Generating, maintaining and exploiting diversity in a memetic algorithm for protein structure prediction

Garza-Fabre, Mario; Kandathil, Shaun; Handl, Julia; Knowles, Joshua; Lovell, Simon

DOI:
10.1162/EVCO_a_00176

Generating, Maintaining and Exploiting Diversity in a Memetic Algorithm for Protein Structure Prediction

Mario Garza-Fabre mario.garza-fabre@manchester.ac.uk
Decision and Cognitive Sciences Research Centre, University of Manchester, Manchester, M15 6PB, UK

Shaun M. Kandathil shaun.kandathil@manchester.ac.uk
Faculty of Life Sciences, University of Manchester, Manchester, M13 9PT, UK

Julia Handl julia.handl@manchester.ac.uk
Decision and Cognitive Sciences Research Centre, University of Manchester, Manchester, M15 6PB, UK

Joshua Knowles j.knowles.1@cs.bham.ac.uk
School of Computer Science, University of Birmingham, Birmingham, B15 2TT, UK

Simon C. Lovell simon.lovell@manchester.ac.uk
Faculty of Life Sciences, University of Manchester, Manchester, M13 9PT, UK

Abstract

Computational approaches to *de novo* protein tertiary structure prediction, including those based on the preeminent 'fragment-assembly' technique, have failed to scale up fully to larger proteins (of the order of 100 residues and above). A number of limiting factors are thought to contribute to the scaling problem over and above the simple combinatorial explosion, but the key ones relate to the lack of exploration of properly diverse protein folds, and an acute form of 'deception' in the energy function whereby low-energy conformations do not reliably equate with native structures. In this paper, solutions to both of these problems are investigated through a multi-stage memetic algorithm incorporating the successful Rosetta method as a local search routine. It is found that specialised genetic operators significantly add to structural diversity and this translates well to reaching low energies. The use of a generalised stochastic ranking procedure for selection enables the memetic algorithm to handle and traverse deep energy wells that can be considered deceptive, which further adds to the ability of the algorithm to obtain a much-improved diversity of folds. The results should translate to a tangible improvement in the performance of protein structure prediction algorithms in blind experiments such as CASP, and potentially to a further step towards the more challenging problem of predicting the three-dimensional shape of large proteins.

Keywords

Protein structure prediction, fragment-assembly, memetic algorithms.

1 Introduction

Proteins are at the heart of cellular function, making possible most of the key processes associated with life. The three-dimensional structure of any given protein is known to be one of the major determinants of its distinctive functional properties.

Thus, protein structure determination is a fundamental step towards the understanding of the function of these important building blocks of life. Gaining insight into the structure-function relationship in proteins can assist, for example, in the design of proteins with novel specific functionalities, in the design of drugs and vaccines, and in the understanding of pathologies characterised by protein misfolding (*e.g.* Alzheimer's and Parkinson's diseases). Given the limitations of experimental methods, however, computational approaches have become the cornerstone of protein structure analysis.

Predicting the three-dimensional structure of a protein molecule, starting only from its amino acid sequence, remains a formidable challenge in computational biology. In recent years, different computational methods have been proposed with a view to tackling this problem, each leveraging observed relationships between amino acid sequences and three-dimensional structures of proteins. For example, the comparative or homology modelling approach (Martí-Renom et al., 2000) is rooted in the idea that proteins that are closely related in evolutionary terms (homologous proteins) are more likely to have very similar global sequences, and therefore very similar structures. However, these methods require the existence of a homologous protein with a known structure, and so they usually cannot be used to infer the structure of an entirely novel sequence. Other approaches, termed *de novo* or template-free modelling methods, seek to overcome this limitation by instead focusing on *local* sequence-structure relationships. These techniques are exemplified by the popular fragment-assembly class of approaches (Simons et al., 1997). Currently, fragment-assembly methods represent one of the most promising approaches to protein structure prediction, having shown a remarkable performance in the biennial critical assessment of protein structure prediction (CASP) experiments (Moult et al., 2014; Kryshtafovych et al., 2014).

Like many methods for protein structure prediction, fragment-assembly approaches use a simplified representation of a protein's tertiary structure, which includes information about backbone torsion angles only. However, the strength of fragment-assembly methods lies in their ability to leverage existing structural information from protein structure databases to reduce the search space. Specifically, local sequence-structure correlations are used to identify a finite set of candidate fragments for every window of residues in the protein chain. In this way, the continuous representation typically used in protein structure prediction is further compressed through the definition of a finite set of possible choices for each residue in the target protein. Fragment assembly thus remodels protein structure prediction as a combinatorial optimisation problem (Papadimitriou and Steiglitz, 1982; Cook et al., 1998), which involves the selection of one of the available fragment choices for each position in the protein chain (see Handl et al. (2012) for a Markov chain analysis of fragment assembly).

Despite the strengths of fragment assembly, there is a general consensus that limitations in fragment quality, inaccuracies of (especially low-resolution) energy functions and the size of the search space make *de novo* prediction a very difficult problem (Das, 2011; Kim et al., 2009). In particular, the prediction accuracy of fragment-assembly methods has been observed to decrease for larger proteins, and particularly those with high contact order (Kryshtafovych et al., 2014). Recent work aimed at improving the quality of search has taken its inspiration from method developments in the optimisation literature, and attempts at improving search performance have included state-of-the-art techniques such as evolutionary algorithms, replica-exchange methods and response surface methodologies (Bowman and Pande, 2009; Brunette and Brock, 2008; Simoncini et al., 2012). While such approaches have been reasonably successful in generating conformations with energy values that are lower than those obtained using,

for example, the well-established Rosetta protocol (Rohl et al., 2004), the corresponding improvements in prediction performance (measured commonly as the structural similarity to the native structure) have been small and often inconsistent, with deteriorations in accuracy observed for a subset of proteins. As a consequence, state-of-the-art approaches to fragment assembly continue to rely on thousands of restarts rather than a smaller number of runs of a more sophisticated search technique. This situation seems counterintuitive in the sense that random restarts are blind to the results of previous executions, which results in redundancies and does not provide an efficient approach to sampling. Yet, recent analysis of the search trajectories of longer Rosetta runs and an advanced sampling protocol indicate that they fail to capitalise on the power of the underlying search heuristics, and are unable to explore a significant number of different folds within a single search trajectory (Kandathil et al., 2016). Such limited exploration of the search space is particularly problematic given the known ruggedness of the energy landscape and its deceptive features, including known inaccuracies in the energy function that make it difficult to correctly differentiate between the quality of different local optima. In view of this, it seems that a successful search technique for fragment assembly will have to incorporate improved mechanisms to generate and retain low-energy structures that correspond to distinctly different folds.

This paper describes a new search heuristic for fragment assembly designed to address these distinctive challenges of the problem. First, given the lack of exploration observed in existing methods, and in view of the ruggedness which characterises the energy landscape, an improved sampling protocol is proposed. The new method aims to attain an appropriate balance between the exploration and the exploitation of the conformational search space. The ratio between exploration and exploitation is known to play a critical role in determining the performance of search algorithms (Črepinšek et al., 2013). In essence, the proposed method is a memetic algorithm (MA), embedding the successful Rosetta protocol as its local search heuristic. An important characteristic component of the proposed MA is the use of specialised genetic operators, which exploit problem-specific knowledge to explicitly encourage the exploration of more diverse protein folds. Secondly, it is the authors' view that, in light of the well-known inaccuracies of low-resolution energy functions (Bowman and Pande, 2009), the generation and preservation of a diverse range of candidate structures is a crucial prerequisite for achieving a robust and competitive performance. Drawing inspiration from the evolutionary multimodal optimisation literature (Das et al., 2011), a stochastic ranking-based procedure is implemented within the proposed MA as a mechanism to strike a trade-off between the optimisation of energy and the identification (and retention) of diverse conformations. The results of this study illustrate conclusively that enhanced optimisation of energy alone results in problematic sensitivity to the accuracy of the energy function, but significant improvements in robustness can be obtained through the explicit consideration of conformational diversity during the search.

The remainder of this paper is structured as follows. Section 2 provides an introduction to protein structure prediction, setting the necessary background for this work. Section 3 provides all details of the proposed memetic algorithm, including the specialised genetic operators implemented. Results for this algorithm demonstrate its ability to identify deep local optima within the search space, but also highlight the issues of optimising energy alone. Section 4 takes these results forward into the description and evaluation of an improved version of the algorithm that uses the stochastic ranking-based procedure to induce a robust balance between exploitation and exploration. Finally, Section 5 discusses the main findings of this study and concludes.

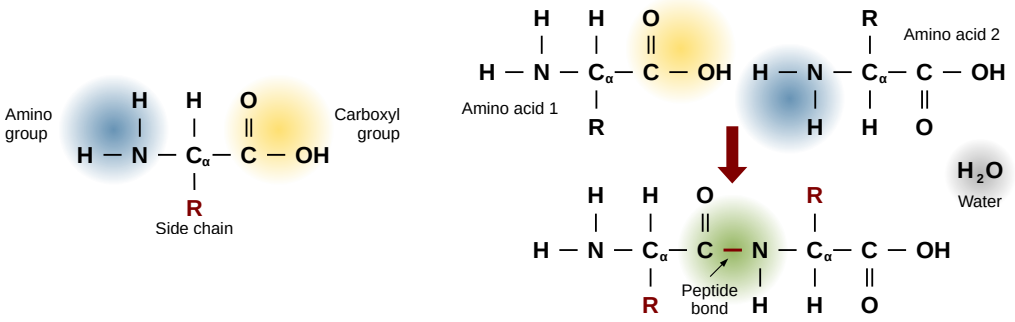

Figure 1: On the left, the general structure of amino acids is shown. On the right, the peptide bond formation process is illustrated. Each amino acid has a central carbon atom (C_α) which is covalently bonded to a carboxyl group (COOH), to an amino group (NH_2), to a hydrogen atom (H), and to a side-chain group denoted by R. There are 20 amino acids commonly found in proteins, each of which has a distinctive R group that is responsible for its particular chemical properties. The peptide bond is formed when the carboxyl group of an amino acid reacts with the amino group of another, releasing a water molecule. The elements of a protein chain are, therefore, amino acid *residues*.

2 Background

Notwithstanding the use of computer-intensive molecular dynamics simulations to model the dynamical *folding process* of proteins, protein (tertiary) *structure prediction* has been treated, fundamentally, since the seminal paper of Anfinsen (1973), as a problem of (static) energy minimisation. Progress in this approach has been gradual, and has been achieved through a combination of staged improvements to energy functions, the availability of greater and greater computational power, and a growing knowledge base of common structural forms. The latter includes the use of fragments of structure in what is termed 'fragment assembly', the preeminent approach to tackling 'new folds', *i.e.* proteins that fall outside a threshold of similarity to other known proteins. These gradual improvements have enabled larger and larger proteins to be 'solved' more accurately through computational prediction, but routine and reliable structure prediction of large proteins remains illusive.

2.1 Proteins

Proteins are fundamental elements of living cells, performing a range of biological functions. They are involved, for example, in transport, structural, enzymatic, hormonal, regulatory, and defensive processes. Amino acids, the building blocks of proteins, are small molecules that follow the general structure presented in Figure 1 (left side). Proteins are linear chains of amino acids, held together by *peptide bonds*, as illustrated on the right side of Figure 1. Hence, protein chains are also referred to as *polypeptides*.

Proteins display complex structures commonly described in terms of three main levels of organisation. The linear sequence of amino acid residues constitutes the *primary structure* of a protein. The *secondary structure* describes the arrangement of amino acids within short stretches of a polypeptide chain into motifs such as α*-helices* and β*-sheets*, connected to each other by chain regions called *loops*. The *tertiary structure* defines the overall folding of the protein chain in three-dimensions, where secondary structure elements are packed into compact domains. Tertiary structure is characterised

by the formation of long-range (high-order) contacts; this describes the fact that regions of the polypeptide chain which are a considerable distance apart in the sequence can be close together in three-dimensional space. The tertiary, three-dimensional conformation typically constitutes the functional state of the protein molecule.[1]

2.2 Protein Structure Prediction

Since the function of a given protein is determined to a large extent by its three-dimensional structure, direct experimental approaches have been used to determine protein structures in the laboratory. Prominent examples of such techniques are *X-ray crystallography* and *nuclear magnetic resonance* (NMR) *spectroscopy*. However, these experimental methods are often time-consuming and expensive, and in many cases require the development of experimental protocols specific to the protein of interest. The relative ease of obtaining protein sequence information has led to a large gap between the availability of sequence and structure information for proteins. As of 28 August 2015, the RCSB Protein Data Bank (PDB) has 111,558 structure entries (Berman et al., 2000; RCSB PDB, 2015), while the UniProtKB/TrEMBL protein sequence database contains 50,011,027 sequence entries (The UniProt Consortium, 2015a,b). The challenge to bridge such an ever-increasing gap has generated considerable interest in exploring computational approaches for protein structure prediction.

It is generally accepted that the amino-acid sequence encodes all the information related to the three-dimensional structure of a protein in a given environment. In other words, it is the specific configuration of amino acid residues in a protein which determines how it folds into a unique and compact three-dimensional conformation, often referred to as the *native state*. Among all the possible conformations that a protein can adopt, it is believed that its native state corresponds to the one with the lowest overall free-energy (Anfinsen, 1973). Hence, the process of inferring the functional, energy-minimising conformation for a protein molecule from its linear sequence of amino acids can be posed as an optimisation problem. This problem is commonly referred to as the *protein structure prediction* (PSP) problem, and represents one of the most active and challenging research areas in the field of computational biology. Let $\mathcal{X}$ be the set of all potential conformations for a given protein sequence, *i.e.* the search (or conformational) space, and let $E : \mathcal{X} \rightarrow \mathbb{R}$ denote a fixed energy model which maps each possible conformation $\mathbf{x} \in \mathcal{X}$ to an energy value $E(\mathbf{x})$. PSP can be more formally stated as the problem of finding the conformation $\mathbf{x}^* \in \mathcal{X}$ such that $E(\mathbf{x}^*) = \min\{\ E(\mathbf{x}) \mid \mathbf{x} \in \mathcal{X}\ \}$.

Methods originally aimed at solving PSP problems have found applications in other biological problem areas. For example, PSP methods have been used in the design of a highly specific endonuclease (Chevalier et al., 2002) and an enzyme that catalyses a Diels-Alder reaction (Siegel et al., 2010), a reaction not catalysed by any known enzyme in nature. The experimentally synthesised enzyme possesses the predicted structure and performs the predicted reaction. The structure prediction method Rosetta has been used in the design of a novel fold (Kuhlman et al., 2003) and the prediction and engineering of sites and affinity for DNA (Havranek et al., 2004) and ligand binding (Meiler and Baker, 2006; Wang et al., 2010). There are also important applications in drug and vaccine design (Whitehead et al., 2012; Azoitei et al., 2011) and medicine, especially in understanding the pathology of diseases such as Alzheimer's, Parkinson's and prion diseases (Chiti and Dobson, 2006). These diseases are characterised by the incorrect folding of certain proteins.

[1] Some proteins are composed of multiple polypeptide chains called subunits, the spatial arrangement of which is described by the protein's *quaternary structure*.

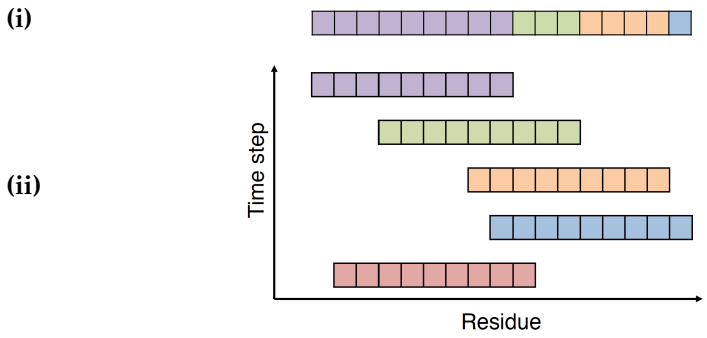

Figure 2: Schematic representation of the process of fragment assembly: **(i)** a short protein chain after five fragment insertions; and **(ii)** the sequence of fragment insertions required to generate the configuration shown in (i). Each coloured block represents structural information (torsion angle values) for a single residue. Individual fragments are coloured differently and are contiguous segments taken from other protein structures. Positions for attempted fragment insertions are chosen at random, and the acceptance of a move is usually governed by the Metropolis criterion. Note that prior fragment insertions can be partially or completely overwritten by subsequent ones.

2.3 Fragment-assembly Methods for Protein Structure Prediction

Methods employing fragment assembly have proven to be the most promising approaches to PSP, starting from only sequence information. They are consistently assessed as providing the best performance in the template-free modelling category of CASP experiments (Tai et al., 2014; Moult et al., 2014). Fragment assembly is based on the principle that, in proteins, the sequence of a short stretch of amino acids (*e.g.* 9 or 10 residues) has a strong influence on the local structure of that region (Simons et al., 1997; Han and Baker, 1996). Thus, it is possible to construct complete structures for a target protein, by assembling short fragments of local structure information. These fragments are typically drawn from proteins of known structure, based on similarity in amino acid sequence (Rohl et al., 2004; Gront et al., 2011). Each window of residues in the target has associated with it a set of fragments. Once a library of fragments has been generated for a given target, fragments are typically assembled using a Metropolis Monte Carlo optimisation scheme (Rohl et al., 2004). The key advantage of this approach is that it does not require the availability of structures for proteins with very similar global sequence (and therefore structure); this is sometimes referred to as template-free prediction. This has enabled fragment-assembly methods to correctly predict the structure of proteins with hitherto unobserved overall topologies (Kuhlman et al., 2003; Jones et al., 2005). Several fragment-based methods for PSP have been proposed over the years (*e.g.* Simons et al., 1997; Xu and Zhang, 2012; Jones, 2001; Lee et al., 2004).

Fragment-assembly methods typically employ a two-phase process. In the first phase, a low-resolution representation of the protein is used in order to rapidly explore the space of possible conformations. The purpose of this phase of prediction is to sample a wide range of plausible overall topologies (folds) for a given target sequence. New conformations are generated by replacing the structural information (configuration of backbone torsion angles) for a stretch of residues in the incumbent structure with that information in a fragment chosen from the fragment library. This process is termed *fragment insertion*, and is illustrated in Figure 2. Following each such insertion,

a relatively inexpensive knowledge-based scoring function is used to assess the quality of the new candidate structure, whose acceptance is commonly determined using the Metropolis criterion (Metropolis et al., 1953). The fragment insertion operation provides a means of exploring various possible conformations for a target. In other words, fragment insertion forms the search operator for the optimisation process. Moreover, fragment libraries describe what choices of structural information are available at any given residue. The set of fragments for a target protein defines, therefore, the search space of the optimisation problem, transforming the otherwise continuous space of structural parameters into a set of discrete configurations. Protein structure prediction, at this low-resolution fragment-assembly phase, is thus a combinatorial optimisation problem which can be stated simply as that of finding the best performing (*i.e.* energy-minimising) combination of the available fragment choices for each residue.

The low-resolution phase is designed to enable rapid conformational variation in the structure. It is in this phase that the overall fold of the structure is determined. The purpose of the second, all-atom phase of prediction is to convert the low-resolution structures into complete models of protein structure, including all side-chain atoms. In this phase, more computationally expensive procedures are used to refine the structure using small perturbations in order to arrive at compact, low-energy structures. This phase of the prediction process usually does not involve fragment assembly, and only performs small refinements to an already compact structure. Since this work is primarily concerned with methods for realising improved exploration of different overall folds, it focuses on the low-resolution fragment-assembly prediction step.

Section 2.4 details the low-resolution fragment-assembly phase of the standard Rosetta protocol, as this forms the basis for the methods proposed in this study. For more detailed descriptions of the Rosetta protocol, the reader is referred to Rohl et al. (2004); Gront et al. (2011); Misura and Baker (2005); Simons et al. (1997, 1999).

2.4 Rosetta

The low-resolution phase of Rosetta starts from a completely extended conformation of the protein chain. As mentioned above, a low-resolution representation of the protein structure is used. This is illustrated in Figure 3. In this representation, the atoms of the side-chain of each amino acid are abstracted as single pseudoatoms placed at the centroid of the side-chain. Bond lengths and bond angles are set to idealised values (these are derived by statistical analysis of experimentally derived structures (Engh and Huber, 1991)), and the backbone torsion angles (ϕ, ψ and ω in Figure 3) are the sole parameters that are varied during the optimisation process. The process of fragment insertion in Rosetta thus entails replacing the backbone torsion angles for a set of residues in the target protein chain with those drawn from a fragment. The structure generated as a result of each fragment insertion (or move) is evaluated using a scoring function (see below), and each move is either accepted or rejected based on the Metropolis criterion, using a fixed value of the temperature parameter, $kT = 2$. This process is repeated by selecting insertion windows in the target sequence uniformly at random, choosing a fragment to insert at that window, and evaluating the move at each step.

The low-resolution phase of Rosetta is composed of four stages. Stages 1 to 3 employ 9-residue fragments (9-mers), whereas stage 4 employs 3-mer fragments. Thus, two separate fragment libraries have to be composed prior to prediction, one for each fragment length considered. Each low-resolution stage also employs a different scoring function, which is a linear weighted sum of ten energy terms (Rohl et al., 2004). The scoring terms are mostly derived on the basis of Bayesian statistical analysis of known

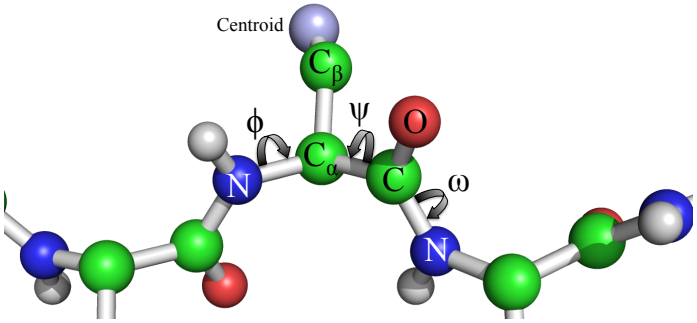

Figure 3: Low-resolution representation of the protein chain used in Rosetta. Atoms are represented as spheres, and bonds between atoms are represented by white sticks. The side-chain is approximated by a pseudoatom placed at its centroid. Backbone torsion angles ϕ, ψ and ω affect the rotation of atoms about the N–C_α, C_α–C, and C–N bonds, respectively. Bond lengths, and the angles between consecutive bonds, are set to idealised values. The torsion angles are the only parameters varied during optimisation.

protein structures (Simons et al., 1997, 1999), and are designed to capture local and global structural properties of real protein structures. These terms capture effects such as solvation (interactions with the protein's environment), and interactions between residues and elements of secondary structure. The weights of each of the scoring terms are gradually increased as Rosetta proceeds through the low-resolution stages, until the weights reach their final values in stage 4. Each low-resolution stage is allocated a certain budget of scoring function evaluations, and the default values can be increased or decreased using a user-supplied multiplier (the *increase_cycles* input parameter).

2.5 Recent Approaches to Improving Conformational Search

In Section 1, it was pointed out that most state-of-the-art structure prediction pipelines make use of a large number of independent restarts of the search heuristic, and that this constitutes an inefficient approach to exploration. In recent years, different methods have been proposed, seeking to leverage the strengths of existing approaches, such as Rosetta, while attempting to integrate more advanced optimisation schemes. For example, probabilistic search frameworks have been proposed, whereby the choice of a local region of conformational space to search is informed by the energy or score values obtained by conformations accessed in initial rounds of sampling (Simoncini et al., 2012; Simoncini and Zhang, 2013). An evolutionary algorithm employing crossover and mutation operators in a fragment-assembly context has also been described (Olson et al., 2013). Saleh et al. (2013) also describe the use of evolutionary and memetic algorithms involving fragment assembly, and show that the memetic algorithm can reduce the likelihood of deep descent into local optima on the energy landscape, which is known to be detrimental to predictive accuracy (Molloy et al., 2013), and has been an issue for some of the above approaches. Owing to inaccuracies in typical low-resolution scoring functions, some first methods have now started to consider structural diversity at strategic points during the search, *e.g.* using clustering or a grid structure (Shehu and Olson, 2010; Molloy et al., 2013). This is the work most closely related to this paper.

Note, however, that there are significant differences in our approach. As will be described in detail in Sections 3 and 4, the main innovative elements in the proposed method are: (i) the use of the successful Rosetta method as a local search operator, thus

drawing heavily upon the fragment libraries, energy functions, and sampling mechanisms already used in Rosetta, which facilitates a direct comparison; (ii) the use of specialised genetic operators that focus on loop regions of the protein chain, explicitly encouraging conformational space exploration (Kandathil et al., 2016); (iii) the incorporation of stochastic ranking as a mechanism to explicitly control the balance between exploration and exploitation during the search; and (iv) the use of novel conformational diversity measures, found during our ongoing work to effectively differentiate between compact structures with different folds (Garza-Fabre et al., 2015).

3 Rosetta-based Memetic Algorithm

The term *memetic algorithm* (MA) was originally coined by Moscato (1989). This term has been widely adopted in the literature to denote a broad class of metaheuristics that extend population-based methods, such as *evolutionary algorithms*, by incorporating problem-specific knowledge, usually in the form of a local search strategy or through the use of specialised search operators (Moscato and Cotta, 2003; Hart et al., 2005). An important number of successful applications of MAs have been reported, covering a wide range of optimisation problem classes and application domains.[2]

In this section, an MA is proposed in the context of the fragment-assembly approach to protein structure prediction. The proposed MA is first introduced in detail in Section 3.1. Then, Section 3.2 describes the experiments performed and presents the results of the evaluation of this method and its comparison with respect to Rosetta, a well-established and one of the most successful sampling protocols in this area.

3.1 Method Design

Our *Rosetta-based memetic algorithm* (RMA) is built upon the framework of *genetic algorithms* (Goldberg, 1989). Incorporation of problem-specific knowledge in the RMA relates to the use of Rosetta as a local improvement search strategy. In addition, the implemented genetic operators exploit information from secondary structure predictions as a means of boosting the exploration of the space of protein folds.

Algorithm 1 outlines the proposed RMA. The RMA consists of four consecutive stages, each of which based on the corresponding stage of the standard Rosetta protocol. In stage 1, lines 1 to 3 in Algorithm 1, an initial parent population is created by generating N fully extended conformations. Each parent individual is then independently processed using stage 1 of Rosetta. Given that van der Waals forces comprise the only information used at this stage to guide the search process, such an initialisation step is aimed at generating a potentially diversified set of valid protein structures. Stages 2, 3, and 4 of the RMA are based on the iterative procedure depicted in lines 5 to 14 of Algorithm 1. These stages differ in the implementation of distinct Rosetta stages as the local improvement strategy, which implies the use of different energy functions and fragment-insertion lengths (refer to Section 2.4 for details). The processing of each of these stages involves a total of G_{max} generations. During the first generation, a new population is obtained by improving the existing parent individuals through local search (based on the respective Rosetta stage). In contrast, subsequent generations produce an offspring population based on mating selection and the application of genetic operators (recombination and mutation). Offspring individuals are then subjected to local improvement. The resulting improved offspring compete against parent individuals in order to survive from one generation to the next (survival selection).

[2]Neri and Cotta (2012) present a review of MAs to address problems in discrete, continuous, large scale, constrained, and multiobjective optimisation, as well as in optimisation in the presence of uncertainties.

Algorithm 1 Rosetta-based Memetic Algorithm (RMA).

Require: Population size (N), number of generations (G_{max})
Ensure: Final population ($\mathcal{P}^*$)
1: $rma\_stage \leftarrow 1$
2: generate initial population $\mathcal{P}$ of fully extended conformations, $|\mathcal{P}| = N$
3: $\mathcal{P}^* \leftarrow rosetta\_local\_search(\mathcal{P},\ \ rosetta\_stage \leftarrow rma\_stage)$
4: **for** $rma\_stage \leftarrow 2$ **to** 4 **do**
5: **for** $generation \leftarrow 1$ **to** G_{max} **do**
6: **if** $generation = 1$ **then**
7: $\mathcal{P}^* \leftarrow rosetta\_local\_search(\mathcal{P}^*,\ \ rosetta\_stage \leftarrow rma\_stage)$
8: **else**
9: $\hat{\mathcal{P}} \leftarrow mating\_selection(\mathcal{P}^*)$
10: $\mathcal{P}' \leftarrow genetic\_operators(\hat{\mathcal{P}})$
11: $\mathcal{P}'^* \leftarrow rosetta\_local\_search(\mathcal{P}',\ \ rosetta\_stage \leftarrow rma\_stage)$
12: $\mathcal{P}^* \leftarrow survival\_selection(\mathcal{P}^* \cup \mathcal{P}'^*)$
13: **end if**
14: **end for**
15: **end for**

3.1.1 Mating Selection

In order to moderate selection pressure and to foster a good mixing of genetic material which favours exploration, the proposed RMA implements a random and panmictic mating selection strategy. That is, parents are selected in random order, without replacement, to form the pairs of individuals that are subjected to recombination. This allows each parent individual to be selected and considered as a source of genetic material exactly once. Offspring produced by recombination are then processed by the mutation operator and the local improvement strategy.

3.1.2 Genetic Operators

The genetic operators of the RMA, both recombination and mutation, exploit available information from secondary structure predictions in order to intensify exploration with regard to those regions of the protein chain that have been predicted to be loops.[3] The design of these operators is premised upon our belief that the increased exploration of the space of possible loop configurations can contribute significantly to the discovery and investigation of different protein folds during the search process. Thus, genetic operators have been implemented in such a way that their application affects only residues located at loop regions. This has the additional advantage of preserving the gains that have been achieved in terms of the optimisation of secondary structure elements.[4] Details on the recombination and mutation operators are provided next:

- The *loop-based recombination* operator takes two parent individuals as input. It produces as output two new offspring individuals by cloning the given parents and, based on a given probability p_c, interchanging the configuration (torsion angle values) of a randomly selected loop region between them. This operation is equivalent to the application of the well-known *two-point crossover* operator where the crossover points are set to the start and end residues of a loop region. The functioning of this operator is illustrated in Figure 4.

[3] Secondary structure predictions are based on PSIPRED 3.3 (Jones, 1999). Dependence on this information does not imply additional computational effort; it is derived during the construction of fragment libraries.

[4] Here, the term secondary structure element is used to refer to non-loop regions of the protein chain.

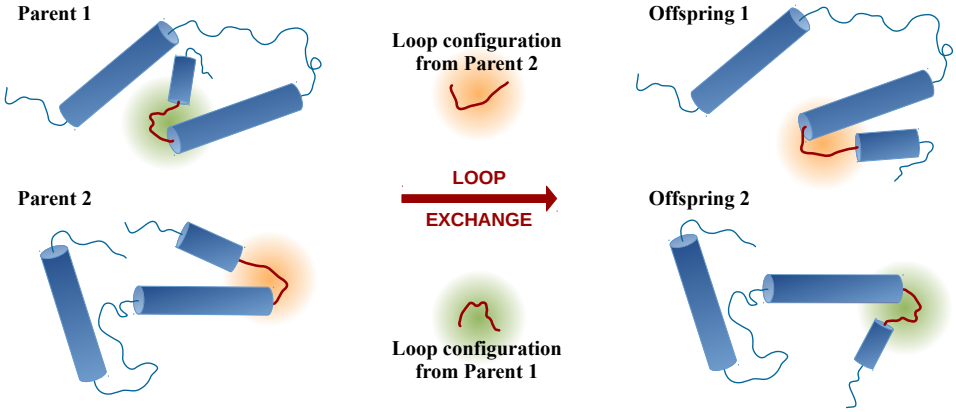

Figure 4: Illustration of the loop-based recombination operator. The two different parent conformations to the left are recombined by interchanging the configuration (torsion angle triplets) of all residues in a randomly chosen loop region. This produces the two offspring shown to the right, representing potentially novel folds.

- Mutation is based on fragment insertions, the same perturbations enforced by the standard Rosetta protocol (see Sections 2.3 and 2.4). During stages 2 and 3 of the RMA, 9-residue insertions are considered, while 3-residue insertions are used during stage 4. All offspring generated by recombination are subjected to mutation. This operator attempts a random fragment insertion at each insertion window, based on a given probability p_m. There exist a total of $W = (\ell - f + 1)$ windows at which a fragment can be inserted, where ℓ denotes the length of the protein sequence and f is the fragment length (*i.e.* $f \in \{3, 9\}$). Hence, by using, for instance, a mutation probability of $p_m = \frac{n}{W}$, $n \leq W$, about n fragment insertions can be expected to occur from the mutation of each individual. In order to constrain the application of this operator to loop regions, only the subset of $W_L \leq W$ insertion windows involving one or more loop residues is to be considered. Moreover, if mutation occurs at an insertion window spanning both loop and non-loop residues, the original configuration of all non-loop residues is preserved.

3.1.3 Survival Selection

At the end of all (but the first) generations in stages 2 to 4 of the RMA, all parent individuals and all produced offspring are considered to be candidates to survive and to form the next generation's population. This survival selection scheme is usually referred to as *plus (+) selection* in the context of evolution strategies (Beyer and Schwefel, 2002). In order to select survivors, the set of $|\mathcal{P}^* \cup \mathcal{P}'^*| = 2N$ individuals is ranked and the N top-ranked solutions are chosen. Here, the energy of the protein conformations encoded by candidate solutions is used as the only ranking criterion.

3.2 Experiments and Results

This section presents the experiments performed and the results obtained during the evaluation of the proposed RMA. First, Section 3.2.1 details the protein targets considered, the performance measures used, and the experimental conditions adopted. Then, Section 3.2.2 reports the results regarding the comparison of the RMA with respect to the standard Rosetta protocol. Finally, Section 3.2.3 discusses the relevance of exploiting secondary structure information within the design of the RMA's operators.

PDB	SS	ℓ	PDB	SS	ℓ	PDB	SS	ℓ	PDB	SS	ℓ	PDB	SS	ℓ	PDB	SS	ℓ
1acf	α-β	125	1cg5B	α	141	1fna	β	91	1lis	α	125	1tig	α-β	88	1wit	β	93
1bgf	α	118	1ctf	α-β	68	1gvp	β	87	1npsA	α-β	88	1tit	β	89	256bA	α	106
1bkrA	α	108	1dhn	α-β	121	1hz6A	α-β	61	1opd	α-β	85	1tul	α-β	102	2chf	α-β	128
1c8cA	α-β	62	1elwA	α	117	1iibA	α-β	103	1rnbA	α-β	109	1vcc	α-β	77	2ci2I	α-β	62
1c9oA	β	66	1eyvA	α	131	1kpeA	α-β	108	1ten	β	89	1who	β	94	2vik	α-β	122

Table 1: The set of 30 protein targets used in the experiments of this study. For each target, this table shows its PDB identifier, its secondary structure classification (SS), and the length of its corresponding amino acid sequence (ℓ).

3.2.1 Experimental Setup

- **Protein targets.** Table 1 lists the set of 30 protein targets used during the experiments conducted. These targets vary in length and secondary structure classification, as detailed in the table. Throughout this study, the different targets will be referred to by their corresponding identifiers from the PDB (Berman et al., 2000).

- **Performance measures.** As a performance measure, the energy of the protein conformations will be used, as computed by the low-resolution score function that Rosetta employs during stage 4 (see Section 2.4). In the experiments reported, energy values have been normalised to the range $[0, 1]$ based on the minimum and maximum values reached for each particular protein target, considering all different methods compared. Energy values are to be minimised in all the cases. Furthermore, a structural measure is used to evaluate the quality of the candidate conformations for a protein. Structural quality is usually assessed in the literature with respect to another reference structure, generally the native structure as determined by experimental methods. Perhaps the most well-known structural quality measure is the *root-mean-square deviation*, or RMSD for short. RMSD is defined as the square root of the mean of the squared deviations (distances) between corresponding atoms in the two structures compared. Calculation of this measure requires the structures to be aligned so that the RMSD is minimised, and it is this minimum RMSD value that is then reported. Such an alignment between the two structures is achieved by the Kabsch algorithm, which provides the correct rigid body rotation in all cases (Kabsch, 1976, 1978). In this study, we calculate RMSD using C_α atoms. Lower values for this measure, expressed in Ångströms (Å),[5] indicate better correspondence to the native structure.

- **Settings for the approaches compared.** During the experiments performed, both Rosetta and the RMA were run to generate a set of 1000 candidate conformations for each protein target. Therefore, a total of 1000 individual Rosetta trajectories were considered, each of which produced a single solution. Recommended settings for Rosetta were used in all the cases.[6] With regard to the RMA, a population size of $N = 100$ was always used. A single execution of the RMA produces a total of N solutions. Hence, 10 independent executions of the RMA were carried out, each using the equivalent computational effort of N trajectories of the standard Rosetta protocol. Additional control parameters of the RMA were set as follows: $G_{max} = 10$, $p_c = 0.1$, $p_m = \frac{1}{W_L}$.

[5] 1Å$= 10^{-10}m$.

[6] As recommended, the settings used for Rosetta are: $increase\_cycles = 10$; $rg\_reweight = 0.5$; $rsd\_wt\_helix = 0.5$; $rsd\_wt\_loop = 0.5$; $use\_filters = true$; $kill\_hairpins = [psipred\ file]$.

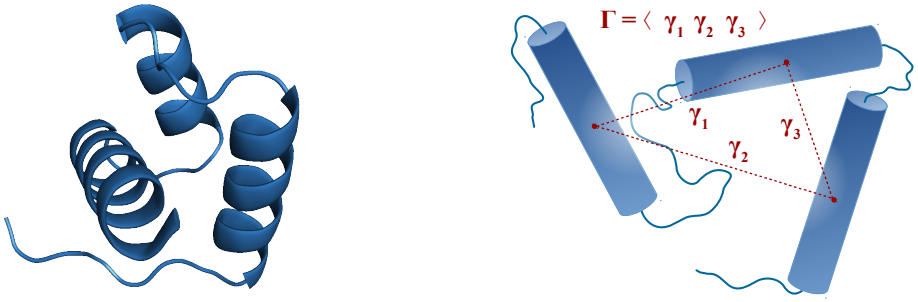

Figure 5: On the left, the figure shows the native conformation for protein 1enh (α protein with 54 amino acid residues). On the right, the computation of the Γ vector is illustrated for a hypothetical protein conformation with three different secondary structure elements. One distance value describes (coarsely) the relative position of each pair of these elements with respect to each other. Distances are computed between the C_α atoms of amino acid residues at the centre of the secondary structure elements.

- **Conformational space discretisation.** The analysis presented in Section 3.2.3 inquires into the ability of the RMA to explore the conformational search space. A small protein, 1enh, and a new measure of conformational diversity are used for this sake, see Figure 5.[7] As illustrated in this figure, a given protein conformation can be coarsely described by its Γ vector, which is composed by a set of distances that account for the relative position of secondary structure elements with respect to each other. A total of $\binom{E}{2}$ distance values define a Γ vector for a protein with E secondary structure elements. Thus, three-dimensional vectors $\Gamma = \langle \gamma_1 \ \gamma_2 \ \gamma_3 \rangle$ are used here to represent folded states for protein 1enh. From this, it follows that the maximum value that can be reached for each dimension γ_i, $i \in \{1, 2, 3\}$ can be derived by computing this measure from a fully extended conformation. Preliminary testing on this particular target indicated that the bulk of conformations explored during the search process produce γ_i values within 50% of this upper bound. Therefore, we focus on the range from 0 to 50% of the maximum possible γ_i values and, in all the cases, this range has been split into a total of 100 subranges in order to achieve a discretisation of the space of all possible Γ vectors. Despite the conceptual simplicity of this approach, distances between secondary structure elements have been found in the authors' ongoing work to be important descriptors of the folded state of a protein molecule (Garza-Fabre et al., 2015).

3.2.2 Comparison with the Standard Rosetta Protocol

In order to investigate the suitability of the RMA, it is imperative to evaluate this method with respect to the standard Rosetta protocol. This is not only due to the fact that Rosetta is widely accepted as representative of the state-of-the-art, having shown a remarkable performance among existing fragment-assembly methods for protein structure prediction; but also, the RMA is based to a large extent on Rosetta, as detailed in Section 3.1. Therefore, a comparison to Rosetta is essential for highlighting the relevance of all other components of the RMA to which Rosetta has been coupled. The results of this comparison are shown in Figure 6, which contrasts the energy and RMSD

[7] Protein 1enh has only three secondary structure elements (all α-helices) separated by two loop regions. This reduced number of secondary structure elements makes this protein a suitable candidate for the analysis conducted, in contrast to all targets listed in Table 1 which involve four or more secondary structure elements.

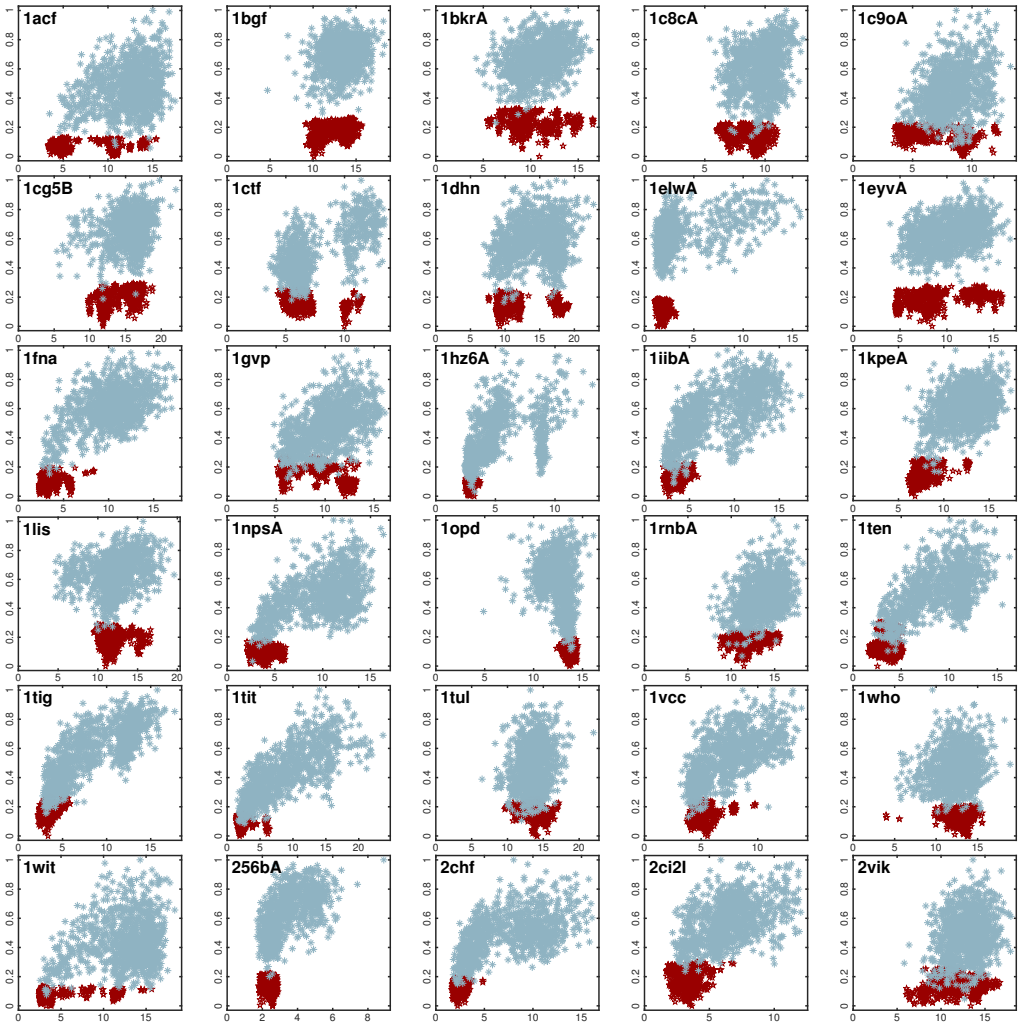

Figure 6: Results achieved by Rosetta (light blue) and the RMA (red) on the set of 30 protein targets listed in Table 1. Each plot contrasts the results scored by the two algorithms in terms of the RMSD (x-axis) and energy (y-axis) criteria. RMSD values are expressed in Ångströms (Å). Energy values have been normalised to range $[0, 1]$ for visualisation purposes. Both energy and RMSD are to be minimised in all the cases.

values scored by Rosetta and the RMA on a set of 30 protein targets (for table-format results, the reader is referred to Section B of the supplementary material).

From Figure 6, it is possible to observe that the RMA exhibits a better performance than Rosetta with regard to energy, a behaviour that remains consistent across all the 30 protein targets considered. The RMA was not only able to produce lower energies than Rosetta, but also to 'narrow' the distribution of energy values scored in all the cases. Given that energy is the only optimisation criterion used to guide the search process, this performance in terms of energy is indicative of the effectiveness of the proposed search strategy. By focusing now on the results for the RMSD measure, the figure shows that for a number of proteins (namely, 1elwA, 1fna, 1hz6A, 1iibA, 1npsA, 1ten, 1tig, 1tit, 256bA, 2chf, and 2ci2I; proteins are referred to throughout by their PDB identifiers),

most (or all) of the solutions generated by the RMA potentially correspond to native-like conformations (as suggested by the low RMSD values below 5Å). For all these targets, note that Rosetta was also able to produce conformations providing low RMSD values during the executions performed. The RMA, however, has clearly increased the likelihood of sampling and better populating those low-RMSD regions of the conformational space. There are some other targets such as 1kpeA and 1vcc for which, in spite of a better *average* performance and overall tendency shown by the RMA (*i.e.* a narrow distribution of points at low energy and low RMSD regions), slightly lower RMSD values have been reached by a few of the Rosetta trajectories. An interesting characteristic common to all of the targets discussed so far is that the energy function seems to be 'well-aligned' to the RMSD measure; that is, for all these targets, lower energies tend to be associated with lower RMSD values, as can be appreciated from the plots. Therefore, in such a favourable scenario, a search strategy that is effective at optimising the energy function can certainly be as effective at locating native-like conformations.

However, the results also reveal that this important alignment between the energy function and the RMSD measure is not as evident when considering some of the remaining protein targets. A clear example of this is protein 1opd, whose corresponding plot in Figure 6 exhibits an almost opposite correlation between the two criteria. As a consequence, an outstanding performance in terms of energy does not necessarily translate into an acceptable performance in terms of RMSD. In the presence of well-known inaccuracies of energy functions (Bowman and Pande, 2009), even the most successful methods for searching the huge conformation space might fail at identifying, preserving and exploiting native-like conformations. Nevertheless, the results also underline that Rosetta's low-resolution energy functions are sufficiently informative to allow the search process to identify different compact (protein-like) folds. For several targets (1acf, 1c9oA, 1ctf, 1eyvA, 1gvp, 1wit, and 2vik), there are some structures that fall within 5 to 6 Å RMSD of the native structure, and for some of these, the RMA seemed to be able to exploit a small energy gradient thus improving the relative frequency of such structures compared to the results of Rosetta. In contrast, Rosetta was more effective at producing lower RMSD values for a series of targets, including 1bgf, 1cg5B, 1lis, and 1opd. This appears to be because the energy gradient for these targets is misleading, and Rosetta's local convergence lends it robustness in this setting.

Finally, it is worth noting from the plots for several targets, namely, 1acf, 1bkrA, 1c9oA, 1ctf, 1dhn, 1eyvA, 1gvp, 1wit, and 2vik, that the distribution of RMSD values obtained by the RMA covers a wide range (with deviations of more than 10Å in most cases) despite the low variation with regard to energy. Such a dispersion in the distribution of RMSD values could be explained, to a certain extent, by the multimodal nature of the energy landscape. Multiple solution clusters shown in these plots can potentially represent different attraction basins. The plots for proteins 1ctf and 1dhn, for instance, suggest the existence of at least two well-defined local basins where the efforts of the RMA were concentrated. Rosetta's results for these targets (as well as for 1hz6A, 1iibA, 1ten, and 1tig) provide additional evidence of this. In the presence of multiple, equally fit energy basins, or when the basins where native-like conformations reside are not rewarded over non-native basins, there is no mechanism that can be exploited in order to assist search algorithms in identifying the most promising basins.[8] This motivates the analysis presented in Section 4, where an alternative selection strategy is equipped into the RMA as a means of encouraging diversification and the retention of a population of solutions that can span multiple basins reached throughout the search process.

[8] In a blind prediction scenario, information from native states, and therefore from RMSD, is unavailable.

3.2.3 Genetic Operators and Secondary Structure Information

The purpose of this section is to discuss the role of the implemented genetic operators, and how the incorporation of secondary structure prediction information within these operators contributes to the exploration behaviour of the proposed RMA.

our variants of the RMA are analysed and compared: (i) RMA without genetic operators, *i.e.* setting $p_c = 0$ and $p_m = 0$;[9] (ii) RMA with standard genetic operators; (iii) RMA with loop-based operators as described in Section 3.1.2; and (iv) RMA with loop-based operators but using incorrect secondary structure information. For the second variant of the RMA, using standard operators, the two-point crossover operator is utilised.[10] The standard mutation is similar to the loop-based mutation introduced in Section 3.1.2, but takes into account all W possible insertion windows (using $p_m = \frac{1}{W}$, consequently) and it is also allowed to affect the configuration of non-loop residues. Therefore, both standard and loop-based operators are functionally similar, differing in whether they can be applied to all residues or they operate on loop regions only. The fourth RMA variant is considered in this analysis to illustrate the crucial role that the quality of the secondary structure predictions plays in the effectiveness of the proposed loop-based operators. In the incorrect secondary structure information provided to this RMA variant, loop regions (as in the real secondary structure information, derived from the native structure) are identified as secondary structure elements and secondary structure elements are identified as loops, representing the worst possible prediction scenario (in terms of loop region identification).

The experiment conducted investigates the extent to which different regions of the conformational space are reached and exploited throughout the search process. This analysis focuses on a small protein target, 1enh, and it is based on the discretisation of the conformational space described at the end of Section 3.2.1. As the result of this analysis, Figure 7 reports the frequency with which each region of this discrete space has been considered while using the four above-mentioned variants of the RMA.[11]

As shown in Figure 7, the first variant of the RMA tends to over-emphasise the exploitation of very compact regions of the conformational space. This is due to the high selection pressure which results from the use of an elitist and extinctive selection scheme, and the lack of genetic operators which are essential mechanisms for promoting exploration. It is evident from the figure that the implementation of standard operators (second RMA variant) has contributed significantly to achieve an increased exploration. Note, however, that the incorporation of problem-specific knowledge into these operators has further improved the exploration capabilities of the algorithm. This can be seen from the results of the third variant of the RMA which uses specialised, loop-based operators. This confirms the underlying hypothesis that guided the design of these operators: that the configuration of loop regions is a major determinant of the arrangement and packing of secondary structure elements (in a fragment-based prediction context), so that the increased exploration in terms of loop configurations should contribute to intensifying the exploration of the space of potential conformations.

[9] Exploration behaviour of the first RMA variant is equivalent to that of Rosetta; the perturbations enforced by the Rosetta-based local search are the only mechanisms used for accessing new candidate conformations.

[10] As stated in Section 3.1.2, loop-based crossover is equivalent to two-point crossover using the first and last residues of a loop as the crossover points. To enable a more reliable analysis, the minimum and maximum separation between the crossover points in the standard operator were set to the minimum and maximum length of a loop region for the protein considered, respectively. This avoids that the differences to be observed in search behaviour can be attributed to the use of different magnitudes for the applied perturbations.

[11] A single execution of each variant of the RMA was performed using a population size of $N = 100$. All candidate solutions accepted during the search process are covered by the results reported. Solutions rejected during the application of the Rosetta-based local improvement strategy have been discarded in all the cases.

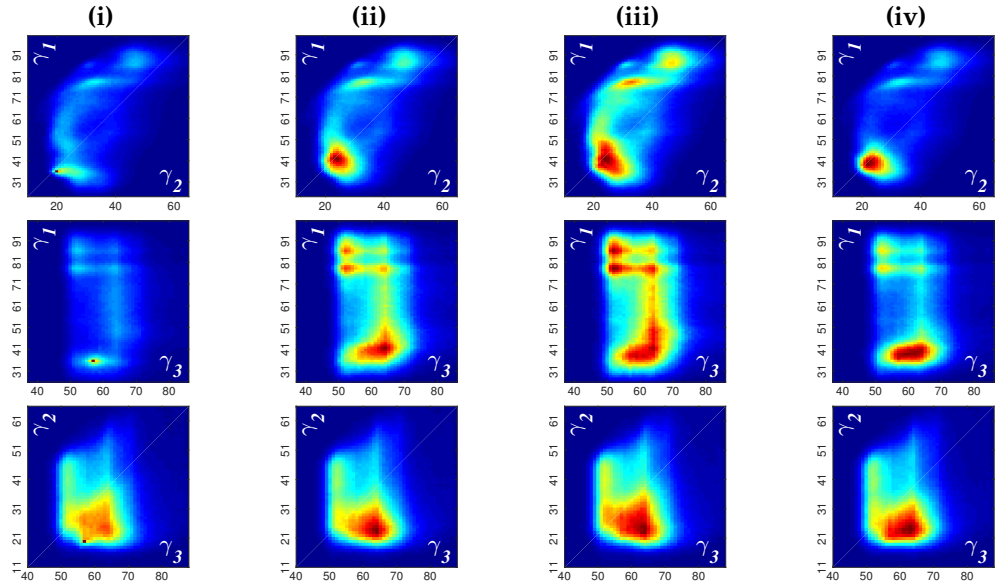

Figure 7: Illustrating the impact of the genetic operators and the exploitation of secondary structure information. Dimensions γ_i, $i \in \{1, 2, 3\}$ of the discretised Γ space for protein 1enh are analysed pairwise. Plots show the frequency with which each region of such discrete space was explored (darker reds refer to higher frequencies). Each column of plots presents the results for a different RMA configuration: **(i)** without using genetic operators; **(ii)** using standard operators; **(iii)** using loop-based operators; and **(iv)** using loop-based operators relying on erroneous secondary structure information.

On the other hand, the results of the fourth RMA variant confirm that the utilisation of erroneous information can be detrimental and may produce even worse results than the use of no information. The fourth (misinformed) RMA variant reports a decreased exploration performance when compared to the second variant which uses standard (uninformed) operators. This is particularly evident when analysing dimension pairs (γ_1, γ_2) and (γ_1, γ_3) of the discretised conformational space. An additional experiment was carried out to investigate whether the results that the RMA exhibited in Section 3.2.2 can be improved by providing this method with the real (rather than the predicted) secondary structure information. No noticeable differences in performance were found. Although the secondary structure predictions used are not completely accurate, these findings suggest that the location and elongation of loop regions, the only information exploited by the RMA, is accurate enough overall.[12] Results of this experiment, as well as detailed information about the accuracy of the secondary structure predictions used, can be found in Section C of the supplementary material.

4 Improving Robustness to Deal with Inaccuracies of Score Functions

In Section 3.2.2, it was found that the original implementation of the RMA, employing energy as the only selection criterion, was effective at producing lower energy and RMSD values than the standard Rosetta protocol for a significant number of the pro-

[12]The average secondary structure prediction accuracy for α, β, and loop residues, and overall 3-state accuracy of targets considered, are $Q_{\alpha}^{\%obs} = 82.62\%$, $Q_{\beta}^{\%obs} = 84.80\%$, $Q_{L}^{\%obs} = 83.01\%$, and $Q_{total} = 83.91\%$, respectively (Rost and Sander, 1993). Further details can be found in Section C of the supplementary material.

tein targets considered. It was also found, however, that existing inaccuracies in the scoring functions have prevented the RMA from consistently generating native-like (low-RMSD) conformations. It is therefore necessary to devise strategies which can increase the robustness of the proposed method in order to cope with this issue.

Given the lack of correlation between the energy and RMSD criteria, and recognising also the multimodal nature of the PSP problem, this section explores an alternative selection mechanism which aims to provide an effective means of regulating selection pressure and enhancing the diversity preservation capabilities of the RMA. This mechanism is motivated by our belief that, in the presence of inaccurate energy functions, the generation of a diverse set of candidate conformations, that can potentially span multiple global and local energy-minima, can constitute a more robust approach (*i.e.* with better chances of achieving native-like structures) than to simply succeed in producing a single global minimum (the general tendency of evolutionary optimisation methods (Shir et al., 2010; Das et al., 2011)). This follows a similar approach of explicit diversity maintenance adopted in some works (Sastry et al., 2005; Goldberg et al., 1992) on dealing with deceptive 'trap' functions (Goldberg, 1987, 1992), or other functions that tend to lead an optimiser away from the best configurations (Watson et al., 1998).

As a population-based approach, the RMA possesses the natural advantage that multiple solutions can be reached during a single execution; in contrast to single-solution-based methods (*e.g.* Rosetta) that rely on performing multiple individual executions (or restarts) in the hope that each of them can discover a different solution. Nevertheless, the RMA needs to be equipped with effective mechanisms to foster the generation and preservation of a diversity of protein conformations in its population.

This section proceeds as follows. In Section 4.1, the alternative selection technique is introduced and its implementation details in the context of the RMA are described. The suitability of this strategy is then evaluated in Section 4.2 with respect to the basic, energy-based RMA and with respect to the standard Rosetta protocol.

4.1 Stochastic Ranking-based Survival Selection

Runarsson and Yao (2000) introduced a rank-based selection mechanism as a means of dealing with constrained optimisation problems. This mechanism, which they called *stochastic ranking*, employs a bubble-sort-like procedure to rank a population of candidate individuals for selection purposes. The characteristic feature of this strategy is the incorporation of a user-defined parameter that represents the probability of using either one or the other of two criteria, namely an objective function and a penalty function (sum of constraint violation), each time a pair of solutions is compared (in order to determine which one is fitter) during the ranking process. Such a parameter, therefore, removes the dependence on hard-to-tune penalty factors, allowing the user to control the balance between the two criteria in order to prevent over- and under-penalisation scenarios. The promising behaviour exhibited by stochastic ranking motivated further research around this proposal, and a number of works have been reported based on this constraint-handling technique (Mezura-Montes and Coello Coello, 2011).

It is possible, nevertheless, to use a similar procedure to achieve the desired balance when considering two arbitrary selection criteria.[13] Algorithm 2 presents such a generalised version of stochastic ranking. It ranks a list of solutions based on a given pair of criteria c_1 and c_2 that, without loss of generality, are assumed to be minimised. The core of the ranking process is detailed in lines 3 to 22 of Algorithm 2, the application of which can be referred to as a 'sweep'. During a sweep, ranking occurs by

[13]Note that this approach could potentially be extended to consider also more than two criteria.

Algorithm 2 Generalised stochastic ranking-based procedure.

Require: Solution list ($H \leftarrow \langle \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_M \rangle$), rank criteria ($c_1, c_2$), bias parameter ($\rho$)
Ensure: Ranked list of solutions (H^*)

```
 1: H* ← H
 2: for i ← 1 to I do
 3:    for j ← 1 to M − 1 do
 4:       chose r uniformly at random in range [0, 1]
 5:       if c1(H*_j) = c1(H*_{j+1}) and c2(H*_j) > c2(H*_{j+1}) then
 6:          swap H*_j and H*_{j+1}
 7:       else if c1(H*_j) > c1(H*_{j+1}) and c2(H*_j) = c2(H*_{j+1}) then
 8:          swap H*_j and H*_{j+1}
 9:       else if c1(H*_j) = c1(H*_{j+1}) and c2(H*_j) = c2(H*_{j+1}) and r ≤ 0.5 then
10:          swap H*_j and H*_{j+1}
11:       else
12:          if r ≤ ρ then
13:             if c1(H*_j) > c1(H*_{j+1}) then
14:                swap H*_j and H*_{j+1}
15:             end if
16:          else
17:             if c2(H*_j) > c2(H*_{j+1}) then
18:                swap H*_j and H*_{j+1}
19:             end if
20:          end if
21:       end if
22:    end for
23:    if no swap operation occurred then
24:       break
25:    end if
26: end for
```

iteratively comparing adjacent individuals. First, lines 5 to 10 stand for the cases where the competing solutions are *indifferent*[14] either with respect to a single criterion, so that the remaining criterion is always used to discriminate between them, or with respect to both criteria, in which case a random decision is made. The general case is described in lines 12 to 20, where the bias parameter ρ denotes the probability of adopting either c_1 or c_2 as the underlying discrimination criterion. The ranking process is completed by applying (at most) I sweeps. Note, however, that this iterative process can stop earlier if no change in the rank ordering of solutions occurs within a complete sweep; this was done in the original implementation of stochastic ranking (Runarsson and Yao, 2000).

As pointed out by Runarsson and Yao (2000), when the maximum number of sweeps I approaches ∞, the ranking will be determined by the criterion favoured by the setting of parameter ρ. Therefore, the right selection bias can be equivalently achieved either by adjusting ρ or by increasing I in response to a given ρ value.[15] By fixing I, parameter ρ thus becomes solely responsible for regulating the strength of the bias. In this study, a value of $I = M$ is considered, where M is the number of solutions to be ranked. Figure 8 illustrates the effectiveness of this method for introducing a bias in the selection process through the use of different values for parameter ρ. As can be seen from this figure, a value of $\rho = 0.5$ provides the best trade-off between the criteria. It can also be seen that even subtle deviations from this value are capable of producing significant biasing effects in order to emphasise one criterion over the other.

[14] In this study, two solutions $\mathbf{x}_1$ and $\mathbf{x}_2$ are said to be indifferent if and only if they present exactly the same value with respect to the evaluation criterion under consideration.

[15] Increasing the number of sweeps I, however, would raise the computational effort required for ranking.

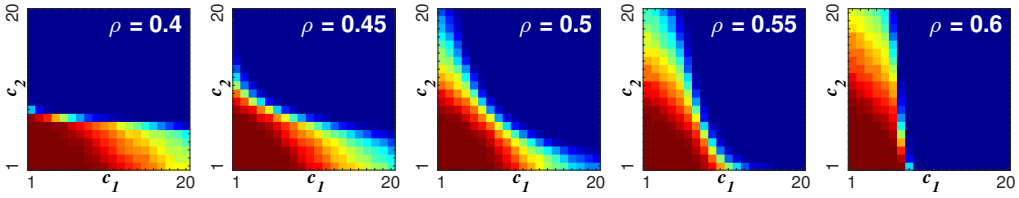

Figure 8: Bias of the stochastic ranking-based selection, as induced by different values of parameter $\rho \in \{0.4, 0.45, 0.5, 0.55, 0.6\}$. A total of 400 solution points represent all possible combinations of values for two hypothetical criteria $c_1, c_2 \in \{1, 2, \ldots, 20\}$. Solutions were ranked, and the top 25% of the solutions based on the ranking obtained were selected. For each ρ value, a total of 1000 independent repetitions of this experiment were performed. Plots show the frequency with which each of the criteria configurations was selected, where darker reds refer to higher selection frequencies.

4.1.1 Selection Criteria and Implementation within the RMA

The above-described stochastic ranking procedure has been adapted in order to replace the basic energy-based survival selection scheme of the RMA (see Section 3.1.3). The following two selection criteria have been considered in this study:

$$c_1(\mathbf{x}) \;=\; energy(\mathbf{x}) \;, \tag{1}$$

$$c_2(\mathbf{x}) \;=\; diversity(\mathbf{x}) \;. \tag{2}$$

The first selection criterion focuses on the quality of individuals, which is measured as the energy of the encoded conformations. It is worthwhile to remember that different energy functions are used by the RMA at different stages of the search.

The second selection criterion evaluates individuals according to their contribution to diversity. The implemented diversity measure operates in phenotype space and aims to improve the exploration and preservation of different protein folds throughout the search process. This measure employs a more-fine grained version of the strategy originally introduced in Section 3.2.1, which requires the computation of the Γ vector as a means of describing the folded state of a protein conformation. As illustrated in Figure 9, the Γ vector accounts for the relative position of secondary structure elements with respect to each other. For each pair of secondary structure elements, four distances are included in vector Γ. Thus, for a protein with a total of E secondary structure elements, the length of the corresponding Γ vector is:

$$G \;=\; 4\tbinom{E}{2} \;=\; 2E^2 - 2E \;. \tag{3}$$

Once the Γ vectors have been computed for all of the candidate solutions to be ranked, the diversity contribution of an individual $\mathbf{x} \in \{\mathcal{P}^* \cup \mathcal{P}'^*\}$ is given by the minimum *root mean square error* (RMSE) between the Γ vector of $\mathbf{x}$ ($\Gamma_{\mathbf{x}}$) and that of another solution $\mathbf{x}'$ within the same set ($\Gamma_{\mathbf{x}'}$). Formally:

$$diversity(\mathbf{x}) \;=\; \min\left\{ \text{RMSE}(\Gamma_{\mathbf{x}}, \Gamma_{\mathbf{x}'}) \;\mid\; \mathbf{x}' \in \{\mathcal{P}^* \cup \mathcal{P}'^*\}, \mathbf{x}' \neq \mathbf{x} \right\} , \tag{4}$$

where

$$\text{RMSE}(\Gamma_{\mathbf{x}}, \Gamma_{\mathbf{x}'}) \;=\; \sqrt{\tfrac{1}{G}\textstyle\sum_{i=1}^{G}\left(\gamma_i^{\mathbf{x}} - \gamma_i^{\mathbf{x}'}\right)^2} \;. \tag{5}$$

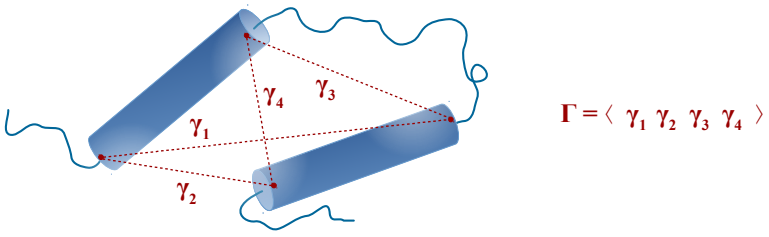

Figure 9: Computation of the Γ vector for a hypothetical protein conformation with two secondary structure elements. Four distances describe the relative position of these elements with respect to each other. Distances are computed to and from the C_α atoms of the first and last amino acid residues of the secondary structure elements.

Note that whereas the energy values used as the first selection criterion are to be minimised, the second, diversity-based criterion is to be maximised. In this way, the aim of incorporating the stochastic ranking-based procedure into the RMA is to achieve a suitable balance between quality and structural diversity in order to drive selection.

An elitist step was introduced in order to prevent the loss of promising solutions regardless of the value chosen for ρ. Initially, the lowest-energy solution from the pool of $|\mathcal{P}^* \cup \mathcal{P}'^*| = 2N$ individuals is selected. All other individuals are ranked on the basis of stochastic ranking in order to determine the remaining $N - 1$ survivors.

4.2 Experimental Results

This section investigates the advantages of replacing the original, energy-based selection mechanism of the RMA with the stochastic ranking procedure introduced in the previous subsection. The two versions of the RMA are compared with respect to each other, and with respect to Rosetta, on a set of 30 protein targets, see Table 1. Results are evaluated in terms of the energy of candidate conformations and the RMSD measure. Details of these criteria, and all settings and experimental conditions adopted for Rosetta and the RMA, have already been defined in Section 3.2.1. Three different values for the parameter of stochastic ranking are analysed, $\rho \in \{0.45, 0.5, 0.55\}$. The results for the energy and RMSD criteria are respectively presented in Figures 10 and 11 (for table-format results, refer to Section B of the supplementary material).

As expected, the use of the stochastic ranking strategy reduced selection pressure with respect to energy. Figure 10 exhibits a clear and consistent tendency in the energy values obtained by the four variants of the RMA. Higher energies tend to be produced as parameter ρ is decreased in order to moderate the emphasis given to the energy criterion (over the diversity criterion) during selection.[16] Note, however, that the energies scored by most configurations of the RMA are significantly better than those obtained by the standard Rosetta protocol. The only exception is the configuration using $\rho = 0.45$, which achieved comparable or worse results than Rosetta in most cases.

The results for the RMSD measure highlight the suitability of the stochastic ranking selection for enhancing the robustness of the RMA. As can be seen from Figure 11, the consideration of structural diversity as an additional criterion to guide selection has increased the likelihood of the RMA reaching and preserving more native-like (low-RMSD) conformations for a number of protein targets (in comparison to the use of

[16]Note that the use of the conventional energy-based selection in the RMA, is equivalent to the use of the stochastic ranking-based selection by setting parameter ρ to its maximum possible value, $\rho = 1$.

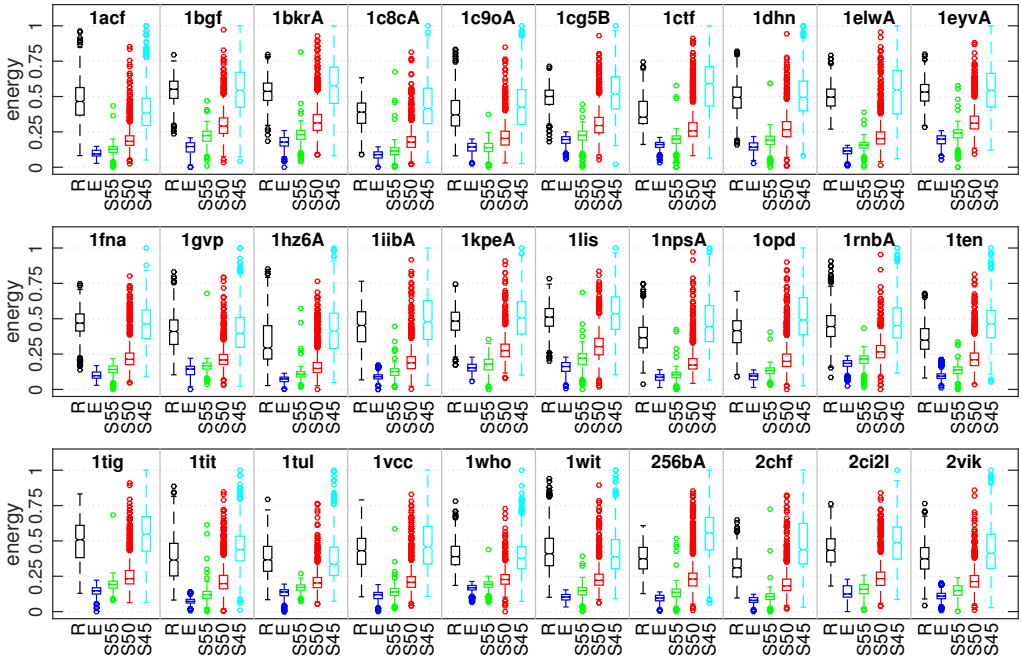

Figure 10: Energy values scored by the standard Rosetta (R) protocol and the proposed RMA using different selection strategies: energy-based selection (E); and stochastic ranking-based selection, for $\rho \in \{0.45, 0.5, 0.55\}$ (denoted respectively as S45, S50, and S55 in the plots). A total of 30 different protein targets is considered. Energy values, to be minimised, have been normalised to range $[0, 1]$ for visualisation purposes.

energy only). The use of $\rho = 0.5$, which provides a balanced trade-off between the two selection criteria, seems to produce the most competitive performance in most cases. The stochastic ranking selection allowed the RMA to outperform the lowest RMSD values achieved by Rosetta in several targets; *e.g.* 1cg5B, 1ctf, 1kpeA, 1who, and 2vik. Conversely, though improving central tendencies in most cases, no configuration of the RMA was able to reach the minimum RMSD values produced by Rosetta for some other targets, such as 1bkrA, 1eyvA, 1lis, 1opd, and 1tul. Such minimum RMSD structures reached by Rosetta tend to be associated with higher energies, as observed from Figure 6, and are therefore difficult for the RMA to retain. A further point of interest relates to the performance that the stochastic ranking-based RMA exhibits when dealing with targets for which the energy function seems to be well-correlated with the RMSD measure (1elwA, 1fna, 1hz6A, 1iibA, 1npsA, 1ten, 1tig, 1tit, 256bA, 2chf, and 2ci2I, see Figure 6). As found in Section 3.2.2, a selection based solely on energy is able to produce satisfactory results under such a scenario. The competitive (and to a certain extent comparable) performance shown by the stochastic ranking-based RMA, illustrates the ability of this strategy to maintain an acceptable degree of success on such moderate-difficulty targets, while having to be more robust under more challenging conditions.

Finally, a relevant subject of analysis concerns the availability of native-like fragments. Fragment-assembly methods rely on the existence of native-like configurations in the conformational space defined by the fragment libraries employed. For some of the targets, *e.g.* 1tul and 1dhn, no native-like structures have been sampled during

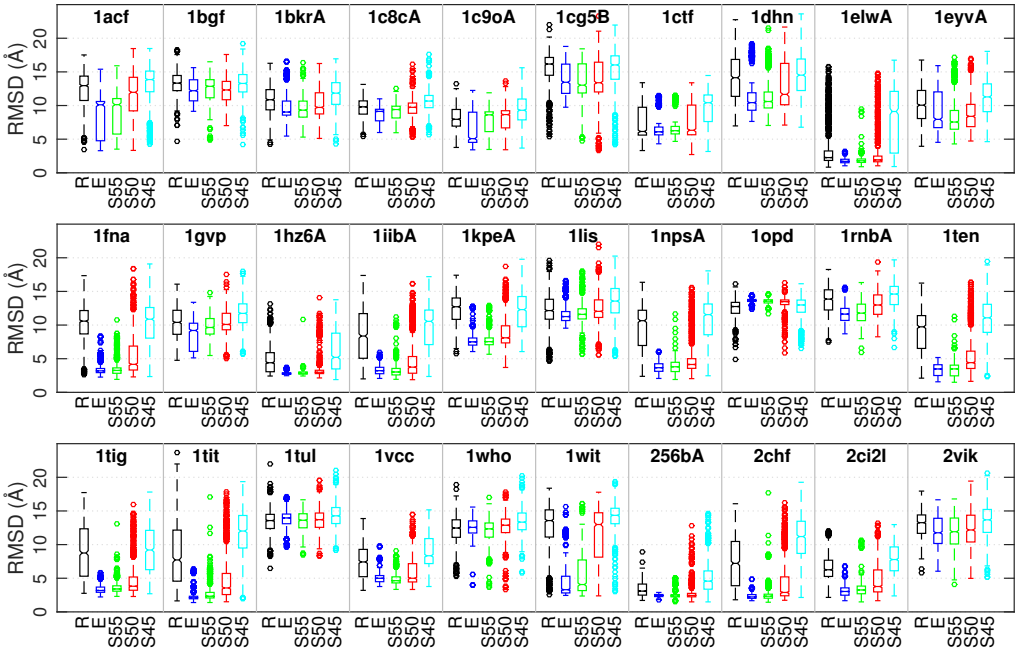

Figure 11: RMSD values scored by the standard Rosetta (R) protocol and the proposed RMA using different selection strategies: energy-based selection (E); and stochastic ranking-based selection, for $\rho \in \{0.45, 0.5, 0.55\}$ (denoted respectively as S45, S50, and S55 in the plots). A total of 30 different protein targets is considered. RMSD values are expressed in Ångströms (Å) and are to be minimised in all the cases.

our experiments, regardless of the search method used. This may suggest that native-like configurations are not covered, or are only scarcely represented, in the fragment libraries adopted for this study, and is an issue which deserves further investigation.

4.2.1 Diversity Generation and Preservation

Similar to Section 3.2.3, we proceed to analyse and discuss the role that both the genetic operators and the survival selection strategy have in terms of diversity generation and preservation. This role is illustrated in Figure 12, which contrasts the distribution of solutions obtained by four variants of the RMA: (i) RMA using energy-based selection, without using genetic operators; (ii) RMA using energy-based selection, using loop-based recombination and mutation; (iii) RMA using stochastic ranking-based selection, $\rho = 0.5$, without using genetic operators; and (iv) RMA using stochastic ranking-based selection, $\rho = 0.5$, using loop-based recombination and mutation.

As shown in Figure 12, without the use of the genetic operators, the energy-driven RMA tends to produce compact, well-defined solution clusters. Each cluster is the result of one of the 10 independent RMA executions performed, each of which produced a total of 100 very similar conformations. The lack of appropriate mechanisms to boost exploration, and the high selection pressure that arises from the use of an elitist and extinctive discrimination strategy based solely on energy, can lead to premature convergence. The inclusion of genetic operators in the energy-based RMA has allowed this method to discover and to exploit more promising attraction basins of the energy landscape, as the results for the second RMA variant suggest. The use of these specialised

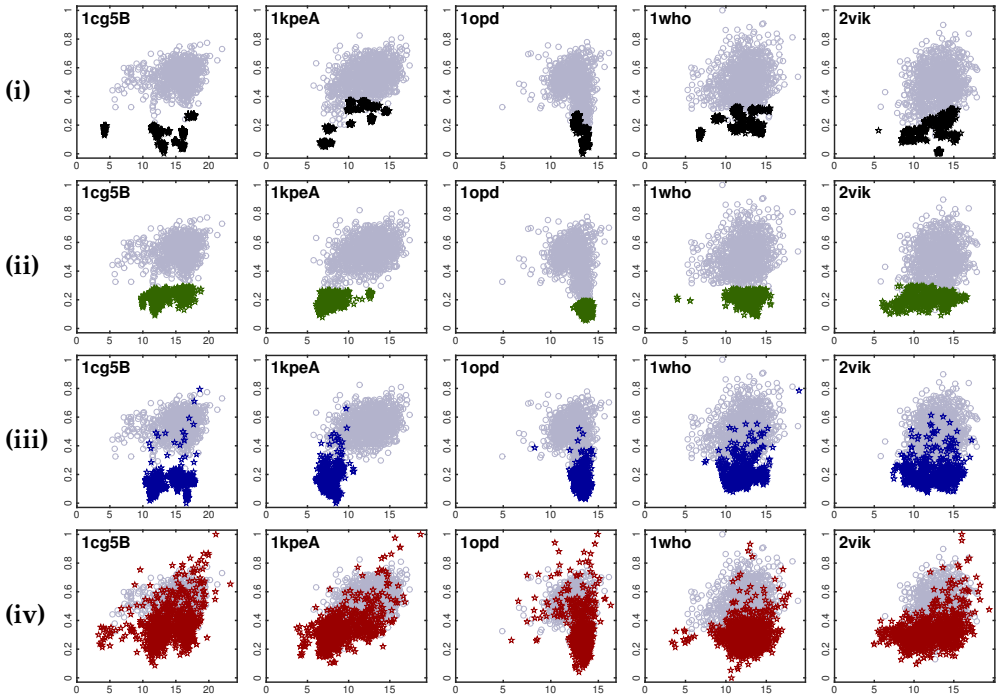

Figure 12: Effects of the genetic operators and survival selection strategies in terms of diversity generation and preservation. Results for five protein targets are shown. Each row of plots in this figure presents results for a different configuration of the RMA: **(i)** energy-based selection, without using genetic operators; **(ii)** energy-based selection, using both recombination and mutation; **(iii)** stochastic ranking-based selection with $\rho = 0.5$, without using genetic operators; and **(iv)** stochastic ranking-based selection with $\rho = 0.5$, using both recombination and mutation. Each plot in this figure contrasts the results scored in terms of RMSD (x-axis) and energy (y-axis). Results obtained by Rosetta are shown at the background as a reference. RMSD values are expressed in Ångströms (Å). Energy values have been normalised to range $[0, 1]$ for visualisation purposes. Both energy and RMSD values are to be minimised in all the cases.

search operators was found previously in Section 3.2.3 to increase conformational space exploration. Similar effects were achieved with the third variant of the RMA. This variant does not employ genetic operators, but replaces the energy-based selection with the stochastic ranking approach. The diversity preservation capabilities of this alternative selection scheme reduce selection pressure with respect to the energy criterion, encouraging a better sampling of the conformational space. Although the individual use of these mechanisms has clearly contributed to RMA's performance, Figure 12 suggests that their combined use (fourth variant) allowed the RMA to further improve both in generating and in retaining a diverse set of solutions throughout the search process.

In order to investigate this further, Figure 13 contrasts the behaviour of the RMA when using the energy- and stochastic ranking-based selection strategies. Behaviour of the RMA throughout the search process is evaluated in terms of the offspring survival rate, convergence, and population diversity, observed as the result of each application of the survival selection process. Results are provided for protein 1cg5B, but similar

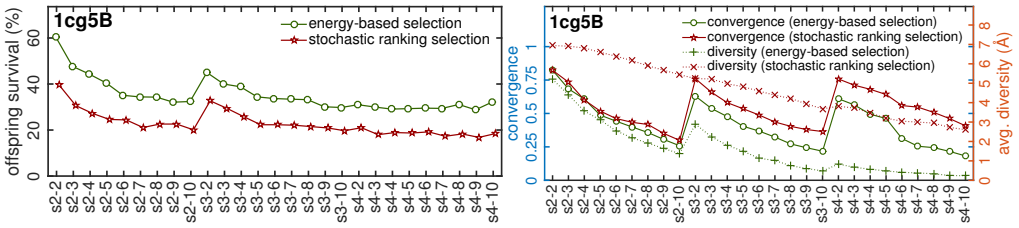

Figure 13: Behaviour of the energy- and stochastic ranking selection schemes (variants ii and iv in Figure 12) on target 1cg5B. To the left, offspring survival rate observed during all 27 applications of the survival selection process (denoted in format *stage-generation*) is shown. To the right, figure shows the convergence (lowest energy in the population) and diversity (average individual contribution, as given by (4)) observed after survival selection. Energies were normalised to range $[0, 1]$ separately for each stage based on the minimum and maximum values observed during this experiment for the corresponding energy functions. Average results of 10 independent runs.

results were observed for additional targets (see Section D of the supplementary material). The use of the stochastic ranking selection prompted a noticeable drop in the number of surviving offspring. As discussed before, the stochastic ranking strategy reduces selection pressure with respect to energy (by preventing over-emphasising this criterion during selection). This slows the convergence speed, as is evident from the convergence curves in Figure 13. However, note that, in spite of this reduced selection pressure in terms of energy, the overall selection pressure of the algorithm rises as a consequence of incorporating an additional discrimination criterion. That is, in order to be selected, candidate individuals need to stand out in terms of the two implemented criteria (as suggested by the analysis of Section 4.1, see Figure 8), which leads to a decrease in the offspring survival rates. On the other hand, Figure 13 shows that whereas the energy-based selection tends to lose diversity and produce a set of very similar (or duplicate) individuals at the end of the search process, the stochastic ranking approach is clearly more effective at maintaining the population's diversity. Diversity preservation is important as a means of producing a more robust solution set consisting of potentially different protein folds discovered during the search process. Moreover, the consequent diversity of genetic material within the population is beneficial and further increases exploration, as it is exploited through recombination.

5 Conclusions

Among all the possible conformations that a protein can adopt, it is believed that its native state, in which it performs its biological functions, corresponds to the one with the lowest overall free-energy (Anfinsen, 1973). From this hypothesis, it follows that predicting structure from sequence is a matter of optimising an energy function with respect to the space of possible tertiary structure configurations. This approach, termed (*de novo*) protein structure prediction (PSP), has been pursued for several decades, and a considerable progress has been made in inferring structures close to the native form, as determined by experimental methods such as X-ray or NMR techniques. This paper focuses on a combinatorial optimisation form of the PSP problem. The fragment-assembly class of methods, which is the most successful approach to *de novo* PSP to date, works with a finite set of tertiary structure fragments, rather than a continuous

space of bond angles. While the approach works very well on some smaller proteins, it is still the case that larger proteins (say of 100 residues and above) generally present a serious challenge. The Rosetta method, a leading example of fragment assembly (which we closely follow here), uses many independent optimisation restarts in order to obtain enough different candidates to be able to make a prediction of structure, and even with this approach it is found to be far from reliable across different protein targets.

This paper has proposed a new sampling protocol for fragment assembly, the Rosetta-based memetic algorithm (RMA). The RMA seeks to overcome the limitations of existing sampling protocols by implementing mechanisms that ensure an appropriate exploration of different protein folds. First, problem-specific knowledge is incorporated into a set of genetic operators that are designed to act on the loop regions of candidate structures. This is based on the understanding that the configuration of loop regions is correlated with the three-dimensional arrangement and packing of secondary structure elements in a fragment-based prediction context, and that a focused exploration of the space of possible loop configurations will translate to an extensive exploration of the space of protein folds. Second, basin-hopping (and appropriate descent into local optima) is further facilitated using the framework of a memetic algorithm that uses the well-established Rosetta protocol as a local search routine. The experiments performed confirm that the new protocol achieved highly competitive results in terms of optimisation performance (*i.e.* minimisation of energy), when evaluated with respect to the standard Rosetta protocol on a large set of protein targets, although this result does not always translate into improvements in prediction performance.

This last finding is not unexpected. In addition to the challenges arising from the size and multi-modality of the search space, protein structure prediction is known to be sensitive to the energy functions used. Whereas state-of-the-art energy functions are often useful in identifying protein-like structures, they are known to have only limited power in pinpointing the most accurate (native-like) folds, a limitation that has not been addressed despite significant research effort focused on the development of more accurate functions. This poses a problem to optimisation protocols which cannot overly rely on the relative rankings between different local optima, which (dependent on the protein) may be more or less 'deceptive' (Goldberg, 1987, 1992). As explicit diversity preservation (niching) has been recognised to be essential in similar scenarios (Sastry et al., 2005; Goldberg et al., 1992; Watson et al., 1998), an alternative selection scheme, based on stochastic ranking (Runarsson and Yao, 2000), was integrated into the proposed RMA with the purpose of regulating selection pressure and enabling diversity maintenance. The results obtained indicate that this modification allows the RMA to display a more robust performance and improve upon Rosetta's performance in terms of the optimisation of both energy and correspondence to the native structure.

In summary, we posit that, due to the enduring inaccuracies of state-of-the-art energy functions, the design of search protocols that explicitly encourage the generation and preservation of diverse folds is a valuable research direction in protein structure prediction. Here, corroborating evidence of this is presented: we described a memetic algorithm that incorporates explicit mechanisms to foster conformational diversity, and we illustrated that this approach can lead to powerful and robust sampling protocols that can offset the problematic bias that is introduced by inaccurate scoring functions. Specifically, the first part of this paper illustrates conclusively that improved optimisation of energy alone, which was successfully achieved, results at times in a problematic sensitivity to the accuracy of the energy function. In the second part of the paper, significant improvements were demonstrated in robustness, however, through the ex-

plicit consideration of diversity during the search (via the stochastic ranking method). Overall, the consideration of structural diversity as an additional criterion to guide offspring generation and selection has increased the likelihood of reaching and preserving more native-like (low-RMSD) conformations for the majority of the targets studied in this work. We intend to exploit this finding in our future work, and will investigate possibilities for further improvement of the proposed approach, *e.g.* by exploring different measures of conformational diversity, alternative mechanisms of diversity maintenance, and further tuning and adjustments to the overall search protocol.

We would like to finish by highlighting the key contributions of this work to the wider research community. From the perspective of protein structure prediction, we present a memetic algorithm for fragment assembly that shows significant promise in comparison to the state-of-the-art technique Rosetta. In line with the core spirit of memetic algorithms, our method uses an established search technique (Rosetta) as a local search strategy, and we design specialised genetic operators and selection schemes to encourage the exploration and retention of diverse conformations. It is our view that exploration performance in general, and memetic algorithms in particular, have been paid insufficient attention in the context of fragment-assembly, and our results help to illustrate the significant improvements that can be achieved by emphasising exploration and the preservation of conformational diversity. Our paper also highlights the possibility of employing stochastic ranking as a general mechanism for diversity preservation. In particular, we use it here to account for the multimodal nature of the optimisation problem, as well as the lack of accuracy in the objective function considered. We expect that our approach will be equally useful in other problem domains, including application areas of multimodal optimisation and problem domains that involve noisy objective functions or other types of uncertainties.

References

Anfinsen, C. (1973). Principles that Govern the Folding of Protein Chains. *Science*, 181(4096):223–230.

Azoitei, M., Correia, B., Ban, Y., Carrico, C., Kalyuzhniy, O., Chen, L., Schroeter, A., Huang, P., McLellan, J., Kwong, P., Baker, D., Strong, R., and Schief, W. (2011). Computation-Guided Backbone Grafting of a Discontinuous Motif onto a Protein Scaffold. *Science*, 334(6054):373–376.

Berman, H., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T., Weissig, H., Shindyalov, I., and Bourne, P. (2000). The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242.

Beyer, H. and Schwefel, H. (2002). Evolution Strategies – A Comprehensive Introduction. *Natural Computing*, 1(1):3–52.

Bowman, G. and Pande, V. (2009). Simulated Tempering Yields Insight into the Low-Resolution Rosetta Scoring Functions. *Proteins: Structure, Function, and Bioinformatics*, 74(3):777–788.

Brunette, T. and Brock, O. (2008). Guiding Conformation Space Search With an All-atom Energy Potential. *Proteins: Structure, Function, and Bioinformatics*, 73(4):958–972.

Chevalier, B. S., Kortemme, T., Chadsey, M. S., Baker, D., Monnat Jr., R. J., and Stoddard, B. L. (2002). Design, Activity, and Structure of a Highly Specific Artificial Endonuclease. *Molecular Cell*, 10(4):895–905.

Chiti, F. and Dobson, C. (2006). Protein Misfolding, Functional Amyloid, and Human Disease. *Annual Review of Biochemistry*, 75:333–366.

Cook, W. J., Cunningham, W. H., Pulleyblank, W. R., and Schrijver, A. (1998). *Combinatorial Optimization*. John Wiley & Sons, Inc., New York, NY, USA.

Das, R. (2011). Four Small Puzzles That Rosetta Doesn't Solve. *PLoS ONE*, 6(5):e20044.

Das, S., Maity, S., Qu, B., and Suganthan, P. (2011). Real-parameter Evolutionary Multimodal Optimization – A Survey of the State-of-the-art. *Swarm and Evolutionary Computation*, 1(2):71–88.

Engh, R. A. and Huber, R. (1991). Accurate Bond and Angle Parameters for X-ray Protein Structure Refinement. *Acta Crystallographica Section A*, 47(4):392–400.

Garza-Fabre, M., Kandathil, S., Handl, J., Knowles, J., and Lovell, S. (2015). Using Machine Learning to Explore the Relevance of Local and Global Features During Conformational Search in Rosetta. In *Genetic and Evolutionary Computation Conference*, pages 935–938. ACM, Madrid, Spain.

Goldberg, D. (1989). *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

Goldberg, D. E. (1987). Simple Genetic Algorithms and the Minimal, Deceptive Problem. *Genetic algorithms and simulated annealing*, 74:88.

Goldberg, D. E. (1992). Construction of High-order Deceptive Functions Using Low-order Walsh Coefficients. *Annals of Mathematics and Artificial Intelligence*, 5(1):35–47.

Goldberg, D. E., Deb, K., and rey Horn, J. (1992). Massive Multimodality, Deception, and Genetic Algorithms. *Urbana*, 51:61801.

Gront, D., Kulp, D. W., Vernon, R. M., Strauss, C. E. M., and Baker, D. (2011). Generalized Fragment Picking in Rosetta: Design, Protocols and Applications. *PLoS ONE*, 6(8):e23294.

Han, K. F. and Baker, D. (1996). Global Properties of the Mapping Between Local Amino Acid Sequence and Local Structure in Proteins. *Proceedings of the National Academy of Sciences of the United States of America*, 93(12):5814–5818.

Handl, J., Knowles, J., Vernon, R., Baker, D., and Lovell, S. C. (2012). The Dual Role of Fragments in Fragment-assembly Methods for De Novo Protein Structure Prediction. *Proteins: Structure, Function, and Bioinformatics*, 80(2):490–504.

Hart, W., Krasnogor, N., and Smith, J. (2005). Memetic Evolutionary Algorithms. In Hart, W., Smith, J., and Krasnogor, N., editors, *Recent Advances in Memetic Algorithms*, volume 166 of *Studies in Fuzziness and Soft Computing*, pages 3–27. Springer Berlin Heidelberg.

Havranek, J. J., Duarte, C. M., and Baker, D. (2004). A Simple Physical Model for the Prediction and Design of Protein-DNA Interactions. *Journal of Molecular Biology*, 344(1):59 – 70.

Jones, D. (1999). Protein Secondary Structure Prediction Based on Position-specific Scoring Matrices. *Journal of Molecular Biology*, 292(2):195–202.

Jones, D. (2001). Predicting Novel Protein Folds by Using FRAGFOLD. *Proteins: Structure, Function, and Bioinformatics*, 45(S5):127–132.

Jones, D., Bryson, K., Coleman, A., McGuffin, L., Sadowski, M., Sodhi, J., and Ward, J. (2005). Prediction of Novel and Analogous Folds Using Fragment Assembly and Fold Recognition. *Proteins: Structure, Function, and Bioinformatics*, 61(S7):143–151.

Kabsch, W. (1976). A Solution for the Best Rotation to Relate Two Sets of Vectors. *Acta Crystallographica Section A*, 32(5):922–923.

Kabsch, W. (1978). A Discussion of the Solution for the Best Rotation to Relate Two Sets of Vectors. *Acta Crystallographica Section A*, 34(5):827–828.

Kandathil, S., Lovell, S., and Handl, J. (2016). Towards a Detailed Understanding of Search Trajectories in Fragment Assembly Approaches to Protein Structure Prediction. *Proteins: Structure, Function, and Bioinformatics*. In press, DOI: 10.1002/prot.24987.

Kim, D. E., Blum, B., Bradley, P., and Baker, D. (2009). Sampling Bottlenecks in *de novo* Protein Structure Prediction. *Journal of Molecular Biology*, 393(1):249 – 260.

Kryshtafovych, A., Fidelis, K., and Moult, J. (2014). CASP10 Results Compared to Those of Previous CASP Experiments. *Proteins: Structure, Function, and Bioinformatics*, 82:164–174.

Kuhlman, B., Dantas, G., Ireton, G. C., Varani, G., Stoddard, B. L., and Baker, D. (2003). Design of a Novel Globular Protein Fold with Atomic-Level Accuracy. *Science*, 302(5649):1364–1368.

Lee, J., Kim, S.-Y., Joo, K., Kim, I., and Lee, J. (2004). Prediction of Protein Tertiary Structure Using PROFESY, A Novel Method Based on Fragment Assembly and Conformational Space Annealing. *Proteins: Structure, Function, and Bioinformatics*, 56(4):704–714.

Martí-Renom, M. A., Stuart, A. C., Fiser, A., Sánchez, R., Melo, F., and Šali, A. (2000). Comparative Protein Structure Modeling of Genes and Genomes. *Annual Review of Biophysics and Biomolecular Structure*, 29(1):291–325.

Meiler, J. and Baker, D. (2006). ROSETTALIGAND: Protein–small Molecule Docking With Full Side-chain Flexibility. *Proteins: Structure, Function, and Bioinformatics*, 65(3):538–548.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., and Teller, E. (1953). Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics*, 21(6):1087–1092.

Mezura-Montes, E. and Coello Coello, C. (2011). Constraint-handling in Nature-inspired Numerical Optimization: Past, Present and Future. *Swarm and Evolutionary Computation*, 1(4):173–194.

Misura, K. M. and Baker, D. (2005). Progress and Challenges in High-resolution Refinement of Protein Structure Models. *Proteins: Structure, Function, and Bioinformatics*, 59(1):15–29.

Molloy, K., Saleh, S., and Shehu, A. (2013). Probabilistic Search and Energy Guidance for Biased Decoy Sampling in *ab initio* Protein Structure Prediction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 10(5):1162–1175.

Moscato, P. (1989). On Evolution, Search, Optimization, Genetic Algorithms and Martial Arts: Towards Memetic Algorithms. Technical Report C3P Report 826, Caltech Concurrent Computation Program, Pasadena, CA.

Moscato, P. and Cotta, C. (2003). A Gentle Introduction to Memetic Algorithms. In Hillier, F. S., Glover, F., and Kochenberger, G., editors, *Handbook of Metaheuristics*, volume 57 of *International Series in Operations Research & Management Science*, pages 105–144. Springer New York.

Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T., and Tramontano, A. (2014). Critical Assessment of Methods of Protein Structure Prediction (CASP) – Round X. *Proteins: Structure, Function, and Bioinformatics*, 82:1–6.

Neri, F. and Cotta, C. (2012). Memetic Algorithms and Memetic Computing Optimization: A Literature Review. *Swarm and Evolutionary Computation*, 2:1–14.

Olson, B. S., De Jong, K. A., and Shehu, A. (2013). Off-lattice Protein Structure Prediction with Homologous Crossover. In *Genetic and Evolutionary Computation Conference, GECCO '13, Amsterdam, The Netherlands, July 6-10, 2013*, pages 287–294.

Papadimitriou, C. and Steiglitz, K. (1982). *Combinatorial Optimization: Algorithms and Complexity*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

RCSB PDB (2015). PDB statistics. Available at `http://www.rcsb.org/pdb/static.do?p=general_information/pdb_statistics/index.html`. Accessed 28/08/2015.

Rohl, C. A., Strauss, C. E. M., Misura, K., and Baker, D. (2004). Protein Structure Prediction Using Rosetta. *Methods in Enzymology*, 383:66–93.

Rost, B. and Sander, C. (1993). Prediction of Protein Secondary Structure at Better than 70% Accuracy. *Journal of Molecular Biology*, 232(2):584 – 599.

Runarsson, T. and Yao, X. (2000). Stochastic Ranking for Constrained Evolutionary Optimization. *IEEE Transactions on Evolutionary Computation*, 4(3):284–294.

Saleh, S., Olson, B., and Shehu, A. (2013). A Population-based Evolutionary Search Approach to the Multiple Minima Problem in *De Novo* Protein Structure Prediction. *BMC Structural Biology*, 13(Suppl 1):S4.

Sastry, K., Abbass, H. A., Goldberg, D. E., and Johnson, D. (2005). Sub-structural Niching in Estimation of Distribution Algorithms. In *Proceedings of the 7th annual conference on Genetic and evolutionary computation*, pages 671–678. ACM.

Shehu, A. and Olson, B. (2010). Guiding the Search for Native-like Protein Conformations with an *ab initio* Tree-based Exploration. *The International Journal of Robotics Research*, 29(8):1106–1127.

Shir, O., Emmerich, M., and Bäck, T. (2010). Adaptive Niche Radii and Niche Shapes Approaches for Niching with the CMA-ES. *Evolutionary Computation*, 18(1):97–126.

Siegel, J. B., Zanghellini, A., Lovick, H. M., Kiss, G., Lambert, A. R., St.Clair, J. L., Gallaher, J. L., Hilvert, D., Gelb, M. H., Stoddard, B. L., Houk, K. N., Michael, F. E., and Baker, D. (2010). Computational Design of an Enzyme Catalyst for a Stereoselective Bimolecular Diels-Alder Reaction. *Science*, 329(5989):309–313.

Simoncini, D., Berenger, F., Shrestha, R., and Zhang, K. Y. J. (2012). A Probabilistic Fragment-Based Protein Structure Prediction Algorithm. *PLoS ONE*, 7(7):e38799.

Simoncini, D. and Zhang, K. Y. J. (2013). Efficient Sampling in Fragment-Based Protein Structure Prediction Using an Estimation of Distribution Algorithm. *PLoS ONE*, 8(7):e68954.

Simons, K. T., Kooperberg, C., Huang, E., and Baker, D. (1997). Assembly of Protein Tertiary Structures from Fragments with Similar Local Sequences Using Simulated Annealing and Bayesian Scoring Functions. *Journal of Molecular Biology*, 268(1):209–225.

Simons, K. T., Ruczinski, I., Kooperberg, C., Fox, B. A., Bystroff, C., and Baker, D. (1999). Improved Recognition of Native-like Protein Structures Using a Combination of Sequence-dependent and Sequence-independent Features of Proteins. *Proteins: Structure, Function, and Bioinformatics*, 34(1):82–95.

Tai, C.-H., Bai, H., Taylor, T. J., and Lee, B. (2014). Assessment of Template-free Modeling in CASP10 and ROLL. *Proteins: Structure, Function, and Bioinformatics*, 82:57–83.

The UniProt Consortium (2015a). UniProt: A Hub for Protein Information. *Nucleic Acids Research*, 43(D1):D204–D212.

The UniProt Consortium (2015b). UniProtKB/TrEMBL Database statistics. Available at `http://www.ebi.ac.uk/uniprot/TrEMBLstats`. Accessed 28/08/2015.

Črepinšek, M., Liu, S.-H., and Mernik, M. (2013). Exploration and Exploitation in Evolutionary Algorithms: A Survey. *ACM Computing Surveys*, 45(3):35:1–35:33.

Wang, C., Vernon, R., Lange, O., Tyka, M., and Baker, D. (2010). Prediction of structures of zinc-binding proteins through explicit modeling of metal coordination geometry. *Protein Science*, 19(3):494–506.

Watson, R. A., Hornby, G. S., and Pollack, J. B. (1998). Modeling Building-block Interdependency. In *Parallel Problem Solving from Nature—PPSN V*, pages 97–106. Springer.

Whitehead, T., Chevalier, A., Song, Y., Dreyfus, C., Fleishman, S., De Mattos, C., Myers, C., Kamisetty, H., Blair, P., Wilson, I., and Baker, D. (2012). Optimization of Affinity, Specificity and Function of Designed Influenza Inhibitors Using Deep Sequencing. *Nature Biotechnology*, 30(6):543–548.

Xu, D. and Zhang, Y. (2012). Ab initio Protein Structure Assembly Using Continuous Structure Fragments and Optimized Knowledge-based Force Field. *Proteins: Structure, Function, and Bioinformatics*, 80(7):1715–1735.