Computational studies of protein–ligand molecular recognition

Computer simulaties van eiwit–ligand herkenning
(met een samenvatting in het Nederlands)

Proefschrift ter verkrijging van de graad van doctor aan de Universiteit Utrecht op gezag van de Rector Magnificus, Prof. dr W. H. Gispen, ingevolge het besluit van het College voor Promoties in het openbaar te verdedigen op donderdag 19 april 2001 des middags te 12.45 uur

door
Malcolm Bjørn Gillies
geboren op 10 juni 1971, te Kopenhagen

promotor: Prof. dr J. P. A. E. Tollenaere
verbonden aan het Utrecht Institute for Pharmaceutical Sciences
van de faculteit Farmacie van de Universiteit Utrecht

co-promotor: Dr E. E. Moret
verbonden aan het Utrecht Institute for Pharmaceutical Sciences
van de faculteit Farmacie van de Universiteit Utrecht

The work described in this thesis was in part made possible by a University of Sydney Travelling Scholarship (Boulton Fellowship).

ISBN 90-393-2688-6

She's a model, and she's looking good.
You'd like to take her home, that's understood.
— Kraftwerk, *The Model*

Contents

1 General Introduction **1**
1.1 Introduction . . . 1
1.2 Structure-based design in practice . . . 2
1.3 Computational approaches . . . 3
1.3.1 Scoring . . . 4
1.3.2 Docking . . . 5
1.3.3 The future . . . 5
1.4 Scope of thesis . . . 6

2 Proton linkage in HIV 1 PR–inhibitor complexes **9**
2.1 Introduction . . . 9
2.1.1 Electrostatic interactions and binding . . . 9
2.1.2 The influence of proton location on electrostatics . . . 10
2.1.3 pH effects . . . 11
2.1.4 Experimental pK_a measurements . . . 12
2.1.5 pK_a prediction . . . 12
2.1.6 Theory . . . 13
2.1.7 HIV-1 protease . . . 15
2.2 Methods . . . 17
2.2.1 Modelling . . . 17
2.2.2 pK_a calculations . . . 18
2.3 Results and discussion . . . 20
2.3.1 Accuracy of pK_a predictions . . . 23

3 Electrostatic complementarity in HIV-1 PR–inhibitor complexes **25**
3.1 Introduction . . . 25
3.1.1 Electrostatic complementarity . . . 25
3.1.2 Master equation approaches . . . 26
3.1.3 Problems scoring HIV PR inhibitors . . . 28
3.2 Methods . . . 30
3.2.1 Poisson-Boltzmann calculations . . . 30
3.2.2 Proton linkage . . . 31
3.2.3 Correlation of electrostatic potentials . . . 31
3.2.4 Other parameters for regression . . . 32
3.2.5 Regression analysis of ΔG . . . 32

3.3 Results . . . 33
3.4 Discussion . . . 35
3.4.1 Comparison with other work . . . 36
3.4.2 Alternative approaches . . . 37
3.4.3 Flexibility . . . 38

4 CBMC: implementation 39
4.1 Introduction . . . 39
4.1.1 Configurational bias Monte Carlo . . . 40
4.1.2 CBMC for polypeptides . . . 41
4.2 Methods . . . 42
4.2.1 CBMC . . . 42
4.2.2 BIGMAC . . . 44
4.2.3 CHARMM22 potential . . . 45
4.2.4 Coupled-decoupled CBMC . . . 47
4.2.5 Implementing CHARMM22 in BIGMAC . . . 48
4.2.6 Validation . . . 51
4.2.7 Alanine dipeptide . . . 52
4.3 Results and discussion . . . 53
4.4 Conclusions . . . 62
4.4.1 Limitations . . . 63

5 CBMC: applications 65
5.1 Introduction . . . 65
5.1.1 Free energy simulations . . . 65
5.1.2 Docking and scoring . . . 66
5.1.3 AUTODOCK . . . 67
5.1.4 Parallel tempering . . . 67
5.2 Methods . . . 69
5.2.1 Models . . . 69
5.2.2 AUTODOCK . . . 72
5.2.3 WHOPPER . . . 73
5.3 Results . . . 74
5.3.1 Docking . . . 74
5.3.2 Parallel tempering . . . 76
5.3.3 Efficiency . . . 78
5.4 Discussion and conclusions . . . 80
5.4.1 Docking . . . 80
5.4.2 Efficiency . . . 81
5.4.3 Prospects . . . 82

6 General discussion 85

Bibliography 89

Summary 105

Samenvatting 107

Curriculum vitæ 111

Dankwoord 113

Chapter 1

General Introduction

1.1 Introduction

Drug design is almost exclusively an activity of the pharmaceutical industry, and given this industrial background, research in the field is characterized by a combination of the pragmatic and the speculative. Novel methods are embraced with great enthusiasm in the hope of finding shortcuts in the hit-and-miss process of discovering and developing leads, but the perspective of traditional medicinal chemistry remains central. Computational approaches have reflected the medicinal chemist's viewpoint from the beginning, first in the form of chemometric methods, and later pharmacophore-based approaches. Recent developments in combinatorial chemistry, high-throughput screening and molecular biology (genomics, bioinformatics, proteomics ...) are the latest trends threatening to turn tradition upside-down.

The rise of "structure-based (drug) design" (also known as "direct design", or even "rational drug design") was accompanied by similar revolutionary pretensions. The longer term consequence of this trend has been the growing importance of the viewpoint of the crystallographer (or structural biologist). This interest in techniques where the structure of the receptor is pivotal has been encouraged by the demonstrable importance of crystallographic data to the development of the highly successful HIV-1 protease (HIV-1 PR) inhibitors [250], and lead compounds for a number of other receptors since [167], coupled to the rapid growth in the number of "receptors in search of a drug" identified by molecular biology [117]. As always, the promise of computational approaches is the ability to make accurate predictions about the activity of novel ligands quickly and cheaply.

A convenient definition of structure-based design is the use of an atomic resolution receptor structure (presumed accurate) to design small ligands with high affinity. Only the *in vitro* biochemical situation is relevant to this viewpoint, and kinetics are neglected. Given a candidate ligand, specific molecular interactions are considered to be responsible for the molecular recognition process. Automated structure-based design procedures work with simplified models in the interest of speed, and in practice do not produce high affinity ligands from scratch. Instead, the aim is to rank collections of compounds as prospective ligands, with the most highly ranked to be considered by the medicinal chemist as possible leads. Design therefore remains a process of iterative improvement, with experimental evaluation required at each step. Once ligands of moderate potency have been found, structural studies of the ligand–receptor complex by X-ray crystallography or NMR give insight into the binding mode, more data for computational efforts, and an improved chance of success in the next iteration.

The picture of molecular recognition which underlies the principles of structure-based design, is born out of macromolecular crystallography. The understanding of protein and DNA structure in terms of hydrogen bonding and shape complementarity has shaped our ideas of how ligand–receptor complexes fit together with high specificity and affinity. The flip side of the incredible success of structural biology is that recognition is seen primarily in static terms, while binding is an equilibrium process where factors such as the displacement of water from the active site are of great importance. A consideration of these other essential components of molecular recognition means expanding the principles of ligand design with techniques and insights from other fields, from the models of chemical physics, to the complex signalling theories of immunology and cell biology.

Despite a wealth of experimental data from structural biology, and an ever-growing level of sophistication in computational methods, we are a long way from a workable theory of molecular recognition. It is generally agreed that well-understood physical laws underlie interactions between biological molecules, but unfortunately this does not imply that testable predictions can be made, for instance of the association constant of an arbitrarily chosen ligand/receptor system. So far, detailed physical models have demanded infeasibly large amounts of computation to predict relevant, observable properties of the system in question. Neither have simplified models, within the reach of computation or analysis, been successful in allowing quantitative prediction. Considering the difficulty of predicting the bulk properties of a pure liquid such as water through simulation, none of this is surprising. Descriptions of molecular recognition derived empirically have been more illuminating, and have also provided alternative methods for quantitative prediction. However, experiment regularly produces unexpected results regarding the affinity or binding mode of novel complexes [53, 153].

1.2 Structure-based design in practice

The development of the HIV-1 protease inhibitors is an informative example of the practice of structure-based design and the analysis of molecular recognition. HIV protease was identified as an aspartic protease soon after the HIV genome was sequenced [189], allowing the first lead compounds to be found quickly without large-scale screening or knowledge of the enzyme structure. Saquinavir, the first drug to be approved by the US FDA, was developed as a substrate analogue, without an important contribution from structural information [250]. All subsequent drugs have been developed with explicit knowledge of the structural basis of HIV protease inhibition, and more than a hundred crystal structures of enzyme–inhibitor complexes determined as part of drug development efforts have been released to the public [236].

Information about the molecular recognition of inhibitor molecules by the protease has been primarily useful as an aid to the intuition of the medicinal chemist. In many cases (for instance, in the development of Indinavir [64]), structures of a complex were used as a basis for modelling studies where possible improvements to the inhibitor structure could be tested against the structural constraints, and additional binding sites could be identified. Knowledge of the structure of the binding site led to such strategies as the use of symmetrical inhibitors [89], and inhibitors designed to displace a structural water molecule bound in the active site [220]. The addition of structural data to the iterative process of development also gave an important opportunity to check if optimization of the ligand structure resulted in the hypothesized molecular interactions. The observation that symmetric inhibitors may bind

to HIV-1 PR in an asymmetric manner led the development effort at Abbott away from these compounds [250]. In general, the intention was that structural insight would guide the development process, allowing it to be completed with fewer cycles of trial-and-error. In view of the rapid appearance of viral resistance, new structure-based computational methods to design agents effective in the face of protease mutation [194] are also likely to attract much interest.

Computational techniques have played a fairly minor role in the development of marketed HIV-1 protease inhibitors, although many details of the story are still trade secrets. Pharmacophores suggested by modelling and crystal structures have been used for prospective searching of databases [128], and a *de novo* technique was used to find appropriate substituents to fill a binding sub-site in one effort at Agouron Pharmaceuticals [80]. The predictions from the available computer-aided design techniques are not reliable enough to make them the guiding principle of the development process, but rather one possible source of ideas for better leads.

So far, analysis of the unprecedented quantity of structural data from HIV-1 PR–inhibitor complexes has not revealed any novel principles of molecular recognition. The enzyme is promiscuous and binding sub-sites can accommodate groups of various sizes [248]. Unlike the natural polypeptide substrate, most inhibitors are hydrophobic, and hydrophobic interactions appear to be the dominant factor in determining affinity [250]. Similar observations were made in a detailed structural and thermodynamic study of the binding of Lysine-X-Lysine tripeptides to the oligopeptide-binding protein OppA [52]. The OppA binding site for the amino acid "X" undergoes little structural change on binding, but adapts readily to different sidechains through the intermediation of bound water molecules and ions. Evidently, the binding equilibrium involves numerous compensatory effects, making prediction of the net balance and the ultimate affinity extremely difficult.

Two unusual features of the HIV-1 PR active site are the presence of a structural water molecule, thought to play a role in substrate binding and catalysis, and shifted pK_as for the two catalytic aspartate residues. Displacement of the water molecule has been used as a strategy for the development of highly specific ligands [128, 129], with one compound in this series, DMP450, still in clinical trials [226]. The shifted pK_as of the aspartate residues are important in determining the affinity and pH-dependence of inhibitor and substrate binding, as almost all ligands hydrogen bond to at least one of these groups [250].

1.3 Computational approaches

Automated methods for structure-based drug design are not mature enough to replace existing practices, but new approaches and computer programs continue to appear. Most work falls into the categories of *de novo* design, docking, and scoring. *De novo* design entails the automatic design of novel ligands intended to bind a given receptor with high affinity, and is the most difficult of the three problems. While *de novo* methods have had some successes, such as in the optimization of a influenza virus sialidase inhibitor [113], most methods have not been subjected to statistically meaningful experimental evaluation. An area where more experimental evaluation has been done is the application of structure-based design techniques to combinatorial libraries. A library of prospective cathepsin D inhibitors designed in this way contained many more compounds with an IC_{50} of micromolar or better than a standard library of the same monomers [122], and a nanomolar thrombin inhibitor was discovered using

a similar approach [168].

Docking and scoring are complementary methods which split the problem of ligand binding into two simpler subproblems. For a receptor of known structure and a ligand, docking involves finding the bound configuration of the ligand and receptor, assuming a specific interaction of reasonable affinity, while scoring involves estimating the affinity of the ligand, given a particular bound configuration. Most docking and scoring methods are empirical approaches involving extrapolation from known ligand–receptor complexes.

The binding process can also be simulated using physically accurate atomic potential functions, and these simulations will yield information about the binding mode and affinity. Currently available force fields are probably sufficient for these free energy simulations, but the calculations are too slow for most applications (see Chapter 4 for further discussion). Practical free energy simulations to compare the affinities of different ligands at a particular receptor have most of the restrictions of empirical scoring approaches, although they are considered to be more accurate. In particular, the bound configuration must be known and used as the starting point of the simulation, or the results converge unreasonably slowly. The computational and technical demands of simulation approaches has meant that docking and scoring methods have received attention as faster, simpler alternatives.

1.3.1 Scoring

Quantitative Structure-Activity Relationship (QSAR) methods were the forerunner of many of the current scoring techniques. Applications of QSAR demonstrated that molecular structure could be quantitatively related to pharmacological activity, and that these relationships could be used in drug development. While the success of QSAR inspired interest in computational approaches to drug design, its restriction to one- and two-dimensional properties of the ligand gave rise to efforts to incorporate three dimensional structure into activity predictions. Growing availability of receptor structural data and the development of docking techniques enabled the interactions between ligand and receptor to be used as the starting point for scoring. Regression-based methods, exemplified by Böhm's scoring functions [29, 30], are a structure-based outgrowth of QSAR techniques.

The theoretical basis of QSAR analysis is the presumed existence of a linear free-energy relationship between a physicochemical descriptor of a molecule (such as log P, the octanol-water partition coefficient) and its affinity for a receptor [94]. These relationships are quantified by regression analysis, and can be highly predictive for closely analogous series of compounds. The Böhm approach replaces the physicochemical descriptors with a collection of quantities calculated from the geometry of the ligand–receptor complex, which are chosen to capture the most important interactions between the two, such as hydrogen bonds. With an appropriate set of interaction descriptors, and a representative set of complexes to perform the regression on, an equation with general applicability should result. Ideally, the coefficients of the equation should be consistent with an intuitive understanding of the molecular recognition process.

Scoring using the master equation approach often uses the same functional form as regression approaches, namely a linear combination of interaction descriptors, although the derivation of these scoring functions is quite different. The underlying assumption is that a highly simplified description of molecular recognition will be sufficient to make useful affinity predictions [4]. The simplifications are typically chosen on the grounds of expediency

and may include: rigid receptor and ligand; neglect of solvation effects; fixed contributions from hydrogen bonds; neglect or rough approximation of entropic effects; simple estimates of hydrophobic binding; continuum approximations to electrostatic energies. The most important assumption is that the free energy of interaction can be decomposed into independent components, such as electrostatic energies, entropies, hydrophobic interactions, and so on. Methods which involve minimization and analysis of force field interaction energies are also in the category of master equation approaches.

1.3.2 Docking

The docking problem has been approached in a variety of ways. Docking is essentially a matter of searching and optimization, and is a straightforward geometric problem when both ligand and receptor are treated as rigid. Consideration of the flexibility of ligand and/or receptor involves many additional degrees of freedom, making heuristics necessary to limit the extent of the search. The quantity to be optimized (the "interaction energy") also presents some difficulties. DOCK, one of the oldest and most widely used programs, is based on a simple on-off model of steric complementarity [126]. Potential functions similar to those used for simulations (such as AMBER and CHARMM) have been popular, but the need for an implicit treatment of solvation has resulted in the development of energy terms similar (or identical) to the scoring functions described above. Small, rigid ligands with high affinity can be docked easily to rigid receptors by any of a variety of methods, but ligands such as peptides which are highly flexible are a challenge to even the most sophisticated docking algorithms.

Essentially important to these techniques are the models of ligand and receptor themselves. At a basic level, these function as a convenient substitute for real-life metal or plastic molecular models, giving immediate visual-spatial insight into structure in an uncomplicated way. Quantitative studies make more demands on the physical realism of the model, including the accuracy of the potential energy function. The most popular protein forcefields have been extensively evaluated for their ability to reproduce conformational equilibria [23], but may be lacking in other aspects which are important for accurate simulations, for instance polarisation, and the strength and directionality of non-bonded interactions. Notably, the most accurate free energy simulations have required additional optimization of potential functions [51, 131]. Perhaps more troubling, the experimental basis of models is sometimes patchy. Structural data from X-ray crystallography or NMR studies are of limited resolution and completeness, and these shortcomings should be considered when building models [104]. Details such as solvation and protonation state need to be addressed, often without any experimental data. All of these uncertainties make the validation of models a vital (if awkward) aspect of work in this field.

1.3.3 The future

The failings of current methods for computing ligand–receptor interactions will be corrected at least in part by developments in related fields. Most importantly, faster computers will allow larger and larger systems to be analysed using brute force methods, as will innovations in numerical algorithms. The accumulation of accurate structural and biochemical data will allow better calibration of force fields and scoring functions, and an improved ability to distinguish

between good and bad models. Improved experimental techniques for measuring pK_as will allow more accurate modelling of protonation states.

The refinement of force fields and (implicit) solvation models will also be an important area of advancement. Efficient and accurate implicit solvation schemes will improve the speed of many types of simulations, and bring some excessively costly calculations within reach. In the absence of accurate experimental data, simulations using more detailed levels of theory can be used in the development of models at coarser levels of description. In this way, quantum chemical calculations provide the parameters for potentials for atomistic simulations, and atomistic simulations provide the parameters for implicit solvent simulations, and so forth [237].

Of course, the impact of new methods is difficult to predict. More efficient simulation and free energy formalisms are very desirable, and the consequences of recent developments in this area, such as linear response approaches [15], advanced Monte Carlo (MC) techniques [77], and generalised ensemble methods [134] are not yet clear. Methods developed in the chemical physics community take time to find applications in biological problems, often due to the extra implementation work needed for the more complex biomolecular models. However, the exceptional properties of biological molecules may require more than just the adaptation of existing chemical and physical knowledge, as research in protein folding has demonstrated. Workable models often depend on careful consideration of how different aspects of the system may be simplified [73, 158]. In this vein hybrid or hierarchical models can be useful, as demonstrated by the MBO(N)D approach to molecular dynamics [45].

An area of recent intense interest has been the development of knowledge-based potentials for scoring [59, 87, 157, 165, 166, 206, 231]. This way of constructing effective potentials was popularized in the field of protein-folding. The growth in the number of ligand–receptor complexes found in the Protein Data Bank (PDB) [25, 26] has made these potentials a practical and perhaps more accurate substitute for regression-based scoring functions. Although the statistical-thermodynamic basis of existing knowledge-based potentials is not entirely rigorous, it is tempting to imagine the possibility of bridging the gap between empirical and physical descriptions of molecular recognition through the concept of potentials of mean force. Further investigation of details such as atom-typing, hydrogen atom placement, and treatment of the solvent, will reveal the possibilities and limitations of these methods.

Developments in computational approaches to molecular recognition are likely to be gradual, considering that the vital ingredient of validation is laborious and time-consuming. The CATFEE initiative [192] for the competitive, blind testing of free energy prediction techniques is an important development in this respect. Advances in experiment, theory and implementation reinforce one another, but feedback between these specialized fields takes time. Looking at the developments over twenty years of docking, fifteen years of protein free energy simulations (with a million-fold increase in simulation time scales along the way) and a decade of scoring, it seems likely that the next ten years will see big improvements in our ability to predict binding affinities.

1.4 Scope of thesis

This thesis describes various computational approaches to the analysis of ligand–protein molecular recognition. The aim is to evaluate a number of methods with particular attention to possible applications in computer-aided structure-based drug design. Addition of extra detail

to computer models regarding protonation and electrostatics, and the use of a novel simulation technique, configurational bias Monte Carlo sampling, are the basis for this attempt to gain more accuracy and insight from scoring and docking methods. HIV-1 protease and its inhibitors were chosen as the model system for these studies because of the ready availability of many X-ray crystal structures, and the special role of protonation in the mechanism of action and inhibition. HIV-1 protease is also a small, straightforward enzyme, as demonstrated by its complete chemical synthesis [49, 198], simplifying the task of modelling.

The starting point for investigating the protonation state and electrostatic characteristics of HIV-1 PR was a series of pK_a calculations on enzyme–inhibitor complexes, using the finite-difference Poisson-Boltzmann approach (Chapter 2). Thermodynamic linkage of the protonation of active site residues with inhibitor binding means that a portion of the free energy change on binding can be ascribed to a change in protonation, and the size of this effect may vary between inhibitors. The results of the pK_a calculations allow models to be built with appropriate protonation, and correction of K_D values for proton linkage effects.

Models of the complexes and corrected K_D data were used to analyse electrostatic complementarity and the relationships between structural parameters and affinity (Chapter 3). A regression analysis drawing on the concepts of the Böhm scoring functions and master equation approaches was used, and a comparison was made with results from other scoring and simulation literature about HIV-1 PR. In conclusion the consequences for our understanding of molecular recognition in this system are discussed.

Some of the shortcomings of scoring functions, particularly the difficulty of estimating the size of the entropy change on binding, are avoided in simulations of ligand–receptor binding using Monte Carlo or molecular dynamics sampling. Configurational bias Monte Carlo (CBMC) sampling is much more efficient than MD or Metropolis MC for simulating the adsorption behaviour of simple polymers, and recently has proven useful for the conformational analysis of simplified polypeptide models. A CBMC technique was developed for the detailed atomic potentials normally used for protein and peptide simulations, and its accuracy and efficiency was validated (Chapter 4).

As well as conventional simulation applications, CBMC can be applied to docking. The efficiency of CBMC in sampling internal degrees of freedom should be advantageous for docking flexible molecules. Docking of a tripeptide inhibitor to a rigid HIV-1 PR model was attempted using CBMC and the results compared with other docking algorithms, such as the MC simulated annealing method of AutoDock [88] (Chapter 5). The parallel tempering method [82], similar to simulated annealing, was used to improve sampling. The efficiency of CBMC for this type of simulation was also further analysed.

Chapter 2

Proton linkage in HIV 1 PR–inhibitor complexes

2.1 Introduction

Structural biology provides a first-rate vantage point for looking at structure-activity relationships of drugs. Detailed and accurate models of receptor proteins make many categories of qualitative and quantitative analysis of ligand–receptor binding possible, from empirical QSAR approaches, to physically realistic simulations. Nonetheless, construction of a model from high-resolution X-ray or NMR data requires a number of assumptions about details absent from the structural data. For models based on X-ray data, the presence and location of protons is missing from the experimental picture. Sometimes precise placement of protons makes no difference to the usefulness of the model, but this is not so if the electrostatic characteristics of the protein are to be considered.

This study aims to calculate apparent pK_as for HIV-1 PR/inhibitor complexes, and make predictions of the nature and size of proton linkage phenomena in these systems. An adequate description of the protonation state is a prerequisite for QSAR studies and accurate modelling of complexes, given that these may represent a sizeable perturbation of "standard" conditions, and a substantial correction to the free energy changes involved. Experimental results showing pH dependence of inhibitor binding indicate this to be the case for a number of the ligands considered.

HIV-1 PR inhibitor data has been used in many QSAR/scoring studies without reference to possible shifts in protonation state in the active site. The pH at which inhibition constants are measured, and at which crystal structures were determined is also rarely taken into account. It may be that the degree of uncertainty inherent in such studies is bigger than the effect of these factors, but on the other hand, their neglect may contribute to the inaccuracy of these methods.

2.1.1 Electrostatic interactions and binding

Molecular recognition observed in crystal structures of ligand–receptor complexes (reviewed from a drug design perspective in [31]) can often be explained in terms of electrostatic interactions. The importance of hydrogen-bonding to recognition is well-known, and there are many examples where a certain pattern of hydrogen-bonding between ligand and receptor is thought to be essential for specificity. Hydrogen-bonding can be adequately described by

classical electrostatic interactions between point charges on atomic nuclei, at least at the level of accuracy expected in general-purpose forcefields. Electrostatic interactions between groups bearing a formal charge (for instance, in salt bridges) also have a clear role in molecular recognition.

The nature of the contribution of electrostatic interactions to the free energy change on binding is less clear. Where these interactions appear to be close to optimal, for instance in the hydrogen-bonding between streptavidin and biotin, it is hypothesized that they are a driving force for association [246], although unravelling the precise contribution has proven difficult [218]. More generally, the net contribution of interactions between polar groups depends on the balance between the strength of the interactions seen in the complex and the solvation energy of the same groups when the species are free in solution, and as seen in studies of thermolysin inhibitors, loss of a hydrogen bond does not necessarily lead to a decrease in affinity [162]. There is no clear consensus on a simple, general description of these contributions, but at a coarser level of description, a term proportional to the polar surface area buried on complex formation appears in some empirical descriptions of the thermodynamics of binding (e.g. [140]).

In a number of cases, the influence of electrostatic interactions on binding kinetics is clear. Electrostatic "steering", where oppositely charged ligand and receptor attract each other at long range, is the mechanism responsible for the diffusion-limited catalytic rate of the oxidation of the superoxide anion by superoxide dismutase [101]. However, these phenomena are not observed in all interactions, and a number of enzyme–substrate systems are known where like charges in ligand and receptor might be expected to impede the binding process [133].

2.1.2 The influence of proton location on electrostatics

The electrostatic behaviour of proteins and peptides is primarily determined by the precise location of covalently bound hydrogens, or more loosely, protons. Hydrogen bonds are directional, so the orientation of hydroxyl rotamers, unclear at the resolution of typical protein X-ray structures, determines the possibilities for interaction. As hydroxyl groups function simultaneously as hydrogen bond donor and acceptor, optimization of the hydrogen bond network which arises when a number of donating and accepting groups interact at an interface, involves the consideration of a large number of possible configurations. Analysis of ligand–receptor interfaces indicates that high-affinity complexes rarely contain hydrogen bond donors which do not make a hydrogen bond [147, 256], suggesting that this situation is accompanied by a significant free energy penalty.

The location of protons also depends on the ionisation state of titratable groups, and the configuration of tautomers. As these protons are also invisible to X-ray diffraction in most practical cases, the formal charge and hydrogen bond donor/acceptor status of many groups in a protein is ambiguous. Experimental pK_as for groups such as carboxylic acids are a guide, but the apparent pK_a of titratable groups in proteins depends on the protein environment, and may not correspond to that measured for model compounds.

Building molecular models of proteins therefore requires information in addition to that provided by X-ray structures. Where this information about hydrogen bonding and protonation is unavailable from experiments, assumptions or predictions must be used instead.

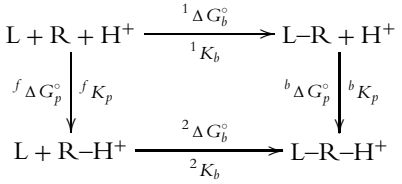


Figure 2.1: Thermodynamic cycle showing a simple linkage scheme for the binding of a ligand (L) to a receptor (R). Superscripts f and b refer to the free and bound states, subscripts p and b refer to the protonation and binding processes, and superscripts 1 and 2 refer to the binding of the deprotonated and protonated forms of the ligand, respectively.

Automated procedures exist for optimization of hydrogen bonding networks (e.g. [103]), and these are adequate for many purposes. Prediction of protonation states is more difficult, and a number of approaches have been proposed. Prediction of apparent pK_as by simulation of electrostatic interactions between residues is the best characterized of these.

2.1.3 pH effects

Biochemical processes are influenced by the pH of the medium in which they take place. At extremes of pH, proteins unfold and enzymes cease to function. Within these extremes, the rate of enzymatic catalysis has been observed to depend on pH, with a maximal velocity at a particular pH. The binding of inhibitors to enzymes can also be pH-dependent. In these situations, where an equilibrium process, such as binding, is pH-dependent, then that equilibrium is thermodynamically linked to the protonation state.

The nature of this linkage can be described in a free-energy cycle and the associated expressions for equilibria (Figure 2.1). This represents the simplest case, with linkage to a single protonation site. When multiple protonation equilibria are linked, this more complex situation is described by the binding polynomial formalism [83, 197]. Linkage can also be described as a binding-related shift in the apparent pK_as of the titratable groups in ligand and/or receptor. When these pK_a shifts are large, the free energy of protonation will be a large component of the free energy of binding at certain pH values. An alternative description is that binding of the ligand is accompanied by the release or uptake of protons, which is associated with a pH-dependent free energy change.

Following the scheme for a single proton linkage given in Figure 2.1, the observed association constant, K_{obs}, as a function of pH, is given by

$$K_{\mathrm{obs}} = {}^1K_b \frac{1 + 10^{(\mathrm{p}K_a)_b - \mathrm{pH}}}{1 + 10^{(\mathrm{p}K_a)_f - \mathrm{pH}}} \tag{2.1}$$

where $(\mathrm{p}K_a)_b = \log_{10}({}^bK_p)$ and $(\mathrm{p}K_a)_f = \log_{10}({}^fK_p)$.

2.1.4 Experimental pK_a measurements

The titration curve of a protein is the result of the contributions of numerous titratable groups. Fitting to this curve gives a series of effective pK_as for the protein, which in the absence of other information cannot be assigned to particular amino acid residues. Additional experimental techniques and knowledge of protein structure may allow complete assignment, as in the case of hen egg-white lysozyme [127]. Unambiguous assignment of an apparent pK_a is possible in an NMR titration experiment, where the pH-dependence of the chemical shift of a group adjacent to the titratable proton is observed (e.g. measurements of hen and turkey egg-white lysozymes [19]).

NMR titration experiments can be performed in the presence of a bound ligand, making measurement of binding-related shifts in apparent pK_a possible. NMR studies of proton linkage along these lines have been done for a small number of HIV-1 protease–inhibitor complexes, examining the titration behaviour of the active site aspartyl dyad [214, 244, 257]. Surface plasmon resonance and isothermal titration calorimetry are also suitable techniques for this type of experiment, if the pK_as can be assigned successfully [159, 254, 255].

Protein titration in the crystal state has been analysed by atomic resolution (< 1.2 Å) X-ray crystallography in an innovative study which appeared in 1999 [24]. Diffraction data collected from RNase A crystals at six different pH values allowed the locations of titrating protons to be resolved, and so pK_a values for most of the histidine residues could be determined. Concerted, pH dependent structural changes involving the rearrangement of hydrogen bond networks could be observed, as well as alternate side chain conformations associated with different protonation states.

2.1.5 pK_a prediction

pK_as for proteins can be predicted quite well by calculations using detailed protein structure models. Absolute pK_as cannot be predicted reliably, but prediction of shifts from experimentally determined standard values (e.g. for model compounds with a single titratable group) is possible. Workable calculations rest on the assumption that electrostatic interactions alone are responsible for pK_a shifts, an approximation which appears to be accurate in comparison to the other simplifications made in the calculation. Unsurprisingly, these electrostatic interactions are strongly influenced by the aqueous environment of the protein, the effect of which is usually approximated by dividing the system into media of low and high dielectric constant, for the protein and solvent respectively.

The most accurate well-characterized protocol for pK_a determination gives predictions with a RMS deviation from experimental values of about 0.7 pK_a units, compared to a typical experimental accuracy of 0.1 pK_a units [12]. Assumption of "average" pK_as for all residues (the "null hypothesis") results in an RMSD of 1 pK_a unit, slightly worse than calculation (though RMSD may not be the best indicator of performance) [12]. Errors in predicted pK_as are not evenly distributed between residues, however, and the predicted pK_a for certain residues may be completely wrong. Given the many uncertainties and approximations inherent in the method, explaining these errors is difficult.

These pK_a prediction calculations can reproduce large pK_a shifts in a number of systems. Often these are residues associated with catalytic activity in an enzyme, such as in hen egg-white lysozyme. Recently, calculations were applied to a system where a large pK_a shift occurs

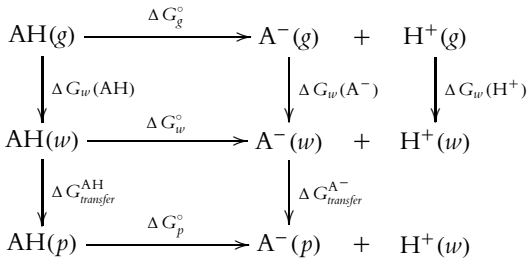


Figure 2.2: Thermodynamic cycles showing breakdown into free energy contributions to the pK_as of AH(w) and AH(p).

on binding, namely HIV-1 protease and a number of its inhibitors [227]. Calculated pK_as agreed well with those that have been experimentally measured. Puzzlingly, both experimental and calculated pK_as gave a prediction of proton linkage while experimental measurement found no pH-dependence of binding. This suggests that the global pH-dependent behaviour of proteins may depend on factors which are difficult to identify.

2.1.6 Theory

The brief theoretical introduction that follows reproduces the formulation found in the review of Briggs and Antosiewicz [35]. A scheme for the free energy changes in the protonation equilibria of proteins is shown in the thermodynamic cycle in Figure 2.2. The group AH is considered in the gas phase, AH(g), as an isolated species in water, AH(w), and as a subunit of a protein, AH(p). The standard free energy change for the deprotonation of AH(w), ΔG°_w, can be defined in terms of the gas phase proton affinity of the group, $-\Delta G^\circ_g$, the solvation free energies of the group in the protonated and deprotonated states, ΔG_w(AH) and $\Delta G_w(A^-)$, and the solvation energy of the proton $\Delta G_w(H^+)$. This in turn defines the absolute pK_a of AH(w) (defining the conversion factor $k \equiv (RT \log_e 10)^{-1}$):

$$pK_a^{(w)} = k\Delta G^\circ_w = k[\Delta G^\circ_g + \Delta G_w(A^-) - \Delta G_w(AH) + \Delta G_w(H^+)]$$

All of these quantities can be predicted by simulation, but no accurate estimates, either calculated or experimental, are available for the proton solvation free energy. As the illustrated thermodynamic cycle is the most workable route to prediction, this means that absolute pK_as such as $pK_a^{(w)}$ can only be determined experimentally. Simulations do provide a reasonably reliable route to relative pK_as, allowing the pK_a of the group in protein environment to be predicted based on the (experimental) isolated pK_a in water, a quantity known from model compounds. The relationship can be seen in the lower cycle in Figure 2.2, and gives:

$$pK_a^{(p)} = pK_a^{(w)} + k(\Delta G^{A^-}_{transfer} - \Delta G^{AH}_{transfer}) = pK_a^{(w)} + k(\Delta G^{AH,A^-}_{protein} - \Delta G^{AH,A^-}_{water}) \qquad (2.2)$$

The terms $\Delta G_{protein}^{AH,A^-}$ and $\Delta G_{water}^{AH,A^-}$ do not appear in the illustrated free energy cycle, and are the free energy differences between protonated and deprotonated states (i.e. the ionization energies) in the protein and water environments, which are more easily accessible to simulation, and are typically used instead of the transfer energies.

The definition of these free energy changes requires some further description. The standard free energy change of protonation of the isolated group in water, ΔG_w°, is related to the equilibrium concentrations of the species by the usual expression

$$\exp \frac{\Delta G_w^\circ}{RT} = \frac{[AH]}{[A^-][H^+]}$$

which using the definitions

$$pH \equiv -\log_{10}[H^+] \quad \text{and} \quad pK_a^{(w)} \equiv k\Delta G_w^\circ$$

gives

$$\log_{10} \frac{[AH]}{[A^-]} = pK_a^{(w)} - pH$$

Assuming a Boltzmann distribution for the relative concentrations of the species [AH] and $[A^-]$,

$$\frac{[AH]}{[A^-]} = \exp \frac{-\Delta G}{RT}$$

giving

$$\Delta G = k^{-1}(pH - pK_a^{(w)}) \tag{2.3}$$

that is, the free energy of the protonated state relative to the deprotonated state at arbitrary pH.

Assuming linear superposition, the combination of Equations 2.2 and 2.3 allows the free energy of an arbitrary ionization state of the whole molecule at a particular pH, relative to the neutral protein, to be written as a function of the model compound pK_as, the ionization energy difference terms, G_{ii}, and an interaction matrix, G_{ij}:

$$\Delta G(pH, x_1, \ldots, x_M) = k^{-1} \sum_{i=1}^{M} x_i \gamma_i (pH - pK_i^{(w)}) + \sum_{i=1}^{M} x_i G_{ii} + \sum_{i=1}^{M-1} \sum_{j=i+1}^{M} x_i x_j G_{ij}$$

for a protein with M ionizable groups, where x_i is 1 if the group i is ionized, and 0 if it is neutral, and γ_i is -1 for acidic groups and $+1$ for basic groups.

The interaction free energy matrix, G_{ij}, can be calculated from the electrostatic free energies of interacting pairs of titratable residues in the protein. Specifically,

$$G_{ij} = \Delta G_{\text{protein},i,j}^{\text{electrostatic}} - (\Delta G_{\text{protein},i}^{\text{electrostatic}} + \Delta G_{\text{protein},j}^{\text{electrostatic}})$$

where each ΔG term refers to the electrostatic free energy change associated with the ionization of a group or pair of groups, in an otherwise neutral system of protein and solvent.

As the free energy changes are calculated relative to the model compound ionization equilibrium, additional terms give the required energy differences as defined in Equation 2.2 (also known as the self-energy terms), and are conventionally assigned to the diagonal elements of the interaction matrix:

$$G_{ii} = \Delta G^{\text{electrostatic}}_{\text{protein},i} - \Delta G^{\text{electrostatic}}_{\text{model},i}$$

with the model compound ionization energy term calculated for an isolated group in water.

Ionization states for the groups of interest can be derived from the Boltzmann average of states over the energy levels. For systems with many titrating groups, a full enumeration is impractical, so a Metropolis Monte Carlo simulation or some variation of a mean-field approximation is used. The apparent pK_a for a group is the pH at which it is on average 50% ionized. A robust method to determine this is simply to perform a series of MC simulations at various pH values.

In this chapter, pK_a predictions were made using the "mesoscopic" continuum model of protein electrostatics, in which the electrostatic potential is calculated from the finite difference solution of the Poisson-Boltzmann (PB) equation. In this model the protein is represented as a low-dielectric solute, bounded by its solvent-accessible surface, embedded in a high-dielectric electrolyte. The charge density of the protein is represented by point charges fixed at the positions of the protein atoms, with ions in the electrolyte obeying a Boltzmann distribution. The physical significance of the protein dielectric constant in this model has been debated, but in a general sense it can be seen as an adjustable empirical screening parameter, which compensates somewhat for the neglect of dipole rearrangement [35]. Another method with similar accuracy to the one used here is the protein dipole-Langevin dipole (PDLD) model [245].

2.1.7 HIV-1 protease

HIV-1 protease (HIV-1 PR) is a viral enzyme vital for the life cycle of the human immunodeficiency virus [124]. The protease is not very specific, but has some preference for cleavage sites in the viral gag and gag-pol polyproteins. The active protease is a symmetric homodimer with the single active site located in a central "tunnel" penetrating the protein (see Figure 2.3). One side of this tunnel is formed by two β-loops (named D1 in the standard nomenclature), one from each monomer, which meet close to the symmetry axis of the enzyme. NMR studies of the apoenzyme in solution show that these loops (or flaps) are very flexible [112], with the otherwise almost inaccessible active site regularly exposed by disordered flap conformations. Crystal structures of the apoenzyme have an ordered, "open" configuration for the flaps, while complexes have the flaps in tighter, "closed" arrangement, with the tips of the flaps moving about 7 Å [248].

The backbone conformation of the enzyme in enzyme–inhibitor complexes is constant between many crystal structures with different inhibitors, and for almost all inhibitors the interactions between ligand and receptor are similar. The binding pockets for the ligand sidechains have been classified and labelled, with sites numbered starting adjacent to the scissile amide bond of the bound substrate. The pocket on the N-terminal side is labelled S1, and on the C-terminal side S1′, following the nomenclature of Schechter and Berger [196]. Like-numbered pockets are formed by identical residues because of the symmetry of the enzyme.

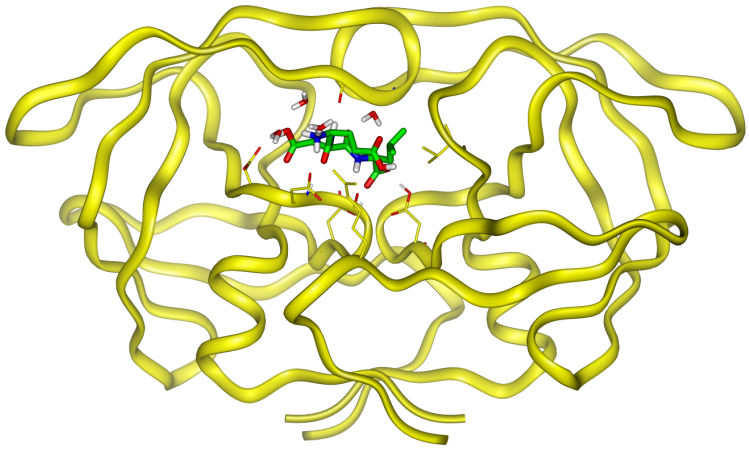


Figure 2.3: Ribbon diagram of the HIV-1 PR dimer with a bound inhibitor

Almost all complexes feature a buried water molecule bridging the NH groups of Ile50/Ile50′ and CO groups at P2 and P1′ of the inhibitor.

The catalytic activity of HIV-1 protease depends on a dyad of aspartic acid residues central in the active site, Asp25 and Asp25′. This makes it a member of the wider family of aspartic proteases, which have a characteristic two-domain structure (though retroviral proteases such as HIV-1 PR are smaller and are the only enzymes known to be symmetric homodimers). As observed in other proteases, the catalytic activity of HIV-1 PR is pH-dependent, with an optimum catalytic constant at pH 5–6 [109, 110]. The pK_as of the catalytic aspartate residues are known to be shifted substantially from the expected values ~4, to a distinct pair of apparent pK_as for the dyad as a whole, measured as ~ 3 and ~ 6 [109, 110, 227]. The distance between the inner oxygen atoms of the coplanar carboxylate groups observed in the crystal structure of the apoenzyme indicates the presence of an acidic proton bridging them [132, 170, 215, 249].

Inhibitors of HIV-1 PR are clinically useful drugs in the treatment of HIV infection. The development of these drugs was largely the result of structure-based ligand design, and crystal structures of HIV-1 PR-inhibitor complexes were of primary importance. Due to this history, there have been an unprecedented number of such crystal structures determined, making the HIV-1 PR system a particularly interesting one for modelling studies of molecular recognition.

Although the inhibitors of HIV-1 PR are structurally diverse, a number of features are common to almost all of them. Many of them were designed using a peptidomimetic strategy, beginning with a transition-state analogue of the natural peptide substrate. Despite this, the inhibitors with high affinity owe this to strong hydrophobic interactions, with relatively few hydrogen bonds. The strength of hydrophobic interactions is determined by the degree of steric complementarity, but describing this quantitatively is difficult.

Nonetheless, an adequate description of the electrostatic contributions to binding, and in

PDB code	inhibitor	res. Å	R factor	pH crys.	K_i nM	pH ass.	ref.
1a30	Glu-Asp-Leu	2.0	0.189	4.2	52000.	4.25	[139]
1aaq	Psi	2.5	0.190	5.0–5.6	4.	6.0	[65]
1hbv	SB203238	2.3	0.177	5.4–5.6	430.	6.0	[105]
1hpv	VX-478	1.9	0.192	5.4	0.6	NA	[123]
1htf	GR126045	2.2	0.193	3.4–5.4	8.	6.0	[102, 115]
1htg	GR137615	2.0	0.190	3.4–5.4	0.21	6.0	[102, 115]
1hvi	A-77003	1.8	0.181	6.2	0.012	6.2	[107]
1hvj	A-78791	2.0	0.158	6.2	0.004	6.2	[107]
1hvk	A-76928	1.8	0.177	6.2	0.011	6.2	[107]
1hvl	A-76889	1.8	0.186	6.2	0.112	6.2	[107]
1hvr	XK263	1.8	0.193	5.4	0.31	5.5	[5]
prmvt101	MVT-101	2.0	0.154	5.4	780.	6.6	[155, 156]
5hvp	acetyl-pepstatin	2.0	0.176	5.0–5.4	35.	5.0	[74, 190]
7hvp	JG-365	2.4	0.146	5.4	0.24	6.5	[220]
8hvp	U-85548e	2.5	0.138	5.4	3.	5.0	[114, 138]
9hvp	A-74704	2.8	0.177	5.4	4.5	4.7	[70]

Table 2.1: PDB codes and details of the HIV-1 PR-inhibitor crystal structures. The prmvt101 coordinates are an improved refinement at higher resolution of the complex in PDB 4hvp, and have not been deposited in the PDB (see text). The K_i assay pH for VX-478 is not documented. K_i values for GR126045 and GR137615 are scaled IC_{50} values.

particular, interactions with the catalytic aspartate residues, is important to understanding the affinity of these compounds. For many inhibitors, binding is known to be pH-dependent [74, 107, 123, 139], indicating that pK_a shifts take place on binding, and in a few cases, the apparent pK_a of the Asps has been determined experimentally for the complex, with substantial shifts observed [214, 244, 257]. A detailed experimental study of inhibitor binding to another aspartic proteinase, plasmapepsin II, using isothermal titration calorimetry found evidence for binding-related pK_a shifts of the catalytic aspartates, as well as the expected proton linkages [255]. The size of the contribution to the total free energy change of binding is unknown for most ligands, and as in the case of the binding of DMP-323 [227] the relationship between observed or calculated pK_a shifts and actual proton linkage may not be straightforward.

2.2 Methods

2.2.1 Modelling

Structures of HIV-1 PR complexed with inhibitor were taken from the Brookhaven Protein Databank [25, 26] and are listed in Table 2.1. Atom names were checked with WHAT CHECK [104] and corrected where necessary. Only structures with a resolution of 3 Å or better, and with an R factor of 0.20 or less, were used. The list of structures used is a revised version of the list of HIV-1 PR complexes appearing in the scoring studies of Eldridge et al. and Moret et al. [69, 161]. The complex with the inhibitor MVT-101, available as structure 4hvp in the

label	D25 Oδ1	D25 Oδ2	D25′ Oδ1	D25′ Oδ2	net charge
00					-2
10	+				-1
01			+		-1
20		+			-1
02				+	-1
11	+		+		0
22		+		+	0
12	+			+	0
21		+	+		0

Table 2.2: Dyad protonation states. Oδ1 and Oδ2 are the inner and outer oxygen pairs of the dyad, respectively. In the apoenzyme and in symmetrical complexes the states **10**-**01**, **20**-**02**, and **12**-**21** are equivalent.

PDB, was reported by the author to be in error [155], and re-published with an improved refinement, so in this case the newer coordinates were used [154]. Minor differences in sequence between the various HIV-1 PR structures did not involve ionizable residues in the vicinity of the active site.

The complexes were prepared in Insight 2000 [160]. Missing sidechains were added and water molecules and ions were removed, except for the groove water when present. Where alternative ligand conformations were given, the one with the highest occupancy was chosen for modelling (1htf: conformation 1; 1htg: conformation 300; prmvt101: conformation b; 5hvp: conformation 1). LIGPLOT analysis [241] of the active site was used to identify probable hydrogen bonds, and this information was used to place the protons of the catalytic aspartate dyad in the fully protonated state. An additional check was performed for possible hydrogen bonds where a carboxylic acid group acted as a donor. Other hydrogen atoms were given initial positions by Insight, and all hydrogen positions were subjected to 500 steps of steepest descent energy minimization (with heavy atom positions fixed) using the CFF91 force field [62, 143].

PARSE charges and atomic radii [209] were assigned for all amino acids. PARSE radii together with CFF91 charges were used for ligand atoms marked as HETATM in the PDB file, with aliphatic CH hydrogen charges merged with the carbon charge to maintain compatibility with the essential hydrogen scheme. Structures as well as atomic radii and charges were translated from Insight format into the necessary form for the pK_a calculation using a number of specially developed programs.

2.2.2 pK_a calculations

A PB titration model which had been previously applied to pK_a calculations on HIV-1 PR was used [227]. In this model, separate sets of partial charges for the neutral and ionized groups are used [11]. The solute dielectric constant was taken to be 4. A dielectric constant of 20 is more reliable on average [12], but the results of the previous study suggest that the lower dielectric constant is more accurate for the desolvated groups of interest in the HIV-1 PR active site [227]. The dielectric constant of the bulk water phase was taken to be 80. The

complex	ligand	model compound	pK_a
prmvt101_b	MVT-101	N-ethylethylamine	10.0
1hbv	SB203238	N-methyl-cyclohexylamine	10.3
1hpv	VX-478	N,N-diethylsulfanilamide	1.8
1htf	GR126045	2,5,5-trimethyl-1,3-thiazolidine-4-carboxamide	5.3
1htg	GR137615	2,5,5-trimethyl-1,3-thiazolidine-4-carboxamide	5.3
1htg	GR137615	2-ethylbenzimidazole	6.2
7hvp	JG-365	N-ethyl-2-methylpyrrolidine	10.6

Table 2.3: Model compound pK_as for ligand titrating groups

ionic strength was set to 150 mM for all calculations.

Additional procedures for dealing with the symmetry of the complex, and tautomerism of the active site aspartates were followed [227]. Each of four possible neutral states for the aspartate dyad (**11** to **21** in Table 2.2) was considered in separate PB energy calculations, and the resulting interaction matrices merged, with energies relative to a fully ionized reference state. Pairs of redundant interaction energies were replaced with single averaged values. Simultaneous protonation of both oxygens of a single aspartate was prevented by inserting a large artificial energy term (210 kJ mol^{-1}) in the interaction matrix. In order to correct for the statistical effect of counting two protonation sites per residue, the model compound pK_a is adjusted by $-\log_{10} 2$. Tautomeric forms of the basic secondary amine of MVT-101, the thiazolidine groups of GR126045 and GR137615, the benzimidazole group of GR137615, and the three carboxylate groups of Glu-Asp-Leu were also included in the calculations following an analogous procedure.

pK_as were determined from the interaction matrices using the cluster method [83] or a Monte Carlo simulation [14] where the cluster method proved unstable, or for the dyad residues when the pK_a fell outside the range of 0 to 15. Model pK_as for the protein sidechains were taken as: C-terminus 3.8; N-terminus 7.5; Asp 4.0; Glu 4.4; Arg 12.0; Lys 10.4; His 6.3; Tyr 9.6; Cys 8.3 [13]. Model pK_as for ligand titratable groups were taken from literature values obtained from the ACDlabs database [3] and are listed in Table 2.3. The pK_a value for the substituted thiazolidine is an empirical prediction made using the ACDlabs software [3]. Titration of the pyridyl groups of the inhibitors was not considered, as CFF91 parameters for protonated pyridine were unavailable. As these groups bind in the S3/S3′ pockets, no significant interaction with the catalytic aspartyl dyad will take place.

The apparent pK_a was defined as the pH at which the average protonation of a group was 0.5. For the symmetric structures (1hvk and 1hvr; the interactions between A-74704 and the aspartate residues in 9hvp were markedly asymmetric), where the aspartate residues are indistinguishable, pK_as for the aspartate dyad were defined as the pHs at which the sum of the average protonation of the various tautomers was 0.5 and 1.5. Monte Carlo simulations at the assay pHs were sampled to determine the relative occupation of the various tautomeric states of the aspartate dyad, and tabulated following the scheme in Table 2.2. Proton linkage energies were calculated using the general form of Equation 2.1.

PDB code	dyad		D29	D30	D60	inhibitor
1a30	3.4,	15.8	<0 <0	1.4, 3.7	3.5, 1.8	10.1, 4.2, >15, >15
1aaq	15.3,	10.0	0.0, 1.2	0.0, 4.6	3.4, 3.0	6.0
1hbv	25.7,	−8.2	1.0, <0	3.6, 2.6	4.4, 3.2	10.4
1hpv	6.8,	14.7	<0, <0	3.8, 2.8	0.4, 3.4	<0
1htf	6.8,	15.3	0.2, <0	2.4, 3.6	3.4, 4.4	<0
1htg	0.2,	23.2	2.0, 3.0	3.2, 3.2	4.2, 1.8	<0, <0
1hvi	9.4,	20.4	<0, <0	3.2, 3.6	0.0, 1.4	–
1hvj	25.4,	5.8	<0, <0	3.4, 2.6	1.6, 5.0	–
1hvk⋆	4.4,	22.7	<0, <0	3.6, 3.4	1.2, 1.8	–
1hvl	16.2,	9.0	<0, <0	3.4, 3.4	2.0, 1.2	–
1hvr⋆	1.2,	21.6	2.8, <0	3.6, 3.6	2.6, 4.8	–
prmvt101	22.7,	−10.8	0.4, 0.6	3.2, 3.8	0.4, 1.4	8.0
5hvp	4.0,	15.4	<0, <0	4.4, 5.0	3.6, 0.2	6.2
7hvp	21.2,	4.2	<0, 0.4	5.0, 2.6	1.0, 4.8	<0
8hvp	25.3,	4.4	3.0, 1.0	0.4, 3.6	3.0, 4.0	5.4, 6.2
9hvp	15.7	−1.3	3.6, 2.2	2.4, 2.4	4.6, 4.2	–

Table 2.4: Calculated pK_as for HIV-1 PR active site residues and ligand titratable groups. Apparent pK_as for the dyad are given in order of increasing value for the symmetric complexes (marked with an ⋆); otherwise values appear for the pairs of residues in the order as in the PDB file.

2.3 Results and discussion

Calculated pK_as are listed in Table 2.4. Results are shown for residues in the active site. Large pK_a shifts are predicted for all complexes, with at least one of the dyad pK_as becoming considerably more basic. This is a logical consequence of the desolvation of the dyad by the inhibitor. The calculations show the second dyad group in 1hbv and prmvt101 as strongly acidic, due to the presence of a positively charged ligand. Ligand titrating groups exhibit large deviations from the model compound values in about half of the cases.

The predicted predominant dyad protonation states at the tabulated assay pH, presented in Table 2.5, indicate that a number of different patterns of preferred protonation are possible. At acidic pH, a fully deprotonated dyad is not expected to be observed for any of the complexes. The commonly described state of a dyad singly protonated on one of the inner (Oδ1) oxygens is fully populated for 1hbv, 1htg, 1hvk, 1hvr, prmvt101 and 9hvp, and predominant for 1hpv, 1hvj, 5hvp and 8hvp. Protonation of both outer oxygens dominates for 1aaq, 1htf, 1hvi, and 1hvl. 1a30 and 7hvp show a preference for a single protonation of an outer oxygen. Population of both singly and doubly protonated states for 1a30, 1htf, 1hvj, 5hvp and 8hvp complicates further modelling studies, as the relative contribution of the two states to binding energetics is unknown.

Similar calculations have been performed previously on the complexes 1hvr and prmvt101 by Trylska et al. [227]. The RMSD for the non-dyad residues between the present and previously published results for 1hvr is 0.35 pK_a units. The direction of the dyad pK_a shifts is consistent for prmvt101, though the magnitudes differ; the previous results predict less

PDB code	pH	predominant protonation		
1a30	4.3	**02** 0.67	**22** 0.26	
1aaq	6.0	**22** 0.99		
1hbv	6.0	**10** 1.00		
1hpv	6.0	**10** 0.46	**02** 0.21	**01** 0.19
1htf	6.0	**22** 0.64	**02** 0.16	**10** 0.13
1htg	6.0	**01** 1.00		
1hvi	6.2	**22** 1.00		
1hvj	6.2	**10** 0.57	**22** 0.35	
1hvk	6.2	**10** 0.56	**01** 0.38	
1hvl	6.2	**22** 1.00		
1hvr	5.5	**10** 0.56	**01** 0.43	
prmvt101	6.6	**10** 1.00		
5hvp	5.0	**01** 0.44	**10** 0.32	**21** 0.14
7hvp	6.5	**20** 0.96		
8hvp	5.0	**10** 0.63	**22** 0.29	
9hvp	4.7	**10** 1.00		

Table 2.5: Predicted predominant protonation states (to 90th percentile) for the catalytic aspartate dyad. The codes (bold type) refer to the scheme in Table 2.2. The most populated states are given in rank order at the pH at which the K_is tabulated in Table 2.1 were assayed. The modal assay pH of 6.0 was chosen for the calculations for 1hpv/VX-478.

extreme dyad pK_as of −6.0 and 10.8. The direction of the shift for the secondary amine group of MVT-101 is reversed, though both results indicate the group will be fully protonated at assay pH. Comparison of the 1hvr dyad predictions is more puzzling. Trylska et al. predict a strong shift to higher pK_a for both dyad values, while the present calculations predict one group will become more acidic.

NMR titration data suggest that the diprotonated state is favoured for the HIV-1 PR–DMP323 complex [257], and it is expected that the analogue XK263 will behave very similarly. However, the picture is confused by the absence of pH-dependence in the K_i of DMP323, a result which Trylska et al. suggest brings into doubt the conclusion that the NMR data indicate a diprotonated state [227]. The inconsistency between the calculations is likely to be due to a difference in the modelling of the initial hydrogen-bonding pattern for the various tautomers. Force field differences may also play a part, as the previous study used QUANTA charges for XK263, along with CHARMM22 bonded energy terms for minimization of hydrogen positions, whereas CFF91 parameters were used in the present calculations. Dyad pK_as for the unliganded HIV-1 PR structures 3hvp [249] and 1hhp [215] (results not shown), where only PARSE charges were used and rotamer positions were not adjusted by hand, were within 0.5 pK_a units of the results of Trylska et al. in all cases.

Inconsistencies of this kind can be avoided by systematic optimization of the hydrogen-bonding networks for the various tautomers, and consideration of possible changes in proton positions in the various deprotonated states. These factors are ignored in the interests of efficiency in the Antosiewicz et al. method [11], resulting in a procedure with computational

demands which scale linearly in the number of titrating groups. The procedure of Yang et al. [258], as implemented by Nielsen [172] in WHAT IF [238], performs a full optimization of proton positions for each interacting pair of titrating groups [103] in all four possible protonation states, giving a more robust procedure at the cost of quadratic computation time requirements. This method was not used in the present study due to the difficulty of implementing the additional forcefield parameters required for the peptidomimetic inhibitors within WHAT IF.

The titration behaviour of the catalytic carboxylate groups in the complex of HIV-1 PR–pepstatin has also been studied by NMR [214]. The pK_a values for the dyad were measured as < 2.5 and > 6.5, with the lower pK_a thus considerable more acidic than the value predicted for acetyl-pepstatin here (4.0) from 5hvp. Pepstatin binds to HIV-1 PR approximately 50 times weaker than acetyl-pepstatin [74], and no structural data is available for the complex. The discrepancy may therefore be due to small differences in binding mode. Both experiment and calculation imply a strong preference for a monoprotonated dyad at the assay pH, although the authors of the NMR study propose that one of the outer oxygens is protonated, in contrast to the prediction in Table 2.5. An isothermal titration calorimetry study of the same system measured a proton uptake of 0.4 protons at pH 6.5 [254]. This is consistent with bound pK_a values of 2.5 and 6.5, and is too low to be consistent with a shift of the more basic pK_a to a value as high as 15.4, if the free enzyme pK_as of 3.5 and 5.5 can be considered accurate.

A subsequent *ab initio* molecular dynamics investigation of the active site protonation state gives an alternative interpretation of the NMR data [182]. Protonation of one of the inner oxygens was found to be more stable than protonation of one of the outer oxygens at the picosecond time-scale investigated, a result also seen in the present calculations. However, a doubly protonated state was calculated to give a better explanation of the observed chemical shifts, suggesting that both dyad pK_as become more basic in the complex. Such a shift, albeit small for the more acidic group, is predicted as shown in Table 2.4. Table 2.5 shows a doubly protonated state, corresponding to the favoured state from the Car-Parrinello molecular dynamics simulation, to be 14% populated at pH 5.0. A doubly protonated dyad would suggest a larger proton uptake than 0.4 at pH 6.5, but other groups may show compensatory shifts.

Various other studies have made predictions of the catalytic dyad protonation state for HIV-1 PR–inhibitor complexes without explicit reference to pH dependence. The crystal structures were analysed with respect to energetic and/or geometric considerations, meaning that predictions are relevant to the pH of crystallization. Rao et al. [187] predict a monoprotonated form for the complex 1hpv, with Asp25 neutral. This is in agreement with the most populated protonation state in the present results. Geometric analysis of molecular dynamics simulations of HIV-1 PR–MVT-101 has been used to predict the predominant protonation state in two studies [81, 97]. The more recent of these (using the prmvt101 data, rather than the earlier 4hvp structure) [81], nominated the **10** configuration in combination with a protonated ligand as most likely, again in agreement with the results in Table 2.5. Semiempirical quantum chemical calculations on the 1hvi complex gave forms **20** and **01** as the lowest and second lowest in energy [223], but did not consider doubly protonated forms, such as the one found to be dominant in these results. A number of models have been proposed based on different analyses of the 8hvp complex. The crystallographers give the **20** configuration as preferred [114], based on hydrogen bonding analysis, while geometric analysis of an MD simulation favoured a diprotonated form [97]. A later MD free energy simulation of ligand

inhibitor	pH^1	K_i^1 M	pH^2	K_i^2 M	ΔG_{link} expt kJ mol^{-1}	ΔG_{link} pred. kJ mol^{-1}
Glu-Asp-Leu	3.1	2.0×10^{-5}	4.2	5.2×10^{-5}	−2.4	−0.19
A-77003	4.7	8.4×10^{-11}	6.2	1.2×10^{-11}	4.8	13.
A-78791	4.7	3.5×10^{-11}	6.2	4.0×10^{-11}	5.4	9.6
A-76928	4.7	7.7×10^{-11}	6.2	1.1×10^{-11}	4.8	4.9
A-76889	4.7	1.0×10^{-9}	6.2	1.1×10^{-10}	5.4	13.
acetyl-pepstatin	4.7	2.0×10^{-8}	5.0	3.5×10^{-8}	−1.4	0.46

Table 2.6: Experimental and predicted proton linkage energies. K_i values measured at two different pHs provide an experimental ΔG_{link}. The prediction for the same pH change is derived from the calculated pK_a shifts of the catalytic aspartate residues. Experimental values are from the references listed in Table 2.1.

stereoisomer preference suggested configuration **01** to be optimal [42]. None of these are in agreement with the predicted dominant state of **10**, but these calculations do suggest a sizable contribution from the **22** form, which is predicted to be 29% populated.

Recalling the thermodynamic cycle in Figure 2.1 and Equation 2.1, the pH-dependence of binding can be determined from the bound and free pK_a values of the species. Experimental results and calculations show that shifts in the active site aspartate pK_as may make a large contribution towards this pH-dependence, however, other groups may also be involved, particularly ligand titratable groups. Accurate *ab initio* calculations of the pH-dependence are not possible in this case, because of the accumulated error from the large number of groups to be considered, and the difficulty of predicting pK_as for small, flexible molecules, such as the ligands described here. Prediction on the basis of experimental data alone is also unreliable in this case, as the complete titration behaviour of free and bound states has not been adequately characterised. Nonetheless, for some complexes, particularly those with non-ionizable ligands, consideration of dyad shifts alone may be sufficient. Table 2.6 lists experimental and calculated proton linkage energies for a number of complexes, with the calculations based on experimental free pK_a values for the dyad of 3.5 and 5.5, and bound values from Table 2.4.

Correspondence between measured and calculated linkage energies is poor, although, with the exception of acetyl-pepstatin, the sign of the change is the same. Large ligand pK_a shifts are expected for both Glu-Asp-Leu and acetyl-pepstatin, so the neglect of these in the calculated energies is one possible explanation for the discrepancy. The calculated linkage energies for the "A-7" series compounds appear to exhibit a trend towards exaggeration, which would be consistent with the tendency of the calculations to overestimate pK_a shifts in general. Furthermore, the absence of pH-dependent binding in the case of the cyclic urea ligands, despite experimental evidence for changes in dyad pK_a (as noted above) underlines the uncertainties of these comparisons; the same results are expected for the analogue XK263 included in this study.

2.3.1 Accuracy of pK_a predictions

Detailed comparisons with experimentally-determined pK_as have pointed out systematic inaccuracies in the approach followed here [12]. Additionally, unrealistically large shifts are

sometimes predicted, which are clearly incompatible with normal protein pH stability profiles. An important source of error is the assumption of a single, rigid protein conformation. Solvent-exposed groups can usually reorient freely, so a single modelled conformation is unlikely to be representative. Changes in ionisation state may also be coupled to conformational changes, and in general only one of these conformations will be observed in the crystal structure. Use of a high dielectric constant for the protein models the effect of the reorientation of fixed dipoles to some extent, and this probably accounts for the greater average accuracy [12].

Explicit consideration of tautomerism, as done here, leads to a notable improvement in accuracy, as does optimization of hydrogen bonding networks [173]. However, a more general approach to conformational rearrangement, in which multiple hydroxyl rotamers are considered [8], does not result in radical further improvements. Sensitivity to the empirical parameters of dielectric constant, partial charges, atomic radii, and small changes in atomic coordinates is probably an important remaining source of problems. A scheme has been developed in which the dielectric constant is adjusted according to the number of solvent-exposed residues [55], and it has been suggested that a model in which the dielectric constant varies continuously through space, with solvent exposed groups assigned a higher value than desolvated groups, will improve matters [35]. This is put in perspective by recent results of unprecedented accuracy achieved using a refinement of Tanford-Kirkwood theory [221], where electrostatic effects were treated with less detail, while extensive side-chain repacking was simulated [98]. A true dynamic treatment of conformational change, for instance grand canonical simulations of proton exchange, may eventually be needed to achieve good predictive accuracy, but such calculations will require further developments in force field parameterization and be computationally intensive. Some initial work in this direction has begun to appear in the last few years [16, 191].

Chapter 3

Electrostatic complementarity in HIV-1 PR–inhibitor complexes

3.1 Introduction

Accurate prediction of the affinity of a ligand for a protein receptor is difficult. Empirical approaches (e.g. "scoring") work quite well for certain classes of small, rigid ligand molecules, but are unreliable for peptides. This is a notable failure, as many endogenous ligands are peptides, and a number of peptidomimetic compounds are marketed as drugs, such as the inhibitors of HIV protease. It is not clear whether this means that the process of molecular recognition of peptide ligands is unusual, or if it only reflects a more general weakness of these empirical models. The case of the HIV protease inhibitors is particularly interesting, as known ligands range from peptides to highly hydrophobic peptidomimetics with reduced flexibility.

One sticking point in the analysis of the interactions in HIV-1 protease–inhibitor complexes has been the role of electrostatic interactions. The aim of this study was to see if careful modelling of electrostatics leads to a clearer picture of their contribution to affinity and specificity in this system. A regression analysis was done to look at the importance of a number of properties calculated from the crystal structures of the complexes, selected considering previous empirical studies of peptide ligands [18, 79, 161]. An measure of electrostatic complementarity was also tested as an independent variable in the analysis.

3.1.1 Electrostatic complementarity

Electrostatic complementarity is a counterpart to the steric complementarity implied by the lock-and-key model of receptor–ligand interaction. One possible definition is that the disposition of charge in the ligand should optimize the electrostatic interaction energy in the bound complex, assuming that the bound geometry will be primarily determined by steric factors. In concrete terms, electrostatic complementarity means that the number of salt bridges and hydrogen bonds should be maximised, with positively charged groups in the receptor approaching negatively charged groups in the ligand and vice versa [169].

Steric complementarity is the most important factor in ligand–receptor binding. Hydrophobic interactions are closely allied, and are thought to be the factor driving association in most cases [145]. In this scheme, electrostatic complementarity is important for determining specificity, and does not make a large contribution to affinity [53, 162]. The binding process is a result of the balance of solvation-related factors. Electrostatic complementarity is a necessary,

but not sufficient, condition for binding, as the interactions between polar groups and water are close to optimal for the free ligand and receptor [53]. While this view is supported by calculations showing that electrostatic interactions have a net destabilising effect for the majority of complexes [204], modelling suggests that it is possible to design ligands with favourable electrostatic binding energy for most receptors, without loss of specificity [44, 118, 119].

Molecular recognition has a physicochemical basis, but proteins and peptides have an evolutionary origin. This implies that the phenomenology of complementarity may have aspects beyond those imposed by the physics of non-covalent interactions [222]. The differences in shape complementarity between various classes of protein–protein interface are an example of this [259], and suggest that the selection criteria of the immune system result in interfaces which have different characteristics to those which have been shaped by a long process of natural selection. These differences can also relate to electrostatic interactions, as seen in studies of catalytic antibodies [17].

Calculation of electrostatic energies for a given complex is straightforward, but simplified models which capture the phenomenology of electrostatic complementarity are nonetheless interesting. Visualisation of the molecular surface, colour-coded by electrostatic potential, allows an intuitive assessment of complementarity [247]. Quantitative elaborations on this idea, which reduces electrostatic complementarity from 3D to 2D, have been proposed for use in docking and QSAR studies. Chau and Dean studied electrostatic complementarity in 34 ligand–receptor complexes, using the correlation coefficient of the ligand and receptor electrostatic potentials at a set of points on the ligand Van der Waals surface, and found significant correlation with negative slope in all but eight cases [40]. Superposition of independent ligand and receptor electrostatic potentials is not physically meaningful, so the relationship between these correlations and electrostatic complementarity is not rigorous. However, the method was shown to give a good yardstick of complementarity for a number of model systems [41].

McCoy, Epa and Colman used an extension of Chau and Dean's approach to investigate electrostatic complementarity at protein/protein interfaces [146]. The most important refinement they applied was to calculate the electrostatic potential using the Poisson-Boltzmann equation and a semi-microscopic dielectric model, thereby including a model of solvation effects. The authors found significant electrostatic complementarity at all twelve interfaces studied. The magnitude of the correlation did not relate directly to the number of salt bridges in the interface, and was not characteristic for a particular class of proteins.

There are some technical difficulties associated with the calculation of correlations between electrostatic potentials. The choice of surface at which the correlation is calculated is rather arbitrary. A common choice is the Connolly molecular surface [46] of the ligand or receptor, though similarity calculations have been made using a volume in the interface, rather than a surface [28]. Calculating the electrostatic potential at the molecular surface may lead to inaccuracies, due to the quantization of the boundary between different dielectrics. Use of a smooth transition between high and low dielectric constants alleviates this effect [39].

3.1.2 Master equation approaches

The mesoscopic continuum electrostatics model introduced in Chapter 2 can be used to estimate the free energies of solvation which are a large component of binding energies. Finite difference Poisson-Boltzmann calculations can be applied to an electrostatic free energy cycle

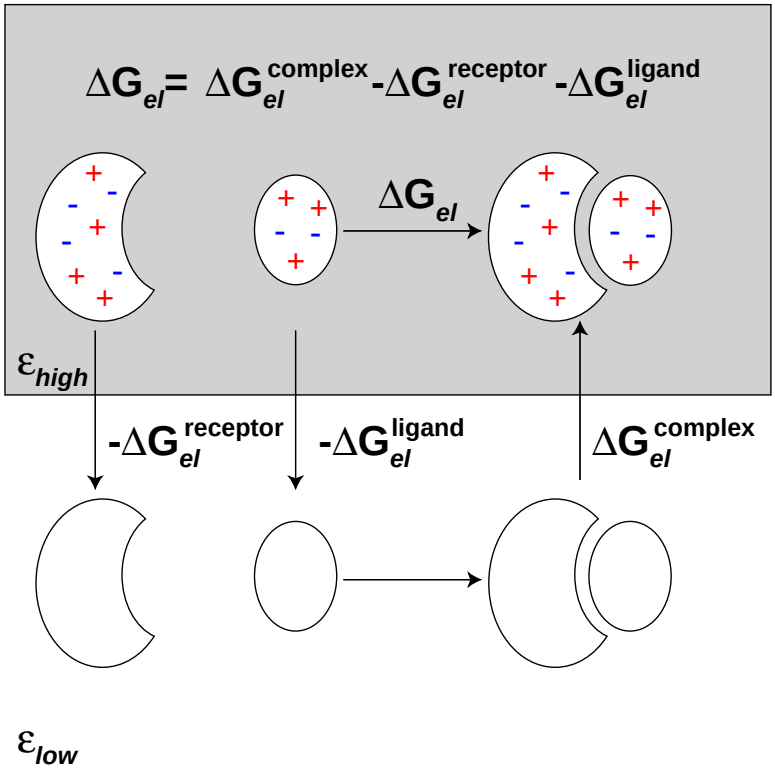


Figure 3.1: Electrostatic free energy cycle for rigid body association.

for rigid body association of ligand and receptor, as depicted in Figure 3.1. Combined with an estimate of the hydrophobic contribution to binding, calculated as a linear relationship with buried surface area, this is the basis for a number of different master equation (or partitioning type) scoring approaches (e.g. [79, 176, 219]). The most successful applications of this approach have been to the calculation of relative binding energies of complexes with mutants [176, 202, 219], but absolute affinities of small molecule and peptide ligands can also be predicted with moderate accuracy [79, 121, 183, 195], usually with consideration of changes of amino acid side chain entropy on binding.

Poisson-Boltzmann calculations to predict solvation and electrostatic interaction energies are quite sensitive to the details of the microscopic charge model. Ambiguities in the placement of protons, as discussed in the previous chapter, can lead to large energy differences, and energies are quite sensitive to small changes in atomic coordinates when a low dielectric constant is used for the protein interior [146]. As with pK_a calculations, the appropriate choice of dielectric constant is not obvious, as there is a tradeoff between physically realistic values which give accurate energies for fully desolvated groups, and higher values which implicitly compensate for electrostatic potential-induced conformational rearrangements [35]. A detailed study of the dielectric relaxation associated with charge insertion in an enzyme active site has highlighted the inconsistency between atom partial charges derived for molecular mechanics simulations performed at a dielectric coefficient of one, and the requirements of continuum models, which gave realistic results with a dielectric coefficient of 4–8 [208]. The PARSE parameter set, optimized for a dielectric of two, goes some way to addressing this problem [209].

3.1.3 Problems scoring HIV PR inhibitors

HIV PR inhibitors have a number of characteristics which are likely to cause problems for scoring algorithms. Foremost, they are extremely flexible molecules, with many freely rotating bonds. The conformational free energy difference between the bound and free ligand may be large, and is difficult to estimate. In addition, the interaction with the receptor appears to be driven by the hydrophobic effect, a contentious phenomenon. Less obvious are the systematic difficulties modelling the electrostatics of the HIV protease–inhibitor system. As discussed in the previous chapter, proton linkage can make a large contribution to the free energy of binding for many of these compounds, and the protonation state of the catalytic residues of the enzyme is unusual. The combination of these uncertainties makes it hard to find out the role of electrostatic complementarity in these interactions.

The general-purpose scoring functions of Böhm [29] and Eldridge et al. [69] reproduce the K_i values of HIV-1 PR ligand–receptor complexes badly. Moret et al. [161] found no significant correlation between Böhm-type scores and experimental affinities for a group of fifteen complexes. No trend to under- or over-predict was noted, and the standard deviation of the error in the predictions was 8.0 kJ mol^{-1}. The Eldridge scoring function is a useful benchmark, as a number of HIV-1 PR complexes were included in the training set, and a detailed analysis of the relative performance with different classes of complex was provided. The Eldridge aspartic protease training set (seventeen complexes) had a root-mean-square residual of 8.6 kJ mol^{-1}, but the ranking of affinities within the group is highly significant. An interesting feature of the regressions is that the intercept and hydrogen-bond coefficient specific to the aspartic protease set diverge noticeably from those of the other sets. The

complex with MVT-101 (PDB code 4hvp) was an outlier in both the Moret and Eldridge analyses.

Without a common test set, a direct comparison of the Böhm and Eldridge scoring functions is not possible. The one HIV-1 PR complex in the published Eldridge test set (A-74704, PDB code 9hvp) is predicted badly, with an error of 14.22 kJ mol^{-1}, and a separate test set of ten endothiapepsin complexes has an RMS residual of 12.58 kJ mol^{-1}, suggesting at least that the Eldridge function is unlikely to be radically better than the Böhm function. In the meantime, Böhm has published an improved scoring function [30] which may do better predicting HIV-1 PR affinities. The four HIV-1 PR complexes in the training set are fit well, with a standard deviation of error of 2.2 kJ mol^{-1}. Evaluation of the Moret HIV-1 PR dataset with the new Böhm scoring function (unpublished results) gave a highly significant correlation with experimental affinities ($r^2 = 0.56$, $P = 0.002$, Spearman $\rho = 0.78$, $P = 0.002$), and an RMS residual of 6.5 kJ mol^{-1} using the standard scaling (LUDI Score $= \log_{10} K_i$).

The VALIDATE scoring function uses a larger set of descriptors than the Eldridge and Böhm functions, and includes a number of explicit energy terms [100]. HIV-1 PR–inhibitor complexes make up 15 of the 51 structures in the training set. A test set of 13 modelled HIV-1 PR–inhibitor complexes with a wide range of activities was included in the evaluation of the parameterization, and was predicted with a RMS error of 5.0 kJ mol^{-1}. The authors qualify this impressive accuracy by noting that the training set is heavily biased towards this type of complex. Performance on a test set of thermolysin inhibitors was considerably worse, with an RMS error of 10.6 kJ mol^{-1}, with the problem ascribed in part to the difficulty of correctly dealing with a flexible ligand. Another noteworthy aspect of the scoring function is the low weight assigned to the electrostatic interaction energy, which contributes only 3% to the model. This might be seen as a result of inaccuracies in the treatment of electrostatic interactions.

As expected, regression analysis restricted to HIV-1 PR complexes performs much better than methods intended to be more broadly useful, presumably due to cancellation of errors caused by the simplified form of the scoring function. For example, the extensive CoMFA analysis of Waller et al. [178, 179, 242] achieves a RMS residual of 2.5 kJ mol^{-1}. A more surprising result is the relatively good performance of the empirical approach of Bardi et al. [18]. Their predictions have a standard deviation of error of 4.6 kJ mol^{-1} using a master equation parameterized from small compound data and protein databases. As with the equation of Froloff et al. [79], opposing terms are present which cancel to a large degree. The more convincing success of Bardi et al.'s approach may be due to the fact that the cancelling terms are primarily dependent on the same solvent accessible surface area changes, so the errors will be correlated.

None of these studies attempted to model the effects of protonation state on binding affinities and scoring parameters. Bardi et al. [18] considered all complexes to be equivalent to pepstatin or KNI-272 in protonation state, and that the proton linkage effect would be small at assay pH. In a 3D-QSAR study, Kulkarni and Kulkarni [125] carefully considered the protonation state of the active site dyad, but a single model was then applied to all complexes. The results in Chapter 2 and of Trylska et al. [227] suggest that these assumptions are incorrect for some ligands. Eldridge et al. [69] mention variation in assay pH as a source of error, implying that more detailed assay data are needed, and Oprea and Marshall make a similar comment with specific reference to HIV-1 protease inhibitors [177]. Moret et al. [161]

acknowledge that varying protonation states of the catalytic aspartate residues of HIV-1 PR are important for scoring, and suggest pK_a calculations as an aid to building more accurate models. Böhm [29, 30] and Waller et al. [242] do not address these issues.

The usefulness of a proton linkage correction term has been demonstrated for the calculation of the effects of point mutations on antibody–protein binding [202]. The change in free energy of binding associated with ten independent point mutations in the protein (hen egg white lysozyme) was predicted using a similar method to that of Froloff et al. [79]. Without consideration of pK_a shifts, the predictions were less accurate than the null hypothesis of zero change in affinity between mutant and native. Consideration of predicted pK_a shifts and the associated proton linkage energies gave a model which was better than the null hypothesis, despite the use of modelled structures for the mutants. In another example, Hansson and Åqvist [96] applied an assay pH-dependent correction to the benchmark data used in a molecular dynamics free energy simulation of HIV-1 PR inhibitors, improving the agreement with simulation-derived results slightly.

3.2 Methods

The HIV-1 PR–inhibitor complex models used in the previous chapter were also used for the present calculations. The protonation state of ligand titratable residues was fixed according to the predicted predominant state at the pH of the assay conditions, following the pK_a calculations. For electrostatic energy calculations, the catalytic aspartate dyad was modelled as doubly deprotonated to allow proton linkage energies to be considered consistently across all complexes.

3.2.1 Poisson-Boltzmann calculations

Finite difference Poisson-Boltzmann calculations were used to determine electrostatic energies and the electrostatic potential at the molecular surface. With the exception of the choice of protein dielectric constant, the procedure followed was that described by McCoy, Epa and Colman [146]. The program *qdiffxs* of DelPhi 3.0 [85, 171] was used to solve the linearized Poisson-Boltzmann equation, with the protein/ligand interior modelled with a dielectric of 8, and the surrounding solvent with a dielectric of 80. The choice of a dielectric constant of 8 for the protein was found to be optimal in a validation of a Poisson-Boltzmann master equation scoring function [195]. The ionic strength was zero. Parse atomic radius parameters were used [209] with CFF91 atomic partial charges [62, 143]. The molecular surface was defined using a probe radius of 1.4 Å. The finite difference calculation used a cubic grid of 201^3 points, with a percentage grid fill of 90%, giving a grid resolution of approximately 3 $Å^{-1}$. Debye-Hückel boundary conditions were used for the grid edges. When calculations were performed with only ligand or receptor present, dummy atoms were used for the absent species, to maintain the scale and orientation of the grid.

The electrostatic free energy change due to rigid body association, ΔG_{el}, was calculated using the free energy cycle in Figure 3.1. Following this scheme, the total electrostatic free energy change is equal to the difference in energies between the complex and the sum of the two isolated species. The electrostatic energy for each species is the sum of the energy needed to bring the point charges together (the Coulombic energy) and the energy to transfer the

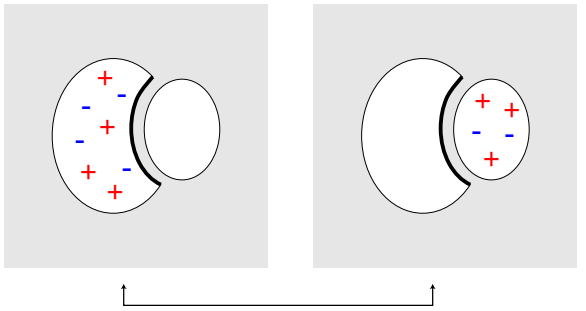


Figure 3.2: McCoy et al. scheme for electrostatic complementarity calculations. The electrostatic potential is calculated on the buried surface indicated by the thick line. The arrow indicates that the correlation is calculated between potentials derived from the two systems. The same procedure is used to calculate the correlation on the buried surface of the other species, and the two coefficients are averaged.

charges from a low dielectric to a high dielectric medium (the solvation, or corrected reaction field energy).

3.2.2 Proton linkage

Two measures of proton linkage were defined for each complex. As the weight of experimental and theoretical evidence indicates that the aspartate dyad must be at least monoprotonated in the bound state, a linkage energy for this obligate protonation at the tabulated assay pH was calculated using the following expression (derived from Equation 2.1).

$$\Delta G_{\text{ion}} = -\text{RT} \log \frac{10^{(\text{p}K_a)_f - \text{pH}}}{1 + 10^{(\text{p}K_a)_f - \text{pH}}}$$

The more basic of the experimental $(\text{p}K_a)_f$ values of the dyad (5.5) was used for all complexes. Consideration of the binding-related shift in the more acidic pK_a of the dyad is more difficult, as there is no clear indication of the relative binding energies of mono- and di-protonated forms, and there appear to be characteristic differences between complexes. In addition, the pK_a shifts calculated in the previous chapter appear to be exaggerated or in conflict with experimental data in some cases. Therefore a simple indicator variable, δ_{ion}, defined as 1 if the more acidic of the calculated pK_as was less than the experimental free pK_a of 3.5, and 0 otherwise, was chosen for further analysis.

3.2.3 Correlation of electrostatic potentials

A set of points on the molecular surfaces of the ligand and receptor, buried in the complex (the "interface"), was calculated using the MS software [46], with a resolution of 4 Å^{-2}, using the PARSE atomic radii, and a probe radius of 1.4 Å. The electrostatic potential at these points was calculated separately for the ligand and receptor using *qdiffxs*, and a correlation coefficient was calculated for each of the two buried surfaces. The Spearman rank correlation coefficient, ρ,

PDB code	ΔG_{exp} kJ mol^{-1}	ΔG_{el} kJ mol^{-1}	ΔG_{ion} kJ mol^{-1}	δ_{ion}	EC	ΔA Å^2	ΔA_{np} Å^2	ΔA_p Å^2
1a30	−24.45	30.61	0.455	1	0.550	−646.6	−398.0	−248.6
1aaq	−47.91	60.75	11.9	0	0.071	−1078.9	−716.9	−362.0
1hbv	−36.32	63.13	11.9	1	0.390	−1105.7	−839.9	−265.8
1hpv	−52.61	49.36	11.9	0	0.364	−986.8	−736.2	−250.6
1htf	−46.18	41.19	11.9	0	0.055	−822.5	−617.9	−204.6
1htg	−55.21	44.75	11.9	1	0.283	−1136.7	−845.0	−291.7
1hvi	−62.33	61.27	14.9	0	0.364	−1272.1	−953.6	−318.4
1hvj	−65.05	55.35	14.9	0	0.176	−1266.8	−964.2	−302.6
1hvk	−62.55	53.43	14.9	0	0.175	−1294.7	−969.8	−324.9
1hvl	−56.79	61.53	14.9	0	0.354	−1276.3	−964.0	−312.3
1hvr	−54.25	40.34	5.78	1	0.190	−1050.2	−859.6	−190.6
prmvt101	−34.85	31.11	21.7	1	0.417	−1285.7	−889.9	−395.7
5hvp	−42.55	53.40	2.29	0	0.288	−1186.0	−832.9	−353.1
7hvp	−54.88	65.12	19.9	0	0.356	−1338.5	−934.5	−404.0
8hvp	−48.64	81.20	2.29	0	0.236	−1326.7	−880.0	−446.8
9hvp	−47.62	42.13	1.22	1	0.331	−1206.9	−909.3	−297.6

Table 3.1: Free energy of binding derived from experimental K_i data and independent variables for regression analysis.

was computed using the *ctest* library [106] of the R statistics program [111]. Each species was considered partially desolvated by the volume of the other species in the complex, as shown in Figure 3.2. The final score of electrostatic complementarity for each complex, *EC*, was equal to minus the average of the two correlation coefficients, as defined by McCoy et al. [146].

3.2.4 Other parameters for regression

The molecular surface areas buried on binding, ΔA, were measured using the MS program, using the standard atomic radii and a probe radius of 1.4 Å. Surface area was defined as "non-polar" (ΔA_{np}) if it contacted carbon or CH hydrogen atoms, otherwise it was defined as "polar" (ΔA_p). The resolutions (*res*) and R factors (R) of the structures, as noted in the Protein Databank files, were also used as independent variables in the regression analysis, to indicate possible effects of systematic shortcomings of the crystallographic refinement. A significant correlation with resolution was noted in a previous study of HIV-1 PR complexes [161]. No parameters related to Van der Waals interactions, intramolecular strain, or configurational free energy were considered. Van der Waals interactions were considered subsumed by the surface area terms, while the other parameters are either difficult to estimate, or approximately the same for all compounds.

3.2.5 Regression analysis of ΔG

The ΔG_{exp} values derived from the experimental K_i values from Table 2.1 of the previous chapter, gave the values of the dependent variable in the regression analysis. Three multiple

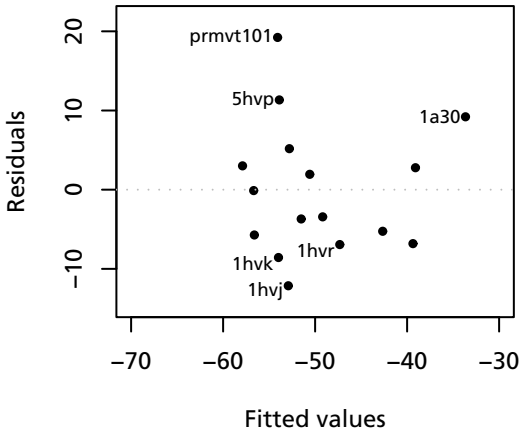


Figure 3.3: Residuals for the regression equation $\Delta G_{exp} = 0.32\Delta G_{el} + 0.04\Delta A - 13$. The six largest values are labelled.

linear regression analyses with stepwise removal of statistically insignificant terms were done using the *lm* function in the statistics program R. The independent variables were (1) ΔG_{el}, ΔA, ΔG_{ion}, δ_{ion}, *res* and R (giving a formulation similar to that of the master equation of Novotny et al. [176]) (2) *EC*, ΔA, ΔG_{ion}, δ_{ion}, *res* and R, and (3) ΔA_{np}, ΔA_{p}, ΔG_{ion}, δ_{ion}, *res* and R.

3.3 Results

Three of the variables listed in Table 3.1 are significantly correlated with ΔG_{exp}, ΔA ($r^2 = 0.32$, P = 0.02), ΔA_{np} ($r^2 = 0.48$, P = 0.003), and δ_{ion} ($r^2 = 0.29$, P = 0.03). The correlation with buried non-polar surface area (and by extension ΔA, which is significantly covariant) is to be expected given the hydrophobic nature of most of these ligands. Of the proton linkage parameters, the only significant covariance is between δ_{ion} and ΔG_{el} ($r^2 = 0.41$, P = 0.01).

Stepwise multiple linear regression following scheme (1) failed to find a regression relationship where the influence of ΔG_{el} was significant. The regression with the independent variables ΔG_{el} and ΔA was significant ($r^2 = 0.43$, P = 0.03), but the coefficient of ΔG_{el} did not differ significantly from zero (P = 0.14). The residuals of this regression equation are plotted in Figure 3.3. prmvt101 is clearly an outlier, as observed for MVT-101 in various previous scoring studies. In this case, inclusion of the proton linkage parameters in the regression did not help in explaining the unexpectedly poor affinity of this compound.

The analysis was repeated with prmvt101 excluded. A regression with the terms ΔG_{el}, ΔA, and ΔG_{ion} was found ($r^2 = 0.74$, P = 0.002) and is listed in Table 3.2. Removal of

	estimate	std error	t value	P(> |t|)
(intercept)	−10.32	9.6	−1.1	0.31
ΔG_{el}	0.38	0.19	2.0	0.07
ΔG_{ion}	−0.79	0.29	−2.7	0.003
ΔA	0.05	0.01	−3.7	0.02

Table 3.2: Best regression following scheme (1) with prmvt101 excluded. $s = 6.1$, multiple $r^2 = 0.74$, $F_{3,11} = 10.4$ (P = 0.002).

	estimate	std error	t value	P(> |t|)
(intercept)	−21.48	8.9	−2.4	0.04
ΔG_{el}	0.51	0.16	3.1	0.01
ΔG_{ion}	−0.53	0.23	−2.3	0.05
ΔA	0.05	0.01	−4.6	0.001
δ_{ion}	9.49	3.36	2.8	0.02

Table 3.3: Best regression following scheme (1) with prmvt101 excluded, and ΔG_{ion} and δ_{ion} corrected for 1hvr. $s = 5.1$, multiple $r^2 = 0.83$, $F_{4,10} = 12.48$ (P = 0.0007).

the ΔG_{ion} term results in a large decrease in the significance of the coefficient of ΔG_{el} (from P = 0.07 to P = 0.17), suggesting that the proton linkage energy is fulfilling its role as as a correction to the electrostatic energy term. No significant regression equation including both the ΔG_{el} and δ_{ion} terms was found.

As discussed in the previous chapter, the binding of the cyclic urea inhibitor DMP323 to HIV-1 PR is not pH dependent. The close analogue XK263 is expected to exhibit similar behaviour, which means that the proton linkage correction term ΔG_{ion} should be zero for the complex 1hvr, in contrast to the calculated value of 6 kJ mol^{-1}. The calculated pK_a shifts are also at odds with previously published experimental work and calculations. ΔG_{ion} and δ_{ion} were both changed to zero to be consistent with this information. A regression equation with the terms ΔG_{el}, ΔA, ΔG_{ion}, and δ_{ion} was the result ($r^2 = 0.83$, P = 0.0007), and is given in full in Table 3.3, with residuals plotted in Figure 3.4. Exclusion of the ΔG_{ion}, and δ_{ion} terms reduces the r^2 to 0.57 and the significance of the ΔG_{el} coefficient to P = 0.17.

Regressions involving *EC* following scheme (2) did not yield any significant relationships. With the corrections to the 1hvr parameters, scheme (3) resulted in an equation with the terms ΔA_{np}, ΔA_{p}, and δ_{ion} ($r^2 = 0.77$, P = 0.0004), shown in Table 3.4. The significance of the

	estimate	std error	t value	P(> |t|)
(intercept)	−25.24	9.7	−2.6	0.02
ΔA_{np}	0.05	0.01	−4.4	0.0008
ΔA_{p}	−0.05	0.02	1.9	0.08
δ_{ion}	10.94	3.3	3.3	0.006

Table 3.4: Best regression following scheme (3) with ΔG_{ion} and δ_{ion} corrected for 1hvr. $s = 5.9$, multiple $r^2 = 0.77$, $F_{3,12} = 13.21$ (P = 0.0004).

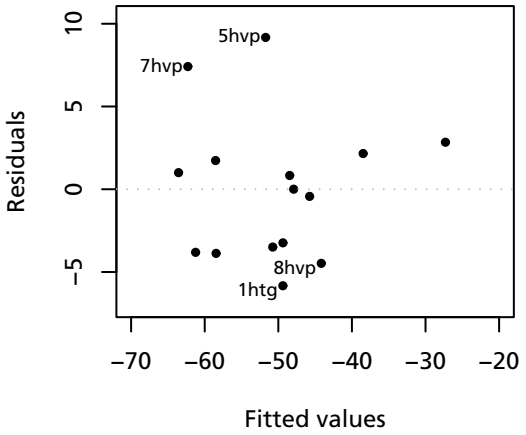


Figure 3.4: Residuals for the regression equation in Table 3.3. The four largest values are labelled.

ΔA_{p} term is marginal.

3.4 Discussion

The combination of electrostatic energy, buried surface area, and proton-linkage correction terms gives a regression equation with reasonable parameters for a data set spanning a wide range of activities and a number of structural classes. These results indicate that the continuum electrostatics approximation can be applied to sets of hydrophobic, peptidomimetic ligands. The difference in correlation coefficient between the regression including ΔA_{np} alone ($r^2 = 0.48$) and the regression including ΔG_{el} ($r^2 = 0.83$) indicates the contribution of electrostatic effects to binding. Inclusion of proton linkage in the analysis is necessary to give reasonable accuracy and statistically significant parameters. Both information on assay pH conditions and ligand-induced pK_a shifts contribute usefully to the model.

The electrostatic complementarity measure defined by McCoy et al. [146] did not prove useful for scoring in this context. This parameter is correlated with the Coulombic interaction energy in the complex, and does not fully encapsulate the information about the desolvation penalties which are so important to the overall magnitude of the electrostatic interaction energy. EC values are fairly low for this data set, with 1a30 having the highest value, reflecting the polar character of the ligand EDL. This parameter is probably more suited for analysis of protein–protein interfaces, where more variation in electrostatic complementarity is encountered, and for classification rather than quantitative applications.

The surface-area-based model of scheme (3) gives a reasonable regression. The dominance

of hydrophobic interactions means that the buried non-polar surface area alone explains close to half the variation in activity between ligands. The linkage term δ_{ion} makes a significant contribution, but the polar surface area appears to be somewhat unreliable as an additional parameter.

There is no ready explanation for the exceptional nature of prmvt101. The discovery of errors in the refinement of the previously deposited coordinates, 4hvp [155] provides an obvious reason why this complex was repeatedly found to be an outlier in earlier studies. However, other factors may be significant, including the low affinity of the compound, the presence of alternative binding modes in the crystal structure, or the relatively high pH (6.6) at which the K_i was measured. The interaction of the cationic secondary amine of the inhibitor with the catalytic aspartic acid groups, causing large pK_a shifts, may also be an important factor. Similar interactions are present in the complex 1hbv with SB203238, which was also found to be an outlier in the scoring study of Moret et al. [161]

3.4.1 Comparison with other work

The coefficients derived in the regression equations are quite stable and agree reasonably well with physical intuition. The coefficient of ΔG_{el} is steady at about 0.4, consistent with the results of Novotny et al. [176] (although a different dielectric constant was used in this study), and not far from the ideal value of one. Similarly, the coefficient of the proton linkage term ΔG_{ion} at approximately -0.5 is commensurate with ΔG_{el}. The buried surface area coefficient, at close to 50 J Å^{-2}, is only about a third as large as that found to be optimal in other studies [176, 195], although this may be in part an artefact of differences in the calculation protocol. It is difficult to give a physical interpretation of the coefficient of δ_{ion} of $\sim$ 10 kJ mol^{-1}. However, for comparison, this is approximately the difference in proton linkage energies at pH 6.0 between that for a pK_a of $\leq$ 3.5 and 6.5 at pH 6.0.

A regression approach allows a more accurate fit of the data at a possible cost to the transferability of the equations derived. The master equation approach of Schapira et al. [195] uses a similar electrostatic energy term to the one used here, together with extensive conformational sampling, and consideration of entropy. The results for a test set of 13 HIV-1 PR–inhibitor complexes (seven of which are included in the data set of this chapter) had an RMSD from experiment of 9.8 kJ mol^{-1}, considerably worse than achieved here, even though the dielectric constant, surface area coefficient and an additive correction constant (i.e. the intercept) were optimized for the data set in question. To allow the present results to be roughly compared with those of Schapira et al., the unscaled ΔG_{el} term was subtracted from ΔG_{exp} and the residual fit to ΔA. This yielded a residual standard error of 9.4 kJ mol^{-1} (prmvt101 excluded), a comparable performance, even though no entropy term was used. Correction with the unscaled ΔG_{ion} decreased the error to 8.1 kJ mol^{-1}, and inclusion of δ_{ion} in the regression gave a further improvement to 6.9 kJ mol^{-1}.

In the absence of data on an identical, independent test set, the work here cannot be compared closely with other published studies. The accuracy falls in the range achieved by the better general-purpose scoring functions, such as those of Böhm [30] and Bardi et al. [18], which is acceptable considering the small number of parameters included in the regressions. However, the regression relationship is probably not very robust, as the necessity of removing prmvt101 from the data set suggests. While the results show that improvement in the

calculation of electrostatic interaction energies is possible for the HIV-1 PR–inhibitor system, increased reliability for flexible ligands, and overall accuracy to the level of less than 6 kJ mol^{-1} appears to remain beyond the reach of general purpose regression and master equation scoring functions.

3.4.2 Alternative approaches

There are a number of alternative approaches to the semi-microscopic Poisson-Boltzmann model for the calculation of solvation and interaction energies. Free energy calculations using Molecular Dynamics or Monte Carlo simulations are the most physically realistic methods, but these are often prohibitively expensive due to slow convergence (see the introduction to Chapter 4 for further discussion). Results from free energy calculations generally agree with experimental measurements (e.g. for relative binding energies of the stereoisomers of an HIV-1 PR inhibitor [187]), but statistical analysis of the accuracy has been hindered by the difficulty of generating large, consistent data sets. It is also difficult to decompose free energy results into entropy and enthalpy components which can be compared with experimental results or empirical predictions. An additional problem is that results may be force field dependent, as seen in simulations with small solute molecules [137]. Current surface-area-based models of hydrophobic binding are unlikely to be as accurate as explicit solvent free energy simulations, though the semi-empirical nature of the charge and radius parameterization means that good results are possible at least with small solutes [209], and as shown by Bardi et al. are competitive with general purpose scoring functions [18].

Linear Interaction Energy (LIE) calculations, an extension of linear response theory, combine features of free energy calculations and empirical approaches [15]. These simulations are considerably faster than free energy calculations, so a statistical comparison with empirical scoring functions is plausible. The performance of LIE and related approaches in predicting small molecule solvation energies has been extensively tested, and was shown to be competitive with empirical, fragment-based methods for predicting $\log P$(octanol/water) [68]. LIE uses one to three empirical parameters to predict ΔG values, and these do not appear to be transferable from one protein–ligand system to another [240], making comparisons more complex. The accuracy of an LIE analysis of HIV-1 PR binding affinities [201] was comparable to regression approaches specialized for this receptor, although an earlier, less sophisticated study of three inhibitors gave much larger errors [96]. The advantage of LIE over conventional free energy calculations was not apparent, as an explicitly calculated entropy term (which needed a long simulation to converge) was included in the later study.

Studies of knowledge-based potentials, developed for protein folding applications, applied to scoring, indicate that this approach is competitive with other scoring methods. One of the first applications was to the system of HIV-1 PR and inhibitors, with one important motivation to develop a method which could bypass the need to consider protonation explicitly. Verkhivker et al. [231] found that the combination of an atomic contact score with an empirical solvation scale was able to differentiate between MVT-101 and JG-365, inhibitors with similar binding modes but differing considerably in affinity. Wallqvist et al. [243] followed a very similar approach, but parameterized their potential using a diverse set of high resolution protein crystal complexes, rather than HIV-1 PR complexes alone. This more general approach gave a reversed ranking for the affinities of MVT-101 and JG-365, but the overall

performance was nonetheless quite good, with a mean absolute error of ≈ 6 kJ mol^{-1}.

More refined methods of analysing molecular electrostatic potentials, or electric fields, may allow better qualitative and quantitative study of electrostatic complementarity. Use of a correlation coefficient calculation is intuitively appealing, but the details are fairly arbitrary. It may be possible to produce more predictive results by using a different surface, or even a volume, to define corresponding points. Other measures of complementarity than correlation coefficients may prove to be better; dipole and higher moments may also make an important contribution. It has also been suggested that the magnitude of the electric field vector at the molecular surface may act as a simple hydrophobicity index, allowing another type of correlation to be assessed [169].

3.4.3 Flexibility

Flexibility of the ligand and receptor is notable in its absence from this analysis. An implicit treatment, as with the use of an increased dielectric constant, will be inadequate when a substantial change in conformational behaviour occurs on binding. Molecular mechanics calculations have suggested that bound conformations are relatively low in potential energy compared to free [33], but the contribution to affinity might still be significant. In such cases, a predominant states model of conformational changes may be useful [84]. Whether this approach or a more complete one is chosen, conformational analysis of the ligand (and possibly the receptor) will be required.

For very flexible ligands, such as peptides, consideration of binding-related entropy changes is important. A static predominant states model does not help to achieve this. Entropy estimates can be calculated from a harmonic approximation to the potential energy surface, or using the "mining minima" method of Gilson and Head [99]. However, when there are many shallow minima, explicit sampling of the equilibrium ensemble is necessary. Sophisticated Monte Carlo techniques, such as configurational bias Monte Carlo, hold promise for doing this more efficiently.

Chapter 4

CBMC: implementation

4.1 Introduction

Empirical analysis of molecular recognition in terms of static complementarity has some substantial shortcomings when applied to drug design. The omission of a good account of solvation effects, or more generally, entropy contributions, results in a poor ability to predict binding affinity for many systems. There have been notable successes in the more limited question of docking (prediction of the preferred configuration of the ligand–receptor complex), but not with highly flexible ligands such as peptides. When empirical models are an inadequate description of the recognition process, an alternative approach is computer simulation with a detailed physical representation of the relevant microscopic system, using the theory of statistical mechanics to predict macroscopic properties.

Many properties of chemical systems can be conveniently predicted by computer simulation. Molecular dynamics (MD) [7] is an elegant approach to many problems [77], which allows the simple calculation of many thermodynamic properties, as well as dynamic ones such as diffusion coefficients. Equilibrium conformations observed in MD simulations can indicate possible modes for molecular recognition, as well as providing an explanation of results from NMR, circular dichroism or other spectroscopic experiments. Theoretically exact methods using MD also exist to calculate free energy changes such as equilibrium (dis)association constants, with an accuracy limited only by the realism of the empirical potential function.

Nevertheless, in many cases MD is unable to reproduce the behaviour of macroscopic, experimental systems, due to the extremely limited size and time scales which can be feasibly simulated [91]. Equilibration of protein folding, for example, is on a time scale of milliseconds for hen egg white lysozyme [186], while simulations of a single, solvated protein molecule have only recently reached the order of one microsecond in duration [66]. The long diffusion times or low successful encounter rates of other processes are a similar impediment. Simulations of diffusion controlled binding processes require time scales of about 100 ns [239], and orders of magnitude more simulation time will be needed where the complex forms with more difficulty.

Potential energy barriers (for instance, the formation of a transition state) are crossed only slowly in MD simulations at constant energy (micro-canonical ensemble; constant-NVE) or constant temperature (canonical ensemble; constant-NVT). In order to accelerate barrier crossing, various non-physical modifications to the potential can be used, short-circuiting impractically slow convergence of simulations which involve an activated process. Examples include umbrella sampling [225], and generalised ensemble techniques (such as en-

tropy sampling) [134]. Monte Carlo (MC) [152] simulations do not reproduce natural dynamics, so they may equilibrate more quickly than MD [200], though they also can become trapped in local minima on the potential energy surface. Many of the schemes used to improve the sampling of MD can also be applied to MC, though with MC there is also the freedom to use special, non-physical trial moves to this end.

Molecular dynamics has been the preferred method in most computer simulations of proteins and peptides. Monte Carlo simulations are of comparable efficiency for small solutes [116] but can become extremely inefficient for chain molecules in solution. On the other hand, even in well-thought-out MD simulations it is often clear that convergence of the properties of interest is not reached, despite complicated sampling schemes. With increasing computer speed, longer simulation times mitigate this problem, but given the many orders of magnitude to be bridged, improved algorithms are desirable. For example, a simulation of the binding of a peptide ligand to ribonuclease S, to predict the effect of a single residue mutation [174], gave unacceptably large statistical errors and failed to converge despite a total simulation time of more than 1 ns, and the use of additional experimental data to guide the simulation.

In part as a result of these limitations, computer simulations for medicinal chemistry applications have not achieved the successes seen in other areas of chemical physics. The only widespread use of free energy simulations is to predict the effects of small changes in ligand molecular structure on the association equilibrium (i.e. by thermodynamic integration or overlapping histogram/free energy perturbation methods [251]). Efficient simulation algorithms are needed to make more difficult problems tractable, with the ultimate aim of enabling the direct simulation of ligand–receptor association (e.g. by a grand canonical ensemble simulation). The aim in this chapter was to implement the configurational bias Monte Carlo (CBMC) algorithm, a method which can greatly improve the efficiency of simulations of flexible polymer chains, for peptides, with an eye for new applications in simulations of ligand–receptor interactions.

A number of force fields exist which were designed for MD simulations of proteins, and which describe the energetics of peptide conformations quite well [23]. Simulation results are nonetheless largely force field-dependent [216], so evaluation and comparison is far easier with the use of a widely-used potential. In any case, determining the parameters for a protein force field is difficult and labour-intensive. In this chapter the implementation of a CBMC algorithm using the CHARMM22 all-atom protein force field [36, 141] is described. Correct use of this potential requires specific attention to the details of the CBMC algorithm, and systematic validation that it is consistent with existing implementations. As the intention is to use CBMC for conformational sampling and free energy simulations, the reproduction of conformational probability distributions was chosen as the criterion for validation.

4.1.1 Configurational bias Monte Carlo

The Monte Carlo importance sampling algorithm introduced by Metropolis et al. [152] is an efficient method for determining ensemble averages of observable, equilibrium properties, requiring only the ability to calculate the potential energy for a given configuration. Successive states of system follow in a Markov chain, with transition from the old state to the new (a step or move) determined in two phases. First, a new, trial, configuration is generated; in the Metropolis scheme this is generated by adding an unbiased random displacement to one of

the particles in the system. Second, an acceptance rule determines if the next state in the Markov chain will be the new configuration, or a repeat of the old configuration. The rule in the Metropolis scheme gives an acceptance probability proportional to the ratio of Boltzmann factors (e^{-U/k_BT}) of the new and old configurations. If ergodicity is assumed (that is, all states which contribute to the ensemble are mutually accessible), then it is straightforward to prove that Metropolis MC samples a canonical ensemble.

Efficiency of the MC scheme is determined by the rate at which successive states explore phase space (the set of configurations which contribute to the ensemble). This is bounded by the size of the change in configuration generated by the trial move, as well as by the probability that the trial move will be accepted (the acceptance ratio). As these two quantities are to a certain extent complementary, maximum efficiency is achieved by balancing them; for Metropolis MC this is generally achieved with an acceptance ratio of about 50%. As with MD, phase space bottlenecks, such as a simple potential energy barrier, impede convergence by trapping the system for many steps. For simulations at high density, most MC schemes suffer from more obvious forms of trapping, marked by an extreme decrease in the acceptance ratio.

Many MC schemes exist which exploit a biased or non-physical generation of trial moves to improve sampling efficiency. Rosenbluth and Rosenbluth [193] introduced a biased sampling method for polymers, where conformations of idealised chain molecules on a lattice were generated using a self-avoiding random walk, with ensemble averages calculated using a specially weighted average. Configurational bias Monte Carlo (CBMC) [76, 180, 207] builds on the principle of Rosenbluth sampling, using a Rosenbluth-like rule for generating trial moves together with an acceptance rule which results in a canonical ensemble.

Smit et al. [211, 212] have used CBMC in simulations of phase equilibria of alkanes, and simulations of adsorption isotherms of alkanes in zeolites, where extremely slow equilibration times (laboratory experiments may take weeks to equilibrate [217]) rule out the use of a naive MD approach. Grand canonical MC (constant-μVT) is effective for short alkanes, but as chain length increases, the acceptance of trial moves approaches zero. CBMC avoids this problem by the biased generation of trial configurations which have low energy external (interactions between chains or with the zeolite) and internal (e.g. bond stretching and bending) interactions.

The first, simple, lattice formulation of CBMC has been extended with time to ever more complex continuum potentials. The CBMC literature on alkanes uses primarily united-atom potentials, with bond stretching, bending, and torsions as bonded terms. Treatment of branched molecules and multiple interdependent energy terms in the bonded potential, requires an elaboration of the basic continuum CBMC algorithm. Realistic simulations of molecules more complex than alkanes — peptides for example — involve the use of a more complicated force field, and these additions, such as the use of multiple torsions, give need for further revisions to the CBMC scheme.

4.1.2 CBMC for polypeptides

CBMC has been applied with some success to the conformational analysis of simple polypeptide models in a vacuum. Schofield and Ratner [199] used a CHARMM-like potential and a variety of non-physical-sampling Monte Carlo techniques, including CBMC of the pep-

tide backbone dihedral angles, biased according to the preferred regions of the Ramachandran map. The combination of these techniques yielded promising results, with good convergence for pentaglycine. The authors state that limiting the use of CBMC to the equilibration of backbone dihedrals is more efficient than using CBMC to generate the entire conformation; however, this approach precludes direct calculation of the excess chemical potential from the Rosenbluth factor (as well as grand canonical and Gibbs ensemble simulations), and makes simulations in the face of strong inter-molecular interactions, such as binding, impractical. The multiple Markov chain method (also known as parallel tempering) [82] is suggested as a route to further improvement upon the non-Boltzmann entropy sampling MC scheme employed.

Deem and coworkers describe a more conventional application of CBMC to peptides, with particular attention to the efficient simulation of cyclic peptides [54, 252, 253]. An AMBER-like potential was used, with the added simplification of fixed bond lengths and angles, and rotation only possible around σ-bonds, reducing the potential to only torsion and non-bonded terms. Equilibration of systems with constrained backbones was assisted by the use of rebridging/concerted rotation. Further improvements in efficiency were achieved with parallel tempering, and the addition of "look-ahead", inspired by the Meirovitch scanning method [148–150]. As backbone and sidechains are equilibrated in separate CBMC moves, insertion of a complete chain into the simulation in one move is impossible, resulting in the same limitations as the method described by Schofield and Ratner [199]. The radical simplification of the forcefield speeds up the simulations considerably, but as angle constraints cause significant changes in the behaviour of the force field [92], results cannot be compared directly with published results using AMBER, and are likely to be less realistic.

A number of other biased-sampling techniques for peptides have been described. Garel and collaborators have developed a recursive, static MC scheme for conformation analysis [20], using a potential similar to Deem, with biasing of backbone torsions according to the Ramachandran plot. A related technique is the pruned-enriched Rosenbluth method [75], although it has only been applied to lattice protein models. Derreumaux has applied diffusion process-controlled MC to a simplified peptide model, with biased exploration of the backbone torsion modes [56, 57], a procedure designed for global search for energy minima. The biased probability Monte Carlo/optimal-bias Monte Carlo minimization procedure of Abagyan and Totrov [1, 2] has similar features, with more extensive use of torsional biasing from empirical distributions.

4.2 Methods

4.2.1 CBMC

CBMC generates conformations from a canonical ensemble efficiently, by avoiding high energy internal and external interactions. Instead of random changes to the coordinates of a molecule, molecules are "grown", atom by atom, at each step choosing a configuration with a bias towards low energies (Figure 4.1). The effect of this bias is then removed by a specially constructed MC acceptance rule. The biased selection is done by generating a number of alternative positions for each new atom, and then choosing one randomly with a probability proportional to the Boltzmann factor of the associated energy.

A simple continuum CBMC algorithm concerns an unbranched chain of l linear segments.

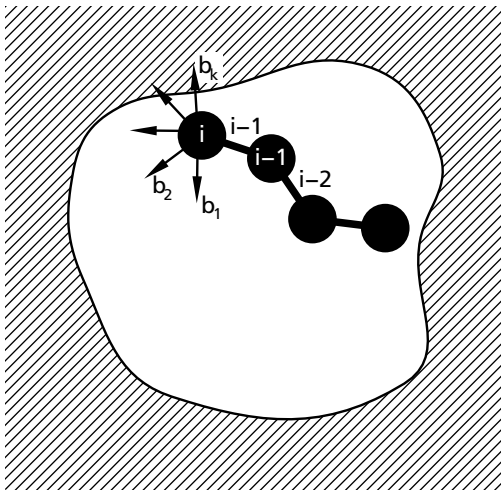


Figure 4.1: Scheme for biased growth of a new polymer chain. The k trial bond vectors, $\mathbf{b}_1, \ldots, \mathbf{b}_k$, are illustrated, one of which will be selected to make bond i from atom i to atom $i+1$. $\mathbf{b}_k$ will be chosen with a low probability, as it has a high potential energy.

Reproducing the formulation of the algorithm presented in [77], the potential energy is divided into U^{bond}, the sum of the bonded terms in the force field, and U^{ext}, accounting for all interactions with other molecules, as well as all non-bonded intramolecular interactions. To perform a Monte Carlo move, both a trial (new) conformation and the old conformation need to be considered. For the new conformation, assuming that $i-1$ bonds along the chain have already been grown, bond i is added as follows:

1. Generate k trial bond vectors, chosen from a Boltzmann distribution according to the bonded interactions with atom i (u_i^{bond}), giving the set $\{\mathbf{b}\}_k = \{\mathbf{b}_1, \ldots, \mathbf{b}_k\}$, where the probability of generating bond vector segment $\mathbf{b}$ is therefore

$$p_i^{\text{bond}}(\mathbf{b})\mathrm{d}\mathbf{b} = \frac{\exp\left[-\beta u_i^{\text{bond}}(\mathbf{b})\right]\mathrm{d}\mathbf{b}}{\int \mathrm{d}\mathbf{b}\exp\left[-\beta u_i^{\text{bond}}(\mathbf{b})\right]}$$

   with $\beta = \frac{1}{k_{\text{B}}T}$

2. One of the k bond vectors is selected with a probability

$$p_i^{\text{ext}}(\mathbf{b}_n) = \frac{\exp\left[-\beta u_i^{\text{ext}}(\mathbf{b}_n)\right]}{w_i^{\text{ext}}(\text{new})}$$

   defining

$$w_i^{\text{ext}}(\text{new}) = \sum_{j=1}^{k}\exp\left[-\beta u_i^{\text{ext}}(\mathbf{b}_j)\right]$$

This becomes the ith bond of the trial conformation

Once the entire molecule has been grown, the Rosenbluth factor can be calculated:

$$W^{\text{ext}}(\text{new}) = \prod_{i=1}^{l} w_i^{\text{ext}}(\text{new})$$

where the Rosenbluth factor of the first atom of the chain is

$$w_1^{\text{ext}}(\text{new}) = k \exp\left[-\beta u_1^{\text{ext}}(\mathbf{r}_1)\right]$$

with $\mathbf{r}_1$ the position of the first atom.

A similar procedure is used to calculate the Rosenbluth factor of the old configuration.

1. A molecule is selected at random and designated "old".
2. The Rosenbluth weight of the first atom in old is calculated

$$w_1^{\text{ext}}(\text{old}) = k \exp\left[-\beta u_1^{\text{ext}}(\text{old})\right]$$

3. The Rosenbluth weights of the remaining $l-1$ bonds are then calculated. For a given segment i the procedure is analogous to that for the new configuration.

   $k-1$ trial bond vectors are generated, which together with the actual bond from $i-1$ to i, make a set of k bonds, $\{\mathbf{b}_o, \mathbf{b}'_2 \ldots, \mathbf{b}'_k\}$, from which the external Rosenbluth weight of the segment can be calculated:

$$w_i^{\text{ext}}(\text{old}) = \sum_{j=1}^{k} \exp\left[-\beta u_i^{\text{ext}}(\mathbf{b}_j)\right]$$

4. The Rosenbluth factor of the old conformation is then

$$W^{\text{ext}}(\text{old}) = \prod_{i=1}^{l} w_i^{\text{ext}}(\text{old})$$

Once the new configuration has been generated, and the old and new Rosenbluth factors have been calculated, the Monte Carlo move is accepted with a probability of

$$\text{acc}(\text{old} \rightarrow \text{new}) = \min\left[1, W^{\text{ext}}(\text{new})/W^{\text{ext}}(\text{old})\right]$$

A proof that the algorithm samples a Boltzmann distribution can be found in [77].

4.2.2 BIGMAC

The program BIGMAC [233, 234] was developed in tandem with a united-atom force field for alkanes and zeolites, optimised to reproduce thermochemical properties such as adsorption isotherms. Bond stretching and bending are modelled by harmonic potentials (an earlier version

of the software used fixed bond lengths), and torsions by a four term power series expansion on $\cos\phi$. Non-bonded interactions consist of a Lennard-Jones potential, and, optionally, a Coulombic term with Ewald correction [72, 135, 136].

Various simulation types in a number of different ensembles are implemented, allowing the calculation of heats of adsorption, Henry coefficients, adsorption isotherms, vapour-liquid coexistence curves and so forth. Most importantly, a number of types of CBMC moves can be used, which in the simplest case give efficient simulations of long-chain alkanes in the canonical ensemble.

BIGMAC uses a number of efficiency optimisations made possible by the freedom of choice in which the force field potential is partitioned between the biased growth steps and the acceptance rule within the CBMC algorithm. Evaluation of bonded potentials is cheap, requiring a constant number of operations ($\mathcal{O}(1)$) per grown atom, so repeated evaluation of these to select favourable internal coordinates is not costly. On the other hand, long-range, non-bonded potential evaluation is much more expensive. The dual-cutoff algorithm [235] is an optimisation which moves evaluation of the most expensive terms to the acceptance rule, so they only need to be calculated once per grown chain .

BIGMAC implements a number of parallel algorithms, giving faster execution on multiprocessor computer systems; a parallel CBMC algorithm [71] is used. The parallel code is written using the portable MPI parallel programming interface [151], so it can easily be run on heterogeneous groups of workstations, or large clusters of low-cost PCs

4.2.3 CHARMM22 potential

The CHARMM computer program for macromolecular simulation, together with its empirical potential parameter sets, has been under ongoing development since the mid-1970s. There has been considerable cross-fertilisation between different protein force fields, and as a result, the AMBER, OPLS, CHARMM and GROMOS empirical energy functions have a lot in common. The CHARMM22 parameter set, dating from 1992, is the result of a number of rounds of improvement, and considerable effort in deriving parameters suitable for a variety of different types of condensed-phase simulations.

The CHARMM22 empirical energy function has the form

$$\begin{aligned} V_{\text{bonded}} &= \sum_{\text{bonds}} k_b(b-b_0)^2 + \sum_{\text{angles}} k_\theta(\theta-\theta_0)^2 + \sum_{\text{UB}} k_{\text{UB}}(S-S_0)^2 \\ &\quad + \sum_{\text{impropers}} k_\phi(\phi-\phi_0)^2 + \sum_{\text{torsions}} k_{\chi_n}[1+\cos(n\chi-\delta_n)] \\ V_{\text{nonbonded}} &= \sum_{i<j}\left(\epsilon_{ij}\left[\left(\frac{r_{ij}^{\min}}{r_{ij}}\right)^{12} - \left(\frac{r_{ij}^{\min}}{r_{ij}}\right)^{6}\right] + \frac{q_i q_j}{\epsilon_1 r_{ij}}\right) \end{aligned}$$

where the sums are over bonded atoms as indicated in Figure 4.2. The bond, angle, Urey-Bradley, dihedral angle, and improper dihedral angle force constants are k_b, k_θ, k_{UB}, k_ϕ, and k_{χ_n} respectively; b, θ, S, ϕ, and χ are bond length, bond angle, Urey-Bradley 1-3 distance, improper torsion angle, and torsion (dihedral) angle. A single torsion angle may have coefficients k_{χ_n} for multiple Fourier terms $n = 1, 2, 3, 4, 6$, each with a phase offset of δ_n (in practice always

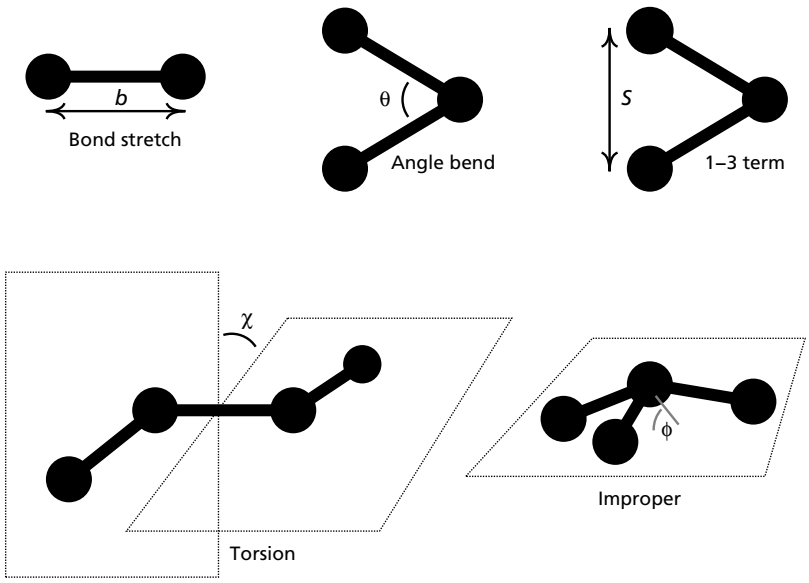


Figure 4.2: Bonded energy terms in the CHARMM22 potential

0 or 180°). All possible bond angle and dihedral angle terms are present, whereas improper torsions and Urey-Bradley terms are only defined for particular groups where necessary.

The non-bonded potential consists of a Lennard-Jones and a Coulomb term. Lennard-Jones terms are described by a well-depth, ϵ_{ij}, and a distance at the Lennard-Jones minimum, $r_{ij}^{\min}$, which between pairs of different atom types are combined using the Lorentz-Berthelodt rules (geometric mean of ϵ_{ij} and arithmetic mean of $r_{ij}^{\min}$). The Coulomb interaction is defined by the charges on the two atoms, q_i and q_j, at a fixed (vacuum) effective dielectric constant ϵ_1. In both terms r_{ij} is the distance between atoms i and j. Non-bonded interactions are over all possible pairs of atoms in the system, with the exclusions of directly bonded atoms, and atoms with two bonds separation. In some cases, a scaled-down well-depth is used for Lennard-Jones interactions when atoms are separated by three bonds (1-4 interactions).

The CHARMM22 functional form differs from the BIGMAC force field in a number of important ways: the use of a harmonic angle potential instead of a harmonic cosine potential; use of improper dihedral terms; use of Urey-Bradley terms; use of a Fourier torsion term instead of power series; use of special coefficients for the non-bonded interactions of 1-4 bonded atoms. The total number of parameters is considerably larger with on the order of a hundred different atom types.

As well as the original implementation, the CHARMM22 potential is also widely used in other molecular mechanics programs. However, many implementations do not match the genuine one completely, due to differences in parameters or terms missing from the potential (typically the special 1-4 terms and Urey-Bradley terms). For this reason, validation is neces-

sary before energies can be compared between ostensibly equivalent programs. The TINKER implementation of CHARMM22 has been carefully validated [184], and was used as the benchmark in this thesis.

4.2.4 Coupled-decoupled CBMC

The BIGMAC CBMC algorithm is incorrect if a pair of atoms defines the central bond in multiple torsion terms. In this algorithm, all bond angles centred on an atom are determined simultaneously, by doing a small spherical coordinate Monte Carlo simulation. This results in angles with the required Boltzmann distribution. After the bond angles have been chosen, the torsion angle is generated by a rejection scheme. However, if multiple torsions are present, bond angle bending terms can no longer be selected independently of torsion terms, and therefore the BIGMAC algorithm is incorrect [144].

This problem is avoided in the BIGMAC force field by ensuring that no more than one torsion term is defined around each pair of bonded atoms, with adjustments to the torsion and bending parameters of branched molecules to compensate. This approach works with the united-atom alkane force field and the types of simulations BIGMAC is designed for, but it is inappropriate for the accurate implementation of an existing, all atom protein force field. Redistributing energy between torsion and bending modes results in discrepancies in conformational energies, and would be tedious for a force field with many parameters, like CHARMM22.

There are various ways to avoid the multiple torsion problem, the most obvious of which are quite inefficient. For instance, the positions of all atoms bonded to a central pair can be generated simultaneously using a Boltzmann rejection scheme, or the torsion potential term can be included in U^{ext} and the number of trial moves increased. A better approach is to decouple the generation of bond angles, torsion angles, and so on, choosing them by sequential, biased selections, thus extending the CBMC idea to the internal degrees of freedom. An additional trick is to couple particular biased selections (for instance, on the torsional and Lennard-Jones energies), so that each biased selection stage sends multiple possible conformations to the next selection step.

The algorithm implemented by Martin and Siepmann [144] generates conformations using coupled torsion and Lennard-Jones selections, while splitting the choice of bond angles into a series of decoupled selections. Each growth step n proceeds from atom f, with the bonded atom which is not being regrown designated p (previous). grow_n is defined as the number of atoms bonded to f which need to be regrown in step n.

Bond angles are chosen in two stages, a and b. In stage a all angles $a - f - p$ are chosen independently for $a = 1 \ldots \text{grow}_n$ by a selection out of $\text{trials}^{\text{bend}}$ trials biased by the bending potential u^{bend_a}. In stage b, the angles $b - f - c$ for $b = 2 \ldots \text{grow}_n$, $c = 1 \ldots b - 1$ are chosen in $\text{grow}_n - 1$ biased selections on u^{bend_b}. The product of the individual Rosenbluth weights gives an overall Rosenbluth weight for the bending selections, w_n^{B}. Stated formally,

$$u^{\text{bend}_a} = u^{\text{bend}}(a - f - p)$$

$$u^{\text{bend}_b} = \sum_{c=1}^{b-1} u^{\text{bend}}(b-f-c)$$

$$w_n^{\text{P}_x} = \sum_{k=1}^{\text{trials}^{\text{bend}}} \exp(-\beta u_k^{\text{bend}_x})$$

$$w_n^{\text{B}} = \left(\prod_{a=1}^{\text{grow}_n} w_n^{\text{P}_a} \right) \times \left(\prod_{b=2}^{\text{grow}_n} w_n^{\text{P}_b} \right)$$

With all bond bending angles fixed, the torsion angle along $f-p$ is selected based on the sum of all relevant torsion potential terms. This biased selection is repeated $\text{trials}^{\text{LJ}}$ times, and from these orientations, the coupled selection on the Lennard-Jones potential is determined. Calculation of the associated Rosenbluth weights is then as follows:

$$w_i^{\text{T}} = \sum_{j=1}^{\text{trials}^{\text{tors}}} \exp(-\beta u_j^{\text{tors}})$$

$$w_n^{\text{L}} = \sum_{i=1}^{\text{trials}^{\text{LJ}}} \exp(-\beta u_i^{\text{LJ}}) w_i^{\text{T}}$$

The acceptance rule for the new configuration after all growth steps is then simply:

$$\text{acc}(\text{old} \rightarrow \text{new}) = \min \left[1, \frac{\prod_{n=1}^{\text{steps}} w_n^{\text{L}}(\text{new}) w_n^{\text{B}}(\text{new})}{\prod_{n=1}^{\text{steps}} w_n^{\text{L}}(\text{old}) w_n^{\text{B}}(\text{old})} \right]$$

The order in which the atoms are grown must be the same for the old and new configurations. There are also some special cases to be considered for the first two (re)growth steps.

4.2.5 Implementing CHARMM22 in BIGMAC

The BIGMAC force field uses a small number of parameters and a handful of atom types. All internal interactions for a particular molecule must be defined in the associated input file. Larger molecules with more complex topologies require unwieldy specifications, and when many parameters are added to the force field, it becomes impractical to prepare input files by hand. At the same time, writing code to automatically process and assign parameters of a force field like CHARMM22 is error-prone, and entails careful testing and validation.

Fortunately, the software package TINKER [184] proved to be a suitable source for most of the routines needed to add the CHARMM22 protein force field to BIGMAC. As the internal representation of molecules and force fields was quite similar in the two programs, it was straightforward to adapt TINKER parameter and topology file reading, and parameter assignment code to work with BIGMAC. At the same time, the BIGMAC energy evaluation functions were rewritten to match the CHARMM22 functional forms. The resulting program, which allows most of the original BIGMAC simulation techniques to be used with TINKER-format topologies and CHARMM22-type potentials, was given the name WHOPPER.

The additional complexity of the force field made changes to the CBMC algorithm necessary. First, for simplicity, bond lengths are fixed at equilibrium values. This constraint is used in most protein MD simulations, and appears not to cause any important change in the behaviour of the system [92]. Second, the improper dihedral, Urey-Bradley, and torsion terms around a particular atom are in general interdependent. Together with the bond-bending and non-bonded terms, they were included in a coupled-decoupled CBMC scheme.

The coupled-decoupled algorithm described by Martin and Siepmann [144] works well when all bond-bending terms centred on an atom are close to orthogonal at their equilibrium values, but becomes inefficient otherwise (unpublished results). In general purpose force fields and protein force fields such as CHARMM22, the observed equilibrium value for bond angles deviates significantly from the minimum of the individual bending terms, with the bending (and potentially other) terms around a central atom counterbalancing each other. Selection of bond angles in two decoupled stages results in the angles selected in the first stage being close to their individual minima, but the angles selected in the second stage being far from equilibrium, giving a low partial Rosenbluth weight. The probability of acceptance becomes very low, and so overall efficiency suffers.

This problem was solved by using the BIGMAC approach of a small MC scheme to choose the bond angles simultaneously (on the basis of the bending, Urey-Bradley, and improper dihedral terms) from an equilibrium distribution, in place of sequential biased decoupled selections. The final scheme is illustrated in Figure 4.3 and is referred to here as MC-CD-CBMC. As bond lengths and angles may be coupled in CHARMM22 via 1-3 Urey-Bradley terms (Figure 4.4), bond lengths cannot be drawn independently from a Boltzmann distribution as was done in BIGMAC. In future, variable bond lengths could be added to the scheme by including these in the small MC scheme.

BIGMAC calculates the Coulomb terms of the force field using Ewald summation [72, 135, 136]. The energy of a periodic system of point charges is split into two components, a "real-space" term equal to the Coulombic energy of the point-charges as if screened by Gaussian potentials, and a "Fourier-space" term or correction, which cancels out the energy due to the screening potentials, and includes the effects of periodicity. In BIGMAC, the real-space term is a pair-wise calculation with a distance cut-off, and is included in the biased selection with the Lennard-Jones potential. The Fourier-space term is more time-consuming to calculate, so it is evaluated once per grown molecule as part of the MC acceptance rule, using a parallelised algorithm.

The total time cost of Ewald summation electrostatic energy evaluation grows as $\mathcal{O}(n^{1.5})$ (where n is the number of interacting charges in the system), while conventional (non-periodic) pairwise evaluation with infinite cutoff grows as $\mathcal{O}(n^2)$. However, pairwise evaluation is considerably faster for systems with few atoms, so this option was added to the

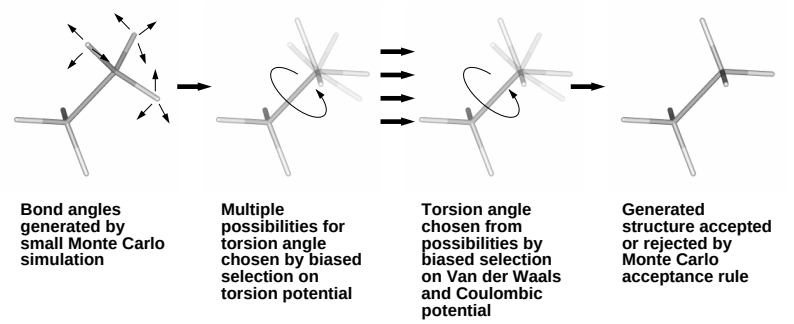
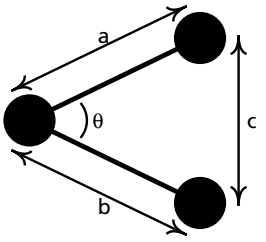


Figure 4.3: Monte Carlo-Coupled-Decoupled-CBMC



$c^2 = a^2 + b^2 - 2ab\cos\theta$
a and b bond stretches,
θ bend angle, and
c Urey-Bradley term

Figure 4.4: Interdependence of force field terms

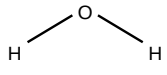
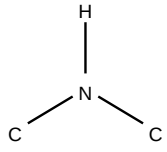
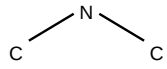
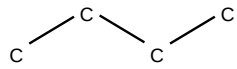
	term	structure	description	CHARMM22 types
a.	bend		TIP3P water	HT; OT
b.	improper dihedral		amide nitrogen	C; NH1; CT1; H
c.	Urey-Bradley 1-3		arginine $N^{\eta 1}$-C^{ζ}-$N^{\eta 2}$	NC; C
d.	torsion		butane	CT3; CT2

Table 4.1: Scheme for validation Monte Carlo simulations of individual bonded force field terms

WHOPPER code. As an experiment, an option of using the unscreened Coulomb potential in the biased selection was also implemented (the difference between screened and unscreened potentials was added to the Fourier-space energy in the acceptance rule). Inclusion of the Fourier-space term in the MC acceptance rule results in some reduction in the acceptance ratio, so shifting some of this energy to the biasing potential may in this way improve efficiency.

4.2.6 Validation

Initial testing of WHOPPER consisted of *ad hoc* comparisons of conformational energies of a series of structures, using TINKER (version 3.7, June 1999) as the reference (results not shown). As a more thorough test of both the force field implementation and the correctness of the CBMC algorithm, simulated probability distributions for each category of bonded potential term were then compared with the exact Boltzmann distributions at 300 K. In addition, the probability distribution of the C_1–C_2–C_3–C_4 torsion angle from an (all-atom) simulation of butane at 1000 K was compared with the distribution from a stochastic dynamics (SD) simulation, in order to check if the Lennard-Jones non-bonded term was correctly implemented, and as an initial validation of the MC-CD-CBMC algorithm for larger, branched molecules. A temperature of 1000 K was chosen to improve equilibration in the stochastic dynamics simulation. The Coulombic non-bonded term was tested in simulations of a united-atom butane molecule where charges of +0.3 and −0.3 were assigned to C_1 and C_4 respectively. The probability distribution of the C_1—C_4 distance was compared between WHOPPER and SD simulations at 1000 K.

The Monte Carlo simulations for the individual force field terms consisted of 320000 MC-CBMC regrowths, following the scheme in table 4.1. The all-atom butane and Coulomb test

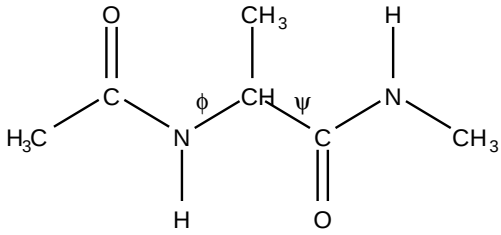


Figure 4.5: Alanine dipeptide, *N*-acetylalanine *N*'-methylamide. Backbone ϕ and ψ torsion angles are indicated.

Monte Carlo simulation consisted of 320000 MC-CD-CBMC regrowths. 100 trial positions were used for the decoupled torsion selections ($\text{trials}^{\text{tors}} = 100$), and 8 trial positions for the decoupled Lennard-Jones selections ($\text{trials}^{\text{LJ}} = 8$). The small MC scheme used 300 MC moves per placed atom. A cut-off of 9 Å was used for Lennard-Jones interactions. No Coulomb term was used for the all-atom butane simulation; tests of the Coulombic term were done with a non-periodic pairwise potential and no cutoff.

The stochastic dynamics simulation was performed using the program *dynamic* from the TINKER package, modified to use the Allen and Tildesley formulation of the velocity Verlet stochastic dynamics integrator [9], and to remove velocity scaling and solvent-accessible surface area calculation code. A friction coefficient (γ) of 6.5 ps^{-1} was used, with a step-size of 1 fs, and bond lengths were constrained to their equilibrium value by use of the "Rattle" algorithm [10]. The SD simulation length was 50 ns and torsion angles were tabulated once every 1 ps.

Estimates of the probability distributions were obtained from the tabulated data using the statistics program R [111]. Kernel density estimation [34, 230] was used with a Gaussian kernel, with the smoothing bandwidth chosen using the direct plug-in method [203]. Where comparisons were made between SD and CBMC results, the larger of the two selected bandwidths was used for both density estimates, to allow a convenient comparison of peak height.

Correlation time estimates for the small MC scheme used to sample bond angles were calculated for the various trigonal and tetrahedral groups present in the tripeptide Glu-Asp-Leu. The size of trial moves was chosen to give an acceptance ratio approximately equal to 0.4. 1500 cycles of two (trigonal) or three (tetrahedral) MC moves, equally distributed between azimuth and elevation moves on a randomly selected atom, were recorded per group, and correlation times for bending angles, 1-3 distances, and improper dihedrals, were estimated from block averages.

All simulations were performed under Linux on a 400 MHz Intel Celeron PC with 64 MB RAM. The g77 FORTRAN 77 compiler (version "0.5.25 19991030 (prerelease)") from the GNU compiler collection was used to compile WHOPPER and TINKER.

4.2.7 Alanine dipeptide

The blocked alanine molecule, (*N*-acetylalanine *N*'-methylamide), commonly known as alanine dipeptide, (see Figure 4.5) is a simple model compound for studying peptide conform-

ation. The probability distribution of the ϕ and ψ backbone torsion angles, namely the Ramachandran plot, describes the most important aspects of the conformational free energy surface. Details of numerous simulations of alanine dipeptide have been published, making it a useful benchmark [37, 224]. Gas phase and aqueous solution results using the CHARMM22 potential have been obtained from stochastic dynamics simulations [86, 210] and MD with adaptive umbrella sampling [213], using a variety of solvation models.

A blocked alanine model was constructed using standard CHARMM22 atom types, along with the charges published in [86]. The published SD simulation was performed using the UHBD program [142], and it is unclear if a full set of CHARMM22 potential terms was used. To ensure consistency of models, an SD simulation using TINKER *dynamic* was therefore chosen as the reference for evaluation of WHOPPER results. Despite the use of bond length constraints in this simulation, the Ramachandran plot was expected to be qualitatively similar to those in the literature.

The MC simulation consisted of 700000 MC-CD-CBMC steps at 1000 K, randomly divided with equal probability between full and partial molecule regrowths (where only a subset of atoms are given a new configuration). Non-periodic pairwise Coulombic interactions were used, with all other parameters as described for the previous simulations. Alanine dipeptide has one stereocentre, the C_α atom; the chirality of conformations generated by CBMC was constrained to be (*S*)-. Torsion angles were sampled every step. Two shorter simulations of 2000 steps were done to allow the acceptance ratios of the three different Coulomb potential evaluation schemes to be compared. The SD simulation length was 50 ns and torsion angles were tabulated once every 100 fs.

Two dimensional kernel density estimation with a bandwidth of 10° was used to give a probability distribution with a degree of smoothing comparable to the histograms in [86]. Convergence of the simulations was followed by plotting the autocorrelation function of the system potential and the characteristic decay time was estimated from block average plots [77]. Free energy minima were located on the benchmark Ramachandran plot, and standard deviations of the probabilities at these minima for both SD and CBMC were calculated by dividing the simulations into five blocks and making density estimates for each block.

4.3 Results and discussion

The probability distributions derived from CBMC simulations for the individual bonded energy terms closely match the exact Boltzmann distributions, with the observed minor discrepancies due to the statistical error of the simulations (Figures 4.6 and 4.7). Larger differences are visible between the MC-CD-CBMC results and SD results for butane (Figure 4.8), but the SD results have noticeable errors. Despite long simulation at a high temperature, with a relatively flexible molecule, the correlation time of the stochastic dynamics simulation is still quite long, and asymmetry in the SD probability distribution is noticeable. The match of distance probability distributions showing the effect of the Coulomb term is also good (Figure 4.9).

The bond angle MC autocorrelation times found for the various groups in Glu-Asp-Leu are shown in table 4.2. While the estimates are rough, they indicate that adding 1-3 and improper terms to the force field only causes a minimal increase in the characteristic decay time. The 300 cycles per placed atom used to sample the bond angle distribution in the other

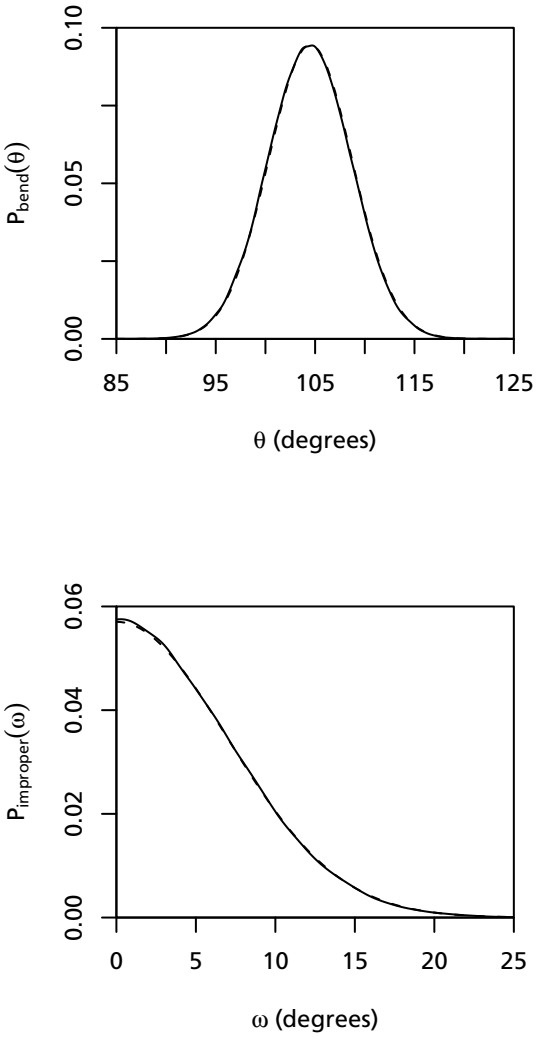


Figure 4.6: Probability distributions for CHARMM22 bonded potential terms at 300 K. WHOPPER observed distribution (solid line) and theoretical Boltzmann distribution (dashed line). Top: TIP3P water angle bending term (bandwidth 0.35°). Bottom: amide nitrogen improper dihedral term (bandwidth 0.58°).

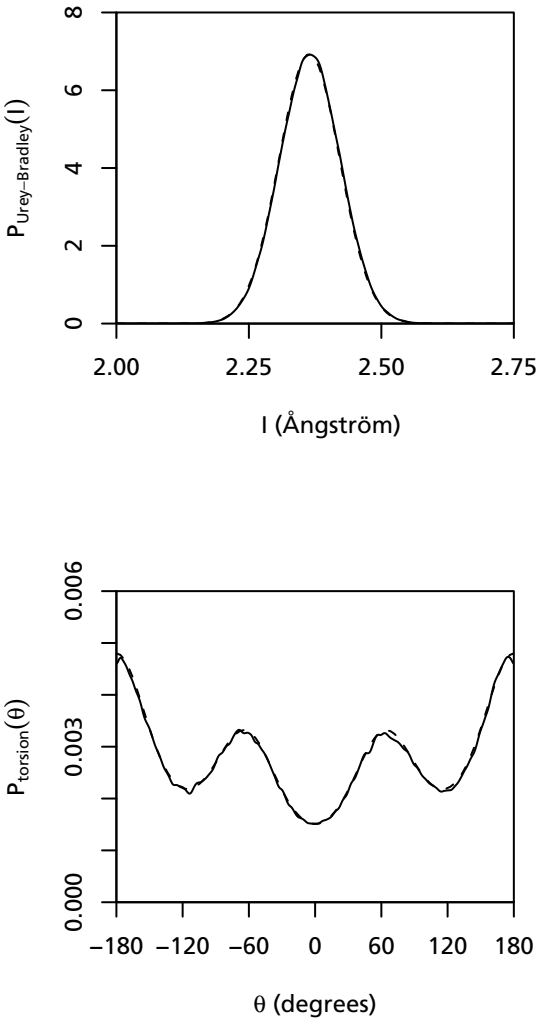


Figure 4.7: Probability distributions for CHARMM22 bonded potential terms at 300 K. WHOPPER observed distribution (solid line) and theoretical Boltzmann distribution (dashed line). Top: arginine N_ζ–C_ζ–N_ζ Urey-Bradley 1-3 term (bandwidth 4.8×10^{-3} Å). Bottom: butane C_1–C_2–C_3–C_4 torsion angle term (bandwidth 2.0°).

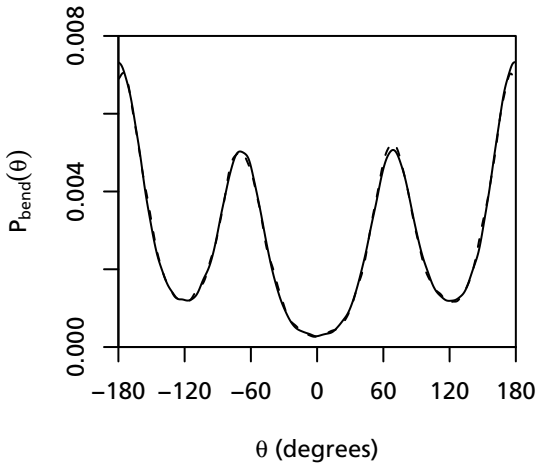


Figure 4.8: Probability distribution for C_1–C_2–C_3–C_4 torsion angle in all-atom butane model at 1000 K. WHOPPER observed distribution (solid line) and distribution from stochastic dynamics simulation (dashed line) (bandwidth 2.7°).

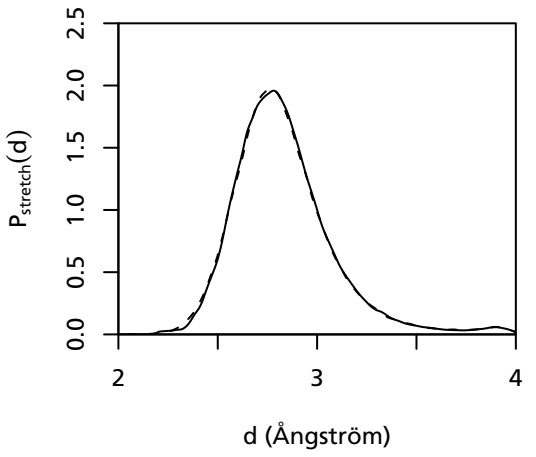


Figure 4.9: Probability distribution for C_1—C_4 distance in charged united-atom butane model at 1000 K. WHOPPER observed distribution (solid line) and distribution from stochastic dynamics simulation (dashed line) (bandwidth 0.018 Å).

	group	τ_c bend (cycles)	τ_c 1-3 (cycles)	τ_c improper (cycles)
a.	NH_3^+–	33	14	–
b.	–CH–	27	17	–
c.	$–CH_2–$	40	20	–
d.	$CH_3–$	33	40	–
e.	–COO	33	8	5
f.	–CO–	27	–	5
g.	–NH–	17	–	13

Table 4.2: Bond-angle MC correlation time estimates for groups in Glu-Asp-Leu

simulations should be more than sufficient to ensure that these terms are sampled from an equilibrium distribution.

The benchmark alanine dipeptide Ramachandran plot from SD simulation (Figure 4.10) appears very well-converged. The peaks are of similar size, location and shape to the published results [86], as expected. Once more, convergence is extremely slow, as shown by the presence of a long "tail" in the autocorrelation plot (Figure 4.13), and a characteristic decay time of 7 ps. It seems likely that the previously published results from a 2 ns SD simulation are inadequately converged, despite the length of the simulation, with a similar slow convergence observed in another SD study of a tripeptide [200]. Due to trapping, this results in a systematic error in relative peak heights, rather than the more obvious presence of "noise" in the results. The simulation here required approximately 20 hours of computer time.

The CBMC simulation shows no long tail in its energy autocorrelation function (Figure 4.13), as it is not so liable to trapping in local minima; the characteristic decay time is 45 MC steps. Convergence of the simulation is slowed by a reduced acceptance rate, and this is reflected in noticeable noisiness in the probability distribution (Figure 4.11), clearly visible in the difference plot (Figure 4.12). Use of the Coulombic energy term has a marked effect on the acceptance rate, reducing it from 0.17 to 0.10 (table 4.3) as strong non-bonded interactions either disturb the bond-bending distributions, or have a non-local effect on torsion distributions via multiple bonds. The CBMC simulation took about six days of calculation time.

The peaks of the benchmark probability distribution are located close to $(-100, 140)$ and $(70, -85)$, corresponding to the β and C_7^{ax} conformations in the usual nomenclature. The probabilities at these peaks are $6.64(\pm0.21) \times 10^{-5}$ deg^{-2} (±s.d.) and $1.41(\pm0.15) \times 10^{-5}$ deg^{-2}, a free energy difference of $12.9(\pm0.9)$ kJ mol^{-1}. The CBMC simulation gives $6.61(\pm0.55) \times 10^{-5}$ deg^{-2} and $1.12(\pm0.10) \times 10^{-5}$ deg^{-2} at the same coordinates. These results confirm that the CBMC results agree with the benchmark to within the bounds of statistical error.

The CBMC simulation does not equal the precision or efficiency of the SD benchmark simulation. Assuming the statistical errors are proportional to $1/\sqrt{n}$, where n is the number of MC moves, a CBMC simulation roughly seven times longer would be required to equal the precision of the benchmark. Taking into account the number of non-bonded trial moves (nchlj), this would mean that the CBMC simulation required as many non-bonded energy evaluations as SD. The number of bonded energy calculations would be on the order of a hundred times higher, as reflected in the relative execution times of the simulations reported

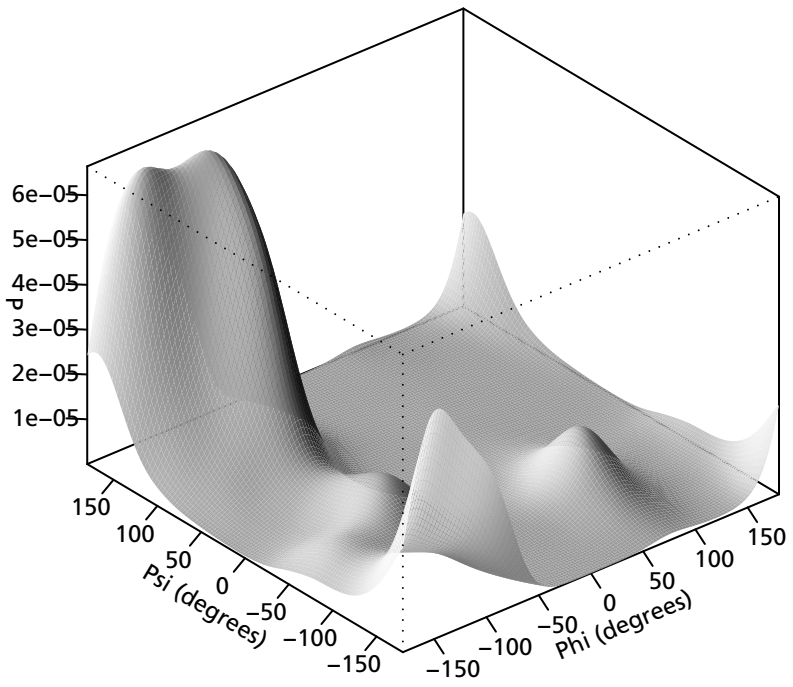


Figure 4.10: Ramachandran plot of alanine dipeptide backbone ϕ, ψ torsion angle probability distribution. 50 ns Stochastic dynamics simulation (benchmark).

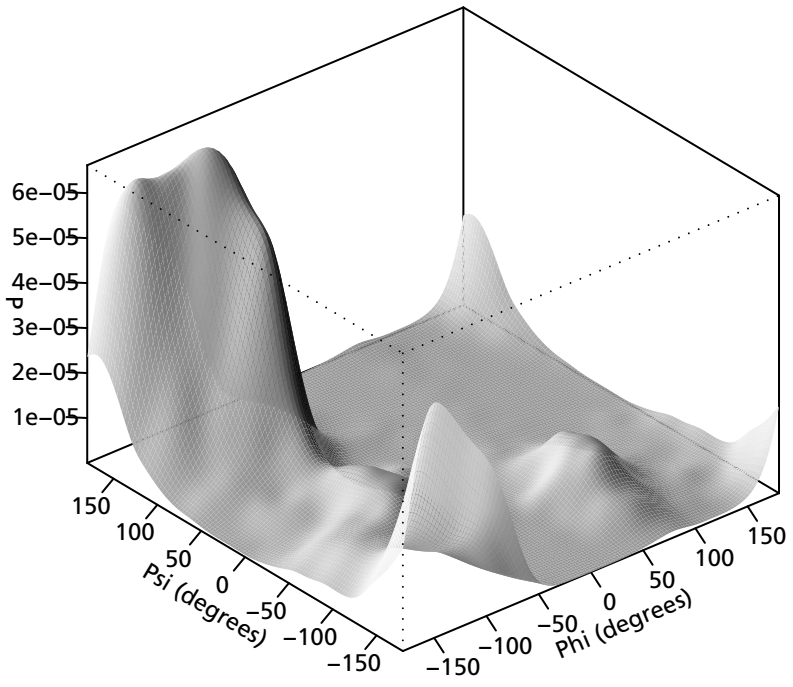


Figure 4.11: Ramachandran plot of alanine dipeptide backbone ϕ, ψ torsion angle probability distribution. WHOPPER MC-CD-CBMC simulation.

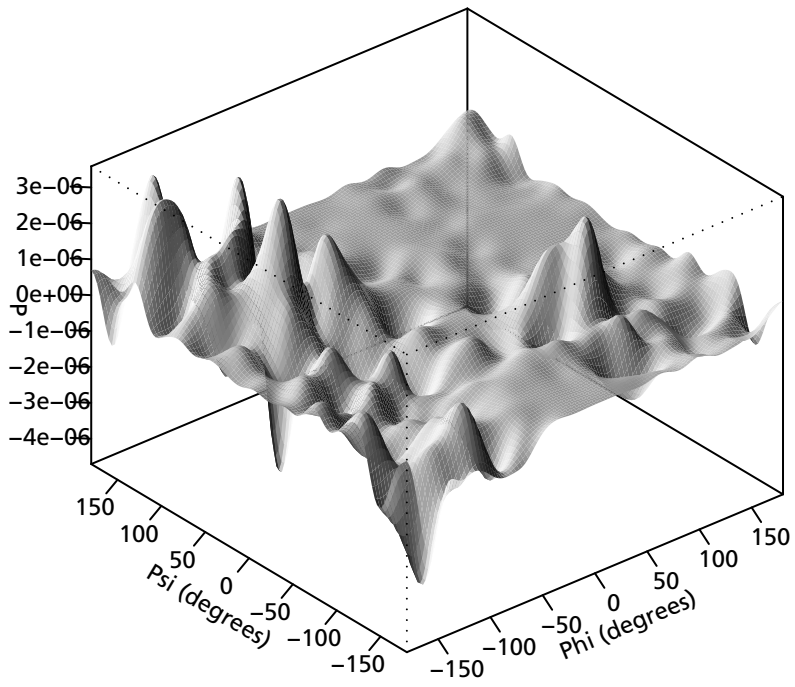


Figure 4.12: Ramachandran plot of alanine dipeptide backbone ϕ, ψ torsion angle probability distribution. Difference between distributions from stochastic dynamics and WHOPPER.

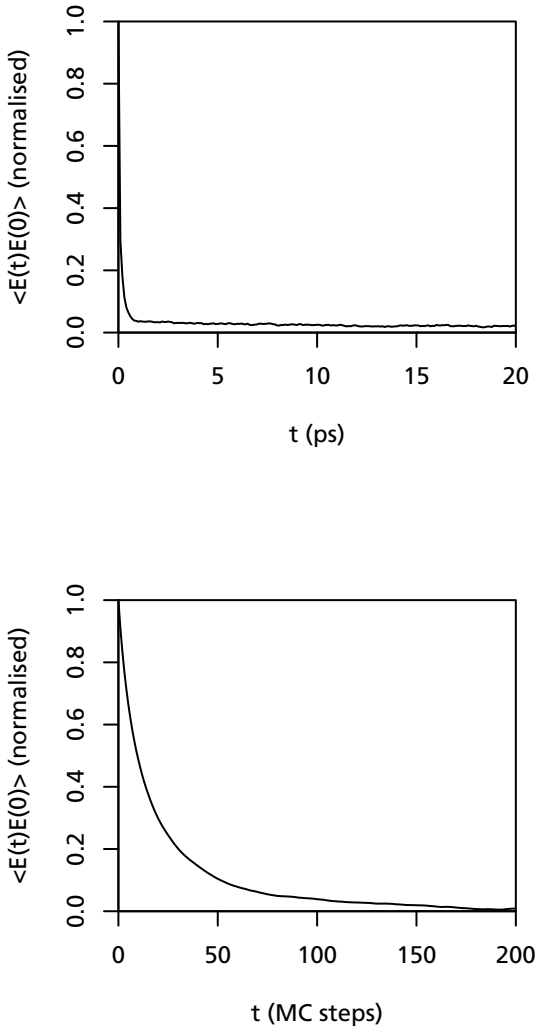


Figure 4.13: Autocorrelations of the potential energy in alanine dipeptide simulations. Top is stochastic dynamics simulation, bottom is WHOPPER.

scheme	acceptance ratio	CPU time (seconds)
no electrostatics	0.17	9210
Ewald	0.10	10200
Coulomb	0.06	9450
Ewald (selection on unshielded potential)	0.06	10060

Table 4.3: Efficiency of Coulomb evaluation schemes for alanine dipeptide

here. However, the cost of bonded energy evaluations will be relatively small with larger systems, making this less of a concern in practice.

At lower temperatures (e.g. 300 K) SD exhibits extremely slow convergence [90, 200], so a comparison of CBMC with other techniques, such as adaptive umbrella sampling, is more appropriate. High energy regions of the free energy surface are poorly sampled by CBMC, but for high Boltzmann weight conformations, CBMC appears to have comparable efficiency to umbrella sampling. For instance, for the calculation of the free energy difference between the two lowest minima on the Ramachandran plot of alanine dipeptide [213], CBMC requires approximately the same number of non-bonded energy evaluations to achieve a result of the same precision. Simultaneous umbrella sampling of numerous internal degrees of freedom is impractical, so for conformational analysis of larger peptides, CBMC should have the same attractive features as parallel tempering, entropy sampling and other generalised ensemble techniques.

Comparison of the different Coulomb evaluation schemes in table 4.3 indicates that the original Ewald implementation in BIGMAC is the most efficient for this system. The additional Fourier-space sum does not seem to use a large amount of extra CPU time compared to the non-periodic pairwise Coulomb evaluation, even for a system of this modest size. This result may not be generally applicable, though strong fluctuations in the Coulomb energy have been noted as a possible source of problems in another biased MC study of peptides [20].

4.4 Conclusions

CBMC algorithms have previously been applied to simple force fields appropriate for highly flexible molecules such as alkanes. Implementation of a standard protein force field demonstrates that in principle CBMC is more widely applicable, though a more involved algorithm is required when there are numerous interdependent terms in the bonded potential. The additional energy evaluations necessary for a coupled-decoupled scheme, and for a small MC scheme, reduce the speed-up obtained to some extent, but the goal of performing fewer expensive (non-bonded) energy evaluations at the cost of more inexpensive (bonded) evaluations remains intact. The efficiency is comparable with umbrella sampling for the evaluation of potentials of mean force, making it interesting for applications in conformational analysis. The speed-ups possible from parallelisation and dual-cutoff optimisation remain attractive.

The CBMC algorithm used in WHOPPER is also suitable for other protein force fields (e.g. AMBER 94 [50]) and even general-purpose force fields with cross-terms, such as MMFF94 [93]. However, the decoupling of torsion angle selection from the other bonded terms will result in inefficiency (namely, low acceptance ratios), if these are appreciably non-orthogonal.

There may also be better formulations of the small MC scheme possible, particularly for non-tetrahedrally-hybridized atoms, or where a combination of stiff and flexible modes are present.

4.4.1 Limitations

The major deficiency of WHOPPER (which is characteristic of many CBMC algorithms) is an inability to perform simulations of cyclic molecules, from simple benzene rings to cyclic polypeptide chains, as the stepwise growth algorithm has no provision to close rings. Concerted rotation schemes for cyclic peptide chains have been described and implemented [54,199,253], and more recently an algorithm which can be applied to arbitrarily large cycles [228], but these involve an expensive polynomial root-finding operation. For small, relatively stiff rings with a few low energy conformations, a simpler approach should be possible. Introduction of aromatic rings into the simulation scheme as rigid subunits seems a reasonable approximation, which would allow simulation of a much broader range of peptides without radical changes to the CBMC algorithm. Other possibilities would be the use of another small MC scheme, or a library of precalculated conformers.

Most Monte Carlo algorithms become inefficient when dealing with systems at high density, such as liquid water at 300 K. Equilibration becomes very slow, and the acceptance ratio of trial moves may become impractically low. Insertion of a molecule into the system (e.g. in a grand canonical simulation) depends upon the presence of a suitably-sized cavity, which for larger molecules and higher densities is very unlikely. A CBMC move under the same circumstances will almost always result in the regrowth of a conformation (nearly) identical to the old one. Explicit simulation of the solvent for a system like a peptide in water is therefore impractical.

A related problem is that the selection of atom positions in a growth step can affect the choices available in all subsequent growth steps, forcing them into energetically unfavourable regions (a similar effect is seen in Metropolis MC when it is performed in torsion space). If partial regrowth moves are possible, the result will be that chain ends will equilibrate more quickly than middle segments. Adding "look ahead" to the CBMC algorithm (along the lines of Meirovitch's scanning method [148,149]) is one way to deal with this problem (see further the discussion in the following chapter), but the use of hybrid MC moves [67], or concerted rotations [63] also helps.

Despite the ability of CBMC to avoid trapping in certain types of local minimum, the presence of numerous non-bonded interactions like hydrogen bonds can stabilise the system to such an extent that correlation times become very long. The methods to increase barrier-crossing mentioned in the introduction, for instance parallel tempering, can also be used with CBMC when this becomes a problem.

Chapter 5

CBMC: applications

5.1 Introduction

The flexibility of polypeptide chains means that thorough conformational analysis is a necessary ingredient of any study at the molecular level. The development of peptide or peptidomimetic ligands by structure-based methods requires insight into 'the bioactive conformation', while predictive simulations require adequate sampling of the relevant conformations. Conformational analysis by unconstrained search is futile, as the numerous internal degrees of freedom combine to give astronomically large numbers of possibilities. Monte Carlo (MC) and molecular dynamics (MD) simulations are more workable methods, but nonetheless, slow equilibration is the rule; a particular problem are the potential energy barriers impeding the inter-conversion of various low energy conformations.

As described in the previous chapter, configurational bias Monte Carlo (CBMC) sampling is an efficient simulation method, which may have advantages for the conformational analysis of peptides. One particular application is docking [126], a simplified form of free energy simulation. In this chapter, the binding of a tripeptide to a rigid protein receptor was used as a test case for CBMC. The performance of the CBMC simulation was compared to an established docking program, AUTODOCK, in reproducing the complex known from an X-ray crystallography experiment. A particular focus was the efficiency of the two methods.

5.1.1 Free energy simulations

The essential aspects of biomolecular recognition are generally amenable to a thermodynamic description in which the equilibrium constant (or equivalently, the free energy difference between the bound and free states) is central. Molecular simulations are no different from physical systems, in that absolute free energies cannot be observed directly. This means that free energy differences can only be derived from (often multiple) simulations which encompass a process that transforms the system between the two end-states (for instance, thermodynamic integration). Phase equilibria have been studied extensively by simulation, and a variety of techniques have been developed to determine free energy differences accurately. Equilibria involving multiple species are more complex, and barriers along the reaction pathway (such as diffusion barriers, or the formation of a transition state) make adequate sampling a problem. Fortunately, simulations are not restricted to physically meaningful reaction pathways, so many of these barriers may be bypassed with specially constructed simulation schemes.

The most successful type of simulations of receptor–ligand (or equivalently, host–guest)

binding equilibria have been predictions of the difference in free energy of binding between two similar ligands for a particular receptor. Simulating the free energy change for the transformation from one ligand to the other, in free and bound states, gives the desired quantity according to a simple free energy cycle [38]. This technique requires a tightly-bound complex which exhibits no dissociation on the simulation time-scale, and the absence of major differences in conformation between the two ligands. The modest difference between the two systems reduces the amount of sampling necessary to follow the transformation process to convergence, and reduces the need for lengthy equilibration to allow the solvent to relax. The effect of inadequacies in the potential (particularly with respect to accurate free energies of solvation) is also minimised.

Adsorption isotherms of linear and branched alkanes in zeolites can be simulated with excellent accuracy using grand canonical ensemble simulations [211]. Direct application of the same methods to the binding of peptide to protein is hampered by a number of factors. Both systems are characterised by slow equilibration, but for zeolite adsorption this is caused primarily by a diffusion barrier, easily overcome by the particle insertion method of the grand canonical simulation scheme. More significant impediments to equilibration in the case of biomolecular binding are the large phase space accessible to both ligand and receptor, which is only slowly explored, and the presence of a relatively high-density solvent, which relaxes slowly. A dense solvent also greatly decreases the efficiency with which particles can be added to the system (particle insertion moves). A general solution to this problem may be found in the use of continuum solvent representations, or new types of Monte Carlo moves, while the problem of flexibility of ligand and receptor can be addressed with CBMC techniques.

CBMC greatly increases the efficiency of simulations of the zeolite adsorption of long alkane chains, by allowing inserted chains to conform to the zeolite matrix, giving net low interaction energy and an increased acceptance ratio. The Henry coefficient, directly related to the excess chemical potential of the inserted species, can be determined at the time of insertion with little additional calculation. Application of CBMC to ligand–receptor binding should similarly lead efficiently to low energy bound complexes. A crude, but nonetheless useful approximation would be to treat the receptor as a rigid species, and perform simulations in vacuum, thereby achieving much faster equilibration. This application of CBMC resembles the well-established idea of docking.

5.1.2 Docking and scoring

An inability to perform realistic simulations to determine the free energy of binding from first principles has led to the development of numerous other computational approaches which qualitatively or semi-quantitatively characterise the binding process. As knowledge of the structure of the bound complex is pivotal for further experimentation, the goal of many specialised computational techniques is to predict it when experimental details (e.g. from NMR or X-ray crystallography) are unavailable. Docking is equivalent to finding the most densely populated state (or family of states) in the ensemble, assuming equilibrium conditions which strongly favour binding. Docking procedures are generally carried out when ligand and receptor are known to bind with high affinity, with a known receptor structure, and sometimes a predetermined ligand conformation. Under these restrictions, a number of docking programs have been developed which can predict the bound complex successfully. A related problem is

the analysis of a bound complex to give a 'score', a rough approximation to the free energy of binding. The approximate nature of scoring limits it to the qualitative comparison of hypothetical ligand–receptor complexes from docking studies or *de novo* ligand design. The shortcomings of empirical approaches to molecular recognition noted in the introduction are evident in most scoring procedures, however there are rarely any better alternatives.

Current docking and scoring approaches have serious limitations when applied to flexible polypeptide ligands. For docking this is a simple consequence of inefficient sampling of peptide conformation. Scoring programs perform badly for all the reasons already mentioned, but are further impeded by formulations and parameterisations appropriate for smaller, much less flexible ligands. As computational studies of peptide ligands are poorly served by these methods, it would be interesting to see if docking methods based on the improved sampling efficiency of CBMC can perform any better.

5.1.3 AutoDock

AutoDock [88] searches for low energy ligand–receptor complexes using a Metropolis MC algorithm together with a simulated annealing procedure. A variant of the Amber united-atom protein potential is used, with efficient grid-interpolation energy evaluation possible due to the assumption of a rigid receptor. Docking is performed without explicit solvent, with an approximate implicit treatment of solvation using a distance-dependent dielectric constant. MC moves consist of rigid body displacements and rotations of the ligand, and changes to designated ligand internal torsions. Version 3.0 of the AutoDock software [163] implements a number of more elaborate features. In addition to Metropolis MC, a genetic algorithm search strategy with a local search optimisation step is available. There is an improved implicit solvation model, using atomic solvation parameters, and energy evaluations use an empirical free energy function, rather than the simplified Amber potential.

Autodock gave a reasonable result for the docking of a relatively flexible, hydrophobic HIV-1 protease inhibitor, XK263, with the lowest energy docked conformation having an RMSD of 0.86 Å from the crystal conformation [164], and also did well with other more straightforward systems, including benzamidine/β-trypsin and streptavidin/biotin [88, 163, 164]. Docking of highly flexible ligands (those with more than eight rotatable bonds, as a rule of thumb) is not possible within a reasonable length of time using the simulated annealing protocol [163], and while the genetic algorithm has been shown to cope better with flexibility in a number of cases, the most flexible ligand successfully docked had eleven rotatable bonds. To work around this limitation, strategies for the docking of peptides have been proposed that use additional assumptions about the nature of binding, for instance by serial docking of fragment sequences [78].

5.1.4 Parallel tempering

Dynamic simulation schemes which sample a canonical ensemble suffer from poor sampling when barriers are present on the potential energy surface. Various non-Boltzmann sampling techniques (generalised ensemble or entropy sampling for instance) simulate a non-physical ensemble with an increased population of higher energy states. The generalised ensemble is constructed to allow the reweighting of simulation results to recover a canonical ensemble.

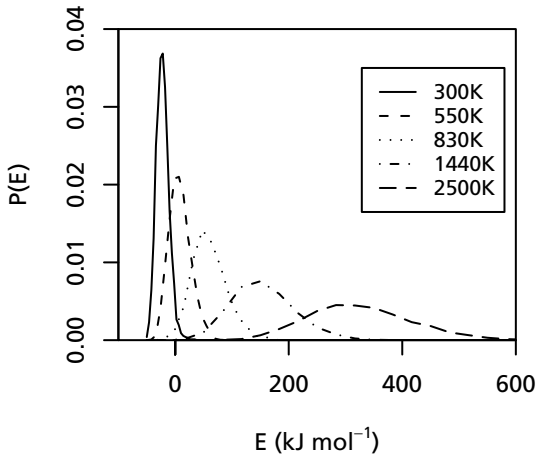


Figure 5.1: Parallel tempering histograms of energies for parallel systems. Regions of overlap indicate the probability of swaps between adjacent systems. Energies taken from conformational analysis of alanine dipeptide.

Related *ad hoc* approaches based on high-temperature equilibration, most notably simulated annealing, also achieve improved barrier crossing behaviour, though the usual implementations do not give a route to a conventional ensemble average. Parallel tempering [82, 95] (also known as the multiple Markov chain or replica-exchange method), is an elegant alternative to generalised ensemble techniques, with much of the simplicity of simulated annealing, as well as the virtues that it is easy to parallelise and uses few adjustable parameters.

Parallel tempering requires multiple independent copies of the simulation run in parallel, each at a different temperature. One simulation runs at the desired temperature, and the remaining systems in a "ladder" of increasing temperatures. At regular intervals during the simulation a pair of systems adjacent on this ladder is chosen at random. A swap of configurations between the two systems is attempted, with an MC acceptance rule using the instantaneous energies of the two systems, along with their difference in temperature, $\mathrm{acc}[(i, j) \rightarrow (j, i)] = \min[1, \exp(-\Delta U_{ij}/k_{\mathrm{B}}\Delta T_{ij})]$ (Figure 5.1). If the temperatures are chosen to give a reasonable acceptance rate (e.g. greater than 10%), each system continues to sample a canonical ensemble at the given temperature, but mixing between the systems produces an enhanced probability of overcoming any free energy barriers.

Verkhivker et al. have applied parallel tempering to the ligand–protein docking problem [232]. Parallel tempering MC was used as a follow-up to docking using the method of evolutionary programming. Either the AMBER force field with an additional implicit solvation term, or a simplified knowledge-based potential function was used. The parallel tempering simulations gave a detailed picture of the free energy landscape near the docked conforma-

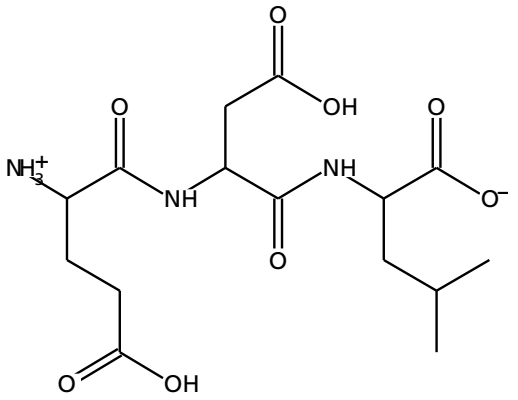


Figure 5.2: The tripeptide HIV-1 protease inhibitor, Glu-Asp-Leu (EDL).

tions, and in some cases were able to correct misdockings which were the result of inadequate sampling in the evolutionary programming docking step.

5.2 Methods

5.2.1 Models

The chosen receptor was the HIV-1 protease dimer, and the tripeptide competitive inhibitor Glu-Asp-Leu (EDL) (which binds with a K_i of approximately 50 μM at pH 4.25 [139]) was the mobile species (Figure 5.2). Global features of the HIV-1 protease dimer and the position of the binding site can be seen in Figures 5.3 and 5.4. The two loops centred around Ile50 are known as the 'flaps', and an 'opening' movement of these resulting in an increased exposure of the active site to solvent has been observed in the apoenzyme. The closed state, which is observed in crystal structures, offers very limited access to the active site, a challenge to docking studies.

Microcalorimetric measurements have indicated that the binding of EDL to HIV-1 protease is exothermic, with a ΔH of $-15.1\,\mathrm{kJ\,mol^{-1}}$ [229]. Combined with an examination of the crystal structure, this suggests specific polar interactions are essential for binding. This is in contrast to other protease inhibitors that are thought to bind primarily because of their hydrophobicity, consistent with the observed positive entropy change on binding. Furthermore, binding of the inhibitors acetyl-pepstatin, indinavir, saquinavir and nelfinavir is enthalpically unfavourable [229]. In the absence of solvent, hydrophobically-driven binding is impossible, making an exothermic binding process more or less a prerequisite for simulation in vacuum.

A model of the protease was built starting from the X-ray crystal structure 1a30 [139] (resolution 2.0 Å, R factor 0.189) from the Protein Data Bank [25,26]. Hydrogen atoms were added and their coordinates minimised using the CFF91 force field [62,143] in MSI Insight 2000 [160], with the protonation state of ionisable residues as tabulated by Insight for pH 4.3 (exceptions noted below). For the simulations with WHOPPER, the side chains of the exterior,

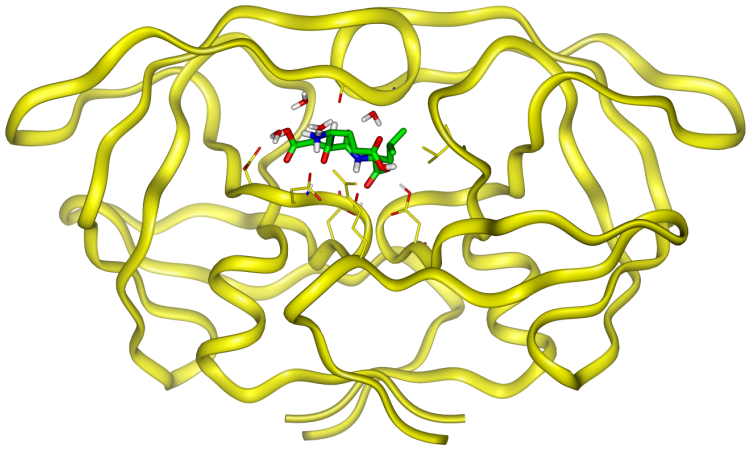


Figure 5.3: Ribbon diagram of HIV-1 protease dimer crystal structure, with bound inhibitor EDL. Coordinates taken from PDB 1a30.

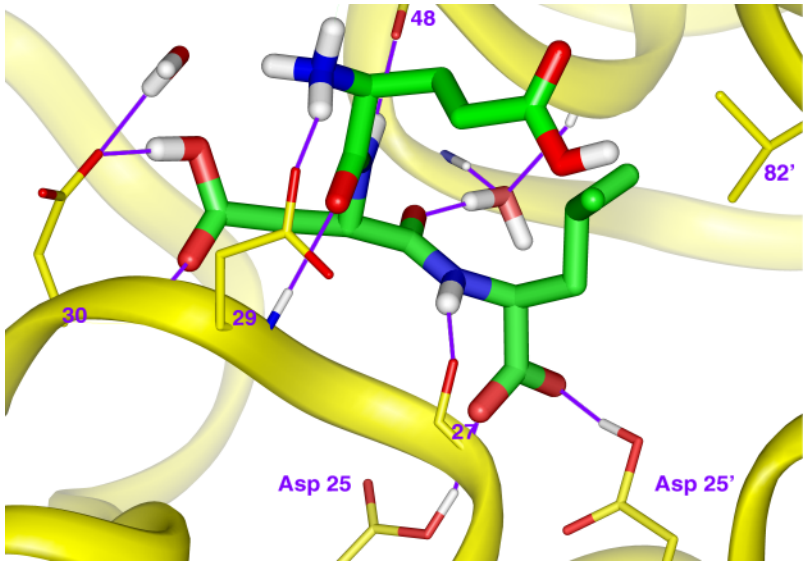


Figure 5.4: Ribbon diagram of HIV-1 protease active site, with bound inhibitor EDL. Coordinates taken from PDB 1a30.

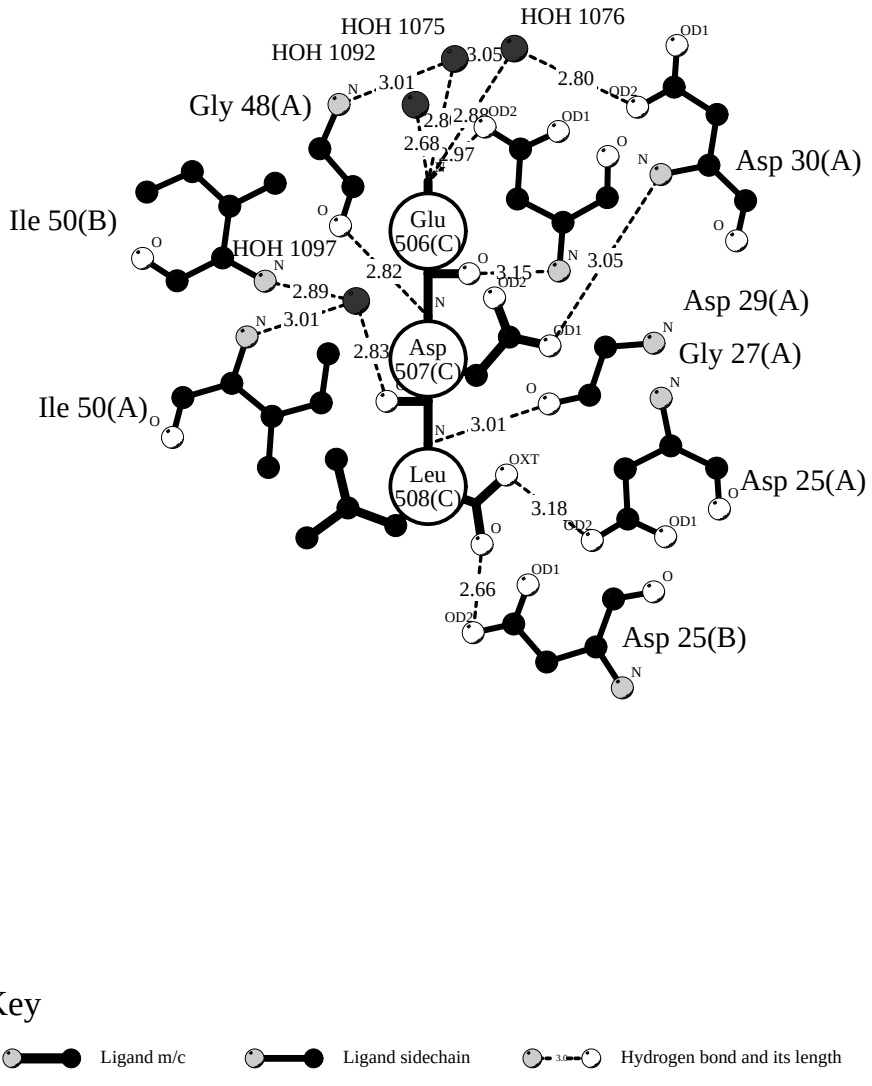


Figure 5.5: Schematic LIGPLOT representation of the HIV-1 protease EDL binding site, as derived from the Protein Databank X-ray crystal structure 1a30. Hydrogen bonds where Asp $O_{\delta 2}$ is a donor have been added by hand.

solvent-exposed residues Lys7, Lys14, Arg41, and Lys41 were removed from both monomers to restore electrostatic neutrality (this was a requirement to ensure the accuracy of the Ewald sum). Binding of the ligand is pH-dependent [139] indicating pK_a shifts in ligand and/or receptor in the bound form. The results from Chapter 2 indicate that at pH 4.3, the N-terminus, C-terminus, and aspartic acid side chain of the tripeptide will be fully protonated, while the glutamic acid side chain will be on average approximately 50% protonated. The catalytic aspartate dyad of the enzyme will be 70% monoprotonated and 30% diprotonated. However, the modelled ionisation state was selected in ignorance of these results, based on hydrogen-bonding analysis published by the crystallographers [139], and the requirement that the ligand be electrostatically neutral. Both Asp25 residues were modelled as protonated and the ionisation state of the tripeptide was modelled as $^{+}$HGlu-Asp-Leu-O^{-} (i.e. the Glu and Asp sidechains are protonated). A LIGPLOT [241] representation of the tripeptide binding site is reproduced in Figure 5.5. The protease coordinates remained fixed in all simulations. All crystallographic water molecules, except the well-known flap water (number 1097 in 1a30) and three other water molecules present in the interface (1075, 1076, 1092) according to LIGPLOT analysis, were removed.

5.2.2 AUTODOCK

Both AUTODOCK versions 2.4 and 3.0 were used, to allow comparison with two different docking protocols. The simulated annealing docking algorithm of AUTODOCK 2.4 is based on a straightforward Metropolis MC scheme, making a comparison with WHOPPER easier. On the other hand, the more elaborate approach embodied in AUTODOCK 3.0 is probably needed for successful docking with a highly flexible ligand, and gives a better impression of the current state of the art.

The Insight models of protein and peptide, with CFF91 charges, were saved in MOL2 format, and standard AUTODOCK scripts and protocols were used to generate the required input files, with the ligand model automatically converted to an essential-hydrogens representation. Grids for potential evaluation had a resolution of 0.375 Å and dimensions of 45 Å × 45 Å × 45 Å. Each docking was repeated 100 times with different random seeds, after which results were clustered and ranked by energy. The simulated annealing protocol consisted of cycles which continued until either 25000 steps were accepted or 75000 rejected. Each cycle was started with the current minimum energy configuration. In a previous study with AUTODOCK, a high starting temperature was needed to give the HIV-1 protease inhibitor XK-263 enough energy to enter the active site [164], so a similar protocol was followed here. The initial cycle of each run was at a temperature of 31000 K, which was multiplied by a factor of 0.955 before each subsequent cycle.The initial maximum displacement of translation steps was 20 Å, and of quaternion rotations 180°; both were multiplied by a factor of 0.985 per cycle. No initial energy limit was specified. The run was stopped after 150 cycles.

The Lamarckian genetic algorithm parameters were left at the AUTODOCK 3.0 default values. These were a population size of 50, a maximum of 1 500 000 energy evaluations, a maximum of 27 000 generations, a survival rate of one elite individual per generation, a mutation rate of 0.02, a crossover rate of 0.8, a window size of 10 generations for picking the worst individual, and α and β values for the mutation Cauchy distribution of 0 and 1. Step sizes were 0.2 Å for translations, 5.0° for quaternions and 5.0° for torsions. Pseudo-

Solis and Wets local search was performed for 300 iterations, with a probability of selecting an individual for search 0.06. The initial value of ρ was 1.0, with the value updated after 4 consecutive successes or failures, and a lower bound of 0.01. A maximum initial energy of 0 was specified, with a a maximum of 10 000 retries. 100 runs were performed.

The dockings were first done with the peptide constrained to the conformation seen in the crystal structure of the complex. This was to ensure that the experimental complex was an energy minimum in the AutoDock model, and that the binding site was sufficiently accessible. In the second round of dockings all σ-bond torsion angles (namely, all backbone ϕ and ψ angles, and all sidechain χ angles) were free to rotate, for a total of 15 rotatable bonds.

AutoDock was run under Irix 6.5.5 on an SGI Origin 200 R10000 180 MHz, and was compiled using the SGI MIPSPro C compiler version 7.30.

5.2.3 Whopper

A parallel tempering scheme was added to Whopper, using the Message-Passing Interface (MPI) [151] to achieve parallel execution of the multiple systems. For the types of simulation described here, the communication overhead required for parallel tempering is low, so n processors for n parallel systems should represent a close to factor n linear speed-up relative to the single processor case. The parallelisation was orthogonal to the parallel algorithms already implemented in Whopper, so existing speed gains can still be achieved in combination with parallel tempering.

The Bigmac CBMC simulation scheme for alkane adsorption in a zeolite matrix was the basis for docking simulations in Whopper. The protein receptor was treated as a fixed collection of atoms, interacting with the mobile, peptide species using Charmm22 Lennard-Jones parameters and charges. All intra-molecular bonded and non-bonded interactions of the protein were ignored. In order to achieve workable execution times, the mobile species was simulated using Charmm19 essential hydrogens bonded and non-bonded parameters, rather than the all-atom potential. Electrostatic interactions appeared to have a serious negative effect on the efficiency of the simulation, so in addition to simulation with the standard parameters, the simulation was also performed with ligand charges scaled by 0.1. Simulations were carried out in the canonical ensemble (constant-NVT) with periodic boundary conditions. Dimensions of the simulation box were chosen large enough (60 Å × 70 Å × 70 Å) to prevent large artefacts from the Ewald sum evaluation of long-range Coulombic interactions [108].

Parameters of the CBMC simulations were as described in the previous chapter. As only one mobile particle was present in the system, no tail correction was applied to the Lennard-Jones interactions. Ewald summation was performed with $\alpha = 0.35$ and $21 \times 24 \times 24$ vectors. Chirality of the C_α atoms of the tripeptide were constrained to be (*S*). The system energy and coordinates were sampled every cycle of 10 CBMC moves. A series of 64 parallel tempering moves were attempted at the end of each MC cycle. The docking simulation consisted of 500 cycles of 10 steps, with 50% complete regrowth and 50% partial regrowth. No additional initialisation/equilibration cycles were used. System temperatures for parallel tempering were 300 K, 350 K, 400 K, 450 K, 500 K, 600 K, 700 K, and 800 K.

Each system was initialized with the ligand displaced 20 Å from its bound position. Complete chain regrowths were performed starting from C_γ of the aspartic acid residue and to speed up the simulation, this start atom was always placed at the coordinates of the corres-

program	lowest energy kJ mol^{-1}	RMSD Å
AutoDock 2.4 (rigid)	−362	0.5
AutoDock 2.4 (flexible)	−129	17.6
AutoDock 3.0 (rigid)	−3	0.2
AutoDock 3.0 (flexible)	+10	8.5
Whopper	+352	47.4
Whopper (scaled q)	−56	1.9

Table 5.1: Summary of docking results. The energy and RMSD (over all ligand atoms) of the lowest energy conformation found, relative to the crystal structure conformation, are listed.

ponding ligand atom from the complex, rather than at random. As the binding site is relatively inaccessible and quite constrained, it is possible to make a reasonable estimate of the degree to which this accelerates the docking. Partial chain regrowths were initiated from a randomly chosen atom, as in the previous chapter, and all atoms, including the Asp C_γ, were free to be regrown in a new position.

An additional series of simulations to measure efficiency in relation to system size were done using all-atom n-alkane models varying in size from CH_4 to $C_{19}H_{40}$. Charmm22 aliphatic carbon parameters were used, with Coulombic interactions neglected. Simulations were of single, isolated chains, at 300 K, with CBMC regrowth moves only. Each simulation consisted of 3000 steps, during which the acceptance ratio was recorded. Other parameters were identical to the previously described CBMC simulations.

Simulations were performed on a cluster of dual-processor 500 MHz Intel Pentium III machines, running under Linux 2.2.14. The LAM 6.3.1 [130] implementation of MPI was used, and programs were compiled using the Portland Group Fortran 77 compiler, version 3.0-4. Inter-node communication for parallel simulation was over switched, full-duplex 100 Mbps Fast Ethernet. Parallel tempering simulations used one processor per temperature.

5.3 Results

5.3.1 Docking

A summary of the results of the dockings is given in Table 5.1. Rigid dockings with both versions of AutoDock successfully found a docked ligand conformation with an RMSD of less than 0.5 Å from the crystal structure as the minimum energy of the simulation. The rigid docking with AutoDock 2.4 found a conformation with an energy 362 kJ mol^{-1} lower in energy than the conformation from the crystal structure, as the latter had a slight steric overlap at the N- and C-termini according to the AutoDock 2.4 potential function. Docked and crystal conformations were much closer in energy as calculated with the AutoDock 3.0 potential function, which has scaled-down Lennard-Jones parameters and applies a smoothing factor to repulsive terms. The docking with version 2.4 was not reliable, as the native structure was only found in one of a hundred runs, and no near-native structures were found. This was despite the use of an elevated starting temperature and an extended cooling programme. AutoDock 3.0 found eight structures in the cluster with the lowest docked energy, indic-

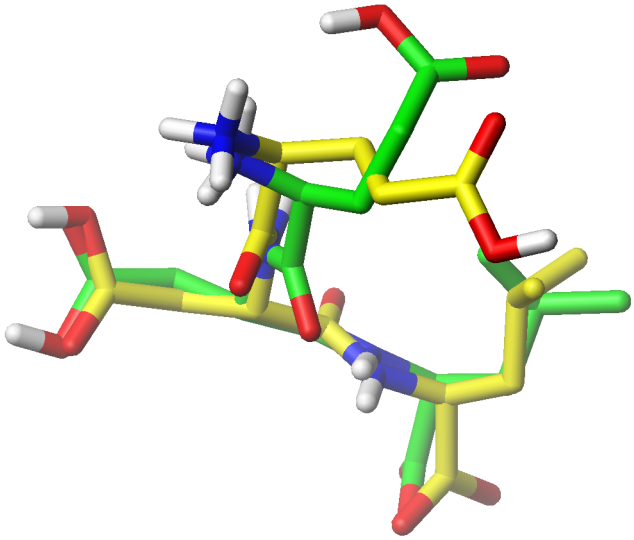


Figure 5.6: WHOPPER lowest energy docked conformation overlaid with crystal structure conformation. The docked structure is drawn in a darker shade.

ating that conformational sampling was good. In addition, the two next most highly ranked clusters, with a total of 12 members, were on average 14 kJ mol^{-1} higher in energy than the minimum, as compared to a jump of 227 kJ mol^{-1} between the highest and second-highest ranking structures for AUTODOCK 2.4.

Dockings in which the ligand torsions were free to move were unsuccessful in finding the native conformation for both AUTODOCK versions. In both cases, there was a complete lack of clustering in the docked structures, with an essentially flat distribution of energies. This suggests that the sampling was far from converged in both cases, even though on the order of 10^9 conformations were evaluated, for a total of 42 hours run time, in the AUTODOCK 2.4 computation.

The scaled-charge WHOPPER simulation converged to a lowest energy docked conformation with an energy $\sim$ 50 kJ mol^{-1} lower than the native conformation. This conformation had an all-atom RMSD of 1.9 Å from the native coordinates, and is shown overlaid with the crystal structure in Figure 5.6. The structure with the lowest RMSD from the native coordinates, 1.2 Å, (considering only the ensemble at 300 K) had an energy 27 kJ mol^{-1} lower than the native conformation. Unsurprisingly, the strong scaling-down of the electrostatic energies changes the shape of the potential energy surface and the location of minima significantly. Nevertheless, the docked structure preserves essentially all the hydrogen bonds seen in the crystal structure, with the only significant deviation in coordinates in the glutamic acid side chain. This side chain is partially solvent exposed in the crystal structure, and is noted by the crystallographers to be disordered [139]. The WHOPPER simulation using standard parameters failed to find any docked conformations.

Despite the low efficiency of the CBMC moves (see below), the docking appears to

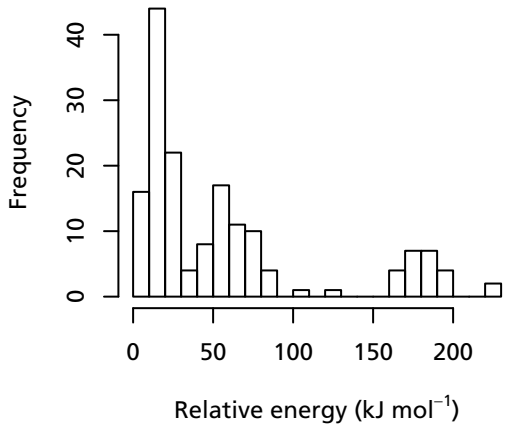


Figure 5.7: Histogram of energies of the 162 unique conformations sampled in the WHOPPER docking simulation at 300 K.

	System temperature (K)						
	300 – 350 – 400 – 450 – 500 – 600 – 700 – 800						
Swap %	26	16	29	55	32	54	59

Table 5.2: Proportion of successful parallel tempering swap moves between the systems.

be well converged. Of the 162 unique conformations sampled at 300 K, 11 cluster within 0.5 Å RMSD from the minimum. The histogram in Figure 5.7 shows a concentration of conformations in the lower energy range.

5.3.2 Parallel tempering

Time series of the four lowest temperature systems are show in Figure 5.8. Exchanges of conformation between systems as a results of parallel tempering moves can be seen as the points where the energy traces touch. The highest temperature trace fluctuates in energy much more strongly than the lower temperature traces, and the parallel tempering scheme makes use of these fluctuations to enhance the sampling of the lower temperature systems, resulting in a rapid decrease in energy at the lowest temperature. The percentages of successful parallel tempering moves are given in Table 5.2. The values are in the range 15% to 60%, showing substantial mixing between the systems. However, longer equilibration would be needed to reliably characterize the swap rate. The docking simulation is too short to allow accurate analysis of the convergence behaviour, but the effect can be seen clearly in the energy autocorrelation of a longer CBMC simulation of alanine dipeptide with and without parallel

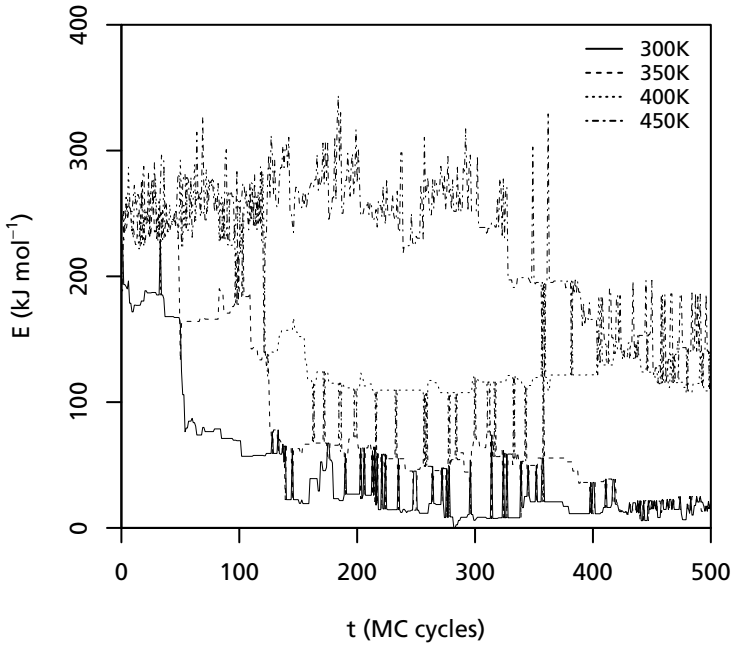


Figure 5.8: Energy time series for the four lowest temperature systems. Energies are relative to the lowest energy found in the simulation. Samples were taken once per cycle of 10 MC steps.

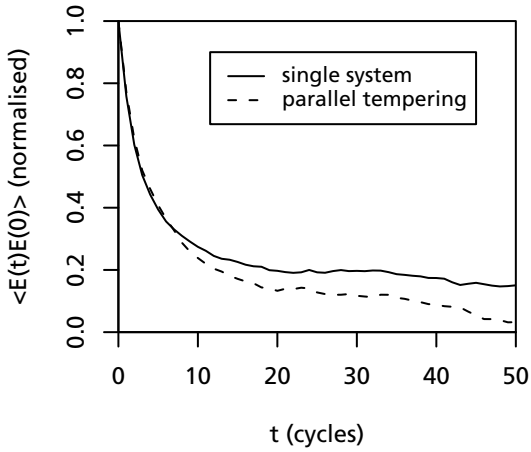


Figure 5.9: Autocorrelation of system energy at 300 K, single system versus parallel tempering. Results taken from CBMC simulation of alanine dipeptide (not shown).

tempering in Figure 5.9 (results obtained using the alanine dipeptide model from Chapter 4). Without parallel tempering, the autocorrelation function decreases rapidly and then flattens out, suggesting that there is a bottleneck slowing the progression between the various low energy states. The parallel tempering autocorrelation function decreases more rapidly for times of approximately 8 cycles or more, with the mixing between high and low temperature systems allowing potential energy barriers to be more quickly overcome.

5.3.3 Efficiency

Docking

The acceptance ratios for the docking simulation with scaled charges are given in Table 5.3. Both the proportion of chain growth completed without overlaps, and the proportion of moves accepted, increases with temperature, as expected. When unscaled charges were used, there were no successful whole chain regrowth moves for the length of the simulation. The acceptance ratio for these moves was so low that there were no successful moves even in a longer simulation ($\sim$ 15000 steps; results not shown).

The computational requirements of the different simulations are summarized in Table 5.4. The number of bonded and non-bonded energy evaluations carried out by WHOPPER are estimates based on the number of trial moves used for the various internal degrees of freedom, and will vary slightly according to the molecular topology. AUTODOCK uses a simplified potential function and the grid-based interpolation method allows non-bonded energies to be evaluated quickly. The CPU time per step is therefore much smaller than for WHOPPER,

	system temperature (K)							
	300	350	400	450	500	600	700	800
whole %	10	12	13	15	16	17	19	19
whole accept	*0*	*0*	*0*	*2*	*0*	*2*	*0*	*2*
partial %	50	56	64	75	82	99	100	100
partial accept %	5	6	8	13	16	23	21	22

Table 5.3: CBMC acceptance ratios for the parallel systems, scaled charge results. The first two rows of results show the percentage of whole chain regrowth moves completed successfully (i.e. not rejected due to steric overlap), and the number of moves accepted (of 1481 attempts). The next two rows give the same figures for partial chain regrowths, with the acceptance ratio given as a percentage in this case.

program	total steps	bonded energy evals	non-bonded energy evals	CPU time seconds
AUTODOCK 2.4 (rigid)	8.9×10^8	–	8.9×10^8	5.2×10^4
AUTODOCK 2.4 (flexible)	9.3×10^8	9.3×10^8	9.3×10^8	1.5×10^5
AUTODOCK 3.0 (rigid)	1.5×10^8	–	1.5×10^8	2.1×10^4
AUTODOCK 3.0 (flexible)	1.5×10^8	1.5×10^8	1.5×10^8	4.9×10^4
WHOPPER	5.0×10^3	$\sim 5.0 \times 10^6$	$\sim 4.0 \times 10^4$	2.8×10^3

Table 5.4: Energy evaluation statistics and CPU use for AUTODOCK and WHOPPER. WHOPPER figures apply to a single system in the parallel tempering simulation. Note that AUTODOCK and WHOPPER CPU times were obtained on different computers.

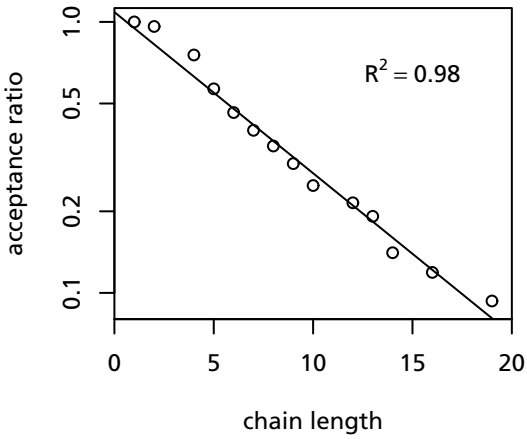


Figure 5.10: Log-linear plot of CBMC simulation acceptance ratio versus alkane chain length. An all-atom model with CHARMM22 parameters was used. The regression line indicates an exponential decrease in acceptance ratio with increasing system size, $r = 1.087e^{-0.137n}$

where no such optimizations have been used. The total number of steps, and the number of energy evaluations performed allow the various algorithms to be compared independently of the performance of a particular computer system or the degree to which the implementations have been optimized. However, as the CBMC algorithm uses relatively more bonded energy evaluations than non-bonded, with the expectation that the former will be less computationally expensive, the net difference in performance from MC will depend strongly on the implementation.

Alkanes

Simulations with a series of all-atom alkane models showed that efficiency, as measured by the acceptance ratio of CBMC whole chain regrowth moves, decreases exponentially with chain length. A log-linear plot of acceptance ratio in relation to chain length (Figure 5.10) indicates that efficiency is poor for $C_{20}H_{42}$ and longer.

5.4 Discussion and conclusions

5.4.1 Docking

The system of HIV-1 protease and inhibitor simulated in this chapter is a particularly challenging one. The problem of ligand flexibility is most important, but the inaccessibility of the binding site, the relatively poor affinity of ligand for receptor, and the unusual protonation

state of the ligand and receptor all contribute. By exploiting *a priori* knowledge concerning the location of the binding site and its protonation state, the CBMC docking simulation using WHOPPER successfully found a low energy docked conformation with the same binding mode as observed in the crystal structure of the complex. Similar conformations were obtained from rigid docking simulations using AUTODOCK, with substantially more reliable results using the genetic algorithm search procedure in AUTODOCK 3.0. The failure of flexible docking using both AUTODOCK versions demonstrates that the conformational freedom of the tripeptide ligand causes a substantial increase in the size of the problem space which needs to be explored.

The size of the problem space was artificially restricted for the CBMC docking by constraining whole chain regrowth to start with the initial atom correctly placed; no additional constraints were placed on the conformation of the ligand. A reasonable estimate of the effect of this constraint on the speed of convergence can be made by considering the total number of possible starting positions in the vicinity of the protein. Assuming that placement of the atom within a volume of 1 $Å^3$ surrounding the position in the native structure will be sufficient (a requirement which will probably be less stringent in practice, as partial chain regrowths may be able to optimize the atom position subsequently), and given a box volume of 100 000 $Å^3$, minus a protein-occluded volume of 26 000 $Å^3$ (calculated with GRASP), then on the order of a factor 80 000 speed up was achieved. Combining this information with the data from Table 5.4 gives an estimate of a simulation length for an unconstrained docking of 4.0×10^8 steps, requiring 4×10^{11} non-bond energy evaluations and 3.2×10^9 bonded-energy evaluations, a workload in the range of about 3–500 times greater than that presented by the AUTODOCK 2.4 simulations. In practice the binding site may be known or easily predicted, so these figures represent a worst case. Use of parallel CBMC within each system will allow further speed-ups; this analysis assumes that the computational costs of the parallel systems are not summed.

As no indication of the speed of convergence of the flexible AUTODOCK simulations is available, no meaningful comparison with the performance of CBMC is possible. However, the CBMC docking algorithm in its present form does not meet its potential, as convergence is still much too slow to allow equilibrium simulations and the calculation of free energies and bulk properties. Pure docking applications may be more practical, but would require extensive optimizations along the lines of AUTODOCK. This would entail the use of a simpler potential function with fewer degrees of freedom in the ligand, making substantial increases in efficiency possible at some cost to the elegance and generality of the approach.

5.4.2 Efficiency

The efficiency of CBMC for all-atom polypeptide simulations was disappointing. Undoubtedly, many parameters require tuning, namely the number of trial moves for the different biased selections, and the second cut-off distance for the dual-cutoff scheme, but such improvements will not radically improve the acceptance ratio. As illustrated by the series of alkane simulations (Figure 5.10), the acceptance ratio decreases exponentially with system size. It is a characteristic of Rosenbluth sampling that the distribution of chain conformations diverges from Boltzmann statistics exponentially as the chain length increases [22]. For simple, linear systems this divergence does not affect the acceptance ratio appreciably except when the chains are extremely long. The failure of CBMC in this case appears to be a general problem for highly branched systems of moderate size and larger, which has gone unnoticed

because of the preponderance of united atom models and relatively short chain lengths in alkane simulations.

The effect of electrostatic interactions on the acceptance ratio also caused efficiency problems. The alanine dipeptide simulations in Chapter 4 showed a similar, but smaller, effect, with the addition of the electrostatic energy term causing a drop in acceptance ratio from 0.17 to 0.10. Strong long-distance interactions appear to present a problem for the CBMC algorithm, an effect which can be explained by the short-sightedness of the method. Electrostatic interactions may cause global changes in ligand conformation, an effect which biased selection on an atom-by-atom basis will overestimate at short range and underestimate at long range. These deviations from the desired statistics will result in the growth of low-weight chains and a low acceptance ratio. The short-range overestimation effect can be compensated by decreasing the contribution of electrostatics to the biasing potential, and considering the balance of the energy in the acceptance rule. In the alanine dipeptide simulation this was observed in the greater efficiency of the Ewald electrostatic evaluation method compared to a simple calculation of pairwise Coulombic interaction energies. Recursive growth methods, as mentioned in the introduction to Chapter 4, may be able to correct the problem of long-range underestimation.

5.4.3 Prospects

There are a number of possible modifications to the simulation which might restore efficiency to a workable level. Substitution of an essential hydrogen atom potential for the CHARMM22 all-atom potential, brings tripeptide simulations within reach, but systems much larger than this are likely to cause difficulties. Drawing trial torsions from an appropriate non-uniform distribution (for instance, based on a Ramachandran plot for backbone torsion angles) will improve efficiency, and reduce the divergence of generated chains from Boltzmann statistics.

The number of degrees of freedom to which CBMC is applied could be reduced, as in the published literature applying CBMC to peptides [54, 199] (see introduction Chapter 4). Alternatively, only partial regrowths could be performed, rather than CBMC insertions of complete chains. In both cases, this leads to the limitations noted in the introduction, and is unsuitable for docking-type simulations. Equilibration is also much slower in this scheme, as atoms in the middle of the chain are still regrown with low efficiency.

Recoil growth

An attractive alternative to CBMC with many similar characteristics is recoil growth [47, 48]. Recoil growth is a dynamic MC scheme which uses a biased growth technique reminiscent of Meirovitch's double scanning method [149, 150], with some resemblance to other recursive growth methods mentioned in the introduction to Chapter 4. There are a number of advantages compared to CBMC, relevant for the efficiency problems that were observed in this study. It is more efficient for long chains, and at higher densities, the latter making it interesting for binding simulations. Two properties which should reduce the problem of low acceptance ratios, are greater efficiency with a small number of trial moves, and early rejection of chains with low Rosenbluth weights.

Recoil growth has some disadvantages. No efficient parallel algorithm is available [48].

The calculation of excess chemical potential is less efficient than in CBMC. Even simple recoil growth schemes have proven to be more difficult to implement than CBMC, and the details for branched systems with complex bonded potentials have not yet been worked out. Even so, recoil growth will undoubtedly be effective when applied to the conformational analysis of simplified peptide models, and has a good chance of finding application in more sophisticated peptide and protein simulation techniques.

Solvation

Given the aim of performing efficient simulations on peptides and proteins, the ability to deal with (aqueous) solvation is essential for biologically relevant results. As mentioned above, use of explicit water solvent molecules is not practical with normal CBMC, so some kind of implicit or continuum representation of solvent effects is necessary. Both continuum electrostatic solvation methods and atomic-solvation-parameter-based models are suitable candidates, and an example of the latter, generalised Born-surface area (GB/SA) [185], is available in the TINKER energy evaluation code.

A solvation potential can be easily added as a correction term in the MC acceptance rule, in the same spirit that the Ewald correction is currently performed in BIGMAC. As evaluation of the solvation term is typically quite expensive, it is logical to perform it at this point in the algorithm (following the discussion of efficiency optimisation in the previous chapter). However, if the reduced acceptance ratio which results should be so low as to seriously compromise the efficiency of the simulation (i.e. in the case where the solvation term dominates the other terms in the forcefield), then some adjustment to the potential used to generate trial moves will be necessary. A simple scaling of the electrostatic energy term might be an effective remedy in such a case.

Elaborations of CBMC have been applied to simple fluid models consisting of binary mixtures of differently-sized particles [32, 60, 61]. It is tempting to consider the possibility of realistic explicit solvation simulations along these lines. The essence of the special MC move used in these studies is the exchange of larger particles with multiple small particles occupying an equivalent excluded volume. Another important development is an efficient scheme for inserting and deleting ions from simulations of dipolar fluids [205], as the extensive ordering of solvent molecules in response to charge makes unbiased approaches impractically inefficient. It remains to be seen if such schemes will translate to efficient algorithms for more complex peptide or protein models, but the possibilities for further research are obvious.

Chapter 6

General discussion

A number of methods for studying the interaction between protein and ligand came into play in the course of this thesis. The intention was to apply and evaluate approaches which are new in the context of drug design, and to consider the aspects of the molecular recognition process which were apparent along the way. The two main subjects of investigation were the role of protonation equilibria and electrostatic interactions, and the application of the configurational bias Monte Carlo simulation technique. In evaluating the results of these studies, both rational and empirical perspectives have a place, but given the patchy nature of our theoretical understanding and our experimental knowledge, it is often hard to draw conclusions.

The most important basis for understanding molecular recognition is the theory of equilibrium processes embodied in (statistical) thermodynamics. There is also increasing interest in kinetic processes, with wider scope than simply substrate–enzyme interactions, but for many biochemical phenomena a purely equilibrium description is sufficient. The investigation of proton linkage in Chapter 2 illustrates how macroscopic binding equilibria can be explained by microscopic molecular models. An intuitive interpretation of the linkage concept which is relevant for describing molecular recognition, is that the protonation state will change on binding to optimize the electrostatic interactions between ligand and receptor, at least as far as the excess chemical potential of the proton (i.e. pH) will permit. Interpretation of the experimental data is difficult, particularly in the case of NMR titration experiments, and some of the calculated pK_a shifts appear to be spurious or unreliable, but the results make clear that different ligands will induce different receptor protonation states on binding. These types of predictions make it possible to better account for the protonation state and proton linkage energies in subsequent modelling studies.

Computer programs to carry out pK_a calculations of the type applied in Chapter 2 have been developed by a number of groups since the method was introduced by Bashford and Karplus in 1990 [21]. Systematic evaluations of the newest developments have not yet been published, but progress is encouraging. Despite the general utility of the method, and the presence of important pK_a shifts in many well-known ligand–protein complexes, use of these methods has not become routine. In part this is because of the disappointing accuracy which could be achieved until recently, but the problem of unwieldy software which requires extensive manual intervention even for simple calculations has also been a source of discouragement. Incorporation of automated pK_a calculation routines in general purpose modelling software, such as WHAT IF [238], should help to popularize these methods.

In Chapter 3 the calculated pK_a shifts, along with experimental data on pH-related effects, were used to fit a scoring function to experimental HIV-1 protease inhibitor activity data. Inconsistencies in the calculated shifts, along with a lack of data regarding the free ligand pK_a values, prevented a quantitative prediction of linkage energies from being used in the analysis. However, the inclusion in the regression equation of information concerning the assay pH conditions, as well as a qualitative descriptor of the pK_a shifts in the receptor, allowed a more accurate scoring function to be found. This demonstrates that experimental data can be used more effectively by more detailed consideration of information regarding assay conditions. It also shows the importance of accurate, complete and consistent experimental data for model building, and underlines the potential of new analytical techniques, such as surface plasmon resonance and isothermal titration calorimetry, which can help in providing it.

One aim of the analysis was to investigate the role of electrostatics in molecular recognition for a data set with predominantly hydrophobic ligands, and at the same time to consider the possibility that the system of HIV-1 PR and peptidomimetic ligands presents an unusual case for molecular recognition. The simple regression model cannot achieve the predictive accuracy of more elaborate methods parameterized for HIV-1 PR, but the results do not point to exceptional properties. Further efforts to explain the activity of the one compound which fell outside the analysis, MVT-101, are needed to round out the picture. The regression relationships derived showed that electrostatic interactions weigh quite heavily in the balance of factors which explain activity. This result suggests that general purpose scoring functions, where in some cases (e.g. VALIDATE [100]) electrostatic terms only make a minor contribution, might be improved by use of more accurate data and methods in this respect.

As seen consistently in other studies using the continuum solvation electrostatics approximation, the sum of electrostatic solvation and interaction energies opposed binding, a contribution which must be overcome by other free energy terms, most importantly the hydrophobic effect. The prediction that systems with a net favourable electrostatic binding energy can be engineered [44, 118] is a fascinating one in this context, and needs to be followed up experimentally. A fine balance between opposing energy terms appears to be characteristic of the binding process, and can also be observed experimentally in the phenomenon of entropy–enthalpy compensation. An analogue exists in molecular recognition, where phenomena such as pK_a shifts, water binding, and global conformational change are seen to compensate for more gross changes, such as the substitution of an interacting group. A consequence of this principle is that predictions which partition the interaction energy into separate components require individual contributing energy terms to be computed with exceptional accuracy in order for the balance to be obtained with any confidence, considerably lowering the upper bound on the reliability of such predictions.

Physically realistic free energy simulations avoid some of the difficulties inherent in scoring approaches, but have their own limitations. Molecular dynamics and Metropolis Monte Carlo simulations of ligand–receptor interactions converge very slowly, limiting the types of predictions which can be made in a reasonable amount of computing time. The failure of MD simulations on a nanosecond time scale to sufficiently sample the intramolecular dynamics of proteins has been highlighted in a detailed comparison between X-ray structures, NMR structural and dynamics data, and MD trajectories [181]. Alternatives to MD which allow equilibrium simulations with faster convergence are extremely desirable, and the success of the configurational bias Monte Carlo technique in the simulation of alkane adsorption in

zeolites suggested the possibility of biochemical applications. A CBMC implementation of the CHARMM22 force field, described in Chapter 4, allows a well-characterised and accurate potential function to be applied to simulations of peptides.

An application of CBMC to the docking of a tripeptide inhibitor of HIV-1 PR was tested in Chapter 5. The particular characteristics of this system, namely the inaccessibility of the active site, and the extensive flexibility of the ligand, make it difficult for many established docking methods, and an interesting test case for the CBMC approach. FlexX, a successful docking program based on an incremental construction algorithm, is able to dock the peptidomimetic ligand VAC, with 17 rotatable bonds, to HIV-1 protease [188], and it would be interesting to test its performance in docking the tripeptide. Successful docking of an octapeptide in the MHC Class I molecule, a problem with similar features, required the use of the dead-end elimination (DEE) approach, a series of algorithms specifically designed to deal with the combinatorial problem of searching for low-energy peptide conformations [58]. CBMC was applied in combination with the parallel tempering sampling method, which has been shown to improve the sampling of MD and MC by enhancing barrier crossing (interestingly, parallel tempering seems to have been independently discovered in the applied statistics, chemical physics and biomolecular simulation communities, more or less simultaneously).

A CBMC docking simulation was successful in finding a near-native docked conformation, while attempts using the AutoDock software failed. The success of CBMC required adjustment of the force field parameters, and docking was achieved in a short simulation by fixing the position of the root atom in the binding site. CBMC did not display the hoped-for efficiency, and the acceptance ratio of CBMC moves was very low even at high temperature. In part, this was a result of the relatively constrained binding site, a situation similar to that found in high density fluid simulations. CBMC is known to suffer from efficiency problems under these conditions, and elaborations of the algorithm, such as the recoil growth procedure [47, 48], are needed to improve matters significantly. However, the low efficiency also appeared to be force field and model dependent, with efficiency decreasing with system size at a much faster rate when highly branched models and the CHARMM potential was used, than for the united-atom alkane models with which CBMC has been used previously. High efficiency may require force fields tailored to the requirements of CBMC, plausible in the light of advances in automated force field parameterization [175], or else modifications to the algorithm. Alternatively, the use of simplified models, as seen in previous work on peptide conformational analysis [54], is a proven route to efficient simulations.

Further development will be needed to make CBMC and related methods generally useful for simulation of peptides and peptide binding. CBMC is unlikely to be competitive with specialized methods for docking, but the possibilities for free energy simulations and conformational sampling remain interesting. Other applications of CBMC in biomolecular simulation are also possible. For example, lipid bilayer molecular dynamics simulations, well known to equilibrate impractically slowly, can be supplemented with CBMC to improve convergence [43]. Advancement in chain-growth algorithms for protein folding, which are closely related to CBMC, also points towards possibilities for the future. It seems likely that "smart" MC methods are likely to join MD as standard techniques in the simulation repertoire before long.

The enzyme HIV-1 protease and its inhibitors served as the main test system in this work. All proteins are not alike, and familiarity with the particularities of a given system of ligand

and receptor is essential for successful modelling. This makes a focus on one example a matter of efficiency. But apart from this consideration, the explosion of attention devoted to structure-based design of HIV-1 PR inhibitors, and the resultant overabundance of crystallographic data, makes it a unique object of study. The availability of large data sets of reasonable accuracy and consistency, all relevant to one protein, has created a wonderful opportunity for models to be evaluated and refined. HIV protease has also proven to provide fertile ground for new developments such as knowledge-based potentials for the scoring of ligand–protein interactions, and the investigation of the molecular basis of drug resistance [6]. HIV-1 PR sequences from patients receiving antiretroviral therapy show polymorphism in 67 of the 99 amino acids [120] making this of vital therapeutic importance, as well as a fascinating subject for investigation.

The observation that the devil is in the details, although a truism for scientific research, sums up the contents of this thesis well. The consequences of many simple physical principles, and the application of numerous established theoretical methods, still remain to be worked out for biochemical system. The volume of experimental data in want of explanation is vast, and growing ever more quickly. Luckily the devil has all the best tunes.

Bibliography

[1] Abagyan, R. and M. Totrov (1994). Biased probability Monte Carlo conformational searches and electrostatic calculations for peptides and proteins. *J. Mol. Biol.*, 235:983–1002

[2] Abagyan, R. A. and M. Totrov (1999). *Ab Initio* folding of peptides by the optimal-bias Monte Carlo minimization procedure. *J. Comput. Phys.*, 151:402–421

[3] Advanced Chemistry Development Inc, Toronto ON, Canada (2001). ACD/I-Lab web service. URL `http://www.acdlabs.com/ilab/`

[4] Ajay and M. A. Murcko (1995). Computational methods to predict binding free energy in ligand-receptor complexes. *J. Med. Chem.*, 38:4953–4967

[5] Ala, P. J., R. J. DeLoskey, E. E. Huston, P. K. Jadhav, P. Y. S. Lam, C. J. Eyermann, C. N. Hodge, M. C. Schadt, F. A. Lewandowski, P. C. Weber, D. D. McCabe, J. L. Duke, and C.-H. Chang (1998). Molecular recognition of cyclic urea HIV-1 protease inhibitors. *J. Biol. Chem.*, 273:12325–12331

[6] Ala, P. J., E. E. Huston, R. M. Klabe, D. D. McCabe, J. L. Duke, C. J. Rizzo, B. D. Korant, R. J. DeLoskey, P. Y. S. Lam, C. N. Hodge, and C.-H. Chang (1997). Molecular basis of HIV-1 protease drug resistance: Structural analysis of mutant proteases complexed with cyclic urea inhibitors. *Biochemistry*, 36:1573–1580

[7] Alder, B. J. and T. E. Wainwright (1958). Molecular dynamics by electronic computers. In I. Prigogine, editor, *Proceedings of the International Symposium on statistical mechanical theory of transport processes (Brussels, 1956)*, 97–131. Wiley, New York, USA

[8] Alexov, E. G. and M. R. Gunner (1997). Incorporating protein conformational flexibility into the calculation of pH-dependent protein properties. *Biophys. J.*, 74:2075–2093

[9] Allen, M. P. and D. J. Tildesley (1987). *Computer Simulations of Liquids*. Oxford University Press, Oxford, UK

[10] Andersen, H. C. (1983). Rattle: A "velocity" version of the Shake algorithm for molecular dynamics calculations. *J. Comput. Phys.*, 52:24–34

[11] Antosiewicz, J., J. M. Briggs, A. H. Elcock, M. K. Gilson, and J. A. McCammon (1996). Computing ionization states of proteins with a detailed charge model. *J. Comput. Chem.*, 17:1633–1644

[12] Antosiewicz, J. and J. A. McCammon (1996). The determinants of pK_as in proteins. *Biochemistry*, 35:7819–7833

[13] Antosiewicz, J., J. A. McCammon, and M. K. Gilson (1994). Prediction of pH-dependent properties of proteins. *J. Mol. Biol.*, 238:415–436

[14] Antosiewicz, J. and D. Pörschke (1989). The nature of protein dipole moments: Experimental and calculated permanent dipole of α-chymotrypsin. *Biochemistry*, 28:10072–10078

[15] Åqvist, J., C. Medina, and J.-E. Samuelsson (1994). A new method for predicting binding affinity in computer-aided drug design. *Protein Eng.*, 7:385–391

[16] Baptista, A. M., P. J. Martel, and S. B. Petersen (1997). Simulation of protein conformational freedom as a function of pH: Constant-pH molecular dynamics using implicit titration. *Proteins: Struct., Funct., Genet.*, 27:523–544

[17] Barbas III, C. F., A. Heine, G. Zhong, T. Hoffmann, S. Gramatikova, R. Björnestedt, B. List, J. Anderson, E. A. Stura, I. A. Wilson, and R. A. Lerner (1997). Immune versus natural selection: Antibody aldolases with enzymic rates but broader scope. *Science*, 278:2085–2092

[18] Bardi, J. S., I. Luque, and E. Freire (1997). Structure-based thermodynamic analysis of HIV-1 protease inhibitors. *Biochemistry*, 36:6588–6596

[19] Bartik, K., C. Redfield, and C. M. Dobson (1994). Measurement of the individual pK_a values of acidic residues of hen and turkey lysozymes by two-dimensional ^{1}H NMR. *Biophys. J.*, 66:1180–1184

[20] Bascle, J., T. Garel, H. Orland, and B. Velikson (1993). Biasing a Monte Carlo chain growth method with Ramachandran's plot: Application to twenty-L-alanine. *Biopolymers*, 33:1843–1849

[21] Bashford, D. and M. Karplus (1990). pK_a's of ionizable groups in proteins: Atomic detail from a continuum electrostatic model. *Biochemistry*, 29:10219–10225

[22] Batoulis, J. and K. Kremer (1988). Statistical properties of biased sampling methods for long polymer chains. *J. Phys. A: Math. Gen.*, 21:127–146

[23] Beachy, M. D., D. Chasman, R. B. Murphy, T. A. Halgren, and R. A. Friesner (1997). Accurate ab initio quantum chemical determination of the relative energetics of peptide conformations and assessment of empirical force fields. *J. Am. Chem. Soc.*, 119:5908–5920

[24] Berisio, R., V. S. Lamzin, F. Sica, K. S. Wilson, A. Zagari, and L. Mazzarella (1999). Protein titration in the crystal state. *J. Mol. Biol.*, 292:845–854

[25] Berman, H. M., J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne (2000). The Protein Data Bank. *Nucleic Acids Res.*, 28:235–242

[26] Bernstein, F. C., T. F. Koetzle, G. J. Williams, E. F. M. Jr, M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi (1977). The Protein Data Bank: A computer-based archival file for macromolecular structures. *J. Mol. Biol.*, 112:535–542

[27] Best, P. and M. Pearson (2000). Docking is painful. *Aust. Vet. J.*, 78:238

[28] Blomberg, N., R. R. Gabdoulline, M. Nilges, and R. C. Wade (1999). Classification of protein sequences by homology modeling and quantitative analysis of electrostatic similarity. *Proteins: Struct., Funct., Genet.*, 37:379–387

[29] Böhm, H.-J. (1994). The development of a simple empirical scoring function to estimate the binding constant for a protein–ligand complex of known three-dimensional structure. *J. Comput.-Aided Mol. Des.*, 8:243–56

[30] Böhm, H.-J. (1998). Prediction of binding constants of protein ligands: A fast method for the prioritization of hits obtained from de novo design or 3D database search programs. *J. Comput.-Aided Mol. Des.*, 12:309–323

[31] Böhm, H.-J. and G. Klebe (1996). What can we learn from molecular recognition in protein–ligand complexes for the design of new drugs? *Angew. Chem., Int. Ed.*, 35:2588–2614

[32] Bolhuis, P. and D. Frenkel (1994). Numerical study of the phase diagram of a mixture of spherical and rodlike colloids. *J. Chem. Phys.*, 101:9869–9875

[33] Boström, J., P.-O. Norrby, and T. Liljefors (1998). Conformational energy penalties of protein-bound ligands. *J. Comput.-Aided Mol. Des.*, 12:383–396

[34] Bowman, A. W. and A. Azzalini (1997). *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations*. Oxford University Press, Oxford, UK

[35] Briggs, J. M. and J. Antosiewicz (1999). Simulation of pH-dependent properties of proteins. In K. B. Lipkowitz and D. B. Boyd, editors, *Reviews in Computational Chemistry*, volume 13, 249–311. Wiley-VCH, New York, USA

[36] Brooks, B. R., R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus (1983). CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.*, 4:187–217

[37] Brooks III, C. L. and D. A. Case (1993). Simulations of peptide conformational dynamics and thermodynamics. *Chem. Rev.*, 93:2487–2502

[38] Brooks III, C. L., M. Karplus, and B. M. Pettitt (1988). *Proteins: a theoretical perspective of dynamics, structure, and thermodynamics*, volume 71 of *Advances in chemical physics*. Wiley, New York, USA

[39] Bruccoleri, R. E., J. Novotny, M. E. Davis, and K. A. Sharp (1997). Finite difference Poisson-Boltzmann electrostatic calculations: Increased accuracy achieved by harmonic dielectric smoothing and charge antialiasing. *J. Comput. Chem.*, 18:268–276

[40] Chau, P.-L. and P. M. Dean (1994). Electrostatic complementarity between proteins and ligands 1. Charge disposition, dielectric and interface effects. *J. Comput.-Aided Mol. Des.*, 8:513–525

[41] Chau, P.-L. and P. M. Dean (1994). Electrostatic complementarity between proteins and ligands 3. Structural basis. *J. Comput.-Aided Mol. Des.*, 8:545–564

[42] Chen, X. and A. Tropsha (1995). Relative binding free energies of peptide inhibitors of HIV-1 protease: The influence of active site protonation state. *J. Med. Chem.*, 38:42–48

[43] Chiu, S. W., M. M. Clark, E. Jakobsson, S. Subramaniam, and H. L. Scott (1999). Application of combined Monte Carlo and molecular dynamics method to simulation of dipalmitoyl phosphatidylcholine lipid bilayer. *J. Comput. Chem.*, 20:1153–1164

[44] Chong, L. T., S. E. Dempster, Z. S. Hendsch, L.-P. Lee, and B. Tidor (1998). Computation of electrostatic complements to proteins: A case of charge stabilized binding. *Protein Sci.*, 7:206–210

[45] Chun, H. M., C. E. Padilla, D. N. Chin, M. Watanabe, V. I. Karlov, H. E. Alper, K. Soosaar, K. B. Blair, O. M. Becker, L. S. D. Caves, R. Nagle, D. N. Haney, and B. L. Farmer (2000). MBO(N)D: a multibody method for long-time molecular dynamics simulations. *J. Comput. Chem.*, 21:159–184

[46] Connolly, M. L. (1983). Analytical molecular surface recognition. *J. Appl. Crystallogr.*, 16:548–558

[47] Consta, S., T. J. H. Vlugt, J. W. Hoeth, B. Smit, and D. Frenkel (1999). Recoil growth algorithm for chain molecules with continuous interactions. *Mol. Phys.*, 97:1243–1254

[48] Consta, S., N. B. Wilding, D. Frenkel, and Z. Alexandrowicz (1999). Recoil growth: An efficient simulation method for multi-polymer systems. *J. Chem. Phys.*, 110:3220–3228

[49] Copeland, T. D. and S. Oroszlan (1988). Genetic locus, primary structure, and chemical synthesis of human immunodeficiency virus protease. *Gene Anal. Techn.*, 5:109–115

[50] Cornell, W. D., P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, Jr., D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman (1995). A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.*, 117

[51] Costante-Crassous, J., T. J. Marrone, J. M. Briggs, J. A. McCammon, and A. Collet (1997). Absolute configuration of bromochlorofluoromethane from molecular dynamics simulation of its enantioselective complexation by cryptophane-C. *J. Am. Chem. Soc.*, 119:3818–3823

[52] Davies, T. G., R. E. Hubbard, and J. R. H. Tame (1999). Relating structure to thermodynamics: The crystal structures and binding affinity of eight OppA-peptide complexes. *Protein Sci.*, 8:1432–1444

[53] Davis, A. M. and S. J. Teague (1999). Hydrogen bonding, hydrophobic interactions, and failure of the rigid receptor hypothesis. *Angew. Chem., Int. Ed.*, 38:736–749

[54] Deem, M. W. and J. S. Bader (1996). A configurational bias Monte Carlo method for linear and cyclic peptides. *Mol. Phys.*, 87:1245–1260

[55] Demchuk, E. and R. C. Wade (1996). Improving the continuum dielectric approach to calculating pK_as of ionizable groups in proteins. *J. Phys. Chem.*, 100:17373–17387

[56] Derreumaux, P. (1997). Folding a 20 amino acid $\alpha\beta$ peptide with the diffusion process-controlled Monte Carlo method. *J. Chem. Phys.*, 107:1941–1947

[57] Derreumaux, P. (1999). Ab initio prediction of polypeptide structure from its sequence. *Comput. Phys. Commun.*, 121–122:139–140

[58] Desmet, J., M. De Maeyer, J. Spriet, and I. Lasters (2000). Flexible docking of peptide ligands to proteins. In D. Webster, editor, *Protein Structure Prediction: Methods and Protocols*, volume 143 of *Methods in Molecular Biology*, chapter 16, 359–376. Humana Press, Totowa, NJ, USA

[59] DeWitte, R. S. and E. I. Shakhnovich (1996). SMoG: de novo design method based on simple, fast and accurate free energy estimates. 1. Methodology and supporting evidence. *J. Am. Chem. Soc.*, 118:11733–11744

[60] Dijkstra, M. and D. Frenkel (1994). Evidence of entropy-driven demixing in hard-core fluids. *Phys. Rev. Lett.*, 72:298–300

[61] Dijkstra, M., D. Frenkel, and J.-P. Hansen (1994). Phase separation in binary hard-core mixtures. *J. Chem. Phys.*, 101:3179–3189

[62] Dinur, U. and A. T. Hagler (1991). New approaches to empirical force fields. In K. B. Lipkowitz and D. B. Boyd, editors, *Reviews in Computational Chemistry*, volume 2, chapter 4, 99–164. Wiley-VCH, New York, USA

[63] Dodd, L. R., T. D. Boone, and D. N. Theodorou (1993). A concerted rotation algorithm for atomistic Monte Carlo simulation of polymer melts and glasses. *Mol. Phys.*, 78:961–996

[64] Dorsey, B. D., R. B. Levin, S. L. McDaniel, J. P. Vacca, J. P. Guare, P. L. Darke, J. A. Zugay, E. A. Emini, W. A. Schleif, J. C. Quintero, J. H. Lin, I.-W. Chen, M. K. Holloway, P. M. D. Fitzgerald, M. G. Axel, D. Ostovic, P. S. Anderson, and J. R. Huff (1994). L-753,524: the design of a potent and orally bioavailable HIV protease inhibitor. *J. Med. Chem.*, 37:3443–3451

[65] Dreyer, G. B., D. M. Lambert, T. D. Meek, T. J. Carr, J. Thaddeus A. Tomaszek, A. V. Fernandez, H. Bartus, E. Cacciavillani, A. M. Hassell, M. Minnich, J. Stephen R. Petteway, and B. W. Metcalf (1992). Hydroxyethylene isostere inhibitors of human immunodeficiency virus-1 protease: Structure-activity analysis using enzyme kinetics, X-ray crystallography and infected T-cell assays. *Biochemistry*, 31:6646–6659

[66] Duan, Y. and P. A. Kollman (1998). Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. *Science*, 282:740–744

[67] Duane, S., A. D. Kennedy, B. J. Pendleton, and D. Roweth (1987). Hybrid Monte Carlo. *Phys. Lett. B*, 195:216–222

[68] Duffy, E. M. and W. L. Jorgensen (2000). Prediction of properties from simulations: Free energies of solvation in hexadecane, octanol, and water. *J. Am. Chem. Soc.*, 122:2878–2888

[69] Eldridge, M. D., C. W. Murray, T. R. Auton, G. V. Paolini, and R. P. Mee (1997). Empirical scoring functions: I. the development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J. Comput.-Aided Mol. Des.*, 11:425–445

[70] Erickson, J., D. J. Heidhart, J. VanDrie, D. J. Kempf, X. C. Wang, D. W. Norbeck, J. J. Plattner, J. W. Rittenhouse, M. Turon, N. Wideburg, W. E. Kohlbrenner, R. Simmer, R. Helfrich, D. A. Paul, and M. Knigge (1990). Design, activity and 2.8 Å crystal structure of a C_2 symmetric inhibitor complexed to HIV-1 protease. *Science*, 249:527–533

[71] Esselink, K., L. D. J. C. Loyens, and B. Smit (1995). Parallel Monte Carlo simulations. *Phys. Rev. E*, 51:1560–1568

[72] Ewald, P. P. (1921). Die berechnung optischer und elektrostatischer gitterpotentiale. *Ann. Phys.*, 64:253–287

[73] Feig, M., P. Rotkiewicz, A. Kolinski, J. Skolnick, and C. L. Brooks III (2000). Accurate reconstruction of all-atom protein representations from side-chain-based low-resolution models. *Proteins: Struct., Funct., Genet.*, 41:86–97

[74] Fitzgerald, P. M. D., B. M. McKeever, J. F. VanMiddlesworth, J. P. Springer, J. C. Heimbach, C.-T. Leu, W. K. Herber, R. A. F. Dixon, and P. L. Darke (1990). Crystallographic analysis of a complex between human immunodeficiency virus type 1 protease and acetyl-pepstatin at 2.0-Å resolution. *J. Biol. Chem.*, 265:14209–14219

[75] Frauenkron, H., U. Bastolla, E. Gerstner, P. Grassberger, and W. Nadler (1998). New Monte Carlo algorithm for protein folding. *Phys. Rev. Lett.*, 80:3149–3152

[76] Frenkel, D., G. C. A. M. Mooij, and B. Smit (1991). Novel scheme to study structural and thermal properties of continuously deformable molecules. *J. Phys.: Condens. Matter*, 3:3053–3076

[77] Frenkel, D. and B. Smit (1996). *Understanding Molecular Simulation: From Algorithms to Applications*. Academic Press, San Diego, USA

[78] Friedman, A. R., V. A. Roberts, and J. A. Tainer (1994). Predicting molecular interactions and inducible complementarity: Fragment docking of Fab-peptide complexes. *Proteins: Struct., Funct., Genet.*, 20:15–24

[79] Froloff, N., A. Windemuth, and B. Honig (1997). On the calculation of binding free energies using continuum methods: Application to MHC class I protein–peptide interactions. *Protein Sci.*, 6:1293–1301

[80] Gehlhaar, D. K., K. E. Moerder, D. Zichi, C. J. Sherman, R. C. Ogden, and S. T. Freer (1995). *De novo* design of enzyme inhibitors by Monte Carlo ligand generation. *J. Med. Chem.*, 38:466–472

[81] Geller, M., M. Miller, S. M. Swanson, and J. Maizel (1997). Analysis of the structure of HIV-1 protease complexed with a hexapeptide inhibitor. Part II: molecular dynamic studies of the active site region. *Proteins: Struct., Funct., Genet.*, 27:195–203

[82] Geyer, C. J. (1992). Practical Markov chain Monte Carlo. *Stat. Sci.*, 7:473–511

[83] Gilson, M. K. (1993). Multiple-site titration and molecular modeling: Two rapid methods for computing energies and forces for ionizable groups in proteins. *Proteins: Struct., Funct., Genet.*, 15:266–282

[84] Gilson, M. K., J. A. Given, and M. S. Head (1997). A new class of models for predicting receptor–ligand binding affinities. *Chem. Biol.*, 4:87–92

[85] Gilson, M. K. and B. Honig (1988). Calculation of the total electrostatic energy of a macromolecular system: solvation energies, binding energies, and conformational analysis. *Proteins: Struct., Funct., Genet.*, 4:7–18

[86] Gilson, M. K., J. A. McCammon, and J. D. Madura (1995). Molecular dynamics simulation with a continuum electrostatic model of the solvent. *J. Comput. Chem.*, 16:1081–1095

[87] Gohlke, H., M. Hendlich, and G. Klebe (2000). Knowledge-based scoring function to predict protein-ligand interactions. *J. Mol. Biol.*, 295:337–356

[88] Goodsell, D. S. and A. J. Olson (1990). Automated docking of substrates to proteins by simulated annealing. *Proteins: Struct., Funct., Genet.*, 8:195–202

[89] Greer, J., J. W. Erickson, J. J. Baldwin, and M. D. Varney (1994). Application of the three-dimensional structures of protein target molecules in structure-based drug design. *J. Med. Chem.*, 37:1035–1054

[90] Guarnieri, F. and W. C. Still (1994). A rapidly convergent simulation method: Mixed Monte Carlo/stochastic dynamics. *J. Comput. Chem.*, 15:1302–1310

[91] van Gunsteren, W. F., P. H. Hünenberger, A. E. Mark, P. E. Smith, and I. G. Tironi (1995). Computer simulation of protein motion. *Comput. Phys. Commun.*, 91:305–319

[92] van Gunsteren, W. F. and M. Karplus (1982). Effects of constraints on the dynamics of macromolecules. *Macromolecules*, 15:1528–1544

[93] Halgren, T. A. (1996). Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *J. Comput. Chem.*, 17:490–519

[94] Hansch, C. and A. Leo (1995). *Exploring QSAR*. American Chemical Society, Washington DC, USA

[95] Hansmann, U. H. E. (1997). Parallel tempering algorithm for conformational studies of biological molecules. *Chem. Phys. Lett.*, 281:140–150

[96] Hansson, T. and J. Åqvist (1995). Estimation of binding free energies for HIV proteinase inhibitors by molecular dynamics simulations. *Protein Eng.*, 8:1137–1144

[97] Harte, W. E., Jr. and D. L. Beveridge (1993). Prediction of the protonation state of the active site aspartyl residues in HIV-1 protease-inhibitor complexes via molecular dynamics simulation. *J. Am. Chem. Soc.*, 115:3883–3886

[98] Havranek, J. J. and P. B. Harbury (1999). Tanford-Kirkwood electrostatics for protein modeling. *Proc. Natl. Acad. Sci. U. S. A.*, 96:11145–11150

[99] Head, M. S., J. A. Given, and M. K. Gilson (1997). Mining minima: Direct computation of conformational free energy. *J. Phys. Chem. A*, 101:1609–1618

[100] Head, R. D., M. L. Smythe, T. I. Oprea, C. L. Waller, S. M. Green, and G. R. Marshall (1996). VALIDATE: a new method for the receptor-based prediction of binding affinities of novel ligands. *J. Am. Chem. Soc.*, 118:3959–3969

[101] Head-Gordon, T. and C. L. Brooks III (1987). The role of electrostatics in the binding of small ligands to enzymes. *J. Phys. Chem.*, 91:3342–3349

[102] Holmes, D. S., R. C. Bethell, H. Cammack, I. R. Clemens, J. Kitchin, P. McMeekin, C. L. Mo, D. C. Orr, B. Patel, I. L. Paternoster, and R. Storer (1993). Synthesis and structure–activity relationships of a series of penicillin-derived HIV proteinase inhibitors containing a stereochemically unique peptide isostere. *J. Med. Chem.*, 36:3129–3136

[103] Hooft, R. W. W., C. Sander, and G. Vriend (1996). Positioning hydrogen atoms by optimizing hydrogen-bond networks in protein structures. *Proteins: Struct., Funct., Genet.*, 26:363–376

[104] Hooft, R. W. W., G. Vriend, C. Sander, and E. E. Abola (1996). Errors in protein structures. *Nature*, 381:272

[105] Hoog, S. S., B. Zhao, E. Winborne, S. Fisher, D. W. Green, R. L. DesJarlais, K. A. Newlander, J. F. Callahan, M. L. Moore, W. F. Huffman, and S. S. Abdel-Meguid (1995). A check on rational drug design: Crystal structure of a complex of human immunodeficiency virus type 1 protease with a novel γ-turn mimetic inhibitor. *J. Med. Chem.*, 38:3246–3252

[106] Hornik, K. (2001). *ctest:* classical tests for R. version 1.2.0

[107] Hosur, M. V., T. N. Bhat, D. J. Kempf, E. T. Baldwin, B. Liu, S. Gulnik, N. E. Wideburg, D. W. Norbeck, K. Appelt, and J. W. Erickson (1994). Influence of stereochemistry on activity and binding modes for C_2 symmetry-based diol inhibitors of HIV-1 protease. *J. Am. Chem. Soc.*, 116:847–855

[108] Hünenberger, P. H. and J. A. McCammon (1999). Effect of artificial periodicity in simulations of biomolecules under Ewald boundary conditions: a continuum electrostatics study. *Biophys. Chem.*, 78:69–88

[109] Hyland, L. J., J. Thaddeus A. Tomaszek, and T. D. Meek (1991). Human immunodeficiency virus-1 protease. 2. Use of pH rate studies and solvent kinetic isotope effects to elucidate details of chemical mechanism. *Biochemistry*, 30:8454–8463

[110] Ido, E., H. P. Han, F. J. Kezdy, and J. Tang (1991). Kinetic studies of human immunodeficiency virus type 1 protease and its active-site hydrogen bond mutant A28S. *J. Biol. Chem.*, 266:24359–24366

[111] Ihaka, R. and R. Gentleman (1996). R: A language for data analysis and graphics. *J. Comp. Graph. Stat.*, 5:299–314

[112] Ishima, R., D. I. Freedberg, Y.-X. Wang, J. M. Louis, and D. A. Torchia (1999). Flap opening and dimer-interface flexibility in the free and inhibitor-bound HIV protease, and their implications for function. *Structure*, 7:1047–1055

[113] von Itzstein, M., W.-Y. Wu, G. B. Kok, M. S. Pegg, J. C. Dyason, B. Jin, T. V. Phan, M. L. Smythe, H. F. White, S. W. Oliver, P. M. Colman, J. H. Varghese, D. M. Ryan, J. M. Woods, R. C. Bethell, V. J. Hotham, J. M. Cameron, and C. R. Penn (1993). Rational design of potent sialidase-based inhibitors of influenza virus replication. *Nature*, 363:418–423

[114] Jaskólski, M., A. G. Tomasselli, T. K. Sawyer, D. G. Staples, R. L. Heinrikson, J. Schneider, S. B. H. Kent, and A. Wlodawer (1991). Structure at 2.5-Å resolution of chemically synthesized human immunodeficiency virus type 1 protease complexed with a hydroxyethylene-based inhibitor. *Biochemistry*, 30:1600–1609

[115] Jhoti, H., O. M. P. Singh, M. P. Weir, R. Cooke, P. Murray-Rust, and A. Wonacott (1994). X-ray crystallographic studies of a series of penicillin-derived asymmetric inhibitors of HIV-1 protease. *Biochemistry*, 33:8417–8427

[116] Jorgensen, W. L. and J. Tirado-Rives (1996). Monte Carlo vs molecular dynamics for conformational sampling. *J. Phys. Chem.*, 100:14508–14513

[117] Joseph-McCarthy, D. (1999). Computational approaches to structure-based ligand design. *Pharmacol. Ther.*, 84:179–191

[118] Kangas, E. and B. Tidor (1999). Charge optimization leads to favorable electrostatic binding free energy. *Phys. Rev. E*, 59:5958–5961

[119] Kangas, E. and B. Tidor (2000). Electrostatic specificity in molecular ligand design. *J. Chem. Phys.*, 112:9120–9131

[120] Kantor, R., R. Machekano, M. J. Gonzalez, K. Dupnik, J. M. Schapiro, and R. W. Shafer (2001). Human Immunodeficiency Virus Reverse Transcriptase and Protease Sequence Database: an expanded data model integrating natural language text and sequence analysis programs. *Nucleic Acids Res.*, 29:296–299

[121] Kasper, P., P. Christen, and H. Gehring (2000). Empirical calculation of the relative free energies of peptide binding to the molecular chaperone DnaK. *Proteins: Struct., Funct., Genet.*, 40:185–192

[122] Kick, E. K., D. C. Roe, A. G. Skillman, G. Liu, T. J. A. Ewing, Y. Sun, I. D. Kuntz, and J. A. Ellman (1997). Structure-based design and combinatorial chemistry yield low nanomolar inhibitors of cathepsin D. *Chem. Biol.*, 4:29700307

[123] Kim, E. E., C. T. Baker, M. D. Dwyer, M. A. Murcko, B. G. Rao, R. D. Tung, and M. A. Navia (1995). Crystal structure of HIV-1 protease in complex with VX-478, a potent and orally bioavailable inhibitor of the enzyme. *J. Am. Chem. Soc.*, 117:1181–1182

[124] Kohl, N. E., E. A. Emini, W. A. Schleif, L. J. Davis, J. C. Heimbach, R. A. F. Dixon, E. M. Scolnick, and I. S. Sigal (1988). Active human immunodeficiency virus protease is required for viral infectivity. *Proc. Natl. Acad. Sci. U. S. A.*, 85:4686–4690

[125] Kulkarni, S. S. and V. M. Kulkarni (1999). Structure based prediction of binding affinity of human immunodeficiency virus-1 protease inhibitors. *J. Chem. Inf. Comput. Sci.*, 39:1128–1140

[126] Kuntz, I. D., J. M. Blaney, S. J. Oatley, R. Langridge, and T. E. Ferrin (1982). A geometric approach to macromolecule–ligand interactions. *J. Mol. Biol.*, 161:269–288

[127] Kuramitsu, S. and K. Hamaguchi (1980). Analysis of the acid-base titration curve of hen lysozyme. *J. Biochem.*, 87:1215–1219

[128] Lam, P. Y. S., P. K. Jadhav, C. J. Eyermann, C. N. Hodge, Y. Ru, L. T. Bacheler, J. L. Meek, M. J. Otto, M. M. Rayner, Y. N. Wong, C.-H. Chang, P. C. Weber, D. A. Jackson, T. R. Sharpe, and S. Erickson-Viitanen (1994). Rational design of potent, bioavailable, nonpeptide cyclic ureas as HIV protease inhibitors. *Science*, 263:380–383

[129] Lam, P. Y. S., Y. Ru, P. K. Jadhav, P. E. Aldrich, G. V. DeLucca, C. J. Eyermann, C.-H. Chang, G. Emmett, E. R. Holler, W. F. Daneker, L. Li, P. N. Confalone, R. J. McHugh, Q. Han, R. Li, J. A. Markwalder, S. P. Seitz, T. R. Sharpe, L. T. Bacheler, M. M. Rayner, R. M. Klabe, L. Shum, D. L. Winslow, D. M. Kornhauser, D. A. Jackson, S. Erickson-Viitanen, and C. H. Hodge (1996). Cyclic HIV protease inhibitors: Synthesis, conformational analysis, P2/P2′ structure-activity relationship, and molecular recognition of cyclic ureas. *J. Med. Chem.*, 39:3514–3525

[130] LAM Team, University of Notre Dame, Notre Dame, IN, USA (2001). Local Area Multicomputer. URL `http://www.mpi.nd.edu/lam/`

[131] Lamb, M. L. and W. L. Jorgensen (1997). Computational approaches to molecular recognition. *Curr. Opin. Chem. Biol.*, 1:449–457

[132] Lapatto, R., T. Blundell, A. Hemmings, J. Overington, A. Wilderspin, S. Wood, J. R. Merson, P. J. Whittle, D. E. Danley, K. F. G. S. J. Hawrylik, S. E. Lee, K. G. Scheld, and P. M. Hobart (1989). X-ray analysis of HIV-1 proteinase at 2.7 Å resolution confirms structural homology among retroviral enzymes. *Nature*, 342:299–302

[133] Ledvina, P. S., N. Yao, A. Choudhary, and F. A. Quiocho (1996). Negative electrostatic surface potential of protein sites specific for anionic ligands. *Proc. Natl. Acad. Sci. U. S. A.*, 93:6786–6791

[134] Lee, J. (1993). New Monte Carlo algorithm: Entropic sampling. *Phys. Rev. Lett.*, 71:211–214

[135] de Leeuw, S. W., J. W. Perram, and E. R. Smith (1980). Simulation of electrostatic systems in periodic boundary conditions. I. Lattice sums and dielectric constants. *Proc. Royal Soc. London A*, 373:27–56

[136] de Leeuw, S. W., J. W. Perram, and E. R. Smith (1980). Simulation of electrostatic systems in periodic boundary conditions. II. Equivalence of boundary conditions. *Proc. Royal Soc. London A*, 373:56–66

[137] Levy, R. M. and E. Gallicchio (1998). Computer simulations with explicit solvent: Recent progress in the thermodynamic decomposition of free energies and in modeling electrostatic effects. *Annu. Rev. Phys. Chem.*, 49:531–567

[138] Lin, Y., X. Lin, L. Hong, S. Foundling, R. L. Heinrikson, S. Thaisrivongs, W. Leelamanit, D. Raterman, M. Shah, B. M. Dunn, and J. Tang (1995). Effect of point mutations on the kinetics and the inhibition of human immunodeficiency virus type 1 protease: Relationship to drug resistance. *Biochemistry*, 34:1143–1152

[139] Louis, J. M., F. Dyda, N. T. Nashed, A. R. Kimmel, and D. R. Davies (1998). Hydrophilic peptides derived from the transframe region of Gag-Pol inhibit the HIV-1 protease. *Biochemistry*, 37:2105–2110

[140] Luque, I. and E. Freire (1998). Structure-based prediction of binding affinities and molecular design of peptide ligands. *Meth. Enzym.*, 295:100–127

[141] MacKerrell, A. D., Jr., D. Bashford, M. Bellott, R. L. Dunbrack, Jr., J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, III, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkiewicz-Kuczera, D. Yin, and M. Karplus (1998). All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B*, 102:3586–3616

[142] Madura, J. D., J. M. Briggs, R. C. Wade, M. E. Davis, B. A. Luty, A. Ilin, J. Antosiewicz, M. K. Gilson, B. Bagheri, L. R. Scott, and J. A. McCammon (1995). Electrostatics and diffusion of molecules in solution: simulations with the University of Houston Brownian dynamics program. *Comput. Phys. Commun.*, 91:57–95

[143] Maple, J., U. Dinur, and A. T. Hagler (1988). Derivation of force fields for molecular mechanics and dynamics from *ab initio* energy surfaces. *Proc. Natl. Acad. Sci. U. S. A.*, 85:5350–5354

[144] Martin, M. G. and J. I. Siepmann (1999). Novel configurational-bias Monte Carlo method for branched molecules. Transferable potentials for phase equilibria. 2. United-atom description of branched alkanes. *J. Phys. Chem. B*, 103:4508–4517

[145] McCammon, J. A. (1998). Theory of biomolecular recognition. *Curr. Opin. Struct. Biol.*, 8:245–249

[146] McCoy, A. J., V. C. Epa, and P. M. Colman (1997). Electrostatic complementarity at protein/protein interfaces. *J. Mol. Biol.*, 268:570–584

[147] McDonald, I. K. and J. M. Thornton (1994). Satisfying hydrogen bonding potential in proteins. *J. Mol. Biol.*, 238:777–793

[148] Meirovitch, H. (1985). Scanning method as an unbiased simulation technique and its application to the study of self-attracting random walks. *Phys. Rev. A*, 32:3699–3708

[149] Meirovitch, H. (1988). Statistical properties of the scanning simulation method for polymer chains. *J. Chem. Phys.*, 89:2514–2522

[150] Meirovitch, H., M. Vásquez, and H. A. Scheraga (1990). Free energy and stability of macromolecules studied by the double scanning simulation procedure. *J. Chem. Phys.*, 92:1248–1257

[151] Message Passing Interface Forum (1994). MPI: A message-passing interface standard. *Int. J. Supercomput. Appl.*, 8

[152] Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller (1953). Equation of state calculations by fast computing machines. *J. Chem. Phys.*, 21:1087–1092

[153] Meyer, E. F., I. Botos, L. Scapozza, and D. Zhang (1995). Backward binding and other structural surprises. *Perspect. Drug Discovery Des.*, 3:168–195

[154] Miller, M. HIV-1 PR–MVT-101 complex crystal structure coordinates. Pers. comm.

[155] Miller, M., M. Geller, M. Gribskov, and S. B. H. Kent (1997). Analysis of the structure of chemically synthesized HIV-1 protease complexed with a hexapeptide inhibitor. Part I: Crystallographic refinement of 2Å data. *Proteins: Struct., Funct., Genet.*, 27:184–194

[156] Miller, M., J. Schneider, B. K. Sathyanarayana, M. V. Toth, G. R. Marshall, L. Clawson, L. Selk, S. B. H. Kent, and A. Wlodawer (1989). Structure of complex of synthetic HIV-1 protease with a substrate-based inhibitor at 2.3Å resolution. *Science*, 246:1149–1152

[157] Mitchell, J. B. O., R. A. Laskowski, A. Alex, and J. M. Thornton (1999). BLEEP—potential of mean force describing protein–ligand interactions: I. generating potential. *J. Comput. Chem.*, 20:1165–1176

[158] Mohanty, D., B. N. Dominy, A. Kolinski, C. L. Brooks III, and J. Skolnick (99). Correlation between knowledge-based and detailed atomic potentials: Application to the unfolding of the GCN4 leucine zipper. *Proteins: Struct., Funct., Genet.*, 35:447–452

[159] de Mol, N. J., M. B. Gillies, and M. J. E. Fischer (2001). Experimental and calculated shift in pK_a upon binding of phosphotyrosine peptide to the SH2 domain of p56^{Lck}. In preparation

[160] Molecular Simulations, San Diego CA, USA (2000). Insight 2000

[161] Moret, E. E., M. C. van Wijk, A. S. Kostense, and M. B. Gillies (1999). Scoring peptide(mimetic)–protein interactions. *Med. Chem. Res.*, 9:604–620

[162] Morgan, B. P., J. M. Scholtz, M. D. Ballinger, I. D. Zipkin, and P. A. Bartlett (1991). Differential binding energy: A detailed evaluation of the influence of hydrogen-bonding and hydrophobic groups on the inhibition of thermolysin by phosphorus-containing inhibitors. *J. Am. Chem. Soc.*, 113:297–307

[163] Morris, G. M., D. S. Goodsell, R. S. Halliday, R. Huey, W. E. Hart, R. K. Belew, and A. J. Olson (1998). Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J. Comput. Chem.*, 19:1639–1662

[164] Morris, G. M., D. S. Goodsell, R. Huey, and A. J. Olson (1996). Distributed automated docking of flexible ligands to proteins: Parallel applications of AutoDock 2.4. *J. Comput.-Aided Mol. Des.*, 10:293–304

[165] Muegge, I. and Y. C. Martin (1999). A general and fast scoring function for protein–ligand interactions: A simplified potential approach. *J. Med. Chem.*, 42:791–804

[166] Muegge, I., Y. C. Martin, P. J. Hajduk, and S. W. Fesik (1999). Evaluation of PMF scoring in docking weak ligands to the FK506 binding protein. *J. Med. Chem.*, 42:3498–2503

[167] Murcko, M. A., P. R. Caron, and P. S. Charifson (1999). Structure-based design. In A. M. Doherty, editor, *Annual Reports in Medicinal Chemistry*, volume 34 of *Annual Reports in Medicinal Chemistry*, chapter 29, 297–306. Academic Press, San Diego, USA

[168] Murray, C. W., D. E. Clark, T. R. Auton, M. A. Firth, J. Li, R. A. Sykes, B. Waszkowycz, D. R. Westhead, and S. C. Young (1997). PRO_SELECT: combining structure-based drug design and combinatorial chemistry for rapid lead discovery. 1. Technology. *J. Comput.-Aided Mol. Des.*, 11:193–207

[169] Náray-Szabó, G. and G. G. Ferenczy (1995). Molecular electrostatics. *Chem. Rev.*, 95:829–847

[170] Navia, M. A., P. M. D. Fitzgerald, B. M. McKeever, C.-T. Leu, J. C. Heimbach, W. K. Herber, I. S. Sigal, P. L. Darke, and J. P. Springer (1989). Three-dimensional structure of aspartyl protease from human immunodeficiency virus HIV-1. *Nature*, 337:615–620

[171] Nicholls, A. and B. Honig (1991). A rapid finite difference algorithm, utilizing successive over-relaxation to solve the Poisson-Boltzmann equation. *J. Comput. Chem.*, 12:435–445

[172] Nielsen, J. E. Calculation methodology used in the WHAT IF pK_a calculation routines. URL `http://www.cmbi.kun.nl/gv/nielsen/pKa/calc_methodology.html`. Unpublished information from PhD thesis

[173] Nielsen, J. E., K. V. Andersen, B. Honig, R. W. W. Hooft, G. Klebe, G. Vriend, and R. C. Wade (1999). Improving macromolecular electrostatics calculations. *Protein Eng.*, 12:657–662

[174] di Nola, A. and A. T. Brünger (1998). Free energy calculations in globular proteins: Methods to reduce errors. *J. Comput. Chem.*, 19:1229–1240

[175] Norrby, P.-O. and T. Liljefors (1998). Automated molecular mechanics parameterization with simultaneous utilization of experimental and quantum mechanical data. *J. Comput. Chem.*, 19:1146–1166

[176] Novotny, J., R. E. Bruccoleri, M. Davis, and K. A. Sharp (1997). Empirical free energy calculations: A blind test and further improvements to the method. *J. Mol. Biol.*, 268:401–411

[177] Oprea, T. I. and G. R. Marshall (1998). Receptor-based prediction of binding affinities. *Perspect. Drug Discovery Des.*, 9–11:35–61

[178] Oprea, T. I., C. L. Waller, and G. R. Marshall (1994). 3D-QSAR of human immunodeficiency virus (I) protease inhibitors. III. interpretation of CoMFA results. *Drug Des. Discovery*, 12:29–51

[179] Oprea, T. I., C. L. Waller, and G. R. Marshall (1994). Three-dimensional quantitative structure–activity relationship of human immunodeficiency virus (I) protease inhibitors. 2. Predictive power using limited exploration of alternate binding modes. *J. Med. Chem.*, 37:2206–2215

[180] de Pablo, J. J., M. Laso, and U. W. Suter (1992). Simulation of polyethylene above and below the melting point. *J. Chem. Phys.*, 96:2395–2403

[181] Philippopoulos, M. and C. Lim (1999). Exploring the dynamic information content of a protein NMR structure: Comparison of a molecular dynamics simulation with the NMR and X-ray structures of *Escherichia coli* ribonuclease HI. *Proteins: Struct., Funct., Genet.*, 36:87–110

[182] Piana, S., D. Sebastiani, P. Carloni, and M. Parrinello (2001). An *Ab-initio* molecular dynamics-based assignment of the protonation state of pepstatin A/HIV-1 protease cleavage site. *J. Am. Chem. Soc.*. Submitted

[183] Polticelli, F., P. Ascenzi, M. Bolognesi, and B. Honig (1999). Structural determinants of trypsin affinity and specificity for cationic inhibitors. *Protein Sci.*, 8:2621–2629

[184] Ponder, J. (1999). *Tinker Users Guide*. Washington University, St Louis, MO, USA. URL `http://dasher.wustl.edu/tinker/`

[185] Qiu, D., P. S. Shenkin, F. P. Hollinger, and W. C. Still (1997). The GB/SA continuum model for solvation. A fast analytical method for the calculation of approximate Born radii. *J. Phys. Chem. A*, 101:3005–3014

[186] Radford, S. E., C. M. Dobson, and P. A. Evans (1992). The folding of hen lysozyme involves partially structured intermediates and multiple pathways. *Nature*, 358:302–307

[187] Rao, B. G. and M. A. Murcko (1994). Reversed stereochemical preference in binding of Ro 31-8959 to HIV-1 proteinase: A free energy perturbation analysis. *J. Comput. Chem.*, 15:1241–1253

[188] Rarey, M., B. Kramer, T. Lengauer, and G. Klebe (1996). A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.*, 261:470–489

[189] Ratner, L., W. Haseltine, R. Patarca, K. J. Livak, B. Starcich, S. F. Josephs, E. R. Doran, J. A. Rafalski, E. A. Witehorn, and K. Baumeister (1985). Complete nucleotide sequence of the AIDS virus, HTLV-III. *Nature*, 313:277–284

[190] Richards, A. D., R. Roberts, B. M. Dunn, M. C. Graves, and J. Kay (1989). Effective blocking of HIV-1 proteinase activity by characteristic inhibitors of aspartic proteinases. *FEBS Lett.*, 247:113–117

[191] Ripoll, D. R., Y. N. Vorobjev, A. Liwo, J. A. Vila, and H. A. Scheraga (1996). Coupling between folding and ionization equilibria: effects of pH on the conformational preferences of polypeptides. *J. Mol. Biol.*, 264:770–783

[192] Roitberg, A. E., R. E. Cachau, and K. A. Fidelis (2000). Catfee: Meeting on critical assessment of techniques for free energy evaluation. URL `http://uqbar.ncifcrf.gov/~catfee/`

[193] Rosenbluth, M. N. and A. W. Rosenbluth (1955). Monte Carlo calculation of the average extension of molecular chains. *J. Chem. Phys.*, 23:356–359

[194] Rosin, C. D., R. K. Belew, G. M. Morris, A. J. Olson, and D. S. Goodsell (1999). Coevolutionary analysis of resistance-evading peptidomimetic inhibitors of HIV-1 protease. *Proc. Natl. Acad. Sci. U. S. A.*, 96:1369–1374

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

[196] Schechter, I. and A. Berger (1967). On the size of the active site in proteases. I. papain. *Biochem. Biophys. Res. Commun.*, 27:157–162

[197] Schellman, J. A. (1975). Macromolecular binding. *Biopolymers*, 14:999–1018

[198] Schneider, J. and S. B. H. Kent (1988). Enzymatic activity of a synthetic 99 residue protein corresponding to the putative HIV-1 protease. *Cell*, 54:363–368

[199] Schofield, J. and M. A. Ratner (1998). Monte Carlo methods for short polypeptides. *J. Chem. Phys.*, 109:9177–9191

[200] Senderowitz, H. and W. C. Still (1998). Sampling potential energy surface of glycyl glycine peptide: Comparison of Metropolis Monte Carlo and stochastic dynamics. *J. Comput. Chem.*, 19:1294–1299

[201] Sham, Y. Y., Z. T. Chu, H. Tao, and A. Warshel (2000). Examining methods for calculations of binding free energies: LRA, LIE, PDLD-LRA, and PDLD/S-LRA calculations of ligands binding to an HIV protease. *Proteins: Struct., Funct., Genet.*, 39:393–407

[202] Sharp, K. A. (1998). Calculation of HyHel10–lysozyme binding free energy changes: Effect of ten point mutations. *Proteins: Struct., Funct., Genet.*, 33:39–48

[203] Sheather, S. J. and M. C. Jones (1991). A reliable data-based bandwidth selection method for kernel density estimation. *J. Royal Stat. Soc. B*, 53:683–690

[204] Sheinerman, F. B., R. Norel, and B. Honig (2000). Electrostatic aspects of protein–protein interactions. *Curr. Opin. Struct. Biol.*, 10:153–159

[205] Shelley, J. C. and G. N. Patey (1994). A configuration bias Monte Carlo method for ionic solutions. *J. Chem. Phys.*, 100:8265–8270

[206] Shimada, J., A. V. Ishchenko, and E. I. Shakhnovich (2000). Analysis of knowledge-based protein–ligand potential using a self-consistent method. *Protein Sci.*, 9:765–775

[207] Siepmann, J. I. and D. Frenkel (1992). Configurational bias Monte Carlo: a new sampling scheme for flexible chains. *Mol. Phys.*, 75:59–70

[208] Simonson, T., G. Archontis, and M. Karplus (1999). A Poisson-Boltzmann study of charge insertion in an enzyme active site: The effect of dielectric relaxation. *J. Phys. Chem. B*, 103:6142–6156

[209] Sitkoff, D., K. A. Sharp, and B. Honig (1994). Accurate calculation of hydration free energies using macroscopic solvent models. *J. Phys. Chem.*, 98:1978–1988

[210] Smart, J. L., T. J. Marrone, and J. A. McCammon (1997). Conformational sampling with Poisson-Boltzmann forces and a stochastic dynamics/Monte Carlo method: Applications to alanine dipeptide. *J. Comput. Chem.*, 18:1750–1759

[211] Smit, B. (1995). Grand canonical Monte Carlo simulations of chain molecules: adsorption isotherms of alkanes in zeolites. *Mol. Phys.*, 85:153–172

[212] Smit, B. and J. I. Siepmann (1994). Computer simulations of the energetics and siting of *n*-alkanes in zeolites. *J. Phys. Chem. B*, 98:8442–8452

[213] Smith, P. E. (1999). The alanine dipeptide free energy surface in solution. *J. Chem. Phys.*, 111:5568–5579

[214] Smith, R., I. M. Brereton, R. Y. Chai, and S. B. H. Kent (1996). Ionization states of the catalytic residues in HIV-1 protease. *Nat. Struct. Biol.*, 3:946–950

[215] Spinelli, S., Q. Z. Liu, P. M. Alzari, P. H. Hirel, and R. J. Poljak (1991). The three-dimensional structure of the aspartyl protease from the HIV-1 isolate BRU. *Biochimie*, 73:1391–1396

[216] van der Spoel, D., A. R. van Buuren, D. P. Tieleman, and H. J. C. Berendsen (1996). Molecular dynamics simulations of peptides from BPTI: A closer look at amide–aromatic interactions. *J. Biomol. NMR*, 8:229

[217] Stach, H., H. Thamm, J. Jänchen, K. Fiedler, and W. Schirmer (1984). Experimental and theoretical investigations of the adsorption of n-paraffins, n-olefins and aromatics on silicalite. In D. Olsen and A. Bisio, editors, *New Developments in Zeolite Science and Technology, Proceedings of the 6th International Zeolite Conference*, 225–231. Butterworth, Guildford, UK

[218] Stayton, P. S., S. Freitag, L. A. Klumb, A. Chilkoti, V. Chu, J. E. Penzotti, R. To, D. Hyre, I. L. Trong, T. P. Lybrand, and R. E. Stenkamp (1999). Streptavidin–biotin binding energetics. *Biomol. Eng.*, 16:39–44

[219] Sussman, F., M. C. Villaverde, and A. Davis (1997). Solvation effects are responsible for the reduced inhibitor affinity of some HIV-1 mutants. *Protein Sci.*, 6:1024–1030

[220] Swain, A. L., M. M. Miller, J. Green, D. H. Rich, J. Schneider, S. B. H. Kent, and A. Wlodawer (1990). X-ray crystallographic structure of a complex between a synthetic protease of human immunodeficiency virus 1 and a substrate-based hydroxyethylamine inhibitor. *Proc. Natl. Acad. Sci. U. S. A.*, 87:8805–8809

[221] Tanford, C. and J. G. Kirkwood (1957). Theory of protein titration curves. I. general equations for impenetrable spheres. *J. Am. Chem. Soc.*, 79:5333–5339

[222] Tapia, O., J. Andrés, and V. S. Safont (1994). Enzyme catalysis and transition structures *in vacuo*. *Faraday Trans.*, 90:2365–2374

[223] Tawa, G. J., I. A. Topol, S. K. Burt, and J. W. Erickson (1998). Calculation of relative binding free energies of peptidic inhibitors to HIV-1 protease and its I84V mutant. *J. Am. Chem. Soc.*, 120:8856–8863

[224] Tobias, D. J. and C. L. Brooks III (1992). Conformational equilibrium in the alanine dipeptide in the gas phase and aqueous solution: A comparison of theoretical results. *J. Phys. Chem.*, 96:3864–3870

[225] Torrie, G. M. and J. P. Valleau (1977). Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *J. Comput. Phys.*, 23:187–199

[226] Triangle Pharmaceuticals Inc. (2000). URL `http://www.tripharm.com/dmp450.html`

[227] Trylska, J., J. Antosiewicz, M. Geller, C. N. Hodge, R. M. Klabe, M. S. Head, and M. K. Gilson (1999). Thermodynamic linkage between the binding of protons and inhibitors to HIV-1 protease. *Protein Sci.*, 8:180–195

[228] Uhlherr, A. (2000). Monte Carlo conformational sampling of the internal degrees of freedom of chain molecules. *Macromolecules*, 33:1351–1360

[229] Velazquez-Campoy, A., M. J. Todd, and E. Freire (2000). HIV-1 protease inhibitors: Enthalpic versus entropic optimization of the binding affinity. *Biochemistry*, 39:2201–2207

[230] Venables, W. N. and B. D. Ripley (1997). *Modern Applied Statistics with S-PLUS*. Springer-Verlag, New York, USA, 2nd edition

[231] Verkhivker, G., K. Appelt, S. T. Freer, and J. E. Villafranca (1995). Empirical free energy calculations of ligand–protein crystallographic complexes. I. knowledge-based ligand–protein interaction potentials applied to the prediction of human immunodeficiency virus 1 protease binding affinity. *Protein Eng.*, 8:677–691

[232] Verkhivker, G. M., D. Bouzida, D. K. Gehlhaar, P. A. Rejto, S. Arthurs, A. B. Colson, S. T. Freer, V. Larson, B. A. Luty, T. Marrone, and P. W. Rose (2000). Deciphering common failures in molecular docking of ligand–protein complexes. *J. Comput.-Aided Mol. Des.*, 14:731–751

[233] Vlugt, T. J. H. (1999). Efficiency of parallel CBMC simulations. *Mol. Simul.*, 23:63–78

[234] Vlugt, T. J. H., R. Krishna, and B. Smit (1999). Molecular simulations of adsorption isotherms for linear and branched alkanes and their mixtures in silicalite. *J. Phys. Chem. B*, 103:1102–1118

[235] Vlugt, T. J. H., M. G. Martin, B. Smit, J. I. Siepmann, and R. Krishna (1998). Improving the efficiency of the configurational-bias Monte Carlo algorithm. *Mol. Phys.*, 94:727–733

[236] Vondrasek, J., C. P. van Buskirk, and A. Wlodawer (1997). Database of three-dimensional structures of HIV proteinases. *Nat. Struct. Biol.*, 4:8. URL `http://www.ncifcrf.gov/HIVdb/`

[237] Vorobjev, Y. N. and J. Hermans (1999). ES/IS: estimation of conformational free energy by combining dynamics simulations with explicit solvent with an implicit solvent continuum model. *Biophys. Chem.*, 78:195–205

[238] Vriend, G. (1990). WHAT IF: a molecular modelling and drug design program. *J. Mol. Graphics*, 8:52–56

[239] Wade, R. C., M. E. Davis, B. A. Luty, J. D. Madura, and J. A. McCammon (1993). Gating of the active site of triose phosphate isomerase: Brownian dynamics simulations of flexible peptide loops in the enzyme. *Biophys. J.*, 64:9–15

[240] Wall, I. D., A. R. Leach, D. W. Salt, M. G. Ford, and J. W. Essex (1999). Binding constants of neuraminidase inhibitors: An investigation of the linear free interaction energy method. *J. Med. Chem.*, 42:5142–5152

[241] Wallace, A. C., R. A. Laskowski, and J. M. Thornton (1995). LIGPLOT: a program to generate schematic diagrams of protein–ligand interactions. *Protein Eng.*, 8:127–134

[242] Waller, C. L., T. I. Oprea, A. Giolitti, and G. R. Marshall (1993). Three-dimensional QSAR of human immunodeficiency virus (I) protease inhibitors. 1. A CoMFA study employing experimentally-determined alignment rules. *J. Med. Chem.*, 36:4152–4160

[243] Wallqvist, A., R. L. Jernigan, and D. G. Covell (1995). A preference-based free-energy parameterization of enzyme–inhibitor binding. Applications to HIV-1-protease inhibitor design. *Protein Sci.*, 4:1881–1903

[244] Wang, Y.-X., D. I. Freedberg, T. Yamazaki, P. T. Wingfield, S. J. Stahl, J. D. Kaufman, Y. Kiso, and D. A. Torchia (1996). Solution NMR evidence that the HIV-1 protease catalytic aspartyl groups have different ionization states in the complex formed with the asymmetric drug KNI-272. *Biochemistry*, 35:9945–9950

[245] Warshel, A. and A. Papazyan (1998). Electrostatic effects in macromolecules: fundamental concepts and practical modeling. *Curr. Opin. Struct. Biol.*, 8:211–217

[246] Weber, P. C., D. H. Ohlendorf, J. J. Wendoloski, and F. R. Salemme (1989). Structural origins of high-affinity biotin binding to streptavidin. *Science*, 243:85–88

[247] Weiner, P. K., R. Langridge, J. M. Blaney, R. Schaefer, and P. A. Kollman (1982). Electrostatic potential molecular surfaces. *Proc. Natl. Acad. Sci. U. S. A.*, 79:3754–3758

[248] Wlodawer, A. and A. Gustchina (2000). Structural and biochemical studies of retroviral proteases. *Biochim. Biophys. Acta*, 1477:16–34

[249] Wlodawer, A., M. Miller, M. Jaskólski, B. K. Sathyanarayana, E. Baldwin, I. T. Weber, L. M. Selk, L. Clawson, J. Schneider, and S. B. H. Kent (1989). Conserved folding in retroviral proteases: Crystal structure of a synthetic HIV-1 protease. *Science*, 245:616–621

[250] Wlodawer, A. and J. Vondrasek (1998). Inhibitors of HIV-1 protease: A major success of structure-assisted drug design. *Annu. Rev. Biophys. Biomol. Struct.*, 27:249–284

[251] Wong, C. F. and J. A. McCammon (1986). Dynamics and design of enzymes and inhibitors. *J. Am. Chem. Soc.*, 108:3830–3832

[252] Wu, M. G. and M. W. Deem (1999). Analytical rebridging Monte Carlo: Applications to *cis/trans* isomerization in proline-containing, cyclic peptides. *J. Chem. Phys.*, 111:6625–6632

[253] Wu, M. G. and M. W. Deem (1999). Efficient Monte Carlo methods for cyclic peptides. *Mol. Phys.*, 97:559–580

[254] Xie, D., S. Gulnik, L. Collins, E. Gustchina, T. N. Bhat, and J. W. Erickson (1998). Thermodynamics and proton uptake for pepstatin binding to retroviral and eukaryotic aspartic proteases. *Adv. Exp. Med. Biol.*, 436:381–386

[255] Xie, D., S. Gulnik, L. Collins, E. Gustchina, L. Suvorov, and J. W. Erickson (1997). Dissection of the pH dependence of inhibitor binding energetics for an aspartic protease: Direct measurement of the protonation states of the catalytic aspartic acid residues. *Biochemistry*, 36:16166–16172

[256] Xu, D., C.-J. Tsai, and R. Nussinov (1997). Hydrogen bonds and salt bridges across protein-protein interfaces. *Protein Eng.*, 10:999–1012

[257] Yamazaki, T., L. K. Nicholson, D. A. Torchia, P. Wingfield, S. J. Stahl, J. D. Kaufman, C. J. Eyermann, C. N. Hodge, P. Y. S. Lam, Y. Ru, P. K. Jadhav, C.-H. Chang, and P. C. Weber (1994). NMR and X-ray evidence that the HIV protease catalytic aspartyl groups are protonated in the complex formed by the protease and a non-peptide cyclic urea-based inhibitor. *J. Am. Chem. Soc.*, 116:10791–10792

[258] Yang, A.-S., M. R. Gunner, R. Sampogna, K. Sharp, and B. Honig (1993). On the calculation of pK_as in proteins. *Proteins: Struct., Funct., Genet.*, 15:252–265

[259] Ysern, X., H. Li, and R. A. Mariuzza (1998). Imperfect interfaces. *Nat. Struct. Biol.*, 5:412–414

Summary

Structure-based drug design is made possible by our understanding of molecular recognition. The utility of this approach was apparent in the development of the clinically effective HIV-1 PR inhibitors, where crystal structures of complexes of HIV-1 protease and inhibitors gave pivotal information. Computational methods drawing upon structural data are of increasing relevance to the drug design process. Nonetheless, these methods are quite rudimentary and significant improvements are needed. The aim of this thesis was to investigate techniques which may lead to improved modelling of molecular recognition and a better ability to make predictions about the binding affinity of ligands. The two main themes were the modelling of acid–base titration behaviour of ligand and receptor, and the application of the simulation technique of configurational bias Monte Carlo (CBMC). The studies were performed with HIV-1 PR and its inhibitors as a model system.

Biological processes are influenced by the pH of the medium in which they take place. Ligand–receptor binding equilibria are often thermodynamically linked to protonation changes in ligand and/or receptor, as seen in the the binding of a number of HIV-1 PR inhibitors. In Chapter 2, a series of sixteen continuum electrostatics pK_a calculations of HIV-1 PR–inhibitor complexes was done, in order to characterize the nature and size of these linkages. The most important effects concern changes in the pK_a of the enzyme active site aspartate dyad. Large pK_a shifts were predicted in all cases, and at least one of the two dyad pK_as became more basic on binding. At physiologically relevant pH, different ligands induced different protonation states, with different tautomeric forms favoured. The fully deprotonated form of the dyad was not significantly populated for any of the complexes. For about a third of the complexes, both singly and doubly protonated forms were predicted to be populated. The predicted predominant protonation states of MVT-101 and VX-478 were consistent with previous theoretical studies. The size of the predicted pK_a shifts for MVT-101 and XK263 differed from a previous study using similar methods. The paucity and ambiguity of available experimental data makes it difficult to evaluate the results fully; however the tendency to exaggerate shifts, as observed in other studies, appears to be present.

"Scoring" is the prediction of binding affinity from the structure of the ligand–receptor complex, according to an empirical scheme. Scoring studies usually neglect or grossly simplify the contribution of protonation equilibria to affinity, so in Chapter 3 proton linkage data was included in a regression analysis of the HIV-1 PR complexes from Chapter 2. Parameters previously shown to correlate with binding, namely electrostatic free energy changes and buried surface areas, were the basis for the analysis, and terms describing proton linkage, in the form of a correction for assay pH and an indicator variable for predicted dyad pK_a shift on binding, were also considered. The complex with MVT-101 was an outlier in the analysis and was excluded. Further analysis demonstrated that the correction for assay pH

made a significant contribution to the regression equation. Amendment of the parameters for XK263 according to the available experimental data led to an improved regression in which the term for calculated pK_a shifts also made a significant contribution. The regression equations obtained had the same form and similar coefficients to scoring functions of the "master equation" type, and fit the experimental data with comparable accuracy.

More physically realistic simulations of ligand–receptor binding using the techniques of molecular dynamics (MD) or Monte Carlo (MC) are potentially more accurate than scoring function approaches. These methods are slow, so the alternative of CBMC, which has been shown to give faster convergence for polymer simulations, was implemented for CHARMM22, an all-atom protein force field (Chapter 4). The correctness of the implementation was demonstrated by comparison with exact and stochastic dynamics (SD) results for individual terms in the force field. The algorithm is more complex than those typically used with alkane force fields, and this has possible consequences for the efficiency. CBMC was used to generate a Ramachandran plot for the alanine dipeptide, and the results were found to be in agreement with those generated by a SD simulation. Analysis of statistical errors suggests that CBMC should be competitive with umbrella sampling for simulating conformational equilibria, particularly when the cost of non-bonded energy evaluations dominates the simulation.

CBMC can be applied to ligand–receptor binding, as demonstrated in grand canonical simulations of alkane adsorption in zeolites. The more limited problem of finding the predominant bound conformation of a flexible ligand given a rigid protein receptor (i.e. "docking") was treated in Chapter 5, using the example of a tripeptide inhibitor which binds to HIV-1 PR. Attempts to perform the docking using the Metropolis MC/simulated annealing and Lamarckian genetic algorithm methods implemented in the program AutoDock failed to reproduce the native configuration (with runs on the order of two days execution time). Docking using CBMC, combined with parallel tempering to further improve sampling, was successful in finding the native binding mode, although this success was dependent on *ad hoc* adjustments to the force field, and *a priori* knowledge of the ligand protonation state and binding site. The efficiency of the method was considerably lower than hoped, with problems due to the force field- and model-dependent coupling between terms in the potential energy function, and the "greedy" nature of the CBMC algorithm.

Various conclusions can be drawn from these studies. Chapters 2 and 3 provide evidence of the importance of protonation equilibria in ligand–protein molecular recognition, and underline the sizable contribution of electrostatic interactions to binding energies. In the face of this finding, neglect of electrostatic terms, as often seen past studies, appears to be counterproductive. The scoring study also shows how experimental data can be used more effectively if factors such as assay conditions are carefully taken into account. Implementation of CBMC for a widely-used protein force field and application of the algorithm to docking (Chapters 4 and 5) represents a proof of concept for a broadly useful simulation technique. Further work will be required to find the right niche for CBMC and fully explore the potential of this and related techniques. A final point is the demonstrated utility of the HIV-1 PR test system which formed the focus of the studies. Abundant structural data has enabled many new approaches to be tested, and further insights are expected from the analysis of unusual cases, such as the anomalous results for MVT-101. As well as the question of scoring, studies of mutation and resistance are likely to attract considerable interest in the future.

Samenvatting

Structure-based drug design wordt mogelijk gemaakt door inzicht in moleculaire herkenning. Het nut van deze aanpak werd aangetoond tijdens de ontwikkeling van de HIV-1 protease remmers, door de belangrijke inbreng van HIV-1 PR–remmer kristal structuren. Op structuur data gebaseerde computer methodes zijn van toenemend belang voor de geneesmiddelontwikkeling. Desalniettemin zijn deze technieken tamelijk grof en verdere verbeteringen zijn nodig. Het doel van dit proefschrift was om technieken te onderzoeken die tot verbeterde nabootsing van moleculaire herkenning kunnen leiden en daardoor tot nauwkeuriger voorspellingen van de bindingsaffiniteit van liganden voor eiwitten. De twee hoofdthema's waren de modeling van zuur–base titratie gedrag van ligand en receptor, en de toepassing van de simulatietechniek *configurational bias Monte Carlo* (CBMC). Het werk werd uitgevoerd met gebruik van HIV-1 PR en remmers als model.

Biologische processen worden beïnvloed door de pH van de omgeving waarin zij plaats vinden. Ligand–receptor bindingsevenwichten zijn vaak gekoppeld aan veranderingen in protonatie toestand van ligand en/of receptor, b.v. in het geval van het bindingsgedrag van bepaalde HIV-1 PR remmers. In Hoofdstuk 2 werd een reeks *continuum electrostatics* pK_a berekeningen van HIV-1 PR–remmer complexen uitgevoerd, om de aard en mate van deze koppelingen te bepalen. De allerbelangrijkste verschijnselen hebben betrekking tot veranderingen van de pK_a's van de aspartaat *dyad* in de *active site*. Grote pK_a verschuivingen werden voorspeld in alle gevallen, en minstens een van de twee pK_a's werd basischer na binding. Bij fysiologische pH geven verschillende liganden aanleiding tot verschillende protonatie toestanden, met verschillende tautomerisatie. Een volledige gedeprotoneerde *dyad* werd alleen met een uiterst geringe waarschijnlijkheid terug gevonden in de resultaten. In ongeveer een derde van de gevallen werden zowel enkel als dubbel geprotoneerde toestanden gevonden. De meest voorkomende protonatie toestand van MVT-101 en VX-478 kwam overeen met eerder gepubliceerde resultaten. De mate van de pK_a verschuivingen van MVT-101 en XK263 was niet gelijk aan de resultaten van een ander onderzoek dat overeenkomstige methodes gebruikte. De schaarste en onduidelijkheid van beschikbare experimentele gegevens maken het moeilijk om de voorspellingen volledig te controleren. Er is wel een trend naar overdrijving van de verschuivingen, zoals in andere onderzoeken wordt beschreven.

"Scoring" betekent het voorspellen van bindingsaffiniteit door middel van een empirische evaluatie van de structurele aard van het ligand–receptor complex. Scoring benaderingen negeren meestal de bijdrage van het protonatieevenwicht aan de affiniteit, en daarom werden in Hoofdstuk 3 protonatiekoppelingsdata toegevoegd aan een regressie analyse van de HIV-1 PR complexen uit Hoofdstuk 2. Parameters waarvan eerder een correlatie met bindingsaffiniteit werd aangetoond, b.v. electrostatische vrije energie veranderingen en veranderingen in moleculaire oppervlakte, vormden de grondslag van de analyse, en termen die met protona-

tiekoppeling te maken hebben, met name een correctie op basis van assay pH en een indicator variabele voor *dyad* pK_a verschuiving, werden ook meegenomen. Het complex met MVT-101 was een uitbijter en werd buiten beschouwing gelaten. Nadere statistische analyse liet zien dat de assay pH correctie aantoonbaar aan de regressie bijdroeg. De regressievergelijkingen die gevonden werden hadden dezelfde vorm en vergelijkbare coëfficiënten in relatie tot scoring functies van het "master equation" type, en kwamen even goed met de experimentele data overeen.

Een meer gedetailleerde nabootsing van de werkelijkheid door middel van moleculaire dynamica (MD) of Monte Carlo (MC) simulaties kan nauwkeuriger schattingen van bindingsconstanten opleveren dan scoring functie benaderingen. Deze methodes zijn helaas zeer langzaam, en daarom werd CBMC, dat een snelle convergentie laat zien bij polymeer simulaties, als alternatieve *sampling* procedure geïmplementeerd voor het CHARMM22 *all-atom* eiwit krachtveld (Hoofdstuk 4). De betrouwbaarheid van de implementatie werd bewezen door resultaten te vergelijken met zowel exacte als stochastische dynamica (SD) resultaten met betrekking tot afzonderlijke termen uit het krachtveld. Dit krachtveld vereist een algoritme met meer complexiteit in vergelijking met typische alkaan krachtveld toepassingen, wat mogelijk gevolgen heeft met betrekking tot de efficiëntie. CBMC werd gebruikt om een Ramachandran grafiek te maken voor alanine dipeptide, en de resultaten kwamen goed overeen met SD resultaten. Analyse van de statistische fouten geeft aan dat CBMC een waardige concurrent zou zijn voor *umbrella sampling*, in het bijzonder in het geval dat de kosten van *non-bonded* energie evaluaties overheersen.

CBMC kan op het ligand–receptor binding probleem worden toegepast, zoals in *grand canonical* simulaties van alkaan adsorptie aan zeolieten werd gezien. Makkelijker is het vinden van de meest voorkomende gebonden conformatie van een flexibele ligand, gegeven een rigide receptor (d.w.z. "docking"), een probleem dat in Hoofdstuk 5 werd behandeld met het voorbeeld van een tripeptide HIV-1 PR remmer. Docking pogingen met de Metropolis MC/*simulated annealing* en *Lamarckian genetic algorithm* methodes van het programma AUTODOCK slaagden er niet in om de juiste configuratie terug te vinden. CBMC docking, in combinatie met *parallel tempering* om de *sampling* verder te verbeteren, kwam wel op een acceptabel antwoord, hetgeen van voorkennis over de protonatie toestand en bindingsplek en ook speciale aanpassingen van het krachtveld afhing. De efficiëntie van de methode was merkbaar minder dan verwacht, met problemen mede veroorzaakt door de krachtveld- en model-afhankelijke koppeling tussen termen in de *potential energy function*, en de "greedy" aanpak van het CBMC algoritme.

Verschillende conclusies kunnen worden getrokken uit dit werk. Hoofdstukken 2 en 3 wijzen op het belang van protonatieevenwichten in ligand–eiwit moleculaire herkenning, en onderschrijven de behoorlijke bijdrage van electrostatische interacties aan bindingsenergieën. Het scoring onderzoek laat ook zien hoe experimentele gegevens beter kunnen worden benut wanneer zorgvuldig rekening wordt gehouden met factoren zoals assay omstandigheden. Implementatie van CBMC voor een veelgebruikt eiwit krachtveld en de toepassing hiervan op docking (Hoofdstukken 4 en 5) laten zien dat CBMC een breed toepasbare simulatietechniek is. Uit nader onderzoek moet blijken welke niche het beste is voor CBMC en wat voor efficiënties haalbaar zijn. Een laatste punt is het aantoonbare nut van het HIV-1 PR test systeem dat hier gebruikt werd. Een schat aan structurele data laat toe om vele nieuwe technieken te testen, en nog meer inzichten zijn te verwachten uit analyse van afwijkende

gevallen, b.v. de uitbijter MVT-101. Naast de scoring problematiek zullen in de toekomst studies met betrekking tot mutatie en resistentie van groot belang zijn.

Curriculum vitæ

The author of this thesis was born on June 10, 1971 in Copenhagen, Denmark. After graduating from high school in 1988, he studied for the degree Bachelor of Science (Honours) at the University of Sydney (Sydney, Australia), majoring in Computer Science, Chemistry, and Pharmacology, with Honours in Pharmacology. The subject of his Honours research project was a computational and behavioural study of cholinergic pharmacology and memory. The degree was conferred with First Class Honours and University Medal in May 1995. He worked for two years as computer programmer and system administrator at connect.com.au Pty Ltd in Sydney, and then moved to the Department of Medicinal Chemistry, Utrecht Institute for Pharmaceutical Sciences, Faculty of Pharmacy, Utrecht University (Utrecht, The Netherlands) in November 1996 to begin a Ph.D. with Dr E. E. Moret and Prof. J. P. A. E. Tollenaere.

Dankwoord

Ik was nooit naar Utrecht gekomen, en was zeker niet gebleven, zonder de inspiratie en eindeloze steun van Ed Moret. Al met de eerste email-wisseling Utrecht–Sydney had ik een beeld van een warme, geestig, geëngageerd mens, en die intuïtie bleek (gelukkig maar) juist te zijn. Ed, je zette alles in om mij te helpen, zowel met de onvermijdelijke, vervelende situaties die samenhangen met vreemdeling zijn, als met het sturen van het onstuurbare, dat wil zeggen het begeleiden van mijn onderzoek. Ik ben je oneindig dankbaar. Ik ambieer om in de toekomst toegevoegd te worden aan de lijst van vreemde figuren die onregelmatig bij jou, Inge, Alex en Bas in Maartensdijk opduiken.

Jan Tollenaere heeft als promotor mij alle vertrouwen gegund, al van het begin toen ik als volstrekt onbekende in Utrecht verscheen. Hij heeft mijn inconsistent werkwijze zonder opmerking geduld, en maakte zich altijd beschikbaar voor consultatie en discussie. Ik wil hem bedanken voor zijn goedaardige toezicht. Nico de Mol heeft mij de kans gegeven om tussendoor wat leuke onderzoek klusjes te doen, dingen die in tegenstelling tot "het grote werk" na een tijd echt af waren. Rob Liskamp heeft het nodige aanzien als leider van de vakgroep en ik heb me dus altijd op een veilige afstand gehouden. Maar ondanks het feit dat mijn onderzoek buiten de mainstream van de groep is gevallen, heb ik altijd het gevoel gehad om een gewaardeerd lid van de club te zijn. Bert Janssen heeft zijn naam onder de eerste uitnodigingsbrief gezet, en is sindsdien een waardevolle onafhankelijke bron voor adviezen geweest.

Ik voelde me al gauw op m'n gemak bij de CMC groep. Hans Hilbers heeft mijn pogingen om het systeembeheer overhoop te gooien geduldig op de werkelijkheid afgestemd. Net als iedereen heeft hij me gesteund in mijn speurtocht om Nederlands te leren, en jarenlang mijn verhaspelde uitspraken moeten verwerken. Met de andere CMC AIOs waren er natuurlijk een belangrijk solidariteitsgevoel en vanzelfsprekend een hoop interessante discussies; voor dit wil ik Alfred, Gertjan en Ellen bedanken. David als ere-CMC'er heeft ook de laatste tijd Z716 een beetje een relaxede plek kunnen maken. Daarnaast hebben studenten en bezoekers voor wat extra kleur gezorgd. Andy Vinter was there at the beginning to get me off the ground. Things may not have turned out as planned, but that first project taught me an awful lot. Patrick was ook iemand die een hoop energie aan de groep toevoegde. Een extra Vlaamse blik op Nederland was voor mij ook een waardevolle aanwinst. Verder wil ik iedereen uit de synthese en farmacognosie hoeken bedanken voor de gezelligheid op werkbesprekingen, dagjes uit, borrels, kerstdiners en andere gemeenschappelijke aangelegenheden.

De molecular simulation cursus op de UvA heeft mijn onderzoek een nieuwe impuls gegeven. Berend Smit, Daan Frenkel en Thijs Vlugt zorgden voor een uiterst inspirerend stuk onderwijs, die uiteindelijk een aardig project voort heeft gebracht. Berend, ik was erg onder de indruk van jouw onbevooroordeelde doch nuchtere aanpak. Thijs, jouw geduldige

antwoorden op mijn vele vragen hebben mij op het goede pad gestuurd. Het is bijna niet te geloven wat je in je onderzoek heb gepresteerd, en eerlijk gezegd heb ik alleen een heel klein stukje eraan gebreid. Alexandre Bonvin hoort ook bij de mensen die mijn promotie hebben gered. Zijn enthousiasme en hartelijkheid gaven mij een steun in de rug, en de mogelijkheid om gebruik te maken van zijn cluster heeft essentiëel bijgedragen aan het verkrijgen van de laatste onderzoeksresultaten.

De lunchtafel droeg bij aan mijn oriëntatie in de Nederlandse samenleving. Marcel heeft mij normen en waarden geleerd maar tegelijkertijd dat ik alles met een grote korrel zout moet nemen. Samen met de CMC lui, Nico, Bert, Anita, en de laatste tijd ook af en toe Frank en Lidija hebben we uiteenlopende wereldproblemen kunnen oplossen, of tenminste krachtig becommentariëren (eerlijk gezegd ben ik vaker toeschouwer dan participant geweest, hetzij wegens een gebrek aan verbale lichtvoetigheid, dan wel vanwege uitlopend ochtendhumeur).

Voor zo'n grote faculteit is het een ware luxe geweest dat de contacten buiten medicinal chemistry toch erg informeel lopen. Ik klopte aan bij PZ (vooral Karlien Agasi is het slachtoffer geweest), DGV (Martijn Pieck), financiën, Linda Hutzezon, en de automatiseringsdienst (Hans Reijn, Leo van der Hark en Ruud van Kooten) zonder een waarschuwing vooraf, en werd bijna altijd meteen te woord gestaan. Bedankt iedereen voor je hulp en verdraagzaamheid. Mijn aanpak bij het Med Chem secretariaat was ook vrij onbeschoft; bedankt Natatia en Lidija dat jullie mij toch altijd vriendelijk hebben geholpen. Kees Beukelman heeft mij ook bijgestaan wat financiële vragen betreft—bij elkaar heb ik dus weinig last gehad van mijn eigen gebrekkige organisatietalent.

Een aantal anderen hebben het leven in Utrecht een stuk aangenamer gemaakt. Olga Janssen heeft voor Rachel en mij twee prachtige woningen gevonden en heeft heel creatief oplossingen kunnen vinden onder moeilijke omstandigheden. Rik en Juul waren onze eerste huisbazen en hebben in eerste instantie als een soort pleeggezin gefungeerd. Martha Kits is jarenlang mijn Nederlands lerares geweest, zonder wie mijn zinsbouw nog alarmerender aspecten zou hebben gehad.

I'm particularly grateful to all the people who I emailed out of the blue with questions and requests, who without exception gave me help without hesitation. Joanna Trylska and Jan Antosiewicz went to great lengths to help me with my pK_a calculations, without which I would have been a couple of chapters short of a thesis; Jim Briggs also gave me essential assistance. Chris Murray and Maria Miller helped me find data I needed on crystal structures, and Jay Ponder was extremely responsive to questions and bug reports about TINKER. TINKER is the most clearly written bunch of FORTRAN77 programs I've seen, and working with it made my life that much easier. Gert Vriend answered WHAT IF questions almost before they were asked (it's in the new version), and Stefano Piana provided me with a fascinating manuscript and very useful discussion on HIV protease ionization states. It's heartening to meet find much good will, and to see that the spirit of academic inquiry is much healthier that I had imagined.

I wouldn't have survived six months in Utrecht without the friends I've made. Luckily enough there's been no shortage of fascinating people to meet, most of whom will hopefully forgive me for being so antisocial the last year or so. From the crowd, I want to single out (in rough order of appearance) Stella, Mary, Kumiko, Hans-Jürgen, Wil, Erica, Dan, Inge, Ollie, Sylvia, Ari and Pirkko. I'm deeply in Tony's debt for a series of excellent Indian dinners (not to mention the free beers). Steve gets a special mention for proving that you can be a crass American, a scholar, and a gentleman all at once. Yves I want to thank for solidarity in the

struggle to maintain decent culinary standards. And if he (and I) hadn't met Nora, my life would have been infinitely duller the last couple of years. Gerrit and Lucia were the most hospitable, open-hearted neighbours possible. Georg was there from the start, and proved to me that promoveren is possible. And Gabriel showed me that big dinners and unstoppable curiosity are the secrets to getting through life.

Fortunately for me, it seems like most of my best friends from way back have escaped from Australia and set up shop in more convenient locations, which has made me feel a bit less like I'm completely afloat in the universe. Keir, it's been good comparing notes with you on the Europe thing, as well as just sitting back for a few beers those few times I wasn't being a stress bunny. Mani, you're an inspiration and a far more thoughtful friend than I deserve. Jeremy, swapping email with you has kept my mind ticking over, and your help the last few months has got me onto a brand new track. Did I ever mention that I worked out you're my evil twin? Irina, you kept on writing even when I was being a slack bastard, and helped me keep things in perspective.

Jill, Gunther, Michael and Emily have put up with my moodiness and my uncanny ability to catch the flu at Christmas time, and let me freeload in London and Marré. Summer evenings under the walnut tree are the pinnacle of civilised existence. Jonathan, you're my shining example of how to do the (academic) hard yards and be a top bloke too. I wish you lived a bit closer than the other end of the world. Mum and Dad, thanks for everything. Knowing you were there and only just managing to hold back from offering help and advice at every turn is what made this possible. It's been four very full years, and I've learned more than seems possible. Rachel, coming to the Netherlands with you was the biggest step I've taken yet, and I don't regret it one bit. Thank you for being there for the trip, and for sticking with it.