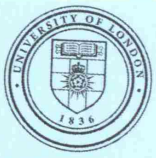


UNIVERSITY OF LONDON THESIS

Degree PhD Year 2007 Name of Author PETTITT, Christopher Steven

Refinement of Protein Structure Models with Multi-Objective Genetic Algorithms

Christopher Steven Pettitt

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

of the

University of London.

Department of Computer Science

University College London

2007

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

I, Christopher Steven Pettitt, 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

Here I investigate the protein structure refinement problem for homology-based protein structure models. The refinement problem has been identified as a major bottleneck in the structure prediction process and inhibits the goal of producing high-resolution experimental quality structures for target protein sequences. This thesis is composed of three investigations into aspects of template-based modelling and refinement.

In the primary investigation, empirical evidence is provided to support the hypothesis that using multiple template-based structures to model a target sequence can improve the quality of the prediction over that obtained solely by using the single best prediction. A multi-objective genetic algorithm is used to optimize protein structure models by using the structural information from a set of predictions, guided by various objective functions. The effect of multi-objective optimization on model quality is examined.

A benchmark of energy functions and model quality assessment methods is performed in the context of automated homology modelling to assess the ability of these methods at discriminating nearer-native structures from a set of predictions. These model quality assessment methods were unable to significantly improve the ranking of threading-based prediction methods though some model quality assessment methods improved model selection for methods which use sequence information alone. The results suggest that structural informational can provide valuable information for distinguishing better models where only sequence information has been used for modelling. The suitability of these energy functions for high-resolution refinement is discussed.

Finally, a stochastic optimization algorithm is developed for refining homology-based protein structure models using evolutionary algorithms. This approach uses multiple

structural model inputs, conformational sampling operators, and objective functions for guiding a search through conformational space. Single- and multi-objective genetic variants are applied to homology model predictions for 35 target proteins. The refinement results are discussed and the performance of both algorithmic variants compared and contrasted.

Acknowledgements

To my family, without whose love and support none of this would be possible.

I would like to thank David Jones for his help, supervision, and encouragement throughout my time at University College London, and especially for the stimulating scientific discussions that directed me out of many a scientific *cul de sac*.

Further thanks go to the members of the Jones Lab, in particular Kevin Bryson and Michael Sidowski for their helpful and entertaining scientific discussions, and to Ching-Wai Tan, Anna Lobley, Melissa Pentony, Daniel Roden, Jon Ward, and Jaz Sodhi. I would also like to acknowledge David Shortle, George Rose, and Douglas Theobald for their helpful and insightful scientific correspondence, and would like to thank Jan Jennings and Lisa Wainer for their emotional support.

Finally, I'd like to thank Liam McGuffin for his encouragement and advice throughout the PhD, and all those who helped with the proofing of this manuscript.

This work was sponsored by the Biotechnology and Biological Sciences Research Council.

Contents

List of Figures

List of Tables

List of Symbols

n	number of parameters
k	number of objective functions of an MOP
m	number of constraints of an SOP or MOP
n	number of decision variables of an SOP or MOP
q	tournament size
N	population size
$\bar{N}$	archive population size
P	population
T	maximum number of generations
$\boldsymbol{e}$	vector of constraints of an SOP or MOP
f	objective function of an SOP
$\boldsymbol{f}$	vector of objective functions of an MOP
P_c	crossover rate
P_m	mutation rate
$\boldsymbol{x}$	decision vector
$\boldsymbol{y}$	objective vector
$\boldsymbol{A}$	global or local Pareto front
$p(\boldsymbol{A})$	set of decision vectors in $\boldsymbol{A}$ non-dominated regarding A
$\boldsymbol{X}$	decision space
$\boldsymbol{X}_f$	set of feasible decision vectors
$\boldsymbol{X}_p$	set of Pareto-optimal decision vectors
$\boldsymbol{Y}$	objective space
$\boldsymbol{Y}_f$	set of objective vectors corresponding to $\boldsymbol{X}_f$
$\boldsymbol{Y}_p$	set of objective vectors corresponding to $\boldsymbol{X}_p$

Abbreviations

BFGS	Broyden-Fletcher-Goldfarb-Shanno
CAFASP	Critical Assessment For Automated Structure Prediction
CASP	Critical Assessment of Structure Prediction
CATH	Protein Structure Classification (Orengo *et al.*, 1997)
DFIRE	Distance-scaled Finite Ideal-gas REference state (Zhou *et al.*, 2002)
GA	Genetic Algorithm
MC	Monte Carlo
MD	Molecular Dynamics
MM	Molecular Mechanics
MOEA	Multi-Objective Evolutionary Algorithm
MOGA	Multi-Objective Genetic Algorithm
MOP	Multi-Objective Optimization Problem
NMR	Nuclear Magnetic Resonance
NSGA	Non-dominated Sorting Genetic Algorithm (Deb, K., 2002)
PDB	Protein Data Bank
REMC	Replica Exchange Monte Carlo
RMSD	Root Mean Squared Deviation
SA	Simulated Annealing
SCOP	Structural Classification Of Proteins (Murzin *et al.*, 1995)
SOP	Single-objective Optimisation Problem
SPEA	Strength Pareto Evolutionary Algorithm (Zitzler *et al.*, 2002)
SVD	Singular Value Decomposition

Chapter 1

Introduction

1.1 Protein structure

Protein molecules are remarkably versatile with respect to the numbers of structures they can form and the resulting functions which these molecular conformations enable. In the mid-1950's Christian Anfinsen discovered that the information determining the tertiary structure of a protein resides in the chemistry of its amino acid sequence. Further experiments, in which he demonstrated that a denatured protein could spontaneously refold from its unfolded to its native state, the $U \rightleftharpoons N$ transition, led to the development of the "thermodynamic hypothesis". This hypothesis states that, under physiological conditions, proteins spontaneously adopt their native conformation, the state at which they are thermodynamically most stable and attain a minimum in Gibbs free energy. For this discovery he was awarded the 1972 Nobel prize in chemistry and later published the seminal work, *Principles that govern the folding of protein chains* (Anfinsen 1973). The field of protein structure prediction grew from this basic premise, that the amino acid sequence solely determines the three-dimensional structure of a protein.

1.1.1 The twenty amino acids

Proteins exhibit an array of diverse functions yet all share a common structural feature; they are all linear polymers of amino acids. There are twenty amino acids provided by the standard genetic code (see Table 1.1) though hundreds more are found in nature. All amino acids contain amino and carboxyl groups, and for the most common α-amino acids, these are attached to the α-carbon (see Figure 1.1). All amino acids (with the

exception of glycine) also have an R group, or *side-chain*, attached to the α-carbon, and it is the composition of the side-chain that confers each amino acid with its chemical properties (see Table 1.1).

Amino Acid	Abbreviation	Letter	Properties
Alanine	Ala	A	Aliphatic, hydrophobic
Arginine	Arg	R	Hydrophilic, basic
Asparagine	Asn	N	Hydrophilic
Aspartic Acid	Asp	D	Hydrophilic, acidic
Cysteine	Cys	C	Sulfurous, hydrophobic
Glutamine	Gln	Q	Hydrophilic
Glutamic Acid	Glu	E	Hydrophilic, acidic
Glycine	Gly	G	Amphiphilic
Histidine	His	H	Hydrophilic, basic, imidazole
Isoleucine	Ile	I	Aliphatic, hydrophobic
Leucine	Leu	L	Aliphatic, hydrophobic
Lysine	Lys	K	Hydrophilic, basic, amine
Methionine	Met	M	Sulfurous, hydrophobic
Phenylalanine	Phe	F	Aromatic, hydrophobic
Proline	Pro	P	Aliphatic, imino acid
Serine	Ser	S	Hydrophilic
Threonine	Thr	T	Hydrophilic
Tryptophan	Trp	W	Aromatic, Hydrophobic
Tyrosine	Tyr	Y	Aromatic, Hydrophobic
Valine	Val	V	Hydrophobic

Table 1.1: The twenty standard amino acids and their properties (Koolman et al. 2005).

There are several ways to classify the amino acids from their physico-chemical properties (Sneath 1966, Grantham 1974, Livingstone & Barton 1993, Mocz 1995, Stanfel 1996), including deriving physico-chemical similarities using amino acid mutation frequencies obtained from the observed amino acid substitutions in alignments of highly similar sequences (Dayhoff et al. 1978, Taylor 1986). In addition to their chemical properties, amino acids have precise structural properties conferred on them

by their stereochemistry.

1.1.2 Stereochemistry

Stereochemistry describes the spatial arrangement of atoms within molecules, and for amino acids, has been well defined in both the classic literature (Pauling 1960) and more recently by small-molecule X-ray crystallography experiments (Engh & Huber 1991). In some cases, interesting functions necessitate the production of strained conformations that result in non-ideal geometries, however, large deviations from ideal geometries are rare, mainly due to the large energy requirements for stretching or bending a bond from its optimal separation, and often indicate errors in the structure accumulated in the crystallographic refinement process. Ideal bond length and bond angle parameters have been used extensively for the validation of protein structures (Laskowski et al. 1993, Hooft et al. 1999), for structure prediction (Šali & Blundell 1993), and as restraints in crystallographic and NMR refinement (Brunger et al. 1998, Murshudov et al. 1997).

1.1.3 The main chain and the ϕ, ψ, ω dihedral angles

Proteins are polypeptide chains that are formed through the joining of amino acids end-to-end during protein synthesis. Amino acids are chiral molecules (except glycine) and contains a common 'core' set of atoms, the central carbon atom (C_α) which forms the chiral center attached to which are a hydrogen, amide nitrogen (NH_2), and carboxyl group ($COOH$) (see Figure 1.1). The linkage of amino acids occurs through the formation of a peptide bond when the carboxyl group of one amino acid condenses with the amino group of the next (Branden & Tooze 1999).

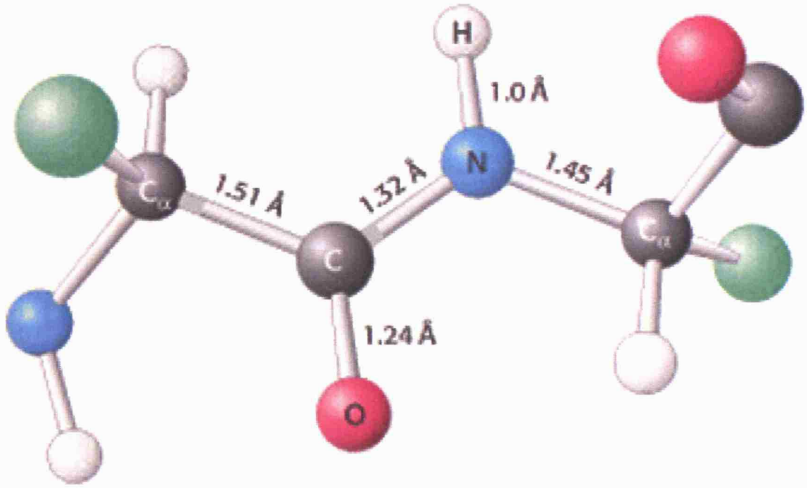


Figure 1.1: The main chain annotated with standard nomenclature (from Berg et al. 2002). The peptide unit consists of a central carbon atom (C_α) to which are bound an amino group (nitrogen (blue) and hydrogen (white)), a carboxyl group (carbon (*C* shown in black) and oxygen (pink)), and the side-chain R groups (green). The main chain is shown centered on the peptide bond (C'-N), and standard bond lengths are annotated.

The remarkable variety of structural conformations that arise from the same basic chemical structure is due to the three degrees of rotational freedom about the bonds. These ϕ, ψ, and ω dihedral angles are the major determinants of the conformational shape of a protein backbone (see Figure 1.2) and together these angles define the conformation of a residue (Lesk 1991). The ϕ angle quantifies the rotation about the $N - C_\alpha$ bond, while the ψ angle describes rotation about the $C_\alpha - C'$ bond. Due to the partially double-bonded character of the peptide bond ($C' - N$) (Corey & Pauling 1953), the six backbone atoms of the peptide unit ($-C_\alpha - CO - NH - C_\alpha -$) are largely co-planar (see Figure 1.2) resulting in restricted rotational movement about the peptide bond. As a result, the primary degrees of freedom are found in the ϕ and ψ dihedral angles of the protein backbone.

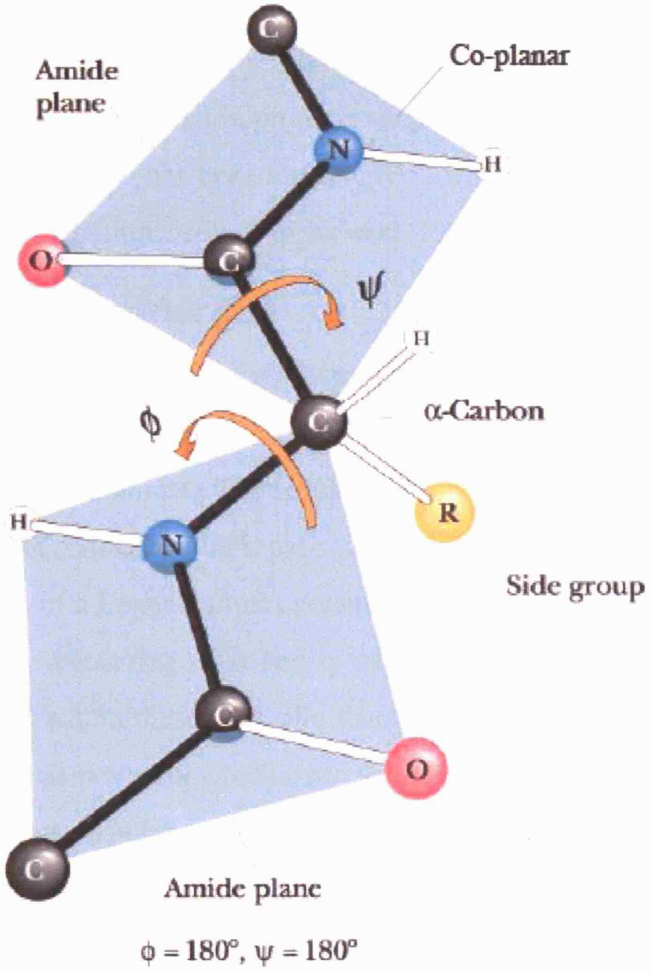


Figure 1.2: The main chain is shown highlighting the primary rotational degrees of freedom, the ϕ/ψ angles (from Berg et al. 2002). The rotational freedom of the peptide bond is restricted by its double bonded character and produces a co-planar arrangement of the peptide unit.

In principle, the number of conformations a main chain can adopt is vast. In reality, the weak non-covalent bonds formed between amino acid backbone and side-chain atoms reduces the conformational space available to the polypeptide chain. Moreover, the stiff repulsive forces that result from an overlap of electron clouds between atoms in the peptide unit of two interacting residues, further limits the number of configurations a main chain may adopt.

1.1.3.1 Configurations of the peptide bond

The ω angle, which describes rotation about the peptide bond, is most often found in one of two states: the *trans* conformation, or the *cis* conformation. This two state restriction arises due to the partially double-bonded character of the peptide bonds.

The additional effect of steric interactions among successive residues means the *trans* conformation, in which $\omega \approx 180°$, is preferred to the *cis* conformation ($\omega \approx 0°$) with 99.96% of conformations observed in proteins adopting the *trans* orientation (Stewart et al. 1990) though this view has been challenged, attributing *cis* conformations with important folding and functional roles (Weiss et al. 1998).

1.1.4 The Ramachandran plot

In 1968, Ramachandran and colleagues (Ramachandran et al. 1963, Ramachandran & Sasisekharan 1968) determined the range of possible clash free ϕ/ψ angle configurations for a dipeptide model. The resulting Ramachandran plot (see Figure 1.3) delineates the range of possible ϕ and ψ pairs which do not result in hard sphere (i.e. the repulsive component of a Lennard-Jones potential (see Section 1.3.3.1) steric overlap. For amino acids, the clustering of ϕ and ψ pairs found in experimentally resolved protein structures lie within these sterically favourable regions of the Ramachandran plot and highlights the powerful predictive capability of this discovery (see Figure 1.3). In fact, the Ramachandran plot is now used extensively to verify and validate the quality of X-ray structures (Morris et al. 1992, Swindells et al. 1995, Kleywegt & Jones 1996, Hooft et al. 1997, Wilson et al. 1998, Lovell et al. 2003). The original Ramachandran boundaries were resolved using the hard sphere model (Richards 1977) and identifies two major populated regions corresponding to backbone dihedral angles most commonly found in α-helices and β-strands (Levitt & Chothia 1976). Since then the Ramachandran plot has been further refined to improve the boundary resolution between energetically favourable and unfavourable regions of the map (Kleywegt & Jones 1996, Hovmöller et al. 2002, Lovell et al. 2003).

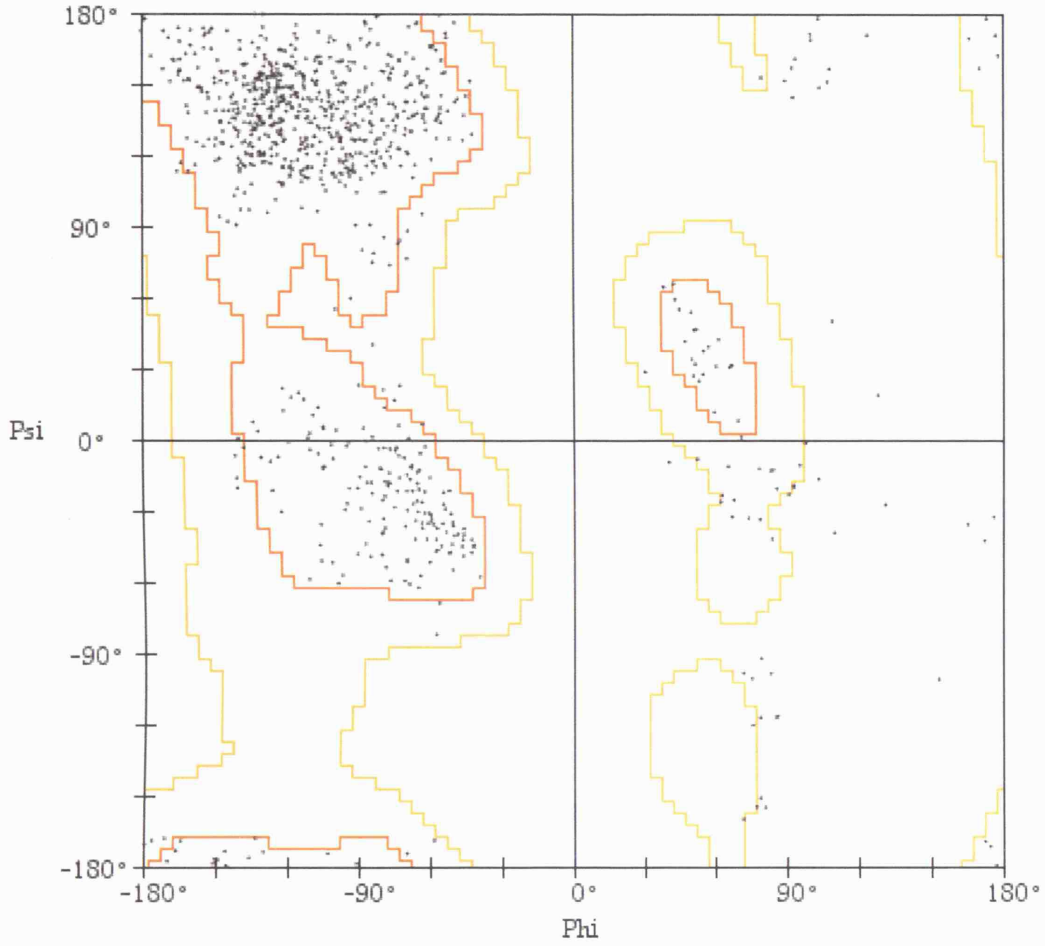


Figure 1.3: An illustration of the combined Ramachandran plots for all twenty amino acids (from Lovell et al. 2003). Regions in orange represent favourable (ϕ, ψ) pairs. The larger regions delineated in yellow show the sterically plausible, though less favourable, (ϕ, ψ) values, while regions outside these two areas are generally unfavourable.

1.1.4.1 Glycine and proline

Glycine has a single hydrogen side-chain and as a result can adopt a large number of clash free conformations not available to the other amino acids. The structural importance of glycine allows the main chain to adopt unusual conformations and is one of the main reasons why a high proportion of glycine residues are conserved in homologous proteins (Branden & Tooze 1999). As a result, the Ramachandran plot for glycine has significantly different characteristics from the other amino acids, while the side-chain of proline residues are covalently linked to the backbone nitrogen, thus greatly restricting in the range of conformations they can occupy.

1.1.5 The side-chains

The twenty amino acids are differentiated by their side-chains, and each side-chain is covalently bound to the C_α atom of the peptide backbone [1]. The asymmetry of the C_α dictates that amino acids (except glycine) are chiral molecules and can therefore exist in either the *L*-isomer or *D*-isomer (Branden & Tooze 1999). The *L*-form is almost exclusively found in biological molecules though there is no real understanding of why one form was favoured over the other. Side-chain atom arrangements are described by between one and four dihedral angles (known as χ angles) depending on the length and size of the side-chain. These χ angles describe the magnitude of rotation about the side-chain bonds (see Figure 1.4).

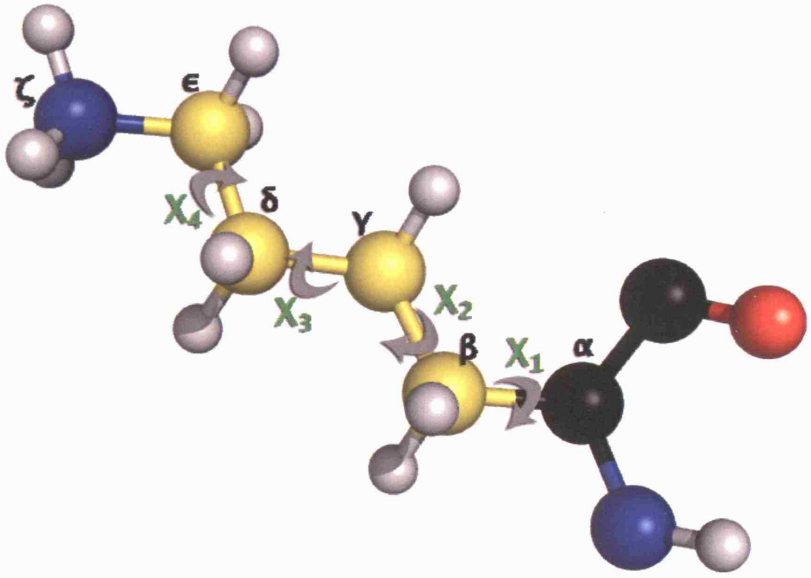


Figure 1.4: Side-chain conformations are defined by their χ dihedral angles, shown in this case, for a lysine residue. The standard side-chain atom labels $(\beta, \gamma, \delta, \varepsilon, \zeta)$ are given without reference to the specific side-chain atom type.

Although side-chains can potentially adopt a wide variety of angular configurations of their χ torsion angles, in reality they generally tend to cluster around a particular subset of values. This phenomenon, known as rotamericity, emerges from the energetic effects that arise from two tetrahedrally coordinated carbon atoms (Ponder & Richards

[1] proline is a special case in which the side-chain also forms a covalent bond with the main chain nitrogen, while glycine lacks any side-chain heavy atoms

1987). Generally, staggered conformations between two tetrahedral or trigonal bound carbon atoms are more energetically favourable and as a result, off-rotamer side-chain conformations are rarely found (Branden & Tooze 1999). The seemingly discrete nature of the χ angle forms the foundation for side-chain modelling programs which sample directly from a set of discrete rotamer states in compiled libraries of conformations found in crystal structures (Bower et al. 1997, Dunbrack 2002). Off-rotamer conformations often have unfavourable torsional energies and van der Waals overlaps, further reducing their frequency of occurrence (Petrella & Karplus 2001). This observation led to the development of rotamer libraries which represent the majority of low-energy side-chain conformations as discrete states (Bower et al. 1997, Dunbrack & Karplus 1993).

1.2 Protein folding

Protein folding describes the spontaneous and reversible disorder $\rightleftharpoons$ order transition that takes place under suitable physiological conditions, and guides the self-assembly process by which a disordered linear polymer adopts its unique tertiary structure. A complete characterisation of the nature of the folding process for proteins is still lacking, yet a generally accepted description of the $U \rightleftharpoons N$ transition has been provided in the form of the thermodynamic hypothesis (Mirsky & Pauling 1936, Anfinsen 1973). This hypothesis states that a solution of unfolded proteins, in which the population assumes a remarkable variety of conformations, will spontaneously be driven towards the conformation that minimises the Gibbs free energy of the system to produce a unique native state in which both intramolecular interactions, and interactions with the surrounding solvent, are optimised on shifting to favourable physiological conditions. This view from equilibrium thermodynamics provides a useful framework for understanding the folding reaction where the transition from the unfolded to the folded state is conceptually similar to a ball rolling down a steep hill into the valley below. Yet despite the powerful simplicity of the thermodynamic hypothesis, there is still no general theory which adequately describes the precise mechanism of the $U \rightleftharpoons N$ transition for proteins.

1.2.1 Levinthal's paradox

It was Cyrus Levinthal who first suggested that the folding transition, leading to the protein's native state, could not be a random search, and he demonstrated this by way of a simple illustration that later became know as "Levinthal's paradox" (Levinthal 1969). He calculated that each bond connecting amino acids could have three possible states, so that for protein containing 101 residues, there are $3^{100} \approx 10^{47}$ configurations. Even with bonds rotating at the subpicosecond speed limit of 10^{-13}, it would take approximately 10^{27} years (longer than the age of the universe) to reach the native state through an unguided search (Zwanzig et al. 1992). However, evidence that proteins fold in milliseconds, and even microseconds (Myers & Oas 2002), suggests that the underlying folding mechanism follows some directed search.

1.2.2 The folding funnel

To explain how the folding process leads from a denatured polypeptide chain to its unique native state Frauenfelder et al. (1991) proposed the folding funnel model of the energy landscape (see Figure 1.5). The shape of the landscape promotes a bias towards low energy states and enables a molecule to form favourable interactions that lower its energy thus promoting chain compaction which drives the molecule towards its native state cooperatively and with gathering speed (Davidson et al. 1995, Miranker & Dobson 1996, Dill & Chan 1997, Honig 1999, Plotkin & Onuchic 2002). This widely held view states that the residue side-chains are predominantly responsibly for folding (most amino acid backbones are chemically equivalent) and are selected by evolution in order to sculpt a landscape topography that results in minimal frustration on traversal to the global minimum. It is the evolutionary selection of sequences that enable these side-chain interactions to successfully navigate the landscape and converge on the native fold while avoiding many other plausible stabilising interactions on a presculpted energy landscape of alternative intrinsically stable domains (Chothia & Finkelstein 1990, Hoang et al. 2004).

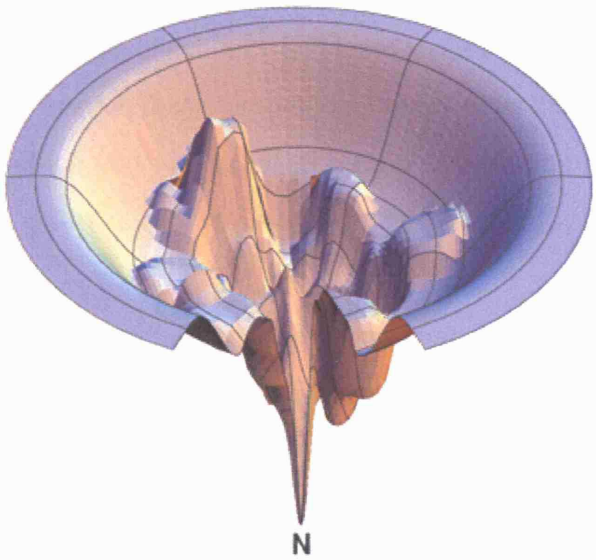


Figure 1.5: The energy landscape model of protein folding suggests a folding funnel shape which directs the folding protein into the native state without the need for a definite pathway. A number of alternative paths may be available in a rough energy landscape though (Dill & Chan 1997).

1.2.3 Energetics

As large molecular systems, proteins exhibit complex energetic behaviour (Creighton 1993). A protein, under physiological conditions, switches rapidly between the unfolded, highly unstable denatured state, and its native conformation. The dynamic equilibrium between states has enabled the study of the energetic determinants of protein folding and allowed physical chemists to dissect out the physical interactions which drive the formation of protein structure (Shortle 1996). Thermodynamic estimates of the free energy difference between the denatured and native state of a protein suggest a narrow range between -5 to -15 kcal/mol consistent with measured values for hundreds of proteins (Gromiha et al. 1999) showing proteins to be only marginally stable.

The free energy of a protein comprises both enthalpic and entropic contributions as well as free energy contributions arising from interactions with the surrounding solvent. The enthalpic contributions to the free energy are due to the stronger and more prevalent noncovalent interactions in the native state than the denatured state (Shortle 1996). This is offset by the entropic penalty of increased order in the native state relative to the favourable disorder of the denatured state (Branden & Tooze 1999). The

enthalpic and entropic contributions can both reach several hundred kcal/mol. The free energy difference between the native and denatured states i.e. the stability of the native state is the difference between these two large numbers. Our limited understanding of the denatured state, which plays an essential role in determining the entropic penalty, severely restricts our ability to characterise or predict the stability of protein structures (Zagrovic et al. 2002).

The interactions that stabilise the native state are much better understood than those in the denatured state. Covalent interactions, except the formation of disulfide bonds, are effectively identical in both the native and denatured states, and hence they contribute little to the stabilisation of the native state. Further, the extremely high energetic penalty for stretching or bending bonds means that the three-dimensional structure of a protein is determined by the much weaker non-bonded interactions that operate through space.

A major interaction between atoms in proteins is the short-range repulsion that develops as electron orbitals overlap when two atoms approach each other. The repulsive energy of these interactions increases extremely quickly with decreasing distance. Typically, these short-range repulsion forces are coupled to the weak, but attractive van der Waals interactions in a single energy potential, most commonly in the Lennard-Jones 6,12 form (Equation 1.1).

$$E(r) = \left(\frac{\sigma_R}{r}\right)^{12} - \left(\frac{\sigma_A}{r}\right)^{6} \tag{1.1}$$

where r is the distance between the two atoms, σ_R and σ_A are the repulsive and attractive constants, respectively.

Another fundamental force is the attraction (or repulsion) between un-like (or like) charges. Electrostatic forces are both strong, of several hundred kcal/mol, and operate over long distances, varying only with the inverse of the atomic separation. The electrostatic interaction between two atoms can be described by Coulombs law (Equation 1.2).

$$E(r_{ij}) = \frac{q_i q_j}{4\pi D \varepsilon_0 r_{ij}^2} \tag{1.2}$$

where r_{ij} is the distance between atoms i and j, q_i and q_j are charges on each atom,

ε_0 is the permittivity of free space, and D the dielectric constant. Because of their strength and long range electrostatic interactions play a major role in protein stability, interactions, and function. However, the electrostatic interaction is heavily modulated by solvent, which shields charged particles from the full force of their electrostatic interaction (Branden & Tooze 1999, Creighton 1993, Leech 2001).

Hydrogen bonds form when two electronegative atoms compete for the same hydrogen atom (Creighton 1993). The hydrogen is formally bound to the donor atom D, while it interacts favourably with the acceptor atom A. In proteins, oxygen atoms frequently participate as acceptors and nitrogens as donors, such as in the canonical secondary structures of the α-helix and β-sheet. Although each hydrogen bond contributes between $2-5$ kcal/mol, they are not thought to be major factors in protein stability, as equivalent numbers of hydrogen bonds form in both the native and denatured state (Branden & Tooze 1999).

The nonpolar side-chains of amino acids have fewer interactions with water than polar side-chains (Table 1.1), which causes nonpolar amino acids to greatly prefer nonpolar environments (Creighton 1993). The preference of non-polar groups for nonpolar, desolvated environments has been termed the *hydrophobic effect*. The free energy gain of removing nonpolar side-chains from the solvent and burying them in the protein core is significant and this *hydrophobic collapse* is thought to be a major factor in driving protein folding (Baldwin 1989, Buckle et al. 1993).

1.3 Protein structure modelling

Modelling and simulation methods have enhanced our understanding of the folding process (Levitt & Warshel 1975), protein energetics (Onuchic et al. 2000), protein design (Kuhlman et al. 2003), and provided modelling tools for predicting the structure of protein sequences with unknown structures (Blundell et al. 1987, Jones et al. 1992, Jones 1997). Generally, the problem of computationally modelling macromolecular structures can be separated into three distinct components: (*i*) representation, (*ii*) conformational sampling, and (*iii*) energy evaluation.

1.3.1 Molecular representation

Molecules can be represented at different levels of detail, from simple lattice-based models (where residues or atoms are assigned to a grid) to highly detailed descriptions with electronic degrees of freedom and explicit or implicit solvent models (Kolinski 2004). Due to the limitations set by current levels of computing power, computational modelling necessarily involves a trade-off between physical accuracy and tractability. Therefore, the most appropriate choice of representation often depends on the specific problem domain under investigation.

The simplest representations in which to explore molecular structures are lattice models where individual residues are represented by single points on a regular lattice. With few degrees of freedom, the number of possible moves for residues on a lattice are restricted to a small subset of the possible conformational positions. These highly incomplete representations allow for fast, but limited, exploration of the available conformational space but are severely restricted in their ability to adequately sample enough of the true conformational landscape (Yue et al. 1995, Park & Levitt 1995, Hinds & Levitt 1994). Models which are continuous and not confined to a grid (off-lattice models) are generally more detailed than their lattice counterparts, and explicitly represent protein-like geometry with greater conformational freedom. The level of detail in these models may vary from a simple C_α chain (Purisima & Scheraga 1984) that traces the path of the polypeptide chain - its fold - through space without explicitly representing other main chain or side-chain atoms, to models containing the full set of main chain heavy atoms (see Figure 1.1) and a C_β atom or virtual side-chain centroid (Levitt 1976, Sun 1993, Park et al. 1997, Mirny & Shakhnovich 2001). The most detailed representation available is that used in quantum-mechanical (QM) models which use electrons and nuclei as the smallest particles in the system. The computational cost of exploring conformational space with QM for even the smallest of proteins is prohibitively large and thus are too expensive for molecular simulations at the current level of computing technology (Becker et al. 2001).

For the majority of applications in protein structure prediction, atomic representations, with either the heavy atoms (united atom) or with heavy atoms including light hydrogen atoms (true all atom), are used. These models are capable of capturing the most important structural details of proteins such as their regular geometry (secondary

structure), molecular volume and surface, hydrogen bonding, and packing density.

1.3.2 Conformational sampling

The conformational sampling problem poses a significant challenge. Even for small proteins, enumerating all conformational states available to the chain is impossible except in cases where the number of amino acid residues is artificially small. Adding further complication, even the simplest potential energy function describes a complex and non-linear energy landscape which requires an exhaustive enumeration of conformational space in order to guarantee finding the lowest energy structure. Consequently, without full knowledge of the (possibly universal) mechanism which drives proteins to their native state, the challenge of conformational sampling becomes the ability to most efficiently select values for the free variable in a model which result in the lowest energy conformation in the shortest possible amount computational time.

The most successful sampling methods in protein folding and structure prediction have been stochastic algorithms such as monte carlo (MC) methods (Kawai et al. 1989) and their derivatives (Li & Scheraga 1987, Zhang & Skolnick 2006), simulated annealing (Kirkpatrick et al. 1983), discrete sampling (Moult & James 1986, Bruccoleri & Karplus 1987), and knowledge-based sampling (Jones & Thirup 1986, Holm & Sander 1991).

1.3.3 Energy functions

The energy of a macromolecular system can be described as a function of its atomic positions and the interaction of those atoms with the surrounding solvent. Calculating this energy effectively is central to the success of the computational modelling of biological macromolecules. Biological molecules are large and dynamic systems, therefore an energy function must be detailed enough to capture the detailed interactions between atoms yet tractable enough to compute many thousands of times is a simulation. For protein structure prediction, an accurate potential energy function must, at minimum, be able to distinguish between native and non-native conformations of protein structures (e.g. the native state should correspond to the global free energy minimum of the potential), and ideally, should be able to rank structure quality with a high energy correlation.

In general, energy functions can be categorised into two broad classes; (*i*) physics-

based energy functions, popular in molecular mechanics (MM) simulations (Karplus & Petsko 1990), and, (*ii*) statistical or knowledge-based potentials (Sippl 1990).

1.3.3.1 Physical Effective Energy Functions (PEEFs)

Physical Effective Energy Functions (PEEFs), or empirical energy functions, use simplified mathematical equations to model the physical interactions that dictate the structure and dynamic properties of biological molecules (Becker et al. 2001). These PEEFs are usually applied to atomic representations (see Section 1.3.1) helping to further increase the tractability of the simulation, and with optimised parameters, can often achieve chemical accuracy.

PEEFs are able to generate an energy surface for bonded and non-bonded interactions and the energy function must be in a form that can be differentiated for use with gradient-based minimization algorithms. The potential energy of a chemical system, $V(R)_{total}$, represented by the structure, R, can be separated into terms describing the internal, $V(R)_{internal}$, and external, $V(R)_{external}$, potential energy (Ponder & Case 2003).

$$[h]V(R)_{total} = V(R)_{internal} + V(R)_{external} \tag{1.3}$$

$$\begin{aligned} V(R)_{internal} = & \sum_{bonds} K_b(b-b_0)^2 \\ & + \sum_{angles} k_\theta(\theta-\theta_0)^2 \\ & + \sum_{dihedrals} k_\chi[1+cos(n\chi+\sigma)] \end{aligned} \tag{1.4}$$

$$V(R)_{external} = \sum_{nonbonded pairs} \left(\varepsilon_{ij} \left[\left(\frac{R_{min_{ij}}}{r_{ij}} \right)^{12} - \left(\frac{R_{min_{ij}}}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{\varepsilon_D r_{ij}} \right) \tag{1.5}$$

where the internal potential energy is associated with covalently connected atoms, and the external terms representing noncovalent, or nonbonded, interactions between atoms. The terms describing the 3D structure are the bond lengths, b; the valence angles, θ; the dihedral or torsion angles, χ; and the distances between the atoms, r_{ij}. Values for these terms are usually obtained from experimental data (Engh & Huber 1991). The remaining parameters such as the value of partial atomic charges, q, bond parameters, b_0, and temperature, K, allow different types of atoms and molecular connectivities to be treated with the same equations (Equation 1.4, Equation 1.5), and it

is the quality of these parameters that ultimately determines the accuracy of the results obtained in computational simulations (Becker et al. 2001).

There are a number of popular force field implementations used in molecular mechanics simulations, each with different compositions of the energy function, though the three major force fields, GROMACS (van Gunsteren & Berendsen 1990, Lindahl et al. 2001), CHARMM (Brooks et al. 1983, MacKerell et al. 1998), and AMBER (Weiner & Kollman 1981, Weiner et al. 1986, Cornell et al. 1995), all assume that the internal forces within proteins can be approximated accurately with pairwise interactions.

1.3.3.2 Statistical Effective Energy Functions (SEEFs)

The statistical analysis of contact distances between non-bonded atom or residues pairs observed in experimentally determined structures was first describe by Tanaka & Scheraga (1976), and has since been used to describe the dominant forces stabilising native globular protein structures. Statistical potentials have been used for a variety of tasks in structural biology (Vajda et al. 1997), including structure validation (Rojnuckarin & Subramaniam 1999), fold recognition (Jones et al. 1992), and loop modelling (de Bakker et al. 2003). These potentials have also become a popular choice for discriminating between native and non-native protein models (Sippl 1990).

Database potentials, expressed as a potential of mean force, W, between two interaction sites i and j, located in the distance range $r \pm \Delta r$ from each other, are usually justified in terms of the Boltzmann relationship

$$W_{ij}(r) = -RTln\left[\frac{P_{ij}(r \pm \Delta r)}{P_{xx}(r \pm \Delta r)}\right] \tag{1.6}$$

where $P_{ij}(r \pm \Delta r)$ is the probability of observing the specific pair (i, j) at separation $r \pm \Delta r$, $P_{xx}(r \pm \Delta r)$ is the corresponding reference probability, independent of residue type, R is the gas constant, and T is the absolute temperature. This justification in terms of the Boltzmann relationship requires a number of assumptions; the choice of the reference state, the application of Boltzmann statistics, the choice of the interaction sites to characterise pairwise interactions, and the validity of dividing the conformational space into finite intervals (Rojnuckarin & Subramaniam 1999, Sippl 1990).

The description of the reference state for a statistical potential is analogous to describing the standard state in thermodynamics which serves as a reference for interpreting the meaning of calculated values. Miyazawa & Jernigan (1985) describe a random mixing approximation to the reference state (termed the quasi-chemical approach), in which the expected number of contacts between a particular pair of species is directly proportional to their relative concentrations. This approach was found to be correct provided that interaction species are all the same size, but failed when considering all-atom representations with detailed side-chain packing (Skolnick, Jaroszewski, Kolinski & Godzik 1997). Recently, Chen & Shakhnovich (2005) proposed the "Gaussian approximation" reference state to adequately account for side-chain packing, and a number of authors have proposed additional reference states based on conditional probabilities (Samudrala & Moult 1998), propensities (Shortle 2003, Fang & Shortle 2005) and ideal gases (Zhou & Zhou 2002).

Most distance-dependent forms of SEEFs formulate their reference states over finite intervals in order to make calculations more tractable. This discretization ensures that statistical potentials are typically less sensitive to interaction details of the protein conformation than their molecular mechanics counterparts, and hence they have much smoother energy landscapes (Lazaridis & Karplus 2000). Consequently, reasonable statistical potential scores can be obtained by direct calculation on the candidate conformation without first minimising the conformation with respect to the statistical potential.

An additional issue resulting from equation 1.6 is that the potential energy is expressed as a sum of pairwise interactions and that each pair of specific residue or atom types (i, j) is assumed to behave independently, regardless of the chain connectivity, constraints imposed by specific sequential neighbors, and context or environmental conditions (Jernigan & Bahar 1996). This confers an additional advantage that, in principle, SEEFs should incorporate implicitly all important interaction contributions to the potential energy (Vajda et al. 1997) and is especially important for poorly understood interactions such as solvation.

One final fundamental assumption is that the experimental data set used to derive a SEEF is large enough to represent the full spectrum of inter-species energetics manifested in protein structures, and various authors have found evidence that the size

and composition of training databases has a large effect on the discriminatory power and specificity of some SEEFs (Furuichi & Koehl 1998, Rooman & Gilis 1998, Zhang et al. 2004).

While PEEFs have the advantage of resting on a firm theoretical basis and can therefore be used to study the contributions of different physical interactions and chemical forces on the folding process, their success in studying large macromolecules has been fairly limited to date. SEEFs have had far wider application, and the greater freedom on their functional form has made these potential energy functions useful for practical applications such as discriminating native from non-native structures, however, their relationship to true free energies remains controversial (Ben-Naim 1997).

1.3.3.3 Decoy sets

To measure the predictive power of a conformation selection function requires that the function be able to achieve two goals. Firstly, the mechanism, such as a molecular mechanics force field or a statistical potential, must be able to discriminate between native-like and non-native "decoy" conformations for the same sequence (Holm & Sander 1992*a*, Park & Levitt 1996, Moult 1997, Samudrala & Levitt 2000), and secondly, it should be able to rank the structures in order of the similarity to the native structure.

A good decoy set must have the following properties; (*i*) it must include a large number of conformations covering a broad spectrum of near-native conformations as well as conformations that have near-native energies but differ in the overall fold (Park et al. 1997, Samudrala & Levitt 2000), (*ii*) there must be some conformations within the native structure energy basin in order to test the detailed discriminatory power of a selection mechanism, (*iii*) the decoy set should contain a large variety of protein classes to avoid over-fitting an energy function to a specific class (e.g. α-proteins), (*iv*) the set should be generated independently from the evaluated scoring mechanisms, to avoid bias towards any particular selection methodology.

A large variety of methods have been used to generate decoy sets that fulfil the above requirements and large repository of these decoy conformations have been assembled for the structure prediction community (Samudrala & Levitt 2000).

Following the work of Novotny et al. (1984), many authors have generated decoy sets by threading an amino acid sequence onto the backbones of proteins of roughly equivalent size but with different folds (Holm & Sander 1992*a*). Decoys have been generated *ab initio* by enumeration on lattice models (Hinds & Levitt 1994), from molecular dynamics trajectories (Wang et al. 1995), Monte Carlo sampling (Monge et al. 1994), and discrete conformational search (Park & Levitt 1996, Samudrala & Moult 1998).

1.4 Structure prediction methods

Genome sequencing projects have provided a wealth of sequence data (more than 55 million unique sequences) for over 260,000 organisms (Kulikova et al. 2006), yet the number of experimentally determined protein structures numbers less than 38,000 (Sussman et al. 1998). Due to the high plasticity of sequence space, deriving accurate functional information from sequence alone is near impossible, as the probability of identifying relevant functional residues falls with the sequence identity between two proteins. Therefore, high-resolution structural models are necessary for characterising protein functions, from low-level descriptions of the fold category (Orengo et al. 1997, Murzin et al. 1995), to high-resolution descriptions of ligand binding sites (Sheng 1996) and functional pockets (Ring et al. 1993). Due to prohibitive costs and experimental difficulties with determining protein structures by X-ray crystallography or nuclear magnetic resonance (NMR), computational approaches are necessary to bridge the widening gap between sequence and structural data. Structure prediction methods generally approach the problem either from an evolutionary perspective (template modelling), or from a physical perspective (*ab initio/de novo* modelling).

The evolutionary approach relies on the observation that proteins can be classified into families that share similar sequences, structures and often related functions. This view has shown that function is usually of primary importance and so proteins evolve only gradually in order to retain their structure and function, often with a similar conservation of much of their sequences. These observations provide the foundation on which template modelling methods, such as comparative modelling and fold recognition, rest. Here, experimentally determined structures are used as a framework, or *template*, on which to build a model for a given query sequence provided

the sequence of a template structure is sufficiently similar to that of the query sequence (Blundell et al. 1987).

Ab initio, or *de novo*, methods rely on insights from thermodynamics as a basis for structure prediction, namely that the native state of a protein corresponds to the global free energy minimum for its sequence (see Section 1.2). In theory, no *a priori* information other than the knowledge of the physical forces (encoded in a objective function) are used to generate a structure prediction by search the resultant energy landscape using some efficient conformational sampling protocol (Bonneau & Baker 2001).

1.4.1 Comparative/Homology modelling

Comparative or homology modelling methods predict the structure of a protein by comparison of the target sequence to similar sequences for which protein structures are available. By the alignment of a sequence against a library of template protein sequences for which structure are experimentally resolved, the optimal alignment can be determined and the target sequence assigned to the structure of the sequence in the alignment.

In the early stages of sequence data analysis, sequences were compared using the pairwise alignment algorithms of Needleman & Wunsch (1970), a dynamic programming algorithm for generating optimal global alignment. Later this technique was extended by Smith and Waterman (Smith & Waterman 1981) to generate optimal local alignments. These early dynamic programming approaches were later superseded by faster techniques for large sequence database searching through the use of approximate methods such as FASTA (Pearson & Lipman 1988) and BLAST (Altschul et al. 1990).

The introduction of the sequence profile and iterative search lead to improved detection of distant sequence homologues in the 1990's. PSI-BLAST (Position Specific Iterative Basic Local Alignment Search Tool) (Altschul et al. 1997) has now become the standard tool for sequence searching due to its speed and sensitivity, and has been show to outperform traditional Smith-Waterman searches for remote sequence homologues (Jones 1999*b*).

The detection of a sequence homologue of known structure enables the construction of a model from the three-dimensional structure of the homologue (known

as the template). In general, the template with the highest sequence identity to the target indicates the most closely related protein, and this high scoring template is often used to build the model. More recently, methods for building models from multiple templates have been developed where more than one related structure is used in the construction of a model (Contreras-Moreira, Fitzjohn & Bates 2003) (for a full review of comparative modelling methods see (Marti-Renom et al. 2000)). Due to the approximate nature of fast sequence searching methods such as BLAST and PSI-BLAST, the sequences of the target and chosen template are often first re-aligned to generate an optimal pairwise alignment. Additional considerations such as the resolution of the template and the protein environment in which the template is found, are also examined to ensure maximal compatibility between the target sequence and the selected structure. Where sequence similarity is low, the structure of the template can also be used to improve the sequence-to-structure fit. By using secondary structure constraints, insertions of gaps into regular secondary structure elements such as α-helices or β-strands are dis-allowed (Sanchez & Šali 2000, Jones 2000).

The techniques for generating comparative models from templates usually employ rigid body assembly (Blundell et al. 1987), modelling by segment matching (Levitt 1992), and modelling by the satisfaction of spatial restraints (Šali & Blundell 1993). The accuracy of the final model is often confirmed by the examining the compatibility of the sequence to template match (Sippl 1993), and by examining the quality of stereochemistry (Laskowski et al. 1993, Hooft et al. 1999, Eisenberg et al. 1997).

1.4.2 Fold recognition

Experiments to assess the accuracy of comparative modelling signalled that the minimum level of sequence identity required for an accurate model was in the range of between 25-30%. This region defined the boundary at which automatic comparative modelling became less effective, and has been labelled the "twilight zone" (Doolittle 1981). In order to detect relationships between a target sequence below this level of sequence identity required novel methods which incorporate structural information.

Fold Recognition grew from the observation that the number of distinct structures in the protein data bank (PDB) grew at a disproportionately slower rate than the database as a whole. This led to the suggestion that only a finite number of fold

topologies were encoded by the millions of protein sequence in nature (Chothia 1992, Johnson et al. 1993). Indeed, it has been found that just 9 different folds (termed superfolds) may account for up to 30% of the known structures (Orengo et al. 1994), and that for up to 70% of new protein sequences, there will be a structure which adopts a similar topology from which a suitable model can be constructed (Jones 2000). As mentioned previously, sequence based search methods are unable to detect some of these templates due to a low sequence identity to any known structure.

Probably the first attempt to relate sequences to folds in the absence of sequence homology was made Ponder & Richards (1987). Following from this work, Bowie et al. (1990) realised that structural information could be used in an analogous way to multiple sequence information in profile methods (Bowie et al. 1991). Through the alignment and scoring of target sequences against the complete library of known folds, a measure of the compatibility of query sequence to each fold could be ascertained. The premise that protein structure is more conserved than sequence led to the assumption that the structural environment surrounding an amino acid would be more conserved that the actual amino acid type itself. By encoding the three-dimensional structural environment into one-dimensional strings or profiles, alignments could then be generated using conventional dynamic programming algorithms.

Although respectable, the methodology of 3D-1D methods - defining structural environments or residue classes - meant that much of the structural context information comprising the specificity of sequence to structure matches was lost. The first solution to this problem was provided by the 'threading' method of Jones et al. (1992). An alignment using the double dynamic programming algorithm of Taylor and Orengo (Taylor & Orengo 1989), is used to score the pairwise interactions at each equivalent residue position by minimising an empirical pairwise distance potential (Sippl 1990), hence taking into consideration the detailed network of interactions between individual residues. This approach is known as threading because it attempts to evaluate the sequence in three-dimensions as it is threaded through a library of structures. In general, the most successful fold recognition methods have been those based around the threading technique (Murzin 1999). A number of related methods for fold recognition were developed throughout the 1990's (Godzik et al. 1992, Bryant 1996, Thiele et al. 1999). Most employ some variation of iterative dynamic programming algorithms to

build models combined with an analysis of pairwise interactions between structurally adjacent residues.

1.4.3 *Ab Initio*/new folds

Fold recognition methods fail to select the correct fold from a library for approximately 50% of cases where no significant sequence similarity exists. In addition, no novel folds can be predicted using these fold recognition techniques. Therefore, template-free modelling methods have been devised to predict the structure of novel folds *ab initio*, from first principles. Traditionally, physics-based approaches were used to explore the conformational space of a target peptide during a folding simulation where iterative sampling of conformations is guided using an energy function. In reality, the vast number of conformations the protein can adopt, and the difficulty of generating accurate energy functions has seen these physics-based approaches largely superseded.

In place of these physical approaches to structure prediction, novel algorithms have been successfully developed for generating low energy conformations using fragments of known structures. The success of these methods may largely be due to the fact the PDB may contain nearly all possible conformational substructures (Du et al. 2003). In general, structural fragments are selected based on their compatibility to the local target sequence and to secondary structure propensities. A small conformational library of possible fragments is generated for each substructure, as often the sequence to structure relationship is not strong enough to describe the fragment completely (Bystroff et al. 1996). The relationship between substructures may be determined by approximate potentials of residue contacts (Aloy et al. 2003). Using these libraries of fragments, large numbers of conformations are then generated using a monte carlo or simulated annealing approach.

Fragment assembly may proceed on a range of scales (Bystroff et al. 2004). One of the first fragment assembly methods, FRAGFOLD (Jones 1997, 2001, McGuffin & Jones 2003*a*), generates conformations using 'super-secondary' structural motifs, and this technique has performed well in CASP experiments (see Section 1.6.1). The most successful strategy to date is the ROSETTA method (Simons et al. 1997, Bonneau et al. 2001, Bradley et al. 2003, Rohl, Strauss, Misura & Baker 2004) which assembles short 9 residue fragments to generate large numbers of conformations. These methods

usually generate between 1000-100,000 structures for a 100 residue target with only a few of these adopting a reasonable fold (between 4-6Å for C_α atoms).

One of the major outstanding problems in the new fold category is the accurate selection of these 'good' models. At present, the most successful technique is the clustering of conformations based on RMSD and energy scores with the selection of cluster representatives (Zhang & Skolnick 2004*c*). Scoring functions based on physical or statistical potentials are not yet accurate enough for accurate model selection due their inability to model the detailed interatomic interactions that occur in the native conformations.

1.5 Structural similarity measures

Determining the success of protein structure prediction algorithms is not a simple task. The criteria for determining what constitutes a good prediction often depends on the predicted modelling difficulty of a target sequence. For comparative modelling targets, where the structures of detected sequences are often highly similar to the structure of the query sequence, specific and detailed measures of similarity are desirable to quantify the accuracy of the global fold and the finer details such as atomic positions of the main-chain atoms and the side-chains. For hard *ab initio* targets, where no homologous structure exists in experimental databases, a less stringent measure of similarity is often required. In these later case, methods which apply strict criteria to the measurement of atom positions in the model and experimental structure are likely to be highly uninformative. Traditionally, exact algorithms have been used to compare the spatial arrangements of atoms in protein structures, though as structure prediction became more widespread these techniques were supplemented with less stringent heuristic methods more suitable to the task of measuring prediction accuracy.

1.5.1 Root Mean Squared Deviation (RMSD)

The need to compare the geometric representations of two protein structures arose in the field of X-ray crystallography, where a method was required to measures the difference between atomic arrangements and measured value atomic coordinates.

The coordinate Root Mean Square Deviation (RMSD) provides a quantitative single-value measure of the structural similarity between two (usually globular) protein

structures (Rao & Rossmann 1973). The calculation of the coordinate RMSD involves first translating each protein so that each centroid is aligned at the origin, followed by the rotating one structure until the optimal transformation is found which minimises the squared deviation of corresponding atoms. More formally, given two structures, each with N atoms represented as ordered vectors, x_k and y_k, the RMSD can be calculated by finding an orthogonal transformation $\mathscr{U}$, and a translation $\mathbf{r}$, such that the residual $\mathbf{E}$ between atomic coordinate distances is

$$\boldsymbol{E} := \frac{1}{N}\sum_{k=1}^{N} |\, \mathscr{U}\boldsymbol{x}_k + \boldsymbol{r} - \boldsymbol{y}_k|^2 \tag{1.7}$$

A number of authors have proposed algorithms for finding the optimal rigid-body superposition of coordinate vectors (MacLachlan 1972, Diamond 1976), though Kabsch (1976) was the first to provide an exact solution (Kabsch 1978).

The use of this metric in non-crystallographic applications has resulted in the identification of a number of application specific problems. The RMSD calculation requires a one-to-one correspondence of points, and so in structural comparisons where the two point vectors are of different lengths, the assignment of point equivalences is a major issue (Irving et al. 2001). Moreover, the statistical significance of the resulting RMSD scores have been called into question; it has been shown that RMSD value is dependent on the sizes of the protein structures (Maiorov & Crippen 1995, Betancourt & Skolnick 2001). For example, ambiguities arise when considering whether an RMSD score of 2Å for a 40 residue protein is better than a 3Å RMSD score for a protein of 150 residues. Solutions to this length dependence problem have been proposed in the form of normalised scores (Carugo & Pongor 2001, Betancourt & Skolnick 2001), though a satisfactory solution for the assignment of equivalent residues is still unsolved. While the RMSD is a common metric for measuring the global structural similarity of protein structure models, heuristic similarity measure were later developed to provide a more informative measure of model quality for predicted protein structures.

1.5.2 Heuristic similarity methods

Heuristic methods were developed to address both the limitations of the RMSD calculation and to provide more powerful analytic techniques for measuring the quality of structure modelled by homology. In general, these structural similarity methods fall

into two categories; (*i*) sequence-independent, and (*ii*) sequence-dependent methods.

Sequence-independent similarity measures evaluate the structural superposition of two protein models without first matching the equivalent residues in the two structures. By ignoring the displacement of residues resulting from the sequence-to-structure alignment stage, a global comparison of the structures could be made without considering the effects of an sub-optimal sequence alignment in the modelling process (Marchler-Bauer & Bryant 1999). These tools were traditionally used to measure fold recognition predictions where the primary aim was to obtain the correct fold or the topology.

In contrast, sequence-dependent methods consider the sequence alignment of two structures, resulting in more stringent evaluation criteria (the sequence assignment must be correct as well as the topology in order to score highly), and these methods generally employ some form of algorithm for determining the optimal subset of atoms that can be superimposed.

Sequence-dependent methods have an advantage over the sequence-independent variety when assessing multi-domain models. In part, this is because there are currently no satisfactory domain detection and parsing techniques, and so the rigid-body superpositions required with sequence-independent methods are highly sensitive to slight movements in the relative domain placements. As a result, a model that contains near-native domains for each individual part of a multi-domain model but which also contains a small rotation in the relative orientation of those domains, will produce an unsatisfactory alignment score after a structural superposition (Sierk & Kleywegt 2004). To illustrate with some examples, Hubbard (Hubbard 1999) produced a coverage graph method by generating multiple structural alignments based around different sets of residue equivalences. The optimal alignment containing the highest number of residue equivalences can then be ascertained from the resulting RMSD plot. These graphs are difficult to interpret and much effort was made to devise an automated measure which produces a single score representing the quality of the model. The CASP experiments (see Section 1.6.1) required a method that was able to accurately assess large numbers of models automatically (Moult et al. 1997). To this end, a number of assessment techniques were generated (Jones & Kleywegt 1999, Murzin 1999, Orengo et al. 1999, Zemla 1999). More recently, a set of heuristic methods, the

MaxSub, Global Distance Test (GDT), and TM-scores, have become popular tools for assessing protein structure prediction.

1.5.2.1 MaxSub

The popular MaxSub score (Siew et al. 2000) was developed after the CASP3 experiment to automate the procedure of model quality assessment, and is a sequence-dependent heuristic method that attempts to identify the largest subset of superimposable C_α atoms within some distance threshold (commonly 3.5Å). The final result of the algorithm is a single normalised score, a variant of the Levitt-Gerstein score (Levitt & Gerstein 1998) (see Appendices), which allows comparisons of models with different selected subsets of residues. A major limitation of the MaxSub score is that only the residues included in the largest substructure are evaluated. This exclusion of spatial information included in the template but outside the selected subset means that model coverage is overlooked.

1.5.2.2 Global Distance Test (GDT)

The GDT (GDT_TS) score (Zemla 2003) identifies multiple maximum (not necessarily contiguous) substructures associated with several distance thresholds (1, 2, 4, and 8Å). The GDT score is then defined as the average coverage of the target sequence of these substructures with the four distance ranges. One of the major problems with the GDT score is that finer distance details within thresholds are partially excluded. This feature is most prominent in the 4-8Å range where small RMSD deviations in atom positions within this range are all given equal weighting in the final score.

1.5.2.3 Template Model (TM-score)

The Template Model, or TM-score, (Zhang & Skolnick 2004*b*) was developed as an extension to the GDT and MaxSub scores, in an attempts to overcome some of the problems associated with the RMSD metric and these two heuristic methods such as their length dependence, or the weighting applied to different regions of the structure by the evaluation procedure.

Using a variation of the Levitt and Gerstein (LG) score (Levitt & Gerstein 1998), the authors develop a single assessment score that has an appropriate balance of alignment accuracy and coverage, and is strongly related to the quality of the final full-length model. The method uses a similar search method to that of Siew et al. (2000),

and Zemla (2003), in order to find a superposition that produces the largest subset of equivalent residues over *all* template-aligned residues.

In the TM-score, modelling errors are normalised by a protein size dependent scale so that the average TM-score of random protein pairs has no bias to the target protein's length, and all residues in a modelled protein are evaluated with this score. The TM-score shows a closer correlation between the initial template alignments and the final models than does the MaxSub score because the TM-score counts the template information of both high accuracy aligned regions and low accuracy aligned regions, while the MaxSub score neglects the alignment information included in the low accuracy aligned regions that could be of assistance in global modeling (Zhang & Skolnick 2004*b*). On the other hand, unlike the RMSD in which the prediction errors are averaged with equal weights for all residues, the TM-score uses the LG-factor that weights the low and high accuracy regions differently. This also allows the TM-score to provide a more sensitive measure than the GDT-score.

1.6 Prediction benchmarks

As a consequence of the growth in the number of experimental resolved protein structures in public databases, many groups began developing algorithms for protein structure prediction in order to tackle the exponential growth in sequences for which no experimental structure was available. In order to determine how well these structure prediction methods were able to predict the structure for an novel sequence, a range of benchmarks were devised to monitor progress in the structure prediction community. Two major benchmarks are the Critical Assessment of Structure Prediction (CASP) blind prediction experiments, and the automated LiveBench experiment.

1.6.1 Critical Assessment of Structure Prediction (CASP)

The Critical Assessment of Structure Prediction (CASP) was initiated in the early 1990's as a large scale community experiment in which participating groups make *bona fide* blind prediction of protein structures. These biennial meetings provide a setting in which the protein structure prediction community can test and evaluate their structure prediction methods. Crystallographers and NMR spectroscopists, in the process of determining a protein structure, publish the amino acid sequence of their targets several

months before the expected completion date of their work, and keep the experimental structure secret. Groups are then required to submit predictions for each of the targets and the predicted results are compared with the experimental structures. The CASP experiments (recently culminating in the CASP6 meeting) are now in their tenth year and have proved a successful way of highlighting both bottlenecks and areas of progress (Moult 2005) in structure prediction methods.

1.6.2 LiveBench automated prediction assessment

In addition to the CASP experiments, LiveBench (Bujnicki et al. 2001) provides an additional resource for assessing automated structure prediction methods. Though less rigorous than the CASP experiments, LiveBench provides a valuable tool for continual large-scale assessment of fold recognition servers. The LiveBench server regularly scans the PDB for newly released targets, thus the experimental structures are available to the prediction community in advance. LiveBench filters from each set of newly released targets, any sequences that can be detected using BLAST (Altschul et al. 1997) with and E-value < 0.1 referring to these as "trivial" targets. This filtering process ensures targets most suited to homology modelling techniques are removed from the target list. The remaining targets are divided into two classes labelled "easy" and "hard". Although the boundaries between these two classes are somewhat arbitrary, in general, a target above a threshold E-value of 0.001 after five PSI-BLAST iterations (Altschul et al. 1997) defines a difficult target. After selecting representatives from the list of targets, the sequences are submitted to the participating fold recognition servers, and the results are then collated and assessed. The assessment protocol in LiveBench employs a range of model assessment methods and the choice of method can affect the ordering of servers in the final rankings. In the latest LiveBench-8 experiment (Rychlewski & Fischer 2005), the Distal-BASIC, Proximal-BASIC and Meta-BASIC methods (Ginalski et al. 2004) achieved the highest sensitivity and selectivity using a profile method which generates a consensus of meta-profiles.

1.7 Model refinement

Recent CASP experiments have shown that in order to achieve accurate high-resolution structure predictions, addition techniques must be developed to refine both template-

based and *de novo* predictions (Moult 2005). Even the best template models often contain errors in the side-chain packing arrangements, distortions in regions that are aligned with the template, distortions resulting from regions without equivalent segments in a template, and errors in loop regions. For a query sequence with more than 40% sequence identity, approximately 90% of the main chain atoms can be modelled with an RMS error of approximately 1Å (Sanchez & Šali 1998), equivalent to a low-resolution (2.5Å) X-ray structure (Clore et al. 1993). When sequence identity falls between 30 and 40%, structural differences become more pronounced and gaps in the alignment become larger. At less than 30% sequence identity about 20% of all residues are found to be misaligned with a resulting RMS error of more than 3Å. This poses a serious challenge for structural modellers who wish to obtain high-resolution model accuracy for the wider scientific and biological community.

To increase both the reliability and accuracy of protein structure prediction requires methods capable of consistently sampling the high-resolution details of native structures. In addition, if these methods are to select low energy native conformations from a set of decoys, potential energy functions capable of recognising the native state as the lowest energy conformation are a pre-requisite. Both sampling strategies and energy functions of this nature are sadly lacking, and as a result, high-resolution refinement remains a formidable challenge. The hope that MD simulations would be able to refine both template-based and *de novo* structural models has yet to deliver results, with previous attempts to apply these methods to the refinement problem generally driving models further away from the native structure and decreasing the accuracy over the starting model. As a consequence of these initial refinement results, many groups chose to exclude a refinement step in the CASP4 experiment (Schonbrun et al. 2002). If the protein community are to achieve the goal of high accuracy modelling increased efforts are crucial if high-resolution structure prediction is to become commonplace.

Recent refinement protocols have mainly employed three techniques; (*i*) the use of constraints, (*ii*) molecular dynamics simulations, and, (*iii*) knowledge-based potentials.

Fan & Mark (2004) applied molecular dynamics to a set of *de novo* models generated by the ROSETTA structure prediction algorithm (Simons et al. 1997) and were able to show that although MD simulations increased the RMS error of models

from that of the starting structure, the final packing of α-helices and the regularization of β-strand geometries were improved after longer simulations. Lee et al. (2001) were able to refine low-resolution structures using a combination of local constraints, molecular dynamics and statistical potentials. By constraining the local geometry of their models to within 2Å of the original configuration, they were able to sample small changes in the structure under the guidance of their potential functions. This approach yielded small but encouraging improvements for the majority of structures refined in this manner, though leaving room for further research and solutions to the refinement problem.

Developing a novel solution to the refinement problem is thus the main focus of this dissertation. An evolutionary computational approach to the structure refinement problem is the main component of the method and a brief introduction to optimisation and genetic algorithms now follows.

1.8 Genetic algorithms

Genetic algorithms (GAs) are a branch of evolutionary algorithms (EAs) which were first described in John Holland's seminal work *Adaptation in natural and artificial systems* (Holland 1975). He presented the GA as an abstraction of biological evolution, and his goal was to study and understand the mechanisms which lead biological systems to evolve and adapt in the natural world. Since that time GA's have been extensively used as search and optimization tools across a wide variety of problem domains and provide a robust set of algorithms with broad applications (Goldberg 1989). The field of genetic algorithms is too broad to cover fully in this short introduction, but the interested reader is directed to the following texts for more details (Back et al. 1997, Goldberg 1989, Vose 1999, Mitchell 1999, Deb 2001). This survey reviews only the areas of GA research that are necessary to understand the work contained in this dissertation.

1.8.1 Search and optimization

In mathematics, optimization refers to the study of problems which seek to find a minimum, or maximum, of some real function by systematically choosing values of variables from within an allowed set. Many real world problems require the solution of optimization problems and the field of optimization is a rich and varied one. Optimization problems can be generally classified as local or global, that is to say, local when the minimum value they seek to find is either a minimum within some neighbourhood of points that need not be (but may be) a global minimum, or, global when the smallest value over the entire range of possible function values is required.

In their most basic form optimization problems require: (*i*) an objective function that one seeks to minimise or maximise, (*ii*) a set of parameters which affect the value of the objective function, and optionally, (*iii*) a set of constraints that defined the range of values a parameter can take. Optimization problems increase in complexity when there is either no objective function to optimize (also known as *feasibility* problems (Boyle et al. 1997)) where the goal is to find a solution that fits a set of constraints, for example, in designing a layout for a circuit board), or when there are multiple objectives that may or may not be compatible with each other (see Section 1.8.4).

There are many techniques for solving linear and non-linear optimization prob-

lems using calculus-based, random, or enumerative techniques. These methods, and even modern derivatives, are essentially local in character and seek to find the best solution in a set of neighbouring points. Moreover, calculus-based methods require the existence of derivatives [1] in order to navigate towards a local minimum. Where the problems under consideration are linear, or non-linear but with well defined objective function gradients, deterministic gradient- and simplex-based methods are often the best choice (Polak 1971, Nelder & Mead 1965).

Difficulties arise when the objective involves many parameters that interact in highly non-linear ways. Objective functions characterised by many local optima, expansive flat planes in multi-dimensional space, points at which gradients are undefined, or when the objective function is discontinuous, pose difficulty for traditional mathematical techniques. In these cases, stochastic and heuristic methods are required to provide true or approximate solutions to the global optimum. Heuristic algorithms, such as simulated annealing (Metropolis et al. 1953) and genetic algorithms, are useful for finding solutions to problems where complete information about the objective function is unavailable. They cannot *guarantee* finding the global optimum of an optimization problem though they can often find a good approximate solution in most cases. Moreover, the parallel nature of GAs (their ability to explore multiple regions of the search space in a single generation) gives them an advantage over traditional search techniques and as a global optimization method, have been shown to perform well on NP-hard problems (Davis 1991, Michalewicz 1999). Due to the stochastic nature of the search, GAs are generally insensitive to their starting conditions, and are less likely to get trapped in a local minimum.

The considerations about the appropriate algorithm to use for a particular problem can be examined with reference to the no-free-lunch theorem.

1.8.2 The no-free-lunch theorem

According to the NFL theorm (Wolpert & Macready 1997*a*,*b*), an algorithm can not exist for solving *all* optimization (or search) problems that is generally (on average) superior to any competitor. In other words, for a given algorithm, any elevated

[1] numerical approximations to the derivative can be calculated when no derivative is available, yet the problem of locality still holds.

performance over one class of problems is exactly paid for in performance over another class.

In light of this theorem, considering whether evolutionary algorithms are superior or inferior to some other class of algorithms is meaningless. As we have seen in Section 1.8.1, the NFL theorem is corroborated by comparing EAs with other types of traditional optimization techniques. Most classical techniques are more efficient at solving linear, quadratic, and other specific problems, whereas EAs are generally more effective at solving problems with are highly non-linear, and non-differentiable (though this also means they are less efficient at solving simpler problems for which classical procedures are more suited).

This result is especially useful for GAs because it suggests that optimal performance can only be achieved on a specific test problem by carefully selecting both the design and parameters for a GA, or more generally, that there is no guarantee that an optimal parameter set can be determined *a priori* (for example, by transferring an optimal set of parameters from one test problem to another).

The rest of this review focuses on genetic algorithms, beginning with a description of the canonical single-objective GA, followed by a review of multi-objective optimization and applications of genetic algorithms to protein structure prediction.

1.8.3 Single objective genetic algorithms

GAs are a class of iterative global search heuristics modelled on the evolutionary processes found in nature. They are randomised (not random) search techniques and use random choice as a tool to guide a highly exploitative search through a coding of parameter space (Goldberg 1989). Unlike many traditional optimization and search methods, they are robust techniques for performing searches in complex spaces and are not limited by restrictive assumptions about the search space.

GA's operate on a finite set of points, called a *population*, with each member, or *individual*, in the population representing a solution in the overall search space. Individuals are usually fixed-length vectors, called *chromosomes*, with each individual vector position (*gene*), holding a particular value (*allele*) that represents a parameter or parameter encoding over some symbolic alphabet. The symbol alphabet used to represent parameters is often binary, though other representations have also been used.

A typical GA follows a simple iterative procedure; (i) initialization, (ii) selection, (iii) reproduction, (iv) termination.

A GA starts with an initial population of individuals that are either randomly generated, or "seeded" so that individuals are placed in areas of the search space where optimal solutions may lie. In the selection stage, a proportion of individuals are selected from the current population to breed a new generation. Individuals in the current population are scored with the *fitness function* and then the selection occurs through a fitness-based mechanism where fitter individuals are more likely to be selected. The majority of selection functions are stochastic and are designed so that a small proportion of less fit individuals are selected in addition to the larger majority of fitter individuals. Once a pool of individuals has been selected, the next step is to produce a new generation of solutions through the process of reproduction. Reproduction involves applying *crossover*, or *recombination*, operators and mutation operators to individuals or pairs of individuals. Pairs of individuals, or *parents*, are selected at random from the pool and then either a *child* is produced by the recombination of two parents, or through the mutation of one parent. This new population then serves as the current population, and the stepwise process of selection and reproduction, which constitutes a single evolutionary generation (Goldberg 1989), continues iteratively until some termination criterion has been reached.

By selecting individuals with high fitness, the relatively fit members are more likely to survive and procreate with the hope that their genes truly result in individuals which represent better solutions. A illustrative representation of the genetic algorithm is shown in Figure 1.6.

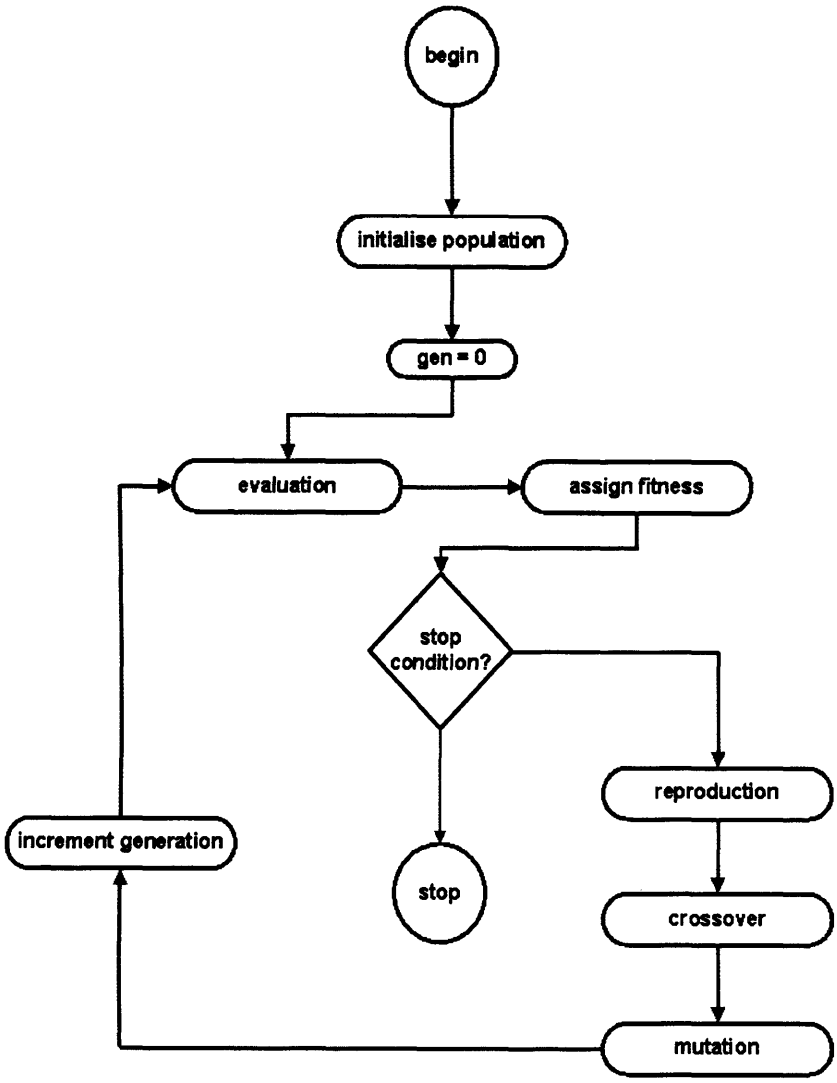


Figure 1.6: The flowchart illustrates the algorithmic steps of a simple genetic algorithm.

1.8.3.1 Representation

The choice of representation for a solution is the most critical factor affecting the success of a genetic algorithm (Mitchell 1999). Although most formal mathematical and theoretical descriptions of genetic algorithms are based on a binary string representation, the symbolic alphabet can easily be extended to non-binary encodings. Real-valued (Goldberg 1990, Wright 1991) and tree representations (Antonisse & Keller 1987) are common in the GA literature as well as a number of other less common representations such as tree encodings (Tackett 1994). The best choice of encoding is naturally problem dependent, and a genetic algorithm can be customised to work optimally on a specific representation. However, it is important that the crossover operator chosen for a particular representation is appropriate as the representation has

a direct effect on performance.

1.8.3.2 Fitness/objective functions

The *objective*, or *fitness*, function is used to evaluate individuals in a population, and describes the space on which the GA search will occur. Therefore, determining the fitness of all possible solutions in the search space describes a hypersurface, termed the *fitness landscape*.

The objective function provides a measure of fitness for each individual with respect to its phenotype, that is, the solution which the chromosome represents. An ideal fitness function should correlate closely with the algorithm's goal, and yet should be able to be computed quickly. Speed of execution is very important, as a typical genetic algorithm must be iterated from hundreds to many thousands of times in order to produce a useable result for a non-trivial problem.

1.8.3.3 Selection schemes

Selection is a primary operator in GAs and is used to improve the average population fitness of solutions during successive rounds of evolution. The selection operator is used to decide which individuals in a population will create offspring for the next generation, and how many will do so. The main purposes of selection are: (*i*) to identify good solutions in a population, (*ii*) to produce multiple copies of good solutions, and, (*iii*) to eliminate bad solutions, thus making room for better individuals. Selection requires a careful balance with the genetic operators so that high fitness, sub-optimal solutions do not dominate the population.

There are numerous selection schemes in the GA literature each providing a unique set of advantages and disadvantage. For example, Holland's original fitness-proportionate "roulette wheel" selection scheme allows individuals to be selected as parents with a probability proportional to their fitness value (Holland 1975). The size of the wheel slices are proportional to the fitness scores of the individuals in the populations. This stochastic selection method can result in a high fitness individual rapidly becoming dominant in a population.

Rank-based selection schemes use only the rank order of the fitness scores of individuals within the current population to determine the probability of selection (Baker 1985). Objective function evaluations are mapped to fitness ranks instead of

absolute fitness scores. This use of rank rather than fitness score avoids giving the largest share of offspring to a small group of highly fit individuals, and thus reduces the selection pressure when the fitness variance is high. Rank selection is often used to prevent rapid convergence typical of roulette wheel selection.

Another common selection scheme is know as tournament selection (Deb & Goldberg 1991). Here a group of q individuals is randomly chosen from the population and may be drawn from the population with or without replacement. This group takes place in a *tournament* where a winning individual is determined depending on its fitness value. The highest fitness individual is then inserted into the next population, and the process repeated λ times until the next population is full. The tournament group size q is a variable parameter with the most common group size of two (binary) individuals.

1.8.3.4 Elitism

Elitism (or an elitist strategy) is a mechanism which ensures that some number of the fittest individuals are preserved in the next generation (De Jong 1975). The use of elitism guarantees that the maximum fitness of the population never decreases from one generation to the next, a result of which enhances the performance of the GA significantly.

1.8.3.5 Genetic operators

The genetic operators are a critical part of any genetic algorithm and form a symbiotic relationship with the encoding strategy. The operators commonly used in classical bit-string encoding of problems are the crossover and mutation operators.

1.8.3.5.1 Crossover The crossover operator creates new solutions by swapping regions of two solutions. Although the creation of solutions which are better than the individual parents is not guaranteed, the chance is much better than random as these individuals have survived rounds of selection and are expected to contain some good regions in their encoding. Crossover operators come in a variety of forms though the simplest is a single-point crossover. Here a single position is selected at random and the parts of the two parents after the crossover point are exchanged. The main motive behind this operator is the exchange of building blocks (schema) between different strings. The single-point crossover has a number of limitations, the most important being its inability to cover all possible schemas (Eshelman et al. 1989). The two-point

crossover is another popular variant which selects two positions at random with an exchange of the segment between them. Many other crossover operators have been used on bit-string encodings with varying success (Spears 1998), and there are no defining criteria for the appropriate operator in any problem definition. Figure 1.7 shows the crossover operator applied to simple bit-strings. Alternative problem encodings, such as tree structures in genetic programming, real valued encodings or multi-dimensional encodings, all require the definition of encoding specific operators and the mechanisms of these methods have not yet been studied extensively.

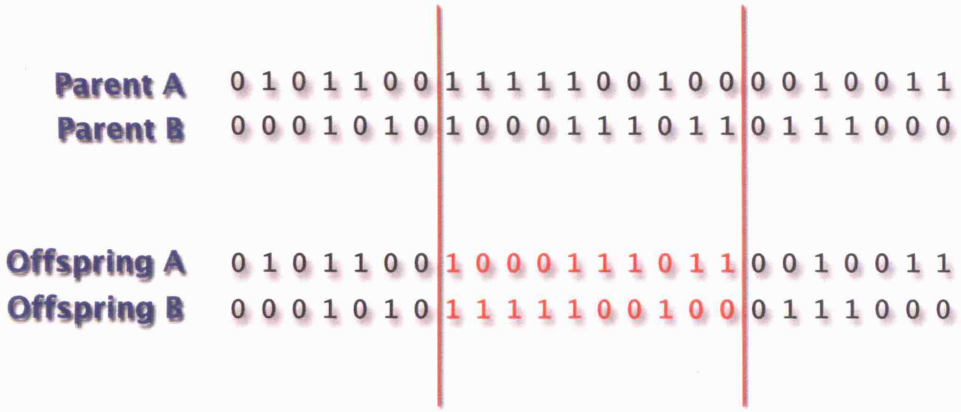


Figure 1.7: A two-point crossover operator is shown. Two points along the chromosome are selected at random and the region spanning these two positions are exchanged between parents to produce new offspring.

1.8.3.5.2 Mutation The mutation operator is commonly viewed as a secondary mechanism in genetic algorithms, with the crossover operator performing the largest part of the variation and search. The mutation operator insures that the population does not get stuck at a particular locus or minimum in the search space by diversifying the individuals in that population. Figure 1.8 shows a simple mutation operator applied to bit-string encodings of a solution.

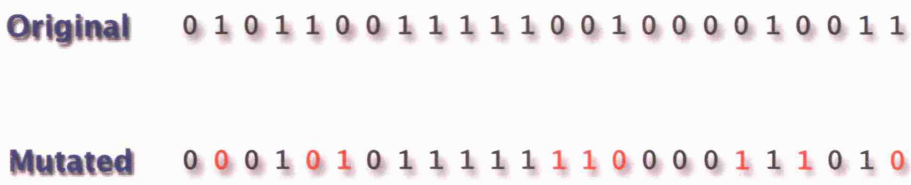


Figure 1.8: A bit-mutation operator is shown. At each position or gene along the chromosome, the value is flipped with probability P_m, producing a new solution.

1.8.3.6 Parameters

The performance of a GA is controlled, in part, by parameter values which govern the size of the population, the recombination and mutation rates, and the selection mechanism. The interaction between parameters is typically non-linear and problem dependent, and single parameters can not be optimized individually making the optimization of a GA more of an art than a science. There is no methodical way of determining optimal parameters though a number of studies concerned with this subject have provided some useful guidelines.

The traditional, and most common, strategy for selecting parameters is known as parameter *tuning*. This is the process by which parameter values are selected before the GA is applied, and the values kept unchanged until the run is complete. The outcome is then recorded and the parameters adjusted until the parameters that provide the best performance are found. This approach requires tuning parameters by hand, and it is both time consuming and, in many cases, impossible to test all combinations of parameter values (Eiben et al. 1999). The most notable study into parameter optimization is De Jong's study on a small suite of test problems (De Jong 1975). Although this work resulted in a set of general parameters that performed well on the set of test problems presented, it may not be the case that these parameters are optimal when transferred to other problems. Indeed the contemporary view is that a GA setup must be optimized for a particular problem (Back et al. 1997).

The alternative approach is parameter *control*, whereby parameters are modified during the course of a run. This approach was first adopted by Grefenstette (1986) who attempted to use genetic algorithms to optimize the control parameters in another

GA. Other parameter control studies have been performed (Schaffer et al. 1989, Davis 1989, 1991) though the consensus from these studies suggests that parameter adaptation introduces additional problems which in turn require optimal parameterization. For example, parameters can be altered according to some deterministic control (for which the function needs to be carefully chosen), or an adaptive control (which requires a feedback mechanism from the search to determine the magnitude and direction of the parameter adjustment) (Eiben et al. 1999).

1.8.3.7 Termination criteria

There are two popular methods for terminating a GA. The first approach is to halt a GA when it has completed a predefined number of evolutionary cycles. This is often adopted for problems of large complexity when it is unlikely that full convergence will be reach. Instead, convergence on an acceptable solution is preferred rather than increasing the computational effort required to improve the solution further.

The alternative approach is to stop the simulation once the population ceases to evolve, in effect the population has converged on either a local, or global, minimum. The pressure towards convergence is driven by fitness, and offers a convenient termination criterion. In fact, after many generations of evolution via the repeated application of reproduction, crossover, and mutation, the individuals in the population will often begin to have similar genotypes. At this point, the GA typically terminates because additional evolutionary cycles will produce little improvement in fitness.

Early population convergence on a local optimum is a problem for GAs and results in sub-optimal solutions for a given problem. One approach to preventing early convergence in this way is through *niching*.

1.8.3.8 Niching

Niching methods are techniques for promoting the formation and maintenance of stable sub-populations within a GA. These methods are designed to ensure that a population contains a diverse set of individuals by altering the selection algorithm to provide diversity within, but not across regions of the search space (Goldberg & Richardson 1987). This can prevent early convergence (see Section 1.8.3.7) by ensuring the GA continues to explore diverse regions of the search space (Thierens et al. 1995).

Niching methods are particularly useful in multi-objective optimization where the

goal is to provide multiple solutions to a problem.

1.8.4 Multi-objective genetic algorithms

Not all real-world problems have a single optimal solution and some require the simultaneous optimization of several incommensurable, and often competing, objectives. In these cases, optimising one objective can often lead to unacceptably low performance in one or more of the other objective dimensions, and so an optimization strategy then involves finding a compromise between competing objectives. The outcome of a multi-objective optimization strategy is a set of trade-offs, known as the *Pareto-optimal* set, named after the Italian economist Vilfredo Pareto. He formulated the concept of Pareto-optimality as a solution to the problem of allocating economic resources and productive output among a group of individuals. He suggested that the optimal allocation of resources had been reached when no-one could be made better off without sacrificing the well-being of at least one person (Pareto 1896). In the context of general multi-objective optimization problems (MOPs), Pareto optimal solutions are optimal in the sense that no other solutions in the search space are superior to them when all objectives are considered.

To illustrate a simple example of a MOP, consider two functions $f_1(x)$ and $f_2(x)$ each with a unique minimum at $f_1(0)$ and $f_2(1)$ where $x = 0$ and $x = 1$, respectively (see Figure 1.9). We can obtain unique optimal solutions for each function individually, however, it is not possible to optimize both functions simultaneously and obtain a single optimal solution. Instead we obtain a set of solutions that represent trade-offs between each objective (in this case the optimal x values are $[0, 1]$). Without further problem specific information it is not then possible to say which of these solutions is better.

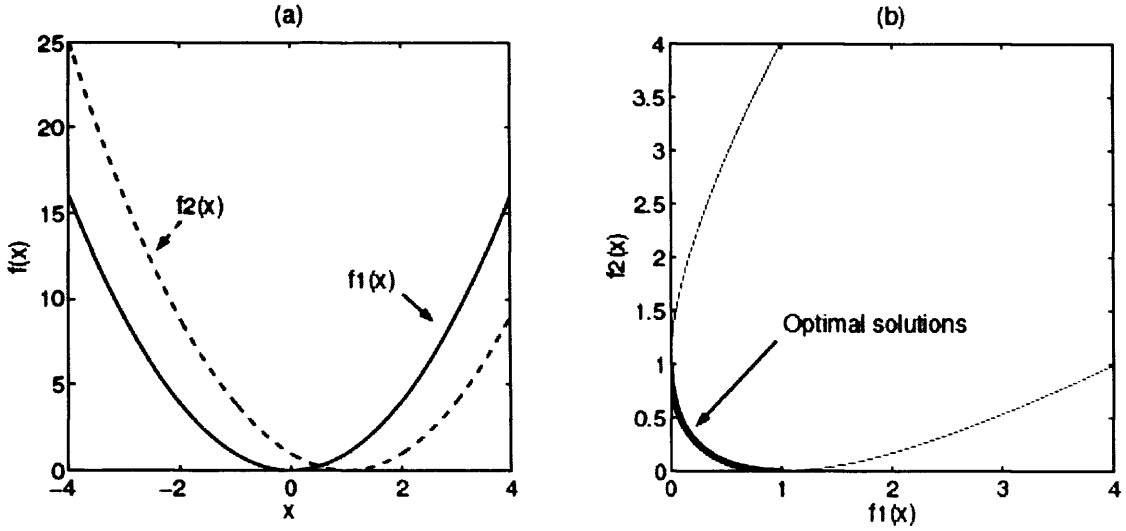


Figure 1.9: (a) Two objective functions, $f_1(x)$ and $f_2(x)$, each have a unique and distinct minimum at $f_1(0)$ and $f_2(1)$. (b) The objective space for the functions $f_1(x)$ against $f_2(x)$ is shown. There are now two optimal solutions which lie along the highlighted front.

1.8.4.1 Basics concepts and definitions

Some basic concepts and formal definitions can now be given with reference to a general multi-objective optimization problem where each objective is given equal importance.

Definition 1.1 (Multi-objective optimization Problem) *A general MOP includes a set of n parameters (a decision vector), a set of k objective functions, and optionally, a set of m constraints. Objective functions and constraints are functions of the decision variables. The optimization goal is to*

$$\begin{aligned} &\textit{maximise} && \boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x}) = (f_1(\boldsymbol{x}), f_2(\boldsymbol{x}), \ldots, f_k(\boldsymbol{x})) \\ &\textit{subject to} && \boldsymbol{e}(\boldsymbol{x}) = (e_1(\boldsymbol{x}), e_2(\boldsymbol{x}), \ldots, e_m(\boldsymbol{x})) \leq \boldsymbol{0} \\ &\textit{where} && \boldsymbol{x} = (x_1, x_1, \ldots, x_n) \in \boldsymbol{X} \\ & && \boldsymbol{y} = (y_1, y_1, \ldots, y_n) \in \boldsymbol{Y} \end{aligned}$$

and $\boldsymbol{x}$ is the decision vector, $\boldsymbol{y}$ is the objective vector, $\boldsymbol{X}$ is denoted as the decision space, and $\boldsymbol{Y}$ is called the objective space. For clarity and simplicity, an assumption is made that all objective functions are to be maximised. To minimise a function f_i, treatment is equivalent to maximising $-f_i$.

The constraints $\boldsymbol{e}(\boldsymbol{x}) \leq \boldsymbol{0}$ determine the set of feasible solutions to a MOP.

Definition 1.2 (Feasible Set) *The feasible set $\boldsymbol{X}_f$ is defined as the set of decision vectors $\boldsymbol{x}$ that satisfy the constraints $\boldsymbol{e}(\boldsymbol{x})$:*

$$\boldsymbol{X}_f = \{\, \boldsymbol{x} \in \boldsymbol{X} \mid \boldsymbol{e}(\boldsymbol{x}) \leq \boldsymbol{0} \,\}$$

The image of $\boldsymbol{X}_f$, i.e., the feasible region in the objective space, is denoted as $\boldsymbol{Y}_f = \boldsymbol{f}(\boldsymbol{X}_f) = \bigcup_{\boldsymbol{x} \in \boldsymbol{X}_f} \{\boldsymbol{f}(\boldsymbol{x})\}$.

In single-objective optimization, the feasible set is completely (totally) ordered according to the objective function f: for two solutions $\boldsymbol{a}, \boldsymbol{b} \in \boldsymbol{X}_f$ either $f(\boldsymbol{a}) \geq f(\boldsymbol{b})$ or $f(\boldsymbol{b}) \geq f(\boldsymbol{a})$. The goal is to find the solution (or solutions) that gives the maximum value of f. However, when considering several objectives, the situation changes: $\boldsymbol{X}_f$ is, in general, not totally ordered, but partially ordered (Pareto 1896).

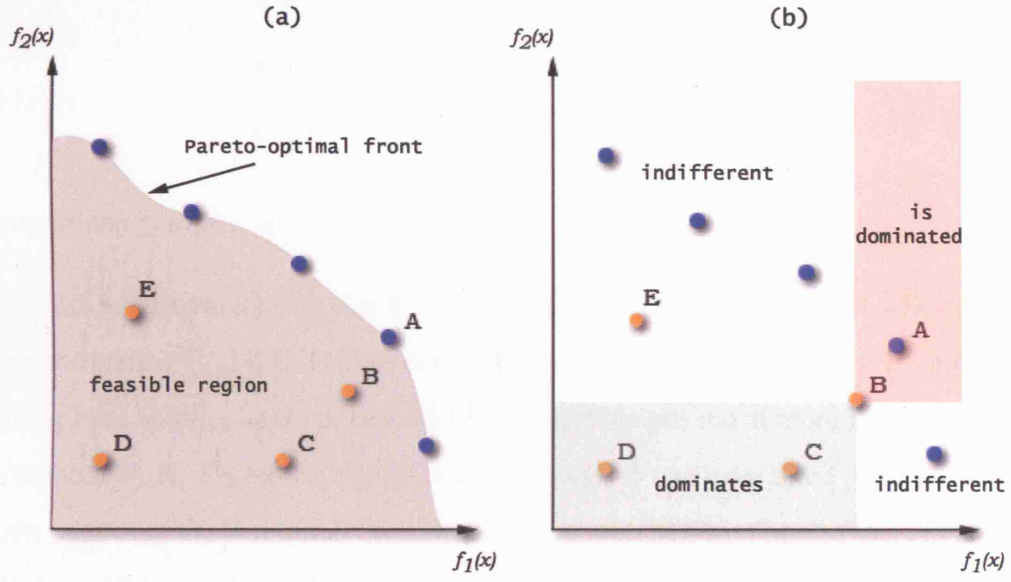


Figure 1.10: (a) A number of solutions to a multi-objective problem are shown in the feasible region of objective space. The function f_1 represents the lightness of the vehicle while f_2 represents safety in this example. The Pareto-optimal front delineates the feasible from infeasible regions, and a number of Pareto-optimal solutions are shown in blue (b) Four regions are highlighted showing their relation to solution B

Consider a multi-objective problem in automobile design with two objectives, to maximise crash resistance for safety and minimise weight for fuel economy. An increase in the crash resistance of the vehicle usually leads to an increase in its weight. We can not find a solution that optimizes both objectives and instead we must find a set of trade-off solutions. An illustration is given in Figure 1.10 showing the lightness of the vehicle (f_1) versus its safety (f_2). The solution represented by point B is better than the solution represented by point C - it provides greater safety at lower weight. It would be even more preferable if it would only improve one objective, as is the case for C and D - despite equal safety, C achieves a lighter weight than D. In order to express this situation mathematically, the relations $=$, $\geq$, and $>$ are extended to objective vectors by analogy to the single-objective case.

Definition 1.3 *For any two objective vectors* $\boldsymbol{u}$ *and* $\boldsymbol{v}$,

$$\begin{aligned} \boldsymbol{u} = \boldsymbol{v} \quad & iff \quad \forall\, i \in \{1,2,\ldots,k\}: u_i = v_i \\ \boldsymbol{u} \geq \boldsymbol{v} \quad & iff \quad \forall\, i \in \{1,2,\ldots,k\}: u_i \geq v_i \\ \boldsymbol{u} > \boldsymbol{v} \quad & iff \quad \boldsymbol{u} \geq \boldsymbol{v} \wedge \boldsymbol{u} \neq \boldsymbol{v} \end{aligned}$$

The relations $\leq$ *and* $<$ *are defined similarly.*

Using this definition, it holds that B >C, C >D, and, as a consequence, B >D. However, when comparing B and E, neither can be said to be superior, since B $\not>$ E and E $\not>$ B. Although the solution associated with E is safer, it weighs much more than the solution represented by B. Therefore, two decision vectors $\boldsymbol{a}, \boldsymbol{b}$ can have three possibilities with MOPs regarding the $=$ relation (in contrast to two with SOPs): $\boldsymbol{f}(\boldsymbol{a}) \geq \boldsymbol{f}(\boldsymbol{b}), \boldsymbol{f}(\boldsymbol{b}) \geq \boldsymbol{f}(\boldsymbol{a})$, or $\boldsymbol{f}(\boldsymbol{a}) \not\geq \boldsymbol{f}(\boldsymbol{b})) \wedge \boldsymbol{f}(\boldsymbol{b}) \not\geq \boldsymbol{f}(\boldsymbol{a})$. The following symbols and terms are used in order to classify these dominance relations.

Definition 1.4 (Pareto Dominance) *For any two decision vectors* $\boldsymbol{a}$ *and* $\boldsymbol{b}$,

$$\begin{aligned} \boldsymbol{a} \succ \boldsymbol{b} \quad & (\boldsymbol{a}\ dominates\ \boldsymbol{b}) && iff \quad \boldsymbol{f}(\boldsymbol{a}) > \boldsymbol{f}(\boldsymbol{b}) \\ \boldsymbol{a} \succeq \boldsymbol{b} \quad & (\boldsymbol{a}\ weakly\ dominates\ \boldsymbol{b}) && iff \quad \boldsymbol{f}(\boldsymbol{a}) \geq \boldsymbol{f}(\boldsymbol{b}) \\ \boldsymbol{a} \sim \boldsymbol{b} \quad & (\boldsymbol{a}\ is\ indifferent\ to\ \boldsymbol{b}) && iff \quad \boldsymbol{f}(\boldsymbol{a}) \not\geq \boldsymbol{f}(\boldsymbol{b}) \wedge \boldsymbol{f}(\boldsymbol{b}) \not\geq \boldsymbol{f}(\boldsymbol{a}) \end{aligned}$$

The definitions for a minimisation problem ($\prec, \preceq, \sim$) *are analogical.*

In Figure 1.10, the light green rectangle encapsulates the region in objective space that is dominated by the decision vector represented by B. The light red rectangle contains the objective vectors whose corresponding decision vectors dominate the solution associated with B. All solutions for which the resulting objective vector is in neither coloured rectangle are indifferent to the solution represented by B.

Based on the concept of Pareto Dominance, the optimality criterion for MOPs can be introduced. Still referring to Figure 1.10, A is unique among B, C, D, and E - its corresponding decision vector $\boldsymbol{a}$ is not dominated by any other decision vector. That means $\boldsymbol{a}$ is optimal in the sense that it cannot by improved in any objective without causing a degradation in at least one other objective. Such solutions are denoted as Pareto optimal.

Definition 1.5 (Pareto Optimality) *A decision vector* $\boldsymbol{x} \in \boldsymbol{X}_f$ *is said to be non-dominated regarding a set* $\boldsymbol{A} \subseteq \boldsymbol{X}_f$ *iff*

$$\nexists\, \boldsymbol{a} \in \boldsymbol{A} : \boldsymbol{a} \succ \boldsymbol{x}$$

If it is clear within the context which set $\boldsymbol{A}$ *is meant, it is simply left out. Moreover,* $\boldsymbol{x}$ *is said to be Pareto optimal iff* $\boldsymbol{x}$ *is non-dominated regarding* $\boldsymbol{X}_f$.

In Figure 1.10 the blue points represent Pareto-optimal solutions. They are indifferent to each other. This makes the main difference to SOPs clear: there is no single optimal solution but rather a set of optimal trade-offs. None of these can be identified as better than the others unless preference information is included (e.g., a ranking of the objectives). The entirety of all Pareto-optimal solutions is called the Pareto-optimal set; the corresponding objective vectors form the Pareto-optimal front or surface. The Pareto-optimal set is only one stage of a successful solution to a multi-objective problem. Usually external information is needed from some decision maker as to which of these solutions is most appropriate for the problem.

Definition 1.6 (Non-dominated Sets and Fronts) *Let* $\boldsymbol{A} \subseteq \boldsymbol{X}_f$. *The function* $p(\boldsymbol{A})$ *gives the set of non-dominated decision vectors in* $\boldsymbol{A}$*:*

$$p(\boldsymbol{A}) = \{\boldsymbol{a} \in \boldsymbol{A} \mid \boldsymbol{a} \text{ is non-dominated regarding } \boldsymbol{A}\}$$

The set $p(\boldsymbol{A})$ *is the non-dominated set regarding* $\boldsymbol{A}$*, the corresponding set of objective vectors* $\boldsymbol{f}(p(\boldsymbol{A}))$ *is the non-dominated front regarding* $\boldsymbol{A}$. *Furthermore, the set* $\boldsymbol{X}_p =$

$p(\boldsymbol{X}_f)$ is called the Pareto-optimal set and the set $\boldsymbol{Y}_p = \boldsymbol{f}(\boldsymbol{X}_p)$ is denoted as the Pareto-optimal front.

The Pareto-optimal set comprises the globally optimal solutions. However, as with SOPs there may also be local optima which constitute a non-dominated set within a certain neighbourhood. This corresponds to the concepts of global and local Pareto-optimal sets introduced by Deb (1999):

Definition 1.7 *Consider a set of decision vectors $\boldsymbol{A} \subseteq \boldsymbol{X}_f$.*

1. *The set $\boldsymbol{A}$ is denoted as a local Pareto-optimal set iff*

$$\forall\, \boldsymbol{a} \in \boldsymbol{A}\colon \nexists\, \boldsymbol{x} \in \boldsymbol{X}_f\colon \boldsymbol{x} \succ \boldsymbol{a} \wedge \|\boldsymbol{x} - \boldsymbol{A}\| < \varepsilon \wedge \|\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{a})\| < \delta$$

   where $\|\cdot\|$ is a corresponding distance metric and $\varepsilon > 0$, $\delta > 0$.

2. *The set $\boldsymbol{A}$ is called a global Pareto-optimal set iff*

$$\forall\, \boldsymbol{a} \in \boldsymbol{A}\colon \nexists\, \boldsymbol{x} \in \boldsymbol{X}_f\colon \boldsymbol{x} \succ \boldsymbol{a}$$

The difference between local and global optima is visualised in Figure 1.11. The dashed line constitutes a global Pareto-optimal front, while the solid line depicts a local Pareto-optimal front. The decision vectors associated with the latter are locally non-dominated though not Pareto-optimal, because the solution related to point A dominates any of them. Finally, note that a global Pareto-optimal set does not necessarily contain all Pareto-optimal solutions and that every global Pareto-optimal set is also a local Pareto-optimal set.

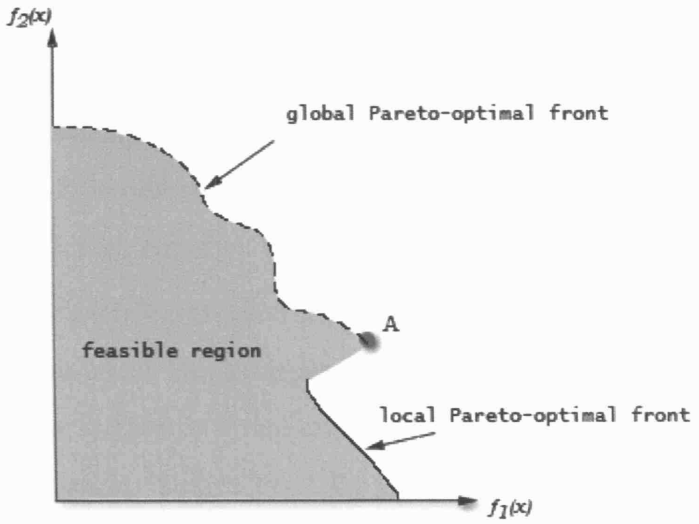


Figure 1.11: The Pareto-optimal front can either be a global front or a local front.

The size and shape of Pareto-optimal fronts generally depends on the number of objective functions and the interactions among the individual objective functions. If the objectives are in conflict, the resulting Pareto-optimal front may have a larger span than if the objectives are more cooperative. However, in most multi-objective optimization problems, the objectives are usually in conflict, with the resulting Pareto-optimal front (local or global) contains many solutions (Deb 1999).

1.8.4.2 Evolutionary algorithms for solving MOPs

Conventional optimization techniques such as gradient-based and simplex methods, as well as some stochastic methods such as simulated annealing, are difficult to extend to multi-objective cases and usually require the reformulation of a multi-objective problem as a single-objective one. In contrast, evolutionary algorithms have been recognised as particularly well suited to multi-objective problems. Their population-based approach allows superior searching because multiple solutions are explored in parallel and eventually taking advantage of any similarities available in the family of possible solutions (Fonseca & Fleming 1995). Additionally, they are less susceptible to the shape of the Pareto front.

There are two tasks that a multi-objective GA should accomplish in solving multi-objective optimization problems: (*i*) guide the search towards the global Pareto-optimal region, and, (*ii*) maintain population diversity in the current non-dominated front.

EA approaches can be generally classified as Pareto-based and non-Pareto-based techniques. Non-Pareto approaches such as aggregating methods combine objectives into a single objective using some aggregating function and are generally successful when the behaviour of the objective functions are well known (Schaffer 1984, 1985). They are generally efficient, but often inferior to Pareto-based methods.

Pareto-based approaches were first proposed by Goldberg (Goldberg 1989) and directly use the definition of Pareto-optimality to move the population towards the Pareto-front. Goldberg suggested a ranking approach which finds the non-dominated individuals in a population, assigns them the highest rank and then removing them from contention. The non-dominated solutions from the remaining individuals are then assigned the next highest rank, and so on, so that chromosomes are then selected for reproducing based on their rank fitness. Since then a number of alternative Pareto-based methods have been developed each with distinct strengths and weaknesses (Fonseca & Fleming 1993, Horn & Nafpliotis 1993, Srinivas & Deb 1994, Zitzler, Laumanns & Thiele 2002, Deb 2002).

1.8.5 Genetic algorithms in protein structure prediction

Genetic algorithms have been used to solve a number of problems in protein structure from *ab initio* protein folding (Dandekar & Argos 1992, Pedersen & Moult 1996, Dandekar & Argos 1994, Elofsson et al. 1995, Pedersen & Moult 1997*a*, Unger & Moult 1993, Dandekar & Argos 1996, Pedersen & Moult 1997*b*, Sun 1995), structure comparison (May & Johnson 1995), structure alignment (Szustakowski & Weng 2000), protein design (Jones 1994), and side-chains optimization (Lee & Subbiah 1991, Tuffery et al. 1991). To date, only one study has been published documenting the use of evolutionary algorithms to refine protein structure models (Contreras-Moreira, Fitzjohn, Offman, Smith & Bates 2003). Contreras-Moreira and coworkers used a genetic algorithm to simultaneously search template and alignment space for 67 targets from the CASP5 experiment. Using a number of models from disparate sources, and alternative alignments for each target, they generated a population of structures by recombination or mutation. The recombination of models involved selection two models are random from the population and exchanging fragments after structurally superposing the structures. A pivot point was selected and the N-terminal of the first

model was then joined at the pivot point with the C-terminal of the paired structure to produce a new structure. Mutation was performed by averaging the coordinates of the superimposed models. Models in the population are ranked by a fitness score composed of a solvation potential and a soft Lennard-Jones contact potential. At the end of each generation, the worst 25% of models are discarded and the process continued until convergence on a single energy score is reached. With this approach they found that in general, recombined models are not significantly different from the best initial model though a handful are better or worse than the starting model in equal measure.

The versatility of genetic algorithms is shown by the extensive use of these search procedures on complex optimization problems that arise in computational biology. In the remainder of this thesis I shall explore further the application of evolutionary algorithms, including both single-objective and multi-objective GAs, to protein structure prediction generally, and specifically, to the refinement of template-based structural models. The next chapter explores the use of multi-objective genetic algorithms in the situation where multiple templates are available for use in refinement.

Chapter 2

Improving Structure Prediction using Multiple Templates

2.1 Introduction

One of the major challenges at the beginning of the post-genomics era is to determine the three-dimensional structures of all known protein sequences, either experimentally or by computational modelling. A decade of DNA sequencing projects have generated a profound wealth of sequence data, including the complete sequence of the human genome (Venter et al. 2001), yet the number of protein structures solved experimentally remains at just a small fraction of the total sequences available (Baker & Šali 2001). With the high cost and technical difficulty involved in resolving structures experimentally, either by X-ray crystallography (X-ray) or nuclear magnetic resonance (NMR), computational modelling has become the most viable approach for generating the structures for these sequences.

Over the last decade, the CASP blind prediction experiments (see Section 1.6.1) have provided the structural community with an accurate assessment of the trends and successes in protein structure prediction. The most noteworthy outcome of these experiments is the superior performance of homology modelling methods for predicting the structure of a protein sequence by exploiting evolutionary relationships between sequences and structures (Moult 2005) over *ab initio* methods which have had a relatively low success rate for medium-to-high resolution structure prediction. Moreover, the growth in the number of protein sequences and experimentally determined structures has propelled knowledge-based procedures ahead in the quest to generate accurate

structural models for protein sequences (Baker & Šali 2001).

Successful template-based (homology) modelling relies on finding sequences with at least one known structure (known as a template) that shares a statistically significant level of similarity to a target sequence, and the accuracy of these models then depends, in part, on the nature of the sequence relationship between the target protein and a known structure. A canonical comparative modelling procedure usually consists of an iterative execution of the following steps; (*i*) select a suitable template (or templates) from a set of evolutionary related sequences, (*ii*) optimally align the query sequence to the residues of the template structure, (*iii*) model the side-chains, and, (*iv*) refine and evaluated the model. Errors introduced in the first two steps of the modelling process account for the majority of modelling errors found in homology models and are a major determinant of the quality of a model generated by a comparative modelling procedure (Sanchez & Šali 1997).

At high levels of sequence identity (more than $\approx$ 50%) models are usually generated with an accurate core region (usually in the range of 1Å to 2Å RMS error for C_α atoms) by copying the coordinates of the template (Sanchez & Šali 1998, Tramontano & Morea 2001). While the alignment stage is usually accurate, more than one template is often required to produce a complete model and the accuracy of these models is often comparable with medium-resolution crystallographic structures (Baker & Šali 2001). Modelling errors at this level of sequence identity result from; an inability to model regions of structure not found in any of the templates, from the incorrect modelling of surface loop regions, and through inaccurate side-chain packing due to small errors in the backbone conformation or from the substitution of an amino acid at a sequence position which can not easily be accommodated by the template framework without introducing steric clashes (Chung & Subbiah 1996).

For distant evolutionary relationships between a target sequence and a template (i.e. sequence relationships found with a high significance score from PSI-BLAST (Altschul et al. 1997)) where the sequence similarity falls to between 30 - 50%, approximately 80% of the target structure is usually shared with a template. This generates models typically in the range of 2 - 3Å. At lower sequence identity (less than 30%), alignment quality sharply decreases and modelling errors are subsequently more pronounced. Here alignment distortions or alignment shifts are a large source of

error, as are incorrectly modelled loop regions and regions with no equivalent residues in the template. Some estimates suggest that structural conservation can range from 90% for close homologues, falling to less than 50% for more remote relationships with less than 30% sequence identity (Venclovas & Margelevicius 2005).

Despite the progress in comparative modelling accuracy over the past decade, further improvements are required to achieve structural models of comparable accuracy to experimentally determined structures. For high sequence identity targets, predictions are often no closer to the experimental structure than the structure of the closest template (Tress et al. 2005), while errors resulting from sub-optimal alignments, sub-optimal template selection, and poor template coverage are additional impediments to accurate modelling at lower sequence identities. Therefore, one of the major bottlenecks to high-resolution structure prediction identified by the CASP experiments is the refinement of template-based models to achieve model qualities matching that of experimentally resolved structures (Ginalski 2006).

2.1.1 Refinement and consensus strategies

Protein structure refinement can be considered either as a separate procedure with which to process the models generated by a template-based modelling method, or, as a fine-tuning of existing template modelling methods. State-of-the-art comparative modelling often uses the latter approach in the form of "holistic" or consensus strategies to build models based on multiple templates or by introducing the use of fragment libraries. These novel methods aim to maximise the information gain from a wide array of available sequence and structural data sources, often by introducing models from other modelling methods, web servers, sequence databases, and structural evaluation methods. This data is then incorporated through an iterative modelling process until convergence of some evaluation score is reached and a final model is generated.

Kosinski et al. (2003) achieved promising results in the CASP6 experiment with their "FRankenstein's Monster" approach and showed that the accuracy of comparative modelling could be improved by a mixture of consensus results from fold recognition methods, model evaluation, and fragment re-assembly using Replica Exchange Monte Carlo (REMC) sampling with restraints derived from the collection of models. Karplus et al. (2003) also use multiple templates and alignments in their UNDERTAKER

program together with a fragment library to improve their template-based models. A genetic algorithm is used to perform a search through a protein's conformational space, where the protein structure is represented as a tree structure and each of the sub-trees represent a structural segment of the protein. Operators are applied to sub-trees enabling fragment replacement, alignment replacement (to replace multiple segments in one step), sub tree re-positioning, and crossover operators. The algorithm optimizes a cost function, which is a linear combination of 24 basic terms, and this method performed well across all categories in the CASP5 experiment (Moult et al. 2003).

By supplementing traditional comparative modelling methods with *ab initio* techniques, structural evaluation methods, and fragment assembly, the traditional boundary between template-based and template-free modelling has become blurred. Yet while these consensus/hybrid approaches have generally been successful at improving the performance of homology-based structure prediction, it is difficult to assess the relative contribution of these additional components to the overall improvements made in recent years. Nor is it simple to determine what contribution the consensus methods themselves have towards the final model quality. It may also be likely that the improvements seen in the recent CASP experiments results from an increase in the number of experimentally resolved structures which provide the foundational knowledge-base for these methods.

2.1.2 Multi-template modelling and refinement

One common feature of these new consensus methods is their use of multiple structures or templates, in the modelling process. Although the idea of including multiple templates have been used extensively in comparative modelling (Šali & Blundell 1993), there are few studies which directly assess what level of improvement multi-template modelling provides, though the intuitive rationale for such usage seems clear; with more evolutionary information available, better and more complete alignments can be generated for modelling (Sauder et al. 2000). Where multiple templates are available, the usual approach is to structurally superimpose the templates in order to generate a multiple-structure based alignment, and then to align the target sequence with the alignment generated from the structural superposition. The structural information is then assumed to be a more reliable indicator of residue conservation in the alignment

(Fiser 2004).

Venclovas & Margelevicius (2005), in CASP6, extended this multiple template approach by collecting 3D models from publicly available servers, and for each target, superimposed these structures with the templates found from sequence alignment. From the pairwise structural alignment of models to the templates, they then extracted a multiple sequence alignment and obtained estimates of alignment reliability for each portion of the sequence using the consensus structural information. Interestingly, the results of this study found that in some cases, models built from multiple templates were not as good as the models produced by simply using the single closest template, though multiple template predictions were very successful for the medium/high homology (*CM/easy*) targets. However, they suggest that because it is often difficult to select the best template, especially for more remote homologues, using multiple templates at least increases the probability that the best template will be selected.

Earlier in CASP5, Contrera-Moreira et al. (2003) used a similar approach which used a genetic algorithm for simultaneously searching template and alignment space. A number of templates were found by searching the template databases against the target sequence, and a number of models with alternative alignments were then generated for each template. A population of these models is grown by selecting pairs of protein models from the set of templates and applying two simple genetic operators to each pair to produce offspring. The recombination operator selects two proteins and a single crossover point and the region from the N-terminal to the crossover point of one protein is then joined to region spanning the crossover point to the C-terminal of the other selected model. Mutation simply involve structurally aligning two models and averaging the coordinates. Models generated by this approach were then scored with an energy function and then the worst 25% of models removed from the population. The algorithm continues until energy score convergence is reached. The authors concluded that in general, the models produced by this method were no better than the best initial model available, though in a handful of cases models were produced that were slightly worse or better than the best starting model.

In light of the inconclusive results of these recent studies, and taking into consideration the increasing numbers of consensus methods using structurally derived data for modelling, it seems a revised assessment of the value of multiple structure

modelling is necessary. More specifically, it is important to determine whether an improvement in model quality can be gained (for practical purposes) by assessing the value of such an approach *in situo*, that is, by using models from currently available automated template-modelling servers. In essence, by determining whether, under realistic conditions, structural information contained within a set of comparative models can be used to improve the quality of structure predictions, a justification can then be made for the use of multi-template/multi-structure approaches to model refinement.

2.1.3 Chapter summary

The focus of this chapter is to investigate whether model quality can be improved using the structural information contained within a collection of homology model predictions. Models are optimized using a multi-objective genetic algorithm in order to explore the effects of various objectives on the quality of the models. The primary objective function is used to optimize a model's structural similarity to the experimental structure. Secondary objectives are added to examine their effects on model quality. Two domain-specific objectives are introduced; a term for optimizing sequence coverage, and a term used to favour a single template-based model during model construction.

There are three motivations for this study; (*i*) to determine whether multi-template/multiple structure modelling is likely to be a fruitful approach to structure prediction for template-based models, (*ii*) to quantify the likely improvements if they are attainable, and (*iii*) to explore how the quality of the model is effected by functions which optmize various desirable model properties under a multi-objective framework. The outcome of this study has implications for future work on the refinement problem in particular, though may also act as a guide for the structural community by indicating whether efforts should focus on improving protein structure models using a multi-template modelling approach, or alternatively to focus on better selection schemes for detecting the single best template and alignment from which to begin refinement. The decision to use of homology models rather than the templates for those models is in order to evaluate the effectiveness of this approach as an automated procedure though the methods ability to refine structures is also evaluated.

The results of this work indicates that in the majority of cases examined

(including many of the more difficult targets), a statistically significant improvement in model quality is obtained by multi-objective refinement. Targets with more distant evolutionary relationships i.e. with low sequence identities, see a greater improvement in the model quality than the high homology (*CM/easy*) targets overall, though most targets were improved over the best initial unrefined model. This suggests that using multiple models for refinement can improve template-based structure predictions, and that a genetic algorithm approach is an appropriate choice of algorithm for this task. Although the results obtained in this study are positive, it remains to be seen whether the multi-template fragment assembling method exhibits similar efficacy using an imperfect scoring function at high-resolution in a true refinement test.

2.2 Methods

2.2.1 Formal definition

Starting with a set of models, M, produced by a template-based modelling approach for a target sequence of interest, and an experimental structure, E, for that target, construct a refined model, m_{ref}, using fragments from any of the N models such that

$$f(m_{ref}, E) > f(m, E) \qquad \forall\ m \in M \tag{2.1}$$

where, $f(m,E) \mapsto \Re$, is a function mapping the structural similarity between a model, m, and an experimental structure, E, to a real value in the closed interval [0,1]. This formalism can be viewed as an optimization problem where the set of all possible fragments contained within the models defines the search space, and the objective is then to recombined these fragments in order to maximise the objective function $f(m,E)$.

For multi-objective optimization, additional objective terms are introduced so that a set of functions, $\boldsymbol{f}(m,E) \mapsto \Re^n$, is obtained. In contrast to the single-objective optimization case where a single "best-fitness" solution is the product of an optimization, multi-objective optimization results in a set of solutions, $\boldsymbol{A}$, in which the objective vector of each solution is non-dominated with respect to all others within the set (see Section 1.8.4.1 for a definition of the non-dominance relation).

2.2.2 A multi-objective genetic algorithm framework

The evolutionary algorithm used in the work is the Strength Pareto Evolutionary Algorithm (SPEA2) (Zitzler, Laumanns & Thiele 2002). The original SPEA algorithm (Zitzler & Thiele 1998*a*) outperformed all non-elitist multi-objective methods in previous tests (Zitzler et al. 2000), and the improved SPEA2 algorithm incorporates an elitist strategy with an improved fine-grained fitness assignment algorithm to further enhance the method's performance.

2.2.3 Modifying the SPEA2 algorithm for model refinement

The general SPEA2 algorithm was adapted for protein structure model refinement by defining the encoding, operators, control parameters, and other algorithmic details. An outline of the modified algorithm is described in Algorithm 1.

Algorithm 1: Modified SPEA2 algorithm for multi-template refinement

input : M (a set template models)
N (the population size)
$\overline{N}$ (the archive population size)
T (the maximum number of generations)
P_c (the crossover rate)
P_m (the mutation rate)
S (structural alignment method)
output: A (a set of non-dominated solutions)

begin

```
/* Step (1): initialization                                    */
```
$P_0 \longleftarrow$ SEED_TEMPLATE(M)
$\overline{P}_0 \longleftarrow \emptyset$
$t \longleftarrow 0$

```
/* perform an initial multiple-structure alignment             */
```
if S = MULTIPLE **then**
  MULTIPLE_STRUCTURE_ALIGN(M)

repeat

```
  /* Step (2): fitness assessment                              */
```
  CALCULATE_FITNESS($P_t, \overline{P}_t$)

```
  /* Step (3): environmental selection                         */
```
  $P_{t+1} \longleftarrow$ ENVIRONMENTAL_SELECTION($P_t, \overline{P}_t$)

```
  /* Step (4): mating selection                                */
```
  $P_{mating} \longleftarrow$ MATING_SELECTION($\overline{P}_{t+1}$)

```
  /* Step (5): variation                                       */
```
  $P_{t+1} \longleftarrow$ VARIATION(P_{mating})

  $t \longleftarrow t+1$

until $t = T$

end

The SPEA2 algorithm uses two populations to provide elitist behaviour; a current

population P, of size N, and an (initially empty) archive population $\overline{P}$, of size $\overline{N}$, that holds all non-dominated solutions. The inputs to the algorithm are a set of template-based models, M, the control parameter values $C = \{N, \overline{N}, T, P_c, P_m\}$, and a preference value for the structural alignment method (see Section 2.2.5).

The algorithm then proceeds as follows; (*i*) in the initialization step the first generation is seeded with models selected at random from the model collection until the population P reaches its maximum size, N. If a multiple structural alignment method is chosen, the complete template-model set is structurally aligned so that all models share a common reference frame, (*ii*) in the fitness assessment step the objective scores are calculated for all members of the population, P, and all non-dominated individuals are then copied into the archive population, $\overline{P}$. If any members of the archive population are dominated or duplicate (in terms of objective scores) these individuals are then removed. If the archive is still larger than the archive size, $\overline{N}$, then further individuals are removed after a clustering technique is used which preserved the features of the non-dominated front. Fitness values are then assigned to members of both populations, (*iii*) mating then proceeds by selecting individuals from a union of archive and regular populations to form a mating pool, P_{mating}, using an environmental selection scheme (Zitzler, Laumanns & Thiele 2002), (*iv*) variation is applied to individuals in the mating pool so that after recombination and mutation, the old population P_t, is replaced with the offspring population, P_{t+1}, and the next iteration begins again at step (*ii*) until a termination criterion is met.

One predominant feature of SPEA2 is its elitist behaviour which is obtained by allowing non-dominated individuals from the archive population, $\overline{P}$, to participate in the genetic operations with the current population, P, in the hope of steering the population towards good regions of the search space (Deb 2001). Elitism ensures that the fitness of a population's best solutions do not deteriorate. Therefore, good solutions are never lost during the course of the algorithm unless a better solution is found to replace it.

2.2.4 Representation

The choice of chromosome representation is paramount for genetic algorithms (see Section 1.8.3.1) and the choice of encoding plays a large role in determining the performance of an EA (Liepins & Vose 1990). Two standard forms of encoding

are available for protein structures; a representation that encodes the structure as a function of the dihedral angles of the main-chain (ϕ/ψ) and/or side-chain χ angles, or alternatively, by representing the atomic coordinates of each residue in Cartesian coordinates.

Representing a protein structure as a function of its dihedral angles requires (at minimum) two degrees of freedom per residue. In Cartesian space each atom is represented explicitly, so depending on the detail of the representation, $L(k \times 3)$ atoms are required to represent a protein, where k is the number of atoms per residues and L is the total number of amino acids.

Although computationally less efficient than the dihedral representation, Cartesian coordinates offer an advantage in that non-continuous chains can easily be represented. For template-based models, the atomic positions of atoms are arranged in space in accordance with the aligned template residues. The major problem with a dihedral representation of non-continuous template models is that in order to reconstruct the backbone from a set of dihedral angles there must be complete chain continuity. Given that many template-based models lack complete chains due to insufficient confidence in the sequence alignment for some residues, or as a result of alignment errors, generating a complete main-chain requires missing fragments in the model to be reconstructed using either loop modelling methods or *ab initio* modelling prior to the calculation and storage of the dihedral angles. Moreover, these modelling procedures are likely to introduce additional sources of errors. A Cartesian representation is therefore adopted in view of the limitations imposed by the nature of most template-based models.

2.2.4.1 Structural representation

A reduced representation of the polypeptide chain is used in this study, with each residue described by the main-chain atoms (N, C_α, C', O) and the C_β atoms. The main-chain atoms are sufficient to represent the necessary topological details of the fold and therefore side-chain atoms are excluded to improve computational efficiency. As the template models used in this study consist solely of the C_α atoms, the main chain atoms are reconstructed using a backbone reconstruction algorithm (Holm & Sander 1991).

2.2.4.2 Chromosome encoding

The chromosome encoding used in this work is a fixed length list of structures where the length of the list L, is equal to the size of the protein sequence under consideration. At each sequence position, i, along the chromosome the list position contains either a structure holding the residue information, in this case a set of atomic coordinates $(N, C_\alpha, C', O, C_\beta)$, or an occupancy marker to indicate that residue position at i contains no structural information in the model. The encoding and representation of a protein within the GA framework is illustrated in Figure 2.1.

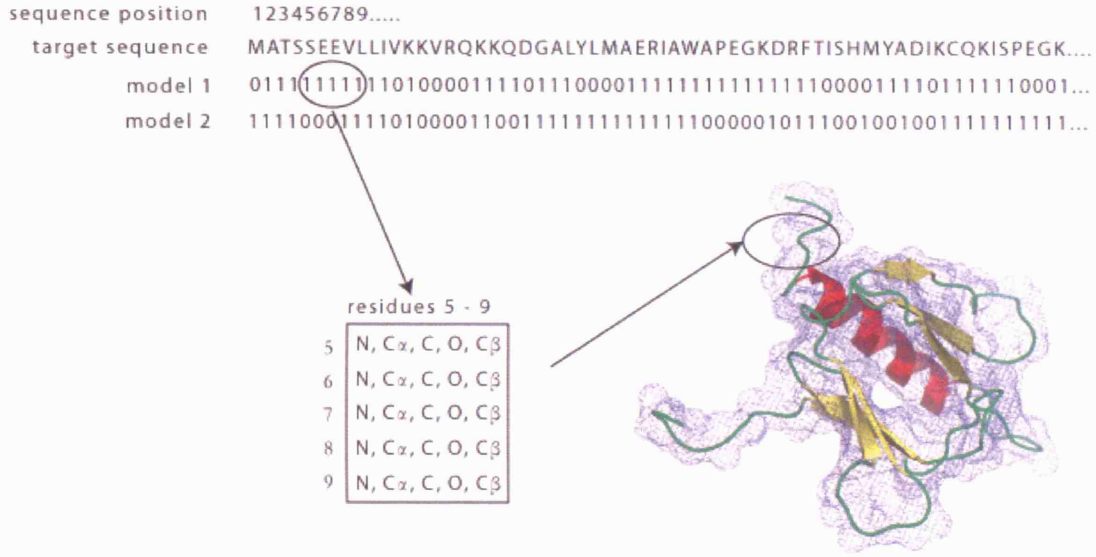


Figure 2.1: An individual chromosome is encoded as a sequential list of residue positions (genes) which can either be populated (1) or empty (0). Each of these gene positions holds a multi-dimensional array of atom coordinates represented as real valued triplets (x, y, z). In the example above, a model with sequence length, L, is shown. Residues 5 to 9 of the model are all populated with coordinates for the atoms, and respresent a portion of the N-terminal loop.

2.2.5 Structural alignments

Each of the template-based models in the input set is produced by a homology modelling procedure which aligns the target sequence to the residues of a template. Depending on the number of database hits and the degree of similarity of these templates to the target, these models are likely to share some general topological features such as a common fold or similar arrangements of secondary structural elements. Based on this assumption, a reduction of the search space available to the

GA can be made through a structural alignment of the models prior to, or during, optimization.

Structural alignments are generated between models so as to align the structures in a common reference frame before the application of the VARIATION() method. Two alignment methods are used; (*i*) the multiple structure alignment method MAMMOTH-mult (Lupyan et al. 2005), and, (*ii*) a sequence-dependent pairwise alignment method, the TM-score (Zhang & Skolnick 2004*b*).

In the case of a multiple structure alignment (MSA), the alignment of the input structures precedes the genetic optimization and is performed only once. This restricts the search space dramatically by "freezing" the structures in a common reference frame for the entire simulation. The alternative pairwise structural alignment method (PSA) requires the calculation of a pairwise alignment for every pair of structures that undergoes recombination and mutation.

2.2.6 Genetic operators

By combining a selection scheme with a crossover operator a GA is able to explore solutions which lie in more promising regions of the search space. In contrast, a selection scheme combined solely with a mutation operator reduces the GA to a stochastic hill climbing algorithm. The real power of the genetic algorithm paradigm lies in the combination of all three features. By constantly selecting good solutions from those generated by the recombination of chromosomes, the GA is able to explore the search space while the mutation operator acts to reduce the chances of convergence on a local minimum.

Genetic operators are applied through the VARIATION() method whereby models from the mating pool are selected for recombination and/or mutation. The VARIATION() algorithm is outlined in Algorithm 2.

Algorithm 2: VARIATION(P_m)

input : P_{mating} (mating pool)

output: P_{t+1} (a new population at $t+1$)

```
begin
    n ⟵ 1
    for i = 1 to max/2 do
        (m_j, m_k) ⟵ TOURNAMENT_SELECT(P_mating)
        (c_i, c_j) ⟵ RECOMBINATION(m_j, m_k)
        if FLIPCOIN() then
            c_i ⟵ MUTATION(c_i)
        if FLIPCOIN() then
            c_j ⟵ MUTATION(c_j)
        P_{t+1}[n] ⟵ c_i
        P_{t+1}[n+1] ⟵ c_j
        n ⟵ n+2
end
```

2.2.6.1 Crossover operators

Crossover, or recombination operators, enable the transfer of structural fragments between individuals in a population under the adopted encoding scheme. A multi-point crossover mechanism is adopted and two modified forms, (*i*) the fragment crossover, and, (*ii*) the fragment crossover with translation, are implemented to allow the exchange of fragments between models.

Fragment crossover is a representation specific implementation of the standard two-point crossover (Goldberg 1989) (see Figure 1.7). Using a similar mechanism to the canonical bit-string two-point crossover, fragment crossover proceeds by selecting two chromosome positions, i and j, at random to define the start and end-points of a fragment along the chromosome. The i position is chosen first then the j position selected by traversing the chromosome from N- to C-terminus by between 2 and 15 residues (again with the fragment length chosen at random with equal probability for each residue length). The regions spanning the start and end positions are then examined in two structures, m and n, which have been chosen for recombination. The

only constraint on the fragment crossover operator is that the i pivot positions are occupied in both structures. The structural region between the two pivot points is then exchanged between a pair of models by swapping the coordinates across structures. This is where the structural alignment step is crucial. Given that the models are off-lattice representations, the precise location of the two fragments after an exchange heavily depends on the orientation and similarity of the folds of the two models. Although the fragments are labelled according to the amino acid sequence (and hence retain their alignment positions in sequence space), the structural fragments may be misaligned with respect to the overall topology after a crossover.

For fragment crossovers with translation, the algorithm proceeds as described for a normal fragment crossover, however, the fragment from the donor model, m, is also translated so that a covalent bond is formed to preserved partial chain continuity with the new acceptor model, n. The translation step positions the fragment (without rotation) so that the bond between the fragment's N-terminus and the model's C-terminus form a bond with a length consistent with ideal values (Engh & Huber 1991). This is illustrated in Figure 2.2. This translation step is designed to reduce stereo-chemical errors by ensuring that partial continuity in the chain is preserved between the model and a newly inserted fragment.

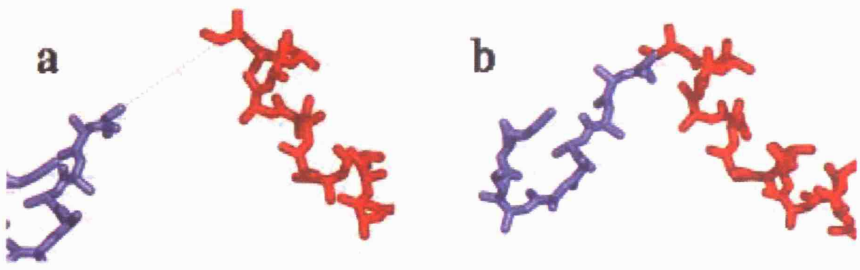


Figure 2.2: The two-point crossover with translation is designed to maintain chain continuity in the model accepting the excised fragment from a donor parent molecule. (a) shows the acceptor model (blue) after a fragment from a donor parent (red) has been inserted. The distance between the C-terminal of the acceptor molecule and the N-terminal of the fragment is reduced through a translation step so that the bond length between fragment N-terminal nitrogen and acceptor molecule C-terminal carbonyl carbon is idealised (b).

2.2.6.2 Mutation operators

Mutation events are expressed as single-position allele mutations. A mutation point, j, is selected at random along the chromosome. If the residue at position j is occupied then a model is selected at random from the original template-model collection and the coordinates exchanged between the chromosome of model m_j and the chromosome of the template-model n_j.

2.2.7 Control parameters

Genetic algorithms, like many machine learning algorithms, vary in performance in accordance with the parameter set used to control the algorithm. A GA's parameters influence each other in a nonlinear manner and so prevents the optimization of individual parameters (Mitchell 1999, Chap. 5.6). Various studies of parameter settings have highlighted some general features for optimal GA performance, namely, a small mutation rate (between 0.005 and 0.01), and a much larger crossover rate (0.75 to 0.95) (De Jong 1975, De Jong & Spears 1990, Grefenstette 1986, Schaffer et al. 1989), and more controversially, a small population size (Grefenstette 1986, Schaffer et al. 1989). However, it is difficult to derive a set of *a priori* principles for parameter choice because performance is often related to the problem type, representation encoding, and performance criteria of each problem.

Here, satisfactory (but not optimal) control parameters are obtained by performing optimization trials on a single protein with the assumption that the transference of the parameter set between proteins is valid. Values for the crossover rate, P_c, and mutation rate, P_m, are determined empirically by performing combinatorial trials where the values of the operator probabilities lie within the interval [0.1 1.0]. For each P_{ci}/P_{xj} combination, where $i = \{0.1, 0.2, ..., 1.0\}$, and, $j = \{0.1, 0.2, ..., 1.0\}$, a number of independent multi-objective optimizations are performed ($n = 100$), each with a different random number seed. The multi-objective GA used for control parameter selection uses a fragment crossover (see Section 2.2.6.1) with multiple structure alignments (see Section 2.2.5). The best performing parameters (see Section 2.2.9.1) are then applied to the remaining multi-objective optimization simulations.

There are also a number of SPEA2 specific parameters that are required by the multi-objective algorithm. The SPEA2 algorithm has a parameterless clustering

algorithm making it an attractive choice for multi-objective optimization. However, the size of the external population $\overline{N}$ must be chosen. The choice of values for the regular population size N and the external population size $\overline{N}$ must be balanced. If $N \ll \overline{N}$ a large elitist selection pressure results and can affect the algorithms ability to converge on the Pareto-optimal front. In contrast, a small external population size $N \gg \overline{N}$ can remove the elitism effect (Deb 2001). The authors of the SPEA2 algorithm suggest a 1:4 ratio between external and regular population sizes (Zitzler, Laumanns & Thiele 2002), and in this work $N = 200$ and $\overline{N} = 800$.

2.2.8 Objective functions

In protein structure prediction and folding algorithms the objective function is usually represented as an approximation to the free energy of a conformation (Lazaridis & Karplus 2000). In this study, an "ideal" objective function is used to determine empirically, what degree of refinement is possible for multi-template modelling if a perfect energy function existed. Instead of a statistical or physical energy function, a structural similarity measure is used to assess a model's quality rather than its energy, where quality is defined as the structural similarity of a model to the native structure. In addition, two other objectives are considered for the MOGA; a coverage term and a model bias term.

2.2.8.1 Structural similarity term

The structural similarity between a model and the native structure is measured using the sequence-dependent heuristic similarity measure, the template-model, or TM-score (Zhang & Skolnick 2004*b*). The TM-score is defined as

$$TM-score = f_{TM}(m,E) = \frac{1}{N_E}\sum_{i=1}^{N_{ali}} \frac{1}{1+(d+d_0)^2} \tag{2.2}$$

where m is the model and E is the experimental structure, N_E is the number of residues of the native structure and N_{ali} is the number of aligned residues in the threading alignment. For a full-length model, N_E and N_{ali} are identical, d_i is the distance of the ith C_α pair between model and native after optimal superposition, and $d_0 = 1.24\sqrt[3]{N-15}-1.8$ (see (Zhang & Skolnick 2004*b*) where this result is calculated). The TM-score is distance weighted so that small C_α- C_α distances are weighted more strongly than larger C_α- C_α distances (Zhang & Skolnick 2004*b*).

The objective maps a structural comparison to a real value in the closed interval [0,1], where a TM-score = 1 indicates two identical structures, while a TM-score < 0.17 indicates random structure pairs. If the TM-score > 0.5 then structures share the same fold.

2.2.8.2 Coverage term

The coverage term is the ratio of model residues to target sequence residues and represents the coverage of the target sequence by the model. The coverage objective function is

$$f_{cov}(m,E) = \frac{N_m}{N_E} \tag{2.3}$$

where N_m is the number of residues present in the model and N_E in the size of the target sequence. In a multi-objective scenario this term affects the genetic search by looking for solutions with a trade-off between more complete models and high similarity scores, assuming that these two objectives are in conflict.

2.2.8.3 Model bias term

The model bias term is used to guide the GA towards solutions which incorporate more of single template-based model. In this way it is possible to examine the solutions which result from a competition between objectives that improve the overall model quality and at the same time try to use as much of a single prediction as possible. The model bias objective is the proportion of residues in a model structure which come from a single starting template over the length of the chromosome.

2.2.9 Performance measures

Performance measures are required for assessing the quality of an approximate Pareto-set and also for assessing a model's structural quality.

2.2.9.1 Multi-objective performance metrics

One difficulty involved in multi-objective optimization is evaluating the performance of the algorithm where the outcome is a set of solutions instead of a single scalar "best-fitness" value that can be subjected to uni-variate statistical tests (Knowles & Corne 2002). In multi-objective optimization problems, the goal is to attain a set of solutions representing the global Pareto-front, although as stochastic optimization techniques the

outcome of a multi-objective optimization algorithm is more often an approximation to the global Pareto-optimal front. Moreover, the solutions in a MOP are multi-variate, and a MOGA produces multiple solutions.

Various metrics have been proposed to determine the quality of multi-objective solutions by comparing the approximate Pareto-set with the true global Pareto-set (absolute metrics). However, when the true Pareto-front is unknown, the performance of a MOGA can be estimated using various quality indicators which compare the relative performance of different MOGA implementations or simulation trials (relative metrics).

In this study, the hypervolume indicator, known as the $\mathcal{S}$ metric, is used to obtain an estimate of the quality of each Pareto-set (Zitzler & Thiele 1998*b*). The hypervolume indicator is described in detail in Appendix A. As the objective functions, $f_1(x), f_2(x), ..., f_n(x)$, are all bound within the closed interval [0,1], the reference point used here to obtain the hypersurface volume is (0,0). The same reference point is used for every multi-objective trial to ensure consistent ordering of solutions.

For parameter optimization (see Section 2.2.7), distributions of $\mathcal{S}$ metric scores are obtained from the 100 optimization runs for each P_{ci}/P_{mj} combination. The distributions are analysed using one-way ANOVA to determine whether there are significant differences between the distribution means, followed by a multiple comparison test using the Tukey method for mean comparisons (Winer et al. 1991) to determine which P_{ci}/P_{mj} combinations offer the best performance.

2.2.9.2 Model evaluation

All models are evaluated using the TM-score (Zhang & Skolnick 2004*b*), though the TM-score measure also calculates other commonly used sequence-dependent similarity scores, the MaxSub score (Siew et al. 2000), and the GDT score (Zemla 2003).

2.2.10 Data sets

2.2.10.1 Homology models

All data is obtained from the recent LiveBench-9 experiment (Bujnicki et al. 2001). Homology models are obtained from the automated fold recognition method mGenTHREADER (McGuffin & Jones 2003*b*) and the sequence-based profile-profile methods Distal-BASIC (Ginalski et al. 2004), the most sensitive method in the LiveBench-

9 experiment. Predictions were made for each of the 188 targets, where each target had available a maximum of 10 model submissions from each server. These 188 targets were filtered by removing targets for which both homology modelling servers submitted less than 10 models, leaving a total of 68 "easy" targets (*CM/easy*) and 28 "hard" targets ((*CM/hard*) (see Section 1.6.2 for a description of the easy/hard classification). For some of the targets the experimental structure was already included in the fold library before the predictions were made, and in such cases, this model was removed.

2.2.10.2 Benchmark data set

The filtered set of 96 LiveBench-9 targets were further reduced to a benchmark data set consisting of 50 targets by selecting targets randomly from the two categories so that 35 "easy" targets and 15 "hard" targets remain. 548 models were available for the 50 targets after the removal of any models built from the native template. The full target list in provided in Appendix B.

2.3 Results

2.3.1 Control parameter selection

Multi-objective trials were performed as described in Section 2.2.7 for each crossover and mutation rate combination. The distributions of $\mathcal{S}$ metric scores obtained after a total of 10,000 simulations were compared using one way ANOVA to determine whether there were any significant differences between the distribution means. Based on the observed ANOVA F statistic ($F = 26.47$), the null hypothesis, H_0, that there is no significant difference between the population means, was rejected (p-value <0.05). Figure 2.3 shows a boxplot of the distributions of $\mathcal{S}$ scores for all 100 P_{ci}/P_{mj} combinations. To determine the form of inequality among the samples, the Tukey-Kramer multiple comparison procedure is used in an *a posteriori* analysis. The Tukey method calculates the minimum significant difference (MSD) for each pair of distribution means, μ_{ij}. If the observed difference between a pair of means is greater than the MSD, it indicates that the m_{ij} pair differs significantly for each other.

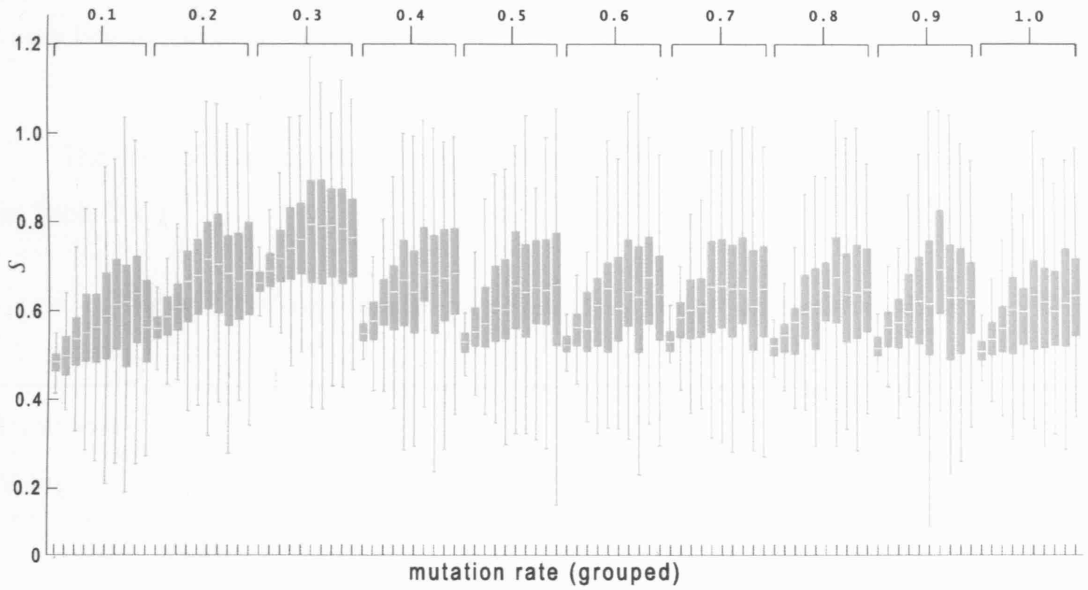


Figure 2.3: Boxplots are shown for $\mathcal{S}$ metric distributions obtained with different crossover and mutate rate combinations. The distributions are grouped by mutation rate with each group labelled $P_m = \{0.1, 0.2, ..., 1.0\}$. Each mutation rate group contains boxplots for the 10 distributions at different crossover rates, where $P_c = \{0.1, 0.2, ..., 1.0\}$.

The Tukey score was statistically significant for the majority of distributions within the group produced with a mutation rate $P_m = 0.3$. Figure 2.4 shows the

distributions of $\mathcal{S}$ metric scores for various crossover probability values at a mutation rate of $P_m = 0.3$. No within-group statistical analysis was performed, instead the optimal crossover rate $P_c = 0.7$ was chosen by visual inspection.

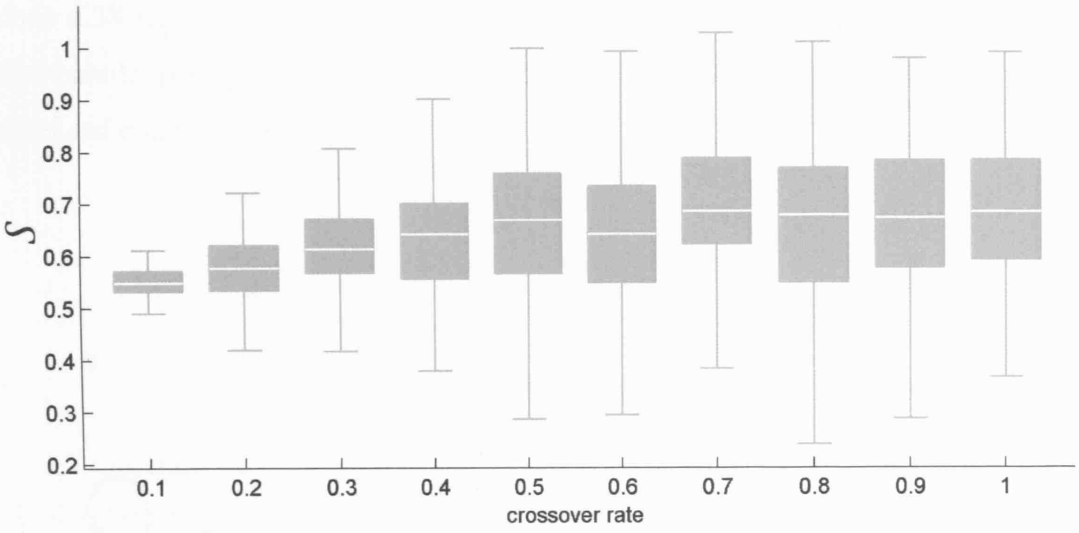


Figure 2.4: The distributions obtained after 100 trials using a mutation rate $P_m = 0.3$ with crossover rates between 0.1 and 1.0. The median $\mathcal{S}$ metric score increases with crossover rates $P_c > 0.4$ with the largest increase found at $P_c = 0.7$.

The final SPEA2 control parameters used in the remainder of the study are shown in Table 2.1.

Table 2.1: Control parameters for the modified-SPEA2 refinement algorithm.

Parameter	P_c	P_m	N	$\overline{N}$	T
Value	0.7	0.3	200	800	100

2.3.2 Selecting the best performance MOGA architecture

Two targets from LiveBench-9, *1nng*, a target from the "easy" category, and *1psy*, from the "hard" category, were selected for testing the multi-objective genetic algorithm implementations with different crossover mechanisms and structural alignment methods. The optimal combination of control parameters, variation operators, and alignment methods, is then used in a benchmark of the multi-objective refinement algorithm.

2.3.2.1 Easy target *1nng*

Chain A of target *1nng* is a 141 residue $\alpha+\beta$ domain in an $\alpha-\beta-\alpha$ fold (SCOP code d.38.1.1). The 1.95Å structure contains a central core helix surrounded by an anti-parallel β-sheet, while the C-terminal contains a short α-helix attached by a large extended coil region (Figure 2.5).

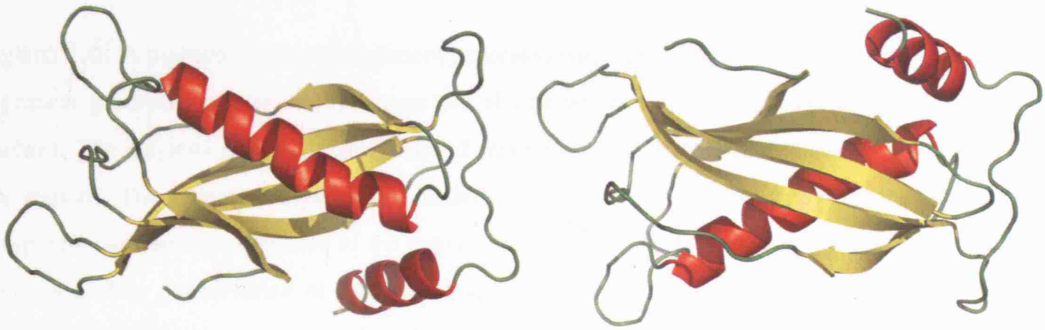


Figure 2.5: Chain A of test target *1nng*, an Acyl-Coa Thioester Hydrolase, is shown as a representative of the easy class in LiveBench-9. This 141 residue $\alpha+\beta$ chain has an $\alpha--\beta--\alpha$ fold as defined by SCOP (d.38.1.1). The topology of the fold is shown in cartoon representation. The leftmost figure shows the central helical structure, while the figure on the right, rotated on the x-axis by 180°, shows the curvature of the β-sheet.

The mGenTHREADER models for this target show a clear conservation of the fold with a strong agreement for the core regions of the protein after a multiple structure alignment (Figure 2.6).

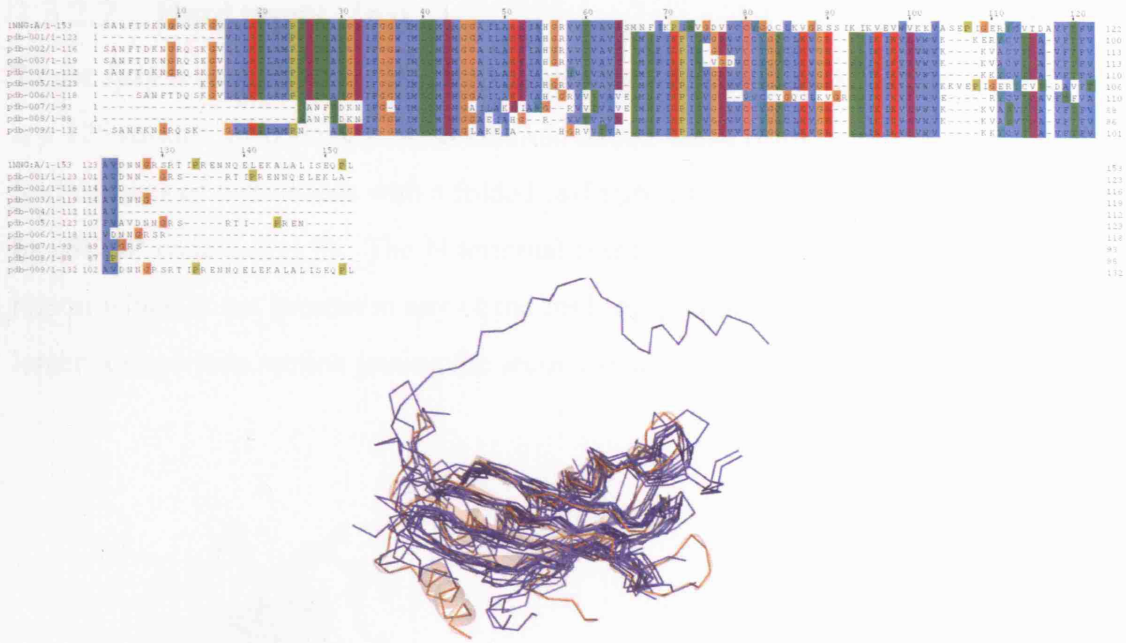


Figure 2.6: A multiple structure alignment generated with MAMMOTH-mult is shown. The sequence alignment generated by the multiple structure alignment shows clear conservation and alignment of residues. The majority of α-helix and β-strand regions are conserved with the poorer alignments in the loop regions. The structural alignment of models (blue) are shown the with native structure (orange). A transparent cartoon representation of the native chain is shown to delineate native regions from models. There is a clear conservation of the structural core within all the template models, with structural variation confined to loop regions. The C-terminal α-helix of the native structure is the only missing secondary structure element not found in any of the template models.

2.3.2.2 Hard target, *1psy*

Target *1psy* is representative of a hard template-modelling target. The native structure is a 125 residue all-α protein of 2S albumin RicC2 taken from *ricinus communis*. The fold consists of 5 α-helices with a folded leaf/right-handed super-helix fold as defined by (SCOP code a.52.1.3). The N-terminal contains a large extended 17 residue coil region which is not present in any of the fold recognition models, along with another larger surface loop section joining the second small helix to the third (Figure 2.7).

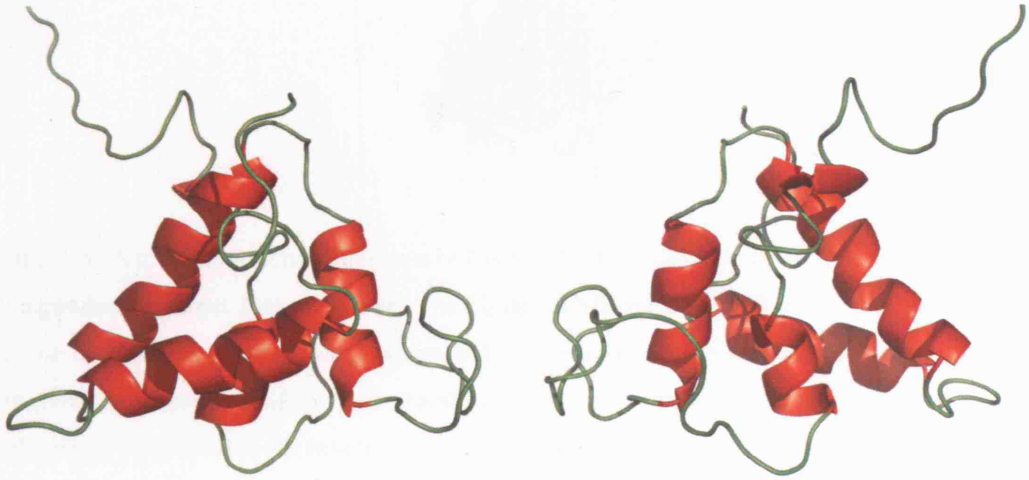


Figure 2.7: The 125 residue all-α structure for chain A of LiveBench-9 target *1psy* is shown in cartoon representation. The chain consists of 5 α-helices folded into a right-handed super-helical fold (SCOP a.52.1.3) with a large N-terminal extended coil region. The figure on the right shows the structure after a rotation on the y-axis of 180°.

A multiple structure alignment of Distal-BASIC models using MAMMOTH-mult is shown in Figure 2.8. The structural alignment shows a distinct lack of clear structural similarities between the Distal-BASIC models. No obvious structural core is conserved between these structures and only small regions show any similarity to the native structure. Moreover, some models contained within the template model sets are from different fold categories.

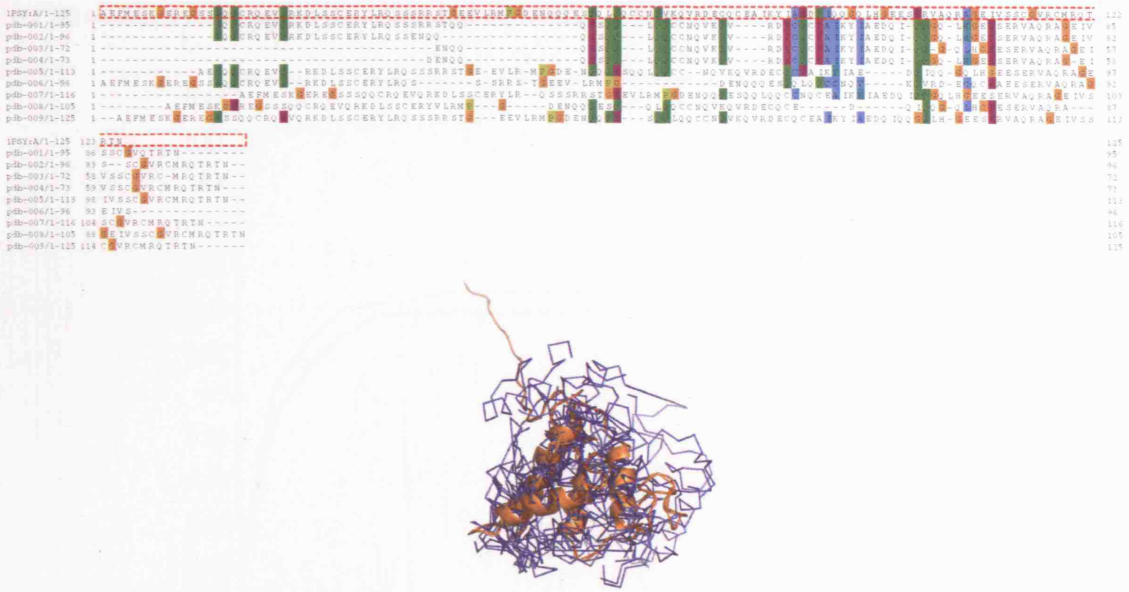


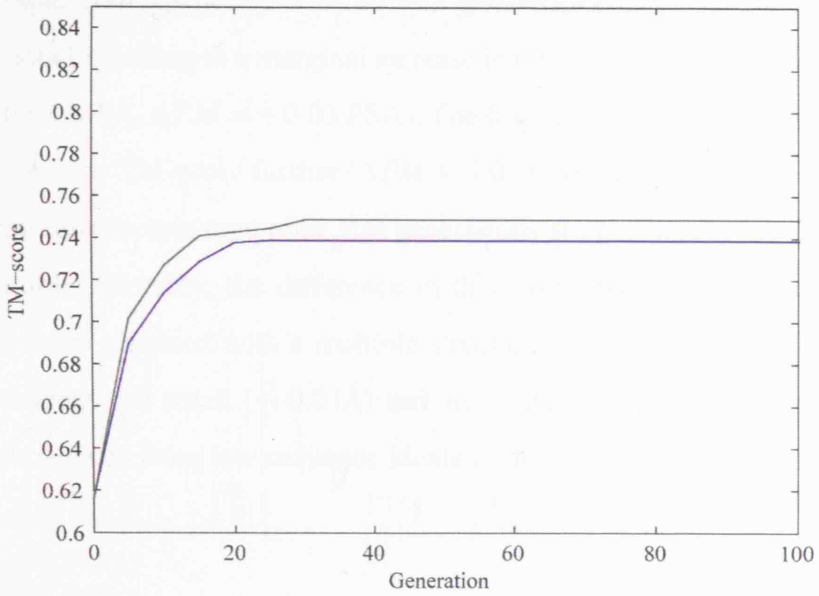
Figure 2.8: A multiple structure alignment of Distal-BASIC models calculated with the MAMMOTH-mult algorithm for target *1psy* are shown. The C_α main chain trace of the 9 models are shown in blue, while the native structure is coloured orange and represented in cartoon form. The upper part of the figure shows the sequence alignment generated by the multiple structure alignment. The poor quality of this alignment is reflected in the structural alignment shown in lower portion of the figure. There is little global or local consensus between regions that can be easily detected from visually inspection.

2.3.2.3 Assessing crossover operators

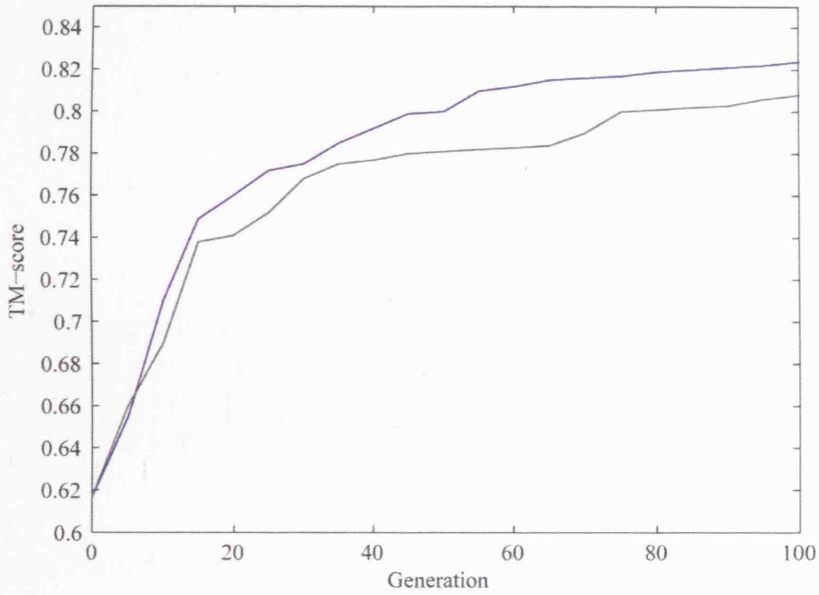
The exchange of structural fragments between models through the variation operators is the dominant feature of this refinement approach, and the performance of the two crossover operators (see Section 2.2.6) is examined by performing multi-objective optimization on the refinement test cases.

The multi-objective GA is used to refine the test targets using the fragment crossover operators and both structural alignment methods (see Section 2.2.5). The similarity and coverage terms are used as objective functions. Figure 2.9 shows the results of multi-objective optimization for target *1nng*. Both crossover mechanisms are able to improve the median TM-score of the population ($\Delta TM = +0.11$ using a multiple structure alignment, $\Delta TM = +0.13$ for pairwise alignments), however, the simple fragment crossover quickly converges after $\sim$30 generations under both structure alignment methods. Using the fragment crossover with translation leads to a greater increase in TM-score ($\Delta TM = +0.19$ for a multiple structure alignment,

$\Delta TM = +0.17$ for pairwise alignments) with no convergence after 100 generations.



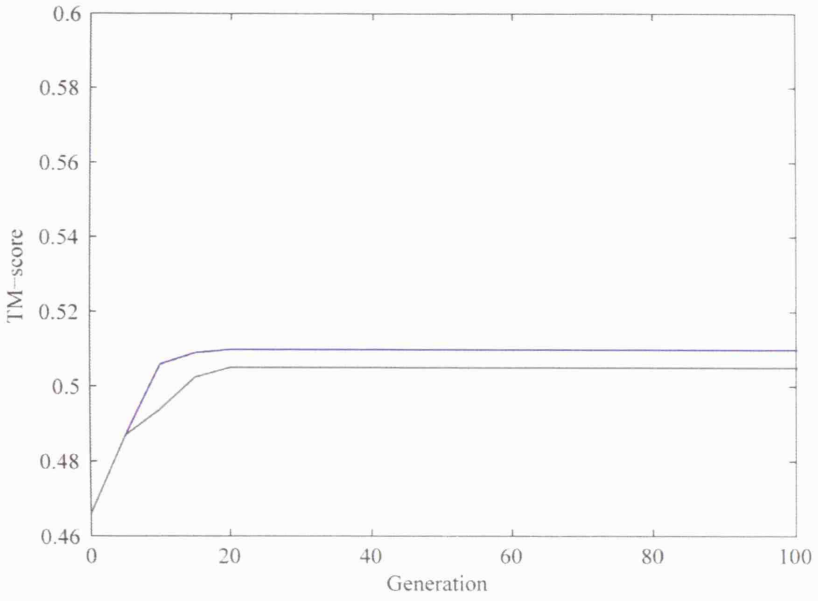
(a) Fragment crossover operator



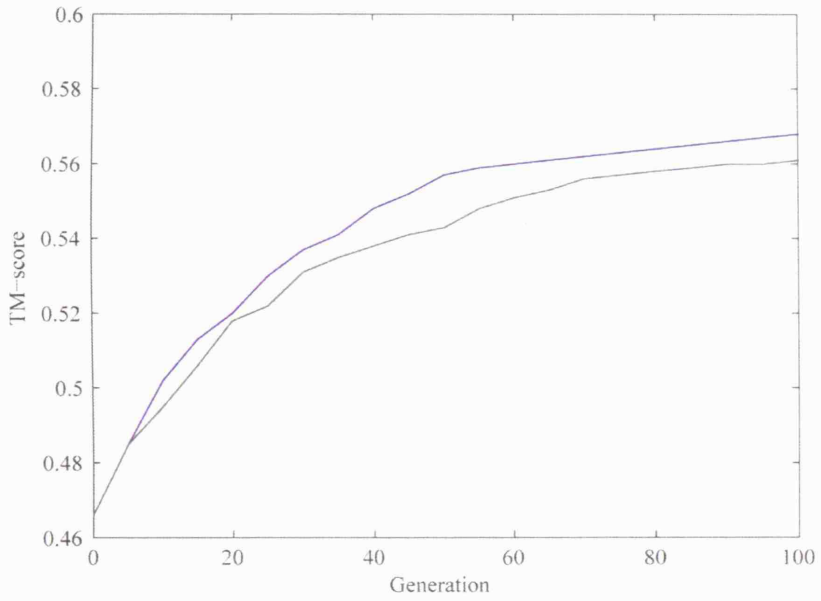
(b) Fragment crossover with translation operator

Figure 2.9: Multi-objective optimizations performed on target *1nng* are shown. Figure (a) shows the median TM-score of the populations during the course of refinement with the fragment crossover operator. Figure (b) shows the optimization results using the fragment crossover operator with translation. The line shown in green represents the median TM-scores when the GA is seeded with a multiple structure alignment, while the blue line indicates results obtained with a pairwise alignment algorithm.

Optimization results for *CM/hard* target *1psy* are shown in Figure 2.10. A similar pattern of rapid convergence is seen at ~25 generations using the simple fragment crossover operator leading to a marginal increase in the median TM-score for this target ($\Delta TM = +0.04$ MSA, $\Delta TM = +0.03$ PSA). The fragment crossover with translation is able to improve the TM-score further ($\Delta TM = +0.10$ MSA, $\Delta TM = +0.09$ PSA) and the algorithm fails to converge after 100 generations suggesting further improvement may be possible. Notably, the difference in this case between the improvements in median TM-score obtained with a multiple structure alignment strategy and pairwise alignment strategy are small ($\sim 0.01\text{Å}$) and are a likely result of the poor quality of input models derived from low sequence identity templates.



(a) Fragment crossover operator



(b) Fragment crossover with translation operator

Figure 2.10: Multi-objective optimizations performed on *CM/hard* target *1psy* are shown. Figure (a) shows the median TM-score of the populations during the course of refinement with the fragment crossover operator. Figure (b) shows the optimization results using the fragment crossover operator with translation. The line shown in green represent the median TM-scores when the GA is seeded with a multiple structure alignment, while the blue line indicates results obtained with a pairwise alignment algorithm.

2.3.3 Assessing refinement under multiple objectives

Using the optimal control parameters, a pairwise alignment strategy, and variation operators consisting of a fragment crossover with translation and mutation operator, multi-objective optimization runs are performed with the similarity term and coverage term, and then with the similarity term and model bias objective.

2.3.3.1 Multi-objective optimization of *1nng*

Pareto optimal sets are obtained after 100 generations of multi-objective optimization on target *1nng*. Figure 2.11 shows the results of refinement using mGenTHREADER models. The Pareto-optimal sets show interesting features of the objective space. In all cases, TM-score improvements can be obtained by selecting solutions from the non-dominated set. The optimization with a model bias term shows that for this target, little TM-score improvements can be gained by adding to the single best template (2.11b), though improvements are possible when less of the best template is used (2.11a). Optimizing both structural similarity and coverage shows that a range of solutions can be constructed in which the TM-score is improved with varying degrees of sequence coverage. At the upper bound of 100% coverage, a model can be constructed that leads to no change in TM-score over the single best template when using the mGenTHREADER models suggesting that regions of the structure are improved which compensate for the errors introduced by the addition of a mis-oriented C-terminal helix.

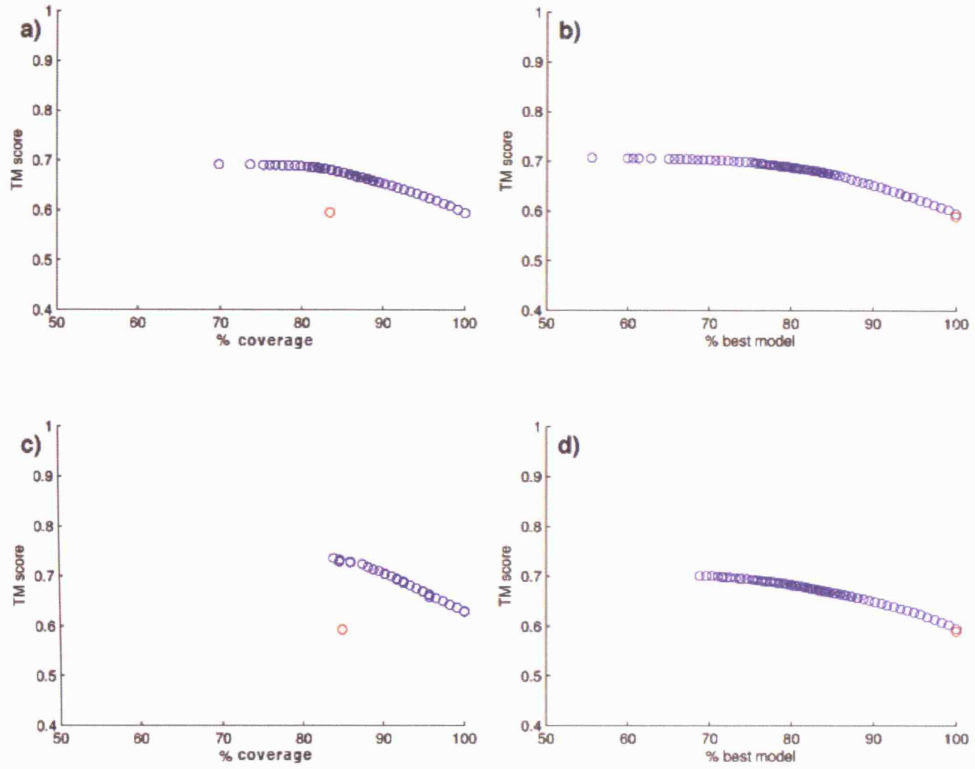
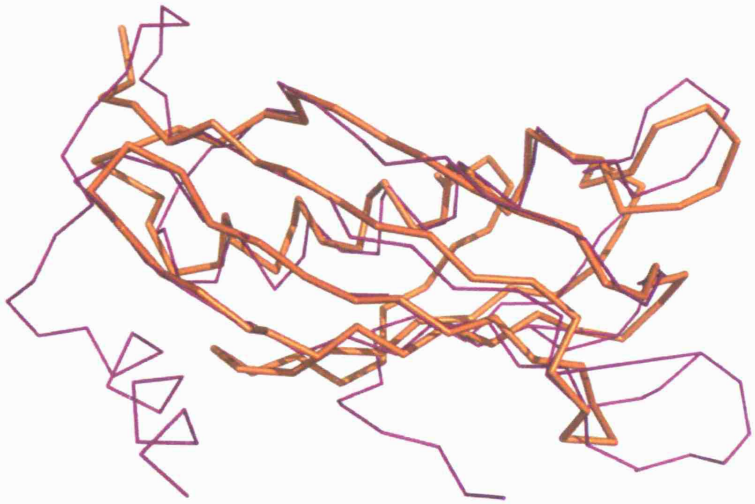
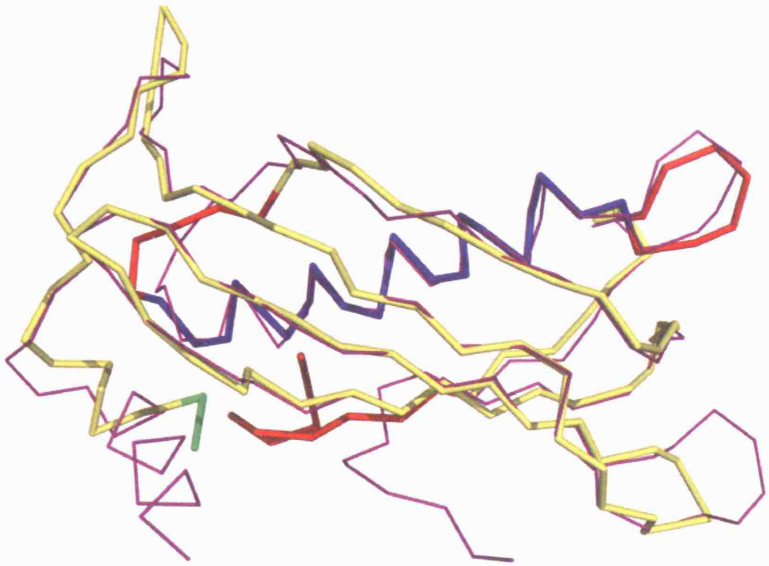


Figure 2.11: The MOGA optimization of target *1nng* leading to the four Pareto-fronts. Figures show (a) the optimization of mGenTHREADER models with similarity and coverage objectives, (b) the optmization of mGenTHREADER models with similarity and model bias objectives, (c) Distal-BASIC models with similarity and coverage objectives, and, (d) Distal-BASIC models with similarity and model bias objectives. The non-dominated solutions are shown in blue which the best starting template is highlighted in red.

Figure 2.12 shows an example of a refined model and the best unrefined structure for *CM/easy* target *1nng* taken from a later benchmark refinement (see Section 2.3.4).



(a) Highest TM-score unrefined mGenTHREADER model (84% coverage)



(b) The highest TM-score refined prediction (93% coverage)

Figure 2.12: The highest TM-scoring mGenTHREADER model for *CM/easy* target 1nng (Figure 2.12a) and an extreme point from the Pareto-front, representing the highest TM-score refined model (Figure 2.12b), after a multi-objective refinement are shown. The protocol used to refine the model using a crossover with translation, pairwise structural alignments, and population sizes $N = 400$, $\overline{N} = 1600$. Models are refined for 100 generations. The model shown by a thick orange line in (a) has a TM-score = 0.62 (RMSD = 4.25Å). After refinement the model shown in (b) has a TM-score = 0.82 (RMSD = 3.25Å). The coloured regions represent the unrefined model from which the fragment was extracted. The native structures in both figures are shown by thin purple lines.

2.3.3.2 Multi-objective optimization of *1psy*

Multi-objective refinement simulations on *CM/hard* target *1psy* are shown in Figure 2.13. In all cases, improvements to the model quality were small due to a poor set of input templates. However, greatest TM-score increases are found when the model coverage and similarity score objectives are optimized, similar to results obtained for the *CM/easy* test case (see Section 2.3.3.1). No increase in TM-score was possible when 100% of the best template was used showing that the GA was unable to improve the model by adding structure to a single template instead requiring a composite of templates to improve the model quality.

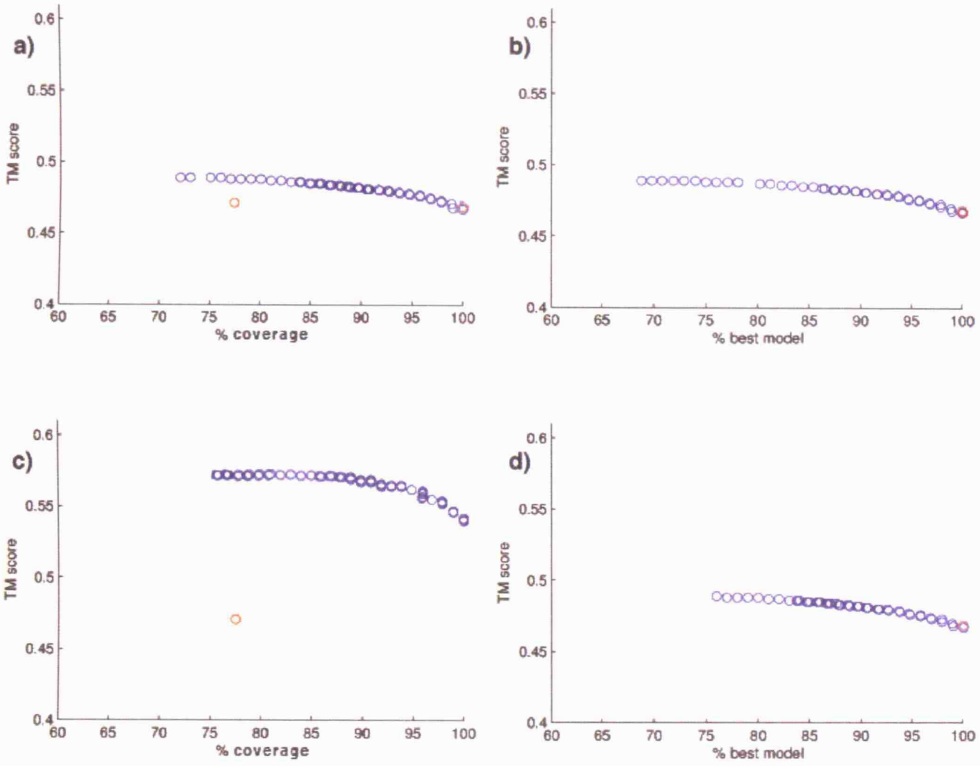
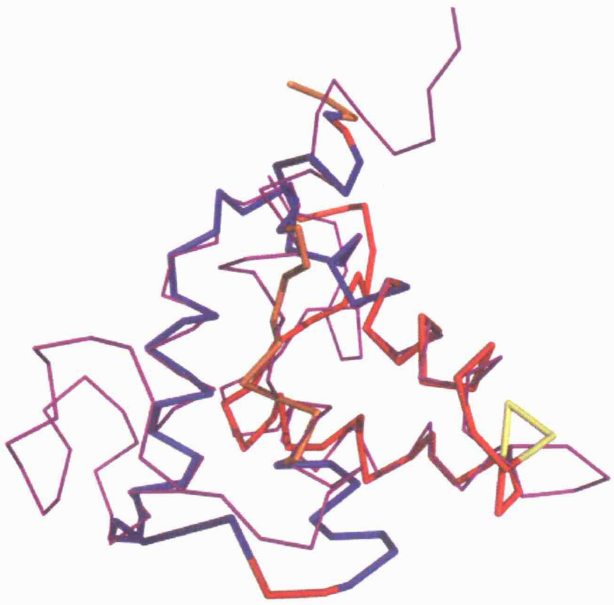


Figure 2.13: The MOGA optimization of target *1psy* leading to the four Pareto-fronts. Figures show (a) the optimization of mGenTHREADER models with similarity and coverage objectives, (b) the optmization of mGenTHREADER models with similarity and model bias objectives, (c) Distal-BASIC models with similarity and coverage objectives, and, (d) Distal-BASIC models with similarity and model bias objectives. The non-dominated solutions are shown in blue which the best starting template is highlighted in red.

Figure 2.14 shows a refined model and the best unrefined structure for *CM/hard* target *1psy* taken from a later benchmark refinement (see Section 2.3.4).

(a) Highest TM-score unrefined Distal-BASIC model (78% coverage)



(b) The highest TM-score refined prediction (80% coverage)

Figure 2.14: The highest TM-scoring Distal-BASIC model for *CM/hard* target 1spy (Figure 2.14a) and an extreme point from the Pareto-front, representing the highest TM-score refined model (Figure 2.14b), after a multi-objective refinement are shown. The protocol used to refine the model using a crossover with translation, pairwise structural alignments, and population sizes $N = 400$, $\overline{N} = 1600$. Models are refined for 100 generations. The model shown by a thick orange line in (a) has a TM-score = 0.41 (RMSD = 7.52Å). After refinement the model shown in (b) has a TM-score = 0.62 (RMSD = 4.17Å). The coloured regions represent the unrefined model from which the fragment was extracted. The native structure in both figures are shown by thin purple lines.

2.3.4 Benchmarking of the multi-objective refinement algorithm

Following the exploration of multi-objective optimization on two test cases, the multi-objective GAs with the best performing architectures are applied to 50 targets in order to benchmark the algorithms and generate solutions for evaluating the multi-template modelling approach. Two multi-objective GAs are used to refine structures for each of the 50 targets using the models submitted by the mGenTHREADER and Distal-BASIC automated servers; the first GA optimizes the structural similarity term and the model bias objective, while the second GA optimizes the structural similarity and coverage objectives. The control parameters are provided in Table 2.1 and used in both GAs, as well as a pairwise alignment scheme. The algorithms terminate when $t = T$. 10 trials are performed per target with each algorithm, and the average $\mathcal{S}$ metric score is used to assess the quality of the approximate Pareto-sets produced by the multi-objective GAs. The quality of the pre-refined models is calculated by considering the set of starting templates as a non-dominated front from which an estimate unrefined $\mathcal{S}$ metric score can be obtained.

Figure 2.15 shows the distributions of mean $\mathcal{S}$ metric scores obtained after multi-objective refinement of all mGenTHREADER predictions. The distributions represent the average $\mathcal{S}$ metric scores (one score per target) for the benchmark data set before and after multi-objective refinement. The distribution curves are estimated using kernel density estimation. To determine if there are differences between the distribution medians, a Kruskal-Wallis non-parametric ANOVA test was performed for non-normally distributed data. After the test the null hypothesis H_0 was rejected (p-value <0.05). A non-parametric multiple comparison test on ranks showed significant differences between the distribution of $\mathcal{S}$ metric scores for the unrefined models and the models refined with a multi-objective GA using structural similarity and model bias terms (non-parametric Tukey test, p-value = 5.2×10^{-4}), as well as between the unrefined model distribution and the distribution obtained after multi-objective refinement with structural similarity and coverage objectives (non-parametric Tukey test, p-value = 4.3×10^{-5}).

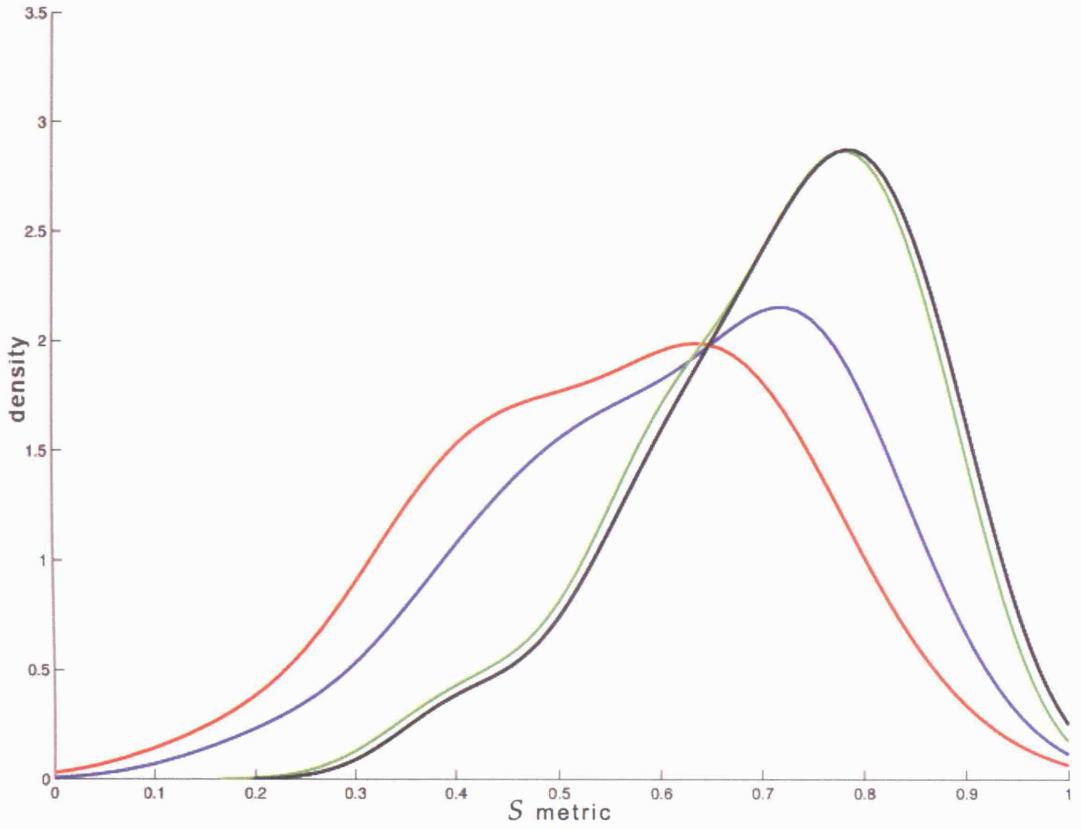


Figure 2.15: Kernel density estimated distribution curves of mean $\mathcal{S}$ metric scores are shown for non-dominated fronts produced by refining mGenTHREADER template-based model predictions for each target with a multi-objective strategy. The estimated unrefined $\mathcal{S}$ distribution is shown in red. The distribution obtained after multi-objective refinement with similarity and model bias objectives is shown in blue, and the distribution of $\mathcal{S}$ scores after multi-objective refinement with similarity and coverage objectives is shown in green. The distribution in black represents the multi-objective optimization with similarity and coverage objectives with a larger population size ($N = 400$, $\overline{N} = 1600$).

A further optimization was performed using the similarity and coverage objectives but with a larger population size $N = 400$, and archive population size $\overline{N} = 1600$. Although there was some improvements in mean $\mathcal{S}$ metric score with a larger population the difference was not statistically significant (non-parametric Tukey test, p-value = 0.18) (see Figure 2.15).

The same analysis, performed with Distal-BASIC input models, is shown in Figure 2.16. For the refinement of Distal-BASIC models a similar pattern to the refinement of mGenTHREADER models was found. After a Kruskal-Wallis test the null hypothesis, that there is no significant difference between any of the distribution medians, was rejected (p-value = 1.3×10^{-4}). A non-parametric Tukey multiple

comparison test found significant differences between the refined distributions and the estimated unrefined distribution, though the difference between the distributions of $\mathcal{S}$ metric scores for similarity and coverage objectives with a population size of $N = 200$ and a larger population size $N = 400$ was again not statistically significant.

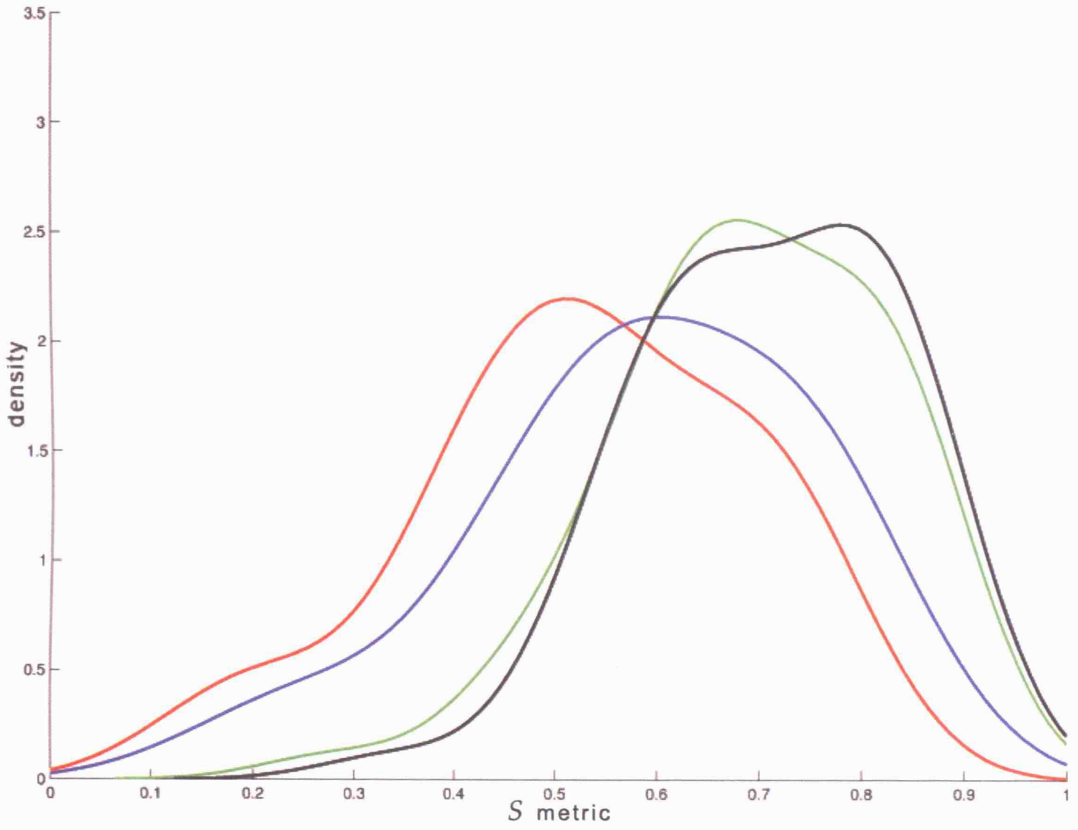


Figure 2.16: Kernel density estimated distribution curves of mean $\mathcal{S}$ metric scores are shown for non-dominated fronts produced by refining Distal-BASIC template-based model predictions for each target with a multi-objective strategy. The estimated unrefined $\mathcal{S}$ distribution is shown in red. The distribution obtained after multi-objective refinement with similarity and model bias objectives is shown in blue, and the distribution of $\mathcal{S}$ scores after multi-objective refinement with similarity and coverage objectives is shown in green. The distribution in black represents the multi-objective optimization with similarity and coverage objectives with a larger population size ($N = 400$, $\overline{N} = 1600$).

2.3.5 Estimation of the upper limits for multi-template refinement using multi-objective GAs

To estimate, from a practical perspective, what degree of improvement in model quality can be obtained by using a multi-template refinement approach, the extreme points (representing the models with the best TM-score in the non-dominated front after each refinement simulation) from the optimized Pareto fronts were selected and compared with the TM-score of the best starting model prediction from the homology modelling servers. Targets were split into *CM/easy* (35) and *CM/hard* (15) categories for the analysis.

Table 2.2 shows the original TM-scores and the TM-scores after refinement for the three GA implementations described in Section 2.3.4 for the 35 *CM/easy* targets in the LiveBench-9 benchmark data set. The mean increase in terms of TM-score for these targets using the optimal procedure (similarity and coverage objectives with a large population) was $\Delta TM = +0.15$ (mGenTHREADER) and $\Delta TM = +0.16$ (Distal-BASIC). Overall, both servers saw a similar mean increases in TM-score with similar standard deviations from the mean.

The same data for the hard targets are shown in Table 2.3. A marked increase in the TM-score was found for all 15 refinement targets with both servers. The mGenTHREADER models were significantly improved with an average increase of $\Delta TM = +0.23$. Distal-BASIC models were also improved gaining an average increase of $\Delta TM = +0.24$. These results indicate that the refinement procedure is slightly more effective at generating better models for more distantly related structures (i.e. harder targets) than for close homologues.

Table 2.2: Extreme points, representing the top TM-score models from non-dominated fronts, after MO refinement of 35 *CM/easy* targets are shown. TM^{start} is the TM-score of best unrefined model. TM^{a} shows the TM-score of the best model generated after refinement with the similarity objective and model bias objective. TM^{b} shows the top TM-score after refinement with the similarity and coverage objectives. TM^{c} shows the top TM-score after refinement with the similarity term and coverage objective with a larger population size.

	mGenTHREADER					Distal-BASIC				
pdb	TM^{start}	TM^{a}	TM^{b}	TM^{c}	$\Delta TM^{\S}$	TM^{start}	TM^{a}	TM^{b}	TM^{c}	$\Delta TM^{\S}$
1j26	0.435	0.454	0.553	0.546	0.111	0.378	0.438	0.522	0.550	+0.172
1j3v	0.716	0.780	0.847	0.873	0.156	0.736	0.806	0.838	0.869	+0.133
1nng	0.616	0.745	0.841	0.830	0.215	0.706	0.744	0.831	0.824	+0.117
1nrk	0.753	0.753	0.852	0.855	0.102	0.191	0.241	0.442	0.489	+0.298
1op4	0.392	0.408	0.638	0.629	0.237	0.413	0.451	0.598	0.605	+0.193
1p91	0.450	0.535	0.592	0.610	0.160	0.411	0.459	0.575	0.592	+0.181
1p9e	0.477	0.578	0.618	0.647	0.170	0.473	0.550	0.629	0.603	+0.129
1pvm	0.580	0.592	0.617	0.651	0.072	0.572	0.596	0.646	0.652	+0.080
1qwr	0.676	0.717	0.748	0.758	0.082	0.741	0.743	0.816	0.825	+0.085
1qxm	0.374	0.411	0.467	0.540	0.166	0.383	0.448	0.496	0.537	+0.155
1r1d	0.688	0.744	0.822	0.831	0.143	0.681	0.773	0.842	0.853	+0.172
1r4w	0.520	0.580	0.654	0.676	0.156	0.522	0.603	0.665	0.669	+0.147
1rku	0.671	0.749	0.835	0.831	0.160	0.666	0.737	0.855	0.873	+0.207
1te5	0.617	0.668	0.746	0.754	0.136	0.579	0.633	0.752	0.805	+0.226
1pmm	0.637	0.701	0.743	0.752	0.115	0.759	0.821	0.894	0.862	+0.103
1r57	0.849	0.855	0.861	0.866	0.017	0.513	0.578	0.646	0.709	+0.197
1v8c	0.405	0.468	0.618	0.662	0.257	0.404	0.442	0.580	0.619	+0.215
1s5a	0.631	0.764	0.860	0.900	0.283	0.776	0.808	0.861	0.871	+0.095
1vhe	0.398	0.443	0.574	0.590	0.192	0.484	0.567	0.624	0.623	+0.139
1q9j	0.735	0.744	0.775	0.785	0.050	0.506	0.536	0.676	0.646	+0.140
1ub9	0.758	0.811	0.794	0.850	0.092	0.728	0.781	0.847	0.868	+0.140
1p97	0.683	0.737	0.828	0.843	0.161	0.684	0.752	0.865	0.874	+0.190
1rli	0.693	0.755	0.808	0.825	0.131	0.712	0.764	0.845	0.810	+0.098
1v4e	0.733	0.817	0.885	0.892	0.159	0.750	0.756	0.838	0.834	+0.084
1uhw	0.564	0.566	0.722	0.723	0.159	0.573	0.592	0.669	0.671	+0.098
1vjx	0.733	0.775	0.821	0.807	0.074	0.732	0.780	0.829	0.844	+0.112
1rz3	0.654	0.758	0.858	0.866	0.213	0.653	0.724	0.801	0.813	+0.160
1sqh	0.356	0.390	0.401	0.401	0.045	0.455	0.533	0.739	0.752	+0.297
1rxx	0.856	0.899	0.912	0.936	0.080	0.550	0.614	0.632	0.639	+0.089
1t35	0.687	0.688	0.801	0.805	0.118	0.376	0.528	0.697	0.726	+0.350
1ri1	0.631	0.689	0.765	0.754	0.123	0.634	0.689	0.740	0.752	+0.118
1vjg	0.730	0.833	0.880	0.866	0.135	0.727	0.776	0.826	0.843	+0.116
1v63	0.498	0.539	0.715	0.740	0.242	0.504	0.576	0.700	0.693	+0.189
1vkh	0.605	0.771	0.800	0.814	0.209	0.617	0.706	0.764	0.790	+0.174
1ujx	0.595	0.676	0.766	0.768	0.173	0.582	0.631	0.721	0.757	+0.176
$<\mu>$	0.611	0.668	0.743	0.756	0.146	0.576	0.634	0.723	0.736	0.159
σ	0.135	0.141	0.125	0.121	0.062	0.144	0.139	0.120	0.114	0.064

§ TM-score difference between the best unrefined template model and best model from refinement protocol TM^{c}

Table 2.3: Extreme points, representing the top TM-score models from non-dominated fronts, after MO refinement of 15 *CM/hard* targets are shown. TM^{start} is the TM-score of best unrefined model. TM^a shows the TM-score of the best model generated after refinement with the similarity objective and model bias objective. TM^b shows the top TM-score after refinement with the similarity and coverage objectives. TM^c shows the top TM-score after refinement with the similarity term and coverage objective with a larger population size N.

	mGenTHREADER					Distal-BASIC				
pdb	TM^{start}	TM^a	TM^b	TM^c	ΔTM	TM^{start}	TM^a	TM^b	TM^c	ΔTM
1hl8	0.434	0.512	0.575	0.600	0.167	0.494	0.529	0.626	0.618	+0.124
1j3w	0.442	0.617	0.733	0.783	0.341	0.484	0.570	0.670	0.723	+0.239
1n8n	0.445	0.522	0.570	0.548	0.103	0.455	0.524	0.646	0.621	+0.167
1nmo	0.195	0.263	0.444	0.446	0.251	0.206	0.234	0.438	0.547	+0.341
1nnv	0.287	0.318	0.535	0.543	0.255	0.221	0.284	0.495	0.527	+0.306
1oap	0.428	0.523	0.758	0.705	0.277	0.372	0.452	0.574	0.609	+0.237
1paq	0.350	0.396	0.688	0.708	0.357	0.537	0.646	0.791	0.793	+0.255
1pfj	0.369	0.524	0.668	0.710	0.341	0.229	0.302	0.591	0.589	+0.360
1psy	0.467	0.497	0.609	0.619	0.152	0.471	0.501	0.583	0.624	+0.153
1q0d	0.333	0.390	0.670	0.690	0.357	0.388	0.420	0.717	0.766	+0.378
1v2y	0.551	0.618	0.737	0.752	0.201	0.549	0.583	0.631	0.660	+0.111
1uww	0.470	0.598	0.673	0.640	0.170	0.513	0.688	0.737	0.785	+0.272
1sr4	0.166	0.190	0.360	0.393	0.227	0.136	0.168	0.280	0.341	+0.205
1se9	0.648	0.684	0.747	0.776	0.128	0.619	0.694	0.767	0.772	+0.153
1v32	0.610	0.677	0.749	0.749	0.139	0.492	0.571	0.682	0.719	+0.227
$<\mu>$	0.413	0.489	0.634	0.644	0.231	0.411	0.478	0.615	0.646	0.235
σ	0.136	0.149	0.119	0.119	0.088	0.147	0.165	0.133	0.121	0.085

§ TM-score difference between the best unrefined template model and best model from refinement protocol TM^c

2.4 Discussion

In this study I have developed a framework in which aspects of a novel modelling approach to protein structure model refinement can be explored. A multi-objective genetic algorithm was used to successfully build composite models from a set of homology modelling predictions by using a fragment assembly approach to model construction. The algorithm was optimized using several test cases, and a detailed analysis of the features arising from multi-objective optimization was also performed on these test targets. Finally, a large scale assessment of the refinement capabilities of the method was examined, and estimates for an upper limit on the model quality attainable using the approach were calculated.

Previous studies have shown that template-based modelling procedures are reaching their natural limits (Contreras-Moreira et al. 2005) while others conclude that further structural refinement of template-based models is required for improving the accuracy of these structures (Moult 2006). Meanwhile, prior to these recent observations, the use of consensus strategies for model building that exploit the well-folded fragments present in a collection of homology models have gained favour as means to enhance structure prediction (Kolinski et al. 2001, Fischer 2003, Kosinski et al. 2005). In light of these recent trends, this study has endeavoured to explore elements of the structure-based approach to homology model refinement using sets of predictions (structural conformations) from automated homology modelling servers as starting points for refinement.

A heuristic search procedure was adopted to examine the feature space that arises from a structural modelling approach in which regions from multiple homology modelling predictions are combined under "ideal" modelling conditions (using information from the native structure as a guide). Therefore, this study should be considered as an exploratory investigation rather than as presenting a novel refinement method. In other words, we can ask questions about the suitability of such an approach to refinement in order to then make decisions about the direction of future research based on the evidence presented in response to those questions. In particular, issues specific to template-based modelling such as the limits to model refinement, the degree of structural coverage attainable, and the effects of such an approach on different modelling target categories, can give weight to the argument that structural information

from multiple models can enhance and improve structure prediction.

Before these questions were examined, a suitable architecture for the GA (the operators, control parameters, encoding, and alignment method) was determined by performing hand optimization of the control parameters and through simulations on simple test cases. The control parameters for the multi-objective GA (Table 2.1) were determined by empirical testing on a single protein (see Figure 2.3 and Figure 2.4) and the choice of the final parameters was validated using statistical ANOVA and multiple comparison tests. The additional components of the GA were selected in order to exploit the structural similarities often found between the structures of evolutionary related sequences by reducing the search space through structural alignments. The multiple structural alignment method, which structurally aligns all starting models once prior to refinement, results in slightly worse performance when compared with the pairwise structural alignment approach (Figure 2.9 and Figure 2.10). However, the choice of crossover operator has a much greater role in affecting the performance of the GA than either structural alignment method alone. The use of the fragment crossover operator results in rapid convergence of the algorithm (Figure 2.9(a) and Figure 2.10(a)) and suggests that the operator imposes rigid constraints on the search procedure. In contrast, the fragment crossover with translation, which allows movement of a fragment from its native position in a model, provides more flexibility in the modelling procedure which the GA exploits successfully to prevent early convergence (Figure 2.9(b) and Figure 2.10(b)).

The optimized multi-objective refinement GA was subsequently applied to two target proteins of varying modelling difficulties. The mGenTHREADER models for the first target (*CM/easy* target *1nng*) displayed large amounts of both structural and sequence consensus (as seen from the sequence alignment and structural alignments shown in Figure 2.6) with all models showing high degrees of similarity to the native structure. In contrast, the Distal-BASIC models for *CM/hard* target *1psy* show very little sequence or structural consensus (see Figure 2.8), and the similarity of these models to the native structure is poor, with only one structure scoring a significant TM-score (TM = 0.41).

In the first investigation, the multi-objective GA was used to explore the effect of sequence coverage on structural quality. A common feature of many template-

based models is the incompleteness of the structures. In the crudest model-building approach, the target sequence is threaded onto the backbone of a template and the coordinates extracted where a confident alignment was generated between the target and template sequence. However, where there is poor alignment quality (often in regions with high structural variability or low functional importance) a gap in the structure is often introduced that can break the continuity of the main chain, leaving regions with no structural coordinates. One of the main reasons for using multiple models for refinement is that there is an increased chance that native-like regions will be found in other evolutionarily related templates. By attempting to optimize both sequence coverage and structural quality with a multi-objective GA it was possible to examine the relationship between these two objectives and their effects on model refinement. The Pareto-fronts, which contain a diverse set of non-dominated individuals after refinement, show a typical convex pattern suggesting that there is a real conflict introduced by the dual optimization of both objectives (Figure 2.11(a), Figure 2.11(c), Figure 2.13(a), and, 2.13(c)). The model quality varies with the degree of sequence coverage, although in most cases the TM-score can be improved significantly by slightly reducing the sequence coverage, or to a lesser extent, by increasing the coverage. An illustration of this effect can be seen in Figure 2.12, taken from a later benchmark refinement on the test target. The refined model (Figure 2.12a) shows a significant increase in TM-score ($\Delta TM = +0.20$) with a slight increase in sequence coverage (93% versus 84%) compared with the unrefined model (Figure 2.12a). The composite nature of the model clearly shows the manner in which the GA has recombined structural fragments, with regions from four different starting structures used to improve both helical and strand regions leading to a higher quality structure than any of the homology models produced by mGenTHREADER. Similarly, the refined test target *1psy* from the harder modelling category shows in minor increase in sequence coverage after refinement (80% versus 78%) and large increase in TM-score ($\Delta TM = +0.21$) (see Figure 2.14).

An alternative question is to ask how much improvement can be made to a single model using fragments of other models within the starting set. An objective was encoded to favour models constructed with the greatest proportion of a single structure (selected prior to refinement as the most native-like structure (as measured by the TM-

score)) and then optimization carried out in conjunction with the structural similarity term. For both test cases, there was lack of improvement over the best structure at the extreme point where 100% of the best model is used, and an increase in model quality is only seen when less of the best structure is incorporated. This suggests that the use of multiple structures can be beneficial in some cases in order to improve a model over the single best template (see Figure 2.11(b), Figure 2.11(d), Figure 2.13(b), and Figure 2.13(d)).

While these detailed studies of multi-objective optimization highlight important features of refinement with multiple models, it is not possible to infer that this approach is consistently better at improving model quality from a small number of cases. Therefore, a large scale benchmark of the GA was performed to derive statistics from the refinements which could be subject to statistical tests. The S metric scores were used to compare the quality of the Pareto-fronts before and after refinement with the multi-objective GAs (Figure 2.15 and Figure 2.16). The distribution of mean S metric scores for mGenTHREADER models (Figure 2.15) and Distal-BASIC models (Figure 2.16) showed that an improvement in the quality of the Pareto-fronts is obtained after sampling with the GA. The larger hypervolumes achieved after refinement showed that an increase in both objective scores is attained relative to the unrefined homology models. Refinement using coverage and structural similarity terms provides the greatest increase in average S metric scores across the data set and increasing the population size to test the limit of this approach shows no significant difference between the population means after applying a non-parameteric test. It is likely that the GA has found the upper limit attainable with this approach using the described methodology.

A quantitative estimate was then obtained for the changes in model quality after refinement by selecting the top scoring model (using the TM-score) from the approximate Pareto-front for each target. These models were then used to estimate the relative improvement over the best starting homology model. By selecting an extreme point (i.e. the point on the Pareto-front with the maximum structural similarity objective score) the model achieving the top similarity score was selected at the expense of the other objective (sequence coverage or model bias), however, as the TM-score is designed to provide greater weight to regions of a structure that are most similar to the native, this approach gave useful estimates of the range of improvements that can be

expected.

Contrasting the average improvements in TM-score found after refinement of both "easy" and "hard" targets showed a larger increase in model quality for harder template modelling targets ($\Delta TM = +0.24$ in the best case refinement of Distal-BASIC "hard" targets compared with $\Delta TM = +0.16$ for the best approach on "easy" targets). One obvious explanation for this result is that the "hard" targets, by definition, are more difficult to model due the evolutionary distance between a target sequence and a detected template. This often leads to less structural conservation and hence poorer coverage and confidence in the aligned regions of sequence, in turn producing poorer models. In fact, the average unrefined TM-score of the best model for the "easy" targets is TM = 0.61 (Table 2.2) compared with TM = 0.41 for harder targets (Table 2.3).

These results are important for a number of reasons; first, they provide empirical evidence that combining the structural information contained within a set of homology models can increase the model quality over the best single model prediction. This therefore gives validity to the use of consensus approaches introduced in recent CASP experiments. Second, the quality of the alignment and model building stages in the homology modelling process do not seem to prevent the GA from finding better quality structures. This is important because although alignment shifts were not permitted in the refinement process the GA was still able to improve structural accuracy.

It is also worth noting that while this study has shown the value of a multiple model approach to refinement, the true limits to model quality were not examined. Although the multi-objective approach has provided evidence that large improvements are possible, to calculate a true limit would require a single-objective optimization of the structural similarity score to ensure the no restrictions are introduced by the dominance relations which arise from trade-offs between objectives in the multi-objective case.

Given that this work has both validated the use of multiple models in refinement and explored the feature space that arises from multi-objective optimization under "ideal" conditions (i.e. using the similarity scores as an objective function) I have shown, at least in principle, that there are potential benefits to such an approach. However, to test the efficacy of this method under true refinement conditions requires the substitution of the "ideal" objective function for an energy function capable of

discriminating good quality models based on the arrangement of atomic positions within the protein structure. The task of examining appropriate energy functions for assessing model quality is the focus of the next chapter.

Chapter 3

Benchmarking Energy Functions

3.1 Introduction

Successful protein structure prediction relies on the existence of energy functions that can accurately describe the stabilizing forces within protein structures. These functions are necessary to further our understanding of the properties of macromolecules as well as aid studies of the protein folding process. For protein modelling specifically, these energy functions must be able to both recognise the native tertiary structure of a protein from all other possible non-native conformations for that sequence, and also provide a relative ranking of non-native states. Energy functions have important applications in all major areas of structure prediction especially fold recognition (FR), homology modelling, and *ab initio* categories.

The majority of energy functions are derived with reference to the thermodynamic hypothesis (Anfinsen 1973), and more recently, the energy landscape theory of protein folding (Wolynes et al. 1995). The thermodynamic hypothesis states that a protein adopts the conformation (or ensemble of conformational states) which places it at global free energy minimum at physiological conditions. Energy functions are then designed so that they either attempt to reproduce the potential energy surface for a particular protein, placing the native conformation at the global free energy minimum, or else they are designed so that the native conformation is scored as the global minimum of a particular scoring function.

The ability to recognise the native state alone is of secondary importance given that reproducing the exact native state (corresponding to an experimentally resolved structure) is rare by using computational methods for all but a few modelling scenarios.

More important is the relative ranking of non-native states that occupy other regions of the energy landscape. An ideal discrimination function should place the native conformation at the global minimum of the function and ensure also that energy scores for non-native conformations correlate well with respect to their "nativeness". This property is usually measured by their structural similarity to the native structure (Cristobal et al. 2001).

The majority of energy functions can be categorised under two broad classes; physics-based functions, and knowledge-based functions. The main difference between these two approaches lies in the background assumptions and underlying principles used in their derivation.

3.1.1 Physical energy functions

Physics-based (empirical) energy functions use a series of approximations to quantum theory and the quantum mechanical descriptions of atoms and molecules in order to model the underlying physical forces which act to drive protein folding and to stabilize proteins (Lazaridis & Karplus 2000). These functions approximate the quantum model using an atomic model with Newtonian mechanics to describe the forces acting on and between particles. These molecular mechanics force fields are thought to represent the true effective energy function (the free energy of the protein and a surrounding solvent). They are typically composed of van der Waals, hydrogen bonding, electrostatics, and covalent terms together with a model of the surrounding solvent (see Section 1.3.3.1).

As the underlying true energy function is unknown, physical energy functions incorporate parameters which can be adjusted to alter the model. The parameters are generally obtained from experimental data, usually from small molecules or *ab initio* calculations (rather than experimentally resolved protein structures), and as a result these functions are also known as empirical energy functions (MacKerell et al. 1998). While physics-based functions are steadily improving, these methods have generally not been favoured for structure prediction due to their computational intensity and the inability of traditional molecular mechanics force fields such as AMBER and CHARMM to discriminate the native from alternative, non-native structures (Novotny et al. 1984, 1988). However, recent improvements to the solvation model by Lazaridis & Karplus (1999*b*,*a*) have greatly improved their native state recognition capabilities,

though recent evidence suggests that these functions are still not adequately modelling the underlying physical forces. One such study has shown discrepancies between experimental results obtained from quantum mechanical calculations of hydrogen bond energies and energies produced by molecular mechanics force fields, suggesting that most empirical energy functions treat hydrogen bonds improperly (Morozov et al. 2004).

While physics-based functions are ultimately required for studying the physical properties of biological macromolecules in solvent environments, as well as for folding pathway and ligand binding studies, statistical energy functions have had surprisingly more success at native state recognition.

3.1.2 Statistical energy functions

Statistical or knowledge-based energy functions are compiled by sampling the information contained within experimentally resolved proteins and are assumed to capture the combined effect of all stabilizing forces both within the protein and between protein and solvent (Sippl 1990). The statistical nature of these functions is reflected in the use of frequency distributions of some particular structural property, usually the distances between pairs of atoms or residues (Sippl 1995). Most statistical potentials are justified in terms of the Boltzmann principle (see Section 1.3.3.2) which relates observed frequencies of a particular spatial feature to the free energy. However, statistical energy functions are not constrained by an underlying physical model and therefore can be formulated in terms of pure statistical methods such as Bayesian statistics (Samudrala & Moult 1998), though these functions can not be used to study physical properties if a relationship to free energy can not be shown (Moult 1997).

In general, knowledge-based potentials have the following three characteristics; (*i*) a defined representation, (*ii*) the selection of a particular restrained spatial feature for extracting frequency distributions, and, (*iii*) a definition of the reference state which represents the random model or the Boltzmann average ensemble.

Early knowledge-based potentials employed a reduced representation of the protein, representing residues with a single interaction site, usually the C_α or C_β atom (Sippl 1995), or two sites located at the C_α or C_β atoms with an addition interaction site at the side-chain centroid (Bahar & Jernigan 1997). These residue-level potentials were

used successfully in fold recognition where the native fold is determined by assessing the quality of the sequence fit onto a particular backbone. The target sequence is usually assessed against a library of known folds from a non-redundant database of crystal structures (Bowie et al. 1991, Jones et al. 1992, Torda 1997). These functions are also used in context of *ab initio* structure prediction where a simplified representation of the protein must be matched in the energy function by ensuring that the forces captured in the potential are robust enough to the systematic errors introduced by a low-resolution protein model (Bonneau et al. 2001). Moreover, as *ab initio* structure predictions are often insufficiently accurate to model the finer details of a particular structure, identifying the correct fold from a set of sampled alternatives is usually the best expected outcome of these modelling methods (Jones 1997).

Simplified knowledge-based potentials are computationally efficient and less sensitive to small displacements, making them robust for low-resolution structure prediction applications like fold recognition where the inclusion of errors in inevitable. However, for protein structure prediction it was recognized that these reduced representations had intrinsic limitations (Mirny & Shakhnovich 1996). Samudrala & Moult (1998) showed that an all-atom discrimination function improved native structure selection by capturing the fine details of atom-atom interactions, while Lu & Skolnick (2001) were later able to obtain a similar discrimination ability with a reduced set of atom types by grouping side-chain atoms. Recently, a number of all-atom and reduced-atom statistical potentials have been developed for structure prediction and perform well at recognizing the native state on most decoy discrimination tests (Kuznetsov & Rackovsky 2002, Zhou & Zhou 2002)

Authors of knowledge-based energy functions have considered many spatial features for compiling potentials such as the distribution of residues between buried and exposed regions of proteins (Bryant & Amzel 1987, Bowie et al. 1991), atomic or residue solvation energies (Eisenberg & McLachlan 1986, Mallick et al. 2002), residue or atom packing densities (Gregoret & Cohen 1990), the environment surrounding individual residue types (Luthy et al. 1992, Eisenberg et al. 1997), similarity between model secondary structure and predicted secondary structure from sequence (Jones 1999*b*,*a*), stereochemical deviation from equilibrium bond lengths and bond angles, (Engh & Huber 1991, Morris et al. 1992, Laskowski et al. 1993), deviations from

favourable main chain dihedral angle distributions (Hooft et al. 1997, Lovell et al. 2003), and torsion angle energies (Gilis & Rooman 1997, Melo et al. 2002), and distance-dependent mean force potentials (Tanaka & Scheraga 1976, Miyazawa & Jernigan 1999, Hendlich et al. 1990, Sippl 1990, Colovos & Yeates 1993, Vajda et al. 1997, Moult 1997, Tobi et al. 2000) .

Distance-dependent potentials capture the detailed inter-atomic interactions as well as interactions between atoms and solvent. So far, these methods have proved to be the most successful at identifying correct models from sets of incorrect structures. Distance-dependent potentials may also be improved by augmenting additional structural features (e.g. residue solvent accessibility) to improve their performance (Kocher et al. 1994, Sanchez & Šali 1998). In a recent large scale evaluation of residue-level statistical potentials and their parameters, Melo et al. (2002) found that a combined potential consisting of residue-level distance-dependent potential with C_β interaction sites and an accessible surface potential performs better than any single feature alone.

3.1.3 Assessment methods

For newly developed energy functions, the traditional means of assessing their efficacy is by assessing their performance at discriminating the native structure from a set of decoy models (see Section 1.3.3.3). Comparisons of scoring functions can then be made by determining how well a particular function performs on different decoys sets in relation to other energy functions (Samudrala & Levitt 2000). As different conformational sampling methods are used to construct the decoy sets, the performance of an energy function on one set can not be used to assume that the scoring function will perform similarly on a different set of structures. Therefore, performing tests on a range of decoy sets generated by a diverse range of conformational sampling methods provides a more reliable evaluation of an energy function's discrimination capability (Park & Levitt 1996, Park et al. 1997, Keasar & Levitt 2003, Tsai et al. 2003).

An alternative assessment approach is to test energy functions by benchmarking them against the models produced by existing structure prediction algorithms. In this way, a practical assessment of the value of using these functions can be discerned for different types of modelling methods, thus enabling the community to identify the strengths and weaknesses of these approaches. To this end, a new sub-category,

Model Quality Assessment Programs (MQAPs), has recently been introduced as part of the fourth round of the Critical Assessment of Fully Automated Structure Prediction (CASFASP4) (Kelley et al. 1999). The MQAP category [1] provides a fully automated structure evaluation protocol and aims to address an outstanding problem in structure prediction, namely the accurate identification of native-like protein structure models. MQAPs assess the quality of models solely on the basis of their coordinates and do not compare the models with an experimental structure to determine their quality.

3.1.4 Chapter summary

This chapter benchmarks a selection of model quality assessment methods, including a novel method MODCHECK, against a set of homology models from various structure prediction algorithms. Using data from the semi-blind automated prediction experiment, LiveBench-9, this study tests the ability of four state-of-the-art programs at distinguishing the best model for each target from a set of predictions, and additionally, their ability to rank these models; the two requirements for an accurate scoring function. Following the benchmarking analysis, confidence estimates for the scores produced by each MQAP are calculated.

The application of structural quality assessment tools to models in this benchmarking study shows that while improvements in ranking were limited for modelling methods that already incorporate structural data, model rankings produced by purely sequence-based methods, including the best profile-profile methods, were improved. This results suggests that, contrary to popular opinion, the inclusion of structural information is useful in structure prediction even when using the best profile-profile methods, though the overall proficiency of these methods is still below the desirable level. The implications of this result for structural refinement are also discussed.

[1] http://www.cs.bgu.ac.il/ dfischer/CAFASP4/mqap.html

3.2 Methods

3.2.1 Model Quality Assessment Programs (MQAPs)

Four model quality assessment programs; MODCHECK, ProQ (Wallner & Elofsson 2003), Solvex (Holm & Sander 1992*b*), and victor/FRST (Tosatto 2005), were evaluated in this study and were selected at random from methods made available for the CAFASP4 MQAP [2].

Solvex is a knowledge-based solvation preference score based on an excluded volume approximation for protein-solvent interactions. Using the solvent contact model, Holm & Sander (1992*b*) derived solvation preference parameters from a database of 63 representative high-resolution proteins. The total solvation preference score, *Solp*, is calculated from the sum of atomic solvation preferences over all side-chain atoms. Solvex was shown to successfully identify the correct structure for a given sequence in three independent tests. Moreover, the method is robust to side-chain conformation errors introduced by modelling procedures and was shown to be able to identify misfolded proteins from the solvation preference evaluation alone, independent of the method used for side-chain placement.

The victor/FRST method is a knowledge-based potential composed of four terms; a distance-dependent pairwise potential, a solvation potential, a hydrogen bonding term for capturing backbone hydrogen bonding preferences, and a torsion angle potential. Each of the terms is combined into a single score using a weighted linear function under the assumption that each of the individual components contains orthogonal information.

ProQ, uses an alternative approach to the statistical energy functions included in the other MQAPs. Wallner & Elofsson (2003) adopt a machine learning approach using a neural network to detect the subtle correlations between a model's coordinates and the information learned from a training set of native structures using a combination of three spatial features; the fraction of atom-atom contacts, the fraction of residue-residue contacts, and solvent accessibility. The neural network generates an output comprising two scores; the LGscore (Cristobal et al. 2001), and the MaxSub score (Siew et al. 2000). These scores give a predicted indication of the structural similarity of a model to the native structure.

[2] http://www.cs.bgu.ac.il/ dfischer/CAFASP4/

3.2.2 MODCHECK

MODCHECK is a structure-based model evaluation function based primarily on classical threading potentials and is designed for assessing the accuracy of template-based modelling predictions. The MODCHECK function consists of two components; (*i*) a set of pairwise potentials of mean force (Hendlich et al. 1990) derived using the inverse Boltzmann equation with frequencies extracted from a set of high-resolution X-ray structures (Jones et al. 1992), and, (*ii*) a solvation potential (Jones et al. 1992).

The pairwise pseudo-energy term, first describe by Sippl (1990) is calculated for each pair of atoms. For specified atoms ($C_\beta \rightarrow C_\beta$ for example) in a pair of residues *ab*, with sequence separation k, and distance interval s, the potential is given by the following expression:

$$\Delta E_k^{ab} = RTln[1 + m_{ab}\sigma] - RTln\left[1 + m_{ab}\sigma\frac{f_k^{ab}(s)}{f_k(s)}\right] \tag{3.1}$$

where m_{ab}, is the number of pairs *ab*, observed with sequence separation k, σ is the weight given to each observation, $f_k(s)$ is the frequency of occurrence of all residue pairs at topological level k and separation distance s, $f_k^{ab}(s)$ is the equivalent frequency of occurrence of residue pair *ab*, and RT is taken to be 0.582 kcal/mol. In this work, short (sequence separation, $k \leq 11$), medium ($11 \leq k \leq 22$) and long ($k > 22$) range potentials have been calculated between C_β atoms only.

The solvation potential for an amino acid residue a, is defined as follows:

$$\Delta E_{solv.}^a(r) = -RTln\left[\frac{f^a(r)}{f(r)}\right] \tag{3.2}$$

where r is the degree of residue burial, $f^a(r)$ is the frequency of occurrence of residue a with burial r, and $f(r)$ is the frequency of occurrence of all residues with burial r. The degree of burial for a residue is defined as the percentage of accessible surface area for a given amino acid relative to its accessibility in a fully extended GGXGG pentapeptide.

One problem in classical threading is finding a suitable way to combine pairwise and solvation energy terms, and also to correct for the effects of protein size. Jones et al. (1992) achieved this correction by calculating Z-scores using the energy values for each hit in the fold library. Here a similar approach is used, but in this case the Z-scores are obtained by carrying out extensive sequence shuffling trials. By comparing

the native threading energies to those from sequence shuffled decoys, biases from both the fold, amino acid composition, and sequence length can be reduced. For a given model, the amino acid sequence is shuffled up to 100,000 times and the pairwise and solvation energy terms evaluated for each sequence shuffled model. To save computer time, the Z-scores of the model are assessed after 1000 shuffles and the calculation is terminated if both the pairwise and solvation Z-scores are lower than 6. The final MODCHECK quality score is derived by simply summing the pairwise and solvation Z-scores.

3.2.3 Automated structure prediction servers

Eight automated structure prediction servers were selected for inclusion from the participating entries to LiveBench-9. Servers were chosen so that there was at least one representative from each of the homology modelling methodology categories; sequence only, threading, or consensus methods. The fold recognition servers provide between 5 and 10 predictions per target in ascending order of predicted accuracy based on their internal scoring mechanisms.

Sequence only methods include FFAS, a sequence-based profile-profile method (Rychlewski et al. 2000), FFAS03, a third generation profile-profile alignment and fold recognition algorithm which uses a similar alignment method to FFAS. The FFAS03 version performs an additional profile-profile matching step and uses a different calculation of the alignment significance score by empirically evaluating the alignment score with respect to a distribution of raw scores obtained for pairs of unrelated sequences (Jaroszewski et al. 2005). ORFeus is another profile-profile alignment methods but includes predicted secondary structure information to create meta-profiles. This methods was found to be more sensitive at detecting remotely homologous relationships than the sequence-based profile method alone (Ginalski et al. 2003).

Four threading-based methods were used; 3D-PSSM (Kelley et al. 2000), an iterative threading method that combines 1D and 3D profiles with secondary structure predictions and a solvation potential; GenTHREADER (Jones 1999*a*), uses a single sequence alignment algorithm for detecting homologs where the sequence-structure fit is assessed using a set of statistical potentials and a neural network jury system

to produce a single assessment score; mGenTHREADER (McGuffin & Jones 2003*b*), is an improved version of GenTHREADER methods and uses PSI-BLAST sequence profiles instead of single sequences for the alignments. Structural alignment profiles and predicted secondary structure are also incorporated to improve the algorithm; SAM-T02 (Karplus et al. 1998, 2001) employs an iterative hidden markov model to generate sequence profiles using and combining predicted secondary structure information and the known secondary structure of detected templates to improve fold recognition.

One consensus method, Pcons-4 (Lundstrom et al. 2001) was included in the analysis. The Pcons-4 method takes predictions made by a selection of sequence and threading methods and returns the best models as judged by a neural network. The predicted score from the initial server and the overall fraction of other structurally similar models from the input collection are used by the neural network to obtain an estimate of the model's quality.

3.2.4 Model similarity measures

Model similarity measures provide a quantitative estimate of the structural similarity between a model and target protein (see Section 1.5). Often the term "model quality" has been used to designate the structural similarity between a predicted model and the native tertiary structure for a particular query sequence. However, it is important to clarify that the term "model similarity" is used in this context, whereas, in contrast, "model quality" refers to the predicted accuracy of a model (its similarity to the native structure) calculated from the model alone without reference to the coordinates of an experimentally resolved structure.

Model similarity is assessed using the MaxSub score (Siew et al. 2000) which calculates the largest subset of C_α atoms that can be superimposed over the native structure at a given distance threshold (3.5Å); the GDT_TS score (Zemla 2003), which calculates the superposition which optimized the percentage of C_α atoms that can be superimposed at distance ranges of 1, 2, 4, and 8Å ; and the 3D-score[1], a method that combines a rigid-body superposition with a contact measure.

[1] http://bioinfo.pl/Meta/evaluation.html

3.2.5 Assessing top model selection accuracy

The ability of an MQAP to select the best/top model is assessed by comparing the results produced by each of the fold recognition methods with the predictions made by the MQAPs. This test does not examine the ability of a fold recognition algorithm to find the best template but instead assesses the FR method's scoring function in comparison to the MQAP.

Each FR method produces an ordered set of models based on the scores given to each of the sequence-to-template assignments found in the list of detected homologous sequence hits. The similarity score, on the other hand, is calculated with reference to the experimental structure for the first FR model prediction from the list. Each of the FR models is then scored again with the MQAPs to produce a new rank order, and the similarity score of the first model in the new ranking is then calculated. Once the two similarity scores for the top model (as predicted by the FR method and MQAP) are obtained for each of the LiveBench-9 targets, a one-sided Wilcoxon matched-pairs signed-rank test is used to calculate whether there is a difference between the population means. The one-sided test ensures that only significant increases (p-value <0.05) in similarity scores using the MQAP program are recorded, where a significant score means that the MQAP improves top model selection over the original FR server ranking. The significance level used in this test is 95%.

3.2.6 Assessing model quality rankings

A non-parametric Pearson's R correlation, tuned on the ranks of the data, was used to compare the MQAP rankings with those provided by the original FR servers. The Spearman's rank test measures the direction and strength of the relationship between the rank order of two variables. To measure the ranking produced by the original FR server the similarity scores for the models are first calculated. These scores are then used to calculate the correlation coefficient, R, between the ranks of the models and the optimal rank order, defined by the similarity scores given to each of the models. The models are then re-ranked by MQAP score and the R values calculated for the new rank order. The quality of the ordering obtained with the MQAPs can then be compared with the original rank presented by each FR method by comparing the Spearman's R coefficient.

The overall ranking of the MQAPs is assessed using the combined set of models. For each target, all models generated by each of the eight servers were pooled and the Spearman's rank correlation R calculated over the set. This combined set was then split into "easy" and "hard" targets, and the Spearman's rank correlations calculated once more.

3.2.7 Determining method confidence estimates

A confidence estimate provides a measure of how reliable a quality assessment score is considered to be. In the case of model quality assessment, a reliable score can be considered as one for which the model has a MQAP score and similarity score above some given set of threshold values. For MODCHECK, a model is considered to be good if it scores a combined Z-score of 6 or greater. In the case of ProQ, models are considered correct if they generate a predicted LGscore >1.5 and a predicted MaxSub score >0.1. For the victor/FRST and Solvex scores, the magnitude of the final score is dependent on the particular protein size. Therefore, a threshold is obtained for each target by calculating the MQAP scores for a random prediction, in this case, by threading a random sequence onto the native backbone.

The confidence value for each MQAP is then estimated by plotting the number of true positives against false positives. True positives are counted as models which obtains a MQAP score above a given threshold and have a corresponding similarity score larger than some predefined threshold (MaxSub score >0.3 (30%) (Rychlewski & Fischer 2005), GDT_TS $\geq 25\%$, and 3D-Score ≥ 40). Models that achieve similarity scores below these threshold values in most cases adopt the wrong fold or topology, or they are bad structural models.

3.2.8 Data sets

All model data for this study were acquired from the fully automated LiveBench-9 experiment for 3D structure prediction (Bujnicki et al. 2001). For each of the eight servers (see Section 3.2.3), between 5 and 10 models were generated per target, and in total 13449 models were collected for the 188 targets. 11880 models were available after removal of corrupt data (i.e. late submissions where models are returned to LiveBench after the end of the experiment deadline or models built from the native template). The targets consisted of protein sequences between 100 and 500 residues

in length and were either single domain structures or single chains from multimeric proteins.

The 188 targets were additionally separated into two subsets using the "easy" and "hard" classification scheme employed by the LiveBench standard (see Section 1.6.2). Briefly, LiveBench defines easy targets as non-trivial sequences which do not share a close sequence homologue at BLAST e-value <0.01 (Altschul et al. 1990). Hard targets are those sequences for which standard sequence methods for homology detection fail to find a similar fold, that is sequences with a PSI-BLAST score with an e-value <0.001 (Altschul et al. 1997). 77 easy targets and 111 hard targets were made available in LiveBench-9.

3.2.9 Model re-construction

The models obtained from the LiveBench server contain solely the C_α atoms and many contain discontinuous chains (due to low confidence sequence alignments at some positions). In order to calculate scores for the model quality assessment methods, the backbone and side-chain atoms were generated on the fixed backbone using the CTrip modelling program (Petrey et al. 2003). No hydrogen atoms were added to the models.

3.3 Results

3.3.1 Improving top model selection

One of the main goals of research in the area of energy functions is to develop energy potentials that can consistently recognize the native tertiary structure for a query sequence and provide a relative ranking of non-native conformations. Here MQAPs are applied to the practical problem of selecting the best model from a set of predictions generated by common homology modelling methods. A one-sided Wilcoxon matched pair signed-rank test was used to measure the top model selection capability of the MQAPs relative to the original FR methods (see Table 3.1). If the difference between the population means was significant (p-value <0.05) then the MQAP was able to select a better model than the server for the LiveBench targets.

Both versions of MODCHECK are able to improve the top model selection for the three sequence-based methods; FFAS, FFAS03, and ORFeus. The improvements were often depended on the similarity score used though MODCHECK consistently improves FFAS and FFAS03 largely independent of the similarity measure. Using the GDT_TS score generates more significant results overall, and ProQ-LG and FRST are able to provide significant top model selection improvements for ORFeus and Pcons-4, respectively. However, Solvex performed consistently poorly in this analysis, as do the victor/FRST function and the LGscore component of the ProQ method. Interestingly, the top model selection for the threading methods (3D-PSSM, GenTHREADER, mGenTHREADER, and SAM-T02) are not improved by any MQAP method.

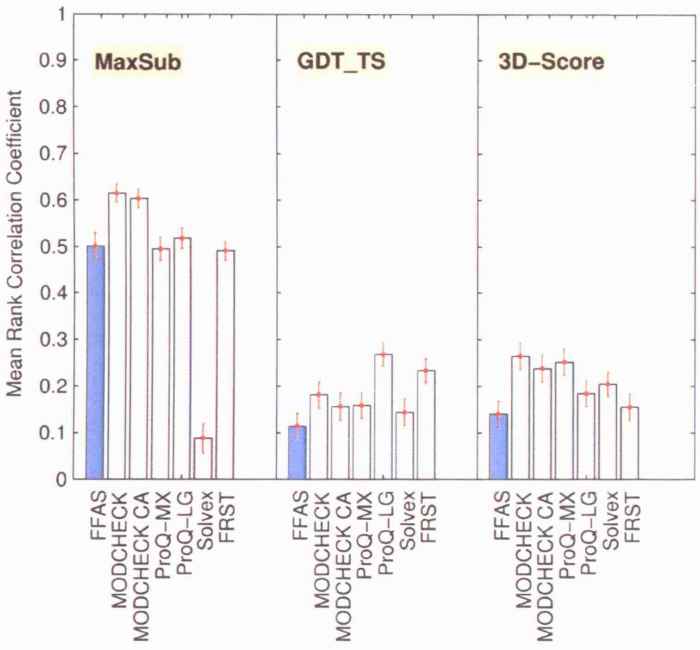
Table 3.1: The p-values from a one-sided Wilcoxon matched pairs sign-rank test are shown. This test compares the similarity score of the top model selected by the original server with the score of the top model selected by the MQAP for all 188 targets. A significant improvement in one method, y (the MQAP) compared with a method x (the original method) is determined at a significance level of 5% (p-value <0.05), and the one-sided test ensures that only significant improvements by the MQAP are reported. Significant results are highlight in bold type.

	MODCHECK	$MODCHECK_{CA}$	$ProQ^{MX}$	$ProQ^{LG}$	Solvex	FRST
MaxSub						
FFAS	0.1736	0.2391	0.9320	0.9987	1.0000	1.0000
FFAS03	**0.0007**	**0.0003**	0.2271	0.1540	0.9999	0.8752
ORFeus	0.2878	0.0985	0.6655	0.3043	0.9999	0.8716
3D-PSSM	0.9858	0.9987	0.9999	0.9999	0.9999	0.9989
GenTHREADER	0.8578	0.8825	0.9976	0.9958	0.9999	0.9962
mGenTHREADER	0.9619	0.9399	0.9888	0.9982	0.9999	0.9871
SAM-T02	0.7083	0.6676	0.7685	0.3328	0.9999	0.9994
Pcons-4	0.2234	**0.0036**	0.3129	0.4413	0.9994	0.0705
GDT_TS						
FFAS	**0.0003**	**0.0054**	0.2455	0.3692	1.0000	0.9999
FFAS03	**0.0067**	**0.0022**	0.1318	**0.0237**	0.8878	0.0875
ORFeus	**0.0156**	**0.0309**	0.2170	**0.0159**	0.9747	0.3578
3D-PSSM	0.2402	0.5946	0.9321	0.9075	0.9999	0.8463
GenTHREADER	0.4393	0.3598	0.7923	0.4934	0.9999	0.6318
mGenTHREADER	0.9284	0.8605	0.8579	0.9424	0.9999	0.9659
SAM-T02	0.7999	0.7204	0.4083	0.1558	0.9754	0.6182
Pcons-4	0.2139	**0.0452**	0.4067	0.1727	0.8248	**0.0116**
3D-score						
FFAS	**0.0094**	0.1640	0.6261	0.9919	1.0000	1.0000
FFAS03	**0.0006**	**0.0006**	0.6601	0.6785	1.0000	0.9999
ORFeus	0.6477	0.6902	0.9993	0.9959	1.0000	0.9999
3D-PSSM	0.3684	0.8172	0.9999	0.9999	1.0000	0.9999
GenTHREADER	0.9939	0.9960	0.9980	0.9988	1.0000	0.9998
mGenTHREADER	0.9975	0.9929	0.9999	0.9999	1.0000	0.9999
SAM-T02	0.4947	0.3502	0.9974	0.9999	1.0000	1.0000
Pcons-4	0.4793	**0.0345**	0.9634	0.9615	1.0000	0.9791

3.3.2 Improving the rank order of models

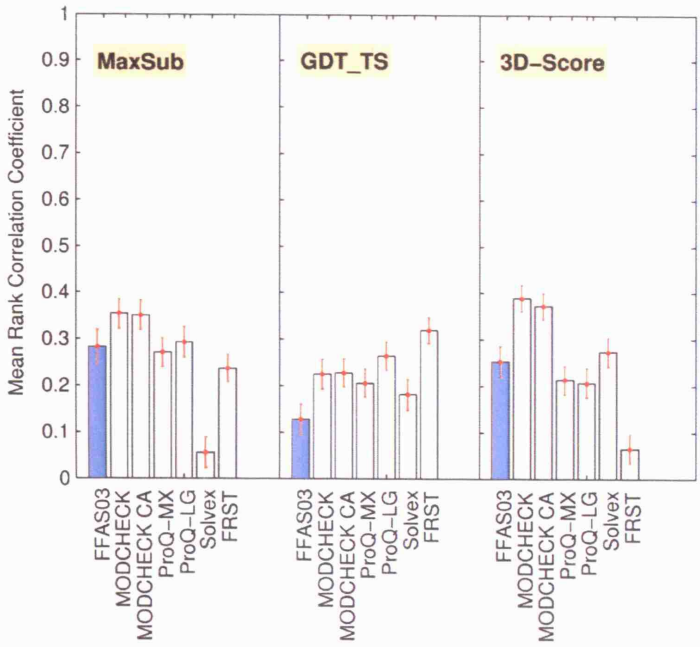
The rank of models produced by the homology modelling servers and the rank obtained with the MQAPs was compared using the Spearman's rank correlation of rank orders. The optimal ordering was defined by the similarity scores for the models (i.e. the optimal rank order is that obtained by scoring the models with an MQAP then re-ordering models in ascending order). Figure 3.1 shows the correlation coefficients describing the rankings produced by applying the MQAPs to the sequence-based methods.

Improvements in rank ordering were obtained with MODCHECK and MODCHECK-CA for all three servers using the three similarity scores. The victor/FRST function performs poorly across all servers but improves the rankings when the GDT_TS score is used. Solvex is also poor at improving the model rankings and does especially badly with the MaxSub score is used to obtain true rank ordering. Almost all methods improve the FFAS and FFAS03 ranking (see Figure 3.1a) though less improvement was seen when MQAPs were applied to ORFeus models (see Figure 3.1c).

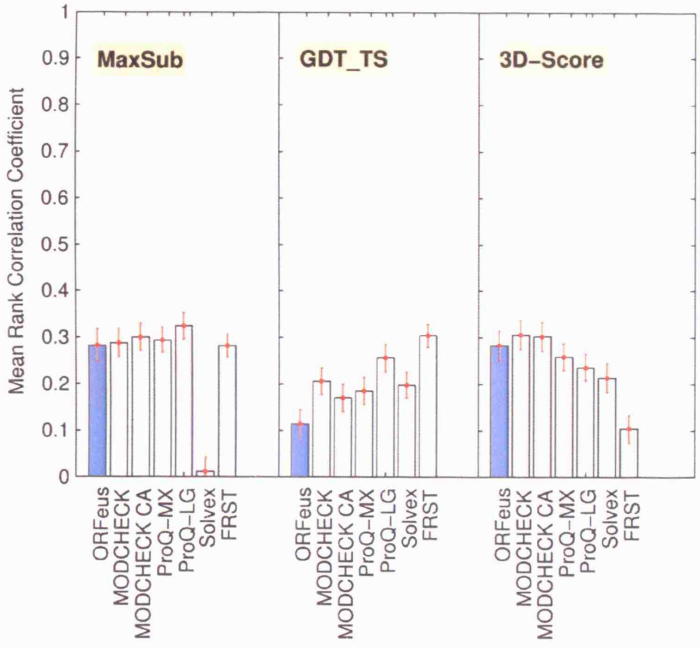


(a) FFAS

Figure 3.1: Continued



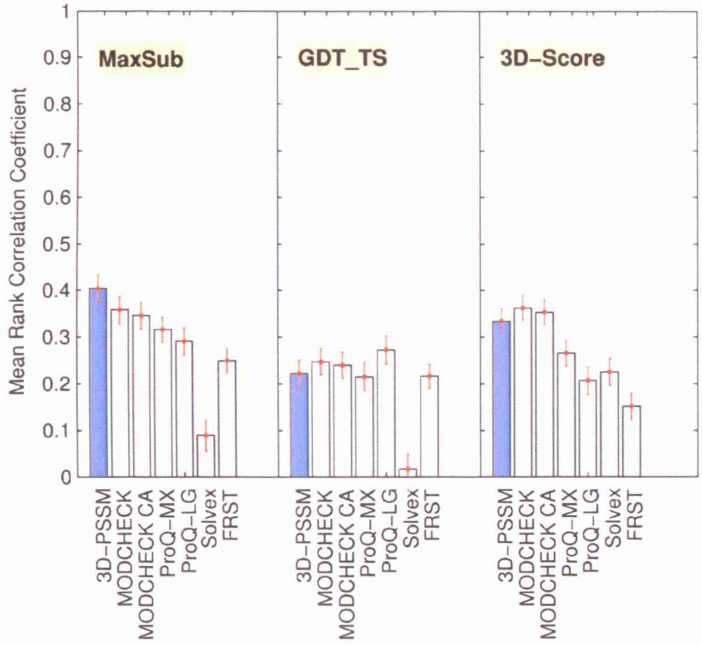
(b) FFAS03



(c) ORFeus

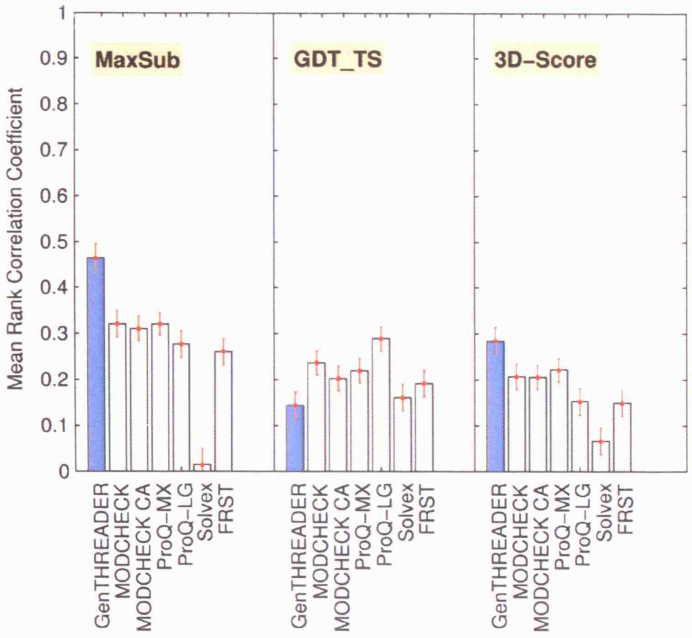
Figure 3.1: The ranking ability of each MQAP method was compared with the ranks provided by the original sequence-based homology modelling servers. The correlation coefficients represent each method's ranking of the models compared to the best possible ordering given by the similarity scores. The MQAPs were applied to the sequence-based methods (a) FFAS, (b) FFAS03, and, (c) ORFeus. Each analysis was performed using the similarity scores MaxSub, GDT_TS, and the 3D-Score. Error bars represent the standard error of the mean.

MQAPs were then applied to the models provided by the threading-based methods. Improvements in the ranking order were obtained when the GDT_TS score was used to provide the optimal ordering for 3D-PSSM (3.2a), GenTHREADER (3.2b), and, SAM-T02 (3.2d) models though the differences between correlation coefficients were marginal. In the majority of cases, the rank ordering of the MQAPs lead to a degradation in model ranking especially when the MaxSub score was used. There are large inconsistencies between the rank orderings obtained with the MQAPs when different similarity scores are used though MaxSub provides the most consistency across servers.

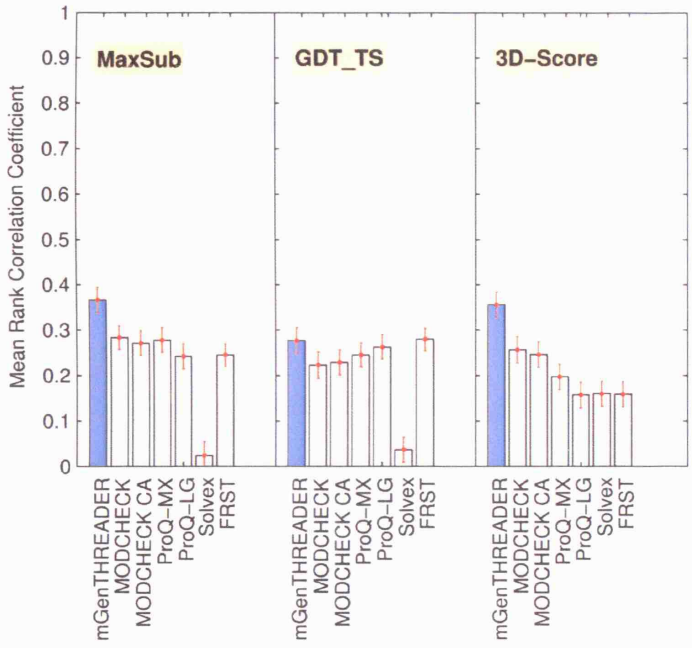


(a) 3D-PSSM

Figure 3.2: Continued

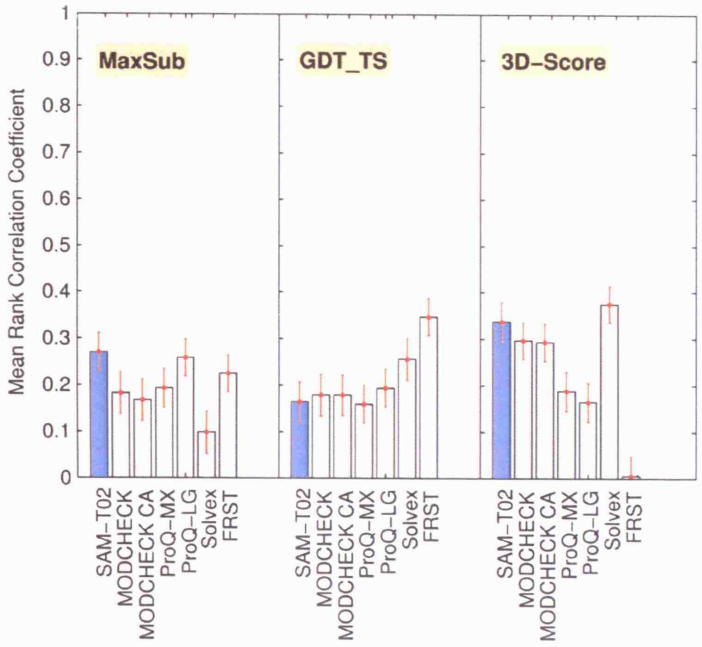


(b) GenTHREADER



(c) mGenTHREADER

Figure 3.2: Continued



(d) SAM-T02

Figure 3.2: The ranking ability of each MQAP method was compared with the ranks provided by the original threading-based homology modelling servers. The correlation coefficients represent each method's ranking of the models compared to the best possible ordering given by the similarity scores. The MQAPs were applied to the threading-based methods (a) 3D-PSSM , (b) GenTHREADER, (c) mGenTHREADER, and, (d) SAM-T02. Each analysis was performed using the similarity scores MaxSub, GDT_TS, and the 3D-Score. Error bars represent the standard error of the mean.

The MQAPs were applied to the consensus method Pcons-4 and improvements in ranking order were found in a number of cases (see Figure 3.3). MODCHECK and MODCHECK-CA improved the rankings regardless of the similarity score used though all servers except Solvex improved the rank ordering slightly when the MaxSub or GDT_TS score was used. Solvex was found to perform badly when the MaxSub score was used to order the models (see Figure 3.1 and Figure 3.2).

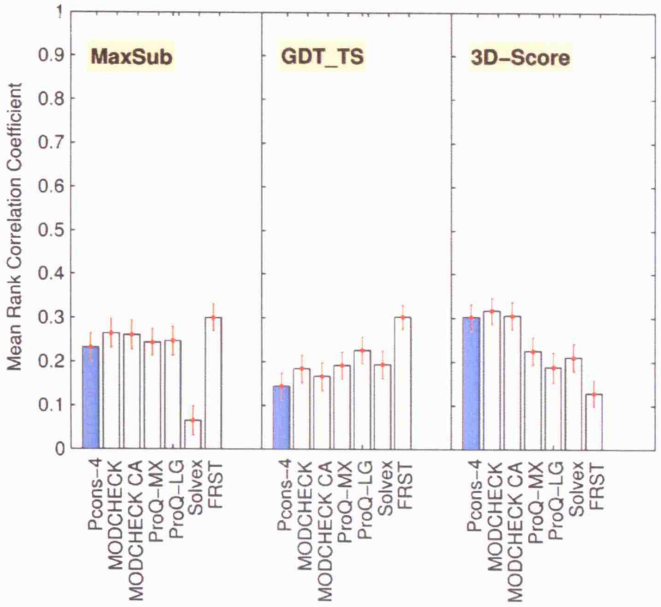


Figure 3.3: The ranking ability of each MQAP method was compared with the ranks provided by the original consensus-based homology modelling servers. The correlation coefficients represent each method's ranking of the models compared to the best possible ordering given by the similarity scores. The MQAPs were applied to the consensus method Pcons-4. The analysis was performed using the similarity scores MaxSub, GDT_TS, and the 3D-Score. Error bars represent the standard error of the mean.

3.3.3 Assessing the general performance using all LiveBench-9 models

The ability of MQAPs to assess model quality based on a "true" quality estimate provided by each similarity score was tested by combining all models from the LiveBench-9 data set (see section 3.2.6). The Spearman's rank correlation coefficient was calculated for each model quality assessment method and described the ability of the method to assign a quality estimate for a model in relation to the "ideal" quality score. Figure 3.4 shows the comparison of MQAP methods using the Spearman's R coefficients.

Overall MODCHECK and ProQ show good ranking abilities while the results obtained with Solvex and victor/FRST were found to be much more variable and dependent on the similarity score used which both MODCHECK variants and ProQ showed more consistency across similarity scores. The results were then split into the "easy" and "hard" categories shown in Figure 3.5. The same pattern of consistency

across similarity scores was found for MODCHECK and ProQ. Larger correlation scores values were found for "easy" targets than "hard" targets regardless of which MQAP method was used (see Figure 3.5a). There was also a noticeable variability between the Solvex and FRST scores between the two categories.

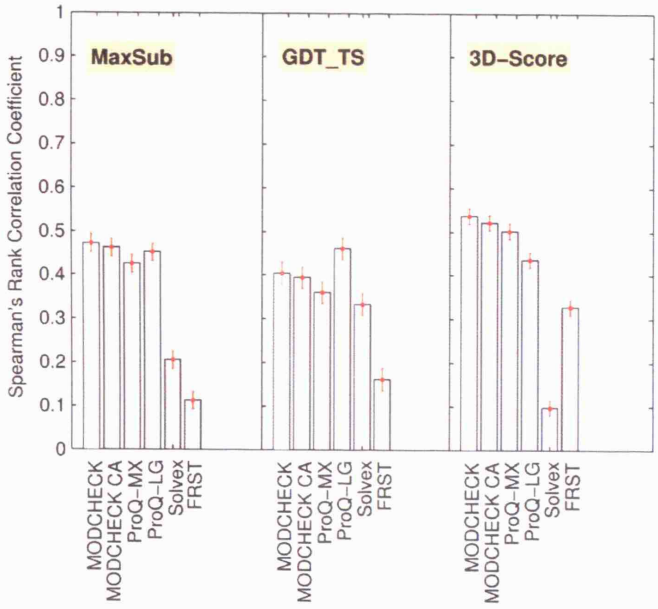
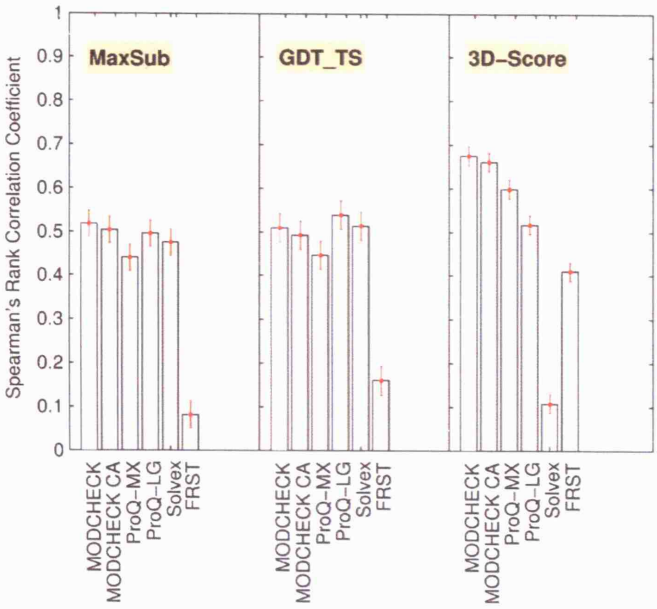
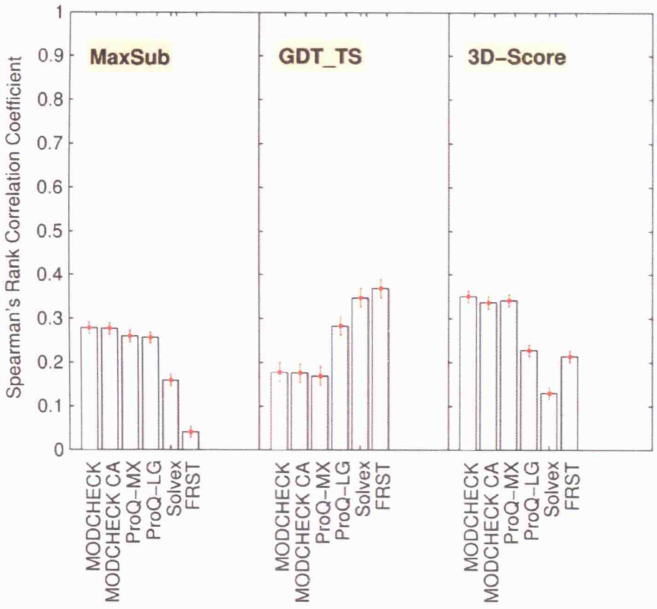


Figure 3.4: The Spearman's rank correlation coefficients of the MQAP methods using each similarity score for the combined set of 11880 models. Error bars (in red) show the standard error of the mean.



(a) Easy targets



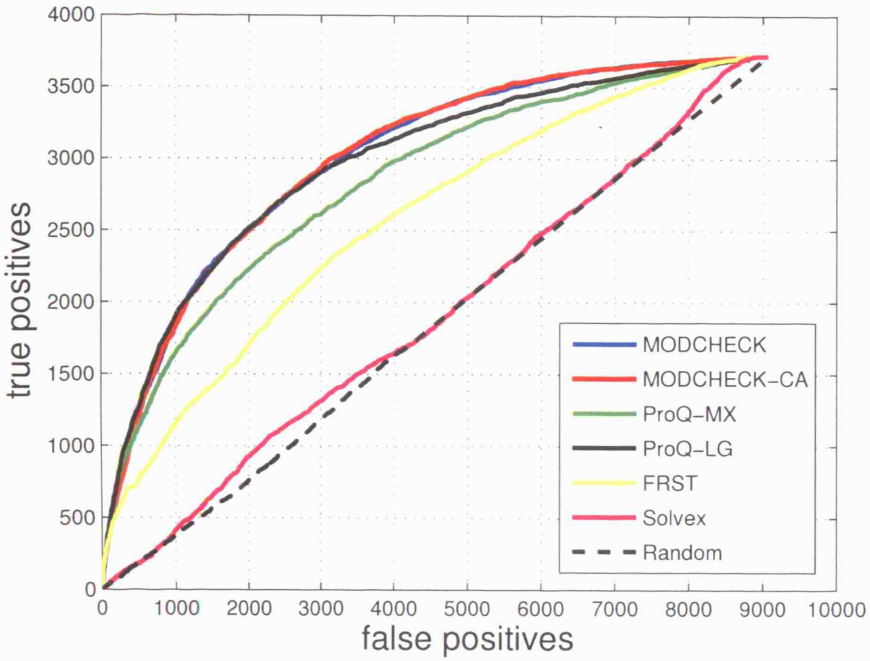
(b) Hard targets

Figure 3.5: The Spearman's rank correlation coefficients are obtained for each MQAP after assessing 11880 models and comparing scores with those provided by the similairty scores. Targets are split by modelling category into (a) "easy" targets, and, (a) "hard" targets. Error bars (in red) show the standard error of the mean.

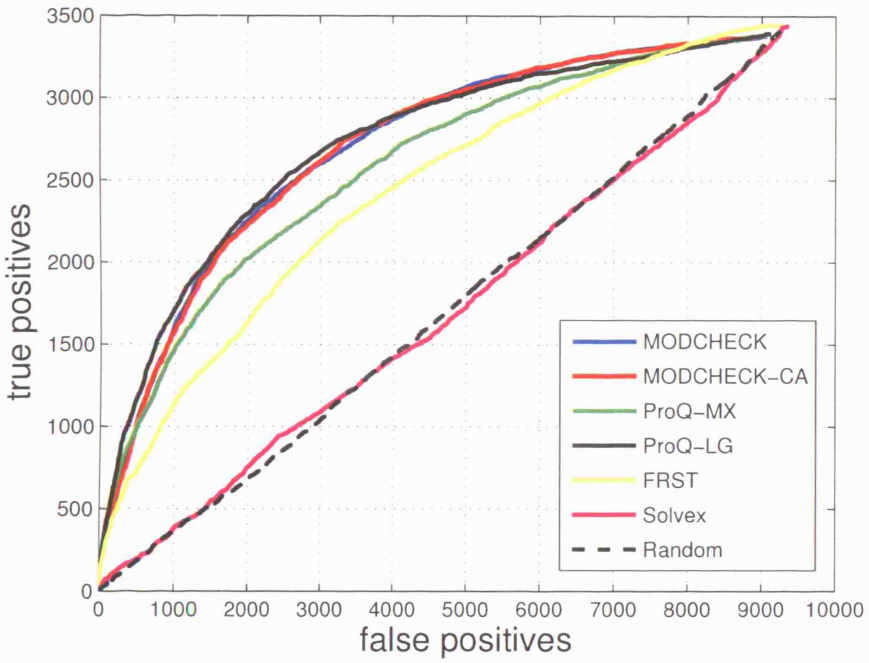
3.3.4 Examining the confidence in MQAP methods

It is important to estimate how confident we can be in the quality scores provided by each MQAP if these tools are to be used in automated structure prediction procedures. Confidence estimates were obtained by plotting ROC curves showing the True Positive Rate (TPR) against False Positive Rate (FPR). The level at which we can be confident a model is "good" can be specified by defining a threshold for the similarity score. Threshold similarity values for defining a True Positive (TP) were; a MaxSub score $\geq$ 0.3 (30%), a GDT_TS score $\geq$ 25%, and a 3D-Score $\geq$ 40.

Figure 3.6 shows the confidence estimates for the MQAP methods using each of the similarity scores, when MQAPs were used to assess the data set of 11880 models. The results of the analysis using the MaxSub score showed that we can be most confident that the MODCHECK and ProQ methods will provide a true prediction more often than other methods tested at this MaxSub threshold. Solvex generates a large number of false positives, and hence we can assume a low confidence in its predictions. Similarly when the GDT_TS score is used, the ProQ LGscore component, MODCHECK, and victor/FRST methods were able to provide high levels of confidence in their scores even though the overall number of true positives was less than when the MaxSub score was used.

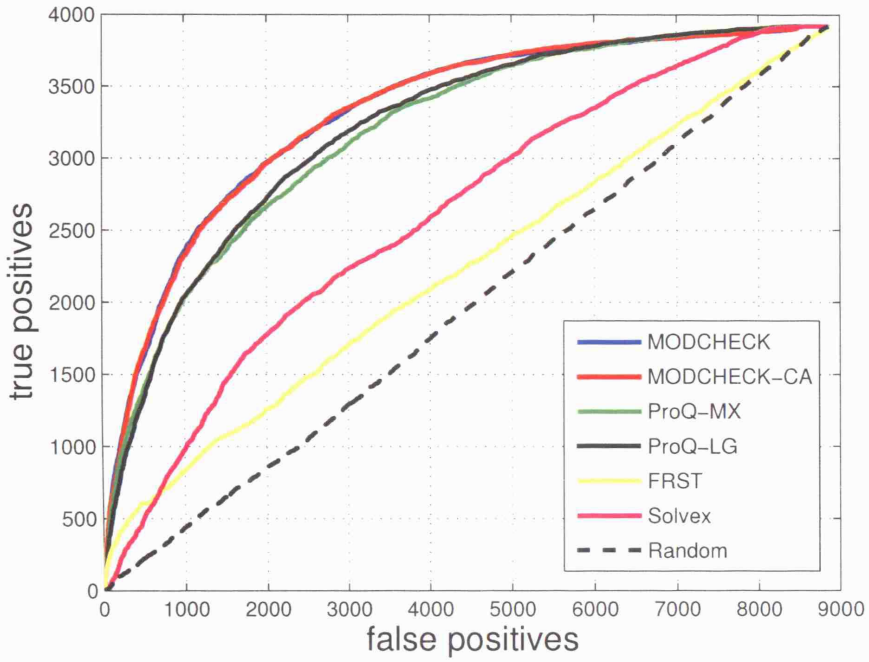


(a) ROC curves obtained with the MaxSub score



(b) ROC curves obtained with the GDT_TS score

Figure 3.6: Continued



(c) ROC curves obtained with the 3D-Score

Figure 3.6: The number of true positives versus false positives for the MQAP methods are shown. The threshold values for defining a true positive for each similarity score are 0.3 (30%), 25% and 40 for (a) MaxSub, (b) GDT_TS, and, (c) 3D-Score, respectively. MODCHECK and ProQ are able to provide more true predictions using all the three similarity measures whereas more variability is observed with victor/FRST and Solvex across the three similarity measures.

3.4 Discussion

The goal of the work presented in this chapter has been to assess the ability of model quality assessment programs to detect superior quality structure predictions from a set of near-native and decoy structures. We look for methods that are able to recognise the correct fold from a set of alternative fold recognition predictions as well as to maximise the model quality of the predictions by discrimination the best models from poorer quality structures. These MQAP methods were selected because they are calculated rapidly and produce a single real-valued quality score for each structural conformation that is easy to interpret. These two features of the scoring functions are necessary for fold recognition in particular, where one of the desired goals is to use these methods for genome-scale automated prediction, but also more generally as desirable properties of a scoring function for all modelling applications.

The first question addressed in this work was whether the top model selected by each MQAP method was better, in terms of model quality, than the model selected by the original server. Or alternatively phrased, if we were to select the model ranked as the best structure by the MQAP would this be a better model than that predicted as best by the original server. The analysis was performed using three similarity scores which are commonly used to rank predictions on the LiveBench experiment, to determine if the choice of similarity score affects the ranking of the models. The p-values from a Wilcoxon Signed-Rank test are used to determine if an MQAP predicts a greater number of higher quality models in the top position than each homology modelling server (see Table 3.1) at a 95% confidence level. The results show that of the 6 MQAP scores the MODCHECK and MODCHECK-CA methods are able to improve the top model quality most often. There is some discrepancy between the results when different similarity scores are used but in general the servers FFAS, FFAS03, ORFeus, and Pcons-4 benefit from MODCHECK scoring. A number of other notable features are discernible; Solvex and victor/FRST show no significant improvement in top model selection ability except for one case where victor/FRST improves the top model selection of Pcons-4 models when the GDT_TS score is used (p-value = 0.01). However, the Wilcoxon test does not provide a measure of the magnitude of improvement and only describes whether there is a significant difference between the population means.

The MQAP methods were then applied to the more general problem of ranking the models in order of quality. Rankings were compared to the original server by means of a correlation coefficient which shows the correlation between the MQAP ordering and an optimal ordering based on the similarity score ranks. The analysis was performed for each server individually. The results show that most MQAP methods improved the ranking of models for the sequence-based methods (Figure 3.1a, Figure 3.1b) with the exception of ORFeus (Figure 3.1c) where Solvex, victor/FRST, and ProQ achieves slightly lower mean correlation coefficient scores for some similarity score rankings (Maxsub and 3D-Score). In contrast, when applied to threading-based and consensus methods, most MQAPs were unable to improve the model rankings over the ranking provided by the server (Figure 3.2). This result was surprising and suggests that these MQAP methods are unable to improve the ranking of models by quality assessment when structural information is already included in the modelling process.

The analysis of these MQAP methods on the individual servers highlighted an interesting relationship between the MQAPs and the model quality scores obtained for the models produced by the different modelling methods. A further analysis was then conducted to determine if additional patterns could be found between the performance of these MQAPs methods and the model data. The models for all the targets were separated into two modelling categories corresponding to a commonly used classification of target difficulty (see Section 1.6.2). A significant difference in performance of the MQAPs between the two modelling classes was found (see Figure 3.5a and Figure 3.5b). MQAP scores for models for targets in the "easy" category were significantly more correlated with similarity scores for those models than in the "hard" category (e.g. a correlation coefficient $R \approx 0.5$ for the easy targets using the MaxSub score for most MQAP methods compared with $R \approx 0.3$ for the harder targets). There is a simple explanation for these results, namely that targets which have close homologues in structure databases are more likely to have better sequence alignments and sequence coverage leading to better, more complete models predicted by the modelling servers. The MQAPs are likely to score these structures more accurately than for harder targets where the likelihood of incomplete and inaccurate models is far greater.

The MODCHECK methods and both ProQ scores showed similar performance across categories independent of the similarity score used whereas the Solvex and

victor/FRST functions show greater variability in their quality assessment abilities (and often worse performance) when different similarity scores are used. However, overall the MODCHECK and ProQ methods attained similar quality assessment performance on this data set (Figure 3.4).

Given that these methods were shown to improve the ranking of sequence-based methods we then wished to derive confidence estimates for each MQAP method to quantify how often a predicted score could be assumed to be accurate. Confidence estimates were determined for the MQAP methods in according with the three similarity scores. MODCHECK and ProQ achieved the greatest accuracy, determined as the ratio of true positive and false positive rates (Figure 3.6). Interestingly, Solvex was close to random (Figure 3.6a and Figure 3.6b).

These results lend themselves to a number of observations; firstly, the ability of these methods to improve the ranking of models for sequence-based methods suggests that, contrary to popular belief, structural information is important and can be used to improve the prediction accuracy of these methods. Secondly, the individual methods fare differently in their ability to accurate predict model quality. MODCHECK, MODCHECK-CA, and ProQ were generally superior to Solvex and victor/FRST and there may be a number of possible reasons. As a solvation potential, Solvex is reliant on the accuracy of both the backbone conformation and the side-chain orientations. However, the models in this study are reconstructed from a $C_\alpha$trace so that any errors introduced by the reconstruction algorithms may lead to a reduction in model quality assessment method accuracy. Compounding this problem is the quality of the fold recognition models. Homology models, especially those which are modelled from sequences with more distant evolutionary relationships, are likely to contain numerous gaps, and these large pockets will have a profound effect on the solvation energy of the molecule. Similarly, the accuracy of the torsion angle potential and the solvation term in the victor/FRST function may be affected by the incompleteness of models, possible to a far greater extent than MODCHECK and ProQ which may be less sensitive to the coverage effects (ProQ uses a neural network which may be less sensitive to the side-chain arrangements and MODCHECK relies solely on the C_β atoms in its calculation). MODCHECK and ProQ also incorporate additional prediction tools in addition to the statistical components such as the sequence shuffling method of MODCHECK,

or the pattern recognition ability of the neural network in ProQ. Both may contribute significantly to the success of these methods over the other test in this study.

Finally, in view of the wider thesis, the results of the study suggest that these MQAP methods are unlikely to be appropriate for use in a refinement procedure due to their performance in this benchmark analysis though they should be considered as useful tools for improving the accuracy of sequence-based fold recognition methods.

Chapter 4

A Multi-Objective Genetic Algorithm for Protein Structure Refinement

4.1 Introduction

Protein structure prediction has advanced substantially in the last three decades, yet the ability to generate accurate high-resolution models for protein sequences with unknown structures is one of the major unsolved challenges in computational biology (Moult 2006). As the divide between the number of sequences produced by genome sequencing projects and the number of structures resolved experimentally continues to grow, computational modelling is becoming an increasingly important strategy for bridging the sequence-structure gap. Algorithms that can reliably produce accurate structural models will not only circumvent the need for expensive experimental resolution methods but will also provide biologists with the necessary data for detailed functional studies, as well as enabling rapid, high-quality annotation of complete genomes (Sanchez & Šali 1998, McGuffin et al. 2006).

At present, comparative modelling generally outperforms *de novo* structure prediction, and where applicable, is the most reliable and accurate method for structural modelling (Baker & Šali 2001). The accuracy of these methods depend not just on the level of sequence identity between a target sequence and a template (Chothia & Lesk 1986), but also on elements of the modelling procedure. Recent estimates suggest that for sequence-to-template matches with high sequence conservation ($> 50\%$ sequence identity) models are generally of high quality and exhibit $\sim 1\mathring{A}$ C_αRMSD from the experimental structure, falling to between $1 - 4\mathring{A}$ C_αRMSD for sequences with 30–

50% sequence identity.

The standard homology modelling procedure consists of four stages; (*i*) identifying suitable homologues that can be used as modelling templates, (*ii*) aligning the query sequence to the template residues, (*iii*) building a model from the sequence-to-templates alignment by copying coordinates from the template or by satisfying spatial restraints, and, (*iv*) evaluating the model (Marti-Renom et al. 2000). State-of-the-art comparative modelling methods are able to detect suitable templates, and, given that a good quality alignment can be made, often produce the correct topological features of the target (Cozzetto & Tramontano 2005). Although the sequence-to-template alignment step is still error prone, especially at lower sequence identities (Venclovas & Margelevicius 2005), the best models produced at high sequence identities are often as good as optimal multi-template models, suggesting that the natural limits for easy comparative modelling targets has been reached (Contreras-Moreira et al. 2005, Ginalski 2006).

4.1.1 Errors in homology models

Most errors in template-based modelling arise predominantly when fitting the target sequence onto a non-native template backbone; a process that can lead to physically unrealistic models. Both incorrect sequence-to-template alignments and amino acid substitutions can produce non-native packing arrangement of side-chains within the core, resulting in models that contain atomic clashes, minor backbone distortions (leading to incorrect side-chain packing), and poor stereochemistry (Morris et al. 1992).

Evolutionary changes such as insertions, deletions, and substitutions, also affect the modelling process by introducing sequence differences between the target sequence and template that must be considered during the alignment and modelling stages. Residue substitutions that are responsible for sequence divergence can alter the conformation of a main chain region even if a correct alignment can be made for the majority of remaining sequence residues. These substitutions can have important conformational effects on a particular main chain segment (e.g. concurring a new function or specificity to the protein) while having no effect on the overall conformation of the fold (Sanchez & Šali 1997). Insertions and deletions are also sources of modelling error because they are, by definition, missing from the parent template. In

most cases, insertions and deletions occur in, but are not limited to, the loop regions of a protein, and modelling these segments requires the use of additional algorithms. Although loop modelling has received considerable attention and is generally quite accurate (Donate et al. 1996, Jacobson et al. 2004), most methods are still limited to modelling loops shorter than ten residues (Fiser et al. 2000), and the modelling of unaligned non-loop regions has only recently begun to be addressed (Rohl, Strauss, Chivian & Baker 2004).

Overall, these modelling errors have a significant effect on the quality of comparative models. Modelling errors not only produce energetically unfavourable conformations but also introduce structural inconsistencies in models such as more rugged and expanded surface areas when compared with the experimental structure. These structural problems also have consequences for the physico-chemical properties of the molecules, producing non-native exposure states for surface residues and altering the electrostatic potential of the molecular surface (Chakravarty et al. 2005). Although medium-resolution models produced by comparative modelling have many applications, further refinement is necessary to drive models towards their energetically most favourable native conformation and to produce models that are accurate to atomic resolution.

4.1.2 High-resolution refinement

Protein model refinement is now one of the primary bottlenecks to high-resolution structure prediction (Moult 2006). Previous refinement attempts have generally failed to improve the accuracy of template-based models, and typically these strategies have lead to a deterioration rather than an improvement in model quality (Koehl & Levitt 1999, Tramontano & Morea 2001). The practical difficulties in refining a 1–3Å comparative model to atomic accuracy have highlighted a number of challenges in the final stages of protein structure prediction, namely; inaccuracies in current force fields and knowledge-based potentials, and difficulties in sampling the vast number of alternatively packed conformations at high-resolution (Bradley et al. 2005, Misura & Baker 2005).

Accurate energy functions that can score the native state as that of lowest free energy in relation to non-native conformations are critical for high resolution

refinement, and all-atom models require a sufficiently detailed energy potential to properly treat all energetic interactions. However, increasing the resolution of an energy function also increases the ruggedness of the potential energy surface, making sampling more difficult. This increase in the complexity of the energy landscape increases the number of potential atomic arrangements in a continuum torsion space thereby limiting the number of conformations that can be sampled on a practical time scale. This requires a highly efficient refinement protocol that can efficiently navigate the corresponding free energy landscape (Lee et al. 2001).

Molecular dynamics has been used to refine protein models though with limited success. Using restricted MD with residue restraints, Flohil et al. (2002) were able to marginally improve one out of three CASP3 targets. Fan & Mark (2004) employed long MD simulations (5-400ns) to refine a number of ROBETTA models. Although the RMSD increased in most cases, long simulations improved the packing of helices and regularizes beta strands. Lu & Skolnick (2003) use statistical potentials to guide a molecular dynamics simulation with local constraints and predicted tertiary contact restraints applied to 67 *ab initio* models. The initial models, ranging from $\sim 2.5 - 10\AA$ C_αRMSD for the unrefined models, were improved by $\sim$0.3Å C_αRMSD (33 cases), though in some cases, improvement as large as 1Å C_αRMSD were achieved.

Recently, some initial progress has been made in refining medium resolution template-based models with a reduced representation and an optimized force field. Using an optimization procedure to sculpt the energy landscape of their potential, the authors were able to produce structures that are often more accurate than the starting template (Zhang & Skolnick 2004*a*). In high-resolution refinement, Qian et al. (2004) improve a number of comparative models by sampling along evolutionary favoured directions obtained by using principal components of the backbone structure variation within a homologous family. By using dimensionality reduction, they partially overcame some of the sampling difficulties arising from a complex potential energy landscape.

Perhaps the most promising result so far is that obtained by Misura et al. (2006). Using their ROSETTA protocol with folding constraints, the authors were able to refine 22 out of 39 template-based models, with the improved models ranked as one of the ten lowest-energy structures. This encouraging result shows the accuracy of their

energy potential and sampling protocol, however, this improvement comes with a large computational cost ($\approx$ 2000 hours per target on a single processor); further highlighting the difficulty of achieving high-resolution models at atomic accuracy.

4.1.3 Chapter overview

This chapter presents two methods for high-resolution template-based model refinement using evolutionary algorithms, where conformational space is explored through a diverse range of conformational sampling operators and the search guided by statistical potentials. Two forms of evolutionary algorithm, the single-objective and multi-objective genetic algorithm, are tested on models submitted for 35 comparative modelling targets in the CASP6 experiment. The ability of these two algorithmic variants; (i) to explore conformational space adequately ensuring the sampling of near-native models, and, (ii) to select models that are both near-native and have lower energy than the best template, is explored. The differences between the two methodologies are compared and the successes and failures of the approach discussed.

4.2 Refinement with evolutionary algorithms

Evolutionary algorithms are used in this study to refine template-based protein models. Both a single-objective GA and a multi-objective GA are applied to a set of full-length all-atom models so that the performance of both algorithms can be evaluated on the refinement problem. Where possible, similar implementation details are used in both algorithms to ensure consistency between the two methods.

4.2.1 Single-objective GAs

The simple genetic algorithm (Goldberg 1989) is used to refine models with a single objective (see Section 4.2.4.1). The simple GA is implemented with a generational replacement strategy (i.e. no members are cloned), binary tournament selection ($q = 2$), a population size $N = 200$, crossover probability $P_c = 0.6$, and mutation probability $P_m = 0.3$. Crossover and mutation control parameter values are discussed in more detail in Section 4.2.6. The GA terminates either after a maximum number of generations have elapsed ($T = 1000$), or when energy convergence is reached. A population of solutions is considered to have converged if the average energy of a population $< E_P >$, is stable for 5 generations (± 3 $kcal/mol$). The single-objective GA is used as a control experiment for comparing with the performance of the multi-objective GAs so that a more complete understanding of the differences between the two approaches can be gained, and the relative advantages and disadvantages of each strategy determined.

4.2.2 Multi-objective GAs

The Non-dominated Sorting Genetic Algorithm (NSGA-II) is used for multi-objective refinement due to its previous performance on a comprehensive benchmark of test problems, and for its ability to handle constraints (Deb 2002). This constraint handling feature is the primary reason for choosing the NSGA-II algorithm over the SPEA2 algorithm used previously in Chapter 2. Both algorithms show similar performance on a large set of test problems though NSGA-II was able to achieve a broader spread of solutions suggesting better overall performance (Zitzler, Laumanns & Thiele 2002). The NSGA-II algorithm is an elitist multi-objective evolutionary algorithm that uses fast non-dominated sorting and a diversity preservation algorithm (via a parameterless niching operator) for obtaining a diverse spread of solutions in each generation. The NSGA-II algorithm uses a binary tournament selection scheme, with recombination

and mutation operators (defined in Section 4.3), and elitism is preserved through a non-dominated sorting step. Control parameters for multi-objective optimization are obtained by trial and error to find a satisfying parameter set that performs well under constraints imposed by the computational run-time.

4.2.3 Representation

By definition, high-resolution structure refinement requires a spatial representation of the polypeptide chain that is able to capture the atomic details of a protein structure. In this work, the polypeptide chain is represented using internal coordinates specified by the rotational degrees of freedom of the backbone (the ϕ/ψ torsion angles) with the ω dihedral angle fixed at $\pm 180°$ for residues in the *trans* conformations and $0°$ for *cis* conformations. In terms of implementation in the GA, the chromosome of each individual is therefore represented as an array of ϕ and ψ angles. Throughout the refinement simulation all bond lengths and angles are fixed at equilibrium values Engh & Huber (1991) and dihedral angles are free to adopt values in the continuous range $[-180° \; 180°]$.

In Cartesian space, proteins are modelled using an all-atom representation (heavy atoms plus amide hydrogen). In order to calculate the energy of each conformation, an individual in the current population is first converted to Cartesian coordinates by reconstructing the backbone from the ϕ and ψ angles. The $N - C_\alpha - C'$ main-chain atoms are generated with a highly efficient reconstruction algorithm (Parsons et al. 2005) followed by the placement of the main chain oxygen and amide hydrogen atoms (Pauling et al. 1951). Side-chains (including C_β atoms) are constructed on the backbone using the fast graph-theoretic side-chain modelling algorithm SCWRL 3.0 with an unmodified backbone-dependent rotamer library (Canutescu et al. 2003). Modifications to a protein structure are performed in both torsion and Cartesian representations.

4.2.4 Fitness functions

Fitness functions are defined for both single- and multi-objective refinement. A high-resolution energy potential is used for scoring individuals in the single-objective GA and supplemented with additional objective functions in the multi-objective GA case for producing a Pareto-set of non-dominated solutions.

4.2.4.1 Energy potentials

For high-resolution refinement, conformations are scored with a knowledge-based statistical potential composed of a distance-dependent statistical potential of mean-force, an orientation-dependent hydrogen bonding potential, and a square-well potential for capturing repulsive interactions and improper steric effects. The form of the energy function is

$$E_{total} = E_{dfire} + wE_{hb} + E_{steric} \tag{4.1}$$

where E_{total} is the total combined energy, E_{dfire} is the energy of pairwise interactions, E_{hb} is the hydrogen bond energy, and E_{steric} is an energy penalty for capturing the effects of steric overlap, w is a weighting factor for the hydrogen bond energy.

Previous results from a model quality assessment benchmark found that none of the energy functions tested were suitable for high-resolution refinement, conferring only marginal improvements to the discrimination of comparative models produced by automated methods (Pettitt et al. 2005) (see Chapter 3). Instead, the all-atom Distance-scaled, Finite Ideal-gas Reference state potential of mean-force (Zhou & Zhou 2002) is used to capture pairwise interactions between atoms as well as implicit atom-solvent interactions. The DFIRE potential is derived using a reference state of ideal gases confined in protein size spheres with a single parameter to capture the finite-size effect of proteins. The DFIRE potential also has a number of desirable properties; it is transferable across different interaction systems, it captures the solvent-induced hydrophobic effect, it is largely database independent, and performs well on decoy discrimination tests (Zhou et al. 2006).

Although the DFIRE energy function is a knowledge-based potential, and thus discrete, its use in a refinement protocol can be justified if the sampling algorithm and search procedure are decoupled. In most gradient-based refinement algorithms the energy function must be differentiable (and hence continuous) so that it can guide a minimization procedure that moves the structure towards a global, or local, energy minimum. In the case of stochastic search algorithms such as GAs, the conformational sampling is independent of the search, so that each conformation in a population represents a discrete "snapshots" on a (generally smoother) energy landscape.

The DFIRE potential, like most statistical mean-force potentials, lacks a hard-

repulsive core. As native structures generally lack energetically unfavourable steric clashes, the repulsive component of distance-dependent pairwise interaction energies (modelled in physical potentials by the 6-12 Lennard-Jones potential) is often under-sampled by statistical procedures unless introduced as an artifact of X-ray crystallographic or NMR procedures. Therefore, to model repulsive interactions, a simple steric term in the form of a hard-core potential, is used to capture improper sterics between pairs of atoms ($C_\alpha - C_\alpha$, $C_\beta - C_\beta$, and $C_\alpha - C_\beta$)

$$E^{steric}_{C_i^x - C_j^x} = \begin{cases} 0 & \text{if } d_{ij} > d_{min}, \\ (d_{min} - d_{ij})^2 & \text{if } d_{ij} \leq d_{min} \end{cases} \tag{4.2}$$

where d_{ij} is the distance between atoms i and j, and $d_{min} = 3.05$ for $C_\alpha - C_\alpha$ and $C_\beta - C_\beta$, $d_{min} = 3.6$ for $C_\alpha - C_\beta$ interactions. Steric energies are calculated for all ij residue pairs except immediate neighbouring residues. Steric clashes between side-chains and between the side-chains and backbone atoms are handled by the side-chain modelling algorithm (see Section 4.2.3).

DFIRE also lacks an explicit hydrogen-bonding term. In this work, backbone-backbone (bb-bb) hydrogen bonds are identified using the method of McDonald & Thornton (1994) and energies are calculated using the orientation-dependent hydrogen bonding potential (Kortemme et al. 2003).

The hydrogen bond weight is determined by optimizing the weight value in order to maximize the average Z-score, $< Z - score >$, of the proteins in the decoy set, where the Z-score is calculated as

$$Z - score = \frac{E_{native} - < E_{decoys} >}{\sigma_{decoys}}$$

where E_{native} is the energy of the native structure, $< E_{decoys} >$ is the average energy of the ensemble, and σ_{decoys} is the standard deviation of the ensemble. Figure 4.1 shows the average Z-score for different hydrogen bond weight values using the Tsai decoy set of 25 all-atom high-resolution decoys (Tsai et al. 2003). The hydrogen bond weight was chosen to be $w = 0.114$ based on this analysis.

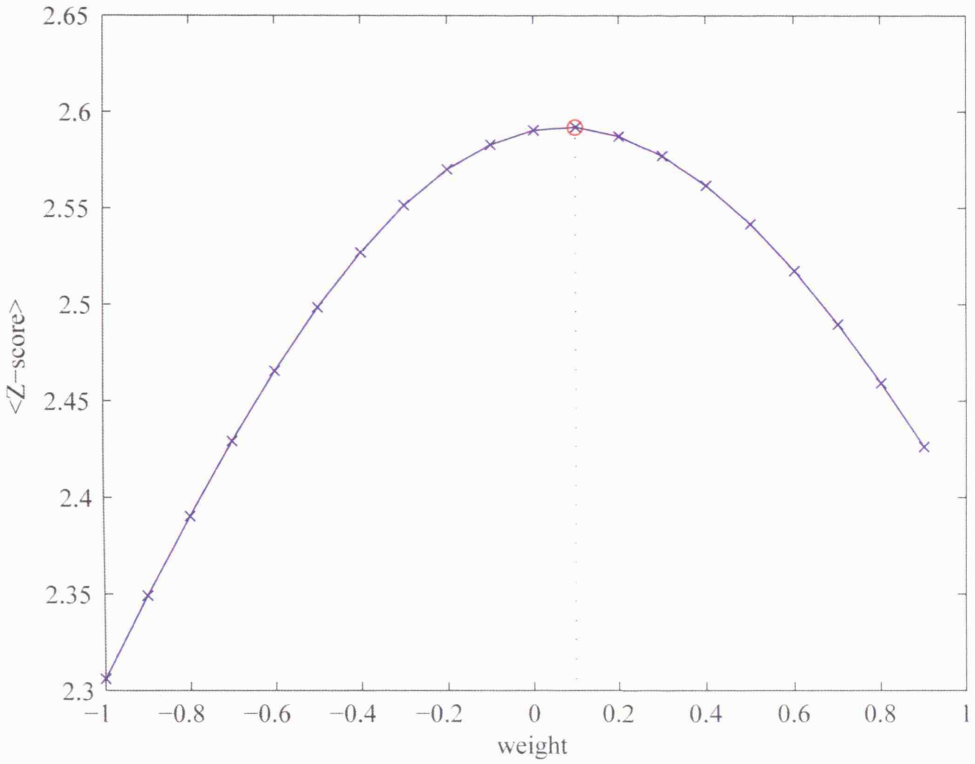


Figure 4.1: The average Z-score is shown at different weight values for the hydrogen bond energy term. The optimal hydrogen bond energy weight is selected in order to maximize the average Z-score over the 25 proteins in the decoys. An optimal weight value was found to be 0.114 (highlighted with a red circle).

4.2.4.2 Divergence function

The parallel nature of genetic algorithms enables them to search multiple regions of the search space simultaneous with the aim of preventing early convergence on local minima thus increasing the probability of finding the global minimum. However, as the complexity of the search space increases the likelihood of sampling local minima increases, and in high-resolution refinement, the energy landscape described by high-resolution models is extremely rugged.

Therefore in multi-objective optimization, a divergence function is used as a supplementary objective to the energy function. The function is devised to prevent early convergence on local minima in the energy landscape introduced either by the unrefined models or by sampling within a restricted conformational neighbourhood. The divergence function is used to drive the search away from the low energy regions occupied by the unrefined models towards more distant parts of the search space.

The conflict that arises between the divergence function (which tries to push the conformational search away from the low energy near-native starting structures), and the energy function (which favours native-like low energy structures) is used to ensure sufficient coverage of conformational space.

The divergence function $f(v,w)$, is defined as

$$f(v,w) = \frac{1}{(1+g(v,w))} - 1.0 \tag{4.3}$$

where $g(v,w)$ is the backbone RMSD between conformation v and the lowest energy unrefined model w, calculated as

$$g(v,w) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(v_{ix}-w_{ix})^2 + (v_{iy}-w_{iy})^2 + (v_{iz}-w_{iz})^2} \tag{4.4}$$

4.2.5 Constrained optimization

The NSGA-II algorithm is able to handle constraints on multi-objective solutions by incorporating constraint scores into the ranking mechanism. Constraints judge each solution to be either feasible or infeasible depending on whether constraints are satisfied, leading to three possible situations in the binary tournament selection scheme used by NSGA-II; (*i*) both solutions are infeasible, (*ii*) only one is feasible, and, (*iii*) both are feasible. The selection of solutions is then made by a simple rule-base method.

1. Choose the solution with the smallest overall constraint violation.
2. Choose the feasible solution.
3. Choose the solution that is not constraint-dominated.

Constraint-dominated solutions are evaluated by adjusting the definition of domination between two solutions i and j, so that

Definition 4.1 *A solution i is said to constrained-dominate a solution j, if any of the following conditions are true.*

1. *Solution i is feasible and solution j is not.*
2. *Solutions i and j are both infeasible, but solution i has a smaller overall constraint violation.*
3. *Solutions i and j are feasible and solution i dominates solution j.*

This definition ensures that any feasible solution has a better non-domination rank than any infeasible solution.

Constraints are imposed on solutions in the multi-objective refinement algorithm by including lower and upper bounds on C_αRMSD deviations from the template. For an individual i, a constraint penalty C_i, is incurred if either a lower bound constraint l, or upper bound constraint u, is violated.

$$C_i = \begin{cases} d_{ij} - l & \text{if } d_{ij} \leq l \\ 0 & \text{if } l < d_{ij} < u \\ u - d_{ij} & \text{if } d_{ij} \geq u \end{cases} \tag{4.5}$$

where d_{ij} is the backbone RMSD between the lowest energy template j, and the current conformation i.

4.2.6 Control parameters

The large computational cost of the GA refinement procedure makes exhaustive control parameter testing infeasible. Therefore, the near-optimal parameter combination from the multi-objective study in Chapter 2 is used in this work. Control parameters are binary tournament selection ($q = 2$), crossover probability $P_c = 0.6$ (reduced by 0.1 from the previous work), mutation probability $P_m = 0.3$, maximum number of generations $T = 1000$, and, population size $P = 200$.

Constraint bounds l and u are assigned as follows; for the refinement of *CM/easy* targets, where a homolog can be found in the PDB using BLAST, the lower bound value $l = 0.2\text{Å}$, while the upper bound constraint $u = 4.0\text{Å}$. For *CM/hard* targets, the lower bound is increased to $l = 0.5\text{Å}$ with an upper bound constraint $u = 6\text{Å}$.

4.3 Conformational sampling operators

For protein model refinement the standard GA operators (crossover and mutation) are reformulated and extended to provide a set of protein specific operators capable of manipulating structural conformations. The crossover operator enables fragment exchange between two conformations, while mutation operators act directly on a single conformation by altering the principle degrees of freedom of the polypeptide chain either by specific (context-dependent) moves, or by making general (context-free) adjustments.

As many of the structural differences between template-based models and native structures occur within specific segments of the protein, refinement requires algorithms capable of altering the local conformation of the main chain between two fixed points while leaving the remaining topology unaltered. A method from inverse kinematics, the Cyclic Coordinate Descent (CCD) algorithm, first presented in protein structure prediction as solution to the loop closure problem (Canutescu & Dunbrack 2003), is used extensively in many of the sampling operators in this work. In fact, due to the simplicity, flexibility, and adequate performance of CCD, a number of recent refinement methods have made use of the algorithm (Offman et al. 2006, Zhu et al. 2006).

As the CCD algorithm is used in many of the context dependent operators, including crossover, a description of the algorithm (and any modifications to it) are first provided, followed by the details of each individual operator.

4.3.1 Restraint-based CCD

The Cyclic Coordinate Descent (CCD) algorithm is an iterative relaxation algorithm for solving inverse kinematics problems (Canutescu & Dunbrack 2003). The algorithm is particularly suited to loop modelling, where a local loop conformation must be adjusted within the constraints of two fixed end-point positions keeping the conformation of the main body of the protein fixed. For a fragment of n residues along the backbone of a model between a fixed N-terminal anchor ($N - C_\alpha - C'$) at position 0 and a fixed C-terminal anchor at position n, the conformation is first perturbed by altering angles between the fixed anchors so that the chain connectivity at the C-terminal anchor is broken. The procedure then involves altering the ϕ and ψ angles iteratively until the

backbone atoms of the n^{th} residue are superimposed on, or within some predefined distance of, the fixed backbone of the C-anchor residue. As a result of the cyclical nature of the algorithm, where each angle is adjusted sequentially from N- to C-terminal, restraints may be placed on angles in turn to restrict the range of values the angles may take.

The modified CCD algorithm for refinement therefore requires; (*i*) a strategy to perturb the original conformation of a modifiable segment, (*ii*) a threshold distance *S*, at which the C-terminal of the adjustable segment is considered within satisfactory range of the fixed C-terminal anchor, and, (*iii*) a method for generating restraints on angles within the modifiable segment.

In the original CCD paper, Canutescu & Dunbrack (2003) used Ramachandran probability maps as orientation constraints where pairs of ϕ/ψ dihedral angles are weighted according to a Gaussian function derived from frequency counts on a 10° by 10° ϕ/ψ grid. For the refinement of template-based models, only medium to small changes in the conformation of a local segment are desirable in the compact state.

The refinement restraints for the modified CCD algorithm used in this study are based on a successful knowledge-based, spatially restrained sampling approach (dePristo, de Bakker, Johnson & Blundell 2003, de Bakker et al. 2006). Following the approach adopted by DePristo, de Bakker, Lovell & Blundell (2003), a discrete state set of ϕ/ψ angles in constructed for each of the 20 amino acids. In their original work, DePristo, de Bakker, Lovell & Blundell (2003) generate large fine-grained state sets with as many as 72^2 states for each residue. The authors note that this choice was made due to previous observations that a strong correlation exists between the size of the state set and its ability to accurately represent native protein structures (Park & Levitt 1995). The rational behind the successful performance of restraint-based conformational sampling is that these methods sample propensity weighted conformations that correspond to favourable low-energy states in the energy landscape.

Here a smaller discrete state set is constructed using the observed ϕ/ψ distributions in the Top500 database[1] of non-redundant protein structures. This set of high-resolution ($\leq 1.8\text{Å}$) structures makes up a hand-curated database of models solved by X-ray crystallography containing few van der Waals clashes, and with a sequence identity

[1] http://kinemage.biochem.duke.edu/databases/top500.php

<60% between any two structures.

To generate the residue-specific discrete state sets, the observed distributions of ϕ/ψ angles are clustered using K-means clustering. An initial number of cluster centres $k = 500$ is used, though during the clustering procedure any singleton and empty clusters are discarded. After clustering, each residue map holds between 473 and 491 ϕ/ψ state representatives. Figure 4.2 and Figure 4.3 show the clustered discrete states of the combined amino acids (excluding proline and glycine), and the discrete states of proline and glycine, respectively.

A 15° by 15° grid is overlaid on the discrete state maps so that each grid square is identifiable by labelled ϕ and ψ coordinates using the lower and upper coordinate form Xx. Restraints are then placed on a ϕ or ψ angle by identifying the Ramachandran grid coordinate Xx for the conformation of the current residue and replacing the ϕ or ψ angle with the angle value taken from a randomly selected cluster centre in the same grid position Xx. If the grid square occupied by a ϕ/ψ pair contains no cluster representatives then the nearest populated grid square is determined by searching progressively further outwards from the Xx position until a state is found.

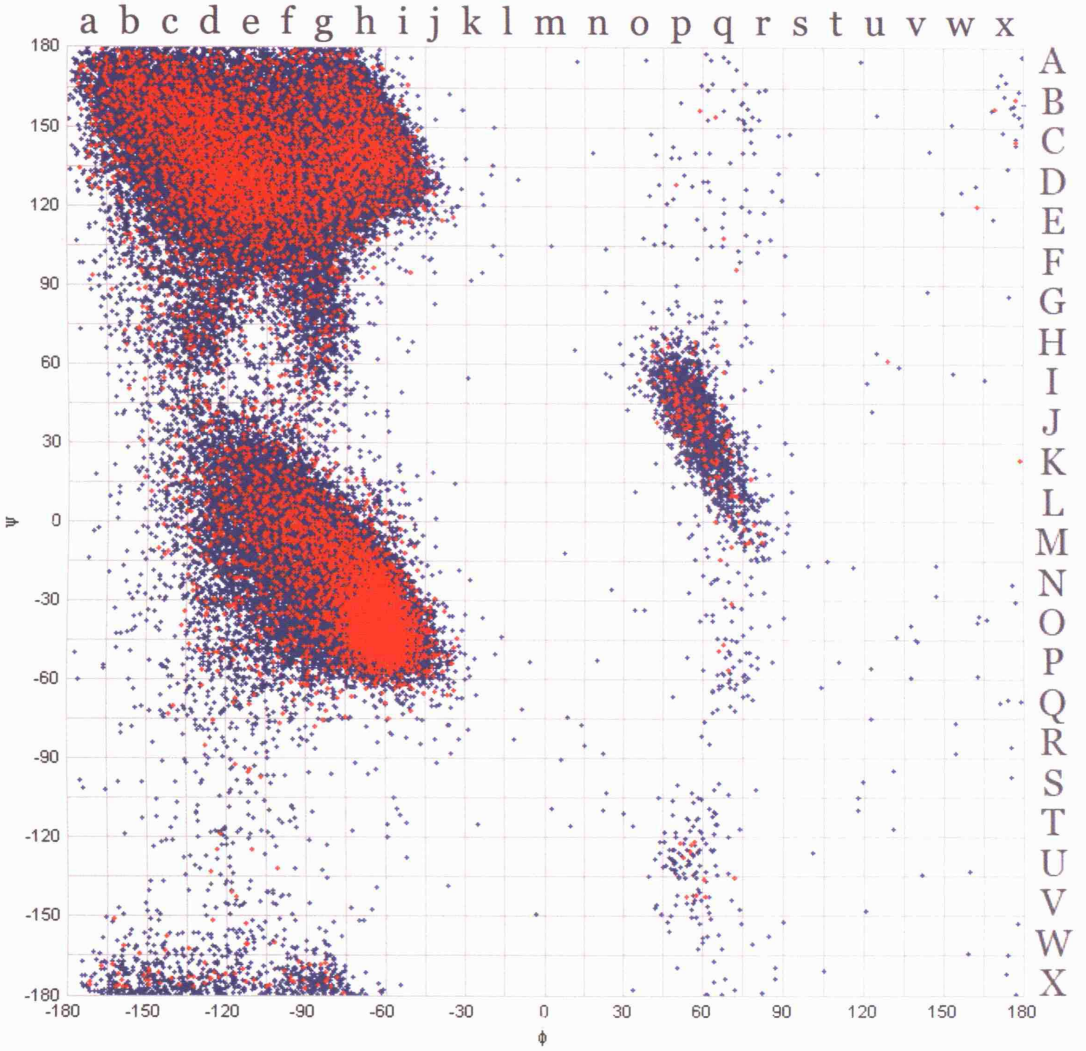


Figure 4.2: The ϕ/ψ distributions and cluster centres for all residues, excluding glycine and proline, in a data set of 500 high-resolution structures are shown. Blue markers indicate individual dihedral states while red markers represents the reduced set of cluster centroids. The Ramachandran plot is partitioned into meso-states at 15° intervals with states labelled in lowercase characters on the ϕ axis and uppercase characters on the ψ axis.

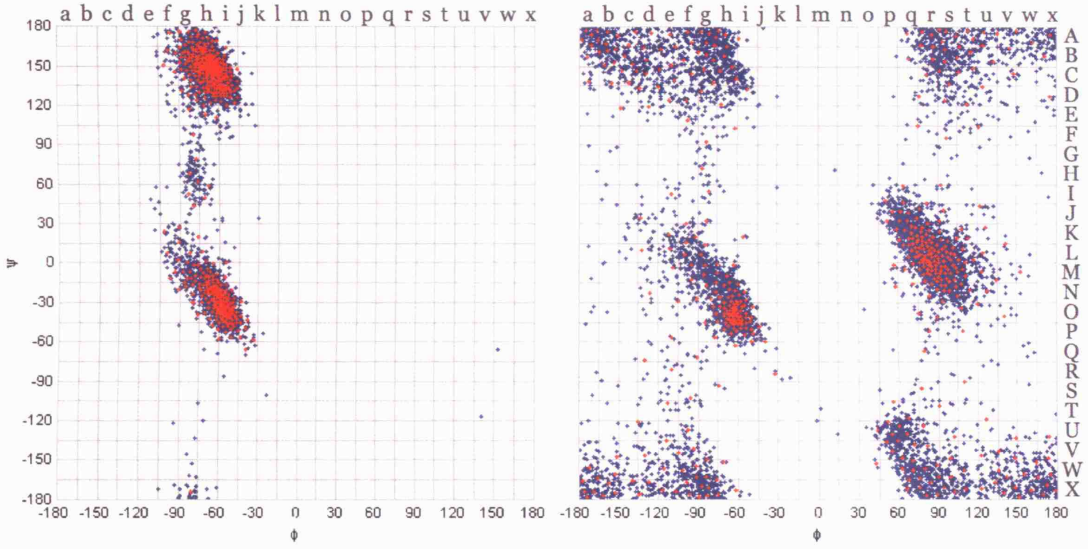


Figure 4.3: The distribution of ϕ/ψ angles and the discrete (clustered) states are shown for both proline (left) and glycine (right). The typical Ramachandran features for these two residues are replicated accurately, with a dense clustering of discrete states in the most favourable regions of each residue plot. Glycine is atypical in that it lacks both a side-chain and β-carbon atom allowing it to adopt a much larger range of conformations than the other amino acids. Proline's range is more restrictive due to the steric restrictions caused by the pyrrolidine ring, where the flexibility in the pyrrolidine ring couples to the backbone (Ho et al. 2005).

4.3.2 Crossover/recombination operators

The recombination operator is a fragment exchange operator based on the principles of binary two-point crossover. In order to exchange a fragment between two conformations $x_i$and x_j, an initial crossover point i, is chosen at random along the sequence. The fragment end-point j, and hence the fragment length, is obtained by extending the fragment from the starting position i by between 2 to 15 residues.

Once the crossover points have been obtained, fragment exchange is then treated as a restrained loop closure problem. The $N-C_\alpha-C'$ atoms at the fragment N-terminal from conformation x_i is superposed on $N-C_\alpha-C'$ atoms at the fixed end of fragment in conformation x_j. The CCD algorithm is used to alter fragment torsion angles so that the atoms at the moving end of the fragment's C-terminal are superposed with the atoms at the fixed end of the x_j.

In order to allow enough flexibility in the fragment to close the loop while also preserving its global conformation, restraints are placed on the torsion angles of the fragment during the CCD. For each consecutive torsion angle, the residue type and Ramachandran grid square Xx are obtained (see Figure 4.2), and restraints placed on the CCD algorithm to ensure any change made to the angle preserves grid coordinate occupancy after adjustment. This restraint ensures a maximum $\Delta\theta \pm 7.5°$ for a torsion angle θ. Five attempts at closure are allowed and the conformation is accepted if the threshold distance S, between the moving and fixed ends is $S \leq 0.001\text{Å}$. The same process is then repeated to incorporate the fragment from x_j into conformation x_i.

4.3.3 Mutation operators

A set of six mutation operators are defined for inducing conformational changes in a compact homology model. Six mutation operators are described for producing either local, or global, conformational adjustments. Local mutations that act within a region defined by two fixed end-points are over-represented in the operator set as small structural changes are more likely to be accepted than moves which affect a global conformational change in the topology.

4.3.3.1 Wiggle

The wiggle operator induces a small local perturbation to the backbone by altering the dihedral angles in a fixed-width region of the protein. A window covering six torsion

angles is positioned at random along the chromosome and a binary mask generated to define which angles will be perturbed. Depending on the permutation generated, between zero and six angles are altered depending on the precise sequence of bits. The magnitude of the perturbation $|\Delta\theta|$, applied to a torsion angle at each alterable position in the mask is obtained by randomly sampling from a Gaussian distribution

$$\Delta(\theta) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(\theta-\mu)^2}{2\,\sigma^2}\right) \tag{4.6}$$

with $\mu = 0$, and

$$\sigma = \theta_{min} + \left[\theta_{min} * \left(1 - \frac{t}{T}\right)^2\right] \tag{4.7}$$

where $\theta_{min} = 5°$, t is the current generation, and, T is the maximum number of generations. The magnitude of standard deviation σ, of the Gaussian distribution is reduced linearly with simulation time so that larger perturbations are more likely to occur in the initial generations followed by progressively smaller angular adjustments as the simulation approaches the maximum number of generations, T. The dynamic adjustment of the standard deviation is designed to reduce unfavourable steric clashes arising from violations of excluded volume constraints in the compact state (Shimada et al. 2001).

4.3.3.2 Single-point mutation

A single-point mutation effects a small, local, and context-independent change at a single residue position. To produce a single-point mutation , a residue position i, along the sequence is selected at random and either the ϕ, ψ, or, ϕ and ψ angles adjusted by a magnitude $|\Delta\theta|$, where $|\Delta\theta|$ is sampled from a Gaussian distribution with $\mu = 0$, and $\sigma = 1°$.

4.3.3.3 Loop mutation

A loop mutation is a context-dependent, restraint-based operator which acts to alter the conformation of either terminal loops or coil regions joining secondary structural elements. To perform a loop mutation, a residue position i, is selected at random along the sequence. The PSIPRED (Jones 1999*b*) secondary structure assignment is used to determine if the residue resides in a predicted coil region if the confidence score for the residue position is greater than four. Otherwise, the secondary structure of the

Table 4.1: Secondary structure classification for ϕ/ψ angle values

SS	ϕ	ψ
helix	$-180° < \phi < -20°$	$-90° < \psi < -10°$
strand	$-180° < \phi < -20°$	$180° > \psi > 20°$
	$-90° < \phi < -10°$	$-180° < \psi < -170°$
coil	otherwise	

residue was classified from the ϕ/ψ angles using Table 4.1 (Kortemme et al. 2003). If position i is predicted to be a coil then the loop is grown by checking the assignments of the immediate neighbours, and if the neighbours are also predicted coils then each residue is added. If either $i+1$ or $i-1$ is not a coil then the assignment continues in only one direction until either the loop reaches a maximum length of 20 residues or the neighbouring residues are assigned to a secondary structure class.

If the loop region lies at the N- or C-terminal of the protein then conformational change is induced by randomly adjusting one third of the loop torsion angles by a magnitude $|\Delta\theta|$ sampled from uniform distribution with range $[-25° 25°]$. If the loop joins elements of secondary structure then the restrained CCD is used. Again, one third of the loop angles are perturbed at random by $|\Delta\theta|$ to induce a break in the chain at the loop C-terminal, and five attempts are then made using the restraint-based CCD to close the loop. A loop is accepted when the threshold distance $S = 0.001\text{Å}$ RMSD.

4.3.3.4 Helix mutation

A helix mutation is used to adjust the orientation of a helix without altering the structure of the helix itself, in effect treating the helix as a rigid body connected to the protein by flexible loops. A helix mutation is performed by randomly selecting a residue within a helical region of the protein based on the secondary structure assignments from PSIPRED (Jones 1999*b*) if the confidence score for the residue position is greater than four. Otherwise, the secondary structure of the residue was classified from the ϕ/ψ angles using Table 4.1, similar to the loop mutation. The start and end residues of the helix are then determined followed by the start and end points of the N-terminal and C-terminal loops connected to the helix. The conformation of these loop regions are adjusted by altering the torsion angles by $|\Delta\theta|$ to break the C-terminal loop. The

CCD algorithm is then used to close the loop leading to the adjustment of the helix orientation.

4.3.3.5 α-helix re-sampling

The α-helix re-sampling operator induces conformational changes to helical regions by sampling from within a restricted range of acceptable Ramachandran states. For a region of sequence predicted as helical using PSIPRED, the start- and end-points i_h and j_h are calculated then expanded to include the nearest neighbouring residue $i_h - 1$ and $j_h - 1$. A number of residues residing within the helical segment are then perturbed by a magnitude $|\Delta\theta|$ sampled from a Gaussian distribution with $\mu = 0$ and $\sigma = 5°$. The helix is then reconstructed using the restraint-based CCD algorithm.

4.3.3.6 β-strand re-sampling

The β-strand re-sampling operator induces conformational changes to β-strand regions by sampling from within a restricted range of acceptable Ramachandran states. β-sheet construction requires the identification of paired strands in order to model the hydrogen bonding network which stabilises these secondary structural elements.

The following four strands combinations are used; β, coil-β, β-coil, β-coil-β. A combination is selected at random for each use of the β-sheet operator and the start- and end- points of the different secondary structure classes then determined. One third of the torsion angles are then perturbed with the magnitude of $|\Delta\theta|$ sampled from a Gaussian distribution with $\mu = 0$ and $\sigma = 10°$ for β-strand residues and $\mu = 0$ and $\sigma = 15°$ for predicted coils. Following perturbation of the segment, the CCD algorithm is used to reconstruct the strand or strand-coil combination using the Ramachandran restraints.

4.4 Data sets and model processing

4.4.1 CASP6 ROBETTA models

The refinement protocols require models with complete structural coverage of the query sequence. In this study, full-length models submitted by the ROBETTA automated prediction server for 35 comparative modelling targets from the CASP6 experiment (Moult et al. 2005, Tress et al. 2005) are used as a refinement test set. Five ROBETTA models are submitted per target and only targets from the easy (CM/easy) and hard (CM/hard) comparative modelling classes are refined. These two classes represent the most challenging categories for refinement as most of the templates are close homologues and the comparative models are often within close range of the native basin (1 - 6Å C_αRMSD). Descriptions of the targets and model information can be found in Appendix B.

4.4.2 Regularization

A reduction in the memory requirements of a simulation can be made by substituting native bond lengths and angles of the covalent bonds for equilibrium values. However, reconstructing the protein backbone from torsion angles alone can leads to large RMSD deviations from the native fold. Holmes & Tsai (2004) suggest that deviations as large as 6Å C_αRMSD are typical for a 150-residue protein structure as a result of small deviations in bond angles values, though bond length differences are found to contribute very little to atomic displacement effects (Holmes & Tsai 2004). Therefore, to accurately reproduce the original fold torsion angles must be adjusted so that the RMS error between the reconstructed backbone and the original model is minimized.

In order to optimise the torsion angles, a quasi-Newton non-linear multi-dimensional optimisation procedure is employed. The Broydon-Fletcher-Goldfarb-Shanno (BFGS) algorithm is a gradient descent method in which successive gradient vectors are analysed in order to approximate the Hessian matrix of a function being minimised. Here the function is the RMSD between backbone atoms of the original native (or models) atomic (backbone) coordinates and the same coordinates constructed under equilibrium geometry (Engh & Huber 1991). The gradient of the RMSD is calculated using the method described by Coutsias et al. (2004). A limited memory version of BFGS algorithm (L-BFGS) (Byrd et al. 1995) is used to perform bound

constrained optimization of the torsion angles. Lower and upper bounds are set at [-180 180] degrees for dihedral angles to reflect the maximum and minimum range of dihedral angles found in proteins.

4.4.2.1 Structure regularization

The 35 experimental structures and the 175 ROBETTA models are regularized, and the energy and structural similarity of the models calculated and compared with the original structures. The energies are calculated using the combined statistical energy function (see Section 4.2.4.1) and the structural similarity of the models is calculated using the RMSD and TM-score (Zhang & Skolnick 2004*b*). SCWRL3.0 is used for side-chain modeling with a backbone-dependent rotamer libary (Canutescu et al. 2003).

4.5 Results

4.5.1 Effects of regularization on structural properties

The fine balance of forces that ensure the native state is marginally stable at the global energy minimum results from the tightly packed arrangement of side-chains within the core of the protein. It is therefore important to ensure that rotamer libraries used with side-chain modelling algorithms can adequately re-construct the native-like arrangement of side-chains adequately without a reduction in the energy of the native state. Similarly, the regularization of models can also lead to deviations from the native torsion angles in order to minimize the RMSD between the native structure and a model after the substitution of bond lengths and bond angles with idealized values. The conformation of the backbone after regularization should therefore be both energetically and structurally similar to the native state if a refinement procedure is to reproduce the native structure from a template with equilibrium geometry and a subset of rotamer states.

Table 4.2 shows the energies of the 35 original native structures and the energies of the native conformations with the side-chains re-modelled using SCWRL3.0. Native side-chain rotamer states are not preserved so that side-chain reconstruction relies solely on a sufficient representation of rotamer states in the backbone-dependent rotamer set (Canutescu et al. 2003). The energies are then calculated once the native conformations have been regularized and the side-chains re-modelled on the altered backbone.

In the majority of cases, re-packing the side-chains using a rotamer library on the native backbone leads to a minor increase in total energy when compared with the crystallographic structure. An analysis of the number of correct rotamer states (defined as the number of χ_1 angles within $\pm 30^\circ$ of the native χ_1 torsion angle) shows that even when side-chains are modelled on the native backbone, SCWRL3.0 is only able to achieve $\sim$70% accuracy. Re-packing the side-chains onto a regularized backbone leads to a further increase in energy with fewer native-like rotameric states ($\sim$ 60%).

The regularization procedure successfully reproduces the topology of ROBETTA models and Figure 4.4 shows the structural similarity scores of all models after regularization. The median and interquartile range (IQR) values for the similarity

Table 4.2: The energies of native structures are calculated for the experimental structure, the crystal structure with the side-chains re-packed using a backbone-dependent rotamer library, and for the regularized native structures with side-chains re-modelled on the altered backbone. The majority of cases show an acceptable RMS error (RMSD $\leq 0.5\,\AA$) between the backbone atoms between the true native and the regularized native structure

Target	Nres	E_{native}	E_{native} [a]	E_{reg} [b]	RMS[c]	χ_1(%)[d]	χ_1(%)[e]
T0204	297	-421.7	-397.9	-383.0	1.16	65.8	41.9
T0229_1	102	-167.5	-163.0	-156.0	0.31	72.2	49.2
T0229_2	24	-16.1	-15.4	-14.5	0.01	72.7	69.3
T0231	137	-206.3	-199.7	-192.0	0.37	76.0	57.4
T0233_1	66	-78.5	-76.3	-73.9	0.12	63.0	56.6
T0240	90	-94.6	-94.6	-87.0	0.20	70.7	61.0
T0244	296	-389.3	-378.3	-355.7	2.64	71.6	34.6
T0246	354	-515.0	-502.8	-453.2	0.45	74.5	65.4
T0247_3	76	-82.6	-81.9	-79.7	0.38	100.0	92.9
T0264	289	-396.5	-395.5	-370.4	0.72	65.9	50.3
T0266	150	-209.1	-206.6	-196.2	0.25	76.0	67.7
T0268_2	109	-140.1	-136.5	-134.3	0.18	70.8	68.2
T0269_1	158	-241.3	-232.9	-227.3	0.47	66.2	46.0
T0271	161	-214.5	-211.5	-197.2	0.22	76.3	58.9
T0274	156	-203.0	-196.5	-188.7	0.41	70.5	45.2
T0275	135	-171.9	-171.2	-162.9	0.32	68.0	53.8
T0276	168	-241.1	-234.2	-230.0	0.14	70.6	66.4
T0277	117	-181.2	-173.6	-172.3	0.22	59.6	51.1
T0282	323	-509.7	-487.0	-478.3	1.71	71.7	43.0
T0196	89	-100.8	-98.7	-92.0	0.15	64.5	66.2
T0199_1	74	-97.4	-96.9	-94.6	0.15	70.0	69.6
T0200	255	-357.8	-343.0	-341.4	0.24	64.3	60.7
T0205	103	-123.6	-121.8	-118.0	0.17	76.7	73.4
T0208	344	-504.3	-488.6	-451.1	2.21	60.4	43.8
T0211	136	-183.5	-178.1	-169.6	0.42	74.2	66.9
T0222_1	264	-377.4	-363.1	-360.6	0.48	68.2	62.6
T0223_1	114	-150.3	-147.1	-57.2	1.55	70.2	41.3
T0232_1	81	-104.5	-103.8	-91.5	0.21	77.6	59.7
T0232_2	146	-200.0	-192.5	-189.1	0.20	73.5	65.0
T0234	135	-182.2	-177.7	-172.0	0.44	77.4	75.4
T0264_2	173	-235.8	-232.4	-220.5	0.23	70.1	62.3
T0265	102	-137.0	-136.8	-128.0	0.32	70.8	67.6
T0267	174	-239.3	-237.8	-232.1	0.26	57.3	51.0
T0269_2	61	-60.4	-58.9	-50.8	0.23	70.6	62.8
T0279_2	121	-169.5	-165.5	-165.0	0.16	71.9	51.7
$\mu(\sigma)$		-220.1	-214.2	-202.5	0.51	70.9	58.8

[a] True native backbone with side-chains re-packed using SCWRL3.0.

[b] Regularized native structure with side-chains re-packed using SCWRL3.0.

[c] Backbone RMSD between the native and regularized structures.

[d] Percentage of native χ_1 side-chain torsion angles correctly modelled with SCWRL3.0 on the native backbone.

[e] Percentage of native χ_1 side-chain torsion angles in correctly modelled regularized native structures.

scores of models after regularization show that the regularization procedure with restricted ω torsion angles is able to accurately reproduce the original topology with minimal error, and the majority of outliers represent models containing >250 amino acid residues. As expected, the backbone RMSD (a measure of global structural similarity) is far more sensitive to small structural deviations than the heuristic similarity methods (TM-score, GDT, MaxSub) which use only subsets of C_α atoms to calculate the similarity score.

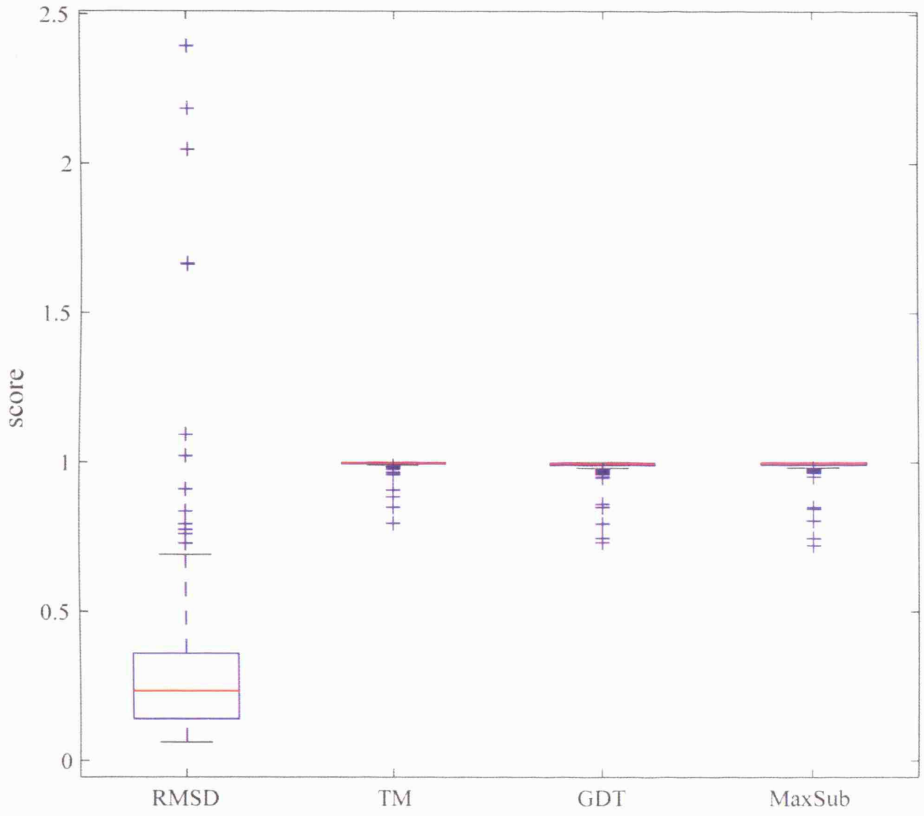


Figure 4.4: The structural similarity scores between original models and the regularized versions are shown for all 175 CASP6 ROBETTA models. The backbone RMSD is more sensitive to global topological differences resulting from regularization than the heuristic similarity measures, though all measures show only minor structural deviations between the backbones of the original models and the regularized models except for some extreme outliers. The median and interquartile range (IQR) values for the four measures are $RMSD = 0.23\,(0.22)$, $TM = 0.99\,(0.00)$, $GDT = 0.996\,(0.01)$, and, $MaxSub = 0.99\,(0.01)$.

The energy differences between regularized models and original ROBETTA models are shown in Figure 4.5. Although regularization introduces only minor structural deviations, these torsion angle deviations subsequently lead a reduction in

model energy after side-chain re-packing, a likely consequence of the use of backbone information in the selection of rotamer states for each residue.

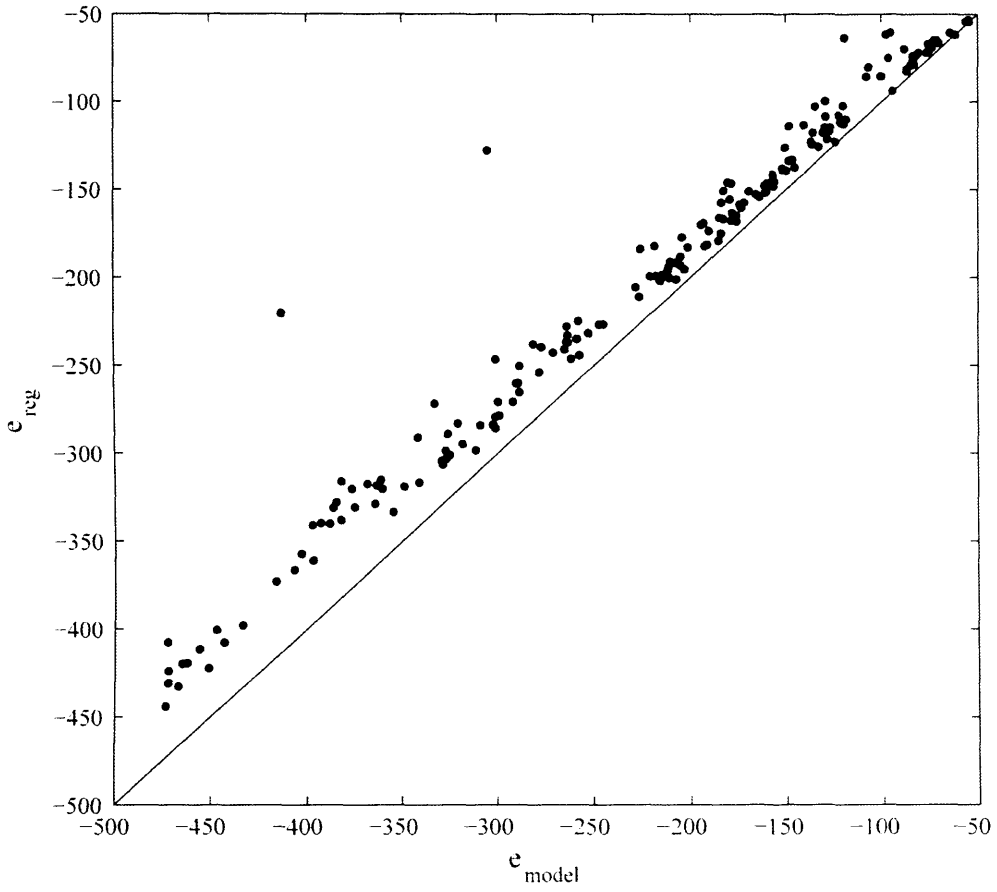


Figure 4.5: The energies of original ROBETTA models after stereochemical regularization and side-chain re-packing using SCWRL3.0 are shown. In the majority of cases, the regularization leads to a increase in model energy after the side-chain conformations are placed on the regularized backbone using SCWRL3.

4.5.2 Single-objective refinement of CASP6 ROBETTA models

For each CASP6 target, 10 refinement runs were performed using the single-objective refinement protocol (see Section 4.2.1). Each refinement was performed until termination either by energy convergence or after the maximum number of generations was reached. Due to the stochastic nature of GAs, the models from each refinement run were pooled before analysis to ensure the statistical significance of the results. Table 4.3 and Table 4.4 show the single-objective refinement results for the *CM/easy* targets and *CM/hard* targets respectively.

Table 4.3: The refinement of *CM/easy* models with a single-objective GA refinement strategy.

		RMSD (Å)			Energy		
Target	*Nres*	T_{rms}[a]	$< rms_{best}^{10\%} >$[b]	$< rms_{min} >$[c]	$< E_{best}^{10\%} >$[d]	$< E_{min} >$[e]	ΔE_N (ΔE_N^{ref})[f]
T0204	297	3.82 (5)	3.74±0.31	3.86±0.12	-371.0±5.25	-380.5±3.26	-421.7 (-383.0)
T0229_1	138	2.91 (2)	2.44±0.16	2.82±0.23	-110.2±2.56	-131.5±2.00	-167.5 (-156.0)
T0229_2	24	1.04 (2)	0.91±0.05	1.27±0.05	-13.0±1.76	-21.9±0.49	-16.1 (-14.5)
T0231	137	3.99 (3)	2.92±0.16	3.78±0.36	-160.0±4.72	-182.4±2.70	-206.3 (-192.0)
T0233_1	66	1.43 (2)	1.04±0.02	1.67±0.11	-75.2±0.99	-80.2±0.23	-78.5 (-73.9)
T0240	90	20.25 (2)	9.87±0.07	10.08±1.03	-76.1±1.03	-77.4±0.04	-94.6 (-87.0)
T0244	296	5.60 (4)	5.18±0.22	6.90±0.24	-314.5±1.50	-333.3±0.22	-389.3 (-355.7)
T0246	354	1.93 (5)	1.94±0.21	2.02±0.09	-427.8±2.25	-434.6±0.58	-515.0 (-453.2)
T0247_3	76	2.89 (3)	2.13±0.03	2.84±0.21	-66.2±2.60	-74.9±0.36	-82.6 (-79.7)
T0264	289	5.85 (1)	6.42±0.11	7.68±0.52	-252.6±5.81	-344.2±0.15	-396.5 (-370.4)
T0266	150	1.80 (3)	1.90±0.12	2.13±0.06	-177.4±2.31	-186.4±0.04	-209.1 (-196.2)
T0268_2	109	2.52 (1)	2.48±0.11	3.03±0.13	-118.4±4.52	-130.4±0.03	-140.1 (-134.3)
T0269_1	158	2.60 (1)	2.83±0.30	4.04±0.31	-170.3±0.62	-215.2±1.03	-241.3 (-227.3)
T0271	161	3.28 (5)	3.32±0.21	6.54±0.50	-183.3±8.13	-211.0±1.81	-241.5 (-197.2)
T0274	156	3.62 (4)	3.55±0.22	4.53±0.26	-172.8±6.27	-195.3±2.37	-203.0 (-188.7)
T0275	135	2.65 (3)	2.62±0.11	3.68±0.61	-134.9±1.90	-163.6±1.73	-171.9 (-162.9)
T0276	168	2.56 (3)	2.39±0.20	2.45±0.20	-208.7±2.10	-218.4±2.01	-241.1 (-230.0)
T0277	117	1.58 (1)	1.54±0.02	1.83±0.11	-167.5±2.23	-178.1±1.50	-181.2 (-172.3)
T0282	323	6.62 (3)	6.30±0.36	9.86±0.42	-310.6±3.82	-411.8±5.07	-509.7 (-478.3)

[a] RMSD of the best ROBETTA model (model rank).

[b] Mean RMSD and standard deviation of the top 10% of pooled lowest RMSD structures.

[c] Mean RMSD and standard deviation of the top 10 lowest energy structures (1 per simulation).

[d] Mean energy and standard deviation of the top 10% of pooled lowest RMSD structures.

[e] Mean energy and standard deviation of the top 10 lowest energy structures (1 per simulation).

[f] The energy gap between the average energy of the lowest energy cluster and the native state (or regularized native).

The mean backbone RMSD ($< rms_{min} >$) and energy ($< E_{min} >$) were obtained by taking the arithmetic mean of the 10 lowest energy models (i.e. the lowest energy model from each of the 10 refinement simulations) while the mean backbone RMSD of best sampled structures ($< rms_{best}^{10\%} >$) was calculated using the top 10% of the 100 lowest RMSD models (the 10 lowest RMSD models from each refinement run).

Table 4.4: The refinement of *CM/hard* models with a single-objective GA refinement strategy

		RMSD (Å)			Energy		
Target	*Nres*	T_{rms}[a]	$< rms_{best}^{10\%} >$[b]	$< rms_{min} >$[c]	$< E_{best}^{10\%} >$[d]	$< E_{min} >$[e]	ΔE_N (ΔE_N^{ref})[f]
T0196	89	4.61 (3)	2.40±0.34	3.44±0.31	-71.5±5.82	-94.5±0.01	-100.8 (-92.0)
T0199_1	74	3.21 (2)	2.02±0.32	2.14±0.13	-81.0±2.86	-88.4±0.71	-97.4 (-94.6)
T0200	255	11.22 (2)	7.48±0.46	16.59±1.47	-182.4±2.87	-274.3±0.00	-357.8 (-341.4)
T0205	103	3.76 (1)	3.41±0.24	3.57±0.17	-129.9±6.83	-148.4±0.03	-123.6 (-118.0)
T0208	344	14.97 (1)	14.91±0.17	19.66±0.07	-174.4±3.36	-412.9±0.03	-504.3 (-451.1)
T0211	136	4.27 (3)	3.99±0.14	4.10±0.02	-132.6±0.45	-151.7±0.29	-183.5 (-169.6)
T0222_1	264	13.91 (4)	13.89±0.32	14.61±0.31	-132.3±7.68	-229.9±0.03	-377.4 (-360.6)
T0223_1	114	9.06 (2)	8.84±0.33	17.14±1.58	-22.5±3.87	-164.2±0.00	-150.3 (-57.2)
T0232_1	81	3.46 (2)	3.39±0.11	3.64±0.13	-80.7±1.31	-95.6±0.27	-104.5 (-91.5)
T0232_2	146	7.41 (4)	5.42±0.25	6.48±0.86	-138.0±0.79	-169.4±0.01	-200.0 (-189.1)
T0234	135	6.17 (4)	5.01±0.25	5.99±0.03	-148.2±7.05	-172.5±0.01	-182.2 (-172.0)
T0264_2	173	6.59 (1)	5.24±0.19	6.40±0.54	-138.4±4.90	-196.2±0.00	-235.8 (-220.5)
T0265	102	6.44 (5)	5.14±0.10	6.94±0.63	-112.1±3.49	-131.5±0.68	-137.0 (-128.0)
T0267	174	2.48 (5)	2.18±0.02	2.99±0.02	-204.5±7.80	-225.9±0.01	-239.3 (-232.1)
T0269_2	61	7.72 (3)	5.88±0.32	7.83±0.92	-36.2±4.67	-57.6±0.25	-60.4 (-50.8)
T0279_2	121	2.77 (1)	2.85±0.12	3.10±0.11	-107.3±3.54	-139.2±0.30	-169.5 (-165.0)

[a] RMSD of the best ROBETTA model (model rank).

[b] Mean RMSD and standard deviation of the top 10% of pooled lowest RMSD structures.

[c] Mean RMSD and standard deviation of the top 10 lowest energy structures (1 per simulation).

[d] Mean energy and standard deviation of the top 10% of pooled lowest RMSD structures.

[e] Mean energy and standard deviation of the top 10 lowest energy structures (1 per simulation).

[f] The energy gap between the average energy of the lowest energy cluster and the native state (or regularized native).

The single-objective GA is able to obtain a cluster of models whose representatives are structurally more similar to the experimental structure than the best ROBETTA models for the majority of targets. However, the average energy ($< E_{best}^{10\%} >$) of models within these low RMSD clusters is often much greater than the average energy ($< E_{min} >$) of the representatives within the lowest energy cluster. For many of the *CM/easy* targets, exploration of low energy regions of the energy function by the GA leads to convergence on deep local minima, resulting in a deterioration in model accuracy after refinement compared with the original ROBETTA models.

After refinement, a sizeable energy gap ΔE_N, remains between the average energy of the lowest energy cluster $< E_{min} >$, and the energy of the native conformation E_{native},

while the energy gap ΔE_N^{ref}, between the lowest energy cluster and the regularized native is often smaller. In some cases, for example target T0229_1 ($\approx 1\AA$ RMSD from the native) and T0232_1 ($\approx 1.5\AA$ RMSD), the energy of the refined models is between 2-5 kcal/mol less than the regularized native, suggesting an inaccuracy in the energy function is the likely source of error.

4.5.2.1 Can ROBETTA models be refined with single-objective GAs?

Figure 4.6 shows the change in RMSD, the $\Delta RMSD$, when either a representative from the lowest RMSD cluster is selected after refinement instead of the best ROBETTA model, compared with the $\Delta RMSD$ resulting from selecting a representative from the lowest energy cluster. This shows the limits of the best possible refinement attainable with the current implementation and parameters for each target compared with the actual refinement produce by selecting the lowest energy conformation from the total set of structures sampled.

For most *CM/easy* targets, only a slight improvement in $\Delta RMSD$ is possible even when selecting models from the low RMSD cluster, whereas for *CM/hard* targets, the lowest RMSD cluster usually contains models that are nearer-native models than the best ROBETTA model. Out of the 19 *CM/easy* targets, 14 targets were improved by selecting a low RMSD model but only 5 targets (T0229_1, T0231, T0240, T0246_3, T0276) were refined using selection by model energy. For the *CM/hard* targets, 15/16 targets had a cluster of structures with improved $\Delta RMSD$ and here 7/16 cases were successfully refined by selecting lowest energy models.

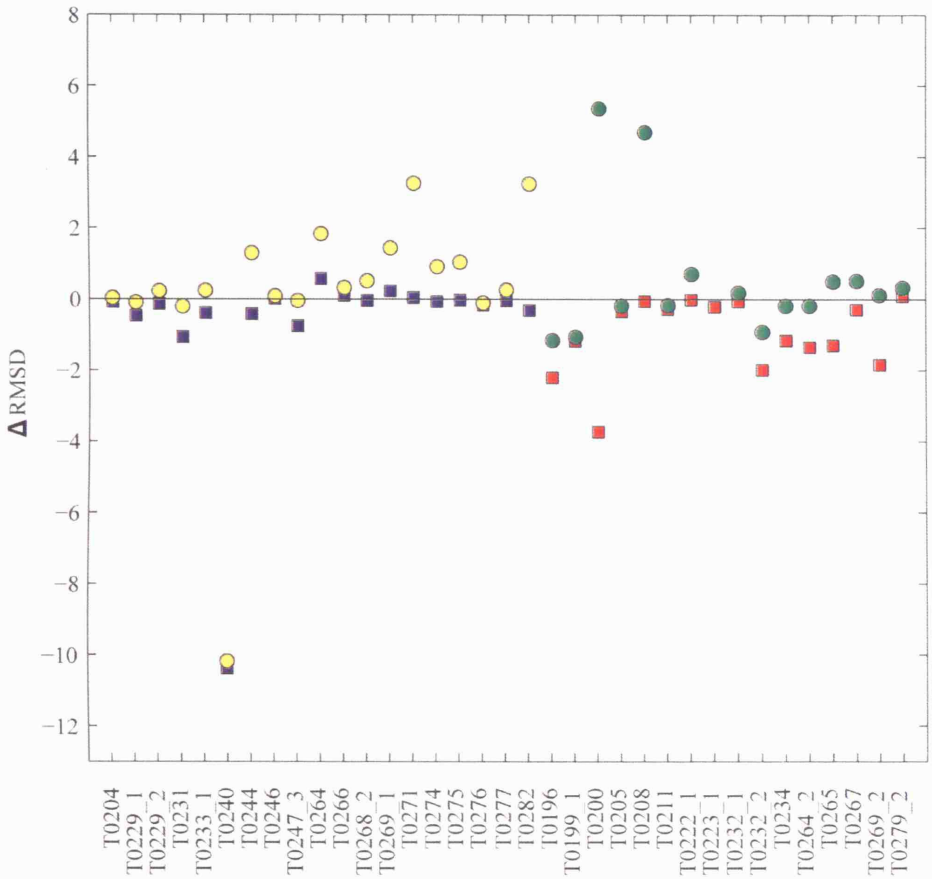


Figure 4.6: The changes in RMSD ($\Delta RMSD$) are shown for CASP6 models after refinement with a single-objective genetic algorithm. The $\Delta RMSD$, the difference between the RMSD of the best original model and the mean RMSD of the top 10% of lowest RMSD models from the pooled refinement runs, is represented by squares (blue for *CM/easy* targets, red for *CM/hard* targets). The $\Delta RMSD$ between the best template and the mean RMSD of the top 10 low energy structures is shown for the *CM/easy* targets (yellow circles) and *CM/hard* targets (green circles).

4.5.3 Multi-objective refinement of CASP6 ROBETTA models

The multi-objective refinement simulations follow a similar refinement protocol as outlined for single-objective refinement with the exception of the model selection stage. In single-objective optimization, the final refined model is selected by energy so that the lowest energy structure generated during the simulation is always submitted as the best model. In multi-objective optimization a number of Pareto-optimal solutions are obtained at the end of each refinement run. The choice of solution from this Pareto-optimal set is then made *a posteriori* and requires either addition domain-specific information or some other criterion with which a decision maker may chose a solution.

Instead of simply selecting the lowest energy model from the set of non-dominated solutions, here the non-dominated set is first reduced from its full size (200 individuals) to a smaller subset of solutions. This reduction is performed by selecting a subset of "knee" solutions (Branke et al. 2005). In the case of two objectives, knee points represent solutions on the Pareto-front in which a small improvement in one objective will cause a large deterioration in the other. Knee points are determined by the finding the individuals with the largest angles between the individual x_i, and its four neighbours $x_{i+2}, x_{i+1}, x_{i-1}, x_{i-2}$, for all solutions in the non-dominated front. For an individual x_i, the angles are calculated between the individual solution and the two neighbours on each side so that four angles are measured per individual ($(\widehat{x_{i-1}, x_i, x_{i+1}})$, $(\widehat{x_{i-1}, x_i, x_{i+2}})$, $(\widehat{x_{i-2}, x_i, x_{i+1}})$, $(\widehat{x_{2-1}, x_i, x_{i+2}})$). The largest of the four angles is then assigned to the individual x_i. The full Pareto-optimal set $\boldsymbol{A}$ is thus reduced to a subset of individuals represented by the largest K angles (where K represents the maximum number of knee solutions to include), and the model with the lowest energy is chosen from this subset of knee solutions.

Table 4.5 and Table 4.6 show the results of multi-objective refinement for *CM/easy* and *CM/hard* targets, respectively.

Table 4.5: The refinement of *CM/easy* models with a multiple-objective GA refinement strategy

		RMSD (Å)				Energy		
Target	*Nres*	rms_T[a]	$< rms^{10\%}_{best} >$[b]	$< rms_{min} >$[c]	$< rms_{knee} >$	$< E^{10\%}_{best} >$[d]	$< E_{min} >$[e]	$< E_{knee} >$[f]
T0204	297	3.82 (5)	3.78±0.13	3.83±0.11	3.82±0.21	-349.8±11.86	-367.9±7.34	-366.3±4.25
T0229_1	138	2.91 (2)	2.34±0.37	3.57±0.20	2.89±0.14	-106.9±7.07	-132.5±4.30	-128.9±3.67
T0229_2	24	1.04 (2)	0.91±0.02	1.58±0.14	1.72±0.21	-11.8±1.65	-22.5±0.55	-18.8±1.20
T0231	137	3.99 (3)	3.29±0.20	3.71±0.10	3.67±0.16	-155.7±6.28	-182.2±2.14	-180.7±0.60
T0233_1	66	1.43 (2)	1.15±0.11	6.19±0.38	6.01±0.59	-73.4±1.48	-81.5±0.71	-79.3±0.90
T0240	90	20.25 (2)	9.75±0.73	10.05±0.36	10.05±0.42	-71.1±11.30	-76.3±0.14	-76.0±0.52
T0244	296	5.60 (4)	4.95±0.14	5.29±0.23	4.68±0.13	-283.8±14.94	-321.5±5.29	-277.3±3.49
T0246	354	1.93 (5)	2.12±0.21	4.05±0.22	3.96±0.11	-399.2±0.85	-426.7±8.14	-414.2±4.63
T0247_3	76	2.89 (3)	2.72±0.45	4.23±0.55	4.20±0.08	-61.6±4.36	-74.9±1.36	-74.0±0.72
T0264	289	5.85 (1)	5.44±0.41	6.52±0.29	5.82±0.10	-244.2±29.56	-326.0±6.23	-320.6±2.78
T0266	150	1.80 (3)	1.84±0.52	4.87±0.44	3.45±0.07	-154.9±4.98	-178.2±4.13	-177.2±2.81
T0268_2	109	2.52 (1)	3.04±0.22	4.02±0.15	3.78±0.12	-94.8±20.42	-128.7±2.12	-126.1±1.41
T0269_1	158	2.60 (1)	2.56±0.14	2.87±0.16	2.83±0.06	-173.8±16.79	-209.3±3.15	-188.4±2.86
T0271	161	3.28 (5)	3.18±0.46	9.05±0.71	8.18±0.21	-168.3±3.45	-214.4±4.24	-212.4±1.22
T0274	156	3.62 (4)	3.51±0.26	6.76±0.21	5.61±0.15	-162.7±1.28	-196.9±4.19	-190.9±2.02
T0275	135	2.65 (3)	2.49±0.10	3.39±0.14	3.07±0.11	-148.4±0.94	-158.5±2.29	-152.8±2.17
T0276	168	2.56 (3)	2.50±0.23	3.39±0.17	3.25±0.07	-189.2±0.19	-225.4±2.19	-221.0±1.78
T0277	117	1.58 (1)	1.75±0.02	1.86±0.06	1.81±0.03	-173.0±3.21	-177.8±1.07	-177.1±1.04
T0282	323	6.62 (3)	5.68±0.45	6.77±0.82	6.09±0.16	-286.9±23.38	-374.5±3.86	-364.2±3.60

[a] RMSD of the best ROBETTA model (model rank).

[b] Mean RMSD and standard deviation of the top 10% of pooled lowest RMSD structures.

[c] Mean RMSD and standard deviation of the top 10 lowest energy structures (1 per simulation).

[d] Mean energy and standard deviation of the top 10% of pooled lowest RMSD structures.

[e] Mean energy and standard deviation of the top 10 lowest energy structures (1 per simulation).

[f] Mean energy and standard deviation of the top 10 knee solutions (1 per simulation).

After multi-objective refinement, 15/19 *CM/easy* targets had a significant sampling of low RMSD models, however only 3 targets (T0231, T0240, T0244) were improved by selecting a model from the lowest energy cluster. Selecting the lowest energy model after filtering the Pareto-set using knee selection increases the number of refined *CM/easy* targets to 6 (T0229_1, T0231, T0240, T0244, T0264, T0279_2). For *CM/hard* targets 15/16 targets had a cluster of lower RMSD models of which 4 were improved by selecting one of the lowest energy representatives, increasing to 5 targets after knee selection.

Table 4.6: The refinement of *CM/hard* models with a multiple-objective GA refinement strategy

		RMSD (Å)				Energy		
Target	*Nres*	rms_T [a]	$< rms^{10\%}_{best} >$ [b]	$< rms_{min} >$ [c]	$< rms_{knee} >$	$< E^{10\%}_{best} >$ [d]	$< E_{min} >$ [e]	$< E_{knee} >$ [f]
T0196	89	4.61 (3)	2.37±0.23	3.08±0.13	2.89±0.04	-70.6±4.95	-92.7±1.05	-90.1±1.21
T0199_1	74	3.21 (2)	2.23±0.15	2.67±0.19	2.45±0.12	-74.2±4.53	-81.0±0.68	-78.1±1.00
T0200	255	11.22 (2)	10.98±0.67	13.47±1.43	13.47±0.15	-143.9±27.88	-234.2±4.13	-233.4±4.20
T0205	103	3.76 (1)	3.51±0.23	4.00±0.21	3.98±0.04	-125.4±12.40	-146.9±2.13	-142.3±2.57
T0208	344	14.97 (1)	12.94±0.81	15.43±0.56	14.30±0.23	-90.2±10.37	-364.9±1.17	-000.0±2.42
T0211	136	4.27 (3)	4.21±0.03	5.54±0.62	4.81±0.10	-125.6±11.68	-152.0±2.33	-149.9±1.06
T0222_1	264	13.91 (4)	13.39±0.25	14.50±0.27	14.50±0.54	-129.3±24.29	-223.4±7.02	-222.0±2.48
T0223_1	114	9.06 (2)	7.77±0.08	9.01±0.91	8.97±0.26	-96.5±30.72	-159.8±3.03	-123.4±3.17
T0232_1	81	3.46 (2)	3.19±0.16	4.74±0.12	4.71±0.03	-83.7±5.51	-93.6±1.40	-90.7±0.80
T0232_2	146	7.41 (4)	4.90±0.04	6.41±0.62	6.01±0.78	-96.0±5.28	-155.2±1.15	-154.6±2.71
T0234	135	6.17 (4)	5.10±0.16	6.39±0.31	6.31±0.59	-131.7±8.45	-166.4±2.11	-161.6±1.15
T0264_2	173	6.59 (1)	5.36±0.23	7.93±0.53	7.00±0.32	-135.6±27.47	-200.6±4.03	-196.8±2.45
T0265	102	6.44 (5)	4.54±0.12	8.66±0.43	7.30±0.50	-110.2±3.15	-128.5±2.01	-126.1±1.78
T0267	174	2.48 (5)	2.51±0.08	4.04±0.12	4.01±0.81	-199.7±7.86	-230.2±3.12	-227.5±3.10
T0269_2	61	7.72 (3)	5.36±0.21	9.10±0.40	8.87±0.62	-39.5±1.84	-53.5±1.03	-50.8±0.93
T0279_2	121	2.77 (1)	2.71±0.11	4.46±0.11	4.00±0.53	-114.6±1.74	-137.7±3.06	-135.3±2.95

[a] RMSD of the best ROBETTA model (model rank).

[b] Mean RMSD and standard deviation of the top 10% of pooled lowest RMSD structures.

[c] Mean RMSD and standard deviation of the top 10 lowest energy structures (1 per simulation).

[d] Mean energy and standard deviation of the top 10% of pooled lowest RMSD structures.

[e] Mean energy and standard deviation of the top 10 lowest energy structures (1 per simulation).

[f] Mean energy and standard deviation of the top 10 knee solutions (1 per simulation).

4.5.3.1 Do multi-objective GAs improve models?

Figure 4.7 shows the improvements over the best ROBETTA template after refinement if models are selected from the lowest RMSD cluster, and the models chosen after the knee selection procedure is used.

The multi-objective refinement procedure is able to sample models with $\Delta RMSD \geq 1.5$Å for many of the *CM/hard* targets, and improvements are seen for some *CM/easy* targets, yet the models selected by knee selection are substantially worse (> 2Å for many *CM/easy* cases) than the best ROBETTA model. This failure to sample nearer-native conformations in the 'easy' comparative modeling cases, and the selection of low energy decoys by the multi-objective GA, suggests that the search

procedure used to explore the energy landscape for non-dominated solutions may be insufficiently constrained to the region surrounding the models. For hard comparative modelling cases, the improved sampling of nearer-native conformations is promising, however, the selection of low energy decoys for many cases is a significant problem.

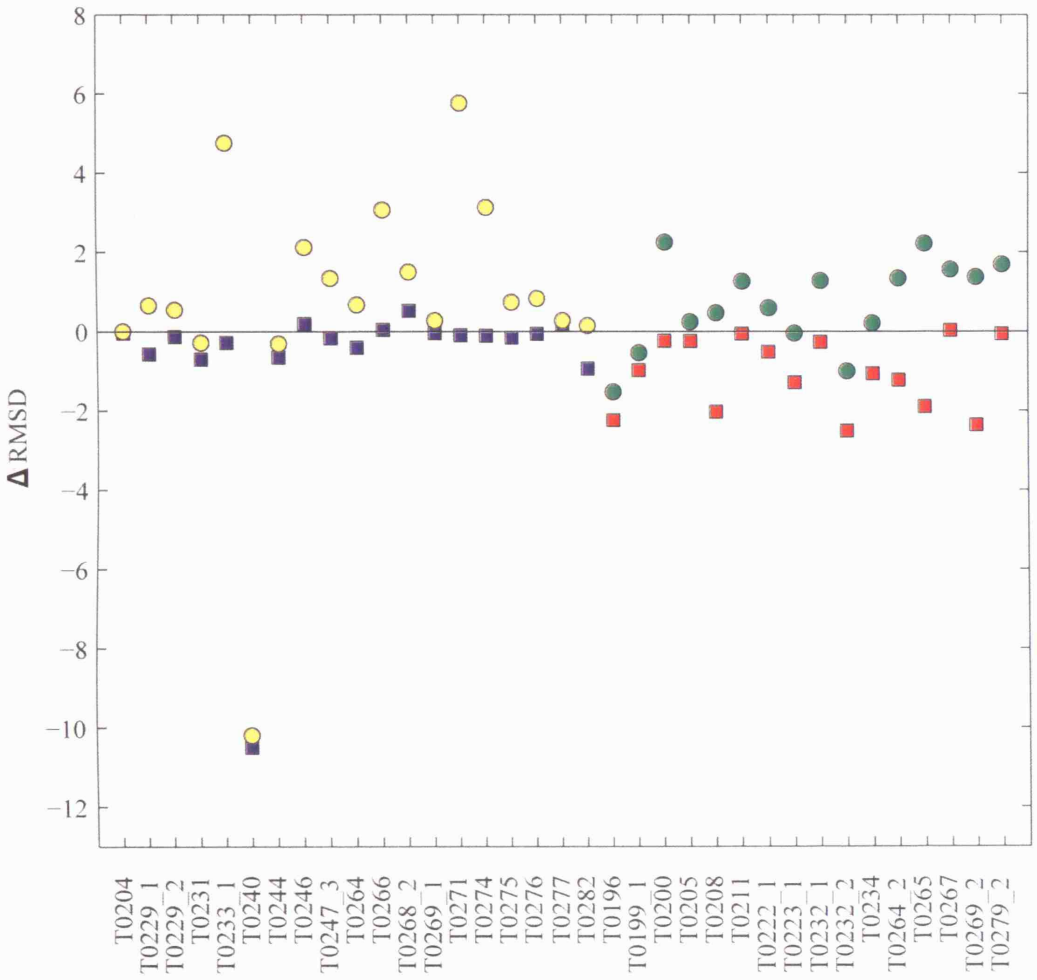


Figure 4.7: The changes in RMSD ($\Delta RMSD$) are shown for CASP6 models after refinement with a multi-objective genetic algorithm. The $\Delta RMSD$, the difference between the RMSD of the best original template and the mean RMSD of the top 10% of lowest RMSD models from the pooled refinement runs, is represented by squares (blue for *CM/easy* targets, red for *CM/hard* targets). The $\Delta RMSD$ between the best template and the mean RMSD of the top 10 low energy structures is shown for the *CM/easy* targets (yellow circles) and *CM/hard* targets (green circles).

4.5.4 Successful refinement cases

To explore the refinement process in more detail, individual refinement cases are examined to understand further how the two evolutionary algorithms traverse the energy landscape during refinement, leading to the outcomes described in Section 4.5.2 and Section 4.5.3.

4.5.4.1 Refinement of *CM/easy* target T0233_1

The *CM/easy* target T0233_1 is a 66 amino acid residue domain with an all-α 4-helix bundle architecture (Orengo et al. 1997). The five unrefined ROBETTA models are built using templates from close homologues and all five models are approximately 1.5Å from the native structure, making this target an ideal candidate for testing the high-resolution refinement ability of the genetic algorithm refinement protocols. The energy range of the regularized ROBETTA models ranges from 68-72 kcal/mol.

The ROBETTA models are refined following both the single- and multi-objective protocol and the energy vs RMSD of the 200,000 models generated using each protocol after 1000 generations of refinement is shown in Figure 4.8. The presence of models with lower RMSD than the best ROBETTA models using both refinement protocols suggests that the conformational sampling operators are at least adequate for sampling conformations within the native basin for this all-α structure. With single-objective refinement a large cluster of low RMSD models is obtained with representatives from this cluster ≈ 1.0Å from the native (see Table 5.2), an improvement of ≈ 0.5Å over the optimal ROBETTA model. However, after 1000 generations of single-objective refinement a lower energy structure is found ≈ 1.6Å from the native state, suggesting a local minimum has been occupied.

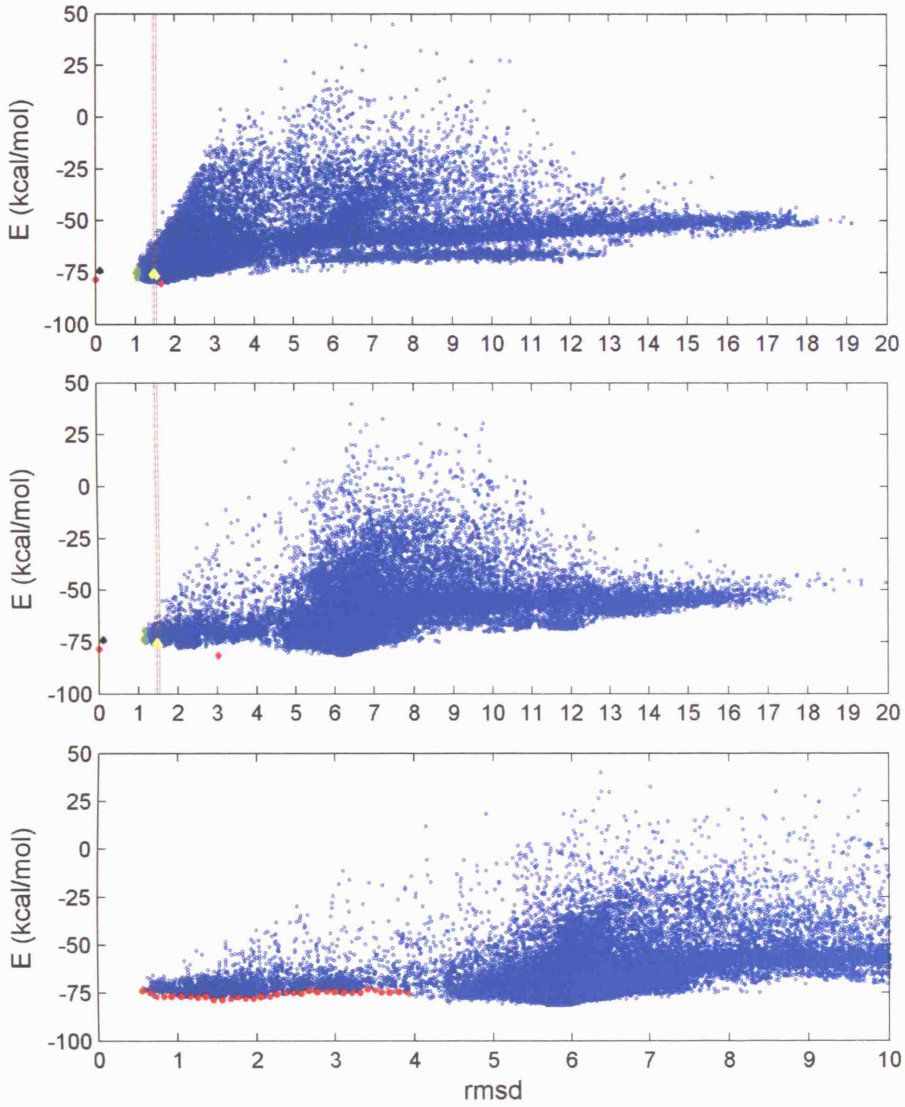


Figure 4.8: The upper and middle plots show the energy vs RMSD of 200,000 models sampled with the single-objective protocol and multi-objective protocol, respectively. The native structure (red) and regularized native (black) are shown, as well as the unrefined ROBETTA models (yellow). The vertical lines intersecting the x axis and the ROBETTA models highlight the RMS distance of each model from the native, while the green points indicate the 10 lowest RMSD models sampled over the course of refinement. For single-objective optimisation, the lowest energy model is shown in magenta while for multi-objective optimization, the magenta point represents the lowest energy models chosen using knee selection. The bottom figure shows energy vs the RMSD of each sampled model from the best ROBETTA template, and the Pareto-optimal set of solutions is highlighted in red.

To examine the range of conformations sampled during refinement and the energetic properties of these models in more detail, the mean population energy $\overline{E}_P$ and the mean population RMSD $\overline{P}_{rms}$ are plotted in Figure 4.9. An initial drop in the average energy of the population ($\approx \Delta E = -5$ kcal/mol) is seen during the first 100 generations with a corresponding increase in the average RMSD of the population $\Delta RMSD = +0.3$, after which the $\overline{E}_P$ is approximately stable ($\sigma = 0.4$).

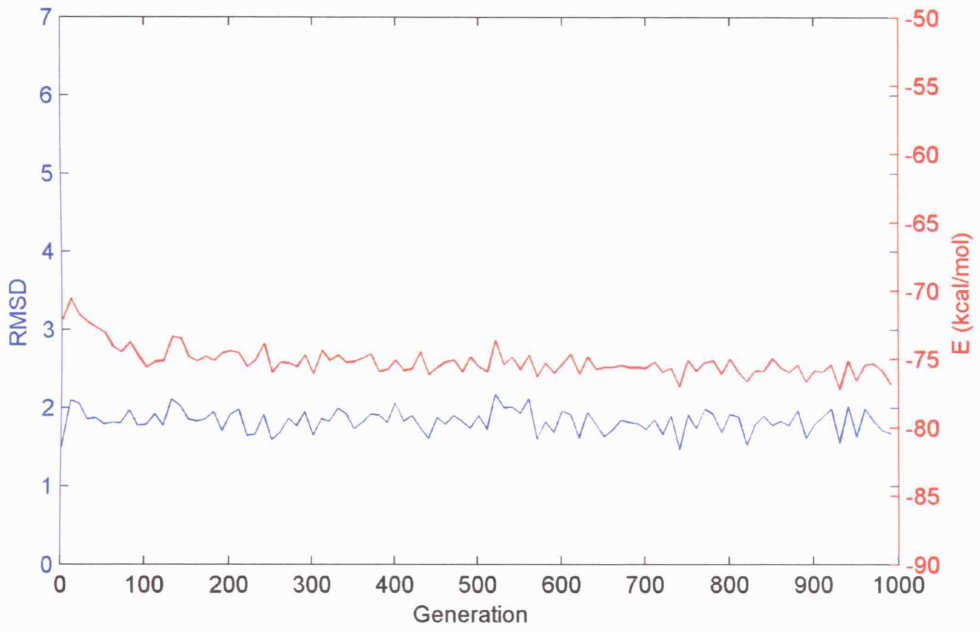


Figure 4.9: The change in mean population energy $\overline{E}_P$, and the mean population RMSD $\overline{P}_{rms}$, during the course of single-objective refinement of *CM/easy* target T0233_1 with a population size $N = 200$, maximum generation number $T = 1000$, crossover rate $P_c = 0.6$ and mutation rate $P_m = 0.3$. The mean population RMSD is shown in blue while the mean population energy is shown in red. Data points are plotted every 10 generations.

The multi-objective refinement strategy shows a similar cluster of low RMSD models at $\approx 1.0\mathring{A}$ (see Figure 4.8) though the energy of these models is not lower than the regularized or original ROBETTA structures. However, the requirement that the multi-objective GA explore the energy landscape to produce a Pareto-optimal set of solutions lead to a navigation away from the region of the near-native ROBETTA models towards a lower energy region at $\approx 6\mathring{A}$ from the native as the GA finds low energy solutions within the upper bound constraint distance of the ROBETTA models (Figure 4.10). Interestingly, the divergence from the topology of the ROBETTA models towards low energy decoy models occurs in the first few generations of

the simulation while a convergence of the mean population energy $\overline{E}_P$ requires a further 300 generations. The lowest RMSD structures are sampled within the first 10 generations though the greater magnitude of the low energy decoy structures affects a drive away from the near-native structures during the remainder of the simulation. Although some low RMSD models remain in the non-dominated set, these models are not representative of knee solutions and were therefore not selected following multi-objective refinement.

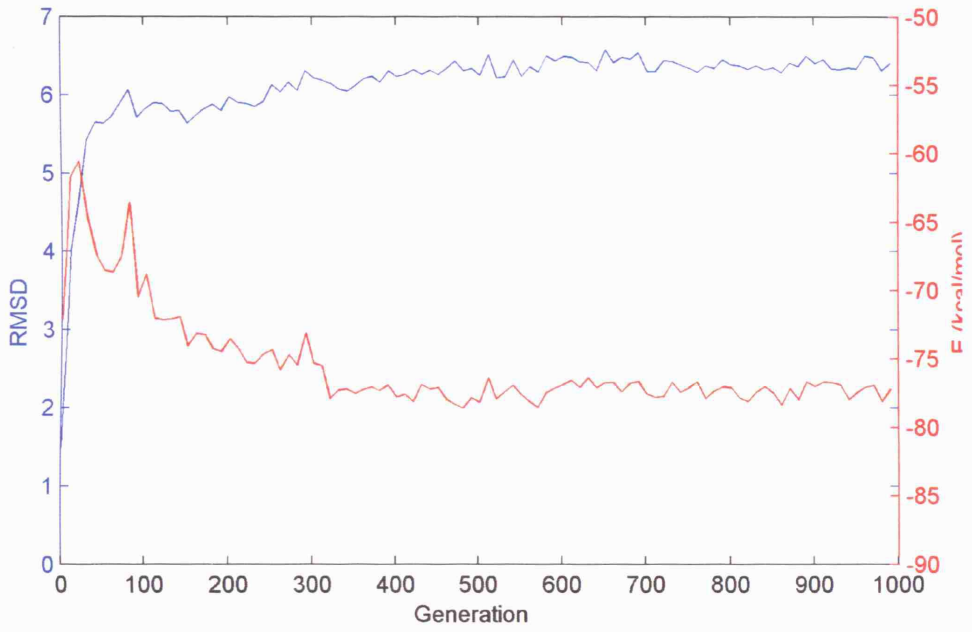


Figure 4.10: The change in mean population energy $\overline{E}_P$, and the mean population RMSD $\overline{P}_{rms}$, during the course of multi-objective refinement of *CM/easy* target T0233_1 with a population size $N = 200$, maximum generation number $T = 1000$, crossover rate $P_c = 0.6$, mutation rate $P_m = 0.3$, lower bound constraint $l = 0.1$, and upper bound constraint $u = 4\text{Å}$. The mean population RMSD is shown in blue while the mean population energy is shown in red. Data points are plotted every 10 generations.

Figure 4.11 shows the structural superposition of representative conformations taken from a single-objective and multiple-objective refinement run on the native structure. A randomly selected representative from the cluster of low RMSD models, the lowest energy model obtained after single-objective refinement, and the lowest energy knee solution from multi-objective refinement. The best ROBETTA template is 1.43Å RMSD (0.84 TM) from the native structure. After single-objective refinement the low RMSD representative's structural similarity score is 1.03Å RMSD (0.90 TM) while the low RMSD representative from the multi-objective simulation scores 1.12Å

RMSD (0.88 TM).

The lowest energy models are shown superposed in Figure 4.12. After single-objective refinement the lowest energy model is 1.12Å RMSD (0.88 TM) from the native, an improvement of ≈ 0.3Å RMSD. The knee selected multi-objective model is 6.19Å RMSD (0.66 TM) from the native, a degradation resulting predominantly from a rotation of the N-terminal helix (labelled C). The combined model energy E_{total} for this model is -79.6 kcal/mol; significantly less that the regularized native ($E_{total} = -73.9$ kcal/mol) but also the true native ($E_{total} = -78.5$ kcal/mol).

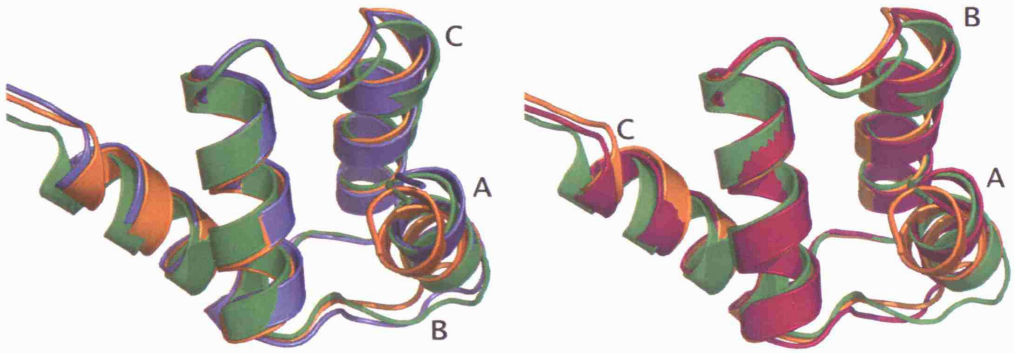


Figure 4.11: Representative refined models taken from the lowest RMSD cluster are shown for single-objective refinement (left) and multi-objective refinement (right). The native structure is shown in green with the best ROBETTA model (orange) and the refined model superposed. Labels A-C highlight regions of structural refinement that improve the quality of the best ROBETTA model. The superpositions were calculated using the maximum likelihood method THESEUS (Theobald & Wuttke 2006).

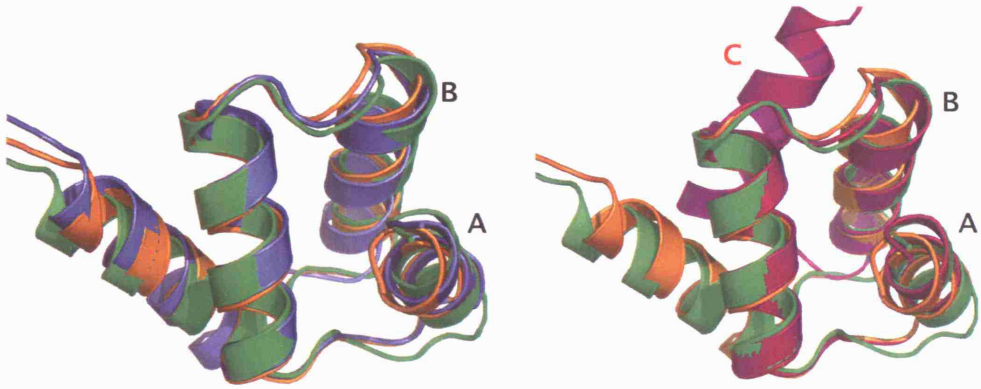


Figure 4.12: Representative refined models taken from the lowest energy cluster are shown for single-objective refinement (left) and multi-objective refinement (right). The native structures is shown in green with the best ROBETTA model (orange) and the refined model superposed. Labels A-C highlight regions of structural refinement that improve the quality of the best ROBETTA model. While the single-objective refinement leads to an improved model, the shift in orientation of the C-terminal helix in the multi-objective refinement model leads to a lower energy structure with greater RMSD from the native (label C). The superpositions were calculated using the maximum likelihood method THESEUS (Theobald & Wuttke 2006).

4.5.4.2 Refinement of *CM/hard* target T0196

The *CM/hard* target T0196, is a 89 residue single domain protein with an 8 stranded β-barrel architecture. The unrefined ROBETTA models are structurally similar to one another with an average $RMSD \approx 4.6\mathring{A}$ from the native structure and an energy range between 81 - 88 kcal/mol. The models were refined first using the single-objective protocol followed by the multi-objective protocols with increased template constraint bounds (see Section 4.2.6).

Figure 4.13 shows the energy vs RMSD plots for a single refinement run using the single-objective and the multi-objective algorithms, respectively. A significant number low RMSD conformations are generated by the sampling procedure during single-objective refinement. While the majority of lower RMSD conformations are at higher energies that the ROBETTA models, the presence of these low RMSD models suggests the conformational sampling operators are adequate for improving these medium resolution models. The 10 lowest RMSD models from the single refinement run (at $\approx 2.5\mathring{A}$ from the native) have marginally lower energies than the regularized ROBETTA models (see Table 4.4) though the lowest energy structure lies at $\approx 3.5\mathring{A}$ and has a marginally lower energy than the regularized native ($\Delta E \approx 2kcal/mol$).

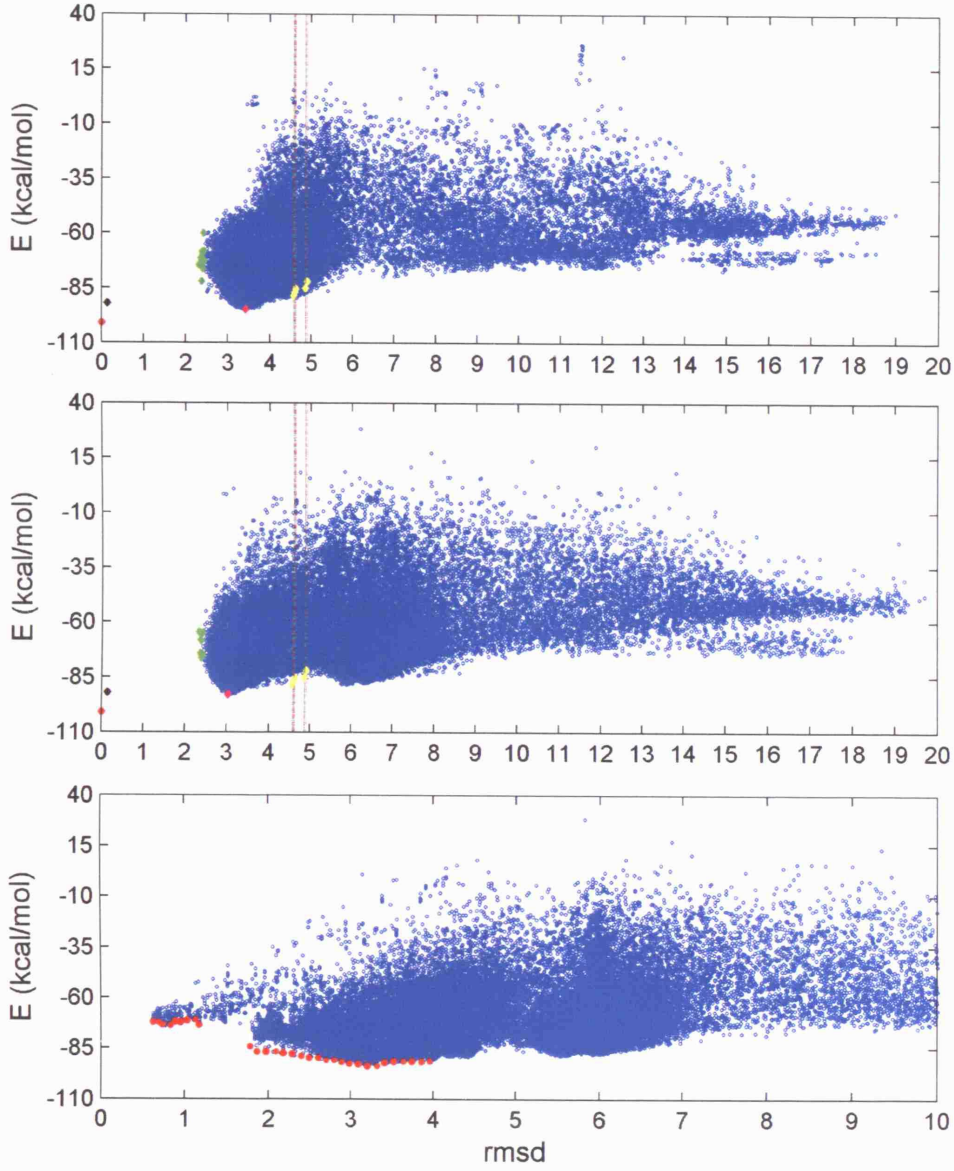


Figure 4.13: The upper and middle plots show the energy vs RMSD of 200,000 models sampled with the single-objective protocol and multi-objective protocol, respectively. The native structure (red) and regularized native (black) are shown, as well as the unrefined ROBETTA models (yellow). The vertical lines intersecting the x axis and the ROBETTA models highlighting the RMS distance of the models from the native, while the green points indicate the 10 lowest RMSD models sampled over the course of refinement. For single-objective optimisation, the lowest model is shown in magenta while for multi-objective optimization, the magenta point represents the lowest energy models chosen using knee selection. The bottom figure shows energy vs RMSD of each sampled model from the best ROBETTA template, and the Pareto-optimal set of solutions is highlighted in red.

A large number of low RMSD models are also sampled with the multi-objective GA with many more near-native models generated than with the single-objective GA. However, the sampled landscape of the multi-objective shows a population of conformations within a deep energy minimum at $\approx 6\mathring{A}$ where the GA has explored conformations in this low energy region during the process of searching for a diverse Pareto-optimal set. This region can be seen in the Pareto-front at $\approx 5.5\mathring{A}$ RMSD from the template (see the lower panel of Figure 4.13), though conformations from this region were not included in the non-dominated set at the end of 1000 generations.

Figure 4.14 shows the single-objective GA's exploration of conformation space during a single refinement run. The average population energy sharply decreases within the first 50 generations at which the rate of change is reduced significantly. The corresponding minimum energy of the population shows a similar trajectory with the GA entering a local energy minimum after ≈ 380 generations. The average RMSD of the population remains at 5Å until it sharply falls to 4Å after 200 generations and oscillates between 3.5Å and 4Å, while the minimum RMSD shows a continuous sampling of low RMSD structures even after 400 generations when there is convergence of the lowest energy conformation.

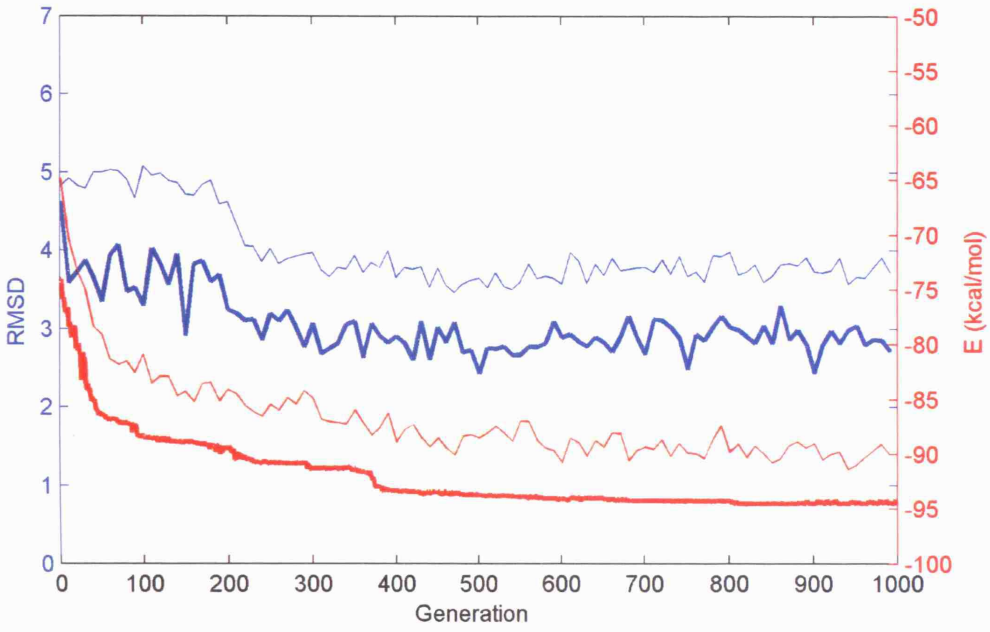


Figure 4.14: The change in mean population energy $\overline{E}_P$, and the mean population RMSD $\overline{P}_{rms}$, during the course of single-objective refinement of *CM/hard* target T0196 with a population size $N = 200$, maximum generation number $T = 1000$, crossover rate $P_c = 0.6$ and mutation rate $P_m = 0.3$. The mean population RMSD is shown in blue (min population RMSD with the thicker line width) while the mean population energy is shown in red (min population energy with thick line width). Data points are plotted every 10 generations.

The multi-objective simulation data in Figure 4.15 shows some interesting features of the population RMSD and energy during the search for a diverse set of Pareto-optimal solutions. Although there is a sharp increase in the average population energy $< E_P >$ in the first 10 generations, followed rapidly by a sharp decrease after 50 generations, the average population RMSD increases. In fact, the trajectory of the average population energy closely follows the minimum RMSD of the population while the average RMSD continues to increase as the simulation approaches termination.

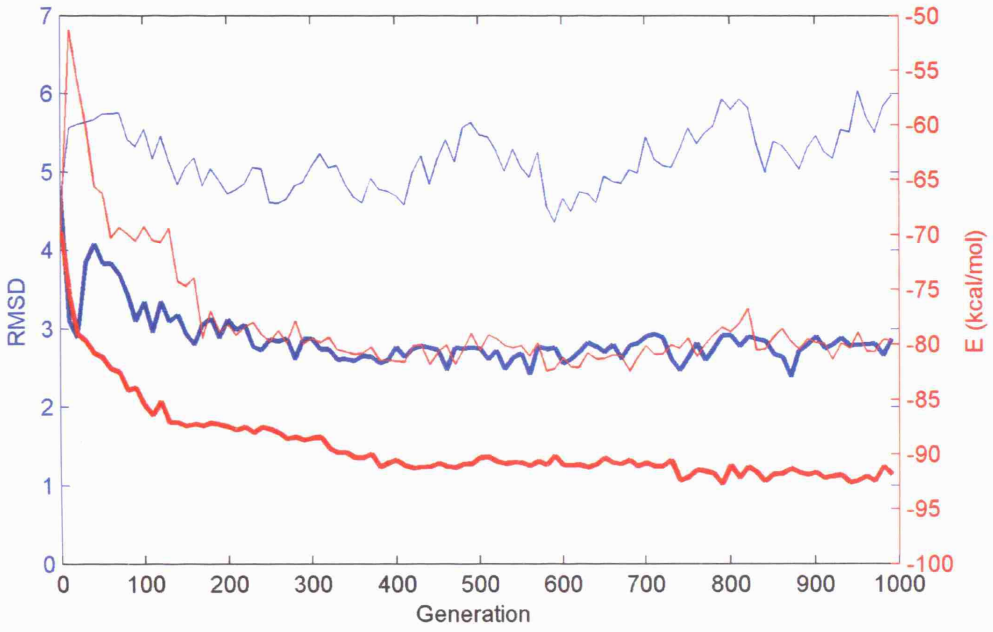


Figure 4.15: The change in mean population energy $\overline{E}_P$, and the mean population RMSD $\overline{P}_{rms}$, during the course of multi-objective refinement of *CM/easy* target T0196 with a population size $N = 200$, maximum generation number $T = 1000$, crossover rate $P_c = 0.6$, mutation rate $P_m = 0.3$, lower bound constraint $l = 0.5$, and upper bound constraint $u = 6\text{Å}$. The mean population RMSD is shown in blue (min population RMSD with the thicker line width) while the mean population energy is shown in red (min population energy with thick line width). Data points are plotted every 10 generations.

Examining cluster representatives of the lowest RMSD models sampled by both protocols shows that many of the improvements obtained are restricted to loop regions (Figure 4.16). The best ROBETTA template is 4.61Å C_αRMSD (0.73 TM) from the native, with many of the native β-strands modelled. After single objective refinement the selected representative has a score of 2.32Å C_αRMSD (0.78 TM) while the representative from the multi-objective refinement scores 2.31Å C_αRMSD (0.77 TM). Most improvements are confined to loop regions, although slight improvement to a β-sheet is seen in the multi-objective model (labelled A in right panel of Figure 4.16).

The lowest energy and knee selected models for T0196 are show in Figure 4.17. The energy selected models are both improvements on the original ROBETTA models ($\Delta RMSD \approx 1\text{Å}$) with the single-objective representative scoring 3.42Å C_αRMSD (0.74 TM) and the multi-objective representative scoring 3.15Å C_αRMSD (0.74 TM). Most improvements to the structure in both cases are in terminal or coil loop regions with

some minor changes to the β-sheet packing (mostly leading to a degradation in the sheet packing as seen from the similar TM-scores of the ROBETTA model and refined structures).

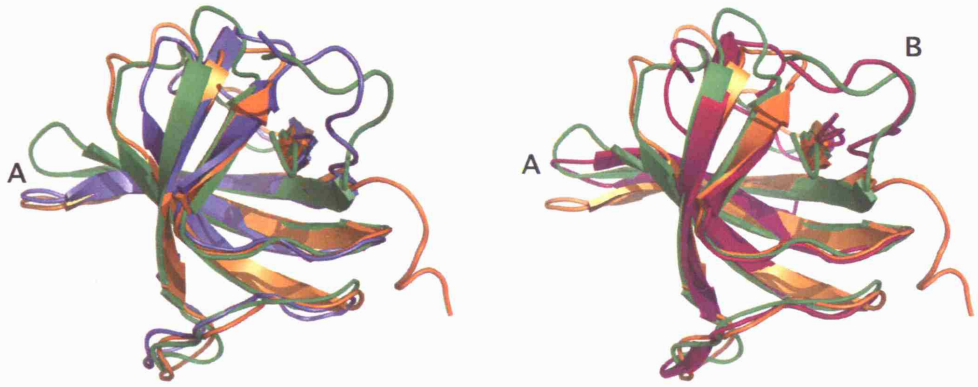


Figure 4.16: Representative refined models taken from the lowest RMSD cluster are shown for single-objective refinement (left) and multi-objective refinement (right). The native structures is shown in green with the best ROBETTA model (orange) and the refined model superposed. Refinement to the β-sheet region (label A) is seen in the multi-objective case with only a slight adjustment to the single-objective model. The C-terminal loop is greatly improved after multi-objective refinement (label B). The superpositions were calculated using the maximum likelihood method THESEUS (Theobald & Wuttke 2006).

Figure 4.17: Representative refined models taken from the lowest energy cluster are shown for single-objective refinement (left) and multi-objective refinement (right). The native structures is shown in green with the best ROBETTA model (orange) and the refined model superposed. The superpositions were calculated using the maximum likelihood method THESEUS (Theobald & Wuttke 2006).

4.5.4.3 Refinement of *CM/hard* target T0199_1

The *CM/hard* target T0199_1 is a 74 residue domain from a heat-inducible transcription repressor protein homolog in *Thermotoga maritima*. The domain has an orthogonal bundle architecture consisting mainly of α-helices with a small β-strand. The regularized ROBETTA models submitted for this target have a wide range of similarity scores and energies, with the closest model at 3.21Å C_αRMSD (-60.76 kcal/mol) from the native and the furthest at 11.63Å C_αRMSD (-60.56 kcal/mol).

Examining the RMSD vs energy plot in the top panel of Figure 4.18 for the single-objective refinement run shows a funnelling of the search towards a deep energy minimum at $\approx 2\mathring{A}$. Even though the GA converges on a low energy non-native structure, the conformation at this point is significantly better than the best starting model (with $\Delta RMSD$ to native >1Å).

The energy vs RMSD plot (middle panel of Figure 4.18) from a multi-objective run again shows an exploration of the energy landscape which includes low energy regions (local minima), some of which are nearer to the native basin than the best ROBETTA models and some which correspond to non-native conformations. The successful use of constrained optimization can be seen from the differences between the regions sampled by the two protocols. In the single-objective refinement a large number of structures are sampled around the 11Å C_αRMSD model which has a similar energy to the best template at 3Å C_αRMSD. However, in the multi-objective case, the sampling is quickly restricted by the constraint penalty to the region of conformational space within the lower and upper constraint bounds ($l = 0.5\mathring{A}$ and $u = 6\mathring{A}$) of the lowest energy template. Although this inhibits sampling around the non-native models, fewer low energy models are sampled with RMSD <3Å in total than with the single-objective GA.

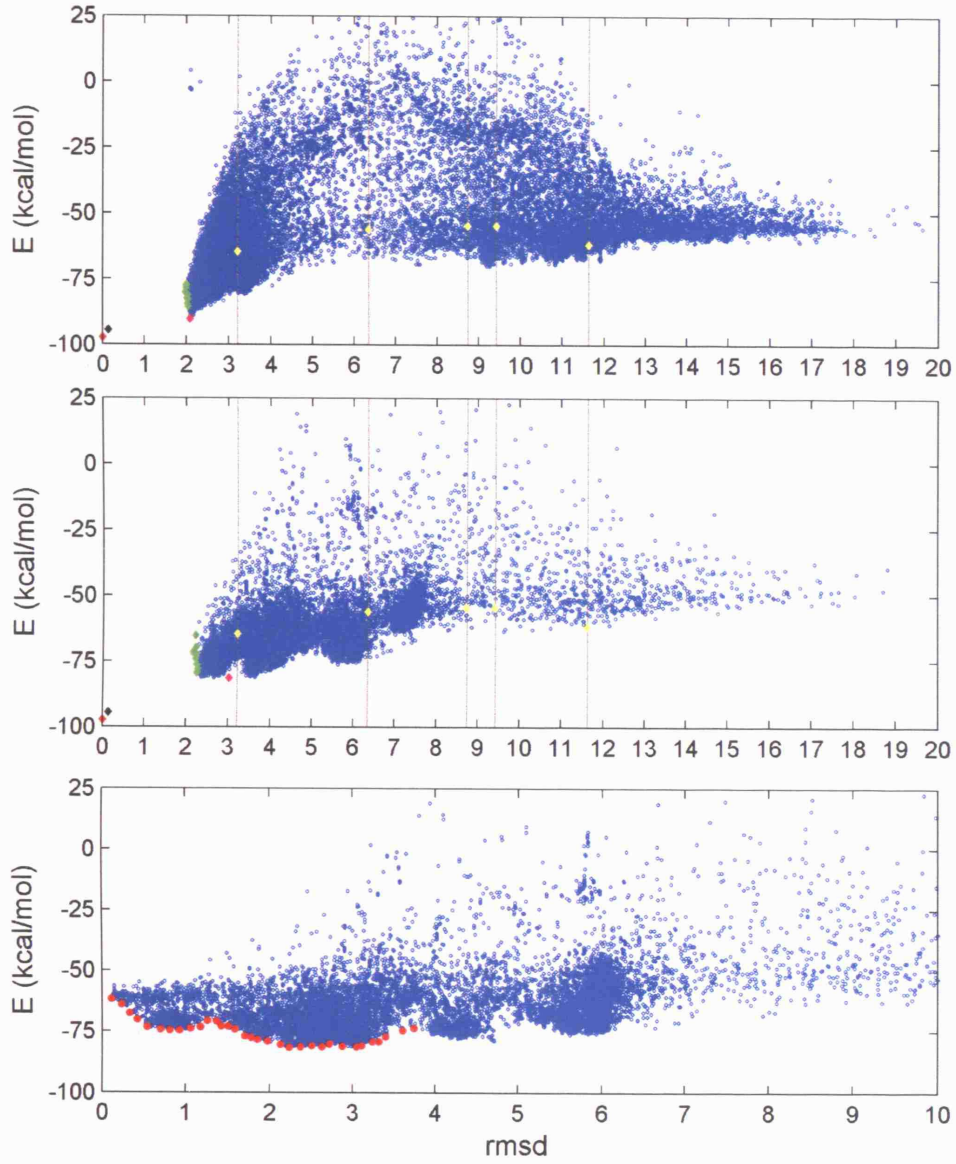


Figure 4.18: The upper and middle plots show the energy vs RMSD of 200,000 models sampled with the single-objective protocol and multi-objective protocol, respectively. The native structure (red) and regularized native (black) are shown, as well as the unrefined ROBETTA models (yellow). The vertical lines intersecting the x axis and the ROBETTA models highlighting the RMS distance of the models from the native, while the green points indicate the 10 lowest RMSD models sampled over the course of refinement. For single-objective optimisation, the lowest model is shown in magenta while for multi-objective optimization, the magenta point represents the lowest energy models chosen using knee selection. The bottom figure shows energy vs RMSD of each sampled model from the best ROBETTA template, and the Pareto-optimal set of solutions is highlighted in red.

The sampling and energetic properties of the populations during single-objective refinement are shown in Figure 4.19. A clear energy convergence takes place within 400 generations after a low RMSD/low energy structure drives the mean population RMSD towards the region of conformational space occupied by the low energy structures. The simulation terminates at ≈900 generations after population energy convergence. In contrast, the multi-objective simulation samples a low RMSD model early in the simulation though the minimum population RMSD rises over the course of the simulation (see Figure 4.20). While the single-objective GA converges rapidly, the mean population energy of the multi-objective refinement run does not converge until after 600 generations, and then, not fully.

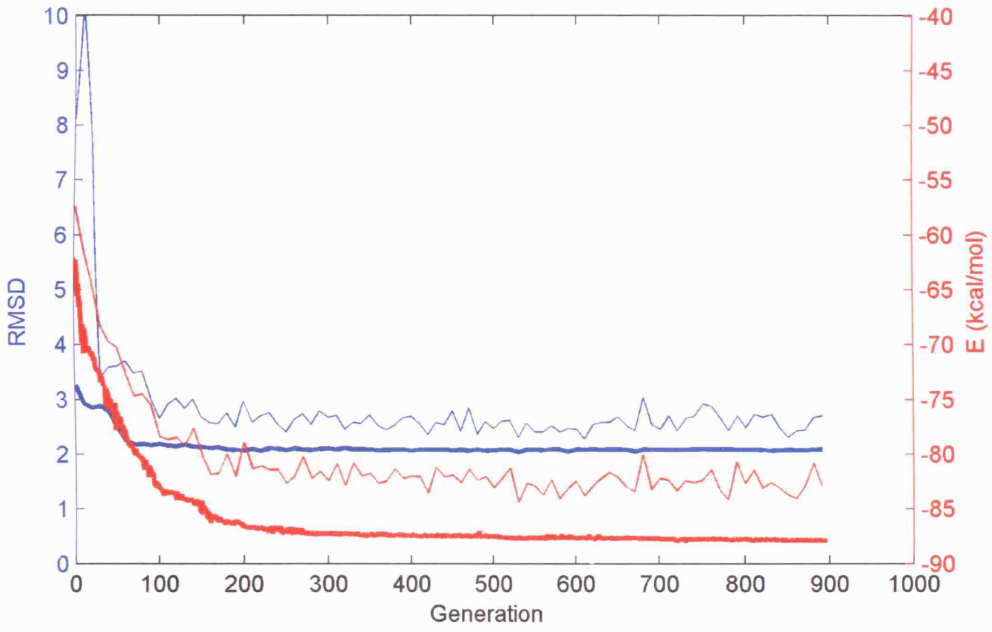


Figure 4.19: The change in mean population energy $\overline{E}_P$, and the mean population RMSD $\overline{P}_{rms}$, during the course of single-objective refinement of *CM/hard* target T0199_1 with a population size $N = 200$, maximum generation number $T = 1000$, crossover rate $P_c = 0.6$ and mutation rate $P_m = 0.3$. The mean population RMSD is shown in blue (min population RMSD with the thicker line width) while the mean population energy is shown in red (min population energy with thick line width). Data points are plotted every 10 generations.

Representatives from the refined low RMSD cluster are shown in Figure 4.21. Regions where models have been refined are labelled A-D and highlight the ability of the conformational search operators to improve both loop region and helical structures.

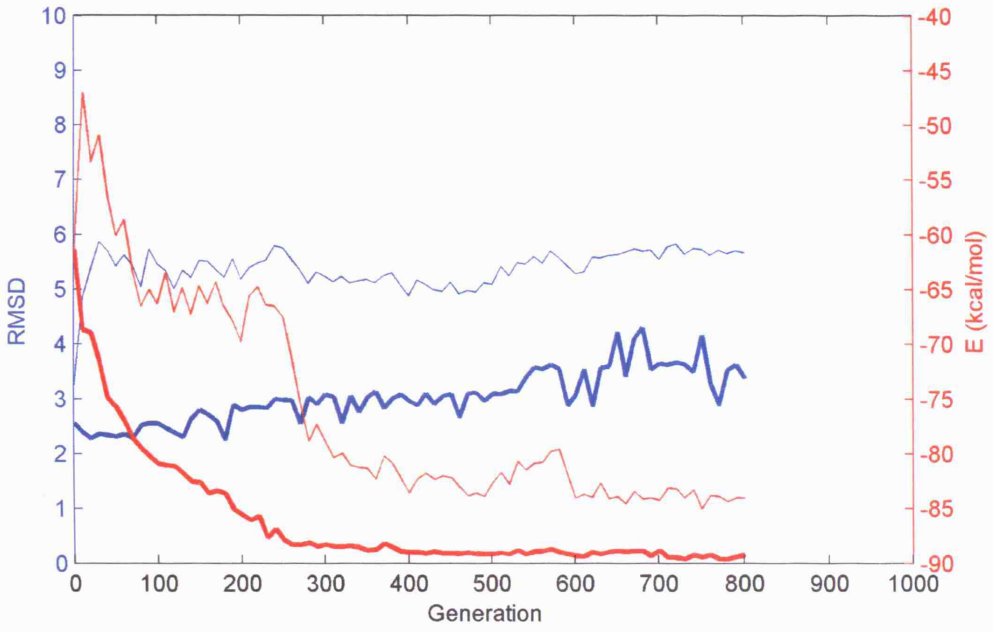


Figure 4.20: The change in mean population energy $\overline{E}_P$, and the mean population RMSD $\overline{P}_{rms}$, during the course of multi-objective refinement of *CM/hard* target T0199_1 with a population size $N = 200$, maximum generation number $T = 1000$, crossover rate $P_c = 0.6$, mutation rate $P_m = 0.3$, lower bound constraint $l = 0.5$, and upper bound constraint $u = 6\text{Å}$. The mean population RMSD is shown in blue (min population RMSD with the thicker line width) while the mean population energy is shown in red (min population energy with thick line width). Data points are plotted every 10 generations.

The best ROBETTA model at 3.21Å C_αRMSD (0.66 TM) from the native is improved by both the single-objective GA with the representative model improved to 1.98Å C_αRMSD (0.76 TM) and the multi-objective GA at 2.15Å C_αRMSD (0.72 TM). The lowest energy models show in Figure 4.22 are also significantly improved.

The lowest energy models for single- and multiple-objective refinement score 2.21Å C_αRMSD (0.73 TM) and 2.47Å C_αRMSD (0.69 TM), respectively. In both cases, the C-terminal helix of the template (label B in Figure 4.22) is re-orientated in the same direction as the native structure though the hydrogen-bonding network is not correctly formed. The lowest energy single-objective model also shows improvements in the small helix (label A) and shift towards a more native orientation in the N-terminal helix (label C).

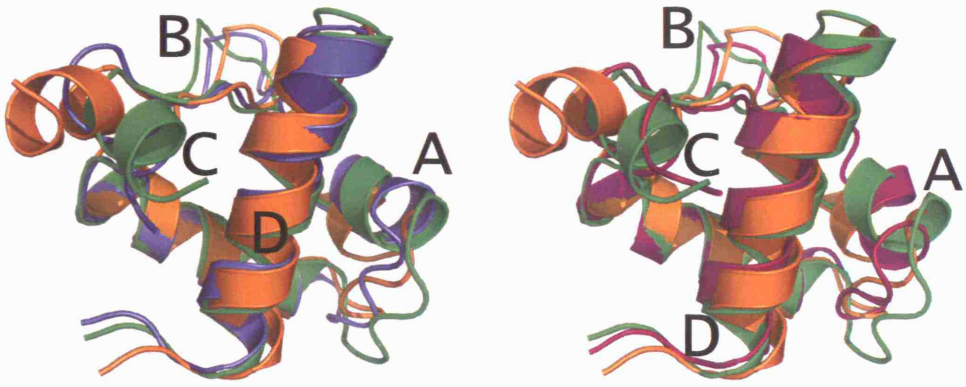


Figure 4.21: Representative refined models taken from the lowest RMSD cluster are shown for single-objective refinement (left) and multi-objective refinement (right). The native structures is shown in green with the best ROBETTA model (orange) and the refined model superposed. Sites with significant refinement are labelled A–D and highlight improvements in both helix orientations and loop conformations. The superpositions were calculated using the maximum likelihood method THESEUS (Theobald & Wuttke 2006).

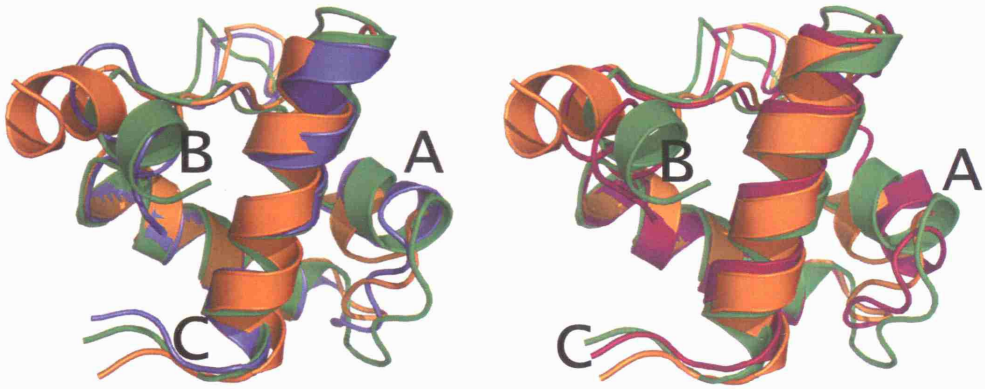
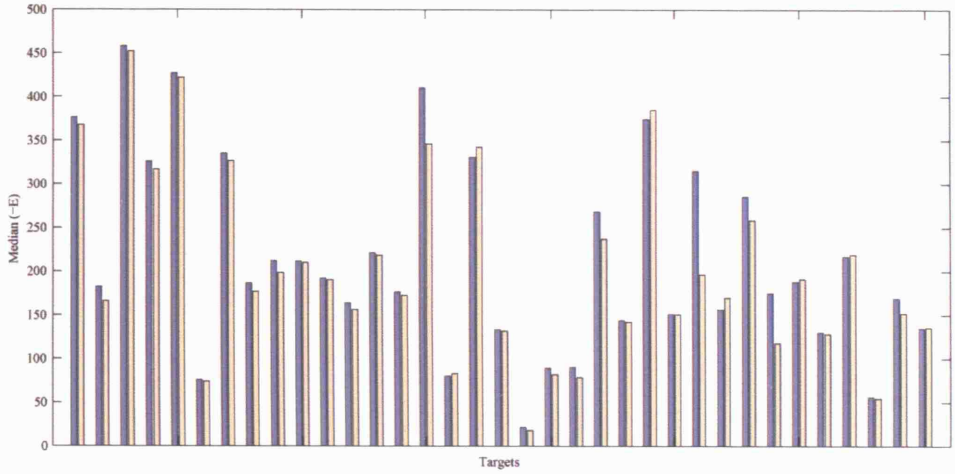


Figure 4.22: Representative refined models taken from the lowest energy cluster are shown for single-objective refinement (left) and lowest energy knee solution for multi-objective refinement (right). The native structures is shown in green with the best ROBETTA model (orange) and the refined model superposed. Sites with significant refinement are labelled A–D and highlight improvements in both helix orientations and loop conformations. The superpositions were calculated using the maximum likelihood method THESEUS (Theobald & Wuttke 2006).

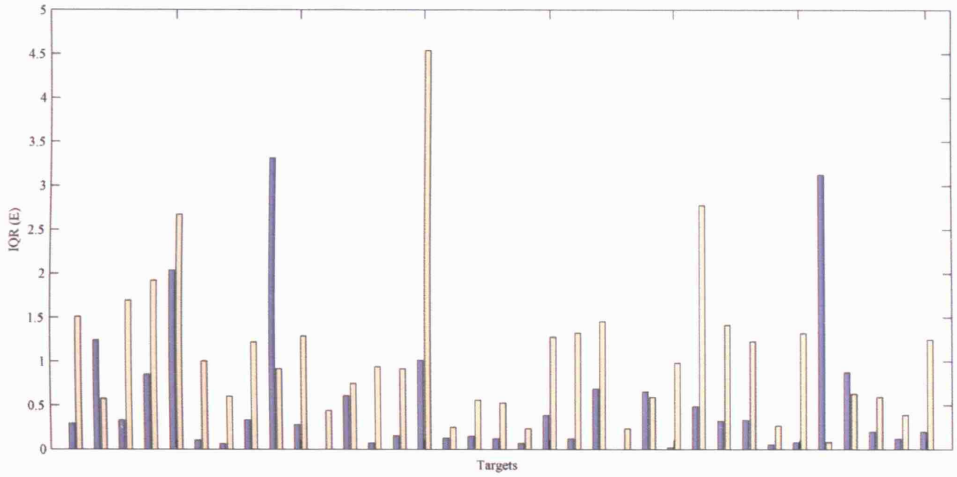
4.5.5 Exploration of the energy function

Inaccuracies in the energy function present significant problems for high-resolution structure refinement and can cause difficulties for any search procedure on a complex and rugged energy landscape. Here the ability of the single- and multi-objective GAs to explore the energy landscape is assessed by measuring the spread of energies (and their corresponding RMSD scores) sampled by each GA variant in a proportion of the low energy structures sampled during a simulation. For each target in the *CM/easy* and *CM/hard* data set, a single simulation was randomly selected from the total number performed, and the conformations generated during the refinement were used for the analysis. The top 10% of the lowest energy structures and their corresponding RMSD scores were used to calculate the median and interquartile range (IQR) value for each distribution.

Figure 4.23 shows the median energies of the low energy samples for each target (Figure 4.23a) and the interquartile range of the same samples (Figure 4.23b). The median values for the single- and multi-objective GAs are similar for most targets however, the interquartile range values are generally much larger for multi-objective optimization than single-objective optimization. When examining the RMSD of the low energy samples (Figure 4.24) a similar result is seen. The median RMSD of the samples is similar for both single- and multi-objective refinement simulations (Figure 4.24a) yet the IQR values are greater in the multi-objective case (Figure 4.24b).

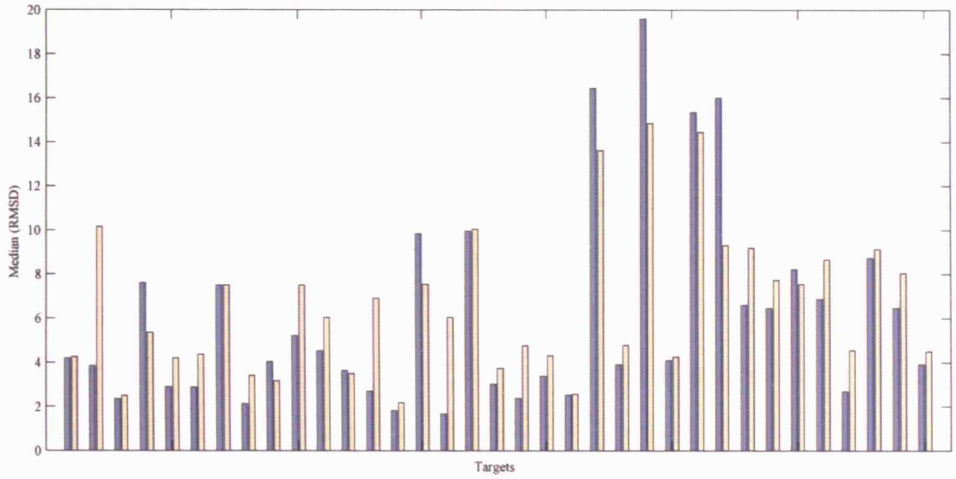


(a) Median energy of the top 10% lowest energy structures sampled during the refinement of CASP6 targets

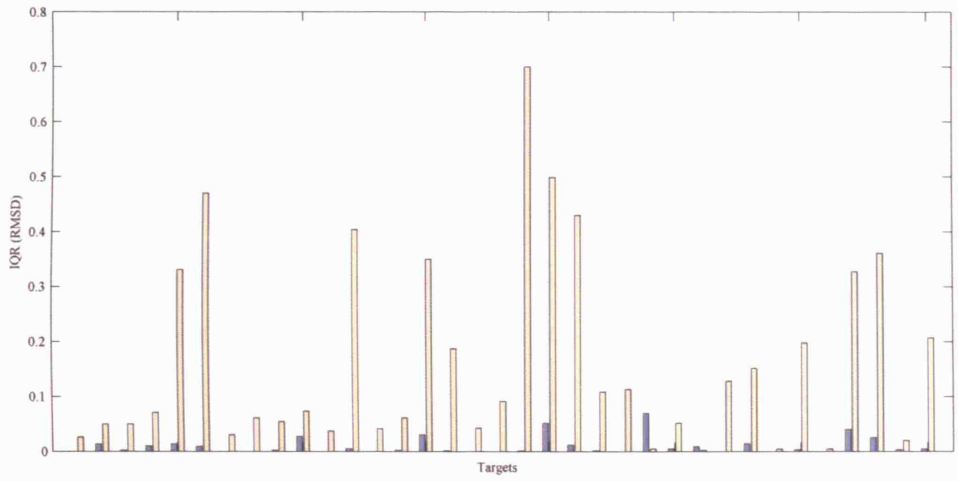


(b) Energy interquartile ranges of the top 10% lowest energy structures sampled during the refinement of CASP6 targets

Figure 4.23: Figure (a) shows the median energy score of the top 10% lowest energy structures sampled during a refinement simulation of each CASP6 target. Bars are coloured in blue and beige to represent single-objective and multi-objective results, respectively. Figure (b) show the interquartile range of the energies of the structures in the top 10% of lowest energy structures sampled during a refinement simulation of each CASP6 target.



(a) Median RMSD of the top 10% lowest energy structures sampled during the refinement of CASP6 targets



(b) RMSD interquartile ranges of the top 10% lowest energy structures sampled during the refinement of CASP6 targets

Figure 4.24: Figure (a) shows the median RMSD score of the top 10% lowest energy structures sampled during a refinement simulation of each CASP6 target. Bars are coloured in blue and beige to represent single-objective and multi-objective results, respectively. Figure (b) show the interquartile range of the RMSDs of structures in the top 10% of lowest energy structures sampled during a refinement simulation of each CASP6 target.

4.6 Discussion

This study has presented a novel approach to the high-resolution protein structure refinement problem using evolutionary algorithms. I have used two forms of evolutionary algorithm, the single-objective and multi-objective GA, and in doing so, have attempted to assess both the efficacy of these approaches to refinement, and the differences between the methodologies.

As a global optimization problem, protein structure prediction is conceptually suited to the single-objective paradigm (if viewed in simplified form) where the native state occupies the global free energy minimum for its sequence. Yet the search space defined by the energy landscape at high-resolution is sufficiently complex, and the conformational search space sufficiently large, that in reality there is a high risk of convergence on local minima under such circumstances. These problems are compounded further when the objective/energy function used to define the landscape, and in the case of heuristic search procedures like genetic algorithms, to guide a search through the landscape, contains inaccuracies. Under these circumstances, is it better to produce a single solution from a refinement simulation or instead to provide multiple solutions from which a decision maker can then choose after further assessment by additional criteria? The application of both single-objective and multi-objective GAs in this study has been, in part, an attempt to answer this question.

The refinements were conducted using an approach that was derived from the insights gained in Chapter 2. In this study, multiple predicted models were used to seed the first population of each genetic algorithm in order to increase the diversity of solutions available at the start of the refinement. In this study only five structures were used per target, providing the GA with five points on the energy landscape from which to begin exploration, though in practice any number of structures can be used to populate the first generation. Refinements were conducted using a common set of operators for manipulating the structural conformations during the simulations and which were complementary in their design to the mechanisms of the traditional genetic algorithm, i.e. crossover requires two solutions for the exchange of fragments which mutation operators affect only a single solution, though in many possible ways.

Multiple refinement simulations were conducted to take into account the stochastic nature of evolutionary algorithms and the statistics were reported using the combined

data. The results of single-objective refinement of the *CM/easy* targets showed that overall very few conformations were sampled that were nearer-native when the unrefined targets were already in 1–3Å from the native. Indeed, only in the cases where structures were less than 100 residues in length did the GA sample improved structures (Table 4.3). The model selected by energy was often of considerably lower energy than the most native-like conformations sampled, and in many cases, more distant from the native in terms of RMSD than the best starting structure; indicating a convergence on a low energy local minimum in the energy function. The single-objective GA performed much better on the *CM/hard* targets with substantially more native-like structures sampled for many of the targets, again with the smaller targets gaining the most improvement (Table 4.4). For these targets, some of structures selected by energy led to improvements over the best starting model though very few of the selected structures represented conformations from the nearest-native cluster (Figure 4.6). The single-objective refinements were, in effect, a test of the accuracy and efficacy of the energy function and sampling operators.

For multi-objective optimization the effects of introducing a second objective were carefully considered in the context of refinement. In previous refinement and modelling algorithms the use of structural constraints (obtained from either the atomic distances or torsion angles found in models or native structures) were used to restrict or constrain segments of a molecule to within some distance or angular threshold in the hope of reducing the search space available to the conformation (Skolnick, Kolinski & Ortiz 1997, Flohil et al. 2002, Kosinski et al. 2003, Zhang & Skolnick 2004*a*, Misura et al. 2006). However, when multiple structures are used, conflicting distance constraints can arise and lead to biases in a sampling procedure towards non-native regions of the energy landscape (Gront et al. 2005). There is no simple *a priori* method for assessing which constraints are inaccurate, so a global measure, the RMSD from the lowest energy unrefined structure, was used to allow conformational flexibility across the structure. Lower and upper limits were defined so that models too similar (< 0.5Å) or too distant (> 4Å for *CM/easy* targets or > 6Å for *CM/hard* targets) from the lowest energy unrefined structure were rejected. The additional objective, which scored a conformation as a function of its structural similarity to the lowest energy unrefined model, was then used in conjunction with the energy function to drive the

search towards regions of conformational space in which structures were less similar to the unrefined models but also lower in energy (with the bounds on the constraints set prior to refinement). Table 4.5 and Table 4.6 show the data for the multi-objective refinements of *CM/easy* and *CM/hard* targets, respectively. In the multi-objective case, a knee solution is also included in the analysis to assess the effects of selecting a single structure by an alternative method instead of the lowest energy model (which represents an extreme point in the Pareto-front under multi-objective refinement, i.e. a solution in which the energy is lowest at the expense of having a corresponding extreme value of the divergence objective).

The multi-objective GA again attained clusters of lower RMSD models (especially for *CM/hard* targets) but had less success in improving models by selecting with energy. This can be seen more clearly in Figure 4.7 where the change in RMSD represented by the best cluster of models and the selected knee solutions are shown. The multi-objective optimizer sampled near-native structures for many targets compared with the single-objective GA (Figure 4.6), however, selection by energy often led to a degradation in model quality.

To explore the reasons for these effects in more detail we then examined individual refinement cases. The energy plots, which show the RMSD and corresponding energies of conformations sampled during simulations, were used to visualise the exploration of the energy landscape (Figure 4.8, Figure 4.13, and Figure 4.18). Comparing the top and middle frames in each figure showed a similar pattern, with multiple troughs at various RMSD ranges in the multi-objective cases (middle frames) where areas of low energy structures were sampled. Indeed, the constraints enable the multi-objective GA to reject structures which are low energy but that also lie outside of the RMSD range defined by the constraints limits. This is shown clearing in the bottom frame of Figure 4.8 where the lowest energy region sampled at 6Å from the best unrefined model is rejected from the Pareto-set of solutions (shown in red) as it violates the upper constraint bound at 4Å. In contrast, the single-objective plots (top frames) lack these multiple troughs and instead show a more dense sampling of a single low energy region.

The contrast between the two GAs is also exhibited by analysing the population energies and RMSDs over the course of the simulations. The average population energies and RMSDs for single-objective refinements (Figure 4.9, Figure 4.14, and

Figure 4.19) showed a similar trend, with a decrease in average energy and RMSD in the first few generations followed by small fluctuations within a narrow energy and RMSD range. For multi-objective refinements (Figure 4.10, Figure 4.15, and Figure 4.20) much greater variation in behaviour was shown, often with a slower decrease in average population energy than the single-objective cases, with a corresponding increase in the average population RMSD over time. These three cases were chosen because their final refinement results were similar for the two types of GA, however, the mechanism by which these two GA variants attain those structures, indicated by both the energy plots and population properties over simulation time, was markedly different. These data suggested that there is a tendency for the multi-objective GA to explore more widely across the energy landscape than the single-objective GA which often converges rapidly, and this possibility was then explored by an analysis of the exploration of the energy function.

The quality of the energy function was a severe constraint on the ability of both GAs to refine structures at high-resolution. A true representation of the energy landscape is required both for guiding the search through conformational space and for discriminating native-like structures. While this study has shown that the conformational sampling method can sample nearer-native structures, the final selection of models by energy often degrades the quality of the model in relation to the best unrefined conformations. This was prohibitive for attaining the ultimate goal of structure refinement though the results of this study highlight other potentially interesting features of the adopted methodology.

The imperfect nature of the pairwise potential component of the energy function was known *a priori* from previous results on decoy discrimination tests (Zhou & Zhou 2002), and was further exhibited by the poor discrimination ability of the energy function in selecting refined models in this study. Moreover, in some cases, the native structure was not recognized as the lowest energy structure for its sequence after refinement (see Table 4.3, Table 4.4, Figure 4.8, and, Figure 4.13). We therefore chose to examine the ability of the two genetic algorithm variants to explore the energy function, and further, to then compare their performance.

The exploration of the energy function was analysed by measuring the coverage of low energy space during the refinement, both in terms of model energies (Figure

4.23) and the corresponding RMSDs of structures within the low energy space (Figure 4.24). Interestingly, the median energy of the lowest energy structures for each target was mostly similar for both single- and multi-objective runs (Figure 4.23a), however, the interquartile range is generally of greater magnitude when the multi-objective GA is used (Figure 4.23b). This is reflected in the RMSD scores for these models, where the median RMSD is again similar across targets for both GA variants (Figure 4.24a) yet the multi-objective GA most often displays a significantly greater interquartile range of RMSD scores than the single-objective GA (Figure 4.24b). In fact, the ability of the multi-objective GA to maintain a diverse set of solutions through the dual mechanisms of selection based on multiple criteria and the inbuilt niching function within the NSGA-II algorithm, enable a wider exploration of conformational space more generally, and in these cases, low energy space specifically.

In conclusion, what these results suggest is that given a more accurate energy function, the multi-objective GA (as implement here) would be a more suitable search strategy for exploring low energy space to increase the chances of sampling within the native basin. Moreover, the set of solutions which results from a multi-objective optimization can present a decision maker with multiple structures as a product of refinement that vary in quality according to the objectives used in contrast to a single model that results from single-objective optimization.

Chapter 5

Conclusions and Future Research

In this thesis I have presented three chapters which have individually examined some aspect of the protein structure prediction problem, and together formed a larger study of approaches to the protein structure model refinement problem. In Chapter 2 and Chapter 4 I have established the efficacy of multiple template modelling for improving low-resolution structures and explored the use of multi-objective optimization strategies in both high-resolution and low-resolution refinement, and in Chapter 3 I have benchmarked a number of model quality assessment programs leading to insightful conclusions about the use of structure to improve sequence-based homology modelling methods, and the likely unsuitability of these methods in high-resolution refinement procedures.

In Chapter 2, I demonstrate that multiple template-based models can be used to produce composite structures that are qualitatively better than any of the individual templates used to build the model. Using a multi-objective algorithm I show how optimizing different features of model quality can produces structures which are qualitatively different from each other by using alternative regions from a set of unrefined model. Although this study was conducted under ideal conditions, the results have shown that multiple template modelling is potentially useful for improving the accuracy of automated homology modelling servers. However, there are still open questions relating to the use of multiple models and/or multiple templates in both homology modelling and structure refinement that require further investigation. I have shown in Chapter 2 that the use of multiple structures built from alternative models can improve the quality of the predictions over and above the best starting structure, especially for low target-template sequence identities. However, a recent

study examining the automated modelling of "twilight-zone" targets found that the use of multiple templates rarely leads to a more accurate structure than the best possible single template at this level of identity (Dalton & Jackson 2007 (in press). This question warrants further investigation in light of the results of Chapter 2. In this regards, there are a number of possible extensions for the multi-objective genetic algorithm approach in Chapter 2; (*i*) the accuracy of the models may be further improved by combining the search through fragments space in structural models with a search through sequence-template alignment space. This could be encoded as a multi-objective problem and may lead to further improvements in the overall quality than the results found in this work, (*ii*) The GA (in its current form) could be extended to generate upper limit estimates for different combinations of templates, and the composite models that are produced then analysed to determine rules governing the optimal selection of the native-like regions in the template set that leads to the optimal model quality.

In Chapter 4, the refinement problem for high-resolution models was explored with both single-objective and multi-objective optimization approaches. A diverse set of conformational modification operators were devised and applied to all-atom structures while a high-resolution energy function guiding the search of both algorithms through the energy landscape. Some successful refinement cases were obtained, even for models within the native basin ($\sim 1\mathring{A}$), though the failure of both algorithms on many of the cases highlights the deficiency of the energy function at this modelling resolution. In terms of searching the energy landscape, a significant difference between the single- and multi-objective GA was shown. Although implicit in the nature of the multi-objective GA, the sampling of the energy function produced solutions that were often more diverse in terms of both energy and RMSD from the native basin during the course of the simulation than the single-objective counterpart. This conclusion has a number of implications for future research and suggests that a multi-objective approach, in addition to providing a set of solutions to a refinement problem rather than a single model, also has the potential to sample more widely on the potential energy surface, and this alone can increase the probability of finding the native basin.

A number of weaknesses and areas for further exploration were highlighted during this research. The limitations of the side-chain modelling algorithm SCWRL were found when attempting to reconstruct the side-chains on both native structures and

regularized natives and models. Only ≈60% of native χ_1 angles were reproduced and this had a detrimental effect on the energies of the re-packed structures. In the context of multi-objective optimization, the divergence function measures the distance (in terms of RMSD) from a single (lowest energy) starting model. This restricts the amount of structural information used in calculating the objective value and leaves the structural information provided by the remaining input models for use by the conformational sampling operators. A more complex divergence function, which utilizes the full set of models may help direct the search more widely on the energy landscape. Finally, the upper and lower constraint bounds used in the multi-objective GA are selected arbitrarily and may restrict the sampling in some refinement cases. For example, if a set of unrefined models for the *CM/easy* category lie at $\sim 5\AA$ from the native, the upper bound constraint at 4Å will prevent sampling in the final 1Å range. A more systematic analysis of the RMSD ranges of template-based models at different sequence identities may be beneficial for setting constraint bounds for specific cases.

In conclusion, this study has contributed to furthering the goal of achieving high-resolution protein structure models and helped advance knowledge in the area of protein structure prediction.

Appendix

Appendix A - Multi-Objective Optimization

Multi-objective assessment

The general goal of multi-objective optimization is to produce an approximate set of solutions that is as close as possible to the optimal Pareto-front and which holds a wide range of diverse solutions as a subset of the complete Pareto-optimal front. In general it is unlikely that a multi-objective optimization will lead to the true Pareto-optimal front and therefore we require a means of stating how good an approximation a non-dominated set of solutions is either relative to another non-dominated set or with reference to the true Pareto-optimal front. An approximation set can be defined as follows:

Definition 5.1 (Approximation Set) *Let $A \subseteq Z$ be a set of objective vectors. A is called an approximation set if any element of A does not dominate or is not equal to any other objective vector in A. The set of all approximation sets is denoted* Ω.

Importantly, this statement does not comprise any notion of *quality*. However, we can define criteria with which to state how good an approximation set is in comparison to other approximation sets in a quantitative manner using quality indicators.

Quality Indicators

A quality indicator is a method for mapping approximation sets to the set of real numbers so that qualitative and quantitative differences between approximate sets can be summarized using a real representation. Although this leads to a loss of information in the dimensionality reduction, these metrics can provide useful means for performing

analyses, equivalent to uni-variate statistical tests, so that precise statements about the quality of approximation sets can be made. A quality indicator can be defined as

Definition 5.2 (Quality Indicator) *An m-ary quality indicator I is a function* $I : \Omega^m \mapsto \Re$, *which assigns each vector* $(A_1, A_2, ...A_m)$ *of m approximation sets a real value* $I(A_1, A_2, ..., A_m)$.

A description of a unary quality indicator, the hypervolume indicator I_H, follows, though other unary, binary, and higher order quality indicators have been presented in the literature (Zitzler, Thiele, Laumanns, Fonseca & Fonseca 2002).

Hypervolume Indicator ($\mathcal{S}$ metric)

The hypervolume indicator, or $\mathcal{S}$ metric, is a unary indicator, defined by Zitzler & Thiele (1998*b*) as the size of space covered by the approximation set A from a given reference point in n-dimensional objective space.

Let $\boldsymbol{X'} = \{\boldsymbol{X_1}, \boldsymbol{X_1}, ..., \boldsymbol{X_k}\} \subseteq \boldsymbol{X}$ be a set of k decision vectors. The function $\mathcal{S}(\boldsymbol{X'})$ gives the volume enclosed by the union of polytopes $p_1, p_2, ... p_k$, where each p_i is formed by the intersection of the following hyperplanes arising out of $\boldsymbol{x_i}$, along with the axes: for each axis in objective space, there exists a hyperplane perpendicular to the axis and passing through the point $(f_1(\boldsymbol{x_i}), f_1(\boldsymbol{x_i}), ..., f_n(\boldsymbol{x_i}))$. In the two dimensional case, each p_i represents a rectangle defined by points $(0, 0)$ and $(f_1(\boldsymbol{x_i}), f_2(\boldsymbol{x_i}))$.
An illustration of the hypervolume indicator can be seen in Figure 5.1.
It is important to note that an efficient calculation is only possible in 2-dimensions and becomes exceedingly difficult in higher dimensional objective spaces.

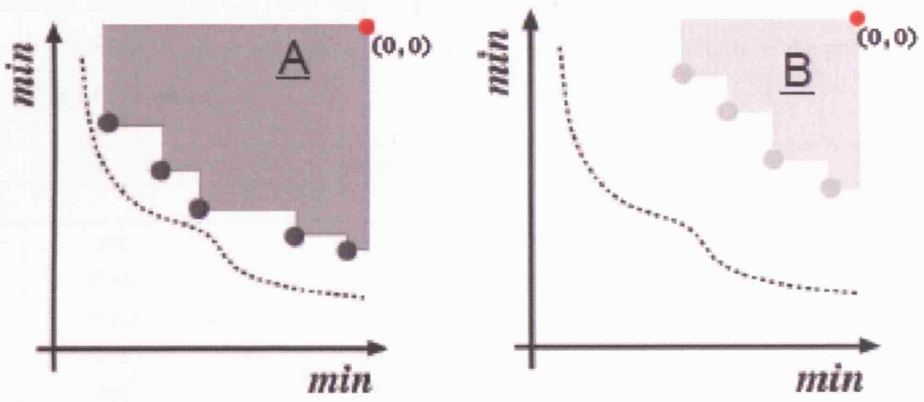


Figure 5.1: The hypervolume area covered by two approximation sets is shown. The leftmost figure shows a large hypervolume coverage by set A with the hypervolume derived from a reference point at (0,0). The rightmost figure shows a second approximation set B with smaller hypervolume area from the same reference point (0,0). In this illustrative case the $\mathcal{S}$ metric $\mathcal{S}(A) = 70\%$ while $\mathcal{S}(B) = 30\%$.

Appendix B - Data sets

Data Sets (Chapter 2)

Table 5.1: Fifty proteins from the LiveBench-9 target list used for refinement are shown. The first 35 structures are classified as *CM/easy* targets while the final 15 are from the *CM/hard* category. The experimental technique used to obtain the structure, the resolution, the length of the protein, and the protein name are all provided.

pdb	technique	res. (Å)	length	protein name
CM/easy				
1j26	nmr	-	112	Peptidyl-Trna Hydrolase
1j3v	X-ray	1.50	289	Gliding Protein-Mglb
1nng	X-ray	1.95	141	Acyl-Coa Thioester Hydrolase HI0827
1nrk	X-ray	2.80	328	Structural genomics UKN (*e.coli*)
1op4	nmr	-	159	Neural-cadherin
1p91	X-ray	2.80	269	ribosomal rna large subunit methyltransferase
1p97	nmr	-	114	Endothelial pas domain protein
1p9e	X-ray	2.40	331	Methyl parathion hydrolase
1pmm	X-ray	2.00	466	Glutamate decarboxylase beta
1pvm	X-ray	1.50	184	Hypothetical protein TA0289 *thermoplasma acidophilum*
1q9j	X-ray	2.75	422	Polyketide synthase associated protein 5
1qwr	X-ray	1.80	319	Mannose-6-phosphate isomerase
1qxm	X-ray	1.70	300	Hemagglutinin component HA1
1r1d	X-ray	2.00	247	Carboxylesterase
1r4w	X-ray	2.50	226	Mitochondrial Glutathione S-transferase
1r57	nmr	-	102	Hypothetical protein (*staphylococcus aureus*)
1ri1	X-ray	2.50	298	Mrna capping enzyme
1rku	X-ray	1.47	206	Homoserine kinase
1rli	X-ray	2.80	184	TRP repressor binding protein
1rxx	X-ray	2.45	421	Arginine Deiminase
1rz3	X-ray	1.90	201	Hypothetical protein (*bacillus stearothermophilus*)
1s5a	X-ray	1.70	150	Hypothetical protein (*bacillus subtilis*)
1sqh	X-ray	2.00	312	Hypothetical protein (*drosophila melanogaster*)
1sr4	X-ray	2.00	261	Cytolethal distending toxin
1t35	X-ray	2.72	191	Hypothetical protein putative lysine decarboxylase
1te5	X-ray	2.00	257	Glutamine amidotransferase
1ub9	X-ray	2.05	100	Putative transcriptional regulator homologue

pdb	technique	res. (Å)	length	protein name
1uhw	nmr	-	109	Dep domain of mouse pleckstrin
1ujx	nmr	-	119	Polynucleotide kinase 3'-phosphatase
1v4e	X-ray	2.28	299	Octoprenyl-diphosphate synthase
1v63	nmr	-	101	Nucleolar transcription factor 1
1v8c	X-ray	1.60	168	MOAD related protein
1vhe	X-ray	1.90	373	Aminopeptidase/glucanase Homologue
1vjg	X-ray	2.01	218	hypothetical protein putative lipase
1vkh	X-ray	1.85	273	hypothetical protein putative serine hydrolase
CM/hard				
1hl8	X-ray	2.40	449	Alpha-L-Fucosidase
1j3w	X-ray	1.50	163	Gliding protein-mglb
1n8n	X-ray	1.95	212	APHA Class B Acid Phosphatase
1nmo	X-ray	2.20	247	Hypothetical protein YBGI (e.coli)
1nnv	nmr	-	110	Hypothetical protein HI1450
1oap	X-ray	1.93	109	Peptidoglycan-associated lipoprotein
1paq	X-ray	2.30	189	Translation intiation factor EIF-2B epsilon
1pfj	nmr	-	108	TFIIH basal transcription factor subunit
1psy	nmr	-	125	Ricinus communis 2s albumin
1q0d	X-ray	2.20	117	NI-containing superoxide dismutase
1se9	nmr	-	126	Ubiquitin family
1uww	X-ray	1.40	191	Endo-1,4-beta-glucanase, alkaline cellulase
1v2y	nmr	-	105	Hypothetical protein ubiquitin-like fold
1v32	nmr	-	101	hypothetical protein AT5G08430
1vjx	X-ray	2.30	273	hypothetical protein putative ferritin-like diiron-carboxylate

Data Sets (Chapter 4)

Table 5.2: The CASP6 targets used for testing the refinement procedures are shown. Only single domain targets or domains consisting of consecutive residues taken from a multimeric protein are included. Targets are taken from the *CM/easy* (19) and *CM/hard* (16) template-modelling categories giving 35 refinement targets in total.

Target/domain	Nres†	Domain	Exp.	Description
CM/easy				
T0204	297	-	X-ray	At5g18200, Arabidopsis
T0229_1	24	001 - 024	X-ray	TM0919, T. maritima
T0229_2	102	037 - 138	X-ray	TM0919, T. maritima
T0231	137	-	X-ray	Glia maturation factor gamma, Mouse
T0233_1	66	014 - 079	X-ray	Anthranilate phosphoribosyltransferase 2, Nostoc sp. pcc 7120
T0240	90	-	X-ray	tonb, E. coli
T0244	296	-	X-ray	galu, E. coli
T0246	354	-	X-ray	3-isopropylmalate dehydrogenase, T. maritima
T0247_3	76	279 - 354	X-ray	aminomethyltransferase, E. coli
T0264_1	116	006 - 121	X-ray	Probable diphtine synthase APE0931, A. pernix
T0266	150	-	X-ray	Hypothetical protein APE2540, A. pernix
T0268_2	109	099 - 297	X-ray	mraW protein, T. thermophilus
T0269_1	158	002 - 159	X-ray	thioredoxin peroxidase, A. pernix
T0271	161	-	X-ray	Hypothetical conserved protein, T. thermophilus
T0274	156	-	X-ray	Probable nitrilotriacetate monooxygenase component B, T. thermophilus
T0275	135	-	X-ray	Hypothetical conserved protein, T. thermophilus
T0276	168	-	X-ray	Conserved hypothetical protein, T. thermophilus
T0277	117	-	X-ray	Probable nucleotidyltransferase, T. thermophilus
T0282	323	-	X-ray	Formiminoglutamase, Vibrio cholerae O1 biovar eltor str.
CM/hard				
T0196	89	-	X-ray	Hypothetical protein, P. furiosus
T0199_1	74	014 - 087	X-ray	Heat shock operon repressor HrcA, T. maritima
T0200	255	-	X-ray	Conserved hypothetical protein, D. radiodurans
T0205	103	-	X-ray	At2g34160, Arabidopsis
T0208	344	-	X-ray	EFR41, E.faecalis
T0211	136	-	X-ray	HR1958, Human
T0222_1	264	002 - 274	X-ray	Spore coat polysaccharide biosynthesis protein spsE, B.
T0223_1	114	001 - 113	X-ray	Putative Nitroreductase, T. maritima
T0232_1	81	006 - 086	X-ray	Atu5508, A. tumefaciens
T0232_2	146	091 - 236	X-ray	Atu5508, A. tumefaciens
T0234	135	-	X-ray	Alr5027, Nostoc sp. pcc 7120
T0264_2	173	122 - 294	X-ray	Probable diphtine synthase APE0931, A. pernix
T0265	102	-	X-ray	Hypothetical transcriptional regulator, S. tokodaii
T0267	174	-	X-ray	Acetyltransferase, T. thermophilus
T0269_2	61	160 - 220	X-ray	thioredoxin peroxidase, A. pernix
T0279_2	121	043 - 163	X-ray	Uroporphyrinogen-III synthase, T. thermophilus

† Number of modelled residues

Table 5.3: The structural similarity of the original CASP6 ROBETTA models are shown prior to regularization and refinement. The structural similarity of each model to the experimental structure is measured using the C_αRMSD and the sequence-dependent heuristic methods, the TM-score.

Targets	Nres	*RMSD (TM)*				
		Model 1	Model 2	Model 3	Model 4	Model 5
CM/easy						
T0204	297	3.94 (0.72)	4.63 (0.71)	4.68 (0.73)	5.36 (0.71)	3.82 (0.73)
T0229_1	102	8.07 (0.66)	2.91 (0.81)	3.77 (0.74)	4.23 (0.79)	6.99 (0.66)
T0229_2	24	1.96 (0.36)	1.04 (0.58)	3.09 (0.35)	2.14 (0.55)	3.80 (0.22)
T0231	137	10.45 (0.74)	7.77 (0.75)	3.99 (0.78)	8.36 (0.76)	7.97 (0.74)
T0233_1	66	1.53 (0.83)	1.43 (0.84)	1.43 (0.84)	1.43 (0.84)	1.47 (0.83)
T0240	90	20.95 (0.33)	20.25 (0.33)	20.80 (0.33)	20.98 (0.33)	21.00 (0.35)
T0244	296	6.86 (0.74)	8.09 (0.71)	8.51 (0.70)	5.60 (0.76)	7.10 (0.74)
T0246	354	2.06 (0.94)	2.13 (0.94)	2.49 (0.93)	2.18 (0.94)	1.93 (0.94)
T0247_3	76	4.04 (0.84)	2.99 (0.82)	2.89 (0.84)	3.75 (0.82)	4.23 (0.81)
T0264	289	5.85 (0.75)	6.27 (0.73)	7.30 (0.76)	8.09 (0.71)	6.70 (0.72)
T0266	150	1.96 (0.87)	1.81 (0.88)	1.80 (0.88)	2.34 (0.85)	2.03 (0.86)
T0268_2	109	2.52 (0.90)	2.66 (0.90)	2.96 (0.87)	2.82 (0.87)	2.89 (0.88)
T0269_1	158	2.60 (0.88)	4.17 (0.85)	3.36 (0.86)	3.86 (0.87)	2.86 (0.87)
T0271	161	4.99 (0.79)	4.98 (0.80)	5.61 (0.78)	4.76 (0.80)	3.28 (0.83)
T0274	156	4.38 (0.80)	3.72 (0.81)	4.31 (0.81)	3.62 (0.80)	4.52 (0.81)
T0275	135	2.82 (0.82)	3.58 (0.72)	2.65 (0.82)	3.26 (0.80)	3.68 (0.72)
T0276	168	2.60 (0.76)	2.84 (0.75)	2.56 (0.77)	3.07 (0.74)	3.26 (0.73)
T0277	117	1.58 (0.88)	1.60 (0.88)	2.14 (0.84)	1.60 (0.88)	1.98 (0.86)
T0282	323	10.72 (0.75)	9.50 (0.77)	6.62 (0.75)	9.84 (0.75)	6.87 (0.73)
CM/hard						
T0196	89	4.58 (0.74)	4.63 (0.70)	4.61 (0.73)	4.91 (0.70)	4.87 (0.71)
T0199_1	74	8.74 (0.62)	3.21 (0.66)	9.42 (0.59)	6.36 (0.58)	11.63 (0.34)
T0200	255	14.62 (0.51)	11.22 (0.59)	16.69 (0.57)	15.72 (0.55)	14.71 (0.54)
T0205	103	3.76 (0.54)	9.13 (0.60)	8.13 (0.39)	8.26 (0.39)	7.93 (0.37)
T0208	344	14.97 (0.41)	17.31 (0.42)	17.34 (0.42)	19.77 (0.41)	15.74 (0.38)
T0211	136	5.11 (0.66)	6.68 (0.59)	4.27 (0.65)	5.92 (0.63)	5.85 (0.56)
T0222_1	264	14.46 (0.30)	14.66 (0.29)	14.44 (0.31)	13.91 (0.34)	14.31 (0.30)
T0223_1	114	16.20 (0.30)	9.06 (0.69)	17.38 (0.29)	16.77 (0.30)	17.36 (0.23)
T0232_1	81	3.62 (0.71)	3.46 (0.73)	3.51 (0.72)	3.47 (0.72)	3.63 (0.71)
T0232_2	146	7.54 (0.61)	10.89 (0.46)	7.51 (0.61)	7.41 (0.58)	7.83 (0.43)
T0234	135	6.56 (0.55)	7.18 (0.54)	6.82 (0.57)	6.17 (0.57)	6.46 (0.56)
T0264_2	173	6.59 (0.65)	7.08 (0.61)	8.96 (0.64)	9.62 (0.57)	8.21 (0.58)
T0265	102	6.99 (0.54)	11.51 (0.46)	8.76 (0.52)	7.83 (0.54)	6.44 (0.49)
T0267	174	4.25 (0.80)	2.95 (0.82)	3.69 (0.78)	3.82 (0.82)	2.48 (0.85)
T0269_2	61	9.22 (0.35)	9.81 (0.35)	7.72 (0.35)	10.22 (0.37)	11.21 (0.34)
T0279_2	121	2.77 (0.74)	3.11 (0.73)	2.98 (0.73)	3.18 (0.72)	3.26 (0.72)

Table 5.4: The combined energies of the ROBETTA models are given for each of the CASP6 targets grouped by their CASP6 category (*CM/easy* and *CM/hard*). The model ordering (1-5) is the same order as provided by the ROBETTA server submissions. Two energies are shown; the energy of original ROBETTA model, and the energy of the ROBETTA model after the backbone has been regularized and side-chains re-optimized using a rotamer library.

Target	Energy, E (Regularized Energy, E_r)				
	Model 1	Model 2	Model 3	Model 4	Model 5
CM/easy					
T0204	-377.18 (-320.32)	-361.10 (-320.28)	-375.42 (-330.81)	-333.62 (-271.83)	-397.22 (-361.08)
T0229_1	-121.77 (-109.11)	-124.20 (-115.12)	-105.01 (-88.97)	-128.39 (-117.29)	-125.19 (-111.01)
T0229_2	-15.15 (-14.92)	-15.58 (-14.34)	-12.41 (13.43)	-14.12 (-13.50)	-9.90 (-9.50)
T0231	-157.49 (-147.68)	-163.91 (-154.25)	-175.96 (-168.00)	-160.16 (-146.46)	-160.94 (-151.71)
T0233_1	-76.71 (-72.02)	-75.81 (-71.33)	-76.02 (-71.91)	-75.86 (-71.43)	-74.88 (-68.71)
T0240	-74.17 (-69.08)	-70.43 (-66.42)	-71.54 (-64.86)	-72.79 (-64.83)	-76.00 (-66.91)
T0244	-326.43 (-289.24)	-301.55 (-279.34)	-321.36 (-283.21)	-327.65 (-298.53)	-299.98 (-270.83)
T0246	-472.14 (-424.05)	-472.11 (-430.82)	-455.87 (-411.58)	-472.44 (-407.80)	-462.29 (-419.36)
T0247_3	-61.41 (-59.60)	-56.27 (-51.15)	-67.84 (-64.12)	-48.86 (-46.59)	-50.97 (-48.97)
T0264	-318.56 (-294.80)	-299.45 (-278.50)	-281.68 (-238.05)	-325.36 (-301.10)	-329.64 (-304.27)
T0266	-192.72 (-182.19)	-191.43 (-181.43)	-184.94 (-165.90)	-190.42 (-173.67)	-191.63 (-181.46)
T0268_2	-329.04 (-306.51)	-355.00 (-333.56)	-327.11 (-303.14)	-349.32 (-319.10)	-341.61 (-316.70)
T0269_1	-210.53 (-208.19)	-215.99 (-209.88)	-210.31 (-206.80)	-208.95 (-206.14)	-208.93 (-204.73)
T0271	-210.87 (-200.20)	-203.26 (-195.41)	-205.60 (-192.88)	-211.38 (-199.85)	-210.58 (-190.99)
T0274	-183.90 (-157.46)	-193.19 (-169.12)	-178.72 (-146.61)	-179.33 (-155.58)	-182.83 (-150.83)
T0275	-156.33 (-145.23)	-158.99 (-146.82)	-157.81 (-145.83)	-160.26 (-151.15)	-165.83 (-152.73)
T0276	-211.91 (-196.11)	-205.22 (-188.26)	-217.96 (-199.08)	-204.62 (-177.33)	-220.89 (-199.49)
T0277	-178.14 (-163.30)	-176.28 (-164.83)	-169.68 (-150.98)	-178.74 (-167.76)	-173.46 (-160.18)
T0282	-382.66 (-338.15)	-403.48 (-357.35)	-393.13 (-339.80)	-388.52 (-340.18)	-364.65 (-328.78)
CM/hard					
T0196	-88.35 (-69.88)	-85.21 (-79.63)	-86.94 (-82.50)	-81.92 (-73.37)	-84.92 (-78.82)
T0199_1	-54.56 (-54.31)	-64.45 (-60.76)	-54.66 (-53.39)	-56.20 (-54.11)	-61.67 (-60.56)
T0200	-228.47 (-205.55)	-218.63 (-182.06)	-265.33 (-241.00)	-258.25 (-224.84)	-263.77 (-237.00)
T0205	-128.53 (-121.39)	-121.67 (-111.51)	-122.45 (-107.76)	-127.12 (-114.73)	-129.47 (-108.49)
T0208	-353.24 (-308.07)	-364.03 (-318.23)	-342.26 (-291.12)	-361.76 (-315.25)	-301.39 (-246.69)
T0211	-95.36 (-60.45)	-96.73 (-75.01)	-150.57 (-126.05)	-134.93 (-102.70)	-118.55 (-109.99)
T0222_1	-226.74 (-210.98)	-271.42 (-242.81)	-257.73 (-244.31)	-226.06 (-183.75)	-263.87 (-233.01)
T0223_1	-100.37 (-85.59)	-108.08 (-85.77)	-97.81 (-61.70)	-83.77 (-73.90)	-94.35 (-93.78)
T0232_1	-84.60 (-78.68)	-86.72 (-81.17)	-83.70 (-76.35)	-83.23 (-78.65)	-80.88 (-72.08)
T0232_2	-137.09 (-122.83)	-130.66 (-117.68)	-156.62 (-148.17)	-136.38 (-124.15)	-145.57 (-137.66)
T0234	-129.53 (-99.60)	-148.53 (-113.82)	-135.92 (-117.53)	-161.69 (-147.90)	-151.71 (-138.76)
T0264	-184.11 (-174.91)	-156.93 (-141.49)	-140.90 (-113.17)	-174.18 (-158.59)	-182.72 (-166.76)
T0265	-113.87 (-28.91)	-106.99 (-80.45)	-119.23 (-63.47)	-124.26 (-122.70)	-119.67 (-112.60)
T0267	-211.30 (-193.89)	-214.50 (-198.40)	-201.72 (-183.03)	-208.48 (-191.53)	-215.84 (-201.94)
T0269	-35.23 (-30.32)	-39.55 (-42.49)	-21.78 (-17.73)	-33.39 (-38.27)	-15.44 (-9.87)
T0279_2	-127.43 (-116.78)	-121.28 (-112.03)	-129.83 (-114.62)	-132.99 (-125.81)	-129.18 (-117.06)

Appendix C - Publications arising from this work

Pettitt, C.S., McGuffin, L.J. & Jones, D.T. (2005) Improving protein structure prediction using 3D model quality. *Bioinformatics*.**21**(17) 3509-3515.

References

Aloy, P., Stark, A., Hadley, C. & Russell, R. B. (2003), 'Prediction without templates: new folds, secondary structure, and contacts in CASP5', *Prot. Struct. Funct. & Bioinf.* **53**(S6), 436–456.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990), 'Basic local alignment search tool', *J. Mol. Biol.* **215**, 403–410.

Altschul, S. F., Madden, T. L., Schaffer, 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 Res.* **25**, 3389–3402.

Anfinsen, C. B. (1973), 'Principles that govern the folding of protein chains', *Science* **181**(96), 223–230.

Antonisse, H. J. & Keller, K. S. (1987), 'Genetic operators for high-level knowledge representations', *Proceedings of the Second International Conference on Genetic Algorithms on Genetic algorithms and their application table of contents* pp. 69–76.

Back, T., Fogel, D. & Michalewicz, Z. (1997), *Handbook of evolutionary computation*, Oxford University Press, Bristol Institute of Physics.

Bahar, I. & Jernigan, R. L. (1997), 'Inter-residue potentials in globular proteins and the dominance of highly specific hydrophilic interactions at close separation', *J. Mol. Biol.* **266**(1), 195–214.

Baker, D. & Šali, A. (2001), 'Protein structure prediction and structural genomics', *Science* **294**(5540), 93–96.

Baker, J. E. (1985), Adaptive selection methods for genetic algorithms, *in* J. J. Grefenstette, ed., 'Proceedings of the 1st International Conference on Genetic Algorithms', Lawrence Erlbaum Associates, pp. 101–111.

Baldwin, R. L. (1989), 'How does protein folding get started?', *TIBS* **14**(7), 291–294.

Becker, O. M., MacKerell Jr, A. D., Roux, B. & Watanabe, M., eds (2001), *Computational Biochemistry and Biophysics*, 1 edn, CRC Press, chapter Atomistic Models and Force Fields, pp. 7–38.

Ben-Naim, A. (1997), 'Statistical potentials extracted from protein structures: Are these meaningful potentials?', *Journal of Chemical Physics* **107**(9), 3698–3706.

Berg, J. M., Tymoczko, J. L. & Stryer, L. (2002), *Biochemistry*, WH Freeman.

Betancourt, M. R. & Skolnick, J. (2001), 'Universal similarity measure for comparing protein structures', *Biopolymers* **59**, 305–309.

Blundell, T. L., Sibanda, B. L., Sternberg, M. J. E. & Thornton, J. M. (1987), 'Knowledge based prediction of protein structures and the design of novel molecules', *Nature* **326**, 347–352.

Bonneau, R. & Baker, D. (2001), 'Ab initio protein structure prediction: progress and prospects.', *Annu Rev Biophys Biomolec Struct* **30**, 173–179.

Bonneau, R., Tsai, J., Ruczinski, I., Chivian, D., Rohl, C. A., Strauss, C. E. M. & Baker, D. (2001), 'Rosetta in CASP4: Progress in ab initio protein structure prediction', *Prot. Struct. Funct. & Bioinf.* **45**(S5), 119–126.

Bower, M. J., Cohen, F. E. & Dunbrack, R. L. J. (1997), 'Prediction of protein side-chain rotamers from a backbone-dependent rotamer library: a new homology modeling tool', *J. Mol. Biol.* **267**(5), 1268–1282.

Bowie, J. U., Clarke, N. D., Pabo, C. O. & Sauer, R. T. (1990), 'Identification of protein folds - matching hydrophobicity patterns of sequence sets with solvent accessibility patterns of known structures', *Prot. Struct. Funct. & Bioinf.* **7**, 257–264.

Bowie, J. U., Luthy, R. & Eisenberg, D. (1991), 'A method to identify protein sequences that fold into known three-dimensional structures', *Prot. Sci.* **253**, 164–170.

Boyle, D. B., Koford, J. S., Scepanovic, R., Jones, E. R. & Rostoker, M. D. (1997), 'Optimization processing for integrated circuit physical design automation system using chaotic fitness improvement method'. US Patent 5,682,322.

Bradley, P., Chivian, D., Meiler, J., Misura, K. M., Rohl, C. A., Schief, W. R., Wedemeyer, W. J., Scheler-Furman, O., Murphy, P., Schonbrun, J., Strauss, C. E. & Baker, D. (2003), 'Rosetta predictions in CASP5: successes, failures, and prospects for complete automation', *Prot. Struct. Funct. & Bioinf.* **53**(S6), 457–468.

Bradley, P., Misura, K. M. S. & Baker, D. (2005), 'Towards high-resolution de novo structure prediction for small proteins', *J. Mol. Biol.* **309**, 1868–1871.

Branden, C. & Tooze, J. (1999), *Introduction to Protein Structure*, 2nd edn, Garland Publishing, New York.

Branke, J., Deb, K., Dierolf, H. & Osswald, M. (2005), Finding Knees in Multi-objective Optimization, *in* 'Proceedings of the Eighth International Conference on Parallel Problem Solving from Nature', Springer, pp. 722–731.

Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S. & Karplus, M. (1983), 'Charmm: A program for macromolecular energy, minmimization, and dynamics calculations.', *J. Comput. Chem.* **4**, 187–217.

Bruccoleri, R. E. & Karplus, M. (1987), 'Prediction of the folding of short polypeptide segments in proteins by systematic search', *Biopolymers* **26**, 137–168.

Brunger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J. S., Kuszewski, J., Nilges, M., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T. & Warren, G. L. (1998), 'Crystallography and NMR system: A new software suite for macromolecular structure determination', *Acta Cryst. D* **54**(5), 905–921.

Bryant, S. H. (1996), 'Evaluation of threading specificity and accuracy', *Prot. Struct. Funct. & Bioinf.* **26**, 172–185.

Bryant, S. H. & Amzel, L. M. (1987), 'Correctly folded proteins make twice as many hydrophobic contacts', *International journal of peptide and protein research* **29**(1), 46–52.

Buckle, A. M., Henrick, K. & Fersht, A. R. (1993), 'Crystal structural analysis of mutations in the hydrophobic cores of barnase', *J. Mol. Biol.* **234**(3), 847–860.

Bujnicki, J. M., Elofsson, A., Fischer, D. & Rychlewski, L. (2001), 'LiveBench-1: continuous benchmarking of protein structure prediction servers', *Prot. Sci.* **10**(2), 352–361.

Byrd, R. H., Lu, P. & Nocedal, J. (1995), 'A limted memory algorithm for bound constrained optimization', *SIAM Journal on scientific and statiscal computing* **16**(5), 1190–1208.

Bystroff, C., Shao, Y. & Yuan, X. (2004), 'Five hierarchical levels of sequence-structure correlations in proteins', *Appl Bioinformatics* **3**, 97–104.

Bystroff, C., Simon, K. T., Han, K. F. & Baker, D. (1996), 'Local sequence-structure correlations in proteins', *Curr. Opin. Biotechnol.* **7**, 417–421.

Canutescu, A. A. & Dunbrack, R. L. (2003), 'Cyclic coordinate descent: A robotics algorithm for protein loop closure', *Prot. Sci.* **12**(5), 963–972.

Canutescu, A. A., Shelenkov, A. A. & Dunbrack, R. L. J. (2003), 'A graph-theory algorithm for rapid protein side-chain prediction', *Prot. Sci.* **12**(9), 2001–2014.

Carugo, O. & Pongor, S. (2001), 'A normalized root-mean-square distance for comparing protein three-dimensional structures.', *Prot. Sci.* **10**(7), 1470–1473.

Chakravarty, S., Wang, L. & Sanchez, R. (2005), 'Accuracy of structure-derived properties in simple comparative models of protein structures', *Nucleic Acids Res.* **33**(1), 244–259.

Chen, W. W. & Shakhnovich, E. I. (2005), 'Lessons from the design of a novel atomic potential for protein folding', *Prot. Sci.* **14**(7), 1741–1752.

Chothia, C. (1992), 'One thousand families for the molecular biologist', *Nature* **357**(6379), 543–544.

Chothia, C. & Finkelstein, A. V. (1990), 'The Classification and Origins of Protein Folding Patterns', *Ann. Rev. Biochem.* **59**(1), 1007–1035.

Chothia, C. & Lesk, A. (1986), 'The relation between the divergence of sequence and structure in proteins', *EMBO J.* **5**, 823–826.

Chung, S. Y. & Subbiah, S. (1996), How similar must a template protein be for homology modeling by side-chain packing methods?, *in* L. Hunter & T. Teri Klein, eds, 'Biocomputing: Proceedings of the 1996 Pacific Symposium', World Scientific Publishing Co, Singapore, pp. 126–141.

Clore, G. M., Robien, M. A. & Gronenborn, A. M. (1993), 'Exploring the limits of precision and accuracy of protein structures determined by nuclear magnetic resonance spectroscopy.', *J. Mol. Biol.* **231**(1), 82–102.

Colovos, C. & Yeates, T. O. (1993), 'Verification of protein structures: patterns of nonbonded atomic interactions', *Prot. Sci.* **2**(9), 1511–1519.

Contreras-Moreira, B., Ezkurdia, M. L., Tress, M. L. & Valencia, A. (2005), 'Empirical limits for template-based protein structure prediction: the CASP5 example.', *FEBS Lett.* **579**, 1203–1207.

Contreras-Moreira, B., Fitzjohn, P. W. & Bates, P. A. (2003), 'In silico protein recombination: enhancing template and sequence alignment selection for comparative protein modelling', *J. Mol. Biol.* **328**, 593–608.

Contreras-Moreira, B., Fitzjohn, P. W., Offman, M., Smith, G. R. & Bates, P. A. (2003), 'Novel use of a genetic algorithm for protein structure prediction: searching template and sequence alignment space.', *Prot. Struct. Funct. & Bioinf.* **53**(S6), 424–429.

Corey, R. B. & Pauling, L. (1953), 'Fundamental Dimensions of Polypeptide Chains', *Proceedings of the Royal Society of London. Series B, Biological Sciences* **141**(902), 10–20.

Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz, K. M. J., Ferguson, D. M., Spellmeyer, D. C., Fox, T., Caldwell, J. W. & Kollman, P. A. (1995), 'A second generation force field for the simulation of proteins, nucleic acids and organic molecules.', *J. Am. Chem. Soc.* **117**, 5179–5197.

Coutsias, E. A., Seok, C. & Dill, K. A. (2004), 'Using quaternions to calculate RMSD', *J. Comput. Chem.* **25**(15), 1849–1857.

Cozzetto, R. & Tramontano, A. (2005), 'Relationship between multiple sequence alignments and quality of protein comparative models.', *Prot. Struct. Funct. & Bioinf.* **58**(1), 151–157.

Creighton, T. E. (1993), *Proteins:Structure and Molecular Properties*, 2nd edn, W.H. Freeman and Company, New York.

Cristobal, S., Zemla, A., Fischer, D., Rychlewski, L. & Elofsson, A. (2001), 'A study of quality measures for protein threading models.', *BMC Bioinformatics.* **2**(1), 2–5.

Dalton, J. A. R. & Jackson, R. M. (2007 (in press)), 'An evaluation of automated homology modelling methods at low target-sequence similarity', *Bioinformatics.* .

Dandekar, T. & Argos, P. (1992), 'Potential of genetic algorithms in proteins folding and protein engineering simulations', *Prot. Eng.* **5**(7), 637–645.

Dandekar, T. & Argos, P. (1994), 'Folding the main chain of small proteins with the genetic algorithm.', *J. Mol. Biol.* **236**(3), 844–861.

Dandekar, T. & Argos, P. (1996), 'Identifying the tertiary fold of small proteins with different topologies from sequence and structure using the genetic algorithm and extended criteria specific for strand regions', *J. Mol. Biol.* **256**, 645–680.

Davidson, A. R., Lumb, K. J. & Sauer, R. T. (1995), 'Cooperatively folded proteins in random sequence libraries', *Nature, Structural Biology.* **2**(10), 856–864.

Davis, L. D. (1989), Adapting operator probabilities in genetic algorithms, *in* J. D. Schaffer, ed., 'Proceeding of the third international conference on genetic algorithms', Morgan Kauffmann.

Davis, L. D. (1991), *Handbook of genetic algorithms*, Van Nostrand Reinhold.

Dayhoff, M., Schwartz, R. M. & Orcutt, B. C. (1978), *Atlas of protein sequence and structure*, Vol. 5, National Biomedical Research Foundation, Silver Spring, Maryland, chapter A model of evolutionary change in proteins., pp. 345–352.

de Bakker, P. I., Furnham, N., Blundell, T. L. & DePristo, M. (2006), 'Conformer generation under restraints', *Curr. Opin. Struct. Biol.* **16**(2), 160–165.

de Bakker, P. I. W., DePristo, M. A., Burke, D. F. & Blundell, T. J. (2003), 'Ab initio construction of polypeptide fragments: Accuracy of loop decoy discrimination by an all-atom statistical potential and the amber force field with the generalized born solvation model.', *Prot. Struct. Funct. & Bioinf.* **51**(1), 21–40.

De Jong, K. A. (1975), An Analysis of the Behaviour of a Class of Genetic Adaptive Systems, PhD thesis, Dept. Computer and Communication Sciences, University of Michigan, Ann Arbor.

De Jong, K. A. & Spears, W. M. (1990), An analysis of the interacting roles of population size and crossover in genetic algorithms, *in* 'Proc. First Workshop Parallel Problem Solving from Nature', Springer-Verlag, Berlin, pp. 38–47.

Deb, K. (1999), 'Multi-objective Genetic Algorithms: Problem Difficulties and Construction of Test Problems', *Evolutionary Computation* **7**(3), 205–230.

Deb, K. (2001), *Multi-Objective optimization using evolutionary algorithms*, Wiley.

Deb, K. (2002), 'A fast and elitist multiobjective genetic algorithm: NSGA-II', *IEEE Transactions on Evolutionary Computation* **6**(2), 182–197.

Deb, K. & Goldberg, D. (1991), 'A comparative analysis of selection schemes used in genetic algorithms', *Foundations of Genetic Algorithms* pp. 69–93.

dePristo, M. A., de Bakker, P. I., Johnson, R. J. & Blundell, T. J. (2003), 'Crystallographic refinement by knowledge-based exploration of conplex energy landscapes', *Structure* **13**, 1311–1319.

DePristo, M. A., de Bakker, P. I. W., Lovell, S. C. & Blundell, T. J. (2003), 'Ab initio construction of polypeptide fragments: Efficient generation of accurate, representative ensembles', *Prot. Struct. Funct. & Bioinf.* **51**(1), 41–55.

Diamond, R. (1976), 'On the comparison of conformations using linear and quadratic transformations', *Acta Crystallogr.* **A32**, 1–10.

Dill, K. A. & Chan, H. S. (1997), 'From Levinthal to pathways to funnels', *Nature Structural Biology* **4**(1), 10–19.

Donate, L. E., Rufino, S. D., Canard, L. H. & Blundell, T. L. (1996), 'Conformational analysis and clustering of short and medium size loops connecting regular secondary structures: a database for modeling and prediction', *Prot. Sci.* **5**(12), 2600–2616.

Doolittle, R. F. (1981), 'Similar amino acid sequences: chance or common ancestory?', *Science* **214**(4517), 149–159.

Du, P., Andrec, M. & Levy, R. (2003), 'Have we seen all structures corresponding to short protein fragments in the Protein Data Bank? an update', *Prot. Eng.* **16**(6), 407–417.

Dunbrack, R. L. J. (2002), 'Rotamer libraries for the 21st century', *Curr. Opin. Struct. Biol.* **12**(4), 431–440.

Dunbrack, R. L. J. & Karplus, M. (1993), 'Backbone-dependent rotamer library for proteins. application to side-chain prediction', *J. Mol. Biol.* **230**(2), 543–573.

Eiben, A. E., Hinterding, R. & Michalewicz, Z. (1999), 'Parameter control in evolutionary algorithms', *IEEE Trans. on Evolutionary Computation* **3**(2), 124–141.

Eisenberg, D., Luthy, R. & Bowie, J. U. (1997), 'VERIFY3D: assessment of protein models with three-dimensional profiles.', *Methods Enzymol* **277**, 396–404.

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

Elofsson, A., Le Grand, S. M. & Eisenberg, D. (1995), 'Local moves: an efficient algorithm for simulation of protein folding', *Prot. Struct. Funct. & Bioinf.* **23**, 73–82.

Engh, R. A. & Huber, R. (1991), 'Accurate bond and angle parameters for x-ray protein structure refinement', *Acta Cryst. A* **47**(4), 392–400.

Eshelman, L. J., Caruana, R. A. & Schaffer, J. D. (1989), Biases in the landscape, *in* J. D. Schaffer, ed., 'Proceeding of the third international conference on genetic algorithms', Morgan Kauffmann.

Fan, H. & Mark, A. E. (2004), 'Refinement of homology-based protein structures by molecular dynamics simulation techniques', *Prot. Sci.* **13**(1), 211–220.

Fang, Q. & Shortle, D. (2005), 'A Consistent Set of Statistical Potentials for Quantifying Local Side-Chain and Backbone Interactions', *Prot. Struct. Funct. & Bioinf.* **60**, 90–96.

Fischer, D. (2003), '3D-SHOTGUN: a novel, cooperative, fold-recognition meta-predictor', *Proteins* **51**(3), 434–441.

Fiser, A. (2004), 'Protein structure modeling in the proteomics era', *Expert Rev Proteomics* **1**(1), 97–110.

Fiser, A., Do, R. K. & Sali, A. (2000), 'Modeling of loops in protein structures', *Prot. Sci.* **9**(9), 1753–1773.

Flohil, J. A., Vriend, G. & Berendsen, H. J. (2002), 'Completion and refinement of 3-d homology models with restricted molecular dynamics: application to targets 47, 58, and 111 in the casp modeling competition and posterior analysis', *Proteins* **48**(4), 593–604.

Fonseca, C. M. & Fleming, P. J. (1993), Genetic algorithms for multiobjective optimization: Formulation, discussion and generalization, *in* S. Forrest, ed., 'Genetic Algorithms: Proceedings of the Fifth International Conference on Genetic Algorithms', pp. 416–423.

Fonseca, C. M. & Fleming, P. J. (1995), 'An Overview of Evolutionary Algorithms in Multiobjective Optimization', *Evolutionary Computation* **3**(1), 1–16.

Frauenfelder, H., Sligar, S. G. & Wolynes, P. G. (1991), 'The energy landscapes and motions of proteins', *Science* **254**(5038), 1598.

Furuichi, E. & Koehl, P. (1998), 'Influence of protein structure databases on the predictive power of statistical pair potentials.', *Prot. Struct. Funct. & Bioinf.* **31**(2), 139–149.

Gilis, D. & Rooman, M. (1997), 'Predicting protein stability changes upon mutation using database-derived potentials: solvent accessibility determines the importance of local versus non-local interactions along the sequence.', *J. Mol. Biol.* **272**(2), 276–290.

Ginalski, K. (2006), 'Comparative modelling of protein structure prediction', *Curr. Opin. Struct. Biol.* **16**, 172–177.

Ginalski, K., Pas, J., Wyrwicz, L. S., von Grotthuss, M., Bujnicki, J. M. & Rychlewski, L. (2003), 'ORFeus: Detection of distant homology using sequence profiles and predicted secondary structure.', *Nucleic Acids Res.* **31**(13), 3804–3807.

Ginalski, K., von Grotthuss, M., Grishin, N. V. & Rychlewski, L. (2004), 'Detecting distant homology with Meta-BASIC', *Nucleic Acids Res.* **1**(32), 576–581.

Godzik, A., Kolinski, A. & Skolnick, J. (1992), 'Topology fingerprint approach to the inverse protein folding problem', *J. Mol. Biol.* **227**, 227–238.

Goldberg, D. E. (1989), *Genetic algorithms in search, optimization, and machine learning*, 2nd edn, Addison-Wesley Professional, Reading MA.

Goldberg, D. E. (1990), Real-coded genetic algorithms, virtual alphabets, and blocking, Tech Rep. 90001, University of Illinois at Urbana-Champaign, Urbana, Illinois.

Goldberg, D. E. & Richardson, J. (1987), Genetic algorithms with sharing for multimodal function optimization, *in* 'Proceedings of the Second International Conference on Genetic Algorithms on Genetic algorithms and their application', Lawrence Erlbaum Associates, Inc., Mahwah, NJ, USA, pp. 41–49.

Grantham, R. (1974), 'Amino acid difference formula to help explain protein evolution', *Science* **185**(4154), 862–864.

Grefenstette, J. J. (1986), 'Optimization of control parameters for genetic algorithms', *IEEE Transactions on Systems, man, cybernetics* **16**(1), 122–128.

Gregoret, L. M. & Cohen, F. E. (1990), 'Novel method for the rapid evaluation of packing in protein structures', *J. Mol. Biol.* **211**(4), 959–974.

Gromiha, M. M., An, J., Kono, H., Oobatake, M., Uedaira, H. & Sarai, A. (1999), 'ProTherm: Thermodynamic Database for Proteins and Mutants', *Nucleic Acids Res.* **27**(1), 286–288.

Gront, D., Kolinski, A. & Hansmann, U. H. (2005), 'Protein structure prediction by tempering spatial constraints', *Journal of computer-aided molecular design* **19**(8), 603–608.

Hendlich, M., Lackner, P., Weitckus, S., Floeckner, H., Froschauer, R., Gottsbacher, K., Casari, G. & Sippl, M. J. (1990), 'Identification of native protein folds amongst a large number of incorrect models - the calculation of low-energy conformations from potentials of mean force.', *J. Mol. Biol.* **216**(1), 167–180.

Hinds, D. A. & Levitt, M. (1994), 'Exploring conformational space with a simple lattice model for protein structure', *J. Mol. Biol.* **243**(3), 668–682.

Ho, B. K., Coutsias, E. A., Seok, C. & Dill, K. A. (2005), 'The flexibility in the proline ring couples to the protein backbone', *Prot. Sci.* **14**, 1011–1018.

Hoang, T. X., Trovato, A., Seno, F., Banavar, J. R. & Maritan, A. (2004), 'Geometry and symmetry presculpt the free-energy landscape of proteins', *Proc. Natl. Acad. Sci. USA* **101**(21), 7960–7964.

Holland, J. H. (1975), *Adaptation in Natural and Artificial Systems*, MIT Press, Ann Arbor, MI.

Holm, L. & Sander, C. (1991), 'Database algorithm for generating protein backbone and side-chain co-ordinates from a c alpha trace application to model building and detection of co-ordinate errors', *J. Mol. Biol.* **218**(1), 183–194.

Holm, L. & Sander, C. (1992*a*), 'Evaluation of protein models by atomic solvation preference.', *J. Mol. Biol.* **225**, 93–105.

Holm, L. & Sander, C. (1992*b*), 'Evaluation of protein models by atomic solvation preferences.', *J. Mol. Biol.* **225**(1), 93–105.

Holmes, J. B. & Tsai, J. (2004), 'Some fundamental aspects of building protein structures from fragment libraries', *Prot. Sci.* **13**(6), 1636–1650.

Honig, B. (1999), 'Protein folding: from the levinthal paradox to structure prediction', *J. Mol. Biol* **293**(2), 283–293.

Hooft, R. W., Sander, C. & Vriend, G. (1997), 'Objectively judging the quality of a protein structure from a Ramachandran plot', *Comput Appl Biosci* **13**, 425–430.

Hooft, R. W. W., Vriend, G., Sander, C. & Abola, E. E. (1999), 'Errors in protein structures', *Nature.* **381**(6580), 272.

Horn, J. & Nafpliotis, N. (1993), Multiobjective optimization using the niched pareto genetic algorithm, IlliGAL Report No. 93005 93005, Urbana: University of Illinois at Urbana-Champaign, Illinois Genetic Algorithms Laboratory.

Hovmöller, S., Zhou, T. & Ohlson, T. (2002), 'Conformations of amino acids in proteins', *Acta Crystallogr D Biol Crystallogr.* **58**(Part 5), 768–776.

Hubbard, T. J. P. (1999), 'RMS/Coverage graphs: a qualitative method for comparing three-dimensional protein structure predictions', *Prot. Struct. Funct. & Bioinf.* **S3**, 15–21.

Irving, J. A., Whisstock, J. C. & Lesk, A. M. (2001), 'Protein structural alignments and functional genomics', *Prot. Struct. Funct. & Bioinf.* **42**(3), 378–382.

Jacobson, M. P., Pincus, D. L., Rapp, C. S., Day, T. J., Honig, B., Shaw, D. E. & Friesner, R. A. (2004), 'A hierarchical approach to all-atom protein loop prediction', *Prot. Struct. Funct. & Bioinf.* **55**(2), 351–367.

Jaroszewski, L., Rychlewski, L., Li, Z., Li, W. & Godzik, A. (2005), 'Ffas03: a server for profile–profile sequence alignments', *Nucleic Acids Res.* **33**(W), 284–288.

Jernigan, R. L. & Bahar, I. (1996), 'Structure-derived potentials and protein simulations', *Curr. Opin. Struct. Biol.* **6**(2), 195–209.

Johnson, M. S., Overington, J. P. & Blundell, T. L. (1993), 'Alignment and searching for common protein folds using a data bank of structural templates.', *J. Mol. Biol.* **231**(3), 735–752.

Jones, A. W. & Kleywegt, G. J. (1999), 'CASP3 comparative modelling evaluation', *Prot. Struct. Funct. & Bioinf.* **S3**, 30–46.

Jones, D. T. (1994), 'De novo protein design using pairwise potentials and a genetic algorithm', *Prot. Sci.* **3**(4), 567–574.

Jones, D. T. (1997), 'Successful ab initio prediction of the tertiary structure of NK-Lysin using multiple sequences and recognised supersecondary structural motifs', *Prot. Struct. Funct. & Bioinf.* **S1**, 185–191.

Jones, D. T. (1999*a*), 'GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences.', *J. Mol. Biol.* **287**(4), 797–815.

Jones, D. T. (1999*b*), 'Protein secondary structure prediction based on position-specific scoring matrices', *J. Mol. Biol.* **292**(2), 195–202.

Jones, D. T. (2000), 'A practical guide to protein structure prediction', *Methods Mol. Biol.* **143**, 131–154.

Jones, D. T. (2001), 'Prediction novel protein folds by using FRAGFOLD', *Prot. Struct. Funct. & Bioinf.* **S5**, 127–132.

Jones, D. T., Taylor, W. R. & Thornton, J. M. (1992), 'A new approach to protein fold recognition', *Nature.* **358**, 86–89.

Jones, T. A. & Thirup, S. (1986), 'Using known substructures in protein model building and crystallography', *EMBO J* **5**(4), 819–822.

Kabsch, W. (1976), 'A solution for the best rotation to relate two sets of vectors', *Acta Crystallogr.* **A32**, 1976.

Kabsch, W. (1978), 'A discussion of the solution for the best rotation to relate two sets of vectors', *Acta Crystallogr.* **A34**, 827–828.

Karplus, K., Barret, C. & Hughey, R. (1998), 'Hidden Markov Models for detecting remote protein homologies.', *Bioinformatics.* **14**(10), 846–856.

Karplus, K., Karchin, R., Barrett, C., Tu, S., Cline, S., Diekhans, M., Grate, L., Casper, J. & Hughey, R. (2001), 'What is the value added by human intervention in protein structure prediction?', *Prot. Struct. Funct. & Bioinf.* **45**(S5), 86–91.

Karplus, K., Karchin, R., Draper, J., Casper, J., Mandel-Gutfreund, Y., Diekans, M. & Hughey, R. (2003), 'Combining local-structure, fold-recognition, and new fold methods for protein structure prediction', *Prot. Struct. Funct. & Bioinf.* **53**(S6), 491–496.

Karplus, M. & Petsko, G. A. (1990), 'Molecular dynamics simulations in biology', *Nature.* **347**(6294), 631–639.

Kawai, H., Kikuchi, T. & Okamoto, Y. (1989), 'A prediction of tertiary structures of peptide by the Monte Carlo simulated annealing method', *Prot. Eng.* **3**(2), 85–94.

Keasar, C. & Levitt, M. (2003), 'A novel approach to decoy set generation: Designing a physical energy function having local minima with native structure characteristics.', *J. Mol. Biol.* **329**(1), 156–174.

Kelley, L. A., MacCallum, R. M. & Sternberg, M. J. E. (2000), 'Enhanced genome annotation using structural profiles in the program 3D-PSSM.', *J. Mol. Biol.* **299**(2), 499–520.

Kelley, L. A., MacCallum, R. M., Sternberg, M., Karplus, K., Fischer, D., Elofsson, A., Godzik, A., Rychlewski, L., Pawlowski, K., Jones, D. T. et al. (1999), 'CAFASP-1: Critical assessment of fully automated structure prediction methods', *Prot. Struct. Funct. & Bioinf.* **S3**, 209–217.

Kirkpatrick, S., Gelatt Jr, C. D. & Vecchi, M. P. (1983), 'Optimization by Simulated Annealing', *Science* **220**(4598), 671–680.

Kleywegt, G. J. & Jones, T. A. (1996), 'Phi/Psi-chology: Ramachandran revisited', *Structure* **4**(12), 1395–1400.

Knowles, J. & Corne, D. (2002), On metrics for comparing non-dominated sets, *in* 'In Congress on Evolutionary Computation (CEC 2002)', IEEE Press, pp. 711–716.

Kocher, J. P., Rooman, M. J. & Wodak, S. J. (1994), 'Factors influencing the ability of knowledge-based potentials to identify native sequence-structure matches.', *J. Mol. Biol.* **235**(5), 1598–1613.

Koehl, P. & Levitt, M. (1999), 'A brighter future for protein structure prediction', *Nature, Structural Biology.* **6**(2), 108–111.

Kolinski, A. (2004), 'Protein modeling and structure prediction with a reduced representation', *Acta Biochim. Polonica* **51**(2), 349–371.

Kolinski, A., Betancourt, M. R., Kihara, D., Rotkiewicz, P. & Soling, J. (2001), 'Generalized comparative modeling (GENECOMP): a combination of sequence comparison, threading, and lattice modeling for protein structure prediction and refinement', *Prot. Struct. Funct. & Bioinf.* **44**, 133–149.

Koolman, J., Röhm, K. H. & Roehm, K. H. (2005), *Color Atlas of Biochemistry*, 2 edn, Thieme.

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', *J. Mol. Biol.* **326**(4), 1239–1259.

Kosinski, J., Cymerman, I. A., Feder, M., Kurowski, M. A., Sasin, J. M. & Bujnicki, J. M. (2003), 'A "FRankenstein's monster" approach to comparative modeling: merging the finest fragments of Fold-Recognition models and iterative model refinement aided by 3d structure evaluation.', *Prot. Struct. Funct. & Bioinf.* **53**(S6), 369–379.

Kosinski, J., Gajda, M. J., Cymerman, I. A., Kurowski, M. A., Pawlowski, M., Boniecki, M., Obarska, A., Papaj, G., Sroczynska-Obuchowicz, P., Tkaczuk, K. L., Sniezynska, P., Sasin, J. M., Augustyn, A., Bujnicki, J. M. & Feder, M. (2005), 'FRankenstein becomes a cyborg: the automatic recombination and realignment of fold recognition models in CASP6', *Proteins* **61 Suppl 7**, 106–113.

Kuhlman, B., Dantas, G., Ireton, G. C., Varani, G., Stoddard, B. L. & Baker, D. (2003), 'Design of a Novel Globular Protein Fold with Atomic-Level Accuracy', *Science* **302**(5649), 1364–1368.

Kulikova, T., Akhtar, R., Aldebert, P., Althorpe, N., Andersson, M., Baldwin, A., Bates, K., Bhattacharyya, S., Bower, L., Browne, P., Castro, M., Cochrane, G., Duggan, K., Eberhardt, R., Faruque, N., Hoad, G., Kanz, C., Lee, C., Leinonen, R., Lin, Q., Lombard, V., Lopez, R., Lorenc, D., McWilliam, H., Mukherjee, G., Nardone, F., Pilar Garcia Pastor, M., Plaister, S., Sobhany, S., Stoehr, P., Vaughan, R., Wu, D., Zhu, W. & Apweiler, R. (2006), 'EMBL Nucleotide Sequence Database in 2006', *Nucleic Acids Res.* **35**(D), 16–18.

Kuznetsov, I. B. & Rackovsky, S. (2002), 'Discriminative ability with respect to amino acid types: assessing the performance of knowledge-based potentials without threading.', *Prot. Struct. Funct. & Bioinf.* **49**(2), 266–284.

Laskowski, R. A., McArthur, M. W., Moss, D. J. & Thornton, J. M. (1993), 'PROCHECK: a program to check the stereochemical quality of protein structures', *J. Appl. Cryst.* **26**, 283–291.

Lazaridis, T. & Karplus, M. (1999*a*), 'Discrimination of the native from misfolded protein models with an energy function including implicit solvation', *J. Mol. Biol.* **288**(3), 477–487.

Lazaridis, T. & Karplus, M. (1999*b*), 'Effective energy function for protein in solution', *Prot. Struct. Funct. & Bioinf.* **35**(2), 133–152.

Lazaridis, T. & Karplus, M. (2000), 'Effective energy functions for protein structure prediction', *Curr. Opin. Struct. Biol.* **10**(2), 139–145.

Lee, C. & Subbiah, S. (1991), 'Prediction of protein side-chain conformation by packing optimization.', *J. Mol. Biol.* **217**(2), 373–388.

Lee, M. R., Tsai, J., Baker, D. & Kollman, P. A. (2001), 'Molecular dynamics in the endgame of protein structure prediction.', *J. Mol. Biol.* **313**(2), 417–430.

Leech, A. R. (2001), *Molecular Modelling: Principles and Applications*, 2nd edn, Prentice Hall, New York.

Lesk, A. M. (1991), *Protein Architecture - A practical approach*, IRL Press.

Levinthal, C. (1969), How to fold graciously, *in* P. Debrunner, J. Tsibris & E. Munck, eds, 'Mossbauer Spectroscopy in Biological Systems', Proceedings of a Meeting held at Allerton House, University of Illinois Press, Urbana, Monticello, IL, pp. 22–24.

Levitt, M. (1976), 'A Simplified Representation of Protein Conformations for Rapid Simulation of Protein Folding', *J. Mol. Biol.* **104**, 59–107.

Levitt, M. (1992), 'Accurate modeling of protein conformations by automatic segment matching', *J. Mol. Biol.* **226**, 507–533.

Levitt, M. & Chothia, C. (1976), 'Structural patterns in globular proteins', *Nature* **261**(5561), 552–558.

Levitt, M. & Gerstein, M. (1998), 'A unified statistical framework for sequence comparison and structure comparison', *Proc. Natl. Acad. Sci. USA* **95**(11), 5913–5920.

Levitt, M. & Warshel, A. (1975), 'Computer simulation of protein folding', *Nature* **253**(5494), 694–698.

Li, Z. & Scheraga, H. A. (1987), 'Monte Carlo-Minimization Approach to the Multiple-Minima Problem in Protein Folding', *Proc. Natl. Acad. Sci. USA* **84**(19), 6611–6615.

Liepins, G. E. & Vose, M. D. (1990), 'Representational issues in genetic optimization', *Journal of Exp. and Theo. Art. Intel.* **2**(2), 4–30.

Lindahl, E., Hess, B. & van der Spoel, D. (2001), 'GROMACS 3.0: a package for molecular simulation and trajectory analysis.', *N* **7**(8), 306–317.

Livingstone, C. D. & Barton, G. J. (1993), 'Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation', *Comput Appl Biosci* **9**(6), 745–756.

Lovell, S. C., Davis, I. W., Arendall, W. B. r., de Bakker, P. I., Word, J. M., Prisant, M. G., Richardson, J. S. & Richardson, D. C. (2003), 'Structure validation by calpha geometry: phi,psi and cbeta deviation.', *Prot. Struct. Funct. & Bioinf.* **50**(3), 437–450.

Lu, H. & Skolnick, J. (2001), 'A distance-dependent atomic knowledge-based potential for improving protein structure selection.', *Prot. Struct. Funct. & Bioinf.* **44**(3), 223–232.

Lu, H. & Skolnick, J. (2003), 'Application of statistical potentials to protein structure refinement from low resolution ab initio models', *Biopolymers* **70**(4), 575–584.

Lundstrom, J., Rychlewski, L., Bujnicki, J. & Elofsson, A. (2001), 'Pcons: a neural-network-based consensus predictor that improves fold recognition.', *Prot. Sci.* **10**(11), 2354–2362.

Lupyan, D., Leo-Macias, A. & Ortiz, A. R. (2005), 'A new progressive-iterative algorithm for multiple structure alignment', *Bioinformatics.* **21**(15), 3255–3263.

Luthy, R., Bowie, J. U. & Eisenberg, D. (1992), 'Assessment of protein models with three-dimensional profiles', *Nature* **356**(6364), 83–85.

MacKerell, A. D., Bashford, D., Bellott, M., Dunbrack, R. L., Eva seck, J. D., Field, M. J., Fischer, S., Gao, J. Guo, H., Ha, S., JosephMcCarthy, D., Kuc nir, L., Kuczera, K.and Lau, F. T. K., Mattos, C. Michnick, S., Ngo, T., Nguyen, D. T., Pro hom, B. Reiher, W. E., Roux, B., Schlenkrich, M., Smith, J. C., Stote, R., Straub, J. W., Tanabe, M., WiorkiewiczKuczera, J. & Yin, D.and Karplus, M. (1998), 'All-atom empirical potential for molecular modeling and dynamics studies of proteins.', *J. Phys. Chem.* **102**, 3586–3617.

MacLachlan, A. D. (1972), 'A mathematical procedure for superimposing atomic coordinates of proteins', *Acta Crystallogr.* **A28**, 656–657.

Maiorov, V. N. & Crippen, G. N. (1995), 'Size-independent comparison of protein three-dimensional structures', *Prot. Struct. Funct. & Bioinf.* **22**(3), 273–283.

Mallick, P., Weiss, R. & Eisenberg, D. (2002), 'The directional atomic solvation energy: an atom-based potential for the assignment of protein sequences to known folds', *Proc. Natl. Acad. Sci. USA* **99**, 16041–16046.

Marchler-Bauer, A. & Bryant, S. H. (1999), 'Comparison of prediction quality in the three CASPS', *Prot. Struct. Funct. & Bioinf.* **S3**, 218–225.

Marti-Renom, M. A., Stuart, A. C., Fiser, A., Sanchez, R., Melo, F. & Šali, A. (2000), 'Comparative protein structure modelling of genes and genomes', *Ann. Rev. Biophys. Biomolec. Struct.* **29**, 291–325.

May, A. C. & Johnson, M. S. (1995), 'Improved genetic algorithm-based protein structure comparisons: pairwise and multiple superpositions', *Prot. Eng.* **8**, 873–882.

McDonald, I. K. & Thornton, J. M. (1994), 'Satisfying hydrogen bonding potential in proteins', *J. Mol. Biol.* **238**(5), 777–793.

McGuffin, L. J. & Jones, D. T. (2003*a*), 'Assembling novel protein folds from supersecondary structural fragments', *Prot. Struct. Funct. & Bioinf.* **S6**, 480–485.

McGuffin, L. J. & Jones, D. T. (2003*b*), 'Improvement of the GenTHREADER method for genomic fold recognition', *Bioinformatics.* **19**(7), 874–881.

McGuffin, L. J., Smith, R. T., Bryson, K., Srensen, S. A. & Jones, D. T. (2006), 'High throughput profile-profile based fold recognition for the entire human proteome', *BMC bioinformatics* **7**, 288.

Melo, F., Sanchez, R. & Šali, A. (2002), 'Statistical potentials for fold assessment', *Prot. Sci.* **11**, 430–448.

Metropolis, N. et al. (1953), 'Equations of state calculations by fast computational machine', *J. Chem. Phys.* **21**, 1087–1091.

Michalewicz, Z. (1999), *Genetic Algorithms + Data Structures = Evolutionary Programs*, Springer-Verlag, New York.

Miranker, A. D. & Dobson, C. M. (1996), 'Collapse and cooperativity in protein folding.', *Curr. Opin. Struct. Biol.* **6**(1), 31–42.

Mirny, L. A. & Shakhnovich, E. I. (1996), 'How to derive a protein folding potential? a new approach to an old problem', *J. Mol. Biol.* **264**(5), 1164–1179.

Mirny, L. & Shakhnovich, E. I. (2001), 'Protein folding theory: From Lattice to All-Atom Models', *Annu. Rev. Biophys. Biomol. Struct.* **30**(1), 361–396.

Mirsky, A. E. & Pauling, L. (1936), 'On the structure of native, denatured, and coagulated proteins.', *Proc. Natl. Acad. Sci. USA* **22**(7), 439–447.

Misura, K. M. & Baker, D. (2005), 'Progress and challenges in high-resolution refinement of protein structure models', *Prot. Struct. Funct. & Bioinf.* **59**(1), 15–29.

Misura, K. M., Chivian, D., Rohl, C. A., Kim, D. E. & Baker, D. (2006), 'Physically realistic homology models built with rosetta can be more accurate than their templates', *Proc. Natl. Acad. Sci. USA* **103**(14), 5361–5366.

Mitchell, M. (1999), *An Introduction to Genetic Algorithms*, 5th edn, MIT Press.

Miyazawa, S. & Jernigan, R. L. (1985), 'Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation', *Macromolecules* **18**(3), 534–552.

Miyazawa, S. & Jernigan, R. L. (1999), 'An empirical potential with a reference state for protein fold and sequence recognition', *Prot. Struct. Funct. & Bioinf.* **36**, 357–369.

Mocz, G. (1995), 'Fuzzy cluster analysis of simple physicochemical properties of amino acids for recognizing secondary structure in proteins', *Prot. Sci.* **4**(6), 1178–1187.

Monge, A., Freisner, R. A. & Honig, B. (1994), 'An algorithm to generate low-resolution protein tertiary structure from knowledge of secondary structure', *Proc. Natl. Acad. Sci. USA* **91**(11), 5027–5029.

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', *Proc. Natl. Acad. Sci. USA* **101**(18), 6946–6951.

Morris, A. L., MacArthur, M. W., Hutchinson, E. G. & Thornton, J. M. (1992), 'Stereochemical quality of protein structure coordinates', *Prot. Struct. Funct. & Bioinf.* **12**(4), 345–364.

Moult, J. (1997), 'A comparison of database potentials and molecular mechanics force fields', *Curr. Opin. Struct. Biol.* **7**(2), 194–199.

Moult, J. (2005), 'A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction', *Curr. Opin. Struct. Biol.* **15**, 285–289.

Moult, J. (2006), 'Rigorous performance evaluation in protein structure modelling and implications for computational biology', *Phil. Trans. R. Soc. B.* **361**, 453–458.

Moult, J., Fidelis, K., Rost, B., Hubbard, T. & Tramontano, A. (2005), 'Critical assessment of methods of protein structure prediction (casp)–round 6', *Prot. Struct. Funct. & Bioinf.* **61**(S7), 3–7.

Moult, J., Fidelis, K., Zemla, A. & Hubbard, T. (2003), 'Critical assessment of methods of protein structure prediction (CASP)-round V.', *Prot. Struct. Funct. & Bioinf.* **53**(S6), 334–349.

Moult, J., Hubbard, T., Bryant, S. H., Fidelis, K. & Pedersen, J. T. (1997), 'Critical Assessment of Methods for Protein Structure Prediction (CASP): round ii', *Prot. Struct. Funct. & Bioinf.* **29**(S1), 2–6.

Moult, J. & James, M. N. G. (1986), 'An algorithm for determining the conformation of polypeptide segments in proteins by systematic search', *Prot. Struct. Funct. & Bioinf.* **1**, 146–163.

Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997), 'Refinement of macromolecular structures by the maximum-likelihood method', *Acta Cryst. D* **53**(3), 240–255.

Murzin, A. G. (1999), 'Structure classification-based assessment of casp3 predictions for the fold recognition targets', *Prot. Struct. Funct. & Bioinf.* **37**(S3), 88–103.

Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995), 'SCOP: a structural classification of proteins database for the investigation of sequences and structures', *J. Mol. Biol.* **247**(4), 536–540.

Myers, J. K. & Oas, T. G. (2002), 'Mechanisms of fast folding proteins', *Ann. Rev. Biochem.* **71**(1), 783–815.

Needleman, S. B. & Wunsch, C. D. (1970), 'A general method applicable to the search for similarities in the amino acid sequence of two proteins', *J. Mol. Biol.* **48**, 443–453.

Nelder, J. A. & Mead, R. (1965), 'A simplex method for function minimization', *Computer Journal* **7**(4), 308–313.

Novotny, J., Brucolleri, R. & Karplus, M. (1984), 'An analysis of incorrectly folded protein models. implications for structure prediction', *J. Mol. Biol.* **177**(4), 787–818.

Novotny, J., Rashin, A. A. & Bruccoleri, R. E. (1988), 'Criteria that discriminate between native proteins and incorrectly folded models', *Prot. Struct. Funct. & Bioinf.* **4**(1), 19–30.

Offman, M. N., Fitzjohn, P. W. & Bates, P. A. (2006), 'Developing a move-set for protein model refinement', *Bioinformatics.* **22**(15), 1838–1845.

Onuchic, J. N., Nymeyer, H., Garcia, A. E., Chahine, J. & Socci, N. D. (2000), 'The energy landscape theory of protein folding: insights into folding mechanisms and scenarios', *Adv. Protein Chem* **53**, 87–152.

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* **5**(8), 1093–1108.

Orengo, C. E., Bray, J. E., Hubbart, T., LoConte, L. & Sillitoe, J. (1999), 'Analysis and assessment of ab initio three-dimensional prediction, secondary structure, and contacts prediction', *Prot. Struct. Funct. & Bioinf.* **S3**, 149–170.

Orengo, C., T., J. D. & Thornton, J. M. (1994), 'Protein superfamilies and domain superfolds', *Nature.* **372**, 631–634.

Pareto, V. (1896), *Cours d'économie politique professé*, à l'Université de Lausanne.

Park, B. H. & Levitt, M. (1995), 'The complexity and accuracy of discrete state models of protein structure', *J. Mol. Biol.* **249**(2), 493–507.

Park, B., Huang, E. S. & Levitt, M. (1997), 'Factors affecting the ability of energy functions to discriminate correct from incorrect folds.', *J. Mol. Biol.* **266**, 831–846.

Park, B. & Levitt, M. (1996), 'Energy functions that discriminate x-ray and near native folds from well-constructed decoys.', *J. Mol. Biol.* **258**, 367–392.

Parsons, J., Holmes, J. B., Rojas, M., Tsai, J. & Strauss, C. E. M. (2005), 'Practical conversion from torsion space to cartesian space for in silico protein synthesis', *J. Comput. Chem.* **26**, 1063–1068.

Pauling, L. (1960), *The Nature of the Chemical Bond and the Structure of Molecules and Crystals: An Introduction to Modern Structural Chemistry*, 3rd edn, Cornell University Press, Ithaca, NY.

Pauling, L., Corey, R. & Branson, H. R. (1951), 'The Structure of Proteins: Two Hydrogen-Bonded Helical Configurations of the Polypeptide Chain', *Proc. Natl. Acad. Sci. USA* **37**(4), 205–211.

Pearson, W. R. & Lipman, D. J. (1988), 'Improved tools for biological sequence comparison', *Proc. Natl. Acad. Sci. USA* **85**, 24444–2448.

Pedersen, J. T. & Moult, J. (1996), 'Genetic algorithms for protein structure prediction', *Curr. Opin. Struct. Biol.* **6**(2), 227–231.

Pedersen, J. T. & Moult, J. (1997*a*), 'Ab initio protein folding simulations with genetic algorithms: simulations on the complete sequence of small proteins', *Prot. Struct. Funct. & Bioinf.* **Suppl 1**, 179–184.

Pedersen, J. T. & Moult, J. (1997*b*), 'Protein folding simulations with genetic algorithms and a detailed molecular description', *J. Mol. Biol.* **269**, 240–259.

Petrella, R. J. & Karplus, M. (2001), 'The Energetics of Off-rotamer Protein Side-chain Conformations', *J. Mol. Biol.* **312**(5), 1161–1175.

Petrey, D., Xiang, Z., Tang, C. L., Xie, L., Gimpelev, M., Mitros, T., Soto, C. S., Goldsmith-Fischman, S., Kernytsky, A., Schlessinger, A., Koh, I. Y., Alexov, E. & Honig, B. (2003), 'Using multiple structure alignments, fast model building, and

energetic analysis in fold recognition and homology modeling.', *Prot. Struct. Funct. & Bioinf.* **53**(S6), 430–435.

Pettitt, C. S., McGuffin, L. J. & Jones, J. T. (2005), 'Improving protein structure prediction using 3D model quality', *Bioinformatics.* **21**(17), 3509–3515.

Plotkin, S. S. & Onuchic, J. N. (2002), 'Understanding protein folding with energy landscape theory Part I: Basic concepts', *Quarterly Reviews of Biophysics* **35**(02), 111–167.

Polak, E. (1971), *Computational Methods in Optimization: A Unified Approach*, Academic Press New York.

Ponder, J. W. & Case, D. A. (2003), 'Force fields for protein simulations.', *Adv. Prot. Chem.* **66**, 27–85.

Ponder, J. W. & Richards, F. M. (1987), 'Tertiary templates for proteins: use of packing criteria in the enumeration of allowed sequence for different structural classes', *J. Mol. Biol.* **193**, 775–791.

Purisima, E. O. & Scheraga, H. A. (1984), 'Conversion from a virtual-bond chain to a complete polypeptide backbone chain', *Biopolymer* **23**(7), 1207–1224.

Qian, B., Ortiz, A. R. & Baker, D. (2004), 'Improvement of comparative model accuracy by free-energy optimization along principal components of natural structural variation', *Proc. Natl. Acad. Sci. USA* **101**(43), 15346–15351.

Ramachandran, G. N., Ramakrishnan, C. & Sasisekaran, V. (1963), 'Stereochemistry of polypeptide chain configurations.', *J. Mol. Biol.* **7**, 95–99.

Ramachandran, G. N. & Sasisekharan, V. (1968), 'Conformation of polypeptides and proteins.', *Adv Protein Chem* **23**, 283–438.

Rao, S. T. & Rossmann, M. G. (1973), 'Comparison of super-secondary structures in proteins.', *J. Mol. Biol.* **76**(2), 241–256.

Richards, F. M. (1977), 'Areas, Volumes, Packing, and Protein Structure', *Annu. Rev. Biophys. and Bioeng.* **6**(1), 151–176.

Ring, C., Sun, E., Mckerrow, J. H., Lee, G. K., Rosenthal, P. J., Kuntz, I. D. & Cohen, F. E. (1993), 'Structure-Based Inhibitor Design by Using Protein Models for the Development of Antiparasitic Agents', *Proc. Natl. Acad. Sci. USA* **90**(8), 3583–3587.

Rohl, C. A., Strauss, C. E., Misura, K. M. & Baker, D. (2004), 'Protein structure prediction using rosetta', *Methods Enzymol* **383**, 66–93.

Rohl, C., Strauss, C. E. M., Chivian, D. & Baker, D. (2004), 'Modelling structurally variable regions in homologous proteins with Rosetta', *Prot. Struct. Funct. & Bioinf.* **55**(3), 656–677.

Rojnuckarin, A. & Subramaniam, S. (1999), 'Knowledge-based interaction potentials for proteins', *Prot. Struct. Funct. & Bioinf.* **36**(1), 54–67.

Rooman, M. & Gilis, D. (1998), 'Different derivations of knowledge-based potentials and analysis of their robustness and context-dependent predictive power.', *Eur. J. Biochem.* **254**(1), 135–143.

Rychlewski, L. & Fischer, D. (2005), 'LiveBench-8: The lareg-scale continous assessment of automated protein structure prediction', *Prot. Sci.* **14**(1), 240–245.

Rychlewski, L., Jaroszewski, L., Li, W. & Godzik, A. (2000), 'Comparison of sequence profiles. strategies for structural predictions using sequence information.', *Prot. Sci.* **9**(2), 232–241.

Samudrala, R. & Levitt, M. (2000), 'Decoys 'r' us: a database of incorrect conformations to improve protein structure prediction', *Prot. Sci.* **9**(7), 1399–1401.

Samudrala, R. & Moult, J. (1998), 'An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction', *J. Mol. Biol.* **275**, 895–916.

Sanchez, R. & Šali, A. (1997), 'Evaluation of comparative protein structure modeling by MODELLER-3', *Prot. Struct. Funct. & Bioinf.* **Suppl 1**, 50–58.

Sanchez, R. & Šali, A. (1998), 'Large-scale protein structure modeling of the Saccharomyces cerevisiae genome', *Proc. Natl. Acad. Sci. USA* **95**, 13597–13602.

Sanchez, R. & Šali, A. (2000), 'Comparative protein structure modelling', *Methods Mol. Biol.* **143**, 131–154.

Sauder, J. M., Arthur, J. W. & Dunbrack Jr, R. L. (2000), 'Large-scale comparison of protein sequence alignment algorithms with structure alignments', *Prot. Struct. Funct. & Bioinf.* **40**(1), 6–22.

Schaffer, J. D. (1984), Some experiments in machine learning using vector evaluated genetic algorithms, PhD thesis, Vanderbilt University, Nashville, TN.

Schaffer, J. D. (1985), Multiple objective optimization with vector evaluated genetic algorithms, *in* 'Proceedings of the First International Conference on Genetic algorithms', pp. 93–100.

Schaffer, J. D., Caruana, R. A., Eshelman, L. J. & Das, R. (1989), A study of control parameters affecting online performance of genetic algorithms for function optimisation, *in* J. D. Schaffer, ed., 'Proceeding of the third international conference on genetic algorithms', Morgan Kauffmann.

Schonbrun, J., Wedemeyer, W. J. & Baker, D. (2002), 'Protein structure prediction in 2002', *Curr. Opin. Struct. Biol.* **12**(3), 348–354.

Sheng, Y. (1996), 'Site-directed mutagenesis of recombinant human beta 2-glycoprotein i identifies a cluster of lysine residues that are critical for phospholipid binding and anti-cardiolipin antibody activity', *J. Immun.* **157**(8), 3744–3751.

Shimada, J., Kussell, E. L. & Shakhnovich, E. I. (2001), 'The folding thermodynamics and kinetics of crambin using an all-atom monte carlo simulation', *J. Mol. Biol.* **308**(1), 79–95.

Shortle, D. (1996), 'The denatured state (the other half of the folding equation) and its role in protein stability', *FASEB J.* **10**, 27–34.

Shortle, D. (2003), 'Propensities, probabilities, and the Boltzmann hypothesis', *Prot. Sci.* **12**(6), 1298–1302.

Sierk, M. L. & Kleywegt, G. J. (2004), 'Deja vu all over again: finding and analyzing protein structure similarities', *Structure* **12**(12), 2103–2111.

Siew, N., Elofsson, A., Rychlewski, L. & Fischer, D. (2000), 'MaxSub: an automated measure for the assessment of protein structure prediction quality.', *Bioinformatics.* **16**(9), 776–785.

Simons, K. T., Kooperberg, C., Huang, E. S. & Baker, D. (1997), 'Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions.', *J. Mol. Biol.* **268**, 209–225.

Sippl, M. (1995), 'Knowledge-based potentials for proteins.', *Curr. Opin. Struct. Biol.* **5**(2), 229–235.

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.', *J. Mol. Biol.* **213**(4), 859–883.

Sippl, M. J. (1993), 'Recognition of errors in three-dimensional structures of proteins', *Prot. Struct. Funct. & Bioinf.* **17**, 355–362.

Skolnick, J., Jaroszewski, L., Kolinski, A. & Godzik, A. (1997), 'Derivation and testing of pair potentials for protein folding. When is the quasichemical approximation correct?', *Prot. Sci.* **6**(3), 676–688.

Skolnick, J., Kolinski, A. & Ortiz, A. R. (1997), 'MONSSTER: a method for folding globular proteins with a small number of distance restraints', *J. Mol. Biol.* **265**(2), 217–241.

Smith, T. F. & Waterman, M. S. (1981), 'Identification of common molecular subsequences', *J. Mol. Biol.* **147**, 195–197.

Sneath, P. H. (1966), 'Relations between chemical structure and biological activity in peptides', *J Theor Biol* **12**(2), 157–195.

Spears, W. M. (1998), The role of mutation and reproduction in evolutionary algorithms., PhD thesis, George Mason University, Fairfax, V.A.

Srinivas, N. & Deb, K. (1994), 'Multiobjective Optimization Using Nondominated Sorting in Genetic Algorithms', *Evolutionary Computation* **2**(3), 221–248.

Stanfel, L. E. (1996), 'A new approach to clustering the amino acids', *J Theor Biol* **183**(2), 195–205.

Stewart, D. E., Sarkar, A. & Wampler, J. E. (1990), 'Occurrence and role of cis peptide bonds in protein structures.', *J. Mol. Biol.* **214**(1), 253–260.

Sun, S. (1993), 'Reduced representation model of protein structure prediction: Statistical potential and genetic algorithms', *Prot. Sci.* **2**(5), 762–785.

Sun, S. (1995), 'A genetic algorithm that seeks the native states of peptides and proteins', *Biophys. J.* **69**(2), 340–355.

Sussman, J. L., Lin, D., Jiang, J., Manning, N. O., Prilusky, J., Ritter, O. & Abola, E. E. (1998), 'Protein Data Bank (PDB): database of three-dimensional structural information of biological macromolecules.', *Acta Crystallogr. D Biol Crystallogr.* **1**(54), 1078–1084.

Swindells, M. B., MacArthur, M. W. & Thornton, J. M. (1995), 'Intrinsic phi/psi propensities of amino acids, derived from coil regions of known structures', *Nature, Structural Biology.* **2**(7), 596–603.

Szustakowski, J. D. & Weng, Z. (2000), 'Protein structure alignment using a genetic algorithm', *Prot. Struct. Funct. & Bioinf.* **38**(4), 428–440.

Tackett, W. A. (1994), Recombination, selection, and the genetic construction of computer programs, PhD thesis, University of Southern California, Los Angeles, CA, USA.

Tanaka, S. & Scheraga, H. A. (1976), 'Medium-and Long-Range Interaction Parameters between Amino Acids for Predicting Three-Dimensional Structures of Proteins', *Macromolecules* **9**(6), 945–950.

Taylor, W. R. (1986), 'The classification of amino acid conservation', *J Theor Biol* **119**(2), 205–218.

Taylor, W. R. & Orengo, C. A. (1989), 'Protein structure alignment', *J. Mol. Biol.* **280**, 1–22.

Theobald, D. L. & Wuttke, D. S. (2006), 'Empirical Bayes hierarchical models for regularizing maximum likelihood estimation in the matrix Gaussian Procrustes problem', *PNAS* **103**(49), 18521–18527.

Thiele, R., Zimmer, R. & Lengauer, T. (1999), 'Protein threading by recursive dynamic programming', *J. Mol. Biol.* **290**, 757–779.

Thierens, D., Elektrotechniek, F. T. W. D., Afdeling, E. S. A. & Leuven, K. U. (1995), *Analysis and Design of Genetic Algorithms*, Departement Elektrotechniek, Faculteit Toegepaste Wetenschappen, Katholieke Universiteit Leuven.

Tobi, D., Shafran, G., Linial, N. & Elber, R. (2000), 'On the design and analysis of protein folding potentials', *Prot. Struct. Funct. & Bioinf.* **40**(1), 71–85.

Torda, A. E. (1997), 'Perspectives in protein-fold recognition', *Curr. Opin. Struct. Biol.* **7**(2), 200–205.

Tosatto, S. C. (2005), 'The victor/FRST function for model quality estimation', *J. Comput. Biol.* **12**(10), 1316–1327.

Tramontano, A. & Morea, V. (2001), 'Assessment of homology-based predictions in CASP5.', *Prot. Struct. Funct. & Bioinf.* **53**(S6), 352–368.

Tress, M., Ezkurdia, I., Grana, O., Lopez, G. & A., V. (2005), 'Assessment of predictions submitted for the casp6 comparative modelling category', *Prot. Struct. Funct. & Bioinf.* **61**(Suppl 7), 27–45.

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.', *Prot. Struct. Funct. & Bioinf.* **52**, 76–87.

Tuffery, P., Etchebest, C., Hazout, S. & Lavery, R. (1991), 'A new approach to the rapid determination of protein side chain conformations.', *J Biomol Struct Dyn* **8**(6), 1267–1289.

Unger, R. & Moult, J. (1993), 'Genetic algorithms for protein folding simulations.', *J. Mol. Biol.* **231**(1), 75–81.

Vajda, S., Sippl, M. & Novotny, J. (1997), 'Empirical potentials and functions for protein folding and binding', *Curr. Opin. Struct. Biol.* **7**, 222–228.

van Gunsteren, W. F. & Berendsen, H. J. C. (1990), 'Computer simulation of molecular dynamics: Methodology, applications and perspectives in chemistry', *Angew. Chem. Int. Ed. Engl.* **29**, 992–1023.

Venclovas, C. & Margelevicius, M. (2005), 'Comparative modeling in casp6 using consensus approach to template selection, sequence-structure alignment, and structure assessment', *Prot. Struct. Funct. & Bioinf.* **61**(Suppl 7), 99–105.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A. et al. (2001), 'The Sequence of the Human Genome', *Science* **291**(5507), 1304–1351.

Vose, M. D. (1999), *Simple Genetic Algorithm: Foundation and Theory*, MIT Press, Ann Arbor, M.I.

Šali, A. & Blundell, T. L. (1993), 'Comparative protein modelling by satisfaction of spatial restraints', *J. Mol. Biol.* **234**(3), 779–815.

Wallner, B. & Elofsson, A. (2003), 'Can correct protein models be identified?', *Prot. Sci.* **12**(5), 1073–1086.

Wang, Y., Zhang, H., Li, W. & Scott, R. A. (1995), 'Discriminating compact non-native structures from the native structure of globular proteins', *Proc. Natl. Acad. Sci. USA* **92**(3), 709–713.

Weiner, P. K. & Kollman, P. A. (1981), 'Amber: Assisted model building with energy refinement. a general program for modeling molecules and their interactions', *J. Comput. Chem.* **2**, 287–303.

Weiner, S. J., Kollman, P. A., Nguyen, D. T. & Case, D. A. (1986), 'An all atom force field for simulations of proteins and nucleic acids', *J. Comput. Chem.* **7**, 230–252.

Weiss, M. S., Jabs, A. & Hilgenfeld, R. (1998), 'Peptide bonds revisited', *Nature, Structural Biology.* **5**(8), 676–676.

Wilson, K. S., Dauter, Z., Lamsin, V. S., Walsh, M., Wodak, S., Richelle, J., Pontius, J., Vaguine, A., Sander, R. W. W., Hooft, V. G. et al. (1998), 'Who checks the checkers? Four validation tools applied to eight atomic resolution structures', *J. Mol. Biol* **276**, 417–436.

Winer, B. J., Brown, D. R. & Michels, K. M. (1991), *Statistical principles in experimental design*, 3 edn, New York: McGraw-Hill.

Wolpert, D. H. & Macready, W. G. (1997*a*), 'No free lunch theorems for optimization', *Evolutionary Computation, IEEE Transactions on* **1**(1), 67–82.

Wolpert, D. H. & Macready, W. G. (1997*b*), 'No free lunch theorems for search', *IEEE Transactions on Evolutionary Computation* **1**(1), 67–82.

Wolynes, P. G., Onuchic, J. N. & Thirumalai, D. (1995), 'Navigating the folding routes.', *Science* **267**(5204), 1619–1620.

Wright, A. H. (1991), 'Genetic algorithms for real parameter optimization', *Foundations of Genetic Algorithms* **1**, 205–218.

Yue, K., Fiebig, K. M., Thomas, P. D., Chan, H. S., Shakhnovich, E. I. & Dill, K. A. (1995), 'A test of lattice protein folding algorithms', *Proc. Natl. Acad. Sci. USA* **92**(1), 325–329.

Zagrovic, B., Snow, C. D., Khaliq, S., Shirts, M. R. & Pande, V. S. (2002), 'Native-like mean structure in the unfolded ensemble of small proteins', *J. Mol. Biol.* **323**(1), 153–164.

Zemla, A. (1999), 'Process and analysis of CASP3 protein structure predictions', *Prot. Struct. Funct. & Bioinf.* **S3**, 22–29.

Zemla, A. (2003), 'LGA: a method for finding 3d similarities in protein structures.', *Nucleic Acids Res.* **31**(13), 3370–3374.

Zhang, C., Liu, S., Zhou, H. & Zhou, Y. (2004), 'The dependence of all-atom statistical potentials on structural training database', *Biophysical Journal* **86**(6), 3349–3358.

Zhang, Y. & Skolnick, J. (2004*a*), 'Automated structure prediction of weakly homologous proteins on a genomic scale', *Proc. Natl. Acad. Sci. USA* **101**(20), 7594–7599.

Zhang, Y. & Skolnick, J. (2004*b*), 'Scoring function for automated assessment of protein structure template quality', *Nucleic Acids Res.* **57**(4), 702–710.

Zhang, Y. & Skolnick, J. (2004*c*), 'SPICKER: a clustering approach to identify near-native protein folds.', *J. Comput. Chem.* **25**(6), 865–871.

Zhang, Y. & Skolnick, J. (2006), 'Parallel-hat tempering: A Monte Carlo search scheme for the identification of low-energy structures', *J. Chem. Phys.* **115**(11), 5027–5032.

Zhou, H. & Zhou, Y. (2002), 'Distance-scaled, finite, ideal-gas reference state improves structure-derived optentials of mean force for structure selection and stabili ty prediction.', *Prot. Sci.* **11**(11), 2714–2726. Erratum PROTSCI 2003 sep 12(9):2121.

Zhou, Y., Zhou, H., Zhang, C. & Liu, S. (2006), 'What is a desirable statistical energy function for proteins and how can it be obtained?', *Cell biochemistry and biophysics* **46**(2), 165–174.

Zhu, J., Xie, L. & Honig, B. (2006), 'Structural refinement of protein segments containing secondary structure elements: local sampling, knowledge-based potentials, and clustering', *Prot. Struct. Funct. & Bioinf.* **65**(2), 463–479.

Zitzler, E., Deb, K. & Thiele, L. (2000), 'Comparison of multiobjective evolutionary algorithms: Empirical results', *Evolutionary Computation* **8**(2), 173–195.

Zitzler, E., Laumanns, M. & Thiele, L. (2002), Spea2: Improving the strength pareto evolutionary algorithm for multiobjective optimization, *in* 'Evolutionary Methods for Design, Optimisation, and Control', CIMNE, Barcelona, Spain, pp. 95–100.

Zitzler, E. & Thiele, L. (1998*a*), An evolutionary algorithm for multiobjective optimization: The Strength Pareto approach., Technical Report 43, Swiss Federal Institute of Technology (ETH), Zurich, Switzerland.

Zitzler, E. & Thiele, L. (1998*b*), Multiobjective Optimization Using Evolutionary Algorithms - A Comparative Case Study, *in* 'PPSN-V', Amsterdam, pp. 292–301.

Zitzler, E., Thiele, L., Laumanns, M., Fonseca, C. M. & Fonseca, V. (2002), Performance assessment of multiobjective optimizers: an analysis and review, Technical Report 139, Institut fur Technische Informatik und Kommunikationsnetze.

Zwanzig, R., Szabo, A. & Bagchi, B. (1992), 'Levinthal's paradox', *Proc. Natl. Acad. Sci. USA* **89**(1), 20–22.