems the boundary element method (Zauhar & Morgan 1985), the generalized Born equation (Constanciel & Contreras 1984) and the finite difference Poisson-Boltzmann method (Klapper *et al.* 1986, Warwicker 1986, Gilson & Honig 1988) have been used together with empirical force fields.

For systems where there is a localized and specific hydrogen bonding between solvent and solute an explicit hydrogen bonding term can be added in Eq (17).

$\Delta G_{solvation}$ can be alternatively calculated as a function of the atomic surface area, a, of individual atoms, i (Eisenberg & McLachlan 1986):

$$\Delta G_{solvation} = \sum_i \Delta\sigma_i a_i \tag{18}$$

The atomic solvation parameters, $\Delta\sigma$, are here specific for every atom and are determined by fitting to experimental solvation data.

2.3.2 Computer simulation methods

Computer simulation methods have undergone rapid development since their introduction in the field of protein science about thirty years ago (McCammon *et al.* 1977). In the last few years, by means of computer simulations the mechanisms underlying many important phenomena have been studied. Examples include water passage via an aquaporin channel (de Groot & Grubmüller 2001), ion transport through potassium channels (Åqvist & Luzhkov 2000), aggregation of lipids into vesicles (Marrink & Mark 2003) and the activation pathway of a photoreceptor protein (Groenhof *et al.* 2004).

Experimental measurements are usually carried out on macroscopic samples that

contain extremely large numbers of molecules. Computer simulations make it possible to generate representative configurations (ensembles) of a macroscopic system from which structural and thermodynamic properties can be derived. Configurations in a simulation are often generated following the time evolution of the system. These methods are called time-dependent methods and rely on the solutions to Newton's equation of motion (classical methods) or to the Schrödinger equation (quantal methods). Alternatively, configurations can be generated by time-independent methods. One such very popular method is Monte Carlo. Quantal methods will not be discussed here, whereas some of the other simulation methods will be briefly introduced in the next sections.

Molecular dynamics

Molecular dynamics (MD) is one major simulation method used to reproduce the evolution in time of a certain system. The central assumption of the method is that nuclei behave as classical particles and Newton's equation can be applied to describe their motion. The force, $\mathbf{F}$, acting on each atom or particle i, is then given by:

$$\mathbf{F}_i = m_i \mathbf{a}_i = m_i \frac{d^2 \mathbf{r}_i}{dt^2} \tag{19}$$

where m is the mass, $\mathbf{a}$ the acceleration, $\mathbf{r}_i$ a vector containing the coordinates of the ith particle, and t the time.

$\mathbf{F}_i$ is also defined as the negative of the derivative of the potential energy V with respect to the position of the ith particle:

$$\mathbf{F}_i = -\frac{\partial V(\mathbf{r}^N)}{\partial \mathbf{r}_i} \tag{20}$$

$V(\mathbf{r}^N)$ is a function of the positions of all N particles in the system and its evaluation is often the most difficult part of a MD calculation. In the previous sections different methods to describe the interactions of the system were introduced. If a MM model is applied, the potential energy, V, in Eq (20) can be replaced by E_{MM} (Eq (16)). Once the potential energy of the system is known, the force acting on each atom can be calculated (Eq (20)) and a new set of coordinates generated (Eq (19)). V must then be re-evaluated and new forces obtained, and so on. By solving iteratively Newton's equation of motion a trajectory, *i.e.* a series of time correlated configurations of the system, is generated. Different algorithms have been developed to numerically integrate the equation of motion (Eq (19)). One of the most commonly used ones is the Verlet

leapfrog algorithm (Verlet 1967). The initial coordinates, in protein simulations, are generally taken from the X-ray structures of the system.

The available energy of a molecule is given by the sum of its kinetic and potential energy. Changes in the potential energy of a system can be though of as movements along an energy surface. Minima on this surface correspond to points were the derivative of the potential energy and therefore the force on all particles is zero. In a MD simulation, molecules are able to explore different energy minima, if the barrier height separating the minima is less than their kinetic energy. Because quite small time steps must be used to integrate Newton's equation of motion, the simulation time is usually in the order of nanoseconds, which is generally shorter than the time scale over which many important phenomena occur, such as large conformational changes. This combined with the use of relatively low temperatures (few hundreds of Kelvin) as simulation temperature, implies that in practice only the local area around the starting configurations will be sampled and only small barriers overcome. It is possible to select a higher initial temperature so that the kinetic energy of the molecule increases and a larger area is sampled. The temperature can then be slowly decreased and the molecule will be trapped in a minimum. In principle, if the cooling is carried out infinitely slowly this minimum will be the global minimum. This approach is called simulated annealing (Kirkpatrick *et al.* 1983, Wilson & Cui 1990).

Monte Carlo

In Monte Carlo (MC) methods, a sequence of configurations is obtained from an initial geometry by changing randomly any degree of freedom of the system, generally the coordinates of a chosen particle. The new geometry is accepted as the starting point for the next perturbation if it is lower in energy than the current one. If the energy is higher, then the Boltzmann factor of the energy difference, $e^{-\Delta E/kT}$ is calculated and compared to a random number between 0 and 1. If the Boltzmann factor is less than this number, then the new geometry is accepted, otherwise the current geometry is used again for the next step. This procedure is called the Metropolis procedure (Metropolis *et al.* 1953) and ensures that the configurations that are generated follow a Boltzmann distribution. The acceptance of configurations with higher energy allows the MC method to escape local energy minima. In order to have a reasonable acceptance ratio for systems with many degrees of freedom, the step size must be fairly small. Therefore, for a macromolecular system many million steps are necessary to explore the local region around a

starting geometry.

MC is a time-independent simulation method, and time-dependent properties of a system, such as transport coefficients, cannot be calculated. In such cases MD has to be used. However, other properties may be calculated using both MC or MD. Usually MC gives more rapid convergence of a simple molecular liquid properties, if the liquid is modelled as a rigid molecule, but it may explore the configurational space (or potential energy surface) of large molecules rather slowly due to the need of small steps. MC is in fact widely used in lattice model simulations, whereas in practice MD is almost always used for systems where there is a significant degree of conformational flexibility. MC may be more effective for conformational changes which jump to a completely different area of the conformational space of a system, while MD advances position and velocities of all particles simultaneously, so it can be very useful for exploration of the local space. Hybrid MD/MC methods have also been developed in which the simulation algorithm alternates between MD and MC (Guarnieri & Still 1994).

Stochastic dynamics

MD methods generate quite detailed information about all the particles in a system. If one is mainly interested in the dynamics of a single particle it is possible to model the surrounding molecules using only their average interactions. This is effectively a coarse-graining process, in which "unimportant" degrees of freedom are removed. The results are stochastic differential equations that describe the motion of one or more particles, such as the Langevin equation (also known as stochastical or Brownian dynamics):

$$m\frac{d^2\mathbf{r}}{dt^2} = -\varsigma\frac{d\mathbf{r}}{dt} + F_{intra} + F_{random} \tag{21}$$

Here, m is the mass of the particle, $\mathbf{r}$ a vector containing the positions of the particle, t the time and ς the friction coefficient. In Eq (21), the force on a particle is considered to arise from three sources. First, a friction term proportional to the atomic velocity of the particle. This can be thought - in the case of a solute in solvent - as the frictional drag on the particle due to the solvent. Second, the normal intramolecular forces (F_{intra}) of the particle, and third, a random component (F_{random}), which averages to zero and can be caused by interactions with solvent molecules. Possibly also external forces, for example an electric field, can be included. For simple particles, the friction coefficient ς can be related to the diffusion constant in the fluid, whereas for large molecules atomic friction coefficients are required, these are usually taken to be proportional to

the accessible surface area of each atom.

One of the main advantages of stochastic dynamics is the reduction of simulation time. This is due to the decrease of degrees of freedom, *i.e.* smaller number of molecules present, and to the fact that longer time steps can often be used. Stochastic dynamics have been widely used to study long-chain molecules and polymers (van Gunsteren *et al.* 1981, Helfand 1984, He & Scheraga 1998a,b), where many interesting phenomena occur over long range time scales, *i.e.* milliseconds and beyond. However, when there are specific interactions between solute and solvent, these can be difficult to include in Eq (21) (Yung-Yu *et al.* 1988).

Calculation of thermodynamic properties

In the previous sections we have seen how molecular simulation methods generate configurations of a certain system. In the following, calculation of thermodynamic properties from such configurations is briefly described.

All thermodynamic functions can be derived from the partition function, Q (introduced in section 2.2.4). In order to numerically calculate Q, one should carry out a summation over all the energy states of a system. When many particles are involved this is practically impossible because these states cannot be simply derived from molecular properties. In section 2.2.4, the concept of ensemble (a collection of replications of the system) was introduced. Some configurations of the ensemble will be more probable than others. In fact, the thermodynamic properties of the system can be evaluated from the average over the most probable configuration of the ensemble. The most probable configuration is given by the Boltzmann distribution (Eq (6)). A simulation generates different configurations M, or replications of a certain system. If the distribution of these configurations of the system follows the Boltzmann distribution, then the average over such finite ensemble can be used to estimate properties of the system. Note that such configurations are also the ones that make the most significant contribution to Q and they can be used to estimate Q. In this context, we can rewrite the expression for the internal energy U and the Helmholtz free energy A in Eq (10) and Eq (14), respectively, in terms of an ensemble average (denoted by $<>$) over M configurations of the system:

$$\langle U \rangle_M = \frac{1}{M} \sum_i^M E_i = \langle E \rangle_M \tag{22}$$

$$\langle A \rangle_M = kT\,ln\left(\frac{1}{M}\sum_i^M e^{E_i/kT}\right) = kT\,ln\langle e^{E/kT}\rangle_M \tag{23}$$

If the configurations are generated by following the time evolution of the system (as molecular dynamics methods), than the average is formally a time average. A time average should produce the same results as an average over an ensemble (the ergodic hypothesis) (Hill 1986). A necessary requirement for producing a representative ensemble is also that the system is at equilibrium.

Thermodynamics properties can be classified as mechanical properties, such as the internal energy, pressure, and heat capacity, and thermal properties, such as entropy, free energy and chemical potential (Hill 1986). Mechanical properties are usually obtained rather easily from a simulation, while thermal properties are difficult to determine accurately. This is a consequence of the fact that mechanical properties involve average over E, while thermal properties involve average over $e^{E/kT}$ as can be seen for U and A in Eqs (22) and (23), respectively. Because states with high energy occur with low probability (Boltzmann distribution), they do not contribute significantly to the average of mechanical properties. Whereas thermal properties depend on the actual value of Q, that is a summation over all states (or configurations) of a system, including the less probable ones, and these will give a significant contribution to the average of thermal properties. Because the less probable states are also most difficult to sample, unless one could simulate for an infinitely long time period, thermal properties are generally very difficult to calculate.

Estimation of errors in computer simulation methods

Because simulations are of finite length, simulation methods always involve a statistical uncertainty. This uncertainty, or error, can be reported in terms of standard deviation of the calculated average value, with respect to the "true" value, and will be inversely proportional to the square root of the number of sampling points. These sampling points, or configurations, should not be correlated. This is usually not the case for nearby points in a simulation. Therefore, in order to estimate a statistical error, the set of points obtained from a simulation is divided into blocks, so that equivalent points of two neighboring blocks are not correlated (Flyvbjerg & Petersen 1989). In order to identify a suitable block size, the variance of the averages calculated over the blocks must be calculated. When the variance for increasingly larger size of blocks reaches a plateau, then the

averages become independent of the block size and the minimum block size of independent sampling points has been determined. Alternatively, in order to estimate the block size, one can directly calculate the autocorrelation function for a certain variable. The autocorrelation should reach a value of zero in an interval of points, or steps, that corresponds to the relaxation interval of such a variable. If configurations are generated by time-dependent methods then the distance between uncorrelated blocks is called the correlation time. The advantage of time-dependent methods is that if the relaxation time of a certain process is known, the required simulation time can be estimated beforehand.

In general, a large ensemble will reduce the error but how well the calculated average resembles the true value depends on how representative the ensemble is. Representative in this context means that configurations in the ensemble must follow the Boltzmann distribution and that those configurations that make a significant contribution to the ensemble average must be sampled. If for example a large number of points is collected from a small part of the phase space (the accessible states of a system in terms of positions and momenta of its particles), then the statistical error may be small but the systematic error will be large, *i.e.* the value will be precise but inaccurate. The underlying problem resides in the fact that it is very difficult to establish if the phase space has been adequately sampled. This issue will be discussed in more details in relation to free energy calculations.

2.4 Free energy calculations

Calculation of thermal thermodynamical properties, such as the free energy is in practice not possible because of the finite length of a simulation and the fact that regions of phase space that make a significant contribution to the free energy are not sampled adequately, as was discussed in the previous sections. However, differences in such quantities can be calculated (King 1993, Straatsma 1996). Two of the most rigorous methods to calculate free energy differences are the free energy perturbation (FEP) and thermodynamic integration (TI) methods. In addition to FEP and TI, the linear interaction energy (LIE) method, the continuum methods and the potential of mean force will be briefly discussed. Thermodynamic cycles will also be introduced because they are often combined with free energy calculations to estimate properties that would otherwise require very long simulations (Kollmann & Merz 1990, Kollman 1993). For clarity, the following discussion refers to the Helmholtz free energy, A, (constant temperature and volume), rather than the Gibbs free energy, G, that is the quantity of choice under

experimental conditions (constant temperature and pressure). However, in practice, in most application the pV work is negligible, and the same equations apply to G as well.

2.4.1 Free energy perturbation

Two systems, X and Y, are considered that are described by two different energy functions E_X and E_Y. The difference in free energy will be:

$$A_X - A_Y = -kTln\frac{Q_X}{Q_Y} \tag{24}$$

Eq (24) can be rewritten in terms of ensemble averages (Zwanzig 1954):

$$A_X - A_Y = kTln\left\langle e^{(E_X - E_Y)/kT} \right\rangle_X \tag{25}$$

where the subscript X indicates that the average is over the ensemble of configurations representative of the initial state X. An analogous expression can be written, where the averaging is over the ensemble corresponding to the final state. The difference between Eqs (23) and (25) is that the exponential now involves an energy difference. As long as this energy difference is comparable with kT, Eq (25) can yield a good estimate of the free energy difference. If X and Y do not overlap in phase space, the phase space of Y will not be adequately sampled when simulating X. In this case the energy difference in Eq (25) will be much larger than kT and the free energy difference will not be accurate. In such cases intermediate states between X and Y can be introduced and described in terms of a coupling parameter λ ($0 \leq \lambda \leq 1$). The simplest approach involves linear interpolation (Frenkel & Smith 1996):

$$E_\lambda = \lambda E_X + (1-\lambda)E_Y \tag{26}$$

and Eq (25) becomes:

$$A_X - A_Y = kTln\sum_\lambda \left\langle e^{(\Delta E_\lambda)/kT} \right\rangle_\lambda \tag{27}$$

Note that the transformation between X and Y by the variable λ may or may not correspond to a physical transformation. The free energy is a thermodynamic state function and its value is independent of the path along which the change is made, as long as it is reversible. It is important that there is enough overlap between successive λ states. In general a change that involves high energy barriers will require much smaller

increments of λ to ensure reversibility of the process with respect to a pathway with a lower barrier. The perturbation is usually performed in both directions, from X to Y and from Y to X to test the quality of the averaging. However, if the simulations are too short, the forward and backward free energy differences may be in good agreement but not accurate (van Gunsteren *et al.* 1993). In such cases the system has no time to equilibrate to the energy function of the next state and one would be averaging always from the same ensemble.

Many free energy calculations involve a change in the molecular topologies. For example, an atom type (and the corresponding bonded parameters) changes. In such case the system can be represented using a single topology or a dual topology. In the former, the topology of the system does not change, but it is at all stages the union of the initial and final states. In the dual topology method, the topologies of the initial and final states are maintained such that both species are present but do not interact with each other. In this case, the energy function describing the interactions of these two species, rather than the energy function of the species itself, is made dependent on the coupling parameter λ.

Generally, a series of simulations with a fixed energy function is performed at every intermediate λ point. Alternatively, in the slow growth method, only one simulation is carried out, where λ is changed at every step. This method requires that the increase in λ is slow enough for the system to remain at equilibrium at all times. This is in practice difficult to achieve and this method is therefore less commonly used (van Gunsteren *et al.* 1993).

2.4.2 Thermodynamic integration

In the thermodynamic integration method the free energy is written as a function of the coupling parameter λ:

$$A(\lambda) = -kTlnQ(\lambda) \tag{28}$$

By differentiating this expression and considering the free energy between two states X and Y we obtain (Kirkwood 1935):

$$A_X - A_Y = \int_0^1 \left\langle \frac{\partial E(\lambda)}{\partial \lambda} \right\rangle_\lambda d\lambda \tag{29}$$

The difference between Eqs (27) and (29) is that in FEP the ensemble average is over finite differences in energy functions, while in TI the average is over a differenti-

ated energy function. This is why, unless the transformation considered involves only very little change, TI is usually the method of choice in free energy calculations of biomolecular systems. In practice, the computational cost for the calculation of the averages with TI and FEP is negligible compared to the cost of generating the ensemble. The same ensemble could in principle be used in both methods to measure reliability of the calculated value.

It should be noted that, formally, the energy E in Eqs (27) and (29) should be replaced by the Hamiltonian H of the system. H can be considered as the total energy of the system, that is the sum of the kinetic and potential energies. However, if only the potential energy of the system changes as a function of λ, the free energy difference will depend in practice only on the potential energy.

2.4.3 Linear interaction energy

The linear interaction energy (LIE) method (Åqvist *et al.* 1994, 2002) postulates that the free energy of binding can be computed from a linear combination of weighted energy estimates of the interactions between the ligand and the rest of the system. In this approach, the binding free energy is calculated from the non-bonded Lennard-Jones (V^{LJ}) and Coulomb (V^{el}) interactions between the ligand and its surroundings in the bound (ligand in the solvated protein binding site) and free (ligand in solvent) state according to:

$$\Delta G = \alpha\Delta\left\langle V^{LJ}\right\rangle + \beta\Delta\left\langle V^{el}\right\rangle \tag{30}$$

The equation contains two empirical parameters, α and β, that are the weight coefficients for the non-polar (Lennard-Jones) and polar (Coulomb) energies, respectively (Åqvist & Hansson 1996, Wang *et al.* 1999).

2.4.4 Potential of mean force

In some cases one might be interested in the change in the free energy as a function of some inter- or intramolecular coordinates, such as the distance between two molecules or the torsion angle of a bond. The free energy change along this coordinate is known as potential of mean force (PMF). Because PMF is calculated for a physically achievable process, the point of highest energy on the free energy profile corresponds to the transition state for the process and it is therefore possible to derive quantities such as rate

constants. The simplest PMF is the free energy change as a function of the separation (r) between two particles. In this case, PMF could be calculated from the radial distribution function, $g(r)$. However, MD or MC methods do not adequately sample low probability regions, where $g(r)$ differs from the most likely value, and it is not possible to calculate PMF using such methods in a straightforward manner. In such cases, a technique called umbrella sampling (Torrie & Valleau 1977) is often used together with PMF.

Umbrella sampling is a method to extend the region of phase space sampled in a single simulation. This method involves the non-Boltzmann sampling of configurations such that the bias introduced can be subsequently corrected for at a later stage. The potential function is modified by a "forcing" potential so that the unfavourable states are now sampled sufficiently. Umbrella sampling is usually performed in a series of stages, each of which is characterized by a particular value of the coordinate and of the forcing potential.

In an alternative approach, free energy perturbation methods, like FEP or TI, can be used with PMF. In this case constraints have to be applied to fix the desired coordinates. A comparison of the two approaches, umbrella sampling and perturbation methods, has shown that proper sampling of the phase space and adequate overlap of phase space are still the major concerns, respectively (Jorgensen & Buckner 1987).

2.4.5 Thermodynamic cycles

Generally, the calculation of the free energy difference of a certain process requires very long simulations. For example, in order to simulate the binding of a ligand to a protein, the ligand and the protein should be brought together from an initial large separation. In the same way, for the calculation of the solvation free energy of a molecule, the latter should be effectively transfered from gas phase (vacuum) to water during the simulation. By using thermodynamic cycles, one can dramatically reduce the simulation time needed. The free energy is a state function and its value is independent of the path used to reach a certain state, as long as the process is at equilibrium. One can therefore calculate the free energy difference (ΔG) between two states using alternative paths which are more convenient in terms of computation time. The solvation free energy, for example, can be calculated using the thermodynamic cycle introduced in Figure 1.

Often one is interested in the difference between the ΔG's ($\Delta\Delta G$), rather than the ΔG itself. For instance, the difference in the free energies of solvation ($\Delta\Delta G_{solvation}$) of

compounds A (ΔG_1) and B (ΔG_2) can be calculated using the thermodynamic cycle in Figure 2 (Jorgensen & Ravimohan 1985):

$$\Delta\Delta G_{solvation} = \Delta G_1 - \Delta G_2 = \Delta G_3 - \Delta G_4 \tag{31}$$

ΔG_3 and ΔG_4 do not correspond to any physical transformation than can be experimentally performed. However, they are theoretically feasible.

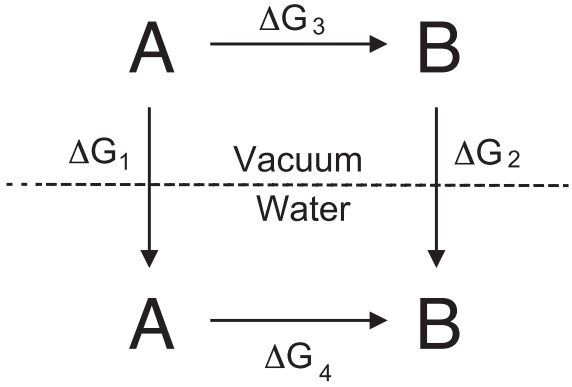

Fig 2. Thermodynamic cycle for the calculation of the relative solvation free energy of two compounds, A and B: ΔG_1 - ΔG_2 = ΔG_3 - ΔG_4.

Analogously to solvation free energies, binding free energies can also be calculated using thermodynamic cycles. In the next sections, two examples are shown (Figure 3and Figure 4) in which a different approach is used to estimate ΔG. An example of the calculation of relative free energies of binding of two different ligands to a protein can be seen in Figure 5. In this case, $\Delta\Delta G$ is given by the difference in the free energy for the mutation of ligand A into ligand B in water (ΔG_3) and in the binding site of the protein (ΔG_4). In particular, this last thermodynamic cycle has been broadly used because a series of different ligands can be tested for their binding properties to a certain protein.

2.4.6 Continuum approach

This approach relies on the decomposition of the free energy for a certain process by means of convenient thermodynamic cycles, and on the use of continuum models for the description of the solvent effects, *i.e.* the solvation free energy. Figure 1 and Eq (17)

illustrate how to compute free energies of solvation of a solute using a continuum model. In the example in Figure 3 a thermodynamic cycle for the calculation of the binding free energy of a ligand L to a protein P in water, using the continuum approach, is shown. Based on such thermodynamic cycle, the free energy of binding can be written as:

$$\Delta G_{b,w} = \Delta G_{b,v} + \Delta\Delta G_{solvation} \tag{32}$$

where $\Delta G_{b,w}$ and $\Delta G_{b,v}$ are the free energies of binding in solution (water) and in the gas phase (vacuum), respectively. Eq (32) essentially means that the affinity in solvent is calculated as the sum of the affinity of the ligand for the protein in an ideal gas (vacuum) and a term that is the *shift* in the affinity due to solvation effects. The latter term ($\Delta\Delta G_{solvation}$) is the difference between the solvation free energies of the protein-ligand complex, PL, and the free protein, P, and ligand, L. Each solvation term is calculated using a continuum solvent model. $\Delta G_{b,v}$ is computed as the difference between the energies of PL and $P+L$ in a vacuum. This energies are usually modelled with a molecular mechanics force field. Entropy differences can be included in $\Delta G_{b,v}$ using empirical estimates such as the side-chain conformational entropy (Pickett & Sternberg 1993). The continuum approach for the calculation of free energies is sometimes referred to as MM-PBSA, because of the use of a MM force field together with a Poisson-Boltzmann (PB) and solvent accessibility (SA) model to estimate the electrostatic and non-polar component, respectively, of the free energy of solvation.

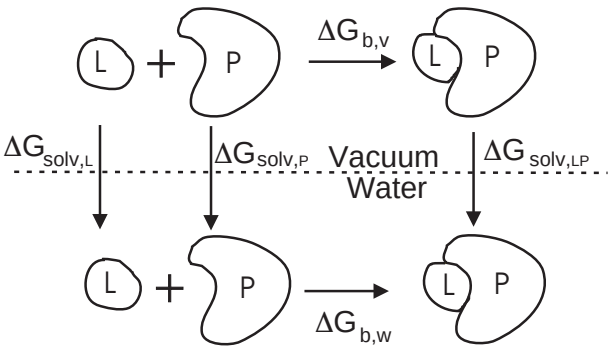

Fig 3. Thermodynamic cycle for the calculation of the binding free energy, using a continuum approach. $\Delta G_{b,w}$ and $\Delta G_{b,v}$ are the binding free energies in solution, *w* (water), and gas phase, *v* (vacuum), respectively; ΔG_{solv} is the solvation free energy; PL, P and L are the complex, the protein and the ligand, respectively.

2.4.7 Estimation of errors in free energy calculations

There are two major sources of error associated with the calculation of free energy from simulation methods. One is related to inaccuracies in the potential energy function (force field) that describes the system (Shirts *et al.* 2003) and the second arises from an insufficient sampling of phase space (van Gunsteren *et al.* 1993). Note that these errors are in general present in any application of simulation methods. Errors arising from insufficient sampling can be furthermore distinguished into two categories. First, sampling errors that depend on poor sampling of the conformational space of a system. In fact, different starting structures will sample different parts of the conformational space. This sampling error is also referred to as "conformational hysteresis". Second, errors that arise from the inability to sufficiently explore the local region of phase space. This is usually known as the statistical error.

The statistical error corresponds to the deviation of the average values of different data series, and it occurs because the ensemble, of finite size, that is sampled is never exactly the same. However, it is assumed that most of the relevant states of the system are included. It can be estimated using the variance of the mean value of a data series, as already discussed in section 2.3.2. The sampling error, on the other hand, represents the fact that the ensemble is not representative of the system, but each data series samples different states available to the system. The statistical error should be, therefore, considered only a lower boundary of the "real" error.

While statistical errors can be easily calculated, sampling errors are more difficult to estimate. Performing a series of simulations with different starting conditions and demonstrating that the statistical error is similar to the sampling error (obtained from the results of different simulations) would be a minimum requirement for demonstrating convergence in the calculations (Pieffet 2005). However, this approach is computationally rather demanding and it is not clear if it would be more efficient than simply increasing the simulation time of the calculation. In the context of FEP and TI, free energy calculation of a certain process can be performed in the forward and backward directions to assess the reversibility of the transformation and identify eventual correlation between simulations at different λ states. A random scheme can also be applied, where the coupling parameter λ is changed randomly between 0 and 1 instead of increasing its value continuously (Pieffet 2005). Different schemes of the dependency of the energy function from λ might be applied as well, and the convergence of the results checked. In fact, depending on the type of transformation and the type of system, one

approach might provide faster convergence than another.

When simulating different λ states, the sampling error might be different at different points. In particular, at intermediate λ points longer simulations might be needed (Pieffet 2005). This would be related to the increase of configurational space accessible to the system at intermediate λ states. The use of soft core potential (Beutler *et al.* 1994) to avoid singularities in the calculations when atoms are created or annihilated seems to be associated with this effect. Use of dummy atoms in a free energy calculation can also be an additional source of sampling errors. Dummy atoms do not interact with the rest of the system, but they do make a contribution to the derivative of the potential energy with respect to λ (for a TI calculation). For the same reason (zero interactions), the conformational space that must be sampled by dummies is generally rather large and might require long simulations.

Many studies have addressed the problem of convergence of free energy calculations in order to better characterize their accuracy and reliability (Straatsma *et al.* 1986, Bishop & Frinks 1987, Straatsma & McCammon 1991, Simonson & Brünger 1992, Pearlman 1994, Nola & Brünger 1998, Villa *et al.* 2003). The difficulty of the system in crossing large energy barriers, often related with the rotation of dihedrals, is generally the main reason for poor sampling. This is why even in cases of rather simple transformations, sampling might not be sufficient. The time scale on which the simulations are performed is simply too short. In such cases different starting conditions will most likely sample different conformations during a simulation. New methods, such as replica exchange approach (Sugita & Okamoto 1999), which has also been proposed in free energy calculations (Sugita *et al.* 2000), could significantly improve the sampling of the system.

Free energy differences can also be estimated using continuum methods, such as MM-PBSA. When only one configuration of the system is used, assignment of a statistical error is not possible. In some cases, different initial configurations of the system have been generated using MD and the free energy calculated for each of them (Wang & Kollman 2000). The error in the free energy of binding is then calculated as variance over the values obtained for every system. However, it has been shown that this approach does not necessarily yield more accurate values than using a single, well equilibrated structure of the complex (Kuhn *et al.* 2005).

2.4.8 Applications to protein systems

One of the earliest applications of the free energy perturbation method was the determination of the free energy required to create a cavity in a solvent (Postma *et al.* 1982). The calculations provided the free energy of cavity formation but also structural and dynamic properties of the water around the cavity. With the increase of computational power from the calculation of relative binding constants of two different ions to a relatively small molecule (Lybrand *et al.* 1986), applications to protein systems have become more and more common. In particular, many examples of relative affinity calculations of inhibitors or ligands to proteins are found, see for instance Bash *et al.* (1987), Rao *et al.* (1992), Miyamoto & Kollman (1993), Ma *et al.* (2002), Talhout *et al.* (2003), Villa *et al.* (2003), Oostenbrink & van Gunsteren (2004). Free energy change associated with amino acid substitutions (Simonson & Brünger 1992, Nola & Brünger 1998, Pieffet 2005) and solvation free energies (Daura *et al.* 1996, Helms & Wade 1997, Villa & Mark 2002) have also been calculated. Because it is possible to significantly reduce the sampling of the system when computing relative free energy differences, examples of these calculations are more abundant. Computation of absolute free energies of binding still remain a major challenge, mainly because of the sampling issue. However, examples of such calculations are also found (Hermans & Wang 1997, Helms & Wade 1998, Chau 2001, Dixit & Chipot 2001, Donnini & Juffer 2004). All the cited applications refer to molecular dynamics free energy calculations, because this is generally the method of choice in simulations of protein systems.

Continuum methods have also been extensively applied to the calculation of free energies of binding. The advantage of these methods is their efficiency because solvent configurations are not required, so that it is possible to test a large set of ligands for their binding abilities (Horton & Lewis 1992, Vajda *et al.* 1994, Froloff *et al.* 1997, Noskov & Lim 2001). To decrease the dependency of the continuum methods on the particular structure used in the calculation, these methods can be combined with MD (Wang & Kollman 2000, Swanson *et al.* 2004).

Free energy calculations provide very useful and detailed information, however they still suffer from many limitations (Reinhardt *et al.* 2001, Chipot & Pearlman 2002). Force field accuracy and sampling errors are some of the major sources of errors, as in most simulation methods. In particular, force field parameters for charged molecules are still much less accurate than for noncharged ones. Therefore, for charged compounds results can be very sensitive to the charge model used (Merz & Kollman 1989, Villa

& Mark 2002). Concerning the sampling issue, investigation of systems consisting of many atoms or more subunits is often limited. This is a consequence of the increasing number of degrees of freedom of the system associated with the number of atoms, or the relative orientation of the subunits of a multimer respect to each other. Systems that undergo relatively large conformational changes upon binding are also difficult to study using free energy calculations, because of the time scale required in such calculations. Solvent accessible binding sites are usually more challenging to study, as presence of solvent in the active site introduces an additional sampling issue. Another source of errors comes from the neglect of ionic strength and possible changes in pK_a of titrating groups upon binding. These are not explicitly modelled by the current methods. Great efforts are being currently applied both to the improvement of models for the description of the system, and to the development of methods to increase conformational sampling in a simulation.

3 Aims of the present study

The main aim of the present study was to understand the applicability of current methods for free energy calculations to protein-ligand association reactions, and subsequently, the relevance of such methods in the functional analysis of protein-ligand complexes. Specifically under investigation was:

1. the capability of continuum and all-atom methods to predict free energies of binding of protein-ligand complexes. Particular attention was given to the calculation of absolute affinities of charged ligands and affinities of ligands with titrating sites.
2. the effect of inclusion of ionic strength in all-atom calculations.
3. limitations of molecular mechanics force fields and level of accuracy required to correctly describe active site conformations.

To achieve the aims, the following were calculated:

1. absolute free energies of binding of 36 complexes (set 1), using continuum methods.
2. absolute free energies of binding of 5 protein-peptide complexes of set 1, using MD and TI and linear interaction energy.
3. relative free energy of binding of two ligands for one particular protein of set 1, with MD, TI, and *ab initio* methods to compute pK_a of the ligands.
4. free energies differences of two ligands at 5 different ionic strengths.
5. energy profiles for the transition between the up and down puckered conformations of an active site proline in vacuum and in the protein environment, using QM/MM methods.

4 Methods

4.1 Calculation of absolute free energies of binding

A continuum approach and an all-atom approach have been used to estimate absolute free energies of binding of 36 protein-ligand complexes and of 5 protein-peptide complexes, respectively. The two approaches are introduced in the next sections.

4.1.1 Continuum approach

A continuum approach was applied for the calculation of the binding free energy of protein-protein, protein-peptide and protein-small ligand complexes. The 36 complexes are listed in Table 1 with the corresponding PDB codes (Berman *et al.* 2000) and experimentally measured affinities.

The thermodynamic cycle in Figure 3 and Eq (32) form the basis of the continuum approach (Janin 1995, Froloff *et al.* 1997, Leach 2001). For easy reference, Eq (32) is here reported:

$$\Delta G_{b,w} = \Delta G_{b,v} + \Delta\Delta G_{solvation} \tag{33}$$

where $\Delta G_{b,w}$ and $\Delta G_{b,v}$ are the free energies of binding in water and vacuum, respectively, and $\Delta\Delta G_{solvation}$ is the difference in the solvation free energy of the protein-ligand complex (PL) and of the free protein (P) and ligand (L). $\Delta G_{solvation}$ is calculated according to the thermodynamic cycle in Figure 1 and Eq (17). All terms in Eq (33) are standard quantities.

The non-polar contribution to $\Delta G_{solvation}$, ΔG_{np}, was calculated as $\gamma\Delta A$, where γ has units of surface tension (energy per unit square area) and ΔA is the change in the solvent accessible area upon solvation.

The electrostatic part of the solvation free energy, ΔG_{el}, is given by the difference between the free energies of creating the charge distribution of the molecule in water, $\Delta G_{el,w}$, and in vacuum, $\Delta G_{el,v}$. In the framework of continuum electrostatics, $\Delta G_{el,w}$ contains a reaction field (W^{RF}) and a Coulombic (W^{C}) component, while $\Delta G_{el,v}$ contains only the W^{C} component. The Coulombic component was calculated applying Coulomb's law, for which it is assumed that the molecular charge distribution is given as a set of (effective) point charges q_i at positions $\mathbf{r}_i$. The final expression for the calcu-

lation of the electrostatic part was (Juffer *et al.* 1997):

$$\Delta G_{el} = W^{RF} + \frac{1-\varepsilon_m}{\varepsilon_m} W^C = \frac{1}{2}\sum_i q_i \phi^{RF}(\mathbf{r}_i) + \frac{1-\varepsilon_m}{\varepsilon_m} W^C \tag{34}$$

where the quantity ε_m is the dielectric constant inside the cavity and reflects the polarization of the solute due to the presence of a polarizable solvent. Note that if the solute is placed in a vacuum, there is no reaction field, and consequently ε_m should be set to 1. $\phi^{RF}(\mathbf{r}_i)$ is the reaction field potential at $\mathbf{r}_i$, which is computed by numerical means.

Under conditions of equilibrium, the association free energy in vacuum is given by (Hill 1986):

$$\Delta G_{b,v} = \mu_{PL} - \mu_P - \mu_L = -NkT \ln\left(V \frac{q_{PL}}{q_P q_L}\right) \tag{35}$$

Here, μ is the (standard) chemical potential, q is the partition function and V is the volume. q contains several independent components, namely a translational (t), rotational (r), vibrational (v) and electronic (e) component. The rotational, vibrational and electronic component are independent of volume, but the translational partition function depends on the size of the accessible space. The electronic component is ignored (that is, it is assumed to be 1). The exact form of the translational, rotational and vibrational partition functions can be found in standard text books (McQuarrie 1976, Hill 1986). Here, it was assumed that the most significant contribution to q(v) is associated with *internal rotational* degrees of freedom, that can be treated by classical mechanics (Hill 1987).

The vibrational partition function can be used to estimate a contribution to the free energy due to conformational flexibility. However, if only one conformation of a certain molecule is used, the conformational space is most probably not sufficiently sampled. In such cases, an alternative approach can be considered where the difference in side-chain conformational entropy ($-T\Delta S_{sc}$) between two states A and B, is given in an entirely empirical manner by:

$$-T\Delta S_{sc} = kT \ln\left(\frac{N_{sc,A}}{N_{sc,B}}\right) \tag{36}$$

Here, N_{sc} is the total number of accessible side chain conformations and it is estimated from the side chain accessible surface area (Pickett & Sternberg 1993). Note that each side chain conformation in Eq (36) is assumed to be equally probable.

All calculations were performed with ICM (Internal Coordinates Mechanics), a molecular modelling program from Molsoft (Abagyan *et al.* 1994). The ECEPP/3 force

field (Nemethy *et al.* 1992) was used, and, for phosphotyrosine and non-peptidic ligands MMFF (Merck molecular force field) (Halgren 1995). Structures were energy minimized using the "regularization" tool of the ICM program, a procedure for fitting a protein model with the ideal covalent geometry of residues (as represented in the ICM residue library) to the atom positions of a target structure. The root mean square deviation (RMSD) of the heavy atoms between the original PDB coordinates and the minimized coordinates was approximately 0.09 nm. In the structure of the SH2-peptide complex 1SHD residues 182 to 185 were missing. These residues were modelled into the structure of 1SHD with ICM. For this purpose another SH2-peptide complex, 1SPS (Table 1), served as a building template.

The reaction field component of the electrostatic work W^{RF} (Eq (34)) was calculated using the boundary element method implemented in ICM (Totrov & Abagyan 2001). The dielectric boundary corresponded to the molecular surface and was defined by the radii of the atoms. No ionic strength was included in the calculations. The calculations were performed with various values of ε_m and γ. The potential energy of the system consisted of a Coulomb and a van der Waals energy term, where interactions between atom pairs separated by less than three bonds were excluded, and a hydrogen bonding and an energy penalty term, the latter associated with the deformation of both dihedral angles and disulphide bonds.

Normal mode frequencies ν_i were calculated with the Gromacs suite of programs (Berendsen *et al.* 1995, Lindahl *et al.* 2001, van der Spoel *et al.* 2005) for the set of SH3 (Src-homology 3) - peptide complexes. ν_i were used to estimate the vibrational entropy difference of binding of the six SH3 complexes. The structures of the free protein, free ligand and of the complex were minimized using conjugate gradient in double precision. If there were no PDB structures of the free protein and free ligand available, the ligand was removed from the complex and both the protein and the ligand were separately minimized. The observed first six normal mode frequencies (these correspond to the overall translational and rotational degrees of freedom) for 1FYN and 1SEM were relatively high ($1-10^{-1}$ cm^{-1} in comparison to $10^{-1}-10^{-2}$ cm^{-1} for the other proteins), possibly indicating that the structures should be further minimized and the frequencies may not be sufficiently accurate.

The loss in side-chain conformational entropy (Eq (36)), was estimated on the basis of the side-chain accessible surface area that becomes buried upon binding, which was computed with ICM.

Table 1. Protein-ligand complexes employed in the continuum calculations. The total charge (e), when different from zero, is specified in brackets, as assigned by ICM. SH3 and SH2 complexes are in bold and italics, respectively. Experimental affinity, ΔG_{exp}, is in kJ mol^{-1}. pY is phosphotyrosine. Peptides' sequences are reported in one letter code.

PDB	Protein	Small molecule Ligand	ΔG_{exp}
1TRD	TIM (+10)	PGH (-2)	-29.3
1AMK	TIM (+2)	PGA (-2)	-24.7
1TSI	TIM (+10)	4PB (-2)	-20.2
6TIM	TIM (+10)	G3P (-2)	-18.5
1AG1	TIM (+10)	HPO4 (-2)	-13.5
5TIM	TIM (+10)	SO4 (-2)	-13.5
4TIM	TIM (+10)	2PG (-3)	-12.4
181L	Lysozyme(mutant) (+8)	Benzene	-21.7
		Peptide Ligand	
1BBZ	ABL-SH3 (-1)	APSYSPPPPP	-33.4
1CKA	C-CRK SH3 (-6)	PPPALPPKKR(+3)	-32.7
1CKB	C-CRK SH3 (-6)	PPPVPPRRRR(+4)	-30.1
1ABO	ABL-SH3 (-1)	APTMPPPLPP	-25.5
1FYN	FYN SH3 (-7)	PPAYPPPPVP	-25.5
1SEM	SEM-5 SH3 (-2)	ACE-PPPVPPRRR (+3)	-24.9
1SPS	SHC-SRC SH2 (+4)	PQpYEEIP (-4)	-47.7
1SHD	C-SRC SH2 (+4)	ACE-pYEEI(E)(-5)	-39.7
1LKK	P56-LCK SH2 (+2)	ACE-pYEEI (-5)	-38.9
1LKL	P56-LCK SH2 (+2)	ACE-pYEEG (-5)	-33.1
1LCJ	P56-LCK SH2 (+2)	EPQpYEEIPIYL (-5)	-32.6
1LCK	P56-LCK (SH3)-SH2(-4)	(T)EGQpYQPQPA(-3)	-29.3
2VAB	H-2KB MHC (-8)	FAPGNYPAL	-49.0
1HHH	HLA-A MHC (-9)	FLPSDFFPSV (-1)	-48.5
2VAA	H-2KB MHC (-8)	RGYVYQGL (+1)	-48.1
1VAC	H-2KB MHC (-8)	SIINFEKL	-47.7
1HHI	HLA-A MHC (-9)	GILGFVFTL	-46.9
1HHK	HLA-A MHC (-9)	LLFGYPVYV	-45.6
1HHJ	HLA-A MHC (-9)	ILKEPVHGV	-37.7
1HHG	HLA-A MHC (-9)	TLTSCNTSV	-37.2
		Protein Ligand	
1EFN	FYN SH3(mutant)(-7.0)	core domain HIV-1 NEF(+2)	-36.6
2PTC	β-Trypsin(+8)	BPTI (+6)	-75.7
2SNI	Subtilisin Novo(+3)	CI-2(-1)	-66.1
1CHO	α-Chymotrypsin(+3)	OMTKY3	-60.3
3HFL	IG*G1 FAB fragment(-1)	Lysozyme (+8)	-59.4
3SGB	Proteinase B	OMTKY3	-53.1
4SGB	Proteinase B	PCI-1(+3)	-49.0
2TGP	Trypsinogen (+8)	BPTI (+6)	-33.1

4.1.2 All-atom approach

Absolute free energies of two SH2-peptide (1SHD, 1LKK) and three SH3-peptide (1BBZ, 1ABO, 1FYN) complexes (Table 1) were computed using molecular dynamics (MD) simulations in combination with the thermodynamic integration (double decoupling) and the LIE methods.

Double decoupling method

The thermodynamic cycle in Figure 4 was applied and the binding free energy, ΔG_b, given by:

$$\Delta G_b = \Delta G_1 - \Delta G_2 \tag{37}$$

ΔG_1 and ΔG_2 are the free energy changes associated with the decoupling of the ligand (and counter-ions) from the solvent (no protein present) and with the decoupling of the ligand (and counter-ions) from the solvent and the protein, respectively. Decoupling in this context means that in the final state the ligand and eventual counter-ions do not "see" the rest of the system and are therefore effectively in a gas phase state. This procedure is referred to as the double decoupling method (Jorgensen *et al.* 1988, Gilson *et al.* 1997).

The free energy difference, ΔG, was estimated using the thermodynamic integration formula in Eq (29) (Beveridge & DiCapua 1989, King 1993, Kollman 1993). The integration was carried out numerically by means of the trapezoidal method. The error in the $< \partial H / \partial \lambda >$ was calculated using a block averaging procedure (Allen & Tildesley 1987, Hess 2002).

To describe the decoupling of the ligand from the system, 8 separate λ points were simulated (0, 0.1, 0.2, 0.4, 0.6, 0.8, 0.9, 1). The ligand intramolecular interactions were described in two different ways in SH2 and SH3 complexes. In the case of the SH2-peptide complexes, all the atoms of the ligand were gradually mutated into dummy atoms as λ evolved from zero to one. A dummy atom is an atom for which the non-bonded interactions (Lennard-Jones and Coulomb) are set to zero. The bonded interactions within the molecule were kept unchanged. The same applied to SH3-peptide complexes, with the difference that Lennard-Jones interactions within the ligand were not switched off.

The ligand of the two SH2-peptide complexes had an overall charge of -5 e. To

ensure electroneutrality of the system at the different stages of the decoupling process, 5 Na^+ counter-ions were decoupled by transforming them into dummy atoms (Figure 4). The Hamiltonian of the system can be written as the sum of the Hamiltonians of the ligand, $H_L(\lambda)$, and of the counter-ions, $H_I(\lambda)$: $H(\lambda) = H_L(\lambda) + H_I(\lambda)$, and the free energy of binding becomes:

$$\Delta G = \int_0^1 d\lambda \left\langle \frac{\partial H_L(\lambda)}{\partial \lambda} + \frac{\partial H_I(\lambda)}{\partial \lambda} \right\rangle_\lambda \tag{38}$$

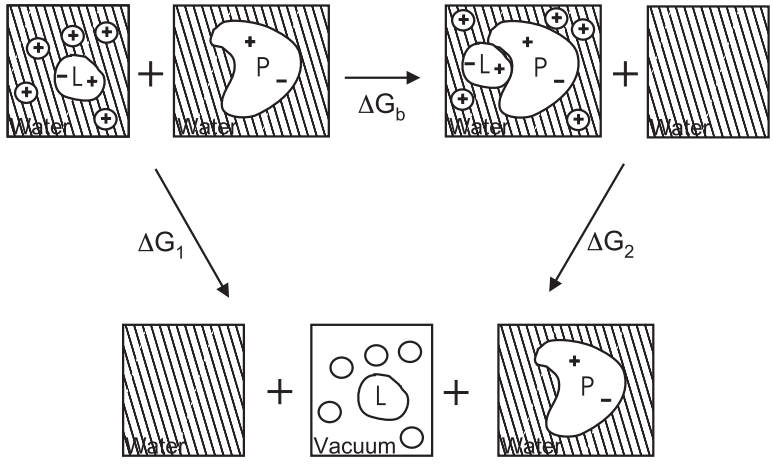

Fig 4. Thermodynamic cycle for the calculation of the binding free energy, ΔG_b, using the all-atom approach (double decoupling method). The negatively charged ligand (-5 e) and five positive counter-ions are decoupled from water (ΔG_1) and from the protein and water (ΔG_2).

If there is no correlation between $\partial H_L(\lambda)/\partial\lambda$ and $\partial H_I(\lambda)/\partial\lambda$, Eq (38) can be written as the sum of two different integrals and the free energy calculated as the sum of the free energy difference of the ligand and the ions (Smith & van Gunsteren 1994). If the free energy difference for the ions is the same in the two decoupling processes, the contribution of the ions to ΔG_b is effectively canceling. The degree of correlation at every λ point can be estimated from the correlation coefficient, $C_{LI}(\lambda)$ (Kreyszig 1999) of the $\partial H(\lambda)/\partial\lambda$.

MD simulations were performed with a standard version of Gromacs (Berendsen *et al.* 1995, Lindahl *et al.* 2001, van der Spoel *et al.* 2005). In order to compute the indi-

vidual contributions of the ligand and ions to the total $\partial H/\partial\lambda$, in the simulations of the SH2 complexes, a modified version of Gromacs was implemented. The GROMOS96 force field (van Gunsteren *et al.* 1996) was used to describe the compounds. Force field parameters for phosphotyrosine were adapted from Smith (Hansson *et al.* 1997, Smith 2002).

The molecules were placed in the centre of a dodecahedron box, which was subsequently filled with Simple Point Charge (SCP) water molecules (Berendsen *et al.* 1981). The number of water molecules varied with the size of the protein, and ranged between 1237 and 6233. A twin range cut-off was used for the Coulomb and Lennard-Jones interactions with cutoff distances of 1.8 nm and 1.4 nm, respectively. Interactions between atoms within 1.0 nm were evaluated every step, while interactions between atoms within the longer cutoff distances were evaluated every 5 steps. For the electrostatic energy, a reaction field correction was applied. Non-bonded interactions between the initial and the final state were interpolated using a soft-core potential (Beutler *et al.* 1994). The soft-core parameter α was set to 0.50 for the SH3 simulations and to 1.51 for the SH2 ones. These parameters were chosen on the basis of test calculations and previous work (Mordasini & McCammon 2000, van der Spoel *et al.* 2001, Villa & Mark 2002). Constant pressure p and temperature T were maintained by a weak coupling of the system to an external bath at 1 bar and 300 K using the Berendsen barostat and thermostat (Berendsen *et al.* 1984) with a coupling time of 1.0 picoseconds (ps) and 0.2 ps, respectively. A leap-frog stochastic dynamic integrator (van Gunsteren & Berendsen 1988) was used for the simulations of the SH2 complexes (1SHD, 1LKK) and the SH3 complex 1FYN, and for the λ=1 simulations of the SH3 complexes 1BBZ and 1ABO. In all other cases a leap-frog integrator was used. The stochastic integrator prevents an accumulation of translational and rotational kinetic energy in the uncoupled system, an artifact of the Berendsen thermostat (Villa & Mark 2002). The bond distances and bond angles of water were constrained using the SETTLE algorithm (Miyamoto & Kollman 1992). All other bond distances were constrained with the LINCS algorithm (Hess *et al.* 1997). Prior to the simulations, the potential energy of each system was minimized using steepest descent, followed by a 10 ps MD simulation with position restraints on the protein, to relax the water molecules. Then, a MD simulation was performed to equilibrate each system before starting the data collection (every ps). The equilibration time was chosen based on the analysis of the drift and the statistical error of $< \partial H/\partial\lambda >$, and on the RMSD of the molecule, and varied between 2 and 4 nanoseconds (ns), with the exception of the free ligand of the SH2 complexes (12 ns). The length of the pro-

duction simulation was chosen such that the free energy difference would reach a stable value (the longest simulation was 9 ns).

A harmonic restraining potential with a force constant k_{pr} was applied to prevent the SH2 and SH3 complexes' ligands from moving out of the binding site during the decoupling. For this purpose, two atom positions were restrained: one close to the centre of mass of the ligand and one close to the centre of mass of the protein. The force constant was $k_{pr} = \lambda * 1000$ kJ mol^{-1} nm^{-2} at $\lambda = 0.4 - 0.9$ and zero otherwise. During the simulation at $\lambda = 0.9$ of the 1FYN complex, k_{pr} was set to 1000 kJ mol^{-1}nm^{-2}: a stronger restraining was required because the peptide had an increased tendency to move out of the binding site. At every step of each λ point simulation, the derivative of the restraining potential with respect to λ was calculated and added to the total derivative of the potential energy with respect to λ (Straatsma & McCammon 1991). The net work of applying the position restraints, *i.e.* restraining the translational motion of a molecule to a position in space at $\lambda = 0.4$ and removing it at $\lambda = 0.9$, was estimated from the effective volume of the molecule (Hermans & Wang 1997) and it was equal to -4.04 kJ mol^{-1} for 1SHD, 1LKK, 1BBZ, and 1ABO, and to -4.56 kJ mol^{-1} for 1FYN.

A correction was applied to account for the different concentration (C) of the solutes in the simulation box and in the standard state. This is given by:

$$\Delta G_{corr} = -RTln\frac{C_B C_A^{\ominus}}{C_A C_B^{\ominus}} \tag{39}$$

where A and B denote the initial and final states, respectively, and $^{\ominus}$ indicates standard quantities. The concentrations of the solutes were estimated from the volume of the corresponding simulation boxes. ΔG_{corr} was -8.8 kJ mol^{-1} for the SH2 and -9.9 kJ mol^{-1} for the SH3 complexes.

The enthalpy difference, $\Delta H^{\ominus}$, was computed as the difference in the potential energy of the whole system between the bound and unbound state. The pressure-volume work $p\Delta V$ was negligible in the simulations. The change of entropy was simply estimated from $\Delta S^{\ominus} = (\Delta H^{\ominus} - \Delta G^{\ominus})/T$.

Linear interaction energy

The LIE method was applied according to Eq (30) with $\alpha = 0.16$ and $\beta = 0.5$. The interaction energies of the free ligand and of the ligand bound to the protein were estimated from the simulations at λ=0 of the double decoupling approach, discussed in the previous section.

4.2 Calculation of relative free energies of binding

The difference in affinity of two inhibitors of triosephosphate isomerase (TIM), 2-phosphoglycolic acid (PGA) and 3-phosphonopropionic acid (3PP), was calculated using an all-atom approach. The form of the inhibitor that binds TIM has a protonated carboxylic group and carries a -2 *e* charge on the phosphate/phosphonomethyl of PGA and 3PP, respectively (species *C* in Figure 6). Upon binding, a proton is donated to the glutamate (Glu) 167 of TIM (TIM-H) and the inhibitors are fully deprotonated in the active site of the enzyme (-3 *e*, species *F* in Figure 6). In the figures the two forms of the inhibitors, the form in solution that binds TIM and the fully deprotonated form in the active site, are referred to as PGA-H/3PP-H and PGA/3PP, respectively. They are referred to in the text as PGA and 3PP, irrespective of their protonation state. The computed relative affinity is however determined for species F, the fully deprotonated form of the substrate, while the experimentally determined relative affinity contains contributions from *all* accessible protonation states of both the ligand and the protein. Consequently, to be able to compare an experimental relative affinity with an computed relative affinity, two additional corrections are necessarily:

1. The computed affinity for species F must be corrected for the free energy that is required to transfer a proton from TIM to species F. This step results in the *predicted* free energy of binding of species C.
2. The experimental affinity must be corrected such that an affinity is obtained that is specifically for species C only. This step results in the *effective* free energy of binding of species C.

The predicted relative free energy of binding or predicted relative affinity can now directly be compared to the relative effective affinity as they now both refer to the same species C.

4.2.1 Predicted relative affinity

Figure 5 details a number of thermodynamic cycles that are employed to estimate the difference in the free energy of binding $\Delta\Delta G\,(pred)$ (*predicted* relative affinity) between PGA (ΔG_1) and 3PP (ΔG_2) to TIM for species C in Figure 6:

$$\Delta\Delta G(pred) = \Delta G_1 - \Delta G_2 = \Delta G_5 - \Delta G_4 \tag{40}$$

where ΔG_1 and ΔG_2 are:

$$\Delta G_1 = \Delta G_{1a} + \Delta G_{1b} \tag{41}$$

$$\Delta G_2 = \Delta G_{2a} + \Delta G_{2b} \tag{42}$$

and ΔG_5 and ΔG_4 are the free energies for the mutation of PGA into 3PP in solution and in the protein, respectively.

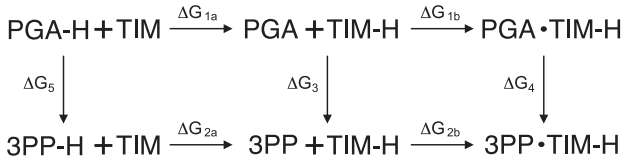

Fig 5. Thermodynamic cycle for the calculation of the relative free energy of binding ($\Delta\Delta G$) of ligands PGA-H and 3PP-H to TIM. PGA-H and 3PP-H are the binding forms of the inhibitors (species C in Figure 6). First, the inhibitors are deprotonated, while the Glu 167 of TIM is protonated (TIM-H). Secondly, the fully deprotonated PGA and 3PP (species F in Figure 6), bind TIM-H.

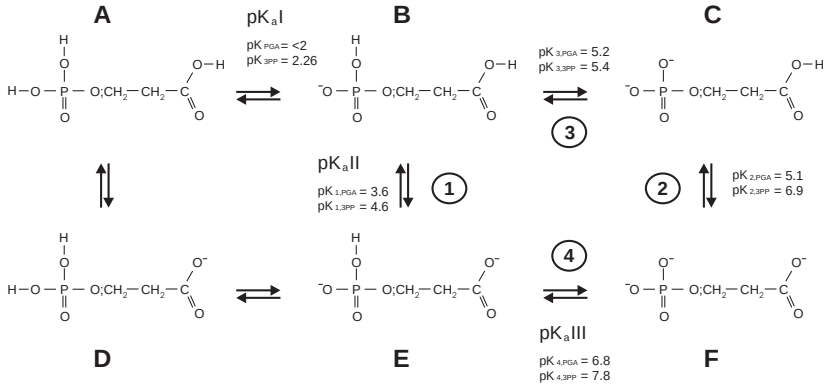

Fig 6. Protonation states of PGA and of its phosphonate analogue, 3PP. An oxygen in PGA is replaced by a CH_2 group in 3PP (indicated as $O;CH_2$). Estimated pK values are reported. pK_a I, pK_a II and pK_a III are the experimental pK_a values and are associated with reactions A→B, B→E and E→F, respectively. C is the binding species to TIM (PGA-H and 3PP-H in Figure 5). For reference with the text, reactions between compounds B, C, E and F are indicated with numbers.

Note, however, that in Eq (40), ΔG_5 is relative to the C form of the inhibitor and Glu 167 of TIM is not protonated, while ΔG_4 is relative to the F form (Figure 6) and the protein is now TIM-H.

The proton abstraction by TIM is described by a second thermodynamic cycle:

$$\Delta G_3 - \Delta G_5 = \Delta G_{2a} - \Delta G_{1a} \tag{43}$$

where ΔG_{1a} and ΔG_{2a} are the free energies for the deprotonation of the carboxylic group of the inhibitors in the -2 *e* form (C in Figure 6) and the simultaneous protonation of the catalytic glutamic acid of TIM (resulting in TIM-H). ΔG_3 is the free energy of mutating PGA into 3PP in water, where both are fully deprotonated (F in Figure 6).

Combining Eqs (40) and (43) we obtain:

$$\Delta\Delta G = \Delta G_{1a} - \Delta G_{2a} + \Delta G_3 - \Delta G_4 \tag{44}$$

The free energies of transformation, ΔG_3 and ΔG_4, are computed with thermodynamic integration and molecular dynamics, whereas (ΔG_{1a}-ΔG_{2a}) is calculated as:

$$\Delta G_{1a} - \Delta G_{2a} = 2.303RT(pK_{1a} - pK_{2a}) \tag{45}$$

where pK_{1a} and pK_{2a} are the pKs relative to the deprotonation of the carboxylic groups of PGA and 3PP, respectively. The method of calculating these pKs is explained in the section *pK calculations* further below. Note that $\Delta G_{1a} - \Delta G_{2a}$ is the correction to the computed relative affinity $\Delta\Delta G(calc)$ that was referred to under correction 1 at the beginning of the Methods section.

4.2.2 Thermodynamic integration and MD calculations

ΔG_3 and ΔG_4 in Figure 5 and Eq (44) were calculated using the thermodynamic integration formula (Eq (29)). 18 independent MD simulations, with λ between 0 and 1 were performed. For every λ point 100 ps of simulation was followed by 200 ps of data collection. During the transformations an oxygen atom was mutated into a carbon atom and all associated bonded and non-bonded interactions were mutated accordingly. To describe the carboxylic moiety of the inhibitors, two sets of atoms were used: the carboxylic group of PGA would gradually become dummy, while the dummy group of 3PP would turn into a carboxylic moiety.

The GROMOS96 force field (van Gunsteren *et al.* 1996) was used to describe the molecules. Force field parameters for the phosphate and phosphonomethyl moiety of

the inhibitors were derived from the force field parameters of phosphoserine (Hansson *et al.* 1997, Smith 2002).

All MD simulations were performed with Gromacs (Berendsen *et al.* 1995, Lindahl *et al.* 2001, van der Spoel *et al.* 2005). The X-ray structure of TIM complexed with PGA (PDB code 1N55 (Kursula & Wierenga 2003)) was used as the starting structure. The inhibitor was placed in a cubic box and the protein-inhibitor complex in a dodecahedric box which were subsequently filled with $\sim$1500 and $\sim$19800 SPC water molecules (Berendsen *et al.* 1981), respectively. All the heavy atoms of the protein were position-restrained with a force constant of 1000 kJ mol^{-1} nm^{-2}. This preserved the active site geometry during simulation. A twin range cut-off was used for the Coulomb and Lennard-Jones interactions with cut-off distances of 1.4 nm. All other parameter were defined as in the MD calculations described in the double decoupling section 4.1.2.

4.2.3 pK calculations

pK values were computed in terms of standard free energies for each proton dissociation reaction, using the Liptak and Shields approach (Liptak & Shields 2001) and the program Gaussian03 (Frisch *et al.* 2004). The standard free energy was given by the sum of the free energy difference in gas and the solvation free energy difference according to the thermodynamic cycle illustrated in Figure 3 and described in Eq (32). In this case the ligand L consisted of a proton. The B3LYP/6-31G* level of theory was applied. A frequency calculation was performed to account for the entropy contribution. To estimate the solvation free energy of each species a solvated calculation was carried out using the polarizable continuum model (PCM) (Miertus *et al.* 1981, Mennucci & Tomasi 1997) of Gaussian03.

PGA and 3PP have three titrating sites and are present in solution as a mixture of differently protonated species, shown in Figure 6. The geometry of each species was optimized in a vacuum at the B3LYP/6-31G* level of theory using Gaussian03 (Frisch *et al.* 2004). Experimental pK_a values (Hartman *et al.* 1975, Heubel & Popov 1979)(pK_a I, pK_a II and pK_a III) were associated with the deprotonation reactions as indicated in Figure 6, based on the comparison with the pK_as of phosphate, phosphonate and carboxylic acid. Species A and D are present at very low concentration at pH 7 (pK_a I $\approx$ 2), therefore, for simplicity, they were omitted from the calculations. Note the use of connotations pK_a and pK to indicate the experimentally and the computationally determined pK_a, respectively. As the calculated absolute pKs are prone to very large

errors and also to take advantage of the available experimental values, shifts in the calculated pKs of reactions of C→F (pK_2) with B→E (pK_1), ΔpK_{2-1}, and of B→C (pK_3) with E→F (pK_4), ΔpK_{3-4}, were considered instead. ΔpK_{2-1} and ΔpK_{3-4} can be thought of as shifts in the pK of a certain protonation equilibria when the overall charge on the molecule changed (Figure 6). pK_2 and pK_3 were then estimated as:

$$pK_2 = pK_1 + \Delta pK_{2-1} \tag{46}$$

$$pK_3 = pK_4 + \Delta pK_{3-4} \tag{47}$$

where pK_1 and pK_4 were given by the experimentally measured pK_a II and pK_a III, respectively. The obtained pK values are reported in Figure 6. pK_2 of PGA and 3PP correspond to pK_{1a} and pK_{2a}, respectively, in Eq (45).

4.2.4 Effective *affinities*

The experimentally determined inhibition constants, $K_i(exp)$, of TIM for its inhibitors, PGA and 3PP are obtained applying the following equation:

$$K_i(exp) = \frac{[TIM][I]}{[TIM \cdot C]} \tag{48}$$

where C is the form of the inhibitor I that binds the enzyme and [I] is the equilibrium concentration of inhibitor, that includes the different ionization forms of I in solution (Figure 6). The *effective* affinity is defined as the affinity of the inhibitor for TIM in a solution containing only the C form. The *effective* affinity, $K_i(eff)$, is related to $K_i(exp)$ by the following relation:

$$\begin{aligned} K_i(eff) &= \frac{[TIM][C]}{[TIM \cdot C]} \\ K_i(eff) &= \frac{[TIM][I]}{[TIM \cdot C]} \frac{[C]}{[I]} \\ K_i(eff) &= K_i(exp) \frac{[C]}{[I]} \end{aligned} \tag{49}$$

For the two ligands, PGA and 3PP, we can derive the following relation between

$K_i(eff)$ and $K_i(exp)$:

$$\frac{K_{i,PGA}(eff)}{K_{i,3PP}(eff)} = \frac{K_{i,PGA}(exp)}{K_{i,3PP}(exp)} \frac{[C_{PGA}]}{[C_{3PP}]} \frac{[3PP]}{[PGA]}$$

$$RTln\Big(\frac{K_{i,PGA}(eff)}{K_{i,3PP}(eff)}\Big) = RTln\Big(\frac{K_{i,PGA}(exp)}{K_{i,3PP}(exp)}\Big) + RTln\Big(\frac{[C_{PGA}]}{[C_{3PP}]} \frac{[3PP]}{[PGA]}\Big)$$

$$\Delta\Delta G(eff) = \Delta\Delta G(exp) + RTln\Big(\frac{[C_{PGA}]}{[C_{3PP}]} \frac{[3PP]}{[PGA]}\Big) \tag{50}$$

where $\Delta\Delta G$ is $\Delta G_{PGA} - \Delta G_{3PP}$.

4.2.5 Determination of equilibrium concentrations

The equilibrium concentrations are calculated from the following equations:

$$[I]_0 = [I] + [TIM \cdot C] \tag{51}$$

$$\big[TIM\big]_0 = [TIM] + [TIM \cdot C] \tag{52}$$

where, $[I]_0$ and $[TIM]_0$ are the initial concentrations of inhibitor (I) and enzyme (TIM), respectively, and TIM·C is the complex between the inhibitor and TIM. The inhibitor I exists in solution in different ionization species (Figure 6), of which the one denoted C binds the enzyme.

The equilibrium concentration of unbound enzyme in solution ([TIM]) can be obtained from Eq (48), and Eq (52) then becomes:

$$\big[TIM\big]_0 = \frac{K_i(exp)[TIM \cdot C]}{[I]} + [TIM \cdot C] \tag{53}$$

At pH 7.6, species A and D of the inhibitors (Figure 6) are present in negligible concentration, and are therefore not considered in the following equations. At equilibrium, [I] can be expressed as a function of [C] as:

$$[I] = [B] + [C] + [E] + [F]$$

$$\big[I\big] = \frac{[C][H^+]}{K_3} + [C] + \frac{K_2[C]}{K_4} + \frac{K_2[C]}{[H^+]} \tag{54}$$

where K_2, K_3 and K_4 are associated with the proton dissociation reactions C→F, B→C and E→F, respectively, and are calculated from the corresponding pK values reported in Figure 6, and $[H^+]$ is calculated from the pH of the solution.

When [I] in Eqs (51) and (53) is expressed as in Eq (54), a system of two equations as function of [C] and [TIM·C] is obtained. The program Mathematica (Wolfram Research 2003) was used to solve this system. The equilibrium concentrations of the other species were then also determined.

4.3 Incorporating ionic strength in free energy calculations

The difference in the free energy of solvation between two triosephosphate isomerase (TIM) inhibitors, 2-phosphoglycolic acid (PGA) and 3-phosphonopropionic acid (3PP), was investigated at five different ionic strengths. Calculations were performed for both the neutral and ionic (-3 *e*) forms of the compounds. The thermodynamic cycle in Figure 2 was applied, where compounds *A* and *B* corresponded to PGA and 3PP. ΔG_3 and ΔG_4 are the free energy differences between PGA and 3PP in a vacuum and in water, respectively, and were calculated using molecular dynamics simulations and thermodynamic integration formula. ΔG_4 was calculated with different concentrations of Na^+ and Cl^- ions in solution, to reproduce different ionic strength concentrations (0, 0.04, 0.06, $\sim$0.1 and $\sim$2 M).

PGA and 3PP were modelled as described in the relative free energy calculations section 4.2.2, with the exception of the carboxylic moiety of the inhibitors, that did not change during the transformations.

All molecular dynamics (MD) simulations were performed with the Gromacs suite of programs (Berendsen *et al.* 1995, Lindahl *et al.* 2001, van der Spoel *et al.* 2005). The inhibitors were placed in a cubic box (edge length approximately 4 nm), which was subsequently filled with $\sim$2170 SPC water molecules (Berendsen *et al.* 1981). All other simulation and free energy calculation parameters were the same as described in section 4.2.2. To estimate ΔG_3, calculations at 0 M were repeated in a vacuum.

4.4 Conformational energies of proline 168 in the active site of TIM

In the atomic resolution structure of liganded triosephosphate isomerase (TIM) (PDB code 1N55 (Kursula & Wierenga 2003)), the side-chain of the active site proline (Pro) 168 occurs in a planar conformation. Potential energies of Pro 168 in three different

conformations, down and up puckered, and planar, were computed in a vacuum and in the protein environment. The down and up puckers of Pro 168 were obtained by fitting these two conformations of the pyrrolidine ring onto the backbone atoms of Pro 168. No optimization of the down and up conformations inside the protein was performed.

To model the proline and its environment correctly, all calculations were performed using the hybrid quantum mechanics/molecular mechanics (QM/MM) approach (Warshel & Levitt 1976). For these calculations, a special version of Gromacs 3.3 (Lindahl *et al.* 2001, van der Spoel *et al.* 2005) with an interface to Gaussian03 (Frisch *et al.* 2004) was employed (Groenhof *et al.* 2004). The proline, including a stretch of the backbone which ranged from the C_α of Glu 167 to the C_α of valine (Val) 169, and nearby residues were described at the B3LYP/6-31G* level of theory. These atoms constituted the QM subsystem while all other atoms were described at the MM level. The computations were repeated with an increasing number of atoms in the QM subsystem: 26 (QM1), 48 (QM2), 60 (QM3), 129 (QM4) and 201 (QM5) atoms. These roughly correspond to spheres centreed at the proline with increasing radii up to about 0.5 nm in QM5. The remainder of the system (MM), consisting of the rest of the protein, the inhibitor (PGA, 2-phosphoglycolic acid), and approximately 20,000 SPC water molecules (Berendsen *et al.* 1981) in a truncated dodecahedron box with periodic boundary conditions, was modelled with the GROMOS96 force field (van Gunsteren *et al.* 1996). The bonds connecting the QM and MM subsystems were replaced by constraints (Hess *et al.* 1997) and the QM part was capped with hydrogen atoms. The forces acting on these cap atoms were distributed over the atom pairs that formed the original bond. The QM system experienced the Coulomb field of all MM atoms and Lennard-Jones interactions between QM and MM atoms were also added.

In addition, potential energy profiles for the down to up pucker transition were calculated. First, the energy profile for the down→up transition was calculated in a vacuum for the proline only, using the Synchronous Transit-Guided Quasi-Newton (STQN) method of Schleger and co-workers (Peng & Schlegel 1994). A total of 12 intermediate points were employed to construct a suitable minimum energy path connecting the down and up puckered conformations. These intermediate conformations were then fitted onto the proline 168 backbone atoms in the protein and in the isolated QM1 system in a vacuum, and energies calculated. Again, no optimization of these intermediate conformations inside the protein was performed.

5 Results

5.1 Calculation of absolute binding affinities

5.1.1 Continuum approach (I)

Free energies of binding of 36 protein-ligand complexes were calculated using the continuum approach. The complexes include protein-small molecule complexes, protein-peptide complexes and protein-protein complexes (Table 1). Four protein families were included - TIM, SH2, SH3 and MHC - among other complexes. For each protein family at least two different members were considered, *i.e.* proteins which are homologous. For instance, in the case of triosephosphate isomerase, the complex identified by the PDB code 1AMK includes a TIM from *L. mexicana*, whereas in the other complexes TIM is from *T. brucei*. Figure 7 shows the calculated standard binding free energies, ΔG_{calc}, *versus* the experimental ones, ΔG_{exp}. The calculated affinities roughly follow the pattern seen in the experimental ones, but their numerical value can differ by more than two orders of magnitude. Only for a subset of the protein ligand complexes, *i.e.* the SH2 and SH3 complexes, ΔG_{calc} was about in the same range as the ΔG_{exp} (between -50 and -20 kJ mol^{-1}). In particular, the ranking order of ΔG_{calc} for the SH3-peptide complexes was correctly predicted (Table 2). However, the calculated ΔG for three complexes, including two of the SH3-peptide complexes and 1AMK, with almost identical ΔG_{exp}, were very different. Also, in about 20% of the cases the calculated affinities were positive, including all the triosephosphate isomerase complexes.

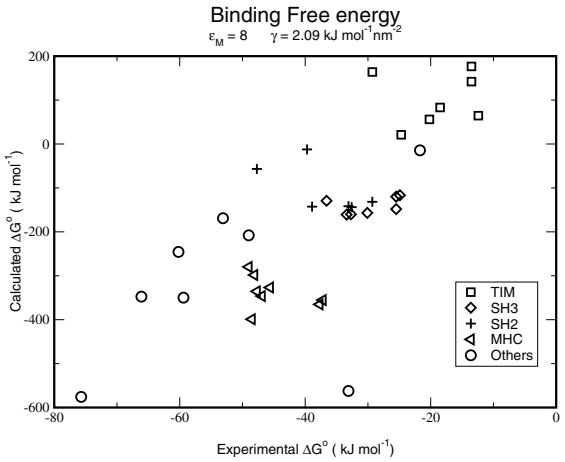

Fig 7. Experimental and calculated binding free energies of 36 protein-ligand complexes. The continuum approach was employed with ε_m=8 and γ=2.09 kJ mol^{-1}.

Table 2. Binding free energy (ΔG) and conformational entropy contributions of SH3-peptide complexes using the continuum approach. ΔS_{sc} and ΔS_{vib} are the side-chain and vibrational entropy difference, respectively. T is the temperature. ΔG and $T\Delta S$ are in kJ mol^{-1}. Abbreviations: experimental (*exp*); calculated (*calc*).

PDB	ΔG_{exp}	ΔG_{calc}	$-T\Delta S_{sc}$	$-T\Delta S_{vib}$
1BBZ	-33.4	-160.8	14.1	15.8
1CKA	-32.7	-160.1	26.7	10.0
1CKB	-30.1	-156.9	29.7	71.2
1ABO	-25.5	-148.2	14.5	29.5
1FYN	-25.5	-119.9	15.3	-20.5
1SEM	-24.9	-116.7	19.2	-37.7

Two important parameters of the continuum method are the dielectric constant of the cavity ε_m and the surface tension γ. These are associated with the electrostatic, ΔG_{el}, and non-polar, ΔG_{np}, terms of the solvation free energy, respectively (see Methods 4.1.1). Because these parameters are rather empirical, the calculations were repeated for several choices of these (ε_m between 2 and 40, and γ between 0.42 and 29.3 kJ mol^{-1}). Both the absolute and relative free energies of five protein families (MHC,

ENP, SH2, SH3 and TIM) in the set of 36 complexes were analyzed, and the values of ε_m and γ that gave the best agreement with experimental measurements within a protein family identified. In the case of absolute affinities, γ=0.42 kJ mol^{-1} gave the best agreement, while ε_m was different for every protein family. When considering relative affinities, γ=0.42 kJ mol^{-1} again and ε_m=40 were the values found for all protein families with the exception of SH3 (γ=8.37 kJ mol^{-1} and ε_m=8). Note that ε_m=40 and γ=0.42 kJ mol^{-1} are the highest and the lowest values of ε_m and γ that were considered.

The values of ΔG_{calc} reported in Figure 7 do not include contributions from the change in conformational freedom of protein and ligand upon binding. The difference in the number of accessible side-chain (*sc*) rotamers in the liganded and unliganded states can be used to estimate to some extent their contribution. When the $-T\Delta S_{sc}$ is added to the calculated ΔG of the 36 protein-ligand complexes, the predicted values shifted towards the experimental ones, but in most cases the calculated ΔG was still one or two orders of magnitude away from the experimental one. In one case (1SHD) the predicted affinity changed sign. In Table 2, the $-T\Delta S_{sc}$ of six SH3-peptide complexes is reported. For these complexes, the vibrational entropy difference of binding, ΔS_{vib}, that is a measure of the contribution to the affinity due to conformational flexibility, was explicitly calculated from the normal mode frequencies of the complex and of the free protein and free ligand (Table 2). Inclusion of $-T\Delta S_{vib}$ shifted the calculated ΔG values towards those found experimentally for all but two of the complexes (1FYN and 1SEM). However, the ranking of the experimental affinities was not reproduced.

In the calculations reported in Figure 7, the structures of both the free protein and the free ligand were derived from the coordinates of the protein-ligand complex. When the computations were repeated using the structures of the free protein and the free ligand, no improvement in the predicted affinities was achieved, but in fact the agreement between experimental and calculated ΔG was lower.

In all cases upon formation of the complex the total solvent-accessible surface of the system decreases ($\Delta A < 0$). The correlation between the decrease in ΔA and the individual contributions to ΔG was calculated. The non-polar contribution to ΔG, is directly proportional to ΔA and shows therefore correlation. However, other contributions, in particular the van der Waals energy difference, were highly correlated to ΔA (correlation coefficient = 0.98).

5.1.2 All-atom approach (I)

The free energies of binding of two SH2-peptide complexes (1SHD and 1LKK) and three SH3-peptide (1BBZ, 1ABO and 1FYN) were calculated using molecular dynamics simulations and thermodynamic integration formula. The predicted affinities are listed in Table 3. All the calculated ΔG values were more negative than the experimental ones. Interestingly, the ΔG_{calc} of 1BBZ and 1ABO were the same as those obtained with the continuum approach, within the limit of the statistical error. With the exception of 1FYN, the trend in the affinities was reproduced. The experimental affinities of the 1ABO and 1FYN SH3-peptide complexes were the same, but the calculated ones differed by 45 kJ mol^{-1}.

ΔG was given by $\Delta G_1 - \Delta G_2$ with ΔG_1 the free energy difference for the decoupling of the ligand from the solvent, and ΔG_2 the ΔG for the decoupling of the ligand from protein and solvent (Figure 4). ΔG_1 and ΔG_2 are reported in Table 3. Even though the relative error of ΔG_1 and ΔG_2 was below 2%, the error in the final ΔG was between 7% and 14%. This was a consequence of the fact that the free energy of binding was calculated as the difference between two large numbers with respect to the final affinity. The difference in enthalpy and entropy of binding is also listed in Table 3. The enthalpy was calculated from the potential energy of the whole system, and the entropy was computed as the difference between the calculated free energy and enthalpy. The relative errors of these number were very high for 1SHD, 1LKK and particularly for 1FYN. With the exception of 1FYN, the entropy difference was always positive and compensated for the positive enthalpy difference, indicating that all the association reactions were endothermic.

ΔG_1 and ΔG_2 are calculated integrating the $< \partial H(\lambda) >$ between the values of the coupling parameter λ, 0 and 1. The error in the $< \partial H(\lambda) >$ was below 3% in almost all cases, with the highest relative errors occurring at higher values of λ when the ligand (and counter-ions) were almost fully decoupled from the rest of the system. The profile of the $< \partial H(\lambda) >$ curves as a function of λ was different for SH3 and SH2 complexes. A well occurred at λ=0.9 in the SH3 profiles, whereas it occurred at λ=0.4-0.6 in those of SH2. The different profile of the curves might be related to the use of different values for the soft core parameter α in the SH3 and SH2 simulations - 0.5 and 1.51 - respectively (Beutler *et al.* 1994, Mordasini & McCammon 2000).

Table 3. Binding free energy (ΔG), enthalpy (ΔH) and entropy (ΔS) of 2 SH2-peptide (in italics) and 3 SH3-peptide (in bold) complexes calculated using an all-atom approach (double decoupling method). ΔG_1 and ΔG_2 are the free energy changes of decoupling the ligand from the solvent, and of decoupling the ligand from the protein and the solvent, respectively (see Figure 4). The ligands of 1SHD and 1LKK complexes were the same peptide, therefore ΔG_1 values were identical. In the case of a charged ligand (1SHD and 1LKK), counter-ions were decoupled along with the ligand. ΔG, ΔH and $T\Delta S$ in kJ mol^{-1}.

PDB	ΔG_{exp}	ΔG_{calc}	ΔG_1	ΔG_2	ΔH_{calc}	$T\Delta S_{calc}$
1SHD	-39.7±0.1	-253.6±31.7	3637.4±20.3	3882.2±24.4	103.9±49.9	357.5±59.1
1LKK	-38.9±0.1	-223.5±31.0	3637.4±20.3	3852.1±23.4	41.2±13.2	264.7±33.7
1BBZ	-33.4±0.2	-165.1±15.2	779.5±12.4	934.7±8.8	312.7±13.2	477.8±20.1
1ABO	-25.5±0.2	-135.9±14.9	652.5±6.5	778.5±13.4	395.2±14.4	531.1±20.7
1FYN	-25.5±0.2	-180.7±11.7	840.9±4.8	1011.6±10.7	-16.7±12.0	164.0±16.8

The ligands of the SH2 complexes were charged (-5 e). To ensure that the system remained electrostatically neutral during all λ points along the transformations, 5 Na^+ counter-ions were decoupled along with the ligand. In order to establish the contribution of this additional decoupling to the calculated free energy, the individual contributions of the ligand (L) and ions (I) to the derivative of the potential energy with respect to λ, $\partial H_L/\partial\lambda$ and $\partial H_I/\partial\lambda$, repectively, were calculated. The free energy of binding of the ligand only (ΔG_L) was then estimated: -168.8 and -131.3 kJ mol^{-1} for 1SHD and 1LKK, respectively (Table 4). The counter-ions contrbuted approximately 80 kJ mol^{-1} to the calculated ΔG. Note also that the ΔG_L values were in the same range as the ones obtained for the SH3 complexes. ΔG_L and ΔG_I were estimated under the assumption that $\partial H_L/\partial\lambda$ and $\partial H_I/\partial\lambda$ were not correlated. This hypothesis was checked, and it was found that while there was no correlation or it was rather weak at λ points 0 and 1, at intermediate λ points the correlation was high, up to -0.8 at λ=0.8. The $\partial H_I/\partial\lambda$ contributed considerably to the total $\partial H/\partial\lambda$ and was certainly not independent from $\partial H_L/\partial\lambda$. The estimated values of ΔG_L and ΔG_I are, therefore, to be taken as indicative only.

Table 4. Free energy contributions of ligand and ions to ΔG_1 (decoupling of ligand and 5 Na^+ ions from solvent), ΔG_2 (decoupling of ligand and 5 Na^+ ions from protein and solvent) and to the calculated binding free energy (ΔG_{calc}). The ligands of the SH2 complexes, 1SHD and 1LKK, were the same peptide, therefore ΔG_1 values were identical. ΔG in kJ mol^{-1}.

	1SHD	*1LKK*
$\Delta G_{1,ligand}$	1801.4	1801.4
$\Delta G_{1,ions}$	1836.0	1836.0
$\Delta G_{1,ligand+ions}$	3637.4±20.3	3637.4±20.3
$\Delta G_{2,ligand}$	1970.2	1932.7
$\Delta G_{2,ions}$	1912.0	1919.4
$\Delta G_{2,ligand+ions}$	3882.2±24.4	3852.1±23.4
$\Delta G_{calc,ligand}$	-168.8	-131.3
$\Delta G_{calc,ions}$	-76.0	-83.4
$\Delta G^*_{calc,ligand+ions}$	-244.8	-214.7

*Standard state correction not included in the ΔG_{calc}.

All the Coulomb and Lennard-Jones interaction energies between the ligand, counter-ions and water were analyzed as a function of λ during the decoupling of the SH2 ligand (1LKK) and counter-ions (Na^+) from solvent. All terms, except interactions between water molecules, approached zero when λ approached 1. The ligand-solvent (L-H_2O) and counter-ion-solvent (Na^+-H_2O) Coulomb interaction energies decreased quite dramatically in going from λ=0 to λ=0.1. A similar change, but less noticeable, is observed for the SH3 ligands. After an initial strengthening, the ligand-Na^+ Coulomb interaction energy became much weaker from λ=0.1 on. We also observed an increase in the electrostatic interaction between water molecules (H_2O-H_2O) caused by the fact that water molecules that were involved in the ion and peptide interactions at λ=0 became available for interactions with other solvent molecules.

Affinities for the association reactions of the SH2-peptide and SH3-peptide complexes were also calculated using the linear interaction energy method (LIE). The affinities of the SH2 complexes were very high (-416.7 and -459.4 kJ mol^{-1} for 1SHD and 1LKK, respectively) as the result of a large difference in the electrostatic terms in the bound and unbound state of the ligand. On the contrary, LIE affinities for SH3 com-

plexes (-24.4, -40.3 and -86.2 kJ mol^{-1} for 1BBZ, 1ABO and 1FYN, respectively) were in the same range as the experimental ones. However, the experimental ranking order of affinities was not reproduced.

5.2 Calculation of relative binding affinities (II)

Relative affinity ($\Delta\Delta G$) of triosephosphate isomerase (TIM) enzyme for two of its inhibitors, PGA (2-phosphoglycolic acid) and 3PP (3-phosphonopropionic acid) has been calculated according to the thermodynamic cycle in Figure 5 and Eq (44). To determine the *predicted* relative affinity ($\Delta\Delta G(pred)$) for the binding form of the inhibitors (PGA-H and 3PP-H in Figure 5 and species C in Figure 6), it is required to correct the *computed* relative affinity $\Delta\Delta G(calc)$ for the fully deprotonated form (PGA and 3PP in Figure 5 and species F in Figure 6), for the free energy that is required to transfer a proton from TIM to species F (Figure 5). This correction is given by ΔG_{1a}-ΔG_{2a} as in Eq (45). The experimental and theoretical affinities are listed in Table 5. The predicted relative affinity of -6.1 kJ mol^{-1}, with PGA being the strongest inhibitor, significantly underestimated the experimental relative affinity $\Delta\Delta G(exp)$ (-17.1 kJ mol^{-}1).

As already mentioned above, ΔG_{1a} and ΔG_{2a} are relative to the deprotonation of the carboxylic groups of PGA and 3PP to yield the fully charged -3 *e* inhibitors. pK values for the corresponding reactions were estimated: 5.1 and 6.9 for PGA and 3PP, respectively. These yielded a free energy difference of -10.3 kJ mol^{-1} for ΔG_{1a}-ΔG_{2a}, with the free energy difference for the deprotonation of PGA more negative, and more favourable, than 3PP.

$\Delta\Delta G(calc)$ was obtained from the difference between the free energy for the transformation of PGA to 3PP in water (ΔG_3) and in the active site of the solvated TIM (ΔG_4) (Figure 5). Each ΔG was computed using molecular dynamics simulations and thermodynamic integration method. The calculated ΔG_3 and ΔG_4 were both positive and of comparable magnitude, 28.3 $\pm$4.3 kJ mol^{-1} and 24.1$\pm$7.2 kJ mol^{-1}, respectively. According to the thermodynamic cycle in Figure 5, ΔG_3-ΔG_4 corresponds to the binding free energy difference of PGA and 3PP in the -3 *e* fully deprotonated forms (ΔG_{1b}-ΔG_{2b}). This difference (4.2$\pm$8.4 kJ mol^{-1}) was not significant, indicating that the -3 *e* form of the inhibitors PGA and 3PP had a comparable affinity for TIM.

It is possible to compute how much specific bonded and non-bonded interactions contribute to the calculated ΔG_3 and ΔG_4. It was found that the contributions to $\Delta\Delta G(calc)$ ($\Delta\Delta G(calc)$=ΔG_3-ΔG_4) from the inhibitor-inhibitor (I-I), inhibitor-solvent (I-H_2O) and

inhibitor-protein (I-TIM) interactions were -2.0, 10.1 and -3.9 kJ mol^{-1}, respectively. Intramolecular interactions of the inhibitors (I-I) and interactions of the inhibitors in the active site with the protein (I-TIM) contributed, therefore, relatively little to $\Delta\Delta G$.

Table 5. Experimental and theoretical results for the affinity of PGA and 3PP for TIM. Effective affinities of the C form of the ligands for TIM ($\Delta G(eff)$) are derived from the experimental affinities of PGA and 3PP ($\Delta G(exp)$). $\Delta G(C \rightarrow F)$ of PGA and 3PP correspond to ΔG_{1a} and ΔG_{2a}, respectively, in Figure 5 and to reaction 2 in Figure 6. The calculated (*calc*) and the predicted (*pred*) relative affinities ($\Delta\Delta G$) are reported in the last two rows. *Exp* data are taken from Lambeir *et al.* (1987) (PGA) and Noble *et al.* (1991) (3PP).

K_i (mM) /	LIGAND		$\Delta\Delta G$
ΔG (kJ mol^{-1})	PGA	3PP	(PGA-3PP)
$K_i(exp)$	0.027±0.005	27	
$\Delta G(exp)$	-26.1±0.4	-9.0	-17.1
$\Delta G(eff)$	-40.7	-15.5	-25.2
$\Delta G(C \rightarrow F)$	29.1	39.4	-10.3
$\Delta G(calc)$			4.2±8.4
$\Delta G(pred)$			-6.1

To further investigate the ligand-protein interactions, 200 ps simulations of PGA and 3PP in the active site of the enzyme was analyzed. The average number of hydrogen bonds between the inhibitors and the protein was 11.5 and 11.7 for PGA and 3PP, respectively, and between the inhibitors and water molecules in the active site 5.1 and 5.0 for PGA and 3PP, respectively. The oxygen atom of the phosphate group was not involved in significant hydrogen bonds. Lennard-Jones and short range (within 0.9 nm) Coulomb interactions energies of PGA and 3PP with the protein were comparable, whereas long range (between 0.9 and 1.4 nm) Coulomb interactions energies were more favourable (of about 15 kJ mol^{-1}) in the TIM-3PP complex than in the TIM-PGA. Even though the total electrostatic interaction energies were more favourable for the TIM-3PP complex, the inhibitor-protein $\Delta\Delta$G component (-3.9 kJ mol^{-1}) indicated that the TIM-PGA complex was slightly more stable than TIM-3PP.

PGA and 3PP are present in solution in different ionization states depending on which of the three titrating sites are protonated (Figure 6). Of these different forms, the

-2 e compound with a protonated carboxylic group binds the enzyme. In the calculations only the binding reaction between the C form of the inhibitor and the enzyme was taken into account. When the affinities are determined experimentally (*exp*), the solution contains all ionization states of the inhibitor, and the ratio of every compound depends on the pH of the solution. Therefore, the equilibria that are considered in the calculations and in the experiment are not the same and one cannot directly compare calculated and experimental affinities. If the ratio of the equilibrium concentration of the binding form of the inhibitor respect with the concentration of inhibitor is known, then a quantity that we refer to as *effective* affinities can be obtained (Eq (49)), that is comparable with the calculated affinity. This can be thought of as the affinity of a solution containing only the binding form C of the inhibitor. The concentrations of compound C of PGA and 3PP were estimated (0.27% and 7.2% of the total inhibitor concentration, respectively) and the effective (*eff*) affinity difference ($\Delta\Delta G$) of PGA and 3PP for TIM (in kJ mol^{-1}) obtained (Eq (50)):

$$\Delta\Delta G(eff) = \Delta\Delta G(exp) - 8.1$$

$\Delta\Delta G$ (*eff*) was -8.1 kJ mol^{-1} more negative than the experimental $\Delta\Delta G$. In other words, PGA had effectively -8.1 kJ mol^{-1} higher affinity to TIM with respect to 3PP than experimentally measured. This was a consequence of the fact that at pH 7.6 the concentration of the binding form of PGA was relatively smaller than 3PP, and PGA was effectively an even stronger ligand for TIM. When the *predicted* relative affinity (-6.1 kJ mol^{-1}) is compared to $\Delta\Delta G$ (*eff*) (-25.2 kJ mol^{-1}) (Table 5) we see that the calculations are underestimating the difference of binding free energy by about a factor of 4.

5.3 Inclusion of ionic strength in free energy calculations (III)

The effect of incorporating explicit ions to mimic ionic strength in free energy calculations using molecular dynamics simulations was investigated. The free energy difference between two triosephosphate isomerase inhibitors (2-phosphoglycolic acid, PGA, and 3-phosphonopropionic acid, 3PP) was calculated at five different concentration of ions (Na^+ and Cl^-) and used as a test case. The transformation from PGA into 3PP is relatively minor, consisting of the mutation of an oxygen atom partially buried in the molecule, into an aliphatic group. There is no change in the net charge of the system.

The acids were neutral at very low pH and deprotonated at neutral pH (-3 e). Both neutral (PGAH, 3PPH) and ionic (PGA, 3PP) forms were investigated. The transformation of the -3 e form in water corresponded to ΔG_3 in Figure 5.

Table 6. Free energy difference (in kJ mol^{-1}) for the transformation of the neutral (PGAH→3PPH) and ionic (PGA→3PP) forms of the inhibitors in water at different ionic strengths (I).

I (M)	PGAH → 3PPH	PGA → 3PP
0.0	-81.2±2.5	-109.6±2.7
0.04	-81.7±2.6	-107.0±9.5
0.06	-82.6±2.6	-112.8±3.8
0.1	-82.3±2.5	-116.4±10.2
2.0	-82.6±2.7	-127.1±7.6

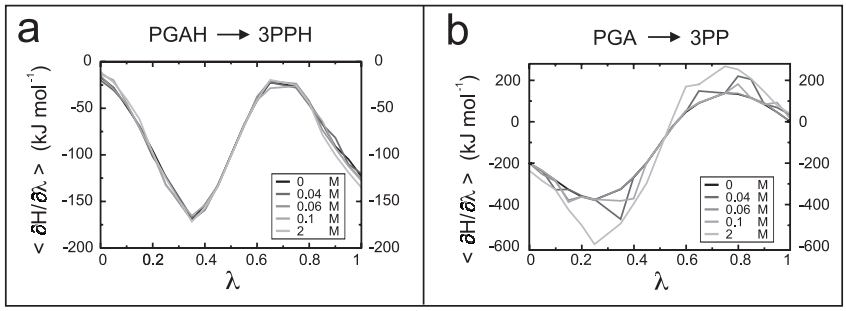

Fig 8. $< \partial H/\partial\lambda >$ (in kJ mol^{-1}) as a function of λ for the transformation of PGAH into 3PPH (a) and PGA into 3PP (b) calculated at ionic strengths 0, 0.04, 0.06, 0.1 and 2 M.

In Table 6 the free energy differences, and in Figure 8 the $< \partial H/\partial\lambda >$ as a function of λ are shown at different ionic strengths (I). For the neutral compounds the free energy differences and the $< \partial H/\partial\lambda >$ profiles did not change significantly over the range of ionic strength considered. The results for the charged species were different. The free energy difference changed up to about 20 kJ mol^{-1} when calculated at different ionic strengths (Table 6). The $< \partial H/\partial\lambda >$ curve at ionic strength 2 M clearly differed

from the one obtained in the absence of ions. Also, the free energy profiles were very noisy when ions were present and the errors in each of the individual $< \partial H/\partial\lambda >$ values were large. When the computation time at each λ point along the transformation was increased from 200 to 10,000 ps it was clear that the presence of ions around the atoms involved in the mutation had a direct effect on the $\partial H/\partial\lambda$. Particularly at low ionic strength, when the number of ions in solution was low, 200 ps was too short to achieve sufficient sampling of ions around the compounds and ions would interact with the inhibitors only occasionally. The distribution of the Na^+ ions around the charged solute was estimated from a 60 ns simulation of PGA in water at ionic strength 0.06 M. There were on average two Na^+ ions interacting with the inhibitors. Individual ions spent on average 40% of simulation time at a distance less than 0.3 nm from the solute and exchanged frequently with the average residence time being about 7 ns. In contrast, the average minimum distance of ions from the neutral inhibitors was about 0.95 nm. Analysis of the contributions to the $\partial H/\partial\lambda$ from specific bonded and non-bonded interactions showed that the presence of a Na^+ ion close to the inhibitor effected the Na^+-inhibitor interaction component of $\partial H/\partial\lambda$ by up to 200 kJ mol^{-1} and the solvent-inhibitor component to a lesser extent (about 50 kJ mol^{-1}).

Calculations were performed using the reaction field method with a 1.4 nm cutoff to model electrostatic interactions. During the simulation, it was possible for an ion to move beyond the cutoff distance. In such case it was effectively not seen anymore by the inhibitor. However, the crossing of cutoff by the ions did not have a significant effect on the $\partial H/\partial\lambda$.

5.4 A strained planar proline in the active site of TIM (IV)

Proline 168 is located in the proximity of the active site of TIM and its side chain adopts a planar conformation in the liganded atomic resolution structure of TIM from *L. mexicana* (PDB code 1N55). Minimization of this structure with various popular force fields, including GROMOS96 (van Gunsteren *et al.* 1996), OPLS (Jorgensen & Tirado-Rives 1988, Jorgensen *et al.* 1996), and AMBER (Wang *et al.* 2004) lead to proline conformations that were not in agreement with the planar conformation observed in the X-ray structure. A more accurate, QM/MM approach was therefore applied to describe such proline and investigate the nature of the planar conformation.

Energies of Pro 168 in the planar and in the down and up puckered conformations were calculated in a vacuum at the QM level, and in the solvated protein environment

using the QM/MM approach (Table 7). The vacuum calculations were performed on a total of 26 atoms, including Pro 168 and adjacent atoms. This QM (sub)system is referred to as QM1. The protein QM calculations were performed on a larger QM subsystem of 129 atoms (QM4), that consisted of QM1 plus a number of nearby residues, while the rest of the system (protein and solvent) was treated at the MM level. This system is referred to as QM4+MM. While in vacuum the planar conformation always had the highest energy, in the protein environment of 1N55, it was the most favourable conformation (Table 7).

Table 7. Relative energies (kJ mol^{-1}) of down, up and planar conformations of Pro 168 in three TIM X-ray structures (1N55, 1NEY and 5TIM) in a vacuum (QM1: 26 atoms) and in the protein (QM4: 129 atoms + MM). The zero energy for each system is taken as the down pucker energy.

PDB	DOWN	UP	PLANAR
Vacuum (QM1)			
1N55	0.0	0.75	13.4
1NEY	0.0	1.2	13.6
5TIM	0.0	1.4	15.6
Protein (QM4 + MM)			
1N55	0.0	-1.0	-10.0
1NEY	0.0	-18.0	-16.0
5TIM	0.0	18.0	29.0

For comparison also energies of Pro 168 in the protein environment of another liganded TIM structure (1NEY) and an unliganded one (5TIM) were computed (Table 7). In the 1NEY structure, the conformation of the down pucker was clearly unfavourable with respect to the up and planar ones, whereas the energies of the planar and up pucker were not significantly different. In the case of the TIM's structure 5TIM, the down pucker was the most favourable conformation. These data are in agreement with the conformations Pro 168 adopts in the correspondent X-ray structures. From a comparison of 12 unliganded and 8 liganded TIM structures with different resolution, it was found that in all unliganded structures Pro 168 is down puckered, while in the liganded ones a near planar conformation of the ring occurs.

In the QM/MM calculations it was necessary to include at least the side chains of residues Glu 129, tyrosine (Tyr) 166 and alanine (Ala) 171, other than Pro 168 (in total 60 atoms, QM3) in order to reach a significant stabilization of the planar conformation with respect to the down and up puckers.

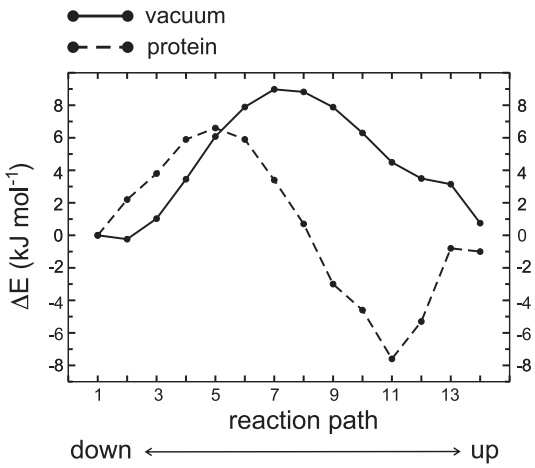

Fig 9. Energy profile for the down→up pucker transition of proline in a vacuum (isolated QM1) and in the protein (QM4 + MM). The zero energy in a vacuum and in the protein is the correspondent down pucker energy.

The energy profile for the down→up pucker transition has been calculated in a vacuum and in the protein environment (Figure 9). In a vacuum, the transition state was the most strained conformation of the intermediate structures between the down and up puckers, *i.e.* the only structure in which more than one dihedral measured less than 10 degrees. The intermediate conformations were then fitted into the protein, and the energy profiles calculated again. From Figure 9 it can be seen that the energy profiles appear very different in a vacuum than in the protein. Because the intermediate conformations of the pyrrolidine ring along the down→up transformation are the same in the protein and in a vacuum, the deviations of the curve depended solely upon the interactions of the proline with its environment. In the protein, the difference in energy between the (vacuum) transition state and the down pucker had decreased to about 6 kJ mol^{-1} and the minimum in energy did not correspond to the down and up puckers, as in a vacuum, but to a structure that was the closest in terms of side chain RMSD to the planar proline of TIM.

QM calculations on the subsystem QM4 (129 atoms) in a vacuum (no protein) were performed and the most important residues for the stabilization of the planar conformation of Pro 168 identified. The isolated QM4 reproduced quite well the QM4 + MM energies, suggesting that the overall stabilization was due to short range interactions only. The conformation of Pro 168 was mostly affected by the side chains of the adjacent residues. In particular, Ala 171, located on the down pucker side of the ring, and Tyr 166 and Glu 129, on the up pucker side of the ring. When the side chain of Ala 171 was left out of the calculations, the down pucker became the most favourable conformation, while leaving out the side chain of Tyr 166 favoured the up pucker. The planar conformation was also less stable when Glu 129 was deleted. This residue does not interact directly with Pro 168, but it is hydrogen bonded to the Tyr 166.

6 Discussion

6.1 Merits and limitations of current free energy calculation methods

6.1.1 Continuum approach (I)

In this study it was of interest to calculate binding affinities of different protein-ligand pairs using the continuum approach. Complexes of a protein with different ligands, of homologous proteins and proteins from different protein families were considered and ligands ranged from a sulphate ion to a whole protein. For each complex, experimental affinities were known and varied between -12 kJ mol^{-1} (4TIM) and -75 kJ mol^{-1} (2PTC). Each contribution in which the binding free energy is decomposed was included in the calculations and both absolute and relative (within a protein family) affinities were considered. The continuum approach performed relatively well for the SH3-peptide complexes, but the agreement with experimental data was in general rather poor.

One major limitation of the method and of empirical calculations in general is the use of parameters that do not always have a clear physical meaning. In particular, it is not easy to assign a value to γ (surface tension) and ε_m (protein dielectric constant) and the intrinsic meaning of these parameters in free energy calculations is still under debate (Juffer & Vogel 2000, Schutz & Warshel 2001). In the calculations carried out for this thesis, the values of γ and ε_m that lead to a somewhat better agreement with experimental values, did not necessarily provide a better description of the systems. In the case of the SH3-peptide complexes, the ranking order was in fact not reproduced with that set of values.

$\gamma\Delta A$, with ΔA the change in accessible surface area upon binding, usually accounts for the hydrophobic effect (or non polar contribution, ΔG_{np}) to the ΔG of binding. In practice this terms includes everything that is not explicitly accounted by other terms such as electrostatics and side chain entropy. The expression $\gamma\Delta A$ is based on hydration free energy data of small molecules (Chothia 1976). However, the transferability to proteins is not always obvious and in order to reach agreement with experiments, the value of γ is often adjusted. Sharp *et al.* (1991) suggested that differences in the sizes

of solute and solvent molecules and the curvature dependence of the hydrophobic effect can be used to define γ in more general terms. Another contribution to ΔG_{np} - that is a function of molecular volume - consists of the van der Waals attraction between solvent occluded interior atoms and solvent. Macromolecules such as proteins differ from smaller molecules in that the interactions of the interior atoms represents a significant favourable contribution to the free energy of transfer, as has been shown by Pitera & van Gusteren (2001). This contribution, that is not accounted for in the continuum model, might explain (some of) the dependence of γ on the particular protein family considered. In this context, γ could be expressed as a function of the molecular volume (Ben-Naim 1978, Juffer *et al.* 1995), rather than the surface.

The protein dielectric constant, ε_m, on the other hand, includes all electrostatic effects that are not explicitly taken into account by the model. These are not only the effect of solvent around the protein, but also reorganization of polar groups and water penetration (Schutz & Warshel 2001). As in the case of γ, these effects are generally dependent on the particular protein considered and the assignment of a set of parameters that are applicable to a wide range of protein-ligand complexes (Schapira *et al.* 1999) is very difficult. Therefore, γ and ε_m are not easily transferable from one protein family to another, considerably reducing the predictive capabilities of such methods.

Many of the continuum approach applications do not take into account conformational changes that occur upon binding (rigid-body approximation). This might explain the positive affinities obtained for the TIM-ligand complexes. Loop 6 of TIM closes the active site upon binding (Wierenga *et al.* 1992) and an important contribution to binding was therefore neglected when this conformational change was not taken into account. Use of different conformations for the ligand-free and ligand-bound state could partially overcome this problem. However, in the calculations considered here the resulting affinities tended to be more disperse when such structures were used. This was in part due to the fact that rather small displacements of atoms can lead to a significant change in the total energy and, in particular, in the van der Waals energy term. This term has been, in fact, ignored in many applications (Vajda *et al.* 1994, Froloff *et al.* 1997, Schapira *et al.* 1999). This approach, of neglecting an explicit van der Waals energy term, has been justified by the fact that the van der Waals energy is proportional to the contact surface area of the complex (correlation up to 0.98) and it would be implicitly included in the hydrophobic term. However, according to the thermodynamic cycle in Figure 3 this is true only for the solute-solvent van der Waals interactions, while the solute-solute van der Waals term should still be included.

When the rigid-body approximation is applied, or even if the structures of the free protein and ligand are used, conformational flexibility is not accounted for in the continuum approach. This contribution can be estimated from the loss of side-chain conformational freedom (ΔS_{sc}), in an empirical manner, or by calculating explicitly the vibrational entropy difference (ΔS_{vib}). For the SH3-peptide complexes, both were estimated (Table 2). Inclusion of this contribution brought the predicted values closer to those found experimentally in almost all cases. Interestingly, the order of magnitude of ΔS_{sc} and ΔS_{vib} was comparable, and in one case (1BBZ) the values almost identical (14.1 and 15.8 kJ mol^{-1}). However, it is not clear if ΔS_{sc} and ΔS_{vib} represent the same contributions. ΔS_{vib} was computed from the normale modes frequencies and these are known to capture harmonic aspects of protein dynamics, while anharmonicity is associated with crossing of higher barriers on the potential energy surface (Hayward *et al.* 1995), such as torsions around bonds. ΔS_{sc}, on the other hand, is estimated from the number of rotamers accessible to the side-chains and is more likely associated with anharmonic modes. Wheter the calculated ΔS_{sc} and ΔS_{vib} are associated with the same modes or instead they are representative of different parts of the conformational space could probably be elucidated by a principal component analysis (Amadei *et al.* 1993), where both harmonic and anharmonic modes can be captured.

Finally, it should be noted that ionic strength was not included in the calculations even though many of the complexes were highly charged. However, it is not very clear how one should satisfy this condition in the continuum approach.

6.1.2 All-atom approach

In the all-atom approach, atomistic simulation techniques are applied to describe the behaviour of the system and derive thermodynamic properties of interest. Typical issues of this approach are the accuracy of the force field and the efficiency of sampling the conformational space of the system. In the following sections, these and other issues will be discussed in the context of the different applications.

Absolute binding free energies (I)

The uncertainty associated with the free energy calculations of the five SH3-peptide and SH2-peptide complexes was, in terms of relative errors, between 7% and 14% (Table 3). This magnitude of error has been reported also in previous calculations of affinities and

solvation free energies (7-10%) (Helms & Wade 1997, Dixit & Chipot 2001, Villa & Mark 2002). However, these should be interpreted only as an estimation of the minimum statistical error. How well the conformational space have been sampled is, in fact, more difficult to assess.

Sampling errors might affect the simulations of the complex, of the protein and of the ligand. SH3 and SH2 do not undergo large conformational changes. However, their ligands - in the complexes considered - were peptides, that are usually very flexible molecules and the sampling of their conformational space might require very long simulations. The C-terminus conformations of the free SH3 peptides in solution (λ=0) were in good agreement with experimental circular dichroism data of proline rich peptides, that adopts a PPII (Williamson 1994) conformation when free in solution. However, the N-terminus was rather unstructured and different conformations were identified, and as for the free SH2 peptide, it was more difficult to validate the simulations and establish how representative they were of the conformational space of the free peptide. Sampling issues also effect the simulations at λ=1, when the peptidic ligands were uncharged Lennard-Jones molecules (SH3) or dummy molecules (SH2) in gas phase. As such, the conformational space to be sampled is even larger. The average structures of the free SH3 peptide in the λ=1 simulations for the decoupling of the peptide from solvent and from the active site of the protein were compared and it was found that both decoupling processes had resulted in essentially the same conformation. However, again, this is insufficient proof of accurate sampling.

In order the transformation is reversible, once the ligand is decoupled from the rest of the system, it should still be able in a (hypothetical) reverse simulation to return to the correct binding mode in the binding site of the protein. Given the size and type of system in this work, this might be an issue. In order to improve the reversibility of the decoupling process, one might apply restraints to the conformation of the ligand as a whole (Hermans & Wang 1997), so that during the decoupling of the peptide from the protein its conformation would not change and the sampling be reduced as well. In a subsequent step the restraints are released. However, in this step the sampling of the conformational space of the free ligand remains an issue. In this work, restraints were applied only to the positions of two atoms (one on the protein and one on the ligand) to prevent the ligand from diffusing out of the binding site during decoupling. In this way, the ligand does not need to sample the whole space available in order to find the binding site on the protein, even though it must still find the correct orientation and conformation in the binding site.

Another issue that should be addressed concerns the fact that the binding sites of SH3 and SH2 are fully exposed to solvent. During the decoupling of the ligand from the active site, water molecules moved into the binding site. It was observed that during such process, the ligand was effectively "pushed out" from the binding site. This was due to the fact that during the transformation, the interactions between the ligand and the rest of the system were slowly switched off, whereas the solvent-protein interactions remained constant. Water molecules would therefore strongly compete with the ligand for the binding site. In the work of Helms & Wade (1998), during decoupling of the ligand in the binding site, water molecules were accordingly coupled. However, in that work the active site was not exposed to solvent and this approach cannot be applied in a straightforward manner in the case of SH3 or SH2 domains.

Enthalpies of binding were calculated from the potential energies of the initial and final systems. Uncertainties in the values of the energy, in particular for large systems, and systematic errors, make computation of these rather inaccurate (Smith & Haymet 1993). The relative errors in the calculated enthalpies were, in fact, rather high for all but the SH3 complexes 1BBZ and 1ABO (Table 3). The entropies of binding were derived as the difference between the free energy and the enthalpy of binding and were positive for all five association reactions. The binding reactions were therefore driven by a large entropy increase in the protein-ligand system, which, in four over five cases, overcame the negative entropy in the surrounding by extracting heat from the rest of the system (Table 3). Experimental data for the binding of SH3 domains with peptide ligands reported negative entropies of binding (Wittekind *et al.* 1994, Renzoni *et al.* 1996). However, when the ligand was a whole protein subunit, rather than a peptide, the measured entropy was positive (Renzoni *et al.* 1996). It was suggested that the loss of conformational freedom of the peptide upon binding could explain the two different thermodynamic measurements. If this is generally true for the binding of SH3, then it is possible that the conformational sampling of the SH3 peptides in the calculations was not sufficient, and there was a contribution missing due to the loss of conformational freedom of the peptide upon binding. This contribution would decrease the calculated free energy, and bring it to a somewhat closer agreement with experimental values.

Affinities of the five complexes were also calculated using the linear interaction energy method (LIE) (Åqvist *et al.* 2002). While for two of the SH3 complexes the calculated affinities were very close to the experimental values, for 1FYN and particularly for the SH2 complexes, the binding free energies were much more negative. Because the charge of these last three systems was much higher than 1BBZ and 1ABO, it seems

likely the predicted affinities were very sensitive to the electrostatic interaction energies. It has already been noted that this method has the disadvantage that for charged systems the interaction energies can be in the order of several thousand and uncertainties in these quantities can be rather large (Åqvist *et al.* 2002). However, even in the case of 1BBZ and 1ABO SH3 complexes, where the computed affinities were close to the experimental values, the ranking order was not reproduced.

In general, description of charged molecules is more difficult than neutral compounds. This is mainly due to the difficulty of treatment of long-range electrostatics. In particular, for the GROMOS96 (van Gunsteren *et al.* 1996) force field applied in the calculations, it was shown that the force field parameters for charged amino acids possibly need further refinement, while the force field performed relatively well for neutral amino acids (Villa & Mark 2002). Therefore, particularly for the case of highly charged compounds, it would be expected that the calculated values are very sensitive to the charge model used.

Relative binding free energies (II)

Relative free energy difference was calculated for the complex of TIM with two of its reaction intermediate inhibitors, PGA and 3PP. TIM undergoes a rather large conformational change at the binding site. Upon ligand binding, loop 6 of TIM "closes" on the active site (the tip of the loop moves by about 0.8 nm) (Davenport *et al.* 1991, Wierenga *et al.* 1992). The application of a thermodynamic cycle for the calculation of relative free energies, in which the enzyme is always bound to the ligands (Figure 5), overcomes the problem of simulating the free protein in solution and therefore of sampling of all the conformational space of the unliganded molecule. In our calculations, conformational sampling was furthermore reduced by the use of position restraints on all heavy atoms of TIM. Because of such restraints, an underlying assumption of these calculations was that the free energies of restraining the active site bound to PGA and 3PP were the same. We could justify this assumption given the similarity of the active site of the enzyme bound to the two ligands for which X-ray structures have been solved (Noble *et al.* 1991, Kursula & Wierenga 2003). Without such restraints it was not possible to maintain the active site geometries observed experimentally in the X-ray structure. It has been previously reported that the conformation of residues in the active site can be strained and molecular mechanics force field might fail to reproduce them correctly (Torrent *et al.* 2002). This is believed to be the case for TIM. In partic-

ular, a strained proline has been identified in the active site (Kursula & Wierenga 2003, Donnini *et al.* 2006), that might (partly) explain the difficulty of simulating the bound enzyme.

From the analysis of the calculated free energy differences of PGA and 3PP in solution and in the binding site of TIM, it appears that the driving force for the preferred binding of PGA to TIM is determined by the solvation properties of the ligands, *i.e.* the difference in the pK of the carboxylic group of the inhibitors (-10.3 kJ mol^{-1}, see Table 5), rather than the active site interactions. Recently, Gloster *et al.* (2007) have reported a study on the inhibition of glycosidase where they underline the importance of solvation and desolvation effects on ligand binding. These two examples provide evidence for the relevance of a rigorous thermodynamic cycle when performing free energy calculations, that takes into account the properties of the ligand in the binding site as well as in solution. Too often, little attention is given to this aspect, while the binding properties in the active site are broadly exploited. This is especially the case in development of new drugs (Jain 2004, Kitchen *et al.* 2004).

Even though the X-ray structures of TIM complexed with the two inhibitors are very similar, there is one important point of discussion that concerns the distance of the carboxylic oxygen of the ligand from the catalytic glutamic acid Glu 167 (according to the numbering of the TIM-PGA X-ray structure 1N55). This distance in the TIM-PGA complex (2.61 Å) suggested the presence of a low barrier hydrogen bond (LBHB) (Cleland *et al.* 1998) rather then a standard hydrogen bond, as in the case of TIM-3PP (2.86 Å). However, the resolution of the TIM-3PP X-ray structure (2.5 Å) is much lower than TIM-PGA (0.83 Å) and it is difficult to draw any conclusions on the nature of this hydrogen bond in the TIM-3PP complex. The Molecular Mechanics (MM) force field applied in the MD calculations did not model the LBHB, but this interaction was treated as a standard hydrogen bond in both complexes. If the LBHB would stabilize the TIM-PGA complex, then a significant contribution would have been omitted from the calculations made in this study. However, the nature of the LBHB interaction as well as its role in the stabilization of reaction intermediates and the magnitude of this contribution are still heavily debated (Cleland *et al.* 1998, Warshel 1998).

One major approximation of the computations concerned the pK calculations. It was not straightforward to assign an error to the calculated pK values. Liptak & Shields (2001) estimated an error of approximately 3 kJ mol^{-1} (about half pK unit) for their pK calculations when a very high level *ab initio* was applied. In these calculations the B3LYP/6-31G* level of theory was used and the absolute pK values deviated greatly

from the experimental ones, while only relative differences were used to estimate pK of the reactions B→C and C→F in Figure 6. pK for reactions B→E and E→F was instead approximated to the second and third experimentally measured pK_a. The effective affinities (based on the estimated concentrations of the binding species of PGA and 3PP) and the difference in the deprotonation energies of PGA and 3PP in Figure 5 (ΔG_{1a}-ΔG_{2a}) were obtained using the calculated pK values. For these calculations it was not possible to assign an error. The numerical values obtained have therefore to be taken as an indication of the tendency of the system to favour one reaction respect with the other, rather than as an absolute quantity.

6.2 Binding affinities: experiment *versus* calculation

6.2.1 Effective *affinities (II)*

The concept of *effective* affinities was introduced to compare computed and experimental affinities, when the latter are obtained from the total concentration of inhibitor, as for example in the case of inhibition constants obtained using competitive inhibition assays. *Effective* affinities are obtained with only the concentration of the inhibitor that actually binds the enzyme, *i.e.* the ionized form of a compound that is recognized by the enzyme. Generally, calculated affinities only consider the binding reaction and do not account for protonation equilibria of the free inhibitors in solution, because these are not so trivial to sample. This is why computed free energies of binding should be compared to *effective* affinities, rather than to the experimental ones. The difference between the *effective* and experimental affinities depends on the ratio of the binding form of the ligand with respect to the total concentration, as can be seen in Eq (49). When this is 1, *effective* affinities coincide with experimental affinities.

PGA and 3PP are two inhibitors of TIM. They have three titrating sites and in solution are present in different ionized forms depending on which site is protonated (Figure 6). At pH 7.6, where the measurement was performed, the binding form of the inhibitors, PGA and 3PP, was only a small fraction of the total inhibitor concentration and *effective* affinities of PGA and 3PP for TIM increased by -14.6 and -6.5 kJ mol^{-1}, respectively, with respect to the experimental values (Table 5). The experimentally measured binding free energies were effectively underestimating the affinity of these inhibitors for TIM. *Effective* affinities can, therefore, be useful for the interpretation of experimental affinities as such and are a necessary means when calculated free energies

are compared to experimental quantities.

6.2.2 Inclusion of ionic strength using explicit ions (III)

In many application of free energy calculations to biomolecular systems (Kollman 1993, Wang *et al.* 2001) the ionic strength has been ignored for reasons of efficiency. It simply takes a very long time to correctly sample ion distribution in solution. However, the chemicophysical environment where the calculations are performed would not be described appropriately and that could lead to erroneous estimates of affinity. Thermodynamic properties of a system such as binding affinity (Lambeir *et al.* 1987, Fedosova *et al.* 2002) and thermal stability (Goto & Fink 1990) are known to be affected by the ionic strength of the environment. Under physiological conditions, where most of the experiments are carried out large numbers of positive and negative ions are always present.

The issue of incorporation of ions in free energy calculations is not only related to the inclusion of ionic strength in the simulation environment. Ions are also included in the system to compensate for the net charge of a solute and maintain, in this way, overall neutrality. However, if counter-ions are included to compensate for the charge of the protein only when computing the affinity of a neutral ligand for a charged protein (Rao *et al.* 1992), inconsistencies in the calculations can occur. In such cases, ions would be present in the simulation of the protein-ligand system, but not in the simulation of the free ligand. It is not clear how this could affect the overall free energy difference within the context of the thermodynamic cycle applied. Moreover, if the total charge of a molecule changes during a transformation, the number of counter-ions must also change in order to maintain neutrality in the system (Dixit & Chipot 2001). As a consequence, one obtains the free energy of a combined process, the association of the ligand and the solvation of the ion (see next Section 6.2.3).

The influence of the distribution of ions around the compound cannot be ignored even for mutations that do not involve a change in the net charge. This effect is most significant if the molecule carries a net charge close to the site mutated, as in the transformation of PGA into 3PP (Table 6). In particular, the precise position of the ions is more important than the number of ions. This implies that when only a small number of counter-ions are included in the system, very long simulations are required to obtain convergence. The calculated free energy will otherwise be heavily dependent on the initial distribution of ions.

Experimentally, binding affinities can be affected by changes in the ionic strength by as much as a few kJ mol^{-1} for a twofold increase of ionic strength (Lambeir *et al.* 1987, Fedosova *et al.* 2002). This falls within the error limits of a free energy calculation. Thus, even if differences in calculated affinities can already be observed when few counter-ions are included in the system (up to 20 kJ mol^{-1}, in this work), the effect, when averaged over all possible distributions is not large. This suggests that in practice to obtain converged results, it is better either to perform the calculations without including counter-ions or to include many ions to ensure sufficient averaging. In the first case one would rely on the response of the solvent to induce an opposing local charge distribution, and in the second case one would have to simulate at high ionic strength.

A calculation at 0 ionic strength is in fact not comparable to a (hypothetical) experiment in the absence of ions. This is because, even though most force fields do not include an explicit ionic strength term, they are parametrized to reproduce experimental data on small molecules over a physiological range of ionic strengths. However, the transferability of the parameters in a force field is implicitly assumed and the models might fail, in particular, in the description of large and highly charged molecules.

Another issue concerns the method to use for the treatment of long range electrostatic interactions. This issue is of relevance here in dealing with ionic solutions where long range electrostatics has a significant effect. In particular the question raised as to whether Ewald summation methods are to be preferred over cutoff together with reaction field correction. It has been previously shown that ion-ion radial distribution functions for solutions of NaCl in water calculated using an Ewald summation or cutoff (1.4 nm) plus reaction field are almost identical (Tironi *et al.* 1995). Results obtained in this study confirmed that the distribution of ions around the charged species was the same when using particle mesh Ewald (PME) or cutoff plus reaction field to describe long range electrostatics. It was also shown that the effect of ions crossing the cutoff and length of cutoff (1.4 or 1.8 nm) did not influence the calculated free energy differences. In addition, it must be noted than when calculating free energies of processes involving creation or deletion of charge with Ewald summation method, problems might raise because lattice sum techniques in general require that the basic unit cell is neutral (Hünenberger & McCammon 1999). In an Ewald summation, a background charge is added to compensate for any net charge in the system. Therefore, during a free energy calculation work will be done against this background charge.

6.2.3 Affinities of charged ligands (I)

During a free energy calculation the system should maintain overall neutrality. If this is not the case then the free energy difference (ΔG) computed contains a contribution coming from the work of creation/annihilation of a charge. This bears little relationship with true reaction conditions, where the ionic strength always compensates for the net charge of a solute. In a similar way, the change of total charge of a certain transformation can be compensated for by adding/removing at the same time one or more ions from the system. For instance, when a charged ligand is decoupled from a solvent, *i.e.* its charges are slowly switched off, a number of ions correspondent to this charge are also decoupled and the total charge of the system remains constant (Figure 4). However, the free energy difference that is computed in this way is the ΔG for the transformation of the ligand plus the ions. If these two contributions to the free energy can be separated, *i.e.* are independent, ΔG of the ligand only can be derived. In fact, the contributions of ligand and ions are always correlated, because they do interact. If correlation is not observed this might be related to the fact that the simulation was not long enough to ensure sufficient sampling of ion distribution. This implies that the ΔG of the ligand and of its counter-ions cannot be separated and the affinity calculated applying the thermodynamic cycle in Figure 4 is the combined affinity of the ligand plus its counter-ions for the given protein. It could be argued that if the contribution of the ions would be the same during the two decoupling processes illustrated in Figure 4 (ΔG_1 and ΔG_2), it would cancel when computing the affinity (ΔG_1-ΔG_2). However, the contributions of ligand and counter-ions to ΔG_1 and ΔG_2 were computed (under the assumption that they are independent), and it was found that the free energy differences of decoupling the counter-ions ($\Delta G_{1,ions}$ and $\Delta G_{2,ions}$) were, in fact, significantly different (Table 4). This can be explained by the fact that the ions interact with the ligand in ΔG_1 and with the protein-ligand complex in ΔG_2. Interestingly, $\Delta G_{2,ions}$ of the SH2 complexes 1SHD and 1LKK were almost identical (7 kJ mol^{-1} difference), suggesting that the interactions of the counter-ions in the 1SHD and 1LKK SH2-ligand systems were comparable. The ligand was, in fact the same in the two complexes and the SH2 domains homologous.

The ΔG_{ligand} and ΔG_{ions} for the decoupling processes in Figure 4 are positive and in the same order of magnitude in all cases (Table 4). This indicates that the system works against the removal of a charge, and that the removal of a charge (ligand or ions) is the larger contributor to the ΔG. This, in turn, suggests that in order to maintain elec-

troneutrality one could compensate for the removal of a charge by adding counter-ions rather than removing them. The contribution made by the addition and the removal of a charge is negative and positive, respectively, and it could be expected that the magnitude of the $\partial H/\partial\lambda$ would decrease significantly if ions would be added, rather than removed, along with the removal of the ligand. In other words $\partial H/\partial\lambda_{ligand}$ and $\partial H/\partial\lambda_{ions}$ would counteract each other. Note that $\partial H/\partial\lambda$ of SH3 complexes are consistently smaller than for the SH2 complexes and accordingly the correspondent absolute errors. The relative errors are, in fact, about the same for SH3 and SH2 complexes, whereas the absolute errors are much higher for SH2. The binding free energy is obtained as the difference between ΔG_1 and ΔG_2 and it is generally one order of magnitude smaller than ΔG_1 and ΔG_2. If the uncertainties in ΔG_1 and ΔG_2 are large than ΔG of binding is most likely to be rather inaccurate. This is why it might prove convenient to compensate for the removal of the charged ligand by adding the same amount of charge to the system. However, it should be taken into account that the contribution of the ions cannot be separated from the calculated free energy difference and that only very long simulations can ensure correct sampling of ion distribution.

Another issue concerns the fact that a charged ligand will most likely affect the protonation state of individual titration sites in a protein. This will have an impact on the overall charge of the system and ultimately affect the affinity. However, current MD simulations use a static molecular charge distribution and the effect of this cannot be taken into account. In fact one should allow for fluctuations in the molecular charge distribution at constant pH (Juffer *et al.* 1997) and this would in turn require a different sampling technique (Baptista *et al.* 1997).

6.3 Occurrence of strained residues in proteins and their significance (IV)

The potential energy surface of the proline shows two minima that correspond to the down and up pucker of the pyrrolidine ring (Ramachandran *et al.* 1970, DeTar & Luthra 1977). However, proline can also occur in a planar conformation, as observed in the liganded atomic resolution structure of TIM (PDB code 1N55). Molecular mechanics force fields are usually parametrized to describe the two minima of the proline (down and up pucker) rather well, but fail to capture the remaining portions of the surface. Using standard MM force fields such as GROMOS96 (van Gunsteren *et al.* 1996),

OPLS (Jorgensen & Tirado-Rives 1988, Jorgensen *et al.* 1996), and AMBER (Wang *et al.* 2004) it was, in fact, not possible to reproduce such a planar conformation. Instead, to describe the features of Pro 168 of TIM, a more accurate QM/MM energy calculation was performed. At least 60 atoms, including the proline and adjacent residues had to be treated at the QM level in order to achieve agreement with the experimental structural data. It was shown that the strain in the proline side chain was induced by the local protein environment and interactions within the protein compensated for the increased internal energy of the proline.

Pro 168 is located near the active site of TIM, at the base of loop 6. This loop acts like a "lid" that in the liganded form of TIM is "closed" on the active site and interacts directly with the ligand, whereas in the unliganded form, it is "open" towards the solvent (Davenport *et al.* 1991, Wierenga *et al.* 1992). In the unliganded conformation, the distance between the ring of Pro 168 and the side chain of Ala 171 (located in loop 6) is sufficiently large for the proline ring to adopt the down puckered conformation. However, upon ligand binding, the loop closes and Ala 171 is then closer to Pro 168, which adopts a planar conformation. This conformational change of proline is observed in all TIM proteins (Pro 168 is a highly conserved residue of TIM (Kursula *et al.* 2004)). In all liganded TIM structures Pro 168 occurs in a planar or near planar conformation, whereas it is down puckered in the unliganded ones. The calculations suggested that the resulting steric repulsion between Pro 168 with nearby residues would be sufficiently high to compensate for the energy that is required for the ring to adopt a planar conformation. Upon loop closure, the proline could store (part of) the energy, which, after completion of the enzymatic conversion, is readily available to open the loop again.

Functional significance of geometric distortion has been previously reported. In particular, Heringa & Argos (1999b) identified many nonrotameric side chains in the proximity of active sites. They suggested that nonrotamericity in protein-ligand complexes was induced by ligand binding and that the increased internal energy might play a role in the formation and release of reaction products. In agreement with such observation, QM/MM calculations on the active site of methane mono-oxygenase and ribonucleotide reductase have shown that the active site conformations were strained and that a protein environment was required for maintaining such strain (Torrent *et al.* 2002). This suggests that the strained configuration stores energy that will be used in later steps of the reaction cycle.

Residues in strained configuration might also represent folding or destabilizing sites (Heringa & Argos 1999a, Rousseau *et al.* 2001, Fuhrmann *et al.* 2004). Strained

residues often occur in regions of tight packing in the protein. The energetic cost of distorting a residue might increase the kinetic barrier to folding, while the increased packing density would likely compensate, in the folded state, for the destabilizing effects.

In order to observe strained residues, the structures need to be solved at high resolution. Therefore, it is not unlikely that the actual number of such conformations is much higher. Because most of current protein force fields fail to capture strain, it might be essential to use a QM/MM approach to refine X-ray structures in which strain is suspected. The use of *ab initio* calculations has been already shown to be an important tool in the refinement of atomic resolution X-ray structures (Schmidt & Lamzin 2002, Schmidt *et al.* 2003).

7 Concluding remarks

In this work the applicability of some of the most common methods for free energy calculations to protein-ligand systems have been investigated, and major limitations highlighted. Continuum and all-atom approaches have been applied to estimate affinities of protein-ligand complexes for which the experimental binding free energies were known.

The main advantage of the continuum method is its computational efficiency. However, it has limited predictive power, which stems from the fact that the calculation parameters cannot be transferred from one protein family to another and the underlying physical model is based upon sometimes unclear approximations. In the calculations carried out in this thesis, the binding free energy was rigorously decomposed according to a thermodynamic cycle, and each component computed. It was of interest to include every contribution to the calculated affinities, rather than fitting the energy function used to describe the system to the known data. The agreement between calculation and experiment ranged from good ($\pm$ 7 kJ mol^{-1}) to very poor (two orders of magnitude difference) within the set of protein-ligand complexes considered. In general, relative affinities of a protein for different ligands are in closer agreement with the experimental values than those of homologous proteins or different protein families. However, the ranking order of the affinities is not necessarily reproduced. In some cases, the agreement within a protein family could be improved by changing the parameters γ and ε_m of the hydrophobic and electrostatic free energy terms, respectively. However, this procedure is possible only when the experimental affinities are known. The inclusion of a volume dependency of γ could possibly account for the van der Waals attraction between solvent occluded atoms and solvent, that was not considered in the approach applied here.

The all-atom approach has the advantage that no initial assumptions are required (except that all interactions are pair-wise) and a much more detailed level of information is obtained. However, it clearly suffers from sampling issues and associated lengthy calculations. In particular, calculations of absolute free energies of binding for complexes involving large conformational changes cannot be considered in either of the two approaches. Relative affinities of two or more ligands for the same protein, on the other hand, are relatively easy to estimate. This is because contributions to the free energy

from the free protein cancel if the difference of affinities is considered. However, it is important to apply a rigorous thermodynamic cycle that accounts for the properties of the protein-ligand complex as well as of the free ligands in solution. Using an all-atom approach for the calculation of relative affinities, it was shown that the driving force for the preferred binding of a strong inhibitor of TIM (PGA) over its phosphonate analogous (3PP) resides in a difference in the pK_a values of the inhibitors, rather than in interactions within the active site of the enzyme.

To our knowledge this was the first time that an all-atom approach was used to predict affinities of peptidic ligands for a certain protein. With the exception of one SH3 complex (1FYN), the ranking order of the binding free energies of five protein-peptide pairs was reproduced. The two SH2-peptide pairs considered had almost identical experimental affinities (-29.7 kJ mol^{-1} and -39.7 kJ mol^{-1}) and accordingly the calculated values were not significantly different. However, the calculated affinities overestimated the experimental values by more than one order of magnitude. Because of the size and flexibility of a peptide, the conformational space to be sampled is rather large. If the loss of conformational freedom of the peptide is not correctly taken into account, an important contribution to the calculated affinities will then be missing. To elucidate this aspect it might prove useful to lengthen one of the peptide simulations. The overall charge of one of the SH3-peptide pairs (1FYN) differed greatly from the others (-7 *eversus* -1 *e*). To neutralize these charges a corresponding number of counter-ions was included in the calculations. It could be of interest to change the electrostatic conditions of the simulations - such as the ionic strength - to check if that could influence the results.

When calculating affinities of complexes with charged ligands and ligands containing titrating sites, the question of which computed affinity should be compared with experimental data is raised. Two of the five complexes employed for the all-atom approach contained a charged ligand. In these cases (counter) ions are included in the calculations to maintain overall neutrality of the system. It was shown here that it may prove impossible to separate the contribution of the ions to the affinity from the other terms, so that effectively a combined affinity (ligand plus ions) is computed, instead of that of the ligand only. Even for mutations that do not involve a change in the net charge the influence of the distribution of the ions around the compound cannot be ignored. This raises a general problem associated with the inclusion of explicit ions, either to mimic ionic strength or to maintain neutrality, in free energy calculations. It was shown that convergence can best be achieved by either incorporating no counter-ions or

by simulating at high ionic strength to ensure sufficient sampling of the ion distribution.

Ligands that contain titrating sites are present in solution as a mixture of different species depending on which site is protonated. The computational cost of taking into account all the protonation equilibria of the ligands in solution would be prohibitive. The approach that was employed here was to estimate the affinity of the binding species of the ligand only. This was referred to as *effective* affinity and can be derived from the experimental affinity on the basis of the concentration of the binding form in solution. *Effective* affinity provides a means by which to compare calculated and experimental binding free energies and by which two or more ligands can be studied in terms of intrinsic binding properties.

An underlying limitation of all models to describe molecular systems resides in the accuracy of the description. Molecular mechanics force fields commonly employed for simulations of protein systems make use of parameters derived from analysis of the properties of short peptides and amino acid analogues, and transferability to proteins is implicitly assumed. In the protein environment, however, strained conformations might occur which exhibit a functional role. If this is the case, force field models fail to correctly describe such a feature. During this work such cases have been encountered, where to correctly represent the planar conformation of a proline side chain in the active site of triosephosphate isomerase, a more accurate hybrid QM/MM model had to be employed. The use of a QM/MM approach was therefore essential to describe the active site of the enzyme and might be found to be an important tool in all cases in which strain is suspected.

To overcome some of the issues raised in this work, such as the occurrence of conformational changes upon binding, treatment of charged molecules or ionic strength, it seems desirable that some sort of compromise between an all-atom and a continuum approach is found. It is envisaged that a future model for ligand affinity calculations should retain a level of detail for the solvent, while at the same time, the protein and the peptide should be described at a less detailed ('mesoscopic') level. Many alternative models based on MD employ a full continuum model for the solvent, but retain an all-atom level of description for the protein, which is a combination of two extremes. In addition, the new model should account for pH effects to handle charged ligands more appropriately.

Bibliography

Abagyan R, Totrov M & Kuznetscov D (1994) ICM - a new method for protein modeling and design. Applications to docking and structure prediction from the distorted native conformation. J Comp Chem 15: 488–506.

Agarwal PK (2005) Role of protein dynamics in reaction rate enhancement by enzymes. J Am Chem Soc 127: 15248–15256.

Albery WJ & Knowles JR (1976) Free energy profile of the reaction catalyzed by triosephosphate isomerase. Biochemistry 15: 5627–5631.

Allen MP & Tildesley DJ (1987) Computer Simulation of Liquids. Oxford Science Publication, Oxford.

Amadei A, Linssen ABM & Berendsen HJC (1993) Essential dynamics of proteins. Proteins: Struct Funct Gen 17: 412–425.

Atkins PW (1998) Physical chemistry. Oxford University Press, Oxford Melbourne Tokyo.

Baptista AM, Martel PJ & Petersen, SB (1997) Simulation of protein conformational freedom as a function of pH: constant-pH molecular dynamics using implicit titration. Proteins: Struct Funct Bioinf 27: 532–544.

Bash PA, Singh CU, Brown FK, Landridge R & Kollman PA (1987) Calculation of relative change in binding free energy of a protein inhibitor complex. Science 235: 574–576.

Ben-Naim A (1978) Standard thermodynamics of transfer. Use and misuses. J Phys Chem 82: 792–803.

Benzon SW (1976) Thermochemical kinetics. Wiley, New York.

Berendsen HJC, Postma JPM, van Gunsteren WF & Hermans J (1981) Interaction models for water in relation to protein hydratation. In: Intermolecular Forces, Pullman, B. ed, Reidel, Dordrecht: 331–342.

Berendsen HJC, Postma JPM, van Gunsteren WF & Nola AD (1984) Molecular dynamics with coupling to an external bath. J Chem Phys 81: 3684–3690.

Berendsen HJC, van der Spoel D & van Drunen R (1995) GROMACS: a message-passing parallel molecular dynamics implementation. Comp Phys Commun 91: 43–56.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat T, Weissig H, Shindyalov I & Bourne P (2000) The protein data bank. Nucleic Acids Res 28: 235–242.

Beutler TC, Mark AE, van Schaik RC, Geber PR & van Gunsteren, WF (1994) Avoiding singularities and numerical instabilities in free energy calculations based on molecular simulations. Chem Phys Lett 222: 529–539.

Beveridge DL & DiCapua FM (1989) Free energy via molecular simulation: applications to chemical and biomolecular systems. Annu Rev Biophys Biophys Chem 18: 431–492.

Bishop M & Frinks S (1987) Error analysis in computer simulations. J Chem Phys 87: 3675–3676.

Born M (1920) Volumen und hydratationswärme der ionen. Zeitschrift für physik 1: 45–48.

Born M & Oppenheimer R (1927) Zur quantumtheorie der molekeln. Ann Physik Leipzig 84: 457–484.

Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S & Karplus M (1983) CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. J Comp Chem 4: 187–217.

Chau PL (2001) Process and thermodynamics of ligand-receptor interaction studied using a novel simulation method. Chem Phys Lett 334: 343–351.

Chipot C & Pearlman DA (2002) Free energy calcualtions. The long and winding gilded road. Mol Sim 28: 1–12.

Chothia C (1976) Hydrophobic bonding and accessible surface area in proteins. Nature 248: 338–339.

Cleland WW, Frey PA & Gerlt JA (1998) The low barrier hydrogen bond in enzymatic catalysis. J Biol Chem 273: 25529–25532.

Constanciel R & Contreras R (1984) Self-consistent field-theory of solvent effects representation by continuum models - introduction of desolvation contribution. Theoretica Chimica Acta 65: 1–11.

Creighton TE (1993) Proteins. Structure and molecular properties. Freeman and company, New York.

Daura X, Hünenberger PH, Mark AE, Querol E, Aviles FX & Gunsteren WF (1996) Free energies of transfer of Trp analogs from Chloroform to water: comparison of theory and experiment and the importance of adequate treatment of electrostatic and internal interactions. J Am Chem Soc 118: 6285–6294.

Davenport RC, Bash PA, Seaton BA, Karplus M, Petsko GA & Ringe D (1991) Structure of triosephospate isomerase-phosphoglycolohydroxamate complex: an analogue of the intermediate on the reaction pathway. Biochemistry 30: 5821–5826.

de Groot BL & Grubmüller H (2001) Water permeation across biological membranes: mechanism and dynamics of aquaporin-1 and GlpF. Science 294: 2353–2357.

DeTar DF & Luthra NP (1977) Conformations of proline. J Am Chem Soc 99: 1232–1244.

Dixit SB & Chipot C (2001) Can absolute free energies of association be estimated from molecular mechanical simulations? The biotin-streptavidin system revisited. J Phys Chem A 105: 9795–9799.

Donnini S & Juffer AH (2004) Calculation of affinities of peptides for proteins. J Comp Chem 25: 393–411.

Donnini S, Groenhof G, Wierenga RK & Juffer AH (2006) The planar conformation of a strained proline ring: a QM/MM study. Proteins: Struct Funct Bioinf 64: 700–710.

Eisenberg D & McLachlan AD (1986) Solvation energy in protein folding and binding. Nature 319: 199–203.

Fedosova NU, Champeil P & Esmann M (2002) Nucleotide binding to Na,K-ATPase: the role of electrostatic interactions. Biochemistry 41: 1267–1273.

Fersht A (1977) Enzyme structure and mechanism. WH Freeman and company, New York.

Fisher E (1894) Einfluss der configuration auf die wirkung der enzyme. Ber Dtsch Chem Ges 27: 2985.

Flyvbjerg H & Petersen HG (1989) Error estimates on averages of correlated data. J Chem Phys 91: 461–466.

Frenkel D. & Smith B. (1996) Understanding Molecular Simualations. Academica Press, New York.

Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, Montgomery Jr JA, Vreven T, Kudin KN, Burant JC, Millam JM, Iyengar SS, Tomasi J, Barone V, Mennucci B, Cossi M, Scalmani G, Rega N, Petersson GA, Nakatsuji H, Hada M, Ehara M, Toyota K, Fukuda R, Hasegawa J, Ishida M, Nakajima T, Honda Y, Kitao O, Nakai H, Klene M, Li X, Knox JE, Hratchian HP, Cross JB, Bakken V, Adamo C, Jaramillo J, Gomperts R, Stratmann

RE, Yazyev O, Austin AJ, Cammi R, Pomelli C, Ochterski JW, Ayala PY, Morokuma K, Voth GA, Salvador P, Dannenberg JJ, Zakrzewski VG, Dapprich S, Daniels AD, Strain MC, Farkas O, Malick DK, Rabuck AD, Raghavachari K, Foresman JB, Ortiz JV, Cui Q, Baboul AG, Clifford S, Cioslowski J, Stefanov BB, Liu G, Liashenko A, Piskorz P, Komaromi I, Martin RL, Fox DJ, Keith T, Al-Laham MA, Peng CY, Nanayakkara A, Challacombe M, Gill PMW, Johnson B, Chen W, Wong MW, Gonzalez C & Pople JA (2004) Gaussian 03, Revision C.02. Gaussian, Inc., Wallingford, CT, 2004.

Froloff N, Windemuth A & Honig B (1997) On the calculation of binding free energies using continuum methods: application to MHC class i protein-peptide interactions. Prot Sci 6: 1293–1301.

Fuhrmann CN, Kelch BA, Ota N & Agard DA (2004) The 0.83 å resolution crystal structure of α-lytic protease reveals the detailed structure of the active site and identifies a source of conformational strain. J Mol Biol 338: 999–1013.

Gilson MK & Honig B (1988) Calcualtion of the total electrostatic energy of a macromolecular system: solvation energies, binding energies and conformational analysis. Proteins: Struct Funct Gen 4: 7–18.

Gilson MK, Given JA, Bush BL & McCammon JA (1997) The statistical-thermodynamic basis for computation of binding affinities: a critical review. Biophys J 72: 1047–1069.

Gloster TM, Meloncelli P, Stick RV, Zechel D, Vasella A & Davies GJ (2007) Glycosidase inhibition: an assessment of the binding of 18 putative transition-state mimics. J Am Chem Soc 129: 2345–2354.

Goto Y & Fink AL (1990) Phase diagram for acidic conformational states of apomyoglobin. J Mol Biol 214: 803–805.

Groenhof G, Bouxin-Cademartory M, Hess B, de Visser S, Berendsen HJC, Olivucci M, Mark A & Robb M (2004) Photoactivation of the photoactive yellow protein: why photon absorption triggers a trans-to-cis isomerization of the chromophore in the protein. J Am Chem Soc 126: 4228–4233.

Guarnieri F & Still WC (1994) A rapidly convergent simulation method: Mixed Monte-carlo/stochastic dynamics. J Comp Chem 15: 1302–1310.

Halgren TA (1995) Merck Molecular Force Field. i.-v. J Comp Chem 17: 490–641.

Hansson T, Nordlund P & Åqvist J (1997) Energetics and nucleophile activation in a protein tyrosine phosphatase. J Mol Biol 265: 118–127.

Hartman FC, LaMuraglia GM, Tomozawa Y & Wolfenden R (1975) The influence of pH on the interaction of inhibitors with triosephosphate isomarase and determination of the pKa of the active-site carboxyl group. Biochemistry 14: 5274–5279.

Hayward S, Kitao A & Go N (1995) Harmonicity and anharmonicity in protein dynamics: a normal mode analysis and principal component analysis. Proteins: Struct Funct Gen 23: 177–186.

He S & Scheraga HA (1998a) Brownian dynamics simulations of protein folding. J Chem Phys 108: 287–300.

He S & Scheraga HA (1998b) Macromolecular conformational dynamics in torsional angle space. J Chem Phys 108: 271–286.

Helfand E (1984) Dynamics of conformational transition in polymers. Science 226: 647–650.

Helms V & Wade RC (1997) Free energies of hydratation from thermodynamics integration: comparison of molecular mechanics force fields and evaluation of calculation accuracy. J Comp Chem 18: 449–462.

Helms V & Wade RC (1998) Computational alchemy to calculate absolute protein-ligand binding free energy. J Am Chem Soc 120: 2710–2713.
Heringa J & Argos P (1999a) Strain in protein structures as viewed through nonrotameric side chains: I. Their position and interaction. Proteins: Struct Funct Bioinf 37: 30–43.
Heringa J & Argos P (1999b) Strain in protein structures as viewed through nonrotameric side chains: II. Effects upon ligand binding. Proteins: Struct Funct Bioinf 37: 44–55.
Hermans J & Wang L (1997) Inclusion of loss of translational and rotational freedom in theoretical estimates of free energy of binding. application to a complex of benzene and mutant T4 lysozyme. J Am Chem Soc 119: 2707–2714.
Hess B (2002) Determining the shear viscosity of model liquids from molecular dynamics simulations. J Chem Phys 116: 209–217.
Hess B, Bekker H, Berendsen HJC & Fraaije JGEM (1997) LINCS: a linear constraint solver for molecular simulations. J Comp Chem 18: 1463–1472.
Heubel P-HC & Popov AI (1979) Acid properties of some phosphonocarboxylic acids. J Sol Chem 8: 615–625.
Hill TL (1986) An Introduction to Statistical Thermodynamics. Dover Publications, Inc, New York.
Hill TL (1987) Statistical Thermodynamics. Dover Publications, Inc, New York.
Hinz HJ (1983) Thermodynamics of protein-ligand interactions: calorimetric approaches. Annu Rev Biophys Bioeng 12: 285–317.
Horton N & Lewis M (1992) Calculation of the free energy of association for protein complexes. Prot Sci 1: 169–181.
Hünenberger PH & McCammon JA (1999) Ewald artifacts in computer simlations of ionic solvation and ion-ion interaction: a continuum electrostatics study. J Chem Phys 110: 1856–1872.
Ippolito JA, Alexander RS & Christianson DW (1990) Hydrogen bond stereochemistry in protein structure and fucntion. J Mol Biol 215: 457–471.
Israelachvili JN (1973) Van der waals forces in biological systems. Quart Rev Biophys 6: 341–387.
Jain AN (2004) Virtual screening in lead discovery and optimization. Curr Opin Drug Discov Devel 7: 396–403.
Janin J (1995) Protein-protein recognition. Prog Biophys Molec Biol 64: 145–166.
Janin J & Chothia C (1990) The structure of protein-protein recognition sites. J Mol Biol 265: 16027–16030.
Jensen F (1999) Introduction to computational chemistry. J Wiley & Sons Ltd, New York.
Jorgensen WL & Buckner JK (1987) Use of statistical perturbation theory for computing solvent effects on molecular conformation. Butane in water. J Phys Chem 91: 6083–6085.
Jorgensen WL & Ravimohan C (1985) Monte Carlo simulation of differences in free energies of hydration. J Chem Phys 83: 3050–3054.
Jorgensen WL & Tirado-Rives J (1988) The OPLS force field for proteins. Energy minimizations for crystalsof cyclic peptides and crambin. J Am Chem Soc 110: 1657–1666.
Jorgensen WL, Buckner K, Boudon S & Tirado-Rives J (1988) Efficient computation of absolute free energies of binding by computer simulations. Application to the methane dimer in water. J Chem Phys 89: 3742–3746.
Jorgensen WL, Maxwell S & Tirado-Rives J (1996) Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. J Am Chem Soc 118: 11225–11236.

Juffer AH & Vogel H (2000) pK_a calculations of calbindin D_{9k}: effects of Ca^{2+} binding, protein dielectric constant, and ionic strength. Proteins: Struct Funct Bioinf 41: 554–567.

Juffer AH, Eisenhaber F, Hubbard SJ, Walther D & Argos P (1995) Comparison of atom solvation parametric sets: applicability and limitations in protein folding and binding. Prot Sci 4: 2499–2509.

Juffer AH, Argos P & Vogel H (1997) Calculating acid-dissociation constants of proteins using the boundary element method. J Phys Chem B 101: 7664–7673.

Kauzman W (1959) Some factors in the interpretation of protein denaturation. Adv Prot Chem 14: 1–63.

King PM (1993) Free energy via molecular simulation: a primer. In: van Gunsteren W, Weiner PK & Wilkinson A (eds.), Computer Simulation of Biomolecular Systems, Escom, Leiden, vol 2, chap. 12.

Kirkpatrick S, Gelatt Jr CD & Vecchi MP (1983) Optimization by simulated annealing. Science 220: 671–680.

Kirkwood JG (1935) Statistical mechanics of fluid mixtures. J Chem Phys 3: 300–313.

Kitchen DB, Decornez H, Furr JR & Bajorath J (2004) Docking and scoring in virtual screening for drug discovery: methods and applications. Nat Rev Drug Discov 3: 935–949.

Klapper I, Hagstrom R, Fine R, Sharp K & Honig B (1986) Focusing of electric fields in the active site of Cu-Zn superoxide dismutase: effects of ionic strength and amino-acid modification. Proteins: Struct Funct Gen 1: 47–59.

Kollman PA (1993) Free energy calculations: applications to chemical and biochemical phenomena. Chem Rev 93: 2395–2417.

Kollmann PA & Merz KM (1990) Computer modeling of the interactions of complex molecules. Acc Chem Res 23: 246–252.

Koppenol WH & Margoliash E (1982) The asymmetric distribution of charges on the surface of horse cytocrome c. Functional implications. J Biol Chem 257: 4426–4437.

Koshland DE (1958) Application of a theory of enzyme specificity to protein synthesis. Proc Natl Acad Sci USA 44: 98–104.

Kreyszig E (1999) Advanced Engineering MAthematics. John Wiley & Sons, Inc.

Kuhn B, Gerber P, Schulz-Gasch T & Stahl M (2005) Validation and use of the MM-PBSA approach for drug discovery. J Med Chem 48: 4040–4048.

Kursula I & Wierenga RK (2003) Crystal structure of triosephosphate isomerase complexed with 2-phosphoglycolate at 0.83-Å resolution. J Biol Chem 278: 9544–9551.

Kursula I, Salin M, Sun J, Norledge BV, Haapalainen AM, Sampson NS & Wierenga RK (2004) Understanding protein lids: structural analysis of active hinge mutants in triosephosphate isomerase. PEDS 17: 375–382.

Lambeir A-M, Opperdoes FR & Wierenga RK (1987) Kinetic-properties of triose-phosphate isomerase from trypanosoma-brucei-brucei: a comparison with the rabbit muscle and yeast enzyme. Biochemistry 168: 69–74.

Leach AR (2001) Molecular Modelling Principles and Applications. Pearson Education EMA.

Lesk AM & Chothia C (1984) Mechanisms of domain closure in proteins. J Mol Biol 174: 175–191.

Leslie AGW & Wonacott AJ (1984) Structural evidence for ligand-induced sequential conformational changes in glyceraldehyde 3-phosphate dehydrogenase. J Mol Biol 178: 743–772.

Lindahl E, Hess B & van der Spoel D (2001) GROMACS 3.0: A package for molecular simulation and trajectory analysis. J Mol Model 7: 306–317.

Liptak MD & Shields GC (2001) Accurate pKa calculations for carboxylic acids using complete basis set and Gaussian-n models combined with CPCM continuum solvation methods. J Am Chem Soc 123: 7314–7319.

Lodish H, Berk A, Zipursky SL, Matsudaira P, Baltimore D & Darnell J (1999) Molecular cell biology. Freeman and company, New York.

Lybrand TP, McCammon JA & Wipff G (1986) Theoretical calculation of relative binding affinity in host-guest systems. Proc Natl Acad Sci USA 83: 833–835.

Ma C, Baker NA, Joseph S & McCammon JA (2002) Binding of aminoglycoside antibiotics to the small ribosomal subunit: a continuum electrostatics investigation. J Am Chem Soc 124: 1438–1442.

Marrink S-J & Mark AE (2003) Molecular dynamics simulation of the formation, structure, and dynamics of small phospholipid vesicles. J Am Chem Soc 125: 15233–15242.

Maseras F & Morokuma K (1995) IMOMM: a new integrated *ab initio* + molecular mechanics geometry optimization scheme of equilibrium structures and transition states. J Comp Chem 16: 1170–1179.

Matthew JB (1985) Electrostatic effects in proteins. Ann Rev Biophys Bioeng 14: 387–417.

Matthews BW (1977) In: Neurath H & Hill RL (eds.), The Proteins, Academic Press, New York, vol. 3: 403–590.

McCammon JA (1998) Theory of biomolecular recognition. Curr Opin Struct Biol 8: 245–249.

McCammon JA, Gelin BR & Karplus M (1977) Dynamics of folded proteins. Nature 267: 585–590.

McQuarrie DA (1976) Statistical Mechanics. Harper and Row, New York.

Mennucci B & Tomasi J (1997) Continuum solvation models. A new approach to the problem of solute's distribution and cavity boundaries. J Chem Phys 106: 5151–5158.

Merz KMJ & Kollman PA (1989) Free energy perturbation simulations of the inhibition of thermolysin: prediction of the free energy of binding of a new inhibitor. J Am Chem Soc 111: 5649–5658.

Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH & Teller E (1953) Equation of state calculations by fast computing machines. J Phys Chem 21: 1087–1092.

Miertus S, Scrocco E & Tomasi J (1981) Electrostatic interaction of a solute with a continuum. A direct utilization of *ab initio* molecular potentials for a prevision of solvent effects. Chem Phys 55: 117–129.

Miyamoto S & Kollman PA (1992) SETTLE: an analytical version of the SHAKE and RATTLE algorithms for rigid water models. J Comp Chem 13: 952–962.

Miyamoto S & Kollman PA (1993) Absolute and relative binding free energy calculations of the interaction of biotin and its analogs to streptavidin using molecular dynamics/free energy perturbation approches. Proteins: Struct Funct Gen 16: 226–245.

Mordasini TZ & McCammon JA (2000) Calculations of relative hydration free energies: a comparative study using thermodynamic integration and extrapolation method based on a single reference state. J Phys Chem B 104: 360–367.

Muller N (1990) Search for a realistic view of hydrophobic effects. Acc Chem Res 23: 23–28.

Murzin A, Brenner S, Hubbard T & Chothia C (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 247: 536–540.

Nemethy G, Gibson KD, Palmer KA, Yoon CN, Paterlini G, Zagari A, Rumsey S & Scheraga HA (1992) Energy parameters in polypeptides. 10. Improved geometric parameters and nonbonded interactions for use in the ECEPP/3 algorithm, with application to proline-containing

peptides. J Phys Chem 96: 6472–6484.
Noble MEM, Wierenga RK, Lambeir A-M, Opperdoes FR, Thunnissen A-MWH, Kalk KH, Groendijk H & Hol WGJ (1991) The adaptability of the active site of trypanosomal triosephosphate isomerase as observed in the crystal structures of three different complexes. Proteins: Struct Funct Gen 10: 50–69.
Nola AD & Brünger A (1998) Free energy calculations in globular proteins: methods to reduce errors. J Comp Chem 19: 1229–1240.
Norledge BV, Lambeir A-M, Abagyan RA, Rottmann A, Fernandez AM, Filimonov VV, Peter MG & Wierenga RK (2001) Modeling, mutagenesis, and structural studies on the fully conserved phosphate-binding loop (loop 8) of triosephosphate isomerase: towards a new substrate specificity. Proteins: Struct Funct Bioinf 42: 383–389.
Noskov SY & Lim C (2001) Free energy decomposition of protein-protein interactions. Biophys J 81: 737–750.
Onsanger L (1936) Electric moments of molecules in liquids. J Am Chem Soc 58: 1486–1493.
Ooi T, Oobatake M, Nemethy G & Scheraga HA (1987) Accessible surface areas as a measure of the thermodynamic parameters of hydration of peptides. Proc Natl Acad Sci USA 84: 3086–3090.
Oostenbrink C & van Gunsteren WF (2004) Free energies of binding of polychlorinated biphenyls to estrogen receptor from a single simulation. Proteins: Struct Funct Bioinf 54: 237–246.
Orengo C & Thornton J (2005) Protein families and their evolution-a structural perspective. Annu Rev Biochem 74: 867–900.
Orengo C, Michie A, Jones S, Jones D, Swindells M & Thornton J (1997) CATH- a hierarchic classification of protein domain structures. Structure 5: 1093–1108.
Pearlman DA (1994) Free energy derivatives: a new method for probing the convergence problem in free energy calculations. J Comp Chem 15: 105–123.
Peng C & Schlegel H (1994) Combining synchronous transit and quasi-newton methods for finding transition state. Isr J Chem 33: 449–454.
Perutz MF (1989) Mechanism of cooperativity and allosteric regulation in proteins. Quart Rev Biophys 22: 139–236.
Pickett SD & Sternberg MJE. (1993) Empirical scale of side-chain conformational entropy in protein folding. J Mol Biol 231: 825–839.
Pieffet G (2005) The application of molecular dynamcis simulation techniques and free energy calculations to predict protein-protein and protein-ligand interactions. Ph.D. Thesis, Groningen Biomolecular Sciences and Biotechnology Institute.
Pimentel GC & McLellan AL (1971) Hydrogen bonding. Ann Rev Phys Chem 22: 347–385.
Pitera JW & van Gusteren WF (2001) The importance of solute-solvent van der Waals interactions with interior atoms of biopolymers. J Am Chem Soc 123: 3163–3164.
Postma JPM, Berendsen HJC. & Haak JR. (1982) Thermodynamics of cavity formation in water. Faraday symposium of the chemical society 94:1725–1733.
Privalov PL & Gill SJ (1988) Stability of protein structure and hydrophobic interaction. Adv Protein Chem 39: 191–234.
Privalov PL & Gill SJ (1989) The hydrophobic effect: a reappraisal. Pure Appl Chem 61: 1097–1104.
Ramachandran GN, Lakshminarayanan AV, Balasubramanian R & Tegoni G (1970) Studies on the conformation of amino acids, XII: Energy calculations on propyl residue. Biochim Biophys Acta 221: 165–181.

Rao BG, Tilton RF & Singh UC (1992) Free energy perturbation studies on inhibitor binding to HIV-1 proteinase. J Am Chem Soc 114: 4447–4452.

Reinhardt WP, Miller MA & Amon LY (2001) Why is it so difficult to simulate entropies, free energies, and their differences? Acc Chem Res 34: 607–614.

Renzoni DA, Rugh DJR, Siligardi G, Das P, Morton CJ, Rossi C, Waterfiel MD, Campbel ID & Ladbury JE (1996) Structural and thermodinamic characterization of the interaction of the SH3 domain from fyn with the proline-rich binding site on the p85 subunit of PI3-Kinase. Biochemistry 35: 1546–15653.

Rousseau F, Schymkowitz JWH, Wilkinson HR & Itzhaki LS (2001) Three-dimentional domain swapping in p13suc1 occurs in the unfolded state and is controlled by conserved proline residues. Proc Natl Acad Sci USA 98: 5596–5601.

Schapira M, Totrov M & Abagyan R (1999) Prediction of the binding energy for small molecules, peptides and proteins. J Mol Recognit 12: 177–190.

Schmidt A & Lamzin VS (2002) Veni, vidi, vici - atomic resolution unravelling the mysteries of protein function. Curr Opin Struct Biol 12: 698–703.

Schmidt A, Jelsch C, Ostergaard P, Rypniewski W & Lamzin VS (2003) Trypsin revisited: crystallography at (sub) atomic resolution and quantum chemistry revealing details of catalysis. J Biol Chem 278: 43357–43362.

Schutz CN & Warshel A (2001) What are the dielectric "constants" of proteins and how to validate electrostatic models? Proteins: Struct Funct Bioinf 44: 400–417.

Sharp KA & Honig B (1990) Electrostatic interactions in macromolecules: theory and applications. Ann Rev Biophys Biophys Chem 19: 301–332.

Sharp KA, Nicholls A, Fine RF & Honig B (1991) Reconciling the magnitude of the microscopic and macroscopic hydrophobic effects. Science 252: 106–109.

Shirts MR, Pitera JW, Swope WC & Pande VS (2003) Extremely precise free energy calculations of amino acid side chain analogs: comparison of common molecular mechanics force fields for proteins. J Chem Phys 119: 5740–5761.

Simonson T & Brünger A (1992) Thermodynamics of protein-peptide interactions in the ribonuclease-s system studied by molecular dynamics and free energy calculations. Biochemistry 31: 8661–8674.

Simonson T, Archontis G & Karplus M (2002) Free energy simulations come of age: protein-ligand recognition. Acc Chem Res 35: 430–437.

Smith DE & Haymet ADJ (1993) Free energy, entropy, and internal energy of hydrophobic interactions: computer simulations. J Chem Phys 98: 6445–6454.

Smith G (2002) http://www.gromacs.org/contributions ffG43a1p: uploaded 16:08 may 15, 2002 by Graham Smith .

Smith PE & Pettitt BM (1994) Modeling solvent in biomolecular systems. J Phys Chem 98: 9700–9711.

Smith PE & van Gunsteren WF (1994) When are free energy components meaningful? J Phys Chem 98: 13735–13740.

Straatsma TP (1996) Free energy by molecular simulation. In: Lipkowitz KB & Boyd DB (eds.), Reviews in Computational Chemistry, VCH Publishers, New York, vol 9: 81–127.

Straatsma TP & McCammon J (1991) Multiconfiguration thermodynamic integration. J Chem Phys 95: 1175–1188.

Straatsma TP, Berendsen HJC & Stam SJ (1986) Estimation of statistical errors in molecular simulation calculation. Mol Phys 57: 89–95.

Sugita Y & Okamoto Y (1999) Replica-exchange molecular dynamcis methods for protien folding. Chem Phys Lett 314: 141–151.

Sugita Y, Kitao A & Okamoto Y (2000) Multidimensional replica-exchange method for free enrgy calculation. J Chem Phys 314: 6042–6051.

Svensson M, Humbel S, Froese RDJ, Matsubara T, Sieber S & Morokuma K (1996) ONIOM: a multilayered integrated MO + MM method for geometry optimizations and single point energy predictions. A test for Diels-Alder reactions and Pt(P(t-Bu)3)2 + H2 oxidative addition. J Phys Chem 100: 19357–19363.

Swanson JM, Henchman RH & McCammon JA (2004) Revisiting free energy calculations: a theoretical connection to MM/PBSA and direct calculation of the association free energy. Biophys J 86: 67–74.

Talhout R, Villa A, Mark AE & Engberts JBFN. (2003) Understanding binding affinity: a combined isothermal titration calorimetry/molecular dynamics study of the binding of a series of hydrophobically modified benzamidinium chloride inhibitors to trypsin. J Am Chem Soc 125: 10570–10579.

Tironi IG, Sperb R, Smith PE & van Gunsteren WF (1995) A generalized reaction field method for molecular dynamics simulations. J Chem Phys 102: 5451–5459.

Torrent M, Vreven T, Musaev DG & Morokuma K (2002) Effects of the protein environment on the structure and energetics of active sites of metalloenzymes. ONIOM study of methane monooxygenase and ribonuclease reductase. J Am Chem Soc 124: 192–193.

Torrie GM & Valleau JP (1977) Nonphysical sampling distributions in Monte Carlo free-energy estimation: umbrella sampling. J Comput Phys 23: 187–199.

Totrov M & Abagyan R (2001) Rapid boundary element solvation electrostatics calculations in folding simualtions: successful folding of a 23 residue peptide. Biopolymers 60: 124–133.

Vajda S, Weng Z, Rosenfeld R & DeLisi C (1994) Effect of conformational flexibility on receptor-ligand binding free energies. Biochemistry 33: 13977–13988.

van der Spoel D, Lindahl E, Hess B, Groenhof G, Mark A & Berendsen HJC (2005) Gromacs: fast, flexible and free. J Comp Chem 26: 1701–1718.

van der Spoel D, van Buuren AR, Apol E, Meulenhoff PJ, Tieleman DP, Sijbers ALTM, Hess B, Feenstra KA, Lindahl E, van Drunen R & Berendsen HJC (2001) Gromacs User Manual version 3.1. Nijenborgh 4, 9747 AG Groningen, The Netherlands. Internet: http://www.gromacs.org.

van Gunsteren WF & Berendsen HJC (1988) A leap-frog algorithm for stochastic dynamics. Mol Sim 1: 173–185.

van Gunsteren WF, Berendsen HJC & Rullmann JAC (1981) Stochastic dynamics for molecules with constraints. Brownian dynamics of n-alkanes. Mol Phys 44: 69–95.

van Gunsteren WF, Beutler TC, Fraternali F, King PM, Mark AE & Smith PE (1993) Computation of free energy in practice: choice of approximations and accuracy limiting factors. In: van Gunsteren WF, Weiner PK & Wilkinson A (eds.), Computer Simulation of Biomolecular Systems, Escom, Leiden, vol. 2, chap. 13.

van Gunsteren WF, Billeter SR, Eising AA, Hünenberger PH, Krüger P, Mark AE, Scott WRP & Tironi IG (1996) Biomolecular Simulation: GROMOS96 Manual and User Guide. BIOMOS b.v., Zürich, Groningen.

Verlet L (1967) Computer experiments on classical fluids: I. Thermodynamical properties of Lennard-Jones molecules. Phys Rev 159: 98–103.

Veselovsky AV & Ivanov AS (2003) Strategy of computer-aided drug design. Curr Drug Targets

Infect Disord 3:33–40.
Villa A & Mark AE (2002) Calculation of the free energy of solvation for neutral analogs of amino acid side chains. J Comp Chem 23: 548–553.
Villa A, Zangi R, Pieffet G & Mark AE (2003) Sampling and convergence in free energy calculations of protein-ligand interactions: the binding of triphenoxypyridine derivatives to factor xa and trypsin. J Comp Aid Mol Design 17: 673–686.
Wand AJ (2001) Dynamic activation of protein function: a view emerging from NMR spectroscopy. Nature Struct Biol 8: 926–931.
Wang J, Wolf RM, Caldwell JW, Kollman P & Case DA (2004) Development and testing of a general Amber force field. J Comp Chem 25: 1157–1174.
Wang W & Kollman PA (2000) Free energy calculations on dimer stability of the HIV protease using molecular dynamics and a continuum solvent model. J Mol Biol 303: 567–582.
Wang W, Wang J & Kollman PA (1999) What determines the van der waals coefficient beta in the LIE (linear interaction energy) method to estimate binding free energies using molecular dynamics simulations? Proteins: Struct Funct Bioinf 34: 395–402.
Wang W, Donini O, Reyes CM & Kollman PA (2001) Biomolecular simulations: recent developments in force fields, simulations of enzyme catalysis, protein-ligand, protein-protein, and protein-nucleic acid noncovalent interactions. Annu Rev Biophys Biomol Struct 30: 211–243.
Warshel A (1998) Electrostatic origin of the catalytic power of enzymes and the role of preorganized active sites. J Biol Chem 273: 27035–27038.
Warshel A & Levitt M (1976) Theoretical studies of enzymatic reactions: dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. J Mol Biol 103: 227–249.
Warshel A & Russell ST (1984) Calculations of electrostatic interactions in biological systems and in solution. Quart Rev Biophys 17: 283–422.
Warwicker J (1986) Continuum dielectric modelling of the protein-solvent system, and calculation of the long-range electrostatic field of the enzyme phosphoglycerate mutase. J Theor Biol 121: 199–210.
Weber G (1975) Energetics of ligand binding to proteins. Adv Protein Chem 29: 1–83.
Wierenga RK, Noble ME & Davenport RC (1992) Comparison of the refined crystal structures of liganded and unliganded chicken, yeast and trypanosomal triosephosphate isomerase. Biochemistry 38: 4389–4397.
Williamson MA (1994) The structure and function of proline-rich regions in proteins. Biochem J 297: 249–260.
Wilson SR & Cui W (1990) Applications of simulated annealing to peptides. Biopolymers 29: 225–235.
Wittekind M, Mapelli C, Farmer BT, Suen K-L, Goldfarb V, Tsao J, Lavoie T, Barbacid M, Meyers CA & Mueller L (1994) Orientation of peptide fragments from Sos proteins bound to the N-terminal SH3 domain of Grb2 determined by NMR spectroscopy. Biochemistry 33: 13531–13539.
Wolfram Research (2003) Mathematica. Wolfram Research, Inc., Champaign, Illinois.
Wüthrich K (1986) NMR of proteins and nucleic acids. Wiley-Interscience, New York.
Yung-Yu S, Lu W & van Gunsteren WF (1988) On the approximation of solvent effects on the conformation and dynamics of ciclosporin A by stochastic dynamics simulation techniques. Mol Sim 1: 369–383.

Zauhar RJ & Morgan RS (1985) A new method for computing the macromolecular electric potential. J Mol Biol 186: 815–820.
Zwanzig RW (1954) High-temperature equation of state by a perturbation method. I. Non-polar gases. J Chem Phys 22: 1420–1426.
Åqvist J & Hansson T (1996) On the validity of electrostatic linear responce in polar solvents. J Phys Chem 100: 9512–9521.
Åqvist J & Luzhkov V (2000) Ion permeation mechanism of the K+ channel. Nature 404: 881–884.
Åqvist J, Medina C & Samuelsson JE (1994) A new method for predicting binding affinity in computer-aided drug design. Prot Eng 7: 385–391.
Åqvist J, Luzhkov VB & Brandsdal BO (2002) Ligand binding affinities from MD simulations. Acc Chem Res 35: 358–365.

Original articles

I Donnini, S. & Juffer, A. H. (2004) Calculation of affinities of peptides for proteins. J Comp Chem 25: 393–411.

II Donnini, S., Villa, A., Groenhof G., Wierenga R. K., Mark A. E. & Juffer, A. H. Understanding a 1000-fold affinity difference: a computational study. Manuscript.

III Donnini, S., Mark A. E., Juffer, A. H. & Villa, A. (2005) Incorporating the effect of ionic strength in free energy calculations using explicit ions. J Comp Chem 26: 115–122.

IV Donnini, S., Groenhof G., Wierenga R. K. & Juffer, A. H. (2006) The planar conformation of a strained proline ring: a QM/MM study. Proteins: Struct Funct Bioinf 64: 700-710.

Original articles are not included in the electronic version of the dissertation.

478. Hillukkala, Tomi (2006) Roles of DNA polymerase epsilon and TopBP1 in DNA replication and damage response

479. Mäki-Petäys, Hannaleena (2007) Conservation and management of populations in a fragmented forest landscape. Behavioural ecology meets population genetics

480. Kärkkäinen, Johanna (2007) Preparation and characterization of some ionic liquids and their use in the dimerization reaction of 2-methylpropene

481. Junttila, Juho (2007) Clay minerals in response to Mid-Pliocene glacial history and climate in the polar regions (ODP, Site 1165, Prydz Bay, Antarctica and Site 911, Yermak Plateau, Arctic Ocean)

482. Sipilä, Laura (2007) Expression of lysyl hydroxylases and functions of lysyl hydroxylase 3 in mice

483. Kivimäki, Anri (2007) Wireless telecommunication standardization processes—actors' viewpoint

484. Räisänen, Liisa (2007) Phage-host interactions in *Lactobacillus delbrueckii*: host recognition and transcription of early phage genes

485. Parviainen, Timo (2007) Ruokohelpiviljelyn optimointi suopohjilla. Turvetuotantoalueiden geologisen ympäristön, pohjaturpeen sekä kierrätyslannoitteiden käytön vaikutus ruokohelpin käyttämiin alkuaineisiin ja satoon

486. Halonen, Raija (2007) Challenges in an inter-organisational information system implementation. Participatory view

487. Välimäki, Panu (2007) Reproductive tactics in butterflies – the adaptive significance of monandry versus polyandry in *Pieris napi*

488. Oinas, Janne (2007) The degree theory and the index of a critical point for mappings of the type (*S*+)

489. Nuortila, Carolin (2007) Constraints on sexual reproduction and seed set in *Vaccinium* and *Campanula*

490. Peltoniemi, Mirva (2007) Mechanism of action of the glutaredoxins and their role in human lung diseases

491. Zheng, Xiaosong (2007) Reference modeling for high value added mobile services

492. Siira, Antti (2007) Mixed-stock exploitation of Atlantic salmon (*Salmo salar* L.) and seal-induced damage in the coastal trap-net fishery of the Gulf of Bothnia. Challenges and potential solutions

Book orders:
OULU UNIVERSITY PRESS
P.O. Box 8200, FI-90014
University of Oulu, Finland

Distributed by
OULU UNIVERSITY LIBRARY
P.O. Box 7500, FI-90014
University of Oulu, Finland

UNIVERSITY OF OULU P.O. Box 7500 FI-90014 UNIVERSITY OF OULU FINLAND

ACTA UNIVERSITATIS OULUENSIS

ISBN 978-951-42-8573-8 (Paperback)
ISBN 978-951-42-8574-5 (PDF)
ISSN 0355-3191 (Print)
ISSN 1796-220X (Online)