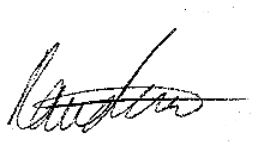
Discrimination of near-native decoy structures using statistical potentials

A thesis submitted for the degree of Doctor of Philosophy at University College London

Rand Hindi
rand.hindi@snips.net
+ 33 (0) 6 26533909
twitter: @randhindi

I, Rand Hindi confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.



Abstract

Being able to select decoy structures that are closest to the native one is essential to any folding simulation. Indeed, modern algorithms use heuristics to quickly sample the conformational space, and as such, will generate a large number of candidate structures.

In this thesis, we create a new statistical energy function to correctly discriminate near-native decoy structures, using three complementary approaches to derive energies from known conformations and decoys.

First, we used a classical definition, where the observed state is modelled by taking a set of 1078 short, well-resolved, non-redundant crystal structures from the PDB, and the reference state is taken as the distribution expected at random. In our second method, which we call “hybrid”, we used the native structures as the observed state, just as in the classical formulation, but this time using the worse generated decoys as the reference state. Finally, our third method, called “decoy-based”, uses only decoys, taking the better than average models as the observed state, and the worse than average as the reference state.

Using the three methods above, we generated potentials to model solvation, hydrogen bonding, and pairwise atomic distances and orientation. We found that overall, combining solvation, atomic distance and orientation using the decoy-based method produced the best results, with a 10% enrichment score of 0.73 versus 0.51 for the classical formulation, and 0.41 for our benchmark potential, DFIRE2.

Our final potential, called the DOS potential, was created by combining the classical, hybrid and decoy-based potentials, and achieved a 10% enrichment score of 0.75 versus 0.41 for DFIRE2.

Acknowledgements

I would like to thank my thesis supervisors, Prof. David T. Jones and Dr Kevin Bryson for their immense patience in helping me complete my thesis. Our hours-long supervisory meetings and debates on various aspects of biology led me to really grasp the intrinsic beauty of bioinformatics, and really stimulated my academic curiosity. I hope these discussions will extend outside the scope of this thesis, and onto practical applications.

I would also like to thank my parents for their continuous support throughout the seven years of studies I have undertaken, and for making it possible for me to stay in London.

Finally, I would like to thank a very close friend of mine, Karim Kaddoura, who has always supported me in every decision I made, and allowed me to follow my ambitions.

Table of Contents

1. GENERAL INTRODUCTION AND BACKGROUND 13

1.1 PROTEIN STRUCTURE 14

1.1.1 Chemical structure of proteins 15

1.1.2 Experimental determination of protein structure 21

1.2 PROTEIN FOLDING 22

1.2.1 Computational protein folding 22

1.2.2 Model quality measures 27

1.2.3 CASP 28

1.3 MOLECULAR DYNAMICS 29

1.4 DECOY SETS 33

1.4.1 Existing decoy sets 33

1.4.2 Potential quality scores 35

1.5 STATISTICAL ENERGY FUNCTIONS 37

1.5.1 Deriving a free energy statistical potential 37

1.5.2 Atomic distance potentials from the literature 40

1.6 HYDROGEN BONDS 41

1.6.1 Hydrogen bond definition 41

1.6.2 Hydrogen bonds in proteins 43

1.6.3 Hydrogen bonding potentials 50

1.7 SOLVATION 54

1.7.1 The hydrophobic effect *54*

1.7.2 Water hydrogen bonds *55*

1.7.3 Explicit models *56*

1.7.4 Implicit models *57*

2. GENERATING A NEAR-NATIVE DECOY SET 59

2.1 INTRODUCTION AND BACKGROUND 60

2.2 METHODS 63

2.2.1 Choice of target proteins *63*

2.2.2 Molecular dynamics runs *66*

2.2.3 Measuring the fitness of an energy function *67*

2.3 RESULTS & DISCUSSIONS 69

2.3.1 Properties of the decoy set generated *69*

2.3.2 Distribution of amino acids *71*

2.3.3 Impact of sequence features on average ARMSD distribution *72*

2.3.4 Impact of simulation parameters on average ARMSD distributions *76*

2.3.5 Measuring compactness of decoys *80*

2.3.6 Performance of existing energy functions *82*

2.4 CONCLUSIONS 83

3. DERIVING A SOLVATION FREE ENERGY POTENTIAL 85

3.1 INTRODUCTION AND BACKGROUND 86

3.2 METHODS 88

3.2.1 Hydrogen atom generation *88*

3.2.2 Protein sets *88*

3.2.3 Solvent model *89*

3.2.4 Statistical potentials 90

3.2.5 Benchmarking potentials 93

3.2.6 Combining energy terms 94

3.3 RESULTS & DISCUSSIONS 95

3.3.1 Generation of the classical reference state 95

3.3.2 Classical potential 98

3.3.3 Decoy-based potential 101

3.3.4 Comparison to potentials from the literature 104

3.4 CONCLUSIONS 107

4. INCLUDING C-H…X HYDROGEN BONDS IN STATISTICAL POTENTIALS 109

4.1 INTRODUCTION 110

4.2 METHODS 112

4.2.1 Protein sets 112

4.2.2 Hydrogen bond definitions 113

4.2.3 Hydrogen generation 116

4.2.4 Hydrogen bonding potentials 116

4.2.5 Performance measures 120

4.2.6 Combining energy terms 120

4.3 RESULTS & DISCUSSIONS 121

4.3.1 Accuracy of the hydrogen atoms generation method 121

4.3.2 Univariate statistical potentials 123

4.3.3 Bivariate statistical potentials 132

4.3.4 Inclusion in the DFIRE2 potential 142

4.4 CONCLUSIONS 143

5. DERIVING DISTANCE, ORIENTATION-DEPENDENT ATOMIC POTENTIALS 145

5.1 INTRODUCTION 146

5.2 METHODS.....148

5.2.1 Protein sets 148

5.2.2 Interaction model.....149

5.2.3 Statistical potentials 150

5.2.4 Combining potential terms 152

5.2.5 Performance measures 153

5.3 RESULTS & DISCUSSIONS.....154

5.3.1 Component analysis of DFIRE2.....154

5.3.2 Distance potentials 155

5.3.3 Angle potentials 157

5.3.4 Dihedral potentials.....164

5.3.5 Comparison to DFIRE2 165

5.4 CONCLUSIONS.....170

6. GENERAL CONCLUSIONS 172

7. BIBLIOGRAPHY.....177

List of Tables

Table 1.1 Amino acids properties 15

Table 1.2 Common decoy sets in the literature 33

Table 2.1 Properties of common decoys sets.....60

Table 2.2 Filters used to select decoy set targets 63

Table 2.3 Properties of the decoy set 69

Table 2.4 Amino acid frequencies in the decoy set and PDB 71

Table 2.5 P-value of the ARMSD difference across CATH classes.....74

Table 2.6 Average residue separation in hydrogen bond 75

Table 2.7 Average scores of the radius of gyration for the MDSET decoy set 81

Table 2.8 Comparison of energy function on the MDSET 82

Table 3.1 Naming conventions for potentials.....92

Table 3.2 Proportions of elements in proteins 96

Table 3.3 10% Enrichment score of the classical SASA potential 100

Table 3.4 10% Enrichment score of the Decoy-based SASA potential 103

Table 3.5 10% Enrichment score of the different potential terms.....104

Table 3.6 Performance of the Decoy-based SASA potential on the HRDECOY set ...105

Table 3.7 Performance scores of the DFIRE2 + SASAD potential 106

Table 4.1 Potentials bin size and range 119

Table 4.2 Naming conventions for potentials.....120

Table 4.3 RMSD of real vs. virtual hydrogens 121

Table 4.4 10% enrichment score for univariate potential terms.....130

Table 4.5 10% Enrichment score for univariate potential combinations.....131

Table 4.6 Preferred regions centres of pairwise features of NH bonds.....134

Table 4.7 Preferred regions centres of pairwise features of CH bonds.....136

Table 4.8 10% enrichment score for bivariate potential terms.....140

Table 4.9 10% Enrichment score for bivariate potential combinations.....141

Table 4.10 Comparison of univariate and bivariate NH potentials.....141

Table 4.11 10% enrichment of DFIRE2 with and without hydrogen bonds.....142

Table 5.1 Potential generation methods.....150

Table 5.2 Potentials residue separation 151

Table 5.3 Geometrical features modelled 151

Table 5.4 Value range for each feature 152

Table 5.5 DFIRE2 energy terms scores.....154

Table 5.6 Distance potentials 10% enrichment for the MDSET 156

Table 5.7 Angle potential 10% enrichment for the MDSET 162

Table 5.8 Dihedral potentials 10% enrichment for the MDSET.....164

Table 5.9 Full potentials 10% enrichment for the MDSET 165

Table 5.10 DOS potential scores.....167

Table 5.11 DOS and DFIRE2 MDSET targets scores 168

List of Figures

Figure 1.1 Amino acids side chains....16
Figure 1.2 Amino acid backbone....17
Figure 1.3 Ramachandran plot of φ / Ψ angles in proteins....18
Figure 1.4 Alpha helix....19
Figure 1.5 Anti-parallel beta sheet....19
Figure 1.6 Tertiary structure of HIV-1 integrase....20
Figure 1.7 Methods of protein structure determination....23
Figure 1.8 Graphical representation of the enrichment score....37
Figure 1.9 From PDB structure to energy function....38
Figure 1.10 Hydrogen bond geometry....42
Figure 1.11 Distribution of hydrogen bond features in proteins....44
Figure 1.12 Cone correction of DHA angles in NH hydrogen bonds....45
Figure 1.13 Distribution of backbone hydrogen bond features....46
Figure 1.14 HA distance in inorganic CH hydrogen bonds....48
Figure 1.15 Coupling of NH-O and CH-O main chain hydrogen bonds....49
Figure 1.16 Switching (SW) truncation function....50
Figure 1.17 Hydrogen bonding potential performance....53
Figure 1.18 Hydrogen bonds in water....55

Figure 2.1 All Atom RMSD of the decoys in the 4state_reduced set....61
Figure 2.2 Average All-Atom RMSD versus Sequence Length....72
Figure 2.3 Average RMSD distributions per CATH class....73
Figure 2.4 All-Atom RMSD versus Simulation time....76
Figure 2.5 ARMSD versus time for different starting structures....77

Figure 2.6 ARMSD versus time for the decoy seeded simulation 78

Figure 2.7 Average radius of gyration vs. ARMSD 80

Figure 3.1 Distribution of protein radius 95

Figure 3.2 Distributions of SASA for random atoms in protein-like spheres 97

Figure 3.3 SASA Classical potentials for Threonine heavy atoms 99

Figure 3.4 SASA Decoy-based potentials for Threonine heavy atoms 102

Figure 4.1 CH donor groups 114

Figure 4.2 Hydrogen bond geometry 115

Figure 4.3 NH bonds classical observed states 124

Figure 4.4 CH bonds classical observed states 125

Figure 4.5 NH bonds potentials 128

Figure 4.6 CH bonds potentials 129

Figure 4.7 NH paired terms distributions 133

Figure 4.8 CH paired terms distributions 135

Figure 4.9 NH classical bivariate potentials 137

Figure 4.10 CH classical bivariate potentials 138

Figure 5.1 DFIRE versus dDFIRE 146

Figure 5.2 Atomic interaction model 149

Figure 5.3 ABC angle distributions 158

Figure 5.4 ABC angle classical potentials 159

Figure 5.5 ABC angle decoy-based potentials 160

Figure 5.6 ABC angle Hybrid potentials 161

Figure 5.7 Performance of DFIRE2 and DOS on targets 1JYH and 3EOI 169

Chapter 1

General introduction and background

1.1 Protein structure

Ever since the original work of Mendel in the 19th century, the mechanisms underlying the transmission and modifications of traits in living organisms has been at the centre of biological research. For the past fifty years, the effects of genetic mutations on the structure of biological molecules have been studied intensively, with the advent of fields such as biochemistry, biotechnology or bioinformatics.

In living organisms, molecules called proteins, which are chains of amino acids folded into a compact and usually unique structure, perform most cell functions. These are generated in the cell by translating a given RNA sequence into its corresponding amino acid sequence.

Protein structure can be described at four levels: its amino acid sequence, its secondary structure, which is the local arrangement of amino acids into sheets and helices, its spatial conformation (tertiary structure), and its quaternary structure which is the complex formed by different amino acid chains.

One of the major challenges in modern biology has been to predict the tertiary structure from the sequence only (*de novo* folding, also called *ab initio* folding). The most accepted hypothesis is that an amino acid sequence will fold into a structure that lies at the bottom of the free energy landscape [Anfinsen 1973]. It is this principle that enables computational algorithms to be used to predict that conformation.

1.1.1 Chemical structure of proteins

Living organisms translate DNA codons into amino acids, which are chained to produce proteins. The covalent bond linking these amino acids is called the peptide bond. Chemically, amino acids share a common backbone, but have different side chains giving them their unique properties. Some common properties are listed in Table 1.1.

Table 1.1 Amino acids properties

Name	Code	Letter	Polarity	Aliphatic	Aromatic	Acidity
Alanine	Ala	A	-	Y	N	Neutral
Arginine	Arg	R	+	N	N	Basic
Asparagine	Asn	N	+	N	N	Neutral
Aspartate	Asp	D	+	N	N	Acid
Cysteine	Cys	C	-	N	N	Neutral
Glutamine	Gln	Q	+	N	N	Neutral
Glutamate	Glu	E	+	N	N	Acid
Glycine	Gly	G	-	Y	N	Neutral
Histidine	His	H	+	N	N	Basic
Isoleucine	Ile	I	-	Y	N	Neutral
Leucine	Leu	L	-	Y	N	Neutral
Lysine	Lys	K	+	N	N	Basic
Methionine	Met	M	-	N	N	Neutral
Phenylalanine	Phe	F	-	N	Y	Neutral
Proline	Pro	P	-	Y	N	Neutral
Serine	Ser	S	+	N	N	Neutral
Threonine	Thr	T	+	N	N	Neutral
Tryptophan	Trp	W	-	N	Y	Neutral
Tyrosine	Tyr	Y	+	N	Y	Neutral
Valine	Val	V	-	N	N	Neutral

Amino acid main chains are composed of 2 carbon atoms, 1 nitrogen atom, 1 oxygen atom, and 2 hydrogen atoms (aside from Proline that only has one main chain hydrogen). The side chain is attached to the central carbon, with the exception of Proline, where a ring is formed between the side chain atoms and the main chain nitrogen. Figure 1.1 shows the different amino acid side chains.

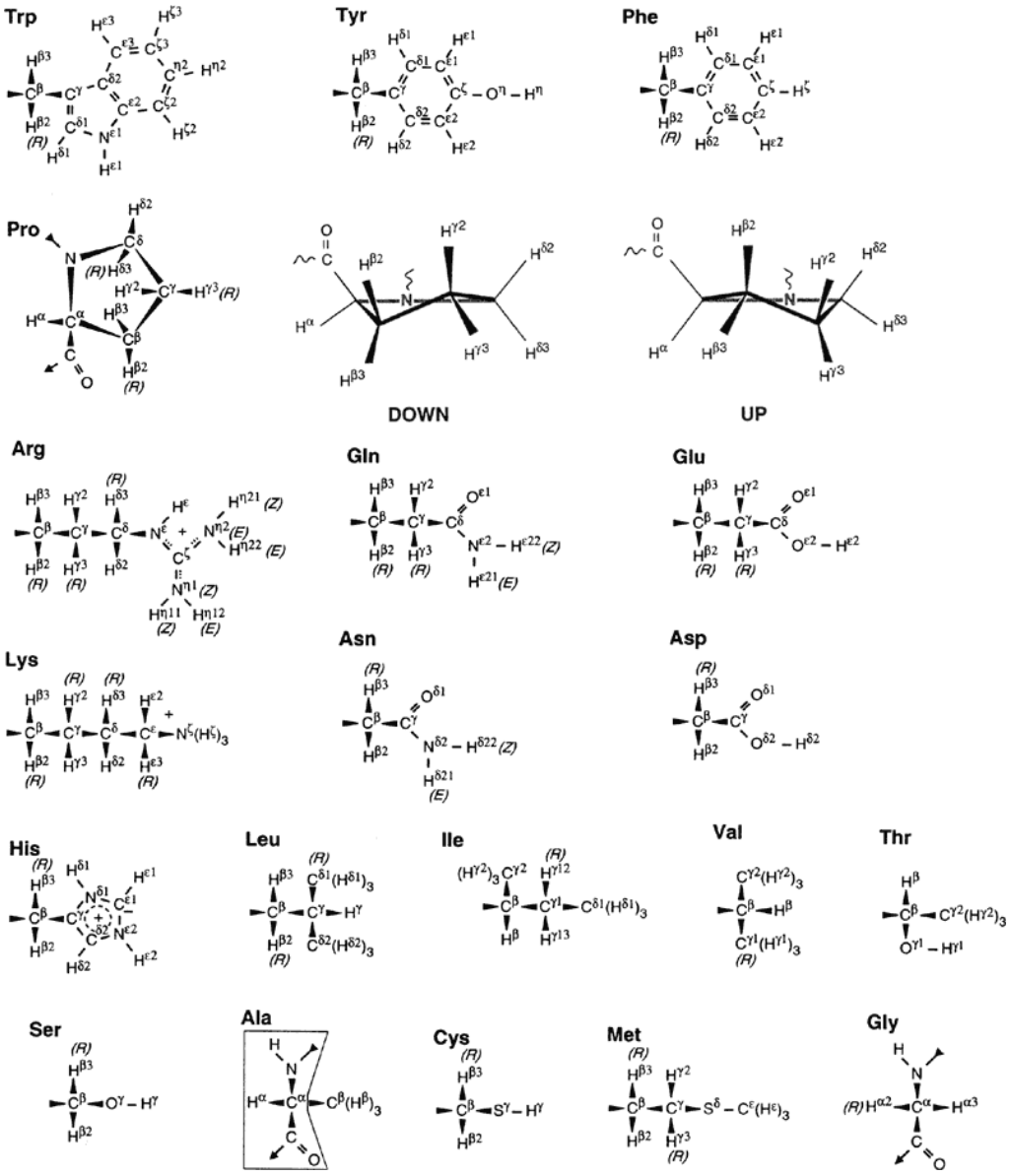


Figure 1.1 Amino acids side chains [Andersen 2010]. Main chains are not represented here, as they all have the same form (shown boxed for the Alanine - Ala), aside from Proline, which is represented fully.

The backbone in a protein can be described either by Cartesian coordinates for each main chain atom, or by the torsion angles, represented in Figure 1.2.

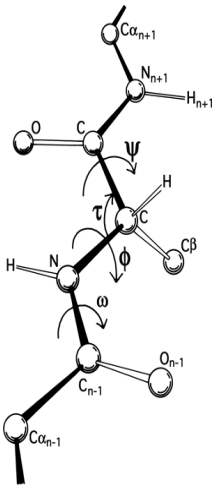


Figure 1.2 Amino acid backbone [Richardson 1981]. We can see that there are three torsion angles: the rotation about the carbonyl carbon and the alpha-carbon (the Ψ angle), the rotation about the alpha-carbon and the nitrogen (the ϕ angle), and finally the peptide bond rotation about the nitrogen and the carbonyl carbon from the residue before it (the ω angle).

In nature, the peptide unit between two residues is planar, meaning that the ω angle is 180 degrees (trans isomer), although in some cases it can be 0 degrees (cis isomer). For the Ψ and ϕ angles, their values are constrained to specific regions, as can be observed in a Ramachandran plot (Figure 1.3).

This arrangement of the backbone into specific structural elements is referred to as the secondary structure of the protein, and is mostly due to the hydrogen bonds occurring between different residue main chains. The two most common types of secondary structure elements are alpha-helices and beta-sheets, which are formed through hydrogen bonding between different amino acid main chain donors and acceptors [Bordo & Argos 1994].

Alpha-helices form when the amine nitrogen from a residue forms a hydrogen bond with the carbonyl oxygen from an amino acid four residues earlier. Its topology is shown in Figure 1.4.

The other common secondary structure element is the beta sheet, composed of two or more strands that are hydrogen bonding in at least 3 places. Depending on the orientation of the N-termini of adjacent residues, they can be categorised as parallel, or anti-parallel. The anti-parallel arrangement is usually preferred since it allows for linear hydrogen bonds. Figure 1.5 shows an example of parallel and anti-parallel strands of a mixed beta sheet.

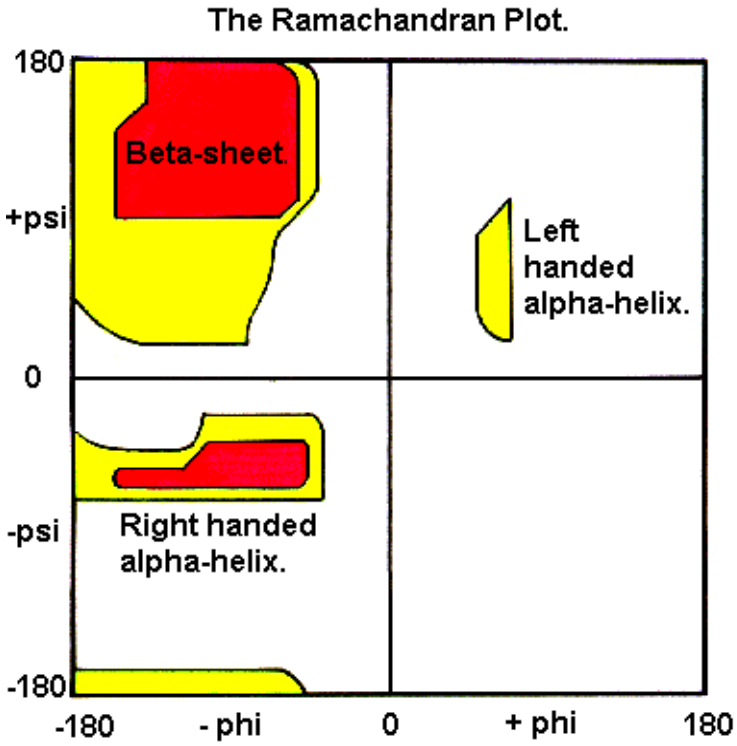


Figure 1.3 Ramachandran plot of ϕ / Ψ angles in proteins [Cooper 1995]. This plot shows strong clusters of ϕ and Ψ values, corresponding to local conformational arrangements known as the secondary structures. These arrangements define the secondary structure of the protein, and are important in stabilising it. Depending on the secondary structure element considered, ϕ and Ψ angles cluster around different values [Ramachandran & Sasisekharan 1968].

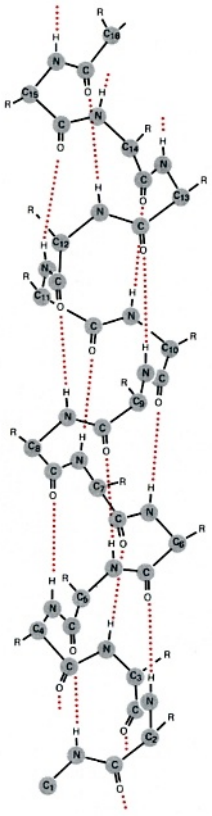


Figure 1.4 Alpha helix [Kimball 2011]. A right handed alpha helix.

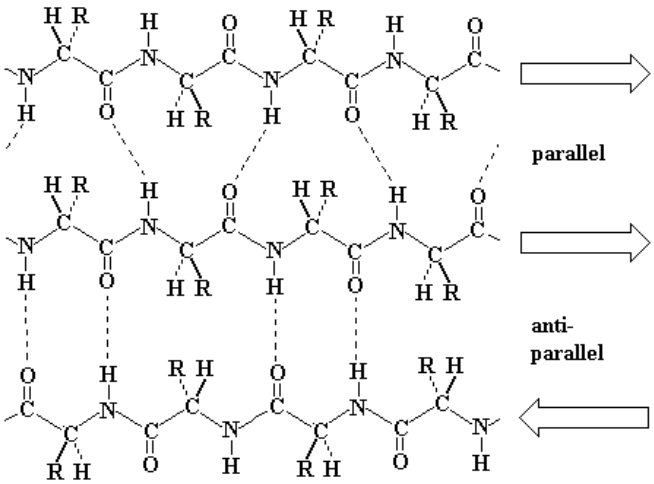


Figure 1.5 Mixed beta sheet [Keates 1998]. Front view, showing the hydrogen bonds (dotted) between NH and CO groups on adjacent strands. Arrows indicate chain direction.

In some proteins, certain conditions can induce switches from one secondary structure element to the other, for example in amyloids [Mimna et al. 2007]. Since the sequence residue separation between hydrogen bonding donors and acceptors is smaller in helices than sheets, the former can be thought of as a local stabilisation of the structure, whereas the latter will have a more global effect [Kamat & Lesk 2007].

The tertiary structure of a protein is defined as the three dimensional arrangement of the atoms forming it, the quaternary structure being the arrangement of the different chains. It is believed that for each sequence there is usually one stable conformation [Anfinsen 1973], and this is what theoretical protein folding tries to reproduce. Figure 1.6 shows the tertiary structure of the HIV-1 integrase protein.

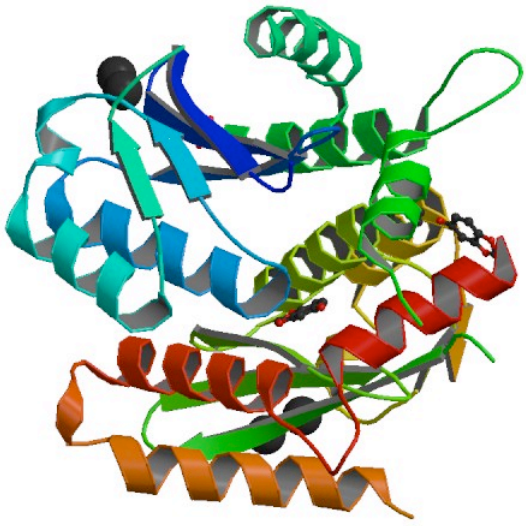


Figure 1.6 Tertiary structure of HIV-1 integrase [Protein Data Bank, 3OVN, 2010]. The twists represent the alpha-helices, while the arrows represent the beta-sheet strands, with their direction.

1.1.2 Experimental determination of protein structure

Protein structures can be resolved experimentally more easily than before, meaning that the number of available structures increases faster with time [PDB statistics, 2010], requiring an efficient database to store and organise them. The largest one is the Protein Data Bank (PDB) [Bernstein et al. 1977], containing more than 60000 structures in 2010.

These structures are usually resolved using X-ray crystallography or Nuclear Magnetic Resonance (NMR), but other experimental methods exist, such as using neutrons instead of X-rays in crystallography, which allows for hydrogen atoms to be resolved. This last method is usually used in combination with conventional x-ray crystallography to produce structures with good resolution, as well as the hydrogen atoms.

The first and most widely used method for determining a protein structure experimentally is by X-ray crystallography, in which a protein structure is determined using the X-ray diffraction patterns from a crystal. Proteins are hard to crystallise well, often producing imperfect crystals, making this method inefficient for large-scale structure determination. The resolution of a structure is the principal parameter to consider when choosing a subset of protein structures, but other ones, such as the R-factor, corresponding to the agreement between the diffraction data and the model used, should be taken into account as well.

Another experimental method called Nuclear Magnetic Resonance (NMR) spectroscopy exploits the magnetic properties of certain atomic nuclei to determine physical and chemical properties of the molecules they are contained in. The advantage is being able to resolve proteins in solution, but the problem is that due to the dynamic nature of those proteins in solution, it generates more than one possible conformation, making it impossible to choose only one position for every atom in the protein.

Since experimental methods are still expensive and cannot be applied to all proteins, researchers have tried to simulate the folding process in order to theoretically predict the tertiary structure. The next section introduces the various categories of computational structure prediction methods, followed by an introduction to molecular dynamics.

1.2 Protein folding

1.2.1 Computational protein folding

Resolving the structure using experimental methods is not practical, as it is slow, expensive, and not always feasible. The gap between the number of proteins that have been sequenced and the ones that have been structurally resolved is increasing fast [Levitt 1998]. Therefore, we need a way to predict this structure quickly from sequence using computational algorithms. This would allow a full analysis of the structural properties of the genes being sequenced, and thus, a better understanding of the interactions with them, leading to better and novel drug designs.

Over the last twenty years, many methods have been developed to address this problem, with various results [Dill et al. 2007]. They can be divided into three categories, depending on the availability of a homolog. The different types of prediction methods are summarised in Figure 1.7.

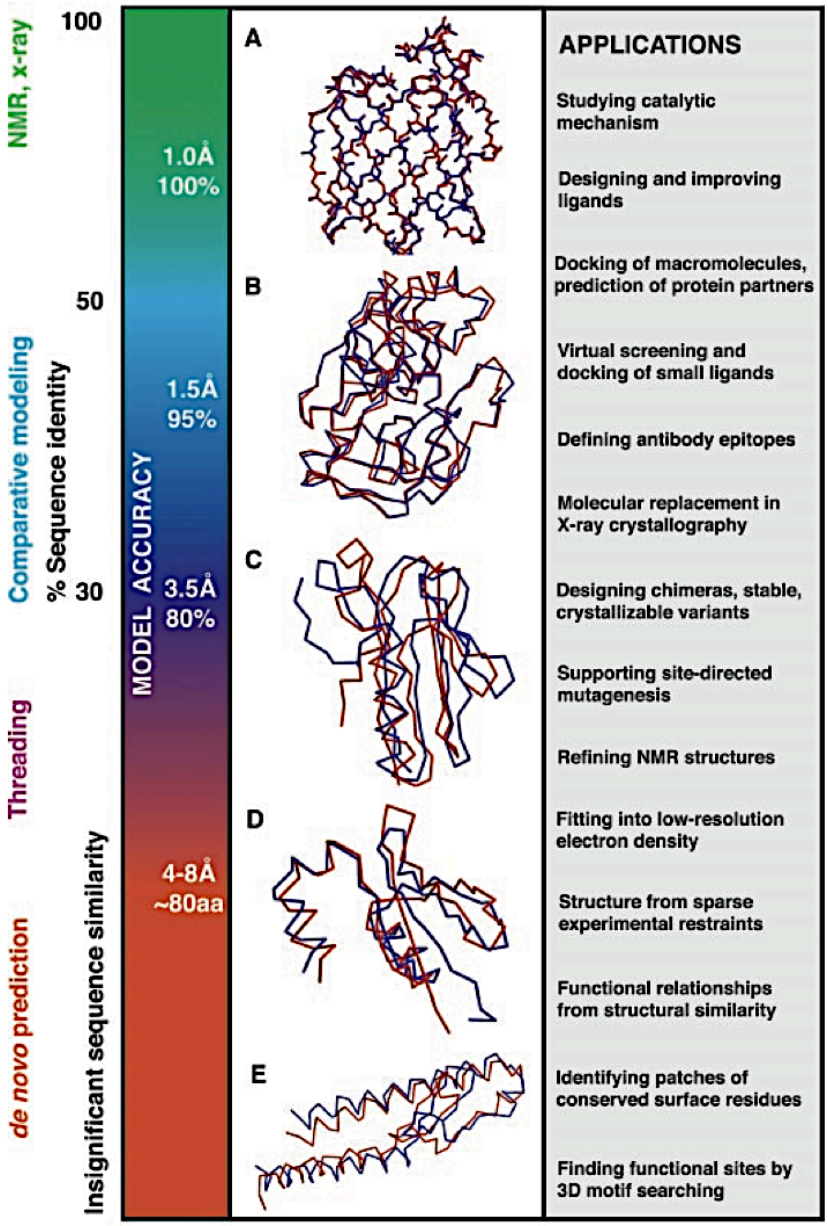


Figure 1.7 Methods of protein structure determination [Baker and Sali 2001]. This figure shows different methods of protein structure determination, with the accuracy that can be expected from it, and the application that could arise from the model created.

The first category is concerned with modelling proteins for which a homolog exists, in which case it is possible to use it as a template for the new protein [Baker & Sali 2001]. As can be seen from the various CASP (Critical Assessment of Structure Prediction) experiments, this is still the best approach whenever it is possible to find a template [Moult 2005].

In order to do so, it is necessary to find a set of similar sequences, using programs such as PSI-BLAST [Altschul et al. 1997], which take the amino acid sequence as input, and return a multiple alignment and a profile by searching a given database. The profile is then compared to each entry in the same database, and the significance of each alignment is returned. This procedure runs iteratively until a number of epochs are reached, or the profile converged. Different criteria can then be used to select a template from the alignments.

The most important factor to consider is the sequence similarity. Comparative modelling usually works when a template has more than thirty percent sequence identity, but some programs, such as MODELLER [Sali & Blundell 1993, Eswar et al. 2006], use more than one template to build a structure. This can be a useful approach when no highly similar structures exist. To build the model, various methods exist, such as rigid region assembly, where the model is built using fragments from the aligned templates and is then refined using a potential by searching for conformations with lower energies. The refinement step is necessary to model highly variable regions such as loops or side chains.

The second most successful method is fold recognition, also called threading [Jones et al. 1992]. When there are no homologs in the PDB with more than thirty percent sequence identity, comparative modelling usually fails to generate a native like model. Because sequences change at a faster rate than structures [Chothia & Lesk 1986], it is possible to find a fold that corresponds to the protein we are modelling, even though there is not a large similarity between sequences. By using an energy function derived from the existing database of experimental structures, it is possible to compare folds and select the most likely one to use as template [Jones & Thornton 1996]. The folds having the lowest energies are selected as candidates, and a similar approach to comparative modelling is then used to build the model. One widely used threading program is GenTHREADER [Jones 1999], the practicality of which has been shown by carrying out structural prediction of entire genomes (although the method did not work for transmembrane proteins).

The third and final prediction category is called *de novo* protein folding (also called *ab initio*). It relies on the observation that all the information needed to fold a protein is contained in its sequence [Anfinsen 1973]. Without relying on a template, *de novo* methods try to build a structure by searching the conformational space with the help of an energy function. There are various methods for generating conformations, the most widely used and successful one being fragment assembly [Moult 2005].

The underlying assumption behind fragment assembly methods is that local structures are more conserved than overall folds, and therefore it is possible to reuse fragments of known structures as a template for modelling.

To do so, a list of existing local conformations is extracted from the PDB, and stored in a fragment database. Fragments are sometimes selected by threading, with confirmation from the secondary structure predictions, allowing for fragments of different lengths to be selected for different motifs. Once a list of fragments have been chosen, the method will go through combinations of them to generate tertiary structures by selecting fragments based on various criteria, such as secondary structure elements, paired beta strands, absence of steric clashes, and low energy features.

If these criteria fail, then a new conformation is generated until it satisfies the above requirements. Since the energy landscape is not smooth, the search method must be able to get out of local minimum traps in order to reach the global free energy minimum. Searching the energy landscape exhaustively would require enormous computing power, so heuristic algorithms are usually used instead. One such algorithm is simulated annealing, where at each step a neighbour of the current state is considered. A probabilistic function then decides whether or not to switch to that neighbour, with the requirement that ultimately the states are driven towards lower energies. This step is then iterated a number of times, until there is a convergence or the number of allowed steps is reached. By using fragments, it is possible to make large moves in one single step, allowing the search to be conducted more rapidly. But due to the heuristic nature of the algorithm, the final structure will often be one that was trapped in a local minimum, since the energy landscape is usually rugged. For that reason, it is very important to consider the energy function as one of the core element of the folding process. Two of the most successful algorithms, according to CASP, are FRAGFOLD [Jones 2001] and ROSETTA [Simons et al. 1997].

Simulated annealing is not the only existing search algorithm though. Another option is based on genetic algorithms [Unger 2004]. The energy function can be the same as the one used with simulated annealing, but the structure is generated by mutating intermediate conformations using different mutation operators, such as crossovers, where two models swap some parts of their structure, or pure mutations where structures are mutated directly by substituting one fragment by another from the selected database.

One major problem with predicting protein folds is the enormous number of possible structures that need to be searched. An experiment has been conducted where a massive network of paralleled computers was used to search the conformational space exhaustively. The conclusions were that current algorithms have sampling bottlenecks, and thus, cannot effectively sample the entire conformational space [Kim et al. 2010]. This implies that whatever the available computing power, and however good the energy function is, there might be cases where the native structure cannot be reached.

1.2.2 Model quality measures

Measuring the quality of a decoy structure can be done by comparing the position of the atoms in the decoys to those in the native structure, provided that the latter is available. The most widely used measure is the root mean square deviation (RMSD) between the atoms in the two structures. A very efficient algorithm for calculating it has been developed recently, using quaternions [Coutsias et al. 2004].

The RMSD is commonly calculated from the position of the CA carbon on the backbone, but using all heavy atoms (ARMSD) can be more suitable when comparing high quality models that have only small differences, and almost identical folds. Moreover, the added information from the side chain positions means that the ARMSD will be sensitive to the quality of the side chain modelling, whereas C-alpha RMSD only detects backbone variations.

The major drawback with RMSD (or ARMSD) is that it is very sensitive to local differences. RMSD works well when decoys are homogeneously deviating from the crystal structure, but not when they have large variations in few places only. This locality problem has been addressed by other measures, and the most commonly used alternative is the Global Distance Test (GDT) [Zemla et al. 1999], where given two superposed structures and a threshold, it returns the length of the largest substructure in which the equivalent residues in the two superposed structures are at a distance below the given threshold, divided by the total number of residues. GDT-TS is an extension of GDT where the average of the GDT for thresholds of 1, 2, 4, and 8 Å are used.

The TM score [Zhang & Skolnick 2004] has a similar formulation, but uses the distance between superposed residues as well as a length-based distance threshold to produce a score between 0 and 1, where 1 is the native structure.

1.2.3 CASP

In order to assess the efficiency of individual methods, a biennial blind test has been created to compare algorithms in different categories. The Critical Assessment of Structure Prediction (CASP) [Bourne 2003, Moult 2005, Kryshtafovych et al. 2009] has run nine times already, showing consistent improvement in protein folding predictions. The resulting competitiveness is boosting the field towards finding new approaches.

Over the years, comparative modelling has produced the best results provided that a template was available, but the frontier between fold recognition and *ab initio* modelling has narrowed, to a point where some *ab initio* methods outperformed threading methods for some protein in the latter category [Moult 2005].

The way in which each category is assessed varies; for comparative modelling, more attention is given to the structure alignments, the side chain building, and the accuracy of fragment that cannot be copied from the template. In fold recognition, the criteria are the recognition of folds, the quality of the model and the structural alignment to the related fold. It should be noted that the fold recognition category has been divided into two subfamilies, namely the homologous fold family, where an ancestor with a similar fold exists, and the analogous fold family, where the fold results from an evolutionary convergence. Finally, for *ab initio* modelling, the two important assessment factors are the fraction of correctly predicted regions, and the success in finding the overall fold.

1.3 Molecular dynamics

Another way of doing *de novo* structure prediction or refinement is using Molecular Dynamics (MD), which takes a force field and a motion algorithm to simulate the Newtonian movement of particles during the folding process.

The trouble is that the large number of atoms considered, combined with the little steps taken each time, makes the algorithm slow. To address this problem, purpose-specific machines for molecular dynamics have been built, optimising the entire process [Shaw et al. 2007]. Although many algorithms exist for calculating the trajectory of atoms, the required number of steps is too big to be tractable in the case of large protein folding simulation. Typically, molecular dynamics simulations use explicit solvent to model the hydrophobic effect that is thought to be the main driving factor behind protein folding. Since most of the simulated system will be water, the cost of running the algorithms on the solvent will be big, and will considerably slow down the procedure, making this unsuitable for large scale folding predictions.

The interaction between the atoms is modelled by a force field, which is usually comprised of five components: a bond stretching term, an angle bending term, a dihedral term, a van der Waals term, and an electrostatic term [Martin 2006]. Parameters for each of these terms are derived from empirical experiments conducted on usually small molecules.

The most commonly used MD simulation programs are AMBER [Pearlman et al. 1995], CHARMM [Brooks et al. 1983, Brooks et al. 2009] and GROMACS [Berendsen et al. 1995, Lindahl et al. 2001, Hess et al. 2008]. The latter is freely available, and thus, largely used in the literature. Available as part of the GROMACS package is the OPLS/aa force field [Jorgensen & Tirado-Rives 1988], which is often used in molecular dynamics studies of proteins. In fact, the parameters of each term are tuned for proteins, and should be used in conjunction with an explicit solvent model. The OPLS/aa force field can be defined as an all atom linear combination of the five terms stated above (Equation 1.1).

$$E_{OPLS} = E_{Bond} + E_{Angle} + E_{Dihedral} + E_{vdw} + E_{coulomb} \quad (1.1)$$

E_{Bond} represents the bond stretching term, as defined in Equation 1.2.

$$E_{Bond} = k_b\left(l - l_0\right)^2 \quad (1.2)$$

In (1.2), k_b is the force constant, and is dependent on the atoms being considered. Only covalently bonded atoms are considered here. l is the length of the bond, and l_0 the equilibrium length. Our second term, the angle bending term, can be expressed similarly:

$$E_{Angle} = k_a\left(\theta - \theta_0\right)^2 \quad (1.3)$$

In (1.3), k_a is the force constant and is dependent on the triplet of atoms covalently bonded. θ is the angle, and θ_0 is the equilibrium angle. The dihedral term is different, and typically hard to model due to the multiple minima that need to be accounted for. Here, a cosine functional form is used, as shown in (1.4).

$$\mathrm{E}_{Dihedral} = \sum_{n=1}^{4} \frac{V_n}{2}\left\{1 + (-1)^{n+1}\cos[n(\phi\text{-}\phi_0)]\right\} \quad (1.4)$$

In (1.4), V_n is a parameter defining each minimum, n is the phase number, ϕ is the dihedral angle, in radians, and ϕ_0 the phase shift. This term is commonly used across all force fields, with differences being mostly in the values of the parameters. The last two terms deal with non-bonded interactions, composed of a Coulomb electrostatic term, and a van der Waals term, which is usually where each force field differs in terms of their functional form. In both terms, only atoms separated by at least 3 covalent bonds are included. A Coulomb potential represents the electrostatic term, shown in Equation 1.5.

$$E_{coulomb} = \sum_{i}^{atoms} \sum_{i \neq j}^{atoms} \frac{q_i q_j e^2}{4\pi\varepsilon_0 r_{ij}} \qquad (1.5)$$

Here, q_i and q_j are the charges on atoms i and j, r_{ij} is the distance between atoms i and j, and e is the Coulomb constant, and ε_0 the dielectric constant. The other non-bonded term in the OPLS/aa force field is the van der Waals interaction, which is modelled using a 6-12 Lennard-Jones potential (Equation 1.6).

$$E_{vdw} = \sum_{i}^{atoms} \sum_{j \neq i}^{atoms} 4\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] \qquad (1.6)$$

In this term, σ_{ij} is the distance at which the potential is 0, r_{ij} is the distance between atoms i and j and ε_{ij} is the depth of the potential well. This formulation is most widely used because of its computational efficiency, and can sometimes be modified to only include atoms that are below a certain cutoff to speed up calculation, such as in CHARMM [Brooks et al. 1983].

In order to generate structures from this force field, we need a search algorithm, of which the molecular dynamics (MD) one is the most widely used. The problem is that if a starting configuration is very far from equilibrium, the forces may be excessively large and the MD simulation may fail. In those cases a robust energy minimization is required. Another reason to perform an energy minimization is to remove any steric clashes and normalize covalent bond lengths and angles. MD simulations are classical, in the sense that they solve the Newtonian equations of motion by running a numerical simulation. The equations of motion define the trajectory of the atoms in the protein, as shown in Equations 1.7 and 1.8.

$$F_i = m_i \frac{d^2 r_i}{dt^2} \tag{1.7}$$

$$F_i = -\frac{\partial V}{\partial r_i} \tag{1.8}$$

Here, *m* is the mass of the atom considered, *r* is the spatial coordinate, *t* is the time, and *V* is the potential function, as defined previously. The simulation requires the force field to be differentiable, and simulates the trajectory by incrementing the time *t* by a little amount. The coordinates generated at each time step forms the trajectory.

1.4 Decoy sets

One of the major components of any *ab initio* algorithm is the energy function, which ideally maps the energy of a conformation to its nativeness, which is how close the modelled conformation is compared to the native one. In order to test this correlation, we need a set of non-native structures with associated RMSDs, for which we will calculate the energy of each structure and calculate various correlation statistics between the RMSD and energy. Such sets are called Decoy Sets, and are typically composed of native structures, together with a preferably large number of misfolded structures (decoys).

1.4.1 Existing decoy sets

Many decoy sets exist in the literature, and have often been designed with specific goals in mind. Decoy sets can be classified in two categories: the first where decoys are generated from folding simulations, and the other where decoys are generated by relaxing the native structure. Since it is usually impossible to get near the native structure in *ab initio* folding, most near-native decoys are generated by relaxing the native structure. A summary of the most common decoy sets is given in Table 1.2 [Park & Levitt 1996, Kaesar & Levitt 1999, John & Sali 2003, Tsai et al. 2003, Rajgaria et al. 2006].

Table 1.2 Common decoy sets in the literature

Decoy set	Targets	Decoys / target	RMSD range (Å)	Generation method
Tsai	78	1400	1 - 13	Rosetta
LMDS	11	439	2 - 14	Random sampling
$4state_{reduced}$	7	665	1.7 - 10	Relaxation of residues
Moulder	20	300	7 - 20	Comparative modelling
HRDECOY	1400	1000	1 - 8	Native Structure Relaxation

In these decoy sets, very few decoy sets have near native structures, but instead have average RMSD values of 4 Å or more. Thus, they would not be suitable for near native decoy analysis, as the number of close to native decoys might not be sufficient to draw significant conclusions.

The main utility of decoy sets is to assess the performance of energy functions at finding the best models. Indeed, if many models are generated, it is important to know how to choose the best ones. Usually, three criteria are required when designing a potential [Gilis 2004]. First, the native structure should have the lowest energy (thermodynamic hypothesis). Second, each structure should have a unique energy (no tied energy ranks), allowing discriminating between individual decoys. Finally, the energy of the decoys should be correlated to their RMSD. If an energy function passes these three tests then it can be considered useful in selecting good candidate conformations.

It has been observed that in most potential functions, the energy of the native structure is not at the global minimum, and that near native decoys will often have the same or lower energies [Brooks & Karplus 1983]. This is due to the fact that there are many viable structures around the native conformation, and discriminating them tends to be difficult with a simplified representation of the protein.

One consideration to keep in mind is that statistical potentials can only discriminate decoys if the bin size used for the potential is small enough to separate them. For that reason, a model within that bin size distance from the native structure is usually considered a good model. To refine near-native decoys using a statistical potential, it is necessary to use smaller bin sizes to account for more subtle differences that would not otherwise be seen by a more approximate energy function, or use more features to filter out more decoys.

Finally, it is important to distinguish decoys that are below six Ångstroms RMSD from the native structure from the ones above since the overall fold of the protein is likely to be incorrect, and even more so when there are no obvious structural defects.

1.4.2 Potential quality scores

When assessing the quality of a potential function using a decoy set, one must choose how to measure the correlation of the energy and similarity to the native structure. After choosing a method to measure the difference between the native structure and the decoy (such as RMSD), a good quality test should take the following into consideration: the energy should be correlated to the nativeness, i.e. the energy should decrease with increased nativeness.

Various statistics exist, such as the Pearson correlation, the Kendal tau and the enrichment score. The Pearson correlation coefficient, defined in Equation 1.9, calculates the degree of linear correlation between two random variables, and is the fraction of the covariance of variables X and Y over their standard deviations.

$$r = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y} \tag{1.9}$$

The problem with this measure is that it assumes a linear correlation between the two variables, which is often not the case when comparing energies and nativeness. Moreover, it is very sensitive to outliers, and can thus show correlation when none exists. One way to overcome linearity problems is to use ranks rather than values, and measure the correlation between them. One of the commonly used rank correlation statistic is the Kendall tau, which compares the correspondence between two rankings. It is defined in Equation 1.10, where P represents the number of concordant pairs, and n the total number of data points.

$$\tau = \frac{2P}{\frac{1}{2}n(n-1)} - 1 \tag{1.10}$$

Although the Kendall Tau addresses the issues of the Pearson correlation, in practice we may want to assess only the very best models, as they are those we want to choose. One measure that specifically addresses that problem is the enrichment score [Tsai et al. 2003]. Given a couple of random variables X and Y and an arbitrary ratio c, the enrichment is the fraction of values in the lowest c percent of X that are also in the lowest c percent of Y, divided by what would be expected from a uniform distribution. Values above one show a better enrichment than that of a uniform distribution. Equation 1.11 shows the functional form of the enrichment.

$$enrich = \frac{N_c}{n \times c^2} \tag{1.11}$$

Here, N_c is the cardinality of the intersection of the c% best values in X and the c % best values in Y (the ratio c is taken between 0 and 1 here), and n denotes the number of items. Figure 1.8 shows the enrichment score graphically.

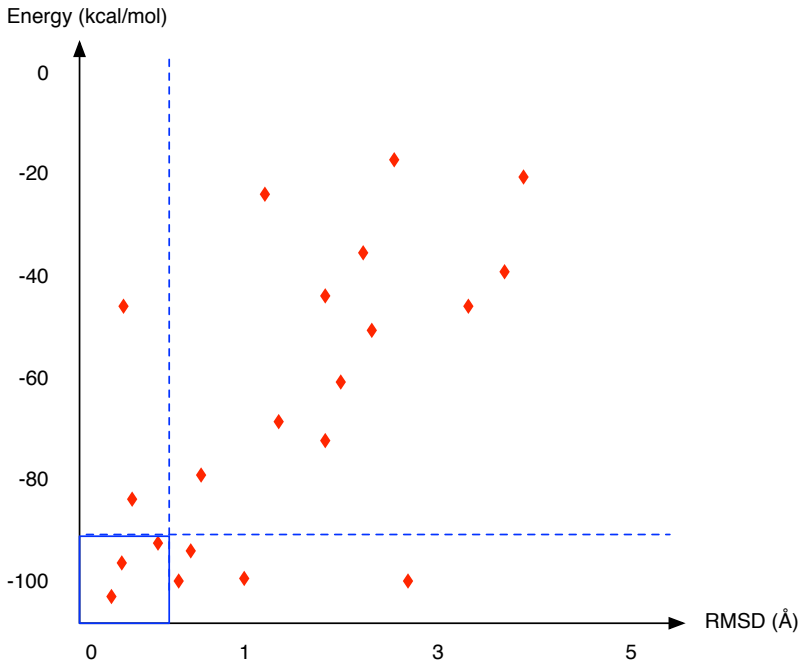


Figure 1.8 Graphical representation of the enrichment score. The enrichment score is simply the proportion of points in the blue square compared to the total number of points to the left and below of the dotted blue lines.

1.5 Statistical energy functions

1.5.1 Deriving a free energy statistical potential

One of the key components of any folding simulation is the potential function. We have seen that in molecular dynamics, a physical energy function is used to model the forces between the atoms. The problem with such potentials is that to extract the free energy of the system, a very thorough sampling of the potential energy landscape must be conducted, which is too computationally expensive. In order to address this problem, the free energy can be derived from statistical mechanics, by approximating the system to a perfect gas, and using an inverse Boltzmann potential [Sippl 1990, Sippl 1995, Hao & Scheraga 1999], which has the following functional form:

$$\Delta G = -kT \ln\left(\frac{f_{obs}}{f_{ref}}\right) \tag{1.11}$$

In (1.11), T represents the temperature in degrees Kelvin, k is the Boltzmann constant, f_{obs} is the frequency of the observed state, and f_{ref} is the frequency of the reference state. The frequency distribution of the observed state is derived from known protein structure, taken from databases such as the PDB (Figure 1.9).

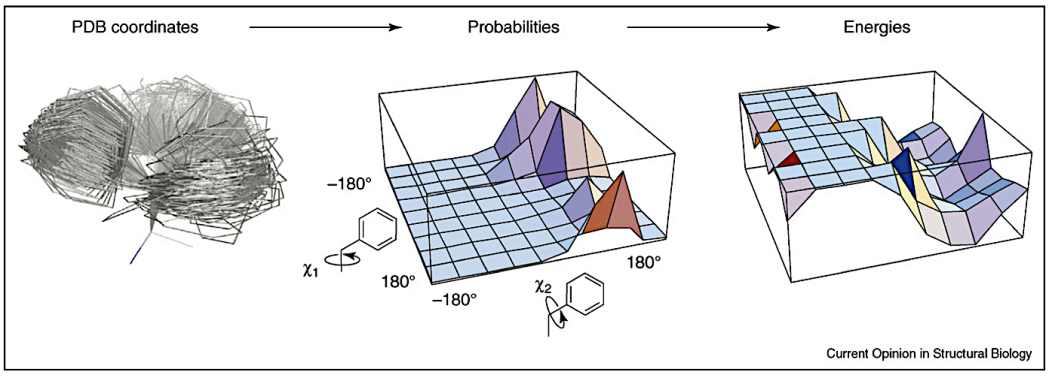


Figure 1.9 From PDB structure to energy function [Boas & Harbury 2007]. Steps involved in extracting a statistical free energy function from a set of PDB protein structures.

Given a large enough set of known protein structures, this knowledge-based approach has been shown to be sufficient to solve protein folding [Zhang & Skolnick 2005], but is not without problems, as many biases can be introduced, producing wrong distributions.

The major problem often comes from the limited number of structures available, which makes rare amino acid interactions harder to observe. Thus, some gap filling protocol must be defined, which could be as simple as adding a pseudo-count to the distribution, using average energy values to fill the gaps [Zhou & Zhou 2002], or even using an empirical approximation in specific cases [Yang & Zhou 2008].

Other biases can be introduced, notably when deriving distance potentials. Indeed, distance distributions above 30 Å will typically be dominated by large proteins, and thus, would not accurately represent small to medium sized ones. Therefore, most statistical potentials use a cut-off before that distance, sometimes in conjunction with a truncating function. Moreover, effects above 15Å are usually negligible, and thus, this is often used as a cutoff to reduce computation time.

The other major element, and active research topic, is the reference state. This can be defined as the probability of observing an interaction at random, or simply represents the distribution of all possible conformations. Since it is hard to generate every possible structure for proteins, a probabilistic approach is usually used instead, allowing us to derive an energy, which is in theory bias free. But the

distribution or functional form of the reference state is highly dependent on the type of interaction being looked at, and often will not even be derivable. Several approaches exist, such as using Bayesian probability models [Shen & Sali 2006], using geometrical models [Zhou & Zhou 2002], or using decoys. The basis behind the later is that the reference state is originally defined as the total conformational space, and thus, given a large enough sample of decoys, could be modelled statistically.

1.5.2 Atomic distance potentials from the literature

Since its initial formulation [Sippl 1990], many pairwise distance potentials were created and applied to the folding problem. In such potentials, a distribution of atomic distances is taken from the PDB, and used in conjunction with a specific reference state to reverse engineer the free energy of interacting atoms. The two most common distance potentials, DFIRE [Zhou & Zhou 2002] and DOPE [Shen & Sali 2006], are very similar in their formulation, differing only in their reference state definition.

DFIRE, which stands for Distance-scaled Finite Ideal-gas Reference state, uses a gas model to approximate the reference state, and normalises the distribution according to the last bin. In its initial formulation, it uses 167 atom types, corresponding to the heavy atoms in each residue, with 20 bins between 0 and 15 Å. In comparison, DOPE uses a Bayesian probabilistic reference state corresponding to non-interacting atoms in a homogenous sphere, with the radius dependent on a sample of native structures. Both potentials perform similarly at decoy discrimination and identification of the native structure [Rykunov & Fiser 2010].

But distance potentials are not the only ones being developed, and more recent potentials such as OPUS-PSP [Lu et al. 2008] or DFIRE2 [Yang & Zhou 2008] integrate atomic orientation of polar atoms. Some potentials, such as the one used in Rosetta [Kortemme et al. 2003], uses a linear combination of different physically and statistically derived terms to compute an overall energy for a given conformation. This approach has proven to be successful, as can be seen from the recent performances of various folding procedures in CASP [Kryshtafovych et al. 2009].

1.6 Hydrogen bonds

1.6.1 Hydrogen bond definition

One of the main non-bonded cohesive forces in proteins is the hydrogen bond, which was first mentioned in 1912 [Moore & Winmill 1912]. Subsequently, it has been widely studied, first in water [Latimer & Rodebush 1920], then in proteins [Pauling 1960] to account for the existence of secondary structures such as the alpha-helix and the beta-sheet.

Controversy exists regarding the nature of hydrogen bonds, more specifically as to its partly covalent nature [Isaacs et al. 1999]. Nonetheless, it is still the strongest non-bonded interaction in proteins, with energies between 5 and 30 kJ/mol.

Hydrogen bonds occur when a hydrogen atom attached to an electronegative one, such as nitrogen or oxygen, interacts with another electronegative atom. In some cases, carbon atoms can act as hydrogen bond donors, especially when they are covalently bonded to another more electronegative atom.

Thus, the hydrogen bond can be represented as a system of 3 atoms: the donor group, composed of a heavy electronegative atom covalently bonded to the hydrogen, and the acceptor, which is also an electronegative atom.

Sometimes, a donor can also be an acceptor, as is the case in certain amino acid side chains, such as histidine, threonine and tyrosine.

In potential energy models, two features commonly describe the hydrogen bond: the distance between the hydrogen and the acceptor, and the angle between the donor, hydrogen atom and acceptor.

This definition has been extended to include the atoms bonded to the acceptor, as evidence suggests a preference on the acceptor side angle as well [Morozov et al. 2004]. A common representation of hydrogen bonding geometry is shown in Figure 1.10.

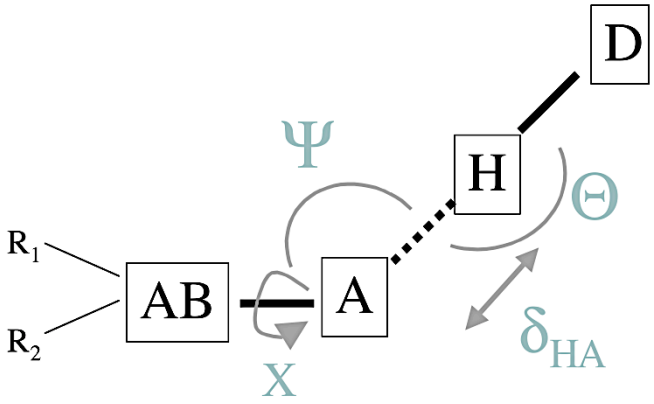


Figure 1.10 Hydrogen bond geometry [Kortemme et al. 2003]. *D* represents the donor atom, *H* the hydrogen, *A* the acceptor and *AB* the acceptor base.

Here, the features described are the H-A distance, the D-H-A angle, the H-A-AB angle, and the A-AB axis dihedral. Due to the partly covalent nature of the hydrogen bond, there is a preference towards linearity for the D-H-A angle and a preference for short van der Waals H-A distances [Pimentel & McClellan 1971].

1.6.2 Hydrogen bonds in proteins

Hydrogen bonds in proteins can be categorized as either strong or weak, the latter still being the subject of controversy. Strong hydrogen bonds are formed between oxygen and nitrogen donor and acceptor groups, both on main chains and side chains. The main chain NH group acts as a strong donor, while the carboxyl CO group acts as a strong acceptor. This peptide hydrogen bond is very important in proteins, as it is responsible for the secondary structure elements as well as stabilizing the core of the protein [Myers & Pace 1996]. In alpha helices, the helical structure is due to the hydrogen bond occurring between the backbone NH group and the carboxyl CO group 4 residues earlier. In beta sheets, each strand is bonded to another one through backbone hydrogen bonds, giving it its flat topology. A study of the fully buried atoms in 57 high-resolution proteins [McDonald & Thornton 1994] has shown that strong hydrogen bonds in proteins have a coupled preference for HA distances around 2 Å and DHA angles around 180º, as shown in Figure 1.11.

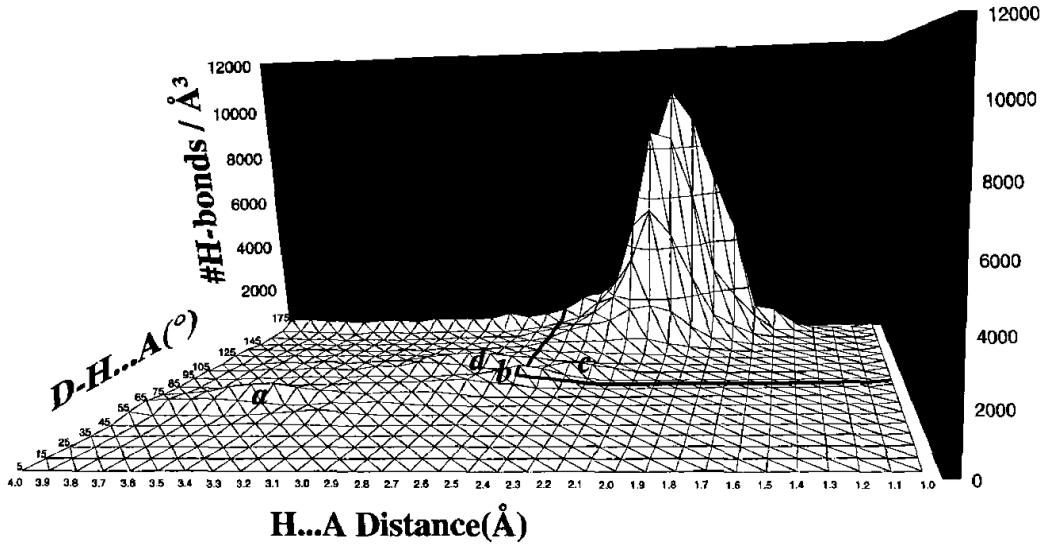


Figure 1.11 Distribution of hydrogen bond features in proteins [McDonald & Thornton 1994]. The (c) peak is the hydrogen bonding interaction, while the (a), (b) and (d) are due to longer-range van der Waals and electrostatic interactions.

Hydrogen bonds in simulations occur when the hydrogen atom enters the van der Waals radius of the acceptor atom, and as such, acts as an “on/off” interaction rather than one that decays with distance, as is the case for other electrostatic interactions.

Therefore, in most hydrogen bond models, a cutoff is used for the HA distance, above which no donor and acceptor systems are considered. This cutoff can vary depending on which study is considered, but mostly remains between 2.4 Å and 3 Å.

The partly covalent nature of the hydrogen bond also puts constraints on the DHA angle which has a preference towards linearity. Although statistical distributions of DHA angles show a peak around 160º, applying a geometrical correction correctly produces the expected 180º preference [Kroon & Kanters 1974]. This correction term is shown in Figure 1.12.

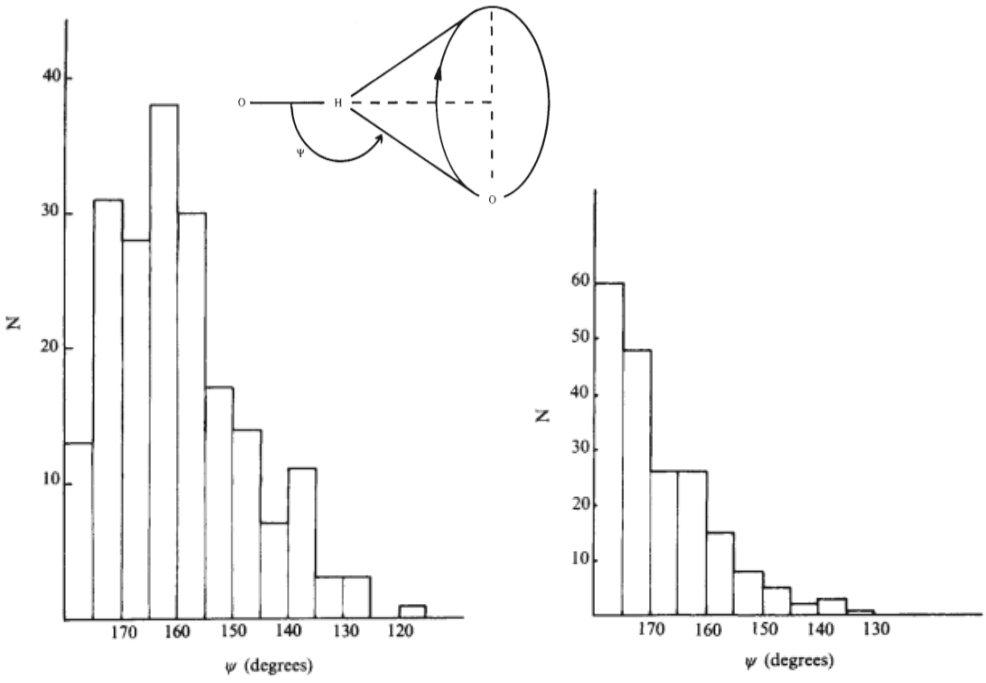


Figure 1.12 Cone correction of DHA angles in NH hydrogen bonds [Kroon & Kanters 1974]. The basis behind the cone correction is that as the opening angle increases, the volume of the cone decreases proportionally to sin DHA. Thus, the probability of finding a hydrogen bond at random is different for different angles, and the distributions must be adjusted accordingly. This is also true for the H-A-AB angle. Using this correction, the preference of NH bonds towards linearity is restored, as can be seen in the right hand side histogram, which is the corrected version of the left one.

Because of this preference towards linearity, most models do not consider hydrogen bonds with angles less than 90º, and some even restrain the system to angles above 120º [Fabiola et al. 2002].

A study of strong hydrogen bonds in proteins using the above formalism has been conducted in order to derive a statistical potential. The corrected distribution of backbone-backbone NH hydrogen bonds has been generated for each secondary structure type (Helix, Sheet, Coil), showing variability in distance, angle and dihedral preferences. This is shown in Figure 1.13.

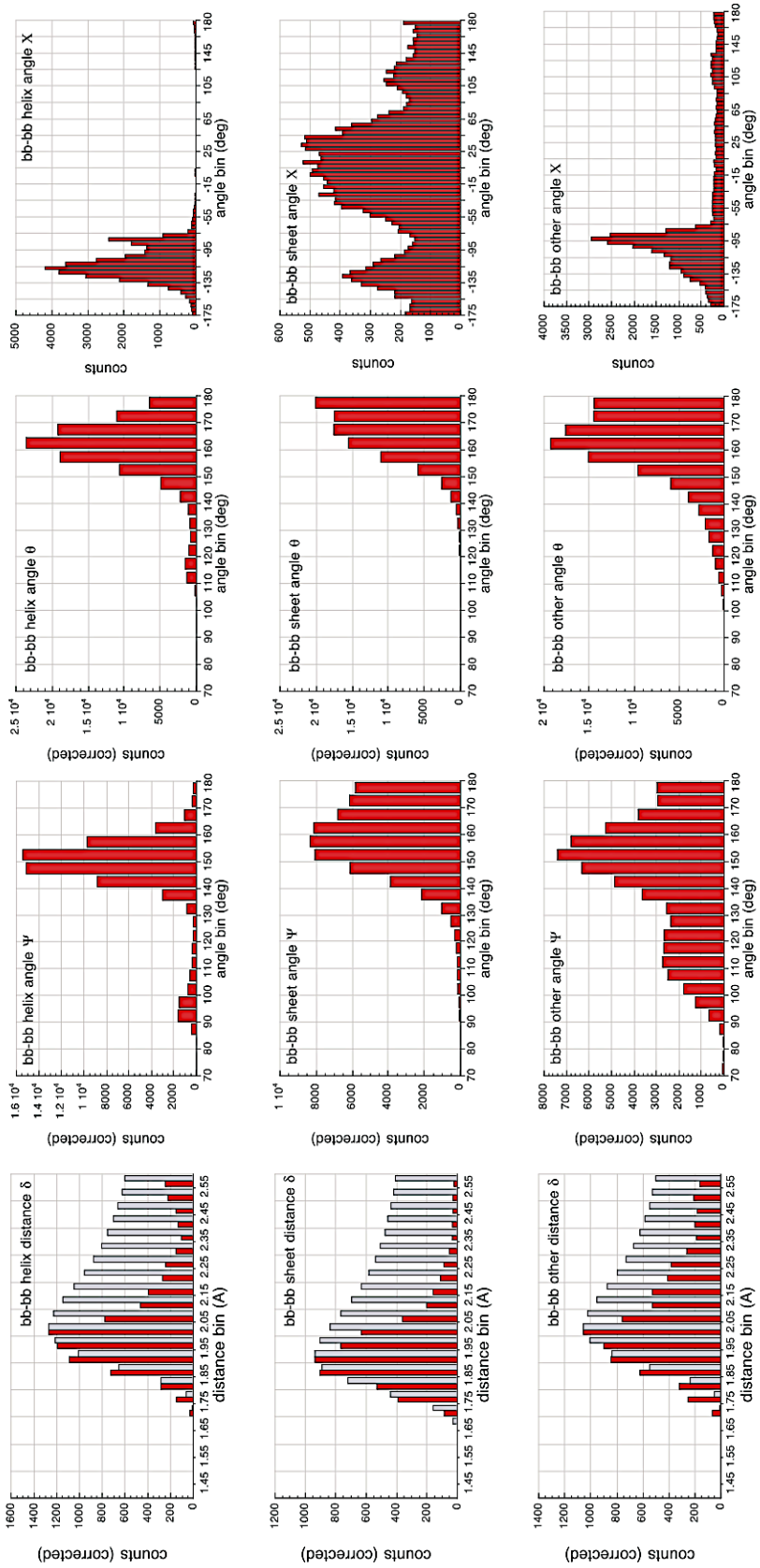


Figure 1.13 Distribution of backbone hydrogen bond features [Kortemme et al. 2003]

We can see in Figure 1.12 that the backbone-backbone (bb-bb) hydrogen bond has its distance peaking at 1.9 Å, and its θ angle peaking at 160º. These values are observed across all secondary structures. As for the Ψ angle, its preference varies from 160º for helices and coils, to 180º for beta sheets. Finally, the X dihedral peaks at -140º in alpha helices, -100º in coils, and -140º, 0º and 125º in beta sheets.

Strong hydrogen bonds are also formed by electronegative side chain groups, but are not involved in secondary structure formation. The analysis of strong hydrogen bond satisfaction shows that almost all buried hydrogen bonds are satisfied, which is not necessarily the case for surface donors and acceptors [McDonald & Thornton 1994].

When the solvent is included in the analysis, most hydrogen bonds become satisfied, with many being in a bifurcated configuration [Jeffrey 2003]. In some cases though, some acceptors may still not be satisfied. Including weak hydrogen bonds involving CH groups completely resolves this, with all donors becoming satisfied [Wahl & Sundaralingam 1997]. Although CH hydrogen bonds in crystals have been widely studied, their existence in proteins is still controversial. Theoretically, a carbon could donate a hydrogen as it is electronegative, but in practice, that force might be so weak that it would not be stable, and thus cannot be considered a major factor. But in cases where the CH group is covalently bonded to a strongly electronegative atom such as oxygen or nitrogen, this bond can become stronger and play an important role in the overall stability of the protein, as well as in protein-protein interfaces [Jiang & Lai 2002]. The more electronegative the atoms next to it are, the stronger the hydrogen bond becomes. The distribution of HA distance for CH hydrogen bonds is shown for chloroform (a), and acetone (b), in Figure 1.14, demonstrating the distance preferences of CH hydrogen bonds.

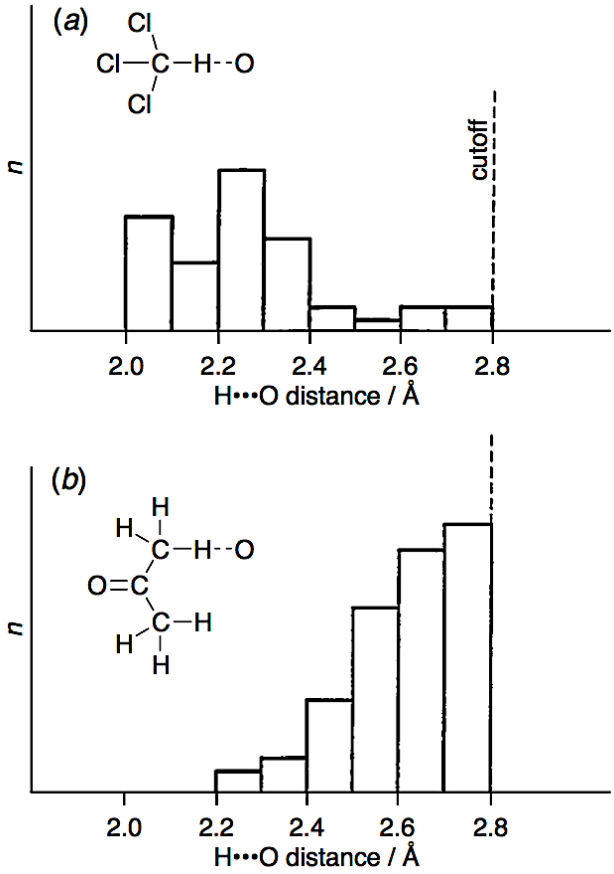


Figure 1.14 HA distance in inorganic CH hydrogen bonds [Steiner 1997]. We can see that in both cases, many CH hydrogen bonds have been observed below the 2.8 Å cutoff used. Moreover, in accordance with theory, the distances for the chloroform bond are shorter than for acetone, as chlorine is strongly electronegative.

In organic crystals, as in proteins, the distribution of the DHA angle for CH bonds follows the one for OH bonds, peaking at 180º once the cone correction has been applied. Additional studies have shown that there is a co-interaction between main chain CH donors and main chain NH donors from adjacent residues, with the linearity of the CH bond increasing as the linearity of the NH bond decreases [Fabiola et al. 1997] (Figure 1.15).

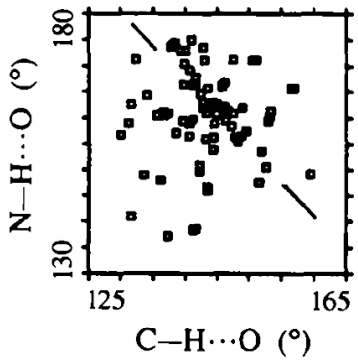


Figure 1.15 Coupling of NH-O and CH-O main chain hydrogen bonds [Fabiola et al. 1997]

Although this shows a correlation between the NH-O and CH-O angles, it does not necessarily imply a causation effect. Indeed, the planarity of the peptide bond means that adjacent CH and NH groups will point in a similar direction, thus making this correlation uniquely due to the position of the oxygen acceptor relative to the two residues it is bonding to.

1.6.3 Hydrogen bonding potentials

Using the formalism introduced previously, various potentials have been developed to assess the energy of a given hydrogen bond. Although initially empirically derived, the increase in the number of crystal structures available has allowed statistical potentials to be derived with high precision.

Approximate potentials can be derived without the hydrogen position from the other atoms in the system, but it is always better to include all possible information. Hydrogens usually do not appear in NMR structures or in conventional crystallography, but do in crystal structures derived from neutron diffraction. Although the number of non-similar structures showing such hydrogens is small, it still allows for an initial analysis to be conducted. Moreover, the position of some hydrogen atoms can be inferred with great accuracy from the chemistry of the group it is attached to, as is the case for the main chain NH group. Since all main chain and some side chain CH and NH hydrogen positions can be inferred, a potential can then be used to assess the quality of a structure by looking at its hydrogen bond network. One empirical potential [Fabiola et al. 2002] is shown in Figure 1.16.

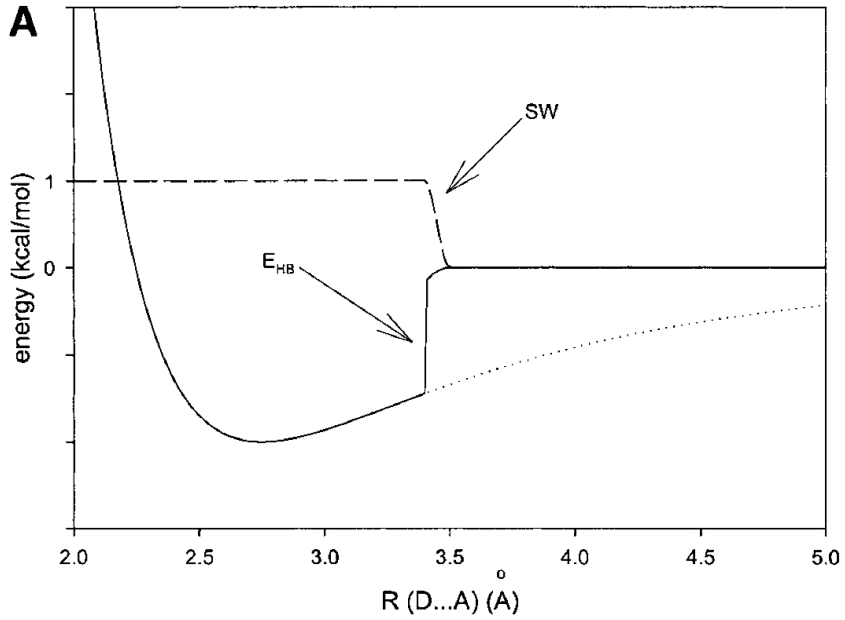


Figure 1.16 Switching (SW) truncation function [Fabiola et al. 2002]. Here, a 6,4 Lennard-Jones potential is used for the DA distance and a switching function to allow smooth truncation above 3.4 Å.

The functional form of this potential is shown in Equation 1.12 below, where SW is the switching function [Brunger 1992].

$$E_{HB} = \varepsilon\left[\left(\frac{\sigma}{r}\right)^{6} - \left(\frac{\sigma}{r}\right)^{4}\right] \times \cos^{4}\left(\theta - \theta_{0}\right) \times SW\left(r\right) \tag{1.12}$$

Here, *r* is the DA distance, σ is the point at which the potential is zero, θ is the D-A-AA angle, θ_0 is the equilibrium D-A-AA angle, and SW is the switching function. Although this potential assumes the hydrogen position is unknown, it can easily be modified to incorporate the other bonding features, as is the case in another potential using a 12-10 Lennard-Jones potential [Gordon et al. 1999]. Such potentials were initially included in force fields for molecular dynamics simulations [Brooks et al. 1983], but are now replaced by more fine tuned electrostatic and van der Waals potential terms [Cornell et al. 1995].

The other way to determine the energy of a hydrogen bond is to use knowledge based potentials, usually in the form of an inverse Boltzmann energy function. In such statistical potentials, the distributions are extracted from a set of non-redundant structures from the PDB. These can either be neutron-diffracted structures, showing the hydrogen position explicitly, or by only considering hydrogens that are geometrically fixed. The four geometrical features commonly used are then represented as separate energy terms, and weighted using an appropriate protocol.

One of the main difficulties with statistical potentials is the necessity to derive a reference state. For the HA distance, a sphere packing model can be used, as shown in Equation 1.13.

$$p_b^{ref}(r) = \frac{r_b^3 - r_{b-1}^3}{r_{\max}^3} \tag{1.13}$$

Here, r is the HA distance, b is the bin being considered, r_b and r_{b-1} are the maximum distances in bins b and $b-1$, and r_{max} is the cutoff distance for the potential. This probability represents the volume occupied by a specific bin relative to the maximum volume considered by the potential. As for the angle potentials, the reference state can be considered as being the cone correction. Some potentials don't explicitly use these normalisation as reference states, but rather use them to correct the observed counts [Kortemme et al. 2003]. Although the hydrogen bonding potential alone is not always sufficient to identify the native structure, it dramatically improves the chances of identifying it when combined with a van der Waals energy term (Figure 1.17).

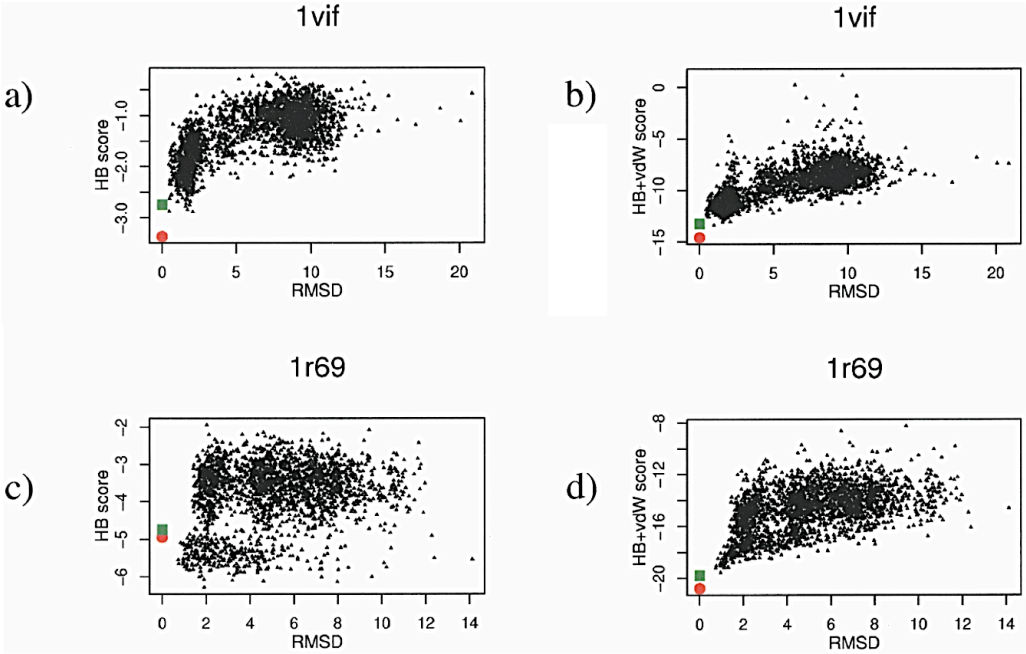


Figure 1.17 Hydrogen bonding potential performance [Kortemme et al. 2003]. Here, a) and c) show the RMSD in Å of the decoys against the hydrogen bonding term alone, whereas b) and d) show the same data with the added van der Waals term. The native structure is circled in red, while the green square shows the native structure after the side chain have been repacked.

Overall, many attempts have been made to model hydrogen bonds, but no approach seem to be predominantly better. Moreover, the existence and usefulness of CH hydrogen bonds in proteins is still controversial, despite being accepted in some smaller crystals.

1.7 Solvation

Modelling solvation has been the subject of numerous studies, not only because of its importance to protein folding, but also because of the remarkable properties of water as a fluid. Traditionally, there are two ways to model water: 1) by looking at each solvent molecule explicitly, and 2) by treating the solvent as a continuum. The latter has the advantage that it is much faster to compute, but the former is more representative of reality.

1.7.1 The hydrophobic effect

Ever since the formulation of the thermodynamic hypothesis of protein folding [Anfinsen 1973], the solvent has been considered the main driving factor underlying protein folding, with other non-bonded interactions only participating in stabilising the protein [Myers & Pace 1996].

The initial question from which this hypothesis emanated was that proteins fold spontaneously into a single conformation, and as such, must have a maximal negative free energy change between its unfolded state and its folded state. During folding, the polar residues in the protein interact with the water through hydrogen bonds, whereas non-polar residues do not. Since water-water interactions are more favourable, the solvent will form a cage around non-polar residues. When two non-polar residues form a contact, the water in between is pushed out towards the bulk solvent, effectively increasing the entropy of the water. This is called the hydrophobic effect, and is believed to be the main force behind spontaneous protein folding.

1.7.2 Water hydrogen bonds

Water surrounding the protein will interact with the polar surface atoms, and form hydrogen bonds with them [Radzicka et al. 1988]. In order to satisfy all donors and acceptors in water molecules, there must be 4 hydrogen bonds forming: two as an acceptor, and two as a donor [Rahman & Stillinger 1973], as shown in figure 1.18.

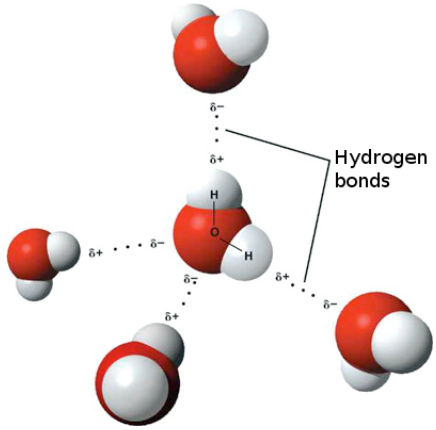


Figure 1.18 Hydrogen bonds in water [Wikimedia 2010]

Because of the properties of water, having polar residues on the surface will allow the solvent hydrogen bonds to be satisfied through bonding with the polar groups on the side chains. It has been observed that when surface water molecules are not satisfied by surface hydrogen bonds, they form weak hydrogen bonds to CH groups [Steiner & Saenger 1993, Steiner 1995]. Thus, weak hydrogen bonds can be seen as complementary to conventional strong hydrogen bonds when interacting with water.

1.7.3 Explicit models

Molecular dynamics simulations, amongst others, often use an explicit representation of water. Although many models exist, the most widely used are the TIP3P [Jorgensen 1981], TIP4P [Jorgensen et al. 1983] and SPC [Berendsen et al. 1981] models.

The first difference between the TIP models and the SPC model is the preferred geometry of the water molecules. In TIP models, it is matched to known properties of water (105º angles), whereas in the SPC model, it assumes an ideal tetrahedral shape (109º angles).

The other major difference is the force field used to model the interaction between water molecules. The TIP models use a 12,6 Lennard-Jones potential for van der Waals interactions and a Coulomb potential for electrostatic interactions. The SPC model uses the same formalism, but adds a corrective term to account for polarization.

Both TIP3P and SPC models use 3 point charges, one on each atom, whereas the TIP4P assumes a fourth point near the oxygen, bisecting the HOH angle. Molecular dynamics mostly uses 3-point charge models because of the decreased computational complexity and simplicity of the model.

Other more complex models exist [Rahman & Stillinger 1973, Bernal & Fowler 1933], with 5 or 6-point charges [Mahoney & Jorgensen 2000, Nada & van der Eerden 2003], but are seldom used in simulation because of the increased computational complexity.

1.7.4 Implicit models

Despite many existing explicit water models, none are computationally practical for large scale full-atom simulations, and thus, simpler, faster models are needed for very long simulations or for *ab initio* folding. These models generally assume a continuous solvent around the protein, and allow for a quick estimation of the solvation free energy. The first implicit solvation potentials used the solvent accessible surface area of heavy atoms to derive their energy contribution [Eisenberg & McLachlan 1986], using the rolling ball algorithm to compute such surfaces [Lee & Richards 1971]. This potential can be written as follows:

$$\Delta G = \sum_{atoms} A_i \Delta\sigma_i \tag{1.14}$$

In this equation, ΔG is the free energy of solvation, A_i is the surface accessibility of atom i and $\Delta\sigma_i$ is the solvation parameter, as defined in the original paper. The problem with this approach is that it only considers five types of atoms (carbon, oxygen/nitrogen, polar oxygen, polar nitrogen, sulphur), and completely excludes hydrogens.

Another approach, part of the EEF1 potential [Lazaridis & Karplus 1999a], uses a Gaussian solvent-exclusion model to calculate the free energy of solvation of a protein. It is written as the sum of each group contribution to solvation, as shown in Equation 1.15 below.

$$\Delta G = \sum_{i}^{atoms} \left[\Delta G_i^{ref} - \frac{2}{3} \sum_{j \neq i}^{atoms} \frac{R_j^3 \Delta G_i^{free}}{r_{ij}^2 \sqrt{\pi \lambda_i}} e^{-\left(\frac{r_{ij} - R_i}{\lambda_i}\right)^2} \right] \tag{1.15}$$

Here, ΔG is the free energy of solvation, ΔG_i^{ref} is the reference solvation free energy taken from a fully solvated small molecule, r_{ij} is the distance between atoms *i* and *j, and* R_i and R_j are the van der Waals radii of atoms *i* and *j*, ΔG_i^{free} is the solvation free energy of isolated group *i,* and λ_i is a correlation length specific to group *i*. Overall, this formulation of the solvation free energy has been successful [Lazaridis & Karplus 1999b] at discriminating decoy structures, and has subsequently been used often in other similar studies.

Chapter 2

Generating a near-native decoy set

2.1 Introduction and background

Protein fold prediction is a very important open problem in biology [Kelly & Sternberg 2009], mostly because of the impact it would have on medical fields and because of the difficulty and expensiveness of current experimental methods. Although theoretical protein folding has been an active field of research for many years, current methods seldom generate a native structure [Zhang 2008, Dill et al. 2008].

Non-native models, commonly called decoys, can be used as a benchmark to test energy functions [Rykunov & Fiser 2010, Gilis 2004]. Selecting the best candidate models from decoys is not trivial, and many methods have attempted to correlate the free energy to the nativeness of the decoys [Aloy & Oliva 2009, Sippl 1995, Ma 2009]. In order to facilitate and standardise the benchmarking of these methods, various sets of candidate structures, or decoy sets, have been created, using various folding packages. Here, we review two of the most relevant sets, summarised in Table 2.1, below.

Table 2.1 Properties of common decoys sets

Decoy Set	Targets	# decoys	All-Atom RMSD range
$4state_{reduced}$	7	665	1.7 – 10 Å
LMDS	11	439	2 – 14 Å

We define a near native decoy as one having an all-atom RMSD to the crystal structure less than 2 Å. The problem with the two sets introduced here is that they do not have many structures below that threshold. Figure 2.1 shows the all atom RMSD distribution for each target in the $4state_{reduced}$ decoy set [Park & Levitt 1996].

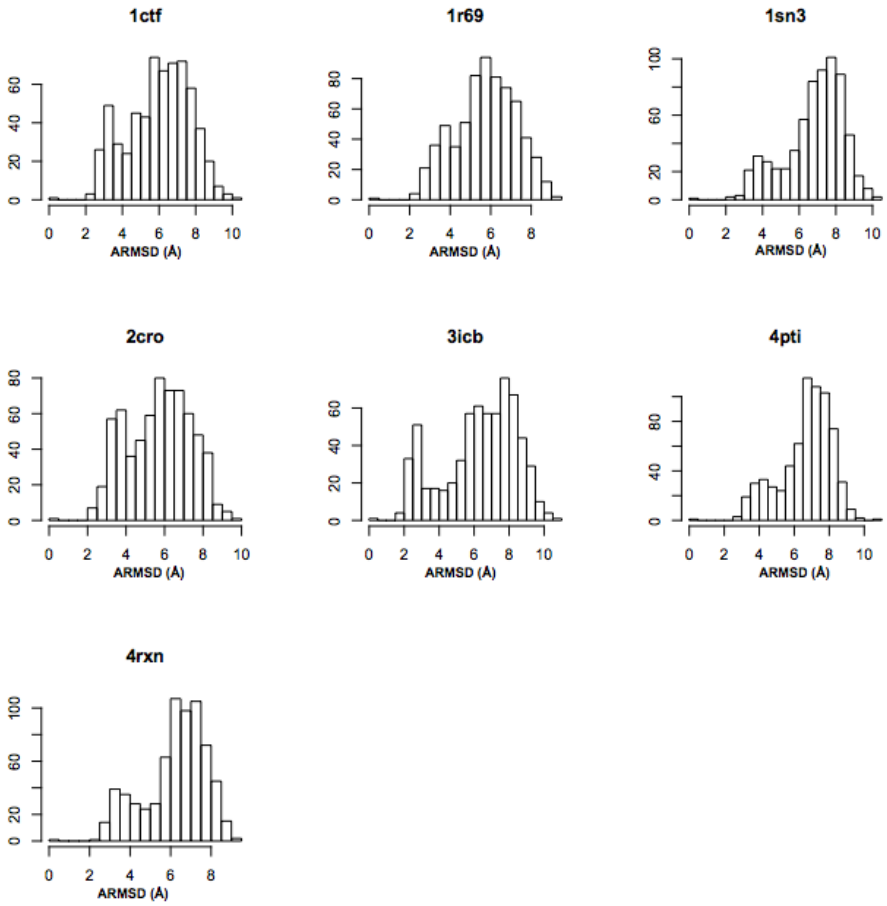


Figure 2.1 All Atom RMSD of the decoys in the 4state$_{reduced}$ set. The average RMSD for this set is around 7 Å, with almost no decoys below 2 Å, aside from the crystal structure. The 4state$_{reduced}$ decoy set was generated using an off-lattice model and a relaxation method. The decoys are compact, have native-like secondary structures and are self-avoiding.

Since we are testing the discriminatory power of existing potentials, we want to avoid biases towards a certain defect that would be generated by a specific decoy generation method, so we have selected another decoy set, the LMDS set [Kaesar & Levitt 1999], which was generated by random sampling of conformational space followed by an energy minimization procedure. By looking at the RMSD distribution for that set, we came to the same conclusion as for the 4state$_{reduced}$, i.e. that there are not enough decoy structures with an RMSD below 2 Å.

These sets are commonly used throughout the literature, and are often taken as benchmarks for new energy functions [Gilis 2004]. What we are interested in, however, is the effectiveness of the current energy functions when applied to near native structures, which are by definition hard to discriminate from the crystal structure, but also from other decoys.

The problem of selecting crystal structures from a set of decoys is trivial in some cases, mostly since they have been generated using a different method, and as such, do not resemble the decoys as much as the decoys resemble each other [Handl et al. 2009]. One of the reasons behind this is the packed surface side chains that are found in crystal structures, but which are sticking out in simulations, meaning the crystal structure will always be more compact than the decoys.

All these considerations have led us to develop our own decoy set, with the intention of keeping the all-atom RMSD at an average of 2.0 Å for each target. To do so, we ran molecular dynamics on a set of target crystal structures, and sampled the trajectories at different temperatures. We chose to use OPLS/aa [Jorgensen & Tirado-Rives 1988] as it is freely accessible as part of GROMACS [Berendsen et al. 1995, Lindahl et al. 2001, Hess et al. 2008].

In this chapter, we will show how molecular dynamics can help us generate a near-native decoy set, which will allow us to test existing potentials on an atomic basis. Our goal is to produce a decoy set in the range of 0-5 Å All-Atom RMSD (ARMSD), which are all well-formed and structurally close to the native structure. We will then analyse the efficiency of the DFIRE2 potential [Yang & Zhou 2008] taken from the literature, by applying it to the newly generated decoy set.

2.2 Methods

In this section, we present the methods used to generate our decoy set and test the potentials. We first define a target selection protocol, which incorporates constraints such as spatial arrangement, ligands, resolution and sequence identity. We then introduce our decoy generation method, which is based on molecular dynamics runs and pairwise RMSD clustering. Finally, we introduce the correlation and fitness measures used throughout this thesis when assessing the discriminatory power of an energy function.

2.2.1 Choice of target proteins

In order to generate a good decoy set, we need to represent the existing folding landscape. Indeed, choosing similar structures will bias the results towards one type of protein, and won't help us when assessing an energy function. Starting from the complete PDB database [Bernstein et al. 1977], we ended up selecting 250 targets, representing the most common folds. The selection process involves a number of steps, which were conducted in the order shown in Table 2.2.

Table 2.2 Filters used to select decoy set targets

Selection Filters	# Structures
PDB dataset	62430
Fewer than 200 residues	48216
Monomeric structures (PISA)	7032
No Ligands	2273
Less than 2.0 Å resolution	1408
No uncommon residue types	897
Less than 25% sequence identity	369
Single domain	250

Selecting structures from the PDB database isn't trivial, and although it can be filtered directly through the website, some other criteria are not accessible and must be implemented separately. Each filter is described below.

PDB dataset

We started by taking the complete set of structure from the Protein Data Bank [Bernstein et al. 1977], as of October 2009.

Sequence length

The trouble with molecular dynamics simulation is the long time it takes to run, which is related to the size of the protein. Since most methods do not try to fold large proteins, and since these would prove more difficult to model, we constrained our data set to relatively small proteins, having fewer than 200 residues per chain.

Monomeric proteins

Multimeric proteins sometimes become unstable when taken out of their original complex. As we do not model protein-protein interactions, nor take into account the lack thereof, we chose to remove all multimeric proteins from our dataset. To do this, we used an online tool called PISA [Krissinel & Henrick 2007], which uses crystallographic information to predict the quaternary structure of the protein.

No ligands

Omitting ligands that should be there will result in unstable conformations during molecular dynamics simulations. Moreover, by not explicitly including ligands in potentials when they are present, the predictive ability of energy functions cannot be properly assessed.

High-resolution structures

We chose only X-ray crystal structures with a resolution better than 2 Å. NMR structures are not present in our dataset because they do not contain the information needed by PISA to derive the quaternary structure of the protein.

No uncommon residue types

The potentials used in this thesis, as well as the ones we will derive, assume that residues being considered are the naturally occurring ones, and thus, all structure containing rare, unknown or mutant amino acids are removed.

Sequence identity

A cutoff of 25% sequence identity was chosen to remove homologs. The sequence alignment and clustering tool BLASTCLUST [Altschul et al. 1997] was used, with a 25% sequence identity threshold.

Single domain

Multi-domain proteins might behave differently when submitted to molecular dynamics. Since we are only interested in small, single-domain proteins, we chose to remove all multi-domain proteins from our dataset. To do so, we used the CATH [Orengo et al. 1997] database, in conjunction with SCOP [Murzin et al. 1995] when no entries were found for a PDB structure.

2.2.2 Molecular dynamics runs

The choice of parameters in molecular dynamics simulations impacts the way the protein will move. In our case, we need to relax the native crystal structure in order to generate enough near-native decoys. Although many molecular dynamics packages would be fit for this experiment, one is commonly used and is freely available, GROMACS. We opted for the default water setting for OPLS/aa, the SPC solvent model.

Each run is conducted in three stages. First, the crystal structure is energy minimised, using 500 steps of steepest descent minimisation. Then the solvent is equilibrated by restraining the position of the protein while running 20 ps of molecular dynamics. Finally, full, unrestrained dynamics is run for 200 ps, and sampled each ps to produce a trajectory file. This final step is repeated at various temperatures, ranging from 250 K up to 400 K, in increments of 25 K.

Although we want the decoys to be different from the native structure, we don't want them to be unfolded, so we discard decoys which are more than 4 Å RMSD away from the native structure. Each decoy is then energy minimised to remove possible steric clashes.

For each of the 250 targets, we generate 1200 decoys, which are then clustered at 0.5 Å RMSD to remove redundant structures. Since we do not want to bias our decoy set towards one target, we randomly sample 500 decoy structures per target (including the native structure), which brings our total number of decoys to 125000.

2.2.3 Measuring the fitness of an energy function

There are various ways of assessing the efficacy of an energy function. One of them being the use of a decoy set to try to correlate the nativeness of each decoy to the energy calculated for it.

Commonly, three criteria are used in determining the quality of an energy function, with respect to a decoy set. First, as a consequence of the thermodynamic hypothesis, it must have the native structure at the minimum of its energy landscape [Anfinsen 1973]. Secondly, decoys should have different energies to be able to differentiate them. Finally, there must be a correlation between the nativeness of the decoy and the energy function.

Depending on the problem we are interested in, some of these criteria might be less important [Gilis 2004]. Specifically, criterion 1 has been shown to be trivial in most cases since the crystal structure, usually considered to be the native conformation, is trivially found from the decoys generated [Handl et al. 2009]. Simply using a compactness term would prove sufficient, as the side chains in the crystal structure tend to be packed towards the centre of the protein, whereas they are sticking out in the solvent in simulations, hence increasing the size of the protein. As for criteria 3, one can either use a standard correlation coefficient such as the Pearson correlation, or use a more robust, less prone to outliers, rank based statistic such as the Kendall tau.

Since we are only interested in selecting the best models, rather than generating them, we will be looking at another statistic, called the normalised enrichment [Tsai et al. 2003], which given a set of decoys, will tell us what proportion of the best N models we would successfully select. This will give us an indication of the power of an energy function at discriminating near-native decoys.

We decided to use the all-atom Root Mean Square Deviation (ARMSD) of the decoy to the crystal structure as the measure of nativeness. Other measures exist such as the GDT or TM score, but are not fit for all-atom comparison, as they have been parameterised on CA atoms only. Equation 2.4 shows the normalised enrichment functional form. This has been modified from Equation 1.1 to have a range between [0, 1].

$$e = \frac{\Omega\left(X_c \cap Y_c\right)}{cN} \tag{2.4}$$

In the above equation, e represents the enrichment score, X_c are the best structures in terms of nativeness, and Y_c are the best structures in terms of energy. The cardinality of the intersection of these two sets yields the number of structures that are both in the best nativeness and best energy sets. Dividing this by the total number of decoys N times the ratio c, gives the enrichment score, which can be interpreted as the probability of finding a structure that is in the given best subset of decoys. Since choosing structures randomly will yield a score of c, anything above that will be better than random, provided the number of decoys is sufficient to allow this statistic to be significant.

2.3 Results & Discussions

We have generated a decoy set comprising of 250 targets, selected from the PDB database. Each target has 500 decoys, which is sufficient to give a statistically significant score in our analysis. We will now describe the properties of our decoy set, and conduct an analysis of the different sequential, structural and experimental features that might bias our decoy set.

2.3.1 Properties of the decoy set generated

Looking at the distributions of ARMSD, we find that targets are normally distributed around 1.7 Å, with a standard deviation of 0.4 Å, and have a sequence length of 125 residues on average. Table 2.3 summarises the properties of the decoy set.

Table 2.3 Properties of the decoy set.

Property	**Value**
Targets	250
Decoys per target	500
Min RMSD between decoys	0.5 Å
Protein in class (Alpha / Beta / AlphaBeta / None)	72 / 78 / 91 / 9
X-ray resolution	< 2.0 Å
Feature	**Average**
ARMSD	1.7 (0.4 std) Å
Residues	125
Atoms	1961
Hydrogen bonds	152

Here, the minimum RMSD between decoys is taken from all heavy atoms. The “None” secondary structure class represents proteins where there are insufficient residues in either classes to categorise them.

This table shows that the proportion of proteins in each secondary structure category is consistent with those in the PDB database, with the exception of

those with few secondary structures, which have been added in larger proportion to be significantly represented in our decoy sets.

Hydrogen bonds are considered only when their planar angles are larger than 90 degrees, and the hydrogen to acceptor distance is smaller than 2.5 Å.

The PDB codes of the proteins included in this decoy set are 133l, 153l, 1aa2, 1aaj, 1acf, 1agi, 1b8e, 1bd8, 1bea, 1bfg, 1bj7, 1bk7, 1bv1, 1bz4, 1c44, 1cei, 1chd, 1dsl, 1e6h, 1e6k, 1ekg, 1enj, 1eq6, 1ew4, 1ey0, 1eyh, 1ezk, 1f32, 1faa, 1fan, 1fas, 1faz, 1fl0, 1gak, 1goa, 1gxq, 1h75, 1hka, 1hoe, 1huf, 1i2t, 1iap, 1ifg, 1igd, 1is5, 1j3a, 1jb3, 1jmw, 1jos, 1jpe, 1jyh, 1kn3, 1koe, 1kqx, 1kxo, 1l01, 1l2p, 1lit, 1lki, 1ln4, 1lou, 1lpl, 1m5i, 1mbm, 1md6, 1mh7, 1mhn, 1mjc, 1mjs, 1mwp, 1mzl, 1na5, 1nko, 1noa, 1ow1, 1oz9, 1p4p, 1pgv, 1pvx, 1pzc, 1q2y, 1q5z, 1qzm, 1qzn, 1r69, 1r8n, 1r9h, 1rj1, 1roa, 1srv, 1t7a, 1tg0, 1txj, 1tzv, 1u9a, 1u9p, 1ua8, 1uj8, 1ulr, 1unp, 1vcc, 1vfq, 1who, 1wm2, 1wou, 1wvh, 1x3o, 1yp5, 1yqb, 1yu5, 1yw5, 1ywp, 1zeq, 1zzk, 2a4v, 2b1k, 2bk8, 2cgq, 2ckx, 2cov, 2cwr, 2cxc, 2czt, 2d59, 2dyi, 2e7v, 2ehg, 2ejx, 2erw, 2evb, 2fd4, 2fi9, 2fjz, 2fk9, 2fl7, 2fph, 2fq3, 2frg, 2fzp, 2gbn, 2gee, 2hdz, 2hlq, 2hp7, 2hwx, 2i1b, 2i6v, 2icc, 2in0, 2iug, 2j71, 2jcp, 2lis, 2nx2, 2o2w, 2o37, 2obi, 2oix, 2op6, 2osa, 2ova, 2ovo, 2p52, 2p5d, 2pcy, 2pko, 2pth, 2ptv, 2q5x, 2qhe, 2qr3, 2qt4, 2r2y, 2rb8, 2rer, 2rh3, 2rk5, 2rkq, 2uwr, 2vc8, 2vga, 2vh7, 2w0g, 2wj5, 2ygs, 2ywd, 2ywk, 2yxm, 2z9t, 2zeq, 2zqe, 2zrr, 3bci, 3bn6, 3bzr, 3cm0, 3co1, 3csp, 3csr, 3ctg, 3d79, 3dfg, 3dj9, 3dvw, 3e21, 3e7u, 3ebk, 3eoi, 3etp, 3eye, 3fh2, 3frr, 3g9b, 3id1, 3id4.

2.3.2 Distribution of amino acids

One of the ideas behind using a large number of targets was to properly represent less frequent residues. In order to verify this, we have computed the frequency of occurrence of each amino acid in the targets we selected, and compared it to the PDB. Table 2.4 summarises the differences. As can be seen, the absolute difference in frequencies is consistently small, and averages 0.6%, which shows that our decoy set is representative of the PDB, in terms of the occurrence of residues.

Table 2.4 Amino acid frequencies in the decoy set and PDB

Residue	Decoy set (%)	PDB (%)
ALA	8	7
ARG	5	4
ASN	4	4
ASP	6	6
CYS	2	3
GLN	4	5
GLU	7	6
GLY	7	7
HIS	2	3
ILE	5	4
LEU	9	8
LYS	7	7
MET	2	2
PHE	4	4
PRO	4	5
SER	6	8
THR	6	6
TRP	1	1
TYR	3	3
VAL	8	7

2.3.3 Impact of sequence features on average ARMSD distribution

Biases towards frequent amino acids is just one of the possible problems that could emerge when generating a decoy set. For example, the average of the average ARMSD of decoys is at 1.7 Å, while the actual average ARMSD per target varies from 1.1 Å, up to 3 Å. This difference can either be due to some proteins being more stable than others, or to a difference in size in which case the ARMSD might be bigger for proteins with more atoms, as they would have more degrees of freedom. If indeed this is the case, then we should observe a correlation between the average ARMSD and the size of the protein. This is shown in Figure 2.2.

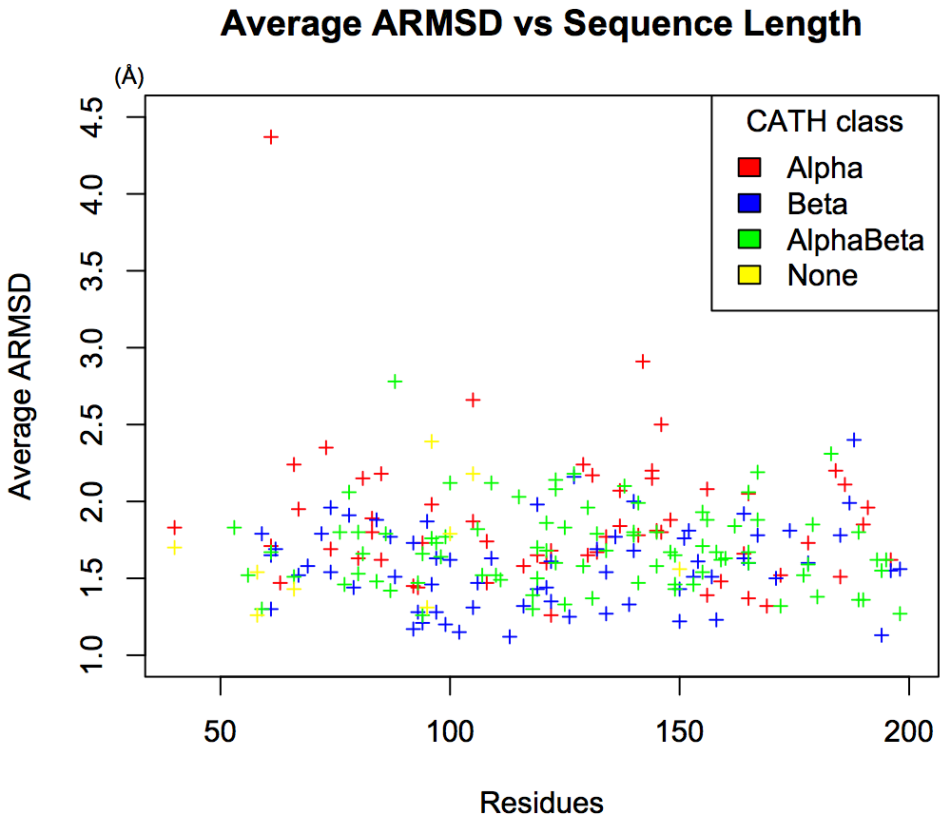


Figure 2.2 Average All-Atom RMSD versus Sequence Length. The ARMSD is calculated based on the 500 decoys in each target. Each point represents a target. Targets have been separated according to their CATH class to show any bias towards a particular type.

We can observe that the ARMSD is uniformly distributed across all sequence lengths, showing that there are near native decoys of all lengths in our decoy set. The only exceptional protein here is 1L2P, which is an outlier in terms of its average ARMSD, and is the only transmembrane protein in our set. Although it is not included in our analysis, we show it here to demonstrate the robustness of our filters. Although the average ARMSD does not depend on the number of residues, there is a difference when considering the average per CATH class. Indeed, the average ARMSD for alpha proteins is larger than for beta ones, with the other two classes in-between. Distributions of mean ARMSD per CATH class are shown in Figure 2.3 below.

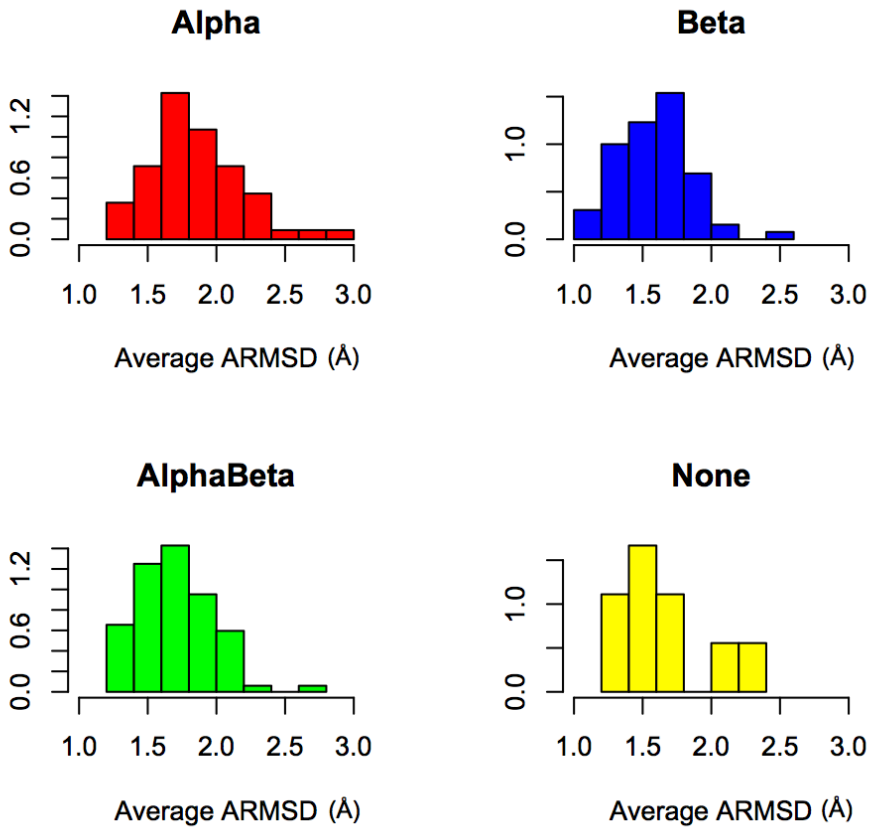


Figure 2.3 Average RMSD distributions per CATH class. Here, the 1L2P outlier has been removed. The means of the distributions above are 1.88 for the alpha class, 1.57 for the beta class, 1.70 for the alphaBeta class, and 1.68 for the none class.

To assess the significance of these differences, we used a 2-tailed t-test on the difference between the mean of the distributions. The p-value matrix is shown in Table 2.5.

Table 2.5 P-value of the ARMSD difference across CATH classes

CATH class	Alpha	Beta	AlphaBeta
Beta	0.0001		
AlphaBeta	0.005	0.008	
None	0.22	0.32	0.83

From this table we can conclude that the difference between Alpha, Beta, and AlphaBeta classes are highly significant, while the difference to the None class is not, most likely because of the small number of observations (there are only 9 proteins in the none class). The difference between alpha and beta classes can be explained by the nature of the secondary structure of the proteins, where alpha-helices stabilise proteins locally and beta-sheets globally. Since the ARMSD is more sensitive to global differences than to local ones, a less globally stable protein will be expected to have a larger average ARMSD. When most hydrogen bonds are located in alpha helices, they only stabilise nearby residues, meaning there is a large freedom of movement for residues not in helices, analogous to the movement of a hinge. In beta sheets, residues between bonding groups are constrained geometrically, so the overall movement of the protein is limited. This locality of hydrogen bonding can be calculated by taking the average residue separation between donor and acceptor groups, and comparing these for Alpha and Beta proteins. A higher residue separation means more residues are located between the residues bonding, and thus, are more constrained. Results are shown in Table 2.6.

Table 2.6 Average residue separation in hydrogen bonds

Class	Mean	Std	Decoys
Alpha	11	5.5	28500
Beta	25	8.1	32500
AlphaBeta	19	6.4	42000
None	17	6.2	4500

Using a 2-tailed t-test, we obtain a p-value less than 0.0001 for the difference between the mean residue separation of Alpha and Beta proteins. This highly significant difference shows that there is indeed more global stability in Beta proteins. Since the ARMSD measure is sensitive to global changes, we can conclude that the difference observed between Alpha and Beta classes are indeed due to the hydrogen-bond network topology in secondary structures. Thus, the difference in the average ARMSD between secondary structure classes can be explained both from the topology of the hydrogen bonding network, and from the geometrical properties of individual classes. Since these differences are expected in real proteins and well modelled by the force field, we can conclude that it is not penalising.

2.3.4 Impact of simulation parameters on average ARMSD distributions

One of the main difficulties with molecular dynamics is to derive a force field that correctly assesses the energy of new structures generated during simulations. Indeed, current force fields are incapable of keeping the native structure as is, and as a consequence will drift away from it, producing decoys (Figure 2.4).

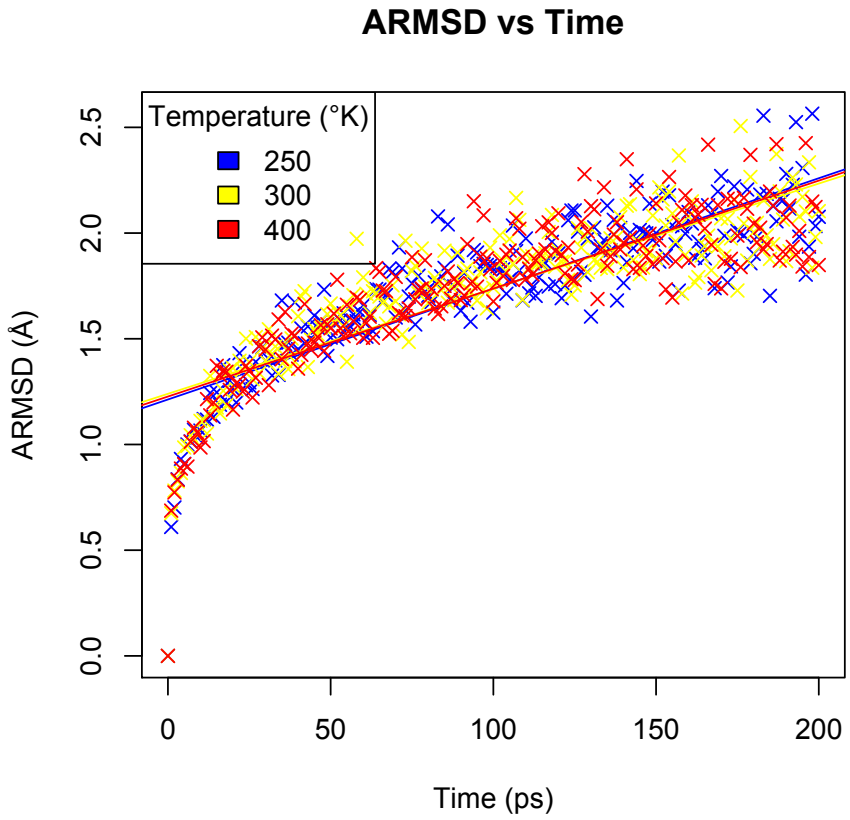


Figure 2.4 All-Atom RMSD versus Simulation time. Here, the ARMSD is taken as the average over all targets. The regression lines are fitted to the points for each temperature.

We observe here that as the simulation progresses, the average RMSD of the trajectories snapshots increases. This effect can be observed independently of the temperature, with regression lines being almost identical, showing that the drift experienced from the native structure does not depend on the temperature of the system. Moreover, this behaviour is observed even when the ARMSD of the decoys are normalised to the average RMSD for that particular target, thus removing possible averaging effects. When we calculated the average Pearson

correlation for each temperature between the Time and ARMSD for each target, we still could not observe any significant difference.

The drift is usually taken to be the time needed for the force field to find a stable structure. In order to test this, we have generated trajectories for the 1HKA protein, both starting from the crystal structure, and starting from a decoy corresponding to a minimised structure from a snapshot at 1 ns in the original simulation. The simulation time was 2 ns, at a temperature of 300 K. Figure 2.5 shows the ARMSD to the native structure versus time for both trajectories.

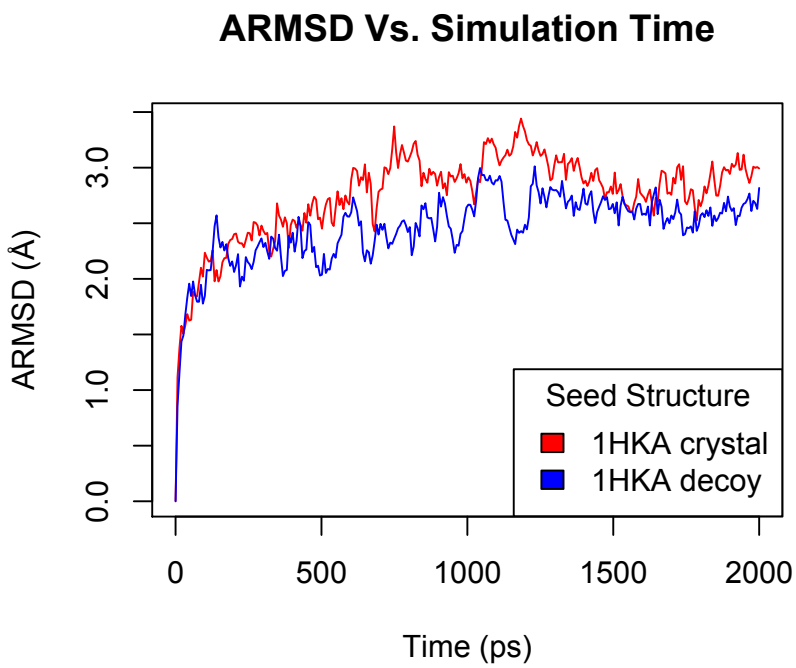


Figure 2.5 ARMSD versus time for different starting structures. Here, the seed structure represents the initial structure used in the molecular dynamics simulation. The ARMSD of the red curve is calculated from the crystal structure, while the blue curve is calculated from the seed decoy structure.

We can see in the above trajectories that regardless of the starting structure, we observe 2 regimes, one where the structure drifts away from the original one (0 to 600 ps), followed by a fluctuation around a mean RMSD value (Time > 600 ps).

The decoy used to seed the second simulation was taken from a stable region in a previous simulation, and as such, the ARMSD should not have drifted so radically. Moreover, when looking at the ARMSD to the crystal structure of the second run, we observe that the trajectory stays stable within a specific range of ARMSD (Figure 2.6).

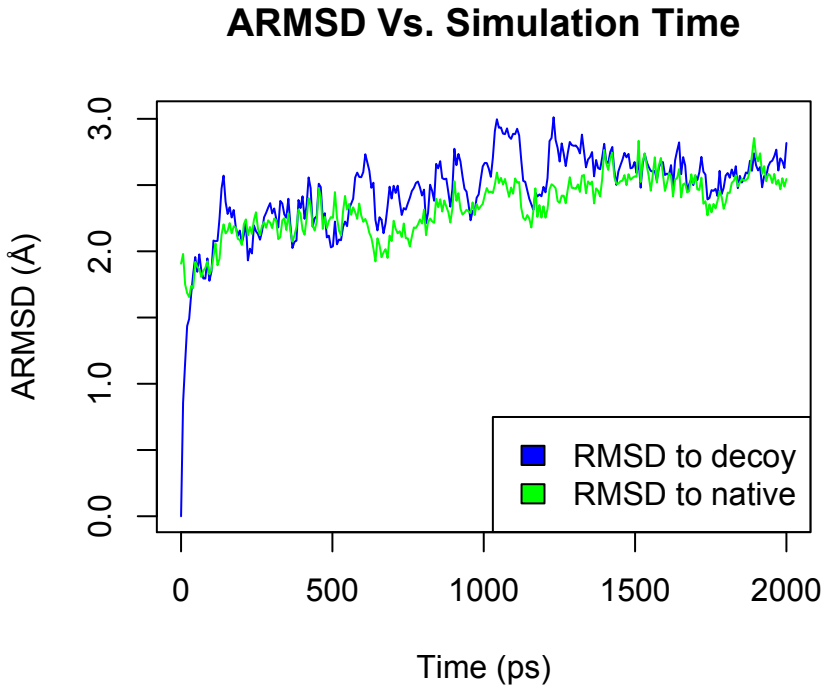


Figure 2.6 ARMSD versus time for the decoy seeded simulation. Here, RMSD to decoy means the seed decoy structure has been used to calculate the RMSD, whereas RMSD to native means the original crystal structure was used.

From a geometrical point of view, the conformational space gets larger as we drift away from the starting structure. This means that randomly sampling around the starting structure will produce more structures at 2 Å than at 1 Å for example. Therefore, the drift experienced is probably due to the random conformational sampling around the starting structure and not to the force field itself.

Moreover, molecular properties are calculated as averaged ensembles, and therefore require more than one structure. Thus, our method is not biased by the time of simulation, but rather accurately represents a random sampling around

the native structure, which is required by the simulation to calculate macromolecular properties.

The only problem though is that since we cannot generate the actual native structure, but only drift from it, the conclusions drawn from this decoy set should not be interpreted as what is needed to fold a protein, but rather as what is needed to keep a protein folded. Furthermore, it shows that the OPLS/aa force field is not able to keep the native structure as it is, and can therefore be improved.

2.3.5 Measuring compactness of decoys

One common assumption is that the native structure is the most compact structure. As such, we could expect the compactness of the decoys to be correlated to their nativeness. To assess this, we compared the ARMSD of decoys to their average ratio of the radius of gyration, expressed as the radius of gyration of the decoy divided by the one of the native structure. This is shown in Figure 2.7 below.

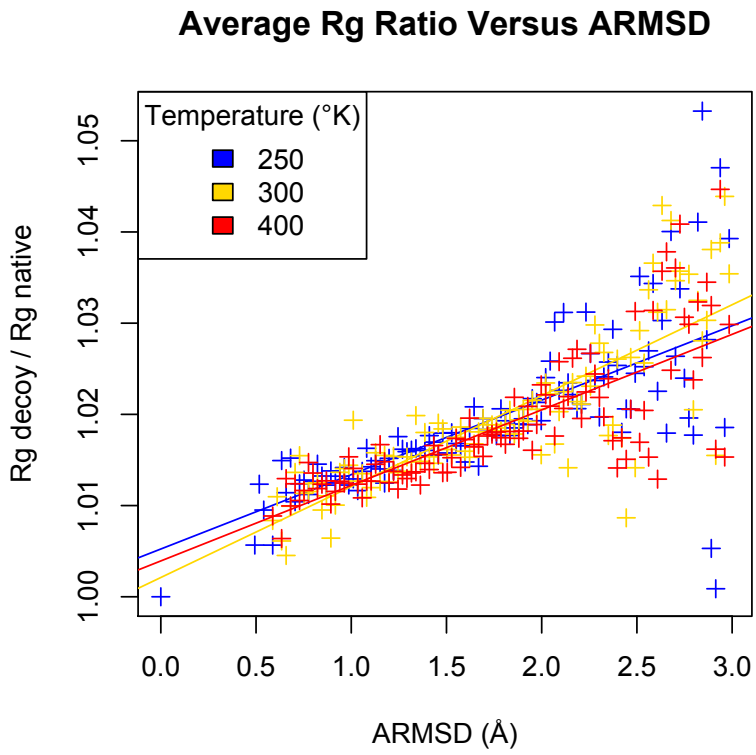


Figure 2.7 Average radius of gyration vs. ARMSD. Here, the average radius of gyration ratio is taken by binning the decoy according to their ARMSD.

The first observation that stems from that plot is that the average radius of gyration ratio is correlated to the average RMSD of the decoy ensembles, for decoys that are less than 2 Å away from the native structure, before diverging and becoming randomly distributed for far away decoys.

Because of the hydrophobic effect, if there are no steric clashes, then the native structures should have a smaller radius of gyration than any other decoys. It is also known that side chains in crystal structures are packed against the protein, whereas they are sticking out in the solvent in simulations. This means that due to differences in the experimental methods used, we will observe a difference in compactness between crystal structures and decoys, that will not be caused by the simulation.

Since finding the native structure is trivial [Handl et al. 2009] and not dependent on the decoy generation procedure used here, we have decided not to use the rank of the native structure as a measure of the quality of a potential.

The second observation is that there is no correlation between the radius of gyration and the simulation temperature, as can be seen from the regression lines that are similar to each other, suggesting that the force field used in the molecular dynamics run is not sufficient to keep the crystal structure packed, even at temperatures of 250K.

Since the temperature does not bias our data set, and since all targets are monomeric and should not interact with ligands, we can conclude that the decoys are unpacking because the force field used is not properly keeping the atoms together. Hence, the radius of gyration is an implicit measure of the error in the force field, and will therefore be used as a benchmark measure. Table 2.7 shows the different average correlation measures for the radius of gyration against the all-atom RMSD, over our generated decoy set, which we call the MDSET.

Table 2.7 Average scores of the radius of gyration for the MDSET decoy set

Measure	Score
10% enrich	0.24
15% enrich	0.28
Pearson R	0.27
Kendall Tau	0.28

From the table above, we can see that the enrichment performs more than twice as well as a random measure for a 10% enrichment, and almost twice as well as random for the 15% enrichment, showing that the radius of gyration is a good benchmark to test energy functions against.

2.3.6 Performance of existing energy functions

Reading through the literature on energy functions, we have found that two of them are commonly used in decoy discrimination studies: DFIRE2 [Yang & Zhou 2008] (all variants), and DOPE [Shen & Sali 2006] which are heavy atom distance and angle statistical potentials. Moreover, these potentials are considered the most successful ones when applied to decoy discrimination. Previous studies have shown that DOPE and DFIRE2 are highly correlated [Rykunov & Fiser 2010] so we will only use the latter as a benchmark.

The energy function we have chosen to test (DFIRE2) is statistical in nature, meaning it was derived from observation rather than empirical forces. Because it uses an inverse Boltzmann formulation, it effectively models the free energy difference that arises when going from a random reference state to a protein-like observed state.

DFIRE2 is distance and angle based, heavy atoms only, and is derived from a subset of the PDB. Here, we assessed the performances of DFIRE2 as well as the radius of gyration, over our MDSET decoy set introduced previously. For each potential, the average enrichment scores at 10% and 15% are shown, as well as the average Pearson R and the average Kendall tau. Results are shown in Table 2.8.

Table 2.8 Comparison of energy function on the MDSET

Measure	R_G	DFIRE2
10% Enrichment	0.24	0.36
15% Enrichment	0.28	0.40
Pearson R	0.27	0.47
Kendall Tau	0.28	0.29

The p-value of the Wilcoxon rank sum test of the difference between the means of the 10% enrichment for the radius of gyration and DFIRE2 was less than 0.001.

2.4 Conclusions

We have seen in this chapter that generating a decoy set using molecular dynamics allows for a good sampling of the near-native basin, with structures generated below 4 Å RMSD, with an average of 1.7 Å.

The decoys were generated from the native structure, which means that the potential used in the molecular dynamics simulation was not able to stabilise it. This implies that however long we simulate the proteins in our decoy set, we will never revert back to the native structure, or it would not have drifted in the first place. Therefore, this decoy set is not representative of methods used to actually fold proteins, and should not be used to assess potentials designed for *ab initio* folding.

One could argue that near-native structures can sometimes be produced by *ab initio* method, and that these should be used as our near-native decoy set. But by looking at the recent CASP results, very few algorithms actually folded protein to less than 2 Å RMSD, and even fewer to less than 1.5 Å RMSD. These algorithms also did not produce near-native structures for many targets, meaning the sample size and total number of decoys would be relatively small. This ultimately implies that if we did generate structures below 2 Å RMSD, we could probably not detect them, as the potentials would have been tuned on decoys with RMSDs mostly above 2 Å. Thus, relaxing the native structure is the only way we currently have to precisely and non-redundantly sample the 0 to 2 Å RMSD range.

Although there is a strong correlation between time of simulation and RMSD, this was found to be not significant in our study, as it represents the conformational sampling, rather than any bias due to the potential, that could then have been trivially spotted. On the other hand, we have observed a significant correlation between the radius of gyration of decoys, and their nativeness, with an average 10% enrichment score of 0.24 over all targets. This score means that the radius of gyration can successfully detect 24% of the 10% best structures as being such. Thus, it is a good benchmark for testing energy functions.

We have then assessed the performance of one of the most successful potentials from the literature, DFIRE2, which scored an average 10% enrichment score of 0.36. DFIRE2 is based on heavy atoms only, and includes four terms, one for the pairwise atomic distance, and three for the orientation between two atoms. In this thesis, we will use DFIRE2 as a base potential in order to develop a more

complete potential that includes hydrogen bonding, solvation, and information from decoy structures.

Chapter 3

Deriving a solvation free energy potential

3.1 Introduction and background

Protein folding is in large part due to the hydrophobic effect, by which atom groups having no affinity to water are buried inside the protein [Anfinsen & Scheraga 1975, Karplus & Weaver 1976]. Thus, when simulating folding, it is important to take this into consideration, which can be done using various methods, one of which is by calculating the free energy associated with the solvent, by either using an implicit solvent model [Einsenberg & McLachlan 1986, Bhattacharyay et al. 2006, Lazaridis & Karplus 1999a], or by using explicit water molecules in conjunction with suitable energy functions [Zielkiewicz 2005, Jorgensen et al. 1983, Hermans et al. 2004].

Although extensive research has been conducted, there still is not a consensus as to which approach is better [Boas & Harbury 2007, Dill et al. 2005]. Using explicit solvent can be computationally expensive, which is why most approaches have looked more closely at implicit solvent models [Privalov & Gill 1988].

Various methods exist to approximate the solvation free energy from a continuum. One of the first successful attempts called the Eisenberg-McLachlan (EM) potential [Einsenberg & McLachlan 1986] used all heavy atoms in the calculation, where the free energy is the sum of the individual contribution of each atom which is defined as its surface accessibility multiplied by a solvation parameter.

Although this model has been widely accepted, other approaches have been used, the most notable, and probably the most popular being the Lazaridis-Karplus (EEF1) potential [Lazaridis & Karplus 1999a]. Although the original paper describes the performance of the CHARMM+EEF1 potential, the solvation term alone has been used in other studies, and added to other potentials to model solvation. In essence, the EEF1 potential tries to model the exclusion and distribution of solvent around protein atoms by deriving a function representing the solvation free energy density.

Solvation potentials are usually not sufficient to discriminate decoys correctly, and thus have to be added to other free energy potentials, such as pairwise atomic distance potentials, which have been extensively studied, and are now being augmented to include other interactions and increase their discriminatory power. Of the existing distance potentials, we found DOPE [Shen & Sali 2006] and DFIRE2 [Yang et Zhou 2008] to be equivalently good at discriminating decoys in

medium to low resolution decoy sets [Rykunov & Fiser 2010]. Thus, in this experiment, we will be using DFIRE2 as the basis of our energy function.

In order to derive our potentials, we have tested various protocols, all of which are statistical potentials based on an inverse Boltzmann potential that allows us to model the distribution of the energy function without knowing its functional form [Sippl 1990]. Statistical potentials require a reference state to be defined, which can be thought of as the distribution expected at random or from unfolded conformations. Here, we study 2 approaches to deriving reference states, one based on decoy structures, and the other based on a statistical model of the expected randomness.

Solvation was modelled implicitly by using the solvent accessible surface area (SASA). From it, we derived a statistical potential to model the free energy associated with the solvent exposure of specific atom types.

We analysed the performances of existing solvation potentials, and found that they did not significantly improve DFIRE2, which is why we decided to derive a new one using a different approach. A the end of this chapter, a comparison will be made between our own potentials and the ones commonly used in the literature.

3.2 Methods

3.2.1 Hydrogen atom generation

In order to generate the hydrogens needed in the calculation of some potential functions, we have used the pdb2gmx tool from the GROMACS package [Hess et al. 2008], along with the OPLS/aa force field [Jorgensen & Tirado-Rives 1988]. This method was applied to the various protein crystal structures and decoys lacking explicit hydrogens.

3.2.2 Protein sets

Crystal structures set

Starting from the PDB dataset [Bernstein et al. 1977], we used the PISCES server [Wang & Dunbrack 2003] to generate a subset of 4370 targets with less than 30% sequence identity, a resolution better than 2 Å, and an R-value better than 0.25. PDB files containing unknown residues or atom types were removed, and the structures minimized using GROMACS. This selection process yielded 713 crystal structures.

Decoy sets

Since we are only interested in near-native decoy discrimination, we have chosen 2 decoy sets with an ARMSD range between 0 and 5 Å. Such sets are typically hard to produce, mostly because current ab initio methods cannot produce models of that quality very often. Hence, the 2 sets used here have a common generation procedure in the sense that they are both based on a relaxation of the native crystal structure.

The first set, MDSET, was introduced in the previous chapter, and was generated using molecular dynamics runs at various temperatures, and sampling along the trajectory. This was done for 250 targets, generating 2000 decoys each time. We then used a QT clustering algorithm with a 0.5 Å RMSD cut-off, which yielded between 600 and 1200 decoys for each target. 500 models were then taken at random for each target to avoid over representing a specific protein when we use them as a training set.

In this decoy set, we have found that the ARMSD correlates to the time after which the snapshot was taken [Figure 2.4, chapter 2], and although this does not invalidate the approach, it limits the interpretation of the results. Therefore, we

have used another decoy set from the literature, which we call the HRDECOY set [Rajgaria et al. 2006], that was generated by constraining the hydrophobic core and charged groups of the proteins according to the expected theoretical behaviour, and varying the amount by which they are allowed to deviate from their original position. The NMR refinement package DYANA [Guntert 2004] was then used to generate the models according to various parameter sets. In total, this decoy set is comprised of 1400 targets, with 500 to 1000 decoys in each. Since we do not need that much data in our study, we have randomly selected 150 targets that we will use in our analysis. These structures are comprised of X-ray structures, and have a balanced distribution of alpha, beta, and alpha/beta proteins. 500 decoys per targets were then randomly selected.

3.2.3 Solvent model

We use an implicit solvent model based on the surface accessibility, defined as the percentage of an atom's surface that is exposed to a probe representing a water molecule. Here, we used a program called NACCESS [Hubbard 1996] with default parameters and a probe radius of 1.40 Å. NACCESS is an implementation of the "rolling ball" algorithm, where the surface accessibility is taken by simulating a ball rolling on the surface of the protein [Lee & Richards 1971].

3.2.4 Statistical potentials

Boltzmann potentials

Knowledge based potentials have the useful property that they allow the calculation of the free energy associated with a structure, without knowing the empirical functional form of the force being modelled. It assumes little about the force itself, aside that it should be represented in a non-redundant set of protein structures and that it can be modelled in the same way as a gas. These types of potentials are commonly derived using an inverse Boltzmann equation.

The more data is in the protein set, the smaller the bins can be, as there is less chance that some would be empty. When this is the case, one has to define a protocol for handling such empty bins. We tested several pseudo-count protocols, and did not find any significant difference between methods when using the potential to discriminate our decoy sets, as we had relatively few empty bins aside from extreme values. Thus, in this thesis, we add a pseudo count of 1 to each bin, except for extreme values, which we decide to ignore completely, and assign to them an energy of 0. For example, in solvent accessibilities potentials, we ignore bins above 60%, while adding a pseudo-count of 1 to bins between 0% and 60%.

The observed state is taken from the set of proteins for which a structure is resolved to a good accuracy. The reference state will usually either be a model describing the probability of a specific bin occurring at random in proteins, or be a set of unfolded protein structures. The set of structures used to derive a statistical potential can be thought of as a training set, and the set used for testing the efficiency of this potential would be called a testing set.

Training and testing sets

In this experiment, we will look at various protocols for deriving the frequencies in the observed and reference states, some of which use decoy structures. Therefore, to avoid testing our potentials on the same data it was derived from, we have split our decoy sets randomly into two subsets, taking around 70% of the targets as the training set, and 30% as the testing set. For the MDSET, there are 180 targets in the training set, and 70 in the testing set. For the HRDECOY set, there are 110 targets in the training set and 40 targets in the testing set. When deriving classical reference states, we have used the native structures

from the decoy sets, and added the structures from the crystal structure data set introduced earlier, which gives us a total of 1003 native conformations.

Potential parameters

In order to derive a solvation statistical potential, we need to define 3 features, namely the bin size, the range of values considered, and the grouping of atoms involved.

We have on average 2000 atoms per protein, giving us, for 1003 proteins, a total of about 2,000,000 atoms. This is the minimum number of observation that we will have in our potentials, since adding decoys will increase it. Considering the large amount of data available, we opted to choose bin sizes that would be fit to represent the small differences between decoys.

For our SASA potential, we used 327 atom types, one for each atom in each residue, with a bin size of 2% in the range [0%, 60%], corresponding to the % surface accessibility of the atom. 60% was used as a cutoff as it is the largest observed SASA in our set of proteins and decoys.

Classical potentials

In this study, we define a classical reference state as being derived from crystal structures and probabilistic models only. The observed state in classical potentials is derived from the distribution of features in a set of crystal structures resolved at a high resolution. The reference state is then derived using an theoretical model corresponding to the probabilities of a specific bin occurring at random.

For our SASA potential, the reference state is not trivial, and as such, we opted for a statistical determination of it. We assumed a globular shape for proteins, and modelled the random chance of a specific atom to be buried by creating a sphere of radius 30 Å, and placing 2000 random atoms in it. This would correspond to a 200 residues long globular protein. Atoms were placed without overlap of their van der Waals radii, and the proportion of each element (N, C, O, S, H) in the sphere was taken from 1000 crystal structures of similar length. The surface accessibility was then calculated using NACCESS in the same way as for real proteins. We generated 10,000 such spheres, giving us a total of 20,000,000 atoms with their respective SASA, and in proportions similar to what would be

found in real proteins. The only difference is that we assumed an ideal-gas model, where no covalent bonds exist between atoms in the sphere created. These random “proteins” were then used to model the reference state, by binning them in the same way as we would bin the real proteins in the observed state.

Decoy based potentials

Our decoy-based potentials are derived from decoys only, intentionally omitting native conformations. The potentials were derived by taking the decoys that are better than average as the observed state, and those that are worse than average as the reference state. The idea behind this formulation is that we noticed that decoys are normally distributed around a mean value, and thus, we are interested in knowing what the difference is between structures that are better or worse (in terms of RMSD) than others. These decoy-based potentials are thus dependent on the method used to generate the structures, and as such, potentials should not be expected to be transferable to other decoy sets. The naming convention used in this chapter for our potentials is given in Table 3.1 below.

Table 3.1 Naming conventions for potentials

Method	Subscript	Observed State	Reference State
Statistical	C	Crystal Structures	Protein-like random spheres
Statistical	D	Better than average decoys	Worse than average decoys

3.2.5 Benchmarking potentials

All-Atom RMSD (ARMSD)

In this study, we measure the degree of nativeness of a structure by taking the root mean square deviation between a model and an experimental structure, using all heavy atoms. This gives us the level of detail needed when considering near-native decoys. One of the main drawbacks of using RMSD measures is that it is very sensitive to local changes. But since all our decoys are created by relaxing the native structure, none of them will be subjected to this effect, as molecular dynamics does not allow for huge spontaneous moves in the short time frame that we simulated, and at the temperatures that we used.

Radius of Gyration

As a benchmark in this study, and as a control to verify that the decoys do not simply expand in all direction (unfold), we have used the radius of gyration, defined as the root mean square distance between atoms in a protein, as the minimal feature to outperform.

Total Surface Accessibility

Another control feature that was used is the total Solvent Accessible Surface Area (SASA) of the protein. Indeed, the more compact, the smaller the exposed area will be, and thus, it is a measure of compactness as well as a measure of how solvated the protein is.

Potentials from the literature

To assess the performances of our potentials, we have compared them against 2 other solvation energy functions taken from the literature, the Eisenberg-McLachlan potential (shortened to EM in this study) [Eisenberg & McLachlan 1986], and the Lazaridis-Karplus EEF1 solvation energy function [Lazaridis & Karplus, 1999a]. Each of them has been used with varying success in other experiments, which is why we included them in our benchmark. Finally, DFIRE2 will be used as the base potential on top of which we will be adding our solvation term.

3.2.6 Combining energy terms

Energy terms are linearly combined using a set of weights optimised using R (*optim* function available in the *stats* library). We optimised the average 10% enrichment score over all targets in the decoy set, using the Nelder and Mead optimisation protocol, and calculated the other statistics using the weights derived from it.

3.3 Results & Discussions

3.3.1 Generation of the classical reference state

In order to derive the probabilities of the SASA potentials reference state, we computed the radius of each protein used to calculate the classical potentials. This is shown in Figure 3.1 below.

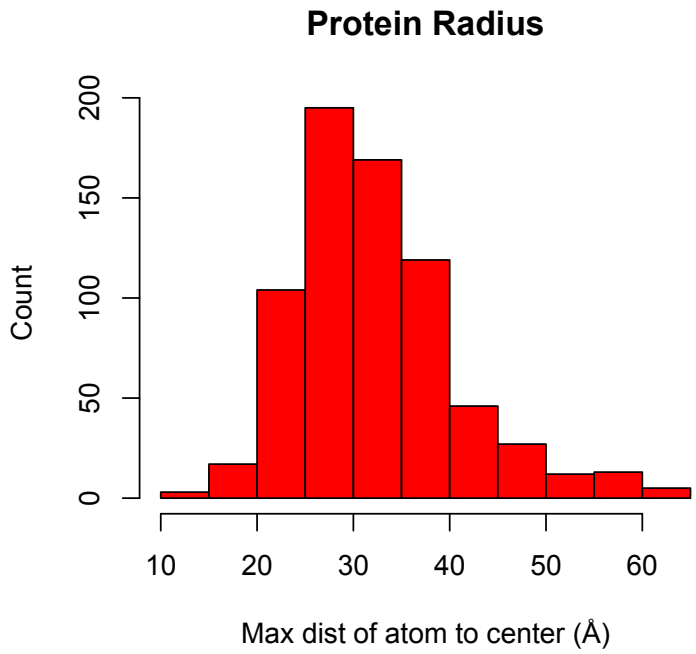


Figure 3.1 Distribution of protein radius. Here, the proteins used are all crystal structures representing the native conformation. The radius is taken as the maximum distance between any atom and the centre of the protein it is in.

As can be seen, the average radius is around 30 Å, and thus, we will use this as the radius of the random spheres we will be creating. Using the same dataset, we calculated the proportions of each element in the proteins. Results are shown in Table 3.2.

Table 3.2 Proportions of elements in proteins

Element	Proportion (%)
N	8.6
C	31.9
O	9.5
S	0.2
H	49.8

These probabilities of occurrence and radiuses are then used to randomly fill spheres of radius 30 Å with 2000 non-overlapping atoms. By doing so, we ensure that the contents of our random protein-like spheres correspond to what would be expected in real proteins. This was then repeated to create 10000 spheres, and the distribution of surface accessibility for each element was computed, as shown in Figure 3.2.

The reference state was then calculated from these distributions by binning them according to the same protocol as the crystal structure atoms in the observed state. From Figure 3.2, we can see that the distribution differs across different atom types. Generally, counts are peaking at 0% SASA, quickly drop, and stabilise around 20% SASA. It then stays relatively flat until 40% SASA, before dropping close to 0 for SASA > 60%. The difference between the different atoms comes from the relative proportions of each SASA counts. For hydrogens, the number of atoms on the surface drops very sharply, while they drop more smoothly for carbon and sulphur, and somewhere in between for nitrogen and oxygen. This is not surprising given that for a fixed number of atoms in a fixed size sphere, the larger atoms will require more overlaps to be buried than small atoms, and thus, this will happen less often at random.

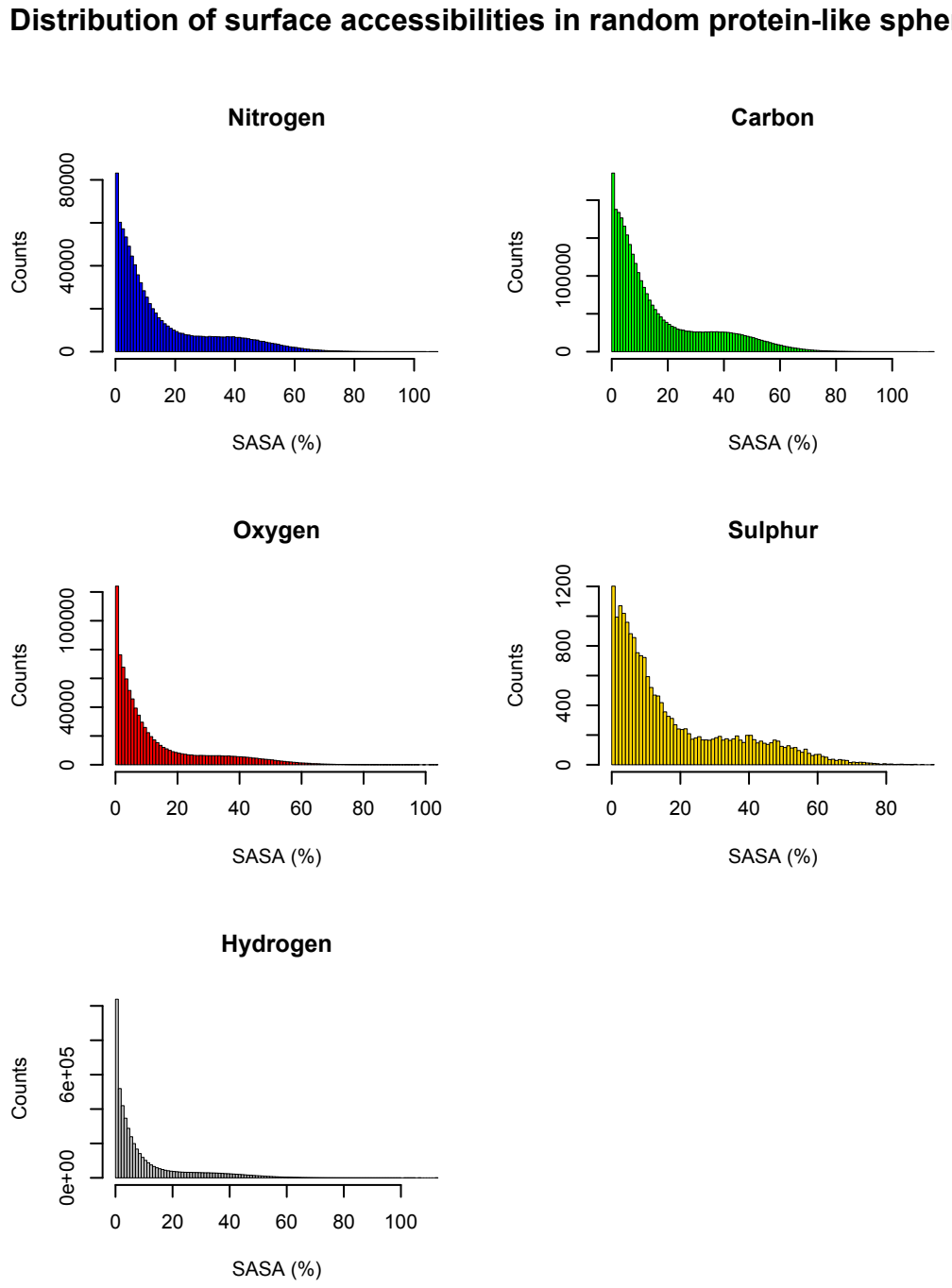


Figure 3.2 Distributions of surface accessibility for random atoms in protein-like spheres. The distributions were generated from 10000 protein-like spheres containing 2000 atoms each, in the same proportions as real proteins.

3.3.2 Classical potential

To generate our classical potential, we took crystal structures from the PDB, and energy minimised them using GROMACS and OPLS/aa in order to remove any steric clashes. The resulting structures were then used as the training set for our classical potentials throughout this thesis.

Using an inverse Boltzmann formulation, and the reference state derived in the previous section, we generated a free energy of solvation potential, using the solvent surface accessibility of atoms. We treated each atom from each residue separately, in order to model more subtle differences that would be due to the chemistry of specific residues.

To illustrate our method, we chose threonine, as it is relatively not too hydrophilic or hydrophobic (hydropathy index = -0.7), has relatively good side chain flexibility, and is a frequently occurring amino acid in natural proteins. Potentials for Threonine heavy atoms are shown in Figure 3.3, for accessibilities between 0 and 60%, which is the maximum observed SASA in our protein set.

Potentials in Figure 3.3 show a consistent preference for fully buried atoms with 0% accessibility. Oxygen and side chain CG2 atoms also have a secondary preference for accessibilities above 13% to 15%. This shows that atoms are only stable in two states, one where they are fully buried, and the other where they are well exposed to solvent, with intermediate exposures not being favourable.

Although we only show potentials for threonine heavy atoms, every of the 327 atoms considered have clear favoured and disfavoured regions, and are used in the calculation of the total solvation free energy.

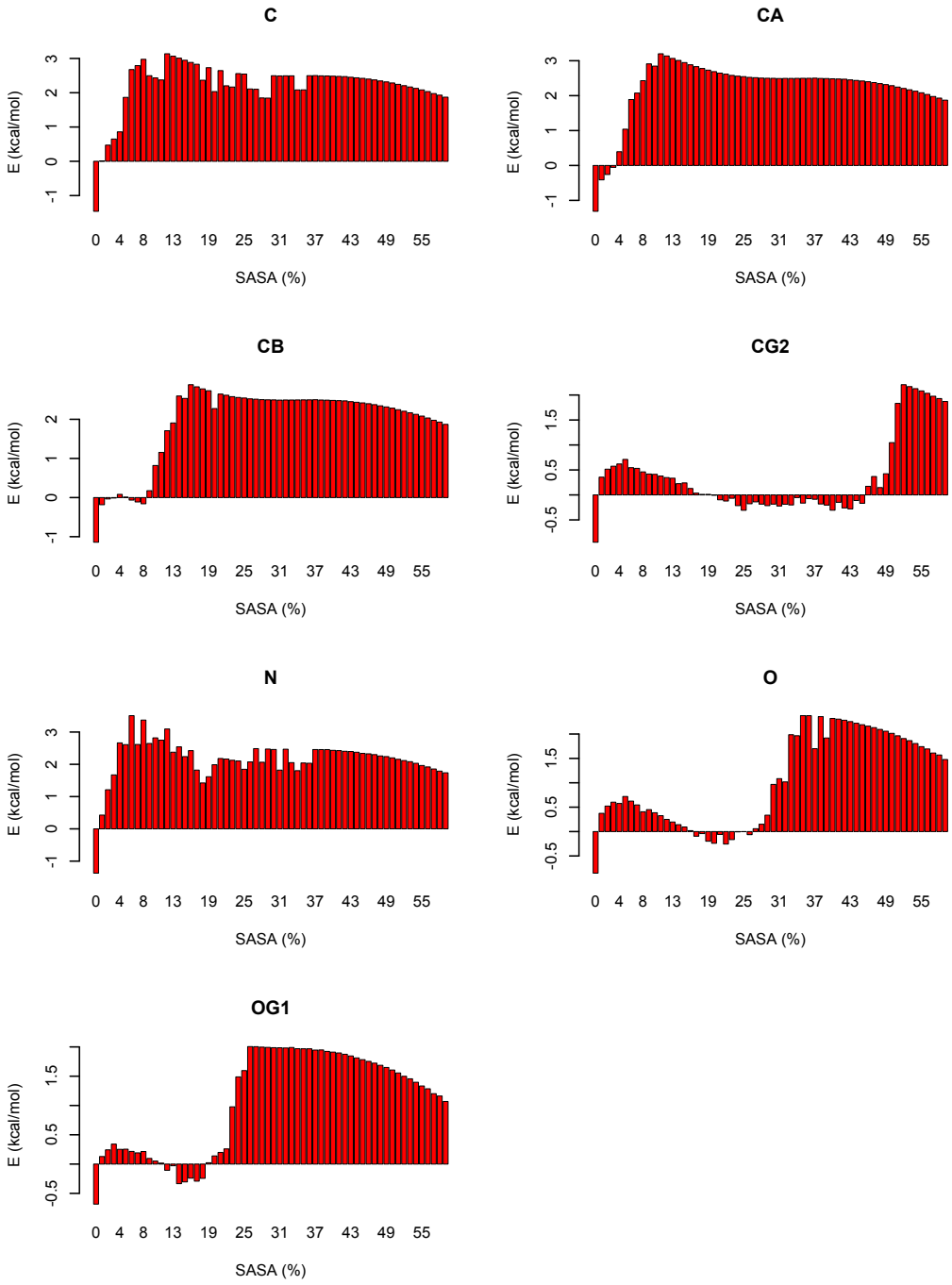


Figure 3.3 SASA Classical potentials for Threonine heavy atoms. The potentials were generated using crystal structures from the PDB, and the reference state derived from random spheres. A pseudo-count was added to all bins in the observed state, which is why we observe some smoother regions.

Using these potentials, we calculated the free energy for our MDSET decoy set, and added the new term to DFIRE2. Results are shown for the MDSET in Table 3.3 below.

Table 3.3 10% Enrichment score of the classical SASA potential

Potential	10% Enrichment	W	p-value (H_0=DFIRE2)
DFIRE2	0.41		
$SASA_C$	0.20	2553	5e-9
DFIRE2 + $SASA_C$	0.41	1568	0.50

The Wilcoxon rank sum test (**W** in the above table) was used to assess the significance of the difference between DFIRE2, and potentials including the classical SASA term ($SASA_C$ and DFIRE2+$SASA_C$). As can be seen, there was no improvement after adding the solvation term. Thus, we can conclude that using a classically derived solvation potential does not help in discriminating near-native decoy structures. This was further verified by generating the classical potential with various pseudo-count protocols, and also by using a constant reference state. In all instances, the potential did not improve DFIRE2.

3.3.3 Decoy-based potential

Since we could not improve DFIRE2 using a classical solvation potential, we derived another one, but this time only using decoys. Here, we define the observed state as the set of decoys that are better than the average RMSD of decoys belonging to the same initial target. Thus, for our 250 targets, we have 250 different means, although they are all around 2 Å RMSD. The reference state is then taken as the set of decoys that are worse than average.

The resulting distribution represents the difference between the runs that ended in good models, and those that ended in bad ones, and should in theory measure the energy gain or loss for specific values of our potential. When we observe values for which the potential is negative, we can conclude that good models are more often occurring in that region, whilst when observing positive values, it's the other way around, and bad models are more represented in that bin.

We illustrate our method by showing the decoy-based SASA potentials derived from Threonine heavy atoms (Figure 3.4 on the next page).

Our first observation is the relatively flat bin for 0% accessibility for all atom types. This means that there is no observable difference between fully buried atom in good and bad models, and thus, we cannot gain information from only considering these atoms.

Given that the 0% bin represents the lowest energy in all of our classical potentials, it is not surprising that we could not improve DFIRE2 by using them.

On the other hand, clear favoured or disfavoured regions can be observed in our decoy-based potentials, such as for the Threonine CB atom, where we have a strong positive peak around 10% accessibility, implying bad decoys have more Threonine CB atoms in that range. Likewise, information is gained from the distribution of surface accessibility of the Threonine main chain N atom, which has a negative peak at 20%, implying good models have more atoms around that value than bad models.

Overall, calculating the energy for decoys using these potential should give us a measure of how good they are, or at least, if they are better or worse than average.

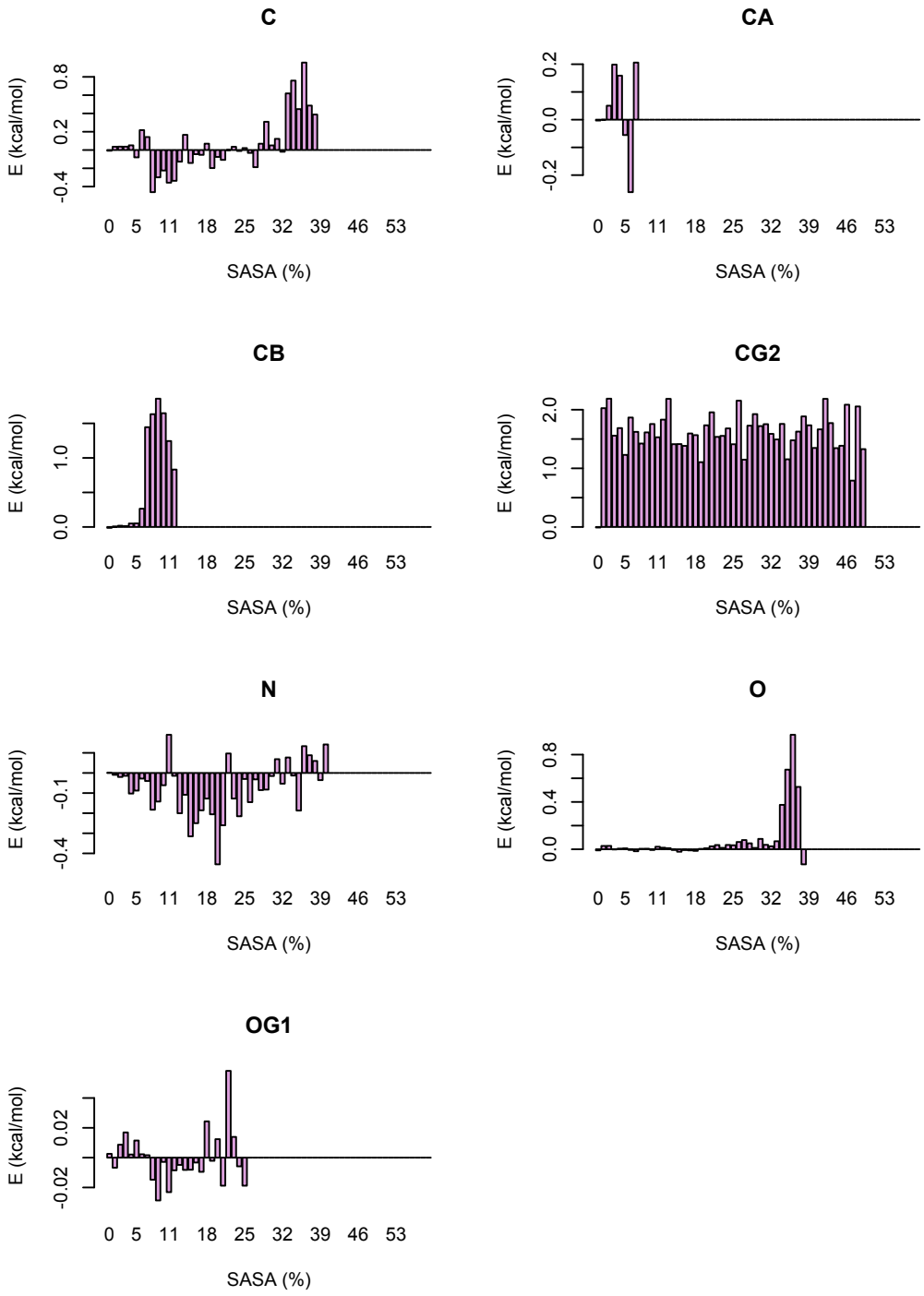


Figure 3.4 SASA Decoy-based potentials for Threonine heavy atoms. The potentials were derived using decoys generated in molecular dynamics run. The observed state is composed of good quality decoys, while the reference state is composed of bad quality ones. Here, when no count is observed in either the observed state or the reference state, or when the frequencies were equal, we assigned an energy of 0.

Using the previously derived decoy-based potentials, we calculated the total solvation energy for each decoy in the MDSET, and compared it to DFIRE2 to assess whether it improved its discriminatory capabilities. Results are shown in Table 3.4 below.

Table 3.4 10% Enrichment score of the Decoy-based SASA potential

Potential	10% Enrichment	W	p-value (H_0=DFIRE2)
DFIRE2	0.41		
$SASA_D$	0.37	1767	0.12
DFIRE2 + $SASA_D$	0.50	1053	0.001

We can see that the $SASA_D$ potential is not significantly different from DFIRE2 in terms of performance. On the other hand, combining the two makes the 10% enrichment score go from 0.41 for DFIRE2 alone, to 0.50 for DFIRE2+$SASA_D$. The difference between these scores was assessed using a 1-tailed Wilcoxon rank sum test, which yielded a p-value of 0.001, making this improvement highly significant. We can therefore conclude that using a decoy-based potential significantly helps to improve the accuracy of DFIRE2 when discriminating near-native decoy structures.

3.3.4 Comparison to potentials from the literature

In order to verify our results, we have compared our potentials to two of the most used ones in the literature, namely the solvation term of the EEF1 potential [Lazaridis & Karplus 1999a] and the EM potential [Eisenberg & McLachlan 1986]. We combined them to DFIRE2 using the same method as for our own potentials, and compared the accuracy of the resulting energy function on the MDSET decoy set. We also included the results for the radius of gyration (Rg) and the total surface accessibility (ASA). The p-values are shown for the difference to DFIRE2 alone.

Table 3.5 10% Enrichment score of the different potential terms

Potential	10% Enrichment	W	p-value (H_0=DFIRE2)
DFIRE2	0.41		
$SASA_D$	0.37	1767	0.12
$SASA_C$	0.20	2553	5e-9
EEF1	0.28	2243	4e-5
EM	0.09	2943	5e-05
Rg	0.25	2302	1e-5
ASA	0.22	2416	4e-7
DFIRE2 + $SASA_D$	0.50	1053	0.001
DFIRE2 + $SASA_C$	0.41	1568	0.50
DFIRE2 + EEF1	0.43	1450	0.25
DFIRE2 + EM	0.42	1512	0.38
DFIRE2 + Rg	0.41	1577	0.49
DFIRE2 + ASA	0.42	1490	0.33

The only potential that significantly improved DFIRE2 was our decoy-based surface accessibility potential. As can be seen, neither the classical version, nor the potentials from the literature could augment the accuracy of DFIRE2 when discriminating near-native decoys.

We have further verified our results by applying it to the HRDECOY decoy set, which is comprised of 150 targets with 500 to 2000 decoys each. Since our method is specific to a given decoy generation method, we split the HRDECOY

set into 2, where 110 targets were used to generate the potentials, and 40 were used as an out sample to test them. Results are shown in Table 3.6.

Table 3.6 Performance of the Decoy-based SASA potential on the HRDECOY set

Potential	10% Enrichment	W	p-value (H_0=DFIRE2)
DFIRE2	0.42		
$SASA_D$	0.35	1069	0.003
Rg	0.23	1334	1e-7
ASA	0.26	1287	1e-6
DFIRE2 + $SASA_D$	0.47	643	0.06

Although the DFIRE2+$SASA_D$ performed better than any other potential for the HRDECOY set, the Wilcoxon 1-tailed test gave a p-value of 0.06, meaning this difference is not quite significant. Since we had a significant increase on the MDSET, we cannot decide purely based on the 10% enrichment, whether this method is useful. To address this, we calculated the Pearson correlation coefficient and the Kendall tau, and compared the results for both potentials, on both the MDSET and HRDECOY set. Results are shown in Table 3.7 below.

Table 3.7 Performance scores of the DFIRE2 + $SASA_D$ potential

	DFIRE2	DFIRE2+$SASA_D$	W	p-value (H_0=DFIRE2)
MDSET				
10% Enrichment	0.41	0.50	1053	0.001
15% Enrichment	0.45	0.54	1143	0.007
Pearson R	0.54	0.60	1292	0.05
Kendall Tau	0.36	0.41	1315	0.07
HRDECOY				
10% Enrichment	0.42	0.47	644	0.06
15% Enrichment	0.49	0.55	600	0.03
Pearson R	0.73	0.82	488	0.001
Kendall Tau	0.55	0.65	421	0.0001

From Table 3.7, we can see that aside from the 10% enrichment on the HRDECOY, and the Kendall Tau on the MDSET, every other statistics show that the improvement of DFIRE2+$SASA_D$ over DFIRE2 is significant. Therefore, we can conclude that adding a solvation term derived from decoys significantly improves the ability of DFIRE2 at discriminating near-native decoy structures.

3.4 Conclusions

In this chapter, we have studied how using solvent accessible surface area as the basis for a knowledge-based solvation potential could significantly improve the ability of DFIRE2 at discriminating near-native decoys. Using a new potential generation method, we were able to improve the discriminative power of DFIRE2 by 5% to 10%, depending on the decoy set chosen, and the score used.

To achieve this, we tried several methods to pick an observed and reference state to be used in the inverse Boltzmann potential. The first is based on a conventional approach using crystal structures as the observed state, and a geometrical model for the reference state.

To accurately represent the randomness expected for surface accessibilities, we randomly generated 10000 spheres with a radius of 30 Å, and randomly placed non-overlapping N, C, O, S and H atoms in the same proportions as real proteins. The resulting distribution of surface accessibilities was then used as the reference state in our classical potential.

The analysis of the distribution of the classical potentials showed that most atoms have an almost binary state, preferring either to be fully buried with surface accessibilities of 0%, or be strongly exposed with accessibilities superior to 15%. This type of interactions tends to limit the accuracy of potentials, as it does not model small variations in the accessibilities, but rather only verifies that the right atoms are solvated or not. Thus, in a near-native decoy set such as the MDSET, the potential failed to improve DFIRE2, as all decoys are well solvated, and have native-like structures, meaning that using such a simple approach would not allow differentiating between closely related good and bad decoys.

To address this issue, we modelled the energy gained from having a better-formed structure (good decoy) compared to having a worse one (bad decoys). Thus, however small the difference is, given a large enough sample size and small enough bin sizes, we can extract the difference between the two, and assess how good a structure is, which should then be equivalent to its nativeness.

This was done by taking the better than average decoys as the observed state, and the worse than average decoys as the reference state. When our potential is negative, then we have more good structures than bad ones in that specific region, and vice-versa for positive values. This approach allowed us to

significantly improve the power of DFIRE2 when discriminating near-native decoys, and this, across multiple decoy sets, using multiple measures.

Chapter 4

Including C-H…X hydrogen bonds in statistical potentials

4.1 Introduction

Before computational methods were accessible, and analysis of large dataset of proteins possible, the thermodynamics and geometry of hydrogen bonds were studied in crystals of small molecules [Sutor 1963, Pimentel & McClellan 1971]. Although the hydrophobic effect is considered the main factor behind protein folding, it has been suggested that hydrogen bonds play a key part in the process [Myers & Pace 1996].

Hydrogen bonds involve the interaction between two electronegative groups, one composed of a heavy atom bonded to a hydrogen atom (the donor) and the other composed of an electronegative heavy atom (the acceptor). Conventional “strong” hydrogen bonds have been extensively studied in proteins [Baker & Hubbard 1984, McDonald & Thornton 1994], and involve the interaction between nitrogen and oxygen donor and acceptor groups (NH-O, OH-O and NH-N). One feature owed to hydrogen bonding is the secondary structure of the protein [Pauling 1960]. Indeed, beta sheets and alpha helices are defined as regions where residues follow a specific hydrogen-bonding pattern. Although various types of helices exist, the most common one found in proteins is the alpha-helix, which consists of 4 or more residues with their main chain nitrogen group hydrogen bonding to the main chain oxygen acceptor 4 residues earlier. Beta sheets on the other hand are adjacent strands of residues where the main chain nitrogen group on one strand hydrogen bonds with the carboxyl oxygen of an adjacent one.

In addition to NH and OH donors, carbon groups involved in hydrogen bonding have been observed and studied in inorganic compounds and membrane proteins [Sutor 1962, Desiraju 1996, Mottamal & Lazaridis 2005]. Since they are at least as common as nitrogen and oxygen groups, they are thought to be participating in stabilising the protein [Fabiola et al. 1997, Wahl & Sundaralingam 1997], as well as interacting with the solvent when no NH bond is [Steiner 1995].

But one major difficulty in studying hydrogen bonds is the lack of available data. Indeed, current crystallographic and NMR methods very rarely resolve hydrogens, and their positions either have to be inferred from geometry, optimised according to a force field, or observed using other experimental methods such as neutron diffraction. As the number of structures resolved using neutron diffraction is small, the only way to analyse hydrogen bonds with

statistical significance and without bias towards a specific force field is to consider only those that are fixed by geometry.

Previous studies have shown that an improvement over existing potentials can be achieved by adding a hydrogen bonding term derived from a statistical potential [Kortemme et al. 2003], but they all assumed linear separability of each geometrical feature, and thus independence. It has been shown that pairwise features cluster [McDonald & Thornton 1994], and as such, they should be treated together. Moreover, most potentials only treat NH bonds, even though it has been suggested that potential CH and NH main chain hydrogen bonds work in tandem in stabilising the backbone [Fabiola et al. 1997, Steiner 1995].

In this experiment, we will study the impact of carbon donating groups on the discrimination of decoy structures, as well as deriving a multivariate statistical potential to account for dependence between bonding features, in order to try to improve the discriminatory power of DFIRE2.

4.2 Methods

In this section, we will introduce the various data sets and methods used to generate hydrogen bond potentials. Moreover, we will enumerate the different groups being considered, and the different features being analysed. The newly derived energy function will then be combined with DFIRE2 to try to improve its discriminatory performances, assessed using various metrics.

4.2.1 Protein sets

Crystal structure sets

In order to study the properties of hydrogen bonds, we need to either use explicit hydrogen positions as seen in neutron diffracted structures, or use a procedure to place them. Considering the small amount of neutron-diffracted structures, we opted for a larger decoy set where fixed hydrogens have been placed artificially. This protein set is the same one as used in generating the solvation potentials, and is comprised of 713 proteins resolved at a resolution of 2 Å or less. For more details on this protein set, refer to chapter 3, sections 3.2.1 and 3.2.2.

Decoy set

In order to assess the performances of the potentials, we tested them on a decoy set composed of native structures and decoy structures exhibiting near-native features. This decoy set has been introduced in a previous chapter, and is labelled MDSET. It has structures with all-atom RMSDs between 0 and 4 Å away from the crystal structure, and exhibits no obvious defects. There are 250 targets with 500 decoys for each, 70 of which are used as out-sample testing set. The results are then validated using the HRDECOY decoy set, comprised of 150 targets, each with 500 decoys. We will use 40 of these targets for testing, and 110 for training. The targets used in this experiment are different from chapter 2, which is why different results for DFIRE2 are observed.

4.2.2 Hydrogen bond definitions

Hydrogen bonding groups considered

As it is impossible to generate the position of rotating side chain hydrogen bonds, we only consider groups where the hydrogen is constrained by the geometry of the residue. This means that no OH, NH2, NH3 or CH3 groups will be included in this analysis, as the position of the hydrogen atoms cannot be inferred by geometry alone. The result is that our potential will only consider specific carbon and nitrogen donors. Only carbons that are covalently bonded to an electronegative oxygen or nitrogen will be considered potential hydrogen bond donors. This is because the presence of a nearby electronegative atom can modify the charge distribution around the carbon, making it a potential donor, as observed in some crystals [Sutor 1963]. Figure 4.1 shows the different CH donating groups in amino acids.

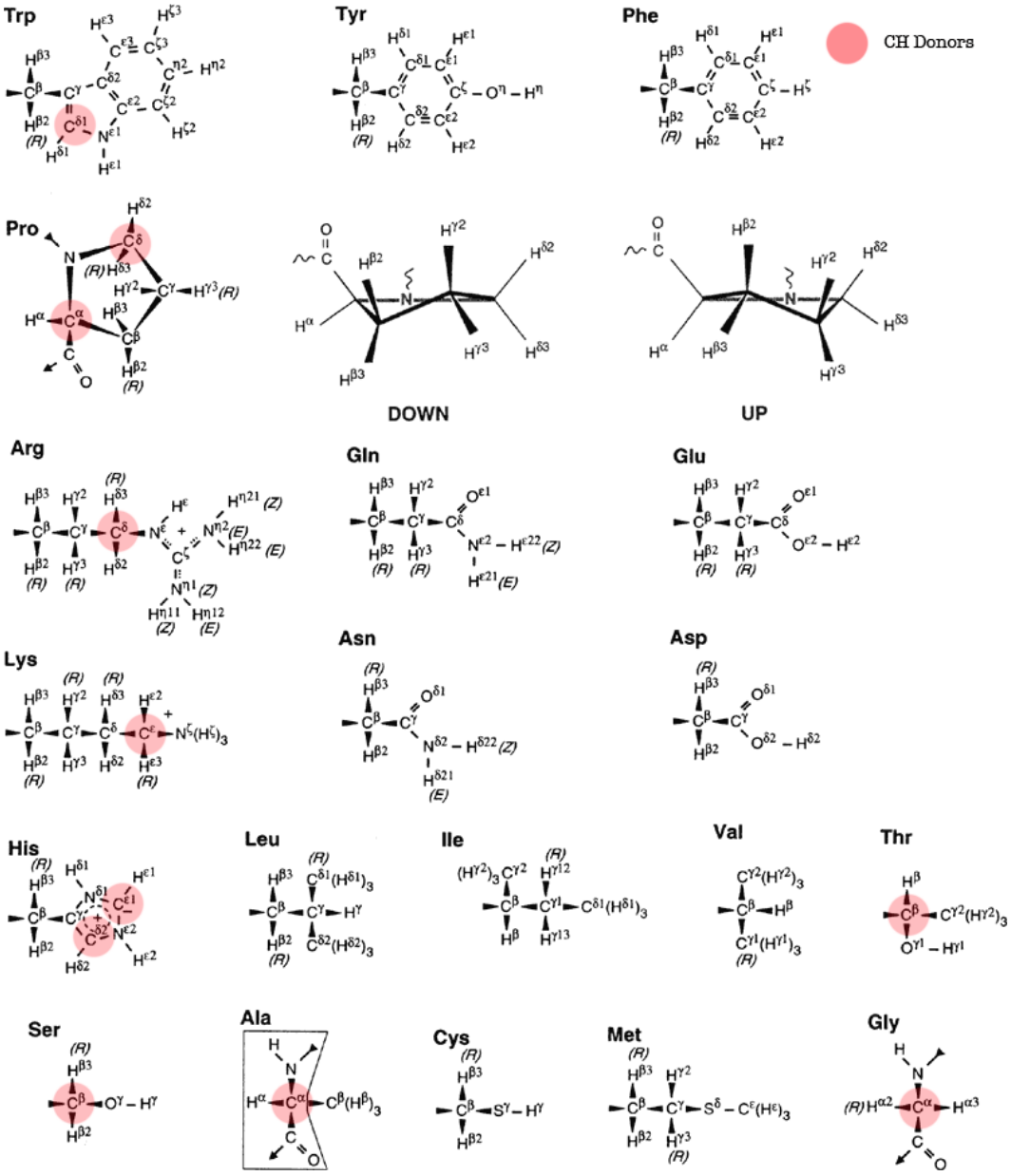


Figure 4.1 CH donor groups. Main chains are not represented, but they all donate hydrogen bonds on their C-alpha main chain carbon, as depicted for alanine. The original image is taken from the Andersen lab at Washington University [Andersen 2010].

Hydrogen bond geometry

Hydrogen bonds can be described using various features, the most common being the hydrogen acceptor distance (HA) and the donor, hydrogen, acceptor angle (DHA). But other possible geometrical features exist, as shown in Figure 4.2 below.

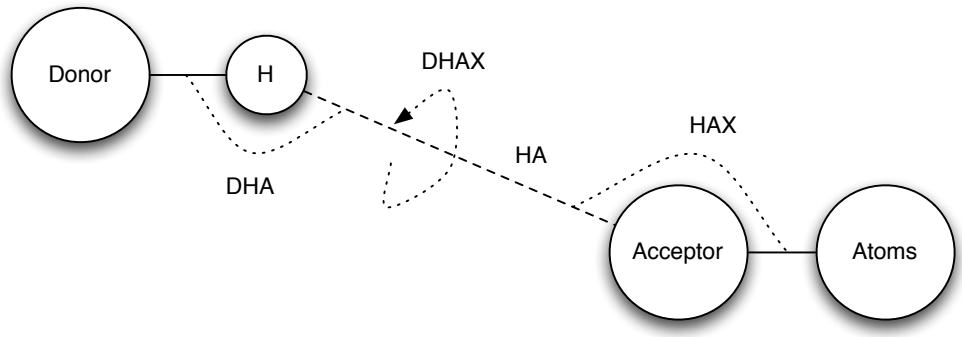


Figure 4.2 Hydrogen bond geometry. In this diagram, "Atoms" represents the weighted average position of all atoms covalently bonded to the acceptor.

As can be noted from this diagram, a hydrogen bond is expressed by 4 features, the DHA angle, the HA distance, the HAX angle and the DHAX torsion angle. For this experiment, we define a hydrogen bond as a pair of non-bonded donor and acceptor groups having a DHA angle larger than 90° and an HA distance below 2.5 Å.

4.2.3 Hydrogen generation

In this study, we generated hydrogen positions for our decoys and crystal structures using the GROMACS pdb2gmx tool. It places the atoms according to the residue topology, thus, when the hydrogen is fixed, it should accurately represent what can be seen in the experimental structure. The efficiency of this method was assessed by comparing the position of hydrogen atoms in neutron-diffracted structures to the same ones with generated hydrogens.

4.2.4 Hydrogen bonding potentials

Using statistical mechanics principles, we can approximate the functional form of an interaction using an inverse Boltzmann potential derived from a sample of protein conformations. Although theoretically based on the probability of the interaction considered in a specific range, given a large number of observations, we can approximate it using the frequency of occurrences in our training set of proteins. The more data is available, the smaller the bins can be, and the less information will be lost by the discretisation of the original probability function. This is shown in Equation 4.1, below.

$$\Delta G_h^n = -kT \ln\left(\frac{f_{obs}^n(t_h, x_h)}{f_{ref}^n(t_h, x_h)}\right) \quad (4.1)$$

In (4.1), h is the hydrogen bond being considered, n represents the feature being considered, T represents the temperature in Kelvin, k is the Boltzmann constant, x_h is the feature value, and t_h represents the type of the hydrogen bond. Feature n can be either the DHA angle, HAX angle, HA distance or DHAX dihedral. Hydrogen bond type t_h is either CH for carbon bonds, or NH for nitrogen bonds. The OH-X bond is not represented since it is impossible to position the hydrogen without making assumptions about the actual geometry of the bond, which is what we are investigating here.

As seen from our study on solvation, we can generate a reference state by taking a random model, or by using decoy structures. Thus, two kind of potentials will be derived here.

The first has a crystal structure based observed state, and a geometrical model of randomness as a reference state. These classical potentials (subscripted C in this thesis) require us to derive the reference state as a probability function modelling the randomness of the feature being considered, which is sometimes hard or impossible to do.

Our second approach, which is conformation generation method specific, uses large number of decoy structures generated by the method being explored, and uses the best decoys as the observed state, and the worse ones as the reference state (subscripted D). Here, best decoy means the decoys have an RMSD below the mean for that specific target, whilst bad decoys mean they have a larger RMSD to native.

Conventional hydrogen bonding statistical potentials assume linear separability between the various geometrical quantities observed, simply calculating a potential for each of them, and taking the weighted sum as the overall hydrogen bonding energy, as expressed in Equation 4.2 below.

$$\Delta G = \sum_{h}^{Hbonds} \sum_{n}^{Feature} w_h^n \Delta G_h^n \tag{4.2}$$

Here, w^n_h is the weight associated with specific hydrogen bonding types and features. Therefore, when optimising weights, we will optimise on 4 parameters for each bond type. Assuming linear separability might not be the most natural approach though, since variables might be clustered around certain regions, thus meaning they are in fact coupled. Assuming that each feature is independent in a random gas state, extending the Boltzmann formulation to account for multiple variables is easy, and is shown in Equation 4.3 below. We call these potentials “multivariate” potentials, as opposed to “univariate” potentials when only one term is used.

$$\Delta G_h = -kT\ ln\left(\frac{f_{obs}(t_h, X_h)}{\prod_{n}^{features} f_{ref}^{n}(t_h^n, x_h^n)} \right) \quad (4.3)$$

Here, *h* is the hydrogen bond being considered and X_h is a vector containing the values for the features being studied. The reference state is as defined in univariate potentials (Equation 4.1). Since we assumed the features are independent in a random conformation, we defined the reference state as being the joint probability of each individual feature, and express it as the product of the probability of each independent feature for that hydrogen bond. For clarity, multivariate statistical potentials are superscripted "+", while univariate (also called linear) potentials are not superscripted.

One major problem though arises from the increased number of observation bins, and thus, we either need to generate more data, use larger bins in our calculations or use pseudo-counts to prevent gaps in the potential. Here, we added a pseudo count to all bins in multivariate potentials by taking the smallest non-zero count in any observed state bin. Combining all four terms would require an enormous amount of data if a precision similar to the linear terms is desired, and thus, will not be part of our study. Here, we will derive potentials with one or two features only, and only consider two types of hydrogen bonds, (CH and NH) without distinguishing between secondary structure location or acceptor type. The weighting scheme is as previously stated, with one weight for each potential term being combined. Table 4.1 gives the bin size and range for each feature.

Table 4.1 Potentials bin size and range

Feature	Type	Bin size	Range	Bins
DHA	Angle	2°	[90, 180]°	45
HA	Distance	0.05 Å	[1.5, 2.5] Å	20
HAX	Angle	4°	[0, 180]°	45
DHAX	Dihedral	8°	[-180, 180]°	45

The reference states are defined as the probability of observing a specific range of value at random. For distances, we use the formalism of DFIRE, but with the bin values given above. For angles, we use the cone correction as the reference state, integrating over the range that we are in. This is given in Equation 4.4 below.

$$p_{ref}^{DHA}(\theta) = \left| \int_{\theta}^{\theta+\Delta\theta} \sin x \, dx \right|$$
$$p_{ref}^{DHA}(\theta) = \left| -\cos(\theta + \Delta\theta) + \cos\theta \right| \quad (4.4)$$

Here, θ is the bin floor, and $\Delta\theta$ is the bin size, with $\theta \in [90, 180]°$. The reference state for the HAX angle is defined similarly. Finally, the DHAX torsion angle has an equal probability for all bins, as no geometrical bias is present. One problem though with these reference states is that they assume the atoms in the protein behave like a gas, which is not the case here.

Table 4.2 summarises the naming conventions for the potentials derived in this study. Throughout, (NH) will refer to NH hydrogen bonds, while (CH) will refer to CH hydrogen bonds.

Table 4.2 Naming conventions for potentials

Protocol	Superscript	Subscript	Observed State	Reference State
Statistical		C	Crystal Structures	Geometrical
Statistical		D	Better than average decoys	Worst than average decoys
Multivariate	+	C	Crystal Structures	Geometrical
Multivariate	+	D	Better than average decoys	Worst than average decoys

4.2.5 Performance measures

Here, as before, we are interested mostly in the discriminatory capabilities of our potential, and as such, we will put emphasis on the 10% enrichment score when analysing intermediate results. Final results will be given for 10 and 15% enrichment scores, Pearson R and Kendall Tau. The performance of potentials is assessed by comparing the energy function values to the All-Atom RMSD.

4.2.6 Combining energy terms

As before, we combined energy terms using R and the *optim* function available in the *stats* library to derive the weights introduced earlier. We optimised the average 10% enrichment score over all targets in the decoy set, using the Nelder-Mead optimization protocol, and calculated the other statistics using the weights derived from it.

4.3 Results & Discussions

In this section, we will review the results of our experiments on hydrogen bonds. First, we will analyse the distributions of the different features for both NH and CH hydrogen bonds. Using these results as well as geometrical considerations, we will derive both linear and multivariate potentials, using the different protocols introduced in the methods section. The results of the best terms will then be compared to DFIRE2 to assess whether it could help increase its performance for decoy discrimination, and whether CH hydrogen bonds are useful in doing so.

4.3.1 Accuracy of the hydrogen atoms generation method

To avoid biasing our results towards specific potentials, hydrogen positions were created geometrically using GROMACS, but without running molecular dynamics. To assess the efficiency of this method, we have removed the hydrogens atoms from a set of 13 neutron resolved structures, and regenerated them to compare their position. The RMSD of the “real” to the virtual hydrogen position are given for geometrically fixed and non-fixed atoms in Table 4.3.

Table 4.3 RMSD of real vs. virtual hydrogens

Protein	Fixed H	Rotating H
1G66	0.59	1.36
1HJE	0.10	1.44
1IXH	0.53	1.48
1L9L	0.38	1.36
1MUW	0.49	1.32
1RTQ	0.43	1.34
1TT8	0.43	1.36
1UCS	0.39	1.40
2B97	0.46	1.37
2BF9	0.12	1.35
2ERL	0.40	1.45
2FDN	0.45	1.51
2VB1	0.11	1.35
Average	**0.37**	**1.39**

We can see in the table above that although the fixed hydrogens are modelled well, with an average RMSD of 0.37 Å, the rotating hydrogens are far less reliable and are on average 1.39 Å RMSD away from their experimental position, as seen in neutron diffracted structures. Therefore, we have decided not to include rotating hydrogens in our studies.

4.3.2 Univariate statistical potentials

Conventionally, only OH and NH hydrogen bonds have been considered in potentials used to discriminate decoy structures. There are different ways to represent hydrogen bonds and assess their energy, most of them being through electrostatic interactions, empirical distance and angle potentials, or statistical potentials derived from known crystal structures. In this last instance, aside from some rare cases, the hydrogen positions are not resolved, and thus, one has to either fix them from geometry, or make assumptions about their position. Considering that we are trying to model an interaction, we want to minimise the number of assumptions we make about it, and thus, we only include bonds that have a fixed geometry.

Hydrogen bonds are conventionally defined by their HA and DHA distance, but in this study we looked at two other features as well: the HAX angle, which is the angle between the hydrogen atom from the donor group, the heavy atom from the acceptor group, and the centroid of the atoms covalently bonded to the acceptor [Kortemme et al. 2003] and the DHAX torsion angle representing the planarity of the hydrogen bond.

In order to derive statistical potentials, we took a set of high-resolution crystal structures from the PDB, and generated decoy structures using molecular dynamics. We then generated two types of potentials for each of the four features (HA, DHA, HAX and DHAX). The first is a classical potential where the observed state is modelled from the crystal structures, and the reference state is taken from the expected frequencies in a random state. The second potential is based only on decoys, and attempts to extract differences in the distributions of better than average and worse than average structures. By better than average, we mean that the RMSD to native is less than the average value, and vice versa for worse than average decoys. Therefore, our observed state will be derived from those good decoys, while our reference state will be derived from the bad ones. The distributions of values for hydrogen bonds in crystal structures are shown in Figure 4.3 for the NH hydrogen bonds, and Figure 4.4 for CH hydrogen bonds.

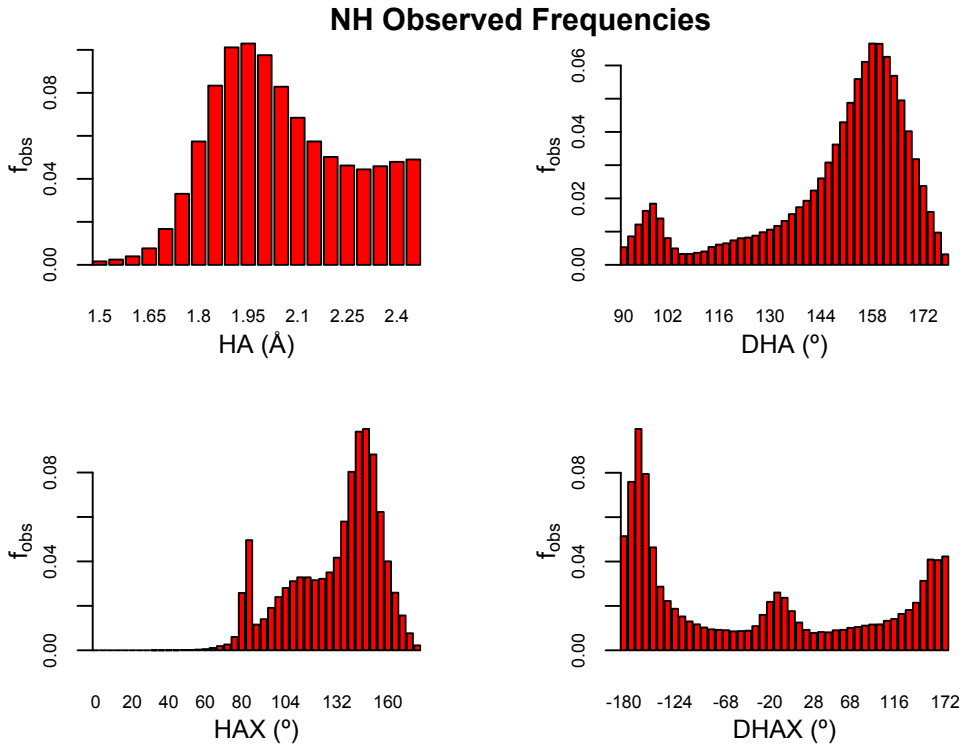


Figure 4.3 NH bonds classical observed states. These distributions were generated from the crystal structures, and only show hydrogen bonds with fixed hydrogen positions.

These distributions for the NH bonds show strong preferences for specific values. For the HA distance, the distribution peaks at 1.9 Å, and remains relatively flat above 2.2 Å. The DHA angle peaks in two places, the main one being at 160º, and the second one, much smaller, at 95º. The other angle, HAX, also peaks in two places, with the main one at 145º, and the second one, much sharper, at 85º. Finally, the DHAX torsion angle peaks in two places, at -160º and 0º.

The same features were calculated for the CH bonds, and are shown in Figure 4.4.

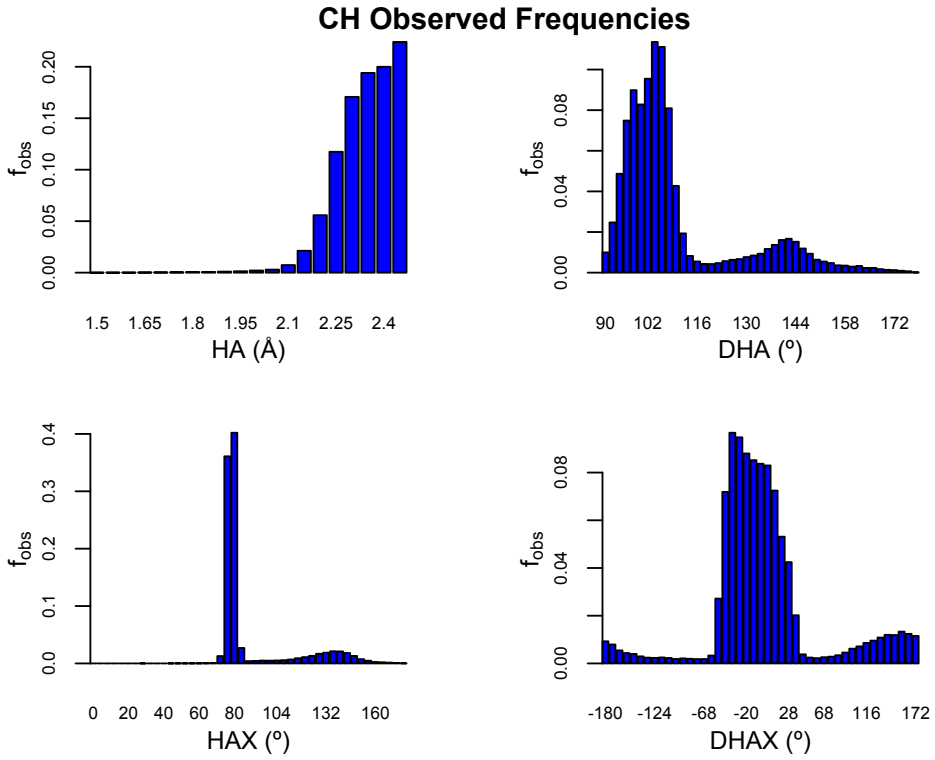


Figure 4.4 CH bonds classical observed states. Some other studies have included bonds with HA distances above 2.5 Å, but we have chosen not to, as there is little evidence that hydrogen bonds exist at such distances.

Here, we can observe that the HA distance do not peak at a specific value, but instead increases and becomes maximal at the boundary. The DHA angle peaks in two places, the main one at 105º, and the other one at 140º. As for the HAX angle, it has a very sharp peak at 80º, followed by a smaller one at 140º. These values are similar to the ones observed in NH bonds, but with different relative frequencies. Finally, the DHAX dihedral is mainly around 0º, with a small peak around 160º.

Overall, both NH and CH hydrogen bonds have preferred values, although the HA distance for the CH bonds are not peaking in the range that we consider in this analysis. Using these different distributions as the observed state of our classical potentials, and using the distributions taken from good and bad decoys, we have generated 16 statistical potentials, one for each feature (HA, DHA, HAX,

DHAX), for each bond type (NH, CH) and for each generation protocol (Classical, Decoy-Based). The energies are shown in Figure 4.5 for NH bonds, and 4.6 for CH bonds.

The potentials displayed on the left hand side of Figure 4.5 and Figure 4.6 are the ones derived from crystal structures, and are similar to those used in the literature. In them, the HA has its lowest energy at 1.9 Å, the DHA at 160º, the HAX at 150º, and the DHAX at -170º. On the other hand, the decoy-based potentials on the right have quite different distributions, and much smaller energies, implying a lesser difference in the distributions of the good and bad decoys. This is expected since all decoys are close to native, and thus, would show subtle differences. To interpret these potentials, one needs to think of them as representing the differences between the good and bad decoys, rather than the difference between real interactions and random ones. Thus, when the energy is positive, there are fewer good decoys found in that region compared to bad decoys, and vice versa for negative energies. In essence, the lower the energy is of a given bin, the larger the proportion of good to bad decoys that will be in that bin.

In NH decoy-based potentials, the HA term strongly disfavours very short distances. The DHA favours 120º angles, as well as linear ones, while disfavouring those below 120º. As for the HAX angle, it is less and less unfavourable until 125º, before becoming favourable around 150º. Finally, the DHAX term favours planar angles, and strongly disfavours those around -60º.

CH hydrogen bond potentials shown in Figure 4.6 are significantly different from the NH ones seen previously. The first difference is for the HA distance energy, where none of the values in the range considered are favourable. This means that a protein would not spontaneously form such hydrogen bonds from an unfolded state. This is observed both for the classical potential, and for the decoy-based one.

The classical DHA potential shows a favourable region between 92º and 112º, and becomes unfavourable for distance above that, with a significant decrease in unfavourableness centred around 140º. The decoy-based DHA potential shows unfavourableness for angles between 100º and 115º, approximately in the same region that the classical potential was favourable. This means that bad decoys are found at the lowest energies of this classical potential, but good decoys aren’t.

The classical HAX angle potential troughs sharply around 77º, while being decreasingly unfavourable as the angle approaches linearity. On the other hand, the decoy-based potential only shows unfavourableness for angles below 75º, and remains relatively flat and negative for angles above.

Finally, the DHAX dihedral has a preference for a wide region between -50º and 50º, and decreasingly becomes unfavourable as planarity is approached. For the decoy-based term, the potential is unfavourable around -60º, and is slightly favourable below -60º and above 50º.

Considering that each of the potentials shown here have very different distributions, we will compare each of them, and combine them to create a full hydrogen bonding potential that will be applicable to the discrimination of near-native decoys. In order to do so, we have selected 50 random targets from the MDSET decoy set, and calculated the 10% enrichment score for each of them. The average enrichment score is then taken as the statistic representing the ability of a potential to correctly identify the best 10% models.

To test the validity of our results, we used the Wilcoxon signed rank test to compare the average score against the random score of 0.10. When combining potentials and comparing them to DFIRE2, or when comparing potentials against each other, we will be using a two-tailed Wilcoxon rank sum test. In both cases, the W statistic and the p-value will be shown. Table 4.4 shows the result for independent potential terms.

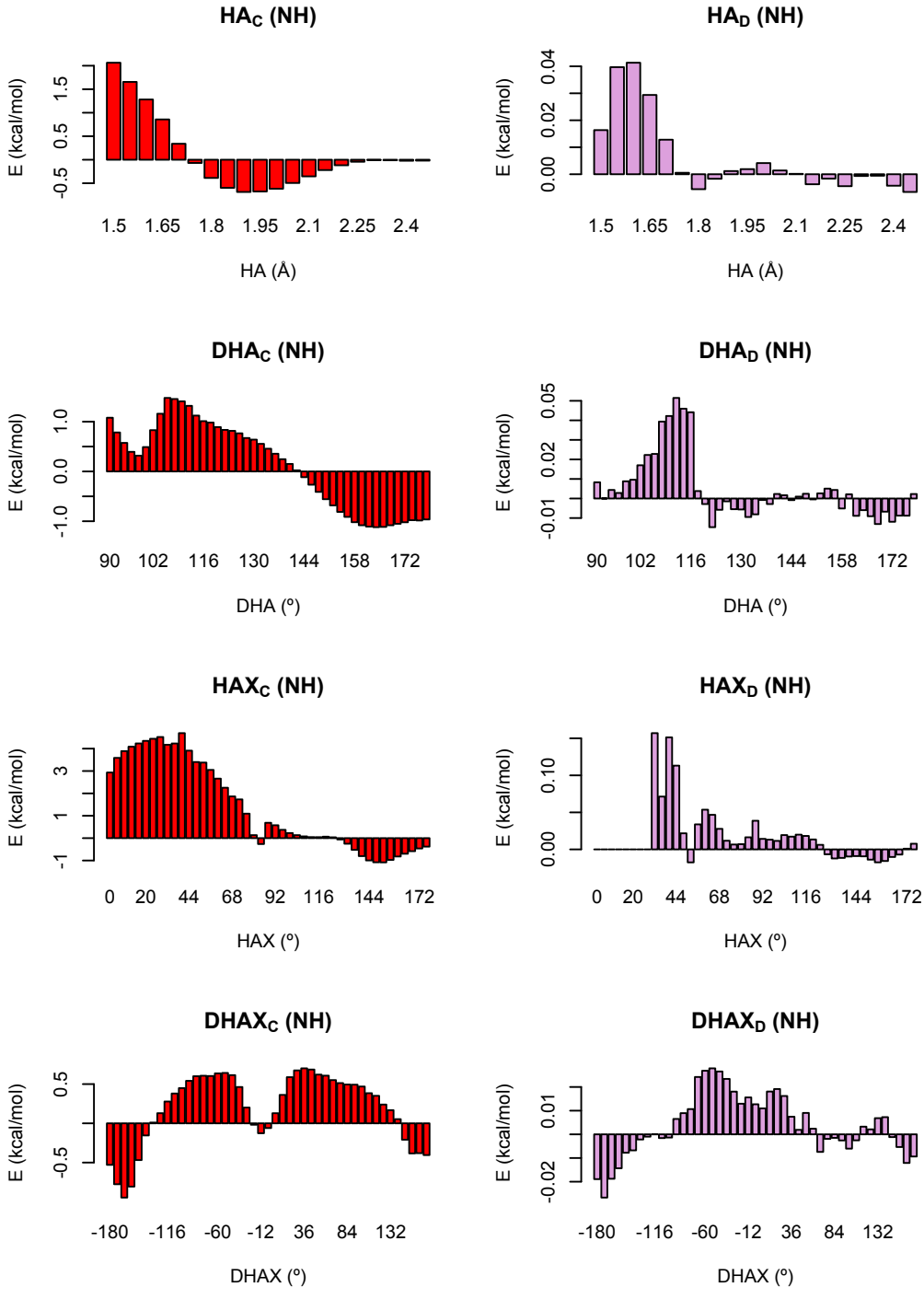


Figure 4.5 NH bonds potentials. The subscript "C" on the left plots represents the classical potentials derived from crystal structures, while the "D" subscript on the right ones represents potentials derived from decoys. These histograms only include fixed hydrogen bonds.

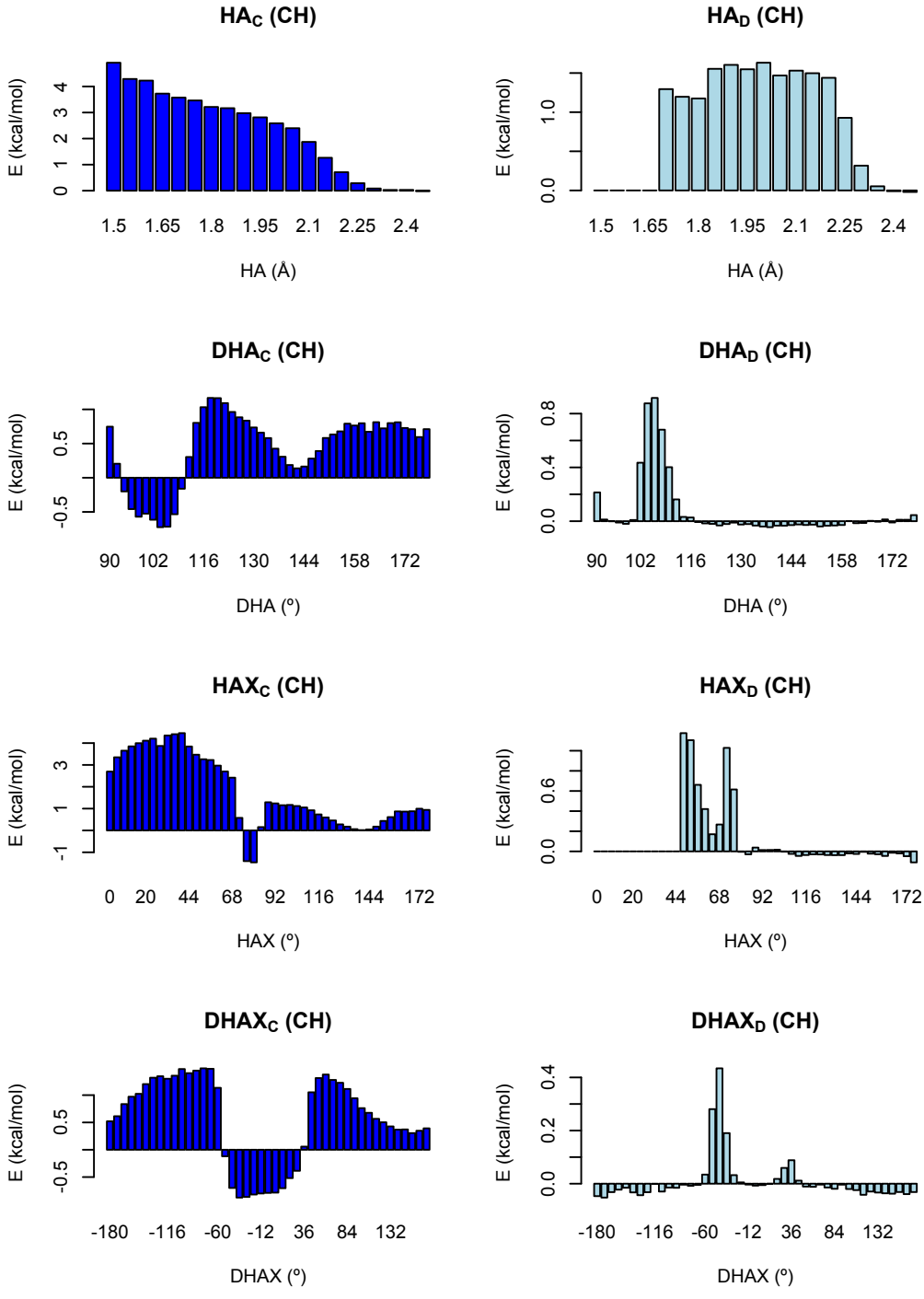


Figure 4.6 CH bonds potentials. Here, potentials shown on the left are derived from crystal structures, while those on the right are derived from decoys alone. These histograms only include fixed hydrogen bonds.

Table 4.4 10% enrichment score for univariate potential terms

Potential	10% enrichment	W	p-value (H_0=0.10)
Classical			
HA_C (NH)	0.18	1017	1e-05
DHA_C (NH)	0.17	1274	5e-07
HAX_C (NH)	0.19	1238	7e-9
$DHAX_C$ (NH)	0.21	1400	1e-8
HA_C (CH)	0.09	264	0.04
DHA_C (CH)	0.08	413	0.01
HAX_C (CH)	0.09	519	0.18
$DHAX_C$ (CH)	0.09	440	0.03
Decoy-based			
HA_D (NH)	0.13	788	0.04
DHA_D (NH)	0.16	1219	8e-6
HAX_D (NH)	0.21	1399	2e-8
$DHAX_D$ (NH)	0.22	1522	3e-10
HA_D (CH)	0.09	348	0.01
DHA_D (CH)	0.12	942	0.08
HAX_D (CH)	0.14	1013	0.0003
$DHAX_D$ (CH)	0.13	1040	0.0003

We can see that all classical NH potentials, as well as the decoy based NH potential are significantly better than random (random is 0.10 for the 10% enrichment). The best performing ones are the classical NH potentials, and the decoy-based NH HAX and DHAX terms. The classical CH terms are all worse than random, while the decoy based ones are slightly better than random for the HAX and DHAX angles.

In order to test the usefulness of including CH bonds in full potential, we have computed 4 different potentials: An NH classical potential, combining all four classical NH term, an NH+CH classical potential, combining all classical NH and CH terms, and two similar combinations of decoy-based terms. Table 4.5 shows the results, with the Wilcoxon (W) statistic and the p-value of the difference to the corresponding NH potential alone.

Table 4.5 10% Enrichment score for univariate potential combinations

Potential	10% enrichment	W	p-value (H_0=NH_X)
NH_C	0.27		
NH_C + CH_C	0.27	1559	0.48
NH_D	0.23		
NH_D + CH_D	0.24	1524	0.40

It follows from this analysis that including CH hydrogen bonds in univariate NH potentials do not improve at all the enrichment score. Thus, we can conclude that linear combinations of CH terms are not useful for discriminating near-native decoy structures.

4.3.3 Bivariate statistical potentials

We have seen that using CH hydrogen bonding terms in a linearly combined hydrogen bond potential did not significantly improve its discriminatory capabilities. In this experiment, we will analyse the effect of grouping potential terms 2 by 2, in order to capture any coupling between them.

First, we will analyse the distribution of each potential terms combination, both for NH and CH hydrogen bonds. We will then combine them, and compare these full potentials to those derived previously. Finally, we will assess the usefulness of including CH hydrogen bonds in these bivariate potentials.

Although the coupling between the HA distance and DHA angle has been studied in the literature [McDonald & Thornton 1994], little emphasis has been put on the coupling between the other 2 terms, namely the HAX angle, and the DHAX torsion. Thus, we start by generating the data points for each pair of features, giving us 6 different distributions: HA-DHA, HA-HAX, HA-DHAX, DHA-HAX, DHA-DHAX, and HAX-DHAX. These are shown for NH hydrogen bonds in Figure 4.7.

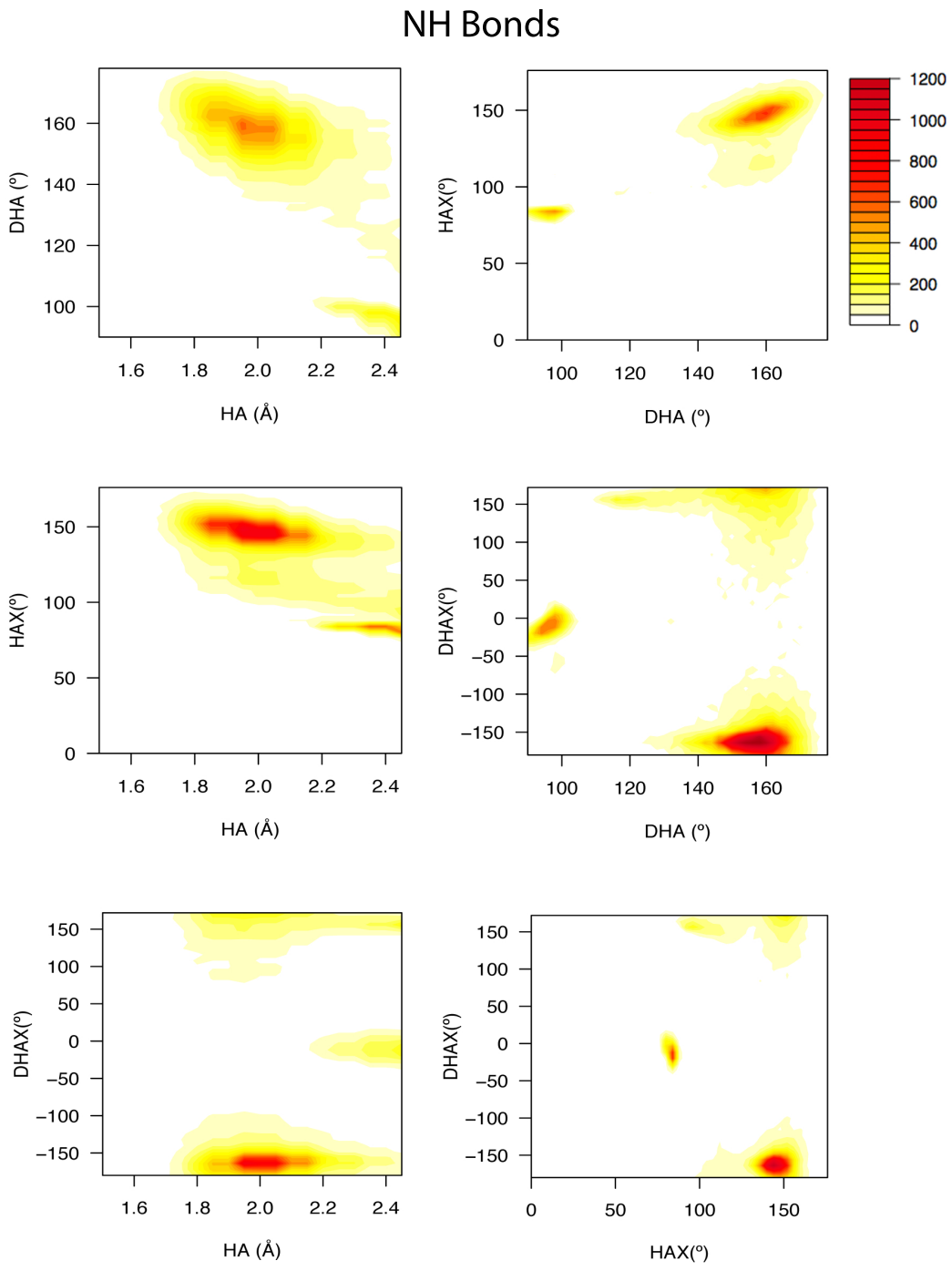


Figure 4.7 NH paired terms distributions The level plots show the counts using heat colours, white being no counts, and red being maximal count.

It is apparent from Figure 4.7 that the terms have a largely preferred region in each case (yellow to red regions). The centres of such regions are given for each pair in Table 4.6 below.

We can also observe secondary peaks, at HA = 2.4 Å and DHA = 100º, HAX = 80º and DHAX = 0º, which are occurring in alpha helices between the i^{th} and i^{th}+4 residues. Given that we overrepresented alpha helices compared to their frequency in nature, we observe more of these interactions that we would expect.

Table 4.6 Preferred regions centres of pairwise features of NH bonds

Feature pair	Centre value (A, B)
HA, DHA	2 Å, 160º
HA, HAX	2 Å, 145º
HA, DHAX	2 Å, -160º
DHA, HAX	160º, 145º
DHA, DHAX	160º, -160º
HAX, DHAX	145º, -160º

The fact that each pair of terms has such strong preference for specific regions means that by treating each of them independently, we might end up loosing information in our potentials, which in turn would decrease the probability of detecting defects in decoys. We will compare the linear and multivariate forms of our hydrogen bonding potential, but first, we will analyse the coupling of terms in CH hydrogen bonds. Using the same protocol, bin sizes and number of samples, we have generated similar level plots for CH hydrogen bonds. The plots are shown in Figure 4.8, next.

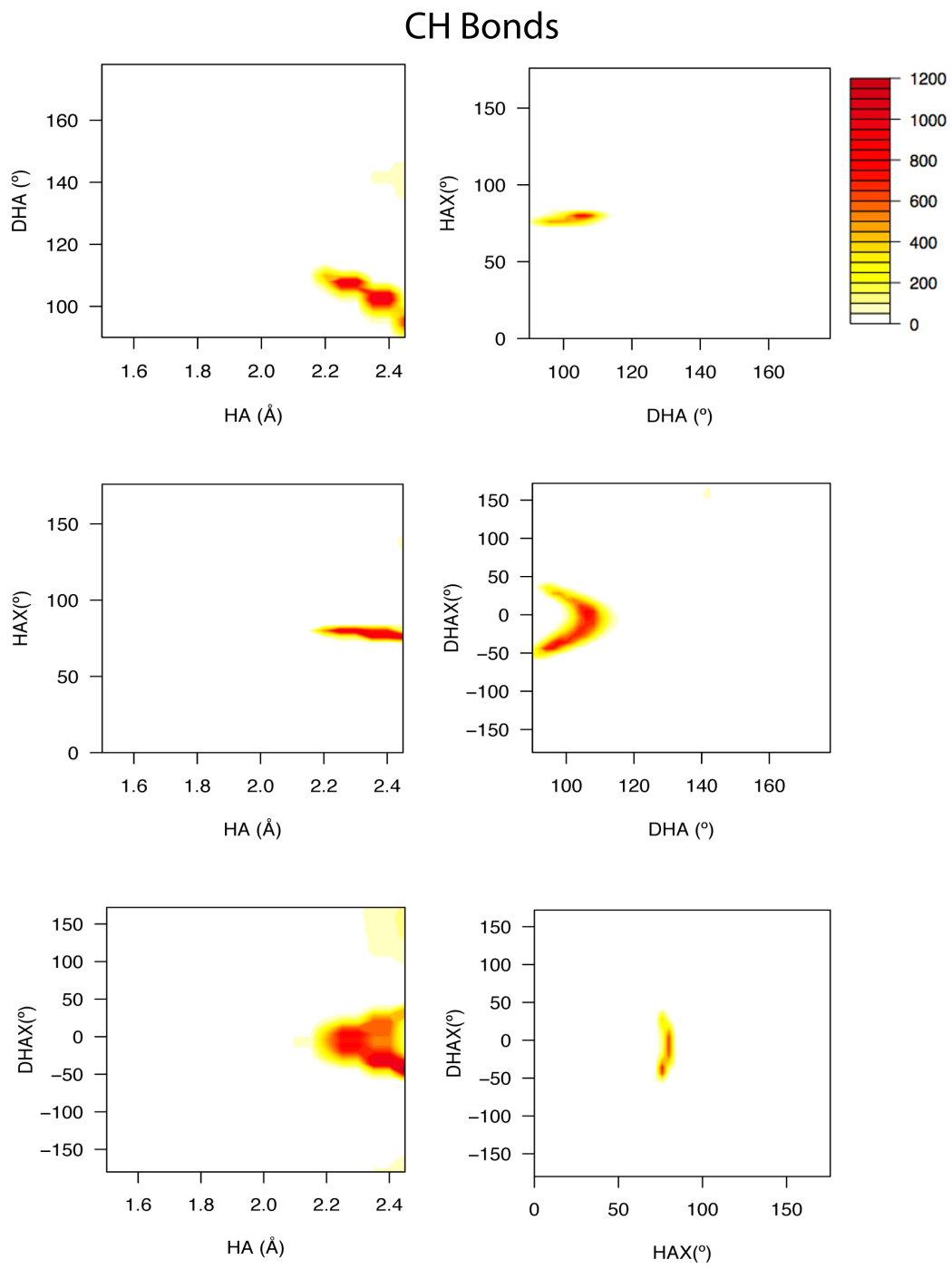


Figure 4.8 CH paired terms distributions The level plots show the counts using heat colours, white being no counts, and red being maximal count.

The patterns showing for the CH hydrogen bonds are very different from those observed for NH hydrogen bonds, which have less concentrated regions in the pairwise plots. First, we can see that the HA-DHA pair shows 2 different clusters, one around 2.25 Å and the other around 2.4 Å. A more detailed analysis showed that the 2.4 Å cluster corresponds to main chain CA carbon atoms, while the other one correspond to side chain CH hydrogen bonds. This difference is most likely due to the strong neighbouring NH hydrogen bond on the main chain, which would then drive the position of the CH bond when forming secondary structures. The centres of the main clusters are given in Table 4.7 below.

Table 4.7 Preferred regions centres of pairwise features of CH bonds

Feature pair	Centre value (A, B)
HA, DHA	2.35 Å, 105º
HA, HAX	2.35 Å, 80º
HA, DHAX	2.35 Å, 0º
DHA, HAX	105º, 80º
DHA, DHAX	105º, 0º
HAX, DHAX	80º, 0º

Using the same protocol as for univariate potentials, we have derived a statistical potential for each of the pairs, for both CH and NH bonds. The level plots of the classical potentials are shown in Figure 4.9 for the NH bonds, and 4.10 for the CH bonds.

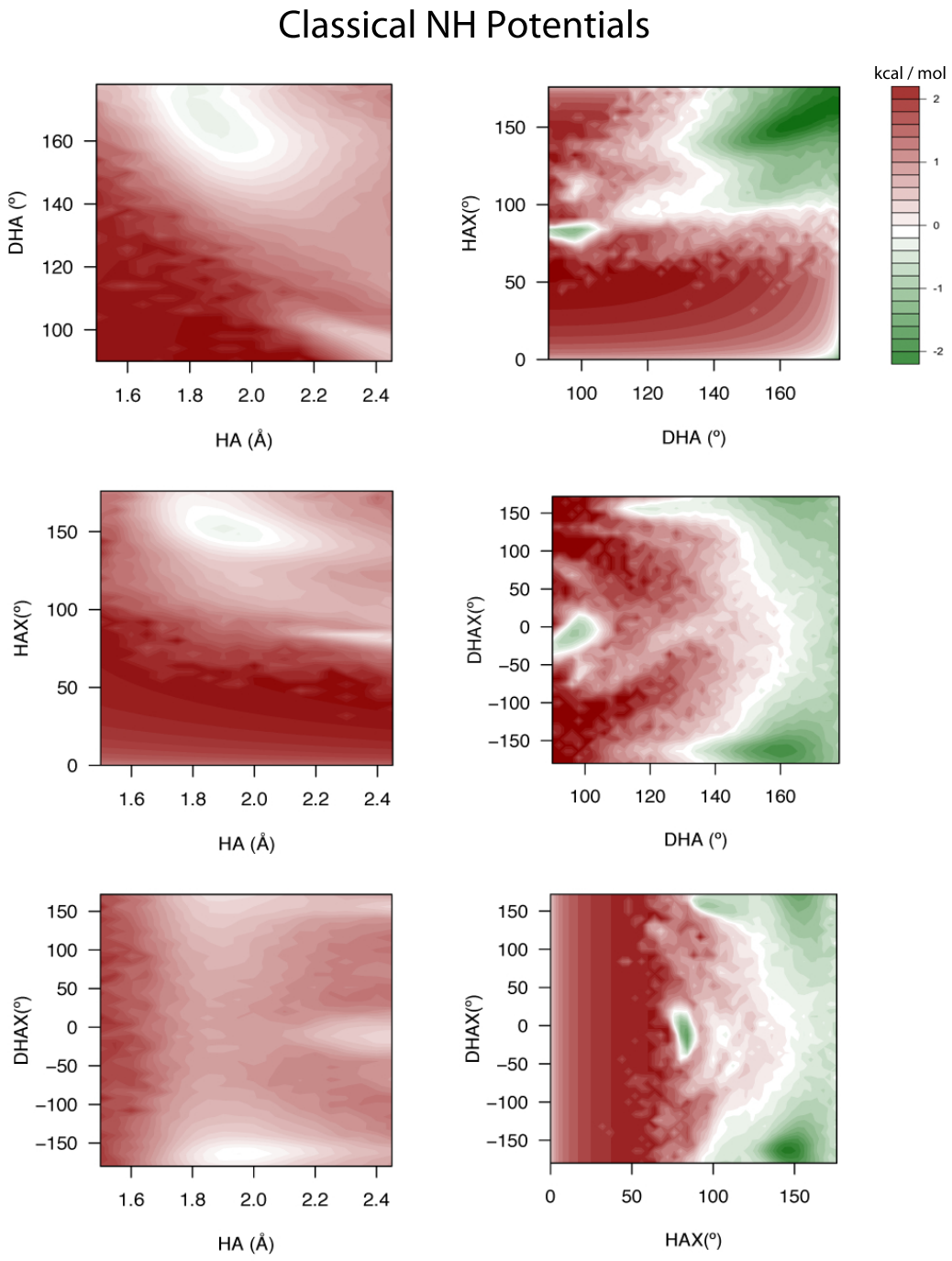


Figure 4.9 NH classical bivariate potentials. The level plots show the energies of the bivariate potentials, ranging from green for favourable energies, to red for unfavourable ones. The darker the colour, the more extreme the energy is on the scale. The same scale has been used for all plots.

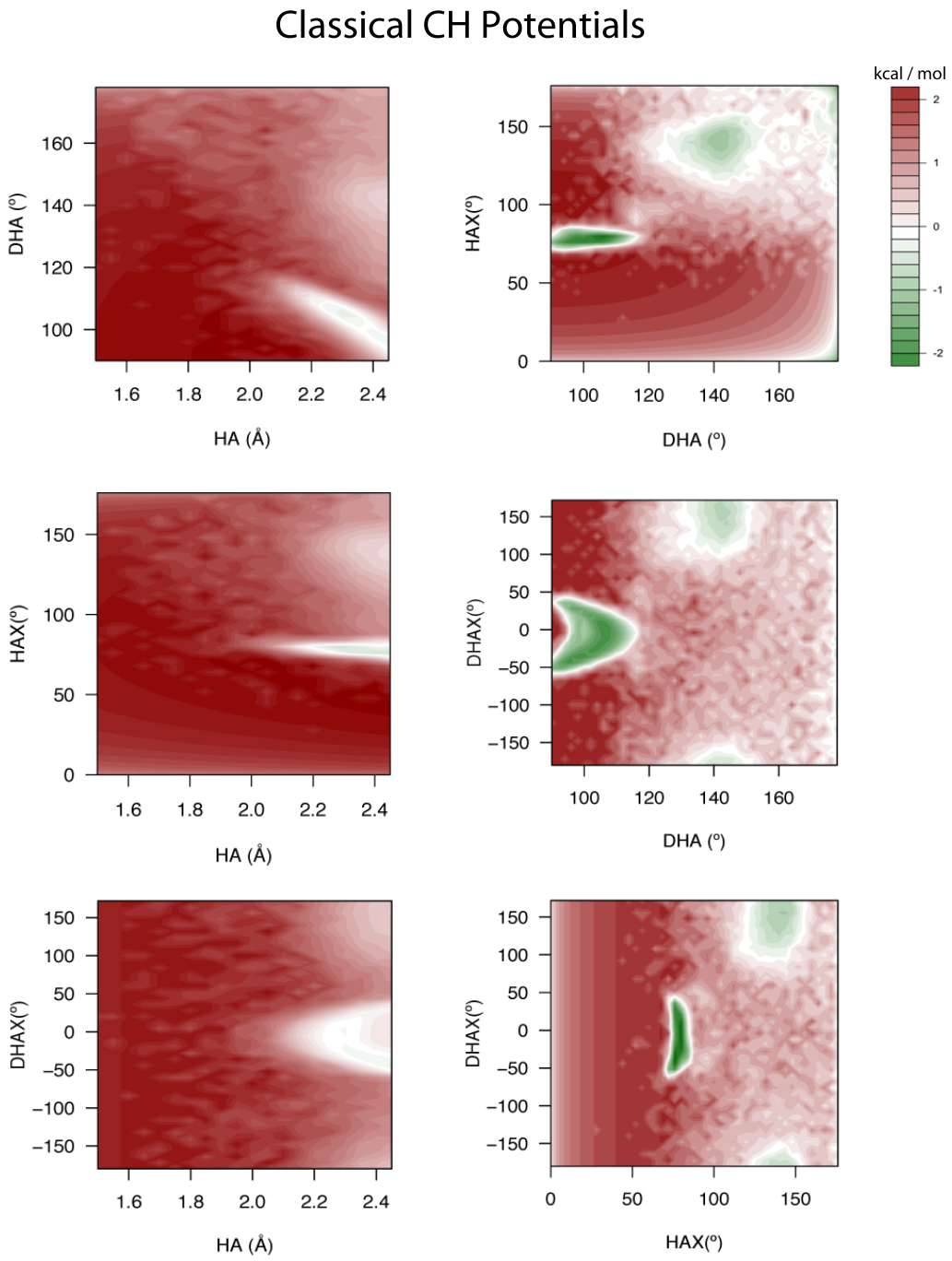


Figure 4.10 CH classical bivariate potentials. The same energy scale has been used as for NH bonds.

There are several interesting patterns emerging from the NH classical potentials. Namely, the terms including the HA distance do not exhibit negative energies as low as terms not including it. This can be explained by the different reference state used, where the sum of the reference state in all bins does not equal 1 (we used the DFIRE reference state for the HA distance). Nonetheless, we can observe preferred regions at HA = 1.9 Å, DHA = 170º, HAX = 155º and DHAX = -170º. For the pure angle terms, we can observe a favourable region in every case, clearly showing the contrast of energies for different values. The DHA-HAX term is more favourable as both DHA and HAX approach linearity. The DHA-DHAX term has a preferred region for linear DHA angles and planar DHAX torsions, but still is favourable for all DHA angles above 160º. A similar pattern is observed for the DHAX-HAX term, with a preference for planar torsions, and HAX angles above 140º.

The classical CH bonds on the other hand show different patterns, with very faint favourable regions in terms involving the HA distance. For angle terms though, we can see several clusters. In the DHA-HAX potential, we can see 3 clusters: one at DHA=100º and HAX=80º, one at DHA=140º and HAX=140º, and one at DHA=180º and HAX=0º or HAX=180º. The last two clusters are very interesting, as they do not show favourableness in univariate potentials introduced earlier. This shows how coupling features can impact the overall potentials. The DHA-DHAX term shows preference for DHA=100º and DHAX=0º, but also for DHA=140º, and DHAX = ±180º. Again, this second cluster does not show in univariate potentials. Finally, the HAX-DHAX term show two clusters, one at HAX=80º and DHAX=0º, and one at HAX=140º and DHAX = ± 180º.

From this analysis, we can see that there is clearly two interactions being captured in classical CH bivariate potentials, while only one is accurately represented in the univariate potentials. This could potentially impact the results of decoy discrimination, which we will test now. Results for the 10% enrichment is shown in Table 4.8 below.

Table 4.8 10% enrichment score for bivariate potential terms

Potential	10% enrichment	W	p-value (H_0=0.10)
Classical			
HA-DHA$_C$ (NH)	0.19	1277	8e-8
HA-HAX$_C$ (NH)	0.24	1393	2e-9
HA-DHAX$_C$ (NH)	0.21	1340	3e-9
DHA-HAX$_C$ (NH)	0.23	1389	2e-9
DHA-DHAX$_C$ (NH)	0.24	1398	2e-9
HAX-DHAX$_C$ (NH)	0.23	1466	5e-9
HA-DHA$_C$ (CH)	0.09	486	0.15
HA-HAX$_C$ (CH)	0.09	519	0.12
HA-DHAX$_C$ (CH)	0.09	365	0.11
DHA-HAX$_C$ (CH)	0.10	593	0.51
DHA-DHAX$_C$ (CH)	0.10	490	0.58
HAX-DHAX$_C$ (CH)	0.10	575	0.71
Decoy-based			
HA-DHA$_D$ (NH)	0.17	1274	5e-6
HA-HAX$_D$ (NH)	0.22	1458	7e-10
HA-DHAX$_D$ (NH)	0.18	1374	4e-10
DHA-HAX$_D$ (NH)	0.18	1161	4e-7
DHA-DHAX$_D$ (NH)	0.17	1162	5e-8
HAX-DHAX$_D$ (NH)	0.21	1479	2e-10
HA-DHA$_D$ (CH)	0.11	613	0.42
HA-HAX$_D$ (CH)	0.12	772	0.11
HA-DHAX$_D$ (CH)	0.11	580	0.66
DHA-HAX$_D$ (CH)	0.12	759	0.06
DHA-DHAX$_D$ (CH)	0.12	587	0.09
HAX-DHAX$_D$ (CH)	0.11	711	0.20

Results in Table 4.8 show that all NH terms are significantly better than random, while most CH terms are either worse, or not very much better than random. In fact, NH terms largely outperforms CH terms, both for classical and decoy based potentials.

The second observation is that the scores for NH terms are generally better than the univariate scores, independently of the terms chosen. In order to compare the bivariate and univariate potentials, we need to generate full potentials. Results for NH and NH+CH potentials are shown in Table 4.9 below.

Table 4.9 10% Enrichment score for bivariate potential combinations

Potential	10% enrichment	W	p-value (H_0=NH_X^+)
NH_C^+	0.27		
NH_C^+ + CH_C^+	0.28	1480	0.31
NH_D^+	0.27		
NH_D^+ + CH_D^+	0.27	1553	0.47

We can observe no significant improvement after adding CH hydrogen bonding potentials compared to NH only potentials, regardless of the protocol used for deriving them. Thus, we can conclude that, as for univariate potentials, adding CH bonds do not improve the discriminatory power of our potential when applied to near native decoys.

In order to decide which of the univariate or bivariate version of our potential to use, we have compared the NH full potentials derived using the univariate terms, versus the bivariate ones. The comparison was made between classical potentials, and separately between decoy-based potentials. Results are shown in Table 4.10.

Table 4.10 Comparison of univariate and bivariate NH potentials

Potential	10% enrichment	W	p-value (H_0=NH_X)
NH_C	0.27		
NH_C^+	0.27	1533	0.42
NH_D	0.23		
NH_D^+	0.27	1150	0.005

We can see that there is no significant difference between univariate and bivariate potentials when derived using a classical reference state. On the other hand, using bivariate terms does improve the decoy-based term, going from 0.23 to 0.27, with a p-value for that difference of 0.005. Thus, we can conclude that bivariate potentials are at least as good as univariate ones, but can also outperform them in specific cases.

4.3.4 Inclusion in the DFIRE2 potential

The DFIRE2 potential is a heavy atom pairwise distance potential including 3 orientation dependent terms. Since it does not explicitly model hydrogen atoms, it fails to capture the contribution of hydrogen bonds to the total free energy of the protein. Thus, adding a hydrogen bonding term might in principle increase the precision of DFIRE2. The results are shown in Table 4.11 below, with the one-tailed Wilcoxon rank sum test for the difference to DFIRE2.

Table 4.11 10% enrichment of DFIRE2 with and without hydrogen bonds

Potential	10% enrichment	W	p-value (H_0=DFIRE2)
DFIRE2	0.41		
DFIRE2 + NH_C	0.42	1521	0.38
DFIRE2 + NH_C^+	0.43	1463	0.28
DFIRE2 + NH_D	0.42	1530	0.41
DFIRE2 + NH_D^+	0.43	1463	0.28

After adding the hydrogen bonding terms to DFIRE2, we could observe no significant difference in the 10% enrichment score, regardless of the type of potential used. Therefore, we can conclude that hydrogen bonding, including or excluding NH bonds, does not significantly improve the existing DFIRE2 potential.

4.4 Conclusions

In this chapter, we have studied various aspects related to hydrogen bonding potentials, and more specifically to the inclusion of CH hydrogen bonds. We have seen that although showing preference for specific regions, there was no improvement when deriving potentials using CH and NH bonds, compared to using only NH hydrogen bonds.

In our first study, we looked at individual NH and CH terms, which all showed specific peaks in the distribution of observations, and in their potentials. Thus, hydrogen bonds were successfully detected by our protocol, but still failed to improve DFIRE2, implying that they did not add any information to it. When taking a closer look at DFIRE2, we see it includes distance and angles components for each pair of heavy atoms. As we are only considering fixed hydrogen position, these would implicitly be modelled from their covalently bonded heavy atoms, and as such would be partially represented in DFIRE2. Moreover, we are only interested in near-native decoys, which are by definition very well formed structures, with many native-like properties, including their hydrogen bonding pattern. The precision gain from adding an explicit hydrogen bonding term, with or without CH bonds, and derived from crystal structures or decoys, would therefore not be significant, as it would be adding redundant information on little varying features.

This redundancy was partially tested by using bivariate potentials, where we paired features describing hydrogen bonds. Since considering each feature separately would be an approximation of reality, by using coupled terms, we should, in theory, observe subtler patterns that would be lost otherwise. Given the relatively small sample size in our crystal structure, we could not couple more than 2 features together, as we would end up having a combinatorial explosion, and many empty bins in our potentials. The analysis of the level plots for the combined terms revealed regions of coupled preference that did not show in the univariate potentials. This was both observed for NH and CH hydrogen bonds, for which the potentials outperformed or did as good as their univariate counterparts. But despite this, neither the classically derived or decoy-based potentials showed a significant improvement over DFIRE2.

This analysis led us to conclude that NH+CH hydrogen bonding potentials are not useful at discriminating near-native decoys compared to pure NH potentials, both when deriving them using decoys or using crystal structures. Moreover, the

information gained from pairwise coupling of features still did not add enough information to compensate for the small variability in the hydrogen bonding patterns of our near-native decoys.

Chapter 5

Deriving distance, orientation-dependent atomic potentials

5.1 Introduction

Conventionally, atomic distance potentials have considered mostly long-range, omnidirectional atomic interactions, with a cutoff distance at 15 Å. Energy functions such as DFIRE [Zhou & Zhou 2002] even increase the bin width as the distance decreases, effectively preventing more detailed modelling of short-range regions. DFIRE2 (and in its first version, dDFIRE), is a significant improvement over DFIRE, as it both models more bins at short distances, but also includes 3 angle terms to account for the orientation dependency of interacting polar atoms. The decoy discriminating performances increased after including the orientation terms, as shown in Figure 5.1.

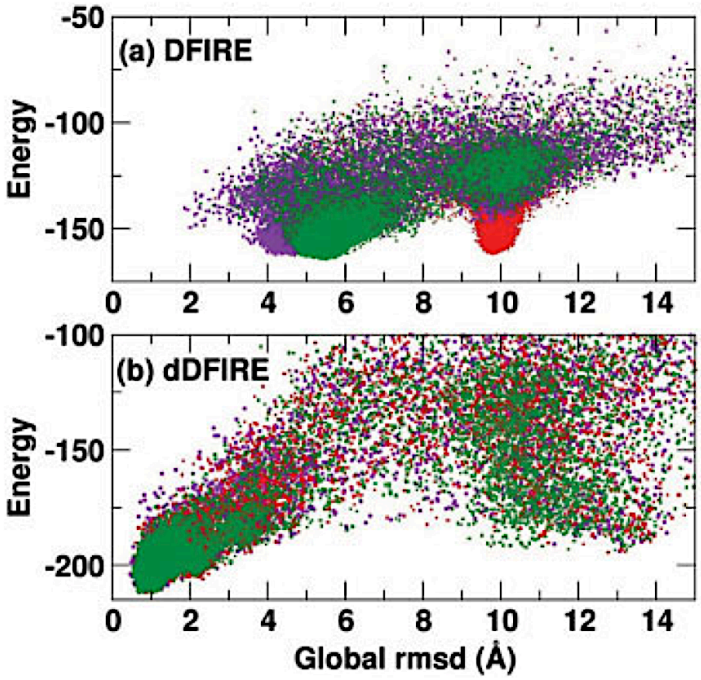


Figure 5.1 DFIRE versus dDFIRE [Yang & Zhou 2008]. This plot shows that dDFIRE performs better for decoys less than 6 Å away from the native structure, but becomes random for the remaining.

There are limitations to this approach though, as the DFIRE2 potential assumes that only polar atoms are orientation dependent. Moreover, it does not include

hydrogen atoms, and as such, can only model heavy atoms, which might not be sufficient for near-native decoy discrimination.

DFIRE2 also uses a 4 dimensional energy function, which for hydrogen bonds in chapter 4 did not provide any significant improvement over a simpler linear combination of each term. For DFIRE2, this combinatorial explosion led to a simplification of the potential, both by using a smaller number of bins (only 6 per angle), and assuming dependencies between the angles modelled.

In this chapter, we have studied how using different potential derivation protocols could improve the performances of DFIRE2 on near-native decoy sets. To do so, we generated potentials for four geometrical features of pairwise atomic interactions, in the same fashion as for hydrogen bonds, meaning one distance term, two angle terms, and one dihedral term. Each of these features was generated from a different method, using combinations of crystal structures and decoys. Moreover, a distinction was made between non-bonded atomic interactions within the same residues, or across different residues.

5.2 Methods

5.2.1 Protein sets

Crystal structure sets

Since this is an all-atom study, we need to generate explicit hydrogen positions for the crystal structures that we use when training the potential. This was done in a similar fashion as for the hydrogen bonding and solvation potential, using the pdb2gmx tool. The crystal structure set selected from the PDB is the same one as used in generating the solvation and hydrogen bonding potentials, and is comprised of 713 proteins resolved at a resolution of 2 Å or less.

Decoy sets

We have considered various decoy sets, but mainly focused our efforts on the MDSET. 70 targets with 500 decoys were selected at random, the remaining 180 being used in conjunction with the crystal structure set to produce the potential distributions. The results obtained are validated using the HRDECOY decoy set, comprised of 150 targets, each with 500 decoys. We will use 40 of these targets for testing, and 110 for training.

5.2.2 Interaction model

One of the hypotheses of this experiment is that interactions between pairs of atoms are similar to hydrogen bonds, in the sense that they are directional, prefer a specific dihedral value, and equilibrate at a certain distance. A non-bonded interaction can be described using four points, as shown in Figure 5.2 below.

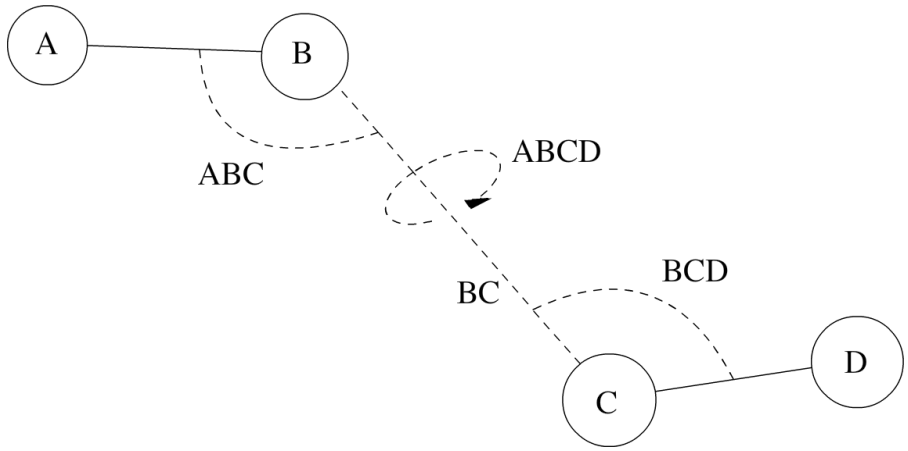


Figure 5.2 Atomic interaction model. In this system, B and C are the atoms interacting, A is the centroid of atoms covalently bonded to B, and D is the equivalent for atom C. ABC represents the angle at atom B, BCD the angle at atom C, ABCD is the torsion angle along the BC axis, and BC is the distance between atoms B and C.

5.2.3 Statistical potentials

We looked at various ways of deriving energies in our potentials, as well as various ways of selecting which atoms are interacting. All of our potentials are derived using the inverse Boltzmann formulation introduced previously in this thesis. The energy derivation protocols used are defined in Table 5.1 below.

Table 5.1 Potential generation methods

Protocol	Subscript	Observed State	Reference State
Classical	C	Crystal structures	Geometrical model
Hybrid	H	Crystal structures	Decoys with RMSD > average
Decoy-based	D	Decoys with RMSD < average	Decoys with RMSD > average

The Classical method is the one conventionally used in the literature, where crystal structures are used as the observed state, and a model of randomness is used for the reference state. Here, we used the method from DFIRE to derive our reference state.

The Decoy-based approach is the same as the one used in previous chapters, and uses better than average decoys as the observed state, and worse than average decoy as the reference state.

The hybrid method is a combination of both, and uses crystal structures as the observed state, and decoys that are worse than average as the reference state.

Each potential was then split into two parts, one representing non-bonded intra-residue interactions, and the other representing non-bonded interactions between atoms in different residues. We decided to separate the two because non-bonded interactions in the same residue are subjected to local topology constraints, and as such, are likely to have a different distribution. The intra-residue term therefore mostly represents the side chain packing of individual residues. The nomenclature used in this study is shown in Table 5.2 below.

Table 5.2 Potentials residue separation

Type	Superscript	Residue separation
Inter-residue		>= 1
Side chain packing (intra-residue)	res	0

The interaction model represented in Figure 5.2 includes 4 geometrical features to represent the interaction between two non-bonded atoms. We will derive a potential for each of them, and refer to them using the formalism in Table 5.3.

Table 5.3 Geometrical features modelled

Name	Type
BC	Distance
ABC	Angle
BCD	Angle
ABCD	Dihedral

These features model the interaction between atoms “B” and “C”, with “A” being the centroid of atoms covalently bonded to “B”, and “D” the centroid of atoms covalently bonded to “C”.

Since we generate a potential for each feature (BC, ABC, BCD, ABCD), for each generation protocol (classical, hybrid, decoy-based) and for both inter and intra-residues, we end up having 24 different energy terms, which we will have to choose from, and combine in order to produce a full potential.

The binning protocol used was derived by taking the minimal bin size that is reasonable given the number of data points we have. Since the crystal structure set used to derive the classical potentials is the smallest, we used it as the basis to define our bins. In this set, there are 713 proteins to which we add the 400 crystal structures from the decoy sets, giving us 1113 proteins from which to

derive the observed state. Each protein has an average of 2000 atoms in them, giving us a total of 2 billion unique pairwise atomic interactions.

Given that tryptophan is the rarest amino acid with a frequency of occurrence of 1.3%, the rarest interaction will be between two tryptophans, which will represent 0.017% of the data points. Thus, we can expect around 340,000 data points to model TRP-TRP interactions. Since tryptophan contains 24 atoms (aside from termini ones), we have 300 unique TRP-TRP interactions, giving us an average of 1133 data points (340,000 atoms divided by 300 interactions) for any two tryptophans. Thus, by using a maximum of 100 bins for the distance potentials, we have an average of 12 data points in each bin, for the rarest interaction occurring in proteins. Any other interaction should be expected to have significantly more data points. Using this logic, we defined the bins as shown in Table 5.4.

Table 5.4 Value range for each feature

Potential	Range	Bins	Bin size
BC	[0, 15] Å	100	0.15 Å
ABC	[0, 180] º	90	2º
BCD	[0, 180] º	90	2º
ABCD	[-180, 180] º	90	4º

This binning protocol was applied to each different potential generation method, for both residue separations considered.

5.2.4 Combining potential terms

As previously, we use the *optim* function in R, with a Nelder-Mead steepest descent search algorithm to find a good set of parameters. This is optimised to maximise the average 10% enrichment score over a training set taken from the MDSET.

5.2.5 Performance measures

To assess the performances of our potentials, we used the average 10% enrichment score over the out-sample set of our decoy sets. The 15% enrichment, Pearson correlation coefficient and Kendall tau are also used as an additional validation in the final assessment.

5.3 Results & Discussions

5.3.1 Component analysis of DFIRE2

DFIRE2 is composed of four energy terms: a distance term and three angle terms. The distance term is equivalent to the original DFIRE, with the difference that the bins all have equal sizes, and thus, short distances are better modelled.

The four terms composing DFIRE2 are referred to as DFIRE for the distance term, and A1, A2, A3 for the angle terms. The performance of each component, as well as DFIRE2, is shown in Table 5.5 for the MDSET. There has been no optimisation done on weights, each term being taken from the published implementation.

Table 5.5 DFIRE2 energy terms scores

Potential	10% Enrichment
DFIRE	0.34
A1	0.35
A2	0.35
A3	0.36
DFIRE + A1	0.38
DFIRE + A2	0.38
DFIRE + A3	0.39
DFIRE + A1 + A2	0.39
DFIRE + A1 + A3	0.40
DFIRE + A2 + A3	0.40
DFIRE2	0.41

We see above that overall, DFIRE2 performs better than any of its constituents, and correctly identifies 41% of the 10% best decoys. This will serve as our benchmark for assessing the performances of our potentials.

5.3.2 Distance potentials

The first potential terms we considered are derived from the distance between two non-bonded atoms. As explained in the methods section, we will consider atoms in the same residue separately from atoms in different residues.

This was done because atoms within a specific residue are constrained to specific positions due to the residue topology. The intra-residue term mostly represents the side chain packing, while the inter-residue non-bonded term represents the long-range van der Waals and electrostatic interactions. We derived three distance potentials for each residue separation, using a different protocol to define the observed and reference states.

The classical potential uses crystal structures as the observed state, and a geometrical reference state modelling the probability of each bin at random. Here, we used the formalism of DFIRE for the reference state.

The decoy-based potential uses the decoys that are better than average as the reference state, and the decoys that are worse than average for the reference state. In chapter 3, we showed that this approach successfully generates a solvation term, so we are now interested in generalising it.

Finally, our third approach uses the crystal structures as the observed state, and the decoys that are worse than average as the reference state.

For each of these potentials, we used the training set for the MDSET, and calculated the 10% enrichment to assess how well they perform at discriminating the near-native decoys in our test-set. We then combined the intra-residue and inter-residue potentials for each method. Results are shown in Table 5.6, along with the Wilcoxon statistic and the p-value of the difference to a random distribution.

Table 5.6 Distance potentials 10% enrichment for the MDSET

Potential	10% Enrichment	W	P-value (H_0=0.10)
BC_C	0.34	1499	1e-9
BC_D	0.53	1594	8e-11
BC_H	0.52	1596	7e-11
BC_C^{res}	0.42	1376	4e-10
BC_D^{res}	0.69	1596	7e-11
BC_H^{res}	0.52	1596	7e-11
$BC_C + BC_C^{res}$	0.48	1596	7e-11
$BC_D + BC_D^{res}$	0.72	1596	7e-11
$BC_H + BC_H^{res}$	0.61	1596	7e-11

We can see that every term derived here is significantly better than random, with scores ranging from 0.34 to 0.72. As in previous chapters, the highest score was achieved by the decoy-based potential.

We can also observe that intra-residue interaction score consistently higher than inter-residue ones. This can be explained by the fact that side chains are easier to unfold than the main chain, and thus, more differences between good and bad structures will be observed in the side chain packing, leading to better discrimination.

Finally, we can notice that combining the inter- and intra-residue terms produce a better score than each term alone. To verify this, we calculated the p-value of the difference between the combination, and the intra-residue term alone, using a 1-tailed Wilcoxon rank sum test. We obtained p-values of 0.05, 0.05 and 0.0001 for the classical, decoy-based and hybrid potentials respectively.

Therefore, we can conclude that combining inter- and intra-residue distance terms is useful for discriminating near-native decoys.

5.3.3 Angle potentials

We have seen that we could successfully derive a combination of distance potentials that would perform reasonably well on a near-native decoy set.

In order to further improve the performance of our full potential, and in order to account for the relative orientation of the different atoms, we derived two angle terms, one for each atom interacting. This is similar to our definition of hydrogen bonds introduced in Chapter 4, but applied to unconstrained, long-range and short-range interactions.

The ABC and BCD angles are expected to be very similar in performance, given that we only consider each pair of atoms once, and as such, the ABC angle would be the BCD angle if we had inverted the order of the two atoms considered. Any difference would be due to the ordering that we arbitrarily chose (in our case, we chose the B atom name to be alphabetically before the C atom), rather that any difference in the interaction.

To illustrate how angles can have a preference, even at longer ranges, we generated plots for the ABC angles in crystal structures, for the Threonine-Threonine interactions between the side chain OG1 and main chain H atoms, and between the side chain CG2 and main chain O atoms. This is shown in Figure 5.3, while the potentials are shown in Figure 5.4 for the classical ones, 5.5 for the decoy-based ones, and 5.6 for the hybrid ones. We chose Threonine because it is neither too hydrophilic nor hydrophobic, and has a flexible side chain with a hydrogen bond donor and acceptor.

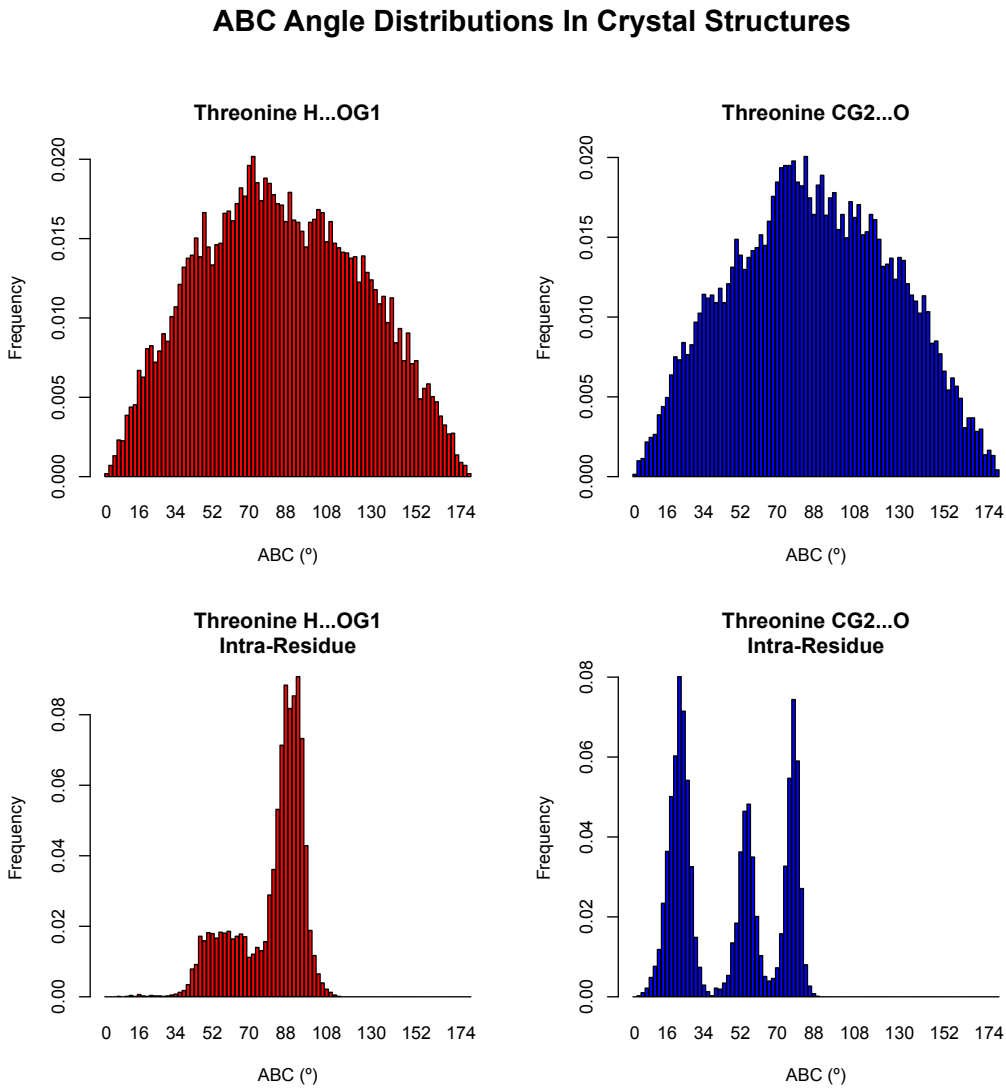


Figure 5.3 ABC angle distributions. This plot shows the distribution of inter- (upper graphs) and intra-residue (lower graphs) ABC angle for the Threonine-Threonine H…OG1 interaction, and the CG2…O interaction. We can observe that the inter-residue distribution is slightly skewed towards smaller angles, while the intra-residue angles have distinct preferences in various regions. The H…OG1 intra-residue interaction shows a preference at 90º, but also has a secondary plateau between 40º and 65º. As for the CG2…O interaction, it has three very strong peaks, at 20º, 55º and 80º. Few potentials consider intra-residue non-bonded angles, but we can see here that they are actually very strongly constrained to specific regions, and should therefore be taken into account. Moreover, we observe no angles at all in some regions, suggesting that steric effects are preventing these angles from being accessible.

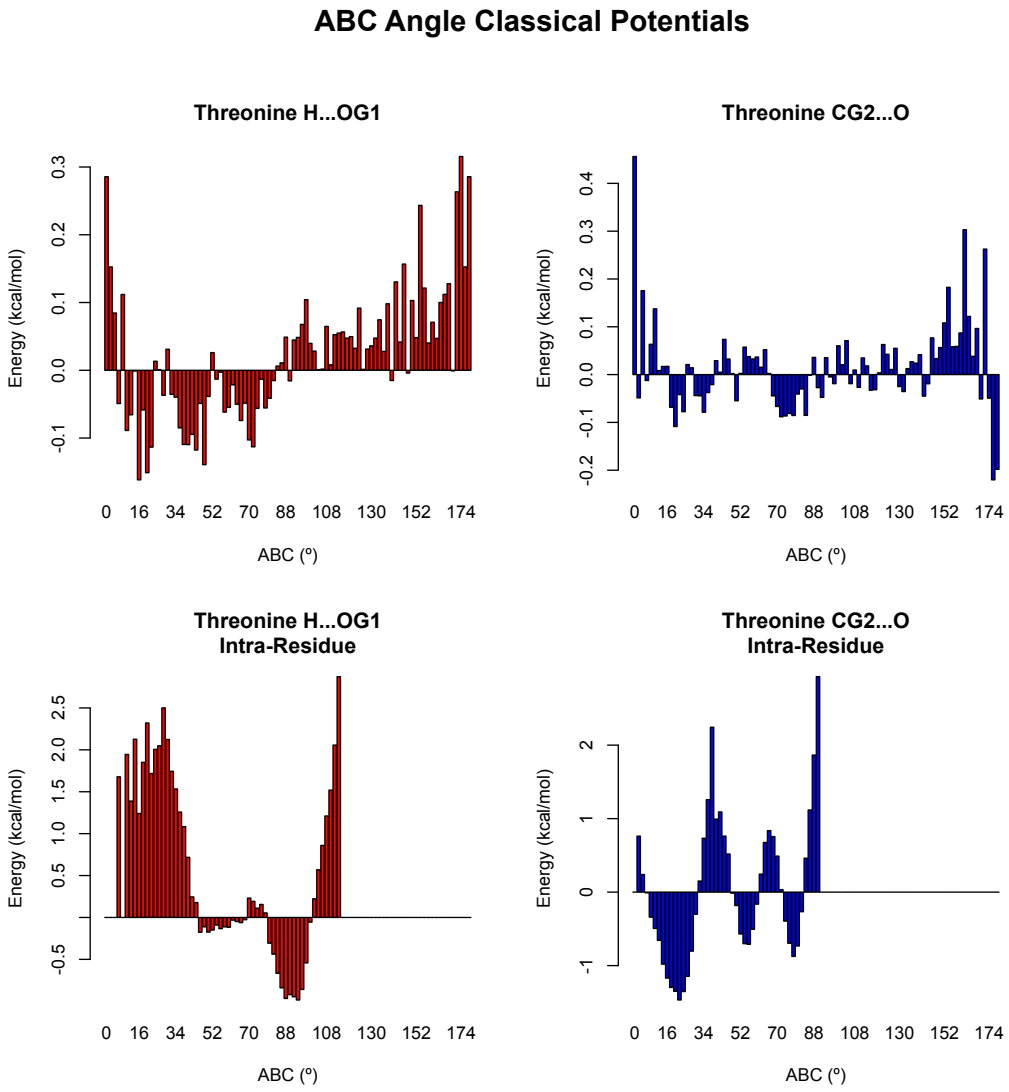


Figure 5.4 ABC angle classical potentials. This plot shows the distribution of inter- (upper graphs) and intra-residue (lower graphs) ABC angle classical potentials for the Threonine-Threonine H…OG1 interaction, and the CG2…O interaction. The classical potentials are derived using crystal structure as the observed state (Figure 5.3) and a random geometric probability model as the reference state. The inter-residue H…OG1 potential has 3 distinct preferred regions at 16º, 50º and 80º, while it disfavours angles above 90º and below 10º. The intra-residue though has a major preference for angles around 90º, and largely disfavours angles below 45º and above 100º. The CG2…O potential has a different distribution, and has a preference for linear angles in the inter-residue potential, while it shows an alternating preference at 20º, 55º and 80º for the intra-residue potential, which is consistent with the observations in Figure 5.3.

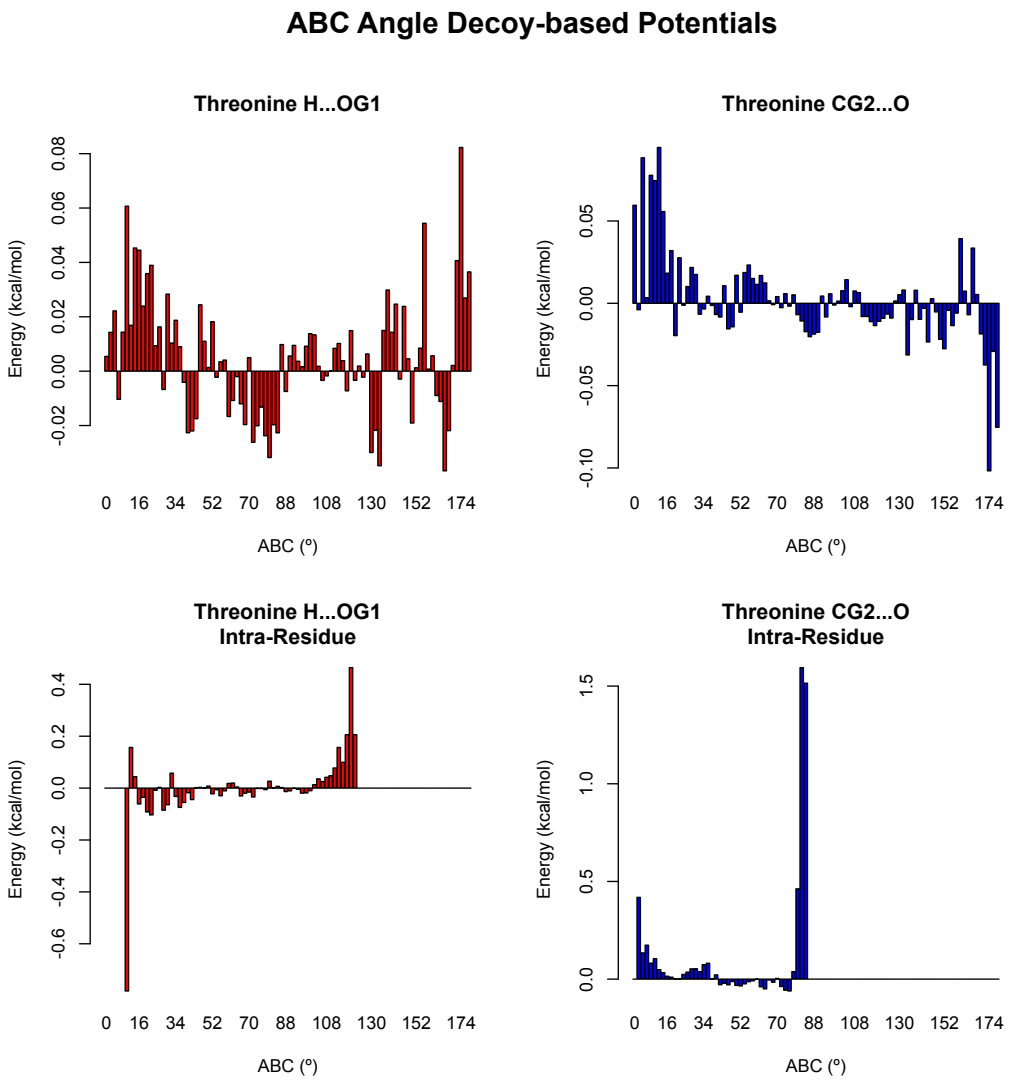


Figure 5.5 ABC angle decoy-based potentials. This plot shows the decoy-based potentials for the Threonine-Threonine H…OG1 interaction, and the CG2…O interaction. The decoy-based potentials are derived using better than average decoys as the observed state and worse than average decoys as the reference state. We can see here that the H…OG1 intra-residue potential has a large preference for very short angles (10º), while it disfavours angles above 90º. The CG2…O potential has a preference for linear angles in inter-residue interactions, while it largely disfavours angles around 80º in the intra-residue ones.

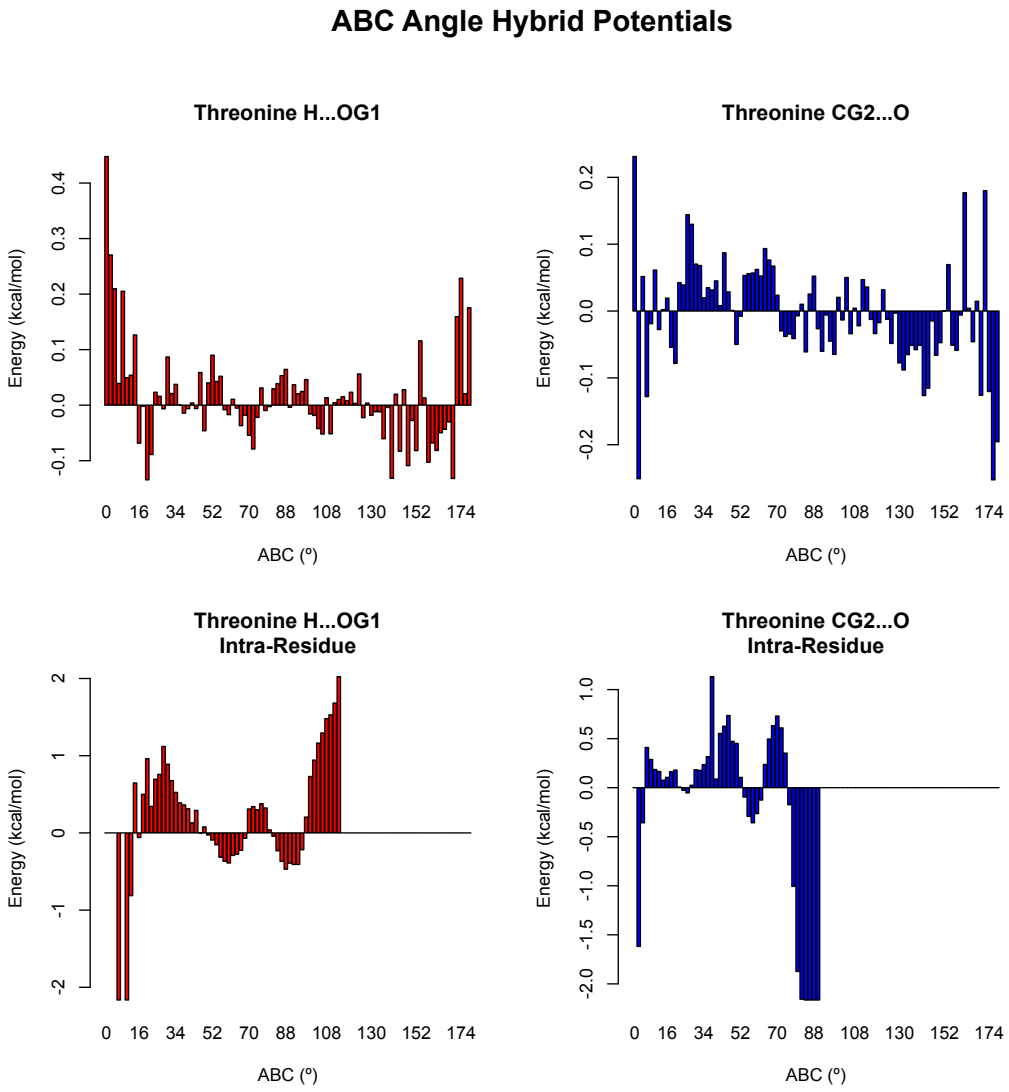


Figure 5.6 ABC angle Hybrid potentials. This plot shows the ABC angle hybrid potentials for the Threonine-Threonine H…OG1 interaction, and the CG2…O interaction. The hybrid potentials are derived using crystal structures as the observed state and worse than average decoys as the reference state. We observe disfavour towards very small angles in the H…OG1 inter-residue interactions. On the other hand, the intra-residue H…OG1 potential shows a large preference for angles around 10º, and has two other favourable regions at 60º and 90º, while strongly disfavouring angles above 100º, and between 16º and 45º. The CG2…O intra-residue potential has a preference for very short angles at 5º, and another between 75º and 90º.

We can see from the examples given previously that all three methods produce very different potentials, suggesting there might be different information to extract from each of them, and thus, it might be useful to eventually combine them to produce a more precise potential. Indeed, the more information is accessible to describe decoys, the more differences will be spotted, and thus, the better we should be able to discriminate them.

In every case seen here, the intra-residue potentials gave more information, as they showed clearer favourable and unfavourable regions, whereas the inter-residue potentials had overall trends observable, but less finer details.

The potentials for the ABC and BCD angles are then used to discriminate decoys in the MDSET. We tested each individual term as well as the combination of the intra- and inter-residue terms. Results for the 10% enrichment, with the p-value to random calculated from a Wilcoxon test, are shown in Table 5.7.

Table 5.7 Angle potential 10% enrichment for the MDSET

Potential	10% Enrichment	W	P-value (H_0=0.10)
ABC_C	0.24	1540	1e-10
ABC_D	0.38	1596	7e-11
ABC_H	0.31	1540	1e-10
BCD_C	0.22	1431	2e-10
BCD_D	0.38	1596	7e-11
BCD_H	0.41	1596	7e-11
ABC_C^{res}	0.37	1596	7e-11
ABC_D^{res}	0.70	1596	7e-11
ABC_H^{res}	0.58	1596	7e-11
BCD_C^{res}	0.34	1540	1e-10
BCD_D^{res}	0.67	1596	7e-11
BCD_H^{res}	0.53	1596	7e-11
ABC_C + ABC_C^{res}	0.40	1596	7e-11
ABC_D + ABC_D^{res}	0.71	1596	7e-11
ABC_H + ABC_H^{res}	0.60	1596	7e-11
BCD_C + BCD_C^{res}	0.35	1540	1e-10
BCD_D + BCD_D^{res}	0.68	1596	7e-11
BCD_H + BCD_H^{res}	0.57	1596	7e-11

Here, ABC represents the angle about the B atom, BCD the angle about the C atom. The p-value is for the difference to the enrichment score of a random distribution.

We can see that, as for distance potentials, the best performing terms are the intra-residue ones, with scores between 0.37 and 0.70. The combination of inter- and intra-residue terms did not have much better scores compared to the intra-residue terms alone.

From these results, we can conclude that using angle potentials is useful at discriminating near-native decoys, even without combining them with a distance potential.

5.3.4 Dihedral potentials

Most potentials in the literature don't include a dihedral term for non-bonded interactions, as they do not consider it as being representative of an existing interaction. DFIRE2 partly models the dihedral angle by including a third angle term, representing the angle between the vectors defined by the atoms and their covalently bonded atoms.

The performance of these potentials was assessed by calculating the 10% enrichment score over the MDSET decoy set, for each term derived using the same protocol that was used for distance and angle terms. The results, as well as the p-value and Wilcoxon test are shown in Table 5.8.

Table 5.8 Dihedral potentials 10% enrichment for the MDSET

Potential	10% Enrichment	W	P-value (H_0=0.10)
$ABCD_C$	0.28	1540	1e-10
$ABCD_D$	0.28	1596	7e-11
$ABCD_H$	0.32	1596	7e-11
$ABCD_C^{res}$	0.36	1485	2e-10
$ABCD_D^{res}$	0.70	1596	7e-11
$ABCD_H^{res}$	0.56	1596	7e-11
$ABCD_C$ + $ABCD_C^{res}$	0.40	1593	9e-11
$ABCD_D$ + $ABCD_D^{res}$	0.70	1596	7e-11
$ABCD_H$ + $ABCD_H^{res}$	0.58	1596	7e-11

As is the case for distance and angle potentials, the best performing single terms are the intra-residue decoy-based and hybrid potentials, with scores of 0.70 and 0.56 respectively. The combination of intra- and inter-residue terms did not significantly score higher compared to the intra-residue terms alone, suggesting that for this decoy set, the intra-residue interactions are a better indicator of the nativeness of the structure.

5.3.5 Comparison to DFIRE2

In order to compare the efficiency of our potential to DFIRE2, and to assess whether including DFIRE2 is useful, we combined the solvation, distance, angle and dihedral terms into several groups. We will combine all terms, including solvation, into three potentials, one composed of all the classically derived terms, one composed of the decoy-based terms, and finally a third one composed of hybrid terms. The solvation potential was regenerated in order to create the hybrid version that was not studied previously.

Results for the 10% enrichment for the MDSET decoy set are shown in Table 5.9. A one-tailed Wilcoxon test was used to assess the p-value of the difference of our potentials to DFIRE2, and is shown in Table 5.9 as well. The FULL potentials are composed of the distance, angles and dihedral terms, and are derived for inter- and intra-residue interactions. We then combine the full inter-residue, intra-residue and solvation terms to produce our complete potential, which we call the Distance, Orientation and Solvation (DOS) potential.

Table 5.9 Full potentials 10% enrichment for the MDSET

Potential	10% Enrichment	W	P-value (H_0=DFIRE2)
DFIRE2	0.41		
$FULL_C$	0.41	1580	0.53
$FULL_D$	0.60	638	3e-8
$FULL_H$	0.59	679	1e-7
$FULL_C^{res}$	0.44	1390	0.15
$FULL_D^{res}$	0.71	243	6e-15
$FULL_H^{res}$	0.63	485	1e-10
DOS_C	0.51	1066	0.002
DOS_D	0.73	194	6e-16
DOS_H	0.67	355	8e-13

We can see from this table that aside from the classical intra- and inter-residue combinations, all other full potentials are significantly better than DFIRE2. In

particular, the combination of the decoy-based intra-residue, inter-residue and surface accessibility yielded the best score, with an average enrichment of 0.73. The best scoring potentials are always the decoy-based ones, followed by the hybrid ones, and finally the classical ones. Thus, we can conclude that using decoys as part of a potential is useful when discriminating near-native decoys.

It is not surprising that the $FULL_C$ term has the same score as DFIRE2, considering they both are formulated and derived in the same way, use the same reference state, and have almost the same combination of terms, the main differences being the number of angle bins, the angle between the vectors in DFIRE2 that has been replaced by the dihedral term here, and the inclusion of non-polar atoms in the angle terms.

DFIRE2 was derived as a multivariate potential, where the three angles and distance are used to produce a 4 dimensional potential. From this experiment, we can see that there is no benefit in the increased complexity compared to the simple linear combination of each independent term, as in the $FULL_C$ potential. This is in line with our conclusion of Chapter 4, where we found that bivariate hydrogen-bonding potentials did not outperform a linear combination of single terms.

To verify our results, we generated the potentials using another decoy set, the HRDECOY set, and combined the DOS_C, DOS_D and DOS_H potentials into a single one, which we simply call the DOS potential. This was done to benefit from the information extracted from native structures, but also from the decoys that are representative of a specific generation method. To avoid over-fitting due to the large number of terms, we optimised using the total value of the DOS_C, DOS_D and DOS_H instead of their individual components, thus only having 3 terms for which to optimise the weights.

We calculated the 10% enrichment, 15% enrichment, Pearson correlation, and Kendall Tau, with the significance of the difference to DFIRE2 assessed using the p-value of the one-tailed Wilcoxon rank sum test. Results are shown in Table 5.10 below.

Table 5.10 DOS potential scores

	DFIRE2	DOS	W	P-value (H_0=DFIRE2)
MDSET				
10% Enrichment	0.41	0.75	164	2e-16
15% Enrichment	0.45	0.75	253	9e-15
Pearson R	0.54	0.79	424	1e-11
Kendall Tau	0.36	0.54	567	3e-9
HRDECOY				
10% Enrichment	0.42	0.47	662	0.09
15% Enrichment	0.49	0.55	618	0.04
Pearson R	0.73	0.83	456	0.0004
Kendall Tau	0.55	0.66	369	9e-6

The analysis of the DOS potential showed that it consistently and significantly outperformed DFIRE2, on both decoy sets and using different correlation measures. The only exception was the 10% enrichment on the HRDECOY set, where the p-value was 0.09.

The difference in performance between the two decoy sets probably comes from the fact that there are 100 less targets used in the HRDECOY set than in the MDSET. Given that the decoy-based potentials are generated for each set, we could expect less details in the HRDECOY one, and thus, a lesser performance. Further work will examine the effect of varying the number of targets used to generate the potentials, and determine the minimum number of structures required to derive precise potentials for near-native decoy discrimination studies.

Nonetheless, we can say that overall, the DOS potential is more useful than DFIRE2 at discriminating near-native decoy structures, as seen in Table 5.11 for the MDEST targets, and illustrated for targets 1JYH and 3EOI in Figure 5.7.

Table 5.11 DOS and DFIRE2 MDSET targets scores

TARGET	DOS	DFIRE2	TARGET	DOS	DFIRE2
133L	0.80	0.34	1UA8	0.88	0.38
153L	0.76	0.36	1UJ8	0.78	0.54
1AA2	0.68	0.24	2B1K	0.80	0.54
1AAJ	0.74	0.20	2BK8	0.88	0.46
1ACF	0.74	0.58	2CGQ	0.72	0.54
1AGI	0.80	0.42	2CKX	0.58	0.54
1EW4	0.82	0.70	2COV	0.68	0.28
1EY0	0.78	0.40	2CWR	0.72	0.28
1EYH	0.80	0.58	2FZP	0.84	0.66
1EZK	0.82	0.40	2GBN	0.76	0.32
1F32	0.70	0.46	2HDZ	0.76	0.58
1FAA	0.72	0.60	2HLQ	0.72	0.20
1JB3	0.66	0.64	2HP7	0.80	0.44
1JMW	0.86	0.40	2OVO	0.66	0.22
1JOS	0.54	0.36	2P5D	0.78	0.46
1JPE	0.66	0.38	2PCY	0.68	0.40
1JYH	0.82	0.06	2PKO	0.80	0.50
1KN3	0.82	0.42	2PTH	0.76	0.86
1MWP	0.84	0.26	2YGS	0.58	0.44
1MZL	0.46	0.08	2YWD	0.78	0.32
1NA5	0.84	0.30	2YWK	0.70	0.30
1NKO	0.74	0.24	2YXM	0.66	0.36
1NOA	0.66	0.26	2Z9T	0.66	0.14
1OW1	0.86	0.64	2ZEQ	0.50	0.30
1TXJ	0.74	0.54	3EOI	0.92	0.26
1TZV	0.82	0.82	3EYE	0.82	0.20
1U9A	0.84	0.52	3FH2	0.88	0.62
1U9P	0.82	0.32	3G9B	0.78	0.28

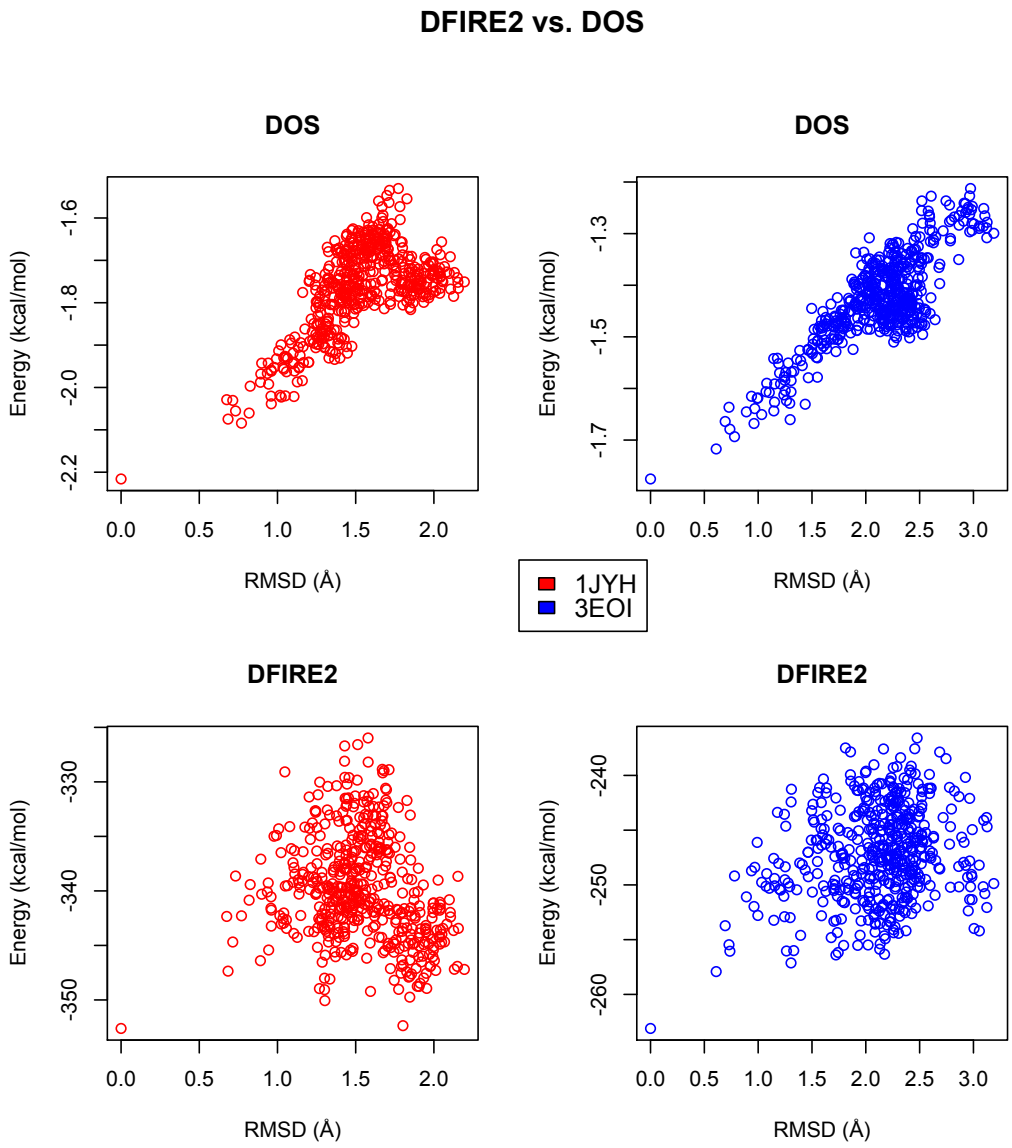


Figure 5.7 Performance of DFIRE2 and DOS on targets 1JYH and 3EOI. We can see that the DOS potential vastly outperforms the DFIRE2 potential on both targets, and at all ranges of All-atom RMSD. We can also note that all potentials successfully place the native structure at the lowest energy point.

5.4 Conclusions

In this chapter, we have studied how distance and orientation potentials could help discriminate near native decoys. In order to do so, we have generated potentials to model four features describing the interactions between two atoms, namely the distance between them, the planar angle about each atom, and the dihedral angle along their axis. This interaction model is similar to the one used for hydrogen bonds in Chapter 3 in terms of the features described.

The potentials were generated for both non-bonded intra-residue interactions, modelling the side chain packing, and for non-bonded inter-residue interactions, representing the van der Waals and electrostatic interactions occurring at longer ranges, thereby modelling the larger scale structural features such as alpha-helices and beta-sheets. We could observe that across all features described here, there were clear preferences for specific regions. This is even truer for intra-residue interactions, where the range of allowed values is constrained due to the residue topology.

To extract the energies from the observed distribution of features, we used three methods, and assessed the performance of each one independently, then combined them to produce a complete potential. The potentials were derived using both crystal structures and decoys. Although the classical formulation performed slightly better than DFIRE2 (0.51 vs 0.41), the decoy-based potential, derived using decoys only for both the observed and reference states, performed much better on the MDSET, with a 10% enrichment score of 0.73.

The combination of the classical, hybrid and decoy-based potential further increased this score to 0.75, meaning that 3 out of 4 of the 10% best decoys will be successfully identified. This performance difference with DFIRE2 can be seen across different measures (0.75 vs 0.45 for 15% enrichment, 0.79 vs 0.54 for the Pearson correlation, and 0.54 vs 0.36 for the Kendall tau), showing that this method is highly effective at discriminating near-native decoys. Moreover, we have tested our potential on a second set of decoys, and it consistently outperformed DFIRE2, scoring as much as 0.83 for the Pearson correlation, and 0.66 for the Kendall tau.

From our results, we can conclude that a combination of distance, angle, dihedral and solvation potentials derived both using crystal structures and decoys, is

useful at discriminating near-native decoys, and outperforms the most successful potentials from the literature.

Chapter 6

General Conclusions

In this thesis, our goal was to develop a new energy function that could be used to correctly identify the best protein models generated by folding simulations.

Folding algorithms such as comparative modelling, and recently even *ab initio* simulations, are generating an increasing number of near-native structures. As such, it is important that a method exists to differentiate between these high quality predictions, so that only the very best models are kept for further analysis. Thus, our work is increasingly gaining value, as the more we approach the near-native region, the more we will need accurate methods to discriminate the best structure.

Many potentials already exists, but we found that none of them performed well when discriminating near-native decoy structures. The best one we found was DFIRE2, which we then selected as our benchmark potential. The DOS potential derived in this thesis has largely outperformed DFIRE2, and is thus more likely to be useful for discriminating decoys in future folding simulations and in blind tests such as the biannual CASP competition.

Of the existing methods, comparative modelling has already been producing sub-2 Å models, and can therefore already benefit from our approach. In some cases, threading will as well, but *ab initio* simulation are still mostly trying to find the right fold, which our method is not designed to do.

One reason for the relatively low precision of existing potentials is the lack of decoy sets published that contains decoys with an RMSD to native below 2Å, as potentials could only have been tested and optimised for that range.

In order to address this problem, we generated our own near-native decoy set using molecular dynamics to relax non-redundant, monomeric crystal structures from the PDB. The trajectory was then sampled at regular intervals, and decoys clustered to remove any structural redundancy. This process was repeated for various temperatures and 250 different proteins. In the end, we kept 125000 decoys between 0 and 4Å, 30% of which were set aside as an out-sample to test our potentials.

Using this decoy set, as well as another near-native one from the literature, we studied the performance of three different potential generation methods. The first, which we call the “classical” method, is the conventional way of deriving statistical potentials, using crystal structures as the observed state, and a probability model as the reference state. This approach generates potentials that theoretically yield the energy difference between native structures and unfolded

ones, and thus, should be expected to easily identify the native structures in the decoy sets, as they would most resemble those used to generate the potentials.

Our second approach, which we call “decoy-based”, uses the mean RMSD of a decoy set to separate good decoys from bad ones, using the former as the observed state, and the latter as the reference state, effectively modelling the structural differences between these two groups. Thus, we can expect it to be more detailed than the classical potentials, although the smaller the differences between decoys, the more structures will be needed to model them, and thus, we can easily end up with potentials that do not exhibit any particular performance.

Finally, our third method, called “hybrid”, uses native structures as its observed state, and the worse than average decoys as its reference state. The assumption behind this approach is that a large enough number of decoys should statistically represent the reference state, and as such we do not need to know its distribution in advance, as is the case in the classical potentials.

These three approaches are then used to generate potential for four different interactions, namely solvation, hydrogen bonding, pairwise atomic distance, and atomic orientation. We found that hydrogen bonding was not useful, and is thus not included in our final potential, called DOS, which is created by combining the classical, hybrid and decoy-based potentials for the distance, orientation and solvation terms. This effectively keeps the precision of the decoy-based approach and the generality of the classical potentials. In a sense, the classical terms add robustness and generality to our potential rather than precision.

In terms of performance, DOS achieved a 10% enrichment score of 0.75 on our near-native decoy set (MDSET), meaning that 3 out of 4 of the 10% best decoys will be successfully identified. In contrast, our benchmark potential from the literature, DFIRE2, only scored 0.41. Furthermore, this performance difference can be seen across different metrics (0.75 vs 0.45 for the 15% enrichment, 0.79 vs 0.54 for the Pearson correlation, and 0.54 vs 0.36 for the Kendall tau). To validate our results, we tested DOS on a second decoy set, and it still consistently outperformed DFIRE2, scoring as much as 0.83 for the Pearson correlation, and 0.66 for the Kendall tau.

Although we only tested our method on two decoy sets, it is easy to derive the DOS potential for any decoy sets. To do so, we suggest using the following protocol (7 steps):

1. Choose a large number of non-redundant targets (> 150) with a sequence identity less than 30%.
2. Generate at least 500 decoys per target, with an RMSD between decoys larger than 0.5 Å.
3. Minimise the structures to remove clashes
4. Generate the decoy-based potentials by taking the decoys that are better than average as the observed state, and worse than average as the reference state
5. Pick a large number of crystal structures (> 1000) and minimise them.
6. Generate the classical and hybrid potentials.
7. Combine the classical, hybrid and decoy-based potentials using a suitable weighting scheme to produce the DOS potential.

There are limitations to our approach that we have not considered though. Namely, we did not try it on non-native decoys sets, and we assumed that all decoys have the same overall fold.

Indeed, the fact that all of our targets have RMSDs normally distributed means that we will have a clear cutoff to use in our analysis. For decoy sets with different distributions of RMSDs and thus probably different fold groups within targets, using the mean as the threshold between good and bad decoys might not be the best choice. One way to work around this problem would be to sort decoys by fold, so that we only compare the good and bad decoys within a specific fold group. This should in theory allow us to select the best structures for each of the different folds, the task of determining the correct fold group being left to another method.

Furthermore, our decoy-based potentials are derived using only one cutoff, meaning that it will assess the energy difference between decoys on either side of the mean RMSD. In reality, we could probably extend this idea, and generate potentials for different RMSD cutoffs, thereby modelling different conformational spaces regions. For example, we could generate a potential for decoy between 0 and 1Å, 1 and 2Å, 2 and 3Å, and so on so forth. We would then either find a way to select one of these potentials or combine them to recreate the overall energy function.

As a general conclusion to this thesis, we can say that using decoys as both the observed state and reference state in statistical potentials is more useful than existing potentials from the literature when used in near-native decoys discrimination studies, and could probably be used on non-native decoy sets if a suitable fold detection method is used to cluster decoys with similar folds.

Bibliography

Allen, L.C. (1975). Simple model of hydrogen bonding. *Journal of the American Chemical Society.* 97 (24) p. 6921-6940.

Aloy, P., & Oliva, B. (2009). Splitting statistical potentials into meaningful scoring functions: Testing the prediction of near-native structures from decoy conformations. *BMC Structural Biology.* 9 (1) p. 71.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., & Lipman.D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research.* 25 (17) p. 3389-3402.

Andersen, N.H. (2010). Amino acids representation. [andersenlab.chem.washington.edu/CSDb/ accessed May 2010].

Anfinsen, C.B. (1973). Principles that govern the folding of protein chains. *Science*. 181 (4096) p. 223-230.

Anfinsen, C.B., Scheraga, H.A. (1975). Experimental and theoretical aspects of protein folding. *Advanced Protein Chemistry*. 29. p. 205-300.

Arab, S., Sadeghi, M., Eslahchi, C., Pezeshk H., & Sheari, A. (2010). A pairwise residue contact area-based mean force potential for discrimination of native protein structure. *BMC Bioinformatics.* 11 (1) p. 16.

Baker, D. (2006). Prediction and design of macromolecular structures and interactions. *Philosophical transactions of the Royal Society of London. Series B, Biological Sciences.* 361 (1467) p. 459-463.

Baker, D., & Sali, A. (2001). Protein structure prediction and structural genomics. *Science.* 294 (5540) p. 93-96.

Baker, E.N., & Hubbard, R.E. (1984). Hydrogen bonding in globular proteins. *Progress in Biophysics and Molecular Biology.* 44 (2) p. 97-179.

Berendsen, H.J.C., Postma, J.P.M., van Gunsteren, W.F., DiNola, A., & Haak, J.R. (1981). Molecular dynamics with coupling to an external bath. *The Journal of Chemical Physics.* 81 (8) p. 3684-3690.

Berendsen, H.J.C., van der Spoel, D., & van Drunen, R. (1995). GROMACS: A message-passing parallel molecular dynamics implementation. *Computer Physics Communications.* 91 (1-3) p. 43-56.

Bernal, J.D., Fowler, R.H. (1933). A theory of water and ionic solution, with particular reference to Hydrogen and Hydrozyl ions. *Journal of chemical physics*. 1 (515).

Bernstein, F.C., Koetzle, T. F., Williams, G. J., Meyer, E. F., Brice, M. D., Rodgers, J.R., Kennard, O., Shimanouchi, T., & Tasumi, M. (1977). The Protein Data Bank: a computer-based archival file for macromolecular structures. *Archives of Biochemistry and Biophysics.* 185 (2) p. 584-91.

Betz, S.F. (1993). Disulfide bonds and the stability of globular proteins. *Protein Science.* 2 (10) p. 1551-1558.

Bhattacharyay, A., Trovato, A., & Seno, F. (2006). Simple solvation potential for coarse-grained models of proteins. *Proteins: Structure, Function, and Bioinformatics.* 67 (2) p. 285-292.

Boas, F.E., & Harbury, P.B. (2007). Potential energy functions for protein design. *Current Opinion in Structural Biology.* 17 (2) p. 199-204.

Bondi, A. (1964). van der Waals Volumes and Radii. *The Journal of Physical Chemistry.* 68 (3) p. 441-451.

Bordo, D., & Argos, P. (1994). The role of side-chain hydrogen bonds in the formation and stabilization of secondary structure in soluble proteins. *Journal of Molecular Biology.* 243 (3) p. 504-519.

Bourne, P.E. (2003). CASP and CAFASP experiments and their findings. *Methods of Biochemical Analysis.* 44 p. 501-507.

Brandl, M., Lindauer, K., Meyer, M., & Suhnel, J. (1999). C-H.O and C-H.N interactions in RNA structures. *Theoretical Chemistry Accounts.* 101 p. 103-113.

Brooks, B., & Karplus, M. (1983). Harmonic dynamics of proteins: normal modes and fluctuations in bovine pancreatic trypsin inhibitor. *Proceedings of the National Academy of Sciences of the United States of America.* 80 (21) p. 6571-6575.

Brooks, B.R., Brooks, C.L. III, Mackerell, A.D. Jr., Nilsson, L, Petrella, R.J., Roux, B., Won,Y., Archontis, G., Bartels, C., Boresch, S., Caflisch, A., Caves, L., Cui, Q., Dinner, A.R., Feig, M., Fischer, S., Gao, J., Hodoscek, M., Im, W., Kuczera, K. (2009). CHARMM: The Biomolecular Simulation Program. *Journal of Computational Chemistry.* 30 (10) p. 1545-1614.

Brooks, B.R., Bruccoleri, R.E., Olafson, B.D., States, D.J., Swaminathan, S., & Karplus, M. (1983). CHARMM: a Program for Macromolecular Energy, Minimization, and Dynamics Calculations. *Journal of Computational Chemistry.* 4 (2) p. 187-217.

Brown, D., & Clarke, J. H.R. (1984). A comparison of constant energy, constant temperature and constant pressure ensembles in molecular dynamics simulations of atomic liquids. *Molecular Physics.* 51 (5) p. 1243-1252.

Brunger, A.T. (1992). X-Plor Version 3.1. A system for crystallography and NMR. *Yale University press*.

Chidambaram, R., & Sikka, S.K. (1968). Bent O-H---O hydrogen bonds in crystals. *Chemical Physics Letters.* 2 (3) p. 162-165.

Chothia, C. (1976). The nature of the accessible and buried surfaces in proteins. *Journal of Molecular Biology.* 105 (1) p. 1-12.

Chothia, C., & Lesk, A.M. (1986). The relation between the divergence of sequence and structure in proteins. *The European Molecular Biology Organization Journal.* 5 (4) p. 823-826.

Clementi, C. (2008). Coarse-grained models of protein folding: toy models or predictive tools? *Current Opinion in Structural Biology.* 18 (1) p. 10-15.

Cooper, J. (1995) *The Ramachandran Plot* [www.cryst.bbk.ac.uk/PPS2/course/section3/rama.html accessed March 2010].

Cornell, W.D., Cieplak, P., Bayly, C.I., Gould, I.R., Merz, K.M., Ferguson, D.M., Spellmeyer, D.C., Fox, T., Caldwell, J.W., & Kollman, P.A. (1995). Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *Journal of the American Chemical Society.* 117 (19) p. 5179-5197.

Coutsias, E.A., Seok, C., & Dill, K.A. (2004). Using quaternions to calculate RMSD. *Journal of Computational Chemistry.* 25 (15) p. 1849-1857.

Dannenberg, J.J. (2010). The Nature of the Hydrogen Bond: Outline of a Comprehensive Hydrogen Bond Theory. *Journal of the American Chemical Society.* 132 (9) p. 3229.

Dehouck, Y., Gilis, D., & Rooman, M. (2006). A New Generation of Statistical Potentials for Proteins. *Biophysical Journal.* 90 (11) p. 4010-4017.

Desiraju, G.R. (1996). The C-H---O Hydrogen Bond: Structural Implications and Supramolecular Design. *Accounts of Chemical Research.* 29 (9) p. 441-449.

Desiraju, G.R. (1996). The C-H.O Hydrogen Bond: Structural Implications and Supramolecular Design. *Accounts of Chemical Research.* 29 (9) p. 441-449.

Dill, K.A., Ozkan, S.B., Shell, M.S., & Weikl, T.R. (2008). The Protein Folding Problem. *Annual Review of Biophysics.* 37 (1) p. 289-316.

Dill, K.A., Ozkan, S.B., Weikl, T.R., Chodera, J.D., & Voelz, V.A. (2007). The Protein Folding Problem: When Will it Be Solved? *Current Opinion in Structural Biology.* 17 (3) p. 342-346.

Dill, K.A., Thomas, M.T., Vojko, V., & Barbara, H-L (2005). Modeling water, the hydrophobic effect, and ion solvation. *Annual Review of Biophysics and Biomolecular Structure.* 34 p. 173-199.

Domagala, M., Grabowski, S., Urbaniak, K., & Mloston, G. (2003). Role of C-H.S and C-H.N Hydrogen Bonds in Organic Crystal Structures - The Crystal and Molecular Structure of. *Journal of Physical Chemistry.* 107 p. 2730-2736.

Eisenberg, D., & McLachlan, A.D. (1986). Solvation energy in protein folding and binding. *Nature.* 319 (6050) 199-203.

Eswar, N., Webb, B., Marti-Renom, M. A., Madhusudhan, M., Eramian, D., Shen, M.-y., Pieper, U. and Sali, A. (2006). Comparative Protein Structure Modeling Using Modeller. *Current Protocols in Bioinformatics*. 15:5.6.1–5.6.30

Fabiola, F., Bertram, R., Korostelev, A., & Chapman, M.S. (2002). An improved hydrogen bond potential: Impact on medium resolution protein structures. *Protein Science.* 11 (6) p. 1415-1423.

Fabiola, G.F., Krishnaswamy, S., Nagarajan, V., & Pattabhi, V. (1997). C-H.O hydrogen bonds in beta-sheets. *Acta Crystallographica Section D Biological Crystallography.* 53 (3) p. 316-320.

Fariselli, P., & Casadio, R. (2001). Prediction of disulfide connectivity in proteins. *Bioinformatics.* 17 (10) p. 957-964.

Fogolari, F., Tosatto, S.C.E., & Colombo, G. (2005). A decoy set for the thermostable subdomain from chicken villin headpiece, comparison of different free energy estimators. *BMC Bioinformatics.* 6 (Mm) 301.

Galeener, F.L. (1985). A model for the distribution of dihedral angles in SiO2 like glasses. *Journal of Non-Crystalline Solids.* 75 (1-3) p. 399-405.

Gilis, D. (2004) Protein decoy sets for evaluating energy functions. *Journal of Biomolecular Structural Dynamics*. 21 (6) p. 725 - 736

Gordon, D.B., Marshall, S.A., & Mayot, S.L. (1999). Energy functions for protein design. *Current Opinion in Structural Biology.* 9 (4) p. 509-513.

Greenwood, J. (2002). The correct and incorrect generation of a cosine distribution of scattered particles for Monte-Carlo modelling of vacuum systems. *Vacuum.* 67 (2) p. 217-222.

Gu, J., Li, H., Jian, H., & Wang, X. (2009). Optimizing energy potential for protein fold recognition with parametric evaluation function. *Journal of computational biology : a journal of computational molecular cell biology.* 16 (3) p. 427-442.

Gu, Y., Kar, T., & Scheiner, S. (1999). Fundamental Properties of the CH---O Interaction: Is It a True Hydrogen Bond? *Journal of the American Chemical Society.* 121 (40) p. 9411-9422.

Guntert, P. (2004). Automated NMR structure calculation with CYANA. *Meth. Mol. Biol*. 278, p. 353-378.

Gunawardena, J. (1996). Statistical Mechanics and Information Theory: report, abstracts and bibliography of a workshop. *HelwettPackard.* June 18.

Handl, J., Knowles, J., Lovell, S.C. (2009). Artefacts and biases affecting the evaluation of scoring functions on decoy sets for protein structure prediction. *Bioinformatics.* 25 (10) p. 1271-1279.

Hao, M.H., & Scheraga, H.A. (1999). Designing potential energy functions for protein folding. *Current Opinion in Structural Biology.* 9 (2) p. 184-188.

Helles, G. (2008). A comparative study of the reported performance of ab initio protein structure prediction algorithms. *Journal of the Royal Society, Interface the Royal Society.* 5 (21) p. 387-396.

Hermans, J., Berendsen, H.J.C., Van Gunsteren, W.F., & Postma, J.P.M. (2004). A consistent empirical potential for water-protein interactions. *Biopolymers.* 23 (8) p. 1513-1518.

Hess, B., Kutzner, C., Van Der Spoel, D., & Lindahl, E. (2008). GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *Journal of Chemical Theory and Computation.* 4 (3) p. 435-447.

Hobohm, U., Scharf, M. Schneider, R., & Sander, C. (1992). Selection of representative protein data sets. *Protein Science.* 1 (3) p. 409-417.

Huang, E.S., Subbiah, S., Tsai, J., & Levitt, M. (1996). Using a Hydrophobic Contact Potential to Evaluate Native and Near-native Folds Generated by Molecular Dynamics Simulations. *Journal of Molecular Biology.* 257 (3) p. 716-725.

Huang, S.-Y., & Zou, X. (2010). Inclusion of solvation and entropy in the knowledge-based scoring function for protein-ligand interactions. *Journal of Chemical Information and Modeling.* 50 (2) p. 262-273.

Hubbard, T. (1996). NACCESS [www.bioinf.manchester.ac.uk/naccess/ accessed October 2009].

Hubner, I.A., Deeds, E.J., & Shakhnovich, E. I. (2005). High-resolution protein folding with a transferable potential. *Proceedings of the National Academy of Sciences of the United States of America.* 102 (52) p. 18914-18919.

Isaacs, E.D., Shukla, A., Platzman, P.M., Hammann, D.R., Barbiellini, B., Tulk, C.A. (1999). Covalency of the hydrogen bond in ice: A direct X-ray measurement. *Physical Review Letters*. 82 (3). p. 600.

Jaynes, E.T. (1957). Information theory and statistical mechanics II. *Physical Review.* (108) 2 p. 171-190.

Jeffrey, G. (2003). Hydrogen-Bonding: An update. *Crystallography.* 9 (2-3) p. 135-176.

Jiang, L., Lai, L. (2002). CH-O Hydrogen Bonds at Protein-Protein Interfaces. *The Journal of Biological Chemistry.* 277 (40) p. 37732-37740.

John, B., & Sali, A. (2003). Comparative protein structure modeling by iterative alignment, model building and model assessment. *Nucleic Acids Research.* 31 (14) p. 3982-3992.

Jones, D.T. (1999). GenTHREADER: An efficient and reliable protein fold recognition method for genomic sequences. *J. Mol. Biol*. 287, p. 797-815.

Jones, D.T. (2001). Predicting novel protein folds by using FRAGFOLD. *Proteins: Structure, Function, and Bioinformatics.* 45 (Suppl 5) p. 127-132.

Jones, D.T., Bryson, K., Coleman, A., McGuffin, L.J., Sadowski, M.I., & Sodhi, J.S., & Ward, J.J. (2005). Prediction of novel and analogous folds using fragment assembly and fold recognition. *Proteins.* 61 (7) p. 143-151.

Jones, D.T., & McGuffin, L. J. (2003). Assembling novel protein folds from super-secondary structural fragments. *Proteins.* 53 (Suppl 6) p. 480-485.

Jones, D.T., Taylor, W.R. & Thornton, J.M. (1992). A new approach to protein fold recognition. *Nature*. 358, p. 86-89.

Jones, D.T. & Thornton, J.M. (1996). Potential-energy functions for threading. *Curr. Opin. Struct. Biol*. 6, p. 210-216.

Jorgensen, W.L., & Tirado-Rives, J. (2005). Potential energy functions for atomic-level simulations of water and organic and biomolecular systems. *Proceedings of the National Academy of Sciences of the United States of America.* 102 (19) p. 6665-6670.

Jorgensen, W.L., & Tirado-Rives, J. (1988). The OPLS Potential Functions for Proteins. Energy Minimizations for Crystals of Cyclic Peptides and Crambin. *Journal of the American Chemical Society.* 110 (6) p. 1657-1666.

Jorgensen, W.L. (1981). Transferable Intermolecular Potential Functions for Water, Alcohols, and Ethers. Application to Liquid Water. *Journal of the American Chemical Society.* 103 (2) p. 335-340.

Jorgensen, W.L., Chandrasekhar, J. Madura, J.D. & Impey, R.W. and Michael L. Klein. (1983). Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics.* 79 (2) p. 926-935.

Kaesar, C., & Levitt, M. (1999). A novel approach to decoy set generation: designing a physical energy function having local minima with native structure characteristics. *Journal of Molecular Biology.* 329 (1) p.159-174.

Kamat, A.P., & Lesk, A.M. (2007). Contact patterns between helices and strands of sheet define protein folding patterns. *Proteins: Structure, Function, and Bioinformatics.* 66 (4) p. 869-876.

Karplus, M., & Weaver, D.L. (1976). Protein folding dynamics. *Nature*. 260 (5550). p. 404-406.

Keates, P. (1998). Lecture 1: Secondary structure of Proteins [www.chembio.uoguelph.ca/educmat/phy456/456lec01.htm accessed August 2010].

Kelly, L.A., & Sternberg, M. J. E. (2009). Protein structure prediction on the Web: a case study using the Phyre server. *Nature Protocols.* 4 (3) p. 363-371.

Kim, D.E., Blum, B., Bradley, P., & Baker, D. (2010). Sampling bottlenecks in de novo protein structure prediction. *Journal of Molecular Biology.* 393(1) p. 249-260.

Kollman, P.A., & Allen, L.C. (1972). Theory of the hydrogen bond. *Chemical Reviews.* 72(3) p. 283-303.

Kortemme, T., Morozov, A.V. & Baker, D. (2003). An Orientation-dependent Hydrogen Bonding Potential Improves Prediction of Specificity and Structure for Proteins and Protein, Protein Complexes. *Journal of Molecular Biology.* 326 (4) p. 1239-1259.

Krissinel, E., Henrick., K. (2007). Inference of macromolecular assemblies from crystalline state. *Journal of Molecular Biology.* 372 (3) p. 774-797.

Kroon, J., Kanters, J.A. (1974). Non-linearity of hydrogen bonds in molecular crystals. *Nature.* 248 (5450) p. 667-669.

Kryshtafovych, A., Krysko, O., Daniluk, P., Dmytriv, Z. & Fidelis, K. (2009). Protein structure prediction center in CASP8. *Proteins.* 77 (Suppl 9) p. 5-9.

Landry, M. (1999) Protein Secondary and Tertiary Structure [www.tulane.edu/~biochem/med/second.htm accessed April 2011].

Lazaridis, T., Karplus., M. (1999a). Effective energy function for proteins in solution. *Proteins.* 35 (2) p. 133-152.

Lazaridis, T., Karplus, M. (1999b). Discrimination of the native from misfolded protein models with an energy function including implicit solvation. *Journal of Molecular Biology.* 288 (3) p. 477-487.

Lee, B., Richards, F.M. (1971). The interpretation of protein structures: Estimation of static accessibility. *Journal of Molecular Biology.* 55 (3) p. 379-400.

Levitt, M. (1998). Potential energy function and parameters for simulations of the molecular dynamics of proteins and nucleic acids in solution. *Computer Physics Communications.* 91 (1-3) p. 215-231.

Levitt, M., Perutz, M.F. (1988). Aromatic rings act as hydrogen bond acceptors. *Journal of Molecular Biology.* 201 (4) p. 751-754.

Li, H., & Zhou, Y. (2005). Folding helical proteins by energy minimization in dihedral space and a DFIRE-based statistical energy function. *Journal of Bioinformatics and Computational Biology.* 3 (5) p. 1151-1170.

Lindahl, E., Hess, B., & Van Der Spoel, D. (2001). GROMACS 3.0: a package for molecular simulation and trajectory analysis. *Journal of Molecular Modeling.* 7 (8) p. 306-317.

Lu, M., Dousis, A.D., & Ma, J. (2008). OPUS-PSP: an orientation-dependent statistical all-atom potential derived from side-chain packing. *Journal of Molecular Biology.* 376 (1) p. 288-301.

Ma, J. (2009). Explicit orientation dependence in empirical potentials and its significance to side-chain modeling. *Accounts of Chemical Research.* 42 (8) p. 1087-1096.

Martin, M.G. (2006). Comparison of the AMBER, CHARMM, COMPASS, GROMOS, OPLS, TraPPE and UFF force fields for prediction of vapor, liquid coexistence curves and liquid densities. *Fluid Phase Equilibria.* 248 (1) p. 50-55.

McDonald, I.K., & Thornton, J.M. (1994). Satisfying hydrogen bonding potential in proteins. *Journal of Molecular Biology.* 238 (5) p. 777-793.

Mimna, R., Camus, M-S., Schmid, A., Tuchscherer, G.,Lashuel, H.A., & Mutter, M. (2007). Disruption of amyloid-derived peptide assemblies through the controlled induction of a beta-sheet to alpha-helix transformation: application of the switch concept. *Angewandte Chemie International Edition.* 46 (15) p. 2681-2684.

Morozov, A.V., Kortemme, T., Tsemekhman, K., & Baker, D. (2004). Close agreement between the orientation dependence of hydrogen bonds observed in protein structures and quantum mechanical calculations. *Proceedings of the National Academy of Sciences of the United States of America.* 101 (18) p. 6946-6951.

Mottamal, M., & Lazaridis, T. (2005). The contribution of C alpha-H.O hydrogen bonds to membrane protein stability depends on the position of the amide. *Biochemistry.* 44 (5) p. 1607-1613.

Moult, J. (2005). A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. *Current Opinion in Structural Biology.* 15 (3) p. 285-289.

Myers, J., & Pace, C.N. (1996). Hydrogen bonding stabilizes globular proteins. *Biophysical Journal.* 71 (4) p. 2033-2039.

Nada, H., & Van Der Eerden, J.P.J.M. (2003). An intermolecular potential model for the simulation of ice and water near the melting point: A six-site model of H[sub 2]O. *The Journal of Chemical Physics.* 118 (16) p. 7401.

Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B., Thornton, J.M. (1997). CATH--a hierarchic classification of protein domain structures. *Structure London England 1993.* 5 (8) p. 1093-1108.

Park, B.H., & Levitt, M. (1996). Energy functions that discriminate X-ray and near native folds from well-constructed decoys. *Journal of Molecular Biology.* 258 (2) p. 367-392.

PDB statistics. (2010). [www.pdb.org accessed January 2010].

Pearlman, D.A., Connelly, P.R. (1995). Determination of the differential effects of hydrogen bonding and water release on the binding of FK506 to native and Tyr82-->Phe82 FKBP-12 proteins using free energy simulations. *J. Mol. Biol*. 248 (3). p. 696-717.

Pimentel, G.C., & McClellan, A.L. (1971). Hydrogen bonding. *Annual Review of Physical Chemistry.* 22 p. 347-385.

Privalov, P.L., Gill, S.J. (1988). Stability of protein structure and hydrophobic interaction. *Advances in Protein Chemistry*. 39. p. 191-234.

Protein Data Bank. (2010). HIV integrase protein, code 3OVN. [www.pdb.org/pdb/explore.do?structureId=3OVN accessed June 2010].

Radzicka, A., Pedersen, L., & Wolfenden, R (1988). Influences of solvent water on protein folding: free energies of solvation of cis and trans peptides are nearly identical. *Biochemistry.* 27 (12) p. 4538-4541.

Rahman, A., & Stillinger, F.H. (1973). Hydrogen bond patterns in liquid water. *Journal of the American Chemical Society.* 95 (24) p. 7943-7948.

Rajgaria, R., Mcallister, S.R., & Floudas, C.A. (2006). A novel high resolution Calpha-Calpha distance dependent dorce dield based on a high quality decoy set. *Proteins.* 65 (3) p. 726-741.

Ramachandran, G.N., & Sasisekharan, V. (1968). Confirmation of polypeptides and proteins. *Advances in Protein Chemistry*. 23. p. 283-438.

Richardson, J.S. (1981). The Anatomy and Taxonomy of Protein Structure. *Advances in Protein Chemistry.* 34 (167) p. 167-339.

Rykunov, D., & Fiser, A. (2010). New statistical potential for quality assessment of protein models and a survey of energy functions. *BMC Bioinformatics.* 11 (1) p. 128.

Sali, A., & Blundell, T.L. (1993). Comparative protein modelling by satisfaction of spatial restraints. J. Mol. Biol. 234, p 779-815.

Samudrala, R., & Levitt, M. (2000). A comprehensive analysis of 40 blind protein structure predictions. *BMC Structural Biology.* 2 p. 3.

Samudrala, R., & Moult, J. (1998). An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. *Journal of Molecular Biology.* 275 (5) p. 895-916.

Schaeffer, R.D., Fersht, A., & Daggett, V. (2008). Combining experiment and simulation in protein folding: closing the gap for small model systems. *Current Opinion in Structural Biology.* 18(1) p. 4-9.

Scheraga, H.A., Khalili, M., & Liwo, A. (2007). Protein-folding dynamics: overview of molecular simulation techniques. *Annual Review of Physical Chemistry.* 58(1) p. 57-83.

Schonbrun, J., and Dill, K.A. (2003). Fast protein folding kinetics. *Proceedings of the National Academy of Sciences of the United States of America.* 100 (22) p. 12678-12682.

Shaw, D.E., Chao, J.C., Eastwood, M.P., Gagliardo, J., Grossman, J.P., Ho, C.R., Ierardi, D.J., Kolossváry, I. (2007). Anton, a special-purpose machine for molecular dynamics simulation. *Proceedings of the 34th annual international symposium on Computer architecture.* 35 (2).

Shen, M., & Sali, A. (2006). Statistical potential for assessment and prediction of protein structures. *Protein Science.* 15 (11) p. 2507-2524.

Simons, K.T., Strauss, C. & Baker, D. (1997). Prospects for ab initio protein structural genomics. *Journal of Molecular Biology.* 306 (5) p. 1191-1199.

Sippl, M.J. (1990). Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. *Journal of Molecular Biology.* 213 (4) p. 859-883.

Sippl, M.J. (1995). Knowledge-based potentials for proteins. *Current Opinion in Structural Biology.* 5(2) p. 229-35.

Steiner, T. (1995). Water molecules which apparently accept no hydrogen bonds are systematically involved in C-H.O interactions. *Acta Crystallographica Section D Biological Crystallography.* 51 (Pt 1) p. 93-97.

Steiner, T. (1997). Unrolling the hydrogen bond properties of C-H.O interactions. *Chemical Communications.* p. 727-734.

Steiner, T. (2003). C-H---O hydrogen bonding in crystals. *Crystallography Reviews.* 9 (2) p. 177-228.

Steiner, T., & Saenger, W. (1993). Distribution of observed C-H bond lengths in neutron crystal structures and temperature dependence of the mean values. *Acta Crystallographica Section A Foundations of Crystallography.* 49 (3) p. 379-384.

Stella, L., & Melchionna, S. (1998). Equilibration and sampling in molecular dynamics simulations of biomolecules. *The Journal of Chemical Physics.* 109 (23) p. 10115.

Sutor, D.J. (1962). The C-H.O hydrogen bond in crystals. *Nature.* 195 p. 68-69.

Sutor, D.J. (1963). Evidence for the existence of C-H O hydrogen bonds in crystals. *Journal of the Chemical Society.* p. 1105-1110.

Taylor, W.R. (2006). Decoy models for protein structure comparison score normalisation. *Journal of Molecular Biology.* 357 (2) p. 676-699.

Tsai, J., Bonneau, R., Morozov, A.V., Kuhlman, B., Rohl, C.A., & Baker, D. (2003). An improved protein decoy set for testing energy functions for protein structure prediction. *Proteins.* 53(1) p. 76-87.

Wahl, M.C., Sundaralingam, M. (1997). C-H---O hydrogen bonding in biology. *Trends in Biochemical Sciences.* 22 (3) p. 97-102.

Wang, G., & Dunbrack, R.L. (2003). PISCES: a protein sequence culling server. *Bioinformatics.* 19 (12) p. 1589-1591.

Wikimedia. (2010). Hydrogen bonds in water. [en.wikipedia.org/wiki/File:3D_model_hydrogen_bonds_in_water.svg accessed February 2011]

Williams, M.A., Goodfellow, J.M., & Thornton, J.M. (1994). Buried waters and internal cavities in monomeric proteins. *Protein Science.* 3 (8) p. 1224-1235.

Yang, Y., & Zhou, Y. (2008). Specific interactions for ab initio folding of protein terminal regions with secondary structures. *Proteins.* 72 (2) p. 793-803.

Yang, Y., & Zhou, Y. (2008). Ab initio folding of terminal segments with secondary structures reveals the fine difference between two closely related all-atom statistical energy functions. *Protein Science.* 17 (7) p. 1212-1219.

Zhang, C., Liu, S., Zhou, H., & Zhou, Y. (2004). An accurate, residue-level, pair potential of mean force for folding and binding based on the distance-scaled, ideal-gas reference state. *Protein Science.* 13 (2) p. 400-411.

Zhang, Y., Skolnick, J., (2004). The Dependence of All-Atom Statistical Potentials on Structural Training Database. *Biophysical Journal.* 86(6) p. 3349-3358.

Zhang, Y., Skolnick., J. (2005). The protein structure prediction problem could be solved using the current PDB library. *Proceedings of the National Academy of Sciences of the United States of America.* 102 (4) p. 1029-1034.

Zhang, Y. (2008). Progress and challenges in protein structure prediction. *Current Opinion in Structural Biology.* 18 (3) p. 342-348.

Zhou, H., & Skolnick, J. (2007). Ab initio protein structure prediction using chunk-TASSER. *Biophysical Journal.* 93 (5) p. 1510-1518.

Zhou, H., & Zhou, Y. (2002). Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Science.* 11 (11) p. 2714-2726.

Zielkiewicz, J. (2005). Structural properties of water: comparison of the SPC, SPCE, TIP4P, and TIP5P models of water. *The Journal of Chemical Physics.* 123 (10) p. 104501.