Multiscale Docking Using Evolutionary Optimisation

David John Huggins

A thesis submitted to the Faculty of Physical Sciences for the degree of Doctor of Philosophy in the University of Oxford

Brasenose College

Trinity Term 2005

Multiscale Docking Using Evolutionary Optimisation

Abstract of a thesis submitted for the degree of Doctor of Philosophy in the University of Oxford

David John Huggins, Brasenose College, Trinity Term 2005

Molecular docking algorithms are computational methods that predict the binding site and docking pose of specified ligands with a protein target. They have proliferated in recent years, due to the explosion of structural data in biology. Oxdock is an algorithm that uses various techniques to simplify this complex task, the most significant being the use of a multiscale approach to analyse the problem using a simple representation in the early stages. Oxdock is shown to be a very useful tool in computational biology, as exemplified by two cases. The first case is the analysis of the NMDA subclass of neuronal glutamate receptors and the subsequent elucidation of their function. The second is the investigation of the newly discovered plant glutamate receptors and the clarification of their natural ligands. The results in both instances open new areas of research into exciting areas of biology.

Despite its effectiveness in solving many problems, Oxdock does fail in a number of circumstances. It is thus important to devise a new and improved method for molecular docking. This is achieved by combining the speed of the multiscale approach with the optimising ability of Evolutionary Programming. This yields an algorithm that is shown to be precise, accurate and specific.

The new algorithm, Eve, is then modified to illustrate its potential in both lead optimisation and *de novo* drug design. These capacities, combined with its ability to predict the location of binding sites and the docking pose of a ligand, highlight the promise of computational methods in solving problems in many areas of biological chemistry.

Acknowledgements

I would like to thank all the people that have helped me to complete this thesis, whether at work or outside it.

Guy and Graham have been excellent supervisors for the three years that I have been working on my DPhil and have provided advice and encouragement whenever it was needed. I must also mention the contribution of Malcolm and Christian, who really helped me to get started by providing me with an intriguing problem and proved instrumental in helping to solve it.

Chris, Mike, Aixia, Meilan, Ben, Mark, Alex, Stella and Biggi have been superb friends to have in the lab and outside it. They have been an important part of my DPhil and I will really miss being able to work in such a nice atmosphere. Extra thanks must go to Ben Webb for helping me so much with the system admin job, which I foolishly undertook like so many before me. I learnt an enormous amount and it was mostly due to Ben.

During these last three years, I have had the pleasure of living with Jason, Andy, Fraser, Mark, Jess and Branwen and Biggi (who deserves two lots of thanks). All of them made my house a nice a place to be and made my graduate life fun. I have also had the help and support of my dear Katie, who kept me happy and kept me going.

David Huggins - 2nd June 2005

Abbreviations

ADMET	– Absorption, Distribution, Metabolism, Excretion and Toxicity
AMPA	– Alpha-Amino-3-Hydroxy-5-Methyl-4-Isoxazole Propionic Acid
ATD	– Amino Terminal Domain
AtGLR	– *Arabidopsis Thaliana* Glutamate Receptor.
CVFF	– Consistent Valence Force Field
DNQX	– 6, 7-Dinitroquinoxaline-2, 3-Dione
EP	– Evolutionary Programming
iGluR	– Ionotropic Glutamate Receptor
HIV	– Human Immuno-Deficiency Virus
LIVBP	– Leucine/Isoleucine/Valine-binding Protein
MCSS	– Multiple Copy Simultaneous Search
mGluR	– Metabotropic Glutamate Receptor
NMDA	– N-methyl D-aspartate
RMSD	– Root Mean Square Deviation

Contents

1 Introduction

Recent predictions estimate that the human genome contains between 25,000 and 40,000 genes. Alternative splicing of these genes means there are likely to be over 100,000 different proteins in the human body. Scientists expend enormous efforts each year in elucidating the structure and function of these proteins. One of the most useful methods for investigating protein function is structural analysis. Since the 1970s, the three-dimensional structures of many proteins have been solved by X-Ray crystallography or, to a lesser extent, nuclear magnetic resonance spectroscopy. As a result, the Protein Data Bank (PDB) [1] now contains over 30,000 entries from organisms all the way between Humans and the Human Immuno-Deficiency Virus (HIV). This increase is shown in Figure 1.1.

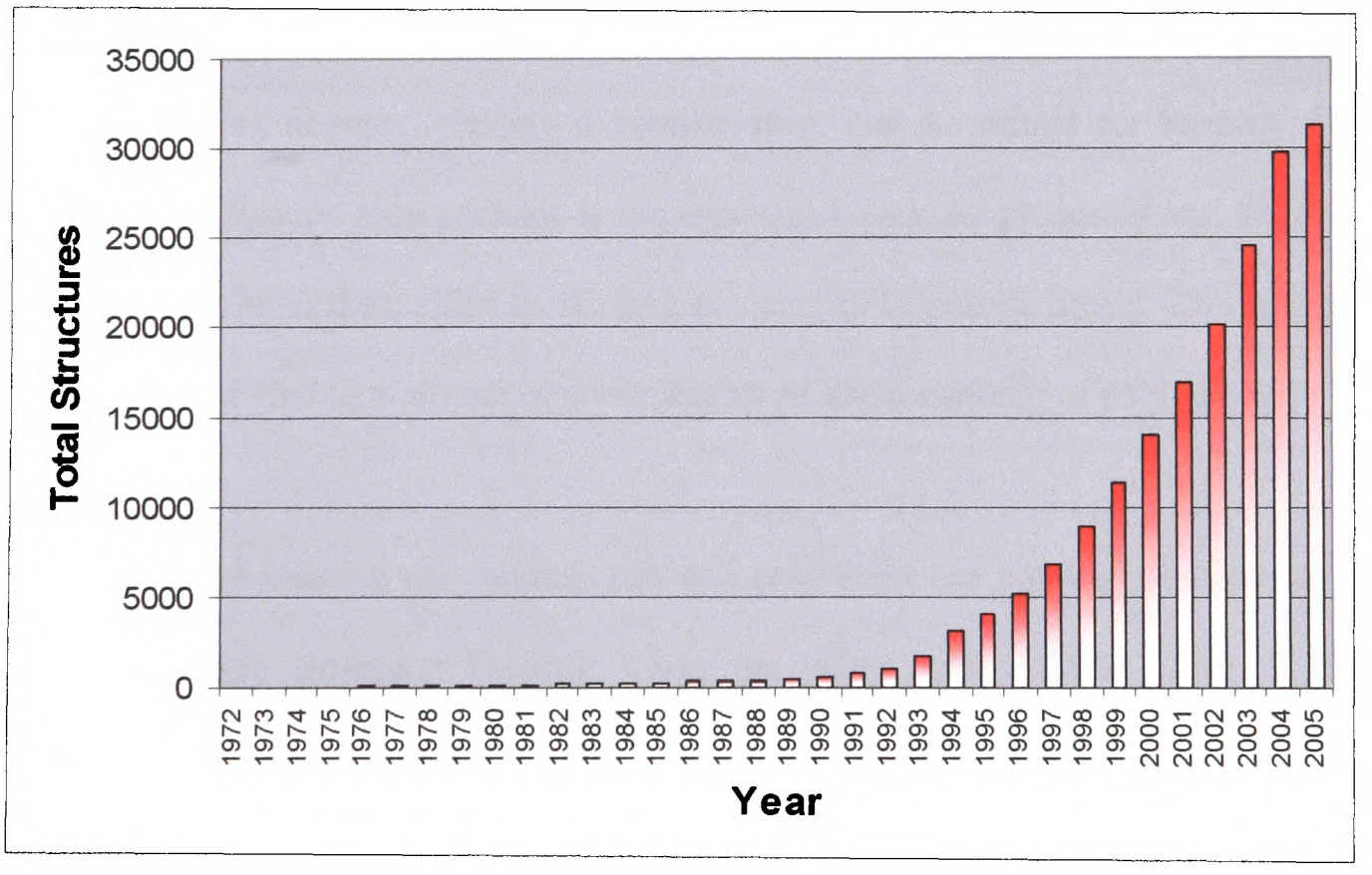


Figure 1.1 - The total number of structures in the PDB between 1972 and May 2005.

In recent years, billions of pounds have been invested in building large synchrotron facilities, which can produce monochromatic radiation for this purpose. The number of available structures appears to be increasing at an exponential rate. If this trend continues, simple calculations suggest that there will be over 2,000,000 structures by the year 2020. Currently, over 60% of the known proteins in the human body have unknown function and thus no identified ligand. It is therefore important that tools are available to investigate protein structures and predict their function.

Many of the thousands of processes that occur in cells involve a small molecule binding to a protein. A lock and key mechanism was suggested by Fischer in 1894 as an analogy for the binding of a ligand to a protein [2]. The notion is that a ligand will bind to a protein only if their three-dimensional structures are complementary. In more recent times, this concept has been replaced by the induced fit model [3]. The complementarity between ligand and protein is still crucial, but both structures will adapt to varying degrees, yielding a conformation that is optimal for binding. The interaction of ligands with proteins is important in metabolic processes and for drug molecules that elicit their effect by binding in place of the natural ligand. Recognizing the ligands that bind to a protein often facilitates an understanding of its function.

The process of locating the binding site and predicting the binding geometry of a ligand is termed molecular docking. There are many computational methods that simulate molecular docking; however, all suffer from the same problem. When attempting to find the correct arrangement of atoms in a protein-ligand complex, there is an infinite number of possibilities. The position, geometry and conformation of a ligand are all continuous variables. This means that in order to explore all possible

docking modes, millions or even billions of calculations must be performed. This has been termed "The Docking Problem". Devising good solutions to this problem creates powerful tools in the process of protein function elucidation

In this thesis, we illustrate the use of molecular docking in protein analysis by considering ionotropic glutamate receptors (iGluRs). Glutamate is the major excitatory neurotransmitter in the central nervous system and as such, glutamate receptors play a vital role in the mediation of synaptic transmission. However, despite their importance in learning and memory as well as a proved link with neuronal diseases such as Alzheimer's, the structure and function of the NMDA subclass of iGluRs is not well understood.

The function of iGluRs has proved even more puzzling in the case of the newly discovered plant glutamate receptors. Phylogenetic analysis of plant and animal sequences shows that these receptors are descended from a common source [4]. This strongly suggests that glutamate receptors evolved at a stage before plants and animals diverged. Furthermore, it suggests that the receptors will have a similar structure and a similar function in both plants and animals. However, questions remain as to how these receptors are both activated and modulated. Molecular docking can enrich biological chemistry by answering these types of questions and is able to make useful and testable predictions.

Molecular docking also has considerable importance within the pharmaceutical industry. Developing new molecules for medical purposes is a multi-million pound industry and *in-silico* screening of drug candidates saves enormous amounts of both

time and money. This process requires a rapid method for calculating both the binding geometry and an estimate of the binding energy for many possible compounds. It is clear that molecular docking is a very important technique and that the marked increase in available protein structures in the coming years will boost the need for improved methods.

The first aim of the work undertaken here is to illustrate the effectiveness of molecular docking by specific example to the problem of glutamate receptor function in both plants and animals. The second aim is to explore possibilities for overcoming the docking problem and create a quick and accurate method for locating the binding site and predicting the binding geometry of a protein-ligand complex. Finally, the numerous uses of this technique will be considered, highlighting its potential in many areas of biological chemistry.

2 Molecular Docking with Oxdock

2.1 Overview

Molecular docking is the process of using a computer to predict the binding geometry of a molecule in the active site of a protein. It is a complex task, which can take an enormous amount of computer time. Oxdock is a molecular docking algorithm, which predicts the binding site for a chosen ligand and protein. It is particularly useful because it uses a multiscale approach to increase the speed of the process. The ligand is represented initially as only one and then as an increasing number of feature-points. Each feature-point represents a cluster of atoms. The multi-scale docking algorithm runs a series of iterations to explore the three-dimensional space of a protein molecule for the most likely binding sites of the ligand. In early iterations, many possible binding sites are found, as all surfaces of the protein are available for ligand binding. As the algorithm proceeds, preferred binding sites are identified. When the algorithm converges, there are only very few overlapping binding positions for the ligand. In this chapter, the operation of Oxdock is explained and an example is used to illustrate its function.

2.2 Molecular Docking

Molecular docking involves the use of a computer to locate a protein binding site and predicting the binding geometry of a ligand. Numerous computational methods have been applied to this problem. The most commonly used are molecular dynamics

(MD), fragment-based approaches, Monte Carlo (MC) simulation and genetic algorithms (GAs).

MD involves simulating the system using Newton's laws of motion. These methods are very accurate at finding energy minima, but are generally poor at crossing large energy barriers. They are thus often used in conjunction with other search methods. Algorithms such as Amber [5] and Charmm [6] use MD. Fragment-based methods involve breaking the ligand into molecular fragments and then rebuilding the ligand in a stepwise manner within the binding site. They are typified by the algorithms FlexX [7] and DOCK [8]. These algorithms are very useful, due to the initial stages creating anchors in the binding site from which the remainder of the ligand can grow. MC functions by randomly selecting a starting position and making random changes in an attempt to find an improved solution. If the acceptance criteria for a new solution are met, the algorithm notes the new position and restarts the cycle. In simulated annealing MC, the "temperature" of the system is lowered during the course of the run by making the acceptance criteria stricter. This facilitates convergence. MC is a commonly used tool in molecular docking. Both Prodock [9] and Glide [10] perform molecular docking using MC. GAs also function using random changes, but the solutions evolve using the genetic operators of mutation, crossover and reproduction. GAs begin with a population of solutions to the problem and these solutions compete, based on their suitability, to reproduce in future generations. GAs are very powerful due to their expertise in searching highly complex spaces. GOLD [11] and Autodock [12] are the two most widely-used GA methods.

Each of these approaches has advantages and disadvantages and thus some of the most successful docking algorithms use a combination of these techniques. Comprehensive reviews of docking methods have been written by Taylor [13] and Brooijmans [14]. Despite the proliferation of molecular docking algorithms, none work consistently well and the best technique tends to vary depending on the problem. One of the most difficult challenges is the location of the binding site on a protein given a known ligand. The enormous search space yields a considerable problem.

The extent to which molecular rearrangement is important in binding is markedly varied in different cases. Many enzymes have a rigid binding pocket that will bind their ligands in the correct pose required for catalysis. Thus, little molecular rearrangement is required. Conversely, in proteins such as ionotropic glutamate receptors, the ligand-binding domains act as a "venus fly trap" [15] and there is a large difference between the bound and the unbound state. Incorporation of protein flexibility is discussed in section 6.3. Docking algorithms that do not consider protein flexibility lose accuracy due to this massive simplification and are less applicable to real problems. Despite this, rigid protein docking is currently more commonplace and can produce excellent results.

Once the three-dimensional structures of proteins became widely available with the advent of the Protein Data Bank and other such repositories, the *in-silico* simulation of protein-ligand binding became a very useful tool. This simulation is termed molecular docking. A huge variety of techniques is used, and molecular docking algorithms have proliferated. Oxdock is a docking algorithm designed and coded in

2001 by Meir Glick and Daniel Robinson. It utilises a multiscale ligand representation to yield both swift and accurate results [16-18].

2.3 Multiscale Representations

One of the most difficult problems involved in molecular docking is the enormous number of degrees of freedom. Both the ligand and the protein are flexible three-dimensional structures and a complete analysis of binding would involve a simulation of translation, rotation and vibrational modes for both species. In the majority of currently available docking methods, the protein is treated as a rigid three-dimensional structure. This greatly reduces the computer time required. Despite this, many modern algorithms do consider either complete or partial protein flexibility in an attempt to make predictions that are more accurate.

In most cases, the ligand is treated as a flexible structure and thus all the induced fit effects act on the ligand only. Despite the simplification of using a rigid receptor and flexible ligand approach, molecular docking is still an enormous task. Proteins that are used as targets of molecular docking can range in volume from approximately 2,000-7,000,000 $Å^3$. This yields a huge search space and a common solution is to use a grid-based method. Many points are created around the protein with a user-defined grid spacing and the ligand is placed only at these points. By choosing larger or smaller grid spacing, the search can be made more or less rigorous. If the search is restricted to a binding site of 1,000 $Å^3$ with a grid spacing of 0.5 Å, there are 8,000 (20 x 20 x 20) grid points to explore. The next step is to consider every possible orientation of the ligand in space. This involves rotation around the x and y axes from 0° to 360° and

around the *z*-axis from 0° to 180°. If the orientation angle is incremented by 20° then there are 2,916 (18 x 18 x 9) orientations for each grid point. The search is further complicated by the requirement to examine every possible conformation of the ligand. If the ligand contains only four rotatable bonds and these are incremented by 30°, there are 20,736 (12 x 12 x 12 x 12) conformations in total. Thus, to search each conformation in all possible orientations around every grid point requires that the ligand is docked 483,729,408,000 times (8,000 x 2,916 x 20,736). The enormous complexity of the task is readily apparent.

The multiscale approach that is used in Oxdock attempts to reduce this complexity greatly. Initially, the ligand is represented as one feature-point, which is placed at the centroid. This feature-point can be considered a virtual atom, with a partial charge equal to the total charge of the ligand and van der Waals parameters equal to the sum of the parameters for all the atoms in the ligand. This single feature-point is docked at every grid point to calculate the interaction energy. The lowest energy grid points are retained for the second iteration. In the second iteration, the ligand is represented as two feature-points (see below). These points are then rotated around the surviving grid points to calculate the interaction energy of each orientation. Again, the lowest energy solutions are retained for the third generation, in which the ligand is represented as three feature-points. This process is repeated until the solution converges. Each feature-point in any given representation is at the centre of all the atoms that it 'owns'. A feature-point 'owns' all the atoms that are closer to it than to any other feature-point (termed a cluster). The parameters of each feature-point are the sums of all the parameters for all atoms in that particular cluster. The first six

feature-point representations of the HIV Reverse Transcriptase Inhibitor Nevaripine can be seen in Figure 2.1

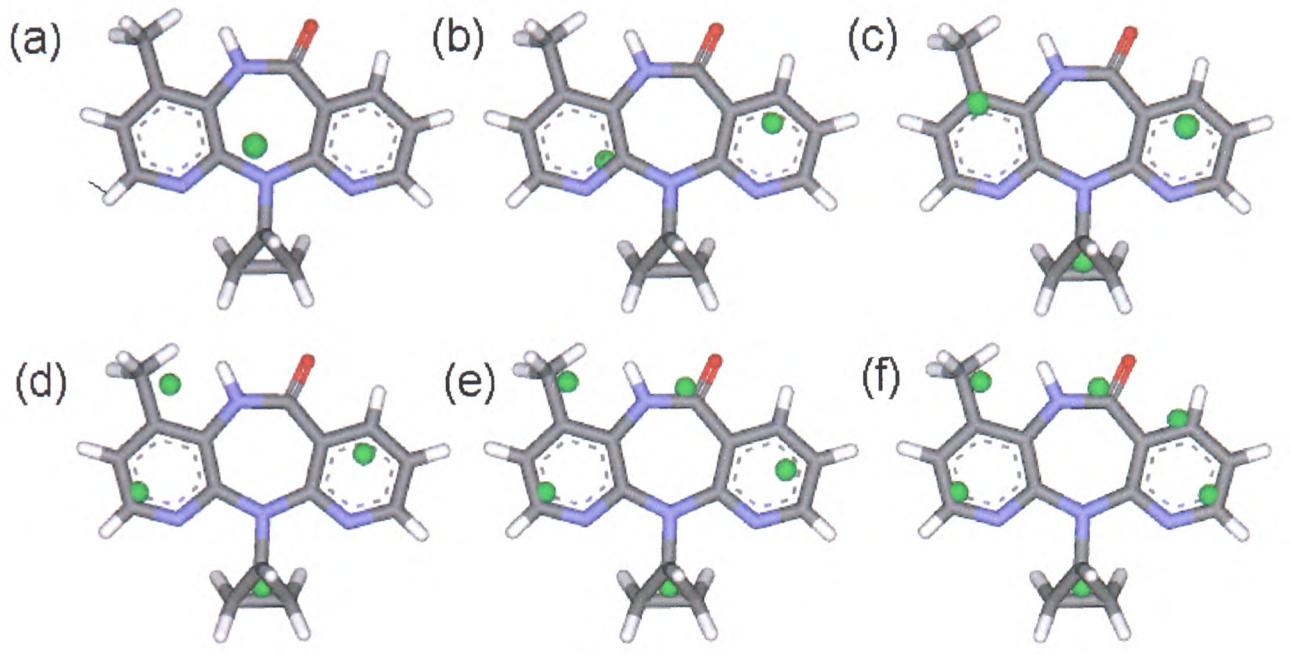


Figure 2.1 - The feature-point representations of Nevaripine for the first six iterations (a to f). The feature-points are shown as green dots.

In any *n*-feature-point representation, the atom cluster that comprises each feature-point is chosen by a *k*-means clustering algorithm. This method involves a number of steps. The one point representation is created at the centroid of the ligand. Each representation is then used to create the next representation, which has one extra feature-point. To create the $(n+1)$th representation the first stage is to select the atom that is farthest from any feature-point in the *n*th representation. This atom becomes a new feature-point. The locations of all $(n+1)$ feature-points are then optimised by the following technique. Each atom is associated with the feature-point to which it is closest. This creates a cluster of atoms for every feature-point and each point is moved to the centroid of its cluster. This process is repeated until it converges, yielding $(n+1)$ feature-points, each at the centroid of an atom cluster.

2.4 Conformer Generation

As shown in the previous section, one of the principal complexities in molecular docking is the existence of many three-dimensional structures for a ligand, based on rotation around the torsions. These structures are termed the conformers of the ligand and, if you assume unquantised rotation, there are an infinite number of conformers for each ligand. It is thus important that many of these conformers are sampled in molecular docking, as in general it will not be the minimum energy conformer that binds to the protein. This problem is solved in Oxdock by defining a torsion step for a given molecule. All the rotatable torsions (a single bond with no partial character between two non-terminal heavy atoms) are calculated and incremented by the torsion step from 0° to 360°. This creates all the possible conformers. The torsion step is user defined but a value of 30° is used as this has been tested and found to yield good results [17]. Even this modest step can produce an enormous number of conformers. For a molecule with only six rotatable torsions, there are $(360/30)^6$ or 2,985,984 conformers. It is simply untenable to dock each of these and the problem must be simplified.

The first simplifying step is to calculate the absolute internal coordinate contribution for each conformer and reject those that are above a certain energy threshold. This threshold is user defined but Oxdock generally uses 20 kcal/mol. This value retains the majority of the conformers, but excludes those that have such a high energy that they are unlikely to have a significant overall binding energy. In cases where this eliminates many of the conformers, the threshold is increased. The internal coordinate contribution is calculated using a simple electrostatic and van der Waals equation as shown in Equation 2.1.

$$Energy = \sum_{atoms} \left(\frac{q_i q_j}{4\pi\varepsilon_0 r^2} + \frac{A_{ij}}{r^{12}} - \frac{B_{ij}}{r^6} \right) + \sum_{torsions} \frac{V_n}{2}\left[1 + \cos(n\phi - \gamma)\right]$$

Equation 2.1 - The equation used to calculate the internal coordinate contribution for a conformer. In the first expression, *i* and *j* are the indices for atoms, *q* is the atomic charge, *r* is the atomic separation and *A* and *B* are the Lennard Jones parameters. In the second expression, *V* is the torsional potential barrier height for the torsion angle *Φ*, *n* is the multiplicity and *γ* is the phase factor.

Unfortunately, this filter removes only a small percentage of conformers. The second stage of filtering is to restrict the total number of conformers to prevent enormous computation time for large ligands. This can be user defined but is kept at 1,000,000 conformers for use in Oxdock. Each of these is calculated by setting the torsion angles at random values. The most important technique for increasing the speed of molecular docking is the concept of conformer clustering. This will be discussed in detail in section 6.4 and is a method for creating a very small subset of conformers to represent the total population. Once the feature-points of the representative conformations have been calculated, it is necessary to define a scoring function to find the lowest energy solutions.

2.5 Scoring Functions

In the context of algorithm design, a scoring function is a routine that maps an abstract concept to a numeric value. In molecular docking algorithms, a scoring function is a calculation of the binding energy of the ligand and the protein. Each docking algorithm has its own scoring functions and there are thus a huge variety used. However, there is a clear division between empirical and theoretical models. This is discussed later in Section 6.6. Oxdock uses a simple theoretical scoring function to approximate the binding energy of the protein and the ligand. It is based on a sum of the van der Waals and electrostatic contributions to the binding. It has the form shown in Equation 2.2.

$$Energy = \sum_{atoms} \frac{q_i q_j}{4\pi\varepsilon_0 r^2} + \frac{A_{ij}}{r^{12}} - \frac{B_{ij}}{r^6}$$

Equation 2.2 - The scoring function used in Oxdock. In the equation, *i* and *j* are the indices for atoms, *q* is the atomic charge, *r* is the atomic separation and *A* and *B* are the Lennard Jones parameters.

The electrostatic term is calculated using a simple coulomb potential with a distance-dependent dielectric. The van der Waals term is calculated as the sum of the repulsive and attractive terms of the Lennard Jones 12-6 potentials. The van der Waals parameters A_{ij} and B_{ij} are taken from the AMBER forcefield [19]. The interaction energies are calculated using a grid-based approach. Before the docking iterations

begin, an energy grid is calculated. A cubic lattice is placed around the protein and grid points are placed at a user-defined resolution throughout the entire grid (a resolution of 0.5 Å is commonly used). The electrostatic potential and an effective van der Waals potential are pre-calculated at each point, based on the distance between the point and each protein atom. During the actual docking, the energy potentials at a point in space are considered as the energy potentials at the nearest grid point. This approach greatly increases the speed of energy calculations in molecular docking.

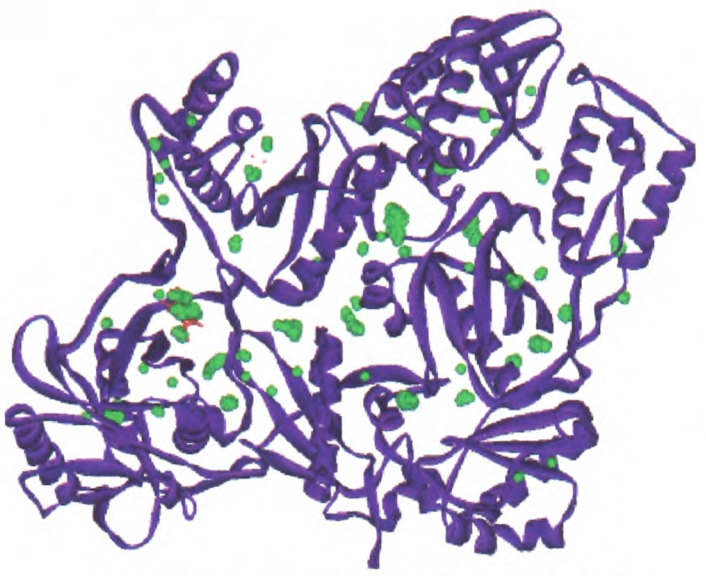
2.6 Applications and Initial Results

Oxdock has been validated for a variety of test cases to prove its applicability to the problem of finding binding sites on protein surfaces [16-18]. However, an example will be given here to illustrate the principle. The complex 1VRT from the PDB (http://www.pdb.org/) between Nevirapine and HIV Reverse Transcriptase can be split into the ligand and protein. Hydrogen atoms are then added and the charges fixed by InsightII using the Consistent Valence Force Field (CVFF) [20]. Both ligand and protein are saved in the car file format and then passed to the algorithm using parameters in Table 2.1.

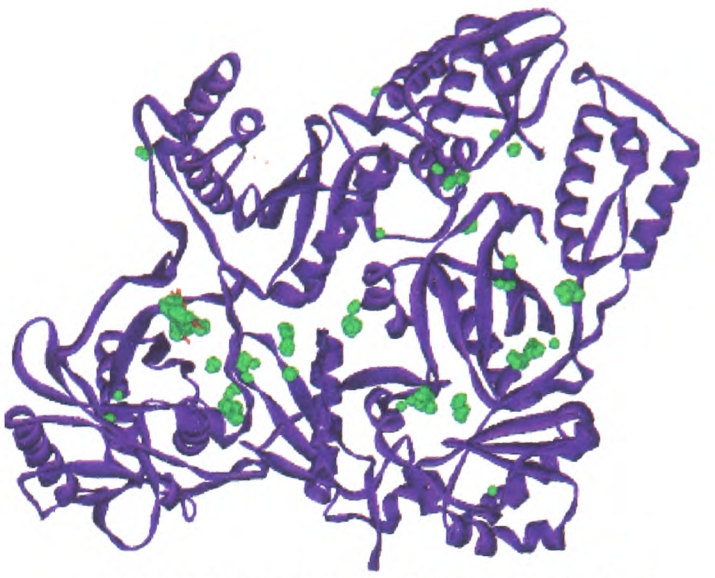
Parameter	Value
Grid Resolution	0.5 Å
Survivors Each Iteration	10 %
Flexible Docking	Yes
Torsion Rotation	30 °
Conformer Energy Threshold	20 kcal/mol
Maximal Grid Points	500

Table 2.1 - Table giving the Oxdock parameters used for the docking of Nevaripine to HIV Reverse Transcriptase.

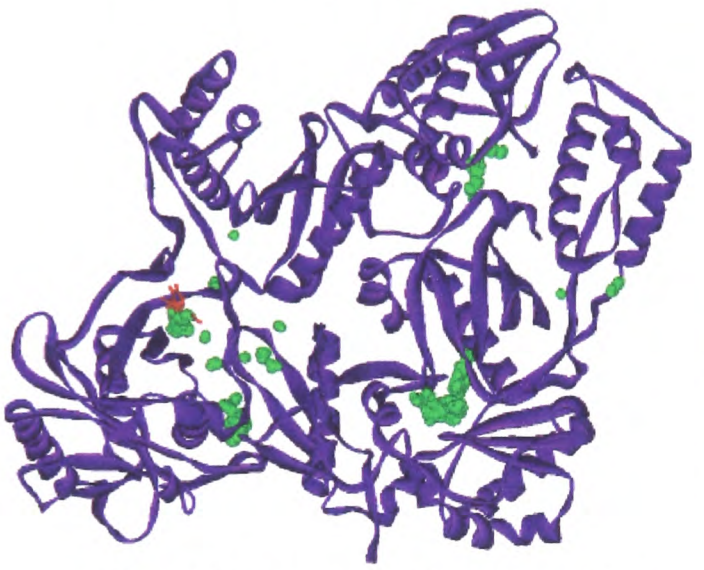
These parameters are standard for use in Oxdock. The results for the first five iterations can be seen in Figure 2.2. The results illustrate that Oxdock locates many good grid points in the first generation with a large cluster in the active site where Nevaripine binds. The poor solutions are slowly removed over subsequent iterations and the fifth and final iteration contains only 58 solutions.



One Feature-point Representation

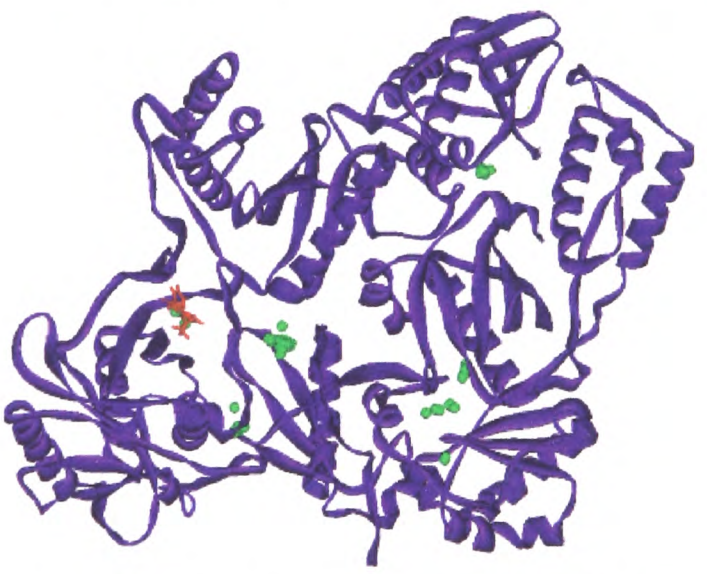


Two Feature-point Representation



Three Feature-point Representation

Four Feature-point Representation



Five Feature-point Representation

Figure 2.2 - The first five iterations of Oxdock with the ligand Nevaripine (shown in red) and the protein HIV Reverse Transcriptase (shown in purple). The surviving grid points are coloured green.

The algorithm converges after the five-point representation and outputs the 50 lowest energy solutions. The lowest energy solution (calculated interaction energy of -63.94 kcal/mol) can be seen in Figure 2.3.

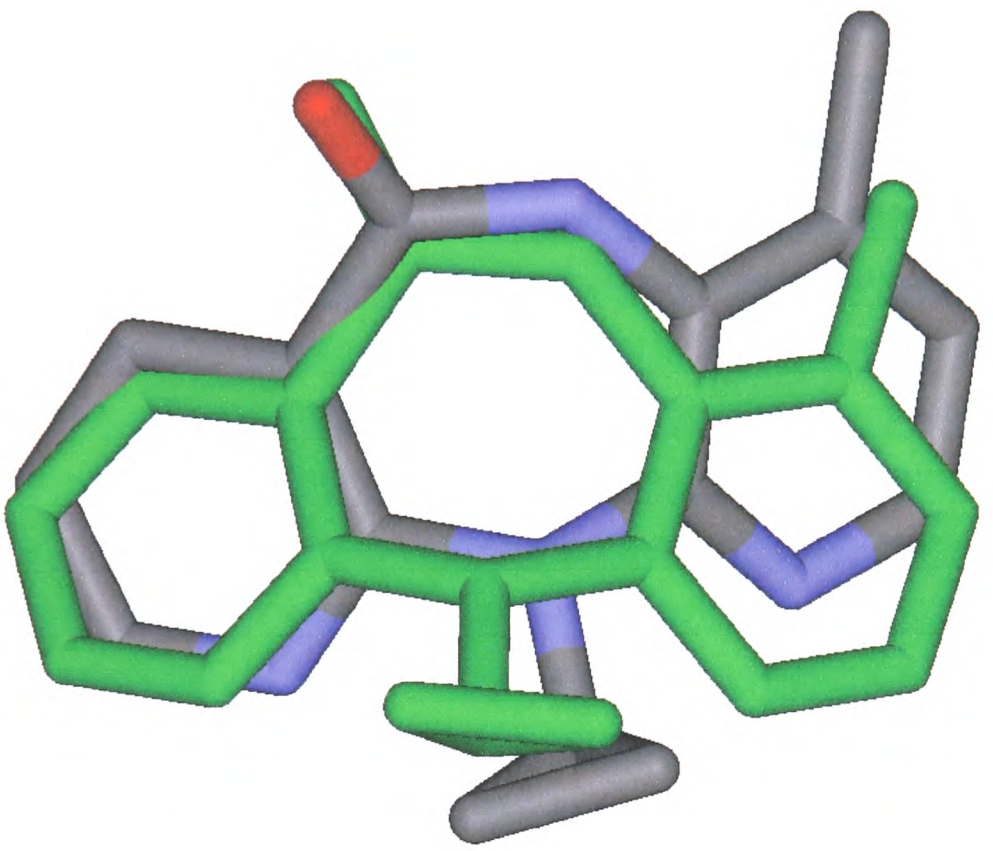


Figure 2.3 - The lowest energy solution for docking Nevaripine with HIV Reverse Transcriptase (PDB ID 1VRT). The crystal structure is coloured green and the best docking solution is atom coloured.

HIV Reverse Transcriptase is a very large protein comprising approximately 1000 amino acid residues and taking up a volume of approximately 432,000 $Å^3$ (90x80x60 Å). The algorithm can make a very accurate prediction of this binding site with a run time of 21 hours and 21 minutes (on a 766MHz processor). This run time is typical of Oxdock. In cases where the ligand has a large number of torsions, the run may take a

few days and in cases where the protein is small or the active site is defined, the run may take a few hours. In most cases, Oxdock converges on a solution in between 6 and 10 hours. This example illustrates the enormous power of Oxdock and hints at its potential in solving biological problems.

2.7 Summary

Oxdock has proved to be an excellent algorithm for the location of active sites on proteins. It can rapidly explore the large and complex search space that is the surface of a protein and converge upon the correct site. This ability is unique amongst docking algorithms, whose main purpose to date has been the elucidation of docking poses and not the location of binding sites. The methodology involves the use of a multiscale approach to the problem. The ligand is represented as one feature-point in the first iteration and then as an increasing number of feature-points in subsequent generations. The energy of each representation is calculated for every grid point using a Coulombic expression with a distance-dependent dielectric for the electrostatic binding and a Lennard Jones 12-6 expression for the van der Waals interaction. The algorithm ignores the highest energy solutions in each generation and thus converges upon the lowest energy solutions. All the possible rotamers are calculated before the run begins and each one that qualifies as distinct becomes an input ligand. Rotamers are defined as distinct if the feature-points are further apart than a defined distance threshold from other feature-points. This process ensures that a distribution of conformers is sampled. This in turn increases the accuracy of the algorithm. The success of Oxdock allows it to be employed to solve genuine biological problems. In chapter three, the algorithm is used to locate the binding site for spermine, which is

known to bind to the extracellular portion of NMDA receptors. The intriguing implications of this result are then explored.

3 Analysing NMDA Receptors

3.1 Overview

Some of the most interesting of all the protein families are the membrane bound receptors that control cell metabolism, development and behaviour. These include hormone receptors, G-protein coupled receptors and neuronal receptors. It is estimated that 60-70% of existing medicines are targeted at these families of proteins [21].

Ionotropic glutamate receptors (iGluRs) are neuronal receptors, which function through the binding of glutamate at synapses in the mammalian brain [22]. At these receptors, binding of glutamate results in gating of the trans-membrane channel, allowing transport of ions such as Ca^{2+} or Na^{+} [23]. The N-methyl-D-aspartate (NMDA) subclass of iGluRs are activated by the combined presence of glutamate and glycine and are thought to exist as a dimer of dimers, with four protein subunits assembling to make a channel through the membrane.

The activity of NMDA receptors can be modulated with a variety of endogenous ligands such as zinc ions [24], polyamines [25], phenylethanolamines [26] and protons [27]. It is thought that the binding sites for these modulators are found in the extracellular amino terminal domain (ATD) of such receptors [28]. However, despite the enormous efforts in mutagenesis and patch clamp experiments on NMDA receptors, the exact assembly of these subunits and the effects of the modulatory species are not well understood.

In this chapter, a combination of homology modelling and the algorithm Oxdock are used to elucidate the structure and the mode of action of NMDA receptors.

3.2 NMDA Receptors

3.2.1 NMDA Receptor Function

Ionotropic glutamate receptors are ligand-gated ion channels. Upon binding glutamate that has been released from a companion cell, ions such as Na+ and Ca2+ pass through a trans-membrane channel in the receptor. This flow of ions results in a depolarisation of the plasma membrane and the generation of an electrical current, that is propagated down the dendrites and axons of the neuron to the next synaptic junction. The iGluR family contains a variety of sub-types that are classified by their receptivity to certain ligands:

- Kainate receptors are implicated in facilitating synaptic plasticity and exist in a huge variety of splice variants.
- Alpha-Amino-3-Hydroxy-5-Methyl-4-Isoxazole Propionic Acid (AMPA) receptors mediate fast synaptic transmission, and are known to have a large N-terminal domain.
- NMDA receptors are activated by the combined presence of glutamate and glycine and are potentiated by the action of the other receptor sub-types. They also have a large N-terminal domain.

The importance of these receptors in regulating the function of the central nervous system is indisputable and huge amounts of experimental data have been collected. This is particularly true for NMDA receptors and a picture of their exact function is slowly being drawn. However, much still remains unknown about both their structure and their function.

3.2.2 NMDA Receptor Structure

NMDA receptors are composed of assemblies of three different classes of subunits [29], NR1, NR2 and NR3. The combination of subunits determines the specific properties of the receptor. The NR1 and NR3 ligand-binding domains are known to bind glycine and the NR2 subunits to bind glutamate. The majority of receptors are composed of a combination of NR1 and NR2 subunits and thus both glutamate and glycine are necessary to facilitate channel opening. However, some receptors are composed of NR1 and NR3 subunits and require only glycine. Each subunit has three trans-membrane regions, one membrane-embedded domain, one bi-lobed domain that forms the ligand-binding site and one bi-lobed amino terminal domain that is thought to be homologous to Leucine/Isoleucine/Valine-Binding Protein (LIVBP) [30].

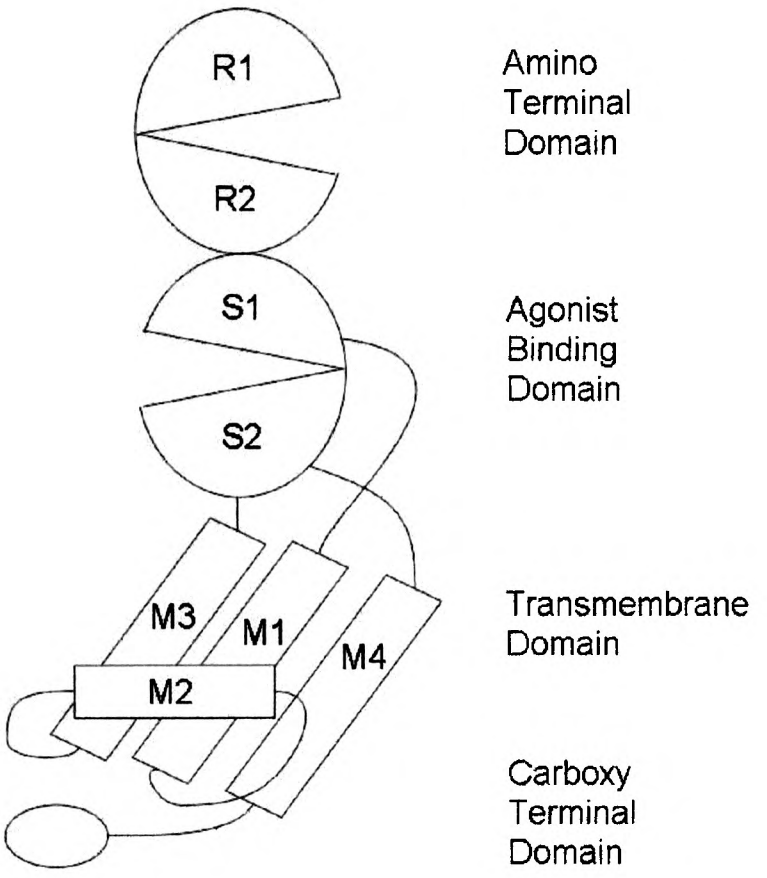


Figure 3.1 - The proposed topology of NMDA receptors in conceptual format. There is a large amino-terminal domain (R1-R2), a large ligand-binding domain (S1-S2), three trans-membrane helices (M1-M3-M4), a re-entrant loop (M2) and a small carboxy-terminal domain.

Previous studies of the ligand-binding domain of the rat ionotropic glutamate receptors GluR2 (PDB ID 1FTJ) [31] have highlighted the region that binds glutamate as the extracellular venus fly-trap domain S1-S2, which is coupled to the M2 helix that regulates the flow of ions.

3.2.3 Modulators of NMDA receptors

3.2.3.1 Protons

Protons inhibit NMDA receptors. At physiological pH, they are inhibited by as much as 50%. The presence of the NR1a subunit seems to be vital for this effect, as receptors containing the NR1b subunit are not affected. This may be due to the presence of an extra exon (Exon 5) in the ATD of the NR1b subunit that is thought to relieve pH inhibition [27]. Acidic residues [32] and histidine residues [33] in the ATD have been implicated in proton inhibition.

3.2.3.2 Polyamines

Polyamines such as spermine relieve the pH inhibition of NMDA receptors and restore normal function. Only receptors containing the NR1a subunit are affected by polyamines and they may thus act by mimicking the action of exon 5, which, like spermine, is known to contain many positively charged residues. Mutation of acidic residues suggests that the polyamine binding site is in the ATD of NMDA receptors [32]. Evidence suggests that polyamines may have a number of effects on NMDA receptors and that there may be more than one polyamine binding site.

3.2.3.3 Aminoglycosides

Amino-glycoside antibiotics are known to increase the action of NMDA receptors and it has been suggested that this occurs by binding to a spermine-binding site and thus relieving pH inhibition (amino-glycosides contain a number of amino groups with a similar spacing to those in polyamines). Only receptors containing the NR1a/NR2b combination of subunits are affected by amino-glycosides [34].

3.2.3.4 Zinc Ions

Zinc ions also inhibit NMDA receptors. They are thought to act by increasing the rate of receptor desensitization and thus decreasing the current passing through the channel [33]. Zinc ions inhibit receptors containing the NR2a or NR2b subunit [24]. However, for NR2a subunits this inhibition is intensified at low pH. It has thus been suggested that zinc ions act by enhancing proton inhibition. Mutagenesis data suggests that that the zinc binding site is in the ATD [35].

3.2.3.5 Phenylethanolamines

Phenylethanolamines such as Ifenprodil inhibit NMDA receptors. Their effect is pH dependent and thus they are thought to act in a similar manner to zinc, by enhancing proton inhibition [36]. Ifenprodil affects only those receptors containing the NR2b subunit. Mutagenesis data suggests that that the Ifenprodil binding site is also in the ATD [26].

3.2.4 NMDA Receptor Action

The basic action of NMDA receptors is well documented. Glutamate and glycine bind to the ligand-binding domains of the multimeric receptor and this leads to domain closure. This is transmitted downstream to the trans-membrane helices, opening a pore in the membrane allowing ions to flow from the synaptic cleft to the neuron cell. However, the large number of modulatory ligands acting at NMDA receptors, combined with a lack of understanding in how ligand binding couples to channel opening, means that the mechanism of action of each ligand is difficult to explain. However, it is useful to note that all of these modulatory ligands are thought to act in the ATD of NMDA receptors. The ATD of such receptors has not been studied in as

much detail as the downstream glutamate-binding region, despite the fact that it is larger and appears to have an interesting biochemistry.

3.3 Homology Modelling

The glutamate-binding domain of NMDA receptors is thought to be homologous with periplasmic binding proteins such as Lysine/Arginine/Ornithine Binding Protein and Glutamine binding protein [37]. These proteins are composed of two domains and have open and shut configurations (agonist bound and unbound). Results from the recent crystal structure studies of the glutamate-binding region of a similar protein, the glutamate receptor GluR2, strongly support this theory [22, 31, 37].

The ATDs of NMDA receptors have not been crystallized but it has been noted that there is homology with other periplasmic binding proteins such as LIVBP [30] and Polyamine Binding Protein [38]. These proteins are also composed of two domains that have agonist bound and unbound configurations. The presence of bound and unbound states could allow binding of modulatory ligands in the ATD.

The creation of a homology model allows elucidation of the function for the ATD of these receptors. Ifenprodil binding has been explored in the past by modelling of the NR2b subunit based on the structure of LIVBP [26]. This work highlights homology not only with LIVBP, but also with the extracellular domain of the metabotropic glutamate receptor (mGluR1) and a hormone binding protein, ANP-C. Both proteins occur naturally as dimers and the extracellular domain of mGluR1 [39] and ANP-C [40] have been crystallized as dimers.

3.3.1 Dimer Formation

Recently, the assembly of two of the glutamate-binding regions has been elucidated by X-ray diffraction [41]. In the closed state, two subunits associate to form a dimer, with two helices from the S1 domain of each subunit contributing to form a four-helix bundle (Figure 3.2).

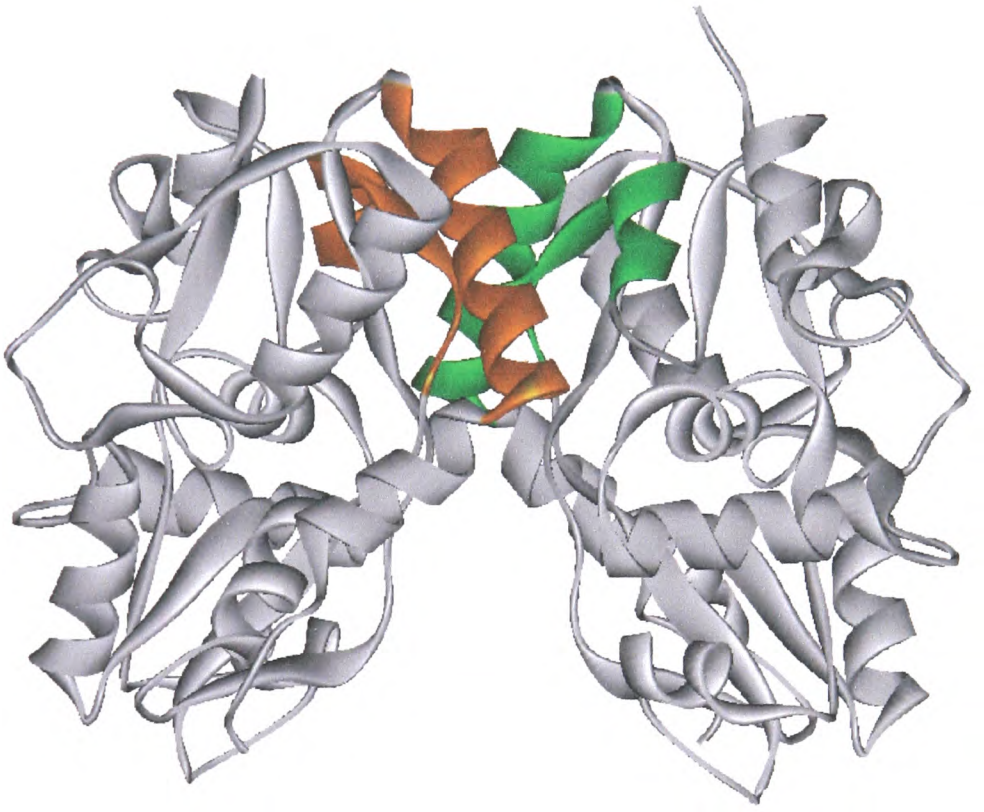


Figure 3.2 - The structure of the dimer of the Glutamate-binding region of GluR2. The four helices comprising the four-helix bundle are coloured in orange and green. It is interesting to note that the residues around these helices are very well conserved across all Glutamate receptors

In the model postulated by Gouaux [41], there are two more states accessible to the dimer. Opening of the channel involves the closure of the S1-S2 cleft to bind

glutamate with the retention of this interface. Essentially the S2 lobes of the subunits swing out and away from one another, pulling apart the helices that block the channel and producing an open state. The third state involves closure of the S1-S2 cleft to bind glutamate and breaking of the S1-S1 domain interaction. This leaves the gate closed but with glutamate bound to the receptor. This is the desensitized state.

The model suggests that the S1 lobes are in close proximity in the closed state and thus the R2 lobes, which are linked to the S1 lobes, will be in close proximity for the closed state. It has been shown that the ATD is vital in functional assembly of AMPA and kainate receptors [42] and that protein-protein interactions in both the S1-S2 and ATD domains are important for subunit assembly [43]. The ATDs might thus be linked together in an analogous way to the S1-S2 glutamate-binding region. Sequence analysis highlights two regions of the ATD that correspond to two helices in mGluR1 and ANP-C that form a four-helix bundle between two subunits (Figure 3.3). Residues in these regions show very high sequence similarity between NR2a and NR2b but these are very different from NR1a. This complementarity means that NR1 subunits would have to combine with NR2 subunits to produce a functioning receptor, as is found in nature. The two crystal structures of mGluR1 are from the Protein Data Bank (PDB ID 1EWT for the apo state and 1EWK for the glutamate-bound state). The mGluR1 monomer sequence, the ANP-C monomer sequence and the human NMDA sequences are aligned using ClustalW [44]. The homology package of InsightII can then be used to assign coordinates to the amino acid sequence of an NR1a/NR2a and an NR1a/NR2b dimer. The models obtained are then refined by CHARMM to optimise the three-dimensional structure [6].

3.3.2 Disulphide Bridging

The model suggests that an NMDA dimer has a structure in which there are protein-protein interactions linking both the S1 and R1 domains. Further evidence comes from mutagenesis data on Cysteine residues [45]. A Cysteine residue in the ATD has been implicated in zinc inhibition but the results suggest that it modulates the zinc effect without binding to zinc. For an NMDA-ATD dimer, these Cysteine residues are located on helices suggested to form the four-helix bundle and the equivalent residues for the mGluR1 dimer are in close proximity (Figure 3.3).

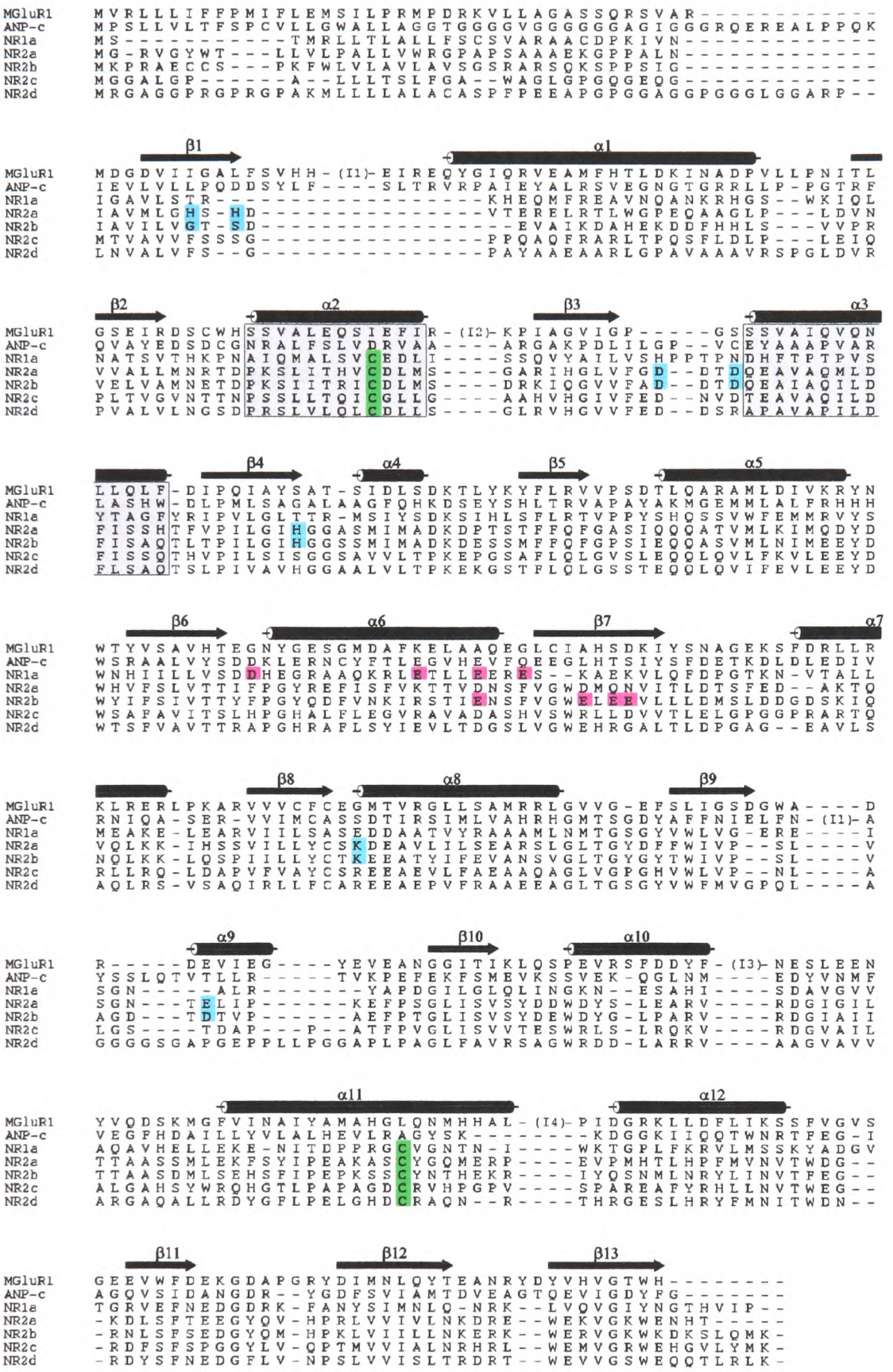

MGluR1
ANP-c
NR1a
NR2a
NR2b
NR2c
NR2d
β1
α1
β2
α2
β3
α3
β4
α4
β5
α5
β6
α6
β7
α7
β8
α8
β9
α9
β10
α10
α11
α12
β11
β12
β13

Figure 3.3 - Multiple sequence analysis of the ATD of NMDA receptor subunits, mGluR1 and ANP-C. The regions that are known to form a four-helix bundle between adjacent subunits are highlighted in the grey box. Residues implicated in Zinc binding are highlighted in blue. Cysteine residues implicated in zinc binding are highlighted in green. Residues implicated in Spermine binding are highlighted in lilac.

This suggests that a Cysteine bridge links the two subunits at this point. Sequence analysis reveals that this residue is conserved across all NMDA receptor subunits and across the AMPA and kainate receptors studied (data not shown). It is interesting to note that the two subunits of the mGluR1 dimer are also connected by a Cysteine bridge, but at slightly different points in the equivalent domains. Models for the open and closed NR1a/NR2a and NR1a/NR2b dimers are constructed, exploiting their homology with the mGluR1 dimer (Figure 3.4). This protein was chosen to create the model due to the existence of both "open" and "closed" X-Ray structures and the similarity between metabotropic and ionotropic glutamate receptors. The insertions found in mGluR1 with respect to LIVBP are removed, as they are not believed to be present in NMDA receptors.

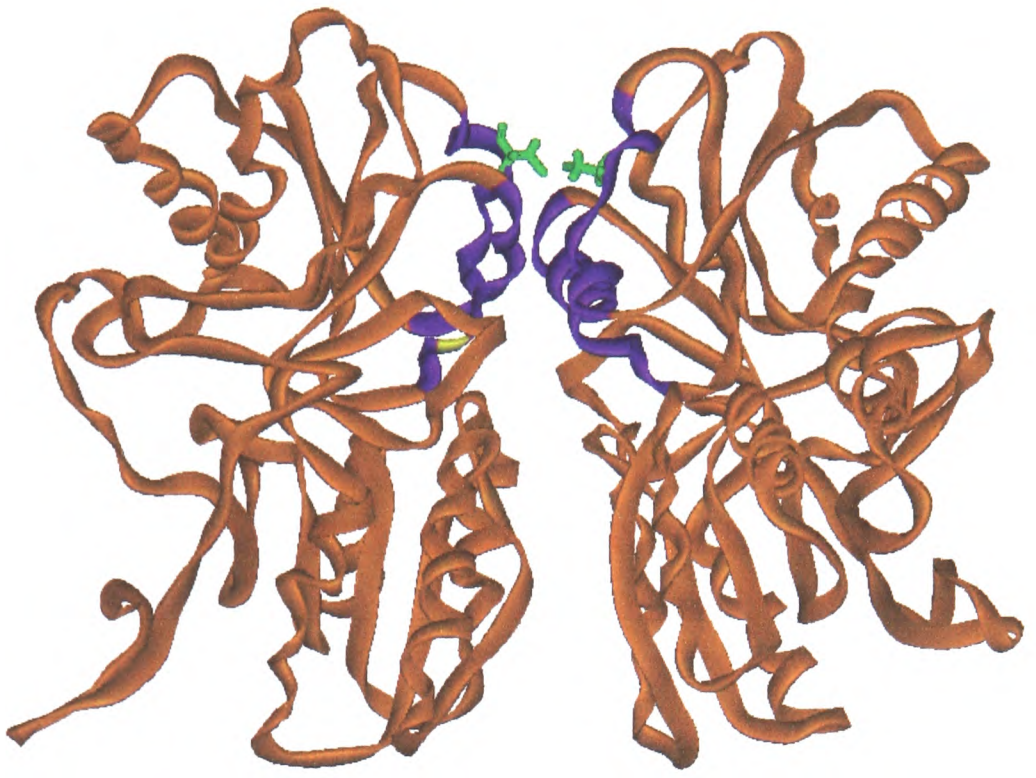


Figure 3.4 - Proposed structure of the extracellular dimer of an NMDA receptor, highlighting the positions of the four-helix bundles (purple) and the disulphide bridge (green) that may help to stabilize this structure.

The structures were then closely analysed to provide information about the function of these receptors.

3.3.3 Zinc Binding

The residues implicated in zinc binding [35] are found in the binding crevice of the NR2a subunit, as shown in Figure 3.5. Mutagenesis data implicates residues from both the R1 and R2 domains. This suggests that zinc binds to the closed state of the ATD. However, the homology model shows that in the closed state the Lysine 233 residue in the R2 domain that is implicated in zinc binding [46] appears to bind to five residues from the R1 domain that are also involved in zinc binding (Histidine 42,

Histidine 44, Aspartate 102, Aspartate 105 and Histidine 128). See Figure 3.5 a and b.

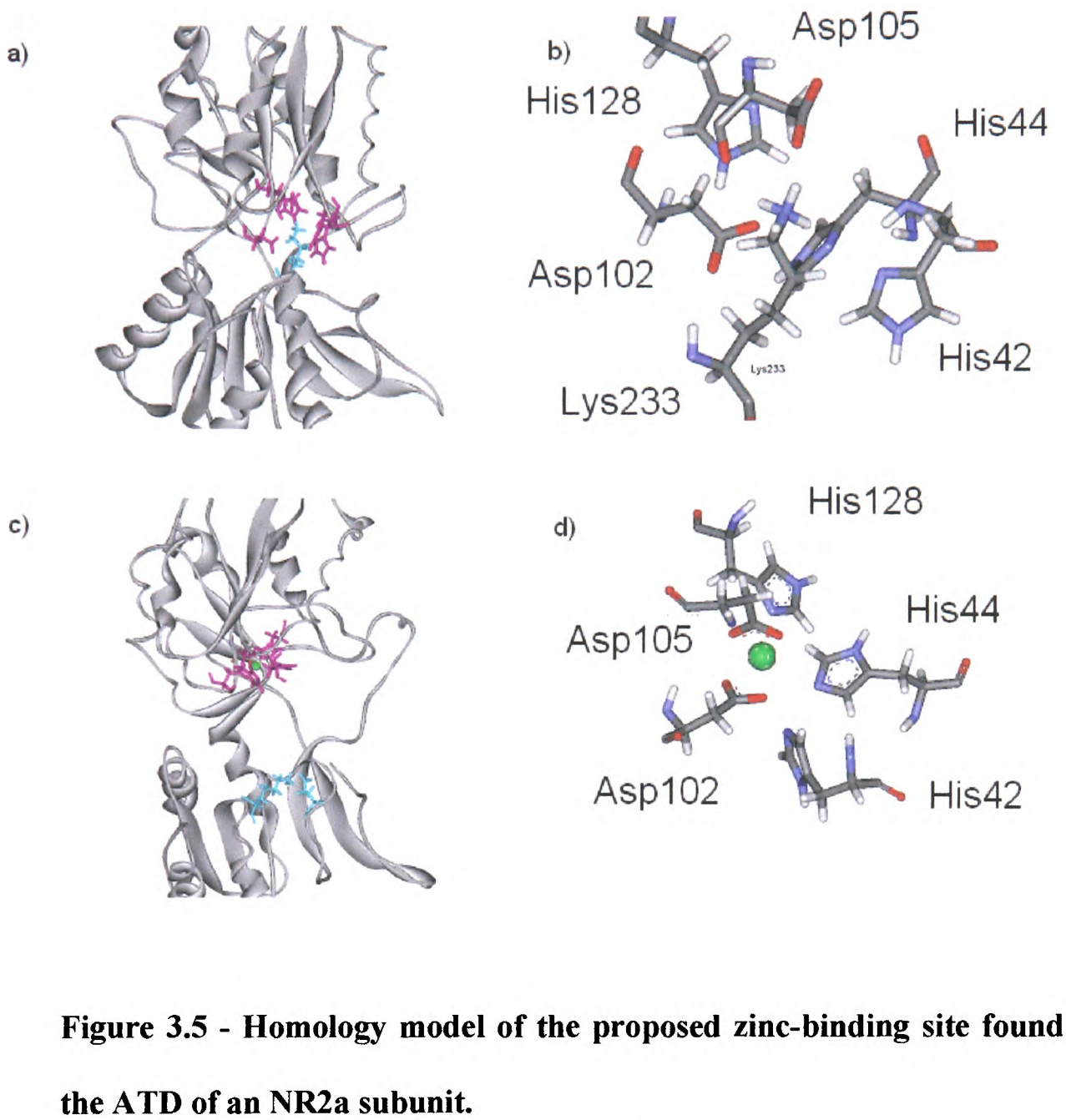


Figure 3.5 - Homology model of the proposed zinc-binding site found in the ATD of an NR2a subunit.

(a) **In the closed state, the interaction between the Lysine 233 in blue and the binding site in pink holds the domains together**

(b) **Lysine 233 in the R2 domain binds to Histidine 42, Histidine 44, Aspartate 102, Aspartate 105 and Histidine 128 in the R1 domain.**

(c) **In the open state, Lysine 233 interacts with Glutamate 266 (both in blue) on the opposite side of the cleft, keeping the domains apart.**

(d) **The Zinc ion is bound to the same five residues in the R1 domain.**

It seems likely that domain opening would remove the Lysine from the binding site created by these residues and allow zinc to bind in its place, as zinc and ammonium ions are of a similar size (88 pm and 125 pm). Thus, a desensitized state is likely to involve an open configuration of the ATD with a zinc ion bound to the R1 domain. Further analysis of the energy-minimised structure suggests that in the open state this Lysine residue forms an interaction with the Glutamate 266 residue from the R2 domain. This Glutamate is the final residue implicated in zinc binding, completing an elegant model (See Figure 3.5c and d).

3.3.4 Ifenprodil Binding

The residues implicated in Ifenprodil binding [26] are found in the binding crevice of the NR2b subunit. Residues from both the R1 and R2 domains seem to be important for binding, as shown in Figure 3.6.

Figure 3.6 - The ATD of the NR2b subunit represented as a grey ribbon. The residues highlighted by mutagenesis to be important in Ifenprodil binding are coloured purple.

This suggests that Ifenprodil would bind to the closed state of the ATD. However, given the similar effects of zinc and Ifenprodil, it seems highly likely that Ifenprodil in fact binds to the open state of the ATD and thus stabilizes the desensitized state. Again, there is a binding site for a Lysine residue (Lysine 234) in the R1 domain and the equivalent residues are implicated in this interaction. However, two of the Histidine residues have been replaced by an aspartate residue in the case of the NR2b subunit and the propensity of zinc to bind with N-donors may explain why zinc does not bind as readily in this case. The relatively long time (in comparison to zinc) taken for Ifenprodil to exert its influence after release begins and stop acting after release ends [26] is attributable to its large size and thus its relative inability to enter and leave the binding site

3.3.5 Spermine Binding

The residues from NR1a that have been implicated in spermine binding [32] are not found in the binding crevice, but on a helix in the R2 lobe on the side of the subunit. The position of the exon 5 insert suggests that it would also be found in this region. Sequence analysis highlights a large number of acidic residues at the interface between the subunits (see Figure 3.3). These negatively charged residues on both sides would give rise to electrostatic repulsion and destabilize the dimer interface. Spermine could bind to these residues between the R2 lobes of NR1 and NR2 subunits, neutralising the interface between them and thus stabilizing the open and closed states of the receptor relative to the desensitized state (in which the interface has been moved due to rearrangement). However, if residues in the NR2 subunit are important in spermine binding then mutagenesis data should highlight this. Little mutagenesis of acidic residues has been done for NR2 subunits but it has been

suggested that mutation of E191, E198 and E200 and E201 in the NR2b subunit affect the action of spermine [47]. These residues can be found at the inter-subunit boundary of the dimer, in close proximity to the helix controlling spermine action in the NR1a subunit. The exact location of the spermine-binding site was thus explored using Oxdock with the NR1a/NR2b dimer. The lowest energy result can be seen in Figure 3., with the ligand at the inter-subunit boundary.

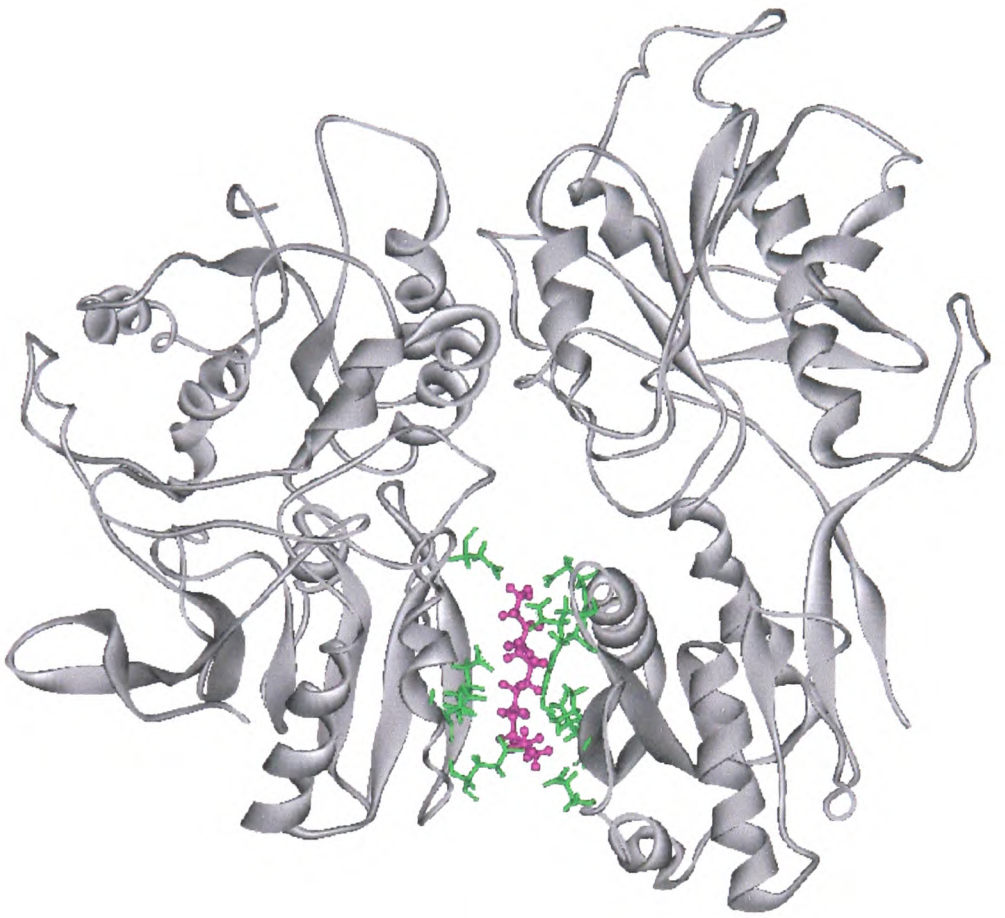


Figure 3.7 - Homology model of the proposed spermine-binding site found at the interface between the NR1a and NR2b subunits. Residues D169, E181, E185 and E192 in the NR1a subunit and residues Y174, Q180, D181, E200, V202 and D211 in the NR2b subunit are displayed in green and the spermine is displayed in pink.

This calculation took approximately 24 hours (on a 766MHz processor). The early stages of the algorithm suggested a binding site at the inter-subunit boundary,

probably due to the presence of a number of negatively charged residues. Further iterations reinforced this idea. The agreement of computational prediction with experimental mutagenesis data strongly suggests that a spermine-binding site is present between the subunits. This idea has not previously been suggested and it has many implications for receptor function. Interestingly, the presence of a ligand-binding site between subunits is also found in the hormone binding protein ANP-C, which has homology with NMDA receptors, as shown in Figure 3.8.

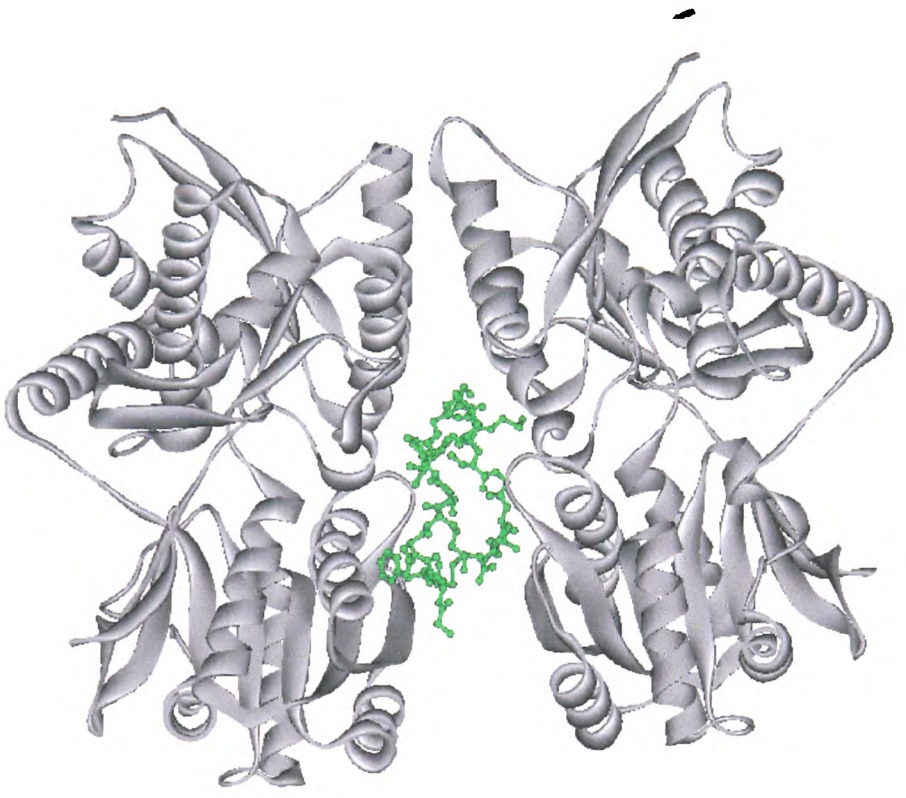


Figure 3.8 - The crystal structure of ANP-C in green with the hormone bound between subunits in an analogous fashion.

Interestingly, the implicated residues from the NR2b subunit are not present in the NR2a subunit. However, it has been suggested that glycine independent spermine stimulation occurs only for receptors containing the NR2b subunit [25]. Such

stimulation increases the magnitude of the current passing through the membrane. The second type of spermine stimulation, glycine dependent, may occur at a different site. Perhaps, given that this increases the affinity of the receptor for glycine only, the binding site is located entirely on the NR1a subunit.

3.4 Opening, Closing and Desensitizing

3.4.1 Mode of Action

Any movement following glutamate binding must transmit upstream to the ATD. If it is assumed that the ATD has a bi-lobed structure like LIVBP and is approximately the same size (S1-S2 is composed of approximately 320 residues and R1-R2 of approximately 350) then a simple mechanism suggests itself. The Cysteine linked four-helix bundle between the R1 domains can act as a hinge for the ATDs, allowing the R2 lobes to open and close whilst the R1 lobes remain in the same position. The ATD is in the closed conformation for the open and closed states of the receptor. Spermine stabilizes these states by binding at the interface between subunits, which is otherwise destabilized by electrostatic repulsion. Desensitization leads to the S1 lobe swinging out and away from the body of the receptor and this can be transmitted upstream if this pulls the R2 domain yielding an open configuration for the ATD. Zinc or Ifenprodil bind to the R1 domain of the open ATD and thus stabilize the desensitized state of the receptor. This process is shown in Figure 3.9.

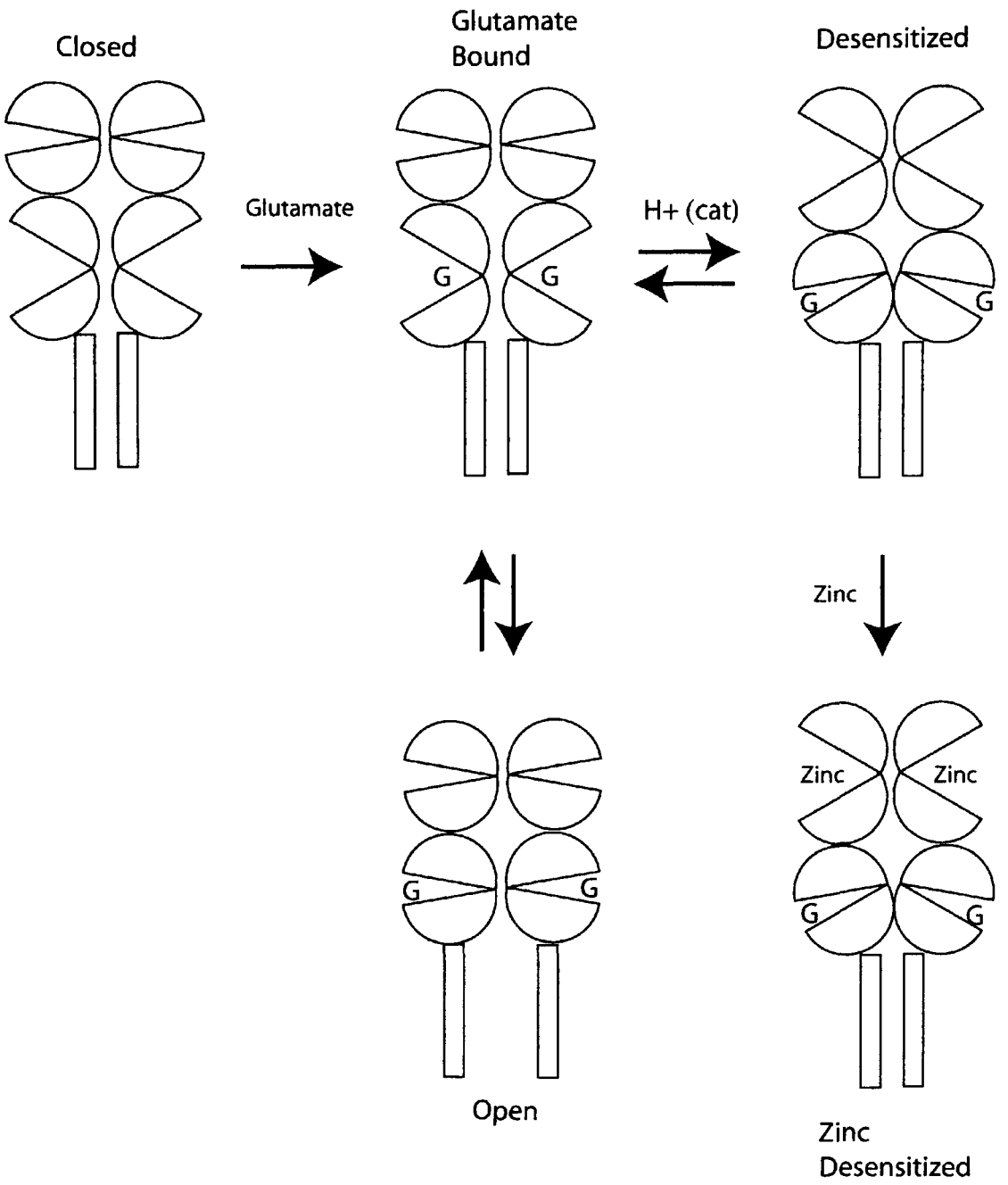


Figure 3.9 - Diagram illustrating the five postulated states of an NMDA receptor and the effect of glutamate and zinc on the equilibrium.

The pH dependence of zinc inhibition of NR2a containing receptors is known to be affected by mutations at the three Histidine residues (Histidine 42, Histidine 44 and Histidine 128) [33]. The mechanism described above, combined with this data, suggests that protonation of one or more of these residues increases the thermodynamic stability of the desensitized state or increases the rate of its formation. The first suggestion seems unlikely. A positive charge on one of these Histidines seems more likely to destabilize a zinc bound state due to electrostatic repulsion. The

second suggestion has merit. The main barrier to R1-R2 domain opening would appear to be the charge separation between the positive Lysine 233 residue and the negative Aspartate 102 and Aspartate 105 residues. However, if one of the Histidines is protonated, which would happen if the pH were decreased, the barrier to domain opening would be decreased and the rate would thus increase. The next step, zinc binding, could then occur by exchange with a bound proton. Effectively, the proton is acting as a catalyst for R1-R2 domain opening and lowers the activation barrier for desensitization (Figure 3.10).

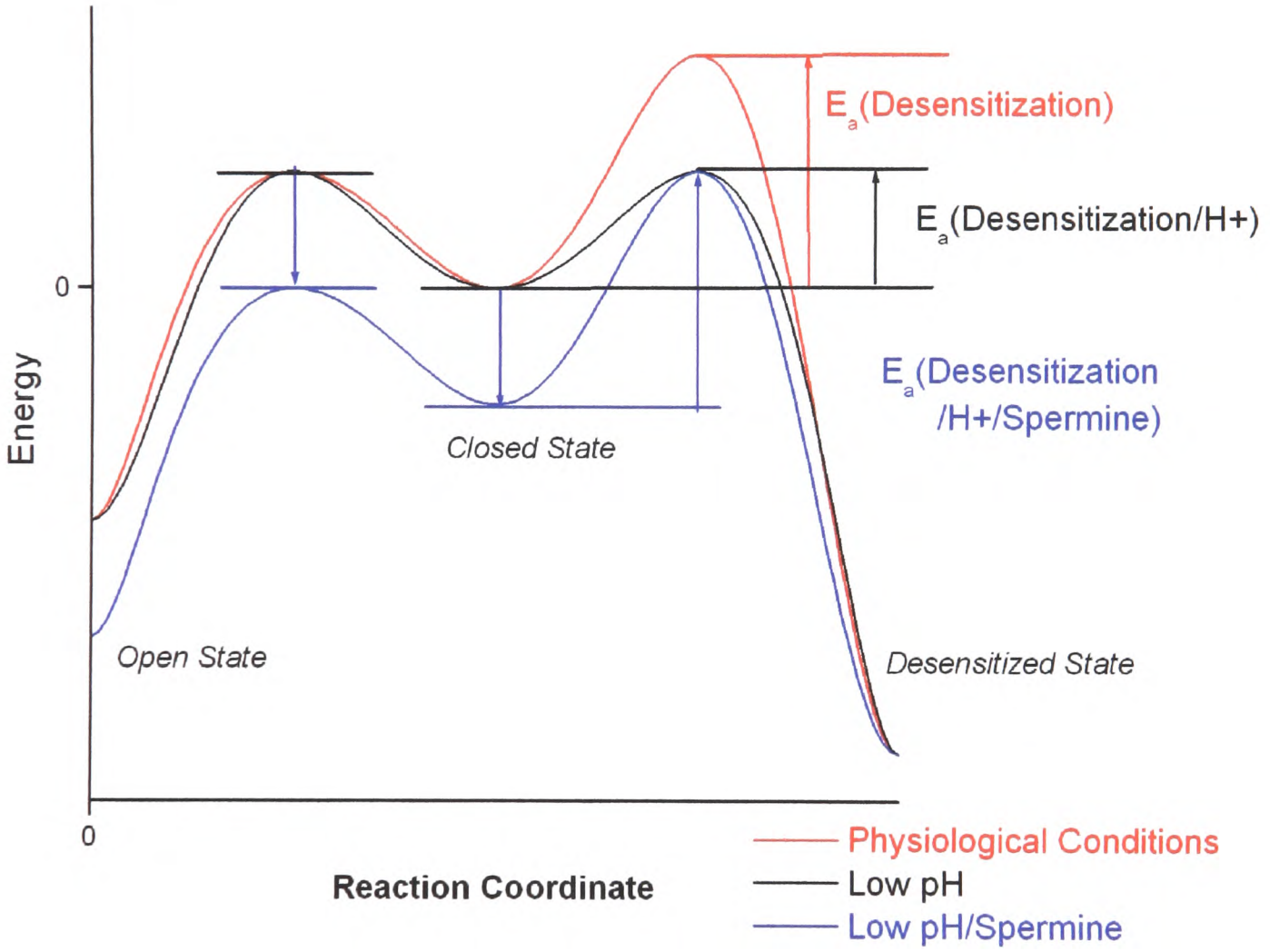


Figure 3.10 - Energy level diagrams of the NMDA receptors. The open and desensitized states lie at lower energies than the closed state but the energy barrier to the desensitized state is greater leading to an initial spike in the open probability of the receptors. Protons lower the

activation barrier for transition to the desensitized state and thus inhibit channel opening. Spermine stabilizes the open and closed states of the receptors and thus raises the activation barrier for transition to the desensitized state. This returns normal function to inhibited receptors.

Spermine can then counter this effect by stabilizing the closed state, thus effectively returning the activation barrier to its original value. This model is completely new and fits with all of the experimental data studied as well as being easily testable.

3.4.2 Implications

The elucidation of this method of action has interesting implications. If Ifenprodil can bind to the open state of the NR2b subunit, it should be possible to design a ligand to bind in an analogous fashion to the NR2a subunit and perhaps the NR1a, NR2c and NR2d subunits. This work also highlights the importance of examining the complete structure of these receptors and not simply the individual subunits. The spermine-binding site at the monomer-monomer interface could not have been found by examining the monomer alone. Indeed, it seems highly likely that another spermine-binding site can be found at the dimer-dimer interface around residues Glutamate 339 and Glutamate 342 of NR1a. These residues have been highlighted by mutagenesis as being important for glycine-independent spermine stimulation and are positioned perfectly to interact with another subunit of the tetramer. The structure of this tetramer may also explain the importance of a further pair of conserved Cysteine residues (Cysteine 308 in NR1a and Cysteine 320 in NR2a) which have been implicated in desensitization but are not thought to bind zinc [45]. These residues are found on the surface of the subunits, at the apex of the ATD. They could together form a disulphide

bridge, linking the two dimers together to yield a functioning tetramer. The stability of the ATD could be exploited by proteolytically severing the tetrameric ATD head of these receptors, allowing crystallization and analysis of the structure.

3.5 Summary

IGluRs are an integral part of the enormously complex organ that is the mammalian brain. We are only beginning to understand the effects of long-term potentiation and synaptic plasticity and there are many years of research ahead. NMDA receptors are one of the most interesting sub-classes of iGluRs. They are known to be central in memory and learning and are considered an important target for neuro-degenerative diseases such as Alzheimer's disease. The many receptor subtypes are structurally and functionally complex and are affected by a large variety of endogenous ligands. The work on homology modelling and ligand-docking yields a consistent model of NMDA receptor action, and highlights a number of experimental validations, as well as a number of useful therapeutic developments [48]. It opens new opportunities in the field and highlights the usefulness of computational models in revealing protein function. The location of the polyamine binding site is a vital part of creating a logical and consistent model of the receptor and Oxdock produces an excellent answer that agrees with experimental findings. The use of molecular docking as a predictive tool is seen again in chapter four in the analysis of putative plant receptor proteins. In this case, the endogenous ligand is unknown and Oxdock makes a prediction that reveals a new exciting area of biology.

4 Investigating Plant Glutamate Receptors

4.1 Overview

The lack of a central nervous system is one of the many features that characterise plants. However, it has been known for many years that plants respond to external stimuli in a system-wide manner. Large-scale movements occur in growing plants to harvest the maximum amount of sunshine (phototropism) and the upward growth of hypocotyl shoots is a response to gravity (gravitropism). Despite this, the exact mechanisms and pathways that give rise to these effects have not been elucidated.

Using a multi-scale docking algorithm in combination with a molecular model of the ligand-binding domain of a putative plant iGluR, the ligand specificity of plant these receptors is investigated. The structure of a rat iGluR has been empirically determined by crystallography [22, 31]. A combination of homology modelling with Oxdock is used to explore the nature of the interaction of the plant iGluRs with glutamate and glycine. Before this work was completed, the ligand for these receptors was unknown, and indeed many plant biologists had suggested that these proteins were not receptors at all.

4.2 The Discovery of Plant iGluRs

In 1998, the genome of *Arabidopsis Thaliana* was searched for protein sequences that may code for Nitrogen sensing proteins [49]. Twenty sequences were discovered, and their marked sequence similarity to iGluR sequences in animals led to them being

termed plant glutamate receptors. The plant sequences show the greatest sequence similarity to the mammalian NMDA receptor subtypes [50-52]. Based on phylogenetic analysis of plant and animal sequences [4], the structure in Figure 3.1 is predicted for these plant receptors. Plant growth studies have suggested the involvement of iGluRs in hypocotyl elongation [53] and have predicted that glutamate-gated channels may be involved in regulating changes in the cytosolic concentration of Ca^{2+} [54]. However, interest has moved away from this area due to the apparent lack of response to the agonist glutamate.

4.3 Homology Modelling of Plant iGluRs

In an attempt to explore the function of the putative plant glutamate receptors, the structure of the extracellular ligand-binding region of an *Arabidopsis thaliana* iGluR subunit (AtGLR) is modelled on the crystal structure of the analogous region of the *Rattus Norvegicus* iGluR (PDB ID 1FTJ) [31]. Armstrong and Gouaux cleverly created the rat construct by fusing together the two portions of the glutamate-binding region (S1 and S2) with a linker sequence. The structure of the rat iGluR, and its interaction with its ligands was determined by crystallography, in addition to detailed functional analyses of the receptor-ligand interactions [22, 31]. Structural and sequence information agree that there are seven key residues in a receptor vital for interaction with the ligand, five of which are shown in Figure 4.1. The crystal structure of the *Rattus Norvegicus* glutamate receptor construct in complex with glutamate is from the Protein Data Bank.

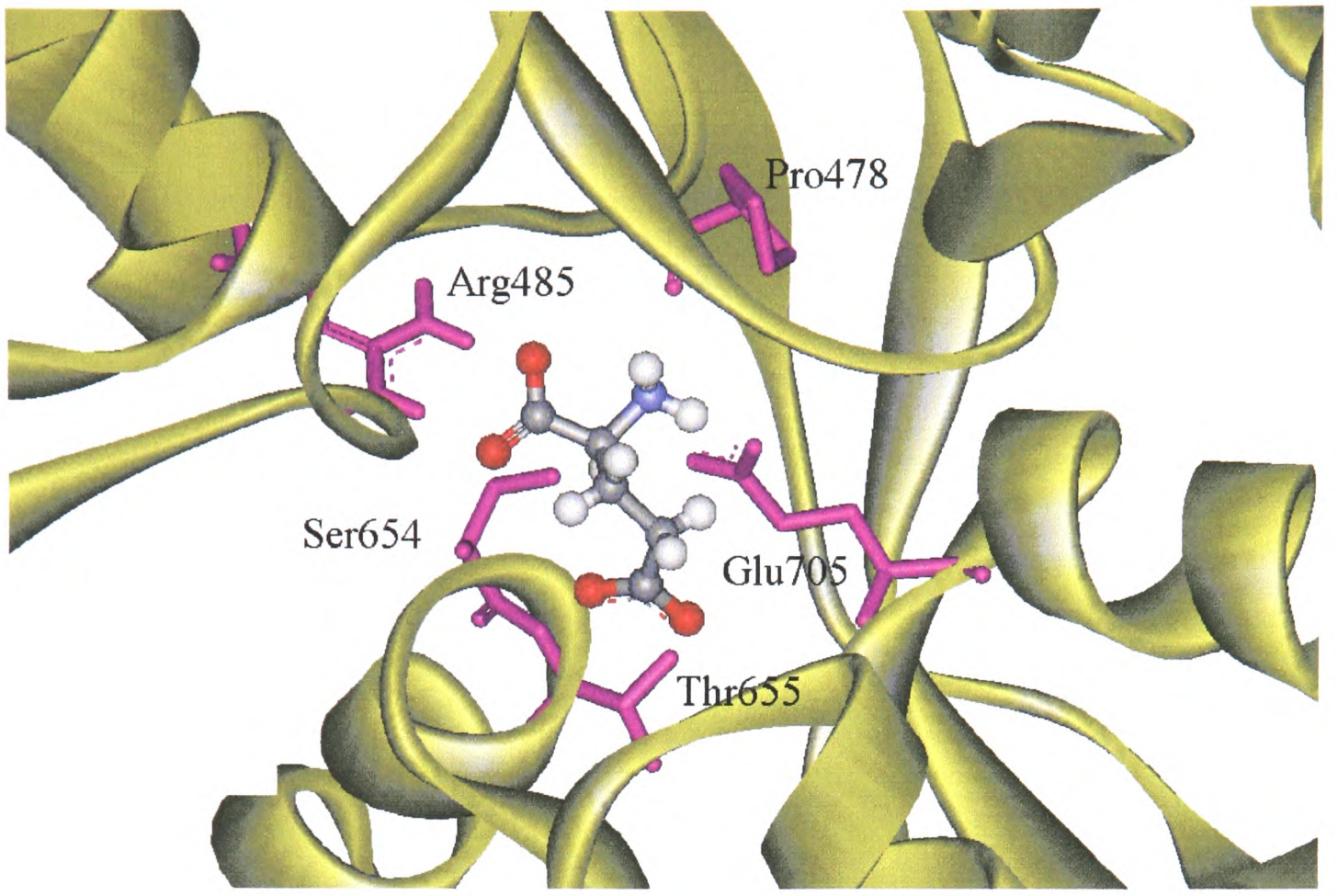


Figure 4.1 - Glutamate bound to the rat receptor construct (PDBID 1FTJ).

Tyrosine 450 (not shown) acts as a "lid" on the binding site but does not interact directly with the ligand. Proline 478, Threonine 480 (not shown) and Arginine 485 are residues in domain one that bind the backbone of glutamate. The Proline 478 backbone carbonyl and the Threonine 480 side chain hydroxyl group hydrogen bond to the backbone ammonium group of the ligand. The Arginine 485 side chain hydrogen bonds to the backbone carboxylate of the ligand. Serine 654, Threonine 655 and Glutamate 705 are residues from domain two, which also bind to glutamate. The Serine 654 hydroxyl group binds to the ligand backbone carboxylate and its amide hydrogen, together with the Threonine 480 hydroxyl group, binds to the terminal carboxylate group of the ligand. Finally, the Glutamate 705 side chain carboxylate

binds to the backbone ammonium group of the ligand. These seven amino acids thus bring about domain closure following glutamate binding because residues from both domains are needed to anchor the ligand. Sequence analysis shows that all NR2 subunits possess a conserved glutamate-binding GST motif, corresponding to Glycine 653, Serine 654 and Threonine 655 in the glutamate-binding site of the Rattus Norvegicus iGluR (See Figure 4.4). More recent work has shown that analogous residues are important for glycine binding in the case of glycine binding to an NMDA NR1 subunit [55]. This is shown in Figure 4.2.

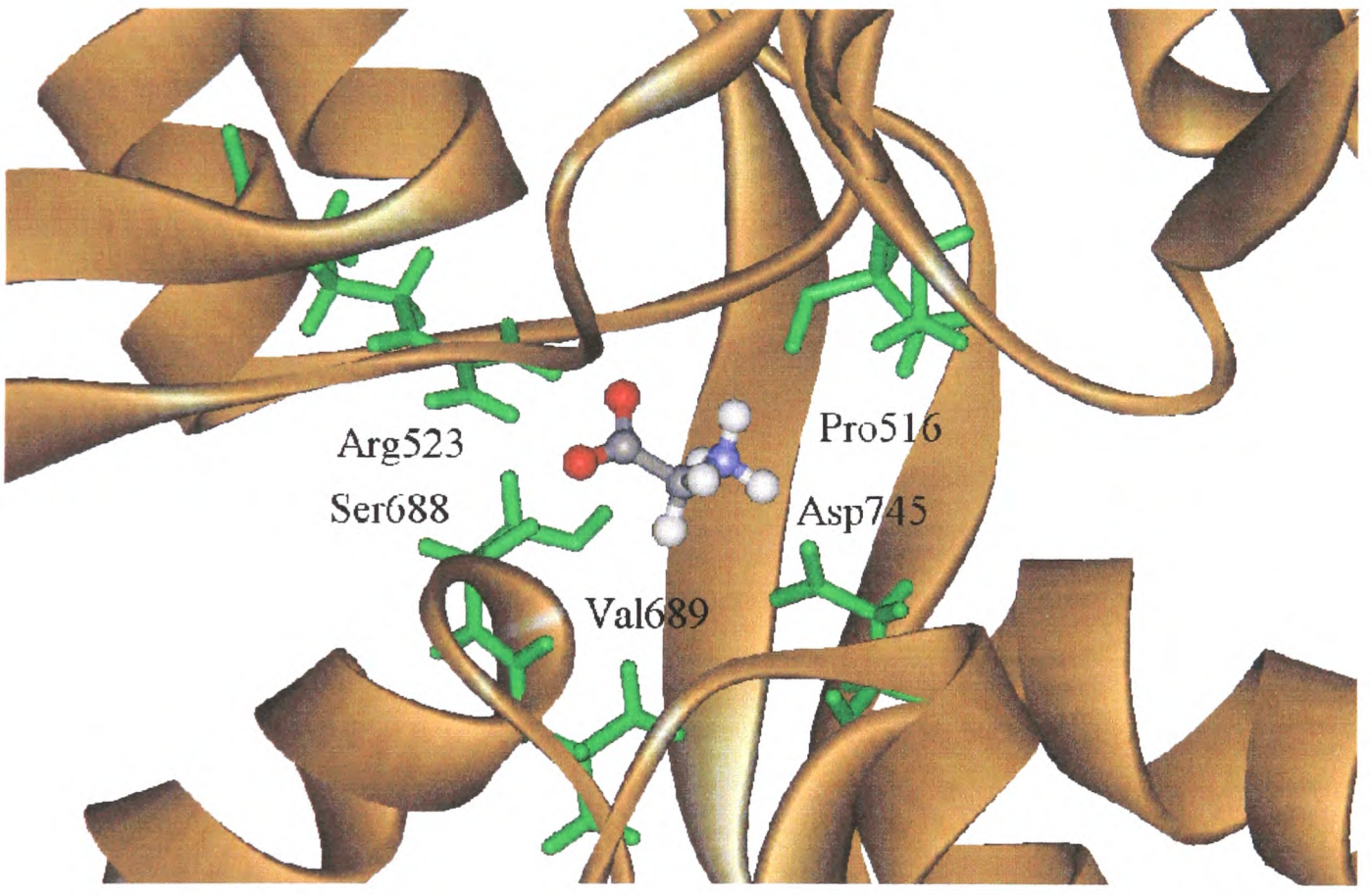


Figure 4.2 - Glycine bound to the rat receptor construct (PDBID 1PB7).

The construct is based on the receptor GlurB (accession number AAA41240) and also contains the two glutamate-binding domains (S1 and S2) joined by a linker region.

The NMDA Receptor NR1 construct in complex with glycine (PDB ID 1PB7) is also from the PDB. The *Arabidopsis thaliana* iGluR, AtGLR2.9, is chosen for the modelling study as it exhibited the greatest sequence similarity to the *Rattus Norvegicus* iGluR, particularly in those regions implicated in ligand binding. The rat glutamate receptor sequence and the plant AtGLR sequences are aligned using both Insight Homology and the Psipred program from the UCL bio-informatics unit [56]. Portions of the alignment are shown in Figure 4.3 and Figure 4.4.

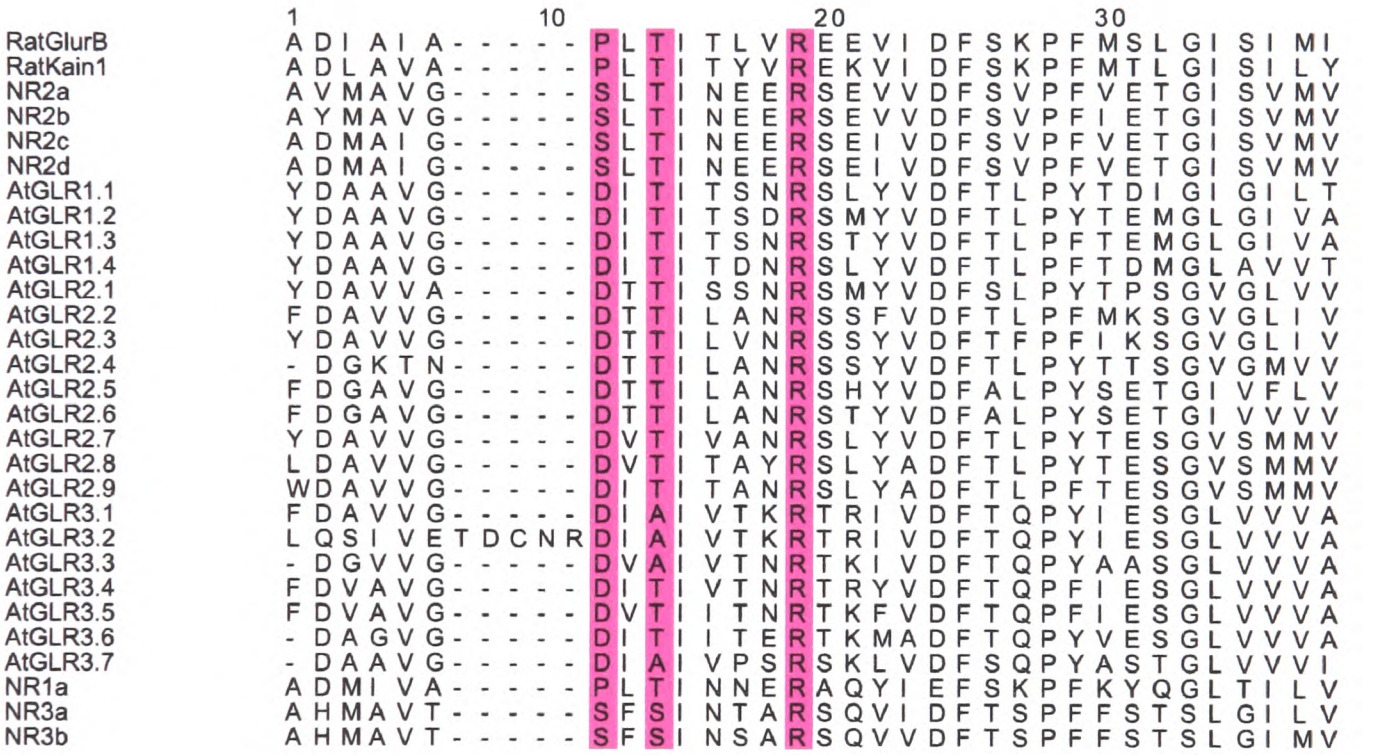


Figure 4.3 - Multiple sequence alignment of iGluR sequences. The peptide sequences (and their corresponding accession numbers) that are used in the alignment included: the *Rattus Norvegicus* iGluRs, RatGlurB (CAA38465), RatKain1 (AAA02873), NR2a (AAC03565), NR2b (AAA41714), NR2c (AAA41713), NR2d (AAC37647), NR1a (AAA16366),

and NR3b (AAL69893); the Homo sapiens iGluR, NR3a (AAL40734); and the Arabidopsis thaliana iGluRs, AtGLR1.1 (AAF26802.1), AtGLR1.2 (BAA96960.1), AtGLR1.3 (BAA96961.2), AtGLR1.4 (AAF02156.1), AtGLR2.1 (AAB61068.1), AtGLR2.2 (AAD26895.1), AtGLR2.3 (AAD26894.1), AtGLR2.4 (CAA19752.1), AtGLR2.5 (CAB96656.1), AtGLR2.6 (CAB96653.1), AtGLR2.7 (AAC33239.1), AtGLR2.8 (AAC33237.1), AtGLR2.9 (AAC33236.1), AtGLR3.1 (AAF63223.1), AtGLR3.2 (CAA18740.1), AtGLR3.3 (AAG51316.1), AtGLR3.4 (AAB71458.1), AtGLR3.5 (AAC69939.1), AtGLR3.6 (CAB63012.1), and AtGLR3.7 (AAC69938.1). Residues in domain one that are important for ligand binding are highlighted in pink.

The two portions of the sequence from AtGLR2.9 corresponding to the two glutamate-binding domains are removed from the plant sequence and joined together by the linker sequence. Examination of the sequence alignment showed that the majority of the secondary structure of the iGluR is retained in the plant sequence; the small insertions and deletions occurred only in loop regions on the surface of the protein. The homology package of InsightII is used to assign coordinates to the amino acid sequence of AtGLR2.9. The model obtained can then be refined by CHARMM to optimise the three-dimensional structure [6].

Analysis of the modelling results highlights residues implicated in glutamate binding in mammalian iGluRs that also appear important for glycine binding in the AtGLRs. For example, the Glutamate and Arginine residues in the binding site appear to be important for ligand binding in both the *Rattus Norvegicus* iGluR (Glutamate 705,

Arginine 485) and the AtGLR (Glutamate 724, Arginine529). Both are perfectly conserved over all receptors investigated (Figure 4.3). Similarly, other amino acids implicated in establishing the three-dimensional structure of the ligand-binding region of domain one, such as Threonine 480 and Isoleucine 481, are conserved between species (Figure 4.3). In contrast, the binding site Proline residue at position 478 in the *Rattus Norvegicus* iGluR is replaced by Aspartate 522 in AtGLR2.9, and appears to bind glycine via a side chain interaction. The replacement of Proline with Aspartate would anchor glycine more strongly within the binding site. This may be due to the importance of long-term effects of binding in plants compared to the swift recycling necessary in the mammalian nervous system.

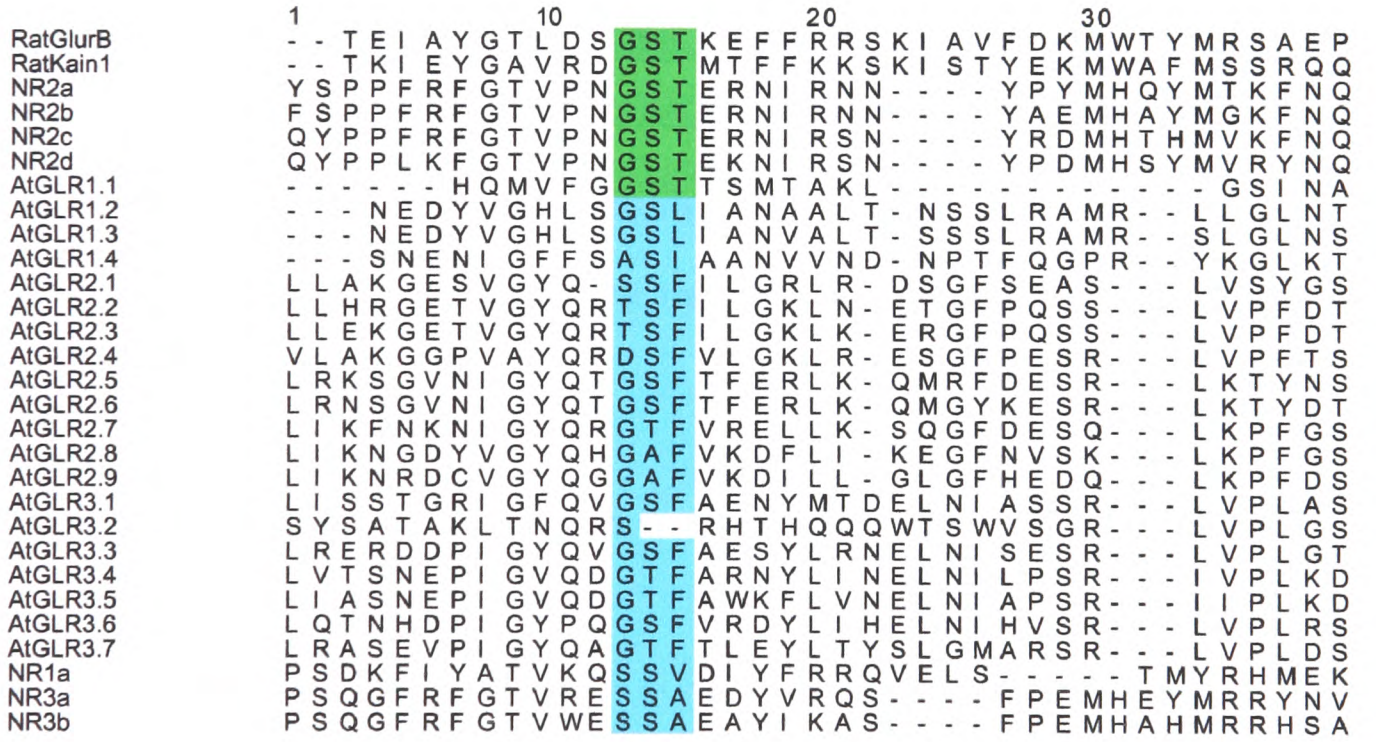


Figure 4.4 - Multiple sequence alignment of iGluR sequences. The peptide sequences (and their corresponding accession numbers) that can be found

in Figure 4.2. Residues in domain two that are important for ligand binding are highlighted in blue.

In domain two, Serine 654 and Threonine 655 of the *Rattus Norvegicus* iGluR are replaced by non-polar Alanine and Phenylalanine residues respectively in the AtGLR. This Alanine 681 residue in the AtGLR is predicted to bind glycine via its backbone nitrogen (as Serine 654 binds glutamate in mammalian iGluRs).

4.4 Ligand Docking with Plant iGluRs

The docking algorithm Oxdock can explore the nature of the interaction of the plant iGluRs with glutamate and glycine. As a validation, Oxdock initially predicts that glutamate binds to the *Rattus Norvegicus iGluR* at the same site as in the crystal structure. In contrast, attempts to dock glutamate to the AtGLR2.9 model failed. Furthermore, attempts to "force" glutamate into the predicted ligand-binding site *in silico* showed that this could not occur naturally, due to significant steric hindrance (see Figure 4.5).

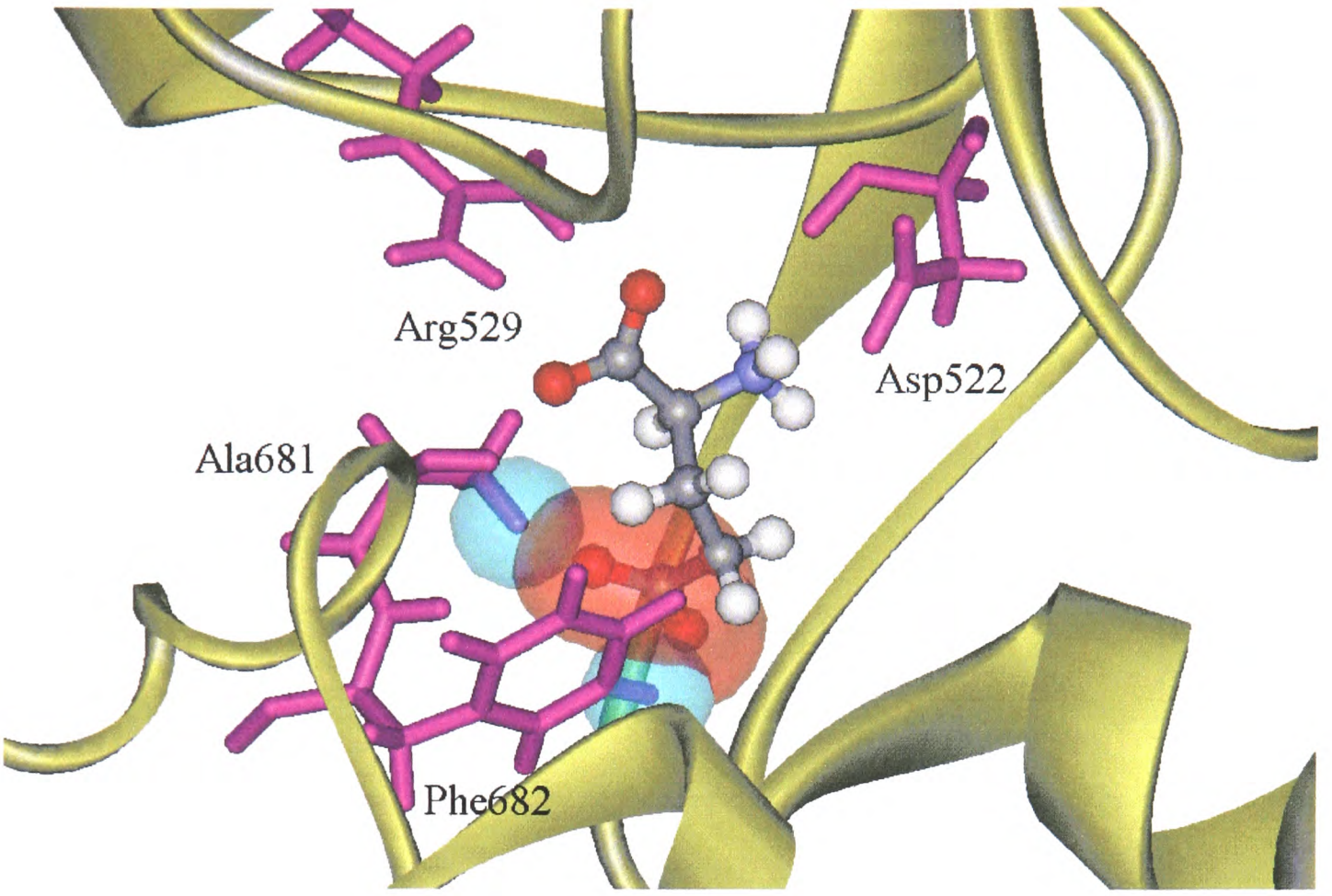


Figure 4.5 - Glutamate artificially placed in the binding site of the plant receptor. The steric hindrance is highlighted by overlapping blue and red spheres.

The steric hindrance is attributable to the replacement of one of the key residues that is required for binding of the side-chain carboxylate group of glutamate (see Figure 4.1). The residue in question, Threonine 655 of the *Rattus Norvegicus* iGluR, is conserved amongst all mammalian iGluRs studied (see Figure 4.4). In AtGLR2.9, Threonine 655 is replaced by a bulky, hydrophobic Phenylalanine residue, which blocks the ligand-binding pocket where glutamate would dock (see Figure 4.5). Threonine 655 is replaced either by Phenylalanine, or by the bulky, hydrophobic branch-chain amino acids Leucine or Isoleucine, in 18 of the 20 AtGLR subunits, and is missing altogether from one other (see Figure 4.2). This suggests that glutamate is not the natural ligand for the majority of AtGLR subunits.

As detailed in Section 3.1, there are two types of mammalian NMDA receptor subunits: glutamate receptors and glycine receptors [57]. Glutamate is bound by NR2 subunits, whereas, glycine is bound by NR1 and NR3 subunits. All NR2 subunits possess a conserved glutamate-binding GST motif, corresponding to Glycine 653, Serine 654 and Threonine 655 in the glutamate-binding site of the *Rattus Norvegicus* iGluR. In contrast, most of the AtGLR sequences are similar to the glycine-binding subunits NR1 and NR3, where Threonine 655 is replaced by amino acids with large, hydrophobic side chains. To determine if glycine could function as a ligand for plant iGluR subunits, glycine is docked to the AtGLR2.9 model. The binding site was defined in a sphere of radius 10 Å centred on the docked glutamate molecule from 1FTJ, as a functioning plant receptor would necessarily have a binding site between the two lobes. This calculation took approximately 30 minutes (on a 766MHz processor) due to the relatively small search space and the small ligand with only one torsion. Glycine preferentially docks with the proposed glutamate-binding site (Figure 4.6 and Figure 4.7).

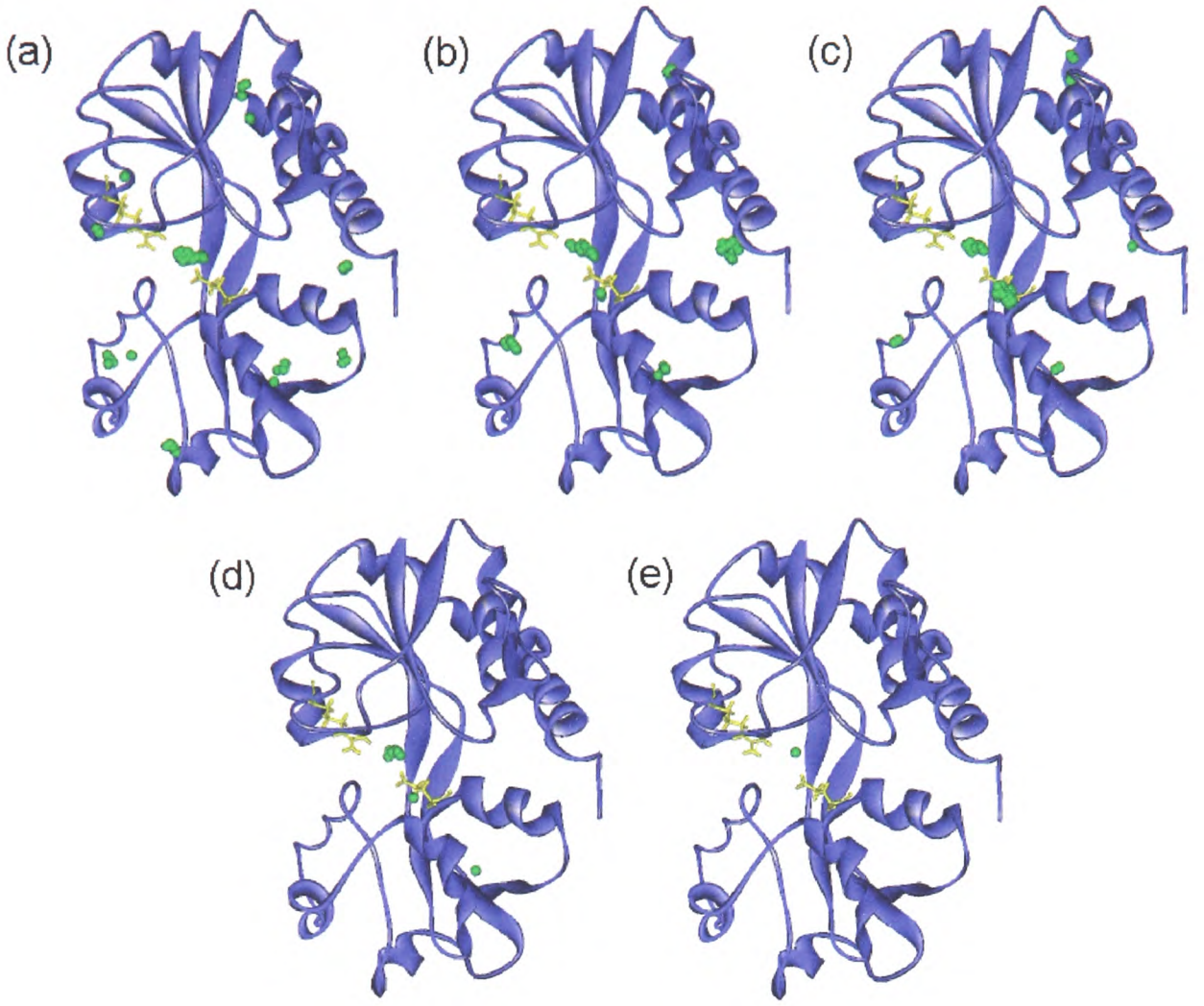


Figure 4.6 - Oxdock docking results for one (a), two (b), three(c), four (d) and five (e) feature-points. Initially, many good solutions are found but the algorithm soon converges to find one single point, in the equivalent of the Glutamate-binding site.

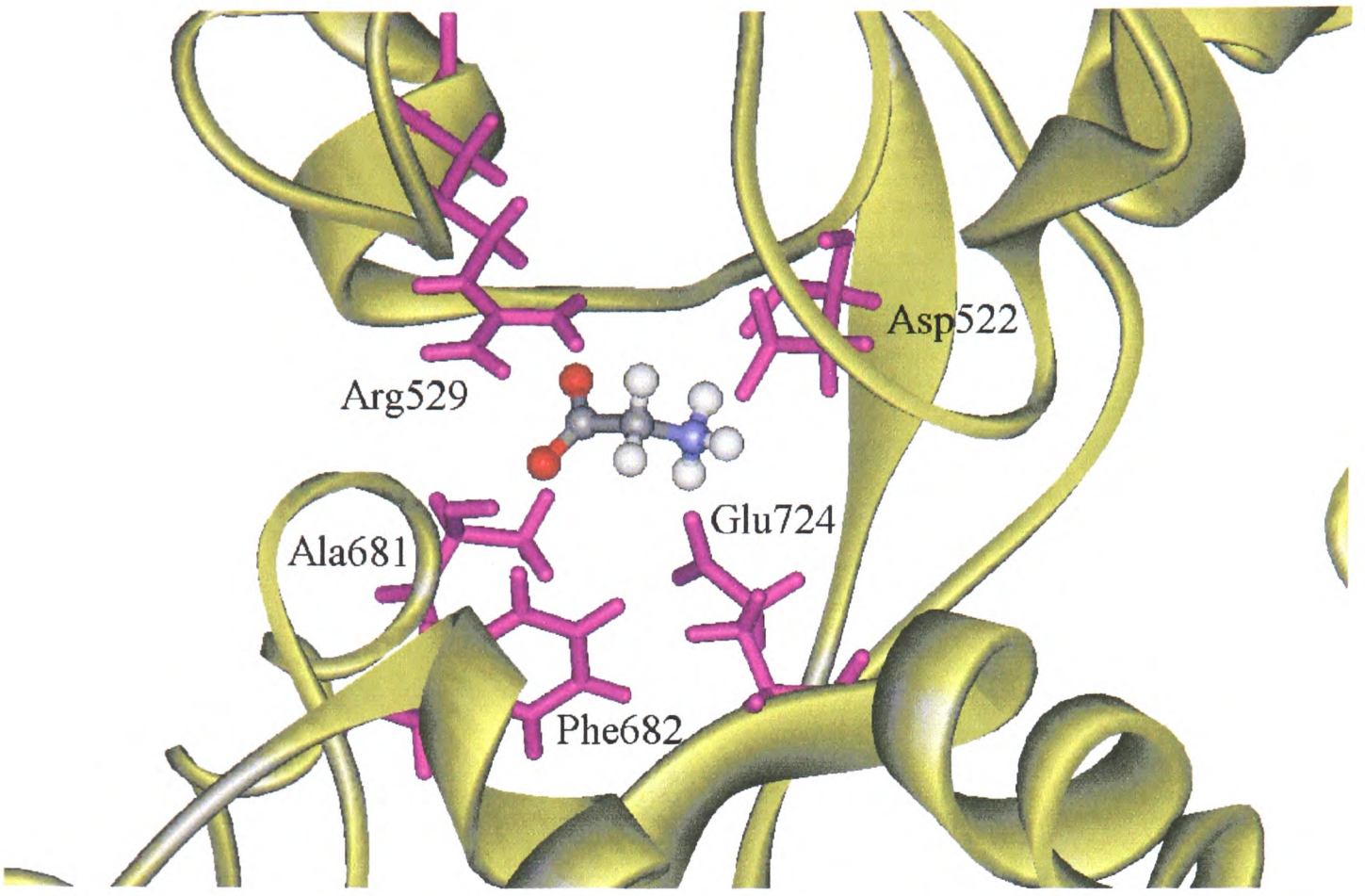


Figure 4.7 - Glycine bound to the plant construct AtGLR2.9 after refinement with CHARMM. The important residues are labelled.

The data strongly suggest that glycine is the natural agonist for the majority of AtGLRs. This previously unknown result makes an exciting contribution to plant science. Targeting these receptors may allow the stimulation or retardation of plant growth. The next stage was to consider the possibility that both glycine and glutamate bind to these receptors, in a similar way to the mammalian NMDA receptors. There is one subunit which contains the glutamate-binding GST motif in the correct region of domain 2 (Figure 4.4) for glutamate-binding.and this is AtGLR1.1, When the ligand-binding domain of AtGLR1.1 is modelled on the *Rattus Norvegicus* receptor, glutamate bound to the same residues as in the mammalian iGluRs (Figure 4.8).

Again, the binding site was defined in a sphere of radius 10 Å. The calculation took approximately 1 hour (on a 766MHz processor).

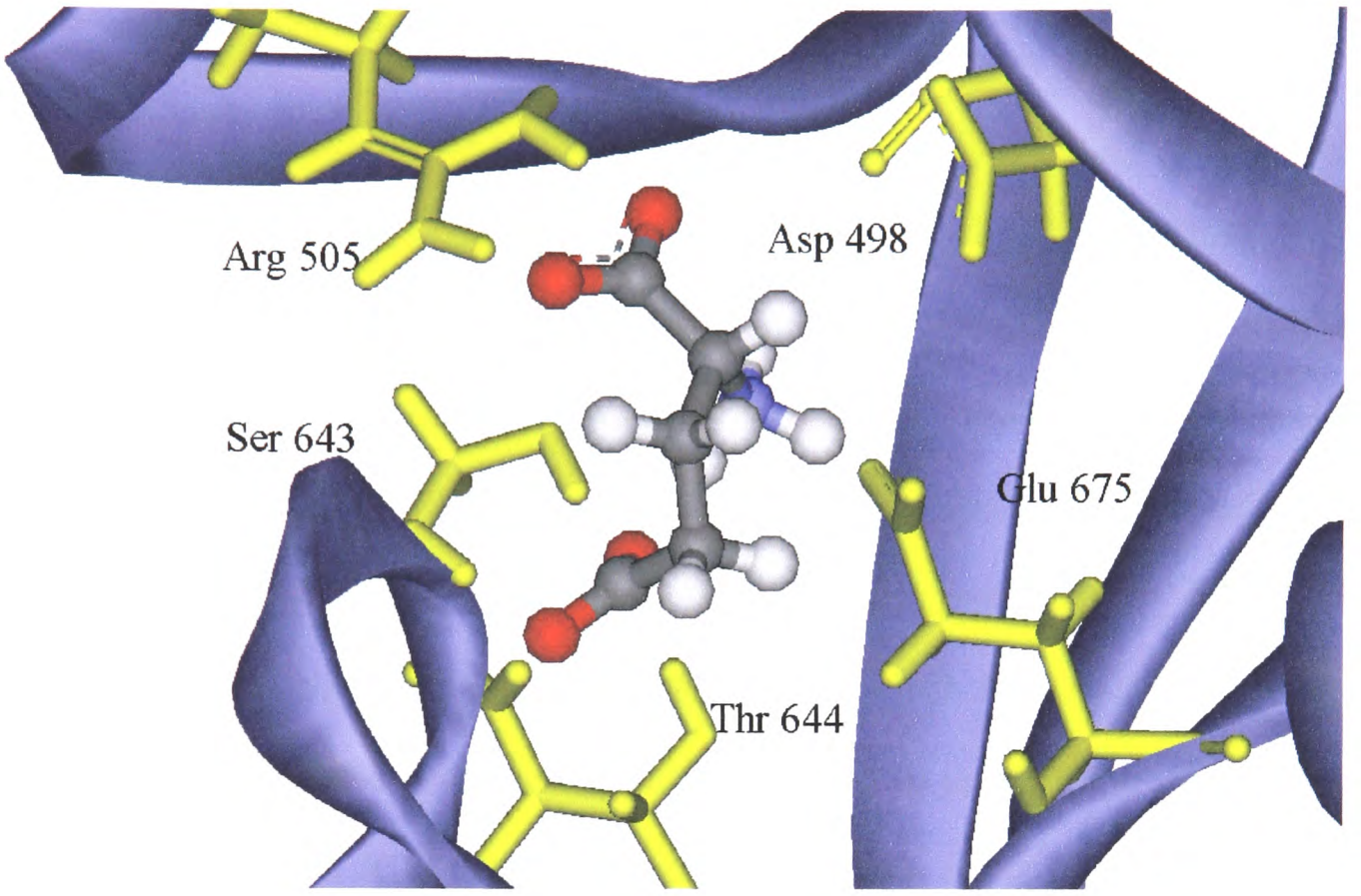


Figure 4.8 - Glutamate bound to the plant construct AtGLR1.1 after refinement with CHARMM. The important residues are labelled.

Thus, while the majority of AtGLR subunits should bind glycine, it seems likely that plant iGluR channels may function with both glycine and glutamate or simply with glycine alone. This prediction has been tested in plant growth studies performed by Campbell and Dubos [58]. The combination of glutamate and glycine increased the levels of cytosolic calcium ions at concentrations equivalent to two orders of magnitude less than those required by these compounds alone. This proves that glycine and glutamate have a synergistic effect on plants and strongly supports the

hypothesis that plant glutamate receptors are activated by these molecules together. Again, this was a completely unknown result and an exciting development in plant biology.

In order to test the hypothesis that glutamate and glycine are eliciting their gating effect at iGluRs, plants have previously been pre-treated with 6,7-dinitroquinoxaline-2,3 dione (DNQX), a compound that inhibits mammalian iGluRs, prior to treatment with the agonists. DNQX treatment completely abolished the increase in calcium ions normally induced by the addition of glutamate, glycine, or both in combination [49]. The ability of DNQX to block the effect of these ligands suggests that they bind to the same receptor as DNQX, an ionotropic iGluR. Modelling with AtGLR1.1 showed that DNQX bound in the predicted ligand-binding pocket. This predicts that DNQX will compete for the plant iGluR ligand-binding pocket with iGluR agonists, as with animal iGluRs. These findings are also consistent with the observation that glutamate and glycine compete with DNQX in the gating of calcium ions and in the regulation of hypocotyl elongation, as found by Campbell and Dubos.

4.5 The Function of Plant iGluRs

Mammalian iGluRs, such as NMDA receptors, function as ligand-gated channels for the transport of ions such as Ca^{2+} or Na^{+} [23]. Plant iGluRs have also been predicted to function as Ca^{2+} channels that are gated by glutamate [50, 59]. There is further evidence suggesting that genes encoding iGluR subunits might be involved in regulating cellular and developmental phenomena in plants. The effect of the iGluR antagonist DNQX on Arabidopsis seedlings has suggested the involvement of iGluRs

in the control of hypocotyl elongation [49]. Treatment of plants with a putative iGluR agonist also suggested an involvement of iGluRs in hypocotyl elongation [53]. Unrelated studies predicted that glutamate-gated channels may be involved in regulating changes in the cytosolic concentration of Ca^{2+} [54]. Over expression of one member of the iGluR family in Arabidopsis implicated at least one iGluR subunit in the control of calcium homeostasis and related developmental phenomena [60].

Glycine is known to be an important modulator of cell-to-cell signalling in the mammalian central nervous system. In addition to the role that glycine plays in stimulating NMDA receptors to initiate calcium signalling, glycine stimulation of NMDA receptors also primes the receptors for clathrin-mediated endocytosis [61]. Thus, glycine functions to modulate not only the immediate response of neurons to neurotransmitters, but it also conditions future responsiveness by altering the occurrence of receptors on the cell surface. Recognition of the roles played by glycine has profoundly changed the understanding of synaptic signalling in the central nervous system [61]. Understanding how glycine and glutamate act synergistically to regulate plant growth may have a similarly profound effect on plant biology.

4.6 The ATD of Plant iGluRs

It is interesting to examine the corresponding portions of the sequences for the ATD of plant receptors. The portions of the sequence corresponding to the two helices implicated in the formation of a four-helix bundle do show a marked similarity within an AtGLR family but vary between families (Figure 4.9). This could allow specificity in the assembly of subunits due to complementarity of the four helices. There are also

a number of conserved Tryptophan residues, which appear in the AtGLR, NMDA and mGluR1 sequences. In the mGluR1 structure, they appear to stabilise the fold of the protein by forming a portion of the hydrophobic core. This suggests that all three protein families will have a similar fold, as does their evolutionary relationship. The Cysteine residue that appears to link the subunits in the case of NMDA receptors is not present. However, this Cysteine is absent in the mGluR1 structure and thus does not appear to be essential.

The site implicated in Zinc binding of the NR2A subunit does not appear to be present in any of the plant receptors. However, there is no evidence that Zinc ions affect these receptors and thus the lack of a binding site is not strange. Analysis of the portion of the sequence corresponding to the proposed spermine-binding site yields interesting results. There are a large number of acidic residues in the majority of the subunits, which would in general favour the binding of the positively charged spermine (Figure 4.9).

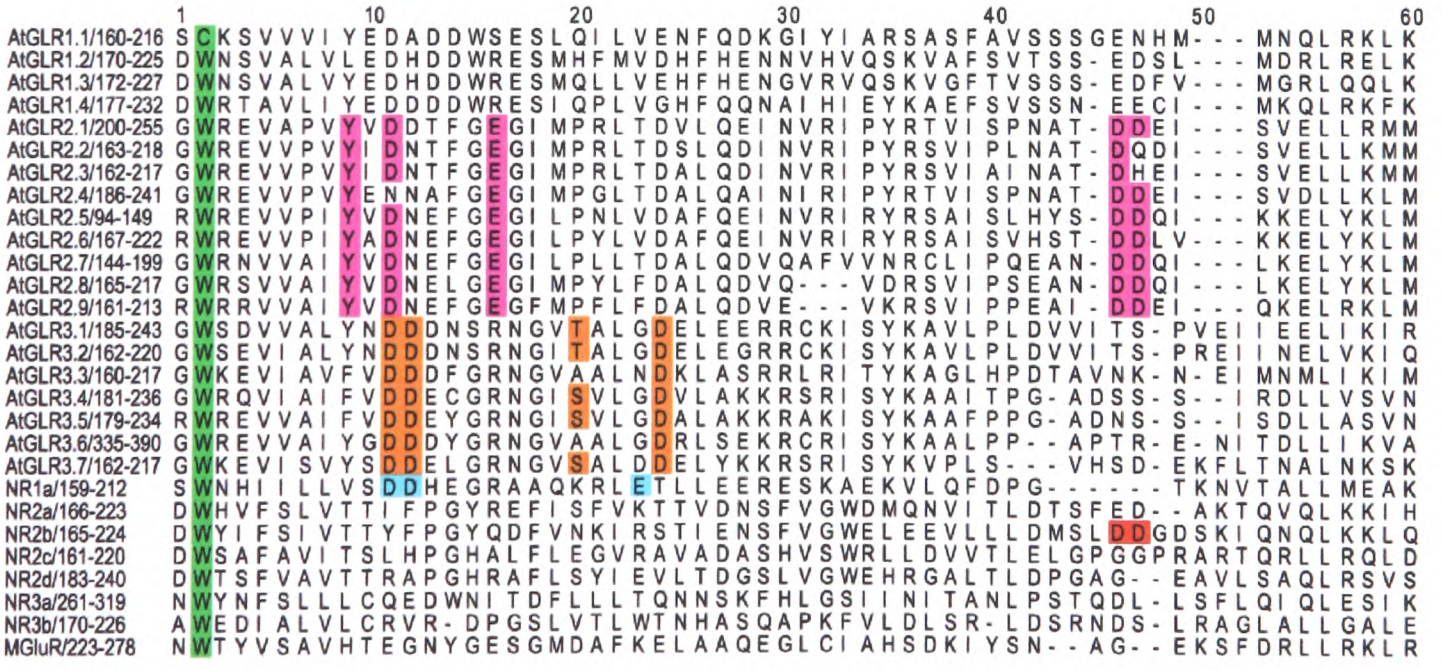


Figure 4.9 - Multiple sequence alignment of the ATD for the plant AtGLR subunits with the Rat NMDA subunits and mGluR1 sequences. The acidic residues implicated by mutagenesis in spermine stimulation are coloured blue in the NR1a subunit and coloured red in the NR2b subunit. The residues in the plant sequences that are suggested to be important by molecular docking are coloured in pink (AtGLR2 family) and orange (AtGLR3 family). One of the almost perfectly conserved Tryptophan residues that may help to stabilise secondary structure is coloured in green.

However, the key factor is the presence of two pairs of conserved aspartate residues, one for the AtGLR3 family that corresponds to the NR1a aspartate residues implicated by mutagenesis data and one for the AtGLR2 family that corresponds to the NR2a aspartate residues implicated by modelling of the dimer. Other residues from NR1a and NR2a are conserved in some but not for all of a given family of AtGLRs, perhaps suggesting that some bind spermine more strongly, or that some do not bind at all. A homology model of a plant ATD dimer is created from the sequences of AtGLR2.1 and AtGLR3.1. These are chosen as they come from different subunit families (so may thus form a functioning dimer) and contain few major deletions (so are more accurately modelled). This dimer is created using the InsightII homology package and then optimised by CHARMM. The potential docking of spermine can then be investigated using Oxdock. Spermine is found to dock at the same inter-subunit boundary as for the NMDA dimer. The result is shown in Figure 4.10.

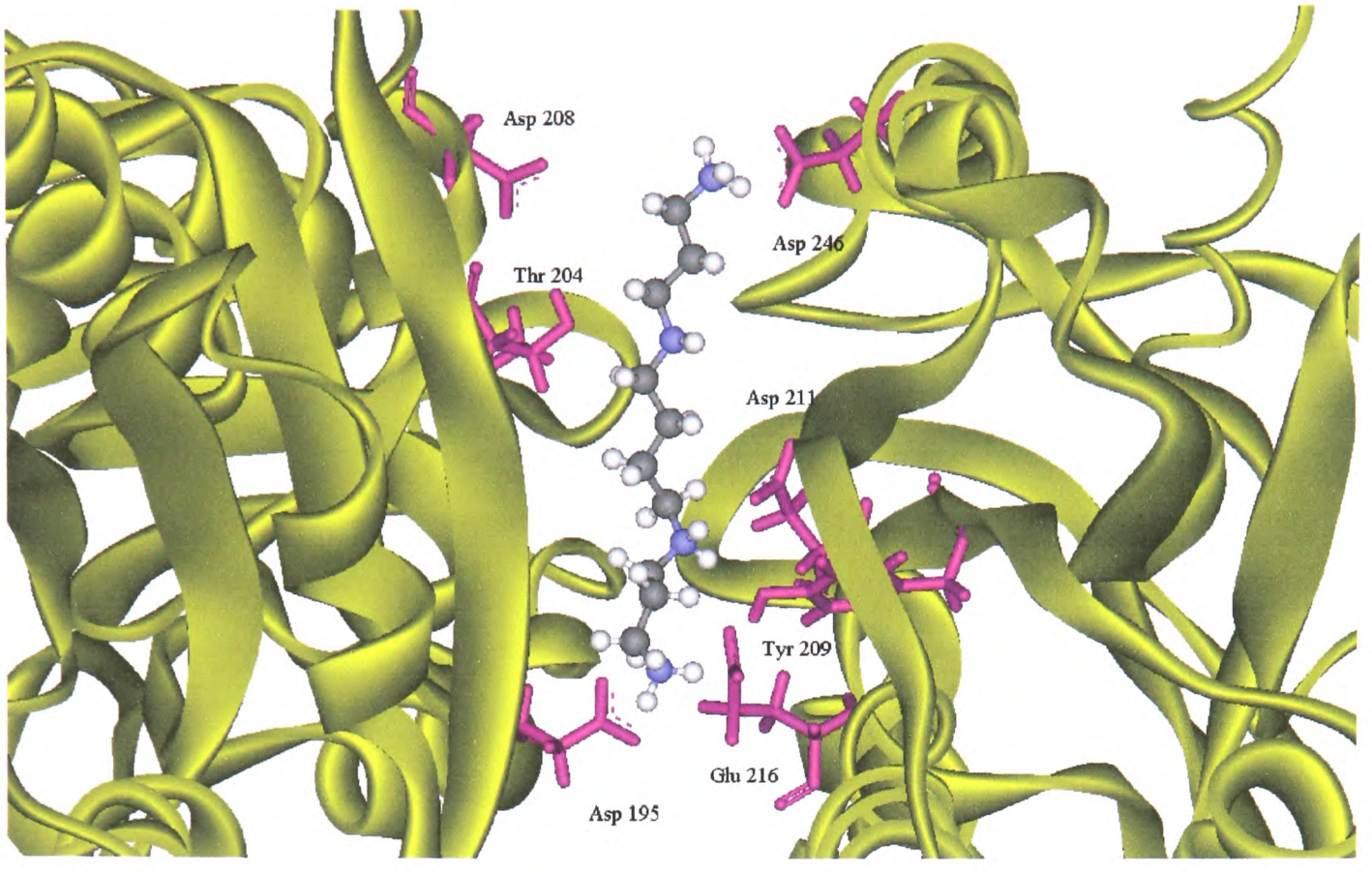


Figure 4.10 - Spermine docked with the AtGLR2.1/AtGLR3.1 dimer. The proximal residues are coloured pink and labelled with name and number.

This modelling work is again substantiated by experimental work by Dubos and Campbell. Plant growth studies initially showed that the aminoglycoside kanamycin affects the growth of plants and that this effect may be dependent on the pH within the cells. The sequence analysis along with the modelling and molecular docking work provided a basis for this finding. The effect can be attributed to the similarity between aminoglycosides and polyamines, as for NMDA receptors. Further work has highlighted the importance of both aminoglycosides and polyamines in regulating plant growth. Thus, spermine and related polyamines may have a similar role in both plant and animal receptors.

4.7 Summary

Data from the middle of the last century suggested that glycine might function as a signalling molecule in plants. In 1939, White found that glycine favoured the growth of tomato roots, in the presence of appropriate combinations of nutrients [62]. Further work, based on White's early observations, suggested that glycine modulated root development (in concert with other amino acids) in a synergistic fashion [63, 64]. However, glycine was thought to function purely as a nutrient, and the potential involvement of glycine as a signal in plants was not considered. Given the recent discovery of NMDA-like iGluRs in plants, and given the role that glycine plays in gating these receptors in animals, glycine signalling may also be important in plants. The interaction of glutamate and glycine with plant iGluRs is thus examined by Oxdock, as well as the glutamate receptor antagonist DNQX.

Taken together, the computational and experimental findings uncover a role for glycine signalling in plants, and show that the synergistic action of glutamate and glycine at NMDA-like receptors predates the divergence of plants and animals. This important result has implications in both evolutionary theory and in the understanding of plant biology. It may also lead directly to the production of new herbicides and plant additives. The work also illustrates how computational modelling, in particular molecular docking, can be used to investigate genuine biological problems and enrich complex topics. However, if the knowledge that is gained by these techniques is to be used practically, new and improved computational methods must be developed.

5 Validation of Oxdock

5.1 Overview

Oxdock has proved to be a very useful tool for molecular docking. In cases when the protein structure and the ligand are known, it is both swift and efficient at prediction of the binding site. It is also able to aid in elucidating what is the natural ligand in cases where this is unknown. The multiscale approach allows poor solutions to be evicted from the set of likely solutions, and increasing the number of feature-points in each iteration allows good solutions to prosper.

Unfortunately, there are cases where Oxdock performs poorly. This is particularly true when looking at very large or very long ligands. When considering these types of ligand molecules, the multiscale approach is an inadequate representation of the entire ligand and the low feature-point representations may preclude finding the correct solution. The grid based approach to energy calculations can also lead to errors. It is assumed that the electrostatic potential and van der Waals effect at any point in space is equal to that at the closest grid point. This is a large approximation and a cause of inaccuracies. Finally, the combination of a multiscale approach and a grid based energy function leads to a poor optimisation method. This in turn (combined with the inexact energy function) means that the calculated binding energies are both inaccurate and non-comparable. These issues are problematic when attempting to calculate an accurate docking pose or compare various ligands against a protein target. In this chapter, these problems are investigated using specific examples.

5.2 Problems with the Multiscale Approach

Despite a major increase in speed, the multiscale approach can cause a number of problems in molecular docking. The swiftness is a major advantage but the reduction in accuracy can lead to either false positives or false negatives.

5.2.1 Long Ligands

One of the problems with Oxdock is the relatively poor performance when considering long (especially long and thin) ligands. This is an unfortunate result of using a multiscale approach and is due to the increasingly inaccurate low feature-point representations of ligands as they grow in size. There is thus a greater probability that a grid point representing a good solution will be removed from the process at an early stage. Problems also arise, as the ends of the ligand move to a much greater degree for long ligands. The effect of these problems can be seen in the case of the complex from PDB ID 1AAQ, which contains a Hydroxy-Ethylene Peptide Analogue, bound to HIV Protease. The parameters are identical to those in Table 2.1 and the results are shown in Figure 5.1.

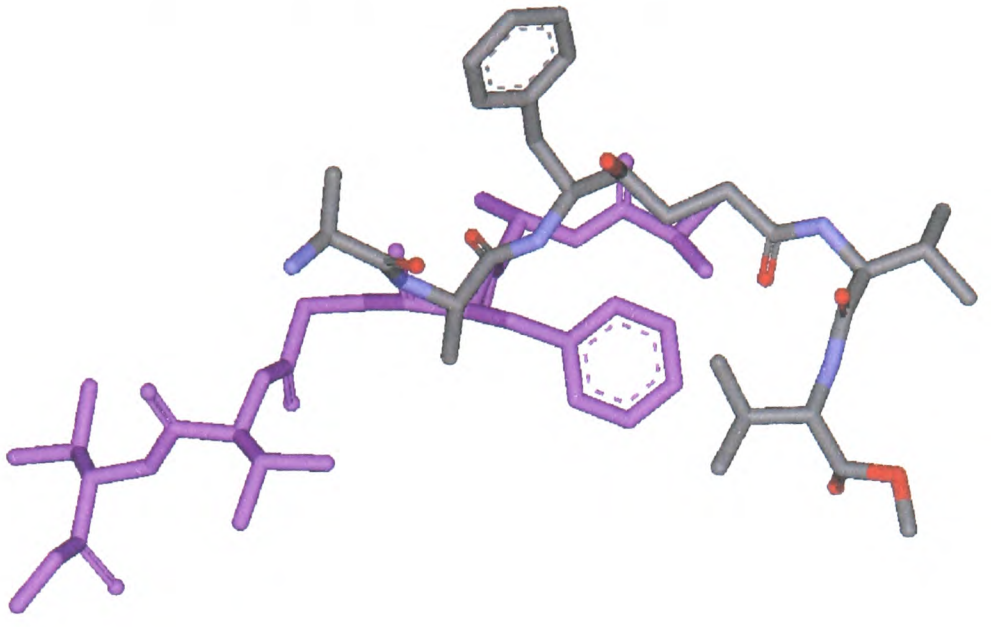


Figure 5.1 - The best solution for the complex 1AAQ between HIV Protease and a Hydroxy-Ethylene Peptide Analogue. The crystal structure is coloured lilac and the calculated position is atom coloured.

The algorithm converges at the five feature-point representation. Analysis of all the results shows that grid points around the ligand centroid are calculated as favourable in early stages of the docking process but are discarded at higher-level representations. Crucially, the algorithm finds the optimal docking positions for the feature-points and not the ligands. This leads to unrealistic solutions. Figure 5.2 helps to illustrate this point.

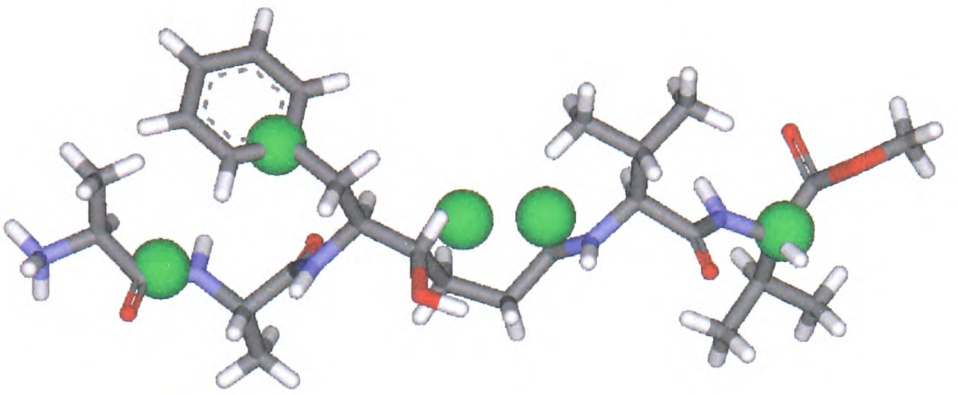


Figure 5.2 - The Hydroxy-Ethylene Peptide Analogue from PDB file 1AAQ showing the five feature-point representation as a series of green spheres.

Both bulky isopropyl groups and the benzene ring are not represented by these feature-points and thus poor solutions, in which large steric clashes occur, can prosper at the expense of good solutions. These groups are only well represented at the eight feature-point representation, as shown in Figure 5.3.

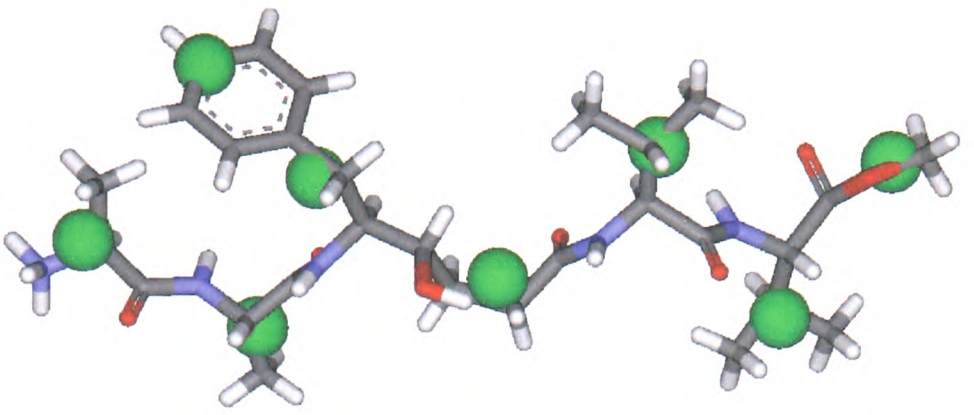


Figure 5.3 - The Hydroxy-Ethylene Peptide Analogue from PDB file 1AAQ showing the eight feature-point representation as a series of green spheres.

This example reveals that it is necessary to consider high-level representations of the ligand in some complex cases to ensure that the correct docking position is found. The multiscale approach is a powerful tool, but one that must be correctly and carefully used.

5.2.2 Wayward Feature-points

One of the more important issues when docking long ligands is that feature-points can be found away from the body of the ligand. Consider the case of 15-keto-prostaglandin F2α in Figure 5.4.

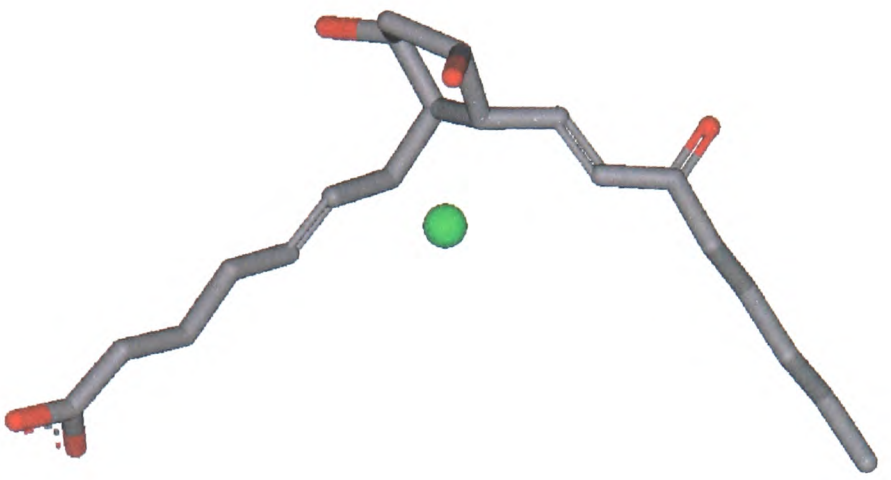


Figure 5.4 - The molecule 15-keto-prostaglandin F2a with the single feature-point represented as a green sphere. Only heavy atoms are shown.

The single feature-point is found at around 2.0 Å from the nearest heavy atom. In a tight binding complex, this may be so close to the protein structure that the high-energy proximal grid points are evicted. The optimal solution will then never be found. This is unacceptable.

5.3 Problems with Docking

5.3.1 Non-Specific Binding

The problem of non-specific binding is one that affects many if not all docking algorithms. For a given protein, there are generally many millions of compounds that will fit within the active site, without significant clashing. There will be thousands that have the potential to form strong van der Waals bonding interactions with the ligand. There are likely to be hundreds of highly charged ligands that will bind to an active site of predominantly the opposite charge. When Oxdock processes a protein-

ligand pair, it will attempt to find the protein site and ligand orientation that maximises the interaction energy. However, it will always find an answer, even if the answer is relatively poor. Consider the case of polyamine-binding protein from the PDB complex 1POT. The results of docking with the actual ligand (spermidine) and a hydrophobic analogue (decane) are shown in Figure 5.5.

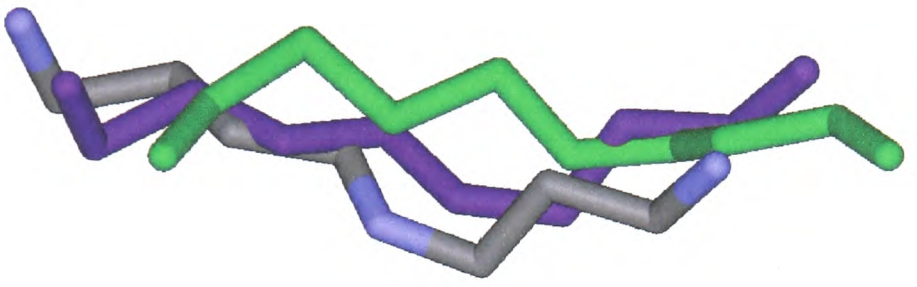


Figure 5.5 - The results from docking the natural ligand spermidine (purple) and a hydrophobic analogue decane (green) with the polyamine-binding protein from PDB 1POT. The crystal structure is atom coloured.

Decane is obviously unlikely to be found in a cell but the importance is the fact that it is predicted to bind at the active site of polyamine-binding protein. The results show that this is due to the favourable van der Waals interactions. Knowledge of chemistry suggests that decane will probably bind to any hydrophobic cavity and is thus a non-specific binder. It is thus important to be able to accurately score a given ligand and eliminate weakly binding and non-specific inhibitors.

5.3.2 False Positives

Another problem that blights docking algorithms is the calculation of false positives: molecules that are predicted to bind strongly to proteins but do not show an *in vivo* effect. Ignoring drug delivery and pharmacokinetic problems, one of the main causes is an oversimplification in the process of calculating the interaction energy. In Oxdock, this is caused by the multiscale approach. Consider again the example used above with spermidine-binding protein from the PDB file 1POT. If the molecule is lengthened to contain two more amino groups, as in Figure 5.6, the electrostatic binding energy increases, but the ligand becomes too long and can no longer fit within the binding site.

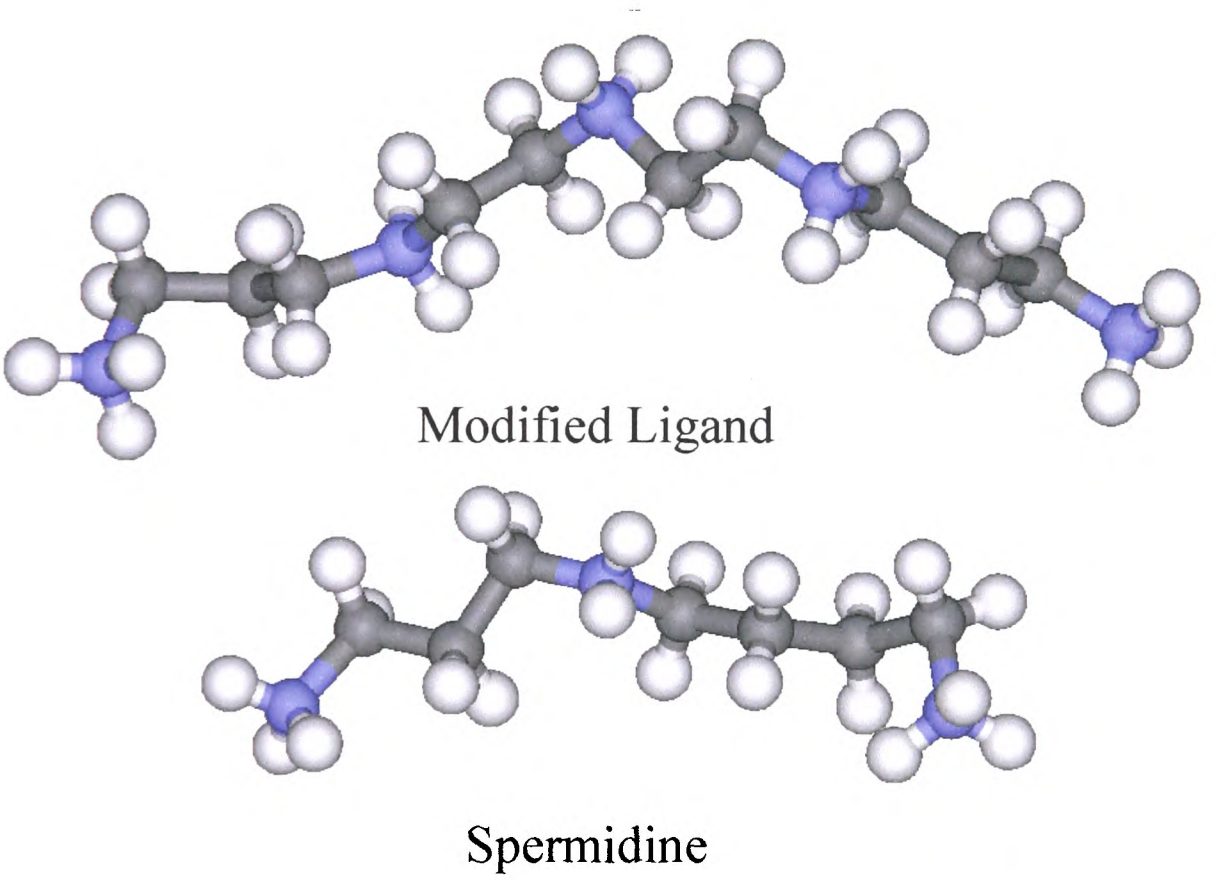


Figure 5.6 - The structures of the two ligands tested with Oxdock to find the optimum binding pose. The modified ligand contains two more amino groups than spermidine.

Oxdock predicts that both ligands will bind within the spermidine-binding site (correctly for spermidine). The predicted binding energies for the final iteration of Oxdock are shown in Table 5.1.

Ligand	Binding Energy (kcal/mol)
Spermidine	-71.03
Modified Spermidine	-91.83

Table 5.1 - The Oxdock calculated binding energy of two molecules with the protein from file PDB 1POT

However, analysis of the results shows significant steric clashing between the modified ligand and the protein at both ends of the binding cavity. Indeed, it is impossible to incorporate such a long ligand into the binding site. Furthermore, the highly charged molecule will interact more favourably with the solvent and thus will lose significantly more energy on desolvation. This effect is discussed in section 6.6.2.4 and would have a marked effect in this case. These findings provide another warning against direct use of results from a multiscale calculation and demonstrate that higher-level calculations are vital in the calculation of binding interaction energies.

5.4 Problems with Grid Calculations

As explained in section 2.5, Oxdock calculates the binding energy of a protein-ligand complex using a grid method. This has the advantage of markedly increasing the speed of docking. However, it is important that the grid has a resolution that is low enough to provide accurate estimates of the energy. Oxdock is thus tested in the following manner. Ten proteins are randomly selected from the protein data bank. For each protein, thirty grids are created, with the centre of each grid lying at the centroid of the ligand active site. Each of the thirty grids has a slightly differing resolution (and is thus a different size), decreasing from 1.5 Å to 0.05 Å in steps of 0.05 Å. Each grid contains 8,000 (20x20x20) grid points. The electrostatic potential and van der Waals potential due to the protein are calculated at each grid point. They are also calculated at the six points that lie at one resolution distance along each axis (positive and negative) and at the six points lying halfway between each of these points and the actual grid point (these sets of points are termed the next points and the midway points respectively). This is illustrated in Figure 5.7.

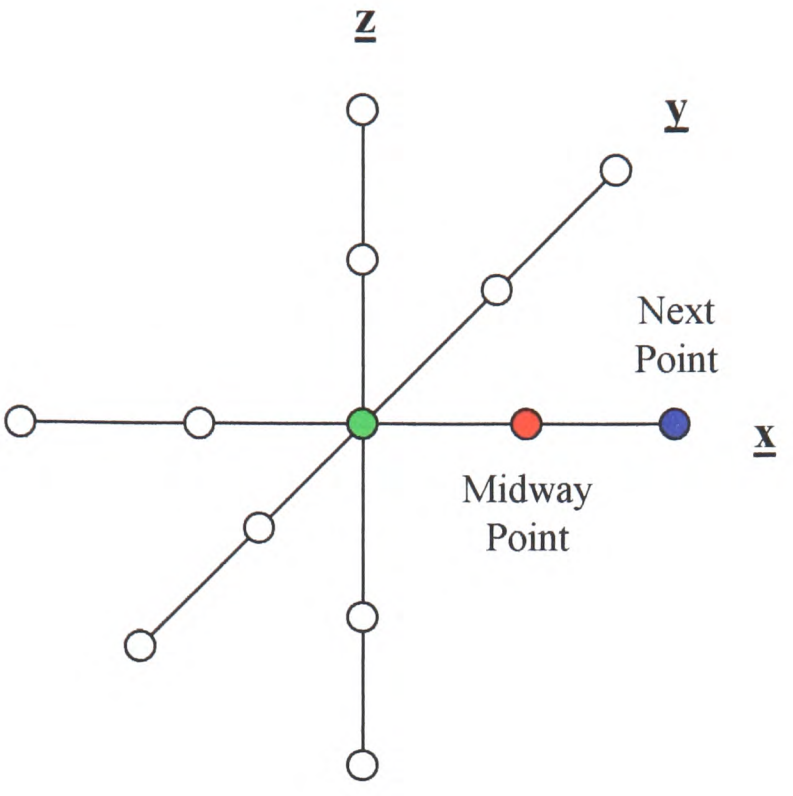


Figure 5.7 - Illustration of the points used in the calculation of the grid-based energy error calculation. The grid point is coloured green, the next point on the x-axis is coloured blue and the midway point between them is coloured red.

The electrostatic and van der Waals energies are calculated by considering a carbon atom with a unit charge placed at every single point (13 points for every actual grid point). The average energy is defined as the average of the grid point energy and the next point energy (this is considered as a prediction of the midway point energy). The error in each case is defined as the deviation between the average energy and the midway point energy. For a 20 x 20 x 20 grid, there are many errors in total (48,000) and thus the average of the errors is calculated and reported for each grid. Two further stipulations are made. Grid points that are further than 20.0 Å from the protein are ignored, as are any grid points that are nearer than 2.0 Å. The first rule ensured that

points are not considered at which the protein had no effect and the second ensured that points are not considered that are within the van der Waals radius of an atom.

5.4.1 Calculation of Electrostatic Interactions

The results for the electrostatic energy can be seen in Figure 5.8.

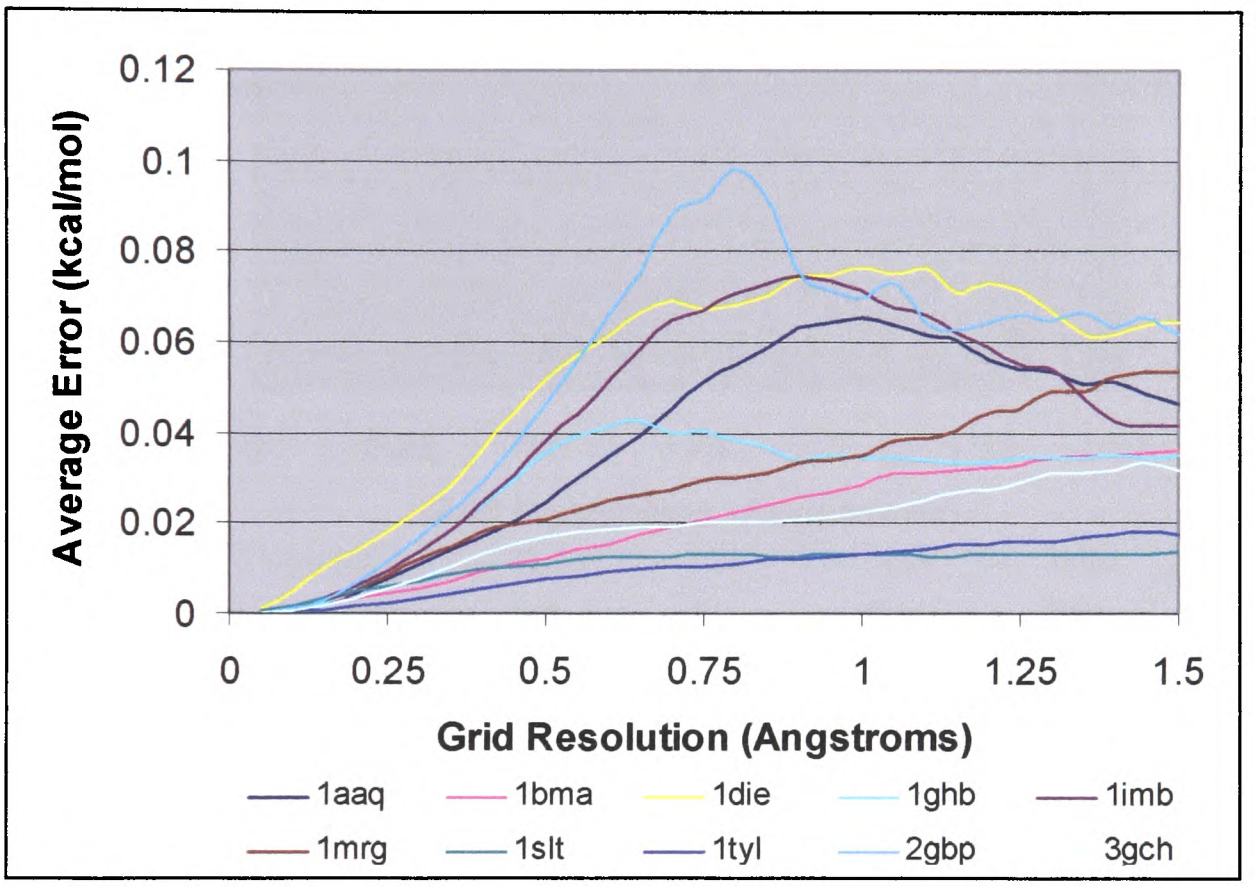


Figure 5.8 - Graph showing how the average error in prediction of the electrostatic binding energy varies with grid resolution. The ten test proteins are coloured as shown in the key.

This error is relatively small even for a 1.5 Å resolution. The error within this range has a maximum of only approximately 0.1 kcal/mol. Binding energies tend to be between 0 and -100kcal/mol. This error is thus unlikely to have a major affect, even for large ligands.

5.4.2 Calculation of van der Waals Interactions

Unfortunately, the average error is considerably more in the case of the van der Waals binding energy, as shown in Figure 5.9.

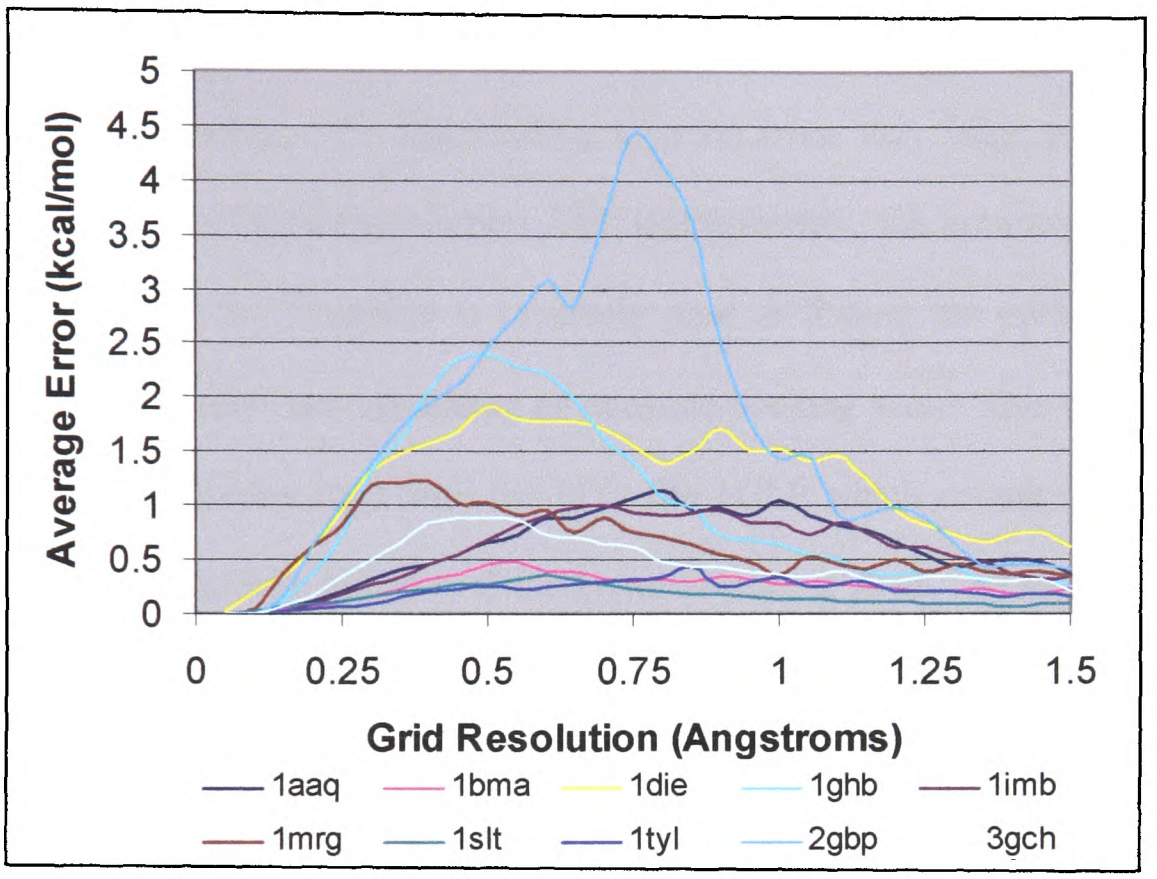


Figure 5.9 - Graph showing how the average error in prediction of the van der Waals binding energy varies with grid resolution. The ten test proteins are coloured as shown in the key.

In this case, the average deviation is as high as 2.5kcal/mol for a 0.5 Å grid resolution. This is for a single carbon atom. In a large ligand with as many as 50 heavy atoms, this could have marked effects on the predicted binding energy. The average error reaches an acceptable level only when the grid resolution is lowered to 0.1 Å. At this point, the average error is below 0.1kcal/mol in all ten cases. Note that the average error appears to be better at higher grid resolutions than at intermediate values. This is attributable to the majority of grid points being far from the protein in these cases.

5.5 Problems with Optimisation

Owing to the enormous search space involved in locating the binding site on a protein surface, Oxdock operates by using low scale representations of the ligand to discover and discard poor solutions early in the docking process. This increases the speed of the algorithm, allowing it to find binding sites on even very large proteins in a reasonable amount of time (see section 2.6). Unfortunately, the accuracy lost in the process means that the algorithm is relatively poor at finding the optimal docking pose of the ligand and thus calculating an accurate docking score. This is illustrated by a docking calculation done with the PDB file 1GLP which contains the protein Glutathione S-Transferase bound to Glutathione Sulphonic Acid. The search space is limited to a sphere of radius 6.0 Å centred on the ligand centroid to illustrate the point whilst reducing the computation time. The lowest energy docking solution can be seen in Figure 5.10.

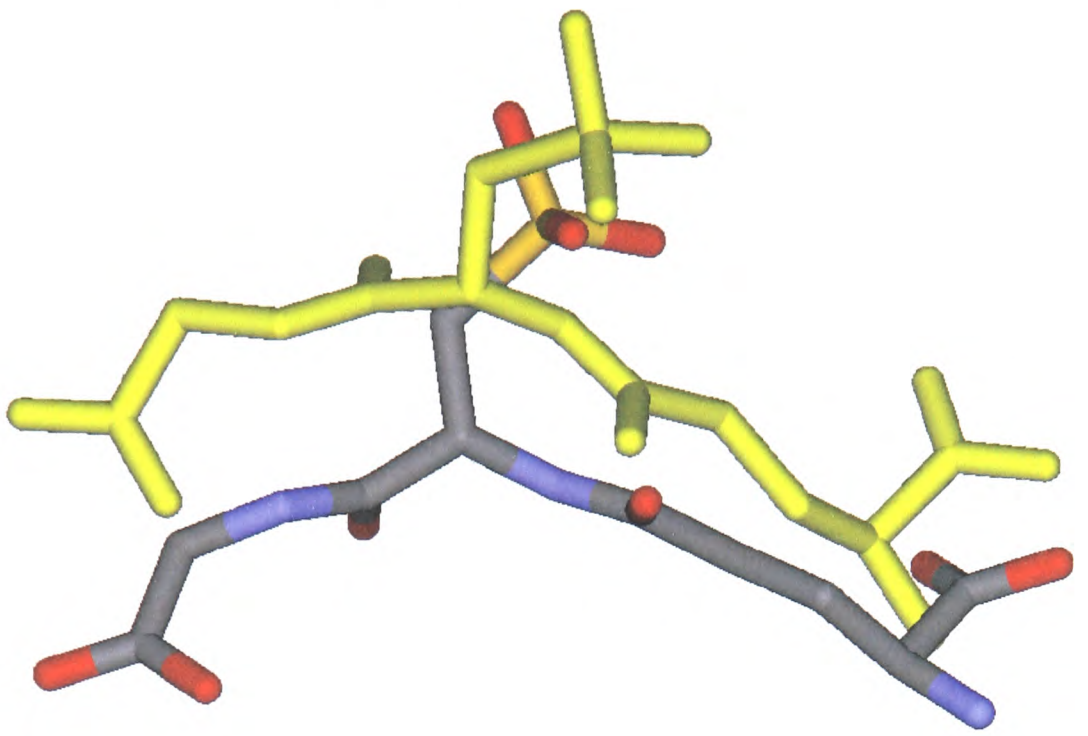


Figure 5.10 - The lowest energy docking solution of the ligand Glutathione Sulphonic Acid bound to Glutathione S-Transferase. The crystal structure is coloured yellow and the calculated structure is atom coloured.

One of the best estimates of accuracy for comparing two sets of coordinates is the root mean square deviation (RMSD). The RMSD of the heavy atom positions is a useful measure of the variation between the crystal structure and the calculated structure as it is a good description of the difference between two sets of coordinates in three dimensions. It is calculated using Equation 5.1.

$$RMSD = \sqrt{\frac{\sum_{atoms} (x_{act} - x_{pred})^2 + (y_{act} - y_{pred})^2 + (z_{act} - z_{pred})^2}{n}}$$

Equation 5.1 - The expression used to calculate the RMSD between the crystal structure and the calculated structure. *x*, *y* and *z* are the coordinates of the atoms, *act* and *pred* are the actual and predicted atoms and *n* is the total number of atoms.

In docking studies, an RMSD value of 2.0 Å or less, is considered to be good and an RMSD value of between 2.0 Å and 3.0 Å is considered to be acceptable. The RMSD in this case is 7.590 Å, well above the acceptable value. The calculated binding energy is -55.64 kcal/mol, which contrasts with the score of -89.64 kcal/mol obtained when the crystal structure is fixed. This highlights the problem with multiscale docking. The calculated docking pose is likely to be fairly dissimilar to the actual docking pose and different enough to mean that the calculated binding energy is inaccurate.

5.6 Summary

Oxdock is a very powerful algorithm and has demonstrated its usefulness in a number of real-world cases, as shown in Chapters 3 and 4. However, a number of issues remain and they are listed below:

- The multiscale approach can prevent a solution being discovered if a feature point is found outside the body of the molecule
- The multiscale approach can select poor solutions to the problem, as it docks the feature points and not the entire molecule.
- The multiscale approach sacrifices accuracy for increased speed. This is excellent if you are attempting simply to find the binding site but is inadequate if you need to predict and score the best docking pose precisely.
- Testing on the grid based energy calculations show that this method can closely predict the electrostatic binding energy at a grid resolution of 0.5 Å (which takes a reasonable amount of computation) but cannot closely predict the van der Waals binding energy at this resolution.
- The grid method of placing ligands means that the search space is quantised and thus optimisation of the solutions is not possible.

These factors strongly suggest that a new approach must be used when attempting to use a multiscale approach to calculate docking poses. In chapters six and seven, the technique of evolutionary programming (EP) is introduced and shown to suit the requirements of the problem perfectly. An algorithm is created which combines multiscale docking and EP to yield a docking method that is optimised for locating the active site of a protein.

6 Development of a New Algorithm

6.1 Overview

Drug discovery is a lengthy and costly process. It takes approximately 15 years to research, patent, develop, formulate and test a new drug molecule. Researchers at Tufts University have estimated the cost at $800 million (£500 million). Any tool that can aid this process is thus valuable. Computational methods are particularly useful in the initial stages, allowing the rational design of potent and specific inhibitors. However, the techniques can be far from accurate and take a significant amount of computer time. New methods are always being developed, and old methods improved in an attempt to solve these difficult problems.

In chapters 2, 3 and 4, we have shown how a multi-scale approach can reduce the complexity of the task of locating binding sites on proteins. The advantage of a multi-scale approach is that it can drastically reduce the time taken for molecular docking whilst retaining accuracy. Evolutionary programming is an optimization method, developed in the 1960s [65], which performs very well in solving complex problems with many possible solutions. It operates using the concept of natural selection. Initially, a population of solutions to the problem are created, and each solution is rated with a fitness value. The solutions are then free to develop over the course of a run, with poor solutions being expunged and good solutions being propagated. The advantage of an evolutionary approach is that it can find excellent solutions to complex problems without resorting to an exhaustive search.

This chapter describes the development of an algorithm called Eve, which combines the speed of a multiscale approach with the optimising power of evolutionary programming. This requires the development of an efficient algorithm and an effective scoring function.

6.2 Evolutionary Programming

For a given problem, the first stage in EP is the creation of a suitable population of possible solutions. This population then evolves, with the best solutions more likely to survive in subsequent generations. In each generation, every solution in the population is evaluated by a fitness function, which calculates its fitness relative to the rest of the population. The exact fitness function determines a surface in n-dimensional space and the purpose of the algorithm is to find the global minimum on this surface. However, in complex cases, the surface is so rugged and so extensive that only local minima are likely to be found. The molecular docking algorithm Eve employs the concept of EP to find the optimal docking pose for protein-ligand complexes.

6.2.1 Solutions

The solutions that comprise a population are strictly defined. In simple cases, the solutions can be represented as an array of numbers. This array is called a chromosome and each unit of the array can have a binary, integer or real numerical value. The single unit is termed a gene. The solutions cannot be represented as an array in more complex cases, but can be considered as structures with certain properties. However, these properties can change, and evolve as the algorithm

proceeds. The solutions in the case of molecular docking are conformations of a ligand at a specific position with a defined orientation.

6.2.2 Fitness

Each of the solutions in the population must be run through a fitness function to calculate its fitness relative to the rest of the population. This fitness is a measure of its suitability as a solution to the problem. The fitness in the case of drug design is generally the calculated binding energy of the protein-ligand complex, though considerations such as ease of manufacture and drug absorption may be considered. The calculated binding energy is ideally an estimate of the Gibbs free energy change (ΔG), but entropic effects are often ignored due to the difficulty of including them, and estimates of the enthalpy (ΔH) are more common. The fitness function used in Eve will be considered in section 6.6.

6.2.3 Reproduction

Reproduction is one of the key steps in EP. The population of solutions in any generation is used to spawn the solutions in the next generation, with the probability of any particular solution surviving being related to its fitness relative to the fitness of every other solution. The reproduction operation is best represented as a roulette wheel. Each solution has a wedge on the wheel and the wheel is spun once for each member of the population. For each spin, the probability of any parent being selected and being replicated in the next generation is equal to its relative fitness.

$$P_n = R_n = \frac{f_n}{\sum_s f_s}$$

Equation 6.1 - Calculation used to calculate probability of survival for each solution. P is the probability of survival of a solution n. R is the relative fitness. f is the fitness of a solution and s is the sum over all solutions.

For example, consider the following case:

	Fitness	Relative Fitness
Solution 1	100	0.20
Solution 2	390	0.78
Solution 3	10	0.02
Total Fitness	**500**	

Table 6.1 - Table of the fitnesses and relative fitnesses of three solutions.

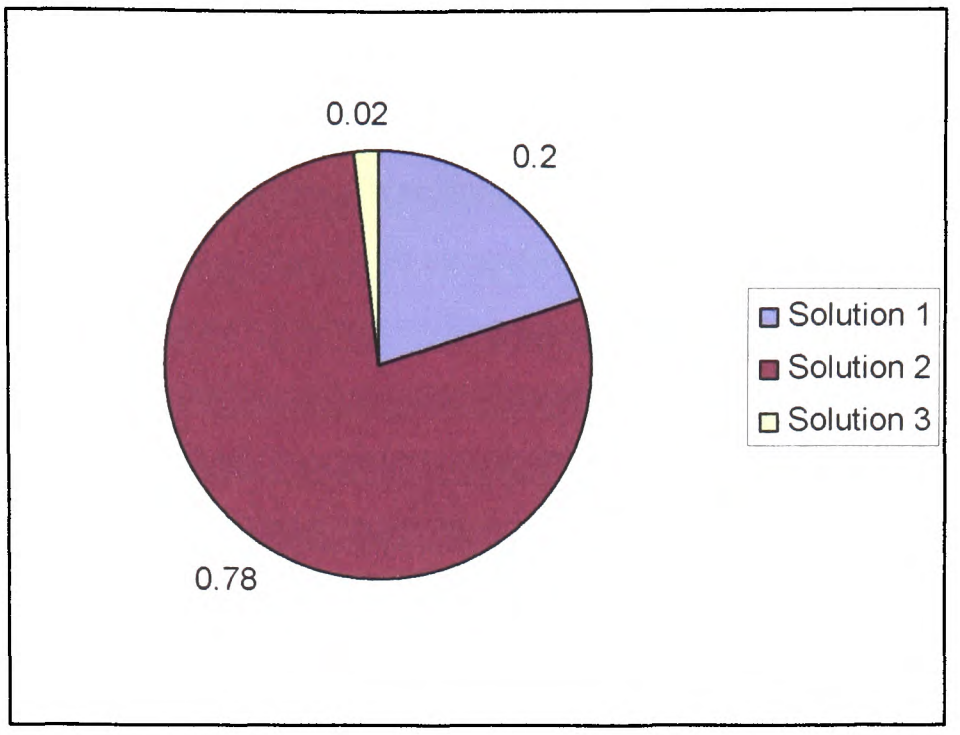


Figure 6.1 - Roulette wheel representation of the relative fitnesses shown in Table 6.1.

6.2.4 Mutations

Mutation is one of the vital elements of evolution and thus of EP. The genes, or the properties of a solution, are variables that can be altered. If the solution is composed of structures, further mutations are possible involving the creation or deletion of these structures. However, as in nature, there is an optimum rate of mutation. Very little change leads to slow optimisation and too much change leads to erratic behaviour and makes it difficult for the solutions to converge upon minima. The optimal rate differs for any given problem and is generally user defined in EP. There are a number of mutations used in Eve and they are discussed in detail in section 6.7.

6.2.5 Algorithm Structure

The combination of these various components means that the course of EP can be represented by the flow diagram in Figure 6.2.

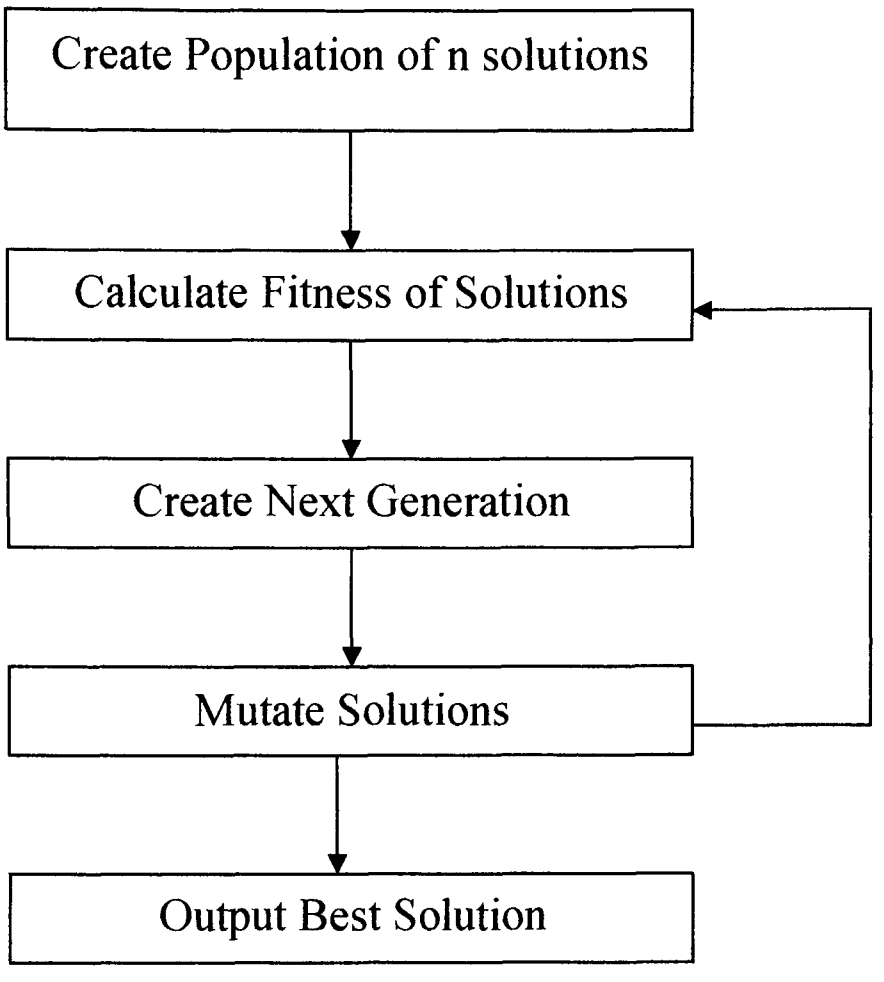


Figure 6.2 - Flow diagram illustrating a typical genetic algorithm.

6.3 Solutions

In this project, the solutions to the problem are ligand molecules, which are composed of atoms. The ligands are input into the algorithm from standard sdf files to yield structures with the properties tabulated in Table 6.2.

Number of Heavy Atoms	The Total Energy
Number of Hydrogens	The Relative Fitness
Number of Torsions	The Centroid Coordinates
Electrostatic Binding Energy	The Two Point Coordinates
Van der Waals Binding Energy	The Three Point Coordinates
The Internal Energy	A List of Conformers
Hydrophobic Binding Energy	A List of Torsions
The Rotor Restriction Energy	A List of Atoms

Table 6.2 - The properties of a ligand stored in Eve.

The individual atoms are also stored and have the properties tabulated in Table 6.3.

Element Type	A List of Bonded Atoms
X Coordinate	Atom Potential Type
Y Coordinate	Partial Charge
Z Coordinate	Hybridization

Table 6.3 -The properties of an atom stored in Eve.

These properties determine the properties of the ligand and allow it to be manipulated in order to optimise the geometry. Both the atoms within the ligand and the ligands

within the population are saved as linked lists. This allows the memory to be dynamically allocated and the atom and ligand data to be swiftly accessed.

Recently developed docking algorithms have begun to consider protein flexibility as well as ligand flexibility [66-68]. This is essential in cases where there is a large change in the active site upon binding and in general, the accuracy and usefulness of an algorithm will be greatly improved if protein flexibility is considered in an accurate way. One of the main problems with attempting to include flexibility of the protein is the massive increase in the search space for an already enormous problem. This issue has been addressed in a number of ways. The first simplification is to consider only proximal amino acid residues to be flexible. The conformation of residues far from the ligand is likely to have a lower effect on binding than those near the ligand. The second simplification is to use a population of conformational sub-states for the protein. Before the main portion of an algorithm runs, the binding site residues are analysed to create an ensemble of low energy protein structures. These conformations become the only structures that are considered by the algorithm.

The algorithm Eve does not consider protein flexibility, partly due to the large increase in run time and partly because one of Eve's main uses is envisaged to be binding site location. Thus, the entire protein would need to be considered as flexible; a problem beyond the ability of modern computer systems. However, the EP algorithm is perfect for incorporating the conformational sub-states model to allow consideration of protein flexibility. Development of the algorithm should be in this direction.

6.4 Conformer Generation

As discussed in section 2.4, it is important to consider the flexibility of the molecule when attempting to find the optimal docking pose for a protein-ligand complex. The problem of sampling many conformers is solved in Eve using the concept of conformer clustering, whereby a small number of conformers are used to represent the entire ensemble. The first iteration, using only one feature-point, requires no clustering, because the one-point representations are identical for every conformer. For the second generation, a two feature-point representation is calculated for each conformer and the distance between the two points is stored. A user-defined threshold is then used to cluster conformers with a similar distance. This cluster is represented by the lowest energy conformer within it. A distance threshold of 0.5 Å is generally used. As an illustration, the five representations from the complex 1BLH between a Beta Lactamase and Phosphonate Inhibitor N-(benzyloxycarbonyl)-amino-methyl-phosphinic acid are shown in Figure 6.3. This molecule has five rotatable torsions and thus with an incremental angle of 30°, the total number of conformers is 248,832 (12^5).

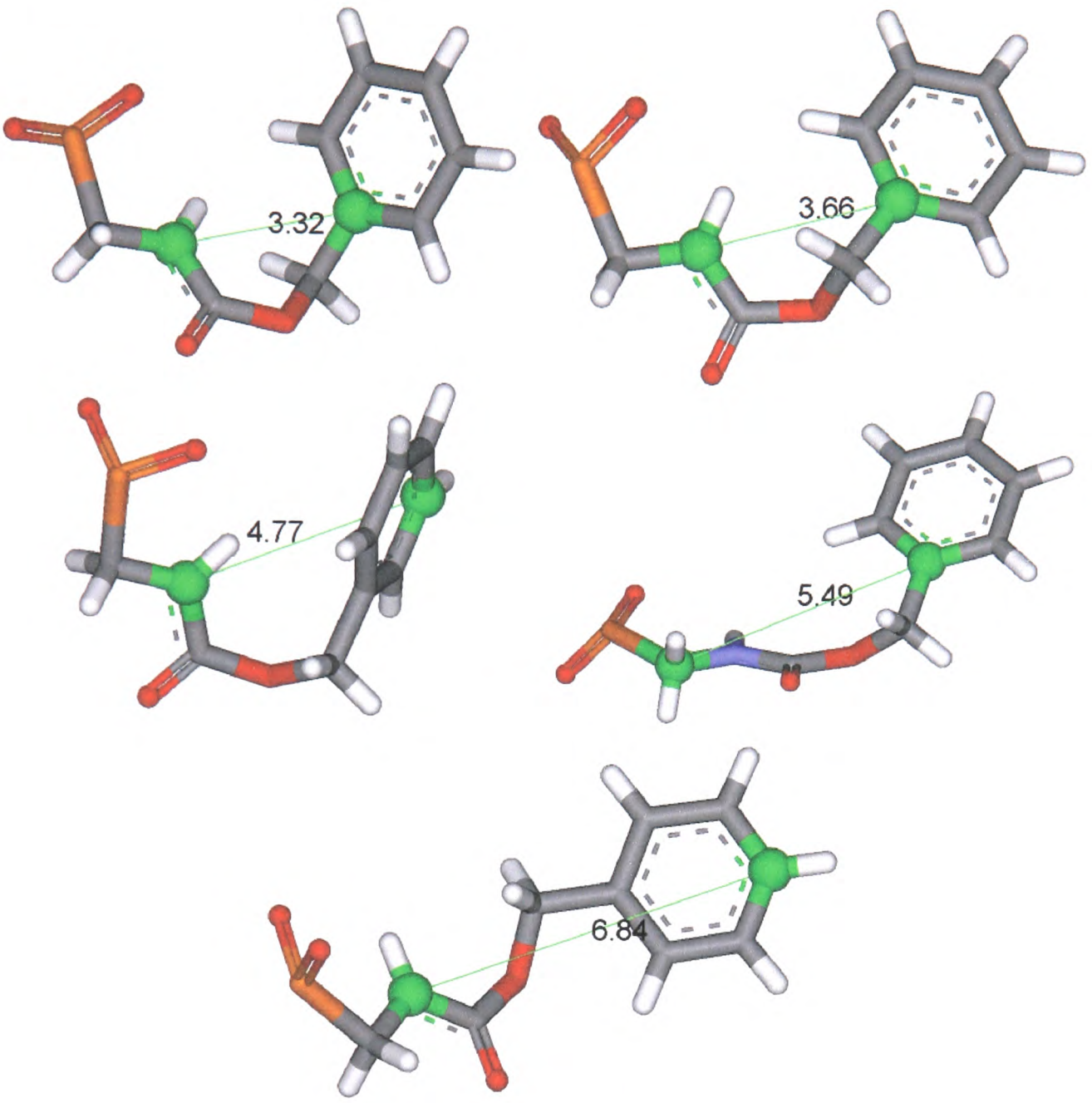


Figure 6.3 - The conformers representing the five clusters for the phosphonate inhibitor in the PDB complex 1BLH. The two feature-point atoms are shown as green spheres and the distance between them is labelled in black.

A comparison between these five conformers and the actual crystal structure is very enlightening. The last of these conformers, with a distance of 6.84 Å between feature-points, is very close to the actual structure. The RMSD between the two is only 1.153 Å. This can be seen in Figure 6.4 and illustrates that in this case, a conformer very

similar to the bound conformer will be sampled, greatly improving the probability that the correct solution will be found. In general, this method ensures that many disparate conformers will be sampled, increasing the probability that a conformer close to the bound conformer will be tested.

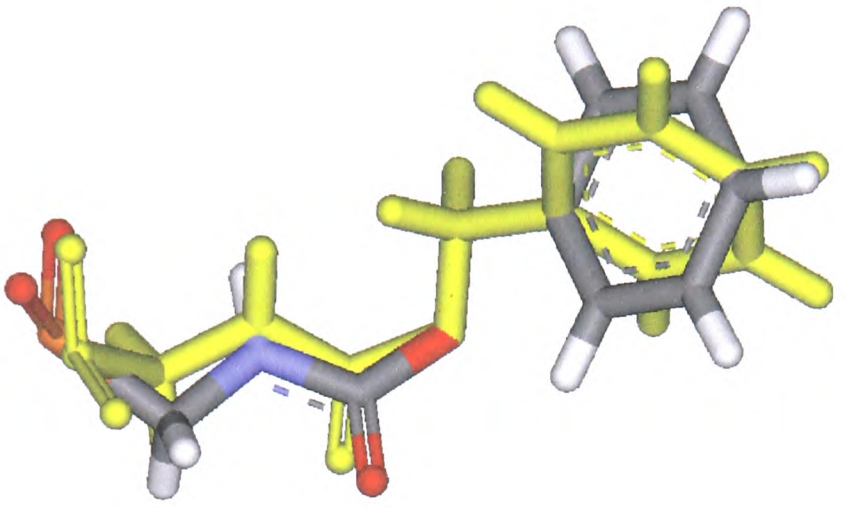


Figure 6.4 - An overlay of the actual crystal structure of the phosphonate inhibitor from PDBID 1BLH (in yellow) and one of the representatives from the clustered conformer search (atom coloured). The RMSD between the heavy atom positions is 1.153 Å.

This technique greatly reduces the number of calculations that must be performed. Only the clustered conformers are docked and the best of the surviving solutions are conformers with an associated grid point and orientation. These solutions then become the first generation for the evolutionary portion of the algorithm. This process is vital to ensure that an ensemble of possible solutions to the problem is considered.

6.5 Modifying the Multiscale Approach

To ensure a swift docking method, the algorithm uses a multiscale approach for the first three feature-points. The geometry of the entire ligand is fixed after three points are defined and thus after this stage, the entire molecule is represented to make the calculations more accurate. This prevents the problems that can occur with long ligands and improves the accuracy of the prediction. However, the method of altering the feature-points differs from that used in Oxdock. In Eve, the feature-points are all at atom positions; the centroid of a cluster of atoms is calculated, but the feature-point is placed on the nearest atom to that position. This ensures that feature-points are never placed outside the ligand, where repulsion may occur. Furthermore, the atom clusters are not chosen using a *k*-means clustering algorithm, as in Oxdock, but by choosing the two atoms that are farthest apart (in the case of two points) or the three atoms that form the largest triangle (in the case of three points). These atoms then "own" the cluster of atoms that are nearest to them. The atom nearest to the centroid of each cluster then becomes the feature-point. This may seem a small issue, but it prevents the problems that occur when a ligand feature point is placed outside the ligand molecule.

6.6 Scoring Functions

As discussed in section 6.2.2, the fitness of a solution is integral to its survival. Thus, the fitness function is vital in shaping the course and direction of evolution. If the function is too complex, it may take too long to evaluate. If the function is too simplistic or is simply inaccurate, the solutions may not be realistic or the final answer may not suit the actual problem. Within Eve, the fitness is defined as the

binding energy between the protein and the ligand. It is thus important to have a good scoring function for calculating protein-ligand interaction energies.

6.6.1 Empirical Scoring Functions

There are huge varieties of energy functions, based on various parameters calculated from experimentally derived binding energies. They are generally accurate for a small dataset of similar compounds and are fairly quick and simple to evaluate. However, they tend to become less precise and thus less useful as the dataset encompasses more disparate molecules. In this project, the ligands could theoretically have any structure and thus an empirical function is less likely to be a good estimate of the binding energy

6.6.2 Theoretical Scoring Functions

Theories of molecular docking vary from simple to complex models. Molecular dynamics simulations and Monte Carlo calculations give good indications of binding energies including both enthalpic and entropic considerations. However, these calculations are often computationally time-consuming and in a case where many such calculations must be performed, they become unfeasible. Static force-field calculations such as those employed in Oxdock are a simpler and accurate replacement. These are generally based on considerations of electrostatic and van der Waals contact forces but their static nature means that entropy is not explicitly considered.

6.6.2.1 Electrostatic Interactions

The electrostatic contribution can be calculated from the Poisson equation, which relates the variation in electrostatic potential (φ) to the charge density (ρ) over a

distance (r) in a medium of uniform dielectric constant (ε). In reduced electrostatic units, this is given by:

$$\nabla^2\phi(r) = -\frac{4\pi\rho(r)}{\varepsilon}$$

Equation 6.2 - The Poisson equation

∇^2 represents the second order differential with respect to x, y and z. For a set of point charges in a constant dielectric medium, this reduces to the much simpler Coulomb equation:

$$\phi(r) = \frac{q_i}{4\pi\varepsilon_0 r}$$

Equation 6.3 - The Coulomb equation giving the electric potential φ at an atom *i*. *q* is the charge on the atom and *r* is the distance from the atom.

Eve uses the Coulomb potential calculated in this way to derive the electrostatic contribution to the binding energy. In the coulomb model, the energy of interaction between two points (*i* and *j*) with charges q_i and q_j, is given by the product of the charge at one, multiplied by the potential at that point due to the other. Unfortunately, this model completely ignores the effect of solvent. In a vacuum, the potential can be considered to occur in this inverse way but in water, the solvent screens the charge-charge interaction between two points and this lowers the interaction energy. One of

the most simple methods for dealing with this problem is to assume that this screening occurs with an inverse dependence on the distance and thus the potential is multiplied by (1/r). This is termed a distance-dependent dielectric. It is commonly employed in docking algorithms and is used in Eve. The electrostatic binding energy is thus calculated using the expression in Equation 6.4.

$$Binding\ Energy = \frac{q_i q_j}{4\pi\varepsilon_0 r^2}$$

Equation 6.4 - Coulombic electrostatic binding energy between atoms *i* and *j*. *q* is the charge on each atom and *r* is the distance between them.

When considering a protein-ligand complex, Eve calculates the interactions between every ligand atom and every protein atom to give the total electrostatic binding energy.

6.6.2.2 Van der Waals Interactions

The van der Waals interaction is due to two competing forces: a repulsive force due to the presence of electrons surrounding the nuclei and an attractive dispersion force due to the induction of complementary dipoles in proximal atoms. These combine to give an overall interaction, which has the following form as atoms are moved farther apart:

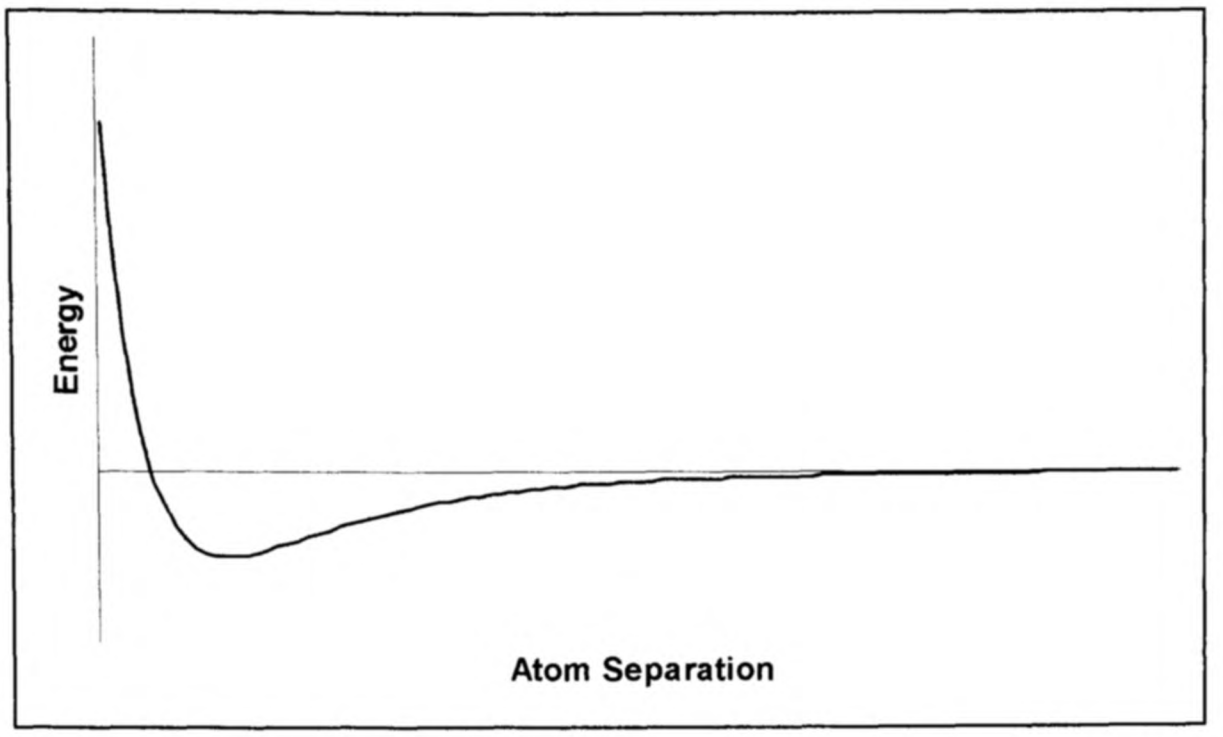


Figure 6.5 - Graph showing how binding energy varies with atomic separation

The van der Waals interaction can be parameterised in a variety of ways. The most common, and the one employed here, is the Lennard Jones 12-6 potential. The interaction energy is given by:

$$Binding\ Energy = 4\varepsilon_{ij}\left(\frac{A_{ij}}{r^{12}} - \frac{B_{ij}}{r^{6}}\right)$$

Equation 6.5 - Lennard Jones derived van der Waals binding energy between atoms *i* and *j*. ε_{ij}, A_{ij} and B_{ij} are the van der Waals parameters and *r* is the distance between the atoms.

The Lennard Jones parameters ε_{ij}, A_{ij} and B_{ij} are empirically determined for and vary for any pair of atoms of identity i and j. The parameters used in Eve are modified

from those used in AUTODOCK [12]. A common technique used to reduce the computation time is a van der Waals cut-off distance. The van der Waals interactions are short ranged when compared to electrostatic interactions and do not have significant effects at long distances. A cut-off of 20 Å is used in Eve and thus only protein atoms within 20 Å of a ligand atom are considered to contribute to the van der Waals binding energy. The calculated van der Waals interaction energy is thus slightly lower it would be, but the very small effect of distant atoms means that the error is likely to be very small.

6.6.2.3 Internal Coordinate Contribution

It is also important to consider the energy contribution from the internal coordinates. If the ligand binds in an unfavourable conformation, this must be included in the total energy. In Eve, the internal coordinate contribution is calculated as a sum of the electrostatic and van der Waals interactions of all atoms that are at least three bonds away. The contribution of the torsion energy is not calculated due to problems with parameterisation, but this will not have a large effect and is often ignored in scoring functions. For any given case, the internal coordinate contribution is defined as the difference between this energy and the minimal internal coordinate energy, which can be calculated at the beginning of the run. The internal coordinate contribution is then the increase in energy required to reach this conformation from the minimum energy conformation.

6.6.2.4 Desolvation Effects

Protein-ligand docking is a complex process involving many steps. One of the most important aspects and one of the most difficult to quantify is the effect of the

desolvation. In the bulk, the ligand and the protein are solvated by water molecules. Many of these water molecules must be removed before binding can occur and this has an energetic penalty associated with it. Any electrostatic interactions between ligand and water or protein and water are lost and there are other subtle effects to be considered. One of the most common models used in molecular docking to account for the effect of solvent is the Generalised Born/Solvent Accessible Surface method. This models the solvent as a continuum that interacts with molecules. The electrostatic interactions can thus be calculated. The remaining effects due to non-polar interactions and cavity formation are assumed to scale linearly with the solvent accessible surface area. Despite these effects, the desolvation contribution is not considered in Eve due to the time taken to perform the calculations. The aim of the current implementation is to swiftly find the binding site this simplification should not prevent that. However, ignoring this contribution will mean that the calculated binding energies will be unrealistic and incomparable to experimental values.

6.6.3 Entropic Effects

When using a force-field approach to evaluate the interaction energy of a protein-ligand complex, a better estimate will generally be obtained by also considering entropic effects, which are implicitly considered in molecular dynamics simulations or Monte Carlo calculations. For this project, the entropic effect is attributed to two factors: rotor restriction and the hydrophobic effect. These are considered in turn.

6.6.3.1 Rotor Restriction

The rotation of a bond is one of the degrees of freedom of a molecule. Once bound to a protein, some or all of the torsion angles are fixed. This increases the order within the molecule, with a resultant decrease in the entropy, disfavouring binding. A

number of studies have been performed to estimate the effect of freezing rotors, after the pioneering work performed by Jencks and Page [69]. Obviously, the actual effect will vary in each case dependent on both the identity of the atoms and their environment as well as the degree of the restriction (the barrier to rotation). However, the following average values have been calculated.

Source Paper	**Method**	**Value (kcal/tor)**
Pickett & Sternberg [70]	Empirical	0.54
Nicholls [71]	Experimental Work on Alkanes	0.45
Luo & Gilson [72]	Theoretical Calculation	0.30
Wang [73]	Theoretical Calculation	0.48
Creamer [74]	Various Experimental	0.5
Lazaridis [75]	Average Experimental	0.4

Table 6.4 - Calculated values of the energetic effect of rotor restriction in published papers.

In this project, a value of 0.4 kcal/torsion will be used. The number of torsions is a property of the ligands and thus this is easy to calculate for any given molecule. The binding residues in the protein molecule may also be fixed to some degree. However, this effect is not considered in the current program due to the difficulty of determining

which residues are fixed in the apo and the ligand-bound state and to what degree they are fixed.

6.6.3.2 Hydrophobic Effect

One of the most important entropic factors in determining binding energy, yet one of the most difficult to quantify, is the hydrophobic effect. This arises from the removal of water molecules from the surface of the ligand when they form interactions with protein residues [76]. Hydrophobic surfaces such as those of phenyl rings or alkyl chains provide a more favourable energy upon desolvation. Thus, the hydrophobic effect is one of the main reasons that oil and water do not mix [77].

The hydrophobic effect actually arises from both entropic and enthalpic effects, and the relative magnitude of each varies with temperature [77, 78] and with the system studied [79]. In general, a water molecule at a hydrophobic surface has less favourable electrostatic interactions, and is ordered to a degree by the surface. This compares to the favourable situation when the water molecule is in the bulk solution, where it is fully hydrogen bonded and free to move around. A number of attempts have been made to quantify this effect [80-82].

However, in this project, a simple quantifier is needed. The effect is parameterised in Eve by considering the entire surface of the molecule and placing roughly equidistant points on the surface of each atom. Each point contributes to the effect if it is close to the protein surface and is thus "buried" in the bound state. The surface of each atom is a sphere with a radius equal to the sum of the van der Waals radius and the approximate "radius" of a water molecule (1.4 Å).

It is noted that the surface a molecule has a quality of hydrophobicity and is not simply hydrophobic or hydrophilic. For example, if the ligand is another water molecule, there will be no hydrophobic effect, and if the ligand is an ammonium ion, the same should be true. Thus, polar and charged ligands will be less affected and the partial charge of an atom may be a good measure of hydrophobicity of its surface. As charge is bipolar and charges near zero should yield a greater hydrophobic effect, it would seem sensible to use a Gaussian function. The hydrophobicity of a surface point is determined by the charge on the atom to which it belongs:

$$H = e^{-h(PC)^2}$$

Equation 6.6 - Estimation of the hydrophobic effect. H represents the hydrophobicity of a point around an atom with partial charge PC. h is a constant.

The constant *h* parameterises the extent to which the hydrophobicity decreases as the charge rises. A value of 10 leads to a sharply decreasing energy as the charge increases, with very little effect as the charge reaches that of a hydrogen atom in a water molecule (+0.41) This meets the criteria of the model. The following approximate method is then used to place equidistant points on the surface of an atom.

Consider the surface of a sphere of radius R. Split the surface into circles by varying the azimuthal angle from -90° to 90° in increments of θ. Each circle will have a radius *r* of Rcos θ. Place points equally spaced around the equator of the sphere. The

distance d between points on the equator becomes the constant distance between points on all circles. In each other circle on the sphere, the incremental angle φ required to produce a distance of d between the points is different and can be calculated by considering the two circles:

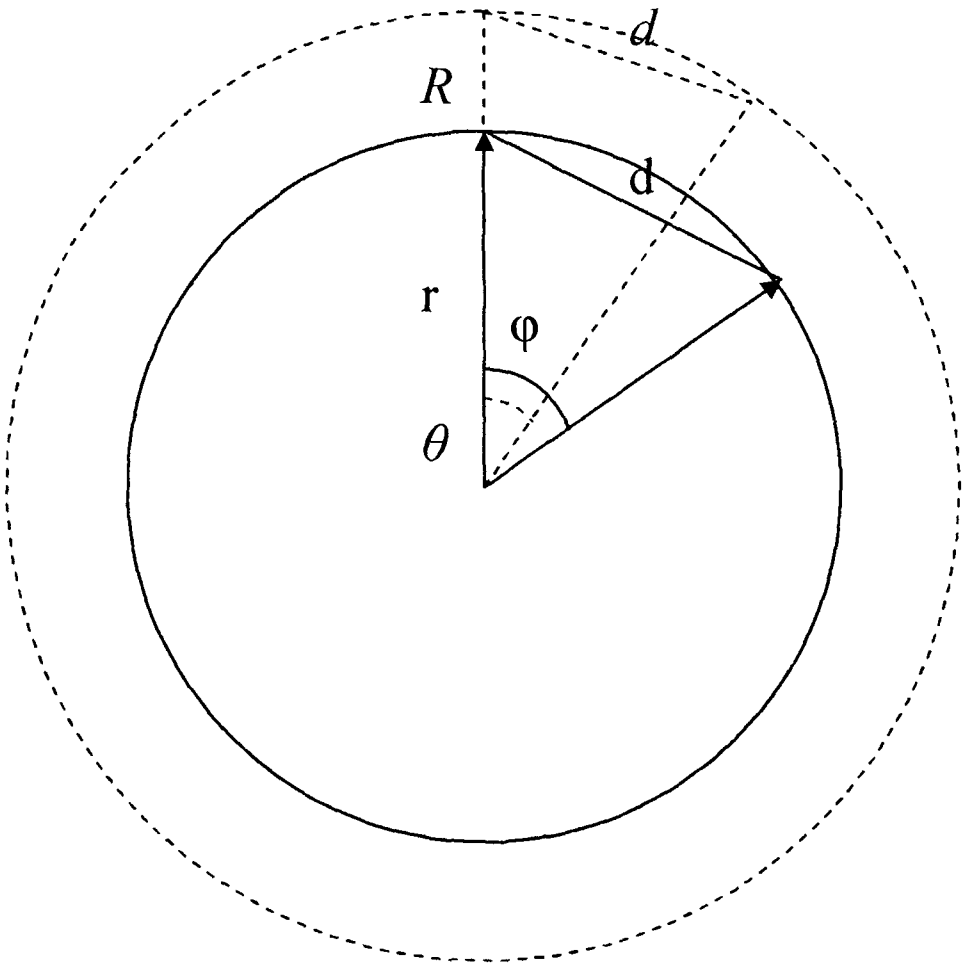


Figure 6.6 - Top view of a sphere showing two latitudinal circles and the angles required to yield an equal edge distance *d*.

The following statements are true:

$$d = 2R\sin(\frac{\theta}{2})$$

$$r = R\cos\theta$$

$$\sin\frac{\varphi}{2} = \frac{d}{2r}$$

Equation 6.7 - Relationships between the edge distance *d*, the radii of two latitudinal circles *r* and *R* and the angles *φ* and *θ*.

Thus, the angle φ can be calculated for any given case:

$$\varphi = 2\sin^{-1}(\frac{\sin 0.5\theta}{\cos\theta})$$

Equation 6.8 - Equation relating the two angles *φ* and *θ*.

This approach operates by placing many rectangular bands of equal height (*H*) around the sphere. Each rectangle is split into squares of length *H*, the number of squares being dependent on the radius of the cross section. Unfortunately, each rectangle will only be symmetrically split if φ is divisible by 360° and thus the solution is not perfect. The surface are will be slightly underestimated or overestimated for each case. However, the errors will tend to cancel out and the solution becomes more exact as the angular increment θ decreases. Due to every point being equidistant from its nearest points, each represents a region of the surface with an area proportional to the total area dependent on the total number of points (*n*).

$$\text{Area per point} = \frac{4\pi R^2}{n}$$

Equation 6.9 - Surface area represented by each surface point. *R* is the atomic radius and *n* is the number of points on the surface.

Only a certain fraction of each atom's area will be counted as buried surface area. The overall hydrophobic effect for each atom can then be calculated as the area of the atom, multiplied by the fraction of the surface that is buried, multiplied by the hydrophobicity of each point multiplied by a constant. This effect can be calculated for each atom and summed to evaluate the total contribution to the binding energy.

$$Energy(kca\ l/mol) = \sum_{atoms} \frac{H \times Hydrophobi\ city \times Atom\ Area \times Surface\ Points}{Total\ Points}$$

Equation 6.10 - Calculation of the hydrophobic binding energy.

The value of the constant *H* is obtained by examining the results of a number of studies [77-83] and reaching a consensus. The values given in these papers for a purely hydrophobic surface correspond to the case of an uncharged atom, and can thus be directly equated to *H*. A value of -0.03 kcal/Å^2 is used. The effect of varying the incremental angle can be seen in Table 6.5 for the molecule diethyl ether for a run of twenty molecules.

Incremental Angle (°)	Average Area (Å)	Min/Max Area (Å)	Standard Deviation (Å)
20	181.39	176.75-184.77	2.072
10	181.48	179.84-184.82	1.733
5	183.76	183.53-184.87	0.344
2	183.77	183.39-184.27	0.237
1	182.75	182.66-182.98	0.082

Table 6.5 - Table showing the effect of decreasing the incremental angle on the accuracy of the surface area calculation.

The slight increase in accuracy with lower values does not counteract the length of time taken to perform the calculation. A value of 10° will be used. The effect of this can be seen in the following examples of some common molecules, their areas and the calculated hydrophobic effect. Owing to the method's slight inaccuracy, the calculated area will change. The average of twenty runs is calculated, along with the range for an incremental angle of 10° and the results are shown in Table 6.6:

Molecule	Average Calculated Area (Å)	Min/Max Area (Å)	Standard Deviation (Å)
Acetate	191.44	186.99-195.49	2.016
Acetone	190.65	186.28-194.67	1.940
Benzene	192.70	180.78-197.59	3.789
Ether	181.48	179.84-184.82	1.733
Methane	141.81	138.32-144.33	1.597
Methanol	168.00	164.48-173.95	2.000
Water	131.03	128.76-133.02	1.050

Table 6.6 - Table showing the actual surface area of a variety of common molecules and the calculated area and standard deviation of the calculation.

The result for an incremental angle of 10° for the Sialic Acid molecule is shown in Figure 6.7:

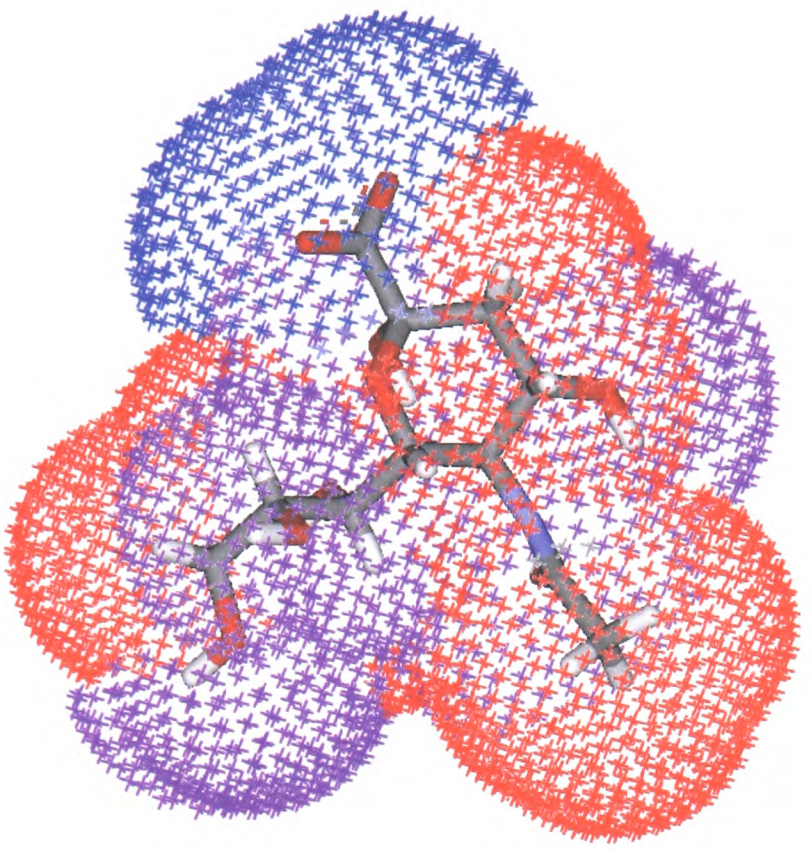


Figure 6.7 - Surface points around the Sialic Acid molecule. The points coloured blue 'belong' to atoms with a high charge (greater than 0.4), the points coloured purple 'belong' to atoms with an intermediate charge (between 0.15 and 0.4) and the points coloured red 'belong' to atoms with an low charge (less than 0.15).

The three surface types are analysed in Table 6.7.

Surface Type	Surface Area (Å^2)	% Area	Binding Energy (kcal/mol)	% Energy
Hydrophilic	8.70	3.88	-0.01	0.24
Intermediate	87.18	38.86	-0.74	17.29
Hydrophobic	128.467	57.26	-3.49	81.30
Total	224.34		-4.24	

Table 6.7 - The surface of Sialic Acid split into hydrophilic (atom charge greater than 0.4), intermediate (atom charge between 0.4 and 0.15) and hydrophobic (atom charge less than 0.15) portions showing the area of each and the contribution to the total hydrophobic binding energy.

As expected, the hydrophobic surfaces (with a low atom centred charge) contribute more to the hydrophobic binding energy. The addition of this effect into the binding energy of each protein-ligand complex should improve the accuracy of the scoring function.

6.6.4 Scoring Function

The overall scoring (fitness) function can be seen in Equation 6.11.

$$
\begin{aligned}
Binding\ \ Energy = {} & van\,der\,Waals\ Interaction \\
& + Electrostatic\ Interaction \\
& + Rotor\ Restriction Term \\
& + Internal\ \ Coordinate\ Contribution \\
& + Hydrophobic\ Interaction \\
& = \sum_{l}\sum_{p} 4\varepsilon_{lp}\left(\frac{A_{lp}}{r^{12}} - \frac{B_{lp}}{r^{6}}\right) \\
& + \frac{q_l q_p}{4\pi\varepsilon_0 r^2} \\
& + \sum_{k} T \\
& + \sum_{i}\sum_{j} 4\varepsilon_{ij}\left(\frac{A_{ij}}{r^{12}} - \frac{B_{ij}}{r^{6}}\right) \\
& + \sum_{i} H \times Hydrophobicity \times Atom\ Area \times Fractional\ Surface
\end{aligned}
$$

Equation 6.11 - The equation used in Eve to calculate the total binding energy. In the expression, *l* is the index of all ligand atoms, *p* is the index of all protein atoms, *k* is the index of all torsions, *T* is the energy for freezing a torsion, *i* and *j* are the indices for the ligand atoms and *H* is the constant for the hydrophobic interaction.

This scoring function calculates the effects of van der Waals interactions, electrostatic interactions and internal coordinate contribution that are considered in Oxdock. However, the torsion energy term is omitted due to difficulty with parameterisation and the electrostatic interaction is calculated using Coulomb's equation with a distance-dependent dielectric. This is a large approximation with no physical basis but

is calculated more swiftly and easily than the commonly used solutions based on the Poisson-Boltzmann equation. In Eve, as in Oxdock, speed is a key attribute of the algorithm and is the justification for these omissions. Two more terms are used in the scoring function that are not present in Oxdock. The rotor restriction term is present to allow comparisons of multiple ligands, where the more rigid molecules will lose less entropy upon binding and will thus interact more favourably. The hydrophobic effect is an important contribution to binding and is often omitted from scoring functions due to the multiple and intractable causes. It is parameterised in Eve by a new method that considers the hydrophobicity of all the surface points that are buried and takes a sum over the entire molecule. This has no theoretical basis, but makes a justifiable and important hydrophobic contribution to the binding energy. The main omission from the scoring function is a term accounting for desolvation effects. This is one of the most complex issues in molecular docking and is addressed commonly using the Generalised Born/Solvent Accessible Surface method. This was discussed in section 6.6.2.4 but is not used in Eve due to the necessity for speed.

The scoring function in Equation 6.11 thus considers many of the factors that affect intermolecular binding energies, but makes many simplifications and omits several terms. It will thus never make accurate predictions of binding interactions. This is unfortunate, but should not prevent the algorithm from completing its most important role, which is the location of active sites. When it is used in practice, more detailed binding energy calculations could be performed on the optimised docking pose to yield predictions that are more accurate. A detailed review of scoring functions has been completed by Gohlke [84].

Within the algorithm, each ligand must be evaluated using this function to calculate its fitness relative to the rest of the population. The probabilities are then calculated using a Boltzmann-like distribution of energies. This represents a realistic situation where many ligands compete for a binding site.

$$P(s) = \frac{e^{-kF(s)}}{\sum_{a} e^{-kF(a)}}$$

Equation 6.12 - Equation to calculate the survival probability (P) of a solution (s) using its fitness (F). The variable (a) is an index used to sum the fitnesses of all the solutions.

The fitness of a solution is equated with the calculated binding energy. The Boltzmann factor k is a variable of the program and can be equated with the temperature of the system in a thermodynamic sense. Thus, a high value of k leads to the increased probability of choosing high-energy solutions. This has the positive effect of allowing the population to surmount energy barriers but the negative effect of allowing poor solutions to prosper. This value can be kept constant, can be decreased during the course of the run or can be used as a momentum to the algorithm. This means that when a solution reaches a local minimum, the value of k rises to allow it to escape the minimum, but then decreases when the solution is "searching" for other minima. In Eve, the user can define a fixed or decreasing value of the Boltzmann factor.

6.7 Mutations

In section 6.2.4 it is stated that there are a number of possibilities for mutations that alter the position of a molecule. These will now be considered.

6.7.1 Modification of a Torsion Angle

The torsion angles within a molecule are vital in determining the three-dimensional structure. Eve thus keeps a list of all the torsions within each ligand and this method rotates one end of the torsion (at random) by a random amount between ± n°, where n is a user-defined angle. The degree of rotation is restricted as a large modification of torsions within a large molecule is likely to alter the structure radically, leading to unfavourable steric clashes. This prevents the algorithm from jumping out of energy minima and allows it to converge slowly but smoothly on optimal solutions. An angle of 60° is used in Eve.

6.7.2 Rotation of the Molecule

The rotation method is used to optimise the position of a molecule within the binding site. An atom is chosen at random as the centre of rotation and a random vector is created that is centred on the atom. The molecule is then rotated by a random amount between ± n°, where n is a user-defined angle. The degree of rotation is again restricted to prevent solutions becoming unviable due to steric clashes. It again prevents the algorithm from jumping out of energy minima and allows it to converge slowly but smoothly on optimal solutions. An angle of 60° is used in Eve.

6.7.3 Translation of the Molecule

The translation method is also used to optimise the position of a molecule within the binding site. A random vector is chosen with a magnitude chosen at random between $\pm n$, where n is a user defined distance, and the entire molecule is translated by this vector. The extent of translation is again restricted to prevent solutions becoming unviable due to steric clashes. This also prevents the algorithm from jumping out of energy minima and allows it to converge slowly but smoothly on optimal solutions. A distance of 0.5 Å is used in Eve.

6.7.4 Survival

The final possibility is that the molecule may simply survive unchanged to the next generation. The optimal ratios of these mutation types are only calculable by exhaustively solving the problem, which defeats the objective of the method entirely. Thus, the ratios must be optimised, based upon some initial guess. The ratios in Table 6.8 have been tested in Eve and found to produce good and consistent results.

Mutation	**Relative Probability**
Torsion Modification	30
Translation	10
Rotation	10
Survival	50

Table 6.8 - The mutations used in Eve and their relative probabilities.

6.7.5 Rotation Matrices

A number of the operations within Eve require that some or all of a molecule be rotated about an axis. The method to rotate the ligand is the prime example. Initially, an atom is chosen at random as the centre of the rotation and a randomly determined orthonormal basis set is placed upon the atom. The rotation of θ radians can be performed on any vector **V** by the following simple steps:

- Rotate **V** onto the *xz*-plane using the matrix **X** and call the result **V'**.
- Rotate **V'** onto the *z*-axis using the matrix **Z** and call the result **V''**.
- Rotate **V''** about the *z*-axis by θ radians using the matrix **θ**.
- Invert the rotation of **V''** from the *z*-axis using the matrix inverse **Z'**
- Invert the rotation of **V''** from the *xz*-plane using the matrix inverse **X'**

The overall operation upon **V** to yield the resultant vector **F** then becomes:

$$\overline{F} = \hat{X}'\hat{Z}'\hat{\theta}\,\hat{Z}\,\hat{X}\,\overline{V}$$

Equation 6.13 - Expression illustrating the combination of simple vector manipulations required to rotate about an arbitrary axis.

This complex combination of matrices can be calculated for the general case of a rotation by θ radians about an axis **V** (v_x, v_y, v_z) and amalgamated into the Rotation matrix **R.** This matrix transformation is taken and modified from that at http://www.devmaster.net/wiki/Transformation_matrices:

$$\hat{R} = \begin{bmatrix} v_x{}^2 + \cos\theta(1 - v_x{}^2) & v_x v_y(1-\cos\theta) - v_z \sin\theta & v_x v_z(1-\cos\theta) + v_y \sin\theta \\ v_x v_y(1-\cos\theta) + v_z \sin\theta & v_y{}^2 + \cos\theta(1 - v_y{}^2) & v_y v_z(1-\cos\theta) - v_x \sin\theta \\ v_x v_z(1-\cos\theta) - v_y \sin\theta & v_y v_z(1-\cos\theta) + v_x \sin\theta & v_z{}^2 + \cos\theta(1 - v_z{}^2) \end{bmatrix}$$

Figure 6.8 - Composite matrix used to rotate about an arbitrary axis.

Each atom is rotated by calculating the vector between its position and the centre of rotation. This vector is then rotated by **R** and the new position calculated relative to the centre of rotation. This process is also used in the method of modifying a torsion angle. However, in this case, the process is complicated due to the necessity of rotating only atoms connected directly to one end of the torsion. This is accomplished by exploring the connectivity. A chain of atoms begins at one end of the torsion and moves through the bonding network. When it reaches a terminal atom or doubles back on itself in a ring system, the atom is flagged and removed from the end of the chain. A flagged atom is never revisited. The process ends when the chain returns to its initial point. All the flagged atoms are then rotated.

6.8 Energetic Parameters

6.8.1 Incremental Charges

When using a forcefield method, it is common to use partial charges on each atom to calculate the electrostatic interaction energy. Thus, the partial charges must be calculated correctly. This is achieved in a quick and simple way by using incremental

charges. For each atom, the charge is modified by a specified value for every atom that is bonded to it. This process is repeated for every atom in every ligand in the population. The incremental charges are taken from the CVFF forcefield [20].

6.8.2 Van der Waals Parameters

The van der Waals ε, A and B values used in Eve are taken from Autodock [12]. These parameters are based on those from the force-fields AMBER [19] and Merck FF [85]. They create an energy-well for each atom-atom pairing that means bonds tend toward their optimal distance with the optimal interaction energy.

6.9 User-Defined Parameters

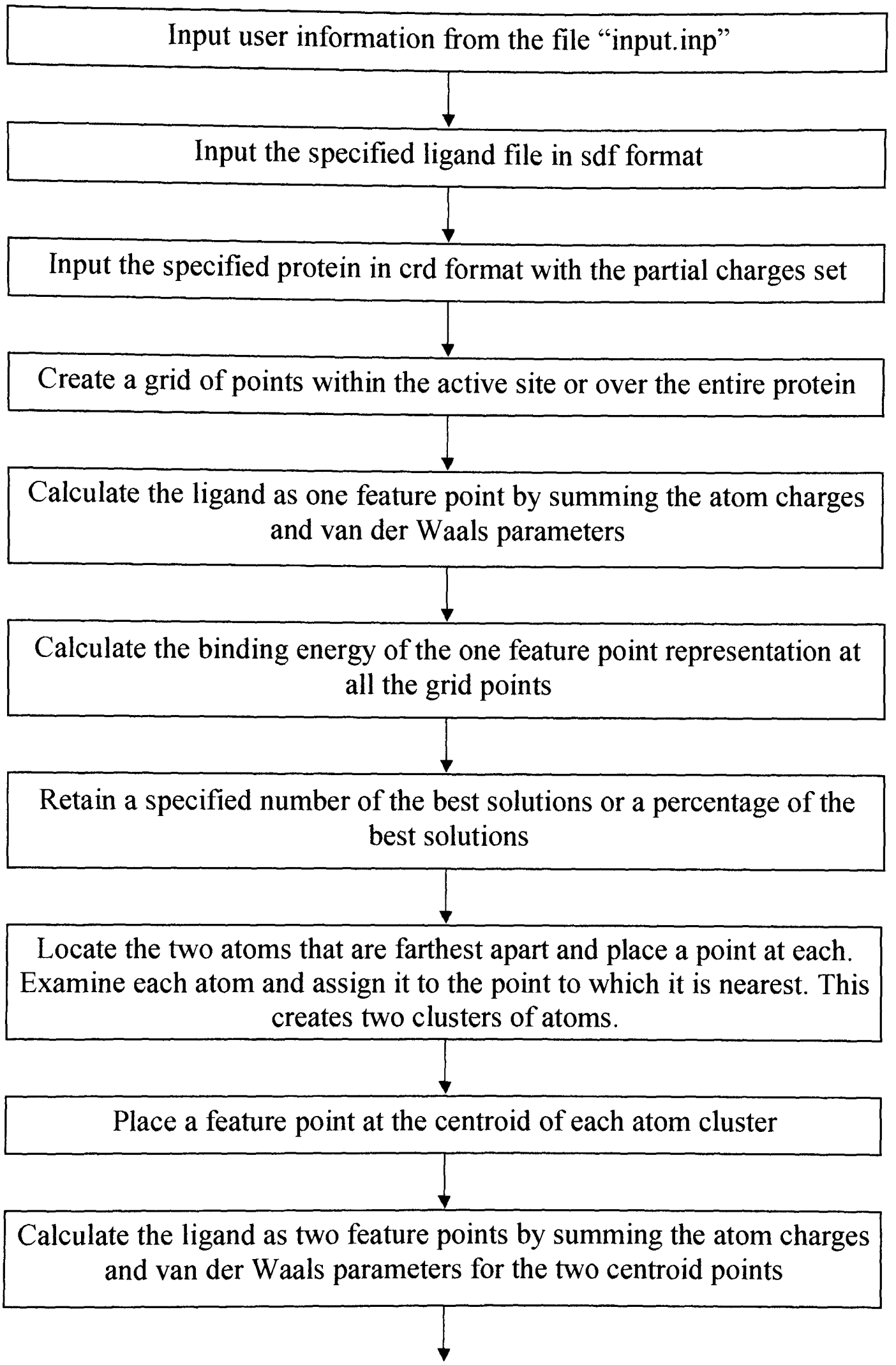
Within a genetic algorithm, there are always a number of parameters that can be optimised to provide a sound speed accuracy trade-off. In this case, there are a huge number:

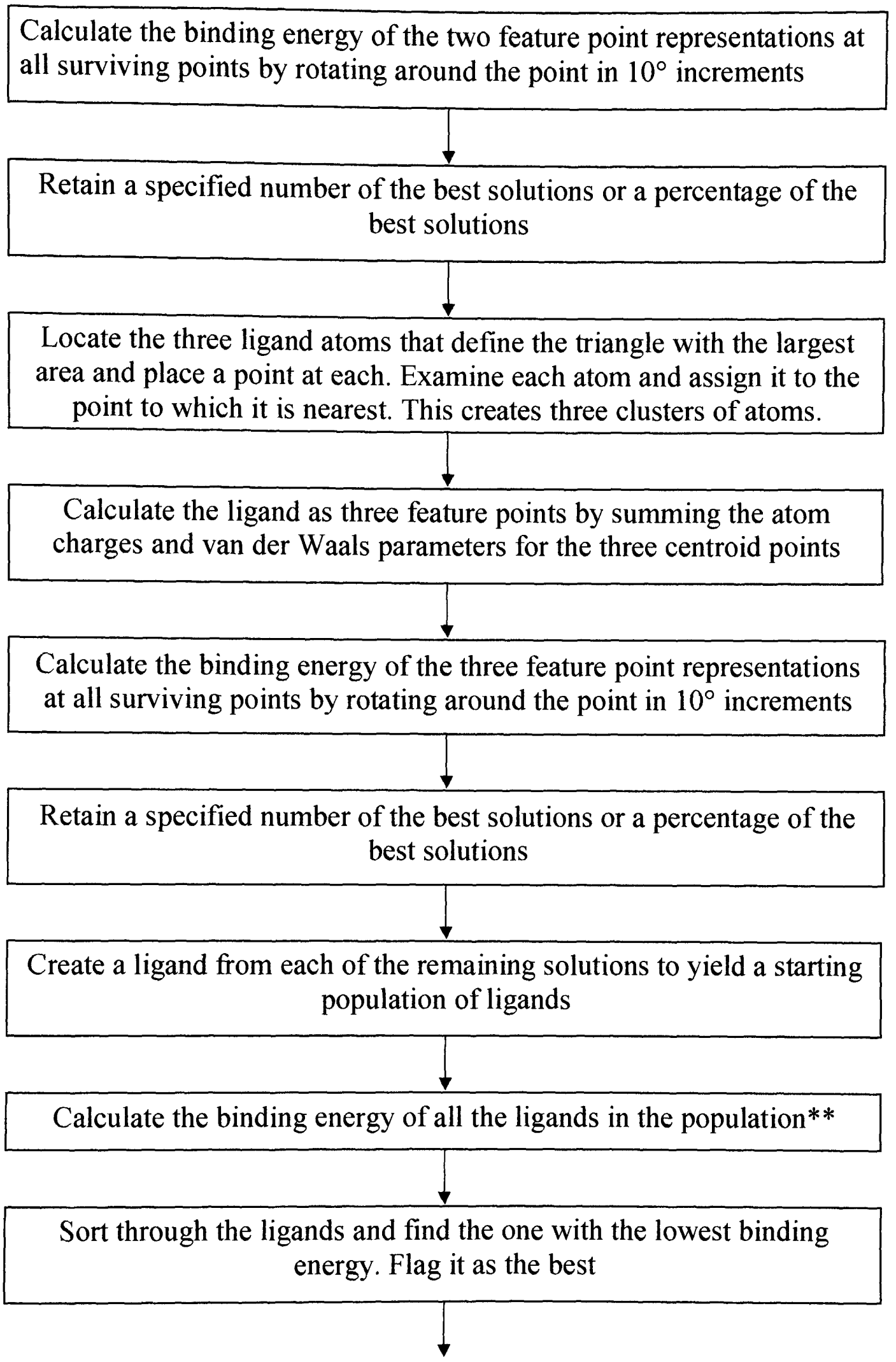
- The incremental angle for conformer generation.
- The energy threshold for conformer generation.
- The distance threshold for conformer clustering.
- The grid resolution.
- The incremental angle for rotation around a point.
- The percentage of solutions that survive after each feature-point.
- The number of generations.
- The Boltzmann factor in the fitness function.
- The relative probabilities for the mutations of a ligand.

Unfortunately, none of these variables has an actual optimum value, as there are infinite cases to which the program can be applied and the best collection of values will vary for each case. However, some parameters remain user-defined and sensible values have been chosen for the remainder. The parameters used in a standard Eve run were chosen by trial and error and can be found in Table 7.1.

6.10 Overview

Molecular docking is a problematical task due to the enormous search space. The algorithm Oxdock simplifies this problem using a multiscale approach to reduce computing time significantly in the initial stages. Unfortunately, despite proving excellent at locating binding sites, Oxdock is relatively poor at optimising ligand binding and scoring docking poses. The algorithm Eve combines the multiscale approach with the optimising ability of Evolutionary Programming. The flowchart in illustrates this.


Input user information from the file "input.inp"
Input the specified ligand file in sdf format
Input the specified protein in crd format with the partial charges set
Create a grid of points within the active site or over the entire protein
Calculate the ligand as one feature point by summing the atom charges and van der Waals parameters
Calculate the binding energy of the one feature point representation at all the grid points
Retain a specified number of the best solutions or a percentage of the best solutions
Locate the two atoms that are farthest apart and place a point at each. Examine each atom and assign it to the point to which it is nearest. This creates two clusters of atoms.
Place a feature point at the centroid of each atom cluster
Calculate the ligand as two feature points by summing the atom charges and van der Waals parameters for the two centroid points


Calculate the binding energy of the two feature point representations at all surviving points by rotating around the point in 10° increments
Retain a specified number of the best solutions or a percentage of the best solutions
Locate the three ligand atoms that define the triangle with the largest area and place a point at each. Examine each atom and assign it to the point to which it is nearest. This creates three clusters of atoms.
Calculate the ligand as three feature points by summing the atom charges and van der Waals parameters for the three centroid points
Calculate the binding energy of the three feature point representations at all surviving points by rotating around the point in 10° increments
Retain a specified number of the best solutions or a percentage of the best solutions
Create a ligand from each of the remaining solutions to yield a starting population of ligands
Calculate the binding energy of all the ligands in the population**
Sort through the ligands and find the one with the lowest binding energy. Flag it as the best

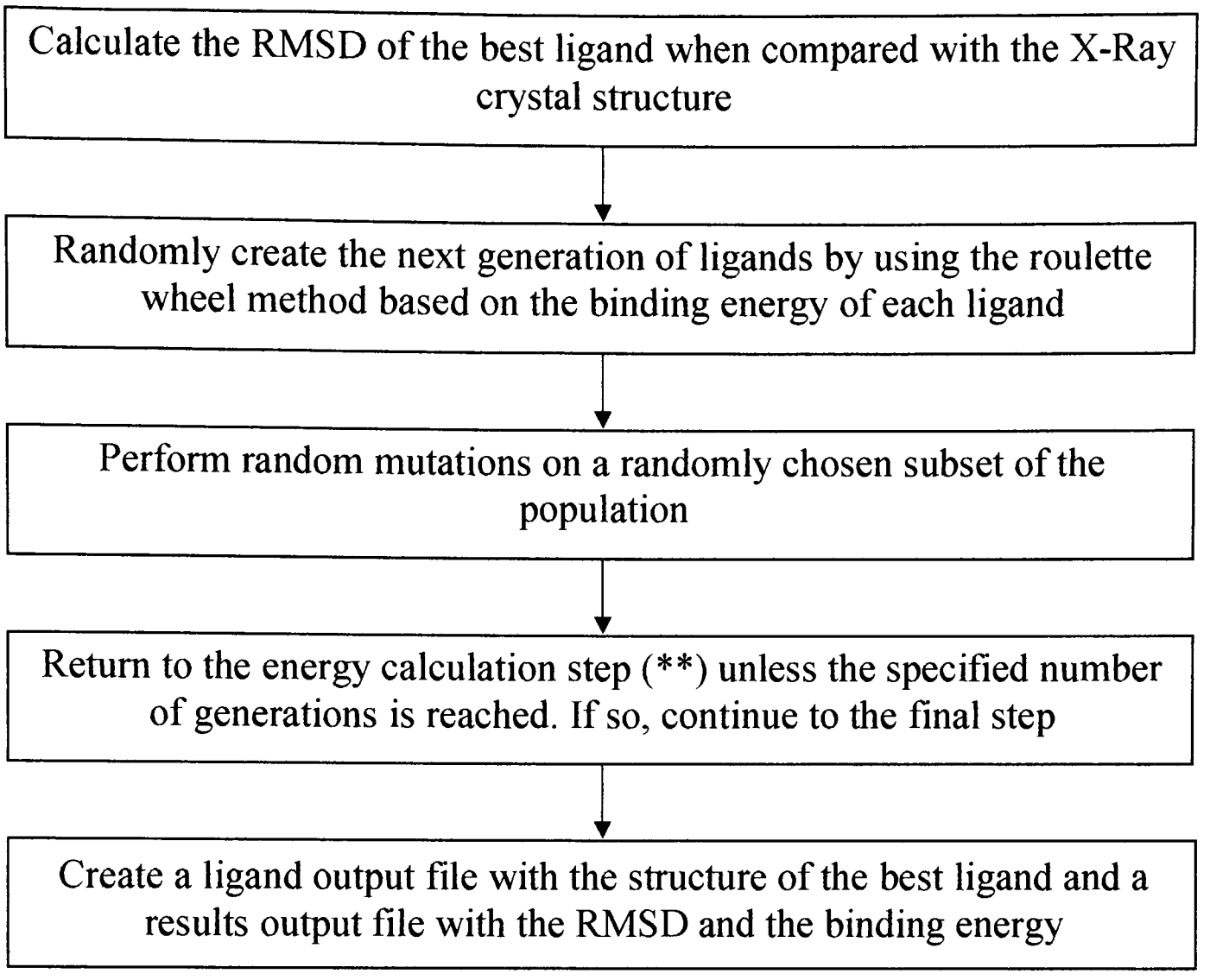


Figure 6.9 - Flow diagram illustrating the Eve algorithm.

The binding energy is calculated using a scoring function that is modified and improved from Oxdock. The electrostatics and van der Waals binding are calculated using the Coulomb equation with a distance-dependent dielectric and the Lennard Jones 12-6 function. The energies are calculated explicitly and not by a grid method. The internal coordinate contribution represents intramolecular forces and includes both electrostatic and van der Waals effects. There are two additional facets to the scoring function. The torsion term considers the loss of entropy on binding and uses data from various studies to parameterise the effect. The hydrophobic term considers the effect of decreasing the amount of solvent accessible surface area that occurs upon

binding. This liberates water molecules into the bulk solution and has an energetic effect that favours binding. This is parameterised by combining calculations of the solvent accessible surface area with data from previous studies.

Eve thus combines a multiscale approach with evolutionary optimisation and a rigorous scoring function to yield a functioning molecular docking algorithm. The improvements to the method allow the problems discussed in section 5.6 to be addressed, and produce an algorithm for the location of active sites and the prediction of binding geometry. Before Eve is used to analyse any biological problems, it must be comprehensively validated. The results of this are discussed in chapter seven.

7 Validation of the New Algorithm

7.1 Overview

In chapters two, three and four, Oxdock is shown to be a useful tool for locating active sites. However, when attempting to calculate the exact docking pose of a ligand with a protein, a more rigorous validation is necessary. The protein databank contains approximately 30,000 proteins (March 2005) and a large number of these contain ligand complexes. To prove the worth of Eve, it is necessary to choose a significant number and variety of these cases to test. The testing involves attempting to predict the docking pose of a ligand accurately given only an estimate of the active site location. The ability of the algorithm to produce low RMSD predictions is vital for its effectiveness. The other essential requirement is that the predicted energy scores are comparable and thus the algorithm is able to highlight lower energy complexes. These two abilities constitute an excellent docking algorithm. In this chapter, the Eve algorithm is extensively validated.

7.2 Methodology

The program developed in the previous chapter is an algorithm that calculates a theoretical binding energy for the complex of a ligand and a protein. It repeats this calculation for a variety of solutions and propagates the best solutions in subsequent iterations. In every generation, a proportion of the molecules are altered in position or structure, allowing optimisation to occur. However, there are a number of important factors to be considered:

- The algorithm must be able to generate reproducible results. It must yield similar results for multiple runs of the same example.
- The algorithm must be adept at searching conformational space to find optimal solutions to the problem. It must be able to find the active site and binding pose of the ligand for a variety of known cases.
- The algorithm must calculate a high binding energy for molecules that are known to bind to a given protein. The crystal structures of a variety of protein-ligand complexes must be scored using the fitness function used in Eve

These tests should help to validate the program and highlight any areas for improvement. Eve reports both the predicted interaction energy in kcal/mol and the heavy atom RMSD of the crystal structure and the calculated structure. The algorithm is validated using the GOLD test set [11]. This test set is used as it represents a specially selected set of ligands and proteins with low-resolution structures that are readily available from the PDB. There are a wide variety of different types of ligands and the results can be easily compared to those of the GOLD algorithm. The GOLD test set has the added bonus that all of the complexes have been analysed in detail for potential problems. This will make analysis of the results much simpler. Three sets of validation runs are used to test the efficiency of the algorithm.

Precision - The test case 1GLP from the GOLD test set is used to test the precision of Eve. The run is repeated for twenty trials. The mean, range and standard deviation of the heavy atom RMSD are reported.

Accuracy - The GOLD test set is used to test the accuracy of Eve. The RMSD between the experimental and calculated positions of the heavy atoms are reported.

Selectivity - Seven ligands that bind to Thrombin are docked to the same structure to evaluate the selectivity of Eve. The lowest energy results for each ligand are reported, along with the results when all seven are docked together.

The 133 ligands in the initial test set were downloaded as PDB files. Each is split into one ligand file and one protein file containing all the amino acid residues and all the heteroatoms. The proteins are converted to crd format, all the hydrogens are added and the partial charges are set using CHARMM [6]. The ligands are converted to sdf files using Web Lab Viewer. The parameters used in the algorithm were chosen based on trial and error in an attempt to produce good and consistent results. The values were kept constant for the validation and can be seen in Table 7.1.

Parameter	Value
Active Site Defined	Yes
Active Site Radius	6.0 Å
Grid Resolution	0.5 Å
Rotation Angle Around Grid Point	10°
Survivors Each Iteration	10 %
Conformer Energy Threshold	0.01 kJ/mol
Temperature Decrease	Yes
Initial Boltzmann Factor	100
EP generations	50
Total Ligands in Population	5000

Table 7.1 – The parameters used in Eve during the validation.

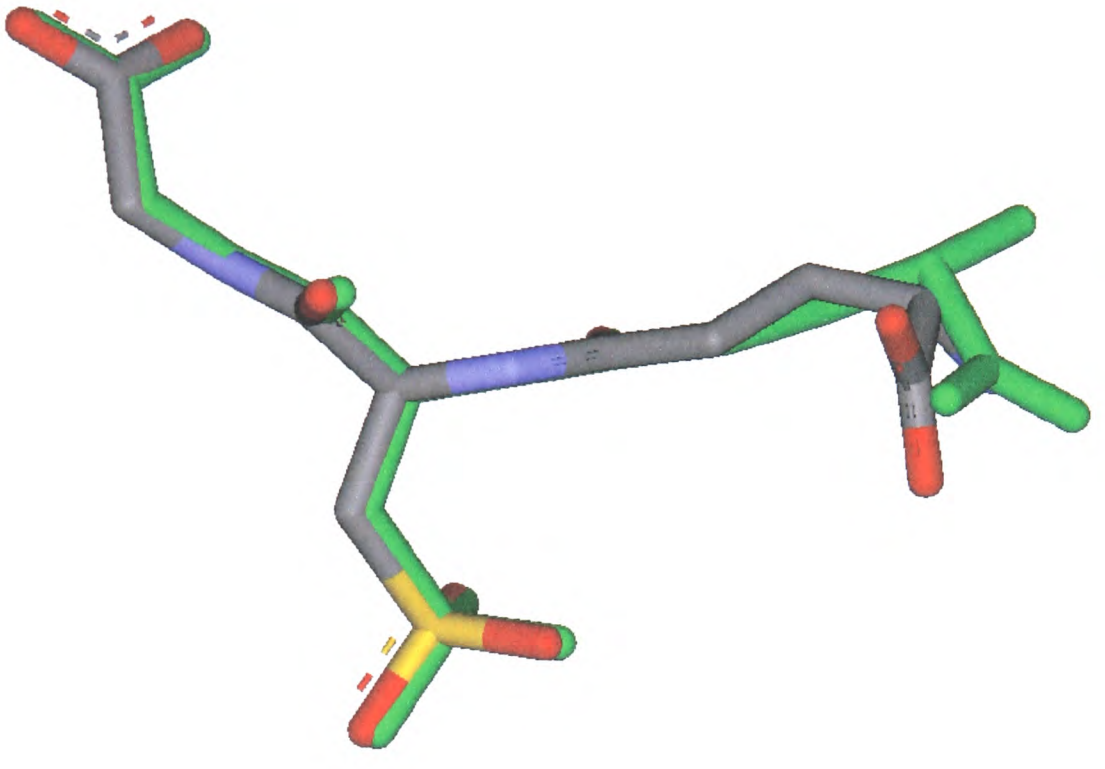
The search space is confined to a 6.0 A° sphere centred on the ligand centroid. This reduces the length of time required for computation and allows the algorithm to search only within the known binding site.

7.3 Results

7.3.1 Precision

Initially, the precision is tested by repeated trials of one test case. The protein-ligand complex 1GLP is used as it contains a relatively large ligand (glutathione sulphonic acid $C_{10}H_{17}N_3O_9S_1$) with ten rotatable heavy atom torsions. This should present a

good case for testing the algorithm. The result of twenty runs of the algorithm on different processors with identical parameters is shown in Figure 7.1 along with an example of the docking.



Trials	20
Mean RMSD (Å)	0.753
Range (Å)	0.723 - 0.822
Standard Deviation (Å)	0.024

Figure 7.1 - The best docking solution (RMSD 0.722 Å) of the complex 1GLP and a summary of all twenty docking results. The crystal structure is shown in green.

The high precision is explicable because the algorithm is deterministic through the multi-scale portion. It will thus yield identical ligand positions and orientations each

time it is run. The non-deterministic EP then swiftly finds the energy minimum near one of the surviving orientations, which has an RMSD of around 0.753 Å.

7.3.2 Accuracy

The high precision of the algorithm removes the necessity of performing multiple runs for each complex. Each pair of ligand and protein is thus run once through Eve and the heavy atom RMSD calculated between the crystal structure and the best result of the algorithm. The results are shown in Table 7.2.

Range (Å)	Eve	GOLD
0<RMSD<0.5	23.6%	8.0%
0.5<RMSD<1.0	35.0%	27.0%
1.0<RMSD<1.5	9.8%	20.0%
1.5<RMSD2.0	5.7%	11.0%
2.0<RMSD<2.5	4.1%	2.0%
2.5<RMSD<3.0	4.9%	3.0%
3.0<RMSD	17.1%	28.0%
Total	123	100

Table 7.2 - Results of the Eve and GOLD test cases, showing the number of results found within each range of 0.5 Å.

Unfortunately, only 123 of the 133 validation complexes can be used in this case due to problems with format of the protein files. Furthermore, only 100 of the GOLD results from this test set were accessible. However, the results from Eve compare very favourably with the available GOLD results, with the great majority of docking poses predicted accurately to within 2.0 Å heavy atom RMSD.

The efficiencies of the algorithm are quite difficult to compare. GOLD docking, on a Silicon Graphics R4400 Indigo II with a 200MHz processor, takes between 3 and 35 minutes to perform, depending on the system studied. However, despite showing better results in terms of RMSD values, Eve docking takes considerably longer. On a 750MHz processor typical run times were between 10 minutes and 20 hours, depending on the system studied. However, further work showed that modifying the multiscale parameters allowed a marked decrease in run time. The rigorous optimisation of these parameters would be an important part of further work on this algorithm.

7.3.3 Specificity

A final test is required to evaluate whether Eve can discriminate sufficiently between disparate ligands and identify the one that binds with the highest interaction energy. Thrombin is used as the test case as it is commonly used as a test in docking and has a number of known inhibitors. The following complexes are thus chosen: 1PPH, 1BRA, 1TNG, 1TNH, 1TNJ, 1TNK, 1TNL. The protein is taken from 1PPH as it contains the active site in the most open conformation. This will clearly bias the results, but should not affect the scores too much, and should allow each ligand to bind. The scores of the

ligands when docked individually are compared in Table 7.3, along with the final solution of a run using all the ligands.

PDB File	**Score (kcal/mol)**
1PPH	-99.01
1BRA	-48.02
1TNG	-47.28
1TNH	-51.16
1TNJ	-51.75
1TNK	-53.39
1TNL	-53.81
All	-94.73 (1PPH)

Table 7.3 - Results of the seven test cases used in the validation of selectivity, giving the calculated energy of binding in each case and for the solution with all the ligands.

The best solution for 1PPH is shown in Figure 7.2, along with the same molecule that is found to be the best solution in the case where all the ligands are tested together. It can be seen that Eve is able to discover the best docking pose in both cases.

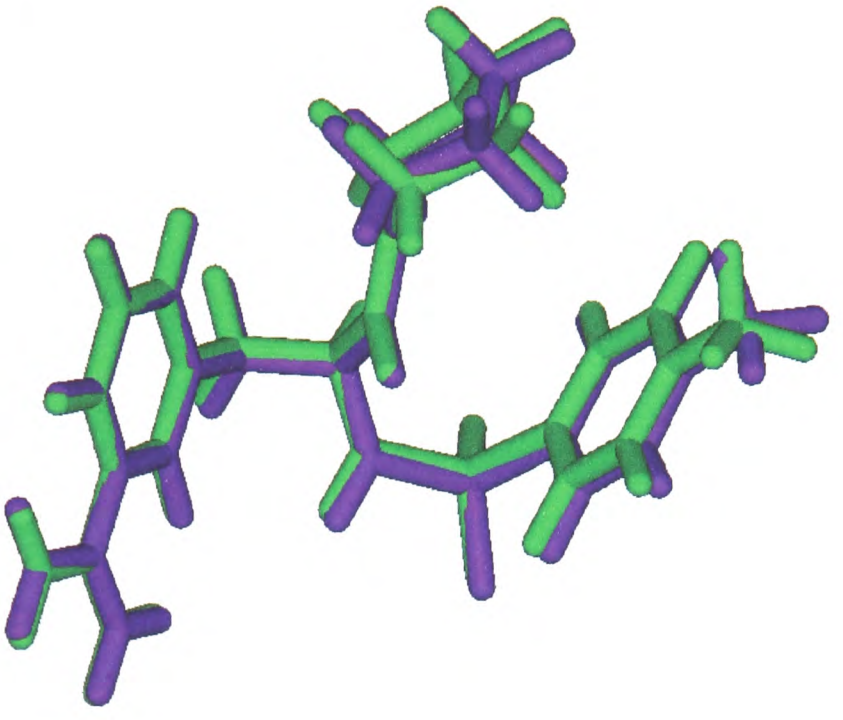


Figure 7.2 - The best docking result of the PDB file 1PPH (in green) compared with the best result when all seven test-ligands are run contemporaneously (in purple).

Unfortunately, the calculated energies are not correct due to ignoring effects such as solvation and desolvation discussed in section 6.6.2.4, but this is not vital if the algorithm can still predict the ligand with the greatest binding energy. In the case of Thrombin, experimental energies show that, as predicted, the ligand from 1PPH has the highest binding energy (-35.52 kJ/mol) with the other ligands having a considerably lower binding energy (in the range -8.50 to -19.22 kJ/mol). The energies do not correlate perfectly between the calculated and experimental values due to a number of issues. Ignoring desolvation effects is one of the main sources of error, as are the simplifications in the scoring function as discussed in section 6.6.4. Another of the problems comes from the use of only one protein structure in the calculation. This affects the calculated binding energy of all the ligands as well as biasing the results.

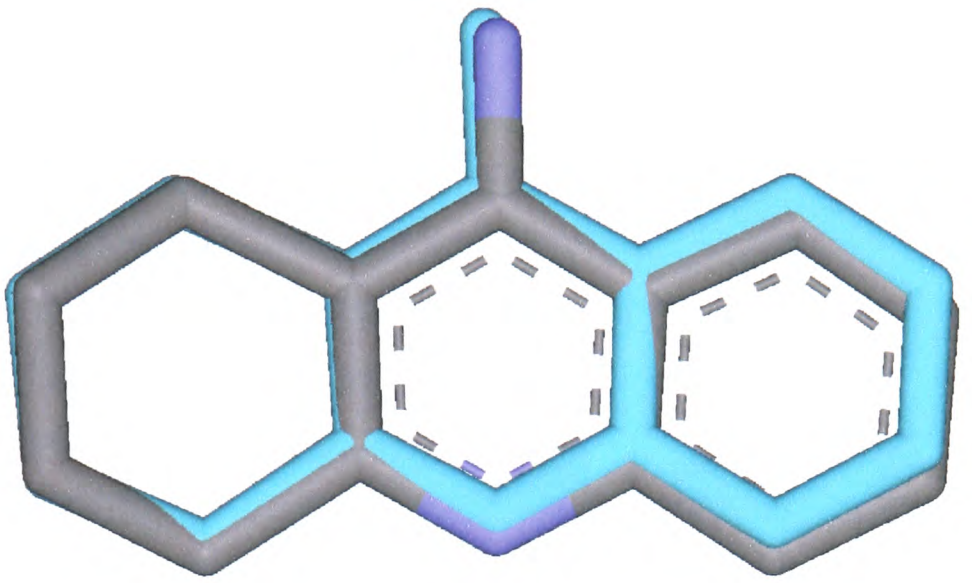
This issue has been partially resolved by the method of cross docking [86]. Every ligand is docked with every protein structure to create a table of predicted docking energies. However, one of the reasons for performing this validation was a proof of methodology and this method is clearly able to search a variety of ligands during the course of one run.

7.4 Difficult Test Cases

Many issues can mean molecular docking algorithms fail to predict the correct docking pose. Three such cases are chosen to illustrate the generality of the Eve algorithm.

7.4.1 Hydrophobic Ligand

The hydrophobic effect can lead to errors in scoring docked ligands. The test case 1ACJ from the Protein Data Bank is used as it contains the hydrophobic ligand Tacrine (1, 2, 3, 4-Tetrahydro-9-Aminoacridine) at low resolution. The run is repeated ten times to allow analysis of the docking results. The results are summarised in Figure 7.3. In this case, there were nine very similar results and one outlier with an RMSD of 0.372 Å, leading to a low standard deviation but a large range. This suggests that perhaps 10 trials were not sufficient in this case.



Trials	10
Mean RMSD (Å)	0.233
Range (Å)	0.194-0.372
Standard Deviation (Å)	0.052

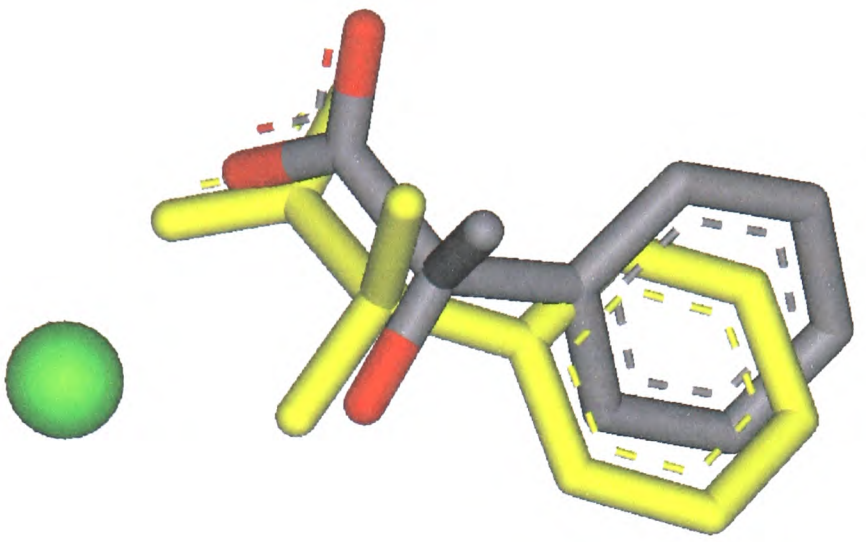
Figure 7.3 - The best docking solution (RMSD 0.194 Å) of the complex 1ACJ and a summary of all ten docking results. The crystal structure is shown in blue.

This result illustrates that Eve can find solutions in cases where the ligand is hydrophobic and has a lower contribution from electrostatic interactions.

7.4.2 Metal Atom in Active Site

Three test cases are used, each containing a different metal ion. The complex with PDB reference 1MDR contains a magnesium ion in the active site, which ligates with

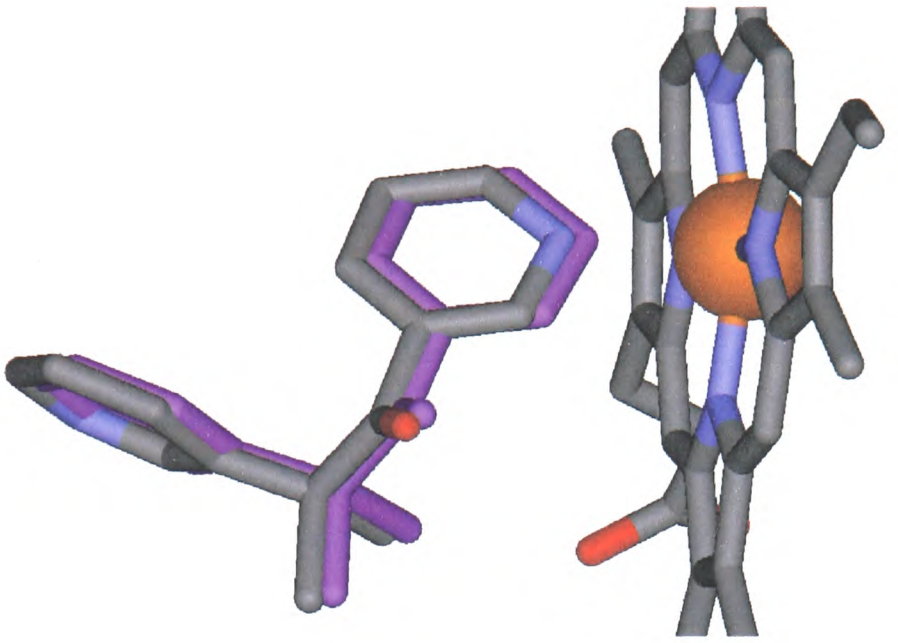
a carboxylate group and a hydroxyl group in the ligand. Eve provides a close approximation to the crystal structure, as can be seen in Figure 7.4.



Trials	10
Mean RMSD (Å)	1.957
Range (Å)	1.092 - 2.473
Standard Deviation (Å)	0.346

Figure 7.4 - The best docking solution (RMSD 1.092 Å) of the complex 1MDR and a summary of all ten docking results. The crystal structure is shown in yellow and the magnesium ion is coloured green.

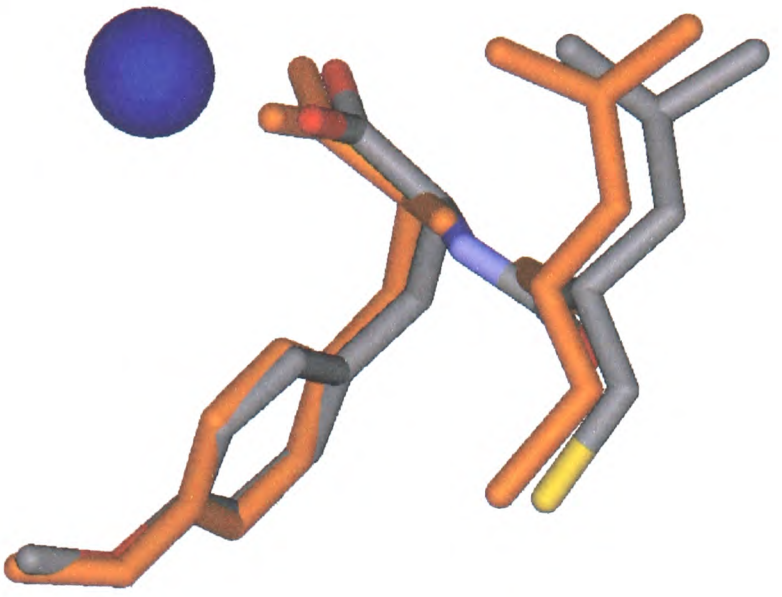
The complex with PDB reference 1PHG contains an iron ion in the protein haem group, which ligates with a ring nitrogen atom in the ligand. Eve again provides a close approximation to the crystal structure, as can be seen in Figure 7.5.



Trials	10
Mean RMSD (Å)	0.459
Range (Å)	0.352 - 0.515
Standard Deviation (Å)	0.052

Figure 7.5 - The best docking solution (RMSD 0.352 Å) of the complex 1PHG and a summary of all ten docking results. The crystal structure is shown in purple and the iron ion is coloured orange.

The complex with PDB reference 1ATL contains a zinc ion in the protein, which ligates with a carboxylate group in the ligand. Eve again provides a close approximation to the crystal structure, as can be seen in Figure 7.6.



Trials	10
Mean RMSD (Å)	0.990
Range (Å)	0.649 - 1.233
Standard Deviation (Å)	0.167

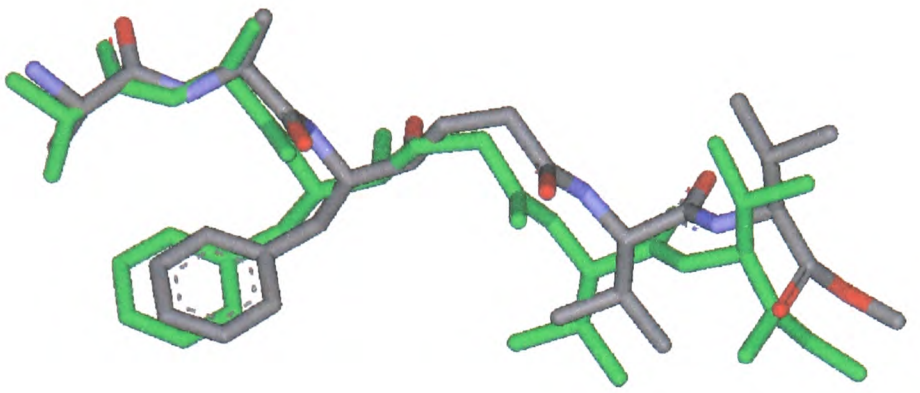
Figure 7.6 - The best docking solution (RMSD 0.649 Å) of the complex 1ATL and a summary of all ten docking results. The crystal structure is shown in orange and the zinc ion is coloured blue.

These three test cases illustrate that Eve can be used in test cases where one of these metal ions is present. The ions calcium, copper, cobalt and manganese can also be used in Eve.

7.4.3 Long Ligand

The complex with PDB reference 1AAQ contains the protein HIV protease, for which the active site is a long hydrophobic cavity. The ligand is a long hydrophobic peptide-

mimic. Eve again provides a close approximation to the crystal structure, as can be seen in Figure 7.7.



Trials	10
Mean RMSD (Å)	1.210
Range (Å)	1.052 - 1.617
Standard Deviation (Å)	0.171

Figure 7.7 - The best docking solution (RMSD 1.052 Å) of the complex 1AAQ and a summary of all ten docking results. The crystal structure is shown in green.

These results show that Eve can be a useful molecular docking tool in cases where the ligand is long. This is often an issue for docking algorithms, as more orientations may need to be searched to ensure that the correct orientation is sampled.

7.4.4 Remaining Problems

Despite the success of Eve in predicting binding geometries in a wide variety of test cases and a good performance in comparison with GOLD, there are a number of problems that still remain, which adversely affect the ability of the algorithm to make accurate predictions. These will now be briefly discussed.

- Unusual Protonation States – Eve assumes that all Glutamate and Aspartate residues are deprotonated and that all Arginine and Lysine residues are protonated. This is not always correct and explains the problems with complexes such as 1EED (RMSD 12.055 Å). This is the worst result in the test set and illustrates the importance of electrostatic interactions in binding proclivity. GOLD also performs very poorly with this test case (Average RMSD 10.06 Å).
- Unusual Molecules – There are some cases that contain rather odd ligands or hetero-groups. The complex from PDB file 6RSA has an RMSD of 7.091 Å and contains a Vanadate group as part of the ligand. The partial charge assignments and van der Waals parameters for this system are suspect and thus the predicted binding geometry is incorrect. GOLD also performs poorly in this case (average RMSD 4.50 Å)
- Protein Irregularities – There are a number of test cases in the GOLD test set that contain problematic X-ray structures. The PDB file 1HEF contains a steric clash with an oxygen and a nitrogen atom separated by 2.03 Å. This leads to a very high RMSD of 6.359 Å.

- Bound Water Molecule – There are a number of PDB complexes in the GOLD test set that contain a water that is bound to the ligand. In GOLD, many of these waters were left in the protein structure. This was not done in Eve and explains the errors in prediction for a number of cases, including PDB files 2CTC (RMSD of 3.383 Å) and 1DIE (RMSD of 4.163 Å).

It would be nice to repeat this work and address some of these issues. Assigning protonation states and retaining vital water molecules would significantly reduce the number of errors in the test set and updating atom parameters for the metal ions would also improve the performance. Any irregularities in the protein files could also be tackled prior to the run or simply removed from the test set.

7.5 Summary

The results of this validation show that Eve is potentially a useful tool for rational design. The algorithm is essentially a docking algorithm inside an evolutionary algorithm. It can thus be used to discover optimal binding poses by mutating the position of the ligand by rotation, translation or torsion modification. The method searches through many conformations and positions to find its optimal geometry. The entire protein surface can be searched to locate the binding site or the active site can be defined by a box or sphere of given dimensions. It is important that the energy function used is accurate and can realistically model the interactions of ligands and proteins. Ligand docking is used to test the algorithm for effectiveness. Protein-ligand complexes are downloaded from the Protein Databank and the proteins and ligands separated. The ligands can then be docked with the protein using Eve to discover:

- If the algorithm can make reproducible predictions of the docking pose.
- If the algorithm can correctly predict the docking-pose for a variety of test cases.
- If the algorithm can select the highest binding energy ligand from a variety of ligands that are known to bind to the target.

A standard test set is used as this allows comparison with other methods and ensures a good spread of examples from a wide variety of cases. The results show that Eve is precise, accurate and specific. Further testing proves that Eve can be used for test cases involving hydrophobic ligands, proteins with metal atoms and long ligands. Now that Eve has been shown to be an accomplished docking algorithm with wide applicability, its various useful applications can be considered.

8 Applications of the New Algorithm

8.1 Overview

In chapter seven, it was proved that Eve is an accurate, precise and specific algorithm. It was shown that Eve could discover the binding pose for a large number of protein-ligand complexes. This chapter moves on to demonstrate that this as a useful ability with a wide range of applications. Firstly, as illustrated in chapters 2, 3 and 4, docking methods can be employed in the identification of binding sites and the elucidation of protein function. Eve can be used in its automatic active site detection mode to help solve this problem. This process has been proved to be effective for Oxdock using a multi-scale approach [58] and is improved with the added optimising power of EP. Secondly, as proved in the validation of Eve, evolutionary docking methods can discriminate between various ligands and chose the most potent inhibitor. This ability is vital for any high-level drug discovery program.

The third useful application is the exploitation of computational ligand probe molecules. Ligand probes can be used to identify binding sites using a process similar to the Multiple Copy Simultaneous Search (MCSS) method [87]. Hydrophobic probes are able to find active sites and areas that may facilitate protein-protein interactions. Probe molecules can also be used to aid in rational drug design by highlighting areas that may allow the introduction of extra functionalities to increase binding affinity [88]. This concept leads by a logical progression to the idea of molecular evolution. Lead optimisation is shown to be an ideal task for EP. The known ligand is given initial coordinates and is altered by changing its structure as well as optimising its

position. A number of possible structural mutations are considered. The stage beyond this one is *de novo* ligand design. The active site is defined by either a box or a sphere and a new ligand is grown inside. Initially, functional groups and rings are tested and then the molecule is sequentially grown to optimise the interactions with the protein. The accuracy of the scoring function has already been proved by the validation procedure, and thus molecules that are predicted to bind strongly are highly likely to be good inhibitors. In this chapter, the applications described above are considered, showing the wide applicability of algorithms such as Eve.

8.2 Locating Active Sites (Known Ligand)

As highlighted in chapters 3 and 4, the process of locating active sites is a very useful technique. New protein structures are being discovered at an increasing rate and many have a known function but unknown active sites. The ability to rapidly search the protein surface to dock a given ligand molecule often allows the active site to be found, providing further insight into both the function and the action of the protein. This process is tested using Eve by considering the test case of 1FTJ with glutamate binding to a rat glutamate receptor. The active site is not defined and thus the entire locale of the protein is searched with a grid resolution of 0.5 Å. The results of the first three feature-point representations can be seen in Figure 8.1.

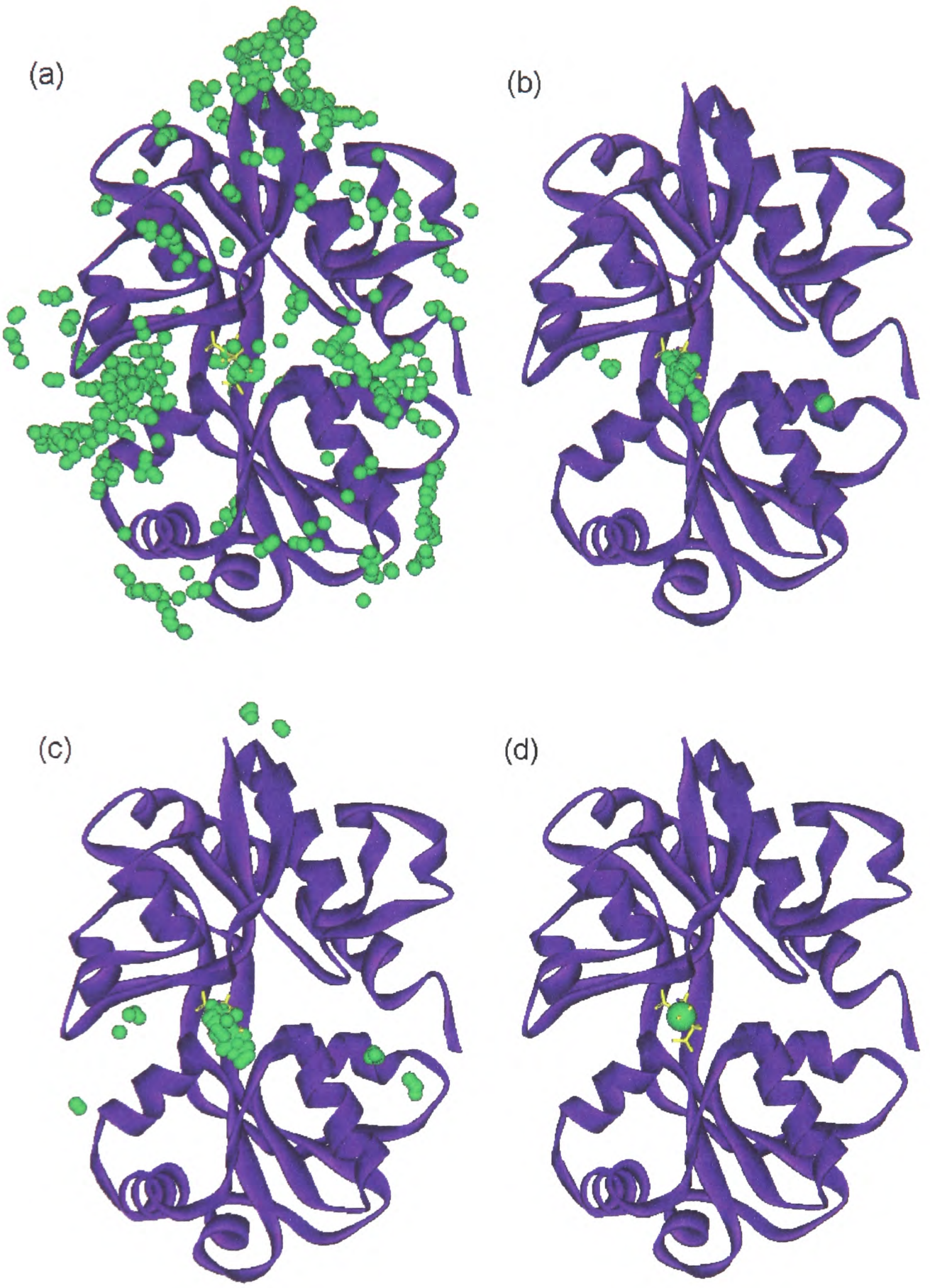


Figure 8.1 - The 5000 lowest energy feature-points from the one feature-point (a), two feature-point (b) and three feature-point (c) representations of glutamate docking with the ionotropic glutamate receptor from the PDB file 1FTJ. The centroid of the lowest energy solution after docking is complete is also shown (d). In all cases, the protein is represented as a

purple ribbon, the points as green dots and the crystal structure of glutamate as a yellow ball and stick molecule.

The active site is pinpointed by the algorithm during the two feature-point representation. The majority of the 5000 solutions that then survive to become the first generation of the evolutionary algorithm are clustered in the active site. The lowest energy solution is shown in Figure 8.2.

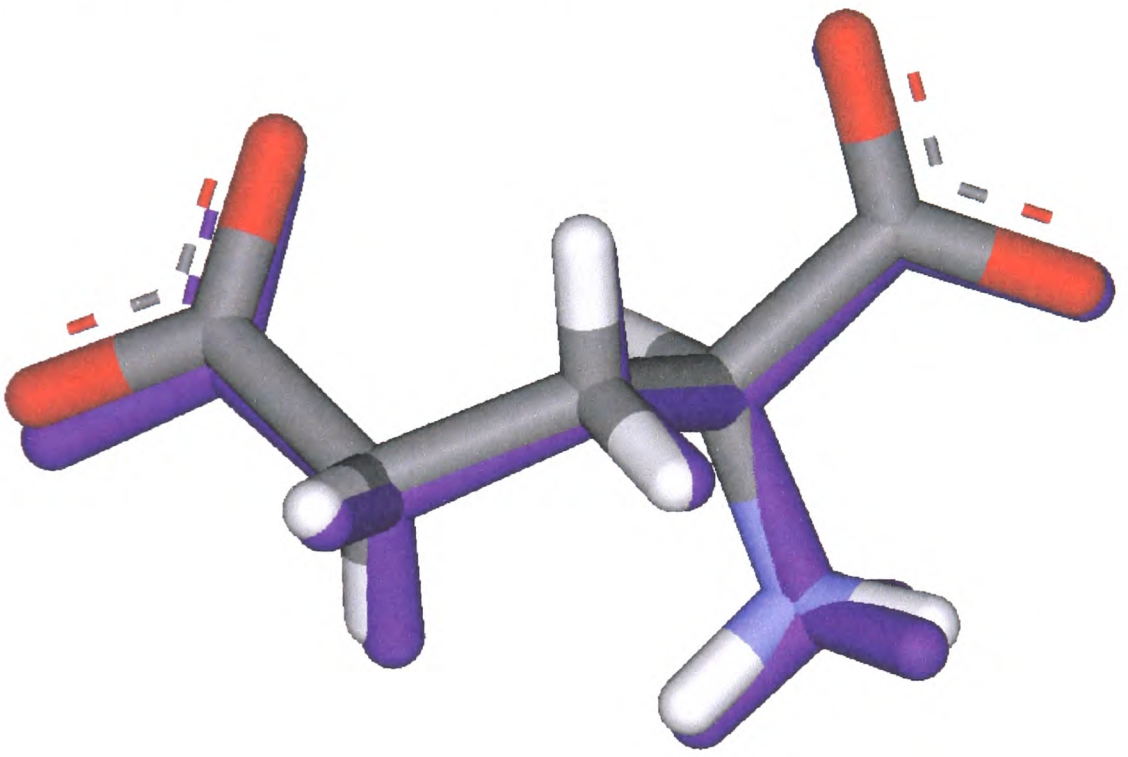


Figure 8.2 - The docked conformation of glutamate bound to the glutamate receptor in PDB file 1FTJ. The crystal structure is coloured purple. The heavy atom RMSD between the actual and calculated structures is 0.181 Å.

This run took approximately 5 hours on a 2.8GHz processor. This is a lengthy docking calculation, but the algorithm searches the entire protein surface and makes a prediction that is in excellent agreement with the X-Ray structure. The results show that Eve, like Oxdock, can be a very useful tool in the analysis of protein structure to

discover the location of binding sites. As illustrated in chapters 3 and 4, this can then aid in the elucidation of protein function

8.3 Locating Active Sites (Unknown Ligand)

As discussed previously, new protein structures are being solved at a great rate. However, in some cases, nothing is known about the function of the protein, or the function is unclear. In these cases, ligand probes may be used to discover the binding site of the protein. The process of employing computational ligand probes was used in the MCSS program [87] and has also been used experimentally [89] by crystallising protein structures in the presence of various small probe molecules. The hydrophobicity of sites has particular importance in binding. When a ligand binds in a hydrophobic pocket, considerable amounts of hydrophobic surface are buried. This favours binding due to the hydrophobic effect, which is discussed in section 6.6.3.2. The docking power of Eve can thus be used to find the location of hydrophobic pockets on the ligand. A hydrophobic molecule can be docked at grid points surrounding the protein; the best solutions can be related to their initial grid point once the run is completed and these returned as the result.

Consider HIV protease with a benzene probe. This protein contains a number of hydrophobic pockets in the binding site. The complex with BILA 1906 (PDB ID 1IDA) is used to provide the protein. The lowest energy 50 grid-points can be seen in Figure 8.3.

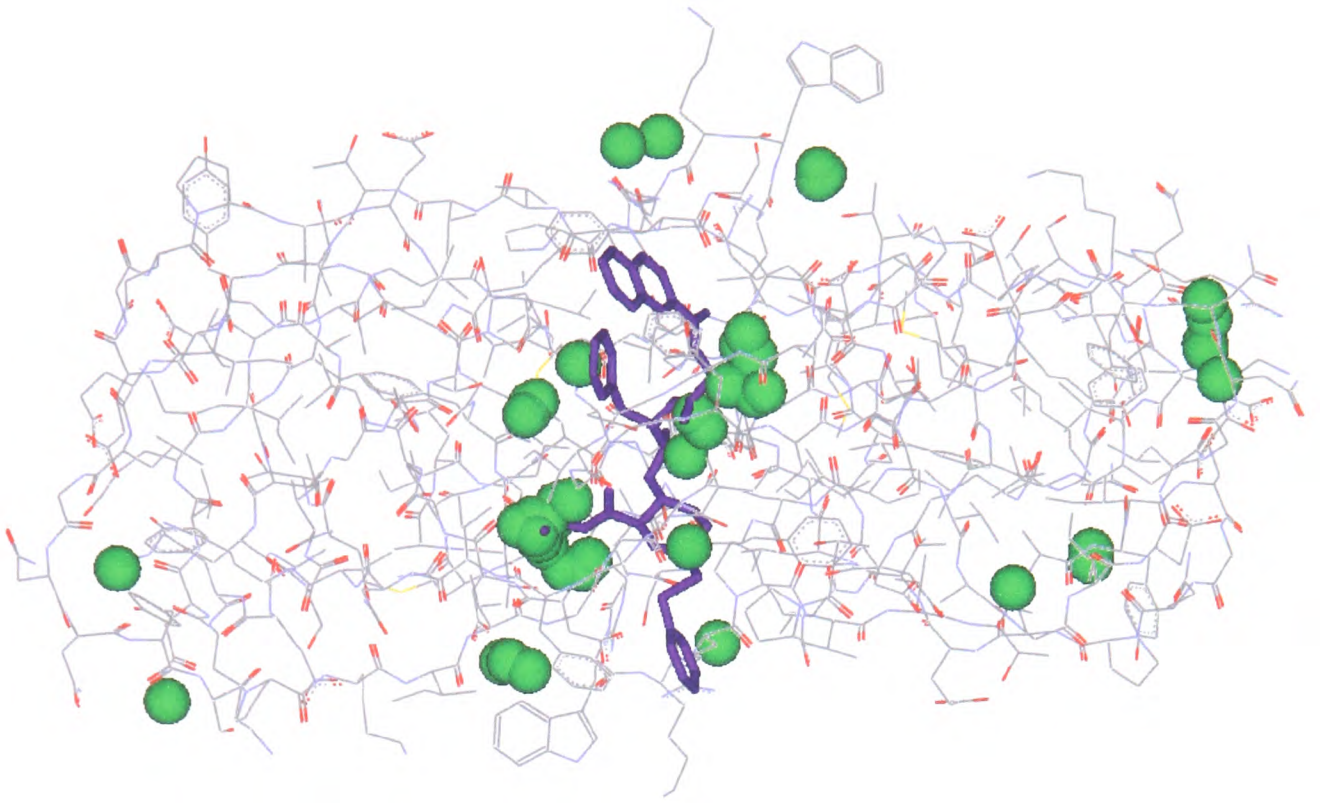


Figure 8.3 - The HIV protease molecule bound with the ligand BILA 1906 (PDB ID 1IDA) in purple. The 50 lowest energy points from the docking with benzene are shown with the centroids as green dots.

The majority of these points are clustered in the binding site, with scattered points found in other clefts. There is a large clustering of points around the two sites where the ligand molecule has a tertiary-butyl and an isopropyl group. This illustrates that docking computational ligand probes can be a useful technique for the location of binding sites. Probes such as this can also be useful in mapping binding sites to facilitate lead optimisation, as shown in the next section.

8.4 Active Site Mapping

The methodology employed in MCSS has been applied successfully to the rational design of the influenza virus neuraminidase inhibitor 4-guanidino-Neu5Ac2en

(Relenza). Neuraminidase is expressed on the viral cell surface and its function is to attach to the host cell. The natural ligand is Sialic Acid. Computational analysis of the Sialic Acid binding site with the program GRID [88], using ligand probe molecules, highlighted a large negatively charged cleft in close proximity to the ligand. This cleft was exploited by the addition of a guanidinium group to the ligand, with marked success. The algorithm Eve is employed in an attempt to replicate these findings. An acetic acid probe is used to test the algorithm and a guanidinium probe is used to find the negatively charged pocket. The protein is from the PDB file 2BAT, which contains Sialic Acid bound to Neuraminidase. The grid is restricted to the binding site inside a 6.0 Å sphere centred on the Sialic Acid centroid. The results are shown in Figure 8.4.

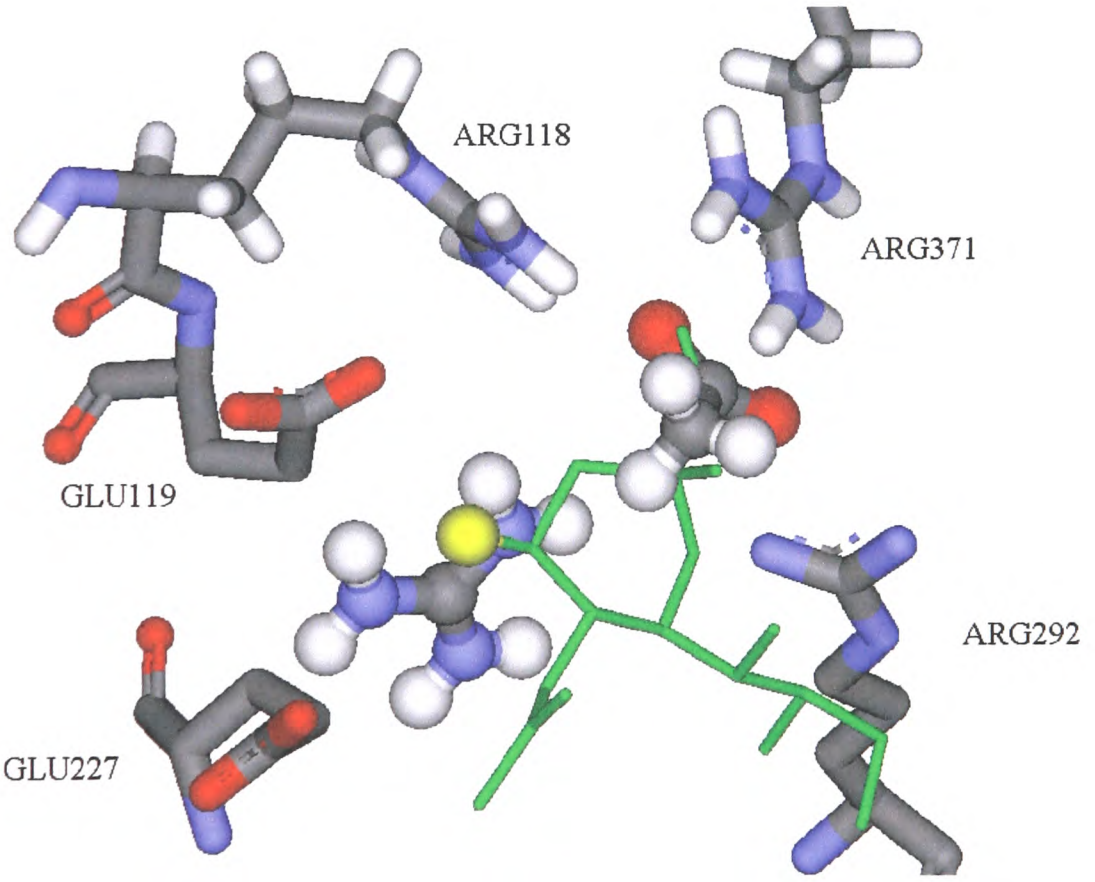


Figure 8.4 - The Sialic Acid binding site of the neuraminidase protein, highlighting five key binding residues. The Sialic Acid molecule is

coloured green and the ligand probe molecule are represented as atom coloured ball and sticks. One of the ligand oxygen atoms is coloured yellow.

The acetate probe molecule bound in the active site between three Arginine residues, in an almost identical conformation to the carboxylate group in Sialic Acid. The guanidine group bound in the negatively charged cleft, between three acidic residues (only two are shown for clarity). The oxygen atom that is coloured yellow in Figure 8.4 is the site that is replaced by a guanidine group in Relenza. Due to the small size of the probe molecules, these calculations take less than 10 minutes on a 766MHz processor. These results show that Eve can be used to map active sites computationally and may be useful in rational design. A logical progression suggests that this process can be automated and used for lead optimisation.

8.5 Ligand Design Methodology

The obvious next step in the use of Eve is to combine ligand docking and ligand growth into the process of ligand design. Instead of simply optimising the position of the ligand, the structure of the ligand can also be optimised. Over many generations, the ligand can grow and mutate, subject to strict rules. De novo design is a widely used technique in medicinal chemistry. Methods to perform this task were suggested in the 1980s and then implemented in the early 1990s [90]. Common algorithms that perform de novo design are LUDI [91], HOOK [92] and Sprout [93]. There are two common methodologies used in molecular design. The first method involves growing a ligand from a starting fragment docked in the binding site, using incremental construction in a similar way to the docking algorithm FlexX [94]. The second

method functions by connecting a number of fragments docked in the binding site by building a molecular skeleton. Both techniques have been shown to work and have produced some interesting and useful results. Congreve et al [95 162] have completed a more complete review of de novo design methods. The molecular design implementation within Eve operates using an incremental construction method to grow a potential inhibitor. The ligand can be altered by the mutations described below, as well as those in section 6.7 which optimise the ligand position, orientation and conformation:

- Atom Deletion – one of the terminal atoms is removed at random.
- Atom Mutation – a random atom is mutated into another atom.
- Addition of a Functional Group – a functional group is added to the molecule
- Addition of a Ring System – a ring system is added to the molecule

These mutations are discussed in the following sections.

8.5.1 Atom Deletion

In some cases, an atom added early in the optimisation process may become disadvantageous later due to rotation or translation of the molecule to an overall more advantageous position. Thus, a list of all the terminal atoms is kept by the program and this method removes one of these atoms at random.

8.5.2 Atom Mutation

This method mutates a randomly selected atom into another atom (with the same or greater maximum valency to prevent hypervalence). The identity of the new atom is

chosen at random by a roulette wheel method. The relative fitness of each atom is one of the program inputs. Many of the atoms and functional groups that are commonly found in organic compounds can also be found in drug-like molecules. However, they occur with widely differing frequency. In an attempt to ensure optimal evolution, a database of 1239 drug molecules is examined by TSAR [96] (a program for analysing molecular structure) to find the relative prevalence of each atom and functional group. The results are used to determine the relative fitnesses for each atom and can be seen in Table 8.1.

Atom	**Frequency**
Bromine	30
Carbon	3500
Chlorine	380
Fluorine	270
Iodine	30
Nitrogen	1020
Oxygen	1550
Sulphur	260

Table 8.1 - The relative fitnesses of the atoms used during mutation in Eve, calculated by analysis of a database of ligand molecules.

However, certain constraints are placed upon the altered atom depending on the identity of the surrounding atoms. The bonds listed in Table 8.2 are not permitted, as they are not stable in aqueous solution.

Non-Permitted Bond	**Name**
O–X	Oxy-halide
O–O	Peroxide
N–X	Nitrogen Halide
S–X	Sulphur Halide
(C=O)–X	Carbonyl Halide
(C=S)–X	Thionyl Halide
(C=N)–X	Imino Halide
(N/O/S)–C–(X/N/O/S)	α - Leaving Group

Table 8.2 - The rules for non-permitted bond used when building a potential ligand using Eve. X represents any of the four halogen atoms: I, Br, Cl and F.

8.5.3 Addition of a Functional Group

The growth of the ligand occurs by addition of atoms, rings and functional groups. A list is kept of all the atoms in the molecule that are addable and one of these is chosen at random. The identity of the new functional group is then chosen at random by a roulette wheel method. However, certain constraints are placed upon the new atom

depending on the identity of the original atom. The non-permitted bonds can be seen in Table 8.2. The relative fitness of each group is calculated using TSAR and can be found below in Table 8.3.

Functional Group	Frequency	Functional Group	Frequency
Ring system	1000	Iodo	30
Carbon atom (sp3)	500	Phosphonate	30
Oxygen atom (sp3)	400	Nitro	20
Oxygen atom (sp2)	1550	Sulphone	20
Amino	1020	Thioamide	20
Amido	575	Thiourea	20
Amidino	500	Disulphur	10
Chloro	380	Dithioester	10
Carboxylate	340	Hydroxylamine	10
Guanidino	300	Phosphate	10
Fluoro	270	Sulphate	10
Ester	260	Sulphonate	10
Sulphur atom (sp3)	260	Sulphonic	10
Peptide Link	200	Sulphur atom (sp2)	5
Methylamino	200	Thiocarboxy	5
Methylformate	200	Thioester	5
Urea	200	Cyanate	2
Carboxyl	140	Isocyanate	2

Sulphonamide	110	Oxime	2
Imide	55	Phosphinic	2
Carbon atom (sp2)	50	Phosphodithionate	2
Ethyne group	50	Phosphothionate	2
Carbon atom (sp)	50	Thiocarboxylate	2
Ethene group	50	Thiocyanate	2
Carbamate	40	Allene	1
Sulphoxide	40	Nitroso	1
Bromo	30	Thioisocyanate	1
Cyano	30		

Table 8.3 - The relative fitnesses of the functional groups used in Eve to build molecules, calculated by analysis of a database of ligand molecules.

The groups are added at a randomly chosen position, providing that they satisfy the bond considerations in Table 8.2. The exact geometry of the bonding must occur within certain constraints. All bond lengths and angles are taken from the CVFF force-field [20] and any unfixed torsions are chosen at random. Obviously, each bond length and angle in every actual molecule has a different magnitude and this method only uses an average length based on the atom type. However, the deviations are rarely too gross and the ligands should have realistic geometries. The calculation of bonding geometry is discussed in section 8.5.6.

8.5.4 Addition of a Ring System

Ring functionalities are very common in drug molecules, as the lack of rotatable torsions in a ring means that there is less unfavourable reduction in entropy upon binding. Aromatic rings are particularly prevalent due to their hydrophobic nature leading to favourable energetics upon desolvation. When adding a ring system, an addable atom is selected at random from the list created for the molecule. There are four types of scaffolds that can be created, depending on whether the atom chosen is already part of a ring system. The four scaffolds are shown in Figure 8.5.

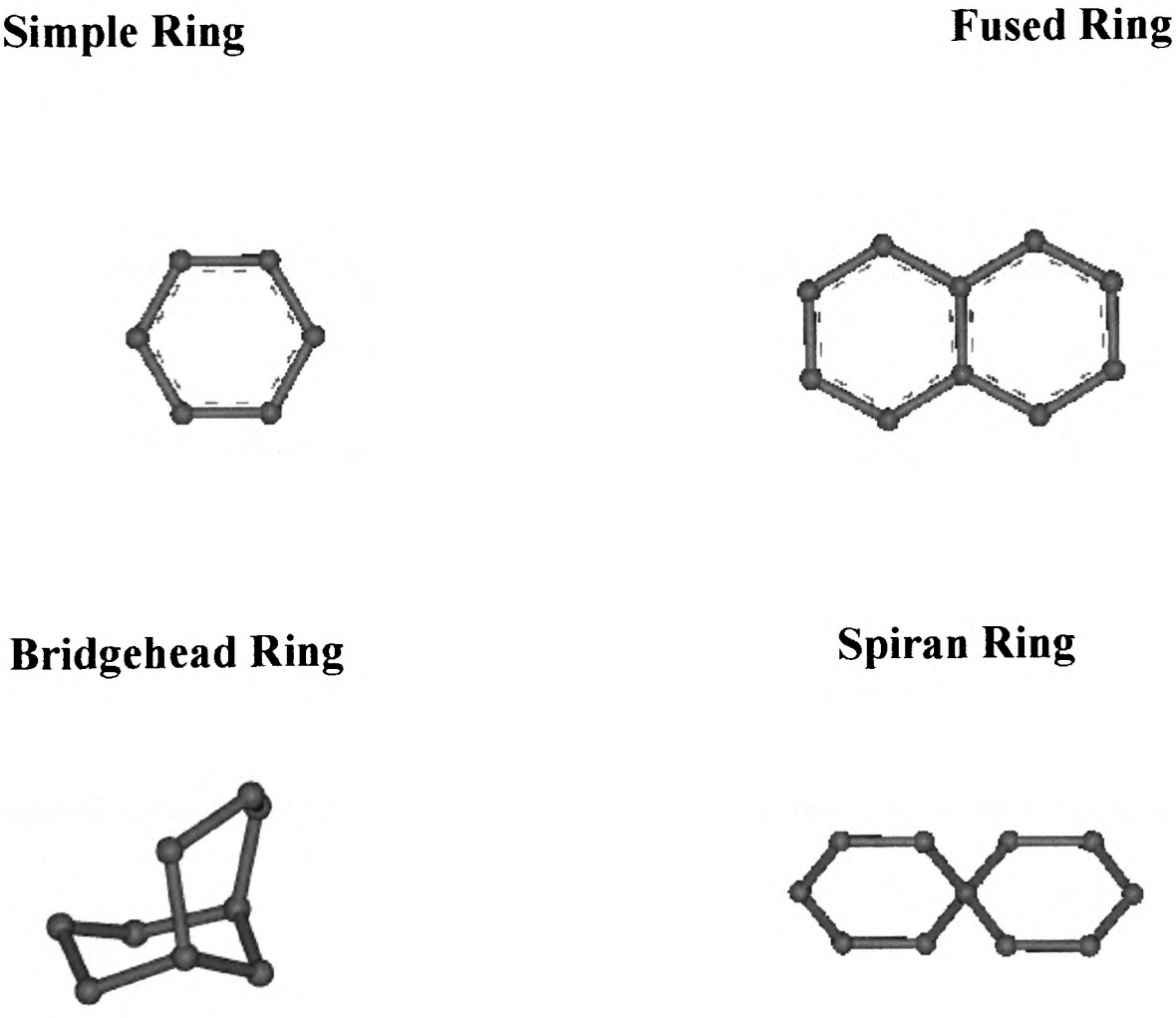


Figure 8.5 - Illustrations of the four ring scaffolds used in Eve to build molecules.

The prevalence of these four ring types is not directly calculable by the TSAR program for calculating molecular descriptors. An estimate is thus made based upon inspection, yielding the following results in Table 8.4

Ring Scaffold	**Relative Fitness**
Simple	700
Fused	400
Spiran	1
Bridgehead	10

Table 8.4 – The relative fitnesses of the ring scaffolds used in Eve to build molecules, calculated by analysis of a database of ligand molecules.

Within each ring scaffold, there are a huge variety of possible ring sizes and types. Only rings with three, four, five and six atoms are considered in Eve, to limit the complexity of the problem. Larger rings could be created by an intra-molecular join or introduced into the program later. TSAR is used to calculate the frequency of the differing ring sizes in the drug molecules studied. The results are shown in Table 8.5

Ring Size	Relative Fitness
Six	2148
Five	665
Four	10
Three	49

Table 8.5 – The relative fitnesses of the ring sizes used in Eve to build molecules, calculated by analysis of a database of ligand molecules.

Once the ring scaffold and ring size have been chosen, there is still a huge variety of possible types available. These are illustrated in Figure 8.6 for the case of a simple six-membered ring:

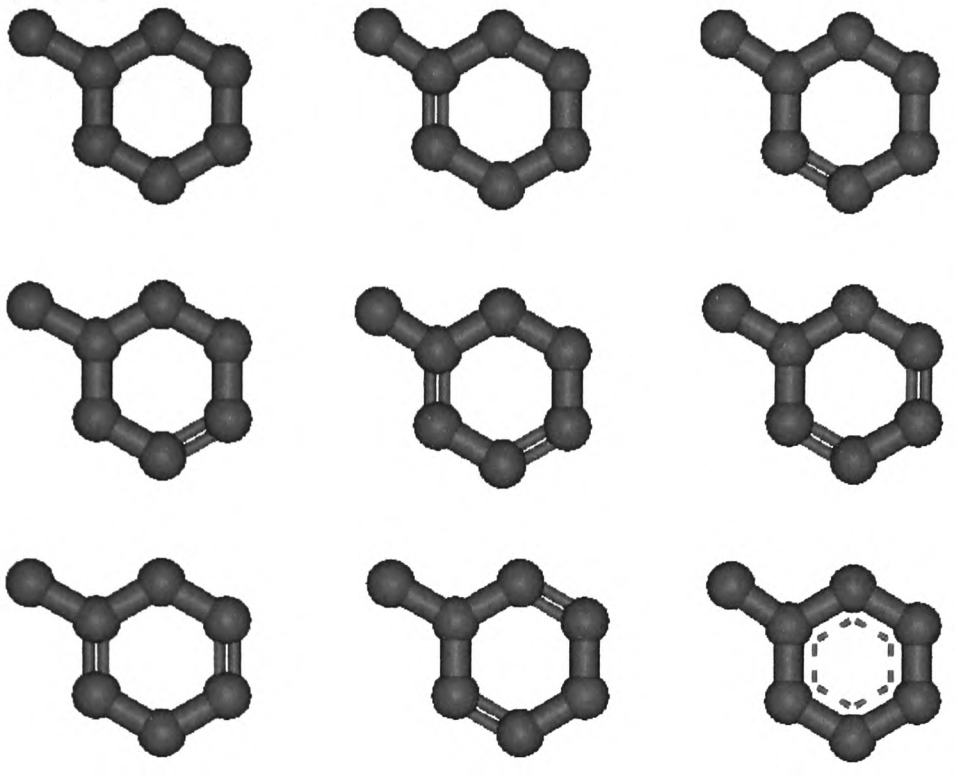


Figure 8.6 - Example of the possible different ring types for a given scaffold and size.

Furthermore, the simple cyclohexane ring system has chair and boat conformations. All of the above types must be considered to provide a complete search of the sample space. The 38 basic ring types can be found in Table 8.6, along with their relative fitnesses.

Ring Type	Relative Fitness	Ring Type	Relative Fitness
Cyclohexane (boat)	11	Cyclopentene 23	7
Cyclohexane (chair)	40	Cyclopentene 34	7
Cyclohexene 12	9	Cyclopentene 45	7
Cyclohexene 23	9	Cyclopentene 51	7
Cyclohexene 34	9	Cyclopentadiene 1234	8
Cyclohexene 45	9	Cyclopentadiene 2345	8
Cyclohexene 56	9	Cyclopentadiene 3451	8
Cyclohexene 61	9	Cyclopentadiene 4512	8
Cyclohexadiene 1234	8	Cyclopentadiene 5123	8
Cyclohexadiene 2345	8	Cyclobutane	7
Cyclohexadiene 3456	8	Cyclobutene 12	1
Cyclohexadiene 4561	8	Cyclobutene 23	1
Cyclohexadiene 5612	8	Cyclobutene 34	1
Cyclohexadiene 6123	8	Cyclobutene 41	1
Cyclohexadiene 1245	7	Cyclobutadiene	1
Cyclohexadiene 2356	7	Cyclopropane	4
Benzene	186	Cyclopropene 12	1
Cyclopentane	32	Cyclopropene 23	1
Cyclopentene 12	7	Cyclopropene 31	1

Table 8.6 – The relative fitnesses of the ring types used in Eve to build molecules, calculated by analysis of a database of ligand molecules. The numbers *ab* after the ring names indicate the position of a double bond

between atoms *a* and *b*, moving around the ring from position 1 to *n* for an *n*-membered ring.

When creating a new ring system, the roulette wheel method is used to determine the ring scaffold, the ring size and the ring type. In every newly created ring, the identity of each unit around the ring is also chosen at random (if it satisfies the valency considerations) using the roulette wheel method. These ring units are again tallied using TSAR to calculate the prevalence of each and the results are shown in Table 8.7.

Ring Unit	**Relative Fitness**
Carbon	12200
Nitrogen	272
Oxygen	2882
Sulphur	126
Sulphoxide (S=O)	8
Sulphone (SO2)	66
Ethenyl (C=C)	12
Carbonyl (C=O)	1665

Table 8.7 - Relative fitnesses of the ring units used in Eve to build ring systems, calculated by analysis of a database of ligand molecules.

Due to the prevalence of ring systems in drug molecules, it is important to ensure that the algorithm correctly places and orients the rings. This is achieved by storing the

geometric parameters for each ring type. Four sets of parameters are kept: the bond orders, the bond lengths, the bond angles and the torsion angles. The bond orders and torsion angles are specific to the ring type and considered independent of the units that make up the ring. Conversely, the bond lengths and bond angles are specific to the units that comprise the ring and are considered independent of the ring type.

8.5.5 Crossover

In EP techniques such as GAs, the method of crossover is used. In the same way that simple EP mimics asexual reproduction, with parent solutions yielding almost identical offspring, GAs mimic sexual reproduction, in which characteristics of two members of a population combine (cross over) to produce the characteristics of the offspring. However, in this case, the variables are not independent and thus crossover is not a simple operation. A crossover operator has been developed for molecules [97] using graph theory and has the advantage of creating a good solution from portions of two bad solutions. This allows good conformations to be sampled more easily and allows the algorithm to cross very high-energy barriers in configurational space. However, it is a complex and lengthy computational task and thus will not be used in this case due to a significant increase in run time of the program.

8.5.6 Geometrical Considerations

The precise three-dimensional structure of a molecule is determined by the favourable orientations of the molecular orbitals involved. When attempting to calculate the binding energy of a protein-ligand complex, it is obviously vital that the structure is correct. Vector algebra is a useful tool in calculating the geometry of molecules. Calculation of atom placement based on bond angles can be performed using dot products.

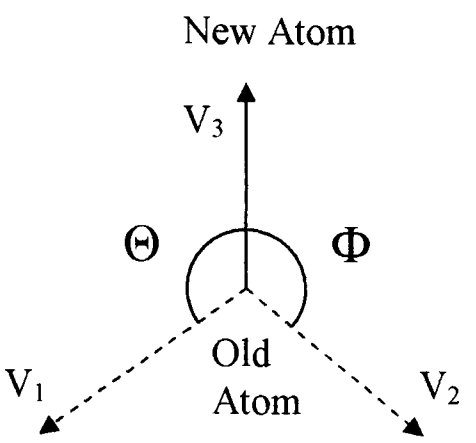


Figure 8.7 - The vectors and angles used in calculating the position of a new atom.

A new atom is being bonded to the old atom that is already present in the molecule and is bonded to two other atoms (at the ends of vectors V_1 and V_2). The new atom can be positioned by calculating the vector V_3 and combining it with the position of the old atom. V_3 can be calculated in the following manner

$$Length\ (V_1) \times Length\ (V_3) \times \cos\ \Theta = V_1.V_3$$

$$Length\ (V_2) \times Length\ (V_3) \times \cos\ \Phi = V_2.V_3$$

$$Length\ (V_3) \times Length\ (V_3) = V_3.V_3$$

Equation 8.1 - Dot products used to calculate the vector of the new bond.

The magnitudes of V_1, V_2 and V_3 are simply the bond lengths and the magnitudes of Θ and Φ are simply the bond angles. Both can be found in tables of data. This allows the

calculation of the vector V_3, which then allows the position of the new atom to be calculated. In many cases (such as bonding to a solitary atom, a primary sp3 or sp2-hybridized atom or a secondary sp3 atom) there is a degree of freedom in the position of the new atom. In these cases, any unfixed angles are randomly determined.

8.6 Lead Optimisation

When considering the rational design of a protein inhibitor, many strategies can be employed. A common technique is to analyse the active site to suggest useful pharmacophores and then look for a scaffold to place these various pharmacophores in the correct orientations. Thus, it is very useful if a scaffold can be found that is known to bind to the protein. The crystal structure of a drug and protein can be split into these two parts and multiple copies of the drug molecule placed in the protein active site. This can then be used as the starting point for the ligand growth algorithm. The fitness of the solution should improve over time until a new and fitter compound is found. This is a particularly attractive use of the algorithm because a ligand scaffold is already present, saving computer time, but the molecule is still free to evolve towards a better solution. The concept is used to test Eve by again using the test case of the Neuraminidase receptor. The PDB file 1A4G contains the protein Neuraminidase in complex with known inhibitor Relenza. Initially, both Relenza and the natural ligand Sialic acid are scored by running Eve with 2000 ligands for 50 generations using fixed starting coordinates. The best docking result for Sialic acid can be seen in Figure 8.8.

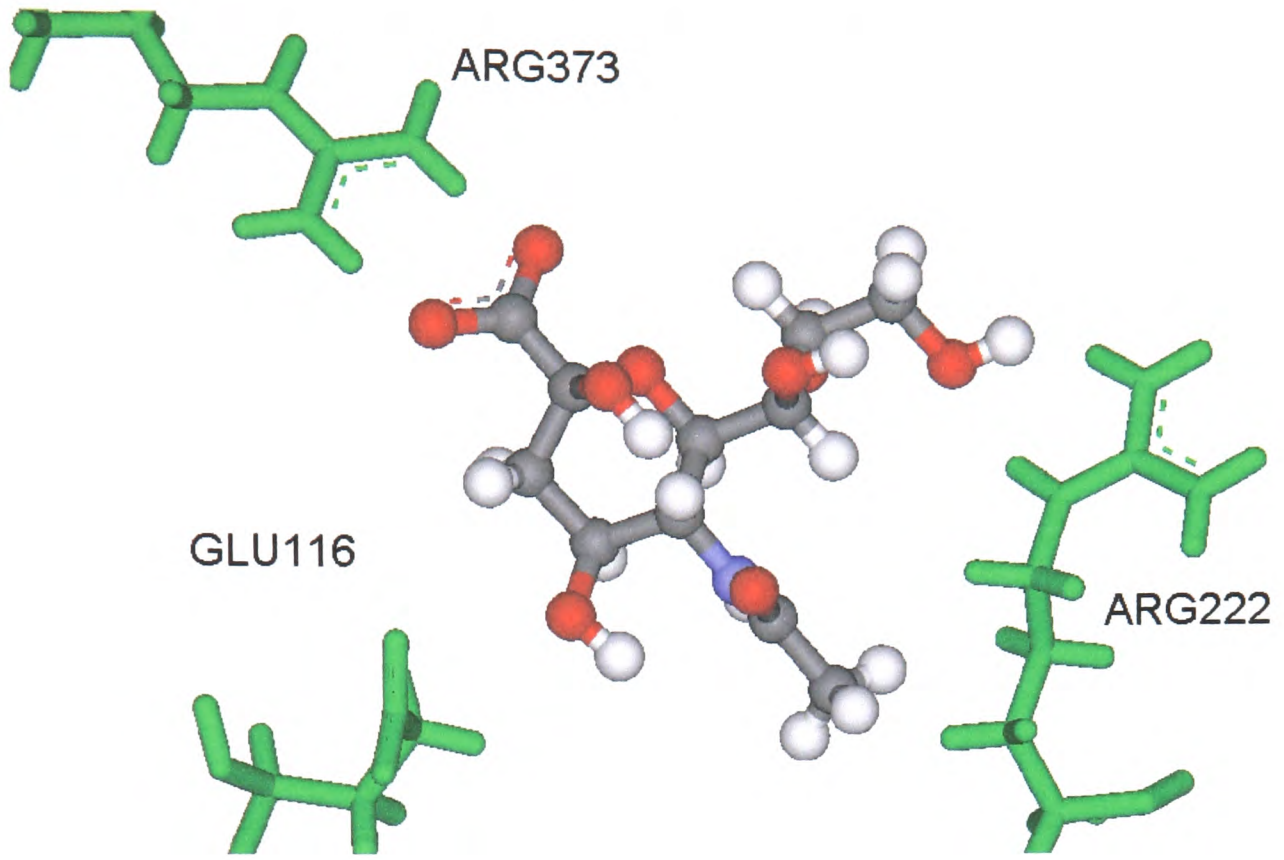


Figure 8.8 - The lowest energy docking position of Sialic acid bound to the neuraminidase receptor. Three of the important binding residues are coloured green. The calculated binding energy is -52.18 kJ/mol.

The lowest energy docking position for Relenza can be seen in Figure 8.9.

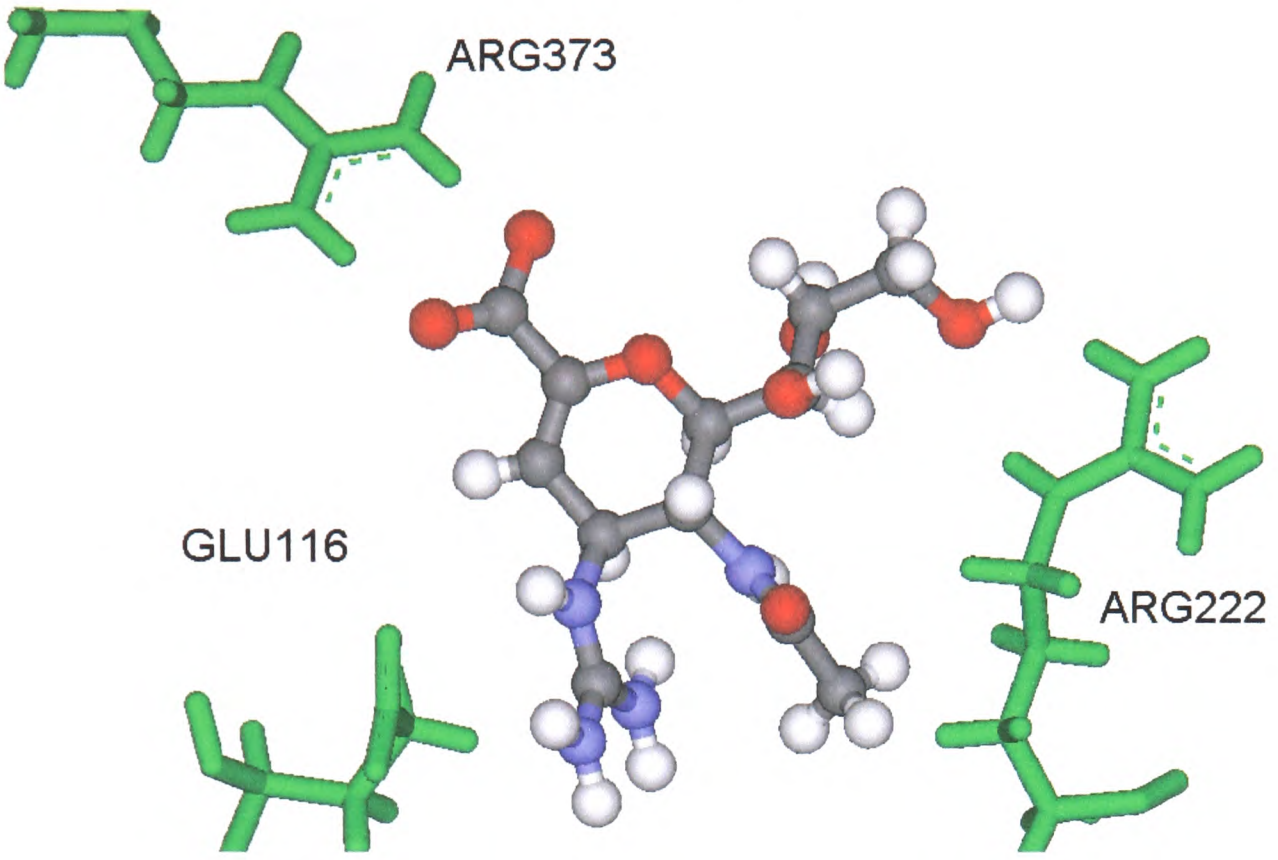


Figure 8.9 - The lowest energy docking position of Relenza bound to the neuraminidase receptor. Three of the important binding residues are coloured green. The calculated binding energy is -67.67 kJ/mol.

The best Sialic acid docking position is then taken as the first generation of an Eve run, in which the ligand structure is allowed to evolve as well as the ligand position. The algorithm is run for 15 iterations with 10,000 ligands. Trials show that 15 generations is enough to yield a good answer to the problem in a reasonable amount of computer time. This run required only 143 minutes on a Pentium IV 2.8GHz processor. The improvement in interaction energy is shown below in Figure 8.10.

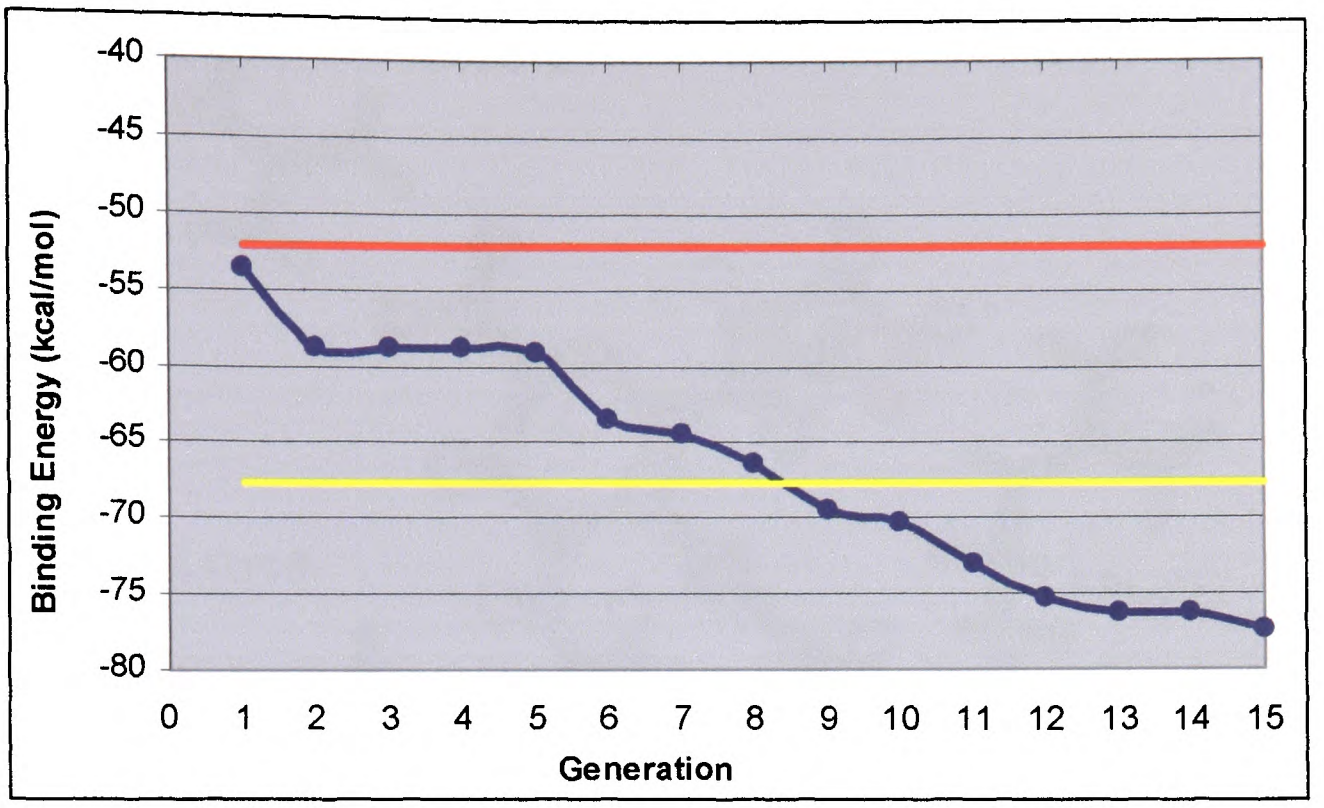


Figure 8.10 - The calculated binding energy (in kcal/mol) of the best ligand for 15 generations of a run exploring the binding site of the complex 1A4G. The energy is shown as a blue line, the calculated binding energy of Sialic Acid is shown as a red line and the calculated binding energy of Relenza is shown as a yellow line.

The lowest energy solution is shown in Figure 8.11 and shows the ligand with three residues from the active site.

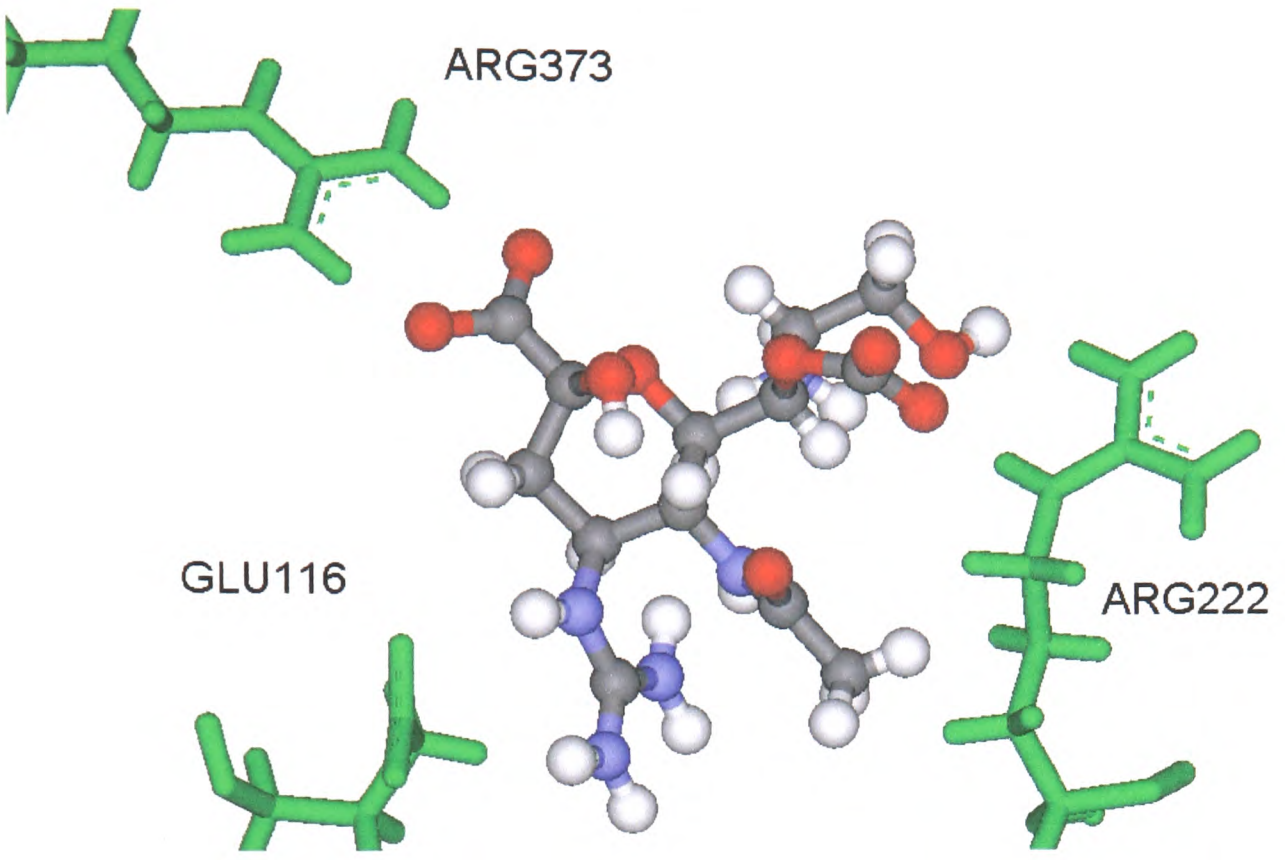


Figure 8.11 - The lowest energy solution from the Eve run, bound to the neuraminidase receptor. Three of the important binding residues are coloured green. The calculated binding energy is -77.41 kJ/mol

Note that the guanidine group is positioned in an almost identical position as for the Relenza molecule in Figure 8.9. There are two further modifications to the structure of Sialic Acid. The first is a carboxylate group added to one of the hydroxyl groups to form a carbonate. This charged group has the potential to form interactions with Arginine 149 and Arginine 222. The second modification is the mutation of an oxygen atom to a nitrogen atom, forming of an amino group in place of a hydroxyl group. This has the potential to form interactions with both Glutamate 274 and Glutamate 275. These changes explain the increased binding energy of the new molecule in comparison with Relenza.

This example illustrates the power for EP in this application. Eve requires only 15 generations to build a potentially better inhibitor than Relenza. With increased run time, this technique could prove useful for any given inhibitor. In the pharmaceutical industry, the use of this method would likely involve multiple runs of the same lead molecule. The best ten results from each run could be evaluated for synthetic viability to yield a number of possibilities for screening at a higher level. The time saved in lead optimisation could then be used in the potentially more complex problems with drug absorption, distribution, metabolism, excretion and toxicity (ADMET). Unfortunately, the use of this method for lead optimisation requires a lead compound as a scaffold to create new and better inhibitors. The more difficult task of *de novo* design will now be considered.

8.7 De Novo Design

The biggest challenge in rational drug design is the creation of novel inhibitors based entirely on the specifics of a protein active site. This has been attempted [98-100] with some success in the past, but remains a difficult task. The problem is similar to that faced in molecular docking; that the search space is so vast that finding "good" solutions becomes a time consuming process. The enormous range of possible molecules makes this a very difficult task but one that is suited to evolutionary programming. As proved in chapter seven, Eve is well suited to the task of finding good solutions in large search spaces and is able to modify the structure of known ligands to optimise their binding interaction, as illustrated in section 8.6. The one remaining issue is where and how the search begins. Eve operates in this respect by creating a population of ligands at random within the defined active site. Ring systems

and functional groups are chosen based on the probabilities discussed in section 8.5. The example in Figure 8.12 shows 50 randomly created ligands in a sphere of radius 15 Å.

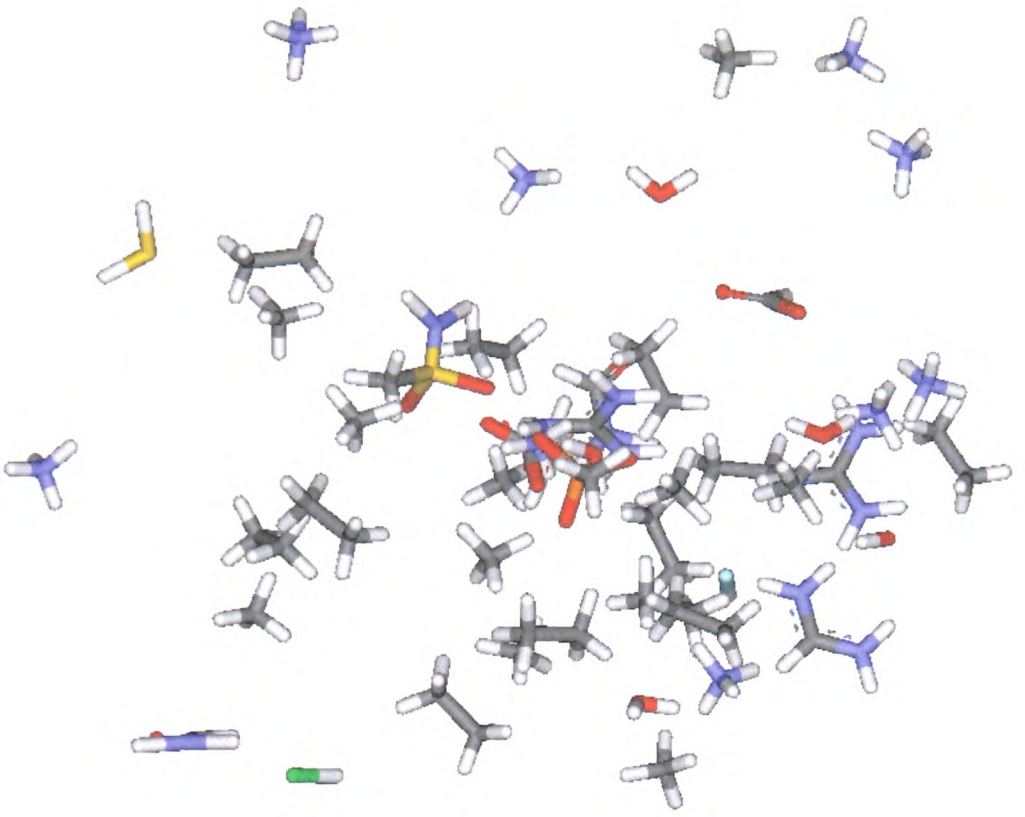


Figure 8.12 - 50 ligands created at random in a sphere of radius 15 Å. The molecules are coloured by atom.

If the ligand structure and position can both be altered, these molecules can evolve towards good solutions. If enough starting positions are sampled, the algorithm quickly locates the extent of the binding site and finds solutions that provide good platforms for further growth. To ensure a large number of low energy solutions in the population, the binding energy of each randomly created molecule is calculated immediately and only solutions below an energy threshold (100 kcal/mol is used) are retained.

This process is tested using the simple example of a glycine receptor in PDB ID 1PB7. The -34.55 kcal/mol estimate of the binding energy of glycine is calculated by running Eve with 5000 ligands for 50 generations using a fixed starting geometry based on the crystal structure. An initial population of 20000 ligands is then created for the active site of the glycine receptor. The lowest energy solution is shown in Figure 8.13

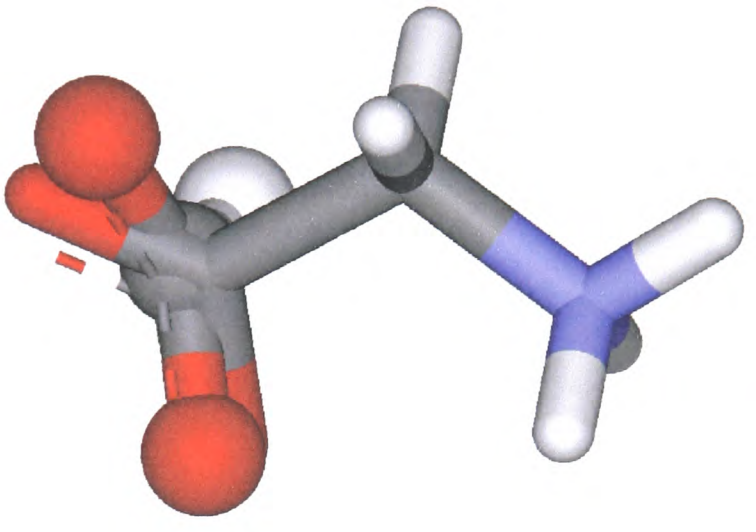


Figure 8.13 - The ligand glycine in its bound conformation (atom coloured with stick representation) and the lowest energy solution from the initial population (atom coloured with ball and stick representation). The molecule is a formate ion with binding energy of -17.04 kcal/mol.

Note the two carboxylate groups in close alignment in both molecules. This result shows that the algorithm can rapidly find good solutions with the ability to form a platform for the growth of excellent solutions. The algorithm is then run for 25 generations in an attempt to design a good ligand for this receptor. The calculated interaction energy of the lowest energy solution decreases during the course of the run as shown in Figure 8.14.

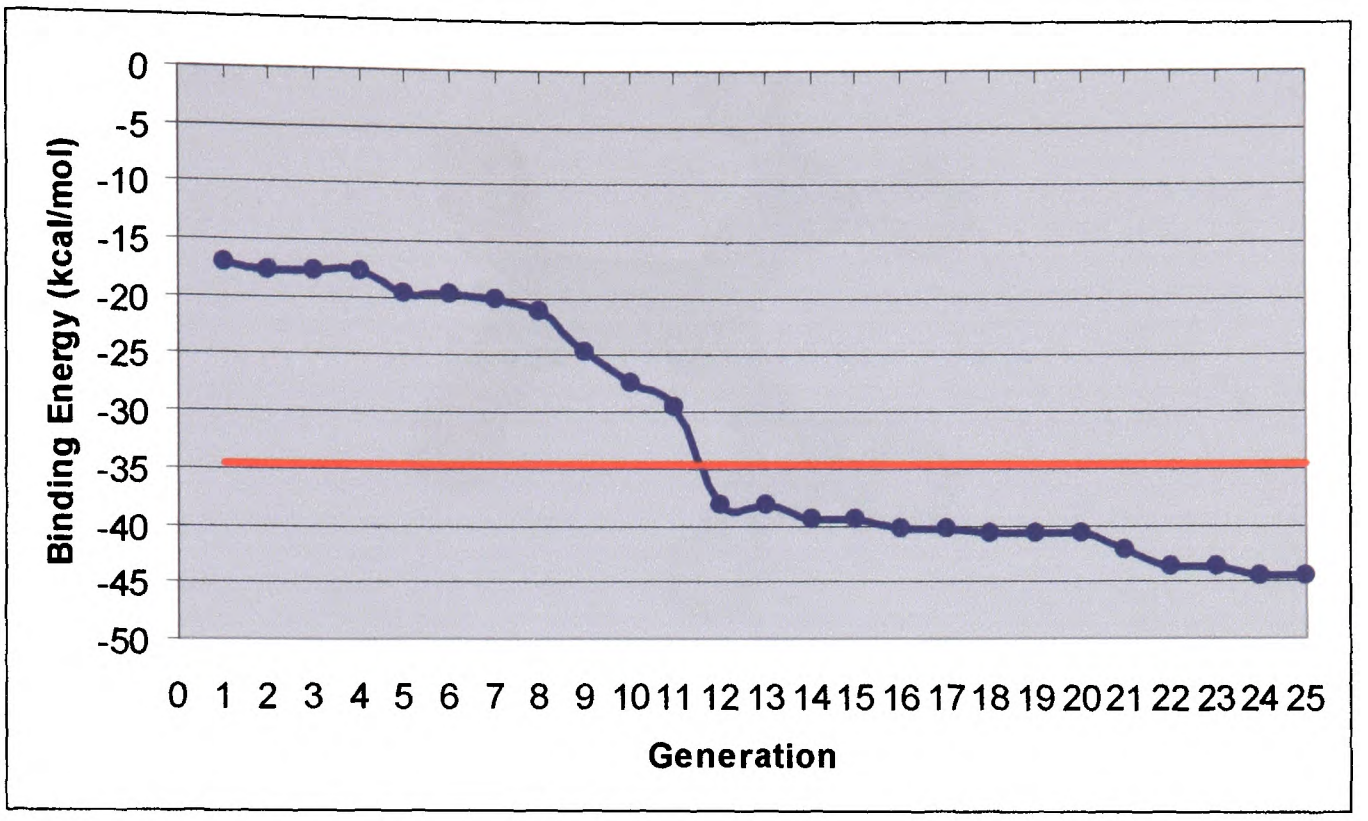


Figure 8.14 - The calculated binding energy (in kcal/mol) of the best ligand for generations 1 to 25 of a run exploring the binding site of the complex 1PB7. The energy is shown as a blue line and the calculated binding energy of glycine is shown as a red line.

The results show that Eve makes a steady improvement in the energy of the best docking solution, as would be expected for such a simple case. However, the algorithm does not find glycine to be the best molecule to dock in this active site. The best solution is shown in Figure 8.15 and has a docking energy of -44.55 kcal/mol.

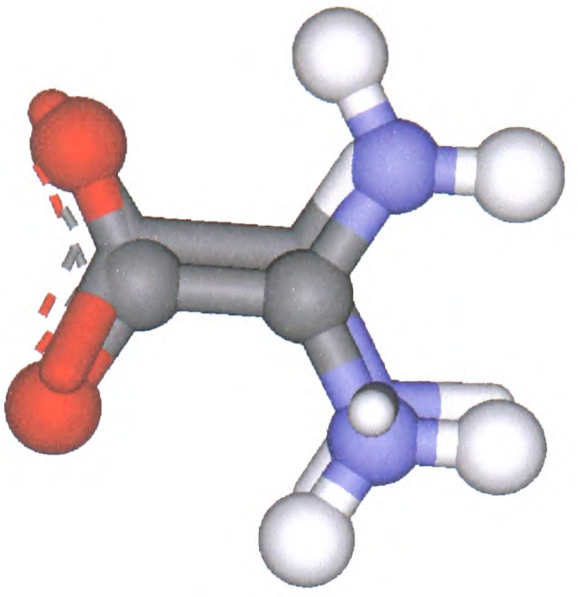


Figure 8.15 - The lowest energy solution for Eve with the test case 1PB7 in ball and stick representation. The crystal structure of glycine is showed in stick form.

The carboxylate groups are very well aligned and the amidino group of the new ligand is in close proximity to the amino group of glycine. This new molecule is amidino-formate and, although it is not registered as a glutamate receptor agonist, it would seem highly likely to bind very strongly to this receptor. The model predicts that it will form interactions with the Arginine 523 and Aspartate 745 residues in the active site. A more accurate calculation of the similarity of these molecules can be performed using the ASP function for molecular similarity within the TSAR program. Using a grid-based similarity method [101], the two molecules have a similarity of 0.626. A value of 0.0 represents complete dissimilarity and 1.0 represents identical molecules. As an example, the amino acids Leucine and Isoleucine have a similarity of 0.922. This result illustrates the ability of the algorithm to search the available space within the active site and make useful and testable predictions of molecules that form strong interactions with the protein.

The next stage is to select a more complex case, but one in which the ligand is not so huge that an estimate of its binding potential cannot be made. An initial population of 10,000 ligands is created for the active site of the glutamate receptor in PDB ID 1FTJ. The combined surfaces for the 50 best starting positions along with the bound glutamate ligand can be seen in Figure 8.16.

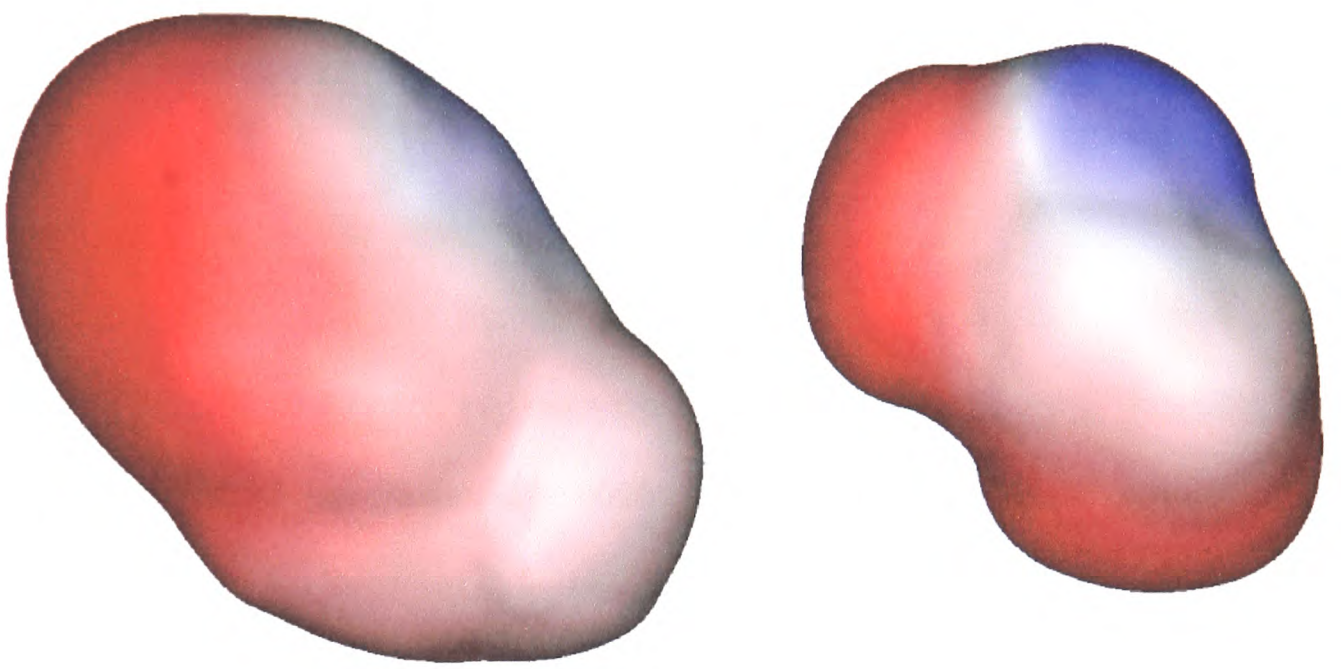


Figure 8.16 - The surface of the glutamate molecule (right) and the combined surfaces of the 50 lowest energy solutions for a run exploring the binding site of the protein 1FTJ (left). The surfaces are coloured by electrostatic potential and have the same geometrical orientation.

Note the electropositive region in the top right of each molecular surface (coloured blue) and the electronegative region in the top left (coloured red). This illustrates that Eve can explore the binding cavity in early iterations and create ligands that exploit the electrostatic potential. The lowest energy result from the fifty best solutions from the first generation is shown in Figure 8.17.

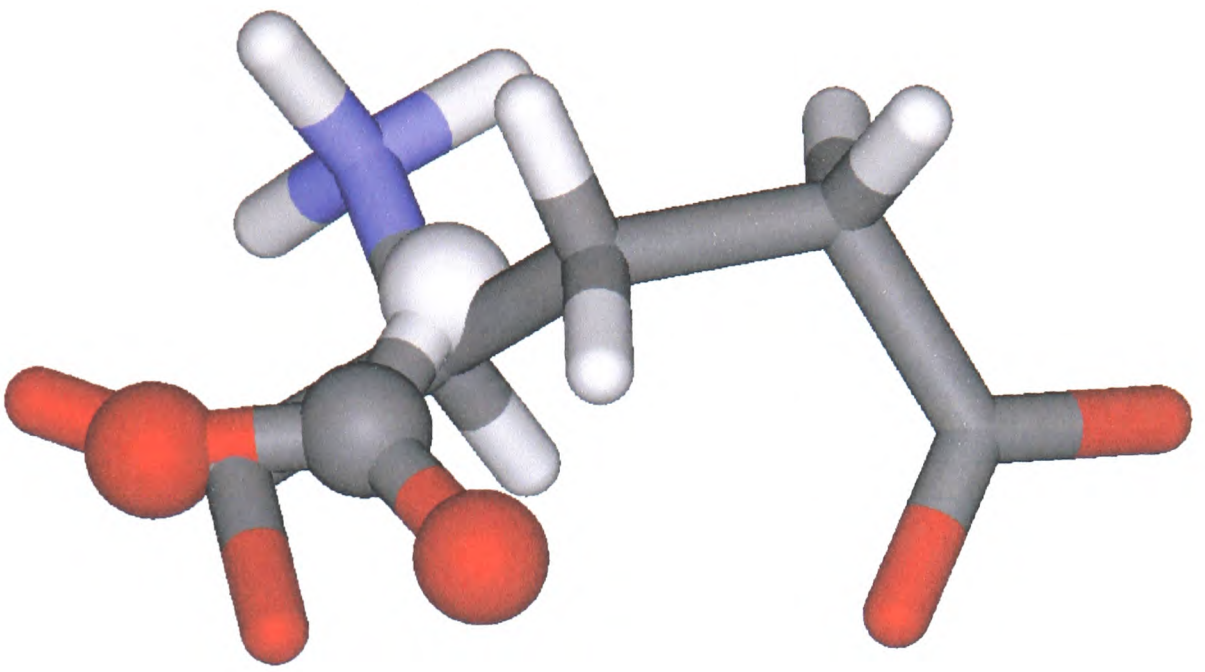


Figure 8.17 - The ligand glutamate in its bound conformation (atom coloured with stick representation) and the lowest energy solution from the initial population of 10000 solutions (atom coloured with ball and stick representation). The molecule is a formate ion with binding energy of -18.92 kcal/mol.

As for the previous example, the algorithm is able to find a useful platform to begin the process of designing a ligand. The algorithm is then run for 50 iterations. The calculated interaction energy of the lowest energy solution decreases during the course of the run as shown in Figure 8.18. The -50.69 kcal/mol estimate of the binding energy of glutamate is calculated by running Eve with 5000 ligands for 50 generations using a fixed starting geometry based on the crystal structure.

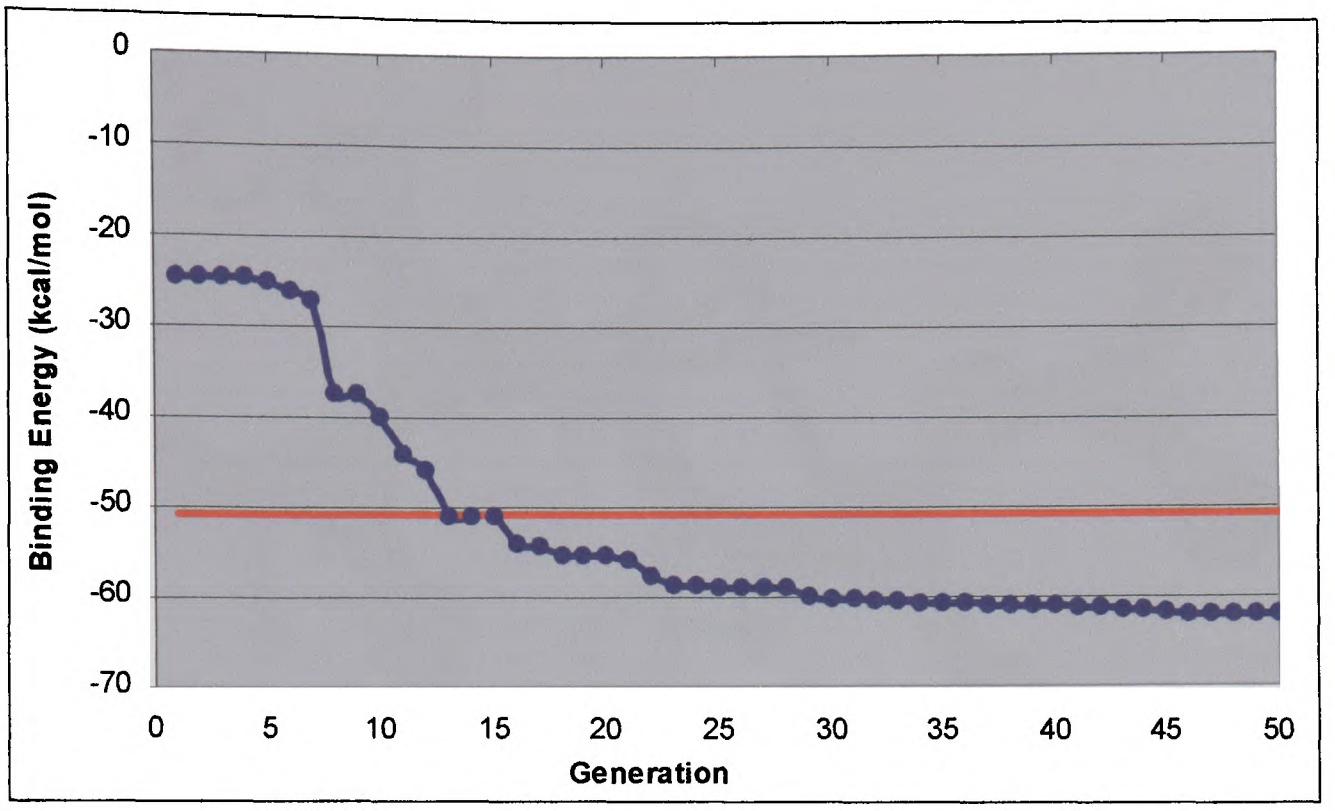


Figure 8.18 - The calculated binding energy (in kcal/mol) of the best ligand for generations 1 to 50 of a run exploring the binding site of the complex 1FTJ. The energy is shown as a blue line and the calculated binding energy of glutamate is shown as a red line.

This run took approximately 4 hours on a 2.8GHz processor. Despite this being a lengthy calculation, if the calculation successfully designs a good inhibitor, the results are worth the time taken. As for the case of the glycine-binding receptor 1PB7, there is steady improvement in the energy of the best docking solution, but the algorithm does not select glutamate as the final solution (it actually finds it but discards it later). However, it again creates a molecule that is predicted to bind more strongly with the binding site than the natural ligand. The best solution is shown in Figure 8.19 and the calculated interaction energy is -62.07 kcal/mol.

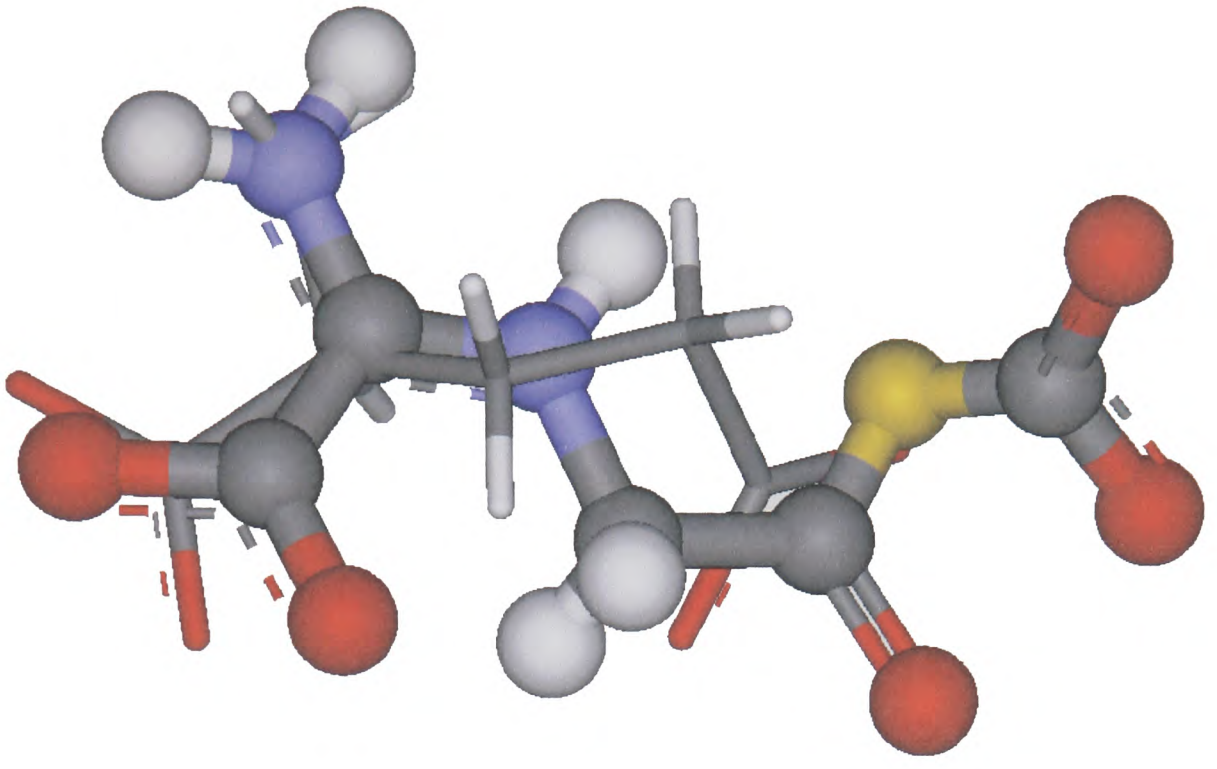


Figure 8.19 - The lowest energy solution for Eve with the test case 1FTJ in ball and stick representation. The crystal structure of glutamate is showed in stick form.

This result is more revealing when the molecular surfaces of the two molecules are examined. These can be seen in Figure 8.20.

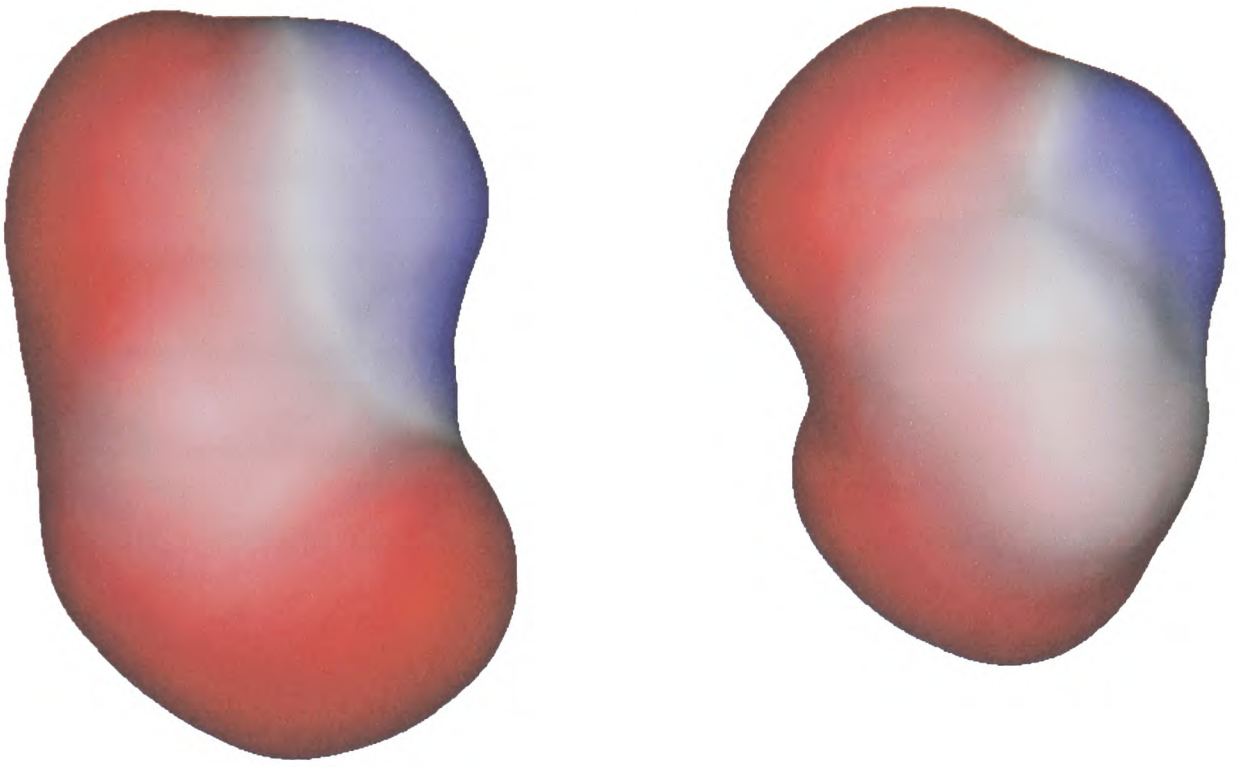


Figure 8.20 - The molecular surfaces of the lowest energy solution (left) and glutamate (right). The surfaces are coloured by electrostatic potential with negative tinted red and positive tinted blue. The molecules have the same orientation.

It is clear that the new molecule has a very similar shape to glutamate and that the electrostatic interactions are likely to be favourable. An analysis of molecular similarity with TSAR using a grid-based method yields a similarity of 0.531 for these two molecules, highlighting their resemblance. If it is synthesisable, the new molecule is likely to be a good agonist for this receptor. Unfortunately, the realms of computational chemistry, although enormous, are not able to prove this satisfactorily, and the result must remain a prediction. However, these two cases do show promise for the use of EP in *de novo* drug design and again illustrate the potential of computational chemistry in leading experimental work.

Sections 8.6 and 8.7 have shown that Eve is able to grow a molecule within the active site of a protein. However, the simplified scoring function means that the results are not predictive of actual binding energy. The omissions discussed in section 6.6.4 will have an effect on the protein-ligand interactions and this would have to be considered in any real world application of Eve. In particular, the lack of a desolvation term leads to the algorithm adding more and more bulk to the molecule in an attempt to create favourable electrostatic binding interactions with the protein. This would have a corresponding effect on desolvation, which it is vital to consider. Furthermore, an analysis of synthetic viability would have to be included in an implementation of this technique.

8.8 Summary

In chapter seven, it was proved that Eve operates as an accurate tool for molecular docking. It is able to predict the correct binding geometry to within a 2.0 Å heavy atom RMSD in over 70% of cases. This ability allows the algorithm to be modified to perform a number of useful applications. Two of these can be used without altering the nature of the code. The first involves using ligand probes to explore the surface of a protein in cases where the binding site is unknown. Hydrophobic probes such as benzene are shown to be good at indicating the probable location of pockets where ligand binding is likely, as hydrophobic effects favour these locations. The second function again uses ligand probes, but in this case, the probes can aid in lead optimisation by flagging areas in an active site where positive binding interactions can be formed. The particular case of Relenza [88] is used to illustrate this concept and the results show an excellent agreement with prior work.

The code for Eve can easily be modified to provide a new algorithm that can be used for two other very useful applications. The first stage of modification is to alter the mutations available within the evolutionary program to allow the structure to be altered as well as the atom positions. Atoms can be changed or removed and various functional groups can be added. The second stage is the use of the TSAR program for molecular descriptor calculation to tabulate the relative prevalence of functional groups. This allows the evolution to be tailored toward more likely solutions. The combination creates an algorithm that optimises ligands to maximise interaction energies. In the first example, lead optimisation is automated, again using the example of the Sialidase inhibitor Relenza. The focus is then shifted to *de novo* design by combining the docking of molecular probes with the use of the ligand optimisation. This produces interesting and testable results.

Chapter seven confirmed that Eve is an excellent tool in molecular docking. This chapter shows that Eve can be used for a wide range of applications, and thus represents a useful tool in computational biology and a step forward in computational drug discovery.

9 Conclusions

The use of computers has revolutionised almost every area of science, giving scientists the ability to simulate physical and biological processes to enhance understanding and suggest new areas of research. Within chemistry, computers have allowed us to visualise and investigate biological molecules. This ability has been exploited in many different biological systems and is being applied to a growing number. Experimental techniques such as X-Ray diffraction, that can be used to determine the structure of large bio-molecules, are becoming more widespread and the PDB has been growing rapidly for many years. There are now over fifty functional synchrotron facilities in the world, with each spending a considerable amount of time harvesting protein structures. This structural data is the starting point for a number of computational approaches to protein function elucidation.

Molecular docking is one of the most important techniques used to analyse these structures, and the current increase in CPU speed will allow faster and better algorithms to be employed. However, at present, the enormous complexity inherent in the docking problem means that computers are not fast enough to calculate binding poses without the use of simplifications. Oxdock is a docking algorithm that uses the concept of a multiscale approach to reduce the complexity of the docking problem. The ligand is docked initially as only one feature-point and surviving solutions are represented as an increasing number of feature-points in subsequent iterations. This approach allows the entire surface of a protein to be rapidly explored, and has proved to be accurate in a number of test cases.

Oxdock has also proved useful in applications to real biological problems. NMDA receptors are a class of ionotropic glutamate receptors that are involved in synaptic transmission in the mammalian brain. They have an interesting biochemistry, and are modulated by a variety of endogenous compounds. A combination of homology modelling and molecular docking with Oxdock suggested how these modulators may act and interact to produce a functional NMDA receptor. Results provided a model in which two venus-flytrap domains are coupled together with a hinge formed at the top, allowing molecules that bind to the topmost amino terminal domain to affect the trans-membrane gating. The model also made predictions that could be tested experimentally.

The effectiveness of marrying computational predictions with experimental evidence was confirmed by the use of Oxdock to explore putative plant receptor proteins. Prior to this work, interest in this subject had waned due to an inability to demonstrate glutamate binding. A combination of homology modelling and molecular docking with Oxdock revealed that the majority of the receptor subunits are unable to bind glutamate due to the replacement of key amino acid residues. Analysis highlighted the similarity with animal NMDA receptor sub-types NR1 and NR3, both of which bind glycine. Further modelling suggested that glutamate and glycine would act synergistically in the functioning of these receptors. This was later proved experimentally, opening new avenues of research into this interesting area.

Whilst the use of molecular docking can provide some very interesting results in an academic investigation, one of the most useful is the ability to predict the binding pose accurately and thus estimate the binding interaction between a ligand and a

protein. Unfortunately, though excellent at locating the binding site when given a ligand and a protein, Oxdock is not good at calculating the exact docking pose and tends to fail in cases with long or hydrophobic ligands. The testing performed on Oxdock also illustrated the importance of a careful use of grid-based energy calculations. A grid spacing of 0.5 Å was insufficient to calculate the van der Waals binding energy accurately and only at 0.1 Å was the energy estimate within acceptable limits. These problems highlighted the need for a new algorithm that combines the speed of multiscale docking with a technique for accurate optimisation.

The newly created algorithm Eve uses the optimising ability of evolutionary programming. The ligand is docked in the first three iterations as one, two and three feature-points to create a multitude of possible orientations. The best of these orientations are then used to create a population of ligands. An evolutionary algorithm then optimises the docking pose by translating and rotating the molecules and modifying the torsion angles. The results of an extensive validation of the algorithm proved that it is precise, accurate and specific. Eve can thus find a use in both the identification of active sites and the calculation of optimal docking poses. As exemplified by the work discussed in chapters 3 and 4, the ability to locate binding sites and to elucidate ligand specificity is a very useful technique in modern biological chemistry. Eve is well suited to solving this problem and represents an improvement over Oxdock. Eve's abilities make it a useful tool for both academic investigation and in high throughput screening.

The subsequent modification of this algorithm has been the incorporation of a method to alter the structure of the ligand as well as its position. This altered the docking

algorithm into a tool for lead optimisation. The results of the lead optimisation are more reliable because the scoring function used to rate the solutions has been validated. Rigorous substantiation of the results was difficult due to the requirement for experimental data, which may be inaccessible or non-existent. However, results suggested that Eve shows promise in this area. Results were also encouraging when Eve was used in *de novo* design. The lead optimisation process was run, but using randomly created small molecular fragments placed in the binding site as the first generation of ligands. This process was shown to be effective for small ligands. Again, it was very difficult to validate predictions due to the difficulty in obtaining experimental proof. However, the results were definitely promising.

Despite the successes of Eve in predicting the docking pose for a variety of protein-ligand complexes, there remain a number of improvements that can be made to the algorithm. The problems discussed in section 7.4.4 could be addressed fairly simply, solving the issues in a number of problematic cases. The efficiency of the algorithm remains poor in comparison with GOLD and an exhaustive parameter optimisation could be performed to decrease the run times whilst sacrificing little or no accuracy. A more comprehensive scoring function should improve the RMSD results and allow Eve to make realistic predictions of the binding energy. A simple scoring function could still be employed for the multiscale portion of the algorithm to lower run times and then rigorous calculations could be performed during the EP optimisation. The new scoring function would include desolvation effects and torsional terms. An improved scoring function should also improve the quality and applicability of the molecular design functionality within Eve. The issues with desolvation discussed in section 8.7 could be addressed, leading to a more realistic description of binding. The

final improvement would be an inclusion of protein flexibility. This would be a vital part of any real world use of Eve and is thus very important to consider. However, as discussed in section 6.3, Eve is well suited to incorporate this protein flexibility using the conformational sub-states model.

The work discussed in this thesis highlights the many uses of computational docking methods in both the academic sphere and the pharmaceutical industry. The location of binding sites and the elucidation of ligand specificity, allows computational models to direct experimental work and this marriage enhances both fields. The ability to locate the binding site of a protein or to discover the natural ligand can elucidate the function of the protein. It is also the first step in rational drug design. The ability to predict the binding pose of a protein-ligand complex accurately and provide a standardised estimate of the interaction energy allows potential drug candidates to be screened for efficacy. The further modification of automated lead-optimisation provides a method for improving existing drug molecules and designing new inhibitors. Eve has proved to be an excellent tool with numerous applications and the potential to solve problems within any area of chemical biology.

References

1. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: The Protein Data Bank. Nucleic Acids Res 28: 235-42 (2000).
2. Fischer E: Einfluss der Configuration auf die Wirkung der Enzyme. Berichte der Deutschen Chemischen Gesellschaft 27: 2985–2993 (1894).
3. Koshland DE: Application of a theory of enzyme specificity to protein synthesis. Proceedings of the National Academy of Sciences of the United States of America 44: 98-104 (1958).
4. Chiu J, DeSalle R, Lam HM, Meisel L, Coruzzi G: Molecular evolution of glutamate receptors: A primitive signaling mechanism that existed before plants and animals diverged. Molecular Biology and Evolution 16: 826-838 (1999).
5. Pearlman DA, Case DA, Caldwell JW, Ross WR, Cheatham I, T.E. , DeBolt S, Ferguson D, Seibel G, Kollman P: AMBER, a computer program for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to elucidate the structures and energies of molecules. Comp. Phys. Commun 91: 1-41 (1995).
6. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M: Charmm - a Program for Macromolecular Energy, Minimization, and Dynamics Calculations. Journal of Computational Chemistry 4: 187-217 (1983).
7. Rarey M, Kramer B, Lengauer T, Klebe G: A fast flexible docking method using an incremental construction algorithm. J Mol Biol 261: 470-89 (1996).
8. Ewing TJ, Makino S, Skillman AG, Kuntz ID: DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases. J Comput Aided Mol Des 15: 411-28 (2001).
9. Trosset JY, Scheraga HA: Reaching the global minimum in docking simulations: a Monte Carlo energy minimization approach using Bezier splines. Proc Natl Acad Sci U S A 95: 8011-5 (1998).
10. Friesner RA, Banks JL, Murphy RB, Halgren TA, Klicic JJ, Mainz DT, Repasky MP, Knoll EH, Shelley M, Perry JK, Shaw DE, Francis P, Shenkin PS: Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. J Med Chem 47: 1739-49 (2004).
11. Jones G, Willett P, Glen RC, Leach AR, Taylor R: Development and validation of a genetic algorithm for flexible docking. Journal of Molecular Biology 267: 727-748 (1997).
12. Morris GM, Goodsell DS, Halliday RS, Huey R, Hart WE, Belew RK, Olson AJ: Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. Journal of Computational Chemistry 19: 1639-1662 (1998).
13. Taylor RD, Jewsbury PJ, Essex JW: A review of protein-small molecule docking methods. Journal of Computer-Aided Molecular Design 16: 151-166 (2002).
14. Brooijmans N, Kuntz ID: Molecular recognition and docking algorithms. Annual Review of Biophysics and Biomolecular Structure 32: 335-373 (2003).

15. Conklin BR, Bourne HR: Homeostatic signals. Marriage of the flytrap and the serpent. Nature 367: 22 (1994).
16. Glick M, Robinson DD, Grant GH, Richards WG: Identification of ligand binding sites on proteins using a multi-scale approach. Journal of the American Chemical Society 124: 2337-2344 (2002).
17. Glick M, Grant GH, Richards WG: Docking of flexible molecules using multiscale ligand representations. Journal of Medicinal Chemistry 45: 4639-4646 (2002).
18. Glick M, Grant GH, Richards WG: Pinpointing anthrax-toxin inhibitors. Nature Biotechnology 20: 118-119 (2002).
19. Veal JM, Wilson WD: Modeling of nucleic acid complexes with cationic ligands: a specialized molecular mechanics force field and its application. J Biomol Struct Dyn 8: 1119-45 (1991).
20. Dauberosguthorpe P, Roberts VA, Osguthorpe DJ, Wolff J, Genest M, Hagler AT: Structure and Energetics of Ligand-Binding to Proteins - Escherichia-Coli Dihydrofolate Reductase Trimethoprim, a Drug- Receptor System. Proteins-Structure Function and Genetics 4: 31-47 (1988).
21. Lundstrom K: Structural genomics on membrane proteins: mini review. Comb Chem High Throughput Screen 7: 431-9 (2004).
22. Armstrong N, Sun Y, Chen GQ, Gouaux E: Structure of a glutamate-receptor ligand-binding core in complex with kainate. Nature 395: 913-917 (1998).
23. Scatton B: The Nmda Receptor Complex. Fundamental & Clinical Pharmacology 7: 389-400 (1993).
24. Chen NS, Moshaver A, Raymond LA: Differential sensitivity of recombinant N-methyl-D-aspartate receptor subtypes to zinc inhibition. Molecular Pharmacology 51: 1015-1023 (1997).
25. Williams K, Zappia AM, Pritchett DB, Shen YM, Molinoff PB: Sensitivity of the N-Methyl-D-Aspartate Receptor to Polyamines Is Controlled by Nr2 Subunits. Molecular Pharmacology 45: 803-809 (1994).
26. Perin-Dureau F, Rachline J, Neyton J, Paoletti P: Mapping the binding site of the neuroprotectant ifenprodil on NMDA receptors. Journal of Neuroscience 22: 5955-5965 (2002).
27. Traynelis SF, Hartley M, Heinemann SF: Control of Proton Sensitivity of the Nmda Receptor by Rna Splicing and Polyamines. Science 268: 873-876 (1995).
28. Herin GA, Aizenman E: Amino terminal domain regulation of NMDA receptor function. Eur J Pharmacol 500: 101-11 (2004).
29. Das S, Sasaki YF, Rothe T, Premkumar LS, Takasu M, Crandall JE, Dikkes P, Conner DA, Rayudu PV, Cheung W, Chen HSV, Lipton SA, Nakanishi N: Increased NMDA current and spine density in mice lacking the NMDA receptor subunit NR3A. Nature 393: 377-381 (1998).
30. Sutcliffe MJ, Wo ZG, Oswald RE: Three-dimensional models of non-NMDA glutamate receptors. Biophysical Journal 70: MP397-MP397 (1996).
31. Armstrong N, Gouaux E: Mechanisms for activation and antagonism of an AMPA-Sensitive glutamate receptor: Crystal structures of the GluR2 ligand binding core. Neuron 28: 165-181 (2000).
32. Masuko T, Kashiwagi K, Kuno T, Nguyen ND, Pahk AJ, Fukuchi J, Igarashi K, Williams K: A regulatory domain (R1-R2) in the amino terminus of the N-methyl-D-aspartate receptor: Effects of spermine, protons, and ifenprodil, and

structural similarity to bacterial leucine/isoleucine/valine binding protein. Molecular Pharmacology 55: 957-969 (1999).

33. Low CM, Zheng F, Lyuboslavsky P, Traynelis SF: Molecular determinants of coordinated proton and zinc inhibition of N-methyl-D-aspartate NR1/NR2A receptors. Proceedings of the National Academy of Sciences of the United States of America 97: 11062-11067 (2000).
34. Masuko T, Kuno T, Kashiwagi K, Kusama T, Williams K, Igarashi K: Stimulatory and inhibitory properties of aminoglycoside antibiotics at N-methyl-D-aspartate receptors. Journal of Pharmacology and Experimental Therapeutics 290: 1026-1033 (1999).
35. Paoletti P, Perin-Dureau F, Fayyazuddin A, Le Goff A, Callebaut I, Neyton J: Molecular organization of a zinc binding N-terminal modulatory domain in a NMDA receptor subunit. Neuron 28: 911-925 (2000).
36. Mott DD, Doherty JJ, Zhang SN, Washburn MS, Fendley MJ, Lyuboslavsky P, Traynelis SF, Dingledine R: Phenylethanolamines inhibit NMDA receptors by enhancing proton inhibition. Nature Neuroscience 1: 659-667 (1998).
37. Sternbach Y, Bettler B, Hartley M, Sheppard PO, Ohara PJ, Heinemann SF: Agonist Selectivity of Glutamate Receptors Is Specified by 2 Domains Structurally Related to Bacterial Amino Acid-Binding Proteins. Neuron 13: 1345-1357 (1994).
38. Williams K, Kashiwagi K, Fukuchi J, Igarashi K: An acidic amino acid in the N-methyl-D-aspartate receptor that is important for spermine stimulation. Molecular Pharmacology 48: 1087-1098 (1995).
39. Kunishima N, Shimada Y, Tsuji Y, Sato T, Yamamoto M, Kumasaka T, Nakanishi S, Jingami H, Morikawa K: Structural basis of glutamate recognition by a dimeric metabotropic glutamate receptor. Nature 407: 971-977 (2000).
40. He XL, Chow DC, Martick MM, Garcia KC: Allosteric activation of a spring-loaded natriuretic peptide receptor dimer by hormone. Science 293: 1657-1662 (2001).
41. Sun Y, Olson R, Horning M, Armstrong N, Mayer M, Gouaux E: Mechanism of glutamate receptor desensitization. Nature 417: 245-253 (2002).
42. Leuschner WD, Hoch W: Subtype-specific assembly of alpha-amino-3-hydroxy-5-methyl-4- isoxazole propionic acid receptor subunits is mediated by their N-terminal domains. Journal of Biological Chemistry 274: 16907-16916 (1999).
43. Ayalon G, Stern-Bach Y: Functional assembly of AMPA and kainate receptors is mediated by several discrete protein-protein interactions. Neuron 31: 103-113 (2001).
44. Thompson JD, Higgins DG, Gibson TJ: Clustal-W - Improving the Sensitivity of Progressive Multiple Sequence Alignment through Sequence Weighting, Position- Specific Gap Penalties and Weight Matrix Choice. Nucleic Acids Research 22: 4673-4680 (1994).
45. Choi YB, Chen HSV, Lipton SA: Three pairs of cysteine residues mediate both redox and Zn2+ modulation of the NMDA receptor. Journal of Neuroscience 21: 392-400 (2001).
46. Fayyazuddin A, Villarroel A, Le Goff A, Lerma J, Neyton J: Four residues of the extracellular N-terminal domain of the NR2A subunit control high-affinity Zn2+ binding to NMDA receptors. Neuron 25: 683-694 (2000).

47. Gallagher MJ, Huang H, Grant ER, Lynch DR: The NR2B-specific interactions of polyamines and protons with the N-methyl-D-aspartate receptor. Journal of Biological Chemistry 272: 24971-24979 (1997).
48. Huggins DJ, Grant GH: The function of the amino terminal domain in NMDA receptor modulation. J Mol Graph Model 23: 381-8 (2005).
49. Lam HM, Chiu J, Hsieh MH, Meisel L, Oliveira IC, Shin M, Coruzzi G: Glutamate-receptor genes in plants. Nature 396: 125-126 (1998).
50. Davenport R: Glutamate receptors in plants. Annals of Botany 90: 549-557 (2002).
51. Lacombe B, Becker D, Hedrich R, DeSalle R, Hollmann M, Kwak JM, Schroeder JI, Le Novere N, Nam HG, Spalding EP, Tester M, Turano FJ, Chiu J, Coruzzi G: The identity of plant glutamate receptors. Science 292: 1486-1487 (2001).
52. Chiu JC, Brenner ED, DeSalle R, Nitabach MN, Holmes TC, Coruzzi GM: Phylogenetic and expression analysis of the glutamate-receptor- like gene family in Arabidopsis thaliana. Molecular Biology and Evolution 19: 1066-1082 (2002).
53. Brenner ED, Martinez-Barboza N, Clark AP, Liang QS, Stevenson DW, Coruzzi GM: Arabidopsis mutants resistant to S(+)-beta-methyl-alpha, beta-diaminopropionic acid, a cycad-derived glutamate receptor agonist. Plant Physiology 124: 1615-1624 (2000).
54. Dennison KL, Spalding EP: Glutamate-gated calcium fluxes in Arabidopsis. Plant Physiology 124: 1511-1514 (2000).
55. Furukawa H, Gouaux E: Mechanisms of activation, inhibition and specificity: crystal structures of the NMDA receptor NR1 ligand-binding core. Embo Journal 22: 2873-2885 (2003).
56. McGuffin LJ, Bryson K, Jones DT: The PSIPRED protein structure prediction server. Bioinformatics 16: 404-405 (2000).
57. Ivanovic A, Reilander H, Laube B, Kuhse J: Expression and initial characterization of a soluble glycine binding domain of the N-Methyl-D-aspartate receptor NR1 subunit. Journal of Biological Chemistry 273: 19933-19937 (1998).
58. Dubos C, Huggins D, Grant GH, Knight MR, Campbell MM: A role for glycine in the gating of plant NMDA-like receptors. Plant Journal 35: 800-810 (2003).
59. White PJ, Bowen HC, Demidchik V, Nichols C, Davies JA: Genes for calcium-permeable channels in the plasma membrane of plant root cells. Biochimica Et Biophysica Acta-Biomembranes 1564: 299-309 (2002).
60. Kim SA, Kwak JM, Jae SK, Wang MH, Nam HG: Overexpression of the AtGluR2 gene encoding an Arabidopsis homolog of mammalian glutamate receptors impairs calcium utilization and sensitivity to ionic stress in transgenic plants. Plant Cell Physiol 42: 74-84 (2001).
61. Nong Y, Huang YQ, Ju W, Kalia LV, Ahmadian G, Wang YT, Salter MW: Glycine binding primes NMDA receptor internalization. Nature 422: 302-307 (2003).
62. White PR: Glycine in the nutrition of excised tomato roots. Plant Physiology 14: 527-538 (1939).
63. Fries N: Limiting factors in the growth of the pea seedling root. Physiologia plantarum 6: 292-300 (1953).

64. Skinner JC, H.E. S: Studies on the growth of excised roots. II. Observations on the growth of excised groundsel roots. New Phytologist 53: 44-67 (1953).
65. Fogel LJ, Owens, A.J. & Walsh, M.J: Artificial Intelligence through Simulated Evolution. Wiley, New York (1966).
66. Leach AR: Ligand Docking to Proteins with Discrete Side-Chain Flexibility. Journal of Molecular Biology 235: 345-356 (1994).
67. Claussen H, Buning C, Rarey M, Lengauer T: FlexE: Efficient molecular docking considering protein structure variations. Journal of Molecular Biology 308: 377-395 (2001).
68. Taylor RD, Jewsbury PJ, Essex JW: FDS: Flexible ligand and receptor docking with a continuum solvent model and soft-core energy function. Journal of Computational Chemistry 24: 1637-1656 (2003).
69. Page MI, Jencks WP: Entropic contributions to rate accelerations in enzymic and intramolecular reactions and the chelate effect. Proc Natl Acad Sci U S A 68: 1678-83 (1971).
70. Pickett SD, Sternberg MJ: Empirical scale of side-chain conformational entropy in protein folding. J Mol Biol 231: 825-39 (1993).
71. Nicholls A, Sharp KA, Honig B: Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons. Proteins 11: 281-96 (1991).
72. Luo R, Gilson MK: Synthetic Adenine Receptors: Direct Calculation of Binding Affinity and Entropy. Journal of the American Chemical Society 122: 2934 - 2937 (2000).
73. Wang J, Szewczuk Z, Yue SY, Tsuda Y, Konishi Y, Purisima EO: Calculation of relative binding free energies and configurational entropies: a structural and thermodynamic analysis of the nature of non-polar binding of thrombin inhibitors based on hirudin55-65. J Mol Biol 253: 473-92 (1995).
74. Creamer TP: Side-chain conformational entropy in protein unfolded states. Proteins 40: 443-50 (2000).
75. Lazaridis T, Masunov A, Gandolfo F: Contributions to the binding free energy of ligands to avidin and streptavidin. Proteins 47: 194-208 (2002).
76. Hummer G, Garde S, Garcia AE, Paulaitis ME, Pratt LR: Hydrophobic effects on a molecular scale. Journal of Physical Chemistry B 102: 10469-10482 (1998).
77. Gill SJ, Dec SF, Olofsson G, Wadso I: Anomalous Heat-Capacity of Hydrophobic Solvation. Journal of Physical Chemistry 89: 3758-3761 (1985).
78. Gallicchio E, Kubo MM, Levy RM: Enthalpy-entropy and cavity decomposition of alkane hydration free energies: Numerical results and implications for theories of hydrophobic solvation. Journal of Physical Chemistry B 104: 6271-6285 (2000).
79. Lum K, Chandler D, Weeks JD: Hydrophobicity at small and large length scales. Journal of Physical Chemistry B 103: 4570-4577 (1999).
80. Zhou HY, Zhou YQ: Stability scale and atomic solvation parameters extracted from 1023 mutation experiments. Proteins-Structure Function and Genetics 49: 483-492 (2002).
81. Levy RM, Zhang LY, Gallicchio E, Felts AK: On the nonpolar hydration free energy of proteins: Surface area and continuum solvent models for the solute-solvent interaction energy. Journal of the American Chemical Society 125: 9523-9530 (2003).

82. Muller N: Search for a Realistic View of Hydrophobic Effects. Accounts of Chemical Research 23: 23-28 (1990).
83. Southall NT, Dill KA, Haymet ADJ: A view of the hydrophobic effect. Journal of Physical Chemistry B 106: 521-533 (2002).
84. Gohlke H, Klebe G: Approaches to the description and prediction of the binding affinity of small-molecule ligands to macromolecular receptors. Angewandte Chemie-International Edition 41: 2645-2676 (2002).
85. Halgren TA: Representation of van der Waals (vdW) Interactions in Molecular Mechanics Force Fields: Potential Form, Combination Rules, and vdW Parameters. JACS 114: 7827-7843 (1992).
86. Kramer B, Rarey M, Lengauer T: Evaluation of the FLEXX incremental construction algorithm for protein-ligand docking. Proteins 37: 228-41 (1999).
87. Miranker A, Karplus M: Functionality Maps of Binding-Sites - a Multiple Copy Simultaneous Search Method. Proteins-Structure Function and Genetics 11: 29-34 (1991).
88. Vonitzstein M, Wu WY, Kok GB, Pegg MS, Dyason JC, Jin B, Phan TV, Smythe ML, White HF, Oliver SW, Colman PM, Varghese JN, Ryan DM, Woods JM, Bethell RC, Hotham VJ, Cameron JM, Penn CR: Rational Design of Potent Sialidase-Based Inhibitors of Influenza-Virus Replication. Nature 363: 418-423 (1993).
89. Ringe D, Mattos C: Analysis of the binding surfaces of proteins. Medicinal Research Reviews 19: 321-331 (1999).
90. Nishibata Y, ltai A: Automatic Creation of Drug Candidate Structures Based on Receptor Structure - Starting Point for Artificial Lead Generation. Tetrahedron 47: 8985-8990 (1991).
91. Bohm HJ: The Computer-Program Ludi - a New Method for the Denovo Design of Enzyme-Inhibitors. Journal of Computer-Aided Molecular Design 6: 61-78 (1992).
92. Eisen MB, Wiley DC, Karplus M, Hubbard RE: Hook - a Program for Finding Novel Molecular Architectures That Satisfy the Chemical and Steric Requirements of a Macromolecule Binding-Site. Proteins-Structure Function and Genetics 19: 199-221 (1994).
93. Gillet V, Johnson AP, Mata P, Sike S, Williams P: Sprout - a Program for Structure Generation. Journal of Computer-Aided Molecular Design 7: 127-153 (1993).
94. Rarey M, Kramer B, Lengauer T: Multiple automatic base selection: Protein-ligand docking based on incremental construction without manual intervention. Journal of Computer-Aided Molecular Design 11: 369-384 (1997).
95. Congreve M, Murray CW, Blundell TL: Keynote review: Structural biology and drug discovery. Drug Discovery Today 10: 895-907 (2005).
96. Modica M, Santagati M, Russo F, Parotti L, De Gioia L, Selvaggini C, Salmona M, Mennini T: [[(Arylpiperazinyl)alkyl]thio]thieno[2,3-d]pyrimidinone derivatives as high-affinity, selective 5-HT1A receptor ligands. J Med Chem 40: 574-85 (1997).
97. Globus A, Lawton J, Wipke T: Automatic molecular design using evolutionary techniques. Nanotechnology 10: 290-299 (1999).
98. Rotstein SH, Murcko MA: GroupBuild: a fragment-based method for de novo drug design. J Med Chem 36: 1700-10 (1993).

99. Pegg SC, Haresco JJ, Kuntz ID: A genetic algorithm for structure-based de novo design. J Comput Aided Mol Des 15: 911-33 (2001).
100. Kamphausen S, Holtge N, Wirsching F, Morys-Wortmann C, Riester D, Goetz R, Thurk M, Schwienhorst A: Genetic algorithm for the design of molecules with desired properties. Journal of Computer-Aided Molecular Design 16: 551-567 (2002).
101. Good AC: The calculation of molecular similarity: alternative formulas, data manipulation and graphical display. J Mol Graph 10: 144-51, 162 (1992).