Investigation of an Immune Algorithm and Differential Evolution to Study Folding of Model Proteins

by

Andrew James Bennett

A thesis submitted to

The University of Birmingham

for the examination of

DOCTOR OF PHILOSOPHY

School of Chemistry
University of Birmingham
December 2, 2009

Abstract

The application of computational search techniques to global optimisation problems is becoming increasingly popular. Search techniques have been used to study the folding of model proteins, with the aim of accurately predicting the native state of a protein from its amino acid sequence. Through modelling, knowledge of the folding process can be obtained.

In this thesis, two search techniques have been applied to a variety of protein models. The development and application of both an Immune Algorithm and a Differential Evolution search technique are described, with the aim of finding the lowest energy conformations of coarse-grained, model proteins. Initially, the two-dimensional HP Lattice Bead Model is investigated, followed by three-dimensional models of varying complexity. The HP Lattice Bead and BLN models, on a diamond lattice are considered, as well as the Dynamic Lattice Model, using backbone torsion angles to define the structure of the lattice.

A modified chain growth constructor is introduced; firstly, to generate the initial population for both search techniques, secondly, to record unoccupied lattice sites of meta-stable conformations to reduce the risk of performing infeasible point mutations during the mutation phase for the Immune Algorithm, and thirdly, to improve the standard of mutations performed by Differential Evolution.

A novel profiling system is introduced based on the theory of genealogy and ancestry by recording the parent of each individual. The method is used to track and evaluate the diversity of populations and assess the impact that genetic operators have on this diversity. The aim of applying this system is three fold: to investigate how effective genetic operators are; to allow a greater understanding of the progress of the optimisations; and to assess the strengths and weaknesses of each search technique investigated.

"The important thing is not to stop questioning."

Albert Einstein

Acknowledgements

I would firstly like to thank my supervisor, Prof. Roy Johnston, for granting me the opportunity to not only undertake postgraduate study here, but also for introducing me to computational chemistry during my undergraduate years, providing the inspiration for this work. His help and support throughout my higher education, both in and outside academia have proven indispensable and his guidance invaluable.

Special thanks go to Dr. Graham Worth for sharing his knowledge in Roy's absence and for not leaving me even more confused after our vector math discussions. Dr. Ben Curley, for not only programming support, but for being a great friend and spending countless hours trying to show me a better way. Dr. Oliver Paz-Borbón for keeping my spirits high in difficult times and for being away so often during our time at Corrisande Road. Adam Cowell for being supportive and a great listener. Dr. Emma Chapman for having a good old moan with me when the pressure increased.

I would like to thank Tom Penfold for his help and assistance with the "sys-admin" and members of the Johnston and Tremayne groups (those not already mentioned), Duncan, Gareth, Graham, Lianna, "Loggy", Nico, Paul, Ramli, Raja and Sam, for making the office a pleasent place to work and for not causing too many distractions (I said too many). The rest of the people on floor 2 for the trips to Staff House and of course the Sunflower Lounge.

Finally, I would like to thank my family and especially Laura, for their continued support and encouragement, particularly as the end drew near.

Publications

Bennett, A. J., Johnston, R. L., Turpin, E. & He, J. Q. *Analysis of an Immune Algorithm for Protein Structure Prediction*, Informatica, 2008. **32** (3), 245-251.

Bennett, A. J., Johnston, R. L. *Investigation of Diamond Lattice Proteins using an Immune Algorithm*, Journal of Chemical Physics, 2009.
submitted.

Bennett, A. J., Johnston, R. L. *Application of an Immune Algorithm for the Study and Analysis of Dynamic Lattice Proteins*, Journal of Physical Chemistry B, 2009.
submitted.

Bennett, A. J., Johnston, R. L. *Using Differential Evolution for the Prediction and Analysis of Protein Structures*, Physical Chemistry, Chemical Physics, 2009.
submitted.

Contents

Glossary

D_H	Hamming Distance.
E_i	Structural conformation energy of an individual.
F	DE mutation rate.
$F*$	Best fitness found as a function of energy.
F_b	Best individual's fitness in the current population.
F_i	Fitness of an individual as a function of energy.
K	DE recombination rate.
μ_{DH}^{L}	Mean Hamming Distance of a population with respect to the lowest energy conformation.
μ_{DH}^{P}	Mean pairwise Hamming Distance with respect to individuals of a population.
μ_{FE}	Mean number of fitness evaluations.
μ_{RMSD}^{L}	Mean RMSD of a population with respect to the lowest energy conformation.
μ_{RMSD}^{P}	Mean pairwise RMSD with respect to individuals of a population.
μ_g	Mean number of generations of a number of algorithm runs.
ω	$C_{\alpha(i)}$-C_i-$N_{(i+1)}$-$C_{\alpha(i+1)}$ dihedral angle for residue i.
ϕ	$C_{(i-1)}$-N_i-$C_{\alpha(i)}$-C_i dihedral angle for residue i.
ϕ_C	Cis-like ϕ torsion angle.
ϕ_T	Trans-like ϕ torsion angle.
ψ	N_i-$C_{\alpha(i)}$-C_i-$N_{(i+1)}$ dihedral angle for residue i.
ψ_C	Cis-like ψ torsion angle.
ψ_T	Trans-like ψ torsion angle.
σ_{FE}	Standard deviation of the number of fitness evaluations.
σ_g	Standard deviation of the number of generations taken over a various algorithm runs.
θ_0^i	Equilibrium bond angle term in the BLN Model.
θ_i	Bond angle term in the BLN Model.
g	Calculation iteration or generation.
g_{max}	Maximum number of generations.
i_p	Selected parent from a population.
i_{max}	Maximum individual age in terms of generations.
i_{r1}	First randomly selected individual.
i_{r2}	Second randomly selected individual.
i_{r3}	Third randomly selected individual.

m_f	Mutation factor for hypermutation used by the Immune Algorithm.
n_{beads}	Number of modelled amino acids.
n_{clo}	Number of clones.
n_{ind}	Population size.
n_{mut}	The number of mutation attempts by the hypermutation operator.
n_{uniq}	Number of unique minima found of a certain fitness.
r_0^i	Equilibrium bond length term in the BLN Model.
r_i	Bond length term in the BLN Model.
s_{FE}	Skewness of the number of fitness evaluations distribution for a number of runs.
s_g	Skewness of the generation distribution for a number of runs.
2D	Two Dimensional.
3D	Three Dimensional.
a.u.	Arbitrary Units.
ACO	Ant Colony Optimisation.
B	Hydropho**B**ic bead for the BLN Model.
BLNM	hydropho**B**ic-po**L**ar **N**eutral Model.
CORN	(-**CO**OH, -**R** group and -**N**H_2) law.
CPU	Central Processing Unit.
DE	Differential Evolution.
DLM	Dynamical Lattice Model.
GA	Genetic Algorithm.
GM	Global Minimum.
H	**H**ydrophobic bead for the HP Lattice Bead Model.
HC	Hypermutation operator and crossover.
HPLBM	Hydrophobic Polar Lattice Bead Model.
IA	Immune Algorithm based on clonal selection.
L	Po**L**ar bead for the BLN Model.
LS	Local Search.
MS	Mixed Strategy Operator.
N	**N**eutral bead for the BLN Model.
P	**P**olar bead for the HP Lattice Bead Model.
PDB	Protein Data Bank.
PES	Potential Energy Surface.

RGA	Recoil Growth Algorithm.
RMSD	Root Mean Square Deviation.
SR	Success Rate measured in %.
VDW	van der Waals.

List of Figures

List of Algorithms

List of Tables

Chapter 1

Introduction

1.1 Proteins

Every living organism contains large biomolecules known as proteins. Proteins are biological building blocks which are responsible for most of the functions of living systems [1]. They are biopolymers constructed from a sequence of amino acid residues that comprise the structure's backbone [2]. Protein characteristics are sequence specific, in that they depend completely on the sequence of amino acid monomers that form the structure in question [3]. The biological function of a protein is determined by this structure [4].

1.1.1 Protein Structure

The structure of a protein, regardless of its function, is constructed by linking many **amino acid** monomer units via amide bonds (peptide bonds). These amino acid (residue) chains can vary in length, with chains of fewer than 50 residues often being called **peptides**, whereas bigger chains are referred to as **proteins** [6]. The term amino acid is shorthand for the common α-amino acid. The generic amino acid has the general formula $NH_2C_\alpha RHCO_2H$ [7].

All amino acids consist of an α-carbon atom (which is asymmetric except in glycine); an **amino** group (NH_2) and a carboxylic **acid** group (CO_2H) [8]. They have non-superimposable mirror images (i.e. are "handed") due to the chirality of the C_α atom.

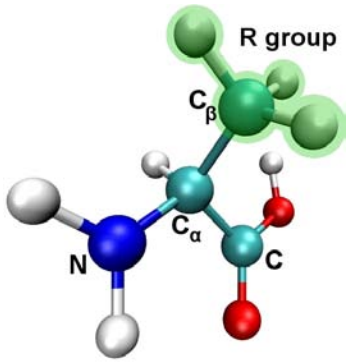


Figure 1.1: A single amino acid, L-Alanine, illustrating the N and the C termini adapted from [5]. The C_β, which is not present in glycine, is the source of the structure's chirality, with its substituents being unique between amino acids. The R group, highlighted in green comprises the C_β and its substituents.

The α denotes the carbon atom to which the side-chain R group is bonded and is highlighted in figure 1.1. With the C_α of glycine being bonded to two hydrogens (i.e. the C_β and its substituents being replaced by a single hydrogen atom), its chirality is lost. Amino acids are present as both L- and D-isomers representing each mirror image, however, natural proteins only contain the L form [3] [9].

Figure 1.2 lists the 20 naturally occurring amino acids [11], associating each with its triple-letter code. It is from this amino acid set that all naturally occurring proteins are derived. It is possible to form proteins incorporating amino acids that are not in figure 1.2, i.e. they are not found in nature, so these are not regarded as natural.

As their name suggests, amino acids are difunctional, in that they contain both a basic amino group and an acidic carbonyl group. It is this difunctionality that allows the peptide bond to be formed by condensing the $-NH_2$ of one amino acid with the -COOH of another, as shown in figure 1.3(a).

The peptide bond is in essence a planar covalent bond. A delocalisation of the nitrogen lone pair and its interaction with the carbonyl group, renders amide nitrogens non-basic. The overlap of the nitrogen p orbital with the Π^* antibonding orbital of the carbonyl group (giving double bond characteristics), restricts rotation around the

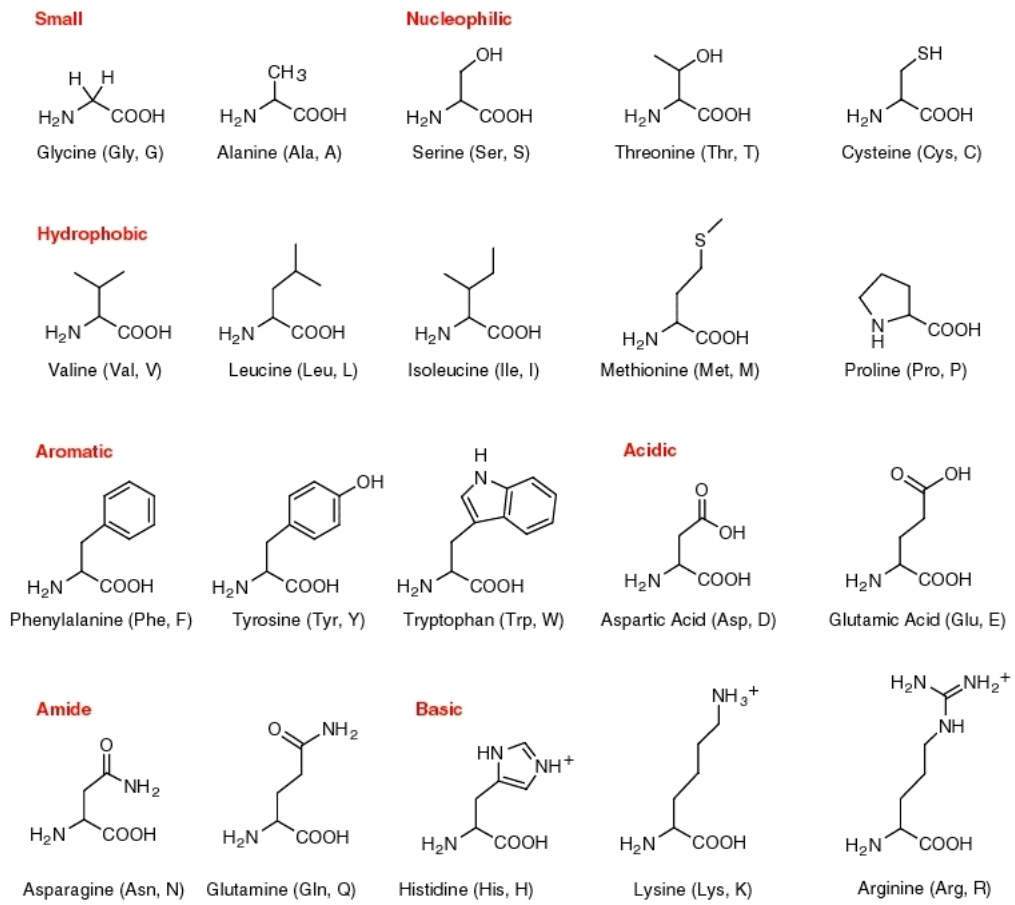


Figure 1.2: Structures of each natural amino acid, single- and three-letter codes (adapted from [10]).

C-N bond, making the peptide unit planar with a 180° H-N-C-O dihedral angle. This is illustrated with both resonance forms in figure 1.3(b). The long repetitive chain of -N-C_α-C(O)- produced as a result of the peptide bonds is known as the **protein backbone**. It is conventional to write the protein sequence from the **N-terminal amino acid** with the free -NH_2 group, to the **C-terminal amino acid** with the free -COOH group. This sequence is known as a protein's **primary structure** [6].

Due to steric hindrance, the backbone of a polypeptide chain assumes preferred, energetically favourable conformations [12]. For each residue, these conformations are characterised by three dihedral angles, ω, ϕ and ψ, providing variability in the protein conformation. A dihedral angle is defined as the angle (positive if clockwise, negative

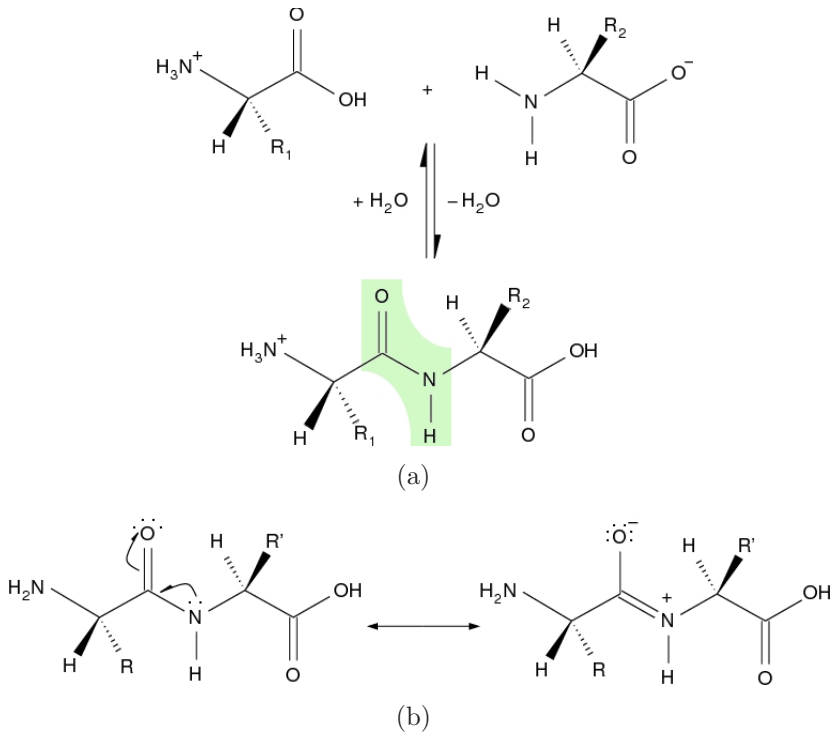


Figure 1.3: (a) Peptide bond formation mechanism portraying expulsion of H_2O (adapted from [5]). Also shown is the cleaving of a peptide bond and the requirement for H_2O to do so. The highlighted region shows the peptide bond and the atoms involved in its formation. (b) The resonance forms of the peptide bond. The partial double bond character of the central C-N bond restricts rotation, rendering a dihedral angle of approximately 180°.

if anti-clockwise) required to rotate around the bond linking the central two atoms in a four atom system, such that the first atom eclipses the fourth. This can be seen in figure 1.4.

For residue i, ω is the dihedral angle between $C_{\alpha(i)}$-C_i-$N_{(i+1)}$-$C_{\alpha(i+1)}$. Due to the restricted C-N bond rotation, $\omega = 180°$ in the case of *trans*, planar peptides. However, *cis* peptides, with $\omega = 0°$ can occur, but are only observed in some proline containing peptides. ϕ and ψ are the dihedral angles between $C_{(i-1)}$-N_i-$C_{\alpha(i)}$-C_i and N_i-$C_{\alpha(i)}$-C_i-$N_{(i+1)}$ atoms along the protein backbone, respectively. Due to the constrained nature of the ω dihedral angle , it is really only the ϕ and ψ angles that provide the backbone with a variety of conformations [9] [12]. The dihedral arrangements can be seen in

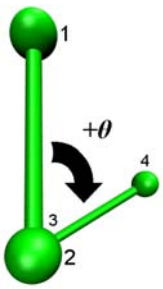


Figure 1.4: The definition of a dihedral angle. Atom two masks atom three with the angle about this bond resulting in an angle between the first and fourth atoms from the same pivot point.

figure 1.5.

A combination of dihedral angles can give rise to certain regular locally folded patterns of peptide backbone [6]. The stability of a protein is governed by the presence of these locally folded arrangements. An α-helix can order as many as 35 residues and can have the greatest influence on structural stability. It is formed when a hydrogen bond exists between the C=O of the n^{th} residue and the N-H of the $(n+4)^{th}$, repeating to form an extended helical conformation. First documented in the work of Pauling *et al.* [13], the α-helix is also known as the 3.6_{13}-helix, where 3.6 is the number of residues per turn and 13 is the number of atoms in the hydrogen-bonded loop. The ϕ,ψ dihedral angles for right-handed α-helices are both approximately -60°.

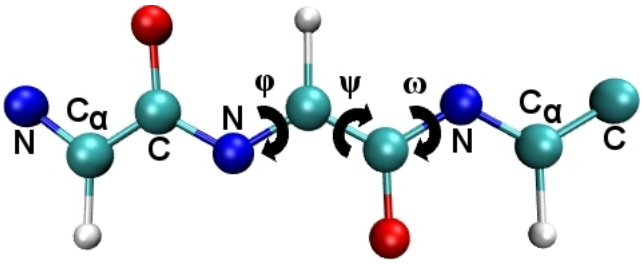


Figure 1.5: A simple protein backbone omitting side chains and hydrogen atoms illustrating the rotations around bonds that give rise to the dihedral definitions.

Another principal helical species is the 3_{10}-helix. Using the nomenclature of the 3.6_{13}-helix, the 3_{10}-helix has a three residue repeat and forms a hydrogen-bond with

the N-H of the $(n+3)^{th}$ residue instead of the $(n+4)^{th}$. The backbone conformational angles are approximately ϕ=-60° and ψ=-30°. A 5_{16}-helix also exists, with both this and the 3_{10}-helix (collectively known as π-helices) being rather strained, and thus, are normally found at the end of an α-helix, or as single turns [7]. Due to both local conformational energy and hydrogen-bond configuration, the 3.6_{13}-helix is considerably more favourable than the 3_{10}-helix.

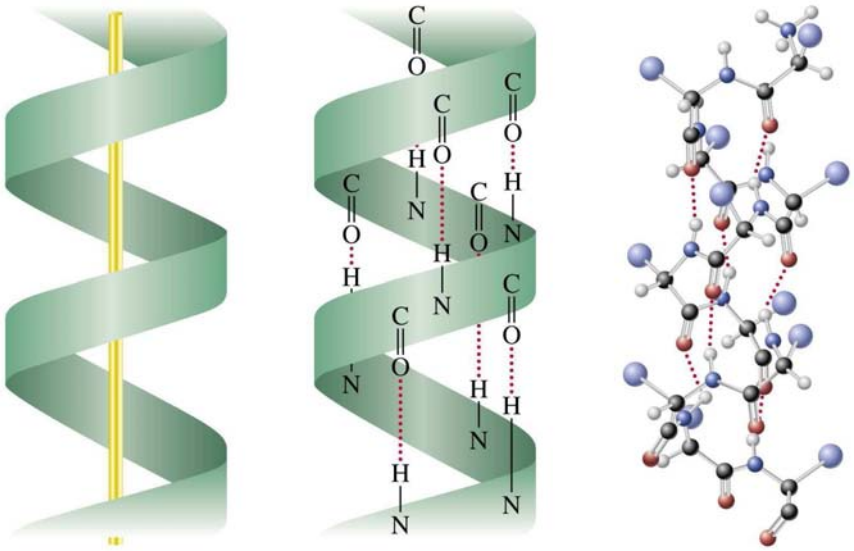


Figure 1.6: A schematic right handed α-helix taken from [14]. It identifies the hydrogen bonds that exist between the oxygen of a carbonyl group and the nitrogen of an amide group to provide its stability.

Ala, Glu, Leu and Met have strong preferences for forming an α-helix, while Gly, Tyr and Ser have strong preferences against this. Although Pro, due to steric hindrance is not a good contributor to helical formation, it is also missing the hydrogen bond donor. However, if Pro is involved in the first helical turn, especially being the initial residue, it can participate in the formation of an α-helix, at the expense of producing a slight bend in the helical axis. Helical species provide protein structure segments with strength and elasticity.

Other than the helical species, the other major structural component found in globular proteins is the β sheet [9]. Often when sections of peptide chain fold back

on themselves, it is common for areas of local structure to adopt this arrangement. These are formed when peptide chains (β strands) line up in a parallel fashion, being stabilised by the formation of hydrogen-bonds between them [6]. Although the chains themselves are parallel, the orientation of the hydrogen bonding between the chains denotes whether a parallel or anti-parallel β sheet is seen or not. In the β sheet, however, the hydrogen bonds are formed between the N-H of one β strand and the C=O of another. For the parallel arrangement, the hydrogen-bond orientation is achieved by both β strands running in parallel from the N to the C terminus. The anti-parallel arrangement corresponds to having a β strand from the N to the C terminus running parallel with one from the C to the N terminus. In both arrangements, the position of the side groups along each strand alternate above and below the sheet, whereas the side groups on neighbouring strands extend to the same side of the sheet and are close in proximity [9].

These principal helical and β-sheet species are known as a protein's **secondary structure**. Other local structures that make up this category include β-turns (also known as reverse-turns or sometimes hairpin-turns) and disulfide bridges. A β-turn is the simplest secondary structure element and requires only three to four residues. It consists of a hydrogen bond between the carbonyl oxygen of the n^{th} residue and the amide N-H of the $(n+3)^{th}$ residue, but rarely between the n^{th} and $(n+2)^{th}$ as such a turn is too strained. This provides a simple way in which to satisfy the hydrogen-bonding capability in a peptide group. However, inspection of this structure reveals that the C=O and N-H groups in these four residues are not in fact making hydrogen bonds with any other backbone atoms. As water molecules are able to donate and accept hydrogen bonds to these groups, β-turns tend to be found on the surfaces of folded proteins [15].

Disulfide bridges are another example of covalent bonding in proteins. They occur when a RS-SR bond is formed between two cysteine residues. Two separate peptide

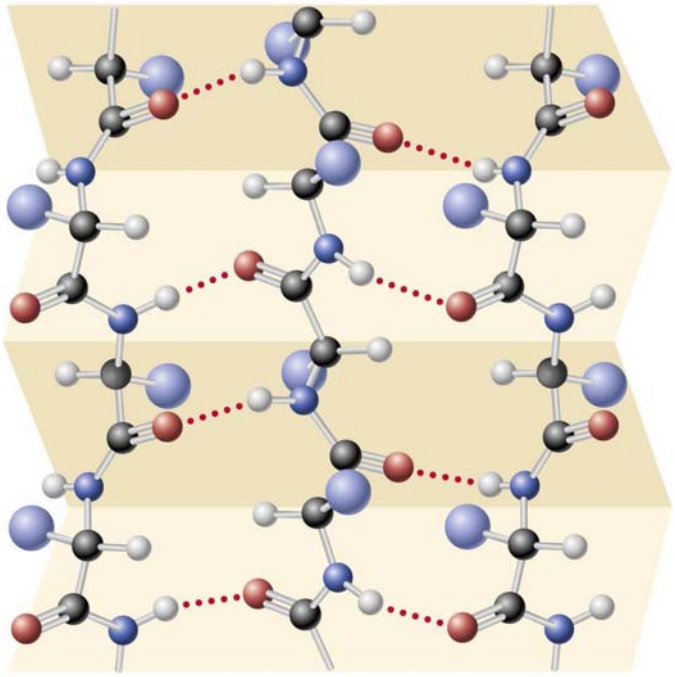


Figure 1.7: An anti-parallel β-sheet taken from [14].

chains can be linked, or a loop within a single chain can result from disulfide bridge formation [6]. The formation of secondary structure is aided by the presence of random coils. They can be described as ordered sections of structure that do not exhibit the repeating nature found with α-helices and β-sheets. [8].

Tertiary structure can be described as ordered regions of secondary structure; sections of random coil connecting α-helices and β-sheets. Hydrogen bonding, van der Waals (VDW) and electrostatic forces contribute to the tertiary structure layout. **Quaternary** structures can be described as super-structures; with individual proteins being held together by intermolecular attractions: VDW and electrostatic forces: for example haemoglobin is a tetramer [8].

1.1.2 Denaturation

Denaturation refers to the unravelling of the tertiary protein structure. Weak intramolecular attractions hold the tertiary structure in place, and with a slight change to the environment, usually the temperature or pH, the tertiary structure can be destroyed, leaving the primary structure intact. With the conditions for denaturation only being mild, the covalent bonds of the primary structure are unaffected. As most but not all denaturation is reversible, spontaneous renaturation can occur. Renaturation involves the restoration of a protein's tertiary structure and biological activity [6]. Denatured proteins are generally not functional, with mutations disrupting the overall structure often leading to drastic functional changes [4].

1.1.3 Ramachandran Plot

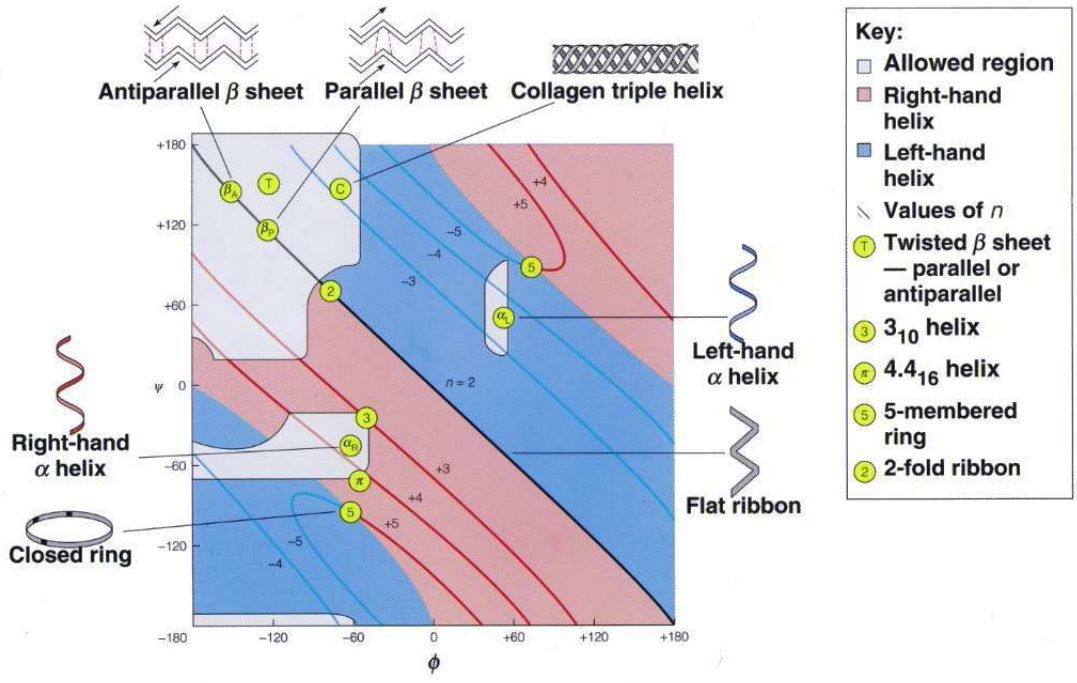


Figure 1.8: A Ramachandran plot taken from [16] illustrating how regions of ϕ,ψ clusters result in specific secondary structures. The contours (labelled +5 to 2 and -5 to -3) signify the number of amino acid residues per turn of a helix; "+" meaning right-handed helices and "-" meaning left-handed helices.

The dihedral angles ϕ and ψ provide the diverse conformational variety in the protein backbone. By plotting the distribution of these dihedral pairs, it can be seen from figure 1.8 that clustering of certain dihedral ϕ,ψ pairs occurs. By plotting ϕ *vs.* ψ in the range -180° to +180°, the clustering illustrates that particular types of secondary protein structure adopt a dihedral pair for each residue within a favourable range [17]. However, due to structural distortions, angle pairs may lie outside these predominant regions and it is difficult to justify whether the conformations found in sparsely populated areas are valid but rare or disallowed conformations [18]. The work of Ramachandran *et al* [19] introduced this type of distribution and in doing so devised the Ramachandran plot. It allows the stereochemistry of the polypeptide chain backbone in a protein structure to be analysed [20]. The plots can be generated in two ways; either from theoretical calculations or from experimental observations [18]. The dihedral pairs produced via protein modelling should occupy favourable Ramachandran regions as much as possbile, and for this reason, in the absence of experimental data, the Ramachandran plot can be seen as a means of assessing the quality of a protein model [12]. With alanine experiencing the least amount of steric hindrance, experimentally its idealised tri-peptide is considered to set the boundaries of allowed space on the Ramachandran surface [20]. However, despite over four decades of research, the exact boundaries of these allowed and disallowed regions, are still under scrutiny [18].

1.1.4 Protein Folding

The protein folding problem is a fundamental problem in computational molecular biology, biochemical physics and chemical biology [21,22]. The problem includes statistical mechanics. However, it also shares a common feature with most biological problems, the effects of evolution [23]. With protein evolution, considerations must be given to how mutational change in the amino acid sequence leads to structural and functional change [24]. The protein folding problem is the prediction of the three dimensional (3D) local spatial arrangement (secondary structure) and the folded conformation (tertiary

structure) adopted by a polypeptide molecule from only the knowledge of its primary amino acid sequence, the one dimensional (1D) structure from which it is built [22, 25]. It is the search for the most biologically active (functional) conformation of a protein (the native state), for a given sequence of amino acid residues. It has been shown to be an NP-hard problem, in that no efficient algorithm can guarantee to find the native state [26]. The relationship between sequence and structure is of critical importance if we are to understand how proteins fold and ultimately highlight the sequence-activity correlation of protein molecules [22, 27].

The reliability of natural proteins to fold to a unique, low energy, most stable state (native state) is related to the presence of a "folding funnel" on the free energy landscape, allowing misfolded proteins to be guided towards the most energetically favourable conformation. To achieve a greater knowledge of protein folding dynamics, the nature of the free energy landscape must be understood [27]. Although progress has been made over many decades, due to the complexity of the problem, it still remains unsolved [25].

1.1.4.1 The Levinthal Paradox

Protein folding dynamics is strongly linked to the **Levinthal paradox** [28]. In 1969 Cyrus Levinthal [29] hypothesised that due to the enormous number of accessible conformations that a protein could adopt, the protein should take an eternity to fold into its native state if it explored its conformational space at random [28,30,31]. If the protein is able to sample 10^{13} different bond configurations per second then it would take 10^{27} years (longer than the age of the Universe) to sample all possible conformations of the protein [32] in order to guarantee finding its native structure. However, due to a protein achieving this in such a short periods of time (typically the millisecond timescale), the folding must in fact be a directed and not a random process [25, 30].

1.1.4.2 Hydrophobic Collapse

Amino acids have either polar (hydrophilic) or non-polar (hydrophobic) side chains. From an early stage, it is said that proteins undergo a process whereby a compact core is formed. The structure folds in such a way as to protect the hydrophobic residues as much as possible from the surrounding solvent environment. This is achieved by encapsulating the hydrophobic residues with one of a hydrophilic nature. The process known as **Hydrophobic Collapse** is illustrated in figure 1.9 and is believed to play a major role in the conformational preferences of biomolecules [23, 33].

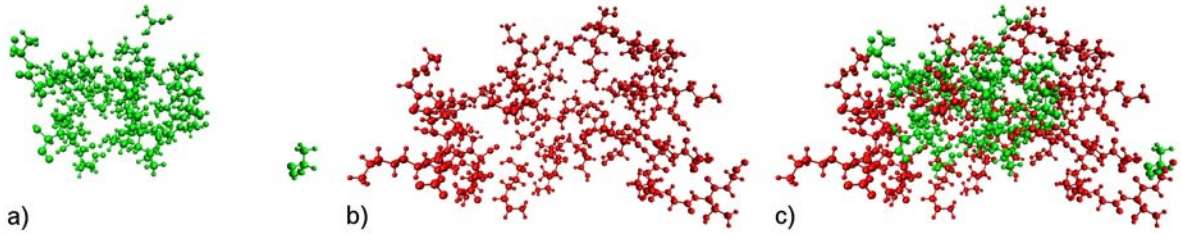


Figure 1.9: Diagram illustrating hydrophobic collapse in the case of 1GHC [34], sourced from [35]. a) shows only the encapsulated hydrophobic core, b) shows only the hydrophilic "casing" and c) shows how the two components form to protect the hydrophobic core. The real protein has folded in such a way as to shield the hydrophobic residues (green) from the solvent by hydrophilic residues (red) as much as possible.

1.1.4.3 Potential Energy Surfaces

The structure and dynamics of a protein system are determined by its underlying potential energy surface (PES) [36–38]. Levinthal's paradox (section 1.1.4.1) assumes that the energy landscape is flat [36]. However, in recent years, the funnel topography of the energy landscape, has become synonymous with the protein folding discussion [39]. The PES of a protein is said to present a funnel [40], with most configuration space being present at the surface of local minimum basins [38]. The local roughness of this funnel reflects the trapping of protein conformations in local potential energy minima [40]. A local minimum is defined as a point, in which any displacement will lead to higher potential energy conformations [7] (via a gradient increase), with displacement

from a transition state leading to lower energy conformations (both have all gradients equal to zero and, thus, are stationary points [41]). Thermal energy is required to overcome these positive gradients (potential energy barriers), with the thermodynamic properties of a system being dependent on the minima being sampled [38]. A PES is independent of temperature (with free energy landscapes considering entropy and, thus, are dependant on temperature) and atomic masses, however, the number of minima increases rapidly with the size of the system [36, 42]. The connectivity of a PES is defined by the characterisation of pathways between minima [7]. The lowest energy conformation is known as the Global Minimum (GM) [38].

1.1.4.4 Thermodynamic and Kinetic Hypotheses

For the Gibbs free energy of a polypeptide system in its normal physiological environment to be its lowest, the three-dimensional structure of its native state should be completely governed by interatomic interactions and hence by the amino acid sequence [31, 43]. In 1973 Christian Anfinsen [43] performed a series of denaturation-renaturation experiments on protein molecules and concluded that the native state of a protein is in fact the GM of the free energy [30]. This was named the **thermodynamic hypothesis** [30, 43].

Opposing this view is one that treats the functional conformation not as the GM but as the structure that is most frequently visited [22]. Under native conditions, this state may of course be meta-stable [22, 31]. This is known as the **kinetic hypothesis** [22].

1.1.4.5 The Metastability Hypothesis

Folding of a polypeptide is kinetically controlled [44]. The folding conformations of several free energy minima are similar and are separated by barriers of various heights. They exhibit similar equilibrium bond angle and dihedral angle distributions based on the configurations belonging to a given minimum, but have different energies from one another. A high proportion of the folded states are metastable and the transition from

one minimum to another is infrequent. This differs from the kinetic hypothesis in that folding is not suggested to occur via a directed pathway, but via several metastable states implying there are multiple pathways for the folding process. This process is governed by the initial conditions of the system [31, 44]. This is known as the **metastability** hypothesis and is illustrated in figure 1.10 [31].

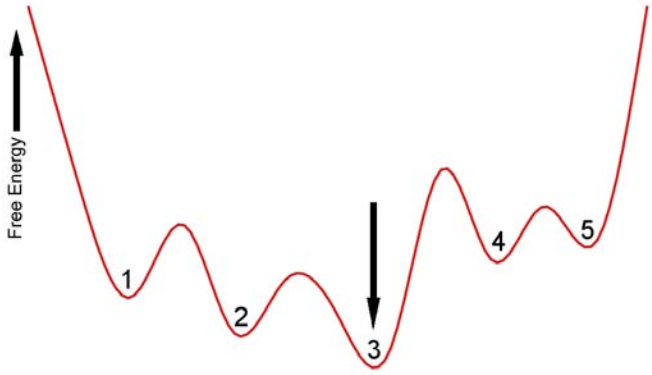


Figure 1.10: An illustration of the metastability hypothesis (adapted from [31]). A free energy profile corresponding to conformational space of a protein molecule. Each minimum, both metastable and global is labelled, with the global minimum (3) highlighted.

1.1.5 Modelling Proteins

There are a variety of protein models which differ in the way in which they approximate the protein molecule and how they treat the interactions between amino acid residues, and, if applicable, with solvents. Due to the enormous complexity and size of protein hypersurfaces, models used to study the protein folding process tend to be simplified [22].

1.1.5.1 HP Lattice Bead Models

The most simplistic of all models, the hydrophobic-polar lattice bead model (HPLBM) [45], has become one of the major tools for studying protein structure [21]. The basis of such a model is that the hydrophobic force is primarily responsible for the determination

of the unique native conformation and therefore the biological function of small globular proteins [23]. Although simple, such models can still capture some essential features of the protein folding problem and provide a basis for thorough theoretical studies [25].

The twenty naturally occurring amino acids can be roughly classified into two categories based on their hydrophobicity [23]. In the HPLBM, these two categories are exploited with amino acids categorised as either **H**ydrophobic (H) or **P**olar (P) residues [21, 23]. The primary amino acid structure of a protein, instead of comprising a sequence of the twenty amino acid alphabet, is therefore represented as a combination Hs and Ps, with each amino acid represented as a uniformly sized bead [46]. The conformations of such a sequence are restricted to a self avoiding walk on a lattice, where lattice sites can only be occupied by a single bead [45]. The presence of a lattice prevents bond lengths and angles from varying and thus both are constant throughout the use of this model [46].

The energy associated with any bead-bead interaction is described as a short range contact between topological neighbours [21, 23]. A topological neighbour is simply a pair of non-bonded beads that lie on adjacent lattice sites, i.e. they are not "sequence neighbours" [22]. Interaction values (ϵ_{ij}) for the possible topological contacts (local interactions) are [22, 24, 47]:

$$\epsilon_{HH} = -1 \quad \epsilon_{HP} = 0 \quad \epsilon_{PP} = 0 \tag{1.1}$$

The conformation energy of the model protein is obtained by summing over these local interactions [22]:

$$E = \sum_{i<j} \epsilon_{ij}.\Delta_{ij} \tag{1.2}$$

where

$$\Delta_{ij} = \begin{cases} 1 & \text{if } i \text{ and } j \text{ are topological neighbours, but not sequence neighbours} \\ 0 & \text{otherwise} \end{cases}$$

In order for the energy of such a model protein to be driven down, H monomers must congregate as much as possible, in turn, producing a larger number of topological contacts, mimicking hydrophobic collapse [23]. With the H-H interaction being the only energetically attractive contribution, it serves as the stabilising interaction for these model proteins. This effective attractive force, as a result, mimics the hydrophobic interaction found in real proteins [46], the driving force for proteins to fold to their native conformations, producing compact cores containing a wealth of hydrophobic residues [22, 23].

1.1.5.2 BLN Model

The hydropho**B**ic-po**L**ar **N**eutral model (BLNM) extends the methodology met in the simple HPLBM. Just like the HPLBM, the BLNM is classed as minimalistic, attempting to capture only essential features of the physical system being modelled. As seen with the HPLBM, BLNM heteropolymers do not contain side groups responsible for intramolecular hydrogen bonding [44].

BLNM proteins contain three different types of "residues" in the backbone of chains in the form of hydropho**B**ic (B), po**L**ar (L) and **N**eutral (N) [48]. In order to replicate hydrophobic collapse (mentioned in section 1.1.4.2), it is important to note that B residues attract other B residues and are the only attractive interaction. The primary amino acid structure of a protein, instead of comprising a sequence of the twenty amino acid alphabet, is therefore represented as a combination Bs, Ls and Ns, with each amino acid represented as a uniformly sized bead. Conformations of such amino acids are arranged in 3D space.

The HPLBM neglects long-range interactions between residues resulting in discrete integer energy values. The disadvantage of not considering these interactions is that the PES can appear flat due to a high number of degenerate minima. By overlooking

long-range interactions, an understanding of the role they play in the protein folding problem may also be missed. To reduce degeneracy, distance dependent long-range interactions must be taken into account. The BLNM considers numerous long-range interactions that contribute to a structure's energy via a continuous, distance dependent method [44]. Both are calculated in a pairwise manner throughout the structure, with R_{ij} representing the distance between bead i and j, and can be seen in equation (1.3) giving rise to the conformation energy (E_i) [41].

$$\begin{aligned}
\mathrm{E}_i \; &= E_{\mathbf{r}} + E_{\Theta} + E_{\Psi} + E_{\mathbf{R}} \\
&= \sum_{i}^{bonds} K_r \left(r_i - r_0^i\right)^2 \\
&+ \sum_{i}^{angles} K_\theta \left(\theta_i - \theta_0^i\right)^2 \\
&+ \sum_{i}^{torsional} \left[A\left(1 + \cos\Phi_i\right) + B\left(1 + \cos 3\Phi_i\right)\right] \\
&+ \sum_{i>j+3}^{nonbondingpairs} 4\epsilon S_1 \left[\left(\frac{\sigma}{R_i j}\right)^{12} - S_2 \left(\frac{\sigma}{R_i j}\right)^{6}\right]
\end{aligned} \tag{1.3}$$

where $S_1 = S_2 = 1$ for B-B (attractive) interactions, $S_1 = \frac{2}{3}$ and S_2 = -1 for L-L and L-B (repulsive) interactions, and $S_1 = 1$ and $S_2 = 0$ for all N containing bead pairs. $\epsilon = 0.0100570$ and is a constant interaction parameter used to determine the energy scale [49].

1.1.5.3 United- and All-Atom Models

As *ab initio* calculations, when used to study the folding of proteins, are computationally too demanding, the potential energy component of the Hamiltonian is represented as an empirical set of equations that describe the bonded and non-bonded interactions

between atoms for atomic systems [50]. These energy functions and their parameters are known as force-fields. Force-fields consist of two major components describing interactions between covalently bonded atoms (such as bond lengths, bond angles and dihedrals) and non-bonded interactions (such as van der Waals interactions, modelled via a Lennard-Jones function and Coulombic interactions) [49,50]. Popular force-fields include AMBER [51], CHARMM [52], GROMOS [53] and OPLS [54].

All-atom force-fields are the most realistic of all protein models. Unlike other models, the detail of the protein structure is not lost by grouping atoms together. They involve a higher level of detail in their representations of protein molecules, as every atom is considered and treated explicitly (including hydrogen). However, the complexity of the force-fields and the number of atoms in the average protein result in studied timescales being shorter than involved in the folding process (limiting to the nanosecond timescale). United atom force-fields, however, treat aliphatic carbons and associated hydrogens as a single particle [50]. It is a style of coarse-graining, and is effective in reducing the complexity of simulations while preserving accuracy [49].

1.2 Natural Immune System

An immune individual is one that exhibits no symptoms of a disease once it has entered its body. Immunology is the study of defence mechanisms that provide resistance against disease [55]. The system employed by our bodies to identify and eliminate external microorganisms is known as the *immune system* [56].

The immune system plays a major role in an animal's survival, in that a large number of specific mechanisms must act on it efficiently and effectively. These mechanisms are optimised for certain roles, for example, a specific microorganism or a range of infecting agents. It consists of a two-tier line of defence, the *innate immune system* and the *adaptive immune system*, with both systems depending upon the activity of *white blood cells* (leukocytes) [55]. The immune system holds a great redundancy of mech-

anisms, such that many combinations can be used against a single agent. A typical mammal is thought to possess around 10^7 - 10^8 different antibody types [56].

The process whereby the immune system is able to selectively recognise foreign invaders is called the **immune response** [56]. In the case of the innate immune system, this response is known as the *innate immune response* and for the adaptive immune system, it is called the *selective immune response* or *adaptive immune response*. Innate immunity is governed predominantly by *granulocytes* and *macrophages*, with its response remaining constant regardless how many times an infectious agent is encountered. However, adaptive immunity, mediated by *lymphocytes*, improves with repeated exposure to a given infection [57]. Whereas the innate immune response is rapid, but can be damaging to tissues, the adaptive immune response may take several days or weeks to develop as a result of its specificity. The adaptive response also has memory, resulting in the response becoming rapid upon subsequent exposure. However, this is not immediate [58]. The adaptive immune response is our only concern here.

1.2.1 Specific Immunity

Lymphocytes, present in the adaptive immune system, are responsible for the recognition and elimination of pathogenic agents (infectious agents or pathogens) [55]. There are two main types of antigen-specific lymphocyte, *B lymphocyte* (B-cell) and *T lymphocyte* (T-cell) [58]. Both B- and T-cells have highly specific antigenic receptors on their surface [55]. Adaptive immunity uses antigen-specific receptors to initiate responses in two stages. First, the antigen is presented to and recognised by the B or T cells [58]. Upon recognition, the lymphocytes interact with an antigenic stimulus (antigen), becoming active, reproducing by means of cell division [55]. Second, an effector response occurs, either due to an activated T-cell (leaving the lymphoid and focusing on the disease site), or due to the release of an antibody (from an activated B-cell into the blood and tissue fluids) [58]. The antibody combining region or paratope is the portion of the antibody molecule used in the identification of other molecules [56].

After exposure to a disease or vaccination, the pathogen (or more commonly germ) proportions immune memory [55].

1.2.1.1 Pattern Recognition

Antigens (substances promoting antibody generation) may be surface molecules present on pathogens, or self antigens composed of cells or molecules of the infected animal. For the immune system to work, antigens must be recognised by *surface receptor molecules* carried by B- and T-cells [55]. The surface receptors of the B- and T-cells are of a certain "shape" that has to be matched by the shape of the antigen. B-cell receptors (BCRs) interact with the antigens that are free in solution, whereas T-cell receptors (TCRs) can only interact and bind with antigens presented by molecules of the host's own body [55, 59].

Upon activation of the B-cell, the antigen B-cell receptors bound to the cell membrane will be secreted in the form of antibodies. The main role of the B-cell is to produce and secrete antibodies in response to pathogenic agents. These antibodies are capable of recognising and binding to a determined protein (in this case broken down portions of the antigen). This secretion and binding of antibodies is a form of signalling to other immune cells to ingest, process and/or remove the bound substances [55].

Pattern recognition occurs at the molecular level and is based on antigens and cell receptors having complementary shapes enabling them to bind together, thereby triggering an immune response [55, 59]. Binding occurs between the receptor of the antibodies (paratopes) and the *epitope* of the antigen [56].

Antibodies posses two paratopes, which are portions of the antibody that are used to identify other molecules [60]. Paratopes and epitopes (regions on other molecules that the paratopes can bind to) are complementary [61]. For an organism to survive, it must be able to produce paratopes that can bind to any epitope. However, as the number of possible epitopes is so large, paratopes must be capable of binding to a whole host of epitopes. The DNA within a cell, contains a large number of building

blocks also allowing for the specifities to overlap, rendering many paratopes able to bind to a single epitope [56]. Even though antibodies are equipped with only a single type of receptor, the antigen may present several epitopes, allowing different antibodies to recognise a single antigen [55]. The closer the match between antibody and antigen, the stronger the molecular binding and the better the recognition [60].

The thymus is an organ in the chest cavity, behind sternum which provides an area for T-cell maturation. B-cell maturation, on the otherhand, takes place in the bone marrow. T-cells can be divided into two main categories: helper T-cells (T_H cells) and killer T-cells (T_K cells) [62]. The antigenic receptors of T-cells are structurally different from B-cell receptors, as they must recognise antigens presented and processed by other cells. T-cells are able to regulate other immune cell activity (T_H cells) and also directly attack cells causing infection to the host (T_K cells) [55].

1.2.2 Self/Nonself Discrimination

Antibody molecules and T-cell receptors produced by lymphocytes have the ability to recognise a variety of molecular shapes; self (self-antigens), non-self (antigens) and artificially synthesised molecules and give rise to immune system completeness [55]. Antibody molecules have antigenic patterns (idiotopes or shapes) that can be recognised by the antigen binding sites of other antibodies. The completeness of the immune system to recognise antigens suggests that all idiotopes will be recognised by at least one antibody molecule [55, 59]. In order to achieve this, the *repertoire* of B- and T-cells available to the immune system must be diverse (obtained by mutation, editing and gene rearrangement), *cross reactive* and *multi-specific*. It is the cross reactivity (recognition of related antigenic patterns) and multi-specificity (recognition of different chemical structures) that allow a repertoire of lymphocytes to have the ability to recognise and bind to a set of antigens that is much bigger than the repertoire itself.

In order for the immune system to function correctly, it must be able to recognise the difference between host molecules (self), foreign molecules (non-self) and indistinguish-

able molecules. Autoimmune diseases are caused by the immune system's inability to make this distinction, allowing immune responses to be engaged against self-antigens. The immune system's capability of not responding to a self-antigen is known as *self tolerance*. Understanding how the immune system accomplishes this distinction is known as the *self/nonself discrimination problem*.

An interaction between an antigen and a lymphocyte does not necessarily mean that the lymphocyte will be activated. A process known as *negative selection* prevents the lymphocyte from being *autoagressive* (self destructive) in that a self-specific lymphocyte will not attack one of its own kind [55].

1.2.3 Affinity Maturation

Over time, T-cell dependent immune responses show an improvement in affinity of antibody for antigen. This process is known as **affinity maturation** [63]. This maturation requires the antigen binding sites of the matured response to be structurally different from those of the primary response.

During clonal expansion, random changes are introduced into the variable region of genes, with one occasionally resulting in an increase in affinity of the antibody. Selection of the high-affinity variants must then occur, creating a pool of memory cells. The diversity of this pool is maintained through hypermutations. However, domination of the response can occur by selection of B-cells with high-affinity mutant receptors. The introduction of random changes results in a large proportion of the mating genes to become non-functional or self-reactive. These self-reactive cells or cells with low-affinity receptors must be removed to prevent them from significantly contributing to the pool of memory cells [55].

1.2.4 The Clonal Selection Principle

The **clonal selection principle** is the term used to describe the basic properties of how the adaptive immune system responds to an antigenic stimulus. Clonal selection

operates on both B- and T-cells and illustrates how only those cells able to recognise the antigenic stimulus will reproduce and be selected over those lymphocytes that do not [64].

When an antigen comes into contact with a B-cell and binds with its antibodies, the B-cell becomes activated and starts to proliferate and differentiate (reproduce) into numerous effector cells [55, 64, 65]. This results in an increase in the number of B-cells that recognise the antigen. Many new identical clones of the parent B-cell are produced via a process called clonal expansion, and undergo somatic hypermutation. These mutated clones produce antibodies that are specific to the invading antigen [64, 65]. As a result, the recognition of the antigen by the B-cell receptors is fine-tuned, determining the binding strength (affinity) of the antibody [57]. The higher the affinity of a B-cell to the antigen, the more likely it is to clone. This results in a Darwinian process of variation and selection, called *affinity maturation* [65].

During B-cell clonal expansion, B-cells undergo somatic hypermutation during reproduction, with B-effector cells being active antibody producers. Each B-cell produces only one type of antibody (monospecific), which is relatively specific to an antigen. Antigenic receptors on the B-cell bind to the antigen, allowing the antigen to stimulate a B-cell with the help of a second signal (co-stimulatory signal) from other immune cells, such as T_H-cells. This stimulation of the B-cell causes it to proliferate (divide) and mature into either terminal (non-dividing) antibody secreting cells called *plasma cells* or long-lived *B-memory cells*. Plasma cells are the most active high affinity antibody secretors. However, dividing B-cells also secrete antibodies but at a much slower rate. Although B-memory cells do not manufacture antibodies, they will rapidly differentiate into plasma cells when exposed to a second antigenic stimulus.

T-cells, however, do not undergo somatic hypermutation during reproduction. Instead of antibodies, T_K cells produce immune signalling cells known as *lymphokines*, notifying other immune cells such as macrophages of infected sites. The effect of mu-

tational and selectional operations during the clonal expansion of B-cells, allows an increase in the diversity and enables the recognition of specific antigens [55].

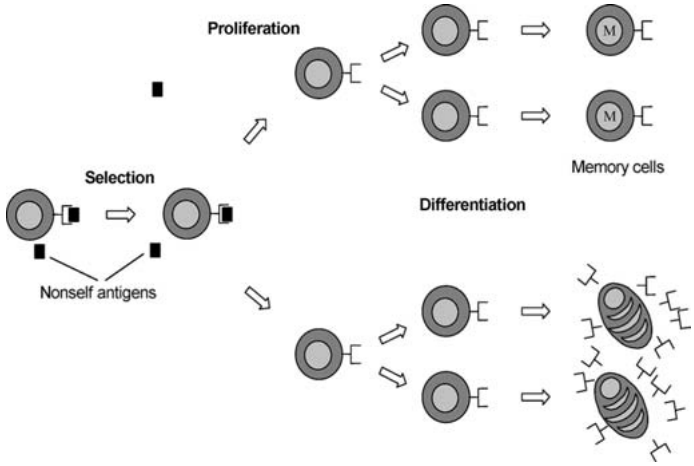


Figure 1.11: A schematic of clonal selection. Immune cells whose receptors recognise and bind with nonself antigens are selected to proliferate and differentiate into memory cells. Taken from [59].

1.3 Evolutionary Computing and Search Techniques

Evolutionary Algorithms (EAs) are popular methods for solving a variety of global optimisation problems [66]. The ability of an individual to adapt is of paramount importance for its survival in a dynamically changing environment. Improving the proficiency of individuals to survive is an optimisation process known as *evolution* [67]. As its name suggests, the model of evolutionary computation is the imitation of natural evolution within a computational search procedure [68], the main concept being the Darwinian theory of survival of the fittest [26, 66]. It is believed that, as Nature evolves, individuals must adapt to their co-evolving environment in order to survive [26]. Survival is achieved through reproduction, where offspring (from two or more parents in artificial systems) contain genetic material from the parents. In some cases, the best characteristics of the parents will be inherited by the offspring. For individuals containing poor characteristics, the fight for survival is lost.

In evolutionary computing, a *population* of individuals is modelled, with a single individual referred to as a *chromosome*, defining the characteristics of the individual in a population [69]. Each element of a chromosome is referred to as a *gene* with the value of a gene known as an *allele*. For each generation of the evolutionary process, individuals compete to reproduce, with individuals having the best chance of survival being the ones having greater chance to do so. By combining sections of the parent's chromosomes, offspring are generated, a process referred to as *crossover*. Individuals may also undergo a *mutation* process where genes within the chromosome are altered (i.e. alleles are changed) [70]. The *fitness* of an individual is a measure of its tendency to survive, usually in the form of a function reflecting the objectives and constraints of the problem to be solved [68]. Both the crossover and mutation processes can affect the fitness of the evolved individual (when compared with the parents) in a positive or negative fashion. In order to maintain the population size, individuals may undergo *culling* or if a strategy of *elitism* is employed, survive until the next generation [67].

1.3.1 Genetic Algorithms

The Genetic Algorithm (GA) [71,72] is a search technique incorporating the principles of genetic (natural) evolution [67, 73]. It uses operators analogous to evolutionary processes, such as mating (or "crossover"), mutation and natural selection [26]. The mating operator exchanges information between individuals in the hope of evolving new and improved solutions to the problem being optimised. The ability of a GA to explore multiple regions of parameter space simultaneously (operating in parallel) is the feature which is responsible for its success in many areas [74]. Information can be dispersed throughout the population of individuals by the exchanging of genetic information via the crossover procedure. The GA's aptitude for identifying favourable regions of search space is shown by its ability to recognise schemata. A good schema corresponds to a set of optimal or near optimal alleles that can propagate through the population, resulting in individuals of relatively high fitness [73].

Algorithm 1.1 Pseudocode for a general GA (adapted from [67]) illustrating how the procedure acts on a population of individuals, repeating mating, mutation and selection until the convergence criteria are met.

1: Let generation, $g = 0$.
2: Create the initial population P_g.
3: **while** not converged **do**
4: Evaluate the fitness of each individual $\vec{P}_{g,n} \epsilon P_g$.
5: $g = g + 1$.
6: Select parents from P_{g-1}.
7: Mate selected parents through crossover to form offspring O_g.
8: Mutate offspring in O_g.
9: Select the new population P_g from the previous population P_{g-1} and the offspring O_g.
10: **end while**

1.3.2 Differential Evolution

Differential Evolution (DE) is a population-based search strategy sharing common features with standard evolutionary algorithms. Like other EAs, DE uses mating, mutation and natural selection to search parameter space in a parallel fashion for the fittest solution to the problem to be solved [75]. However, one main difference between DE and other EAs is that mating and mutation occurs in a single step [67]. DE is self-organising in that the difference between randomly chosen individuals is used to perturb an existing individual. This perturbation occurs for every member of a population [75]. Another difference is that the selection process is deterministic in that a direct comparison between parent and child exists such that the fitter is allowed into the population.

Although DE is a relatively new EA [66], it can offer fast convergency, robustness and simplicity. This, combined with the ease of implementation and a small number of search parameters, has resulted in its use for complex optimisation problems [76]. DE will be met in more detail in section 2.3.

Algorithm 1.2 Pseudocode for a general DE taken from [67] illustrating how the procedure acts on a population of individuals, repeating mating, mutation and selection until the convergence criteria are met.

Let $g = 0$ and initialise p_r and γ.
Initialise a population C_g of N individuals.
while not converged **do**
  for each individual, $\vec{C}_{g,n}(n = 1, ...N)$ **do**
    select n_1, n_2, $n_3 \sim U(1, ..., N)$, with $n_1 \neq n_2 \neq n_3 \neq n$.
    select $i \sim U(1, ...I)$
    for $j = 1, ...I$ **do**
      if $U(0, 1) < p_r$ or $j = i$ **then**
        $O_{g,nj} = C_{g,n3j} + \gamma(C_{g,n1j} - C_{g,n2j})$
      else
        $O_{g,nj} = C_{g,nj}$
      end if
    end for
  end for
  Select the new population C_{g+1} of N individuals.

$$\vec{C}_{g+1,n} = \begin{cases} \vec{O}_{g,n} & \text{if } \mathcal{F}_{DE}(\vec{O}_{g,n}) \leq \mathcal{F}_{DE}(\vec{C}_{g,n}) \\ \vec{C}_{g,n} & \text{otherwise} \end{cases}$$

end while

1.4 Swarm Algorithms

From studying the social behaviour of individuals in swarms or colonies, came the design of efficient optimisation and clustering algorithms [67]. The study of bird flocks led to the design of Particle Swarm Optimisation (PSO) algorithms [77], with the study of the foraging behaviour of ants resulting in Ant Colony Optimisation (ACO) algorithms [78].

1.4.1 Ant Colony Optimisation

Ants work on the concept of *stigmergy* in that they are able to indirectly communicate through environmental interactions. During the foraging process, ants indirectly communicate with each other by laying pheromone trails on the ground, influencing decisions to be made by other ants. It is this simple form of communication that gives rise to the complex behavioural patterns of the entire ant colony [21]. Ant Colony Op-

timisation (ACO) is a population based approach inspired by the foraging behaviour of ants used to solve combinatorial optimisation problems. Fundamentally, it is an iterative process in which a population of agents (in this case "ants") repeatedly construct candidate solutions to a given problem by probabilistically using a combination of heuristic information and the "pheromone trails" left by previous ants [21, 78].

Algorithm 1.3 Pseudocode for a general ACO (adapted from [79]) illustrating how the procedure acts on a colony of ants, repeating solution creation, updating of the pheromone level to emphasise favourable solutions and reduction of the pheromone level to eventually remove unfavourable solutions for each ant, until the convergence criteria are met.

```
1: Assign the same initial pheromone value to each edge of the graph, and randomly
   place an ant in a location of the search space.
2: while Not Converged do
3:    for each ant, a do
4:       Probabilistically move a over the space to build a solution to the problem.
5:       Evaluate the fitness of the solution obtained by a.
6:       update the pheromone level of each edge by reinforcing good solutions.
7:       Reduce the pheromone level of each edge.
8:    end for
9: end while
```

1.4.2 Particle Swarm Optimisation

PSO is like a GA, in that it is initialised with a population of random solutions. It is based on memory and awareness, with a particle being a potential solution to a problem. Each particle is assigned a random velocity and is then flown through hyperspace. However, if a particle deviates from the swarm it is pulled back into formation. Each particle keeps a track of its coordinates in hyperspace, its current fitness and its best fitness so far. Globally, the overall best fitness and its location are also recorded, with individuals being drawn towards it as the system iterates [77].

1.5 Immune Algorithms

An Immune Algorithm (IA) is a population based search technique that incorporates the principles of the natural immune system. Many implementations of IAs exist, such

Algorithm 1.4 Pseudocode for a general PSO (adapted from [79]) illustrating how the procedure acts on a population of particles, repeating current and best comparison, current and neighbourhood comparison, velocity determination and position update for each particle, until the convergence criteria are met.

```
1: Randomly initialise a population of particles and set the current best, b.
2: while Not Converged do
3:   for each particle, p do
4:     Evaluate the fitness, f of p.
5:     if f_p > f_b then
6:       b = p.
7:     end if
8:     if f_p > f among all neighbours then
9:       neighbourhood best, n = p
10:    end if
11:    Determine the velocity, v of p in its current trajectory as a function of the
       difference between p's previous best and current positions, and the difference
       between n and p's current position.
12:    Update p's position by adding its previous position to v.
13:  end for
14: end while
```

as Artificial Negative and Artificial Positive Selection algorithms. However, here only the Clonal Selection Algorithm is discussed.

1.5.1 Clonal Selection

Clonal selection focuses on the ability of the B- and T-cells to adapt in order to match and kill an antigen. A Clonal Selection Algorithm (CSA) [80] is a population based search technique that mimics the ability of the immune system of adapting B-cells to new types of antigen, powered by processes such as clonal selection and affinity maturation by hypermutation [81]. New genetic information is added to a population using a mutation phase that acts on a series of clones of individuals of the population. The clones allow for local regions of search space of each individual to be explored. The mutation phase involves inversely proportional hypermutation, where a clone exhibits a series of point mutations inversely proportional to its affinity towards the antigen.

Algorithm 1.5 Pseudocode for a general CSA (adapted from [82]) illustrating how the procedure acts on a population of individuals, repeating cloning, mutation, selection and memory storage until the convergence criteria are met.

1: Randomly initialise population of individuals P and memory set M.
2: **while** Not Converged **do**
3: Calculate the fitness of each member of P with objective function or pattern.
4: Clone and mutate n highest fitness members of P.
5: Select j highest fitness members from P and place in M (replace in M if new fitness is higher).
6: Replace I lowest fitness members of P with randomly generated individuals.
7: **end while**
8: Output M.

1.6 Generic Operators

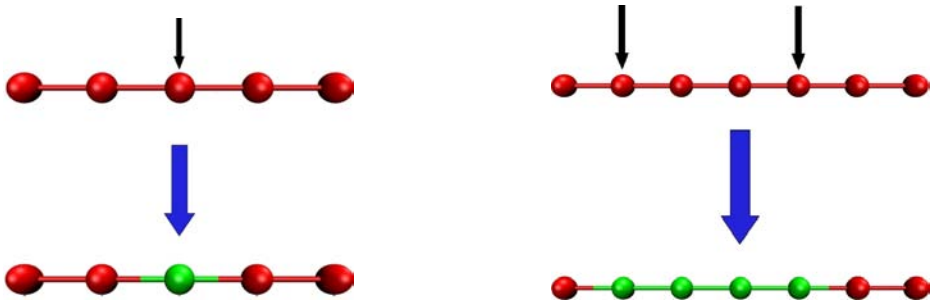
1.6.1 Mutation

Whereas the crossover operator allows the exchange of genetic material between individuals, *mutation* allows the introduction of new genetic material [67, 70, 73]. The primary aim of a mutation operator is to ensure that a full range of genetic material is available to the search mechanism. If no new genetic material is introduced, a possibility of stagnation arises due to the diversity of the population being poor. As shown in figure 1.12(a), new genetic material is introduced by making one or more random changes to the genes of an individual [83] and thus will add diversity to the genetic characteristics of the population [84].

Mutation falls under two categories; **static** where the genetic change is completely random and **dynamic** where a small change is performed about the initial value [73]. An important consideration is the mutation probability. A small probability value is usually used to ensure only minor distortions occur to good solutions. However by initialising a large mutation rate and decreasing it over time, a much larger search space is covered at the beginning of the search due to larger distortions, with much smaller distortions occurring as solutions begin to approach the optimum [67].

Selection of the genes for mutation also fall under two categories; **random mutate**,

where genes are selected completely at random (refer to figure 1.12(a)) and **inorder mutate**, where two genes are randomly selected, with each gene within this range being mutated as shown in figure 1.12(b).



(a) (b)

Figure 1.12: (a) Schematic representation of a point mutation (adapted from [73]). Shown is a point mutation where a gene of an individual is selected (represented by the black arrow) and its value changed. (b) Schematic representation of inorder mutation (adapted from [67]). The black arrows show the selected genes. It should be noted that point mutations occur between these genes.

1.6.2 Selection

When considering search methods, individuals of the population represent potential solutions to the optimisation problem. Each generation produces a new set of individuals and therefore new potential solutions. Generally a new population is determined via the process of crossover, mutation and elitism [67]. The selection operator may provide an opportunity for the better candidate solutions to reside in the population and the poorer ones to be removed [85]. In the case of crossover, individuals of a higher fitness should have a greater opportunity to reproduce, thus allowing offspring to be formed using combinations of genetic material from higher fitness (fitter) individuals. In the case of mutation, the logic may be somewhat different. In order to preserve a high fitness individual, it may be favourable to provide it with less chance of mutation, i.e. removing them from the selection procedure. During the elitist phase, only the fittest individuals may be selected for the next generation ensuring that a decrease in

the best fitness is not seen [84].

Selection techniques exist as either **explicit** or **implicit** fitness remapping. Explicit fitness remapping uses the normalisation of the fitness value between the values of 0 and 1. An example of this is **proportional selection**, where probabilities of selection (P) are calculated by dividing an individual's fitness by the sum of all fitnesses of the population, as shown in equation (1.4). The disadvantage with this form of selection is the population diversity can be limited as certain individuals may start to dominate offspring production [67].

$$P_i = \frac{\mathcal{F}_i}{\sum_{n=1}^{N} \mathcal{F}_i} \tag{1.4}$$

where i represents an individual in the population and $\mathcal{F}_i$ is its fitness.

A common method employing this proportionality is *roulette wheel* selection [71,72]. Visualising a roulette wheel, the segment widths reflect an individual's probability with selection being performed by "spinning" the wheel. If the probability of the selected individual is greater than a random value between 0 and 1, the chosen individual is tested, otherwise another selection is made [73, 86].

In contrast, implicit fitness remapping takes advantage of using the actual fitness value of the individual for selection. *Tournament selection* [71, 72] adopts this strategy where a number of individuals are selected from the population to take part in a tournament [87]. The individual with the highest fitness wins the tournament [88]. In the case of crossover, two tournaments must be held, one for each parent. The disadvantages of tournament selection in its simplest form are that an individual could be selected more than once and that an individual can mate with itself to produce offspring which are identical to the starting structure. The advantage is that as unfit individuals will tend not to win a tournament, they will consequently not be selected and will probably not contribute to the formation of offspring via mating. Tourna-

ment selection often prevents the highest fitness individuals from governing mating, as tournaments can also take place between pairs of low fitness individuals.

Rank-based selection involves ordering of the population in order of fitness [86, 87] to determine the relative probability of selection [89]. As an advantage, high fitness individuals will not dominate the selection process due to a fitness independent probability calculation. An example of rank-based selection is *non deterministic linear sampling*, where the selection operator selects an individual at random from the previously ordered population. *Nonlinear* selection operators tend to select the better individuals of a population risking premature convergence.

Random selection as the name suggests, occurs when fitness is completely ignored and individuals are selected in a completely random fashion. This type of selection is independent of any fitness remapping procedures [67].

1.6.3 Mating

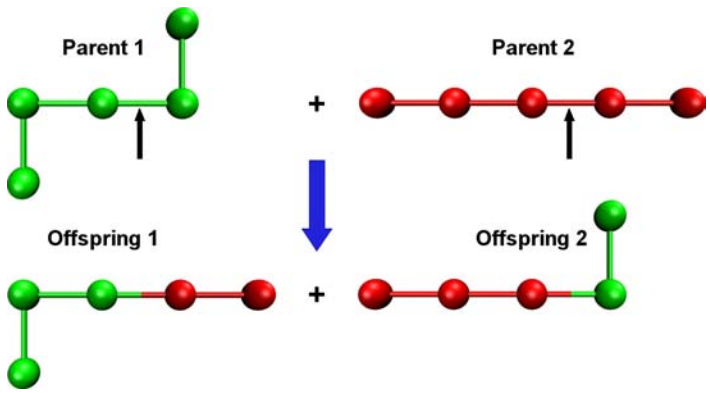


Figure 1.13: A schematic representation of one point crossover (adapted from [73]), highlighting the single point where each parent chromosome is cut (represented by the black arrow) to form offspring. Offspring are generated by combining complementary genes from the parents.

The purpose of mating is to use two or more parents to produce offspring. Selection operators introduced in section 1.6.2 and sometimes probability is used to determine which parents will take part [67]. Mating is a term used to describe the transfer of

genetic information from more than one parent to their offspring [83]. In evolutionary computation, mating is represented by a procedure known as *crossover*. Different methods of crossover exist, the simplest being **one-point crossover** which is illustrated in figure 1.13. Each of the parent chromosomes are cut at a single identical point. As a result of this separation, complementary genes from each of the parents (i.e. the first part of one parent with the second part of the other and vice versa) are combined to produce offspring [73].

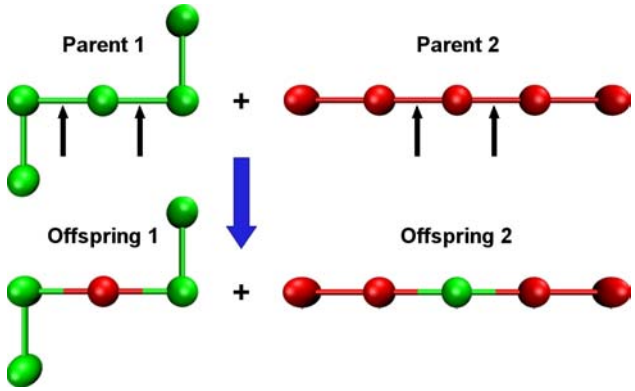


Figure 1.14: A schematic representation of two point crossover (adapted from [73]), highlighting the two points where each parent chain is cut (represented by the black arrows) to form offspring. Offspring are generated by combining complementary genes from the parents.

Two-point crossover as seen in figure 1.14, involves the determination of a segment of each parent chromosome by making two point selections. Genetic information is then transferred to produce offspring by replacing the central selected segment of one parent with the same segment of the other parent, and vice versa. This may result in more chromosome alteration than one-point crossover does.

Other forms of crossover include **uniform crossover** where offspring are generated by taking a certain number of genes from each parent with no restriction on where in the chromosome they can be taken from [73].

Chapter 2

Methodology

Unless otherwise stated, all algorithmic implementations were created by the author using the C++ programming language and the standard template libraries (STL). All data plots were created using gnuPlot: an interactive plotting program, with protein conformations rendered using the Visual Molecular Dynamics (VMD) [90] package. Root Mean Square Deviation (RMSD) is calculated using an interpretation of the Kabsch algorithm [91, 92], with the system independent random number generator taken from numerical recipes in C++ [93]. Descriptions use genetic terminology where appropriate. All calculations were performed using the University of Birmingham's high performance computing service [94].

2.1 The Basic Immune Algorithm

The complete IA proposed here is based on previous work by Cutello et al. [65] and the clonal selection principle as described in section 1.2.4. IAs are population based search techniques and require a population of individuals (or B-Cells) in order to search (in our case) areas of the PES simultaneously and efficiently. Here the B-Cells are modelled as individuals in a population representing trial solutions to the protein folding problem of finding the most energetically favourable 3D conformation (secondary and tertiary protein structure) from only knowledge of the primary sequence. The antigen in the case of the IA is the lowest energy conformation (highest fitness). In some models used

in this work, as is the case in the clonal selection principle, the antigen is unknown and so the algorithm must learn about this species as the calculation proceeds.

In order to provide a strong defence against foreign body attack, B-Cells undergo cloning upon recognition of an antigen, even if the antigen has never been encountered before. Once the initial population of individuals is randomly generated (details presented in section 2.1.1), the B-Cells are cloned and mutated repeatedly (in the form of generations) until a viable solution to the problem is found. Each generation (g) also consists of a selection operator which probes that generation's population, preparing a new population to be exploited by such operators in future generations.

Upon initialisation of the algorithm, various parameters are required in order to provide the algorithm with valuable information on how to approach the problem in hand. As our search problem is the protein folding problem, we must first provide the algorithm with a primary sequence of residues. It is then the job of the algorithm to decipher which 3D conformation provides the lowest energy. Before we subject our primary sequence to the algorithm operators, we must provide our algorithm with a little more information. With the search procedure being population based, we must also provide a population size (n_{ind} or number of individuals) as seen with GAs. Once a population of size n_{ind} is randomly generated, the generation scheme is entered. As will be made apparent in sections 2.1.5 and 2.2.2, a cloned population must be present for our manipulation operators to act upon. For this to be the case, the number of times each individual is to be cloned number of clones (n_{clo}) must be specified. In the original methodology proposed by Cutello et al. [65], to reduce computational time, a mutation factor (m_f) is applied in order to calculate the number of mutation attempts allowed in the mutation phase (seen in more detail in section 2.1.5.1). With B-Cells having a finite lifetime, the algorithm must reflect this and the maximum individual age (i_{max}), in terms of generations, before the individual is removed from the calculation, must be defined. Finally, the maximum number of iterations (g_{max}), or the length of

time the calculation is given to progress, must be defined (in terms of g). Once this information is catered for, the population of individuals is subjected to our algorithm operators as outlined in algorithm 2.1.

Algorithm 2.1 Pseudocode for the IA illustrating how the procedure acts on a population of individuals, with each individual being subjected to repeat rounds of cloning, mutation and selection until the lowest known energy conformation is found or the maximum number of generations has been exceeded.

```
1: Assign g_max, n_ind, n_clo, g and m_f
2: Generate the initial population containing n_ind individuals.
3: Create empty populations for clones and mutants.
4: g is zero.
5: while Lowest energy conformation has not been found or g ≤ g_max do
6:    Produce n_clo clones of each individual and separate to the cloned population.
7:    Hypermutate each clone and separate to the mutated population.
8:    Hypermacromutate each clone and separate to the mutated population.
9:    Age each individual in the mutated and current population.
10:   Select individuals for the next generation.
11:   Increment g.
12: end while
13: Output lowest energy conformation.
```

2.1.1 Generating the Initial Population

In order for any algorithm to function, a starting point needs to be generated in the form of a calculated trial solution, or a series of trial solutions for the case of population based search techniques. In the work presented here, the search techniques are all population based methods and require numerous trial solutions to provide the infrastructure of the search itself.

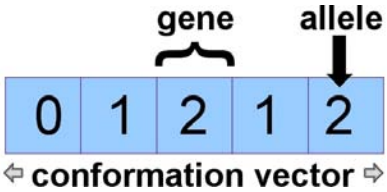


Figure 2.1: Schematic of a conformation vector containing five genes, each of which has an integer allele.

For these search methods, it is the chromosome that completely defines the conformation, energy and fitness of any individual in a population. A population of individuals can be thus generated simply by considering a decision matrix at each locus. Randomly generating a value for each gene to adopt, will provide a complete chromosome and in turn a complete trial solution. The chromosome in any case is typically represented as a conformation vector containing the random values made available from the decision matrix. An example five gene conformation vector can be seen in figure 2.1, illustrating that a gene is a container for an allele represented as an integer in this work, with the conformation vector itself (the chromosome) being a container for the five genes.

There is an issue that must be considered when randomly generating a population of individuals in this way for the problems visited here. Structural validity is of prime importance in order to keep within the context of the science involved. Atomic repulsions prevent atoms from being in too close proximity, and also even in real space, atoms cannot occupy the same point. For this reason the protein models which have been developed take these into account and therefore the algorithms must interpret these constraints, even for simple lattices. When generating an individual, more than one atom cannot occupy a single lattice site, even in the case of the dynamical lattice model (DLM) where residues are deemed to have infeasible positions if a single atom is out of place. Therefore, checks need to made whilst generating an individual. It is for this reason that a recoil growth algorithm (RGA) is employed by the constructor to correct any structural infeasibility as the chain is built. Pseudocode for the RGA can be found in algorithm 2.2.

The RGA guarantees valid conformations and has the ability to produce compact arrangements. The process starts with the randomly generated conformation vector, which is manipulated as the chain grows, if the investigated allele produces invalidity. The RGA's first task is to suggest alternative alleles, leaving the preceding chain

Algorithm 2.2 Pseudocode for the RGA illustrating how the procedure acts on an individual's chromosome, revisiting previous atoms in a reverse order until a valid structure is obtained.

```
1: while Alleles still need to be investigated do
2:   if Last atom placement attempt failed then
3:     Test atom placement using all elements of the decision matrix, identifying
       available ones.
4:   end if
5:   if Available elements of the decision matrix exist then
6:     if Current allele is available then
7:       place atom.
8:     else
9:       Generate another random allele from decision matrix and place atom.
10:    end if
11:    Move to next atom.
12:  else
13:    Atom failed to be placed, move to previous atom
14:  end if
15: end while
16: Calculate energy and fitness.
```

segment as intact as possible if the randomly generated one is unsuitable. Only if no other alleles solve the invalidity issue, the constructor revisits previous chain positions in a reverse manner and attempts to find another route through conformational space until all atoms have been placed, rendering the conformation valid. For this to happen, the RGA must distinguish between available and unavailable alleles. As a result, once a valid conformation is discovered, the algorithm is aware of the elements of the decision matrix available for future mutation possibilities in relation to the final structure determined. This concept is known as mutation memory and is explained in detail in section 2.2.1.

Once the chromosome is valid (once having tested and placed all atoms), the energy and fitness are then calculated according to the model being used. A population of individuals are created using the same protocol, with the procedure repeated until n_{ind} valid individuals have been generated.

2.1.2 Fitness

The fitness (F_i) of a chromosome is a measure of its quality with respect to the function being optimised [68,95]. For a maximisation problem, high fitness would reflect a high function value and in the case of a minimisation problem, a low value of the function would be obtained [73]. Mimicking the thermodynamic hypothesis described in section 1.1.4.4, the problems presented here are all minimisation problems. Our fitness values are directly proportional to E_i according to equation (2.1). Fitness values may also be used to rank population members for use with other operators, such as selection, as described in section 1.6.2.

$$F_i = -E_i \tag{2.1}$$

2.1.3 Ageing

In section 1.2.4, it was stated that antibodies provide the basis for the natural immune system. The antibodies, which undergo mutation and cloning, must carry out their specific antigen pattern recognition until they die. In terms of the CSA, the ageing operator broadens diversity in a population by allowing B-Cells that have not successfully mutated (no improvement in fitness) to be removed after a length of time. The ageing operator's function effectively removes potential problem individuals that may hinder the discovery of the GM.

In order to take into account the mean lifetime of a B-Cell, individuals include an integer age counter [65]. The algorithm-independent ageing operator provides a mechanism to age each member in the current population by a single generation. In computational terms, the age counter, which is initiated with a value of zero for each new individual and for energetically improved individuals via other operators, is incremented by one for each generation that the individual survives.

The ageing operator is a means of instilling diversity into the population by preventing stagnation and reducing local minimum trapping. This is achieved after the mutation phase by removing any B-Cells with an age greater than the maximum B-Cell age, preventing any B-Cells that are "too old" from entering selection for the next generation. This operator is also independent of fitness and therefore even if a fit member of the population has not undergone an energetic improvement during the mutation phase for a i_{max} number of generations, it too is deleted from the population.

2.1.4 Cloning

In biology, a clone is a cell, cell product, or organism that is genetically identical to the unit or individual from which it was derived. B-Cells exist as clones in order to provide a stable defence against antigens by instantaneous recognition of their specific patterns. In order to provide a wealth of opportunities for the immune system to not only recognise but fight the antigens, clones are introduced. In terms of the clonal selection algorithm, the clones allow duplication of individuals, providing many copies of local structures within a population.

In order to model this idea, the B-Cells of a population are also cloned every generation. A clone is defined as simply an individual having the same chromosome and age as another. It is then this information that allows the clone to have access to other individual traits such as structure, conformation energy and fitness. The clones provide a stable foundation for mutations to occur, leaving the original population untouched. This provides the algorithm with numerous opportunities to improve each member of the original population, performing mutations on each. Cloning occurs by simply making a copy of each individual a specific number of times. Each clone is then subjected to both of the standard mutation operators, as described in section 2.1.5.

2.1.5 Mutation

A mutation allows for the introduction of new genetic matrial into a population [70]. In the IA, two types of mutation occur every generation and act only on the cloned population. The first is hypermutation which is described in section 2.1.5.1. The second is hypermacromutation, which is described in section 2.1.5.2. As a result of the mutations, a mutated population is generated, which is used by the selection operator. In all lattice problems, static mutation usually occurs by exchanging one allele for another, resulting in a different conformation, often leading to a different energy and fitness. The two operators described here utilise this and exploit it in different ways. A successful mutation is defined as one that gives rise to a different valid conformation. Model specific mutations, such as "crank shaft", "snake rotation" [22] or "corner-change" [95] used by the HPLBM, are not considered in this work.

2.1.5.1 Hypermutation

Hypermutation, as the name suggests, is an overactive mutator in that point mutations are attempted until it either succeeds or a maximum number of attempts has been reached. This method of mutation, combined with the idea of cloning, allows many adjacent points on the PES to be searched. It is a problem-independent operator with regards to lattice models due to a point mutation being executed by changing an allele value, as previously described.

The allowed number of mutation attempts is a function of chain length (n_{beads}), F_i and the current population's best fitness (F_b). A factor m_f is incorporated to scale the number of mutations (n_{mut}) made available to the mutator in order to optimise the mutation process computationally. n_{mut} is calculated as described in equation (2.2).

$$n_{mut} = \begin{cases} \left(\left(1 + \dfrac{F_b}{1}\right) \times \beta\right) + \beta & \text{if } F_i = 0 \\ \left(\left(1 + \dfrac{F_b}{\text{fitness}}\right) \times \beta\right) & \text{if } F_i > 0 \end{cases} \tag{2.2}$$

where, $\beta = n_{beads} \times m_f$, with m_f lying within the range 0.1 to 1.0 inclusive.

2.1.5.2 Hypermacromutation

Hypermacromutation is supported by the idea of inorder mutation introduced in section 1.6.1. As previously mentioned, two random genes are selected from the chromosome, with point mutations performed at every gene within the range. Whereas hypermutation provides only a single entry of new genetic material, hypermacromutation provides the individual with a host of new genetic material, allowing exploration of completely unrelated areas of the PES. Again mutations are static, in that alleles are changed completely at random from an available set of values. Although no calculation is involved in determining the number of mutation attempts, as with hypermutation, the mutation does however, stop at the first constructive attempt.

For a mutation to be sustained within the range of genes, it must give rise to a valid structure at each intermediate stage. If this is not the case, then the allele reverts back to its original value. Before moving on to the next gene within the range, all possible mutations are explored. Again the final chromosome, after all mutations have been performed within the range, must result in a valid conformation. Figure 2.2 illustrates the process involved in performing a hypermacromutation, with the black arrows representing the randomly selected range and the red alleles indicating a point mutation has been made. It should be noted that, the sequential point mutations within the range can occur in either a forward or backward direction, which is also randomly determined. The black allele within the range indicates an attempted change that resulted in an invalid conformation up to that point. As a result the allele is left unchanged.

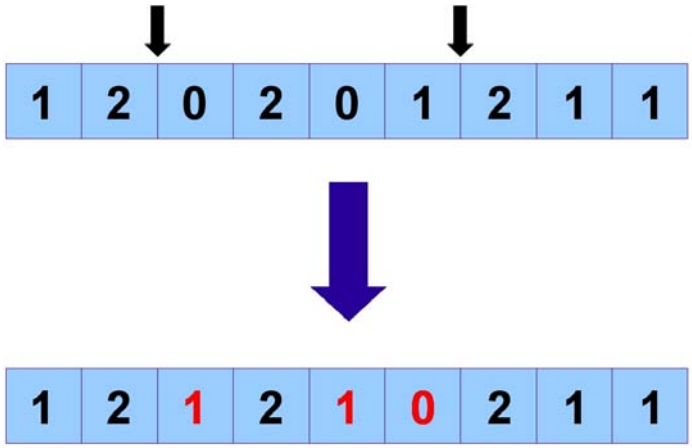


Figure 2.2: Schematic of hypermacromutation used by the IA. The small black arrows represent the randomly chosen range, with the alleles in red indicating a point mutation has been made. If no point mutation has been made within the range, the allele remains black and represents an attempted change resulting in an invalid conformation. point mutation occur sequentially in a randomly chosen forward or backward direction.

2.1.6 Selection

The selection method employed shares characteristics with rank-based selection met in section 1.6.2. In rank based selection, population members are ordered by fitness, such that the first member of the population will be the most fit. The difference between this selection method and traditional rank based selection is that no probabilities are considered. A pure elitist strategy is adopted, ensuring that the best from each generation is selected for the next one. It is due to the ageing operator met in section 2.1.3, that a pure elitist strategy works in conjunction with artificial immunity. With classic GA methods, choosing the "cream of the crop" may lead to premature convergence as the fittest individuals will dominate future populations. As individuals can "die out" after failing to be improved by mutation, lower fitness individuals can still be considered for the new population.

Before selection actually takes place, the current population is merged with the mutated population, identical individuals are removed and again the population is sorted in order of descending fitness. This may provide a single population which is

larger than n_{ind} in capacity. In order to be fair, all individuals that share the same fitness values are treated equally, as no one is better than the other. For this to be reflected in selection, the fitness of the individual occupying the n_{ind} position in the population is recorded, with individuals having this fitness being identified. Any individuals having a fitness greater than this are automatically selected and placed in the population for the next generation. The remaining available positions in this new population are filled by randomly selecting individuals from the identified (via fitness) individuals. The end result is a new population of size n_{ind}.

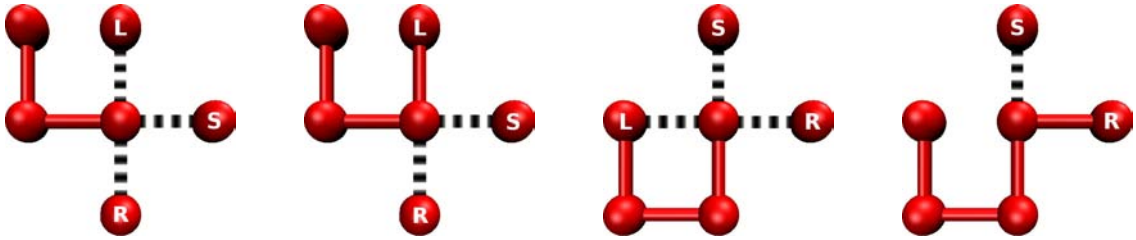
2.2 Immune Algorithm Extensions

With the intention of improving the methodology and search efficiency, simple additions to existing operators, and new operators have been investigated in this work.

2.2.1 Mutation Memory

When performing mutations on chromosomes, a problem with lattice models is that atom placements may result in structure invalidity. This occurs commonly when a mutation attempts to place beads on lattice sites that are already occupied. It is this idea that has inspired the introduction of mutation memory. It is initially invoked in the constructor, employing an RGA that backtracks [96] through previous loci and repairs any structural infeasibility. Before a bead is to be placed, all values in the decision matrix are explored in order to identify which values can be adopted by the gene. From this short list, a random decision is made, producing a self avoiding conformation up to that chain position. These available decisions are marked accordingly, producing a possible mutation list that does not include any currently used allele. The decisions from the list are made available to the mutation operators with an allele change resulting in a feasible conformation up to that locus. Figure 2.3 illustrates how the problem arises and is overcome. It can be seen from figure 2.3(a) that all elements

of the decision matrix are available and so an unhindered decision can be made as to where to place the next bead (figure 2.3(b)). If we consider figure 2.3(c), it should be noted that when the decision matrix is explored, one choice is unavailable. Placing test beads identifies this and removes the conflicting element from selection in the decision matrix. Once a decision has been made from among the remaining elements, it too is marked to prevent further selection in the decision matrix.



(a) (b) (c) (d)

Figure 2.3: A schematic showing the various stages in mutation memory in the 2D HPLBM. (a) All options are available from the decision matrix. (b) A random choice is made from the available decisions (left in this case). (c) test beads highlight already occupied lattice spaces (left in this case). (d) Problematic left decision is no longer available and a random choice is made from the remaining options (right in this case).

The information gained as a result of the RGA is retained and then utilised by the mutation operators. As described in section 2.1.5, a mutation is simply an allele change from a list of available options, which in turn will affect the conformation according to the new chromosome. In the first instance, the options available to the mutation operators have been derived from the constructor, where, in future generations, the information stored would be from the previous generation's mutation phase. A mutation occurs by consulting the decision matrix for available decisions, with one being selected and the chromosome updated. As the self avoiding conformation is constructed, again the full range of elements in the decision matrix for each locus are explored, thus updating an individual's mutation memory for the next generation [97].

2.2.2 Crossover

As detailed in section 1.6.3, mating is an opportunity for genetic information to be exchanged between individuals of a population. Favourable genetic information, once transferred, may improve the fitness of individuals, whilst an unfavourable exchange may result in a fitness reduction. Section 1.6.3 features methods of mating, explicitly crossover, with one-point crossover being used in this work.

The conformation vector, as detailed in section 2.1.1, defines how crossover operates between pairs of members of the population. A cut is made at a randomly generated point along the chromosome, represented by a random element selection of the conformation vector in our model chromosome. Once the selection has been made, all elements up to that point are exchanged between the two individuals. The new combination of conformation vector elements gives rise to two new offspring.

Crossover is employed by the IA and acts only on the cloned population, as the mutation operators did. Any valid offspring produced as a result of crossover are stored in an independent population, as for mutated individuals. As the operator acts on a cloned population, each individual present in the population is only given one chance at participating. It is also futile for identical individuals to mate with each other. It is for this reason that an additional constraint has been added to prevent identical clones from undergoing crossover. The procedure involved in selected individuals for crossover is outlined in algorithm 2.3.

Algorithm 2.3 Pseudocode for the crossover selection procedure ensuring no like clones are used by the operator.

```
1: while individuals require crossover do
2:    repeat
3:       select random individual (R^a) to use as the first parent.
4:    until individual is yet to be used for crossover
5:    repeat
6:       select random individual (R^b) to use as the second parent.
7:    until R^a_pos mod n_clones ≠ R^b_pos mod n_clones and R^b is unused.
8:    perform crossover using selected parents
9: end while
```

As previously discussed in section 1.6.3, the crossover mechanism used here also uses two parents with both offspring being considered for validity.

2.2.3 Local Search

The IA is equipped with two methods of local search [98], a *point mutation neighbourhood* search and a *macromutation neighbourhood* search [21, 78, 99]. Both can be used as alternatives to the standard IA operators hypermutation and hypermacromutation and act on the cloned population of individuals. The fundamental ideas of these methods are similar, in that the point mutation neighbourhood search performs point mutations and macromutation neighbourhood search operates over a range of genes. With the local search methods utilising mutation, they are an alternative method used to introduce new genetic material into a population as described in section 1.6.1. For a point mutation to be successful, it must result in an improvement in fitness of the individual. If this is the case, then the mutation becomes permanent, giving rise to a new intermediate conformation.

Instead of stopping the search process at the first constructive mutation, (the strategy used by the hypermutation operator), after visiting random sequence positions, each point mutation requires an energy and fitness calculation. If the resultant fitness is an improvement on the original, the search continues from that point until all genes in the sequence have been explored. In contrast, if the resultant fitness is lower or remains unchanged, then the search will continue from the last mutation that rendered an improvement in fitness, again until all genes have been investigated. This idea is mimicked by the macromutation neighbourhood search. As for the standard macromutation operator developed in the original IA, a range of genes is randomly selected from the chromosome with the same approach towards mutation as for the point mutation neighbourhood search. Again, only an improvement in fitness for that specific point mutation will yield a permanent change to the chromosome. It should be noted that, whereas in point mutation neighbourhood search gene mutation occurs in a random

order throughout the chromosome, macromutation neighbourhood search visits each gene within the selected range in order, whether that be in a reverse or conventional fashion.

Algorithm 2.4 Pseudocode for the local search operator, where i and j are either the first and last locus (point mutation neighbourhoo) or the beginning and end of a range of loci (macromutation neighbourhood).

```
1: while i ≤ j do
2:   repeat
3:     select random available allele.
4:     apply change to conformation vector temporarily
5:     calculate fitness of individual
6:     if fitness is improved then
7:       commit allele permanently
8:     else
9:       revert back to original allele
10:    end if
11:  until no available alleles remain for locus i or improvement in fitness is observed
12:  increment i
13: end while
```

2.2.4 Mixed Strategy

The mixed strategy (MS), like other alternative genetic operators described here, replaces the traditional hyper and hypermacromutation operators. The MS provides an array of mutation methods within a single mutation operator, with each separate mutation operation being selected probabilistically. The probability of selecting each mutation operator is initially uniform ($P_{op} = 0.25$). The probability for each operator is recalculated every generation, based on the number of successful mutations performed. If $P_{op} > 0.6$ for any operator, the probabilities are reset, giving each operator equal probability of being selected. The genetic operations available to the MS are the hyper-, hypermacromutation (traditional), point- and macromutation-neighbourhood search (local search) operators.

A cloned population is traditionally subjected to two mutation phases per generation. The MS adheres to this rule, being followed by a crossover phase once the

MS mutation phase has terminated. As a result, once a population is cloned, the individuals of the cloned population undergo MS followed by crossover phase.

Algorithm 2.5 Pseudocode for the mixed strategy operator.

```
for Each clone do
    Randomly select operator based on probability (P_op).
    Perform selected mutation on individual.
    if Mutation is successful then
        Increment selected operator counter.
    end if
end for
Compare operator success.
if Operator is most successful then
    P_op = P_op + (P_op × 0.01)
else
    P_op = P_op - (P_op × 0.01)
end if
```

2.3 Basic Differential Evolution

DEs [75] are population based search techniques that have proven successful when applied to continuous search spaces [100]. They require a population of individuals to allow for efficient, simultaneous searching of energy landscapes. An individual is represented by a single conformation vector that allows access to the structure, energy and fitness. The fundamental approach to DE involves the selection of a parent individual (i_p) followed by the selection of three random individuals (i_{r1}, i_{r2} and i_{r3}) from the population.

The DE must be initialised with a specific n_{ind}, which must remain constant throughout its lifetime, defined by g_{max}. Each generation concerns sequentially using each member of the current population as a parent individual. This idea prevents any currently known genetic material from being overlooked in the combined mutation and recombination step explained in detail in section 2.3.1. Once the parent individual is identified, the three random individuals are used to provide unknown new genetic material (with regards to the selected parent). To make the experimental genetic ma-

terial diverse, each random individual cannot be the same as another, nor can any be the same as the selected parent. Once the three random individuals have been chosen, all four selected individuals must then engage in mutation and recombination.

It must be reiterated that n_{ind} cannot fluctuate. To combat this, once mutation and recombination have produced a trial solution, the trial must be compared to the parent individual, with the fittest remaining in the population as described in more detail in section 2.3.2. Pseudocode for the DE presented here can be seen in algorithm 2.6.

Algorithm 2.6 Pseudocode for DE illustrating the procedure involved for selecting parents and random individuals.

```
Assign n_ind, g_max, recombination rate (K) and mutation rate (F).
Generate a population of n_ind individuals.
g = 0.
while Lowest known energy conformation has not been found or g ≤ g_max do
   for i = 1 to n_ind do
      Assign i_p = i.
      Randomly select i_r1, i_r2 and i_r3 such that none are equal to each other nor to
      i_p.
      Generate i_O from i_p, i_r1, i_r2 and i_r3
      if i_O fitness > i_p fitness then
         Assign i = i_O
         if i_O fitness > i_b fitness then
            Assign i_b = i_O.
         end if
      end if
   end for
   increment g.
end while
Output lowest energy conformation.
```

For each parent that generates an improved trial solution, the trial solution is compared to the current best individual (i_b), which it replaces if its fitness is greater. If the current best individual is of higher fitness, then it is left unchanged.

2.3.1 Mutation and Recombination

Unlike GAs and the extended IA presented here, DE offers a combined mutation and mating step known as mutation and recombination. This method requires two parameters, F and K, to scale the effects of the process on the individuals. As the individuals in the DE are encoded and represented by the same conformation vector as for the IA, manipulating a chromosome is achieved in an identical fashion. The values of the genes can be exchanged for other values that the model allows access to.

Mutation and Recombination occur in a single step defined by a simple scheme. Equation (2.3) contains both K and F factors used to scale the process. The function itself requires a parent individual (i_p) to be selected from the population of individuals. The parent provides the base genetics to be manipulated by the rest of the terms in the function. If we initially consider the equation as a whole, three random population members must be chosen from the population, i_{r1}, i_{r2} and i_{r3}. It is apparent by considering the terms individually, that the mutation term (containing F) is not affected by any of the genetic information held by the parent. In contrast, the recombination term (containing K) is governed by parental genetics. The resultant genetics of each of these terms must then be added to the originally selected parent to produce the trial solution (or offspring) i_O:

$$i_O = i_p + K\left(i_{r1} - i_p\right) + F\left(i_{r2} - i_{r3}\right) \tag{2.3}$$

With DE proving highly successful in continuous space, the way in which the terms are calculated must be specifically defined to take care of the discrete problems used in this work. For continuous problems, dynamic mutations could exist for each gene, in which a small change to the bond angles could be executed. However, as our bond angles are discrete (by the use of integer alleles), and in the case of the DLM so are the torsion angles, a discrete approach must be taken.

The conformation vector defines the 3D conformation for an individual. By adding and subtracting these vectors, we should be able to generate an intermediate vector. In continuous problems, the alleles could literally be added or subtracted and scaled using the appropriate factor. If we have a finite decision space (a discrete problem) however, this is not possible. In this work, a clockwise approach has been adopted. Simply adding or subtracting the alleles may render the new allele outside our discrete finite range after scaling. If this is the case, the value must re-enter the clockface to produce a valid allele within the range, as shown in figure 2.4.

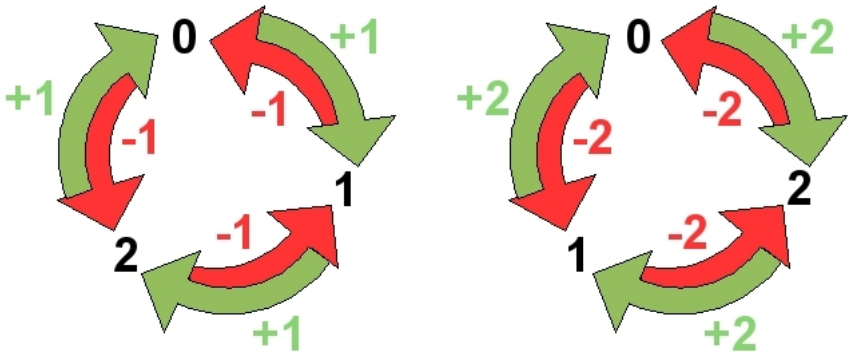


Figure 2.4: Schematic representation of the clockface idea used for the addition and subtraction of alleles of a chromosome adopted by the combined mutation and recombination operator of the DE.

With this clockface notion in mind, a simple addition or subtraction of any two conformation vectors is shown in figure 2.5. It should be noted that, figure 2.5 illustrates addition and subtraction occurring between an identical pair of conformation vectors. The resultant conformation vectors however, are very different indeed.

2.3.2 Selection

The selection procedure in DE is much simpler than for the IA. In the IA, a separate selection procedure involves merging the current population with the mutated one, and in the case of other evolutionary algorithms, probabilistic effects determine the fitness of individuals to be selected. In DE once a mutation phase has been completed, a

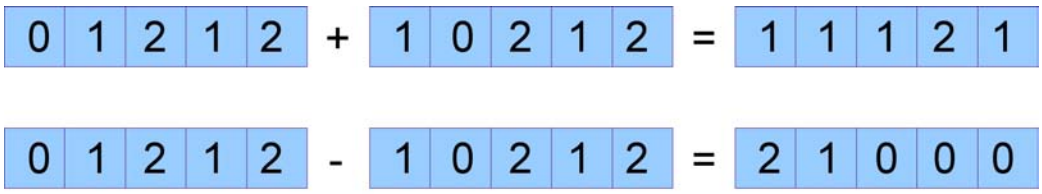


Figure 2.5: A sample addition and subtraction as performed by the DE during the mutation and recombination phase. Note how performing opposing operations gives rise to a completely different conformation vector.

simple comparison of fitness between the parent and the newly generated offspring is performed. If the parent has higher fitness, then the offspring is discarded. However, if the offspring is fitter, then the parent is replaced by this newly created individual. This method ensures that lower quality offspring cannot enter the population. However, this method also provides an opportunity for high energy individuals to enter the population by replacing parents of even higher energy.

2.3.3 Differential Evolution Extentions

As described for the IA in section 2.2, generating an initial population requires each individual to be a self avoiding conformation. Utilisation of the RGA is essential to prevent the loss of the segments of local structure due to a bead misplacement. As described in section 2.1.1, generating an initial population requires testing bead placements in the lattice environment to discover available adjacent lattice sites, in order to make the least number of point changes to the conformation vector to render the structure a self avoiding walk.

Once an individual has left a mutation phase, it may also be beneficial to prevent the loss of what might be favourable genetic information that may ordinarily be lost as a result of an invalid conformation. This idea has inspired the implementation of the RGA post mutation. As a result of calling the RGA after a mutation phase on invalid individuals, all individuals after leaving the mutation phase will enter the fitness comparison stage between parent and trial solution. This will prevent the possibility of

useful genetic information being lost that would ordinarily be discarded, therefore, not having the chance to propagate through a population, due to regions of invalid local structure.

2.4 Model Encoding

A previous study has illustrated how a local coordinate system offers better performance than a global one for studying protein folding [101]. In this work, a local coordinate system is used to define the folding conformation of the model proteins.

2.4.1 Static Lattice Bead Models

As outlined in sections 1.1.5.1 and 1.1.5.2, the HPLBM and BLNM are simplistic protein representations. They define how residues are to be depicted and how conformation energies are to be calculated. The models themselves may be accompanied by a lattice framework which helps to define structural characteristics, such as bond lengths and angles.

2.4.1.1 The Square Lattice and its Coordinate System

The HPLBM used in this work is an example of a static lattice bead model, using the simplest potential. As outlined in section 1.1.5.1, its simplified protein representation includes only two bead types on either a square or diamond lattice. Each sequence residue is represented by a single uniformly sized bead. This depiction allows for an easier investigation of the protein system by reducing its complexity, provided potentials are used that reflect characteristics of a real system.

Table 2.1 lists the possible alleles representing a square lattice that can be adopted by each locus to produce a conformation vector. In order to provide the genes of the chromosome with alleles, a decision matrix is used. It is apparent from table 2.1 that each locus in the chain has the same number of decisions available resulting in a decision matrix of size $n \times n$. In order to simplify the interval between bond angles, an

Allele	Bond Angle (θ)
0	-90.00°
1	+ 90.00°
2	0.00°

Table 2.1: Possible alleles and corresponding bond angles for the HPLBM on the square lattice

integer representation is used. An array of integers makes up the conformation vector which represents the two dimensional (2D) spatial arrangement of the beads. Together with a list of beads, i.e. the primary sequence, the conformation vector is what defines a model 2D protein structure.

It should be noted that, in order to increase efficiency and drastically reduce search space without omitting valuable, unique minima (not rotationally, translationally or reflectively related to another), the first two beads of a sequence are always fixed at (0.0, 0.0) and (1.0, 0.0) in 2D. As this is the case, the conformation vector itself is actually $(n_{beads} - 2)$ in length, with each bond being a single arbitrary quantity in size. Bond angles are restricted to either 0° or 90°, with torsion angles being neglected completely. This rigidity restricts the bonds themselves to lie along either the $\pm x$ or $\pm y$ axes.

In this work, a local coordinate system is used to define the folding conformation of the model proteins, that is the position of bead n is defined relative to beads $(n-1)$ and $(n-2)$. The bond joining the $(n-1)th$ and nth beads can be in a left, right or straight ahead direction relative to the bond joining the $(n-2)th$ and $(n-1)th$ bead, corresponding to an integer representation of 0, 1 and 2 respectively as shown in table 2.1. The protein conformation is therefore expressed as a conformation vector, containing a list of 0s, 1s and 2s. In order to place bead n, $\vec{orig}$ (the vector between the $(n-1)th$ and the $(n-2)th$ beads) is calculated to translate it to the axis origin. This is done by subtracting $(\vec{n-1})$ from $(\vec{n-2})$. It is then rotated by θ around the

z-axis according to $(n-2)th$ position of the conformation vector to produce $\vec{rot}$. $\vec{orig}$ and $\vec{rot}$ are then added together to provide the relative change in direction , which when added to $(n-\vec{2})$ reverses the translation. This results in a new bead position and a bond angle representative of the allele in the chromosome.

2.4.1.2 The Diamond Lattice and its Coordinate System

A diamond lattice has been chosen for simple 3D conformations, as it leads to similar structural motifs to real proteins.

Allele	Bond Angle (θ)	Torsion Angle (ϕ)
0	109.47°	0.00°
1	109.47°	−120.00°
2	109.47°	+120.00°

Table 2.2: Possible alleles and corresponding bond and torsion angles for the HPLBM and BLNM on the diamond lattice

As seen for the square lattice, integer representations are used to simplify the characteristics of model diamond lattice proteins. Table 2.2 describes the bond and torsion angles represented by each allele for static bead models on the diamond lattice used in this work. A decision matrix of size $n \times n$ is once again observed from this table as each gene can adopt a value of either 0, 1 or 2, resulting in a conformation vector.

For the same reasons that the position of the first two beads were fixed in the 2D case, the first three beads of the diamond lattice are also fixed, due to the introduction of a torsion angle, thus creating the positions (0.0, 0.0, 0.0), (1.0, 0.0, 0.0) and (cosθ, sin θ, 0.0) respectively. For the same reason, the length of the conformation vector in this case is $(n_{beads} - 3)$. The torsion angles are the only variable in defining the diamond lattice, as opposed to them being ignored in the 2D case, thus all bond angles are 109.47°. As our structure is mounted on a regular diamond lattice, all bond lengths are equal, as seen with the 2D case. To be consistent with previous work [49, 102, 103],

bond distances are constrained to a length of 3.4 arbitrary units (a.u.) by scaling once the geometry has been calculated.

Diamond lattices are created by repeating alternate tetrahedral bead arrangements composing face centred cubic Bravais sub-lattices A and B. Lattice points in A are tetrahedrally coordinated to 4 nearest neighbours belonging to lattice B and *vice versa*. As a result, topological contacts can only be produced between beads lying on different lattices, separated by 4+2n positions along the sequence.

Once the nth bead is placed such that the nth-$(n-1)th$ bond eclipses that of the $(n-2th)$-$(n-3)th$, a simple rotation about the $(n-2)th$-$(n-1)th$ bond by ϕ according to the allele at position $n-3$ of the coordination vector, produces a torsion angle reflective of the ones found for the diamond lattice. To produce this torsion, the requirement of the previous three beads is such that the $(n-2th)$-$(n-3)th$, $(n-2)th$-$(n-1)th$ bonds are 3.4 a.u. in length and the $(n-3)th$-$(n-2)th$-$(n-1)th$ bond angle is 109.47°. The coordinates must all be translated to the origin by subtraction of $(n-2)th$ coordinate from each coordinate set. By taking the triple cross product of the three translated vectors, a vector ($\vec{p}$) perpendicular to the $(n-2)th$-$(n-1)th$ bond in the $(n-3)th$-$(n-2)th$-$(n-1)th$ plane is produced. By addition of the sin(ϕ-90) x component and cos(ϕ-90) y component to $\vec{p}$, a mirrored bond angle of 109.47° is produced. By adding this new $\vec{p}$ to the $(n-2)th \vec{-} (n-1)th$ translated bond vector, the new bead position is such that a reverse translation can be performed prior to torsional rotation about the $(n-2)th \vec{-} (n-1)th$ translated bond vector.

The BLNM potential described in section 1.1.5.2 considers four different energy terms. For this work the BLNM is coupled with a diamond lattice and as a result, some of the terms used to described the off-lattice model are simplified. From equation (1.3), the BLNM potential initially considers energetic contributions from bonding beads. As r_i and r_0^i have the same magnitude, the bonding contribution is rendered redundant. When considering the bond angle energy term, θ_i and θ_0^i also have the same magnitude

due to the bond angle restriction imposed by the lattice. This results in the bond angle term not being a contributing factor to a model BLN protein's conformation energy.

As previously explained from table 2.2, the lattice also restricts torsion angles. For this reason, torsional contributions can be considered in two ways: by use of the torsional energy term; or by integrating bead torsional relationships into the nonbonding pair energy term. This work considers the torsional contribution as a separate entity as stated in equation (1.3), due to this term considering bead types to a greater extent than the nonbonding pairs energy term. The reduced (as a result of the lattice) nonbonding pair lattice based term can be seen in equation (2.4):

$$\begin{aligned} V_{N-[B,L,N]} &= 4\epsilon E_h \left(\frac{a}{R_{ij}}\right)^{12} \\ V_{L-[B,L]} &= 4\epsilon E_l \left[\left(\frac{a}{R_{ij}}\right)^{12} + \left(\frac{a}{R_{ij}}\right)^{6}\right] \\ V_{B-B} &= 4\epsilon E_h \left[\left(\frac{a}{R_{ij}}\right)^{12} - \left(\frac{a}{R_{ij}}\right)^{6}\right] \end{aligned} \tag{2.4}$$

where, ϵ is an interaction parameter defining the energy scale and is set to 0.0100570, a is set such that V_{B-B} is a minimum at the first nearest neighbour and has a value of 3.029, R_{ij} is the distance between two beads due to a pairwise interaction, and E_h is equal to 1, with E_l defined as $\frac{2E_h}{3}$.

2.4.2 Dynamic Lattice Bead Model

As previously described in section 1.1.1, the protein backbone consists of repeating sequences of N, C_α and C atoms. The DLM developed by Kobe et al. [104] does not neglect this fact by treating residues as a single bead. In fact, it exploits the backbone configuration, treating only side chain atoms collectively. With the chemical bonding properties of the protein backbone being the main focus of this model, the

bond lengths between atoms of sequential amino acids and the bond angles between the atom connections are fixed and are summarised in table 2.3.

	Variable	Value
Bond Length	N-C_α	1.47 Å
	C_α-C	1.53 Å
	C-N	1.32 Å
Bond Angle	N-C_α-C	110°
	C_α-C-N'	114°
	C-N'-C_α'	123°

Table 2.3: Bonds and bond length data taken from [104] for backbone atoms of the DLM.

As previously described in section 1.1.1, the protein backbone is defined by three torsion angles. Applying biological constraints, ω is restricted to 180°, with the remaining torsion angles, ϕ and ψ having varying values that differ from one amino acid to the next. A reduced set of torsion angles from the many accessible by individual amino acids, was obtained by Kobe et al. [104], analysing 403 protein structures from the protein data bank (PDB) [105]. Performing cluster analysis on the resulting Ramachandran plots allowed for a number of highly utilised ϕ,ψ angle pairs for each amino acid to be determined.

For the purpose of calculating the conformation energy, a C_β bead is introduced at a distance of 1.53 Å from the C_α. In the case of glycine, a virtual C_β is adopted. The side chain centre is modelled as a single bead with their distances from the C_α corresponding to the side chain radius ($d_{sc}/2$). The side chains themselves are modelled as hard spheres with volumes analogous to the VDW volumes of real side chains listed in table 2.4.

Applying the bond length and bond angle parameters mentioned, results in the C_β atoms being placed in the corner of a distorted tetrahedron comprising residue backbone atoms with the C_α at the centre, as depicted in figure 2.6. The significance

Amino Acid	ϕ,ψ Angle Pairs	d_{sc} (Å)
A	(-135,150), (-65,-35), (-75,140)	4.0
C	(60,35), (-110,140), (-75,30), (-120,45)	4.4
D	(55,35), (-95,140), (-110,60), (-70,-30)	4.4
E	(-105,135), (-70,-35)	4.7
F	(-115,140), (-70,-35), (-115,40),	5.1
G	(95,-165), (85,10), (-125,170), (-70,-30)	3.6
H	(60,35), (-115,135), (-80,-25),	4.9
I	(-125,130), (-65,-45), (-95,-5), (-95,130)	4.9
K	(-105,140), (-70,-30)	5.1
L	(-100,135), (-70,-35)	4.9
M	(-110,135), (-70,-35)	4.9
N	(55,40), (-105,130), (-80,-20)	4.5
P	(-60,-25), (-75,165), (-65,140)	4.4
Q	(60,40), (-70,-30), (-110,140)	4.8
R	(-70,-35), (-110,135)	5.2
S	(-135,150), (-70,-25), (-80,150)	4.1
T	(-130,160), (-100,135), (-95,-5), (-65,-40)	4.5
V	(-130,145), (-100,130), (-115,0), (-65,-40)	4.7
W	(-110,140), (-60,-45), (-80,-25)	5.4
Y	(-115,135), (-75,-30)	5.2

Table 2.4: Angle pairs (taken from [104]) defining DLM backbone characterisation and side chain diameters for each of the 20 natural amino acids.

of adopting this strategy in accordance with the -COOH (carboxylic acid group), the -R group (indicating the side chain) and -NH_2 (the amine group) CORN (-**CO**OH, -**R** group and -**N**H_2) law [9] configuration is that the correct L-amino acid chirality is achieved, mimicking the geometries of real amino acids as highlighted in section 1.1.1. It is the ϕ,ψ angle pairs that generate the conformations of model dynamical lattice proteins.

2.4.2.1 Constraints

Due to side chains being modelled as hard spheres, side chain-side chain overlap is not permitted and will lead to structure invalidity. In order to account for the volume of the protein backbone, a distance constraint is applied between C_α atoms in that they cannot be closer than 3 Å. This allows for more realistic folding, preventing the protein

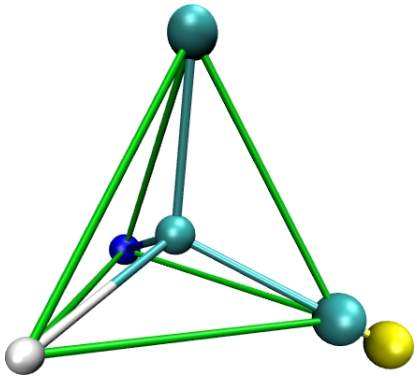


Figure 2.6: A simple CORN tetrahedron with carbon atoms represented in light blue, the backbone nitrogen atom represented in dark blue, the hydrogen in white and the side chain centre in yellow. The residue dependent C_α-SC distance is equal to $d_{sc}/2$ from table 2.4.

backbone from being too strained.

2.4.2.2 Potential

As previously stated, the C_β beads are solely responsible for the conformation energy of model dynamical lattice proteins. A simple sum over all pairwise interactions involving these beads gives rise to the total conformation energy (E_T) as shown in equation (2.5).

$$E_T = \min \sum_{ij}^{n} E_{ij} \tag{2.5}$$

E_{ij}, defined in equation (2.6) represents the interaction between C_β^i and C_β^j, with r_{ij} being the distance between them. The interaction is a smooth approximation to a stepwise function with a distance cutoff of 8.0 Å, i.e. $E_{ij} = 0$ for $r_{ij} > 8.0$ Å.

$$E_{ij} = e_{\mu\nu} \left\{ \frac{\tanh\left[\frac{8.0 - r_{ij}}{2}\right]}{2} + 0.5 \right\} \tag{2.6}$$

$e_{\mu\nu}$, described in equation (2.7), identifies which two residues are interacting and the extent of the interaction by use of a series of three interaction constants. $\epsilon(i,j)$ is the interaction between the i^{th} and j^{th} residue. $e_{\mu\nu}$ also includes the solvent effects

on specific residues: $\epsilon(S,i)$ represents the solvent interaction with the i^{th} residue and likewise $\epsilon(S,j)$ is the solvent interaction with the j^{th} residue. The interaction constants used here were parameterised by Settanni et al. [106] and are presented in table2.6.

$$e_{\mu\nu} = \epsilon(i,j) + \epsilon(S,i) + \epsilon(S,j) \tag{2.7}$$

2.4.2.3 The Co-ordinate System

As with the simple static lattice models a decision matrix is used to provide the genes of the chromosome with alleles. However, in the case of the DLM the matrix is not of the standard $n \times n$ size, as each allele can adopt a specific number of values depending on which amino acid is present at the locus of the chromosome. If we take the first two amino acids, alanine and cysteine, the number of decisions available to each is different, with alanine having three, as with the previously described lattice models, and cysteine having four. These values are quoted in table 2.4. For the purpose of ϕ, ψ selection, each angle pair is given a numerical value, allowing genes to represent a pair of angles with a single integer value. Table 2.5 clarifies how angle pairs are defined for the cysteine residue.

Allele	ϕ,ψ Angle Pairs
0	(60,35)
1	(-110,140)
2	(-75,-35)
3	(-120,45)

Table 2.5: Angle pairs (taken from [104]) for the cysteine residue, with corresponding alleles used in the DLM.

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	-0.0269																			
C	-0.0142	-0.1509																		
D	0.0222	0.0027	-0.0130																	
E	0.0308	0.0047	0.1587	0.0712																
F	0.1415	-0.0500	0.0167	0.0458	-0.1098															
G	0.0386	0.0635	-0.0015	0.0205	-0.0774	0.0023														
H	0.0392	-0.0293	0.0255	-0.0472	0.0240	-0.0322	0.0046													
I	-0.0261	-0.0960	0.1188	0.1026	-0.0423	0.0310	0.0175	-0.1753												
K	0.0760	0.0372	-0.0675	-0.2113	0.0302	-0.0050	0.0137	0.1177	0.0498											
L	-0.0540	0.0776	0.0078	0.0740	-0.0988	-0.0787	0.0849	-0.0672	-0.0431	-0.1246										
M	0.0070	0.0239	-0.0995	0.0116	-0.0614	-0.0241	-0.0603	-0.0472	0.0807	0.0179	0.0393									
N	0.0036	-0.0158	0.0208	-0.1343	0.1070	0.0377	-0.0005	0.1172	-0.0636	0.1172	-0.0070	-0.0064								
P	-0.0760	0.0134	0.0180	0.0272	0.1145	-0.0043	0.0615	0.0253	-0.0798	0.0125	-0.0205	0.0713	0.0381							
Q	-0.1203	0.0294	0.0724	0.0274	0.0853	0.0572	-0.0068	0.0530	-0.0691	-0.0792	0.0176	-0.0199	-0.0460	0.0141						
R	0.0154	0.0202	-0.1486	-0.1610	-0.1207	-0.0010	-0.0315	0.0528	0.0892	-0.1136	0.0646	0.0225	-0.0228	0.0233	0.0624					
S	0.0358	0.0047	-0.1300	-0.0895	0.058	-0.0257	-0.0469	0.0288	0.0214	0.2016	-0.0236	-0.1070	0.0085	0.0671	0.0733	-0.0618				
T	-0.0831	-0.0511	-0.0667	0.1258	-0.0492	0.0526	-0.0890	-0.0541	0.0180	0.1139	0.0194	-0.1003	0.0623	0.0087	0.0340	-0.0576	0.0138			
V	-0.0658	-0.0066	0.0702	-0.0609	-0.0185	-0.0406	0.0911	-0.0908	-0.0146	-0.0724	0.0465	0.0365	-0.0351	-0.0155	0.0528	0.0927	0.0414	-0.0637		
W	0.0878	0.0129	0.0575	0.0093	0.0210	0.0055	0.0123	-0.0631	0.0281	0.0058	0.0380	-0.0487	-0.0345	-0.0733	0.0007	-0.0076	0.0234	-0.0504	0.0118	
Y	-0.0367	0.0389	0.0111	0.0521	-0.1059	0.0249	-0.0507	-0.0505	0.0235	-0.0265	-0.0785	-0.0117	-0.0536	-0.0089	0.1021	-0.0219	-0.0317	0.0482	-0.0826	0.1666
Sol	-0.0053	-0.0850	0.0737	0.0575	-0.0901	0.0387	-0.0201	-0.0478	0.0317	-0.0448	0.0164	0.0186	0.0801	0.0121	0.0141	0.0202	0.0006	-0.0556	-0.0462	-0.0923

Table 2.6: Interaction constants representing the interaction strength between each amino acid and solvent [106].

With regard to mutation, a ϕ, ψ angle pair can be modified by a single integer change in the chromosome. Again, during the structural build phase, using the idea of mutation memory described in section 2.2.1, only alleles that give rise to a valid structure up to that point in the chain are available to the mutation operator. Once a valid structure has been created, the mutation memory is updated, making it only applicable to the new conformation.

2.5 Structural Similarity Measures and Population Diversity

A frequent problem in the application of evolutionary algorithms to optimisation problems is premature convergence (population stagnation) [85]. This arises from the population in the computational process settling on a suboptimal state, rendering the operators incapable of producing improved individuals in successive generations [107]. One of the fundamental issues in evolutionary computing is therefore maintaining the population diversity [108].

In order to quantify how diverse a population is, a measure must be employed to assess how similar individuals are within a population. The first method used in the proposed IA for assessing structural similarity is the Hamming Distance (D_H). For D_H to apply, our chromosomes must be of the same length. Additionally, in this work our protein sequences are identical. For two chromosomes, C_a and C_b, $D_H(C_a, C_b)$ is the number of loci in which the alleles differ as shown in equation (2.8) [109]:

$$D_H(C_a, C_b) = \sum_{i=1}^{N} 1 - \delta_{ab}^i \tag{2.8}$$

where

$$\delta_{ab}^i = \begin{cases} 1 & \text{if } C_a^i = C_b^i \\ 0 & \text{otherwise} \end{cases}$$

In the case of the HPLBM described in section 2.4.1, where our protein conformations are defined as a series of directional choices at lattice positions, our chromosomes

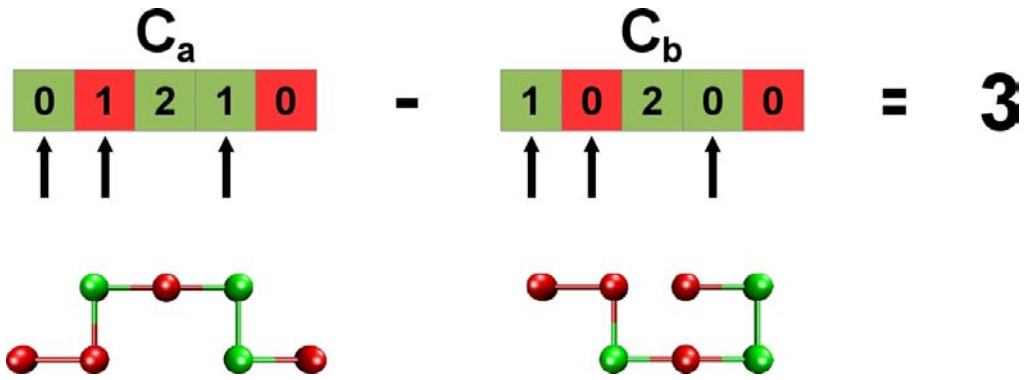


Figure 2.7: Schematic representation of Hamming Distance for the case of the HPLBM. It illustrates how the chromosome relates to its structure and how the structures differ as a result, highlighting inconsistent alleles (black arrows). It should be noted that the chromosome is shorter than the structure itself as the first two beads have fixed positions in this case. The colours present in the conformation vector correspond to the bead type at that locus.

contain genes of alleles corresponding to the directions chosen. D_H is therefore specifically defined as shown in figure 2.7 with inconsistent alleles highlighted identifying the differences between the chromosomes.

D_H is very effective at determining chromosome similarity in this way and is also computationally inexpensive. However, to determine three dimensional conformational similarity it lacks the ability to accurately quantify small structural changes in local coordinate systems. If $D_H = 1$ with the difference lying at either terminus of the chromosome, then the low D_H will reflect a small structural change. However, if the difference is at a more towards the centre of the chromosome, then this low D_H can actually reflect a large structural diversity between two individuals. In either case D_H is incapable of detecting whether structures are merely rotationally or reflectively related.

In contrast, providing superposition of two three dimensional structures (S_a and S_b) is made, RMSD can be used to determine accurately how structurally different the two conformations are, even those related through rigid rotations and translations. RMSD is the most natural and most frequently used method to accomplish this and

utilises ordinary Euclidean distances in $3N$-dimensional space [110]. By comparing the distances between the atoms of two structures ($S = r_1...r_N$) at identical loci, a numerical value can be obtained of unit distance. This is summarised in equation (2.9). If two structures are identical then a value of zero is obtained. In the case of real proteins and the DLM, only the distances between C_α atoms are considered.

$$D_{RMSD}(S_a, S_b) = \min \sqrt{\frac{1}{N}\sum_{i=1}^{N}(r_i^a, r_i^b)^2} \tag{2.9}$$

2.6 Algorithm Genealogy

Obtaining raw statistics about a series of calculations provides vital information regarding the most energetically favourable determined conformation. However, these statistics fail to give an insight into how a search technique manages to arrive at its end point. By tracking a run, recording information as the calculation proceeds [111], not only provides information regarding the final generation and hence the minima found, but information can be gained as to how the calculation manages to arrive at its conclusion.

In this work, a very simple application of this idea has been employed. As the standard mutations, the mixed operator and local search operators only require a single parent to produce a single mutant, by recording the parent from which the mutant came over the generations, a full history (or ancestry) of each individual present in the population can be obtained.

This simple idea involves extending the individual data structure by another variable. A tag is initially assigned to each member of the population. Once an individual is created, a tag is assigned (X-X), containing no information regarding the current generation and/or position in the current population. Once the individual has entered either the first generation, a successive generation or if it has arisen due to the birthing phase, the tag is updated. The tag records the current generation and the current

position in the population that the individual in question occupies (gen-pos). Once mutated, the tag of an individual has the mutation operator appended to it in order to publicise which operator has given rise to the new conformation. This information remains with the individual until the start of the following generation. By waiting until the following generation to update the tag, the individual retains information about its parent throughout the mutation process. At the end of each generation, the population is printed with corresponding tags, population position and fitness. By the end of the calculation, every population has been printed and a full ancestry of each individual can be traced, producing a profile.

Table 2.7 is an example profile from a single run of the L27 sequence HPLBM on the diamond lattice. A separate program (written by the author) is used to analyse the profile in order to prevent complications during the calculation process and help keep central processing unit (CPU) time to a minimum. In this work various types of data are accessible using the profiles from successful and unsuccessful calculations. From the data shown in table 2.7, pairwise comparisons of D_H and RMSD between individuals, as well as D_H and RMSD comparisons between individuals and the GM, can be conducted. The data also allow for simple F_i profiles as well as basic statistics involving mutation operations to be developed. The entire run from start to finish can be profiled, allowing various conclusions to be drawn.

However, there is a drawback to this method. To get the most from this method, relies on the fact that the genetic operators used require a single parent to generate a single offspring. As previously mentioned, the standard mutation methods, the mixed operator and local search apply to this constraint. Therefore, any algorithm run involving genetic operations using a crossover method (resulting in a non-linear nature [26]) would result in an incomplete ancestry of its individuals.

Population Position	Tag	Conformation Vector	Fitness
Generation: 1			
0)	X-X	12020021012110210	2
1)	X-X	21200121002011012	1
2)	X-X	02122021110012201	1
3)	X-X	10012122200102011	1
4)	X-X	20112022011210120	-0
5)	X-X	01222120002102212	-0
6)	X-X	12001211102012210	-0
7)	X-X	20121012122222120	-0
8)	X-X	20200102112211200	-0
9)	X-X	21022102011111021	-0
Generation: 2			
0)	1-7-M	02210111011011000	4
1)	1-0-H	12020021012110010	3
2)	1-2-H	02122021110012001	3
3)	1-7-M	20121012001101211	3
4)	1-0-M	12002222200110210	2
5)	1-0-H	11020021012110210	2
6)	1-0-M	12021122200110210	2
7)	1-3-M	02100001000102011	2
8)	1-0-M	12020021012110000	2
9)	1-0-H	12120021012110210	2
Generation: 3			
0)	2-0-H	02110111011011000	4
1)	2-3-H	10121012001101211	4
2)	2-0	02210111011011000	4
3)	2-2-H	02120021110012001	3
4)	2-0-M	02221221011011000	3
5)	2-7-M	02100001000100111	3
6)	2-0-M	02210111011011110	3
7)	2-8-M	12020021012110012	3
8)	2-3-H	20111012001101211	3
9)	2-0-H	02210211011011000	3
Generation: 4			
0)	3-2-H	02210121011011000	5
1)	3-1-H	10121012101101211	5
2)	3-6-H	02210121011011110	4
3)	3-8-H	10111012001101211	4
4)	3-1-H	10121022001101211	4
5)	3-1	10121012001101211	4
6)	3-7-M	11220021012110012	3
7)	3-5-M	20100001000100111	3
8)	3-7	12020021012110012	3
9)	3-7-M	12020021012021012	3
Generation: 5			
0)	4-2-H	02210121011011210	5
1)	4-5-M	10121220001101211	5
2)	4-0	02210121011011000	5
3)	4-1	10121012101101211	5
4)	4-3-H	10111212001101211	4
5)	4-2-H	02210121011011120	4
6)	4-1-H	10121212101101211	4
7)	4-0-H	02210121011011100	4
8)	4-2	02210121011011110	4
9)	4-0-H	01210121011011000	4
Generation: 6			
0)	5-1-H	10121200001101211	5
1)	5-7-H	02210121011011000	5
2)	5-1-H	10122220001101211	5
3)	5-7-M	02210121011011210	5
4)	5-1-M	02210120201101211	5
5)	5-1	10121220001101211	5
6)	5-0-H	02210121011011211	4
7)	5-0-H	02210111011011210	4
8)	5-0-M	02210121011012001	4
9)	5-1-H	00121220001101211	4
Generation: 7			
0)	6-4-H	02210120001101211	7
1)	6-8-H	02210121011012201	5
2)	6-6-H	02210121011011210	5
3)	6-1-H	02210121011011002	5
4)	6-8-H	02210121011012011	5
5)	6-9-H	10121220001101211	5
6)	6-1	02210121011011000	5
7)	6-1-M	02210121011011111	4
8)	6-3-M	02210121011011122	4
9)	6-3-M	02210121011012010	4

Table 2.7: An example profile illustrating data recorded for a population size of 10 for sequence L27 of the HPLBM on the diamond lattice using the common parameter set.

2.7 Algorithm Statistics

To help develop an understanding of how the algorithms used cope with the protein models studied in this work, standard statistical methods have been employed to assess their consistency. The mean number of generations (μ_g), mean number of fitness evaluations (AFE), standard deviation of the number of generations (σ_g), standard deviation of the number of fitness evaluations (σFE), skewness associated with the number of generations (s_g) and skewness associated with the mean number of fitness evaluations (s_{FE}) are calculated for each job (consisting of 100 separate runs of the algorithms).

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{2.10}$$

The mean, as calculated in equation (2.10) allows a statistical or geometric average to be observed. The standard deviation (σ) measures population dispersion, indicating how close values tend to be. A small σ implies that values are closely related, whereas a high σ indicates values are spread about the mean. The equation for calculating σ can be seen in equation (2.11).

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (x_i - \mu)} \tag{2.11}$$

Skew (s) measures asymmetry about the mean value, allowing easy visualisation of where a large proportion of values lie. In the cases proposed here, a positive skew will indicate low values of either fitness evaluations or generations (favourable), whereas a negative skew will indicate a large number of fitness evaluations and a long algorithm duration (unfavourable) in comparison to the mean. Skew is defined as:

$$s = \frac{\sum \left[(x_i - \mu)^3\right]}{\sigma^3} \tag{2.12}$$

Chapter 3

HP Bead Model on the Square Lattice

As there are many techniques for searching low energy protein structures, common sequence sets are often used to allow a direct comparison of efficiency and performance. These sequences are known as benchmark sequences. For this study, a set of well investigated protein benchmark sequences have been considered: the tortilla HP benchmark sequences [112], for use with the HPLBM on the 2D square lattice. They range in length from 18 to 50 beads and are listed in table 3.1. The table also includes the energy, E^*, of the putative global minimum (or minima, since all of these structures have degenerate global minima) for each sequence.

Sequence ID	Length	GM Energy	Bead Sequence
HP18a	18	-9	$PHP_2HPH_3PH_2PH_5$
HP18b	18	-8	$HPHPH_3P_3H_4P_2H_2$
HP18c	18	-4	$H_2P_5H_2P_3HP_3HP$
HP20a	20	-9	$HPHP_2H_2PHP_2HPH_2P_2HPH$
HP20b	20	-10	$H_3P_2(HP)_2HP_2(HP)_2HP_2H$
HP24	24	-9	$H_2P_2(HP_2)_6H_2$
HP25	25	-8	$P_2HP_2(H_2P_4)_3H_2$
HP36	36	-14	$P_3H_2P_2H_2P_5H_7P_2H_2P_4H_2P_2HP_2$
HP48	48	-23	$P_2H(P_2H_2)_2P_5H_{10}P_6(H_2P_2)_2HP_2H_5$
HP50	50	-21	$H_2(PH)_3PH_4P(HP_3)_3P(HP_3)_2HPH_4(PH)_4H$

Table 3.1: 2D benchmark HP sequences, chain lengths, and corresponding energies [112].

By considering the 2D test case, the performance of the IA can be assessed and

directly compared to a previous study by Cutello *et al.* [65]. This implementation of the search technique utilises the RGA throughout for construction and mutation of individuals.

3.1 Parameter Determination

Comparing search techniques across different platforms, architectures and computers can be difficult, with an array of different factors affecting a process running on a CPU. In order to compare search techniques for efficiency and eliminate such factors, μ_{FE} is an average measure of how many times the fitness is calculated over a series of runs. A single run is defined as a single calculation over a specified number of generations, with a single solution being obtained at the end. A series of runs will be referred to as a job, with μ_{FE} being calculated as an average over the series.

The primary goal, as explained in section 1.1.4, is to find the energetically optimum structural arrangement or the native state. In order to quantify the level of success between runs that discover the native state of the model sequence, μ_{FE} values are compared. The lower the μ_{FE}, the more efficient the run is considered to be, as fewer fitness calculations have been performed. As a potential solution leaves a genetic operator, its fitness is calculated, so μ_{FE} is a direct measure of how many potential solutions have been assessed throughout the period of the calculation.

In order to maximise success rate (SR) and minimise μ_{FE}, all areas of parameter space must be explored. As explained in section 2.1, various parameters affect the outcome of each run. The n_{ind}, i_{max} and n_{clo} values used in the calculation help determine how effective each run will be, regardless of which genetic operators are used. Only when using the standard hyper-macro-mutation operator does m_f become a factor in determining a run's outcome. For this study, the same values of m_f have been taken as in [65].

The investigation of optimal parameter sets involved only the use of the standard

mutation operators, the hypermutation and hypermacromutation operators. Optimal parameter sets were obtained for a variety of n_{ind} values. The work of Cutello et al. [65] showed how various parameters affected SR when using small n_{ind} values (n_{ind} = 10) using an IA for protein structure prediction. In contrast, the work of Cox et al. [22] illustrated better GA performance for high populations (n_{ind} = 200) for the same problems. In this work, n_{ind} adopts values ranging from 10 to 200, taking values of 10, 25, 50, 100 and 200.

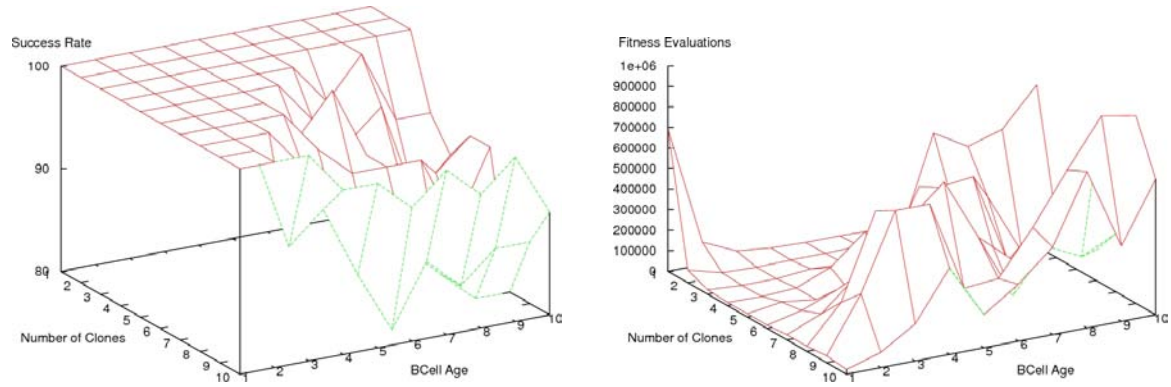


Figure 3.1: Investigation of parameter space, showing the fluctuation in (a) SR and (b) AFE as a function of both n_{clo} and i_{max}.

The parameterisation protocol adopted here mimics the procedure presented by Cutello et al. [65]. To understand how both i_{max} and n_{clo} affect the outcome of each calculation over a range of n_{ind} magnitudes, i_{max} and n_{clo} were combinatorially varied from 1 to 10, with n_{ind} adopting the values listed above. A combination of 500 different parameter sets have been used for each benchmark sequence up to 25 beads in length.

Figure 3.1 maps how SR and μ_{FE} differ as a function of both n_{clo} and i_{max} for sequence HP-20a. Perfect SR is achieved for $i_{max} = 1$, regardless of the magnitude of n_{clo}. However, upon increasing i_{max}, the magnitude of n_{clo} must be decreased in order to maintain SR. By increasing i_{max}, a population is more likely to prematurely converge, as sub-optimal conformations may remain in the population for a longer period of time. As clones are produced, their genetic material is not wasted. By allowing unimproved

individuals to have much shorter lifetimes, problematic (not GM-like and no longer optimisable) individuals can be removed from the population, possibly without loss of their low energy contributions.

Figure 3.1(b), illustrates how increasing n_{clo} may ultimately provide high levels of success, but also increase the magnitude of μ_{FE}. The optimum parameter space is significantly reduced, with combinations of both high magnitudes of n_{clo} and i_{max} reducing calculation efficiency. By providing more clones, the magnitude of μ_{FE} will probably increase, as more individuals are likely to successfully mutate.

Magnitudes of μ_{FE} for the highest magnitudes of SR were collated and graded for overall performance. As a result of this preliminary testing, the results presented below (table 3.2) were obtained using $i_{max} = 4$, $n_{clo} = 3$ and $n_{ind} = 10$. It should be noted that, although these parameters were not optimum for sequence HP-20a, they were in fact the best across all the sequences. All results quoted are averaged over 30 independent runs.

	No Memory B-Cells		Memory B-Cells		Cutello et al.	
Sequence ID	**SR**	**AFE**	**SR**	**AFE**	**SR**	**AFE**
HP-18a	100	89,578	100	117,251	100	69,210
HP-18b	100	40,167	100	200,740	100	41,724.2
HP-18c	100	87,761	100	72,270	100	87,494.5
HP-20a	100	15,221	100	30,414	100	23,710
HP-20b	100	26,207	100	312,405	100	18,085.5
HP-24	100	26,580	100	49,616	100	69,816.7
HP-25	100	79,042	100	95,123	100	269,513.9
HP-36	63	4,867,993	90	3,082,014	100	2,032,504
HP-48	3	6,318,721	3	4,195,086	56.67	6,403,985.3
HP-50	50	4,904,031	96	853,706	100	778,906.4

Table 3.2: Comparison of SR and AFE for the Birmingham IA with and without the use of memory B-Cells with the IA results from [65].

It is apparent from table 3.2 that, although the use of memory B-Cells [65] hinders the discovery of global minima for some of the smaller sequences, it enhances the search for the larger, more difficult to find, sequences. The memory allows mid to high fitness

conformations to remain in the population for a longer number of generations. For larger sequences, this allows a more detailed exploration for certain areas of the PES, permitting the memory B-Cells to converge towards the global solution much sooner. In contrast, for smaller sequences the mid to high fitness range is much smaller, thereby preventing a rapid exploration of the PES by retaining unfavourable segments of local structure for a larger number of generations. Generally, the use of memory B-Cells allows a more diverse inspection of the PES, due to a greater number of the degenerate conformations being found. This is achieved as favourable fragments of local structure are not rapidly disposed of during the retirement process, thereby hindering efficiency.

3.2 Global Minima

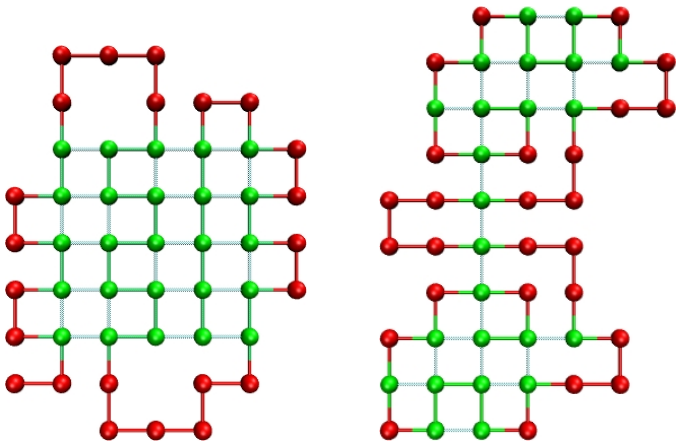


Figure 3.2: GM conformation for sequence (a) HP-48, E^* = -23 and (b) HP-50, E^* = -21. Topological contacts are highlighted in cyan with H beads in green and P beads in red.

The compact structural arrangement present in sequence HP-48 is apparent from figure 3.2(a). With the driving force being the hydrophobic topological contact, the compact hydrophobic cores give rise to a high fitness conformation. The 5×5 hydrophobic core presents a problem to the IA (or other optimisation algorithms [27]) in achieving convergence, as a single misplaced hydrophobic bead will result in only a metastable conformation. The problem does not exist for the HP-50 sequence (figure

3.2(b)), due to the presence of two small hydrophobic cores coupled by a chain of hydrophobic beads. This explains the increase in SR and the decrease in μ_{FE} necessary for HP-50, compared with HP-48 and (when using memory B-cells) even the much shorter HP-36 sequence [27]. The work of Cutello *et al.* supports this idea [65], as similar magnitudes of μ_{FE} for these problematic sequences can be seen, with a much lower success rate for HP-48 than for any other instance.

The structural motifs present in HP-48 (similar to those observed in HP-36), represent the topological contacts of a "β-sheet" arrangement. However, the increased magnitude of SR observed for HP-50 may be attributed to two hydrophobic cores resulting from "α-helical" topological contact arrangements.

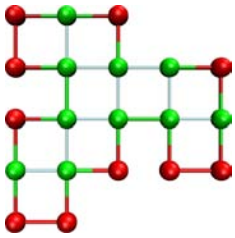


Figure 3.3: GM conformation for sequence HP-20a, E^* = -9 a.u. Topological contacts are highlighted in cyan.

Figure 3.3 shows the GM conformation for sequence HP-20a. Comparing only those sequences that resulted in perfect SR, a common feature is that a terminal H bead is embedded in the conformation. Sequences of the same size see an increase in μ_{FE} with regard to the extent at which this H bead is encapsulated. Sequences HP-18b and HP-20a exhibit the lowest magnitudes of μ_{FE} for their size. They also exhibit a lower level of encapsulation of the innermost (HP-20a) terminal H bead. The increase in μ_{FE} is seen to be linked to the difficulty in placing these terminal beads, in order to result in the correct number of topological contacts. Many more sub-optimal conformations exist (within 1 energy unit) than the number of GMs. Many of these embedded H beads contribute more than one topological contact. A sub-optimal conformation may not lead to the GM, unless complete rearrangement of its conformation occurs. This would

involve not only overcoming a large energy barrier, but also maintaining a sufficient number of topological contacts to remain in the population.

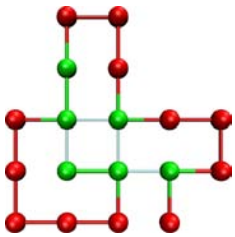
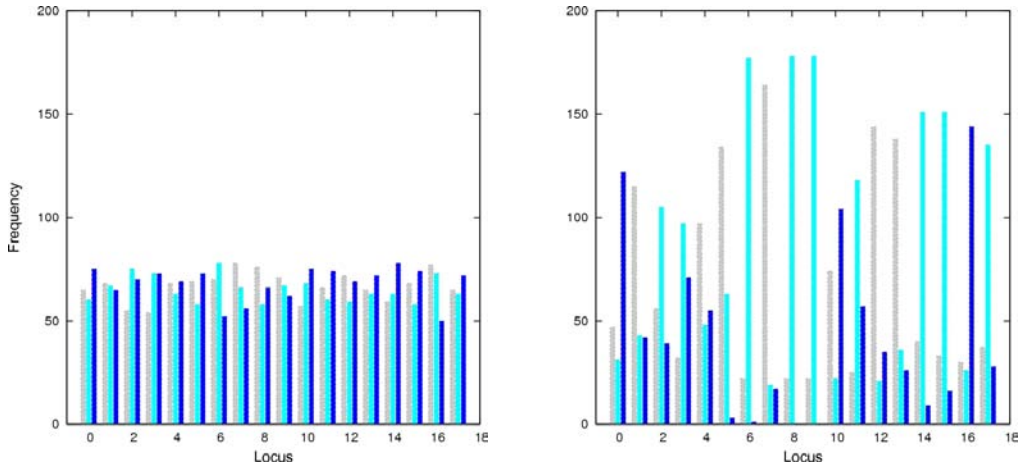


Figure 3.4: GM conformation for sequence HP-18c, E^* = -4 a.u. Topological contacts are highlighted in cyan.

The GM conformation of sequence HP-18c (figure 3.4) shows only 4 topological contacts. The reduction in topological contacts present in this GM when compared to others, may contribute to it being the only conformation to exhibit a reduction in μ_{FE} when using memory B-Cells. However, for sequences up to 25 beads in length, it is the only sequence that contains a single embedded N-terminal (the first) H bead. As this embedded H adopts a fixed lattice position, the construction of the GM conformation is simpler than having to place the embedded chain once the rest of the conformation is already in place. Mutations are more likely to result in valid conformations by fixing this bead, with sub-optimal conformations having a closer relationship to the GM (3 topological contacts for the first 12 beads). By building the conformation from the other direction (placing the P bead first), only a single topological contact exists over the first 13 beads. Therefore, sub-optimal low energy conformations will not produce topological contacts between the correct beads. This would involve overcoming a large energy barrier in order to unfold and re-fold to adopt the GM conformation. By allowing these unfavourable sub-optimal conformations to live longer, premature convergence may occur, or at least an increase in the magnitude of μ_{FE} if the GM were to be found. This is illustrated in that the use of memory cells for the other small sequences (up to 25 beads in length), result in a decrease in efficiency.

3.3 Introducing Search Profiling

For much larger population sizes, ensuring population diversity can be problematic for many search techniques. In this section, a single run, with population size 200 for sequence HP-20a has been analysed. The global minimum was found in generation 28, at which point the algorithm was terminated due to meeting the search criteria. In order to help us understand the progress of the optimisation and ultimately to improve the methodology, monitoring population diversity and the progress of the algorithm is beneficial.



(a) (b)

Figure 3.5: The frequency of alleles at each locus along the model protein chain for the initial population (a) and the final population (b), 0 (grey), 1 (cyan) and 2 (blue).

Figure 3.5(a) assigns a colour to each of the three possible direction decisions (corresponding to alleles in a genetic sense) made when placing each successive bead. It can be seen that initial structure generation, using the constructor, is indeed statistically uniform, showing the frequency of available choices at each locus of the model protein chain to be very similar. In contrast, figure 3.5(b) illustrates how this statistical distribution is skewed in the final population (generation 28), in that the IA has concentrated its search to a much narrower region of the PES. It should also be noted that, position 6 in the chain has a very low frequency of the straight ahead choice (dark

blue), because (for most population members) previous direction decisions preclude (for structural and/or energetic reasons) this choice from being made at this chain position.

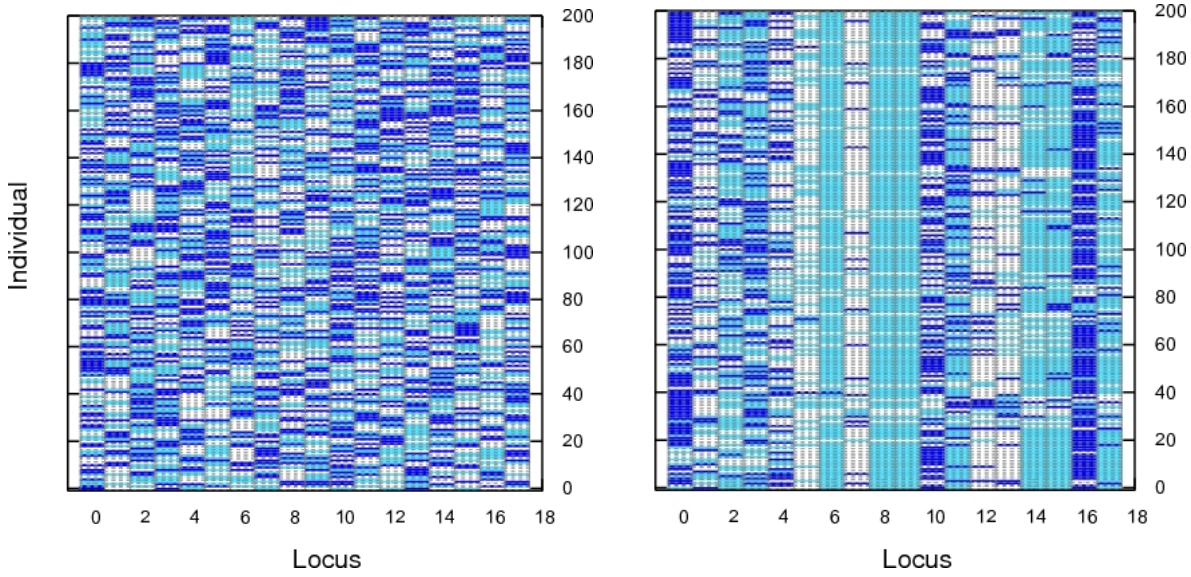
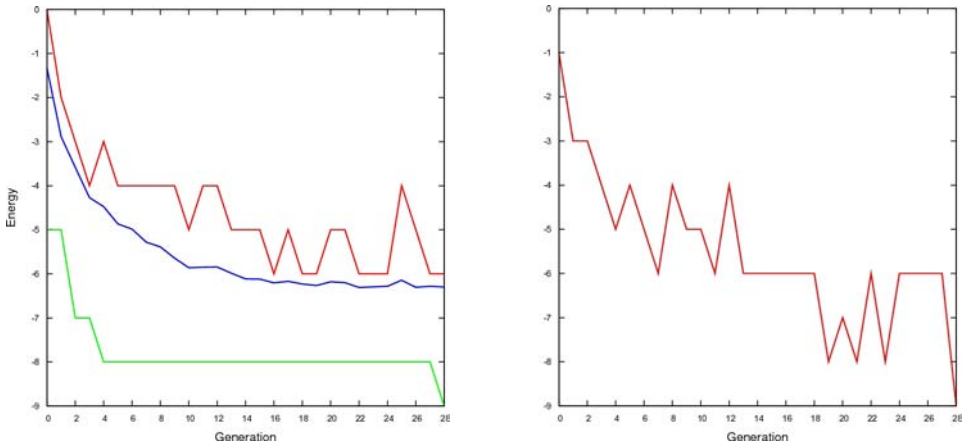


Figure 3.6: Graphical representation of an initial population (a) and final population (b) of B-Cells, left (grey), right (cyan) and straight ahead (blue). Population members are sorted by descending fitness, with structures of the highest energy at the bottom of the plot.

Figure 3.6 shows a graphical representation of the initial and final populations of the calculation. By plotting the conformation vector for each individual, the population can be quickly compared for diversity. Individuals are ordered by descending fitness and the colour scheme is similar to the allele frequency distribution shown in figure 3.5, but with white replacing grey for the left choice. It is clear that initially the population has high diversity (in agreement with the allele frequency plot shown above), with the algorithm preserving favourable regions of local structure (corresponding to schemata in a GA sense) as the calculation converges. A more detailed analysis of the final population shows that there are often correlations (or anti-correlations) between directions at specific loci, with certain combinations giving rise to favourable energies or infeasible structures, respectively.

A typical successful search should see a gradual decrease in the mean energy for a

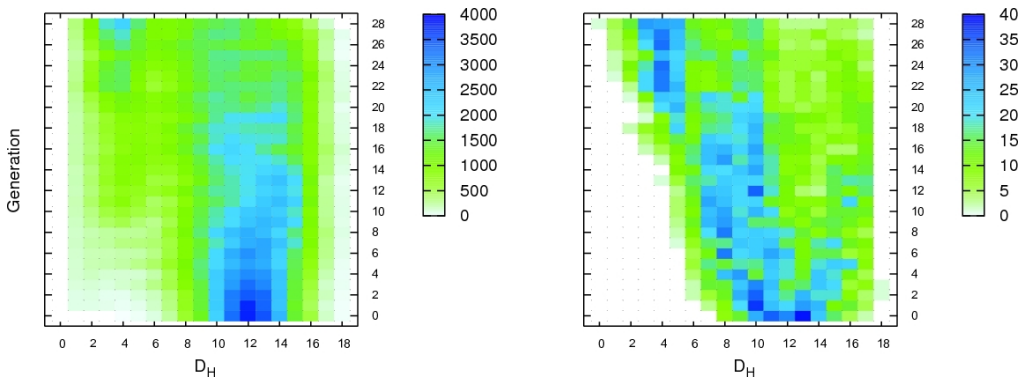


(a) (b)

Figure 3.7: (a) The change in energy throughout the calculation, showing (a) the lowest (green), highest (red) and mean (blue) energies and (b) the energy pathway taken by the GM individual per generation.

generation over time. Figure 3.7(a) shows how the highest, lowest and mean energies fluctuate as a function of generation. A steady decrease is observed as the mutation operators work to drive down the energy of sub-optimal conformations. By plotting the energy of the resultant GM individual per generation (from birth), emphasis can be placed on the fact that conformations must unfold and overcome energy barriers in order to ultimately adopt their most energetically stable arrangement. This is supported in figure 3.7(b), as the energy of this individual suffers many energy increases as a result of mutation. It is also important to note, that the resultant GM individual, remains alive throughout the calculation. Due to ageing, this need not be the case. As mentioned above, the final bead placement involves multiple topological contacts. The GM individual undergoes an energy decrease of 3 energy levels corresponding to the final bead placement. If sub-optimal conformations of energy -6 a.u. exist within the population, it is unlikely that they will mutate to form the GM. It is for this reason that premature convergence is to be avoided by using a diverse population.

For simple protein models such as the HP lattice bead model, D_H can be used as a simple measure of similarity between structures in the population. By calculating the



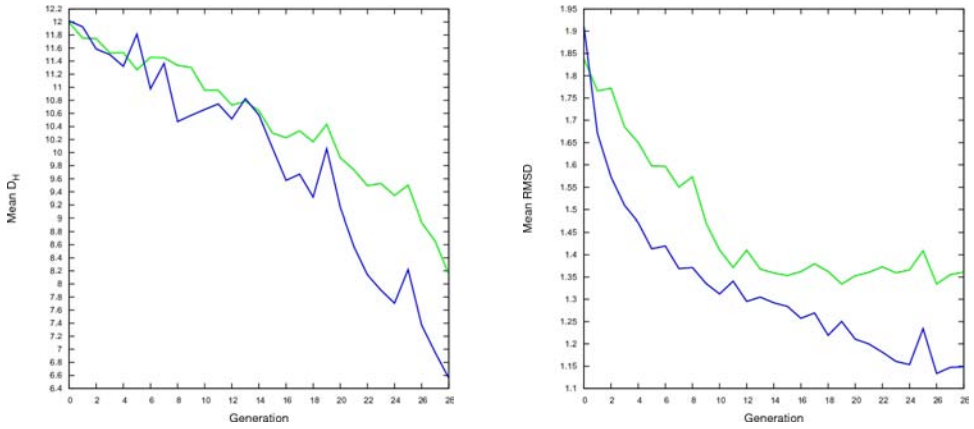
(a) (b)

Figure 3.8: The density of Hamming distances, D_H, between individuals of a population (a) in a pairwise manner and (b) the GM conformation throughout the calculation.

D_H between members of the population in a pairwise manner (as the population size is 200, there are a total of 19,900 pair Hamming distances) and calculating the frequency of each D_H magnitude, density plots can be generated illustrating the dominant D_H magnitudes per generation (figure 3.8(a)). The diversity scale shows a gradual change from white (low density) to blue (high density). Diversity is expected to be high for the initial generation, due to a random allele distribution (figures 3.5(a) and 3.6(a)). This is reflected by a large density for high magnitudes of D_H. As a calculation progresses and the search begins to become more directed (as favourable regions of local structure begin to dominate a population), the most dense region is expected to shift down the D_H scale. As structural diversity shows a more uniform spread (beginning around generation 20), the search focuses on a much more concentrated area of the PES. As the GM region of the PES is approached, this density is expected to decrease again (generation 26). For this case, it confirms that the calculation has not discovered the GM by chance, but a directed search strategy has been employed. It can be seen how the diversity of the population changes as the calculation approaches the GM, which is found in generation 28.

By performing the same density analysis of individuals with respect to the GM, a similar trend is expected. A reduction in D_H frequency arises due to a reduced number

of comparisons (200 for n_{ind} of the same magnitude). The coloured density scale reflects the same relative magnitudes of density, with the maximum density usually of a smaller magnitude. In the hope that individuals in a population begin to adopt GM regions of local structure, the density should gradually decrease over time. This is indeed the case in figure 3.8(b), supporting the idea of a directed search process.



(a) (b)

Figure 3.9: The fluctuation in mean (a) D_H and (b) RMSD as a function of generation between individuals of a population in a pairwise manner (green) and the GM (blue).

Although D_H and RMSD are different measures of structural similarity, they can be used to asses how effective a search technique is. Figure 3.9 illustrates how, as the diversity of the population decreases, so does that of the mean pairwise Hamming Distance with respect to individuals of a population (μ^P_{DH}), the mean Hamming Distance of a population with respect to the lowest energy conformation (μ^L_{DH}), the mean pairwise Hamming Distance with respect to individuals of a population (μ^P_{RMSD}) and the mean RMSD of a population with respect to the lowest energy conformation (μ^L_{RMSD}). By plotting these four quantities, the relationship between conformation vector and 3D conformation can be quantified. This illustrates that in terms of both conformation vector and 3D geometry, over time the individuals of a population exhibit similar structural traits, becoming more like each other and the GM. Surges in these values are either caused by mutation or birthing phases.

3.4 Conclusions

Parameterisation of any search technique is important in order to obtain the greatest performance. Whilst the IA is efficient at searching the PES of smaller sequences, larger sequences still pose a problem. Although implementation of a modified constructor for use in the mutation phase of the IA has not always given greater success rates (especially for more challenging sequences), it has allowed for a more efficient search to be performed, in some cases, showing a decrease in the number of fitness evaluations performed.

The position of terminal beads in the GM of these simple 2D model proteins, dictates algorithm efficiency and success. The more embedded a terminal bead is, combined with the number of topological contacts it must create, can render a search to be less efficient, leaving an individual with large energy barriers to overcome. The presence of a complex hydrophobic core (for the larger sequences) involves many H beads making multiple topological contacts with other beads.

Large sequences exhibiting multiple, smaller hydrophobic cores have larger magnitudes of SR, as the misplacement of a single H bead is not problematic for the search, as seen with single core structures. The structural motifs, resulting in core formation, may explain the observed fluctuations of SR.

The use of algorithm genealogy (introduced in section 2.6) allows a greater understanding of the algorithm's ability to explore areas of the PES of these simple model proteins. The complexity of the problem requires populations to be diverse, with monitoring of this diversity required to improve the methodology of the search technique. Profiling has the potential to identify exactly which operators are beneficial in this type of search technique, and indeed the pitfalls of using others. Areas of favourable local structure along the chain can be assessed, illustrating the important allele combinations that contribute to low energy structures.

Chapter 4

Diamond Lattice Proteins

As described in chapter 3, initial tests for the IA were performed on the well understood, highly investigated benchmark sequences of the HPLBM on the square lattice. This provided a strong base for comparison of algorithm performance and efficiency across numerous search methods within short timescales.

A natural progression from a basic two dimensional structure would be to employ the same model on a three dimensional lattice. For this reason, the diamond lattice, as described in section 2.4.1, is an obvious choice over the less realistic cubic lattice. Using the same simple HP potential as used for the square lattice, the efficiency and performance of the algorithm for exploring a three dimensional search space for structure prediction in relatively short time frames can be investigated. It should be noted, however, that the benchmark sequences for the diamond lattice used here, obtained from [49], are different from the three dimensional benchmark sequences used for the HP cubic lattice system obtained from [112]. They are in fact a selection of sequences taken from a grid search presented in [49].

This study comprises two categories of three dimensional HP sequences. The first category consists of the high degeneracy sequences listed in table 4.1. The second category consists of the low degeneracy sequences presented in table 4.2. The significance of considering both high and low degeneracies is that the ability of search methods to search PESs with many minima or very few minima can be assessed. Although some

sequences are treated separately, they may be in fact mirrored strings of others listed, with the exception of H6 and L9, which are palindromic (i.e. self-mirrors). The importance of investigating the mirrored strings allows insight into how the IA and other search methods treat the growing of chains of beads in both the forwards and reverse direction and how the search space is explored as a result of this.

Sequence ID	Bead Sequence	GM Fitness	Mirror	Degeneracy
H1	PHHPHHHHHHHHHHHHHPHP	5	-	5585
H2	PPHHPHHHHHHHHHHHHHPH	5	H3	7683
H3	HPHHHHHHHHHHHHHPHHPP	5	H2	7683
H4	PPPPHHHHHHHHHHHHHHHH	5	H5	8221
H5	HHHHHHHHHHHHHHHHPPPP	5	H4	8221
H6	PHPHHHHHHHHHHHHHHPHP	5	H6	8345
H7	PPHPHHHHHHHHHHHHHPHH	5	H8	9628
H8	HHPHHHHHHHHHHHHHPHPP	5	H7	9628
H9	PPHPHHHHHHHHHHHHHHPH	5	H10	12372
H10	HPHHHHHHHHHHHHHHPHPP	5	H9	12372

Table 4.1: Three dimensional HP sequences of high degeneracy, corresponding GM fitness, references to mirrored sequences and degeneracies [49]. The H in the sequence ID signifies that these sequences are of high degeneracy.

4.1 Parameter Determination

In order to parameterise the IA for 3D protein models, the same protocol as presented for the 2D case has been employed. n_{clo} and i_{max} have been combinatorially varied from 1 to 10 across a variety of magnitudes of n_{ind} (10, 25, 50, 100 and 200). For each value of n_{ind}, m_f adopts values in the range zero to one with an interval of 0.1, in order to understand how the full range of values affects the mutation phase and thus the final outcome. Due to the mutation memory employed during each generation, it is possible that the value of n_{mut} made available to the hypermacro mutation operator may in fact exceed that of the actual possible number of mutation attempts allowed by the mutation memory operator. In such cases, n_{mut} is assigned the value of actual number of possible mutation attempts derived by the mutation memory operator, thus

Sequence ID	Bead Sequence	GM Fitness	Mirror	Degeneracy
L1	HPHPPHHHHHHHHHHHHPHH	7	L17	2
L2	HPHHPHHPHHPHHHHHHHHH	7	L23	2
L3	HPHHPHHPHHHHHHHHHPHH	7	L16	2
L4	HPHHPHHPHHHHHHHHHHPH	7	L14	2
L5	HPHHPHHHPHHHPHHHHHHH	7	L21	2
L6	HPHHPHHHPHHHHHHHHHPH	7	L13	2
L7	HPHHPHHHHHHHHPHHHPHH	7	L15	2
L8	HPHHPHHHHHHHHHPHHHPH	7	L11	2
L9	HPHHPHHHHHHHHHHPHHPH	7	L9	2
L10	HPHHHPHHHPHHHHHHHHPH	7	L12	2
L11	HPHHHPHHHHHHHHHPHHPH	7	L8	2
L12	HPHHHHHHHHHPHHHPHHPH	7	L10	2
L13	HPHHHHHHHHHHPHHHPHPH	7	L6	2
L14	HPHHHHHHHHHHHPHHPHPH	7	L4	2
L15	HHPHHHPHHHHHHHHHPHPH	7	L7	2
L16	HHPHHHHHHHHHHPHHPHPH	7	L3	2
L17	HHPHHHHHHHHHHHHHPPPH	7	L1	2
L18	HHHHPHHPHPPHHHHHHHHH	7	L22	2
L19	HHHHHHPPHPHHHPHHHHHH	7	L20	2
L20	HHHHHHPHHHPHPPHHHHHH	7	L19	2
L21	HHHHHHHPHHHPHHHPHPH	7	L5	2
L22	HHHHHHHHHPPHPHHPHHHH	7	L18	2
L23	HHHHHHHHHPHHPHHPHPH	7	L2	2
L24	HPHPPHHHHHHHHPHHHHHH	7	L42	4
L25	HPHPPHHHHHHHHHHPHHHH	7	L37	4
L26	HPHHPHHHPHHHHHHHHPHH	7	L33	4
L27	HPHHPHHHHHHHHHPHHPHH	7	L32	4
L28	HPHHHHHPHPPHHHHHHHHH	7	L45	4
L29	HPHHHHHHHPHPPHHHHHHH	7	L43	4
L30	HPHHHHHHHHPPHPHHHHHH	7	L41	4
L31	HHPPHPHHPHHHHHHHHHHH	7	L47	4
L32	HHPHHPHHHHHHHHHPHHPH	7	L27	4
L33	HHPHHHHHHHHHPHHHPHPH	7	L26	4
L34	HHHPHPPHHHHHHHHPHHHH	7	L36	4
L35	HHHHPHHHHHPHPPHHHHHH	7	L39	4
L36	HHHHPHHHHHHHHPPHPHHH	7	L34	4
L37	HHHHPHHHHHHHHHHPPHPH	7	L25	4
L38	HHHHHPHPPHPHHHHHHHHH	7	L46	4
L39	HHHHHHPPHPHHHHHPHHHH	7	L35	4
L40	HHHHHHPHPPHHPHHHHHHH	7	L44	4
L41	HHHHHHPHPPHHHHHHHHPH	7	L30	4
L42	HHHHHHPHHHHHHHHPPHPH	7	L24	4
L43	HHHHHHHPPHPHHHHHHHPH	7	L29	4
L44	HHHHHHHPHHPPHPHHHHHH	7	L40	4
L45	HHHHHHHHHPPHPHHHHHPH	7	L28	4
L46	HHHHHHHHHPHPPHPHHHHH	7	L38	4
L47	HHHHHHHHHHHPHHPHPPHH	7	L31	4
L48	HPHPPHHHPHHHHHHHHHHH	7	-	6

Table 4.2: Three dimensional HP sequences of low degeneracy, corresponding GM fitness, references to mirrored sequences and degeneracies [49]. The L in the sequence ID signifies that these sequences are of low degeneracy.

reducing n_{mut}.

Whereas many 2D benchmark sequences were subjected to parameterisation (section 3.1), only the most degenerate sequence of the low degeneracy set (L48) was examined in this study.

All calculations were assigned $g_{max} = 1000$. A single job consisted of one hundred runs of the IA, with the random number generator uniquely seeded in increments of ten, with an initial seed value of ten.

Population Size (n_{ind})														
10			25			50			100			200		
i_{max}	m_f	n_{clo}	i_{max}	m_f	n_{clo}	i_{max}	m_f	n_{clo}	i_{max}	m_f	n_{clo}	i_{max}	m_f	n_{clo}
1	0.1	1	1	0.1	1	1	0.1	1	1	0.1	1	1	0.1	1
2	0.2	2	2	0.2	2	2	0.2	2	2	0.2	2	2	0.2	2
3	0.3	3	3	0.3	3	3	0.3	3	3	0.3	3	3	0.3	3
4	0.4	4	4	0.4	4	4	0.4	4	4	0.4	4	4	0.4	4
5	0.5	5	5	0.5	5	5	0.5	5	5	0.5	5	5	0.5	5
6	0.6	6	6	0.6	6	6	0.6	6	6	0.6	6	6	0.6	6
7	0.7	7	7	0.7	7	7	0.7	7	7	0.7	7	7	0.7	7
8	0.8	8	8	0.8	8	8	0.8	8	8	0.8	8	8	0.8	8
9	0.9	9	9	0.9	9	9	0.9	9	9	0.9	9	9	0.9	9
10	1.0	10	10	1.0	10	10	1.0	10	10	1.0	10	10	1.0	10

Table 4.3: Parameter combinations used to determine the optimal set. The cells highlighted in yellow mark the parameter value contributing to the optimal set. It should be noted that the number of calculated mutations exceeds the number of possible mutations and for populations of size 50 and 100 renders the mutation factor obsolete for values above 0.1.

Table 4.3 summarises the parameter combinations used for the low degeneracy structure set from which the optimal parameter set was determined. The highlighted cells mark the parameters that contribute to the set that provide the highest SR with the lowest μ_{FE}.

Table 4.3 shows that many values of m_f are "optimal". As the number of physically possible mutations are calculated (by the RGA), it is a possibility that this value is less than n_{mut} (calculated by the hypermutation operator). In such cases, n_{mut} will adopt a value equal to that derived by the RGA. This is indeed the case concerning all

values of m_f greater than 0.1, as the number of mutation attempts is directly affected by m_f as explained in section 2.1.5.1.

When n_{ind} is set to 10, 25 and 200, it can be seen from table 4.3 that m_f adopts the smallest value of 0.1, in order to achieve a maximum SR and a minimum μ_{FE}. By increasing m_f to values greater than 0.1, the value of n_{mut} becomes overridden by the number of possible mutations available. This, in turn, results in calculations proceeding in an identical manner for identical values of i_{max} and n_{clo} with varying m_f.

For the remaining values of n_{ind} (50 and 100), this effect is reversed, in that, if m_f adopts its smallest possible value, a maximum SR is not achieved. In order to achieve a maximum SR combined with a minimum μ_{FE}, m_f can adopt any value greater than 0.1 for identical values of i_{max} and n_{clo}, as the calculations proceed in an identical manner as previously explained.

For n_{ind} of 50 and 100, m_f of 1.0 (to maximise the number of mutation possibilities for other genetic operators) is considered to be a contributing factor in determining the optimal parameter set, with a value of 0.1 clearly determined for other magnitudes of n_{ind}.

Figure 4.1 illustrates how SR fluctuates as a function of i_{max} and n_{clo} for $n_{ind} = 200$ and $m_f = 0.1$. The highest success rate corresponds to $i_{max} = 1$ and $n_{clo} = 3$ for this particular population size, as also highlighted in table 4.3.

Figure 4.1 also illustrates how values of i_{max} greater than the optimal value of 1 actually affect the progress of a calculation. It is apparent that, for the diamond lattice, small values of i_{max}, when used in conjunction with n_{clo}, increase performance as far as SR is concerned. As explained in section 2.1.3, the primary purpose of the ageing operator is to instill diversity into a population by removing an individual if it has failed to mutate after i_{max} generations. In order to maintain diversity and keep the turnover of potential candidate solutions high, a small value of i_{max} can be seen as optimal for any n_{ind} investigated here.

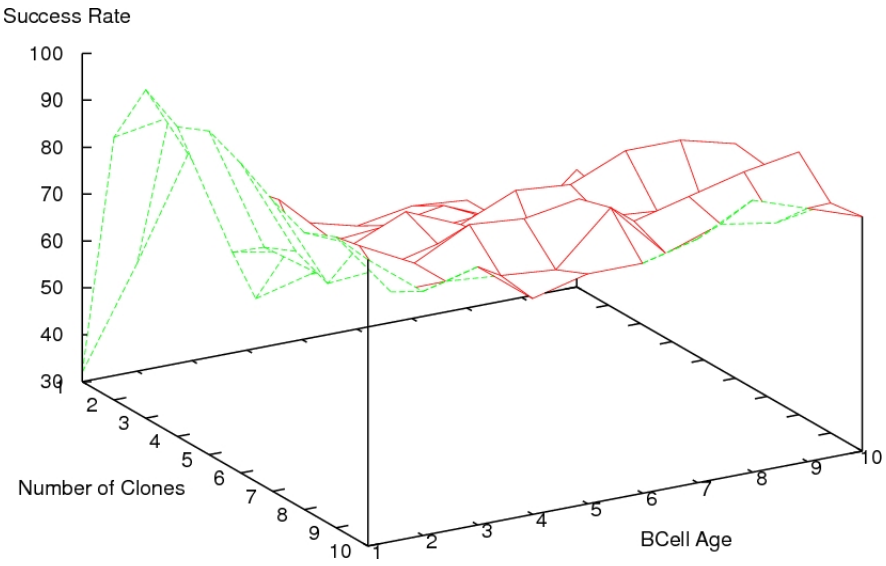


Figure 4.1: Profile of maximum individual age and number of clones with regard to success rate for a population size of 200 and mutation factor of 0.1. This illustrates which parameters give rise to the highest success rate and shows how success rate fluctuates as a function of parameter values.

According to table 4.3 and figure 4.1, a general decrease in n_{clo} is seen with an increase in n_{ind}. The work of Cox et al. [22] demonstrated that the larger the pool of candidate solutions subjected to the genetic operators, the more efficient a calculation would be. This is due to a wealth of genetic information present in the pool of candidate solutions for the genetic operators to work with. In order to keep this wealth of genetic information high, we see generally, that for smaller values of n_{ind}, a larger n_{clo} is required. This gives the genetic operators greater access to favourable structural conformations to work with that may already be present in the current population of individuals for larger n_{ind}, hence the requirement for a lower n_{clo} for larger n_{ind}.

4.2 Global Minima and Algorithm Efficiency

As explained in section 1.1.5.1, a single unitless energy contribution originates from a topological contact. A topological contact exists between two or more non-bonded H beads when they lie on adjacent lattice sites. The sum of these energy contributions

gives rise to a conformation energy. In order to achieve the GM conformation energies specified in table 4.1 for highly degenerate structures and table 4.2 for low degeneracy structures, our GMs must exhibit 5 and 7 topological contacts, respectively.

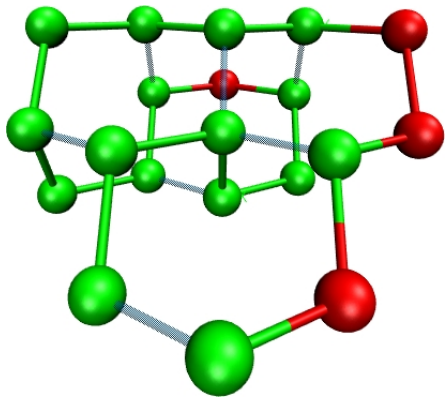
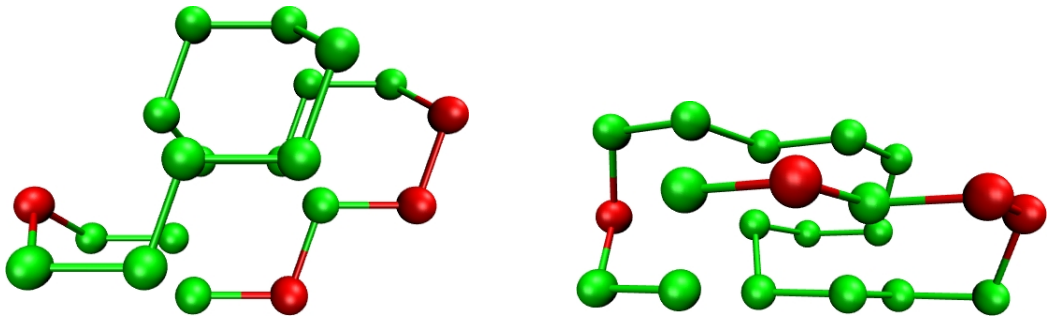


Figure 4.2: A view of the low degeneracy GM L24, illustrating the individual energy contributions in transparent cyan. It should be noted that, as the structure is from a sequence with low degeneracy, the expected energy is -7 a.u., hence it has a fitness of +7. The chosen view allows all topological contacts to be seen.

Figure 4.2 illustrates the individual energy contributions from topological contacts. The structure is of the L24 sequence from the low degenerate set and hence the expected energy of -7 a.u., and fitness of +7 is observed. Figure 4.3 illustrates one of the GMs found by the IA for the L1 sequence from table 4.2. It should be noted that only one GM is pictured from two different views, as the other GM is simply a mirror image of the one shown. Figure 4.3(a) identifies the presence of a honeycomb arrangement due to the restrictions placed on the conformation by the diamond lattice. Figure 4.3(b) illustrates the layered pattern of the beads, comparable with that of a beta sheet protein structure. Example GM for each of the low degenerate sequences featured in table 4.2 can be found in appendix B.2

In order to assess how effective and efficient the IA is as a search method for simple and complex (in terms of number of GMs) PESs, both high and low degeneracy sequences have been subjected to the IA, using the standard mutation operators.



(a) (b)

Figure 4.3: A sample low degeneracy GM (L1), shown from two angles: a) simple view illustrating the honeycomb structure adopted as a result of the diamond lattice, b) another view highlighting the presence of "beta sheet" layers.

Figure 4.4 shows simple plots constructed using the data gathered from 100 runs of the IA for the optimal parameters listed in table 4.3 for the low degenerate sequences featured in table 4.2. The left hand column illustrates how both SR (red) and μ_{FE} (green) vary for each sequence with ascending n_{ind}. The right hand column shows the same trend for identical values of n_{ind} with regard to both the number of unique minima (n_{uniq}) (red) and μ_g (green). If we simply consider SR, it is apparent that we see an overall increase in percentage success as we increase n_{ind}. Population based search techniques rely on an array of genetic material in order to search many areas of the PES simultaneously. As we increase the population size, the variety of genetic material present in a population is increased, thus allowing a greater conformational space to be searched per generation. This results in a greater SR.

Population Size (n_{ind})														
10			25			50			100			200		
i_{max}	m_f	n_{clo}	i_{max}	m_f	n_{clo}	i_{max}	m_f	n_{clo}	i_{max}	m_f	n_{clo}	i_{max}	m_f	n_{clo}
1	0.1	7	1	0.1	5	1	1.0	5	1	1.0	3	1	0.1	3

Table 4.4: The optimum parameters used to compare different genetic operators for varying population sizes. The values have been obtained by inspection of the values presented in table 4.3.

As the goal is to produce results of high SR coupled with low μ_{FE}, μ_{FE} can be

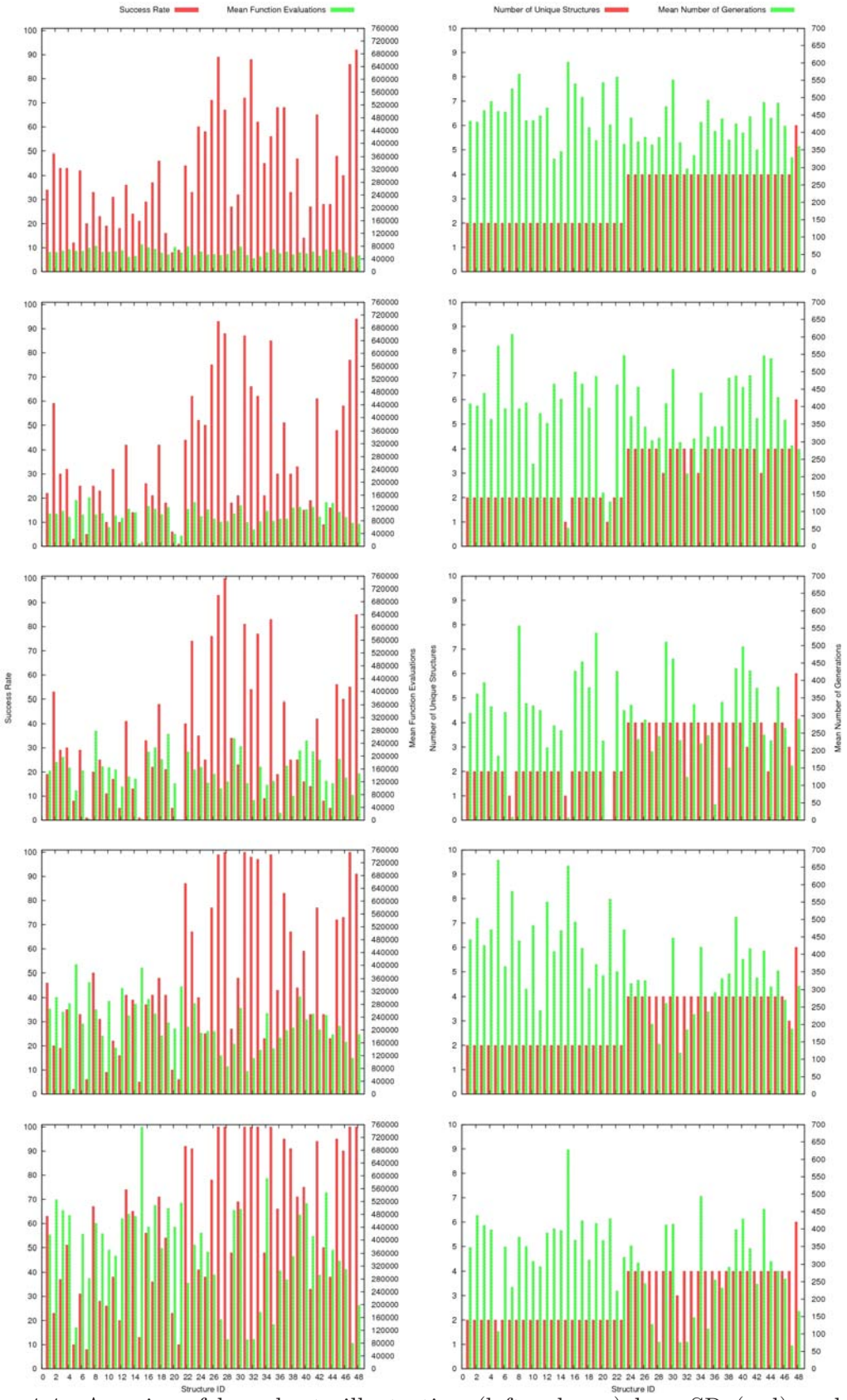


Figure 4.4: A series of bar charts illustrating (left column) how SR (red) and μ_{FE} (green) and (right column) n_{uniq} (red) and μ_g (green) changes with ascending population size for 100 runs when using the optimal parameter set and the low degenerate sequences featured in table 4.2.

seen to vary in a similar manner. It should be noted that μ_{FE} contributions are only made by successful runs of the IA and then averaged over the number of successful runs performed. As previously discussed, the number of fitness evaluations is incremented every time a fitness calculation is performed and thus reflects the number of valid conformations visited in order to reach the GM. Therefore, it is possible to observe a low μ_{FE} along side a low SR and contrastingly a higher μ_{FE} with a higher SR.

Table 4.5 lists the μ_g, σ_g and s_g for n_{ind} of 200 for the standard mutation scheme employed by the IA. For comparative reasons, table 4.5 also exhibits μ_{FE}, σ_{FE} and s_{FE} under the same conditions for all HP Diamond sequences quoted in table 4.2. In order to gain more insight into how the IA is performing, the σ_g is calculated from the μ_g and likewise, the σ_{FE} from the μ_{FE}. In both cases the standard deviations are of the same order of magnitude as the mean values themselves. As the IA searches the PES in a stochastic manner, it is expected that both σ_g and σ_{FE} should be of similar size to their means. As with any σ, highly skewed distributions will lead to σ being significantly different from that of the mean. For this reason, the s_g and s_{FE} have been calculated to quantitatively explain the values of σ. It should be noted that the values of s_g and s_{FE} are identical and therefore represent identical distributions. Identical distributions reflect the linear relationship between a generation and a fitness evaluation. This is to be expected, as proceeding through a calculation requires producing valid conformations as a result of a mutation, increasing the number of fitness evaluations each iteration.

If sequence L2 is considered, it can be seen that both s_g and s_{FE} are 0.05. Taking into account μ_g and σ_g, our data is characteristic of a normal distribution and, in terms of generations, a large proportion of the IA runs would find a GM around generation 430 with our most and least efficient runs discovering the GM at around (430 - 290) and (430 + 290) generations respectively. However, if we consult figure 4.4, it is apparent that sequence L2 has one of the poorer success rates for n_{ind} of 200. In contrast, if we

Sequence ID	μ_g	σ_g	s_g	μ_{FE}	σ_{FE}	s_{FE}
L1	346.71	259.84	0.46	415623.28	311353.44	0.46
L2	438.39	289.25	0.05	525426.08	346546.79	0.05
L3	410.72	348.13	0.32	492291.48	417078.01	0.32
L4	397.58	299.76	0.44	476427.09	359033.30	0.44
L5	106.70	140.52	1.22	127984.90	168290.30	1.22
L6	348.83	266.42	0.15	418045.96	319112.31	0.15
L7	234.37	320.03	1.08	280996.62	383403.02	1.08
L8	376.92	293.15	0.44	451779.20	351202.71	0.44
L9	350.14	302.55	0.51	419637.17	362427.70	0.51
L10	308.19	296.09	0.69	369386.00	354701.70	0.69
L11	293.47	293.97	0.71	351749.89	352144.38	0.71
L12	388.50	335.52	0.37	465743.10	402062.80	0.37
L13	400.86	273.83	0.35	480377.68	328013.34	0.35
L14	395.76	266.65	0.38	474252.43	319391.21	0.38
L15	627.84	297.68	-0.60	752432.23	356642.05	-0.60
L16	367.42	259.43	0.65	440337.71	310774.60	0.65
L17	423.50	289.84	-0.19	507668.80	347315.31	-0.19
L18	311.97	279.00	0.74	374006.70	334305.92	0.74
L19	415.85	237.03	0.18	498571.57	284068.80	0.18
L20	367.26	246.10	1.05	440184.26	294836.95	1.05
L21	429.60	318.77	0.19	515039.30	382014.99	0.19
L22	222.83	224.11	1.30	267166.06	268493.65	1.30
L23	320.19	234.44	0.66	383827.18	280884.50	0.66
L24	351.78	268.53	0.65	421442.43	321529.13	0.65
L25	304.07	305.73	0.86	364359.81	366094.47	0.86
L26	244.41	285.29	1.00	292971.07	341743.07	1.00
L27	127.56	121.85	1.57	153037.57	146001.69	1.57
L28	76.27	71.66	1.84	91580.07	85854.42	1.84
L29	411.47	290.49	0.29	493027.62	347922.64	0.29
L30	413.73	311.97	0.30	495776.37	373679.12	0.30
L31	75.48	73.05	2.60	90641.38	87527.94	2.60
L32	76.46	118.26	4.34	91804.65	141685.27	4.34
L33	147.13	141.32	1.68	176495.41	169357.30	1.68
L34	493.79	302.51	0.10	591756.60	362397.52	0.10
L35	114.78	118.08	1.92	137691.23	141435.94	1.92
L36	254.95	310.68	1.26	305651.39	372202.24	1.26
L37	231.82	226.21	0.91	277938.13	271020.97	0.91
L38	292.19	261.04	0.76	350395.39	312859.89	0.76
L39	398.54	310.27	0.30	477831.85	371846.73	0.30
L40	428.53	268.39	0.31	513659.81	321591.98	0.31
L41	343.45	256.71	0.44	411566.63	307468.65	0.44
L42	243.30	242.30	1.09	291746.28	290337.95	1.09
L43	456.84	317.93	0.11	547560.42	380920.33	0.11
L44	308.23	281.77	0.71	369494.65	337593.65	0.71
L45	280.28	227.26	0.88	335976.81	272256.54	0.88
L46	258.42	268.98	0.96	309850.65	322308.38	0.96
L47	65.92	83.12	2.90	79216.62	99642.27	2.90
L48	164.62	165.36	1.29	197462.49	198157.97	1.29

Table 4.5: Statistics for the three dimensional HP sequences of low degeneracy, showing the mean number of generations, the standard deviation and skewness from the mean, mean number of fitness evaluations as well as the standard deviation and skewness from the mean taken over 100 runs. Values quoted are for a population size of 200 and only considering standard IA mutation schemes.

consider any of the most readily found sequences, L26, L27, L28, L31 (including mirror images L33, L32, L45 and L47, respectively) and L48, all of which have a relatively high SR, both s_g and s_{FE} exhibit strong positive skew (in the range 1.25 - 4.35). This suggests that for the IA to achieve a high SR and lower than median μ_g and μ_{FE} under these conditions, the sequence itself plays an important role in determining how efficient the task of GM discovery is.

Taking into account the number of minima on the PES of energies quoted in table 4.2, a general trend is visible. Figure 4.4 illustrates how the number of GMs affects SR. It is clear that, for all n_{ind}, the total number of GM conformations a sequence has, affects the rate of success, with SR behaving in the same stepwise manner as the number of GM. From table 4.2, we know that sequences L1 to L23 all have a degeneracy of 2, sequences L24 to L47 have a degeneracy of 4 and sequence L48 has a degeneracy of 6. The maximum SR for all n_{ind} for the same sequence ranges, varies in a similar manner. Therefore, it should be noted that the degeneracy shares a direct relationship with SR, as we would expect. When the IA begins to probe the PES, the greater the degeneracy, the greater the chance of finding a GM. The search is expected to take longer and have less probability of success if the degeneracy is lower.

Figure 4.5 presents SR, μ_{FE}, number of unique minima found and μ_g for the IA, utilising the standard mutation operators for n_{ind} = 10, 25, 50 100 and 200, for the highly degenerate HP sequences from table 4.1. It shows that SR is considerably higher for all values of n_{ind} adopted. The degeneracies of these sequences are considerably higher than any of those featured in table 4.2. As previously explained, SR is heavily dependent on degeneracy for these types of problems, which is responsible for the flawless SR observed in figure 4.5. A key feature of these plots, in comparison to the ones seen in figure 4.4, is that, coupled with the consistency of SR, we see a much lower μ_{FE}, indicating the calculations' high efficiency when searching the PES.

As previously explained, the SR percentages are determined over one hundred runs

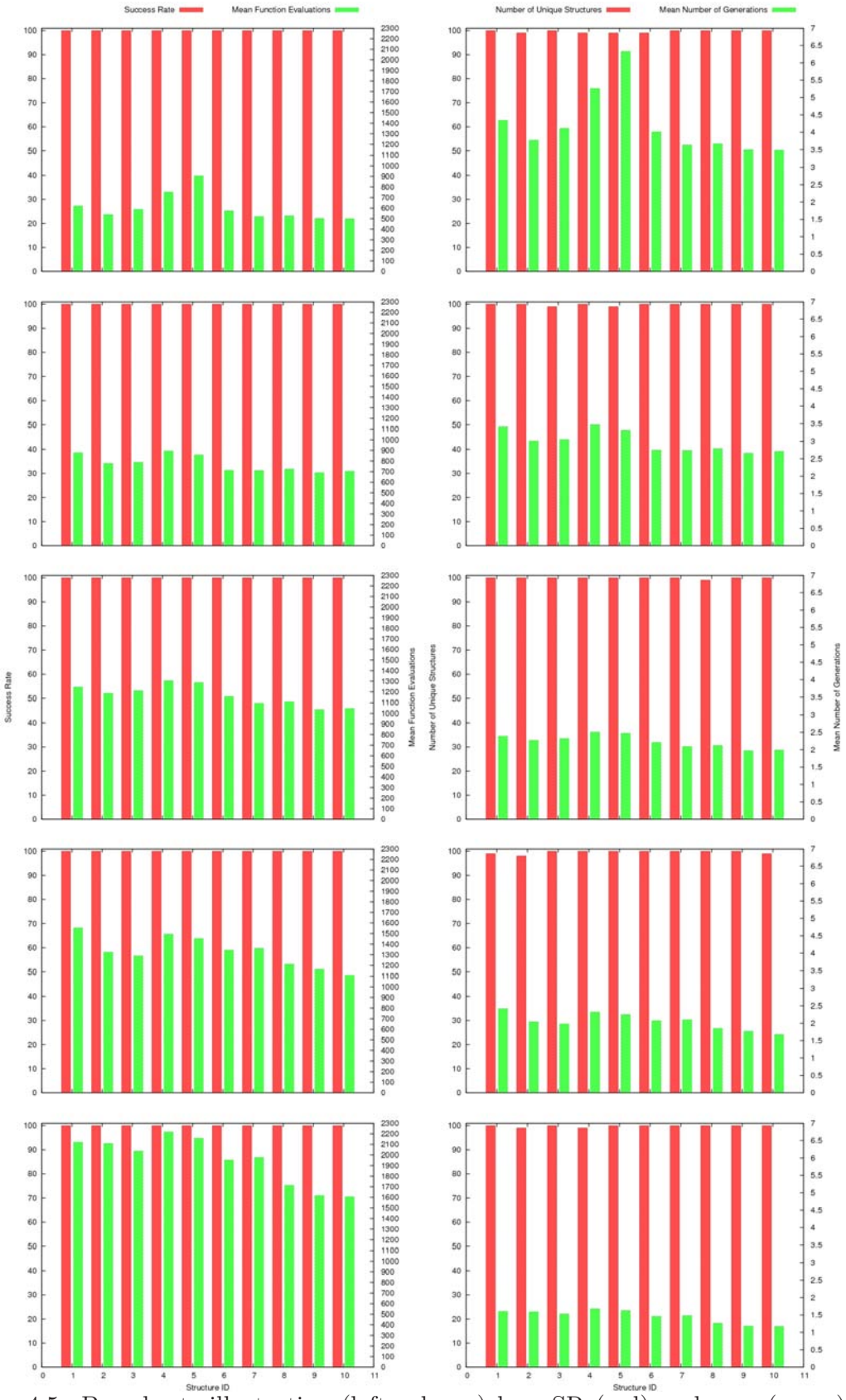


Figure 4.5: Bar charts illustrating (left column) how SR (red) and μ_{FE} (green) and (right column) n_{uniq} (red) and μ_g (green) changes with ascending population size for 100 runs when using the optimal parameter set and the high degeneracy sequences of table 4.1.

of the IA. The complexity of the PESs for the sequences in table 4.2, compared to those in table 4.1, can be assessed by the number of unique GMs discovered during the calculations. We see that for the case of low degeneracy, 75% to 100% of the GMs are discovered by the IA for varying sizes of n_{ind}. With respect to the sequences of high degeneracy, for each run of the IA, each sequence and n_{ind} we witness a unique minima discovery of 99% to 100%. This indicates that, for the highly degenerate sequences, there is a wealth of minima present on the PESs and that again, the degeneracy has a direct effect on the SR and efficiency of stochastic search techniques.

The wealth of minima present on the PES for these sequences cannot be attributed to the chain length, as all the HPLBM sequences investigated here for the diamond lattice contain twenty beads. The H-P ratio is also identical as each sequence has the composition H_4P_{16} and, therefore, cannot contribute to the different PESs for these sets of sequences. As explained in section 1.1.4, there must be a sequence-structure relationship with regard to activity and behaviour of protein molecules. This is indeed the case here, as it is obvious that the different behaviours are witnessed between high and low degeneracies and can only be linked to the position of H beads in the sequence.

Sequence ID	μ_g	σ_g	s_g	μ_{FE}	σ_{FE}	s_{FE}
H1	1.60	1.07	-0.01	2119.98	1291.44	-0.02
H2	1.59	0.92	-0.38	2108.08	1113.27	-0.38
H3	1.53	0.95	-0.08	2036.02	1155.74	-0.08
H4	1.68	1.10	-0.04	2216.54	1335.12	-0.04
H5	1.63	1.13	0.02	2157.97	1367.88	0.03
H6	1.46	1.06	0.03	1952.05	1284.92	0.02
H7	1.48	0.83	-0.29	1976.22	1010.32	-0.29
H8	1.26	0.91	0.09	1712.38	1094.37	0.09
H9	1.18	0.91	0.10	1616.40	1104.16	0.09
H10	1.17	0.91	0.04	1604.49	1100.10	0.04

Table 4.6: Statistics for the three dimensional HP sequences of high degeneracy, showing the mean number of generations, the standard deviation and skewness from the mean, mean number of fitness evaluations as well as the standard deviation and skewness from the mean taken over 100 runs. Values quoted are for a population size of 200 and only considering standard IA mutation schemes.

Table 4.6 demonstrates mean (μ), standard deviation (σ) and skew (s) for both generations and fitness evaluations for the highly degenerate sequences listed in table 4.1 for a $n_{ind} = 200$, using the standard genetic operators. As previously stated, a drastic decrease in μ_{FE} is observed due to the number of minima present on the PES when compared to that of the sequences listed in table 4.2. For the same reason, we notice a uniform distribution of both s and σ. This illustrates that there are no outliers in terms of calculation duration and number of fitness evaluations and that the results gathered for the sequences of high degeneracy show consistent efficiency.

4.3 The Effect of Population Size on Algorithm Efficiency

For ease of use and comparison, a "one size fits all" set of parameters has been selected, using inspection from the optimal parameter sets described in section 4.1. The common parameter set does not include n_{ind}, as comparisons on how population size is affected by various genetic operators will be visited later. The common parameter set is given in table 4.7.

Population Size (n_{ind})														
10			25			50			100			200		
i_{max}	m_f	n_{clo}	i_{max}	m_f	n_{clo}	i_{max}	m_f	n_{clo}	i_{max}	m_f	n_{clo}	i_{max}	m_f	n_{clo}
1	0.1	5	1	0.1	5	1	0.1	5	1	0.1	5	1	0.1	5

Table 4.7: The common parameters used to compare different genetic operators for varying population sizes. The values have been obtained by inspection of the values presented in table 4.3.

In section 4.1, it was shown that, for various values of n_{ind}, the optimal parameters differ somewhat, in order to compensate for the lack of genetic material present in a population at any one time. If comparisons in n_{ind} were to be made with these parameters, then the old phrase "a fair test" would not be upheld. In order to make fair comparisons between results of varying n_{ind}, the common set of parameters in

table 4.7 has been used to reduce the variety in conditions that the IA is subjected to.

Figure 4.6 illustrates how SR, μ_{FE}, n_{uniq} and μ_g varies as a result of using the common parameters outlined in table 4.7 for the sequences of low degeneracy in table 4.2. Upon initial inspection of SR, it is apparent that a general increase is seen as we increase population size for certain sequences. Success behaves in a stepwise manner, as sequences generally with a greater number of GM conformations exhibiting higher maximum SR. This is expected due to the number of GM regions on the PES being greater for sequences exhibiting more GM conformations. However, it should be noted that, moderate success is only seen for each sequence when $n_{ind} = 10$.

Other search techniques [74, 78, 95, 113–116] tend to see an increase in SR with an increase in n_{ind}. These search techniques differ from the IA in a number of ways. The work of Cutello et al [65] illustrated how small values of n_{ind} are beneficial for an IA with the HPLBM on the 2D square lattice. Large values of n_{ind} allow for more areas of a PES to be explored simultaneously. This allows for energy to be minimised (in the case of a minimisation problem such as this) at a faster rate, in terms of gs. Table 4.8 lists the sequences of low degeneracy that have two GM conformations (one if mirror images are neglected), as well as the positions on the chain that contribute two topological contacts. As the sum of topological contacts determine the conformation energy for this model, each position listed in the table contributes -2 a.u. to the overall energy of each GM (of energy -7 a.u.).

Omitting L2 and its mirror L23, it is apparent that all sequences involve a terminal residue contributing two topological contacts in the GM conformation. If the final bead (labelled 20) exhibits this behaviour, then the $(GM - 1)$ sequences must all have an energy of -5 a.u. (making 5 topological contacts). If we consider the data for n_{ind} of 200, where the magnitudes of SR show the greatest diversity, all failed runs (not contributing to SR) result in a energy of -6 a.u. (or fitness of 6). Based on the observation that the final bead placement increases the fitness from 5 to 7 (the GM

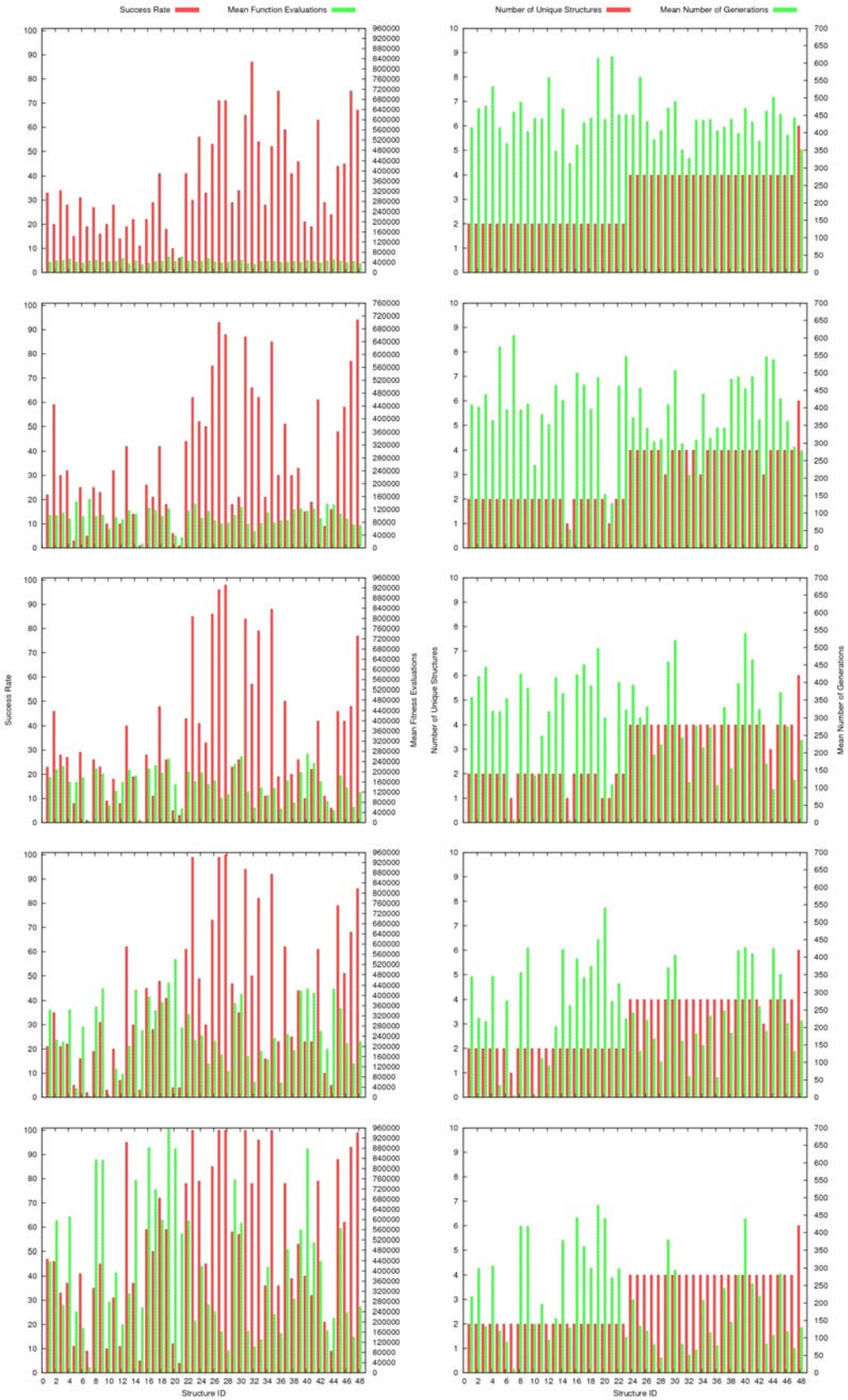


Figure 4.6: Bar charts illustrating (left column) how SR (red) and μ_{FE} (green) and (right column) n_{uniq} (red) and μ_g (green) changes with ascending population size for 100 runs when using the common parameter set and the low degenerate sequences featured in table 4.2.

ID	double topological contacts First	Second	Third	First Contact	SR (%)
L1	1	3	20	8	47
L2	0	-	-	14	46
L3	1	20	-	11	33
L4	1	20	-	11	37
L5	1	12	-	11	11
L6	1	20	-	11	41
L7	1	15	20	8	9
L8	1	20	-	8	35
L9	1	20	-	8	45
L10	1	9	20	8	10
L11	1	20	-	8	31
L12	1	9	20	8	11
L13	1	20	-	8	95
L14	1	20	-	8	37
L15	1	15	20	11	5
L16	1	20	-	8	59
L17	1	3	20	11	50
L18	1	9	20	7	72
L19	9	20	-	9	59
L20	9	20	-	10	12
L21	1	12	-	9	4
L22	1	9	20	8	78
L23	-	-	-	11	100

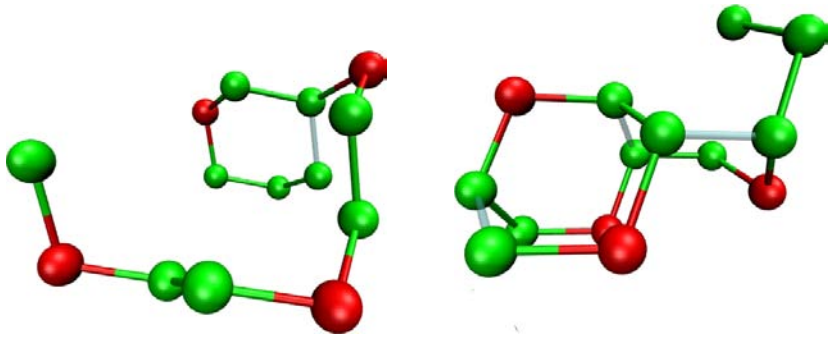
Table 4.8: Sequences exhibiting two GM conformations and the bead position involved in making two topological contacts. Sequences are numbered sequentially, such that the bead labelled position 1 is the first bead for a sequence and the last bead for its corresponding mirror.

fitness), then these conformations require an energy barrier to be overcome. This is achieved by unzipping the topological contacts already present (raising the energy) to produce new ones in such a way as to allow the final bead placement to create two new contacts. It seems that the combination of the aggressive selection operator (choosing the best from a population), the ageing operator and large values of n_{ind} does not allow for these sub-optimal minima (having fitnesses of 6) to be sufficiently overlooked, thus hindering the search process. This combination directs the search process away from the GM resulting a failed search.

If a conformation has not seen an improvement in fitness as a result of a mutation, i_{max} of an individual is exceeded and it is removed from the population. This seems to favour small values of n_{ind}, as a moderate level of success is witnessed for all sequences. The small size of the population may allow for problematic fitnesses (values of 6) to be completely removed, increasing the turnover of lower fitness conformations. This will in turn, increase the ability of the IA to probe new areas of the PES, areas that may indeed lead to the GM conformation. Larger population sizes contain a greater number of problematic fitnesses, forcing the search down a dead end, resulting in failure.

Considering sequences L2 and L23, the table shows that no double topological contacts are formed. This allows for the GM conformation to be found from conformations of fitness = 6. For large population sizes, this results in a dramatic increase in success for sequence L23. However, its mirror (L2), also containing no double topological contacts, suffers in terms of SR. Table 4.8 also lists the first contact made upon folding the chain sequentially according to the GM conformation. These contacts only refer to the initial contact made as the chain folds around itself and not the first bead in the chain to form a contact with beads further down the sequence. It should be noted that, for sequence L2, the first contact is made by bead 14 upon chain growth, i.e. fourteen beads must be placed before a single contact is made. By forming contacts at an earlier stage of chain growth (before the 14th bead), in order to produce the GM conformation, these must be completely unzipped and new ones reformed at the correct stage. This may involve overcoming a large energy barrier, especially if forming the contacts too soon results in a relatively high fitness conformation.

Figure 4.7 shows the first fourteen atoms for both the GM conformation and the highest fitness conformation from a failed run for $n_{ind} = 200$. The first fourteen atoms have been selected according to table 4.8, as this is the minimum number of atoms that need to be placed in order to produce a topological contact according to the GM conformation of sequence L2. It should be noted that a single contact is made after



(a) $F_{HP} = 1$ (b) $F_{HP} = 3$

Figure 4.7: The first fourteen beads of both the (a) GM conformation ($F_{HP} = 7$) and (b) highest fitness conformation ($F_{HP} = 6$) from an unsuccessful run of sequence L2. Fourteen beads have been placed in each case illustrating the first contact formed in the GM. Note how the GM has only one contact ($F_{HP} = 1$) and the sub-optimal minima has three $F_{HP} = 3$. Topological contacts are shown in transparent cyan.

placing all fourteen atoms. However, if we consider the highest fitness conformation from a failed run, it can be seen that, after placing fourteen beads, three topological contacts have already been made. In order to probe the correct area of the PES, after fourteen beads alone have been placed, the contacts would need to be broken (corresponding to an increase in energy from -3 a.u. to 0 a.u.) before placing the atoms again to produce a contact at position 14 (corresponding to an energy of -1 a.u.). The size of this energy barrier also depends on the fact that no other contacts have been made after placing beads fifteen to twenty. Considering these complete minima alone (all twenty beads), a barrier of 6 energy levels would need to be overcome in order to fall into the well of the -7 a.u. GM.

The difficulty of recovering from a search error of this magnitude, can be quantified by considering how early on in the search the IA gets stuck in this local minimum on the PES. All runs of the IA for this n_{ind} managed to find the sub-optimal conformations (of fitness = 6), in the early stages of the search. The larger number of contacts made before the fourteenth bead, and thus the sub-optimal conformation were all discovered in no more than five generations. This illustrates that for this sequence especially,

the search technique does have the ability to drive down the conformation energy very quickly. However, the nature of the sequence results in the IA getting trapped in the well of a local minimum.

The diamond lattice requires the first three beads to be fixed in place, thus allowing only the remaining seventeen beads to change lattice positions. Fixing three H beads (as for L23) as opposed to a H, a P and another H (as for L2), may explain the greater magnitude of SR. However, considering conformations that either exhibit no double topological contacts or conformations that utilise double topological contacts at the first and final bead positions (therefore omitting L5, L19 and their mirrors L21 and L20 respectively), a trend exists concerning SR between mirrored pairs. In a majority of cases, higher success rates are seen for sequences that make a topological contact earlier during chain growth of the GM. However, higher magnitudes of SR are seen for the same pairs when the longest chain of H beads occurs earlier in the sequence.

The nature of the diamond lattice prevents a H bead, when found at the first position, from producing a topological contact with a H bead on a neighbouring lattice site, found before the sixth position. For the case of L23 (SR = 100), the longest chain of H beads is found at the beginning of the sequence. The chain is nine H beads in length, resulting in no possible topological contact being formed within the first five H beads. Only a small number of possible contacts can be formed up to the end of this H chain, considerably reducing the flexibility of the complete chain to form the required number of contacts for the global minimum arrangement.

The initial population, which is the starting point for any population based algorithm, provides the starting position for any search. The population is created using the RGA described in section 2.1.1. By inspecting each adjacent lattice position to the previous bead, the RGA does have the ability to produce compact starting conformations. As with any bead placement method, it is considerably easier to produce a non-compact segment of structure in the early stages of chain growth as opposed to the

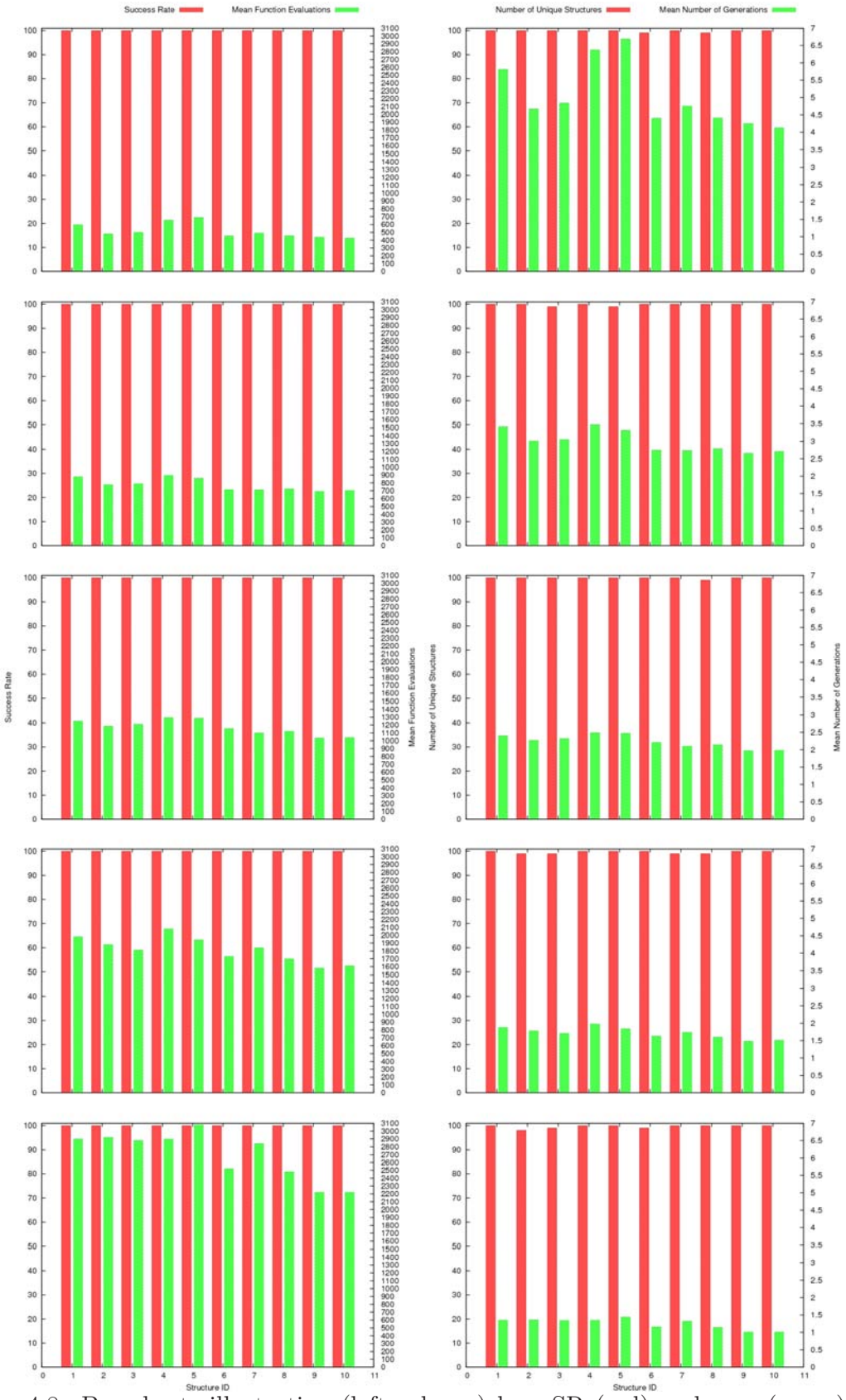


Figure 4.8: Bar charts illustrating (left column) how SR (red) and μ_{FE} (green) and (right column) n_{uniq} (red) and μ_g (green) changes with ascending population size for 100 runs when using the common parameter set and the high degenerate sequences featured in table 4.1.

end. The few possible contacts that can be produced within this short segment if H beads may cause a problem for finding the GM region of the PES, since the initial contact is made after the first P bead has been placed. The longest run for this sequence using a $n_{ind} = 200$, took 406 generations. Obtained by profiling (to be discussed in section 4.5), over the first 395 generations, only once did the conformation not exhibit a contact within the first nine H beads. The same is seen for a much shorter run, one of 67 generations in length. A premature contact is witnessed until generation 65. For this particular sequence at least (and its mirror, L2 demonstrated in figure 4.7), the production of a topological contact too early, seems to be the limiting factor in finding the GM. What also should be noted is that the number of generations needed to find the GM in this case, once no premature contact is observed, is very small indeed. The shape of the PES in this region seems to be a narrow funnel, in that once the initial nine beads do not make any contacts with each other, the energy can be easily driven down by using them to form other contacts with beads further down the chain to result in the GM conformation.

Figure 4.8 shows the performance of the IA for the sequences of high degeneracy from table 4.1, using the common parameter set. As expected, the same behaviour is seen as witnessed when using the optimal parameter set in section 4.2. Due to the degeneracy of each sequence, the IA has no trouble in finding the GM conformation. In fact, a near perfect SR is witnessed for each sequence for each magnitude of n_{ind}. By increasing the magnitude of n_{ind}, the magnitude of μ_g is decreased, showing that in terms of generations, an increase in efficiency is obtained. Due to the increase in the genetic material present per generation, a larger number of areas of the PES are able to be searched simultaneously. This allows one of the many minima to be found more rapidly. For these high degeneracy sequences, this is advantageous and works for the configuration of operators present in the IA, although problems have been seen for sequences of low degeneracy. The efficiency for this set of sequences, however, is

hindered by the increase in magnitude of n_{ind}, as shown by the increase in μ_{FE}. As the population size is increased, the likelihood for a self avoiding conformation to be generated as a result of a mutation is higher. This results in a larger increase in μ_{FE} per generation. As the degeneracy of each sequence is high, small population sizes are enough to successfully search the PES for the GM conformation.

4.4 Comparison with the Genetic Algorithm

Parameterisation for the HPLBM on the diamond lattice was discussed in section 4.1. Previous work has shown that the GA responds well to a higher n_{ind} [117]. The GA used here [49], utilises $n_{ind} = 200$, a much larger magnitude than used for the IA, with inspiration from this work resulting in the testing of larger population sizes during the parameterisation process. It was shown that, although higher levels of success were achieved for some sequences for larger population sizes, $n_{ind} = 10$ delivered a more consistent SR for all sequences. The GA and IA use g_{max} of 10,000 and 20,000, respectively, for the results shown here.

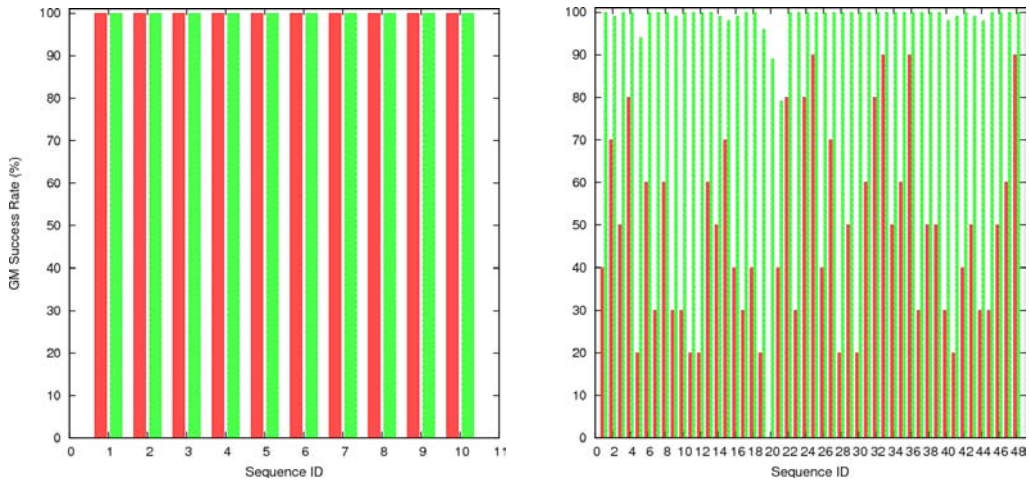
4.4.1 The HP Lattice Bead Model

Figure 4.9 illustrates how the SR for the GA [49] differs from that of the IA using the optimum parameters described in section 4.1. As witnessed during parameterisation, the sequences of high degeneracy have perfect success rates, as shown in figure 4.9(a). As expected, these results show no difference to the GA performance from the data provided, due to the PES of each sequence containing so many global minima. With regard to the sequences of low degeneracy, the success rates in figure 4.9(b) show an improvement in magnitude with regard to the GA for all sequences.

Table 4.9 lists SR for both the GA and IA, as well as supplementary data for the IA. The table illustrates how the IA can out-perform the GA in terms of success for the parameters used. However, as the supplementary data is not published for the GA,

ID	SR_{GA}	SR_{IA}	μ_{FE}	n_{uniq}	μ_g
L1	40.00	100.00	282798.20	2	2006.06
L2	70.00	100.00	234183.94	2	1663.98
L3	50.00	100.00	284273.99	2	2017.75
L4	80.00	100.00	297937.90	2	2116.22
L5	20.00	89.00	879019.23	2	6241.42
L6	60.00	100.00	298139.19	2	2117.03
L7	30.00	100.00	589328.47	2	4179.63
L8	60.00	100.00	445496.18	2	3162.69
L9	30.00	100.00	552677.51	2	3924.52
L10	30.00	99.00	502067.48	2	3563.26
L11	20.00	100.00	359403.15	2	2551.97
L12	20.00	100.00	547371.66	2	3884.20
L13	60.00	100.00	358035.27	2	2542.33
L14	50.00	99.00	601935.39	2	4276.41
L15	70.00	97.00	666927.87	2	4728.83
L16	40.00	100.00	410538.78	2	2914.98
L17	30.00	100.00	314136.57	2	2228.56
L18	40.00	100.00	224647.58	2	1596.52
L19	20.00	97.00	578410.32	2	4109.46
L20	0.00	90.00	924279.07	2	6567.17
L21	40.00	81.00	945870.07	2	6714.56
L22	80.00	100.00	250919.98	2	1783.90
L23	30.00	100.00	319623.60	2	2271.18
L24	80.00	100.00	148630.09	4	1054.90
L25	90.00	100.00	151951.39	4	1078.86
L26	40.00	100.00	106214.62	4	753.83
L27	70.00	100.00	71664.63	4	508.64
L28	20.00	100.00	117517.49	4	835.06
L29	50.00	100.00	418958.58	4	2976.55
L30	20.00	100.00	388567.63	4	2759.76
L31	60.00	100.00	101875.56	4	723.19
L32	80.00	100.00	59363.42	4	421.38
L33	90.00	100.00	116121.24	4	824.04
L34	50.00	100.00	190772.37	4	1354.99
L35	60.00	100.00	173465.95	4	1232.73
L36	90.00	100.00	128353.58	4	910.06
L37	30.00	100.00	138755.38	4	984.84
L38	50.00	100.00	355901.74	4	2529.27
L39	50.00	100.00	187829.84	4	1334.85
L40	30.00	100.00	648638.47	4	4609.56
L41	20.00	99.00	412037.67	4	2925.91
L42	40.00	100.00	119939.62	4	851.20
L43	50.00	100.00	413168.20	4	2934.22
L44	30.00	100.00	466863.13	4	3317.57
L45	30.00	100.00	235926.16	4	1675.98
L46	50.00	100.00	257693.96	4	1831.66
L47	60.00	100.00	72481.31	4	513.94
L48	90.00	100.00	65881.94	6	467.55

Table 4.9: GA SRs [49] and SR, μ_{AFE}, n_{uniq} and μ_g for the IA for the sequences of low degeneracy using the HPLBM.



(a) (b)

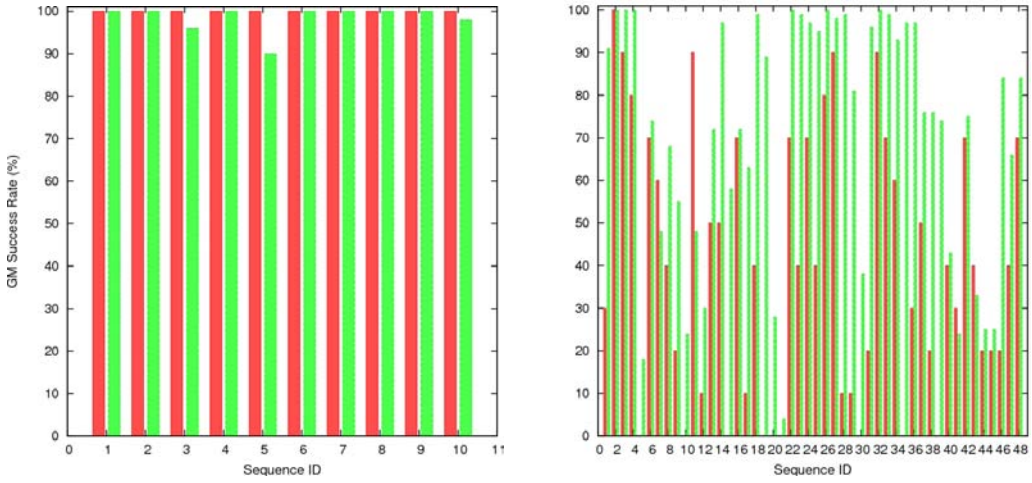
Figure 4.9: Success rate plots for the HPLBM on the diamond lattice for the sequences of (a) high degeneracy and (b) low degeneracy for the GA (red) [49] and the IA (green).

it has been provided for the IA to aid comparisons with possible future work.

4.4.2 The BLN Model

Figure 4.10 illustrates how the SR for the GA [49] compares to that of the IA using the optimal parameters described in section 4.1. However, one modification has been made to the parameter set used by the IA. As the parameters determined previously were for the HPLBM on the diamond lattice, the BLNM exhibits a more noisy PES, and therefore to compensate, m_f has been increased from 0.1 to 1.0. This increases the number of attempted mutations per individual, as described in equation (2.2).

Figure 4.10 shows that for the sequences of high degeneracy, comparable SRs are achieved. However, for sequences H3, H5 and H10, equivalent magnitudes of SR are not witnessed for the GA and the IA. The lowest energy conformations obtained as a result of an unsuccessful run, maximise the distance between P beads, as governed by the repulsive interactions of the BLNM potential. However, the first possible topological interaction occurs too early upon chain growth. This results in a different compact core arrangement from the GM for a sequence and thus, would require the conformation to



(a) (b)

Figure 4.10: Success rate plots for the BLNM on the diamond lattice for the sequences of (a) high degeneracy and (b) low degeneracy for the GA (red) [49] and the IA (green).

completely unfold, breaking all contacts to reform the initial contact between the two correct beads. This will involve overcoming a large energy barrier, five energy levels in height according to the HPLBM potential alone.

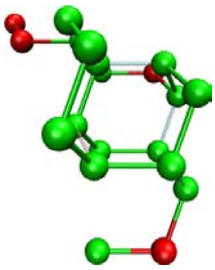


Figure 4.11: The GM conformation of sequence H3, with the initial topological contact formed through chain growth highlighted in pink, with all remaining topological contacts highlighted in cyan.

Figure 4.12 illustrates the four lowest energy conformations obtained by the unsuccessful runs for sequence H3. Figures 4.12(a), 4.12(b) and 4.12(c) all share a single correctly positioned topological contact with the GM. All conformations make the five topological contacts required by both the HPLBM and the BLNM. However, to minimise the repulsive interaction between P beads, the compact core arrangement

prevents their distances from being maximised.

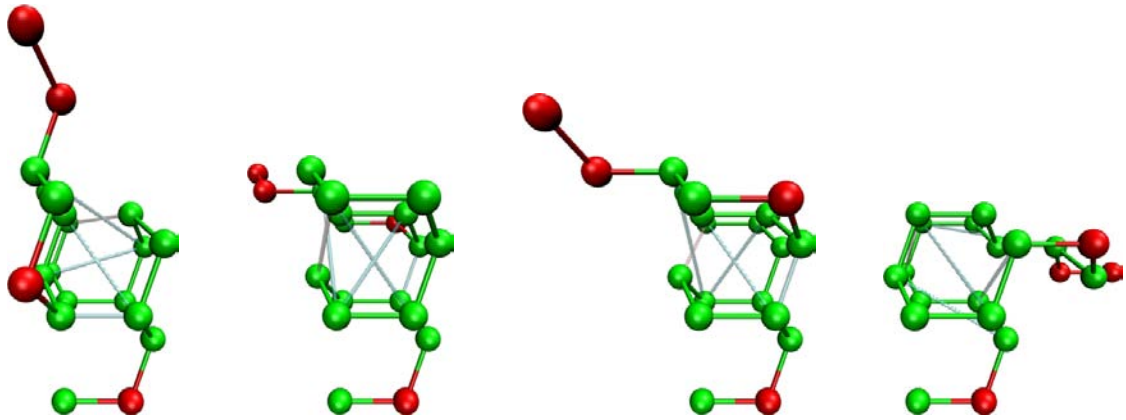


Figure 4.12: Lowest energy conformations of the four failed runs for sequence H3. GM topological contacts have been maintined to highlight the difference. (a) F_{BLN} = -1.45783 (b) F_{BLN} = -1.45783 (c) F_{BLN} = -1.45783 (d) F_{BLN} = -1.45789. The initial "topological contact" formed through chain growth is highlighted in pink with all other "topological contacts" highlighted in cyan.

Table 4.10 lists the SR for both the GA [49], and the IA, as well as the supplementary statistics μ_{FE}, n_{uniq} and μ_g, for the BLNM on the diamond lattice.

The IA saw a consistent improvement in the success compared to the GA when negotiating the PES of HPLBM proteins. The reduced success observed with the BLNM for both search techniques, is attributed to the increase in noise of the energy landscape, resulting from the distance dependent potential it employs. Some sequences (L7, L11, L41 and L43) are best searched using the GA as a better SR is observed. However, the sequences themselves show no obvious trend explaining why this is the case.

The improvement in SR can only be attributed to the methodology of the two techniques. Local minimum trapping is an issue with any optimisation method. The IA combats this with the use of small population sizes and an ageing operator. The GA relies on selection, mutation and large population sizes to prevent population stagnation. Reinitialisation of the population can occur if population convergence is observed for

ID	$\mathbf{SR}_{GA}$	$\mathbf{SR}_{IA}$	μ_{FE}	n_{uniq}	μ_g
L1	30.00	91.00	569046.57	2	5870.69
L2	100.00	100.00	131816.54	2	1360.71
L3	90.00	100.00	261220.31	2	2722.08
L4	80.00	100.00	494285.06	2	5176.53
L5	0.00	18.00	854608.16	2	8801.16
L6	70.00	74.00	733840.35	2	7628.50
L7	60.00	48.00	955246.64	2	10002.54
L8	40.00	68.00	710927.66	2	7419.41
L9	20.00	55.00	874713.21	2	9195.36
L10	0.00	24.00	1016629.45	2	10575.16
L11	90.00	48.00	729616.93	2	7695.33
L12	10.00	30.00	737162.26	2	7723.20
L13	50.00	72.00	673034.80	2	6957.94
L14	50.00	97.00	543840.28	2	5700.00
L15	0.00	58.00	927987.06	2	9703.05
L16	70.00	72.00	734053.01	2	7730.68
L17	10.00	63.00	768782.58	2	7945.44
L18	40.00	99.00	335587.46	2	3516.53
L19	0.00	89.00	605247.67	2	6349.76
L20	0.00	28.00	1023271.14	2	10760.28
L21	0.00	4.00	1068772.00	2	10901.75
L22	70.00	100.00	330867.41	2	3443.43
L23	40.00	99.00	379676.65	2	4016.58
L24	70.00	97.00	363593.48	2	3760.97
L25	40.00	95.00	485697.11	2	4999.50
L26	80.00	100.00	179932.92	2	1886.30
L27	90.00	98.00	405294.17	2	4218.62
L28	10.00	99.00	316419.67	2	3316.06
L29	10.00	81.00	696310.30	2	7226.35
L30	0.00	38.00	773244.84	2	8150.55
L31	20.00	96.00	498957.45	2	5183.13
L32	90.00	100.00	394887.20	2	4148.83
L33	70.00	99.00	286138.34	2	2952.05
L34	60.00	93.00	560344.66	2	5854.43
L35	0.00	97.00	592423.16	2	6242.80
L36	30.00	97.00	447130.74	2	4673.81
L37	50.00	76.00	551794.44	2	5755.19
L38	20.00	76.00	699265.88	4	7305.30
L39	0.00	74.00	787468.32	2	8235.08
L40	40.00	43.00	875202.04	2	9147.11
L41	30.00	24.00	639384.41	2	6658.87
L42	70.00	75.00	604582.70	2	6329.13
L43	40.00	33.00	777679.87	2	8115.45
L44	20.00	25.00	1018592.72	2	10527.04
L45	20.00	25.00	914405.64	2	9554.84
L46	20.00	84.00	624375.29	4	6468.57
L47	40.00	66.00	766663.78	2	7866.78
L48	70.00	84.00	585509.14	2	6031.78

Table 4.10: GA SRs [49] and SR, μ_{AFE}, n_{uniq} and μ_g for the IA for the sequences of low degeneracy using the BLNM.

the GA. Minima are recorded and entered back into the population. The IA attempts to treat this on the fly, providing an opportunity for any individuals (of both low and high energies) that survive the ageing process to remain in the population. This allows a combination of favourable and unfavourable genetic material to be present, with further mutation possibilities for individuals of high energy. If insufficient individuals have survived ageing, birthing can allow new and old genetic material to be investigated simultaneously. As ageing reduces population stagnation and allows old and new genetic material to co-exist, it may result in the increase in success rate observed for the IA.

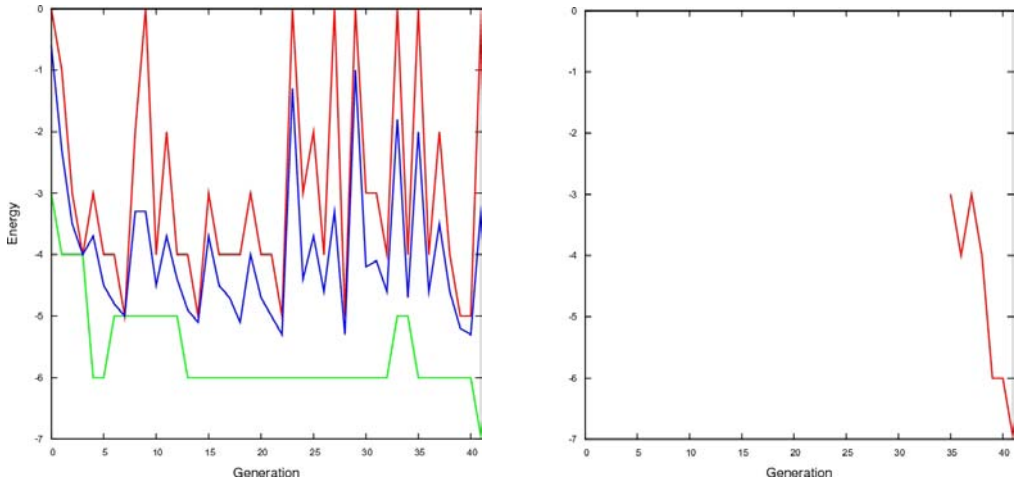
4.5 Profiling the HP Diamond System

It was shown in figure 4.9, how the success rates fluctuate between sequences of low degeneracy for the HPLBM on the diamond lattice for both the GA and the IA. With respect to the IA, sequence L21 showed the worst performance, with SR = 81%. This section focuses on profiling a successful and a failed run with respect to this sequence. Although sequence L21 was not the worst case for the GA, profiling is only performed with respect to the IA. As described in section 2.6, profiling prefers a parent-individual ratio of 1:1. Thus the GA cannot be as successfully profiled due to the use of the crossover operator requiring two or more parents to produce two or more offspring. The parameters used in this work are the optimum parameters determined in section 4.1 for a $n_{ind} = 10$ as explained in section 4.4.

4.5.1 A Successful Case

Here a successful search is under scrutiny, with the GM found in generation 41.

Figure 4.13(a) illustrates how the highest, lowest and mean energies fluctuate as a function of generation. Upon initial inspection, the profiles appear noisy. The level of noise may be attributed to the IA incorporating an ageing operator, allowing indi-

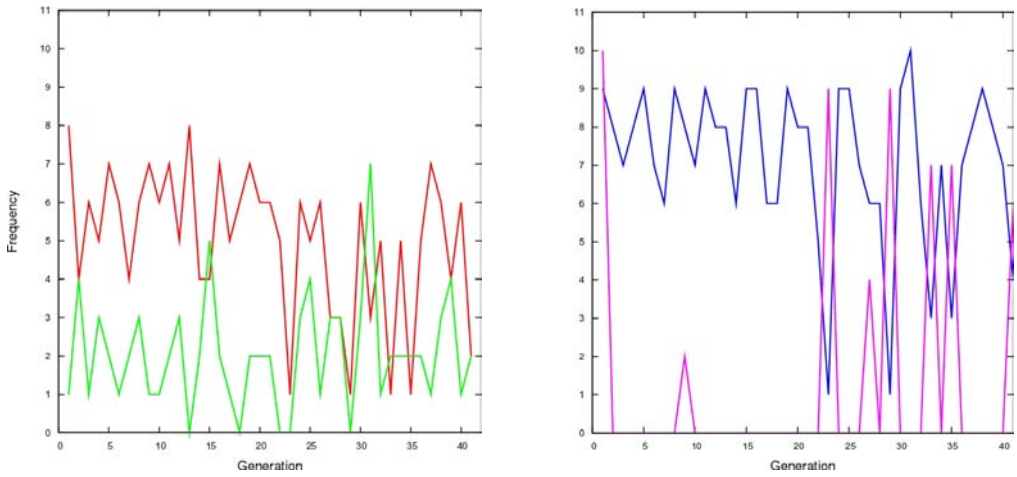


(a) (b)

Figure 4.13: Fitness as a function of generation for the (a) lowest (green), mean (blue) and highest (red) as a function of generation and (b) the GM for sequence L21. A full plot is not observed for the GM as the individual was born in generation 35.

viduals to be removed if an improvement in fitness does not result from a mutation. This would allow for a decrease in energy measures unlike with other search techniques. Generations 9, 23, 27, 29, 33, 35 and 41 all witness a dramatic increase in the highest energy value, with the mean values exhibiting more of a shift for generations 23, 27, 33, 35. The increase in energy arises due to the number of births (2, 9, 4, 9, 7, 7, 6 respectively, figure 4.14) occurring in those generations. Generations that fail to invoke a birthing phase, do not exhibit such a drastic decrease in highest or mean energies. Figure 4.13(b) shows no activity for the resulting GM individual until generation 35. This shows that the GM individual was produced during birthing of this generation, with a rapid descent into the well of the GM funnel. This may imply that this region of the PES is narrow and deep.

Larger deviations in the mean are found when a population only contains mutated or new individuals. Figure 4.14(b) illustrates that for generations 23, 27, 29, 33, 35 and 41, mutations exist for all of the individuals that survived the previous generation. Figure 4.14(a) illustrates that the number of hypermutations performed, frequently

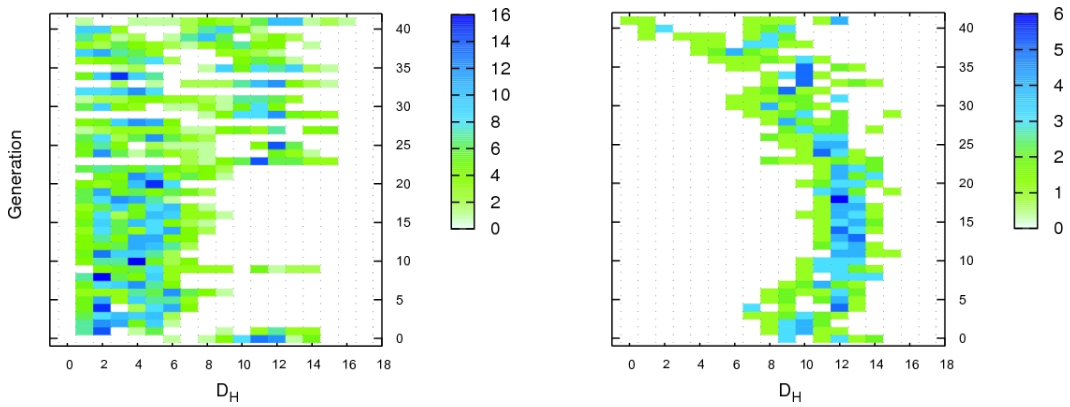


(a) (b)

Figure 4.14: (a) The number of mutations as a function of generation for the hypermutation (red) and hyper-macro-mutation (green) operators. (b) The number of total mutations (blue) and the number of births (magenta) as a function of generation.

exceeds the number of hypermacromutations. This is expected as the hypermutation operator only performs a single point mutation, whereas the hypermacromutation operator performs a number of point mutations across a range of loci. Performing only a single point mutation is more likely to result in a valid conformation than when performing more. However, performing many point mutations is necessary to hop from one region of the PES to another, especially if the mutations take place near the centre of the conformation.

Comparing the D_H between individuals in a population in a pairwise manner, aids the understanding of how the individuals relate to each other with respect to their conformation vectors. Figure 4.15 illustrates the number of individuals that exhibit a particular D_H as a function of generation. The IA incorporates an operator that removes any degeneracy (with regard to D_H) within a population. For this reason, figure 4.15(a) illustrates that no individuals exhibit $D_H = 0$. Throughout, the population remains fairly diverse with respect to D_H, with diversity surges exhibited for the birthing generations. With major birthing contributions exhibited from generation 23, it is the

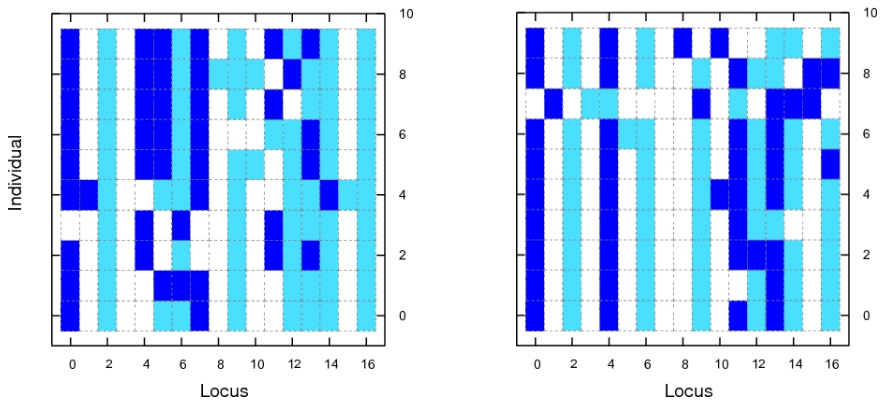


(a) (b)

Figure 4.15: (a) The diversity within a population as a function of generation with respect to (a) other individuals in the population in a pairwise manner and (b) the GM.

regular introduction of these new individuals that provided the population diversity. Unfortunately, as the populations are as diverse as for the initial generation, according to this plot the success may be attributed to a fortunate sequence of mutations and not necessarily due to a directed search. However, in figure 4.15(b), the D_H density does begin to decrease with respect to the GM from generation 23 onwards. Although new material is regularly subjected to the population for and after this generation, the existing individuals allow their central conformations to proliferate and thus dominate the population. High energy conformations are usually added during the birth phase as they are randomly generated (for this case, no new individual had an energy lower than -3 a.u.). The probability that an existing conformation (one of already lower energy), may remain in the population (i.e. push out the high energy conformations) is significantly increased, resulting in population domination. However, if a mid-energy new-born individual does enter the population and successfully mutates, a change in dominant local structure for the following population may be achieved (as in generation 23).

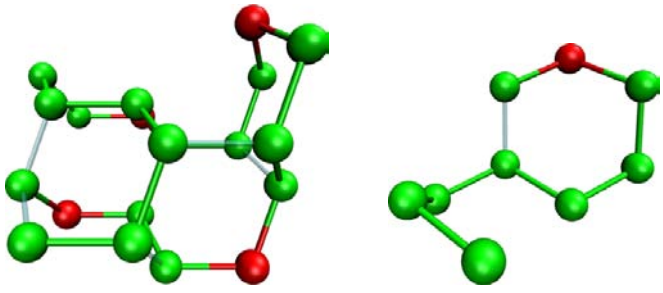
Figure 4.16 shows the conformation vectors of the populations for generation 21 and 24. The GM of sequence L21 exhibits its first topological contact (upon chain growth)



(a) (b)

Figure 4.16: Conformation vectors (0 = white, 1 = cyan and 2 = blue) of individuals for generations (a) 21 and (b) 24 with individuals listed in order of fitness, with the highest fitness in position 0.

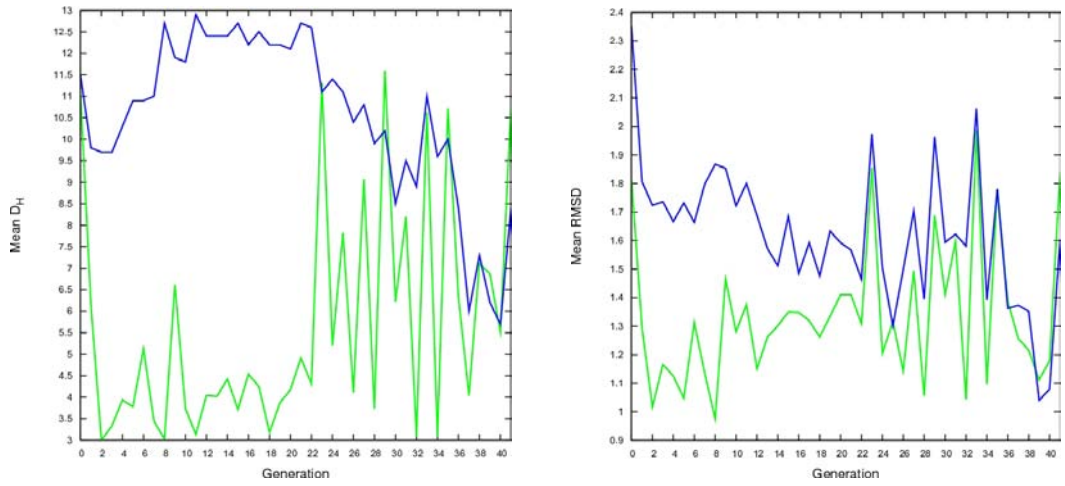
between loci 3 and 8 (figure 4.17(b)). For a conformation to exhibit this contact, a mutation process may involve breaking all contacts whilst maintaining sufficient new contacts (once refolded) to remain in the population. With hypermutation performing only a single point mutation and with the number of hypermacromutations being lower per generation, overcoming a large energy barrier may be unlikely, requiring a birth phase to change the region of search space. Producing this first contact is the key to success for this particular sequence. As the near-terminal beads of the model protein are more mutatable, the central configuration (in this case, loci 3 to 8) determines the region of the PES that the search probes.



(a) (b)

Figure 4.17: (a) The GM conformation of sequence L21. (b) The first nine beads of the L21 sequence. Topological contacts are shown in cyan.

Figure 4.18 shows how μ^P_{DH}, μ^L_{DH}, μ^P_{RMSD} and μ^L_{RMSD} fluctuate as a function of generation. Figure 4.18(a) shows a much lower μ^P_{DH} than μ^L_{DH} until generation 22. The change in magnitude witnessed after generation 22, results from the many birthing phases, producing between 4 and 9 new individuals. However, although the birthing phase was invoked, a decrease in μ^L_{DH} is observed. This signifies that the search is no longer random, but is more directed, and continues to be so for the remaining generations. After generation 35 (the GM individual birth phase), μ^L_{DH} exhibits a large decrease, resembling the magnitude of μ^P_{DH}. This indicates that the diversity of the population has begun to settle, with individuals showing a stronger relationship with each other and the GM. This is expected for a directed search, however, as the final generation sees another birthing phase, both measures increase.



(a) (b)

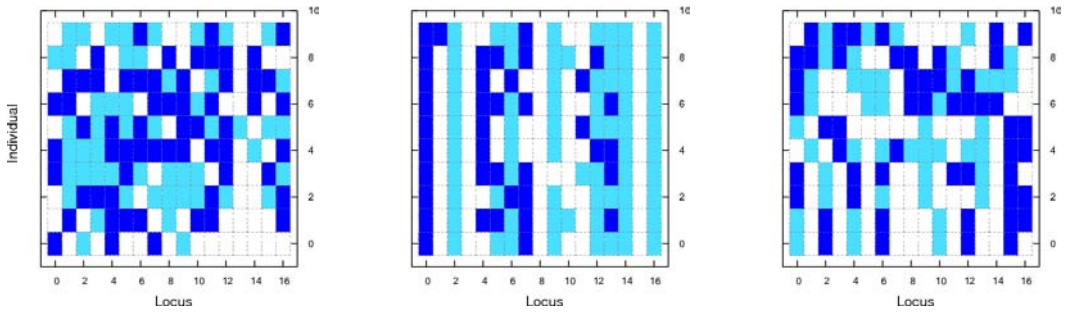
Figure 4.18: Mean (a) D_H and (b) RMSD as a function of generation with respect to individuals in the population (green) and the GM (blue).

In terms of 3D conformation, a greater similarity exists between members of the population and the GM. Both μ^P_{RMSD} and μ^L_{RMSD} see a drastic decrease over the first two generations. As hypermacromutations help to drive down the energy of non-compact conformations, the level of compactness increases and so a decrease in the mean RMSD is observed. By measuring the mean RMSD, birthing phases are still

apparent, shown by the local maxima. However, as the number of GM present on this surface is small, it provides a good insight into how related the individuals are to each other and the GM. Although D_H is a quick, simple measure of conformation vector similarity, it is misleading in this case. A more stable D_H (showing a consistent, large difference between the GM and pairwise measures) than RMSD is observed. This suggests that the conformations are more closely related to each other than they are to the GM. However, on inspection of the RMSD measure, it is apparent that the individuals become less related to each other and more related to the GM in terms of 3D conformation. With large population changes (via birthing), both μ^P_{RMSD} and μ^L_{RMSD} fluctuate significantly. However, after generation 35, the search continues to improve once a GM related conformation is obtained, thus reducing both μ^P_{RMSD} and μ^L_{RMSD}.

Figure 4.19 shows the conformation vectors of the initial, mid and final populations. As the initial population contains only new-born individuals, no common regions of local structure exist, illustrating the random distribution of alleles and that no bias exists in population initialisation. By generation 20, there is a domination of specific alleles at certain loci, illustrating a lower population diversity (figure 4.15(a)). Loci 1, 3, 8, 11 and 13 show the correct allele domination in comparison to the GM. However, as the initial contact is formed between the third and eighth loci, with loci 4, 5, 6 and 7, not exhibiting GM alleles, the conformations still lie in different regions of the PES. Low energy conformations across all shown conformation vectors exhibit a minimal number of trans-like torsion angles (allele = 2), as a cis-like torsion angle is preferred in order to generate a compact structure and therefore the required number of topological contacts.

For the IA to successfully search the energy landscape of HPLBM proteins on the diamond lattice, mutation alone is not sufficient to prevent local minimum trapping. Birthing phases must be invoked in order to raise the mean energy for a generation,



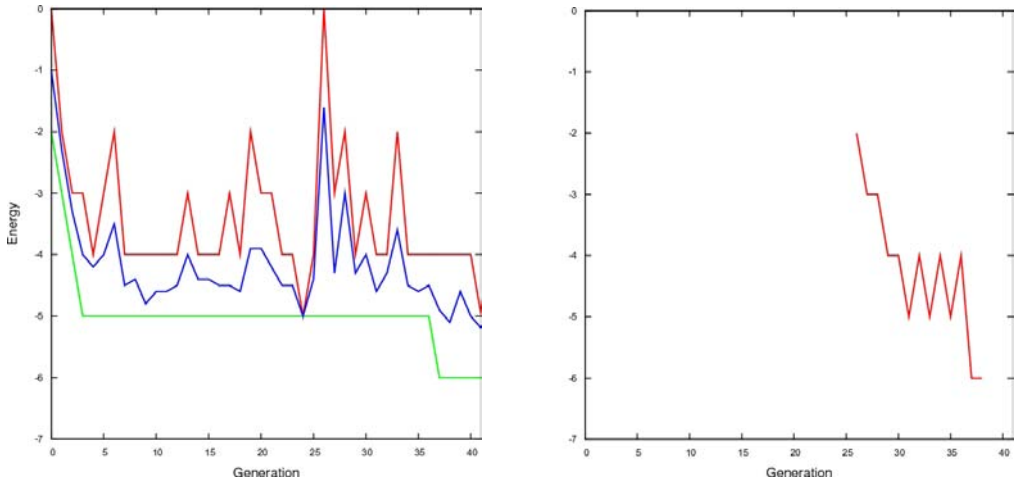
(a) (b) (c)

Figure 4.19: Conformation vector mappings (0 in white, 1 in cyan and 2 in blue) for sequence L21 for (a) the initially constructed population, (b) the population for generation 20 and (c) the final population (including the GM). Individuals are ordered with respect to their fitnesses, with the highest fitness individual at position 0 in the population.

to shift the search to a new area of the energy landscape. In order for the new-born individuals to survive a generation, their energy must contribute to the stabilisation of the mean, i.e. they cannot be high energy conformations (due to the nature of the selection operator). Although the total number of mutations performed remains quite stable (unless a birthing phase is required), the type of genetic material provided as a result of each mutation type differs. Hypermutations are beneficial when trying to improve the energy of already compact, low energy conformations, whereas hypermacromutations provide larger reductions to the energy of non-compact, high energy conformations (shortly after birthing phases). Populations remain fairly diverse for these simple lattice proteins, with an unstable pairwise D_H density, due to the frequent birthing phases and mutations of near-terminal loci. Birthing phases may lead to a change in dominant central regions of local structure, resulting in the diversity of a population showing a drastic increase. In order for the search to negotiate the PES successfully, the sequence of alleles, determining the correct initial topological contact formed (upon chain growth), must be present to result in an efficient, directed search process.

4.5.2 An Unsuccessful Case

By analysing an unsuccessful run for the same sequence, under the same conditions, possible reasons for its failure can be highlighted. Here data is shown for the same range of generations as for the successful case, with the lowest energy conformation not seeing further improvement after generation 37.



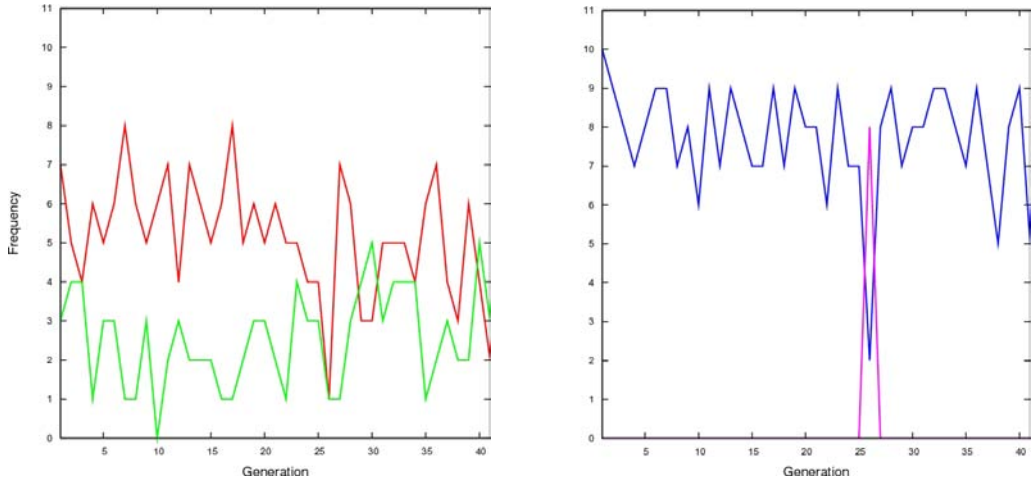
(a) (b)

Figure 4.20: Fitness as a function of generation for the (a) lowest (green), mean (blue) and highest (red) as a function of generation and (b) the lowest energy conformation obtained for sequence L21.

Figure 4.20(a) shows how the highest, lowest and mean energies fluctuate as a function of generation for the failed case. A more stable approach to the lowest energy conformation is seen for the lowest energy per generation, with less fluctuation of the highest and mean energies in comparison with the successful case. Birthing gives rise to a surge in mean and highest energy for a generation, with only a single birthing phase initialised at generation 26 (figure 4.21(b)). Whereas fluctuation in both the highest and mean energies is expected due to near-terminal mutations, the stability of the mean energy for generation prior to the birthing phase, suggests that these near-terminal mutations are dominant. As conformational changes to the centre of the sequence are desirable to investigate different funnels of the PES, the presence of the

birthing phase suggests that the search may have become trapped in a local minimum.

As observed for the successful case, birthing phases are usually accompanied by mutants of all of the remaining individuals from the previous generation. This is upheld here and is seen for generation 26 in figure 4.21(b).

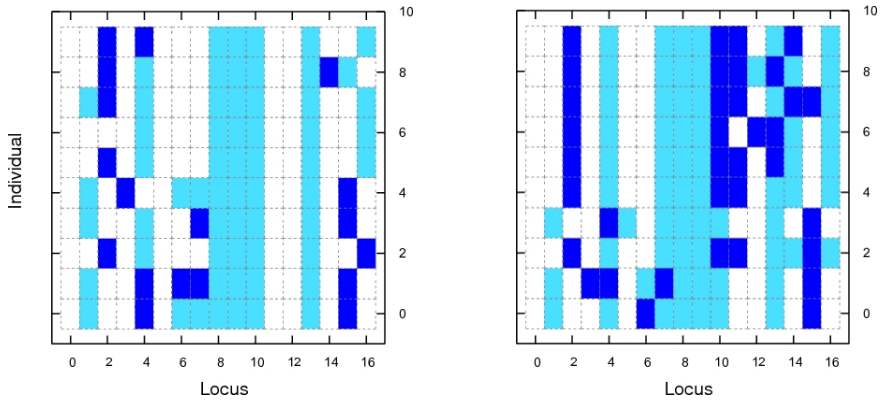


(a) (b)

Figure 4.21: (a) The number of mutations as a function of generation for the hyper-mutation (red) and hyper-macro-mutation (green) operators. (b) The number of total mutations (blue) and the number of births (magenta) as a function of generation.

Figure 4.22 shows the conformation vectors of populations for generations 17 and 23. A change to the dominant alleles at loci 7 and 10 are observed. However, from figure 4.20(a), no improvement to the lowest energy is seen over these generations, only a peak to the mean and highest values. This suggests that although a change to the central conformation may shift the focus of the search to a different region of the PES, the importance of forming the correct initial topological contact (loci 3 and 8 for the L21 case), and thus the correct allele combination up to the eighth locus, is of highest priority for HPLBM sequence on the diamond lattice.

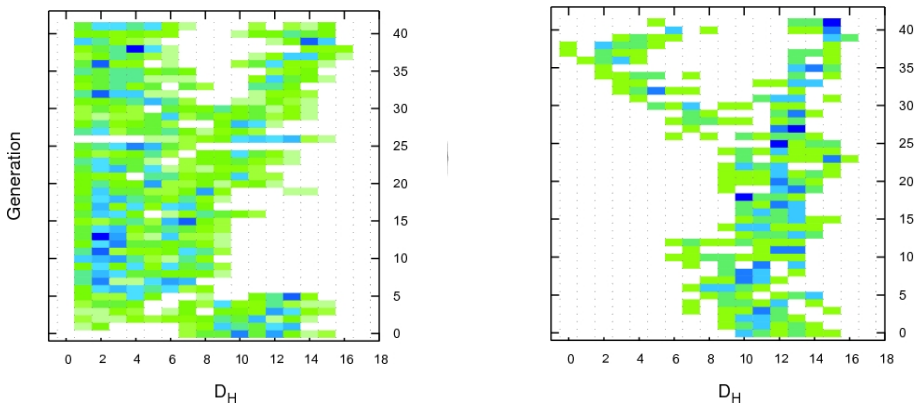
Figure 4.20(b) illustrates that the resulting lowest energy conformation individual was born in generation 26. Seven mutations occurred in generation 28 resulting from this grandparent. This illustrates how a new-born individual has the potential to



(a) (b)

Figure 4.22: Conformation vectors (0 = white, 1 = cyan and 2 = blue) of individuals for generations (a) 17 and (b) 23 with individuals listed in order of fitness, with the highest fitness in position 0.

dominate proceeding populations, focusing the search onto a different region of the PES.



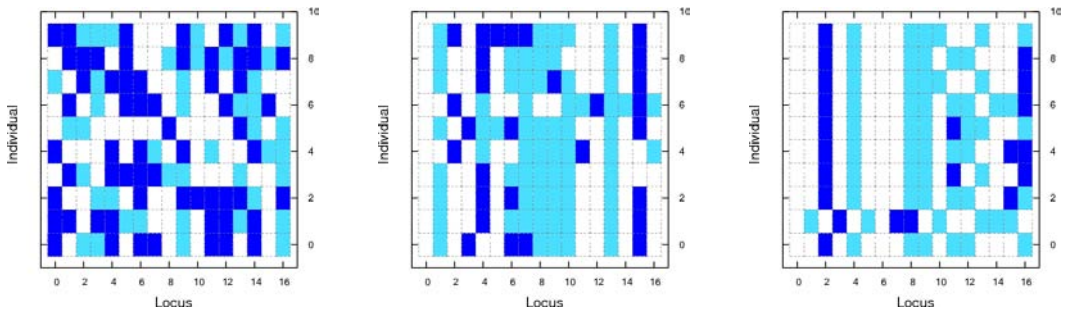
(a) (b)

Figure 4.23: (a) The diversity within a population as a function of generation with respect to (a) other individuals in the population in a pairwise manner and (b) the lowest energy conformation for the first 41 generations.

The diversity within the population with regard to the individuals and the lowest energy conformation is described in figure 4.23. As expected, the initial diversity is large, with the population predominantly adopting D_H values in the region of 10. Although the diversity with respect to the lowest energy conformation remains quite

high, with respect to the population itself, a dramatic decrease is seen over the first twenty generations. This decrease mimics that seen for the mean energy, suggesting that the individuals are closely related in terms of conformation and energy. The diversity increases at generation 20, reflecting the mutation pattern that gave rise to the change in dominance of alleles at loci 7 and 10. This diversity begins to stabilise, with an increase seen for the new-born individuals produced in generation 26.

With a change in dominant local structure observed again, the diversity begins to increase, branching off into two significant regions of high and low density. The branching effect (initialised in generation 26) illustrates that the population is now searching two independent regions of the PES, with two groups of five closely related conformations. This branching effect is also observed with respect to the lowest energy conformation in figure 4.23(b). This suggests that whilst descending into the funnel of the lowest energy conformation, another search is in progress elsewhere. The level of branching decreases after generation 37 when the lowest energy conformation is removed from the population, with its level of dominance over the population decreasing in proceeding generations.



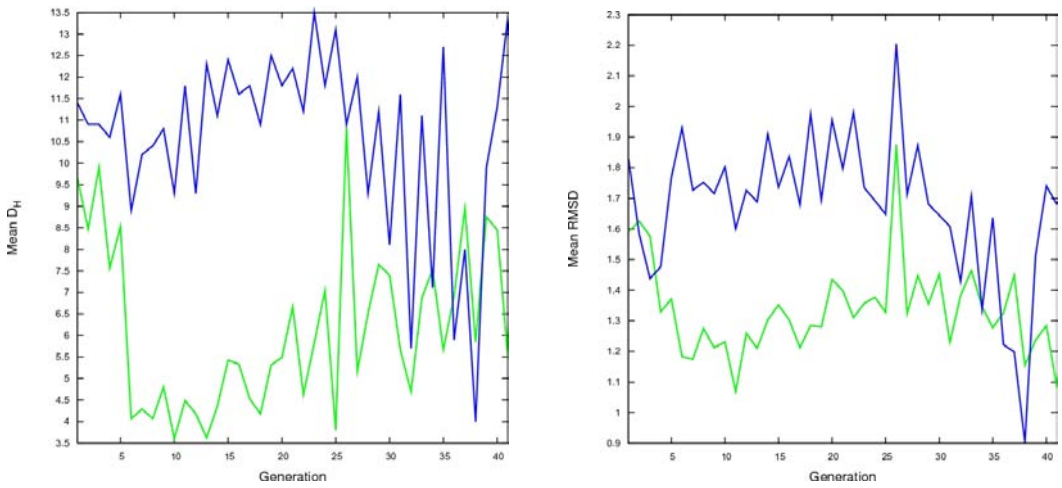
(a) (b) (c)

Figure 4.24: Conformation vector mappings (0 in white, 1 in cyan and 2 in blue) for sequence L21 for (a) the initially constructed population, (b) the population for generation 20 and (c) the final population (including the lowest energy conformation). Individuals are ordered with respect to their fitnesses, with the highest fitness individual at position 0 in the population.

Generation 43 exhibits another birthing phase, introducing nine new individuals and

a single mutated individual originating from the lowest energy conformation in generation 41. The population dominating conformation for the following 40 generations, shares the first eleven loci with this individual. As mutations for these generations predominantly occur from the twelfth locus, it suggests that the search is trapped in a broad, deep funnel of the PES. The arrangement of these loci does not reflect that of the GM, and therefore the topological contact between the third and eighth loci does not exist.

Figure 4.24 illustrates how the populations have changed, from the initial population to the population in generation 41. The random distribution of alleles present for the initial population, indicates that the individuals are generated randomly. However, by generation 20, central and terminal loci show dominant alleles. By generation 41, loci 0 to 9 adopt the same alleles for all but one individual, with seven sharing the tenth locus. It is this conformation segment that remains dominant for a further 40 generations.



(a) (b)

Figure 4.25: Mean (a) D_H and (b) RMSD as a function of generation with respect to individuals in the population (green) and the lowest energy conformation (blue).

When comparing the two case studies, fluctuations in μ^P_{DH} for the unsuccessful case appear similar to those of the successful one. A dramatic decrease in magnitude is

observed, once favourable, low energy local structure begins to proliferate. It was shown in section 4.5.1, that the erratic behaviour of μ^P_{DH} was caused by the birthing phases invoked at certain generations, with the magnitude of μ^P_{DH} remaining fairly stable for other generations. However, more erratic behaviour is observed for the unsuccessful case. Only a single birthing phase exists (generation 26), with dramatic deviations in the magnitude of μ^P_{DH} for other generations. This suggests that in terms of D_H, the population is more diverse for this case than for the successful one, with regular mutations actually occurring from either one terminus or the other over eight loci.

Figure 4.23(a) illustrated a larger population diversity than in figure 4.15(a) (the successful case). Although maximising population diversity is beneficial for large populations, failure here may be attributed to small populations exhibiting too much diversity (in terms of D_H), preventing the search from exploring a region of the PES thoroughly. Comparing the magnitudes of μ^L_{DH} between the successful and unsuccessful cases, may not be accurate, due to the final low energy conformation being different, and, hence, different regions of the PES being explored. However, it does illustrate, due to the increase in diversity, how the unsuccessful case is not able to channel the individuals towards a single region of the PES as well as the successful case. This is supported by the fact that the lowest magnitude of mean D_H with respect to both measures is never as low as for the successful case.

In terms of RMSD, the diversity issue is supported, with figure 4.25(b) exhibiting smaller deviations than in figure 4.18(b) (omitting birthing generations). A large fluctuation in the magnitudes of both μ^P_{RMSD} and μ^L_{RMSD} as a result of a birthing phase, is observed as for the successful case. As seen for the D_H measures, a larger deviation exists between the magnitudes of μ^P_{RMSD} and μ^L_{RMSD} (up to generation 26 and after generation 37), illustrating that in terms of 3D conformation, the search is unable to gradually reduce the search space explored.

Success or failure seem to be independent of how the energies fluctuate during the

search process. The formation of the initial GM topological contact, seems to be the limiting factor in the search for these model proteins. The successful case demonstrated that a new-born individual with GM characteristics (correct alleles between the ninth and c-terminal loci), can result in rapid success in determination of the GM. However, although conformations that adopt this local structure exist for the failed case, the level of diversity within a population is greater, thus hindering a thorough search of PES regions. This results in a search that never finds the GM, continually getting trapped in local minima.

4.6 Methods of Generating and Exchanging Genetic Material

Different search techniques utilise different methods of generating genetic material. The GA uses a crossover operator to exchange structural information between individuals, with Ant Colony Optimisation (ACO) using a local search technique. The traditional IA, however, uses no such method to exchange genetic material, only two mutation operators (hyper- and hypermacromutation), resulting in different degrees of mutation. Table 4.11 lists the levels of success achieved when combining the traditional mutation operators with those used in other search techniques. With the intention of improving SR and reducing μ_{FE}, the IA has been coupled with a hypermutation operator and crossover (HC), MS [118] and local search (LS), with figures quoted, obtained using the optimum parameters for a $n_{ind} = 10$.

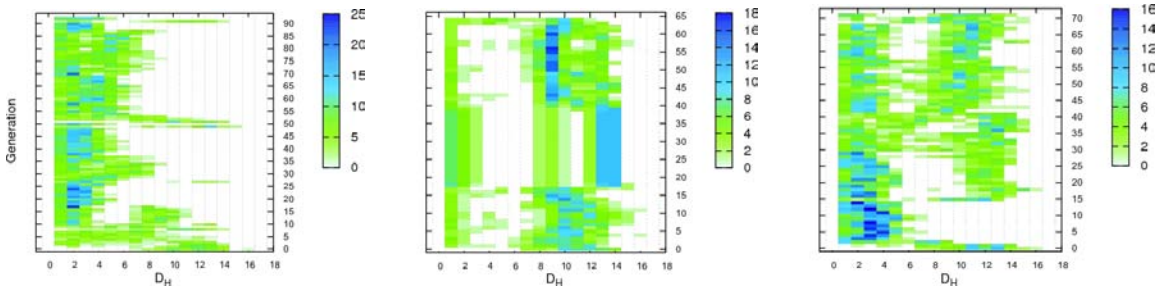
Across the three mutation schemes, HC observes the largest fluctuation in SR. It was shown in sections 4.4.1 and 4.5, that a balance of introducing new genetic material between one and many loci proves successful in GM structure determination. However, by replacing hypermacromutation with the crossover operator (two operators that are good at jumping from one region of a PES to another), the reduction in new genetic material poses a problem for the IA.

ID	SR_{HC}	AFE_{HC}	SR_{LS}	AFE_{LS}	SR_{MS}	AFE_{MS}
L1	61.00	1120171.83	40.00	465828.15	96.00	530815.91
L2	57.00	1083405.17	32.00	301168.21	87.00	667642.56
L3	52.00	1263662.11	35.00	384010.34	69.00	764430.72
L4	41.00	1148947.95	25.00	390071.24	74.00	905922.85
L5	20.00	1264131.20	27.00	521143.70	75.00	868675.17
L6	53.00	999421.00	32.00	420899.68	76.00	740488.52
L7	47.00	1216283.93	45.00	443687.37	92.00	732982.14
L8	25.00	1088644.40	39.00	360772.56	59.00	1040639.01
L9	23.00	1289367.56	40.00	407077.25	43.00	1006285.55
L10	31.00	1115768.64	53.00	346620.11	83.00	842758.20
L11	22.00	1293961.59	34.00	450993.41	60.00	997695.13
L12	46.00	1291774.54	62.00	394022.32	92.00	654755.89
L13	50.00	1180132.50	31.00	412367.25	85.00	692178.61
L14	35.00	1034830.57	30.00	340026.86	64.00	1068114.96
L15	52.00	1107875.25	42.00	488804.73	86.00	703046.36
L16	39.00	1003511.97	35.00	397342.05	61.00	1033721.18
L17	69.00	1097133.94	54.00	364246.42	94.00	630448.85
L18	36.00	895165.52	42.00	324864.35	80.00	593099.50
L19	14.00	1459164.35	39.00	395839.25	30.00	1194287.00
L20	7.00	993965.57	21.00	404071.80	19.00	1278810.15
L21	11.00	1543178.09	39.00	399334.87	38.00	1048867.34
L22	72.00	1092377.12	50.00	400940.62	94.00	539208.84
L23	66.00	1122720.13	25.00	408555.72	75.00	771504.10
L24	73.00	1216172.94	65.00	407129.89	100.00	498804.38
L25	69.00	1124075.55	60.00	416610.13	97.00	460607.69
L26	92.00	831637.52	57.00	381864.47	100.00	360436.25
L27	100.00	333113.38	49.00	422002.79	100.00	194922.28
L28	69.00	997774.31	65.00	371220.21	98.00	428049.30
L29	30.00	1140912.30	54.00	375525.35	62.00	804661.12
L30	24.00	944194.54	66.00	358686.34	65.00	943049.12
L31	94.00	619263.52	87.00	303841.73	100.00	264879.66
L32	94.00	491849.23	48.00	388413.20	100.00	136741.55
L33	92.00	752648.96	56.00	460240.48	100.00	378748.84
L34	84.00	920868.38	59.00	356777.98	100.00	262025.95
L35	59.00	1021379.67	63.00	383807.63	90.00	449144.54
L36	81.00	921977.11	65.00	351807.15	100.00	371398.02
L37	71.00	920141.43	57.00	427254.21	99.00	452320.85
L38	33.00	891857.72	59.00	370118.10	49.00	993913.65
L39	52.00	837316.73	55.00	319452.87	82.00	913414.46
L40	14.00	1153580.57	55.00	337939.81	54.00	860352.18
L41	29.00	1133014.55	64.00	351078.04	59.00	1041626.44
L42	73.00	838022.53	65.00	426284.98	100.00	342778.83
L43	28.00	1237300.57	46.00	385209.84	56.00	855474.35
L44	23.00	953578.69	54.00	376608.22	59.00	925593.83
L45	55.00	1057153.56	54.00	303164.85	79.00	801787.17
L46	35.00	1120161.54	61.00	335574.47	66.00	1015532.75
L47	95.00	695429.53	85.00	375098.68	100.00	222622.16
L48	87.00	947293.71	76.00	330669.19	100.00	239812.04

Table 4.11: SRs using the IA coupled with different mutation schemes, hyper-macro-mutate and crossover (HC), local search (LS) and the mixed operator (MS).

The LS, showing comparable results to the HC for some sequences, is more successful for others. The approach of the LS is different from both the traditional operators and the HC, in that a mutant can experience numerous point mutations when subjected to each of the mutation operators. Both operators thoroughly search local regions of the PES for a given individual, with more consistency being achieved with regard to the magnitudes of SR and μ_{FE} when compared to HC.

By combining all mutation schemes (both traditional and new) in a probability based operator, the MS exhibits performance near to that seen for the traditional operators. Searches performed have access to all of the mutation techniques, from single point mutations to crossover.



(a) (b) (c)

Figure 4.26: (a) The diversity within a population as a function of generation with respect pairwise D_H for individuals in the population for the (a) hyper-mutation and crossover, (b) local search and (c) mixed strategy operators. Generations shown result in the lowest sub-optimal conformation found.

Figure 4.26 illustrates how the diversity of the population fluctuates as a function of generation for the three mutation schemes. It was discussed in section 4.5 how the hypermutation operator can help to drive down the energy of low fitness conformations. Although the behaviour exhibited for the HC scheme is similar to that seen for the traditional operators, crossover only allows favourable regions of local structure to dominate populations, without performing a mutation. This seems to be counter productive for the IA. Performing only single point mutations, preventing significant change to a conformation vector, results in local minimum trapping and poor success

rates.

By performing only multiple point mutations per LS operation, coupled with small population sizes (seen to be beneficial to the IA with the traditional operators), seems to result in branching of the population diversity (with respect to D_H). As the LS mutation operators produce either numerous point mutations from N to C terminus or within a range of loci (point mutation neighbourhood and macromutation neighbourhood search, respectively), too much disruption to the conformation may lead to the GM not being readily discovered.

The branching effect is also observed for the MS. This may be due to the use of LS, contributing to a large population diversity. As the traditional operators are also accessible, the level of success is more comparable to the use of those alone. As the MS is also followed by a crossover phase, different degrees of mutation can occur, as well as basin hopping. It may be the use of crossover that limits the success and increases the magnitude of μ_{FE} for this mutation scheme for small population sizes, as observed for the HC.

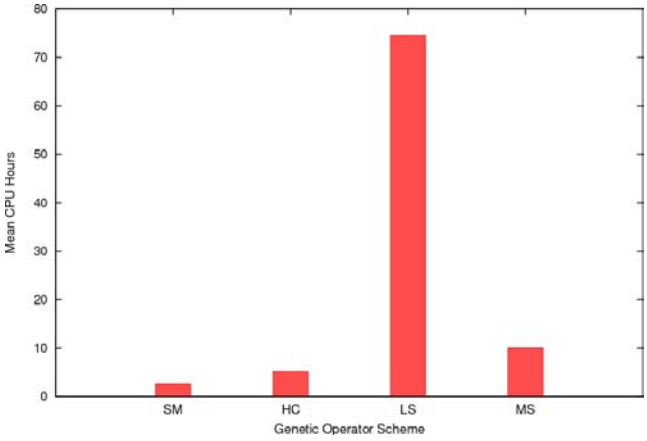


Figure 4.27: Mean CPU times for each genetic operator scheme.

Figure 4.27 shows the mean number of CPU hours used to determine the SRs quoted in table 4.11, and for the standard mutation operators, averaged over all low degeneracy sequences. As successful runs are less likely to run for g_{max} generations, their mean

CPU time is expected to be reduced. The plot confirms that the standard mutation operators (the most successful) ran on average for the least number of CPU hours. Slightly larger magnitudes are observed for HC and MS, reflecting their reduced SR. However, as both mutation phases present in the LS operator involve making numerous point mutations, the CPU suffers as a result, being the least efficient of all four mutation schemes. If LS could be improved (in terms of SR), the poor efficiency would not make the scheme a viable alternative to the standard mutation operators.

4.6.1 Preferential Global Minima

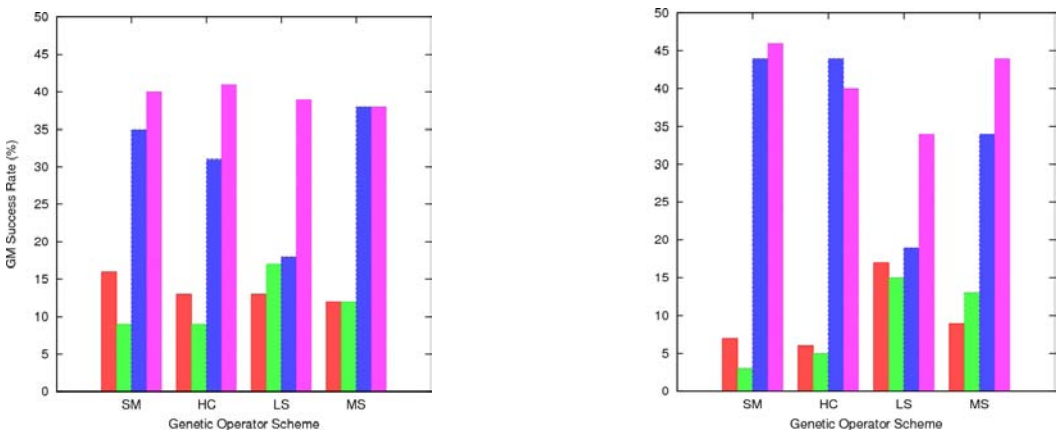


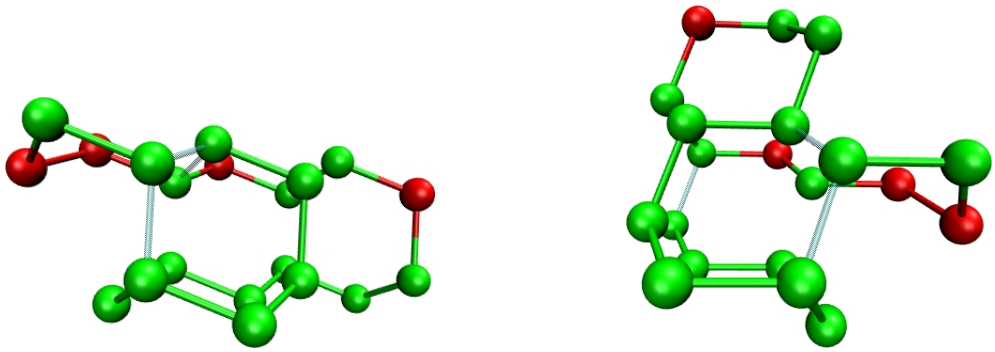
Figure 4.28: The GM distributions for mirrored sequences (a) L31 and (b) L47. Minima represented by red and green are reflectively related and likewise are the blue and magenta minima.

As each sequence set involves a degree of degeneracy, the frequency of each degenerate GM contributes to the SR. Statistically, one would expect each minimum to be found an equal number of times. However, considering sequence L31 (and its mirror L47), this is not the case. This particular sequence and its mirror have been chosen due to their high SRs across the mutation schemes.

Figure 4.28 shows the frequency of each GM (including mirrors) for sequences L31 and L47. The first two GMs (red and green, mirror images) illustrate a much lower frequency than the final two (also mirror images). With one mirror set being heavily

favoured over the other (in both the forwards and reverse sequence definition), the structures are expected to drastically differ in terms of topological contact arrangement.

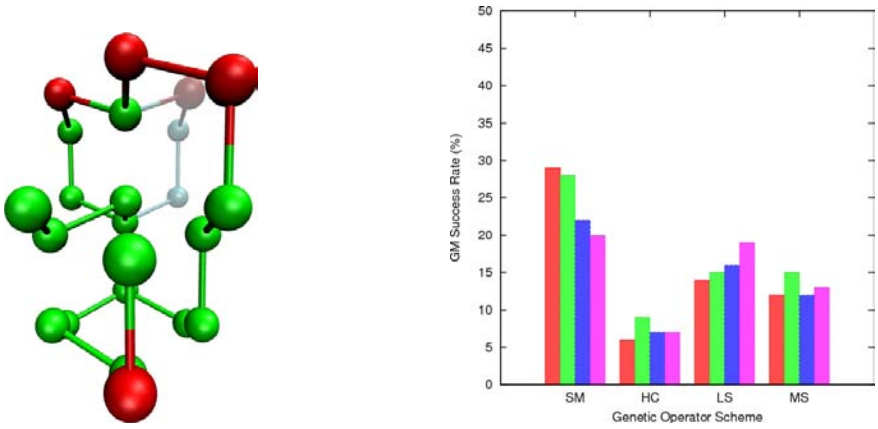
Figure 4.29 shows the two GM conformations (not including mirror images) for sequences L31 and L47. The GM least frequently found involves both terminal beads making a topological contact with each other and two other beads (each terminal bead contributing two topological contacts). For the more dominating conformation, the two terminal beads do not interact with each other, with the terminal H of the long hydrophobic tail producing only a single contact. It was observed for sequences L1 - L23, that higher SRs can be achieved with no beads contributing more than a single topological contact. It seems however, that the positions of the beads involved in making the double contacts help determine levels of success for each GM.



(a) (b)

Figure 4.29: The two GM conformations (not including mirror images) for sequences L31 and L47 showing the (a) lower frequency and (b) higher frequency conformations. Contacts produced using terminal beads are shown in transparent cyan.

In contrast, sequence L30 and its mirror L41 (having two GMs, not including mirror images), have GMs that differ by a single kink in the chain. The kink does not affect the arrangement of topological contacts, and as a result, a near-statistical distribution in frequencies of the GM is observed (see figure 4.30).



(a) (b)

Figure 4.30: (a) The two GM conformations (not including mirror images) for sequences L30 and L41, with the structural diversity shown in transparent cyan. (b) The GM distributions for mirrored sequences L30 and L41. Minima represented by red and green are reflectively related and likewise are the blue and magenta minima.

4.7 Conclusions

The work presented here demonstrates that the IA is successful in determining GM conformations with regard to the HPLBM on the diamond lattice. By investigating parameter space, a set of optimum values were determined for various population sizes across all sequences (of both high and low degeneracy). This, in combination with the use of a common parameter set, allowed an optimum population size to be determined. By using the optimum population size, combined with its optimum parameter set, the most efficient results were obtained for both sequence sets, averaged over one hundred runs.

With the optimum population size being one twentieth of that seen for the GA, the maximum number of generations was increased in order to illustrate the potential of the IA for this model. By increasing the number of generations to 20,000, high to perfect success rates were observed for all sequences, something not seen for the GA. By considering the worst case (in terms of success), for both sequence sets, profiles were taken to illustrate the strengths and weaknesses of the search technique under

both successful and unsuccessful circumstances.

It has been shown how the highest, lowest and mean energies fluctuate as generations pass, with major contributions from birthing phases. The introduction of randomly generated, high energy conformations into a population, shows a dramatic increase in the mean. Although these high energy conformations are not favourable, if one undergoes a favourable mutation (to result in an energy decrease), regions of its conformation are able to dominate future populations. The mean energy of post-birthing populations is shown to decrease dramatically, with the increase in use of hypermacromutations being seen. Hypermutations have been shown to help decrease the energy of already compact, low energy conformations, whereas hypermacromutations have seen to be most useful in driving down the energy of mid to high mean energy populations. This has been attributed to the ability of the hypermacromutation operator at performing multiple point mutations within a range of sequence loci, with the hypermutation operator performing only a single point mutation. The single point mutation results in a new valid conformation being more likely, due to the reduced disruption to the original conformation when compared to the hypermacromutation operator.

The search for the GM is heavily governed by the formation of the initial topological contact (upon chain growth). It has been shown that by producing a contact too early, in order to recover, the search must gradually unfold the conformation (via mutation) and maintain sufficient contacts to save the individual from being removed from the population. If this is not possible, in order to drastically change the direction of the search, and in some cases, prevent local minimum trapping, the birthing phase is vital to this search technique.

The levels of diversity within a population have been shown to be attributed to sections of dominating local structure. Diversity constantly fluctuates due to the presence of mutants within a population. Diversity is seen to surge when a birthing phase

occurs, as the new-born individuals are randomly generated, with no bias towards to existing population members. A population dominated by new-born individuals combined with a small number of mutants, results a population diversity maximum. Large diversities are observed in the absence of birthing, due to a change in the dominant region of local structure for a population. This may occur due to a series of favourable mutations, producing more energetically favourable conformations. Population diversity shows a minimum when a common region of local structure to the population is located towards the centre of the conformation vector and few terminal mutations occur. Central mutations are limited, in that successful mutations about these central loci, producing self-avoiding conformations, are difficult to achieve.

Although D_H is a simple, quick measure of conformation vector similarity, RMSD provides more information with regard to 3D conformation. The short-comings of D_H have been shown with regard to how it can be misleading with respect to individuals within a population and the GM. As the level of diversity within a population can be characterised with respect to both D_H and RMSD, populations that frequently exhibit too much diversity can prevent the search from thoroughly exploring regions of the PES, that would otherwise invoke a birthing phase. This results in the search getting trapped in local minimum from which it cannot escape.

Although the use of other mutation schemes yielded a decrease in SR, it has been shown that weighting different mutation operators provided promising results when compared to the traditional methods.

Chapter 5

Dynamic Lattice Bead Model

Whilst protein models on a regular lattice allow for comparisons of algorithm efficiency within relatively short timescales, the regularity of the lattice cannot give a true insight into the preferable backbone configurations of real proteins. The DLM, as described in section 2.4.2, utilises the actual backbone torsional angles, ϕ, ψ and ω, to determine the lattice on which the protein lies.

PDB ID	Length	Energy	Residue Sequence
1AL1	12	*-1.5624*	ELLKKLLEELKG
1A1P	13	*-1.5178*	ICVVQDWGHHRCT
1AKG	16	*-1.1730*	GCCSLPPCALSNPDYC
1L2Y	20	*-0.8784*	NLYIQWLKDGGPSSGRPPPS
1D9J	20	*-1.3445*	KWKLFKKIGIGKFLHSAKKF
1B19:A	21	*-3.5435*	GIVEQCCTSICSLYQLENYCN
1G04	26	*-1.2427*	GNDYEDRYYRENMYRYPNQVYYRPVC
1ANP	28	-2.5157	SLDRSSCFTGSLDSIRAQSGLGCNSFRY
1AML	40	-4.9146	DAEFRHDSGYEVHHQKLVFFAEDVGSNKGAIIGLMVGGVV
1QHK	47	-3.8985	GNFYAVRKGRETGIYNTWNECKNQVDGYGGAIYKKFNSYEQAKSFLG

Table 5.1: DLM sequences [104] and corresponding lowest energies. Energies shown in italics are the GM energies found as a result of a branch and bound systematic search, otherwise they are the lowest energies found by evolutionary techniques [117].

Table 5.1 lists the PDB IDs, sequences [104] and lowest energies obtained [117], for which the performance and ability of search techniques is assessed with regard to the DLM.

Table 5.2 lists the SR, μ_{FE}, n_{uniq} and μ_g for the DLM when searched using the IA, with optimal parameters determined in section 4.1, for $n_{ind} = 10$. Each run was given

g_{max} = 20,000, with all values averaged over one hundred runs. Mid to high success is achieved for sequences up to twenty six beads in length. The IA for sequences 1ANP and 1QHK failed to determine the GM conformation and the lowest energy conformation, respectively. However, sequence 1AML was successfully searched for the lowest energy conformation determined by evolutionary techniques ([117]).

In order to determine why perfect success was not achieved, sequence 1B19:A, the only sequence for which the GM is known (via branch and bound) and that didn't result in perfect success, will be analysed using the data obtained from the method described in section 2.6. In order to gain a complete picture, a successful run, followed by an unsuccessful run will be considered and compared.

PDB ID	F*	SR	μ_{FE}	n_{uniq}	μ_g
1AL1	1.56250	100.00	500.45	1	5.06
1A1P	1.51775	100.00	20601.26	2	217.68
1AKG	1.17301	100.00	110595.70	1	1172.60
1L2Y	0.87846	100.00	114240.86	2	1152.30
1D9J	1.34457	100.00	46744.93	1	475.70
1B19:A	3.54363	68.00	491752.13	1	6136.77
1G04	1.24277	100.00	42276.17	1	433.49
1ANP	2.37581	1.00	575035.00	1	6095.00
1AML	4.91472	19.00	907232.78	1	9373.26
1QHK	3.78268	2.00	1424157.50	1	14522.00

Table 5.2: SRs, fitnesses, μ_{AFE}, n_{uniq} and μ_g for the IA for the DLM sequences in table 5.1 using the DLM.

Figure 5.1 consists of two energy profiles illustrating how the energies fluctuate with respect to highest, mean and lowest energies in a population, and the GM as a function of generation. Figure 5.1(a) initially shows a gradual decrease in the highest, mean and lowest energies for a generation. However, a surge in highest energy is observed in generations 17, 33, 50 and 57. This may arise for several reasons: either the mutations performed are unfavourable in terms of energy (not necessarily in terms of the GM conformation) or a birthing phase has been invoked, due to the ageing operator not

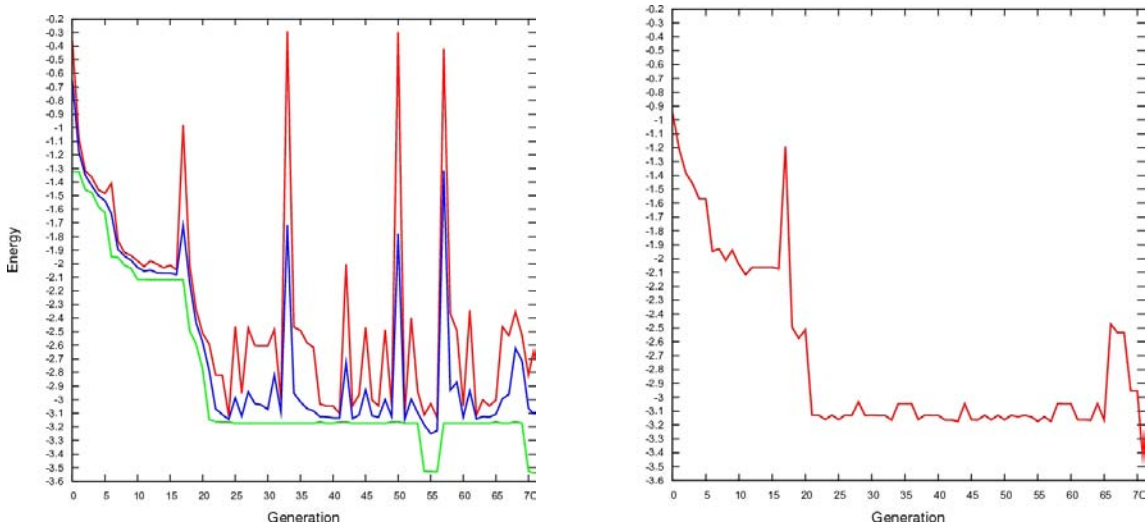
being able to retain the population size as a result of poor quality mutations. The effect on the mean suggests either numerous unfavourable mutations were performed (i.e. not just one), or a birthing phase created numerous individuals.

In terms of the lowest energy per generation, this seems to retain a steady decrease initially, settling after twenty one generations. A trough is observed between generations 54 and 56 inclusive, illustrating a decrease in the lowest energy present for those generations.

Figure 5.1(b) illustrates how the energy path of the resulting GM individual, initially sees a steady decrease, with a surge at generation 17. This supports the idea that perhaps for this generation, a birthing phase is invoked to a lesser extent, and that a number of unfavourable mutations were performed. As this surge is seen for the resulting GM pathway, this energy increase only contributes to the shift in the mean, as the energy for this conformation is still not the lowest for that generation.

What is apparent, is that although an ageing operator is utilised by the IA, the resulting GM individual is never produced from a birthing phase, but from a successful mutation from the previous generation. As the surge in energy is not witnessed for this individual at generations 33, 50 and 57, the individual only contributes to the mean, implying that other conformations of much higher energy are responsible for the peak in mean energy. A smaller increase in energy is witnessed for generation 68 in both profiles, indicating that this individual is responsible for the peak, and that unfolding of this conformation has taken place to some degree as a result of mutation. Smaller increases in energy throughout the calculation indicate small regions of unfolding of the conformation.

Figure 5.2 illustrates how the number of mutations and the number of births fluctuate as a function of generation. Figure 5.2(a) demonstrates that the number of hyper- and macromutations performed does not generally increase or decrease over time. This suggests that both mutations contribute to new genetic material in a uniform fashion

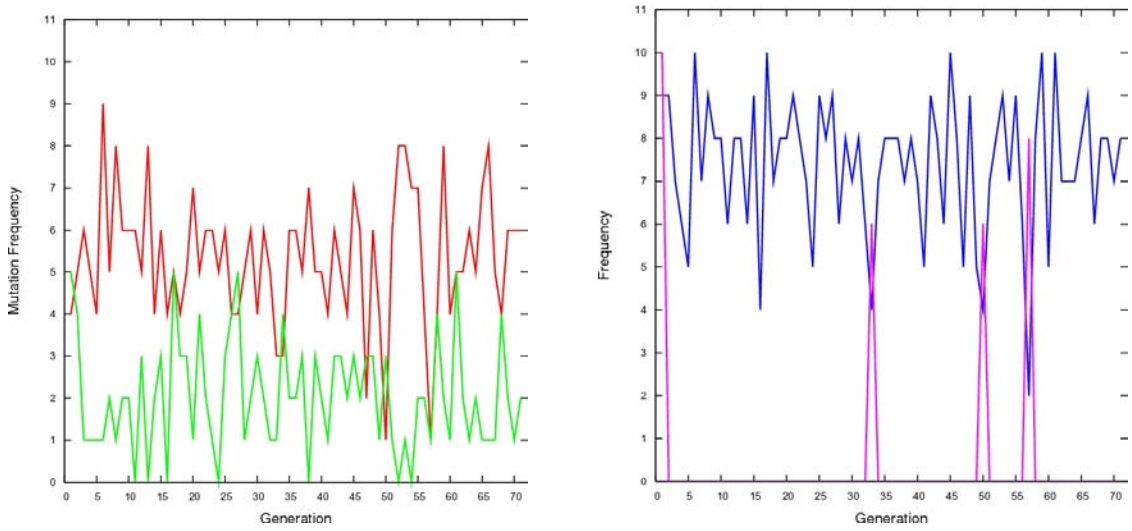


(a) (b)

Figure 5.1: Energy profiles for the IA, showing energy as a function of generation (a) with the lowest energy shown in green, mean shown in blue and highest shown in red. (b) for the GM conformation.

over time. However, what it does show is that hypermutations are more likely to produce higher energy conformations for a generation, as generally more are present per population. This could be expected, as hypermutations only result in $D_H = 1$ between parent and child, reducing the risk of producing invalid conformations. The presence of hypermacromutations does however, allow the search to jump between regions of a PES, as point mutations are performed across a range of loci of the conformation vector.

Figure 5.2(b) shows how the birthing phase was invoked for generations 33, 50 and 56, producing 6, 6 and 8 new conformations, respectively. Comparing these generations to how the energy decreases over the initial handful of generations, birthing frequently produces higher energy, possibly less compact conformations. This explains the sudden increase in mean and highest energies for those generations in figure 5.1(a), and also why the lowest energy conformation did not see an energy increase to the same degree. In contrast, generation 17 also saw a surge in energy for the population. As the birthing phase was not initialised for this generation, this can either be attributed to



(a) (b)

Figure 5.2: Mutation profiles for the IA with showing how the number of (a) hyper (red) and macro (green) mutations and (b) the total number of mutations (blue) and the number of births (magenta) fluctuate as a function of generation.

a poor mutating operation or that the lower energy conformations were simply too old and were therefore removed. Only generations 6, 17, 33, 50 and 57 see all existing individuals undergoing a mutation, all of which lead to an increase in energy for the highest energy conformation.

The trough observed between generations 54 and 56, may be attributed to the increase in the number of hypermutations performed and a simultaneous decrease in the number of macromutations performed. By mutating only a single locus, the probability of producing a lower energy conformation is increased, as minimal disruption is made to the existing conformation.

Generation 42 also experienced a large increase in energy. As nine mutations were performed (six hyper- and three hypermacromutations), six conformations all shared the same parent, with one being the parent (not a mutant). The five mutated children all exhibited an increase in energy, thereby giving rise to an increase in highest and mean energies for that generation. Generation 42 is shown in table 5.3. A drastic change in the mean and highest energies is observed from generation 66. The data

presented so far is unable to explain this feature.

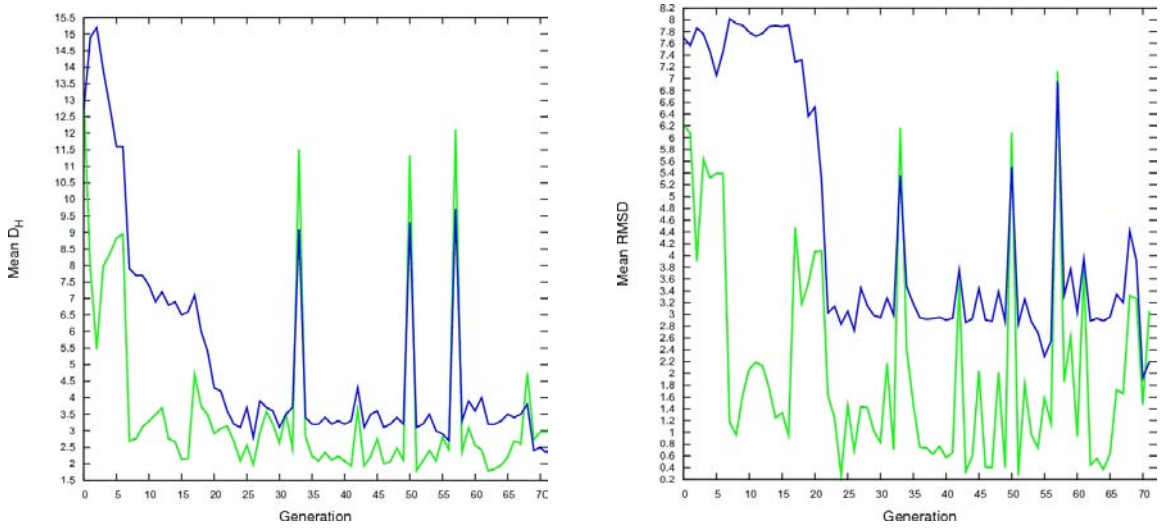
Population Position	Tag	Conformation Vector	Fitness
Generation: 42			
0)	41-2-H	230120231111002100130	3.16431
1)	41-1	000120231111002100130	3.16285
2)	41-5-M	200120231111002100130	3.16152
3)	41-6-H	130120231111002100130	3.14435
4)	41-5-H	230120231111002100110	3.13189
5)	41-1-H	002120231111002100130	2.57896
6)	41-1-M	000120231111002101021	2.48262
7)	41-1-H	000120231111002101130	2.36629
8)	41-1-M	000120231230102101030	2.07111
9)	41-1-H	000120231011002100130	2.00499

Table 5.3: The forty second generation of the profile created for the successful case of sequence 1B19. Five individuals share the same parent, with another individual being the unmutated parent from generation 41. Fitness is quoted here rather than energy, as it is used as a measure of individual quality. Fitness is simply the negative of the energy.

Figure 5.3 shows how both the mean D_H and mean RMSD fluctuate as a function of generation. In figure 5.3(a), a large difference exists between μ_{DH}^{P} and μ_{DH}^{L}. With respect to the individuals themselves, the sudden decrease that is witnessed over the first two generations, implies that the individuals are becoming more alike in terms of their conformation vectors. This arises due to a dominant segment of local structure being present in the central positions of the conformation vector. By generation 7, the fluctuation in pairwise mean D_H has settled, resulting in smaller variations due to many mutations occurring at the termini of the conformation vectors.

Due to the birthing phase present in generations 33, 50 and 57, peaks are witnessed, affecting both μ_{DH}^{P} and μ_{DH}^{L}. Omitting these generations, the most noticeable fluctuations are seen again for generations 6, 17 and 42, where all population members were mutated. For generations 17 and 42, hypermacromutations have been performed (5 and 3, respectively), with a majority (4 and 2, respectively) resulting in high energy conformations in comparison to the hypermutations performed. Out of the four hy-

permacromutations resulting in higher energy conformations in generation 17, three of them produced major changes to the centre conformation vector. This suggests that the search withdrew from the folding funnel for that individual and entered another upon mutation. This means that, although hypermacromutations are good at jumping from one region of the PES to another (depending on the length of the segment to be mutated and the position it is in the chain), hypermutations are better at producing improvements in energy due to smaller disruption being caused to the conformation.



(a) (b)

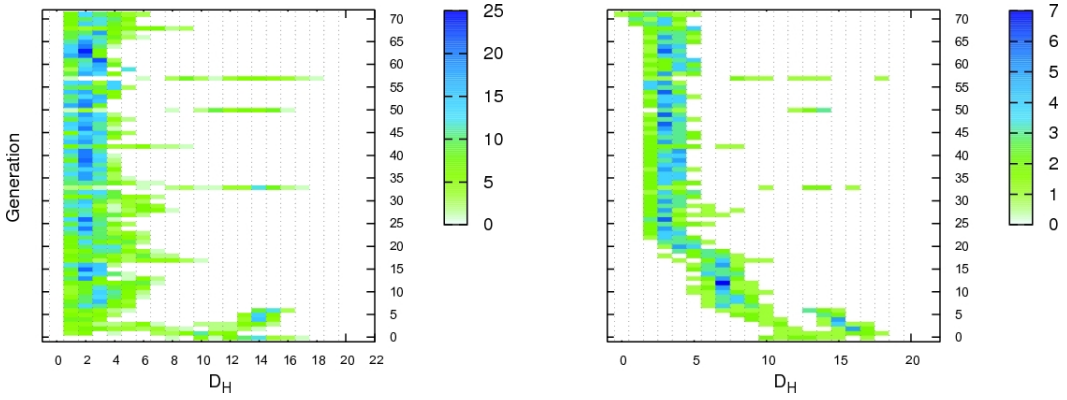
Figure 5.3: Variation of (a) mean D_H and (b) mean RMSD as a function of generation with respect to other individuals in a population (green) and the GM (blue).

μ_{DH}^{L} is larger than μ_{DH}^{P} until the final stages of the calculation, due to the favourable segment of local structure present in the GM, situated at the centre of the conformation vector, not being present in the individuals of other populations. From generation 68, the individuals begin to exhibit large changes to the centre of the conformation vector (new genetic material), eventually adopting that present in the GM. This is reflected in the sudden increase in μ_{DH}^{P} in generation 68, with μ_{DH}^{L} witnessing a decrease, one resulting in its magnitude being less than for the pairwise case.

Figure 5.3(b) shows an initial large deviation between the pairwise RMSD measure and the RMSD with respect to the GM across the first twenty one generations. This

is attributed to the configuration of the conformation centre and how it more closely resembles that of the GM after generation 21. The decrease in μ^P_{RMSD} seems to also fall at generation 21, as the newly adopted, more favourable region of local structure present at the centre of the conformation begins to dominate the population. Table 5.4 illustrates the change in the central local structure and possibly signifies the transition from one funnel of the PES into another.

The surges in mean RMSD in generations 33, 50 and 57 are observed due to the mutation and birthing combination, with the general fluctuation in mean RMSD being caused by mutations of the loci more towards the termini of the conformation. As seen for the mean D_H profile, generation 68 sees an increase in mean RMSD with regard to both the pairwise and the GM measures, reflecting a modification to the central configuration of the conformation. This modification results in a quick convergence to the GM



(a) (b)

Figure 5.4: Population diversity mappings for each individual with respect to (a) each other in a pairwise manner and (b) the GM conformation.

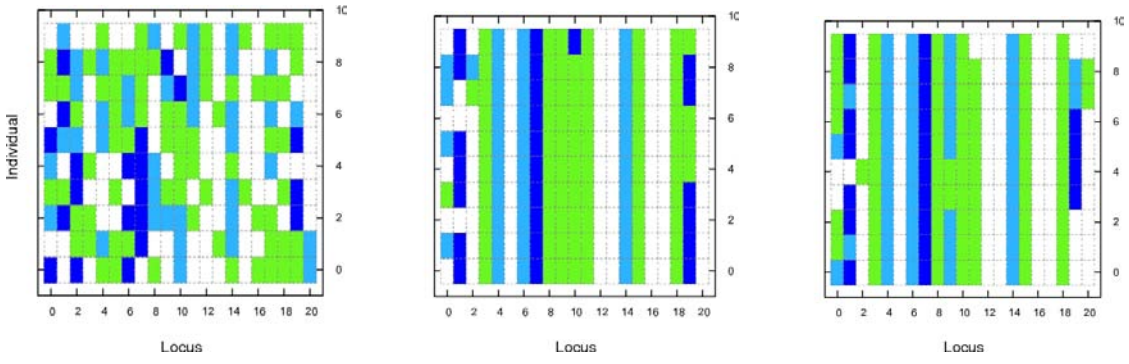
Figure 5.4 illustrates how the population diversity fluctuates with respect to the individuals of a population in a pairwise manner and with respect to the GM. Figure 5.4(a) illustrates that individuals are related by small D_H values. The diversity is large for the first six generations, due to the population exhibiting differences at the centre

Population Position	Tag	Conformation Vector	Fitness
Generation: 20			
0)	19-0-H	200120231001102100120	2.77614
1)	19-0	200120231001002100120	2.59145
2)	19-0-H	200120231001002100121	2.58085
3)	19-1	200120231232002100120	2.57481
4)	19-1-H	200120231212002100120	2.57167
5)	19-1-H	200120231232002100121	2.56421
6)	19-2-H	200120231131002100120	2.54035
7)	19-3-H	200120231131002100121	2.52975
8)	19-0-H	200120231001002100110	2.52922
9)	19-1-M	200120231232002100110	2.51258
Generation: 21			
0)	20-7-H	200120231111002100121	3.14608
1)	20-9-M	200120231111002100110	3.1291
2)	20-0-H	230120231001102100120	2.78874
3)	20-0	200120231001102100120	2.77614
4)	20-2-H	200120231001102100121	2.76554
5)	20-8-M	200120231001102100110	2.71391
6)	20-0-H	200120231001102100130	2.70454
7)	20-8-M	200120231001002100100	2.63425
8)	20-9-M	200120231232002100100	2.61761
9)	20-2-H	200120231001002100120	2.59145
Generation: 22			
0)	21-1-M	200120231111002100130	3.16152
1)	21-0	200120231111002100121	3.14608
2)	21-1-H	230120231111002100110	3.13189
3)	21-1-H	000120231111002100110	3.13043
4)	21-1	200120231111002100110	3.1291
5)	21-1-M	330120231111002100110	3.11137
6)	21-0-H	100120231111002100121	3.09866
7)	21-0-H	201120231111002100121	2.9982
8)	21-0-H	200120231211002100121	2.94355
9)	21-3-H	200120231001102100100	2.81894

Table 5.4: Generations 20-22, illustrating how in one generation, a favourable region of local structure can dominate a population. Fitness is quoted here rather than energy, as it is used as a measure of individual quality. Fitness is simply the negative of the energy.

of the high energy conformation vectors. Diversity surges are witnessed at generations 17, 42, 50 and 57 due to all veteran individuals being mutated and the presence of a birthing phase. The high density regions for the remaining generations stem from mutations being applied to the loci towards the termini of the conformation vector, with the same conformation vector centre dominating the population. The diversity spread in generation 68 arises from successful mutations at the centre of the conformation vector for some individuals.

Figure 5.4(b) shows a small increase in population diversity with respect to D_H, the population and the GM in generations 25, 27 and 28. In generations 27 and 28, the standard terminal mutations are responsible for the diversity. However, for generation 25, the central schema of the conformation vector exhibit a unique pattern for one individual. The individual has the arrangement present in the GM, but is short lived as it fails to mutate in generation 26 and thus is removed by generation 27. The configuration of the conformation centre for the other individuals of generation 25 is identified as not being that different from the GM. This supports the fact that for the DLM, mutations about the centre of the conformation are difficult to achieve.

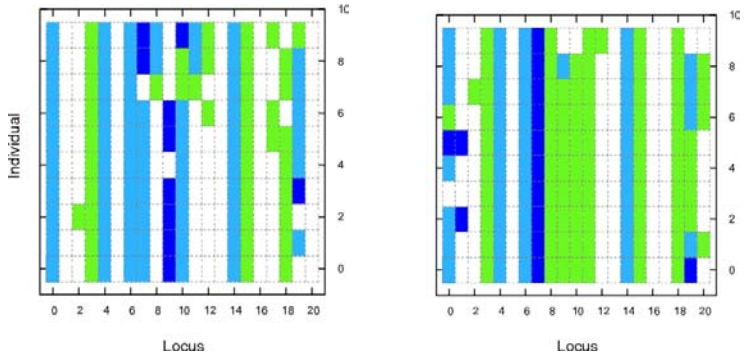


(a) (b) (c)

Figure 5.5: Conformation vectors for the (a) initial (b) half-way and (c) final populations. 0 = white, 1 = green, 2 = cyan and 3 = blue. Individuals are ordered by fitness, with the highest fitness individual shown at position 0.

Figure 5.5 shows the conformation vectors of individuals for various populations

throughout the calculation. The initial population (figure 5.5(a)) indicates the random generation of individuals via the construction phase. Alleles are as random as possible, but governed by the vacancy of residue sites determined by the RGA. It is this randomness that gives rise to the initial large diversity in the population and the large magnitudes of mean D_H and mean RMSD. Figure 5.5(b) illustrates how favourable regions of local structure have propagated through the population by generation 35, with the main contribution to population diversity coming from the alleles close to the terminal chain positions. Figure 5.5(c) represents the final generation (generation 71, where the GM is found) and shows how it shares regions of local structure with generation 35. In the range of loci 3-18, the only variation between some individuals is locus 9. As discussed above, it is the determination of this favourable allele at position 9 that proved a difficult task for the IA. As for generation 35, the reduced diversity in this population with regard to both RMSD and D_H, arises from the alleles adopted near to and including the terminal residues.

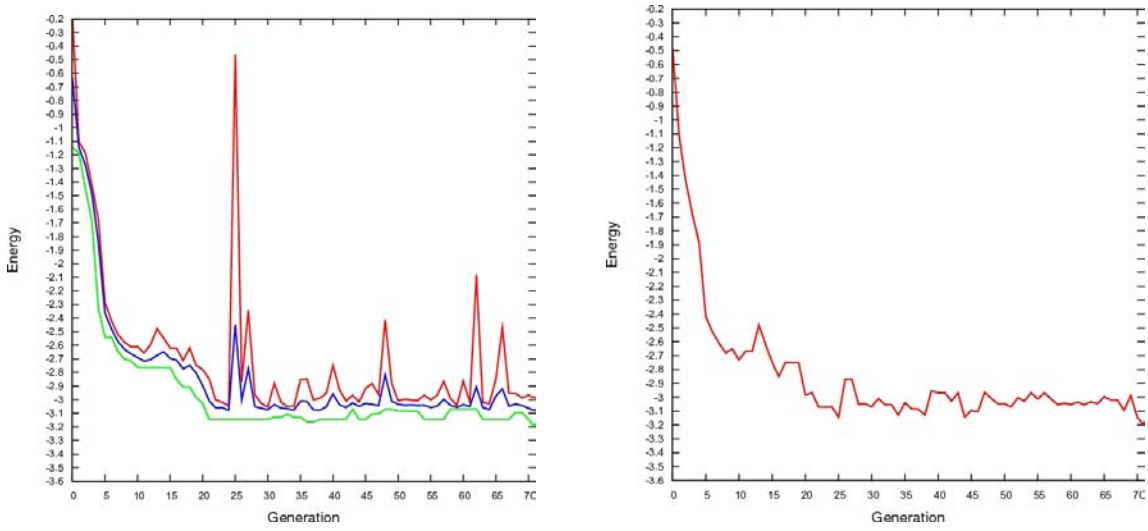


(a) (b)

Figure 5.6: Conformation vectors for generation (a) 17 and (b) 22, illustrating the change in dominant central local structure. 0 = white, 1 = green, 2 = cyan and 3 = blue. Individuals ordered by fitness, with the highest fitness individual shown at position 0.

The unfolding of the conformations present in generation 17, led to an increase in the mean and lowest energy conformations for that generation. By generation 22, a drastic decrease in all energy measures was seen, representing a change in the dominant

central local structure (figure 5.6). As the energy landscape was negotiated efficiently in generation 17, this suggests that possibly the shape of the PES in this region is in fact shallow. In contrast, the dominant central local structure present in generation 22, required a larger number of generations before descending into the GM region. As only a single change to the centre of the conformation allowed the GM to be reached, with lower energies exhibited for the population members, it is possible that this region of the PES is somewhat deeper than that seen for generation 17. The single mutation to the centre of the conformation, which results in rapid GM exporation, suggests that this region may be part of the same broad funnel on the landscape, a sub-optimal well, separated from the GM by a moderate energy barrier.



(a) (b)

Figure 5.7: Energy profiles for the IA with showing energy as a function of generation: (a) with the lowest energy shown in green, mean shown in blue and highest shown in red; (b) the lowest energy conformation for the failed case.

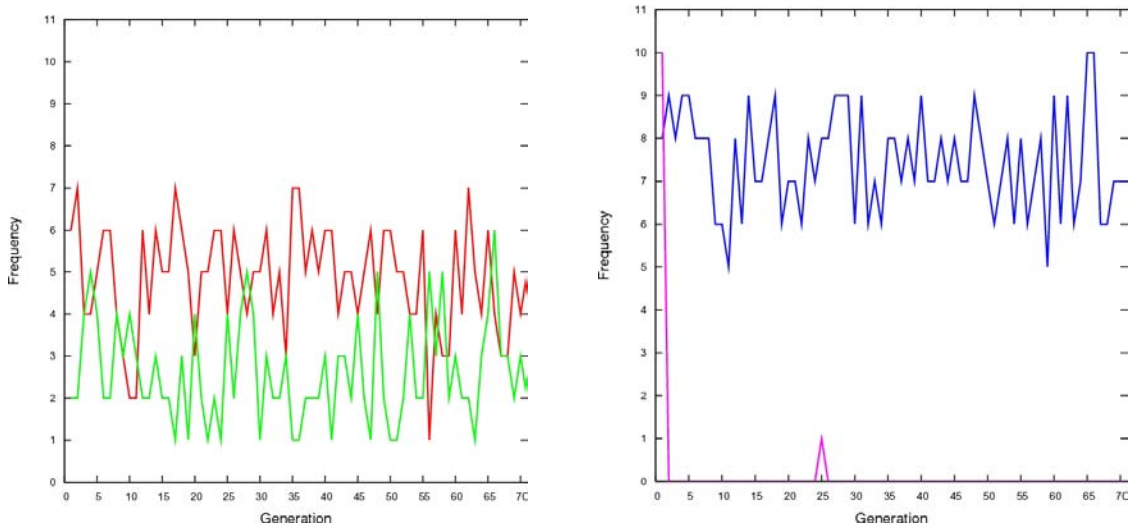
In order to compare the successful run analysed previously, analysis of an unsuccessful run is required. Figure 5.7 illustrates how the highest, mean and lowest energies fluctuate as a function of generation, as well as how the energy changes for the lowest energy conformation obtained. With respect to the successful case, the initial decrease in energy for all measures, up to generation 11 exhibits a steep gradient. In the early

stages of a calculation, hypermacromutations tend to dominate, as the conformations exhibit relatively high energies. In order to drive the energy of the system down quickly, performing many point mutations in a single operation can accomplish this.

Although the lowest energy for a population remains fairly stable, as with the successful case, surges are witnessed in the mean and highest energies recorded, due to mutations and/or births of new individuals. A sudden increase in the highest energy is witnessed for generation 25, causing an increase in the mean. As the energy value closely resembles that of the starting energies, a birth phase is suspected to have occurred. A single new individual is released into the population (figure 5.8(b)). Not only does birthing maintain the population size, it also introduces new genetic material into the population. This allows the search space to be opened up, possibly allowing new areas of the potential energy surface to be discovered. In order for this to happen, the new individuals must either be able to compete with current population members in terms of energy, or there must be enough births to dominate the population. The sudden decrease in lowest and mean energies witnessed for generation 26, illustrates how a birthing phase that introduces only a single, high energy conformation, is quickly removed, due to the more favourable genetic material being present in the population.

Generation 27 also witnesses an erratic increase in the highest and mean energies. Figure 5.8(a) illustrates that four hypermacromutations were performed in this generation, all of which produced conformations that were of higher energy than that of the highest in generation 26. It was shown for the successful case that energy surges were seen when all individuals were mutated. For this generation, all but one were mutated, with four hypermutations contributing to an energy decrease. This suggests that by providing more than a single mutation per operation, there is a greater risk of increasing the conformation energy.

Generations 48, 62 and 66 all witness an increase in the mean and lowest energy. As for generation 27, all generations involve nine mutations, with macro-mutations



(a) (b)

Figure 5.8: Mutation profiles for the IA with showing how the number of (a) hyper (red) and macro (green) mutations and (b) the total number of mutations (blue) and the number of births (magenta) fluctuate as a function of generation for the failed case.

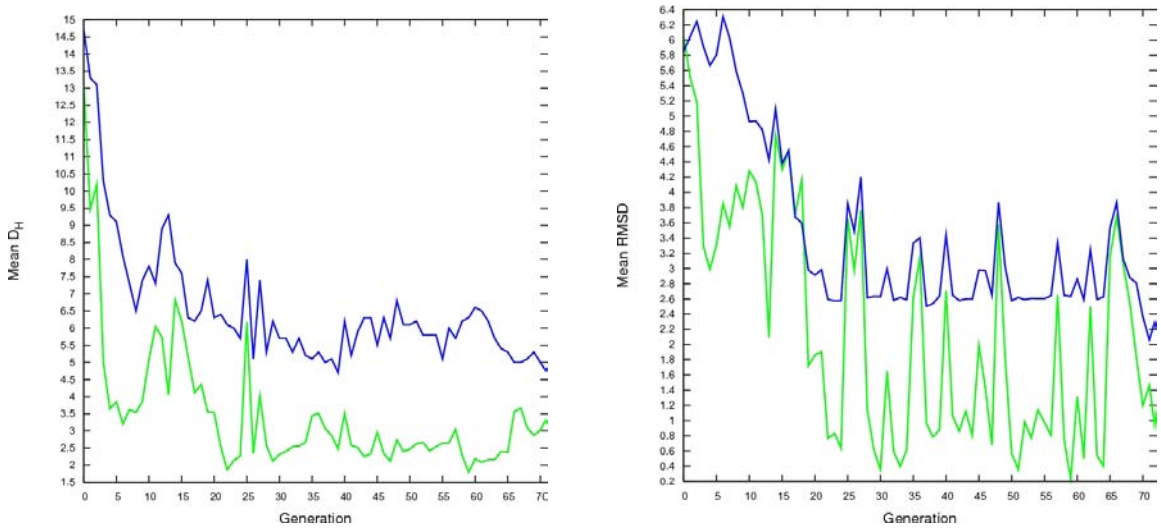
producing the highest energy conformations. As witnessed for the successful case, all individuals are mutated in generation 66 (figure 5.8(b)). In generations 11 and 59, a trough is witnessed for all three energy measures as a result of only five mutations occurring, the lowest for the first 71 generations shown here.

In contrast to the successful case, the failed case presented here does not witness a drastic energy increase with respect to the lowest energy conformation (figure 5.7(b)). The successful case reveals the beginning of a change in population-dominating local structure at the centre of the conformation vector for generations 17 and 66. This suggests that unfolding and re-folding occurred, emerging from one funnel, and entering another. This is not witnessed for the first 71 generations of this unsuccessful case, suggesting that the failure may be attributed to the difficulty encountered in emerging from a sub-optimal folding funnel.

The number of births in the first 71 generations, is one twentieth of that for the successful case. However, in both cases, the introduction of new material, as a result of a birth, failed to shift the domination of local structure within the population. All

mutants are related to the mature individuals, and therefore cannot be attributed to success or failure.

Generations 18 to 22, see a gradual decrease of all energy measures in figure 5.7(a). With respect to the successful case, central conformational changes were witnessed from generation 17 to 22 and 66 to 71. This energy change also reflects a central conformational change. However, the change only occurs at position 9. This change from 1 to 2, corresponds to a transition from a ϕ_T, ψ_T, arrangement, to a ϕ_T, ψ_C arrangement. The central conformational change in the successful case, also results in the same mutation. However, the tenth position (the CYS residue), exhibits a ϕ_T, ψ_T arrangement, as opposed to a ϕ_C, ψ_C arrangement for the unsuccessful case.



(a) (b)

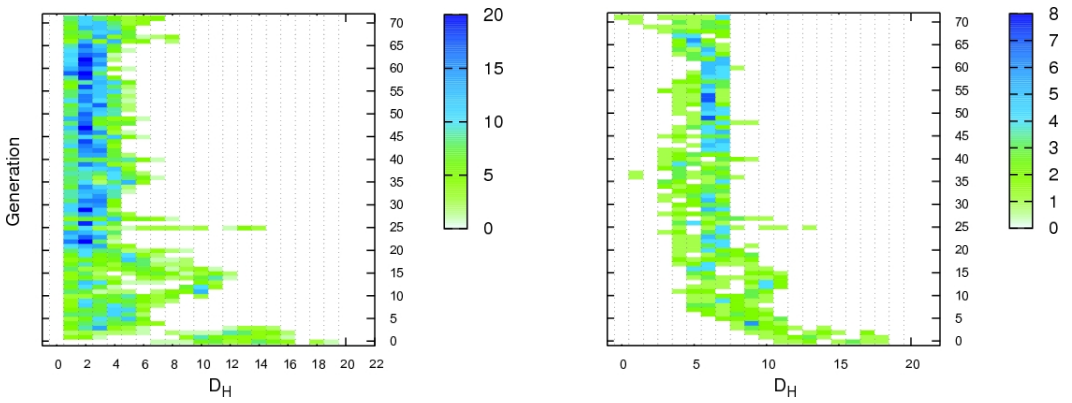
Figure 5.9: (a) Variation of (a) mean D_H and (b) mean RMSD change as a function of generation, with respect to other individuals in a population (green) and the lowest energy conformation (blue).

Figure 5.9(a) shows how the mean D_H fluctuates with regard to the individuals in the population and the lowest energy conformation obtained. Figure 5.9(b) illustrates the information in the same context with regard to RMSD. As witnessed for the successful case, the initial large magnitude of mean D_H for both measures indicates that the individuals are diverse with respect to each other and the lowest energy conforma-

tion, due to the random generation of conformation vectors. As the diversity decreases over the first few generations, the individuals begin to adopt a common region of central local structure. The increase in diversity from generations 7 to 19 reflects large deviations from the central local structure. Another point mutation that dominates the population occurs at position nine in the chain, giving rise to a diversity increase from generation 22 to 26. There is a local maximum in generation 25, due to the introduction of a new individual from birthing. All remaining fluctuations (post generation 26) are due to near terminal mutations, hence the population diversity becomes more stable.

A similar trend is seen for the mean RMSD. However, the relationship in terms of 3D conformation is more predominant, due to a small D_H change at the centre of the conformation, may in fact reflect a large conformational change in terms of 3D space. The more steady decrease in mean RMSD up to generation 20, with regard to the lowest energy conformation obtained, reflects the population settling on a common central region of local structure, similar to that found for the lowest energy conformation. However, the peaks at generations 14 and 15 reflect the intermediate conformations that were searched before the change in common central local structure for the population was complete. The peak present for generation 25 is representative of the invoked birthing phase, with the larger peak at generation 27 reflecting a mutation change in the ninth position for three individuals.

Figure 5.10(a) illustrates how the population diversity fluctuates with respect to each individual in the population for the failed case. The initial large diversity reflected in the large D_H values, arises due to the initialisation of the population with randomly selected alleles. The population diversity drops as individuals begin to adopt low energy contributing regions of local structure. The increase in diversity between generations 15 and 19, reflects the final seven loci at the C-terminus experiencing a local structure change. The subsequent decrease in diversity reflects the domination of the most



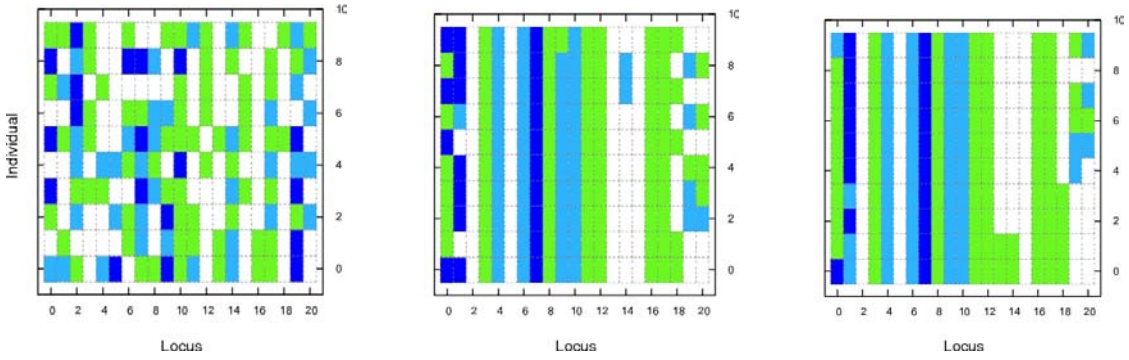
(a) (b)

Figure 5.10: Population diversity mappings for each individual with respect to (a) each other in a pairwise manner and (b) the lowest energy conformation.

favourable segment of local structure in the population. The closer the structural changes are to the centre of the conformation, the more the diversity increases. As it is common for the terminal loci to mutate, the more uncommon central loci mutations allow the diversity to surge. The more individuals that exhibit a central change, the more the diversity is affected. This is observed around generation 35, and more so around generation 65. At generation 25, the large diversity increase reflects the addition of a new-born individual into the population. As the newly added individual is not fit enough to survive another generation, it is removed from the population, and hence the diversity increase is short lived.

The diversity (relative to the lowest energy conformation) is shown in figure 5.10(b). Some of the same characteristics are seen as for the pairwise case, however, as the individuals are measured against the lowest energy conformation, simple mutations to individuals are not as predominant. Compared to the successful case, the most dense regions exhibit higher D_H values. This is as expected, as we are no longer comparing our population to the GM, only the lowest energy conformation found. The number of sub-optimal minima, locally related to the lowest energy conformation found, is larger than those related to the GM (a much broader part of the folding

funnel). Generally the magnitudes of D_H for the unsuccessful case, reflect those seen for the successful case between generations 7 to 17, before performing desirable central conformational changes. This suggests that the search itself is not able to negotiate the PES successfully, in order to probe the correct funnel and find the GM. As the lowest energy conformation determined in generation 71 is not superceded by one of even lower energy for the remaining 19,929 generations, it suggests that the central conformation adopted by the individuals, prevents the search from emerging from the funnel in which it is trapped. This suggests that the shape of the funnel is not necessarily shallow, but possibly deep and broad.

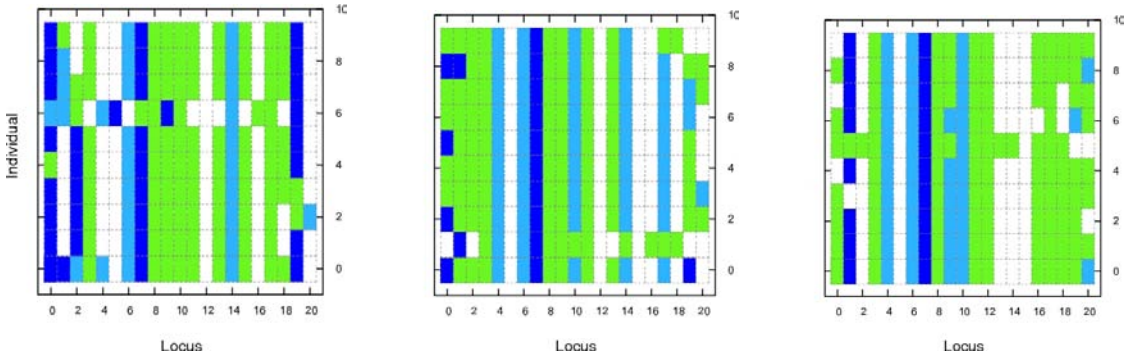


(a) (b) (c)

Figure 5.11: Conformation vectors for (a) the initial (b) half-way and (c) final populations. 0 = white, 1 = green, 2 = cyan and 3 = blue. Individuals are ordered by fitness, with the highest fitness individual shown at position 0.

Figure 5.11 illustrates the conformation vectors for the initial, half-way and final generations for the unsuccessful case. Again, as witnessed for the successful case, the initial generation exhibits randomness, with no dominating regions of local structure. By the thirty fifth generation, dominant alleles exist for all loci, with variants exhibited at the termini. In comparison with the successful case, the final generations also exhibit structural similarities. Blocking occurs as loci share the same dominant alleles for positions 0 to 9, 11, 13, 19 and 20. However, supporting the mutations described, positions 10, 12 and 14 to 17, for all individuals, do not share any alleles with the

final population for the successful case. It seems that the central positions are more difficult to mutate, and in contrast the near terminal residues exchange alleles more freely (greater mobility), i.e. it is the central conformation that determines the folding funnel in which the technique is to search.



(a) (b) (c)

Figure 5.12: Conformation vectors for the (a) third (b) thirteenth and (c) twentieth populations. 0 = white, 1 = green, 2 = cyan and 3 = blue. Individuals are ordered by fitness, with the highest fitness individual shown at position 0.

Figure 5.12 illustrates the conformation vectors for generations 3, 13 and 20. These generations are significant, as they highlight the first occurrence of a new dominant central local structure between positions 6 and 10, inclusive. Although numerous alleles are the same, the near terminal regions of local structure are more free to mutate in following generations. They represent the three main regions of central local structure throughout the first seventy one generations of the calculation. For the conformations present in generations 3 and 13, the energy barriers are quickly overcome, resulting in the change witnessed in generation 20. This suggests that they are part of a larger funnel, with only a small energy barrier to overcome in order to descend into the main well. As this region of local structure, present in generation 20, remains throughout the calculation for thousands of generations, this suggests that it contributes greatly to the low energy of protein conformations for this model and that it lies within a broad deep minimum. By generation 295, positions 13, 14 and 15 dominate the population,

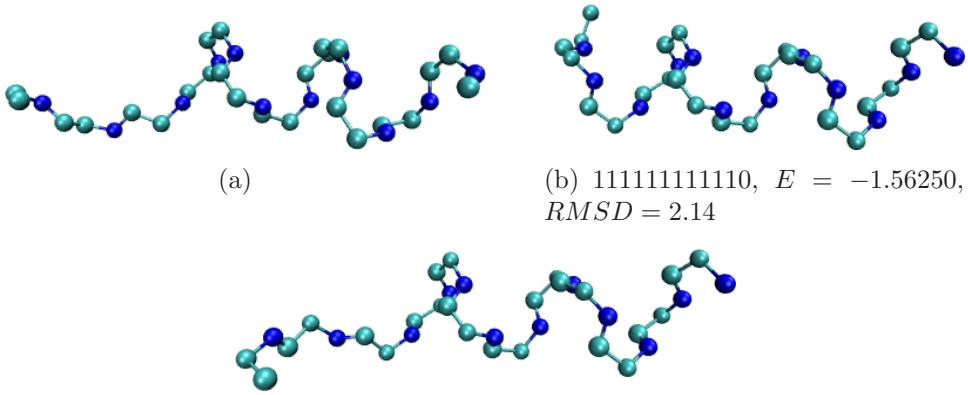
adopting alleles = 1 until generation 2768 (with all new individuals existing in generation 2769). This suggests that the folding funnel may be surrounded by large energy barriers that the IA is unable to emerge from.

5.1 Global Minima

The RMSD values quoted in the following sections use the method adopted by Kobe et al. [104] [119]. Figure 5.13, shows the structure of the 1AL1 protein from both the PDB and GM conformation obtained by the IA. The PDB structure comprises a 3.6_{13}-helix, which is also adopted by the model protein. The C-terminal tail exhibited by the real structure is not present in the model structure due to a combination of reduced torsional space and the distance dependent potential. By manually changing the conformation vector (corresponding to a ϕ_C, ϕ_C to ϕ_T, ϕ_T change for Lys and vice versa for Gly for the final two residues), a C-terminal tail can be produced, as visually close to the real structure as the model torsional space permits (shown by a decrease in RMSD from 2.14 to 0.94). However, the DLM potential recognises this as a sub-optimal conformation (figure 5.13(c)).

Due to the distance constraint employed by the DLM potential (8.0 Å), the α-helix for sequence 1AL1 only recognises interactions between residue i and residues $i+2$ to $i+4$ inclusive. The interaction between the final Glu-Lys residues ($e_{\mu\nu} = -0.1221$), contributing more to the lower conformation energy than the final Glu-Gly repulsive interaction ($e_{\mu\nu} = 0.1167$ with a greater C_α distance), is the driving force for the extra helical turn present in the GM.

All interactions are shown in figure 5.14 for various residue combinations. With the potential being distance dependent, the distance should be maximised between repulsive interactions and minimised between attractive interactions. Due to the position of Leu residues throughout the structure, the Leu-Leu interaction ($e_{\mu\nu} = -0.2142$) favours C_α positions at a similar point in the helical turn (figure 5.14(a)). The same is



(a)

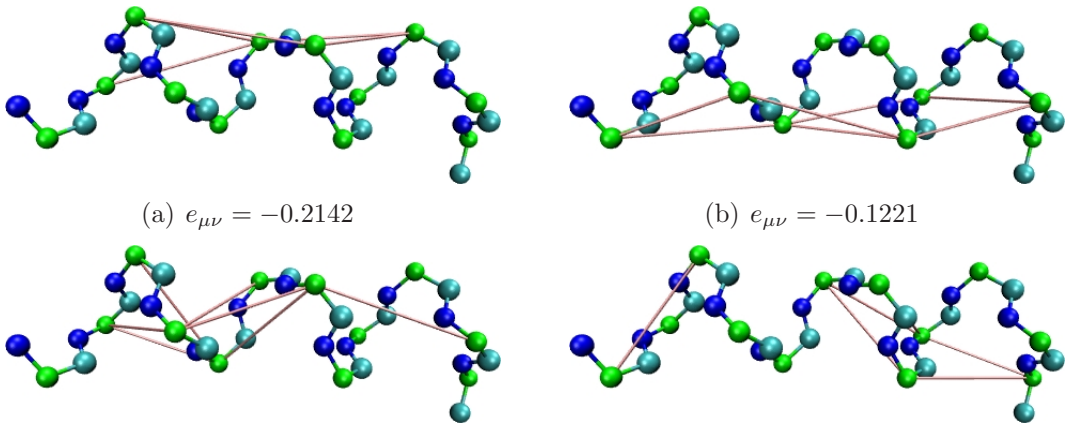
(b) 111111111110, $E = -1.56250$, $RMSD = 2.14$

(c) 111111111101, $E = -1.39302$, $RMSD = 0.94$

Figure 5.13: (a) 1AL1 structure from the PDB. (b) GM conformation found by the IA. (c) Modified GM conformation. Orientation is not N to C terminus, but to show the best agreement. Both (b) and (c) are accompanied by their conformation vectors and RMSD values.

observed for interactions between Glu and Lys Residues. Due to the density of Lys-Leu interactions, the same behaviour cannot be observed. To maintain favourable distances between the interacting residues, the C_α positions occur prodominantly within one and a half helical turns, with the exception of the final Lys. In contrast, to maintain a maximum distance between repulsive interactions, with the exception of Gly (the C-terminal residue), interactions are found between opposite sides of the helical turn.

Figure 5.15 shows both the PDB entry and the GM for sequence 1B19. Two 3.6_{13}-helices are linked via a small residue chain. Residues 6 and 11 (both Cys) form a disulfide bridge, allowing the chain to break its first helical arrangement. No disulfide bridge is present between Cys residues at positions 7 and 20. Some degree of helical arrangement is present for the model protein. However, the disulfide bridge interaction does not exist between residues 6 and 11, due to the C_β distance being too large for the DLM potential to recognise. Cys residues at positions 6 and 7 do have an interaction with the Cys residue at position 20, contributing to the low conformation energy, with this being responsible for the chain folding back on itself. The restriction

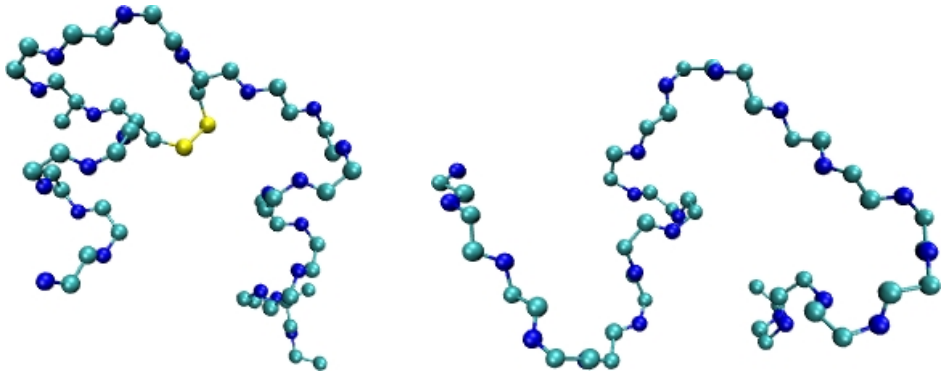


(a) $e_{\mu\nu} = -0.2142$ (b) $e_{\mu\nu} = -0.1221$

(c) $e_{\mu\nu} = -0.0562$ (d)

Figure 5.14: Attractive interactions for the GM of 1AL1 sequence between (a) L-L, (b) E-K and (c) K-L residues with (d) showing all repulsive interactions. All interactions are shown in pink with C_α atoms highlighted in green.

in torsional space is responsible for the poor helical arrangements observed in the model protein. The diversity present with respect to the geometry is attributed to the Cys-Cys interactions that are not present in the real conformation and vice versa.



(a) (b)

Figure 5.15: (a) 1B19 structure from the PDB with the disulfide bridge highlighted in yellow. (b) GM conformation found by the IA.

5.2 Extending The Dynamic Lattice Model

The work of Kobe et al. [104], demonstrated that by taking a selection of 403 structures [12] from the PDB, centroids of ϕ, ψ clusters from a Ramachandran plot, can be

calculated in order to determine an average angle pair representative of the cluster itself. As explained in section 1.1.3, torsion angles favour certain regions of Ramachandran space. While performing a k-means cluster analysis [120] on more ϕ, ψ angle pairs for each residue, will not shift the allowed regions of Ramachandran space, it may allow for a higher resolution of the average angle pair, and thus different values may be obtained.

The PDB contains tens of thousands of protein structures. However, in order to produce a high quality data sample, various factors have been considered. As identical chain segments exist between structures, they must be removed in order to produce a fair unbiased angle pair distribution. Structures must be at least 50 residues in length, with 95% chain identity (no more than 5% similarity). As many experimental techniques are used for structure characterisation, a choice exists as to the source of the data. As X-ray diffraction can produce high resolution results, selection has been restricted to structures determined using this method. A resolution of no more than 2.0 Å must be used, with experimental data to support this. The R-Factor must be no more than 0.15 (15%), to represent good agreement between observed and calculated diffraction intensities. The structure must have been determined at low temperature (no more than 180 K) to keep atom motion to a minimum. To reduce biasing even further, only PDB entries that contain single structures (one chain) have been used, containing no DNA, RNA or hybrids of both. These search criteria (more strict than in [12]) resulted in 482 unique PDB entries.

In order to extract ϕ, ψ distributions, the backbone angles for each structure were calculated and sorted by residue. This allows Ramachandran plots (figure 5.16(a)) for each residue, based solely on the PDB data set to be produced. By performing k-means clustering (with a point distance of no more than 3.0) on the angle pairs from the Ramachandran plots for each residue, clusters occupying favourable regions of Ramachandran space can be generated (figure 5.16(b)). Once the clusters are generated, the centroids can be calculated, to result in a reduced set of ϕ, ψ angle pairs (one

representing each cluster). Clusters are ordered by size (number of contributing angle pairs) with no more than 20 clusters considered for each residue. This, in effect reduces continuous Ramachandran space into a discreet space of 20 angle pairs. Centroids are rounded to the nearest 5° [104].

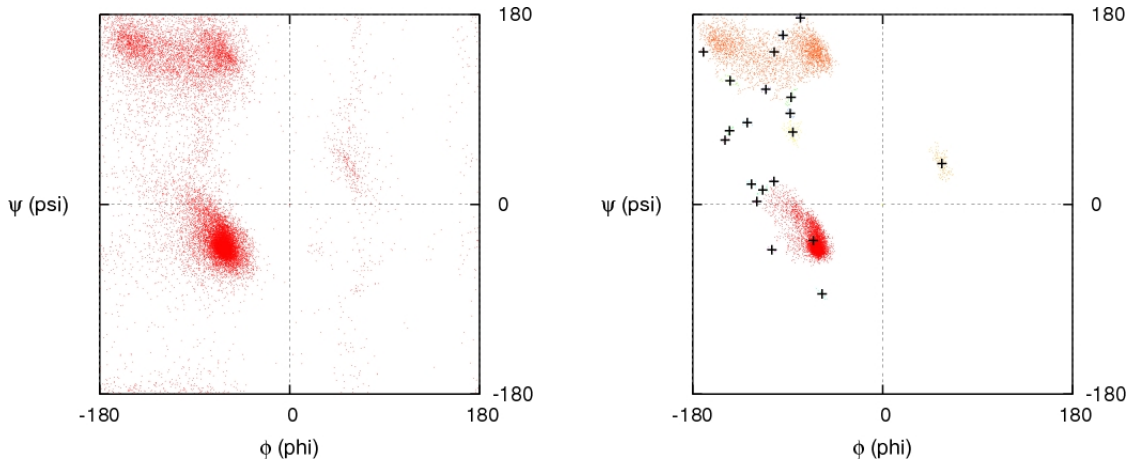


Figure 5.16: ϕ, ψ distributions for the Alanine residue illustrating (a) Ramachandran space and (b) the clustering angle pairs giving rise to cluster centroids.

Angles were selected by inspection, with magnitudes being as closely related to the existing angle pairs [104] as possible. Although 20 centroids have been calculated, the number of angles made available (defined by Kobe et al.) has been maintained for comparative reasons. This results in small deviations from the existing angle pairs by no more than 5 - 10° in most cases, however, some angles show a much larger deviation. The original angle pairs cover a majority of available Ramachandran space, resulting in angle pairs not being biased towards β-sheet or α-helical regions, hence the choice to mimic them as much as possible. New angle pairs determined here are shown in table 5.5.

Table 5.6 compares the SR and energies of the most stable conformations gained for each torsion angle set. The modified torsion angles not only improve the energies of sequences 1AKG, 1L2Y and 1D9J, but they also show comparable energies for other

Amino Acid	ϕ,ψ Angle Pairs
A	(-145, 115) (-65, -35) (-100, 145)
C	(-70, -15) (-105, 155) (-65, -40) (-120, 20)
D	(55, 40) (-85, 125) (-130, 35) (-75, -25)
E	(-105, 140) (-70, -35)
F	(-115, 140) (-70, -35) (-130, 70)
G	(75, -155) (85, 5) (-105, 160) (-70, -30)
H	(60, 40) (-130, 140) (-70, -30)
I	(-110, 130) (-65, -40) (-90, -5) (-75, 155)
K	(-105, 140) (-70, -30)
L	(-100, 135) (-70, -35)
M	(-120, 135) (-65, -35)
N	(60, 35) (-100, 130) (-80, -15)
P	(-65, -25) (-85, 135) (-65, 150)
Q	(55, 45) (-70, -30) (-15, 140)
R	(-70, -30) (-110, 140)
S	(-110, 150) (-70, -25) (-70, 120)
T	(-110, 145) (-80, 110) (-80, -25) (-90, -55)
V	(-130, 90) (-110, 130) (-110, -5) (-65, -40)
W	(-130, 145) (-65, -35) (-75, -5)
Y	(-115, 140) (-75, -30)

Table 5.5: Revised angle pairs taken from a sample PDB data set for the DLM backbone for each of the 20 natural amino acids.

sequences. High SR suggests that the change in torsion angles results in reproducible conformations for different runs of the IA. Lower RMSD values are observed for sequences 1AKG, 1B19, 1G04 and 1ANP, illustrating that although an improvement in conformation energy is not observed (except for sequence 1AKG), an improvement in structural similarity to the PDB conformation is (shown by the reduction in RMSD). Smaller sequences generally show a closer relationship to the PDB sequence than larger sequences. Due to an increase in the number of amino acids per chain, a general increase in RMSD is observed.

5.2.1 Minima Using Modified Angles

Figure 5.17 shows a direct comparison between the PDB entry, the GM conformation using the original angles, and the lowest energy conformation using the modified angles

PDB ID	E^*_{orig}	SR_{orig}	$RMSD_{orig}$	E^*_{mod}	SR_{mod}	$RMSD_{mod}$
1AL1	-1.56250	100.00	2.14	-1.56140	100.00	2.14
1A1P	-1.51775	100.00	3.62	-1.25122	100.00	5.57
1AKG	-1.17301	100.00	4.05	-1.42336	100.00	3.29
1L2Y	-0.87846	100.00	3.96	-0.93497	100.00	9.49
1D9J	-1.34457	100.00	5.05	-1.43968	49.00	6.74
1B19:A	-3.54363	68.00	5.94	-2.62982	5.00	5.72
1G04	-1.24277	100.00	10.12	-1.13887	99.00	9.60
1ANP	-2.37581	1.00	8.29	-2.19046	2.00	7.91
1AML	-4.91472	19.00	7.38	-4.49711	2.00	8.68
1QHK	-3.78268	2.00	11.05	-3.40506	1.00	12.52

Table 5.6: SRs, lowest energies and RMSDs (to PDB structure) for both the original [104] and modified angle sets for the IA using the DLM sequences in table 5.5.

for the IA for sequence 1AKG. It is an example of how an improvement in energy (from 5.17(b) to 5.17(c)) is observed when exchanging one angle set for another (resulting in a different conformation vector). Although detailed structural features found in figure 5.17(a) and also found in figure 5.17(b), are not observed in figure 5.17(c), upon initial inspection, an improvement to the general geometry is observed. However, an improvement in energy should reflect an improvement in structural detail for an accurate model. The backbone kinks (which model the α-helix) are not as predominant for the lower energy conformation. A regular α-helix can be modelled using similar ϕ, ψ angles for each residue (as observed for sequence 1AL1, deviations of around 5°). The reduced α-helix characteristics in figure 5.17(c), suggests irregularity of backbone angles for that structure segment. However, according to the RMSD values, the conformation is more closely related than that generated using the original torsion angle set.

The 1AKG sequence contains 4 cysteine residues. As explained in section 1.1.1, disulfide bridges can be formed between two cysteine residues of the same sequence. The compact arrangement and stability of the 1AKG sequence, is attributed to these two disulfide bonds. One of the largest attractive interaction parameters for the DLM (table 2.6) exists between two cysteine residues. Although the variable distance be-

tween the side chains (derived from other backbone angles), may not render the interaction to be its most favourable for this particular sequence, it is the interaction between these residues (particularly the first and last) that provides the driving force for this conformation, for both torsion angle sets.

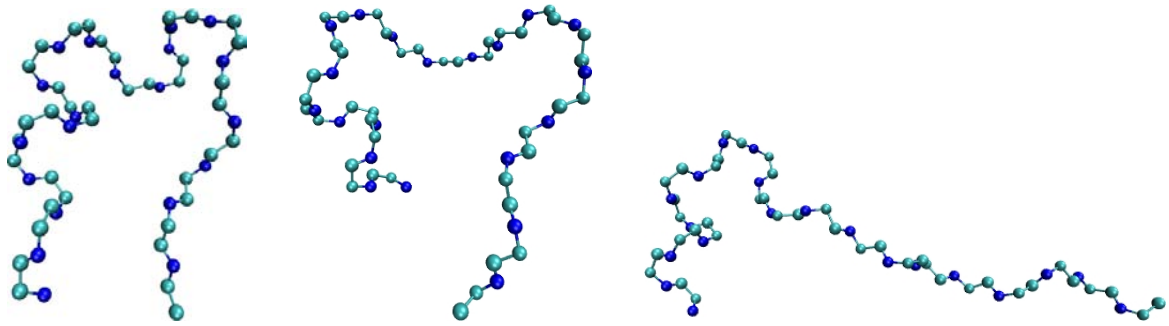
(a) (b) (c)

Figure 5.17: Graphical representations of the 1AKG structure from the (a) PDB, (b) IA using the original backbone angles ($E^* = -1.17301$, RMSD = 4.05) and (c) IA using the modified backbone angles ($E^* = -1.42336$, RMSD = 3.29).

Sequence 1L2Y shows an improvement in energy between figures 5.18(b) and 5.18(c). However, although an improvement in energy is observed, the geometry is less like the PDB entry in figure 5.18(a). The long tail (from the first Pro onwards in figure 5.18(c)), does not mimic characteristics of the real protein chain. The driving force for the formation of this model protein seems to be the interactions between the fourth and eleventh residues, specifically the numerous Leu-Ile and Leu-Gly interactions. This is due to their close proximity and the attractive interaction parameters, observed in table 2.6.

As previously stated, the interaction parameters do not differ between the modified and original models, only the backbone torsion angles. The problem with using the DLM for protein structure determination is three-fold. The search technique is not guaranteed to find the GM (in terms of energy). If it does find the GM it must resemble the real conformation. If it does not resemble the real conformation, a higher energy solution might (section 5.1). Either the search technique, the potential or the

torsion angle set may be a contributing factor to the level of success. The issue observed here, is one involving the DLM potential, due to a lower energy conformation being found but which is less like the PDB entry. For this particular case, changing the torsion angles for a different set has proven successful in determining a lower energy conformation, but unsuccessful in matching the experimental structure.



(a) (b) (c)

Figure 5.18: Graphical representations of the 1L2Y structure from the (a) PDB, (b) IA using the original backbone angles ($E^* = -0.87846$, RMSD = 3.96) and (c) IA using the modified backbone angles ($E^* = -0.93497$, RMSD = 9.49).

5.3 Conclusions

In this work, the IA has been used to investigate the PES of a number of DLM proteins, in the hope not only to discover the GM, but also to generate realistic conformations. The original backbone torsion angles were determined by Kobe et al. [104], with modified torsion angles shown here, calculated using k-means clustering of ϕ, ψ pairs for each residue.

The IA has demonstrated that it is able to discover the GM (determined by exhaustive branch and bound searches) efficiently for DLM model proteins. The DLM is able to reproduce protein conformations close to those of real proteins in some cases. However, as chain size increases, the DLM representation becomes increasingly dissimilar to the real conformation, due to either the failings of the potential, model or search method.

Results obtained from profiling suggest that success is based on population diversity. Large contributions to population diversity arise from mutations of the central schema of the conformation vectors. The central loci present more problems during mutation than terminal loci due to the increased disruption to the existing conformation and restricted mobility. For this reason, they are predominantly responsible for changing the direction of the search for small population sizes, with premature convergence occurring when this cannot be achieved. Smaller contributions to population diversity arise from mutations of loci at either terminus.

Although the stability of a disulfide bridge is reflected in the interaction parameter of Cys, the reduced torsional space does not always allow the distance between the C_β atoms to be small enough to take advantage of the interaction strength. This seems to be a determining factor in whether a model protein will adopt the conformation of the PDB protein derived from NMR and X-ray diffraction.

Reducing torsional space is unfavourable for some model proteins to exhibit the structural characteristics found in real conformations. A set of torsional angles obtained using strict search criteria fails to improve the efficiency and results of the IA using the DLM. By providing a new set of torsion angles, from a more restricted set of structures from the PDB, the energies of some model proteins and the RMSDs of others may be reduced. However, the energy does not necessarily reflect a model protein's ability to adopt the real conformation provided by the PDB. In some cases where a reduction in energy is observed, the model conformation becomes less like the real one.

Chapter 6

Differential Evolution

DE differs from other search techniques due to the combined mutation and mating operator that acts each parent in a generation. Whereas other search techniques utilise a selection phase for mating of parents, a DE sequentially treats every individual of a population as a parent. The DE was introduced in section 2.3 along with the two factors (F and K) that can affect the rate of convergence for this combined mutation and mating phase.

The models used in this work are discrete due to the complexity and nature of the protein folding problem. To drive down CPU times, lattice bead and united atom models have been employed for use with the DE. In a conventional DE, with continuous variables, F and K can adopt continuous values between 0 and 1.0 inclusive. However, the discrete representations resulting in integer only alleles in the gene pool, prevent F and K from adopting such continuous values, restricting both magnitudes to 1.0.

As with any other search technique, n_{ind} and g_{max} can also be varied. For testing and comparative reasons, in this work n_{ind} and g_{max} take the same values as for the IA (chapters 4 and 5). Due to the nature of the combined mutation and mating operator integrated into the DE, a reduction in μ_{FE} is expected and, therefore, longer calculation durations are also considered.

6.1 HP Lattice Bead Model

Section 1.1.5.1 introduced the simplistic HPLBM, involving simple 2D and 3D model representations of protein molecules. For initial testing and to understand how the DE copes with simple 2D and 3D structure determination, the HPLBM is the first example which has been subjected to the DE.

6.1.1 The Square Lattice

Section 2.4.1.1 highlights that the HPLBM, coupled with a simple 2D square lattice, imposes heavy restrictions on the conformation of model proteins, reducing the search space and providing an ideal test case for new search methods. The heavily studied sequences in table 3.1 also provide a good test set for comparing search methods.

In order for a direct comparison to be made between the DE and the IA (chapter 3), results for the common variables between both methods; n_{ind} adopting values ranging from 10 - 200, inclusive, and g_{max} of 1000, have been reported in figure 6.1 in an identical format to the IA results shown previously.

Figure 6.1 clearly identifies which values of n_{ind} give rise to higher SR when using the simple combined mutation and mating operator for the DE search method. For small n_{ind} (< 50), zero success rates are observed, with very poor SRs for the remaining magnitudes of n_{ind}. The wealth of genetic material present when using larger values of n_{ind} is much greater than for small values. As expected, the level of success increases with an increase in n_{ind}, as observed with other search techniques. With an increase in genetic material, more areas of the PES can be searched simultaneously, increasing the opportunity for search techniques to discover the GM.

The effect of searching different areas of the PES is quantified by the number of unique minima found over one hundred runs for each set of variables. It is apparent from figure 6.1 that as we increase the magnitude of n_{ind}, an increase in n_{uniq} is also seen, although a plateau is observed between the larger magnitudes of n_{ind} (100 - 200).

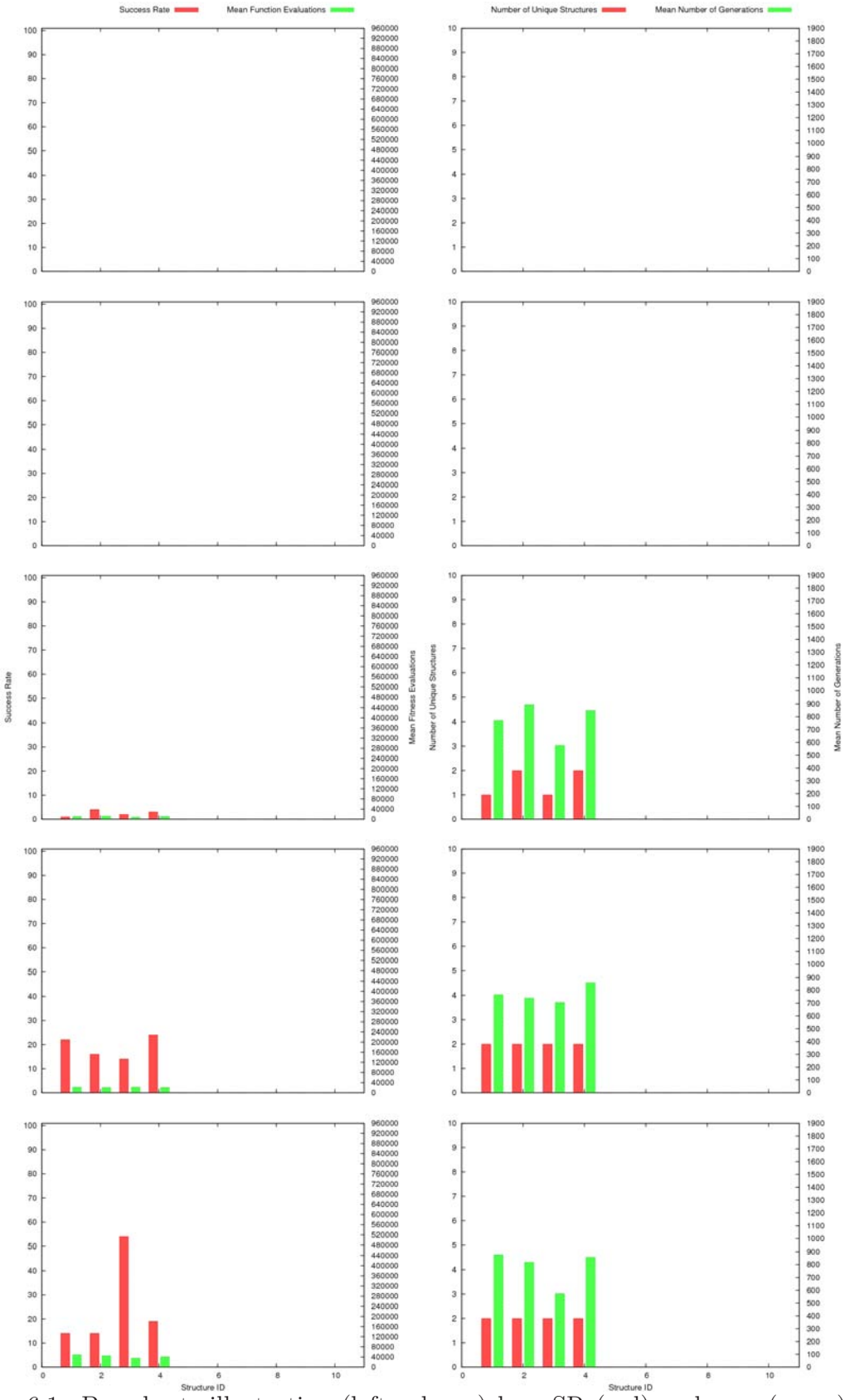


Figure 6.1: Bar charts illustrating (left column) how SR (red) and μ_{FE} (green) and (right column) n_{uniq} (red) and μ_g (green) change with increasing population size for 100 runs of 1000 generations for the sequences featured in table 3.1, without the use of the RGA for repair.

The increase in n_{uniq} arises from the different areas of the PES that can be explored as a result of a population's genetic material being more varied. However, although we see a general increase in SR, μ_g remains relatively stable. This illustrates that an increase in efficiency is not seen for the search technique itself as a result of increasing n_{ind}, but again more areas of the PES can be explored at any one time, leading to an increase in the number of runs that converge.

The PES of a model protein is dependent on the number of amino acid residues being modelled. The search space increases exponentially with the length of the primary sequence. It is obvious from figure 6.1 that the DE cannot cope with primary sequences larger than twenty beads. This demonstrates that for small sequences, corresponding to a less congested PES, the DE is able to search for the GM successfully when presented with a wealth of genetic material. In contrast, larger sequences correspond to a more complex PES, presenting many more sub-optimal minima for the DE to overcome. The DE in this case fails to discover GMs, even when presented with an array of genetic information. Many more opportunities for mutation are available for larger sequences (greater than 20 beads in length), therefore, the likelihood of getting trapped in local minima is increased.

It is expected that μ_g should fluctuate with μ_{FE} for each value of n_{ind}. It is also expected that an increase in n_{ind} should yield an increase in μ_{FE}. This is indeed the case, due to the number of valid conformations generated per generation of the DE. If n_{ind} increases, a greater number of valid conformations (a greater number of areas of the PES) are explored per generation, thereby increasing μ_{FE}. Again, it is this increase in μ_{FE} that contributes to the increase in SR.

With other search techniques, mutation and mating are performed separately, allowing intermediate conformations to be stored in the population. The combined mutation and mating operator employed by the DE forbids individuals to be stored in the population until the entire two-phase process is complete. The problem with this type

of genetic operator is that the subtraction of the third random individual from the second (mutation phase) may give rise to a valid conformation. Also the subtraction of the first random individual from the parent (mating phase) may also give rise to a valid conformation. These possible solution candidates are then combined to produce a trial solution. During this whole process, two areas of the PES have been completely overlooked (the mated and the mutated). It may be that one or both of these unprobed areas could have opened up a more efficient pathway towards finding the GM or indeed be the GM itself. By not adding these individuals to the population, the exploration of these areas of the PES is prohibited and possible candidate solutions may be overlooked.

In order to add an individual to a population in DE, the offspring must be fitter than the current parent in order to replace it. In DE, the offspring conformation is influenced by the parent as well as by three randomly selected invdividuals. By allowing the intermediate conformations to occupy a position in the population, it could happen that an individual has been promoted to the population without having interaction with the parent (the mutation phase). If this were the case, then the methodology behind DE would be compromised, and in essence, the DE would be more like other evolutionary search techniques currently used, having separate mutation and mating phases.

Figure 6.2 illustrates the same data types as featured in figure 6.1, for the DE now coupled with an RGA (to be referred to as the RGA-DE) during the combined mutation and mating phase. The DE is coupled with an RGA as described in section 2.2.1. The presence of the RGA ensures that the combined genetic operator produces offspring that are self avoiding walks. This is accomplished by performing as few changes as possible to the conformation vector after visiting each locus sequentially.

Initially, it is apparent that, even for the smallest magnitude of n_{ind}, a tiny level of success is achieved. Again, as expected, as we increase n_{ind}, for certain lengths of

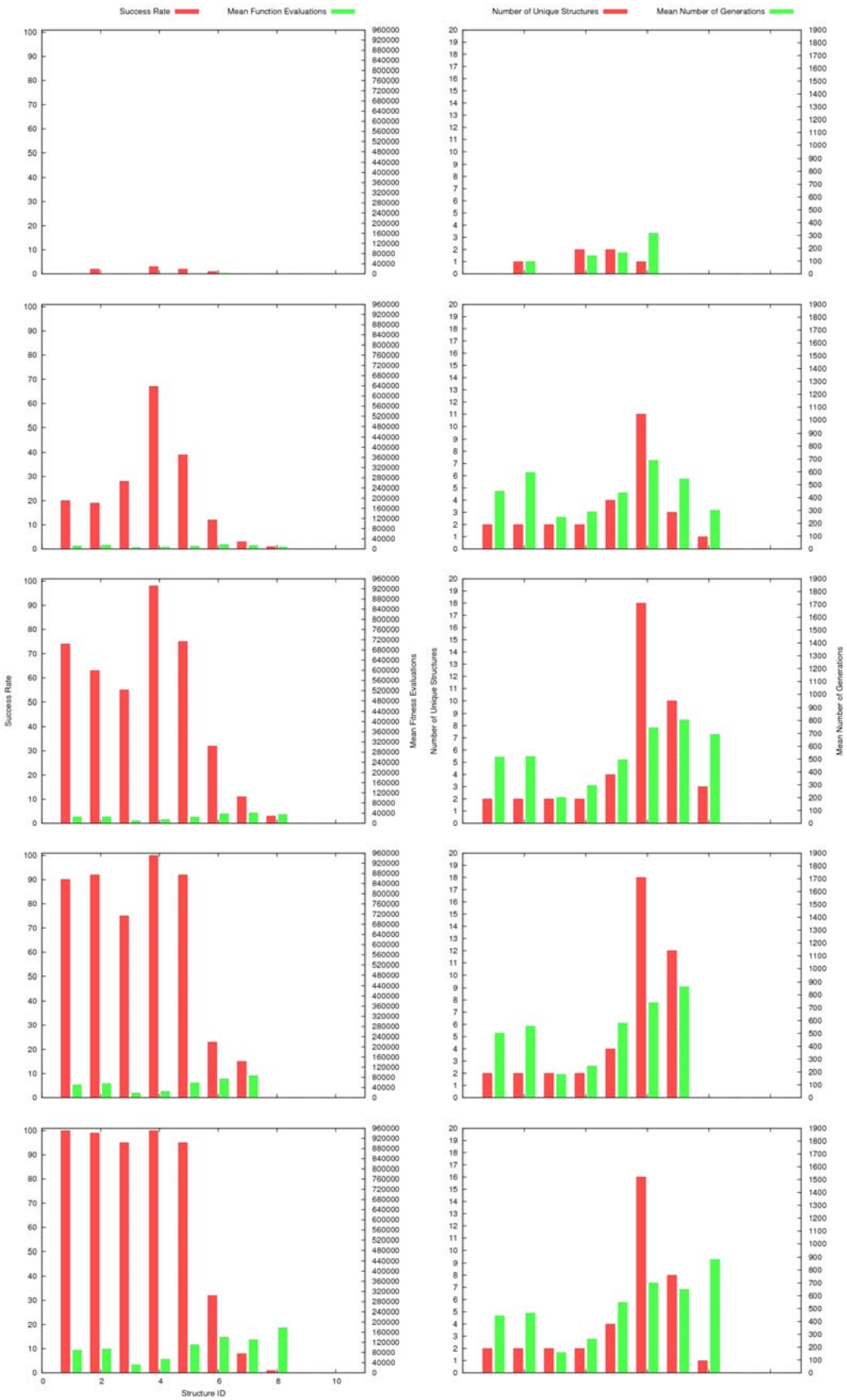


Figure 6.2: Bar charts illustrating (left column) how SR (red) and μ_{FE} (green) and (right column) n_{uniq} (red) and μ_g (green) changes with increasing population size for 100 runs of 1000 generations for the sequences featured in table 3.1 with the use of the RGA for repair for the HPLBM on the square lattice.

primary sequence, a sudden increase in SR is observed. As before, the abundance of genetic information present in larger populations, in comparison to that for smaller values of n_{ind}, yields a greater SR.

Due to the guaranteed production of a self avoiding trial solution, many more areas of the PES can be explored at any one time. The minimal changes being made to the conformation vector allow local searches of the invalid trial solution to be undertaken. This allows the adoption of a self avoiding walk of a nearby sub-optimal structure to the trial solution. The nature of the DE allows any new trial solutions to take part in any subsequent genetic operations on other parents. The information gained as a result of a genetic operation can then be passed on to other trial solutions, moving the search to other areas of the PES. The number of areas of the PES covered by this search is far greater for the DE when incorporating the RGA than when omitting it. This is reflected in success being achieved for larger primary sequences as seen in figure 6.2.

Upon inspection, the same trend is seen for the increase in μ_{FE} as was seen for the stand-alone DE in figure 6.1. Due to the incorporation of the RGA, the general increase in μ_{FE} as we increase n_{ind} is due to the larger number of fitness evaluations made per generation. The larger n_{ind} values statistically provide a greater number of successful genetic alterations. The magnitudes of μ_{FE} across all sizes of n_{ind} are somewhat larger for the DE with the RGA than without, due to the brute force nature of the RGA providing genetic alterations. Each successful genetic alteration requires a conformation energy and fitness calculation as explained in chapter 2. The greater the number of fitness calculations per generation, the higher μ_{FE} will be over all generations.

As observed for the stand-alone DE, the RGA-DE shows very little fluctuation in μ_g. The magnitude of n_{ind} is responsible for the increase in SR for this set up and not the ability of the process to recover from local minimum trapping over time. A problem still exists with the DE's ability to prevent sub-optimal minimum trapping for larger

sequences ($n_{beads} \geq 36$). It is not uncommon for a model protein to have to overcome energy barriers (partially unfold) in order to find its native state. Trial solutions for the DE overwrite the parent individual if an improvement in fitness is achieved, preventing the two from co-existing. By not handling the mutated individuals separately from the current population (as seen in the IA), nor giving high energy conformations an opportunity to re-enter a population by selection, favourable regions of local structure may be replaced prematurely by a thermodynamically more stable (but sub-optimal) conformation. Updating the population in this way also prevents further exploration of local areas of the PES of the former parent individual.

The previous figures illustrate that coupling the RGA with the DE has proved invaluable in determining low energy conformations for these model proteins. By modifying the conformation vectors of individuals, as explained in section 2.3.3, this search technique has access to a wider variety of local structures that may have been lost due to an invalid conformation produced via the genetic operator. By allowing these modified individuals a chance to enter the population, the individuals themselves can actually provide a pathway to different regions of the PES for exploration. The information retained in the population as a result provides access to GM areas of the PES, in turn producing higher success rates.

Without the use of the RGA, the success rates are very low, even for small chains of beads. This reflects the importance of having a large turnover of valid conformations as a result of genetic operations. The likelihood of producing an improvement in fitness over the parent with a higher turnover of valid conformations is significantly greater than for considerably fewer successful genetic operations.

It is expected that by allowing a search to operate for a larger number of generations, the magnitude of SR should be greater. Figure 6.3 illustrates that by increasing g_{max} in steps of 1000 from 2000 - 5000, the overall magnitude of SR does indeed increase. As the primary goal is to achieve high success, increasing g_{max} not only fulfils this

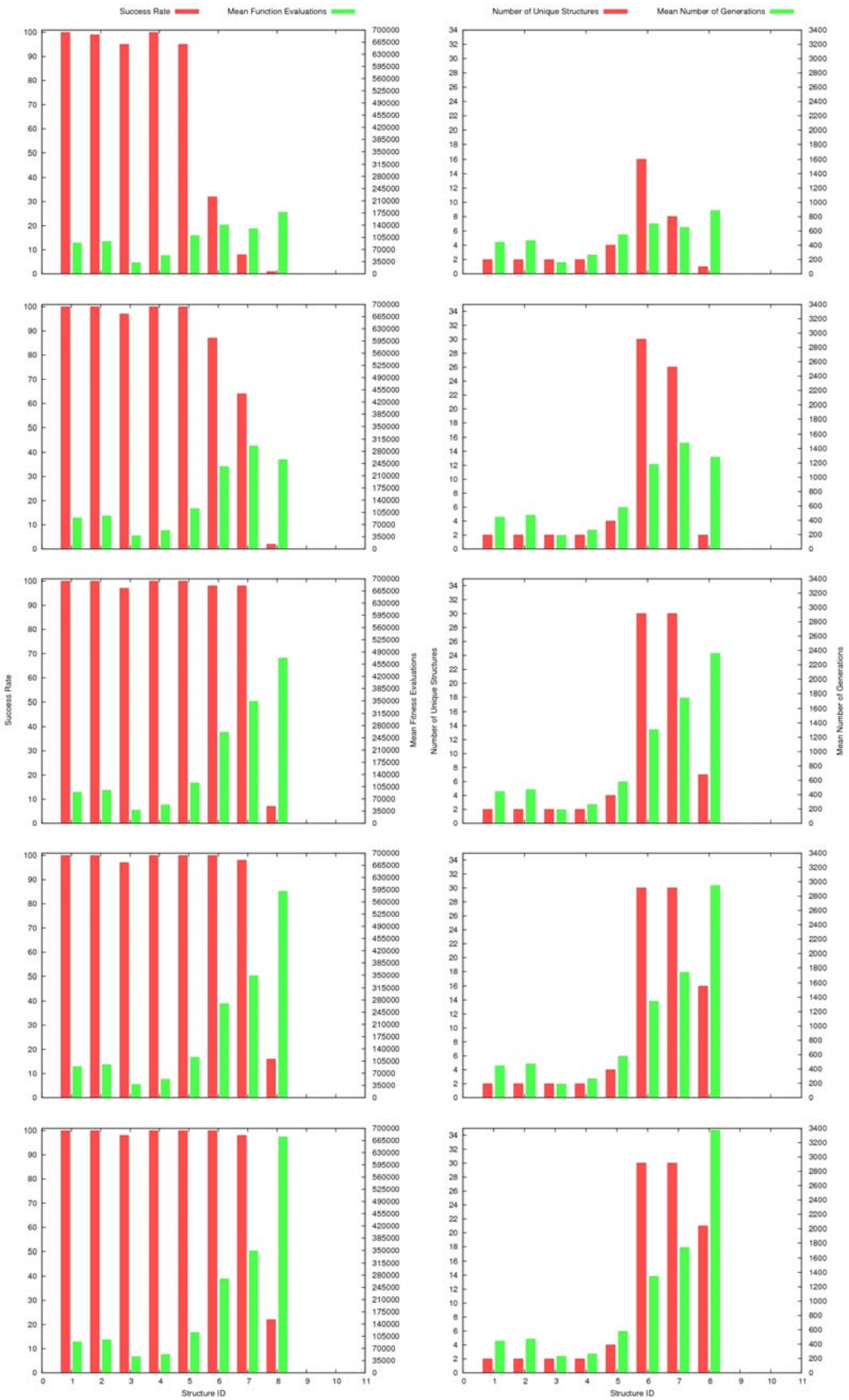


Figure 6.3: Bar charts illustrating (left column) how SR (red) and AFE (green) and (right column) n_{uniq} (red) and μ_g (green) change with an increasing number of generations with a population size of 200 for 100 runs for the sequences featured in table 3.1 for the HPLBM on the square lattice.

requirement, but also gives us an insight into the search capability of the DE and its ability to recover from local minimum trapping. This does, however, require the integration of the RGA with the DE to function, with the plots shown in figure 6.3 utilising this option.

However, the efficiency of this search technique is in question. The magnitude of g_{max} doesn't necessarily determine the efficiency. If a genetic operator is poor, then the number of successful genetic operations should be low (as seen with the stand-alone DE), reflecting a low μ_{FE}. The magnitude of μ_{FE} reflects the number of valid conformations visited in order to reach the GM. If we compare the performance of the DE with the IA from chapter 3, we can see a considerable increase in μ_{FE} for the DE with the HPLBM on the square lattice. In comparison to the IA, the DE struggles with efficiency for all sequences and with SR for chain lengths in excess of 36 beads. For small sequences ($n_{beads} \leq 25$), although inefficient, the DE proves successful in determining low energy conformations, however the search technique is not thorough enough to probe GM regions of the PES for larger sequences.

6.1.2 The Diamond Lattice

As highlighted in chapter 4, a natural progression from a flat, 2D protein representation would be a more realistic 3D one. Again, due to structural motifs, and to aid comparison with other search techniques, the diamond lattice is the preferred choice for this particular search method. As for the IA, the benchmark sequences used for the HPLBM on the diamond lattice are those of high and low degeneracy, listed in tables 4.1 and 4.2, respectively.

The inspiration for testing the structures of high degeneracy arises from the desire to investigate the DE's ability to explore congested PESs (in terms of GMs). We saw with the IA that successful determination of the lowest energy conformation was achieved very quickly. Figure 6.4 describes how SR, μ_{FE}, μ_g and n_{uniq} vary for the sequences of high degeneracy (table 4.1), varying magnitudes of n_{ind}, adopting values

from 10 to 200.

As previously concluded, due to the increase in genetic material present per generation, we see an increase in SR as we increase n_{ind}. However, it should be noted that although the PES contains a vast number of GMs for these sequences, the stand-alone DE still tends to struggle to find GMs for smaller n_{ind}. As only one genetic operator exists for the DE (a combined mutation and mating operation), the production of new genetic material as a result of the operation is significantly reduced in comparison to the IA with its separate mutation operations, even for low n_{ind}.

Upon increasing the magnitude of n_{ind} from 10 to 25, the DE begins to behave in a similar manner to the IA. Due to the amount of genetic material contained in a population, the DE sees a drastic increase in SR. However, in comparison to the IA, the DE's capabilities are still inferior. Sequence L1 from table 4.1 is the only sequence that does not yield a perfect result at $n_{ind} = 50$. In fact, sequence H1 exhibits the worst performance in terms of SR, yielding 2%, 68% and 99% for $n_{ind} = 10$, 25 and 50, respectively. Remaining magnitudes of n_{ind} yielded 100% for all sequences of high degeneracy. This suggests that, as all sequences of high degeneracy are 20 beads in length, the positions of the H beads play a role in determining the shape and complexity of the PES in terms of numbers of sub-optimal minima present and their corresponding fitnesses.

It can be concluded from figure 6.4, that $n_{ind} = 100$ is required to truly obtain an identical performance (in terms of SR) between the stand-alone DE to the IA. The next quantity to compare is μ_{FE}. In order to satisfy our second criterion, μ_{FE} must be at its lowest, accompanying a high SR. Considering $n_{ind} = 100$ for the DE gave rise to equivalent success rates as for the IA. If we compare μ_{FE}, for this value of n_{ind}, for the two search methods, it is apparent that the DE operates with a third of the efficiency of the IA. The IA provides μ_{FE} in the region of 1000 - 1,300, with the DE visiting around 3,000 - 7,000 conformations during its search. In terms of the DE, the larger

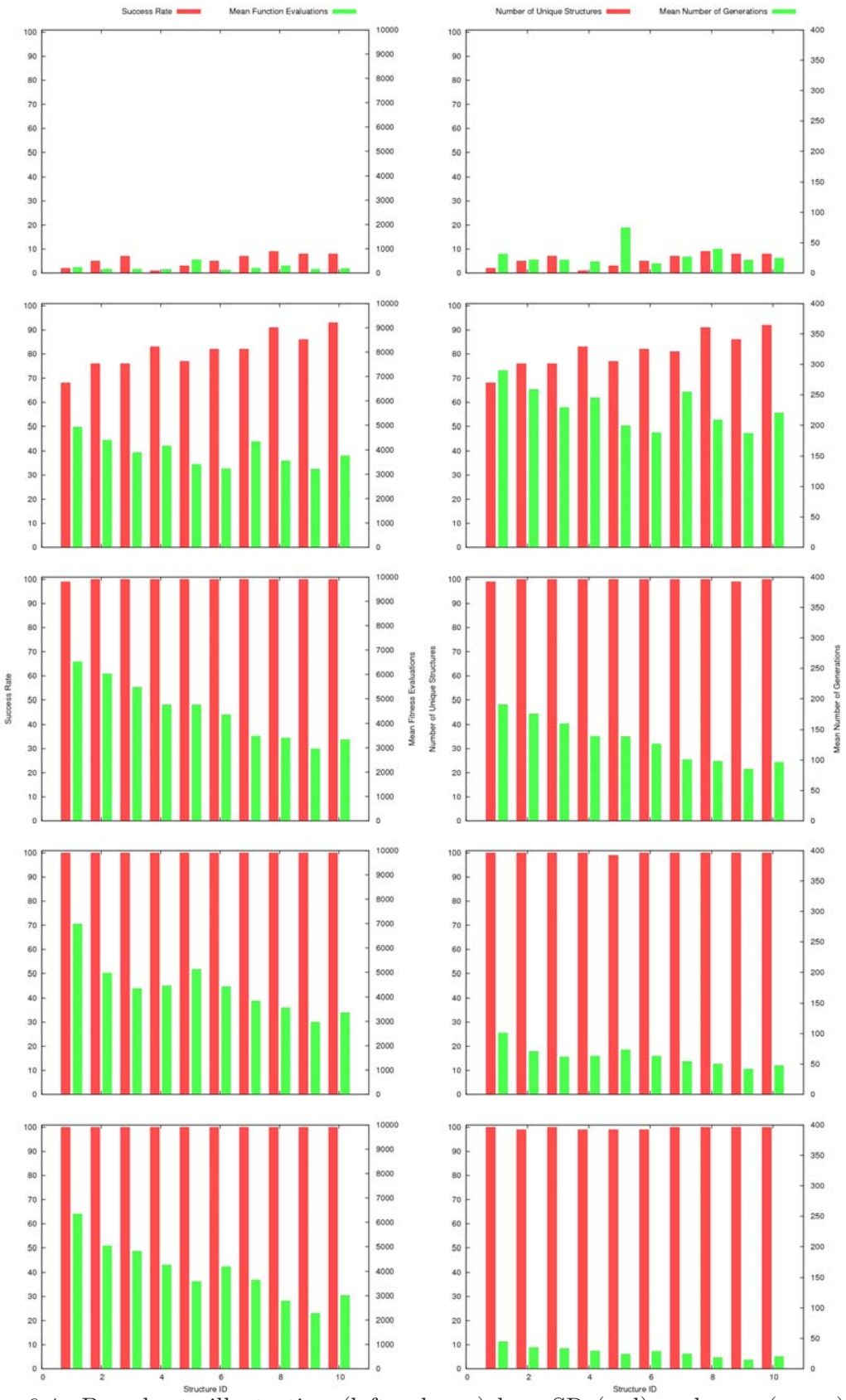


Figure 6.4: Bar charts illustrating (left column) how SR (red) and μ_{FE} (green) and (right column) n_{uniq} (red) and μ_g (green) change with increasing population size for 100 runs of 1000 generations for the sequences featured in table 4.1 for the HPLBM on the diamond lattice.

population sizes result in the highest SR, with $n_{ind} = 200$ yielding lower values of μ_{FE} when compared to those obtained for $n_{ind} = 100$.

The DE struggles with efficiency, visiting more conformations before finding the GM when compared to the IA. The combined mutation and mating procedure adopted by the DE may be condensing too many steps into one, not allowing consideration of intermediate conformations. By neglecting intermediate conformations and producing vast changes to a parent individual, a detailed search across the PES is not performed. This results in a sudden leap to an unrelated region of the PES, not preserving the segments of local structure that initially gave rise to the low energy parent.

Another issue is premature convergence. As explained for the HPLBM on the square lattice in section 6.1.1, premature convergence may be occurring due to a combination of the search not being thorough as well as the parents being overwritten by newly created trial solutions. In the case of the IA, new individuals were combined with current individuals prior to the selection procedure. Both new and old individuals had a chance of being selected for the next generation. The standard genetic operators also allowed a more detailed search to be conducted across the PES by utilising hypermutation (point mutation) and hypermacromutation (an example of inorder mutation) operators that act independently. However, the vast number of GMs present on each PES for the sequences of high degeneracy fail to allow this type of analysis. In order to provide insight into how effective the algorithm is for the HPLBM on the diamond lattice, a more testing set of sequences, those of low degeneracy, need to be investigated using the DE method.

It was observed in figure 6.4 that the larger magnitudes of n_{ind} resulted in the optimal combination of SR and μ_{FE} for sequences of high degeneracy. The sequences of low degeneracy from table 4.2 have previously given more of an insight into a search technique's ability to explore the search space due to a lack of GM on the PES. Figure 6.5 illustrates how successful the DE is in searching for the lowest energy conformations

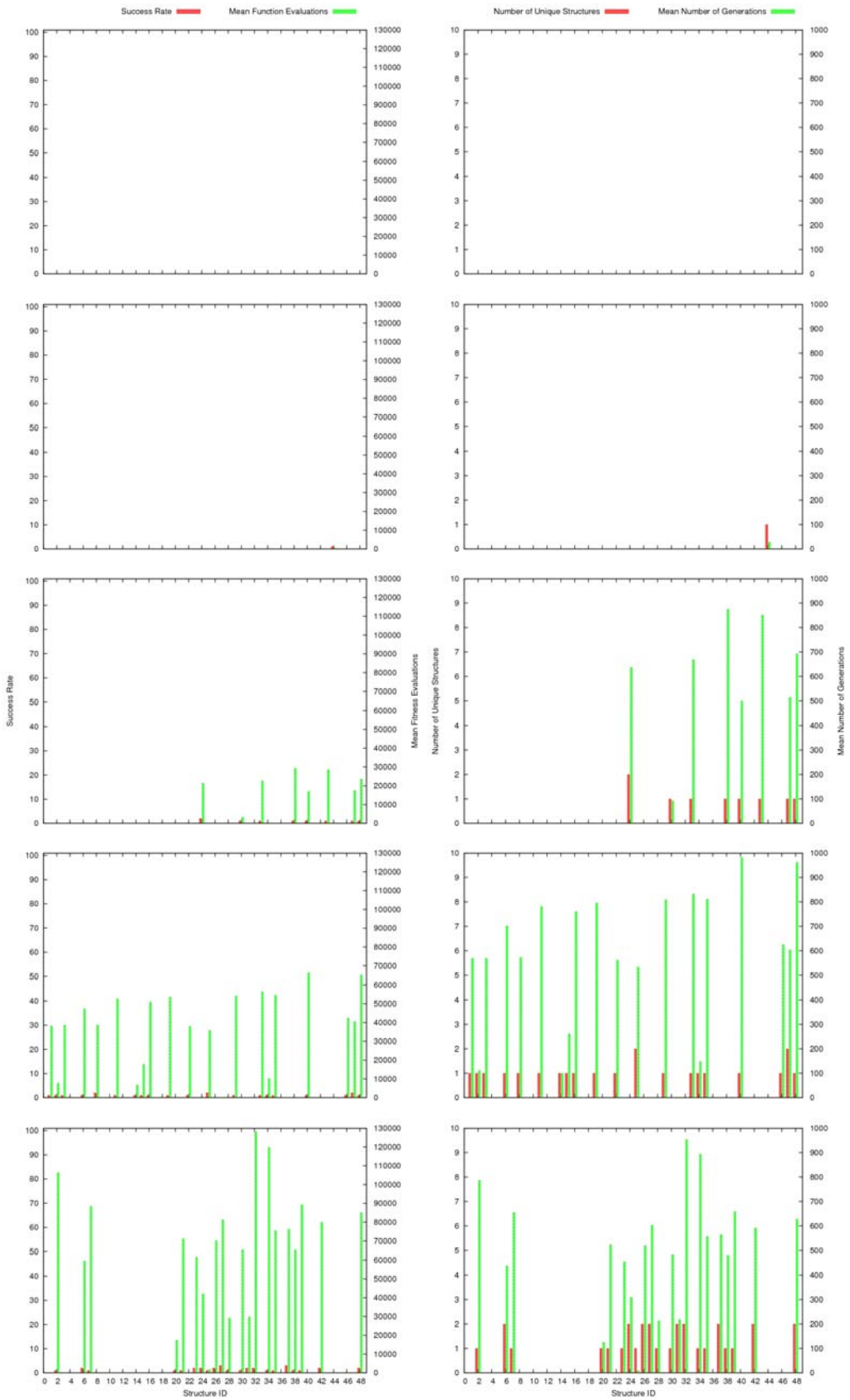


Figure 6.5: Bar charts illustrating (left column) how SR (red) and μ_{FE} (green) and (right column) n_{uniq} (red) and μ_g (green) change with increasing population size for 100 runs of 1000 generations for the sequences featured in table 4.2 for the HPLBM on the diamond lattice.

for each of these sequences. It is observed that the stand-alone DE shows poor success, or none at all for all values of n_{ind}.

The number of minima present on the PES for these sequences is significantly less than for the sequences of high degeneracy, hence the poor success rates. As seen with the HPLBM on the square lattice, the introduction of the RGA provided more valid conformations per generation and showed drastic improvements.

Figure 6.6 shows how SR, μ_{FE}, μ_g and n_{uniq} change for a fixed n_{ind} of 200 and g_{max} varying from 1000 - 5000, after incorporating the RGA, as described in section 6.1.1. The results show the same trend as for the HPLBM on the square lattice. A dramatic improvement in SR is seen due to the number of valid conformations produced from the combined genetic operator. This implies that the complexity of the problem is not to blame for the poor SR, but rather the ability of the search technique to produce valid conformations as a result of the genetic operation.

For the HPLBM, the energy function is not distance dependent, resulting in flat regions on the PES. This means that identical fitnesses will be seen for sequences sharing a hydrophobic core with different P bead placements. This, in turn, results in the funnel of the energy landscape being very broad for sub-optimal conformations and narrow for the GM. The sheer quantity of sub-optimal minima, may prevent a successful search from taking place. By running this search technique for a longer period of time, it is evident from figure 6.6 that the flat regions of the PES at various fitnesses can be overcome and a search for the GM can be pursued. This is reflected in the increase in SR as g_{max} increases.

Compared to the work carried out on the IA for the HPLBM on the diamond lattice, the RGA-DE is still outperformed by its artificial immune system rival (using $g_{max} =$ 20,000 and $n_{ind} = 10$). Considering that the magnitude of g_{max} has been increased and that the RGA has been incorporated to increase the number of valid conformations created by the genetic operator, the complexity of the problem seems too great for

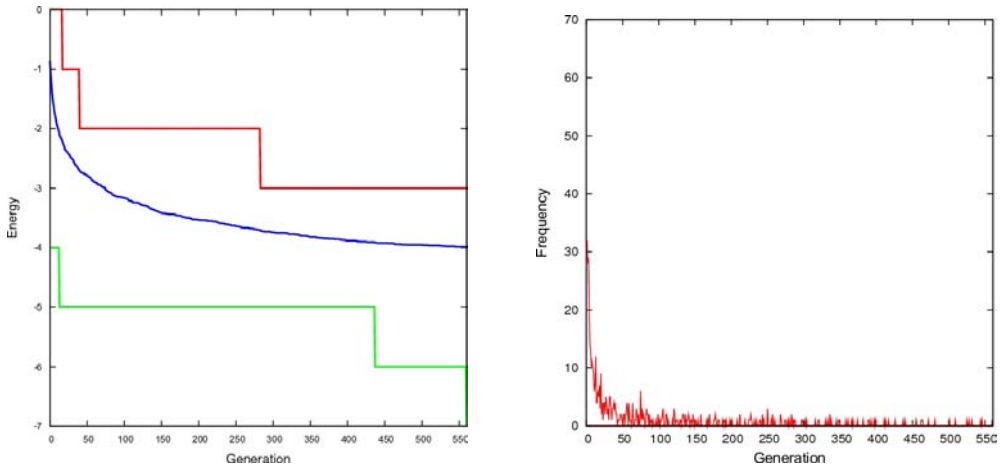
Figure 6.6: Bar charts illustrating (left column) how SR (red) and μ_{FE} (green) and (right column) n_{uniq} (red) and μ_g (green) change with an increasing number of generations with a population size of 200 for 100 runs for the sequences featured in table 4.2 for the HPLBM on the diamond lattice using the RGA.

the RGA-DE to compete with the IA in terms of SR. However, the DE does show promise when increasing the search time (number of generations). The levels of SR are generally lower, with the maximum μ_{FE} being two thirds of that witnessed for the IA. It is clear that the DE favours large values of n_{ind}, unlike the IA. This means that, to remain comparable to the IA, the DE can be further optimised to achieve higher levels of success, while suffering a slight decrease in efficiency, as measured by μ_{FE}. It should be noted that, in order to achieve this comparable performance, the DE has been equipped with the RGA. The stand-alone DE is flawed, in that the methodology itself is not sufficient to produce successful results. When comparing the stand-alone DE to that coupled with the RGA, it is the significant perturbation caused by the RGA that gives rise to such success and not the DE methodology.

Sequence L37 for both the stand-alone DE and the RGA-DE provided success for $n_{ind} = 200$ and $g_{max} = 1000$. Taking a profile of a successful run from each provides insight into how the search progresses in both cases. Figure 6.7(a) shows the energy profile for a successful run for the stand-alone DE. It is apparent that the clock-face method, explained in section 2.3.1, is sufficient to produce individuals of fitness 5 (energy -5 a.u.) early in the search process. Initially, the fittest individuals exhibit an energy of -4 a.u.. This is quickly reduced to -5 a.u. by generation 12. However, the fittest individual does not exhibit a lower energy for over 400 more generations. As the energy profile suggests, and since that the sequence only has four GM conformations, an energy of -6 a.u. is favourable with regard to finding the GM. It is obvious that the mean energy (blue line) for each generation is continually reduced, suggesting that more and more individuals exhibit a lower energy per generation.

Figure 6.7(b) plots the number of successful mutations (mutations resulting in a fitter trial individual than the parent and thus replacing the parent) per generation for the stand-alone DE. The initial sharp decrease in mean energy illustrated here, is supported by the high number of successful mutations early in the calculation. The

starting sequences for the stand-alone DE are generated using the RGA to give the best chance of beginning the search with structurally compact individuals. However, the number of successful mutations plateaus very early, resulting in fewer than five successful mutations per generation of 200 individuals. As fitter individuals tend to have more compact bead arrangements, the stand-alone DE struggles to increase the level of compactness as the mean energy decreases per generation. The ability of the combined mutation mating operator to improve an individual's fitness seems very limited.



(a) (b)

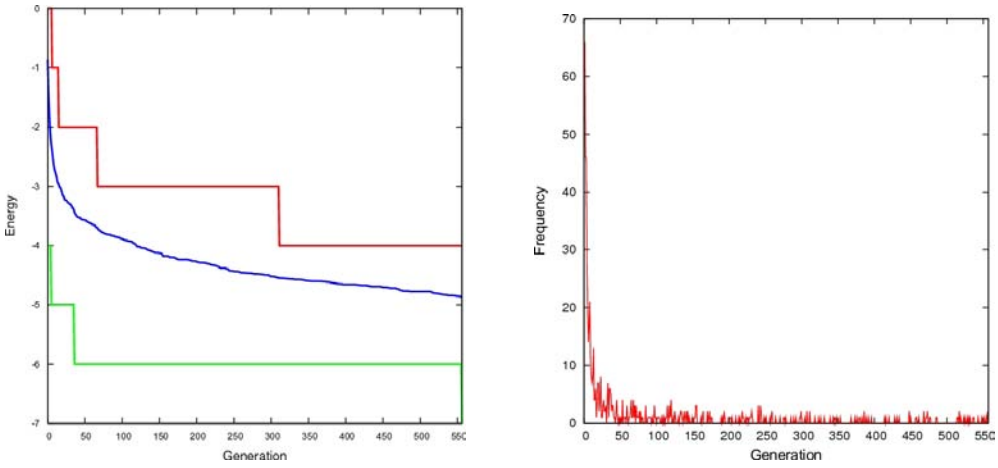
Figure 6.7: (a) A fitness profile for the stand-alone DE with the lowest energy shown in green, mean shown in blue and highest shown in red. (b) Mutation profile illustrating how many successful mutations resulting in an improvement in fitness are performed per generation.

In contrast, figure 6.8(a) illustrates the same energy information for a successful run of sequence L37 for the RGA-DE. When compared to the stand-alone DE, it is obvious that the brute-force nature of the RGA provides a population with fitter individuals. It should be noted that the initial highest and lowest energies do not differ to those of the stand-alone DE for the same sequence. In terms of fitness, both techniques start from equivalent positions. The RGA-influenced mutation procedure allows the search technique to quickly remove low fitness conformations, populating a generation with

more competitive individuals. This is reflected in the sudden decrease in energy for both the best and worst individual in a population. The rapid improvement of these energies, allows a decrease in the mean when compared to the stand-alone DE. The search probes further into the folding troughs of the PES, resulting in a greater chance in finding the GM.

Figure 6.8(b) shows the total number of successful mutations in the same manner as in figure 6.7(b). The number of successful mutations is again much higher initially. Whilst the number of mutations is comparable to that seen for the stand-alone DE, the improvement in the mean fitness suggests that the RGA is able to drive down the energies of already stable conformations. As previously explained, the RGA produces compact structures for the initial population from random conformation vectors in both algorithm interpretations. By providing the RGA with already compact structures (i.e. during the mutation phase), it is able to improve the fitness (and lower the energy), such that the chance of discovering a GM conformation is increased. This is reflected when comparing the levels of success for this with the stand-alone DE, having SRs equal to 39.0% and 3.0%, respectively (for $g_{max} = 1000$ and $n_{ind} = 200$).

In order to compare how effectively both DE interpretations search for the GM, figure 6.9 describes the search process by the frequency of D_H (with respect to the GM) for individuals in a population for all generations. Figure 6.9(a) considers the stand-alone DE and figure 6.9(b) considers the DE coupled with the RGA. The bands illustrate the number of individuals that have a particular D_H with respect to the GM. D_H ranges from 0-17 for the twenty bead sequences considered for the diamond lattice, as the first three beads adopt fixed positions. The darker the band, the greater the number of individuals. It should be noted that, the larger the D_H, the more unrelated the individuals are to the GM in terms of their conformation vector. Figure 6.9(a) exhibits little change in D_H density as the calculation proceeds. D_H consistently ranges from 6-15, with the highest density found in the range 10-13. However, after



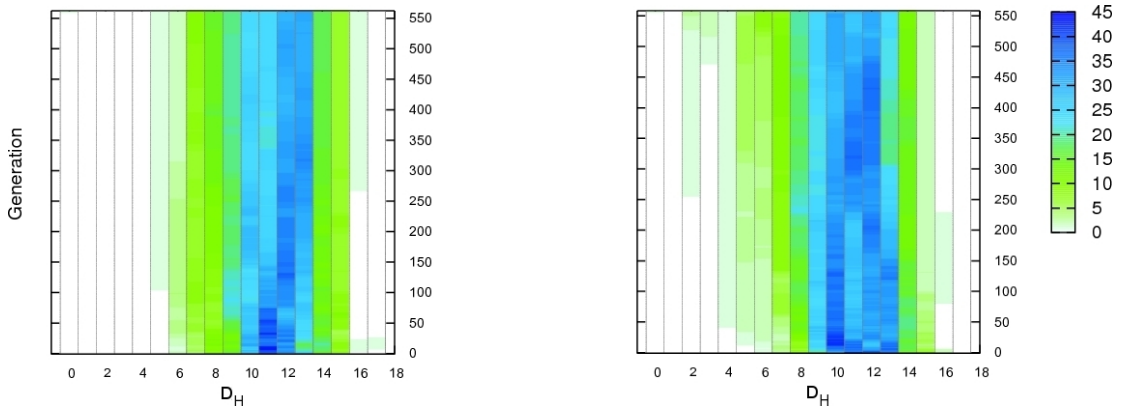
(a) (b)

Figure 6.8: (a) A fitness profile for the RGA coupled DE with the lowest energy shown in green, mean shown in blue and highest shown in red. (b) Mutation profile illustrating how many successful mutations resulting in an improvement in fitness are performed per generation.

around 30 generations, the density at $D_H = 6$ begins to increase. This corresponds to a decrease in energy for the worst individual, as shown in figure 6.7(a). As the search begins to improve structural arrangements, the population becomes more similar to the GM. As there are only four GM (not including mirror images), the search begins to delve deeper into the folding funnels. However, $D_H = 5$ is the lowest seen for the stand-alone DE until the GM is found. This suggests (and is reflected in the SR = 3.0%) that the GM was in fact found accidently, and that the directed element of the search resulted in local minimum trapping.

In contrast, the profile shown in figure 6.9(b) for the RGA coupled DE, shows consistent density in the region 9-13, even in the early stages of the search. This corresponds to the lower mean energy shown in figure 6.8(a) than seen for the energy profile for the stand-alone DE. Due to the incorporation of the RGA, individuals consistently exhibit $D_H = 4$ from generation 40. As the calculation proceeds, a shift in the density begins, with more individuals with $D_H = 5$ and 6. This again reflects the gradual decrease in the mean energy from the profile shown and consequently illustrates the ability of

the RGA-DE to probe the depths of the funnelled landscape. After generation 250, $D_H = 2$ are observed. This implies that a more directed search is employed by the RGA-DE than for the stand-alone DE. It should also be noted, that as well as this DE configuration exhibiting a lower D_H boundary, very few individuals are seen to have high D_H values (16 and 17). Again, this contributes to the lower mean energy seen in the profile.

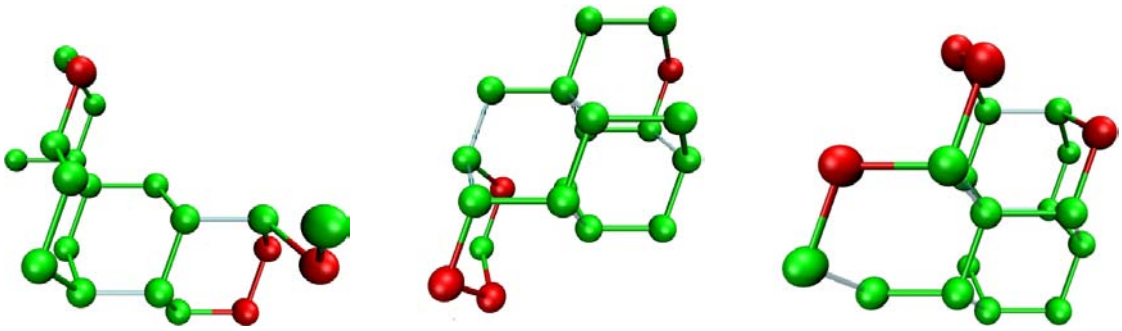


(a) (b)

Figure 6.9: Profiles showing the summation of conformations exhibiting D_H values with respect to the GM and how this changes per generation for (a) the stand-alone DE and (b) the DE coupled with the RGA.

Figure 6.10 shows the GM conformation for sequence L37 and a precursor conformation (for the stand-alone DE in figure 6.10(a) and for the RGA-DE in figure 6.10(b)). As explained previously, the profile in figure 6.9(a) suggested that the search wasn't directed, but in fact the GM was stumbled upon. The precursor conformation in figure 6.10(a) is structurally diverse in terms of compactness and energy. In contrast, the precursor conformation in figure 6.10(b) may be illustrative of the directed search pathway of the RGA-DE, as its energy is two energy levels lower and it is more compact.

As explained in section 2.5, D_H poses certain restrictions on assessment of structural similarity. In order to overcome these limitations, by comparing the RMSD between



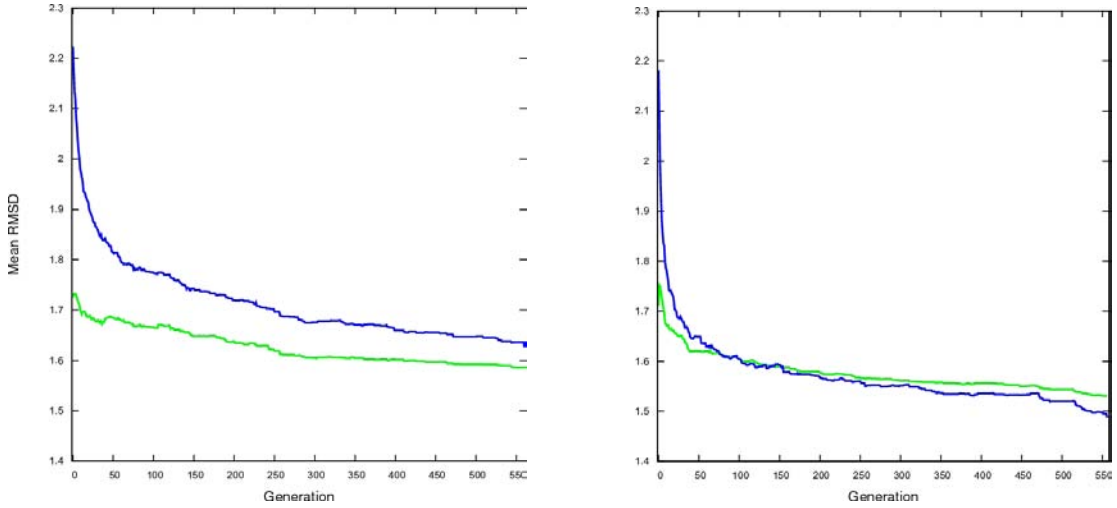
(a) (b) (c)

Figure 6.10: Two precursor GM conformations for the HPLBM on the diamond lattice (a) for the stand-alone DE, $F_{HP} = 4$, (b) for the RGA coupled DE $F_{HP} = 6$ and the GM conformation (c) $F_{HP} = 7$. Topological contacts are shown in transparent cyan.

population members in a pairwise manner, 3D conformations can be measured for structural similarity, not just the conformation vectors. Figure 6.11 illustrates how the mean RMSD within a population and how the population members compare to the GM. Figure 6.11(a) in particular shows this information for the stand-alone DE, whereas figure 6.11(b) illustrates this information with regard to the RGA-DE. For both configurations, it is clear that the mean RMSD with respect to the GM is initially larger than the pairwise values. This suggests that, in terms of 3D conformation, the individuals in a population are more closely related than they are to the GM. In both cases, it is apparent that, as the calculation proceeds, both the μ^P_{RMSD} and the μ^L_{RMSD}, decrease. This suggests that the search itself reduces the number of regions of the PES in which to concentrate its efforts.

However, with regard to the stand-alone DE, the higher mean energy per generation from figure 6.7(a), results from higher energy, less stable conformations in the population. It was hypothesised that the stand-alone DE did not find the GM via a directed search process, but rather via a random one. Figure 6.7(a) supports this idea, in that μ^L_{RMSD} is always higher than the pairwise value, illustrating that the conformations in a population are more closely related to each other than to the GM. In contrast,

for the RGA-DE, in support of the mean energy from the profile from figure 6.8(a), it can be seen, from around generation 80, that the RGA allows for the individuals to become increasingly related to each other and to the GM. Over time, this feature is improved, in that the population members continue to become more closely related to the GM than they do to each other. This is an example of a directed search process, one having the ability to probe a region of the PES, until the bottom of the funnelled landscape is reached, and thus the GM is found.



(a) (b)

Figure 6.11: Profiles showing how mean RMSD changes per generation for (a) the stand-alone DE and (b) the RGA coupled DE. The pairwise mean RMSD for individuals in a population is shown in green, with the mean RMSD with respect to the GM shown in blue.

6.2 BLN Model

Many conformations on a PES may differ by only a single bead placement. As the HPLBM energy function is not distance dependent, many of these related conformations may be equal in terms of energy and fitness. This poses a problem for search techniques in that more compact conformations may appear identical in terms of search criteria to less compact ones. The consequence of this on the PES, is that some regions

	Population Size (n_{ind})																				
	10					25					50					100					200				
ID	**F***	**SR**	μ_{FE}	n_{uniq}	μ_g	**F***	**SR**	μ_{FE}	n_{uniq}	μ_g	**F***	**SR**	μ_{FE}	n_{uniq}	μ_g	**F***	**SR**	μ_{FE}	n_{uniq}	μ_g	**F***	**SR**	μ_{FE}	n_{uniq}	μ_g
H1	-1.46291	1	111.00	1	14.00	-1.45765	1	5025.00	1	317.00	-1.45759	1	27292.00	1	821.00	-1.45759	41	55737.92	4	854.21	-1.45759	26	110019.92	4	835.53
H2	-1.46190	1	12.00	1	0.00	-1.45778	1	27.00	1	0.00	-1.45778	5	20069.20	3	613.80	-1.45778	45	53232.42	4	806.84	-1.45778	35	110982.00	4	840.54
H3	-1.46457	1	12.00	1	0.00	-1.45876	1	12564.00	1	792.00	-1.45778	5	17607.40	3	531.40	-1.45778	51	54683.78	4	826.86	-1.45778	41	117689.36	4	891.82
H4	-1.45745	1	162.00	1	22.00	-1.45478	1	1233.00	1	70.00	-1.45455	2	22713.00	2	674.00	-1.45455	57	50928.28	8	758.71	-1.45455	66	114236.90	8	851.87
H5	-1.45680	1	478.00	1	61.00	-1.45505	1	3602.00	1	211.00	-1.45455	7	22080.71	5	662.57	-1.45455	53	51972.92	8	774.98	-1.45455	62	115581.56	8	861.56
H6	-1.46372	1	173.00	1	22.00	-1.46009	1	7426.00	1	455.00	-1.45635	8	19545.25	6	598.00	-1.45635	69	50332.50	10	761.36	-1.45635	72	114603.02	10	866.00
H7	-1.46043	1	112.00	1	13.00	-1.45641	1	5668.00	1	358.00	-1.45615	3	27620.00	2	837.33	-1.45615	35	51058.02	4	777.34	-1.45615	29	113969.93	4	868.06
H8	-1.46125	1	74.00	1	7.00	-1.45728	1	15365.00	1	924.00	-1.45615	2	21543.00	2	682.50	-1.45615	39	53164.43	4	809.94	-1.45615	34	111154.26	4	846.20
H9	-1.46271	1	41.00	1	4.00	-1.45831	1	3347.00	1	200.00	-1.45625	9	23856.00	7	719.44	-1.45625	63	49944.25	10	751.61	-1.45625	77	112345.59	10	844.64
H10	-1.46355	1	198.00	1	24.00	-1.45648	1	3683.00	1	218.00	-1.45625	6	23410.33	4	705.50	-1.45625	64	50724.95	10	761.39	-1.45625	64	113579.57	10	854.25

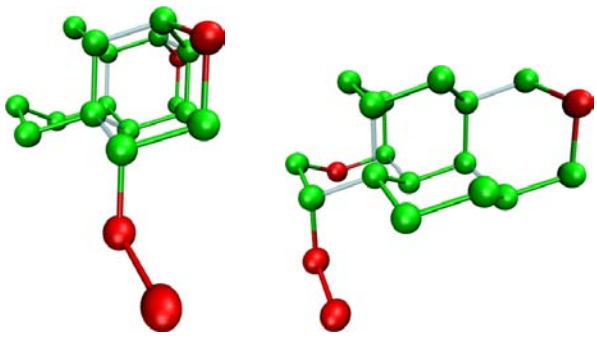
Table 6.1: Statistics for all diamond lattice sequences of high degeneracy using the DE with for BLNM. F* is the best fitness found for a sequence.

	Population Size (n_{ind})																								
	10					25					50					100					200				
ID	**F***	**SR**	μ_{FE}	n_{uniq}	μ_g	**F***	**SR**	μ_{FE}	n_{uniq}	μ_g	**F***	**SR**	μ_{FE}	n_{uniq}	μ_g	**F***	**SR**	μ_{FE}	n_{uniq}	μ_g	**F***	**SR**	μ_{FE}	n_{uniq}	μ_g
H1	-1.45769	1	452.00	1	43.00	-1.45759	5	5372.00	2	212.80	-1.45759	44	20192.90	4	401.81	-1.45759	92	49849.82	4	496.47	-1.45759	100	118760.00	4	591.79
H2	-1.45805	1	1632.00	1	161.00	-1.45778	5	5682.00	2	225.40	-1.45778	38	21708.57	4	432.15	-1.45778	92	46490.04	4	462.89	-1.45778	100	103252.00	4	514.26
H3	-1.45797	1	2142.00	1	212.00	-1.45778	2	6102.00	2	242.00	-1.45778	32	19675.43	4	391.46	-1.45778	84	48155.57	4	479.53	-1.45778	99	115268.66	4	574.33
H4	-1.45478	1	532.00	1	51.00	-1.45455	1	3802.00	1	150.00	-1.45455	35	20284.85	8	403.65	-1.45455	92	43556.34	8	433.54	-1.45455	100	106436.00	8	530.17
H5	-1.45455	1	722.00	1	70.00	-1.45455	8	4864.50	5	192.50	-1.45455	54	15719.59	8	312.35	-1.45455	88	38437.22	8	382.35	-1.45455	100	85730.00	8	426.64
H6	-1.45899	1	202.00	1	18.00	-1.45635	7	9959.14	5	396.28	-1.45635	69	16693.30	10	331.82	-1.45635	100	41609.00	10	414.07	-1.45635	100	85262.00	10	424.30
H7	-1.45634	1	1322.00	1	130.00	-1.45615	2	11027.00	2	439.00	-1.45615	34	26715.23	4	532.26	-1.45615	82	65604.43	4	654.02	-1.45615	100	142766.00	4	711.82
H8	-1.45644	1	772.00	1	75.00	-1.45615	4	4895.75	2	193.75	-1.45615	39	16823.79	4	334.43	-1.45615	83	44794.77	4	445.92	-1.45615	100	97728.00	4	486.63
H9	-1.45827	1	1552.00	1	153.00	-1.45625	4	6264.50	4	248.50	-1.45625	67	17523.64	10	348.43	-1.45625	100	37784.00	10	375.82	-1.45625	100	80566.00	10	400.82
H10	-1.45625	1	2562.00	1	254.00	-1.45625	10	8087.00	7	321.40	-1.45625	58	16927.00	10	336.50	-1.45625	98	37948.93	10	377.46	-1.45625	100	81504.00	10	405.51

Table 6.2: Statistics for all diamond lattice sequences of high degeneracy using the DE coupled with the RGA for the BLNM. F* is the best fitness found for a sequence.

where these conformations lie, appear flat. In order to reduce the flatness of these PES regions on the diamond lattice, the BLNM potential has been coupled with those sequences of low and high degeneracy previously investigated by the HPLBM. The GM fitnesses quoted here are identical to the values gained by work carried out on these sequences using the branch and bound technique systematic search [49].

Table 6.1 lists the best BLNM fitness found ($F*$), SR, μ_{FE}, μ_g and n_{uniq} for all sequences of high degeneracy for various magnitudes of n_{ind} over 100 runs of the stand-alone DE, for the BLNM. It should be noted that n_{uniq} is not necessarily the number of unique GM found, but the the number of unique conformations found of the best fitness quoted in the table.



(a) (b)

Figure 6.12: Two conformations found by the stand-alone DE for sequence H2 with topological contacts shown in transparent cyan. (a) GM, $F_{BLN} = -1.45778$ and $F_{HP} = 5$ (b) lowest energy conformation for $n_{ind} = 10$, $F_{BLN} = -1.46190$ $F_{HP} = 5$. Note how the HPLBM recognises both as GM conformations.

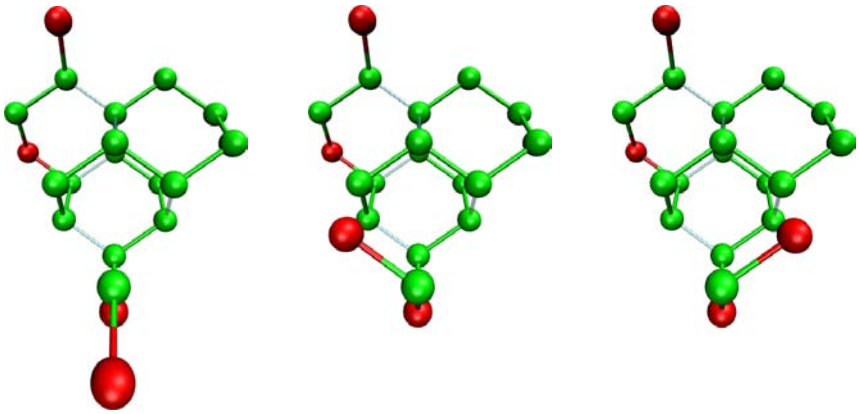
Considering SR alone, it is apparent that the BLNM behaves in the same manner as for other protein models. We notice that our levels of success increase as we increase the magnitude of n_{ind}. However, considering $n_{ind} = 10$ and 25 (highlighted in yellow), $F*$ values do not match the GM fitnesses for the sequences of high degeneracy [49], except for sequence H2 at $n_{ind} = 25$. This is attributed to the reduced amount of genetic material present in a population for low n_{ind}. It has been shown previously that the stand-alone DE finds the production of a sufficient number of self avoiding

walks problematic (as a result of the combined mutation, mating operation). However, for the case of the HPLBM on the diamond lattice, this still resulted in some success, albeit poor, for these magnitudes of n_{ind}. As the sequences tested using the BLNM potential on the diamond lattice are the same as for the HPLBM, this reduction in SR is attributed to the increase in complexity of the PES due to the reduction in smoothness brought about by the BLNM potential. The consequence of this is that more sub-optimal minima are present on the PES. For $n_{ind} = 10$ and 25, the search technique is incapable of avoiding local minimum trapping. This results in a low SR for the best fitness conformation obtained, and, more importantly, SR = 0 with respect to the GM.

For sequence H2, the GM is found using $n_{ind} = 25$. However, as $\mu_g = 0$, the construction phase of the initial population (using the RGA), constructed the GM conformation. This illustrates that the RGA is capable of generating valid, compact conformations.

Figure 6.12(a) shows the constructed GM conformation for $n_{ind} \geq 25$, with figure 6.12(b) showing the lowest energy conformation found for $n_{ind} = 10$. As the HPLBM potential allows the PES to exhibit flat regions, the two conformations shown are in fact both GM. As previously described, the PES of a BLNM protein exhibits more noise and therefore, these HPLBM GM have different energies. It is important to appreciate that the DE is able to search such a noisy surface for $n_{ind} > 25$.

Considering the data for $n_{ind} \geq 100$, it appears that all GM are successfully found by the stand-alone DE. It should be noted that, the higher success rates are witnessed for the sequences with a greater number of minima. Due to the increased complexity of the PES compared with the HPLBM, the number of GM present on the surface is much lower (> 1000 times). The many minima present under the HPLBM are now treated individually due to the different positioning of the P beads and density of the conformations.



(a) (b) (c)

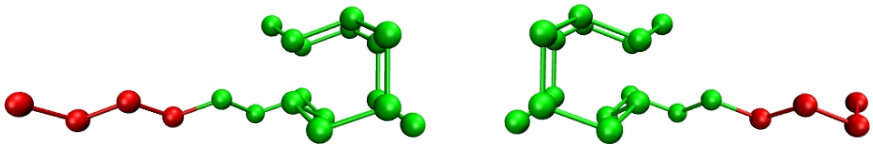
Figure 6.13: An illustration as to how the fitnesses now differ due to P placement between HPLBM GM. (a) GM, $F_{BLN} = -1.45759$, $F_{HP} = 5$ and $\rho = 428.848$ (b) high energy conformation conformation $F_{BLN} = -1.47983$, $F_{HP} = 5$ and $\rho = 410.497$ (c) higher energy conformation conformation $F_{BLN} = -1.48050$, $F_{HP} = 5$ and $\rho = 410.540$. Note how the HPLBM recognises all three as GM conformations.

Figure 6.13 shows the GM conformation found for $n_{ind} \geq 50$ using the stand-alone DE. It also illustrates how changing the position of the final bead (P in this case), gives rise to very different energies (and therefore fitnesses). Figure 6.13(a) highlights how the energy is driven down, by making the distance of the final P from surrounding Hs, as great as possible. The fitness of the conformation in figure 6.13(b) is a little lower (reflecting a higher energy) due to the different distance of the final P bead from the nearby H beads. This is due to the repulsive term (applicable between H and P beads) of the potential described in section 2.4.1.2. The highest energy is calculated for the conformation in figure 6.13(c). The outermost H bead on the right of the conformation does not have an equivalently positioned bead on the left of the conformation. It is the position of this bead that increases the repulsion effect due to the position of the final P, resulting in a higher energy and lower fitness for the conformation.

Comparing the results gained here for the BLNM to those for the HPLBM, both for the diamond lattice, it is evident that the flatness of the PES affects how a search progresses and, ultimately, its success. The funnels present in the landscape of the

BLNM proteins are much narrower, with the wells being less flat. The wealth of minima present in the HPLBM for sequences of high degeneracy, and the stepwise appearance of the surface, govern the difficulty of a search technique to succeed. It should be noted that, although the DE is relatively unsuccessful, with $n_{ind} \leq 25$ for the BLNM, the search technique and methodology alone are sufficient to search the PES for the sequences of reduced degeneracy (listed in table 4.1) reasonably well.

Table 6.2 lists $F*$, SR, μ_{FE}, n_{uniq} and μ_g for the sequences of high degeneracy for the RGA-DE. We have shown that for other protein models, the presence of the RGA increases SR. This, however, is not seen for the BLNM. For $n_{ind} = 10$, an improvement in fitness is seen, but in many cases, the GM is not discovered (highlighted in yellow). The only anomaly for this magnitude of n_{ind} is sequence H5, for which the DE does in fact discover the GM.



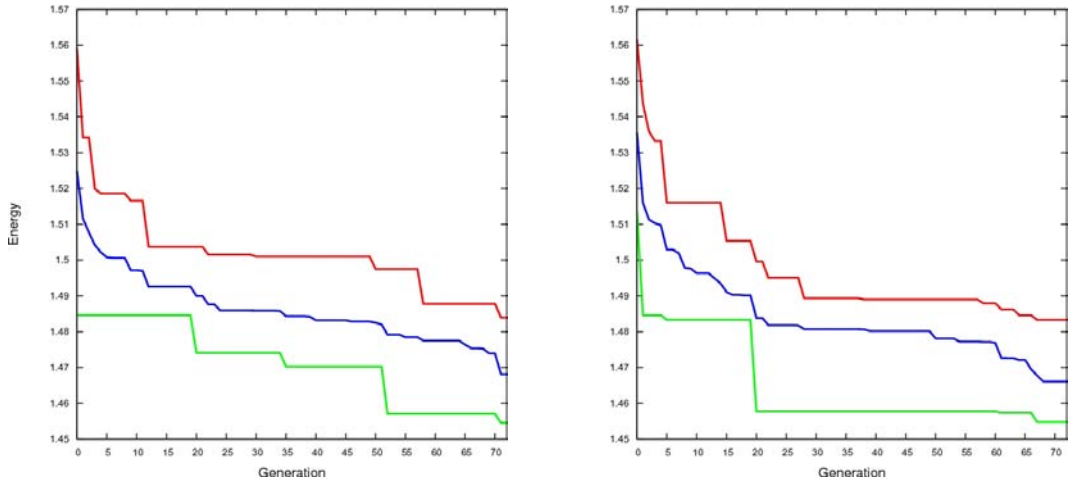
(a) (b)

Figure 6.14: Resultant conformations for sequence H5 using $n_{ind} = 10$ for the DE coupled with the RGA and the BLNM potential. (a) The GM found of $E_{BLN} =$ 1.45455, $E_{HP} =$ -5 (b) A sub-optimal conformation found as a result of a failed run $E_{BLN} =$ 1.45480, $E_{HP} =$ -5.

Figure 6.14 shows two resultant conformations found by the RGA-DE for $n_{ind} = 10$. It can be seen that the conformations satisfy GM criteria for the HPLBM. However, figure 6.14(b) is only a sub-optimal conformation when considering the BLNM, whereas figure 6.14(a) is a GM. As both conformations exhibit the GM requirement for the HPLBM, the terminal P bead in figure 6.14(b) is not correctly positioned to result in the lowest conformation energy. In order for the energy to be lower, the distance between these bead types should be maximised.

In order to reach the GM conformation, it is sometimes necessary for the protein

to unfold to reach the correct region of the PES, where further compacting can take place. In the case of figure 6.14(b) , unfolding is not required to reach the GM, merely the final bead (in the case of H5 it is a P) needs to be repositioned.

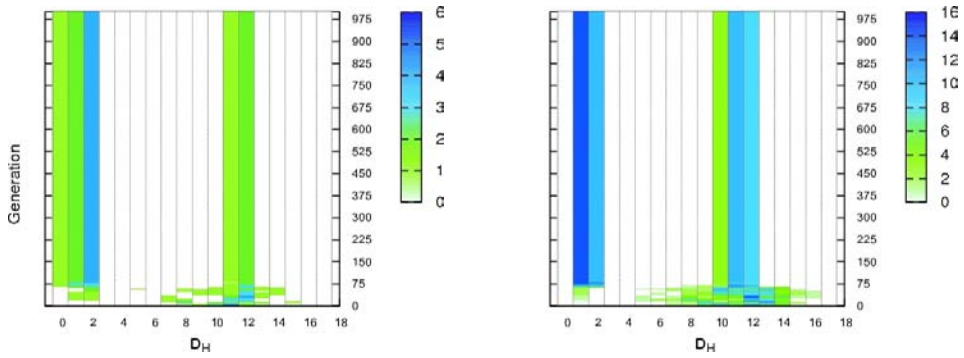


(a) (b)

Figure 6.15: Fitness profiles for (a) the successful run and (b) for an unsuccessful run, showing highest (red), mean (blue) and lowest (green) energies per generation. It should be noted that (b) is truncated due to no change in data.

Figure 6.15 shows energy profiles for both the successful and a failed run for sequence H5 found by the RGA-DE for $n_{ind} = 10$. It can be seen that a more steady decrease in the best energy is witnessed for the successful run, whereas for the failed run, a sudden decrease in best energy is witnessed around generation 20. It is important to note that the mean energy for the unsuccessful run is closer to the highest energy per generation than for the successful run. This implies that more conformations exhibit energies closer to the highest than to the lowest energy values obtained. In contrast, the mean energy for the successful run lies mid-way between the highest and lowest values. This suggests a more even spread of energies throughout the population. In terms of the DE, a more steady decrease in energy is preferred to allow constant updating of the population. As witnessed for the unsuccessful run, the lowest energy structure seen, remains in the population for around 45 generations. This is due to the algorithm not

being able to improve upon the fitness seen for that specific individual, as a result of mutation. Figure 6.15(b) suggests that perhaps the search is trapped in a sub-optimal minimum, and that the nature of the mutation operator is preventing recovery. It should be noted that, although this figure terminates at generation 72, it has in fact been truncated to remove constant energy data.

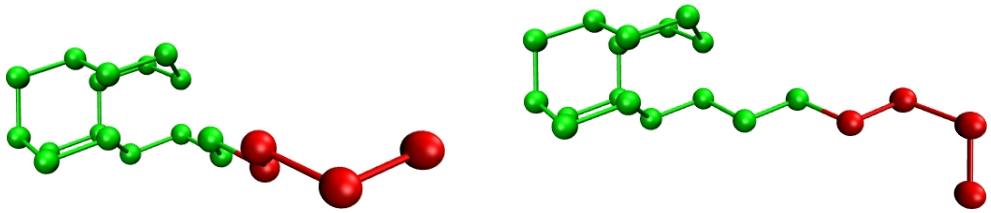


(a) (b)

Figure 6.16: Profiles showing the summation of conformations exhibiting D_H values (a) with respect to the lowest energy conformation and (b) with respect to other individuals of the population, and how they change per generation.

For the unsuccessful run discussed here, very different behaviour is seen for the D_H profiles in figure 6.16. Although the lowest energy conformation was found in generation 69, the data describes the entire run, in order to assess activity beyond this generation. Prior to generation 69, an expected distribution of D_H is seen, with much higher densities observed for $D_H = 11$, 12 and 13. However, a number of conformations exhibit $D_H = 1$ in figure 6.16(b), i.e. these conformations only differ at a single locus in the conformation vector. Considering that a value of $n_{ind} = 10$ is used, this is undesirable, as it reflects poor population diversity. Figure 6.16(a), shows one conformation that is related to the lowest energy conformation by $D_H = 1$. However, once the lowest energy conformation is found, no activity seems to occur after this generation in terms of D_H. Due to the number of generations taken to find the lowest energy conformation for this run being so small, this supports the idea that the search is unable to improve

fitness of the lowest energy individual.



(a) (b)

Figure 6.17: Resultant, sub-optimal conformations for sequence H5 using $n_{ind} = 10$ for the DE coupled with the RGA and the BLNM potential. (a) The lowest energy conformation found of F_{BLN} = -1.45480, $F_{HP} = 5$ (b) A precursor conformation found for this failed run F_{BLN} = -1.45733, $F_{HP} = 5$. They differ by a $D_H = 1$.

Figure 6.17 illustrates that the lowest energy conformation, (figure 6.17(a), a different orientation shown to figure 6.14(b)) exhibits $D_H = 1$ along the protruding polar-terminating tail, in comparison to its precursor conformation in figure 6.17(b). Of course, in terms of the HPLBM, both conformations are considered as GM. However, the distance dependent BLNM potential, also considers repulsive interactions. The repulsion between the polar chain and the compact hydrophobic cluster gives rise to these energy differences and renders the structure in figure 6.17(b) as more unstable.

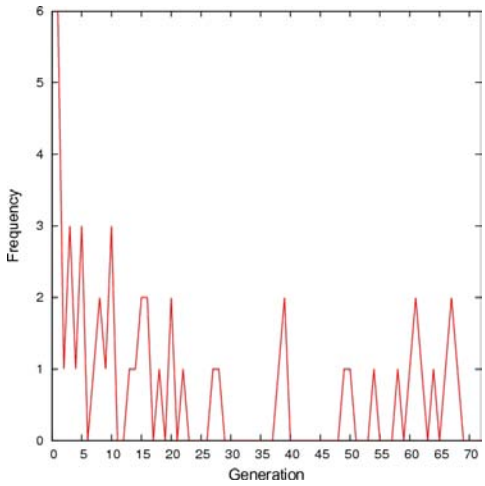


Figure 6.18: A mutation profile for sequence H5 for the failed run showing the number of mutations performed that resulted in an improvement in fitness per generation. The plot has been truncated to remove the lack of data beyond the points shown.

Figure 6.18 shows the number of mutations performed per generation that resulted in an improvement in fitness, thus replacing the individual in the population. It should be noted that this plot has been truncated to remove the lack of activity after generation 69. As previously highlighted, the failure of this run may be attributed to the search getting trapped in a local minimum.

The functionality of the RGA was introduced in section 2.1.1. Structural improvements (i.e. producing a valid conformation from a previous non-self avoiding conformation) are performed by making as few changes to the conformation vector as possible to render the conformation valid. For a single point mutation (resulting in $\Delta D_H = 1$), the conflicting bead must either be towards the non-fixed terminus of an otherwise compact conformation, or lie at any other point of a less dense conformation. A large number of mutations are performed in the initial stages of a search due to the structure compactness, being less than for later generations. The evidence suggests that the fitness of the sub-optimal lowest energy conformation found in this run, was not improved due to the RGA and the standard mutation scheme being incapable. By increasing the level of compactness too rapidly, the ability of the technique to recover is hindered if the sub-optimal configuration being investigated lies in a deep region of the PES.

If the magnitude of n_{ind} is increased, it is obvious from table 6.2, that the magnitude of SR is increased, as seen for the stand-alone DE. As we increase n_{ind}, the level of genetic material present in a population is also increased. It seems that, in order for the RGA-DE to be competitive in terms of SR to the IA, large magnitudes of n_{ind} are also required.

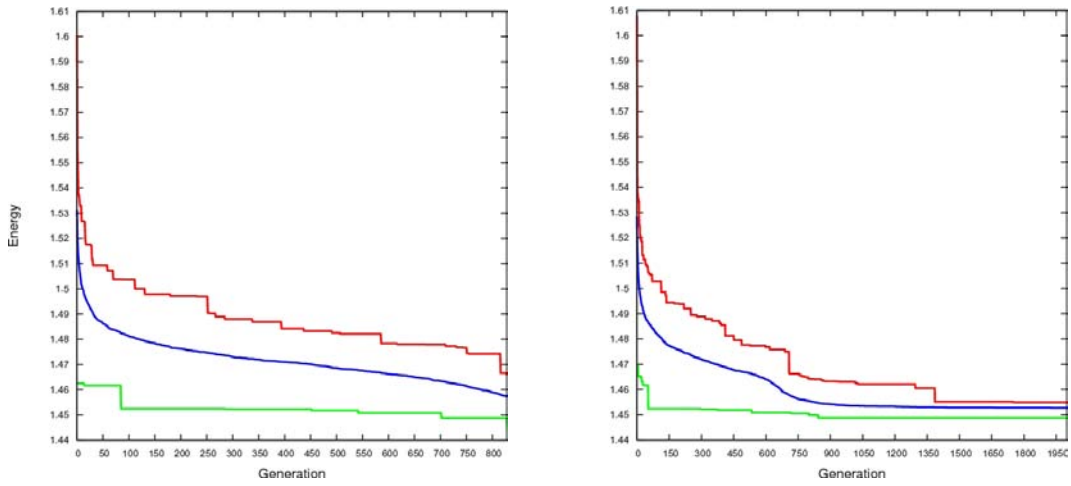
It was observed for low degeneracy sequences, that an increase in the amount of genetic material present in a population led to an increase in SR. It was also shown that the use of the stand-alone DE did not prove beneficial in finding the GM. Table 6.3 shows the results obtained for the sequences of high degeneracy for $n_{ind} = 200$ when

	Maximum Number of Generations (g_{max})																					
	1000					2000					3000					4000					5000				
ID	F*	SR	μ_{FE}	n_{uniq}	μ_g	F*	SR	μ_{FE}	n_{uniq}	μ_g	F*	SR	μ_{FE}	n_{uniq}	μ_g	F*	SR	μ_{FE}	n_{uniq}	μ_g	F*	SR	μ_{FE}	n_{uniq}	μ_g
L1	-1.44382	9	134735.33	2	671.66	-1.44382	21	238544.85	2	1190.71	-1.44382	43	370876.41	2	1852.37	-1.44382	54	440039.03	2	2198.18	-1.44382	69	538381.71	2	2689.89
L2	-1.44217	56	156269.85	2	779.33	-1.44217	91	189626.17	2	946.12	-1.44217	91	189626.17	2	946.12	-1.44217	91	189626.17	2	946.12	-1.44217	91	189626.17	2	946.12
L3	-1.44260	10	126782.00	2	631.90	-1.44260	39	256689.17	2	1281.43	-1.44260	73	364130.76	2	1818.64	-1.44260	92	430536.78	2	2150.67	-1.44260	98	459230.57	2	2294.14
L4	-1.44393	12	154035.33	2	768.16	-1.44393	20	212372.00	2	1059.85	-1.44393	27	289639.03	2	1446.18	-1.44393	35	384504.85	2	1920.51	-1.44393	46	504997.65	2	2522.97
L5	-1.44593	8	150377.00	2	749.87	-1.44593	15	191268.66	2	954.33	-1.44593	15	191268.66	2	954.33	-1.44593	15	191268.66	2	954.33	-1.44593	15	191268.66	2	954.33
L6	-1.44436	9	117668.66	2	586.33	-1.44436	15	184308.66	2	919.53	-1.44436	16	210277.00	2	1049.37	-1.44436	19	285612.52	2	1426.05	-1.44436	20	314282.00	2	1569.40
L7	-1.44593	9	74824.22	2	372.11	-1.44593	27	221439.03	2	1105.18	-1.44593	32	266664.50	2	1331.31	-1.44593	40	350137.00	2	1748.67	-1.44593	49	447565.26	2	2235.81
L8	-1.44436	6	97435.33	2	485.16	-1.44436	6	97435.33	2	485.16	-1.44436	7	149887.71	2	747.42	-1.44436	7	149887.71	2	747.42	-1.44436	7	149887.71	2	747.42
L9	-1.44393	9	135379.77	2	674.88	-1.44393	16	216489.50	2	1080.43	-1.44393	17	232684.35	2	1161.41	-1.44393	18	254957.55	2	1272.77	-1.44393	19	289844.10	2	1447.21
L10	-1.44769	5	138362.00	1	689.80	-1.44769	8	184327.00	2	919.62	-1.44769	8	184327.00	2	919.62	-1.44769	11	309256.54	2	1544.27	-1.44769	11	309256.54	2	1544.27
L11	-1.44436	12	116602.00	2	581.00	-1.44436	13	138294.30	2	689.46	-1.44436	15	180842.00	2	902.20	-1.44436	15	180842.00	2	902.20	-1.44436	15	180842.00	2	902.20
L12	-1.44769	9	147424.22	2	735.11	-1.44769	24	224110.33	2	1118.54	-1.44769	26	246902.00	2	1232.50	-1.44769	35	375499.14	2	1875.48	-1.44769	41	449411.75	2	2245.04
L13	-1.44436	4	55202.00	2	274.00	-1.44436	5	114802.00	2	572.00	-1.44436	5	114802.00	2	572.00	-1.44436	6	203435.33	2	1015.16	-1.44436	8	385477.00	2	1925.37
L14	-1.44393	16	119039.50	2	593.18	-1.44393	17	132107.88	2	658.52	-1.44393	17	132107.88	2	658.52	-1.44393	17	132107.88	2	658.52	-1.44393	17	132107.88	2	658.52
L15	-1.44593	3	122535.33	1	610.66	-1.44593	7	203030.57	2	1013.14	-1.44593	11	318329.27	2	1589.63	-1.44593	14	403544.85	2	2015.71	-1.44593	17	479566.70	2	2395.82
L16	-1.44260	16	134727.00	2	671.62	-1.44260	18	142990.88	2	712.94	-1.44260	18	142990.88	2	712.94	-1.44260	18	142990.88	2	712.94	-1.44260	18	142990.88	2	712.94
L17	-1.44382	10	111702.00	2	556.50	-1.44382	16	189239.50	2	944.18	-1.44382	21	255287.71	2	1274.42	-1.44382	26	342694.30	2	1711.46	-1.44382	27	365127.92	2	1823.62
L18	-1.44160	5	125002.00	1	623.00	-1.44160	12	215935.33	2	1077.66	-1.44160	12	215935.33	2	1077.66	-1.44160	14	271787.71	2	1356.92	-1.44160	19	437612.52	2	2186.05
L19	-1.44206	10	104262.00	2	519.30	-1.44206	14	158059.14	2	788.28	-1.44206	15	174255.33	2	869.26	-1.44206	16	208002.00	2	1038.00	-1.44206	16	208002.00	2	1038.00
L20	-1.44206	11	84347.45	2	419.72	-1.44206	18	181302.00	2	904.50	-1.44206	30	313015.33	2	1563.06	-1.44206	44	432770.18	2	2161.84	-1.44206	54	515753.85	2	2576.75
L21	-1.44593	11	110856.54	2	552.27	-1.44593	20	179152.00	2	893.75	-1.44593	21	191544.85	2	955.71	-1.44593	22	218911.09	2	1092.54	-1.44593	22	218911.09	2	1092.54
L22	-1.44160	6	96502.00	2	480.50	-1.44160	15	210122.00	2	1048.60	-1.44160	16	225914.50	2	1127.56	-1.44160	17	250449.05	2	1250.23	-1.44160	17	250449.05	2	1250.23
L23	-1.44217	48	150635.33	2	751.16	-1.44217	71	176593.54	2	880.95	-1.44217	72	179885.33	2	897.41	-1.44217	73	187769.12	2	936.83	-1.44217	73	187769.12	2	936.83
L24	-1.44270	100	114300.00	2	569.49	-1.44270	100	114300.00	2	569.49	-1.44270	100	114300.00	2	569.49	-1.44270	100	114300.00	2	569.49	-1.44270	100	114300.00	2	569.49
L25	-1.44342	100	123398.00	2	614.98	-1.44342	100	123398.00	2	614.98	-1.44342	100	123398.00	2	614.98	-1.44342	100	123398.00	2	614.98	-1.44342	100	123398.00	2	614.98
L26	-1.44307	6	135568.66	2	675.83	-1.44307	37	284012.81	2	1418.05	-1.44307	60	363065.33	2	1813.31	-1.44307	79	441885.54	2	2207.41	-1.44307	91	499054.74	2	2493.26
L27	-1.44475	13	141694.30	2	706.46	-1.44475	92	275995.47	2	1377.96	-1.44475	99	288812.10	2	1442.05	-1.44475	99	288812.10	2	1442.05	-1.44475	99	288812.10	2	1442.05
L28	-1.44342	24	119118.66	2	593.58	-1.44342	52	210902.00	2	1052.50	-1.44342	67	272079.61	2	1358.38	-1.44342	83	355402.00	2	1775.00	-1.44342	88	384467.90	2	1920.32
L29	-1.44270	3	101468.66	2	505.33	-1.44270	14	259944.85	2	1297.71	-1.44270	49	430144.85	2	2148.71	-1.44270	70	513299.14	2	2564.48	-1.44270	84	576037.71	2	2878.17
L30	-1.44270	29	152726.13	2	761.62	-1.44270	70	220962.00	2	1102.80	-1.44270	78	251894.30	2	1257.46	-1.44270	81	266322.98	2	1329.60	-1.44270	83	279551.39	2	1395.74
L31	-1.44160	6	172402.00	2	860.00	-1.44160	15	225122.00	2	1123.60	-1.44160	23	331036.78	2	1653.17	-1.44160	31	421253.61	2	2104.25	-1.44160	32	435933.25	2	2177.65
L32	-1.44475	40	142702.00	2	711.50	-1.44475	85	219166.70	2	1093.82	-1.44475	93	241855.76	2	1207.26	-1.44475	96	254291.58	2	1269.44	-1.44475	97	260247.36	2	1299.22
L33	-1.44307	8	114752.00	2	571.75	-1.44307	12	161518.66	2	805.58	-1.44307	13	184340.46	2	919.69	-1.44307	16	278702.00	2	1391.50	-1.44307	17	313437.29	2	1565.17
L34	-1.44094	2	122402.00	1	610.00	-1.44094	26	310732.76	2	1551.65	-1.44094	69	426051.27	2	2128.24	-1.44094	84	466635.33	2	2331.16	-1.44094	88	486245.18	2	2429.21
L35	-1.44166	16	119639.50	2	596.18	-1.44166	28	194444.85	2	970.21	-1.44166	33	237377.75	2	1184.87	-1.44166	36	281568.66	2	1405.83	-1.44166	42	374011.52	2	1868.04
L36	-1.44094	16	146002.00	2	728.00	-1.44094	86	255043.86	2	1273.20	-1.44094	96	276377.00	2	1379.87	-1.44094	97	281210.24	2	1404.04	-1.44094	98	287187.71	2	1433.92
L37	-1.44342	99	103717.15	2	516.57	-1.44342	99	103717.15	2	516.57	-1.44342	99	103717.15	2	516.57	-1.44342	99	103717.15	2	516.57	-1.44342	99	103717.15	2	516.57
L38	-1.44514	22	131674.72	4	656.36	-1.44514	54	227150.14	4	1133.74	-1.44514	60	254685.33	4	1271.41	-1.44514	65	288322.00	4	1439.60	-1.44514	69	322062.86	4	1608.30
L39	-1.44166	9	129824.22	2	647.11	-1.44166	17	200943.17	2	1002.70	-1.44166	22	264365.63	2	1319.81	-1.44166	23	287558.52	2	1435.78	-1.44166	26	352548.15	2	1760.73
L40	-1.44094	10	133022.00	2	663.10	-1.44094	55	276725.63	2	1381.61	-1.44094	77	334913.68	2	1672.55	-1.44094	86	368706.65	2	1841.52	-1.44094	90	391186.44	2	1953.92
L41	-1.44270	14	142202.00	2	709.00	-1.44270	26	210125.07	2	1048.61	-1.44270	31	256060.06	2	1278.29	-1.44270	31	256060.06	2	1278.29	-1.44270	32	273170.75	2	1363.84
L42	-1.44270	99	97917.15	2	487.57	-1.44270	99	97917.15	2	487.57	-1.44270	99	97917.15	2	487.57	-1.44270	99	97917.15	2	487.57	-1.44270	99	97917.15	2	487.57
L43	-1.44270	16	134539.50	2	670.68	-1.44270	39	228494.30	2	1140.46	-1.44270	57	314402.00	2	1570.00	-1.44270	63	349189.30	2	1743.93	-1.44270	66	373547.45	2	1865.72
L44	-1.44094	26	149186.61	2	743.92	-1.44094	93	241651.46	2	1206.24	-1.44094	96	248520.75	2	1240.59	-1.44094	98	256650.97	2	1281.24	-1.44094	98	256650.97	2	1281.24
L45	-1.44342	26	155394.30	2	774.96	-1.44342	53	219715.20	2	1096.56	-1.44342	63	260938.50	2	1302.68	-1.44342	72	309965.88	2	1547.81	-1.44342	75	333340.66	2	1664.69
L46	-1.44514	16	132477.00	4	660.37	-1.44514	48	232422.83	4	1160.10	-1.44514	53	255386.90	4	1274.92	-1.44514	54	262179.77	4	1308.88	-1.44514	56	286462.71	4	1430.30
L47	-1.44160	23	154549.82	2	770.73	-1.44160	42	198925.80	2	992.61	-1.44160	44	215938.36	2	1077.68	-1.44160	45	227948.66	2	1137.73	-1.44160	49	278863.22	2	1392.30
L48	-1.44286	100	101142.00	2	503.70	-1.44286	100	101142.00	2	503.70	-1.44286	100	101142.00	2	503.70	-1.44286	100	101142.00	2	503.70	-1.44286	100	101142.00	2	503.70

Table 6.3: Statistics for all diamond lattice sequences of low degeneracy using the RGA-DE for the BLNM. F* is the best fitness found for a sequence. Rows highlighted in yellow indicate sequences that did not show an increase in SR due to the increase in g_{max}. Rows highlighted in green indicate the sequences that could not be improved due to having SR = 100.

using the RGA-DE.

According to the table, however, SR for a number of sequences were unable to improve for larger g_{max}. Sequences, L24, L25 and L48 resulted in a SR of 100%, with all GM discovered for $g_{max} = 1000$. However, for sequences L37 and L42 (mirrors of L25 and L24, respectively) we were unable to find the GM conformation every time. Both sequences resulted in a SR = 99% for $g_{max} \geq 1000$. As all statistics are averaged over one hundred runs of the DE, this signifies that only a single run, regardless of an increase in g_{max}, failed to determine the GM conformation. Figure 6.19 illustrates the fitness profiles for both a successful run (GM found in generation 827, figure 6.19(a)) and the unsuccessful run (lowest energy conformation found in generation 842, figure 6.19(b)), with $g_{max} = 2000$.

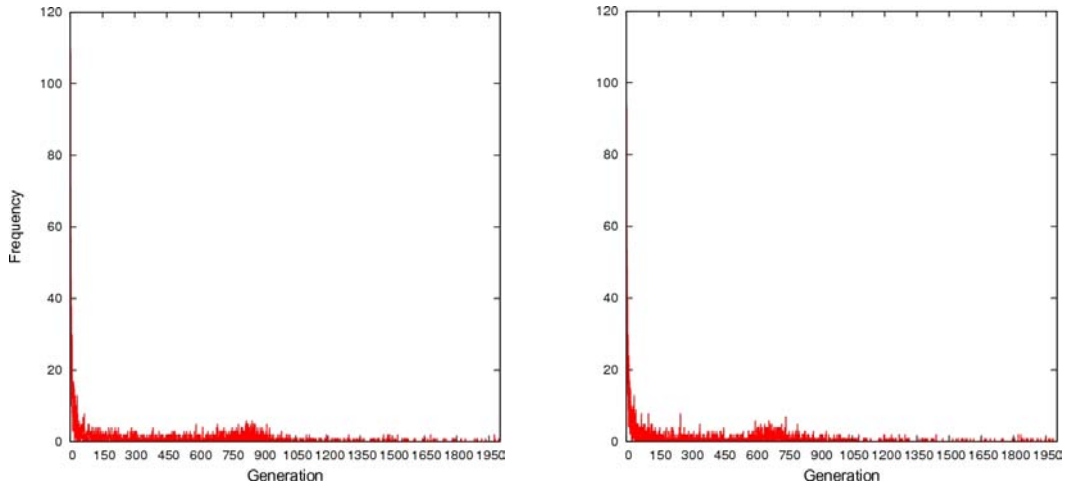


(a) (b)

Figure 6.19: Fitness profiles for (a) the successful run and (b) for an unsuccessful run for sequence L37, showing highest (red), mean (blue) and lowest (green) energies per generation..

In both cases, the highest and lowest energies for the starting population are less for the successful case than for the unsuccessful one. As the initial population provides a starting point for all all population based search techniques, the integrity of the individuals present will begin to drive the search in a certain direction. Upon inspection,

the unsuccessful run exhibits a close relationship between the highest, lowest and mean energies much sooner than for the successful run. This may be attributed to the search getting trapped in an unfavourable (not GM related) energy well present on the PES, in which the favourable regions of local structure begin to dominate the population. For the case of figure 6.19(b), soon after the lowest energy conformation is achieved, the mean energy does not tend to change. As the highest energy decreases for another 400 generations, this implies that only a small number of mutations are successful per generation, involving only the most unstable conformations within a population. This suggests that the search itself became trapped in a local minimum well from which it could not escape. Data are shown for $g_{max} = 2000$, with the same run failing for a further three thousand generations.

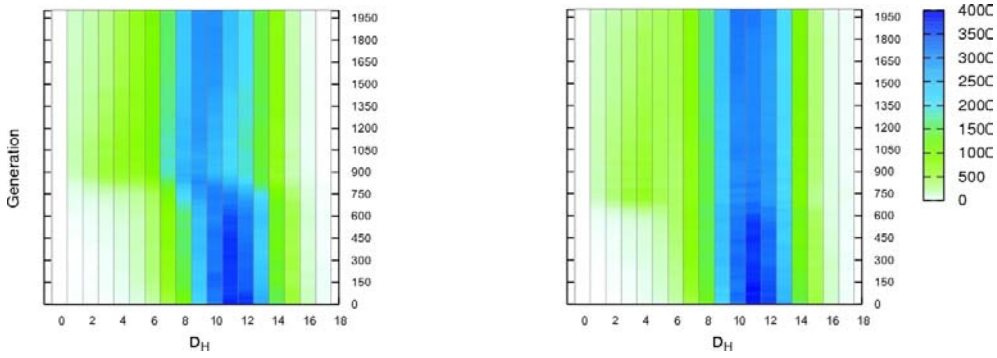


(a) (b)

Figure 6.20: Mutation profiles for (a) the successful run and (b) for a failed run, illustrating the number of successful mutations performed per generation for sequence L37. (a) is extended past the GM generation for comparative reasons.

With the mean energy stabilising after finding the lowest energy conformation for the unsuccessful case, the number of successful mutations per generation comes into question. Figure 6.21 illustrates this for both the successful case (figure 6.21(a)) and the unsuccessful case (figure 6.21(b)). As expected, the initial number of successful

mutations per generation is high and decreases rapidly. It can be seen, from both figures that a small increase in the rate of successful mutations occurs at around generation 750. This corresponds in both cases to a sharper decrease in mean, highest and lowest energies, and is supported by figure 6.19. Once the GM (for the successful case) and the lowest energy conformation (in the unsuccessful case) are found, the calculations behave in a similar manner in terms of the number of successful mutations performed.

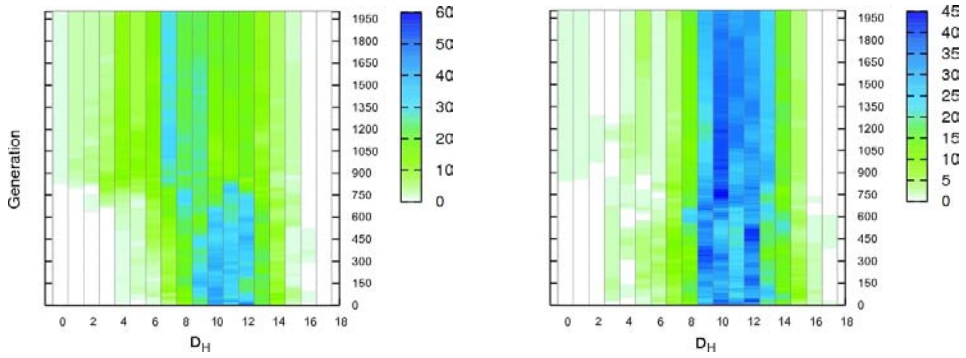


(a) (b)

Figure 6.21: Pairwise D_H profiles showing the population density for individuals of a population when compared to each other for (a) the successful run and (b) for a failed run for sequence L37. (a) is extended past the GM generation for comparative reasons.

Figure 6.21 illustrates how the individuals of a population relate to each other in terms of D_H for both cases. It is apparent that, for both cases, the D_H maximum density lies in the range 9 - 13. However, for the unsuccessful case, this maximum density region never shifts, as it does for the successful case. The shift in population diversity for the successful case is favourable if the conformations present in a population provide a path to the GM. This shift occurs around 50 generations before discovering the GM, and is responsible for the sudden success for this sequence. An expansion of D_H range is seen in both cases, as the population begins to adopt favourable sections of local structure.

Maintaining population diversity is the key to prevention of local minimum trapping. However, if the population is not diverse, or this diversity shifts, it is beneficial to



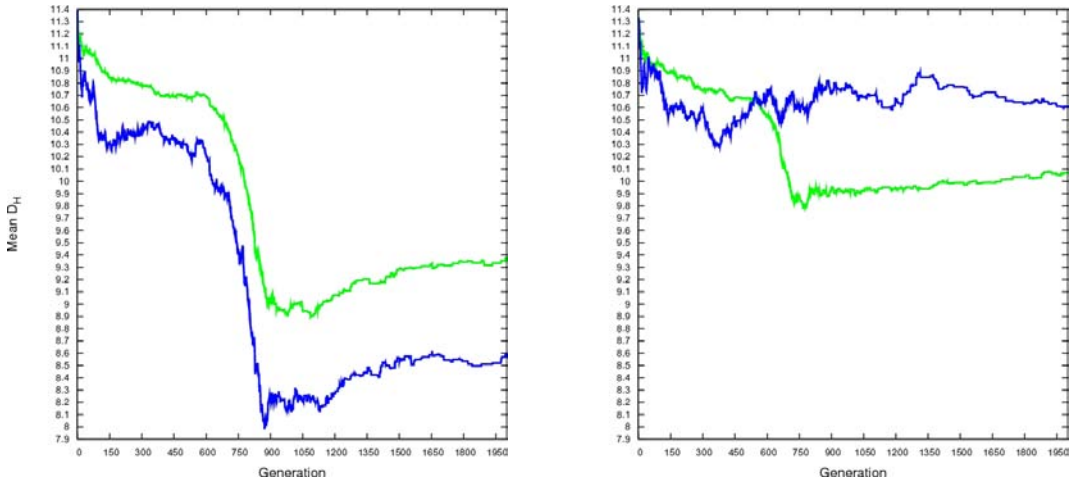
(a) (b)

Figure 6.22: GM D_H profiles showing the population density for individuals of a population when compared to the best found conformation for (a) the successful case and (b) for an unsuccessful case for sequence L37. (a) is extended past the GM generation for comparative reasons.

know whether the shift is positive or not. Figure 6.22 shows how diverse the population is with respect to the GM in terms of D_H. Again the outcomes are very different for the two situations. As before, with the pairwise case, the most dense region spans D_H = 9 - 13 for the failed case, shown in figure 6.22(b). Unfortunately, this range never fluctuates throughout 2,000 generations. However, in the case of the successful run, a slightly shorter range is seen for the initially most dense region. A D_H range of 8 - 12 exists up to the point in the calculation where the surge of mutations occurs (around generation 750). These mutations offer favourable regions of local structure that are beneficial to finding the GM conformation. This is illustrated by the disappearance of the most dense region, and the gradual transition to acceptance of more GM-like conformations.

Although an increase in the number of successful mutations per generation is seen for the failed case, they do not contain favourable regions of local structure that could possibly contribute to the GM. This is evidenced by the lack of shift in population diversity with respect to the GM, with the run exhibiting a very small percentage of conformations that lie close to the GM. Figure 6.22(b) supports the hypothesis that the search became trapped in a local minimum and was unable to recover.

For the successful run, figures 6.21(a) and 6.22(a) illustrate how the individuals in a population become more like each other, subsequently becoming more like the GM. It is evident that a directed search process took place, and not a lucky mutation that gave rise to the GM.

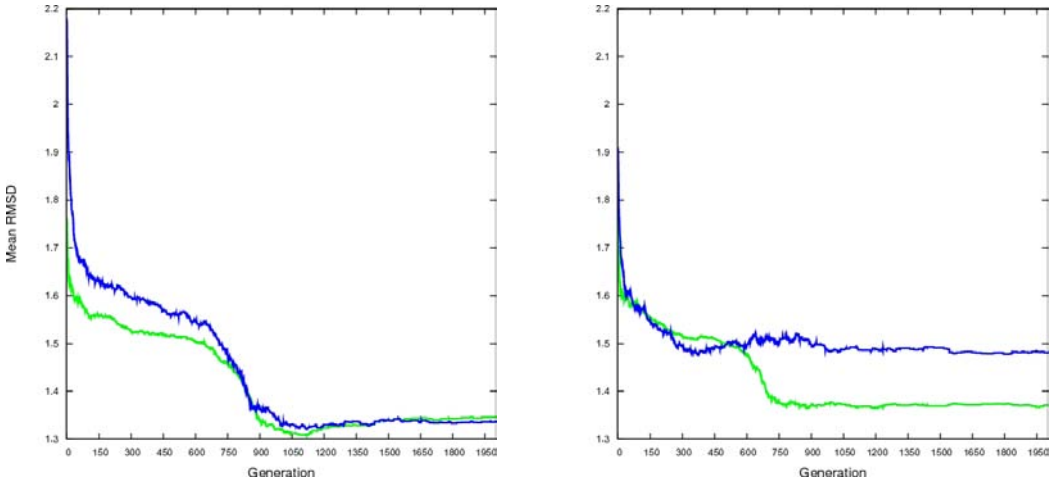


(a) (b)

Figure 6.23: D_H profiles showing the mean values when individuals in a population are compared to each other (green) and the lowest energy conformation (blue) for (a) the successful run and (b) for a failed run for sequence L37. (a) is extended past the GM generation for comparative reasons.

Figure 6.23 illustrates the mean D_H with regard to individuals in a population and the lowest energy conformation. For directed search processes, a gradual decrease in mean D_H is expected. In terms of D_H, figure 6.23(a) supports this idea, illustrating how a sharp decrease in D_H, both with respect to each individual and between each individual and the GM, is apparent up to the "GM-found" generation. This sharp decrease in mean D_H occurs when an increase in successful mutations occurs in figure 6.20(a). For a run to be successful, both how the individuals relate to each other and how they relate to the GM are critically important. For success, the similarity of the population towards the GM should be greater on average than between individuals. This is illustrated by the mean D_H with respect to the GM, being lower than that for the individuals themselves. In contrast, the failed run looks promising up to around

generation 400. The same features are observed as for the successful case up to this point. A decrease in mean D_H with regard to both the GM and the individuals themselves is witnessed. However, the increase in the number of successful mutations per generation, sees the shape of the run change dramatically. The individuals start to become more closely related, however, with respect to the lowest energy conformation, the relationship begins to grow further apart. Unfortunately, the explored region of the PES does not contain a GM.

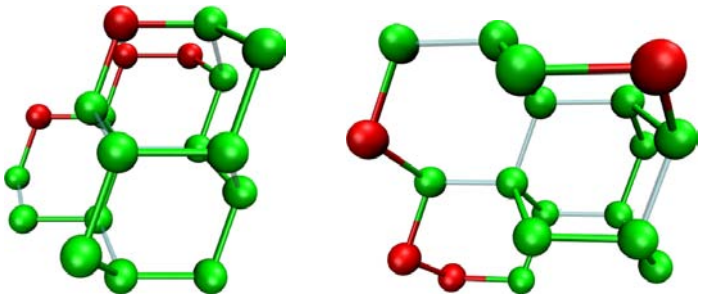


(a) (b)

Figure 6.24: RMSD profiles showing the mean values when individuals in a population are compared to each other (green) and the lowest energy conformation (blue) for (a) the successful run and (b) for a failed run for sequence L37. (a) is extended past the GM generation for comparative reasons.

Figure 6.24 illustrates how mean RMSD changes with respect to the lowest energy conformation and within the population. It is not necessarily expected for the shape of these profiles to be the same as figure 6.23. However, by sharing features, these profiles are able to support the conclusions as well as provide an insight into how effective D_H is at quantifying structural change. In figure 6.24(a), it is apparent that up to the "GM-found" generation, the individuals of the population are more structurally similar to each other than to the GM. However, in the case of figure 6.24(b), at around generation 600, the individuals become more closely related to each other (lower RMSD) and to

the lowest energy conformation found in generation 842.



(a) (b)

Figure 6.25: Lowest energy conformations for (a) the successful run and thus the GM and (b) for a failed run showing a sub-optimal minimum. Pairwise atom distances give 407.87 and 411.78 degrees of compactness respectively. Conformation vectors 02110121011220122 and 20102010010001022 respectively, are shown.

Comparing the lowest energy conformations themselves, figure 6.25, shows that the unsuccessful search did in fact probe a non-GM region of the PES, getting trapped in a sub-optimal minimum and not being able to recover. Even as far as the HPLBM is concerned, the conformation in figure 6.25(b), is not a GM. If we consider the pairwise atom distance as a measure of compactness, the GM is more compact than the sub-optimal minimum, as expected. The conformation vectors differ by $D_H = 12$ and RMSD = 1.5521. In order to produce the conformation in figure 6.25(a), the first topological interaction must be produced between H beads 1 and 9. The sub-optimal conformation in figure 6.25(b) requires the first topological contact to be produced between H beads 1 and 7. In order to produce the bonding required, figure 6.25(b) would have to completely unfold, breaking all topological contacts and overcoming a large energy barrier of 6 energy units (in terms of the HPLBM).

6.3 Dynamic Lattice Model

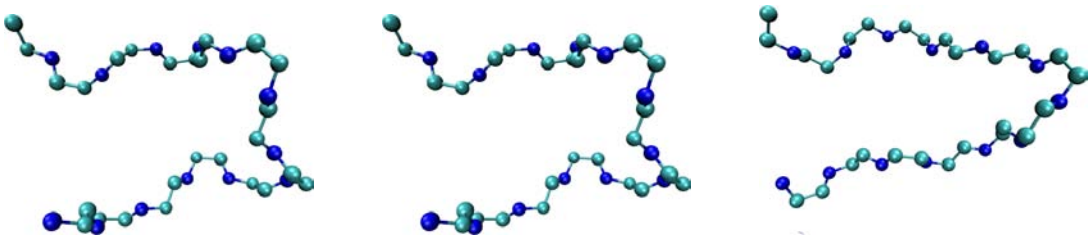
For the square lattice, the placement of the initial two beads does not contribute a change to the conformation energy and thus they were fixed. Likewise, for the diamond lattice the first three beads were fixed. This resulted in the size of the conformation

vectors being two less than the sequence length for the square and three less than the sequence length for the diamond lattice. For the case of the dynamic lattice model, the protein chain is characterised by ϕ , ψ torsion angle pairs, as described in section 2.4.2. The first atom in the chain involved in a torsion is the nitrogen of the second residue. This makes a torsion angle with the initial nitrogen, and thus adopts the ψ angle specified by the initial residue. For this reason, the length of the conformation vector for DLM proteins is the same as for number of residues in the chain.

Table 6.4 shows how the levels of success achieved for the RGA-DE change for a fixed n_{ind} of 200 and various magnitudes of g_{max}, for the sequences listed in table 5.1. As seen for the BLNM, by increasing the magnitude of g_{max}, SR increased for some sequences. The table shows that, for this particular model, the magnitudes of SR reach their peak for runs of $g_{max} = 2000$. This suggests, as seen for the BLNM, that the search is getting trapped in a sub-optimal minimum of the PES. The search space available to the DLM proteins is dependent on the residues in the sequence. It was shown in section 2.4.2 that the residues can adopt a varying number of torsion angle pairs. In some cases this may lead to an increase in the complexity of the search space, when compared to the previous models studied. The fitnesses observed in table 6.4 are comparable to those obtained using evolutionary algorithms [117], with the exception of 1QHK (highlighted in yellow).

It should also be noted that the search claims to find two unique minima that exhibit a fitness of 1.51775 for sequence 1A1P. In terms of conformation vector, the search does in fact find two unique GM. However, as previously explained, the first torsion angle (ψ) requires the first two nitrogen atoms to be placed, with a torsion defined by the first and second residues. As with real proteins, the first ϕ angle is defined by backbone carbon atoms 2 and 3, the first carbon atom is not involved in producing a torsion angle. The conformation vectors of the two unique individuals are 0302200312130 and 3302200312130, so they differ by a single bit change at the

first position. According to the DLM torsion angles in table 2.4, the first and final torsion entries for Ile are (-125,130) and (-95,130), respectively. The ψ torsion angles are identical, and thus so are the minima in terms of 3D conformation, as shown in figure 6.26.



(a) (b) (c)

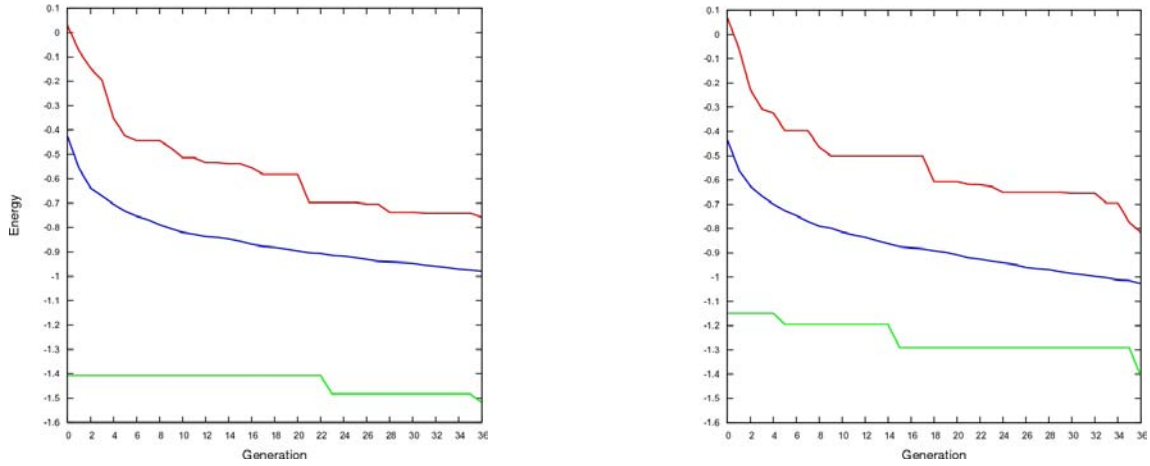
Figure 6.26: The GM obtained by the RGA-DE with conformation vectors (a) 0302200312130 and (b) 3302200312130. (c) The experimental structure from the PDB.

Sequence 1A1P has a SR = 99% for all values of g_{max}. As previously hypothesised, this may be due to the search becoming trapped in a local minimum, resulting in never being able to find the GM, regardless of how many generations the calculation is run for. For the unsuccessful run, the lowest energy conformation was discovered in generation 35, with analysis focusing on these generations, with a comparison to a successful case of the same duration.

Figure 6.27 illustrates how the highest, lowest and mean energies fluctuate as a function of generation for both a successful case and the unsuccessful case for sequence 1A1P. As observed for all energy profiles, both the mean and highest energies for a population undergo a dramatic decrease over the first few generations, as mutation operators drive down the energies of uncompact starting conformations. However, the successful case shows unfavourable behaviour with respect to the mean, as the population are skewed to higher energies. The unsuccessful case illustrates a mean lying almost equidistant from the highest and lowest energies for a population. This suggests that both high and low energy conformations are present in the population in equal quantities. However, although the successful case does not exhibit this behaviour,

	Maximum Number of Generations (g_{max})																				
	1000					2000					3000					4000					5000				
ID	F*	SR	μ_{FE}	n_{uniq}	μ_g	F*	SR	μ_{FE}	n_{uniq}	μ_g	F*	SR	μ_{FE}	n_{uniq}	μ_g	F*	SR	μ_{FE}	n_{uniq}	μ_g	F*	SR	μ_{FE}	n_{uniq}	μ_g
1AL1	1.56250	100	4001.14	1	17.93	1.56250	100	4001.14	1	17.93	1.56250	100	4001.14	1	17.93	1.56250	100	4001.14	1	17.93	1.56250	100	4001.14	1	17.93
1A1P	1.51775	99	18230.80	2	89.15	1.51775	99	18230.80	2	89.15	1.51775	99	18230.80	2	89.15	1.51775	99	18230.80	2	89.15	1.51775	99	18230.80	2	89.15
1AKG	1.17301	100	14388.01	1	69.93	1.17301	100	14388.01	1	69.93	1.17301	100	14388.01	1	69.93	1.17301	100	14388.01	1	69.93	1.17301	100	14388.01	1	69.93
1L2Y	0.87846	78	26814.84	2	132.06	0.87846	78	26814.84	2	132.06	0.87846	78	26814.84	2	132.06	0.87846	78	26814.84	2	132.06	0.87846	78	26814.84	2	132.06
1D9J	1.34457	96	12949.96	1	62.73	1.34457	96	12949.96	1	62.73	1.34457	96	12949.96	1	62.73	1.34457	96	12949.96	1	62.73	1.34457	96	12949.96	1	62.73
1B19:A	3.54363	100	18900.00	1	92.49	3.54363	100	18900.00	1	92.49	3.54363	100	18900.00	1	92.49	3.54363	100	18900.00	1	92.49	3.54363	100	18900.00	1	92.49
1G04	1.24277	86	18497.34	1	90.47	1.24277	86	18497.34	1	90.47	1.24277	86	18497.34	1	90.47	1.24277	86	18497.34	1	90.47	1.24277	86	18497.34	1	90.47
1ANP	2.37581	2	178802.00	2	892.00	2.37581	2	178802.00	2	892.00	2.37581	2	178802.00	2	892.00	2.37581	2	178802.00	2	892.00	2.37581	2	178802.00	2	892.00
1AML	4.91472	5	163322.00	1	814.60	4.91472	14	202873.42	1	1012.35	4.91472	14	202873.42	1	1012.35	4.91472	14	202873.42	1	1012.35	4.91472	14	202873.42	1	1012.35
1QHK	3.75158	2	188302.00	1	939.50	3.78268	1	273802.00	1	1367.00	3.78268	1	273802.00	1	1367.00	3.78268	1	273802.00	1	1367.00	3.78268	1	273802.00	1	1367.00

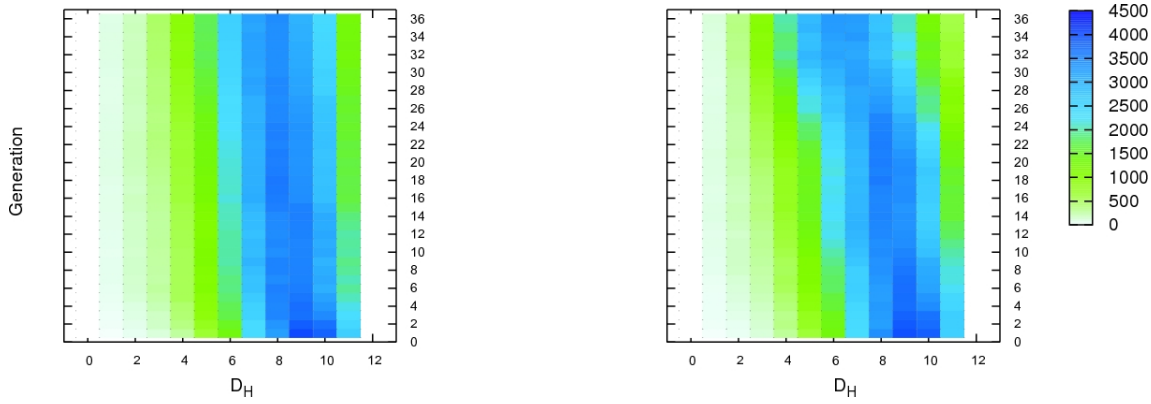
Table 6.4: Statistics for all dynamic lattice sequences using the DE coupled with the RGA for the DLM. F* is the best fitness found for a sequence.



(a) (b)

Figure 6.27: The fluctuation of highest (red), mean (blue) and lowest (green) energies in a population as a function of generation for (a) the successful case and (b) the unsuccessful case.

equal quantities of high and low energy conformations may exist, with a single, much lower energy conformation present than for the unsuccessful case. For both cases, the high and mean energies exhibit similar values throughout, indicating that this is indeed the case.

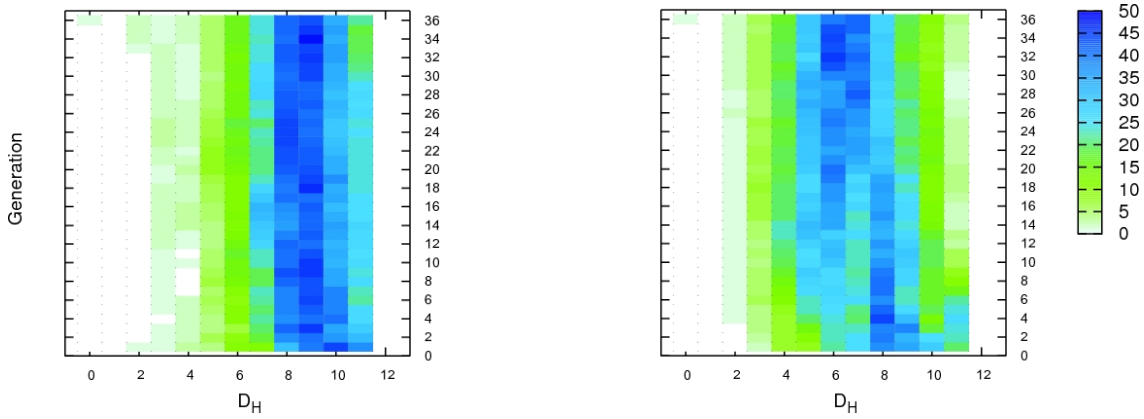


(a) (b)

Figure 6.28: How population diversity with respect to D_H fluctuates as a function of generation for (a) the successful case and (b) the unsuccessful case.

The spread of population diversity is characterised in figure 6.28 for both cases. Both cases demonstrate an initial high population diversity with respect to D_H, with

magnitudes of D_H lying in the region 7 - 11. Once the relatively non-compact starting conformations undergo mutation, the diversity shows a decrease as expected. The successful case illustrates how a large population diversity is maintained with respect to D_H, demonstrating consistant magnitudes between 6 and 10 after this point. However, the unsuccessful case shows a continual drop in diversity approaching generation 36. This suggests that, not only is a diverse population beneficial to the DLM for large magnitudes of n_{ind}, but the GM may have been found quickly as a result of a lower energy conformation present from the start. The continuous decrease in diversity for the unsuccessful case, suggests that the search is trapped in a local minimum.

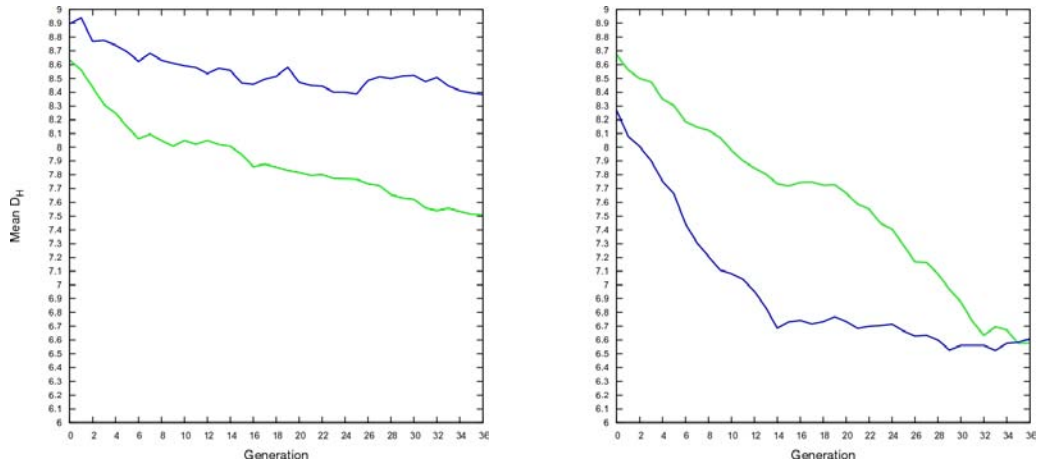


(a) (b)

Figure 6.29: How population diversity with respect to D_H fluctuates as a function of generation for (a) the successful case and (b) the unsuccessful case when compared to the lowest energy conformation.

By considering the population diversity with regard to the lowest energy conformation (the GM for the successful case), the successful case illustrates a greater degree of diversity than the unsuccessful case. This supports the theory of local minimum trapping for the unsuccessful case, with the populations exhibiting a close relationship to the lowest energy conformation from the outset.

Very different behaviour is observed when measuring μ_{DH}^{P} and μ_{DH}^{L} for both cases. As expected, the magnitudes of μ_{DH}^{P} and μ_{DH}^{L} start high, and decrease dramatically once the non-compact starting individuals begin to mutate. However, with respect to

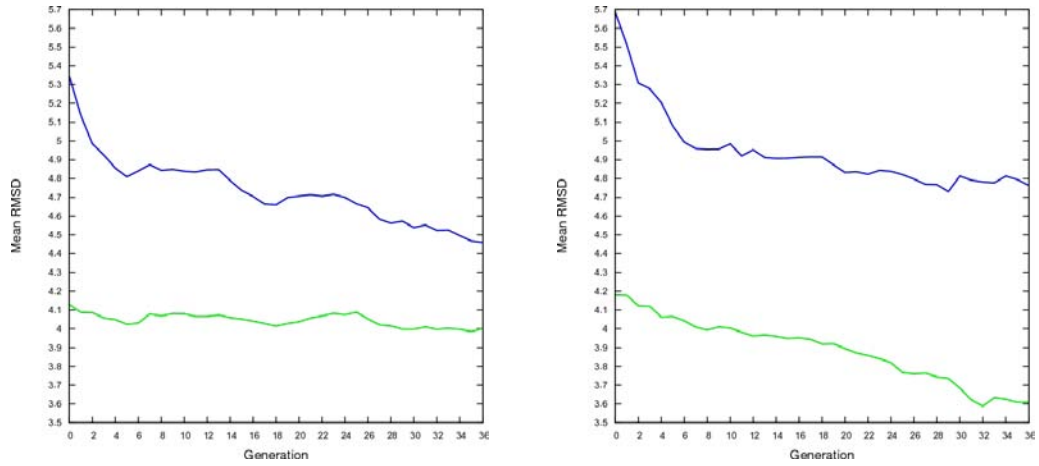


(a) (b)

Figure 6.30: The fluctuation of D_H with respect to individuals in a population (green) and the lowest energy conformation found (blue) as a function of generation for (a) the successful case and (b) the unsuccessful case.

the lowest energy conformation found, the initial population for the unsuccessful case, shows a stronger relationship to that of the GM for the successful case. However, as seen with other models, individuals in a population should be more related to each other than to the lowest energy conformation until convergence begins. Although the successful case does not show convergence with respect to the mean D_H, the two measures are distinctly different, with μ_{DH}^{L} being greater than μ_{DH}^{P}. However, the unsuccessful case tells a different story. The magnitude of μ_{DH}^{L} is lower than that of μ_{DH}^{P}. Again, this supports the idea of local minimum trapping, but also suggests that the individuals were more related to the lowest energy conformation than they were to each other. This demonstrates that, although the starting individuals are randomly generated, for this particular case, the quality of the starting material hindered the search process.

Based on what is seen for μ_{DH}^{P} and μ_{DH}^{L}, the mean RMSD plots do not appear as expected. The initial values of both μ_{RMSD}^{P} and μ_{RMSD}^{L} appear as expected, with a decrease observed over time. However, the successful case shows a near constant magnitude of μ_{RMSD}^{P}, illustrating that dominating regions of local structure do not



(a) (b)

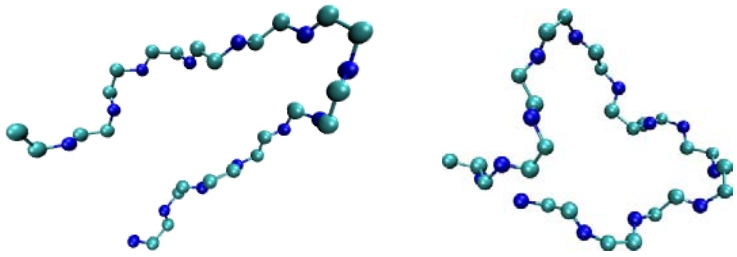
Figure 6.31: The fluctuation of RMSD with respect to individuals in a population (green) and the lowest energy conformation found (blue) as a function of generation for (a) the successful case and (b) the unsuccessful case.

propagate through the population sufficiently to induce a stronger relationship in 3D conformation. The magnitude of μ^L_{RMSD} does, however, decrease more dramatically over time, suggesting that, although the individuals remain as diverse as each other with respect to 3D conformation, they actually exhibit more GM-like characteristics over time. For the unsuccessful case, however, a more dramatic decrease in μ^P_{RMSD} than for μ^L_{RMSD} is observed, illustrating that the individuals within a population exhibit peer traits more quickly than they do 3D traits of the lowest energy conformation. This confirms that population convergence is an issue for the unsuccessful case.

Figure 6.32 shows the GM conformation obtained by the successful run, and the lowest energy conformation obtained by the unsuccessful run. Note how the density of the GM is higher than for the unsuccessful lowest energy conformation, as expected.

6.4 Conclusions

In this study, DE has been applied to the protein folding problem. The results presented here reflect averaged statistics obtained from one hundred runs of the algorithm, using various parameters over two implementations. It has been shown that DE is successful



(a) (b)

Figure 6.32: Lowest energy conformations found by RGA-DE for (a) the successful case (GM), $E = 1.51775$ and $\rho = 307.706$ and (b) the unsuccessful case, $E = 1.40788$, $\rho = 308.803$.

at searching PESs in order to find lowest energy conformations across a range of protein models and PES types.

Whilst this search technique is successful, it should be noted that acceptable success levels were only achieved by incorporating the RGA to correct conformations that are otherwise not self-avoiding. Although the number of successful mutations per generation is not improved as a result of using the RGA, what is improved is the mean energy for a generation. The effect of this, on searching conformational space, is that lower energy regions of the PES (i.e. further into the folding funnel) are considered, and, thus, more directed search strategies can take place. In turn, the likelihood of probing a GM region of the PES is considerably increased, resulting in a higher success rate.

Initially the RGA-DE was coupled with a simple 2D HPLBM. Although high levels of success were achieved for this model, this technique failed to effectively search the PESs of any sequence greater than 36 beads, resulting in zero success. This suggests that this implementation shows promise for the 2D case. However, a number of reasons may prevent the this technique from reaching its goal. The larger the sequence the larger the search space, meaning that this implementation impedes the coverage of the PES for larger sequences.

Following the 2D case, the HPLBM was coupled with a diamond lattice. By extending the search to 3D geometries, a similar trend was witnessed to the 2D case for 3D diamond sequences of high degeneracy. By increasing the magnitude of n_{ind}, the coverage of the PES was also increased, resulting in higher levels of success. However, for smaller magnitudes of n_{ind}, poor success rates were observed, something not seen for these sequences in the IA diamond lattice study. The ability of the combined mutation and mating operator to focus its search on other areas of the PES is limited when considering small n_{ind}.

Sequences of lower degeneracy have a simpler PES, in that the number of GM is significantly lower (although more difficult to randomly determine). The likelihood of discovering the GM on a PES of this type is reduced for this reason. Whereas for the IA small magnitudes of n_{ind} were suitable for achieving success for sequences of high degeneracy, for this DE implementation, these values of n_{ind} cannot provide enough genetic information to direct the search towards the GM.

It should be noted that the IA performs more genetic operations per generation than the DE, in that two mutation operators work independently to provide a mutated population that is later compared to the current population for that generation. The DE works differently, by performing a single combined mutation and mating operation, and overwriting any parent from the population if there is an increase in fitness. This reduces the amount of genetic information accessible to the search method per generation, and, thus, the DE requires a larger number of generations to reach its goal. A disadvantage of having less genetic information to mutate per cycle, results in the DE being unable to find all GM for some sequences when run for the same length of time as the IA, especially for small magnitudes of n_{ind}. However, it has been shown that once g_{max} has been increased, the DE is able to fully explore the PES and provides good SR, finding all GM for sequences of low degeneracy.

By simply recording the parents of each individual, as well as the conformation

vectors, it has been shown that profiles can be produced, and data can be collected to build up a picture of how the DE reaches its end point. By considering the pathway chosen by the search technique, it has been shown how the strengths and weaknesses of such a technique can be assessed, and how they relate to specific model proteins.

The importance of bonding within a protein conformation also contributes to the success or failure of the DE. It has been shown that, by producing topological contacts not present within the GM, the search may result in failure. By producing topological contacts too early, not only would the conformation have to completely unfold (and overcome a large energy barrier), but once the population begins to share low energy segments of local structure, can in fact result in local minimum trapping.

The DE is successful at searching the PES for DLM proteins. However, as with any model protein, the chain length becomes a factor with regard to efficiency and success. The DE seems not to exhibit an increase in success once g_{max} is increased beyond 2,000. This suggests that the DE has been pushed to its limit under the current methodology and that a different approach may be required to see an increase in success. Population convergence has been invoked to explain the plateau of success. It has been shown how the D_H and RMSD measures can provide different information with regard to success and failure for DLM proteins. Population convergence can be quantified using both measures, for relationships between individuals and to the lowest energy conformations.

Chapter 7

Conclusions and Future Work

Both the IA and the DE have proven successful in determining the GM conformations of various model proteins. The success rate is dependent on the parameters chosen for each search technique and the size of the protein chain. Generally a lower SR is observed for larger sequences, due to the exponential increase in the size of the PES and its complexity. Controversially, the IA benefits from much smaller magnitudes of n_{ind} than used for other search techniques, due to the presence of an ageing operator. The DE however, shows an improvement in performance (in terms of μ_{FE}) as the size of the population is increased, due to its reliance on a wealth of genetic material, used to replace members of a population with individuals of lower energy.

By recording the parents of each individual, an insight into how both techniques cope with searching the PES has been achieved. Both success and failure have been attributed to population diversity, with unsuccessful searches arising due to local minimum trapping. For large magnitudes of n_{ind}, it has been shown that large diversities in the population are beneficial. Controversially, for smaller population sizes, a population can be too diverse when using non-traditional genetic operators. Although this prevents a thorough search of the PES, and ultimately lower success, efficiency has been seen to improve. A reduction in the population diversity may be achieved if individuals are not guaranteed a mutation opportunity and that performing a mutation is determined probabilistically for these operators.

Smaller magnitudes of n_{ind} present a greater level of detail with regard to analysis, due to each individual contributing more to population diversity measures. The diversity within a population has been attributed to the location of mutations within a protein chain. Terminal mutations provide diversity noise and occur more frequently due to a reduced level of disruption to the conformation. The degree of mutation towards the centre of a chain, has been shown to cause difficulty for conformations of low energy and can result in low population diversity. However, these mutations are important if the search is to explore other regions of the PES. This obstacle may be avoided if a protein chain was not grown from one terminus, but grown from the centre out.

As profiling provides information about the relationship between individuals within a population and the GM, disconnectivity analysis [27, 121–123], using principal components [124–126], would be beneficial in determining the shape of the folding funnels for an entire search and various populations throughout.

It has been shown that the implementation of the DE in this work, can regularly suffer from local minimum trapping. Erratic searching of the PES (due to the combined mutation and mating operation) has prevented successful determination of the GM and thorough searching of PES regions. An improvement may be observed if this methodology were to be combined with that of other search techniques (that benefit from large population sizes, e.g. GA), intelligently switching between the two search methods if the population begins to stagnate [127].

Although united atom models introduce an increased complexity, some GM conformations exhibit structural features of real proteins. However, on the whole, structural similarity is low between model and real protein conformation. The use of a reduced torsional space is not beneficial when also considering a simplified interaction mechanism between residues. As changing the torsion options for each residue did not improve matters, it is hoped that by reducing the torsional space of amino acid triples

(whereby an amino acid is given a series of torsional options dependant on which residues lie adjacent to it in the chain), an increase in structural similarity may be seen.

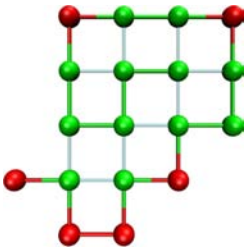
As a ϕ, ψ pair is determined via clustering of Ramachandran regions, another approach may be to weight the selection of torsion angles based on the size of the cluster that determined the centroid. This would allow for many regions of Ramachandran space to be represented without the loss of knowledge that highly populated regions, that give rise to α-helix or β-sheet structural preferences exist.

Providing a reduced torsional space dramatically reduces search space and increases algorithm speed. Without complicating the torsion angle selection during chain growth, an improvement to the backbone configuration may be observed if torsion angles are classified as cis-like or trans-like. This would result in four categories for each ϕ, ψ angle pair, and sixteen regions of Ramachandran space if also considering the quadrants. To further extend this theory, the weighting procedure could be applied once appropriate cis- or trans-like data had been gathered from the PDB and clustered in the same way.

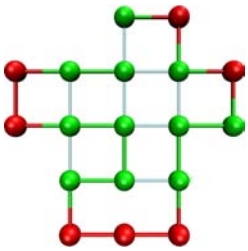
For Cys containing sequences, often, a disulfide bridge which is present in a real protein is not reproduced by the model interpretation. By introducing a long range potential, improvements to the backbone configuration may be observed, so modelling a separate interaction between Cys residues may be beneficial.

Appendix A

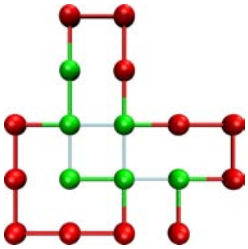
HP Bead Model on the Square Lattice



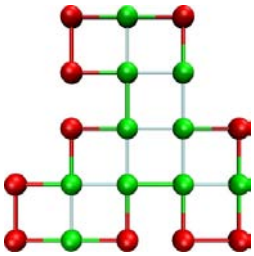
(a) HP-18a, $E_{HP} = -9$ a.u.



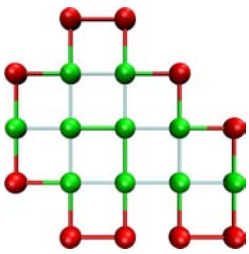
(b) HP-18b, $E_{HP} = -8$ a.u.



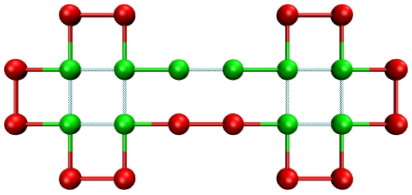
(c) HP-18c, $E_{HP} = -4$ a.u.



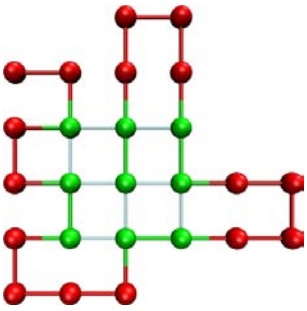
(d) HP-20a, $E_{HP} = -9$ a.u.



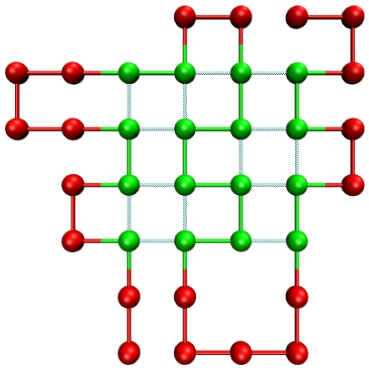
(e) HP-20b, $E_{HP} = -10$ a.u.



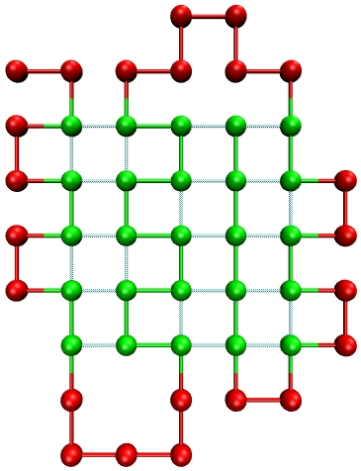
(f) HP-24, $E_{HP} = -9$ a.u.



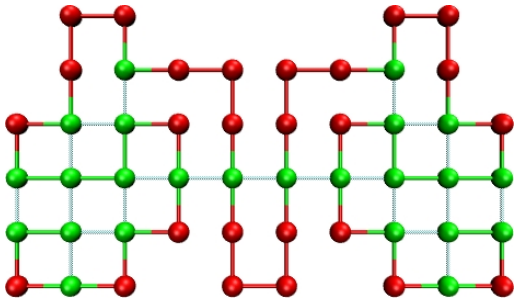
(g) HP-25, $E_{HP} = -8$ a.u.



(h) HP-36, $E_{HP} = -14$ a.u.



(i) HP-48, $E_{HP} = -23$ a.u.



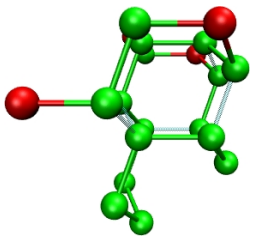
(j) HP-50, $E_{HP} = -21$ a.u.

Figure A.1: Most frequently found exmaple GM conformations of benchmark sequences for the HPLBM on the 2D square lattice.

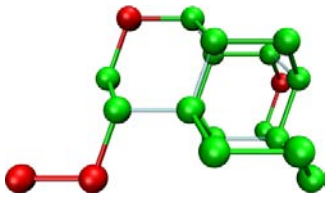
Appendix B

The Diamond Lattice

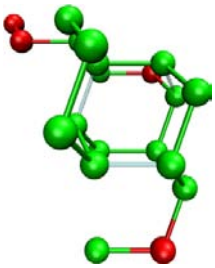
B.1 High Degeneracy Global Minima



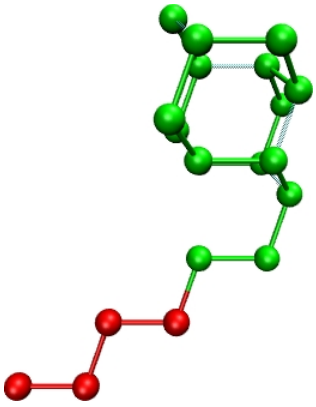
(a) H1, $E_{HP} = -5$ a.u., $E_{BLN} = 1.45759$ a.u.



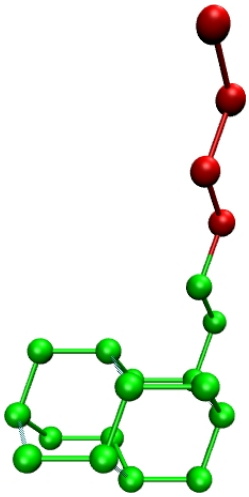
(b) H2, $E_{HP} = -5$ a.u., $E_{BLN} = 1.45778$ a.u.



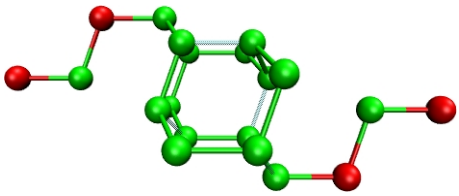
(c) H3, $E_{HP} = -5$ a.u., $E_{BLN} = 1.45778$ a.u.



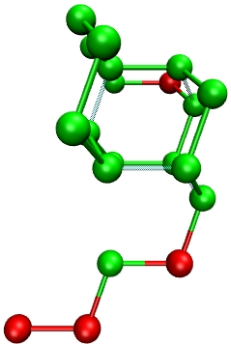
(d) H4, $E_{HP} = -5$ a.u., $E_{BLN} = 1.45455$ a.u.



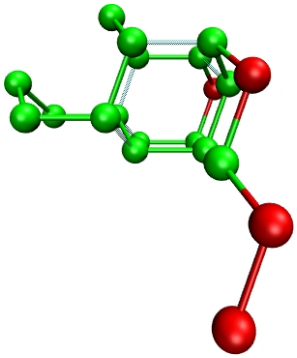
(e) H5, $E_{HP} = -5$ a.u., $E_{BLN} = 1.45455$ a.u.



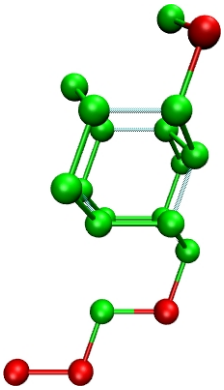
(f) H6, $E_{HP} = -5$ a.u., $E_{BLN} = 1.45635$ a.u.



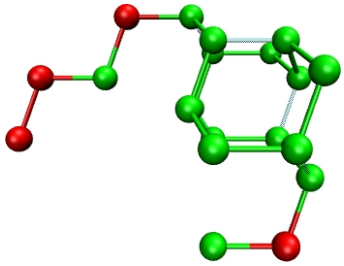
(g) H7, $E_{HP} = -5$ a.u., $E_{BLN} = 1.45615$ a.u.



(h) H8, $E_{HP} = -5$ a.u., $E_{BLN} = 1.45615$ a.u.



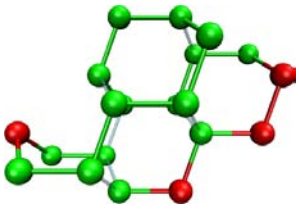
(i) H9, $E_{HP} = -5$ a.u., $E_{BLN} = 1.45625$ a.u.



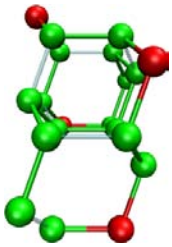
(j) H10, $E_{HP} = -5$ a.u., $E_{BLN} = 1.45625$ a.u.

Figure B.1: Most frequently found example GM conformations for the sequences of high degeneracy for both the HPLBM and the BLNM.

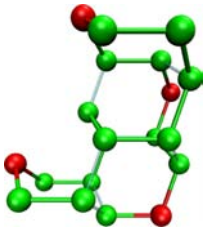
B.2 Low Degeneracy Global Minima



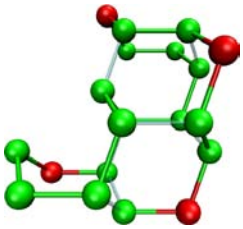
(a) L1, $E_{HP} = -7$ a.u., $E_{BLN} = 1.44382$ a.u.



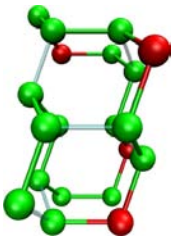
(b) L2, $E_{HP} = -7$ a.u., $E_{BLN} = 1.44217$ a.u.



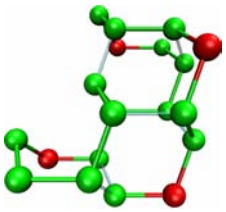
(c) L3, $E_{HP} = -7$ a.u., $E_{BLN} = 1.44260$ a.u.



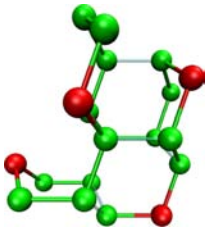
(d) L4, $E_{HP} = -7$ a.u., $E_{BLN} = 1.44393$ a.u.



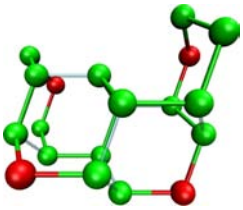
(e) L5, $E_{HP} = -7$ a.u., $E_{BLN} = 1.44593$ a.u.



(f) L6, $E_{HP} = -7$ a.u., $E_{BLN} = 1.44436$ a.u.

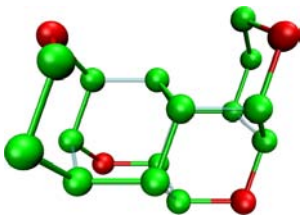


(g) L7, $E_{HP} = -7$ a.u., $E_{BLN} = 1.44593$ a.u.

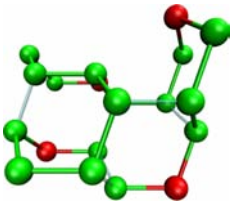


(h) L8, $E_{HP} = -7$ a.u., $E_{BLN} = 1.44436$ a.u.

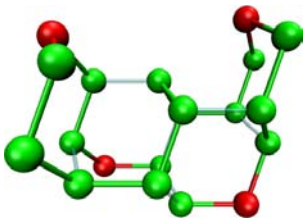
Figure B.2: GM for sequences L1 - L8 for both the HPLBM and the BLNM.



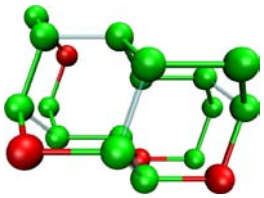
(a) L9, $E_{HP} = -7$ a.u., $E_{BLN} = 1.44393$ a.u.



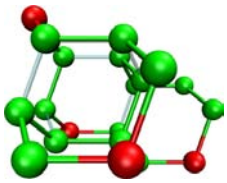
(b) L10, $E_{HP} = -7$ a.u., $E_{BLN} = 1.44769$ a.u.



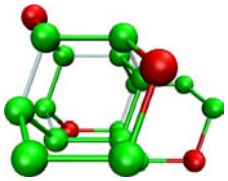
(c) L11, $E_{HP} = -7$ a.u., $E_{BLN} = 1.44436$ a.u.



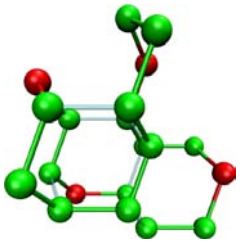
(d) L12, $E_{HP} = -7$ a.u., $E_{BLN} = 1.44769$ a.u.



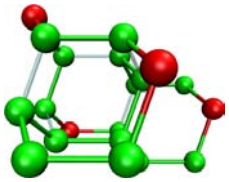
(e) L13, $E_{HP} = -7$ a.u., $E_{BLN} = 1.44436$ a.u.



(f) L14, $E_{HP} = -7$ a.u., $E_{BLN} = 1.44393$ a.u.

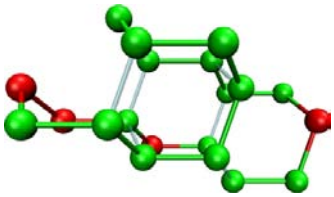


(g) L15, $E_{HP} = -7$ a.u., $E_{BLN} = 1.44593$ a.u.

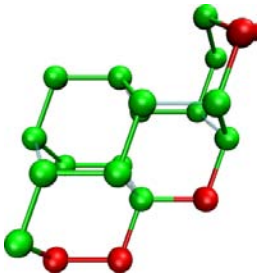


(h) L16, $E_{HP} = -7$ a.u., $E_{BLN} = 1.44260$ a.u.

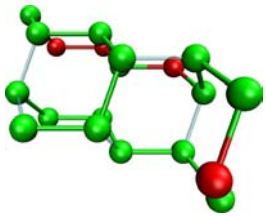
Figure B.3: GM for sequences L9 - L16 for both the HPLBM and the BLNM.



(a) L17, $E_{HP} = -7$ a.u., $E_{BLN} = 1.44382$ a.u.

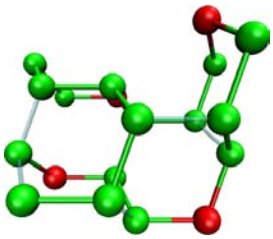


(b) L18, $E_{HP} = -7$ a.u., $E_{BLN} = 1.44160$ a.u.

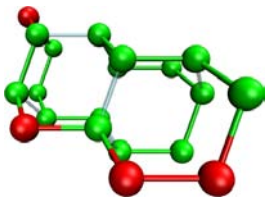


(c) L19, $E_{HP} = -7$ a.u., $E_{BLN} = 1.44206$ a.u.

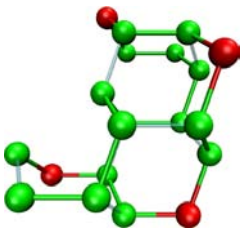
(d) L20, $E_{HP} = -7$ a.u., $E_{BLN} = 1.44206$ a.u.



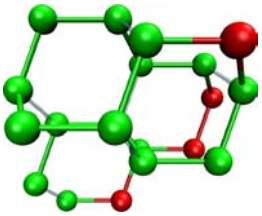
(e) L21, $E_{HP} = -7$ a.u., $E_{BLN} = 1.44593$ a.u.



(f) L22, $E_{HP} = -7$ a.u., $E_{BLN} = 1.44160$ a.u.

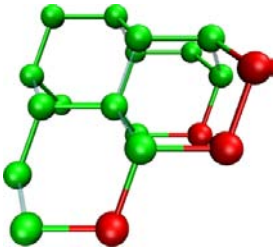


(g) L23, $E_{HP} = -7$ a.u., $E_{BLN} = 1.44217$ a.u.

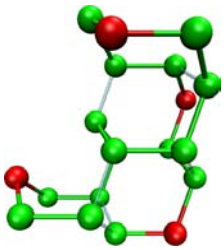


(h) L24, $E_{HP} = -7$ a.u., $E_{BLN} = 1.44270$ a.u.

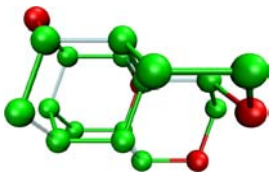
Figure B.4: GM for sequences L17 - L24 for both the HPLBM and the BLNM.



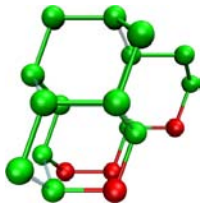
(a) L25, $E_{HP} = -7$ a.u., $E_{BLN} = 1.44342$ a.u.



(b) L26, $E_{HP} = -7$ a.u., $E_{BLN} = 1.44307$ a.u.



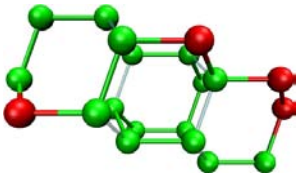
(c) L27, $E_{HP} = -7$ a.u., $E_{BLN} = 1.44475$ a.u.



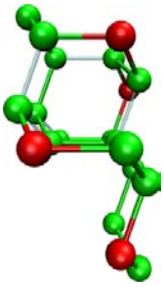
(d) L28, $E_{HP} = -7$ a.u., $E_{BLN} = 1.44342$ a.u.

(e) L29, $E_{HP} = -7$ a.u., $E_{BLN} = 1.44270$ a.u.

(f) L30, $E_{HP} = -7$ a.u., $E_{BLN} = 1.44270$ a.u.

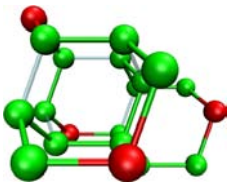


(g) L31, $E_{HP} = -7$ a.u., $E_{BLN} = 1.44160$ a.u.

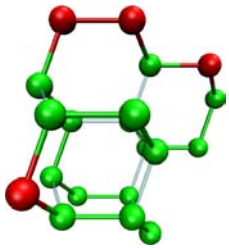


(h) L32, $E_{HP} = -7$ a.u., $E_{BLN} = 1.44475$ a.u.

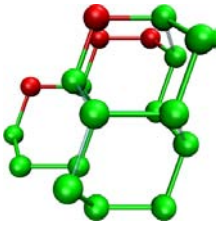
Figure B.5: GM for sequences L25 - L32 for both the HPLBM and the BLNM.



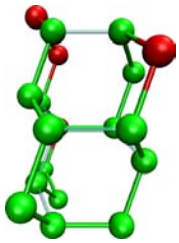
(a) L33, $E_{HP} = -7$ a.u., $E_{BLN} = 1.44307$ a.u.



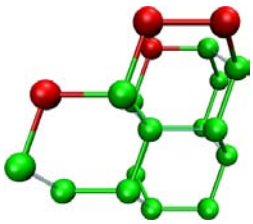
(b) L34, $E_{HP} = -7$ a.u., $E_{BLN} = 1.44094$ a.u.



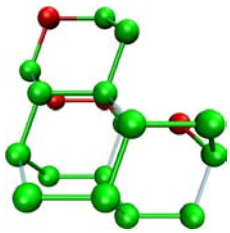
(c) L35, $E_{HP} = -7$ a.u., $E_{BLN} = 1.44166$ a.u.



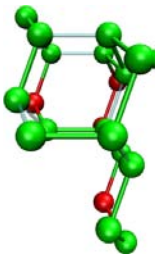
(d) L36, $E_{HP} = -7$ a.u., $E_{BLN} = 1.44094$ a.u.



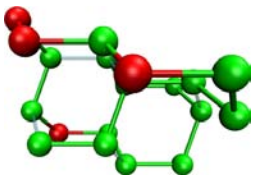
(e) L37, $E_{HP} = -7$ a.u., $E_{BLN} = 1.44342$ a.u.



(f) L38, $E_{HP} = -7$ a.u., $E_{BLN} = 1.44514$ a.u.

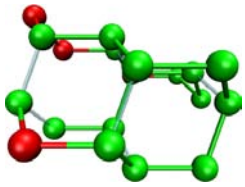


(g) L39, $E_{HP} = -7$ a.u., $E_{BLN} = 1.44166$ a.u.

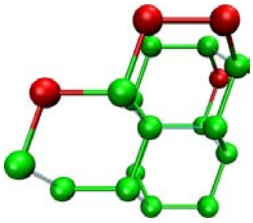


(h) L40, $E_{HP} = -7$ a.u., $E_{BLN} = 1.44094$ a.u.

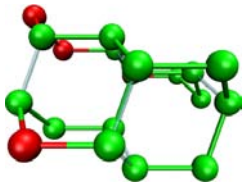
Figure B.6: GM for sequences L33 - L40 for both the HPLBM and the BLNM.



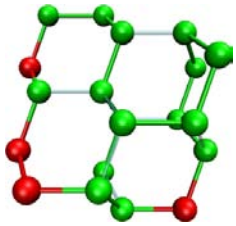
(a) L41, $E_{HP} = -7$ a.u., $E_{BLN} = 1.44270$ a.u.



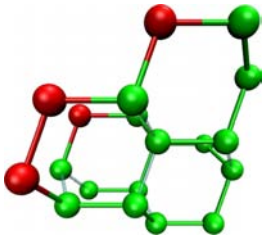
(b) L42, $E_{HP} = -7$ a.u., $E_{BLN} = 1.44270$ a.u.



(c) L43, $E_{HP} = -7$ a.u., $E_{BLN} = 1.44270$ a.u.

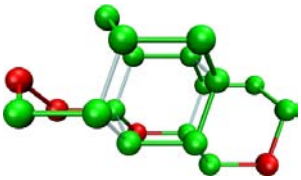


(d) L44, $E_{HP} = -7$ a.u., $E_{BLN} = 1.44094$ a.u.

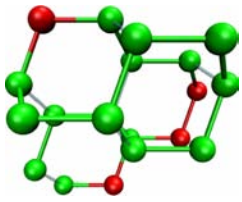


(e) L45, $E_{HP} = -7$ a.u., $E_{BLN} = 1.44342$ a.u.

(f) L46, $E_{HP} = -7$ a.u., $E_{BLN} = 1.44514$ a.u.



(g) L47, $E_{HP} = -7$ a.u., $E_{BLN} = 1.44160$ a.u.



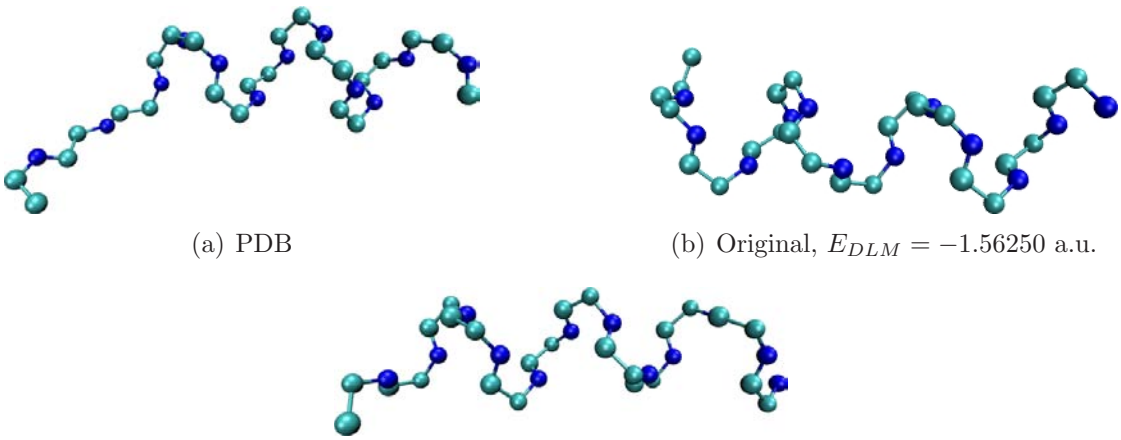
(h) L48, $E_{HP} = -7$ a.u., $E_{BLN} = 1.44286$ a.u.

Figure B.7: GM for sequences L41 - L48 for both the HPLBM and the BLNM.

Appendix C

Dynamic Lattice Model

C.1 Global Minima

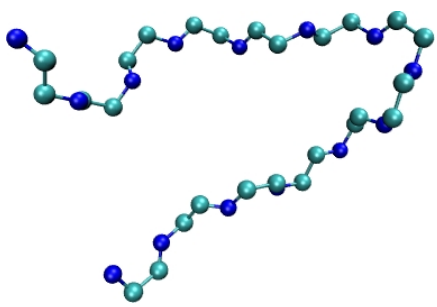


(a) PDB

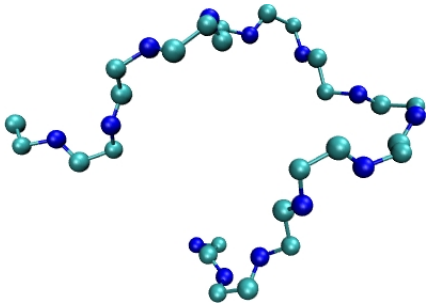
(b) Original, $E_{DLM} = -1.56250$ a.u.

(c) Modified, $E_{DLM} = -1.56140$ a.u.

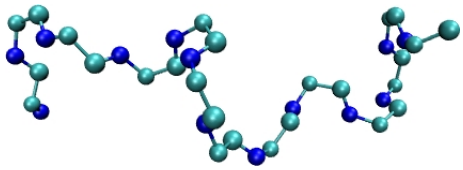
Figure C.1: 1AL1 PDB and GM conformations.



(a) PDB

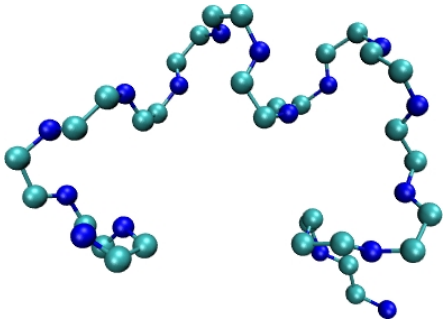


(b) Original, $E_{DLM} = -1.51775$ a.u.

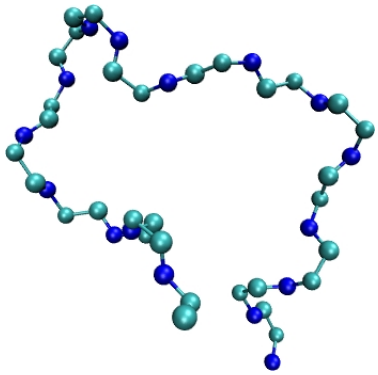


(c) Modified, $E_{DLM} = -1.25122$ a.u.

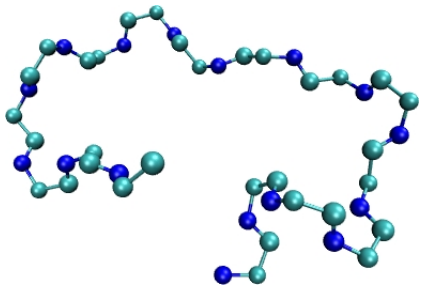
Figure C.2: 1A1P PDB and GM conformations.



(a) PDB

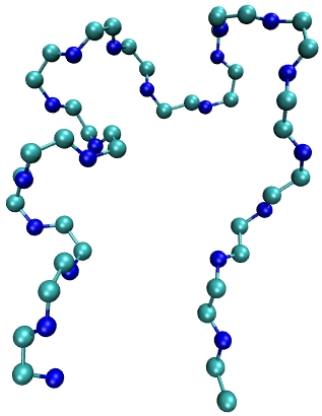


(b) Original, $E_{DLM} = -1.17301$ a.u.

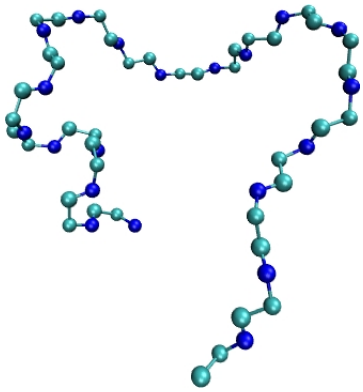


(c) Modified, $E_{DLM} = -1.42336$ a.u.

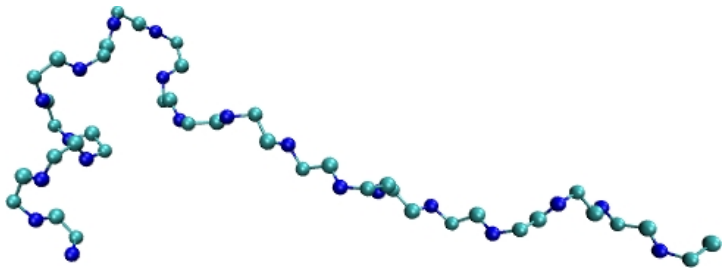
Figure C.3: 1AKG PDB and GM conformations.



(a) PDB

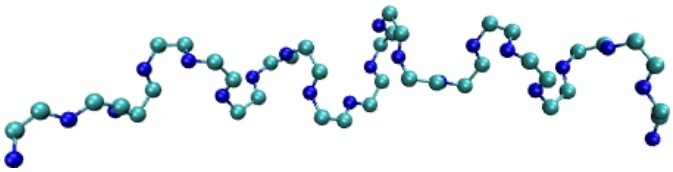


(b) Original, $E_{DLM} = -0.87846$ a.u.

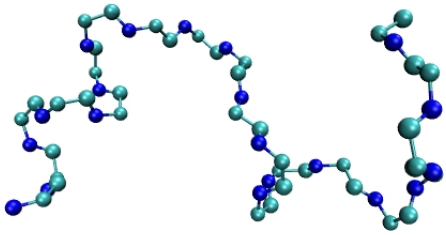


(c) Modified, $E_{DLM} = -0.93497$ a.u.

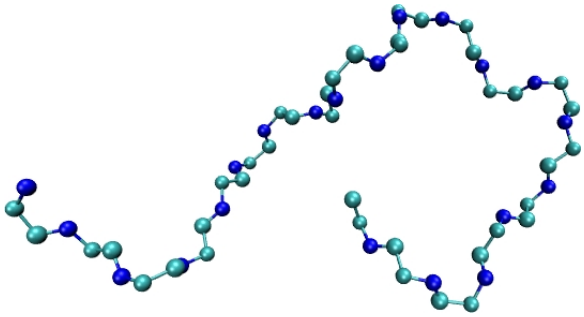
Figure C.4: 1L2Y PDB and GM conformations.



(a) PDB

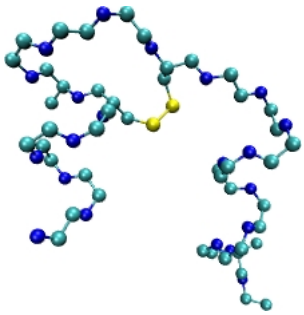


(b) Original, $E_{DLM} = -1.34457$ a.u.

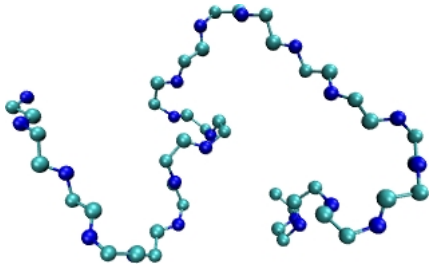


(c) Modified, $E_{DLM} = -1.43968$ a.u.

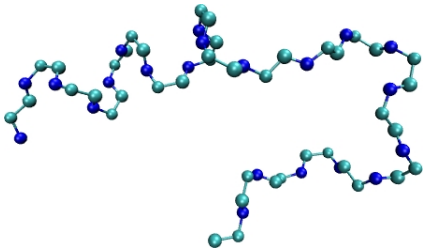
Figure C.5: 1D9J PDB and GM conformations.



(a) PDB

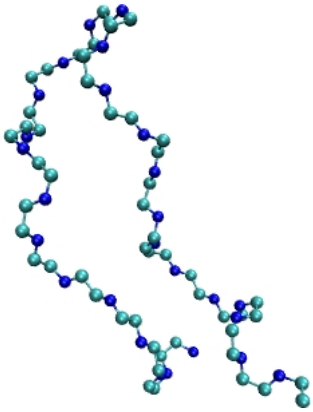


(b) Original, $E_{DLM} = -3.54363$ a.u.

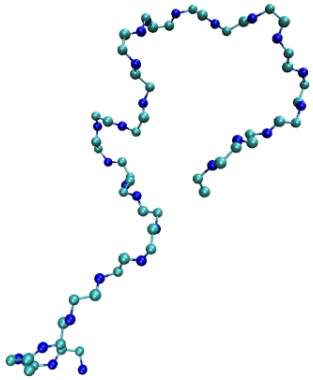


(c) Modified, $E_{DLM} = -2.6298$ a.u.

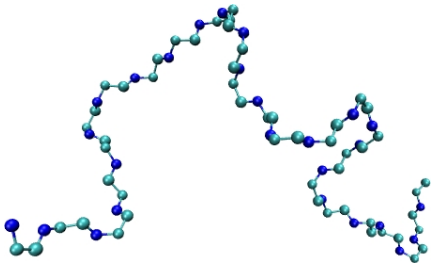
Figure C.6: 1B19:A PDB and GM conformations.



(a) PDB

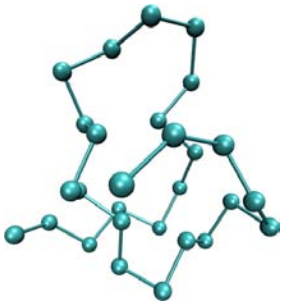


(b) Original, $E_{DLM} = -1.24277$ a.u.

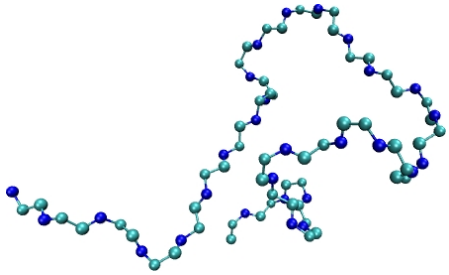


(c) Modified, $E_{DLM} = -1.13887$ a.u.

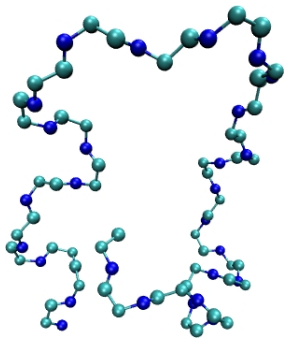
Figure C.7: 1G04 PDB and GM conformations.



(a) PDB

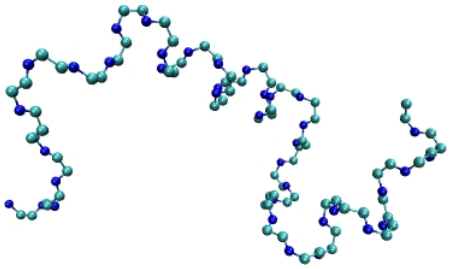


(b) Original, $E_{DLM} = -2.37581$ a.u.

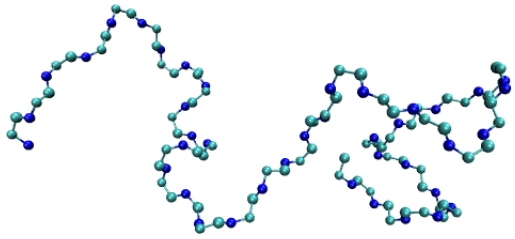


(c) Modified, $E_{DLM} = -2.19046$ a.u.

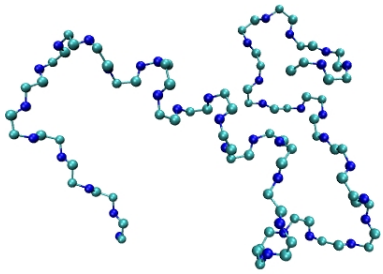
Figure C.8: 1ANP PDB and GM conformations.



(a) PDB

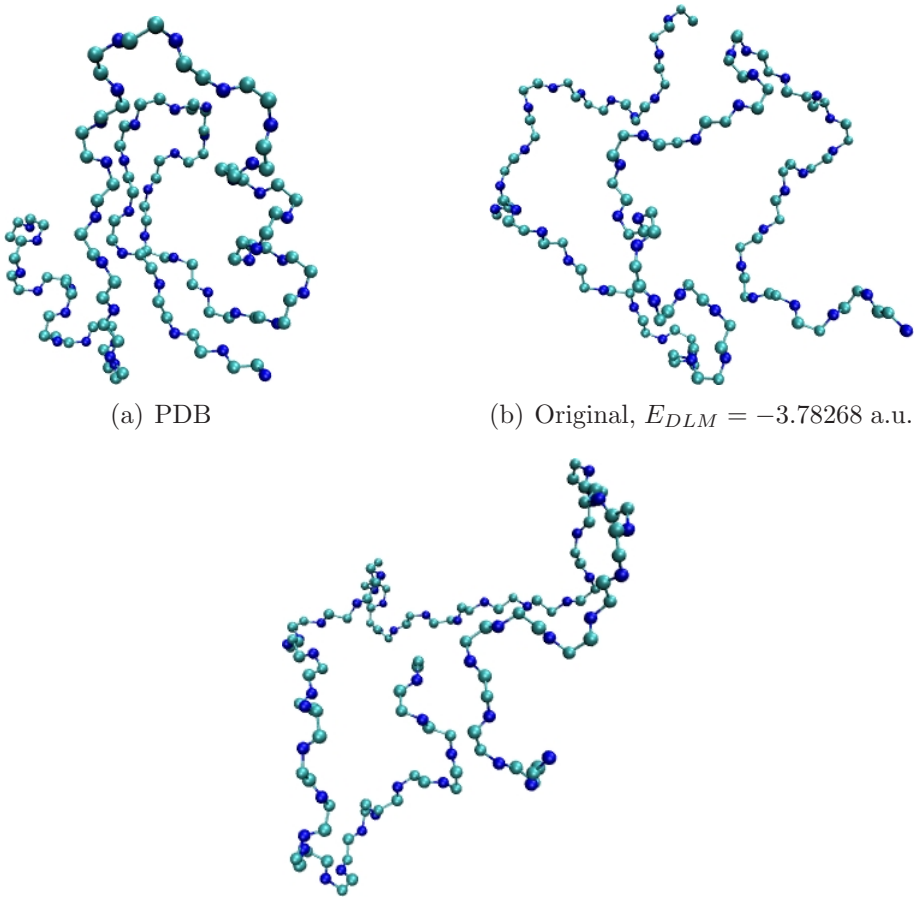


(b) Original, $E_{DLM} = -4.91472$ a.u.



(c) Modified, $E_{DLM} = -4.49711$ a.u.

Figure C.9: 1AML PDB and GM conformations.



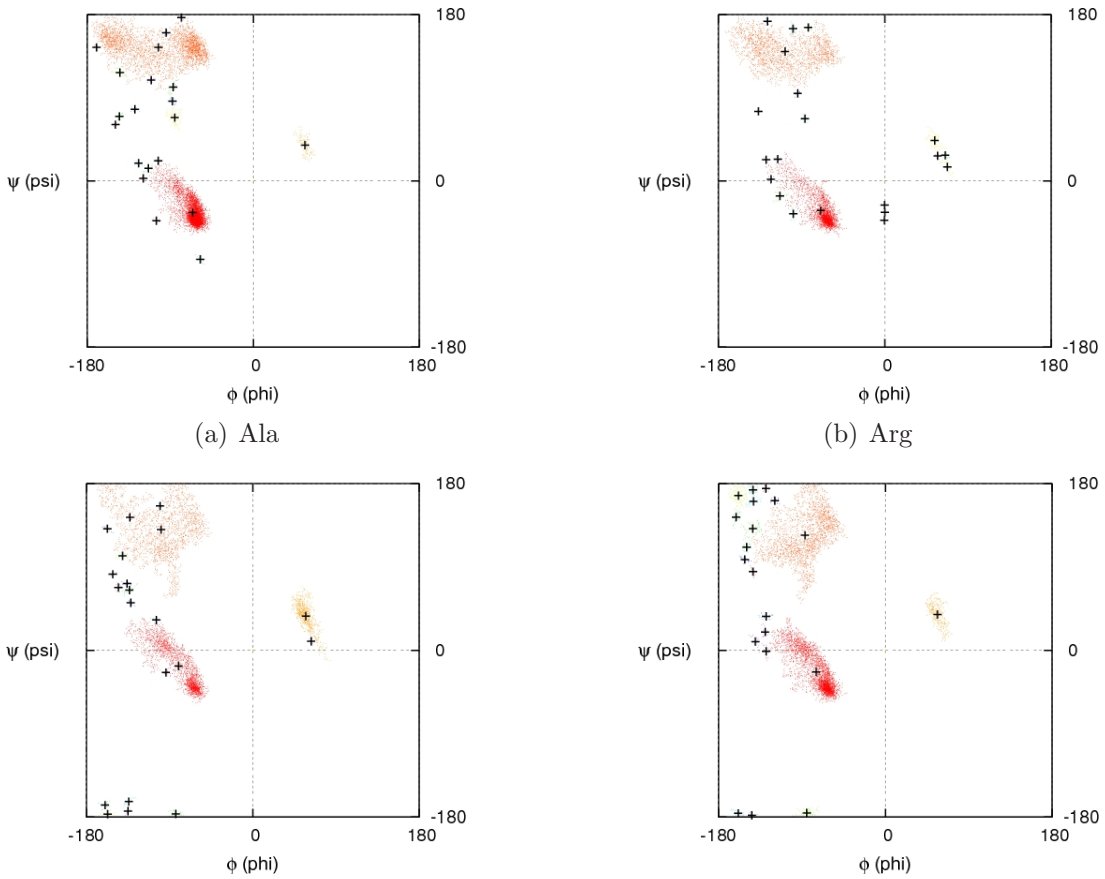
(a) PDB

(b) Original, $E_{DLM} = -3.78268$ a.u.

(c) Modified, $E_{DLM} = -3.40506$ a.u.

Figure C.10: 1QHK PDB and GM conformations.

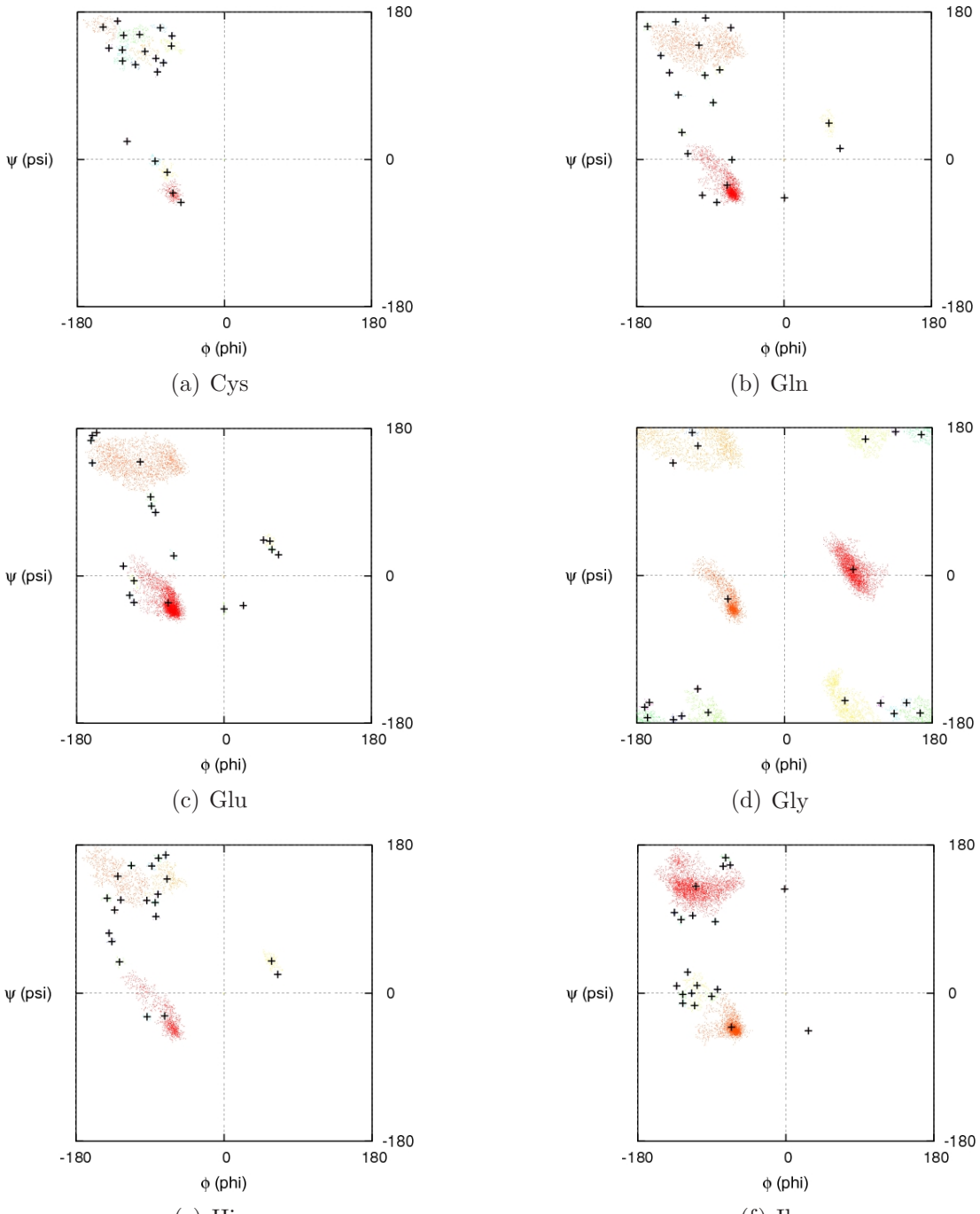
C.2 Ramachandran Clusters



(a) Ala (b) Arg

(c) Asn (d) Asp

Figure C.11: Ala - Asp Ramachandran clusters.

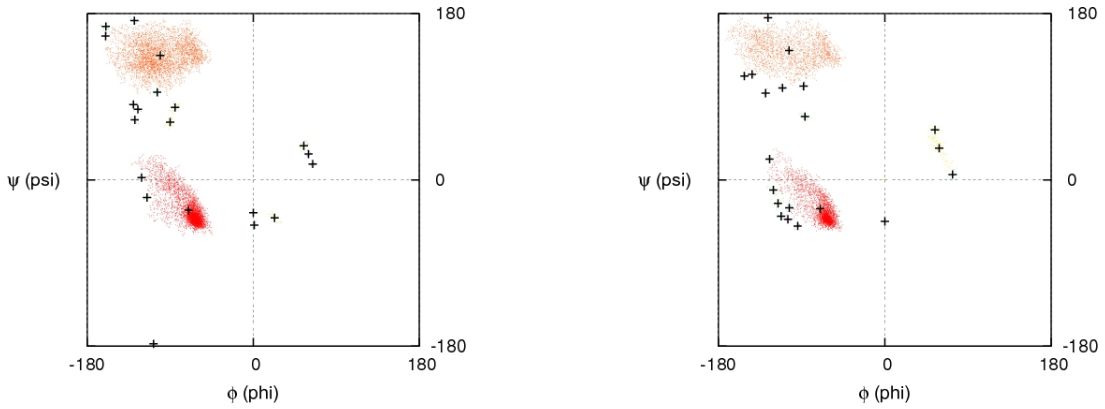


(a) Cys (b) Gln

(c) Glu (d) Gly

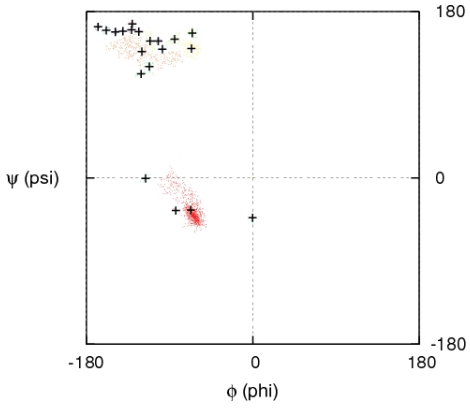
(e) His (f) Ile

Figure C.12: Cys - Ile Ramachandran clusters.

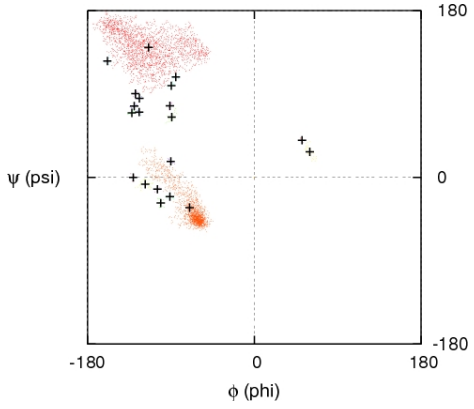


(a) Leu

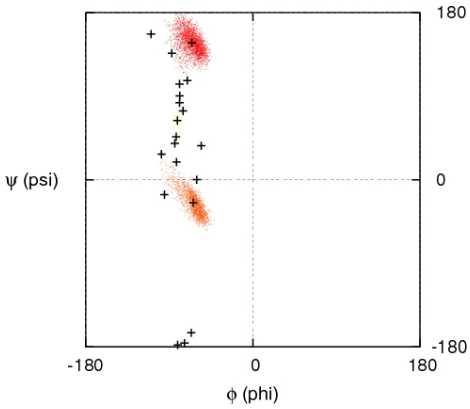
(b) Lys



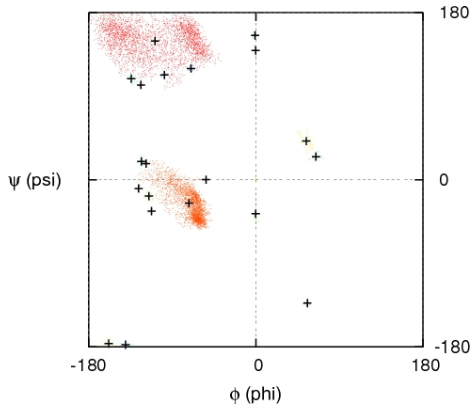
(c) Met



(d) Phe

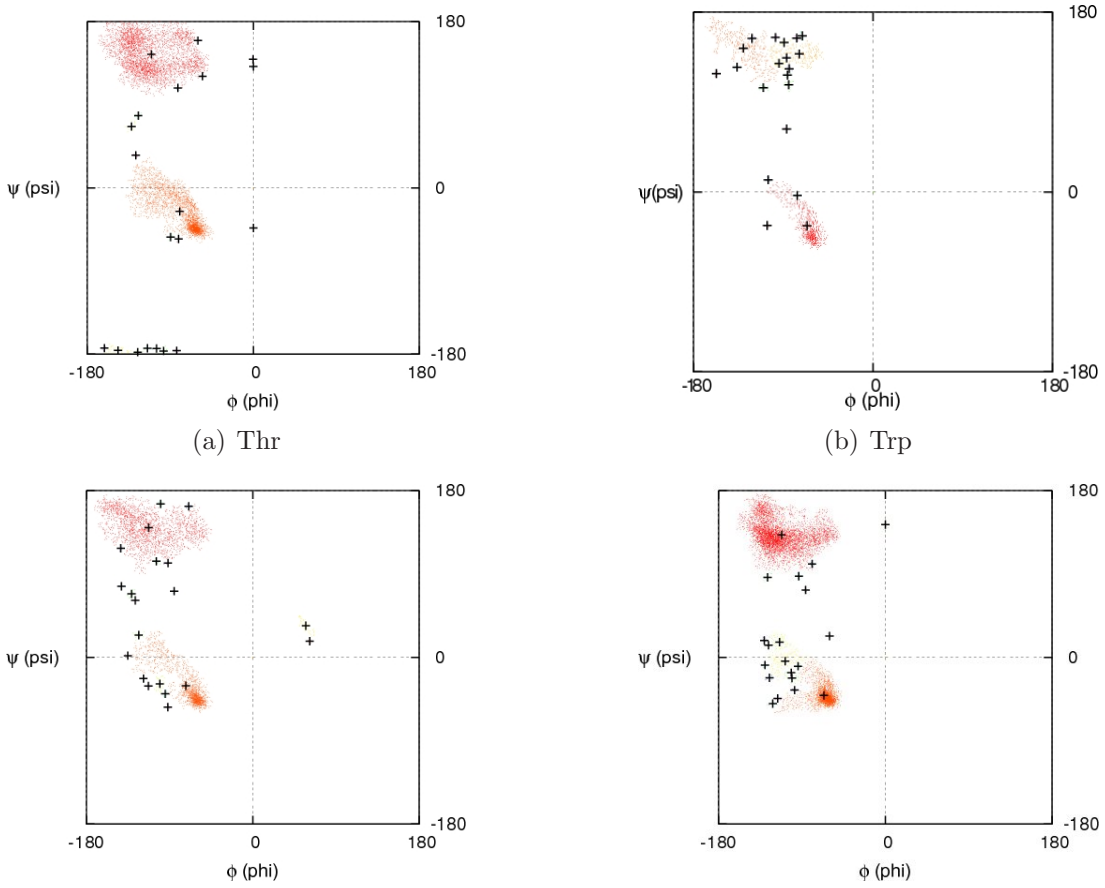


(e) Pro



(f) Ser

Figure C.13: Leu - Ser Ramachandran clusters.



(a) Thr (b) Trp

(c) Tyr (d) Val

Figure C.14: Thr - Val Ramachandran clusters.

Bibliography

[1] Frauenfelder, H. *Energy Landscape and Dynamics of Biomolecules*. J. Biol. Phys., 2005. **31**, 413–416.

[2] Cheung, M. S., Chavez, L. L., & Onuchic, J. N. *The Energy Landscape for Protein Folding and Possible Connections to Function*. Polymer, 2004. **45**, 547–555.

[3] Lesk, A. M. *Introduction to Protein Architecture*. Oxford University Press, 2003.

[4] Sadowski, M. I. & Jones, D. T. *The Sequence-structure Relationship and Protein Function Prediction*. Curr. Opin. Struc. Biol., 2009. **19**, 357–362.

[5] Backofen, R. & Will, S. *A Constraint-Based Approach to Fast and Exact Structure Prediction in Three-Dimensional Protein Models*. Contraints, 2006. **11**, 5–30.

[6] McMurray, J. *Organic Chemistry, Fifth Edition*. Brooks Cole, 2000.

[7] Wales, D. J. *Energy Landscapes With Applications to Clusters, Biomolecules and Glasses*. Cambridge University Press, 2003.

[8] Jones, M. *Organic Chemistry Second Edition*. W. W. Norton and Company, 2000.

[9] Richardson, J. S. *The Anatomy and Taxonomy of Protein Structure*. Adv. Protein Chem., 1981. **34**, 167–339.

[10] NEB. *Amino Acid Structures*.

http://www.neb.com/nebecomm/tech_reference/general_data/amino_acid_structures.asp.

[11] Krasnogor, N., Pelta, D., Lopez, P. M., Mocciola, P., & de la Canal, E. *Genetic algorithms for the protein folding problem: A critical view*. In Alpaydin, C. F. E., editor, *Proceedings of Engineering of Intelligent Systems*. ICSC Academic Press, 1998 .

[12] Kleywegt, G. & Jones, T. A. *Phi/Psi-chology: Ramachandran Revisited*. Structure, 1996. **4** (12), 1395–1400.

[13] Pauling, L., Corey, R. B., & Branson, H. R. *The Structure of Proteins: Two Hydrogen-Bonded Helical Configurations of the Polypeptide Chain*. P. Natl. Acad. Sci. USA, 1951. **37** (5), 205–211.

[14] Tro, N. J. *Introductory Chemistry*. Prentice Hall, 2003.

[15] Petsko, G. A. & Ringe, D. *Protein Structure and Function*. New Science Press Ltd., 2003.

[16] Mathews, C. K. & van Holde, K. E. *Biochemistry*. Prentice Hall, 1990.

[17] Sims, G. E., Choi, I., & Kim, S. *Protein Comformational Space in Higher Order ϕ-ψ Maps*. P. Natl. Acad. Sci. USA, 2005. **102** (3), 618–621.

[18] Anderson, R. J., Weng, Z., Campbell, R. K., & Jiang, X. *Main-Chain Conformational Tendencies of Amino Acids*. Proteins, 2005. **60**, 679–689.

[19] Ramachandran, G., Ramakrishnan, C., & Sasisekharan, V. *Stereochemistry of Polypeptide Chain Configurations*. J. Mol. Biol., 1963. **7**, 95–99.

[20] Gunasekaran, K., Ramakrishnan, C., & Balaram, P. *Disallowed Ramachandran Conformations of Amino Acid Residues in Protein Structures*. J. Mol. Biol., 1996. **264**, 191–198.

[21] Shmygelska, A. & Hoos, H. *An Ant Colony Optimisation Algorithm for the 2D and 3D Hydrophobic Polar Protein Folding Problem*. BMC Bioinformatics, 2005. **6** (30), 1–22.

[22] Cox, G. A., Mortimer-Jones, T. V., Taylor, R. P., & Johnston, R. L. *Development and Optimisation of a Novel Genetic Algorithm for Studying Model Protein Folding*. Theor. Chem. Acc., 2005. **122**, 163–178.

[23] Tang, C. *Simple Models of the Protein Folding Problem*. Physica A, 2000. **288**, 31–48.

[24] Lau, K. F. & Dill, K. A. *The Theory of Protein Mutability and Biogenesis*. Biophysics, 1990. **87**, 638–642.

[25] Cejtin, H., Edler, J., Gottlieb, A., Helling, R., Lao, H., & Philbin, J. *Fast Tree Search for Enumeration of a Lattice Model of Protein Folding*. Theor. Chem. Acc., 2002. **116** (1), 352–359.

[26] Khimasia, M. M. & Coveney, P. V. *Protein Structure Prediction as a Hard Optimisation Problem: The Genetic Algorithm Approach*. Mol. Simulat., 1997. **19**, 205–226.

[27] Cox, G. A. & Johnston, R. L. *Analyzing Energy Landscapes for Folding Model Proteins*. J. Chem. Phys., 2006. **124** (20), 163–178.

[28] Mukherjee, A. & Bagchi, B. *Correlation Between Rate of Folding, Energy Landscape, and Topology in the Folding of a model Protein HP-36*. J. Chem.Phys., 2003. **118** (10), 4733–4747.

[29] Levinthal, C. *How to Fold Graciously. Mossbauer Spectroscopy in Biological Systems*. In DeBrunner, J. & Munck, E., editors, *Proceedings of a meeting held at Allerton House*. University of Illinios Press, Urbana, Illinois, USA, 1969 pages 22–24.

[30] Govindarajan, S. & Goldstein, R. A. *On the Thermodynamic Process of Protein Folding*. P. Natl. Acad. Sci. USA, 1998. **95**, 5545–5549.

[31] Honeycutt, J. D. & D.Thirumalai. *Metastability of the Folded States of Globular Proteins*. volume 87, 1990 pages 3526–3529.

[32] Zwanzig, R., Szabo, A., & Bagchi, B. *Levinthal's Paradox*. P. Natl. Acad. Sci. USA, 1992. **89**, 20–22.

[33] Newcomb, L. F., Haque, T. S., & Gellman, S. H. *Searching for Minimum Increments of Hydrophobic Collapse: Flexible Dinaphthyl Carboxylates*. J. Am. Chem. Soc., 1995. **117**, 6509–6519.

[34] Cerf, C., Lippens, G., Ramakrishnan, V., Muyldermans, S., Segers, A., Wyns, L., Wodak, S. J., & Hallenga, K. *Homo- and Heteronuclear Two-Dimensional NMR Studies of the Globular Domain of Histone H1: Full Assignment, Tertiary Structure, and Comparison With the Globular Domain of Histone H5*. Biochemistry-US, 1994. **33**, 11,079–11,086.

[35] Bernstein, F. C., Koetzle, T. F., Williams, G. J., Jr., E. E. M., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T., & Tasumi, M. *The Protein Data Bank: A Computer-based Archival File For Macromolecular Structures*. J. Mol. Biol., 1977. **112**, 535.

[36] Miller, M. A. & Wales, D. J. *Energy Landscapes of a Model Protein*. J. Chem. Phys., 1999. **111** (14), 6610–6616.

[37] Levy, Y., Jortner, J., & Becker, O. M. *Solvent Effects on the Energy Landscape and Folding Kinetics of Polyalanine*. P. Natl. Acad. Sci. USA, 2001. **98** (5), 2188–2193.

[38] Wales, D. J. *A Microscopic Basis for the Global Appearance of Energy Landscapes*. Science, 2001. **293**, 2067–2070.

[39] Levy, Y. & Becker, O. M. *Energy Landscapes of Conformationally Constrained Peptides*. J. Chem. Phys., 2001. **114** (2), 993–1009.

[40] Onuchic, J. N., Socci, N. D., Luthey-Schulten, Z., & Wolynes, P. G. *Protein Folding Funnels: The Nature of the Transition State Ensemble*. Fold. Des., 1996. **1** (6), 441–450.

[41] Komatsuzaki, T., Hoshino, K., Matsunaga, Y., Rylance, G. J., Johnston, R. L., & Wales, D. *How Many Dimensions are Required to Approximate the Potential Energy Landscape of a Model Proteins*. J. Chem. Phys., 2005. **122**, 084,714.

[42] Wales, D. J. & Doye, J. P. K. *Stationary Points and Dynamics in High-Dimensional Systems*. J. Chem. Phys., 2003. **119** (23), 12,409–12,416.

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

[44] Honeycutt, J. D. & D.Thirumalai. *The Nature of Folded States of Globular Proteins*. Biopolymers, 1992. **32**, 695–709.

[45] Lau, K. F. & Dill, K. A. *A Lattice Statistical Mechanics Model of the Conformational and Sequences Spaces of Proteins*. Macromolecules, 1989. **22**, 3986–3997.

[46] Hirst, J. D. *The Evolutionary Landscape of Functional Model Proteins*. Protein Eng., 1999. **12** (9), 721–726.

[47] Frauenkron, H., Bastolla, U., Gerstner, E., Grassberger, P., & Nadler, W. *Development and Optimisation of a Novel Genetic Algorithm for Studying Model Protein Folding*. Phys. Rev. Lett., 1998. **80** (14), 3149–3152.

[48] Kim, S. & Lee, J. *The Energy Landscape of a BLN Protein with βHairpin Shape*. J. Korean Phys. Soc., 2004. **44** (3), 589–593.

[49] Rylance, G. J. *The Visualisation, Exploration and Analysis of Energy Landscapes*. Ph.D. thesis, University of Birmingham, Birmingham, UK, 2007.

[50] Tieleman, D. P., MacCallum, J. L., Ash, W. L., Kandt, C., Xu, Z., & Monticelli, L. *Membrane Protein Simulations with a United-Atom Lipid and All-Atom Protein Model: Lipid-Protein Interactions, Side Chain Transfer Free Energies and Model Proteins*. J. Phys-Condens. Mat., 2006.

[51] Weiner, S. J. *A New Force Field for Molecular Mechanical Simulation of Nucleic Acids and Proteins*. J. Am. Chem. Soc., 1984. **106**, 765–784.

[52] Brooks, B. R., Bruccoleri, R. E., Olafson, D. J., States, D. J., Swaminathan, S., & Karplus, M. *CHARMM: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations*. J. Comp Chem., 1983. **4**, 187–217.

[53] Scott, W. R. P., Hünenberger, P. H., Tironi, I. G., Mark, A. E., Billeter, S. R., Fennen, J., Torda, A. E., Huber, T., Kruger, P., & van Gunsteren, W. F. *The GROMOS Biomolecular Simulation Program Package*. J. Phys. Chem. A, 1999. **103**, 3596–3607.

[54] Jorgensen, W. L. & Tirado-Rives, J. *The OPLS Potential Functions for Proteins - Energy Minimisations for Crystals of Cyclic-peptides and Crambin*. J. Am. Chem. Soc., 1988. **110**, 1657–66.

[55] de Castro, L. N. & Timmis, J. *Artificial Immune System: A New Computational Intelligence Approach*. Springer, 2002.

[56] Farmer, J. D., Packard, N. H., & Perelson, A. S. *The Immune System, Adaption, and Machine Learning*. Physica, 1986. **22**, 187–204.

[57] Delves, P. J. & Roitt, I. M. *The Immune System: First of Two Parts*. New Engl. J. Med., 2000. **343** (1), 37–49.

[58] Parkin, J. & Cohen, B. *An Overview of the Immune System*. Lancet, 2001. **357**, 1777–1789.

[59] de Castro, L. N. & Timmis, J. I. *Artificial Immune Systems as a Novel Soft Computing Paradigm*. Soft Comput., 2003. **7**, 526–544.

[60] Timmis, J., Neal, M., & Hunt, J. *An Artificial Immune System for Data Analysis*. BioSystems, 2000. **55**, 143–150.

[61] Delves, P. J. & Roitt, I. M. *The Immune System: Second of Two Parts*. New Engl. J. Med., 2000. **343** (1), 108–117.

[62] Springer, T. A. *Adhesion Receptors of the Immune System*. Nature, 1990. pages 425–433.

[63] Smith, K. G. C., Light, A., Nossal, G. J. V., & Tarlintn, D. M. *The Extent of Affinity Maturation Differs Between the Memory and Antibody-Forming Cell Compartments in the Primary Immune Response*. The EMBO J., 1997. **16** (11), 2996–3006.

[64] Timmis, J., Knight, T., de Castro, L. N., & Hart, E. *An Overview of Artificial Immune Systems*. Computation in Cells and Tissues: Perspectives and Tools for Thought, 2004.

[65] Cutello, V., Nicosia, G., Pavone, M., & Timmis, J. *An Immune Algorithm for Protein Structure Prediction*. IEEE T. Evolut. Comput., 2007. **11** (1), 101–117.

[66] Chong, S. Y. & Tremayne, M. *Combining optmisation Using Cultural and Differential Evolution: Application to Crystal Structure Solution from Powder Diffraction Data*. Chem. Commun., 2006. pages 4078–4080.

[67] Engelbrecht, A. P. *Computational Intelligence: An Introduction*. John Wiley and Sons Ltd., 2002.

[68] Larranaga, P., Kuijpers, C. M. H., Murga, R., Inza, I., & Dizdarevic, S. *Genetic Algorithms for the Travelling Salesman Problem: A Review of Representations and Operators*. Artif. Intell. Rev., 1999. **13**, 129–170.

[69] Pedersen, J. T. & Moult, J. *Protein Folding Simulations with Genetic Algorithms and a detailed Molecular Description*. J. Mol. Biol., 1997. **269**, 240–259.

[70] Bäck, T. & Schwefel, H. *An Overview of Evolutionary Algorithms for Parameter Optimization*. Evol. Comput., 1993. **1** (1), 1–23.

[71] Holland, J. H. *Adaption in Natural and Artificial Systems*. MIT Press, 1992.

[72] Goldberg, D. E. *Genetic Algorithms in Search, Optimisation and Machine Learning*. Addison-Wesley, 1989.

[73] Johnston, R. L. *Evolving Better Nanoparticles: Genetic Algorithms for Optimising Cluster Geometries*. Dalton T., 2003. **22**, 4193–4207.

[74] Unger, R. & Moult, J. *Genetic Algorithms for Protein Folding Simulations*. J. Mol. Biol., 1993. **231**, 75–81.

[75] Storn, R. M. & Prove, K. V. *Differential Evolution - A Simple and Efficient Heuristic for Global Optimisation over Continuous Spaces*. J. Global Optim., 1997. **11** (4), 341–359.

[76] R.Bitello & Lopes, H. S. *A Differential Evolution Approach for Protein Folding*. IEEE Symposium on Computational Intelligence and Bioinformatics and Computational Biology (CIBCB '06), 2006. pages 1–5.

[77] Eberhart, R. C. & Kennedy, J. *A new optimizer using particle swarm theory*. In *Proceedings of the 6th International Symposium on Micromachine and Human Science*. Nagoya, Japan, 1995 pages 39–43.

[78] Shmygelska, A. & Hoos, H. *An Improved Ant Colony Optimisation Algorithm for the 2D HP Protein Folding Problem*. Lect. Notes. Comput. Sc., 2003. **2671**, 993–1010.

[79] de Castro, L. N. *Immune, Swarm and Evolutionary Algorithms*. volume 3, 2002 pages 1464–1468.

[80] de Castro, L. N. & Zuben, F. J. V. *Learning and Optimization Using the Clonal Selection Principle*. IEEE T. Evolut. Comput., 2002. **6** (3), 239–251.

[81] Garret, S. M. *How Do We Evaluate Artificial Immune Systems?*. Evol. Comput., 2005. **13**, 145–178.

[82] Timmis, J. *Artificial Immune Systems - Today and Tomorrow*. Nat. Comp. Ser., 2007. **6**, 1–18.

[83] Pedersen, J. T. & Moult, J. *Genetic Algorithms for Protein Structure Prediction*. Curr. Opin. Struc. Biol., 1996. **6**, 227–231.

[84] Lloyd, L. D., Johnston, R. L., & Salhi, S. *Strategies for Increasing the Efficiency of a Genetic Algorithm for the Structural Optimization of Nanoalloy Clusters*. J. Comput. Chem., 2005. **26**, 1069–1078.

[85] Sokolov, A. & Whitley, D. *Unbiased tournament selection*. In *GECCO '05: Proceedings of the 2005 conference on Genetic and evolutionary computation*. ACM, New York, NY, USA. ISBN 1-59593-010-8, 2005 pages 1131–1138.

[86] Collins, R. J. & Jefferson, D. R. *Selection in Massively Parallel Genetic Algorithms*. In *Proceedings of the 4th International Conference on Genetic Algorithms*. Morgan Kaufmann, 1991 pages 249–256.

[87] Chakraborty, U. K., Deb, K., & Chakraborty, M. *Analysis of selection algorithms: A markov chain approach*. Evol. Comput., 1996. **4** (2), 133–167. ISSN 1063-6560.

[88] Miller, B. L. & Goldberg, D. E. *Genetic Algorithms, Tournament Selection, and the Effects of Noise*. AIP Conf. Proc., 1995. **9**, 193–212.

[89] Miller, B. L. & Goldberg, D. E. *Genetic Algorithms, Selection Schemes, and the Varying Effects of Noise*. Evol. Comput., 1997. **4** (2), 113–131.

[90] Humphrey, W., Dalke, A., & Schulten, K. *VMD – Visual Molecular Dynamics*. J. Mol. Graphics, 1996. **14**, 33–38.

[91] Kabsch, W. *A Solution for the Best Rotation to Relate Two Sets of Vectors*. Acta Crystallogr. A, 1976. **32**, 922–923.

[92] Ho, B. K. *RMSD: Root Mean Square Deviation*.

`http://boscoh.com/protein/rmsd-root-mean-square-deviation`.

[93] Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. *Numerical Recipes in C++: The Art of Scientific Computing*. Cambridge University Press, 2002.

[94] BlueBEAR. *The University of Birmingham Supercomputing Service*.

`http://www.bear.bham.ac.uk/bluebear`.

[95] Song, J., Cheng, J., Zeng, T., & Mau, J. *A Novel Genetic Algorithm for HP Model Protein Folding*. PDCAT, 2005. pages 935–937.

[96] Coutsias, E. A., Seok, C., & Dill, K. A. *Protein Structure Predcition using Evolutionary Algorithms Hybridized with Backtracking*. In *Lect. Notes Comput. Sci.*, volume 2687. Berlin, Germany, 2003 pages 321–328.

[97] Bennett, A. J., Johnston, R. L., Turpin, E., & He, J. Q. *Analysis of an Immune Algorithm for Protein Structure Prediction*. Informatica, 2008. **32** (3), 245–251.

[98] Johnson, D. S. *How Easy is Local Search*. J. Comput. Syst. Sci., 1988. **37** (1), 79–100.

[99] Shmygelska, A., Aguirre-Hernández, R., & Hoos, H. *An Ant Colony Optimisation Algorithm for the 2D HP Protein Folding Problem*. Lect. Notes Comput. Sci., 2002. **2463**, 369–428.

[100] Chong, S. Y., Seaton, C. C., Kariuki, B. M., & Tremayne, M. *Molecular versus Crystal Symmetry in Tri-substituted Triazone, Benzene and IsoCyanurate Derivatives*. Acta Crystallographica Section B, 2006. pages 864–874.

[101] Krasnogor, N., Hart, W. E., Smith, J., & Pelta, D. *Protein Structure Prediction with Evolutionary Algorithms*. In Banzhaf, W., Daida, J., Eiben, A. E., Garzon, M. H., Honavar, V., Jakiela, M., & Smith, R. E., editors, *GECCO '99: Proceedings of the 1999 conference on Genetic and evolutionary computation*, volume 2. Morgan Kaufmann, Orlando, Florida, USA, 1999 pages 1596–1601.

[102] Guo, Z. & D.Thirumalai. *Kinetics of Protein Folding: Nucleation Mechanism, Time Scales and Pathways*. Bioploymers, 1995. **36**, 83–102.

[103] Guo, Z. & D.Thirumalai. *Kinetics and Thermodynamics of Folding of a de Novo Designed Four-helix Bundle Protein*. J. Mol. Biol., 1996. **263**, 323–343.

[104] Dressel, F. & Kobe, S. *Global Optimisation of Proteins Using a Dynamical Lattice Model: Ground States and Energy Landscapes*. Chem. Phys. Lett., 2006. **424**, 369–373.

[105] Hobohm, U. & Sander, C. *Enlarged Representative Set of Protein Structures*. volume 3, 1994 pages 522–524.

[106] Settanni, G., Micheletti, C., Banavar, J., & Maritan, A. *Determination of Optimal Effective Interactiosn Between Amino Acids in Globular Proteins*. cond-mat/9902364, 1999.

[107] Leung, Y., Goa, Y., & Xu, Z. *Degree of Population Diversity - A Perspective on Premature Convergence in Genetic Algorithms and its Markov Chain Analysis*. IEEE T. Neural Networ., 1997. **8** (5), 1165–1172.

[108] Zhu, K. & Liu, Z. *Levinthal's Paradox*. In Boulicaut, J., Esposito, F., Giannotti, F., & Pedreschi, D., editors, *Proceedings of the 15th European Conference on Machine Learning*. Springer, Pisa, Italy, 2004 pages 537–547.

[109] He, M. X., Petoukhov, S. V., & Ricci, P. E. *Genetic Code, Hamming Distance and Stochastic Matrices*. B. Math. Biol., 2004. **66**, 1405–1421.

[110] Wallin, S., Farwer, J., & Bastolla, U. *Testing Similarity Measures with Continuous and Discrete Protein Models*. Proteins, 2003. **50**, 144–157.

[111] Haberson, S., Harris, K. D. M., Johnston, R. L., Turner, G. W., & Johnston, J. M. *Gaining Insights into the Evolutionary Behaviour in Genetic Algorithm Calculations, with Applications in Structure Solution from Poweder Diffraction Data*. Chem. Phys. Lett., 2002. **353**, 185–194.

[112] Hart, W. & Istrail, S. *HP Benchmarks*.

`http://www.cs.sandia.gov/tech_reports/compbio/tortilla-hp-benchmarks.html`.

[113] R. Ramakrishnan, B. R. & Pekny, J. F. *A Dynamic Monte Carlo Algorithm for Exploration of Dense Conformational Spaces and Heteropoloymers*. J. Chem.Phys., 1997. **106**, 2418–2425.

[114] Beutler, T. & Dill, K. *A Fast Conformational Search Strategy for Finding Low Energy Structures of Model Proteins.*. Protein Sci., 1996. **5**, 2037–2043.

[115] Kirkpatrick, S. & Gellat, C. D. *Optimization by Simulated Annealing*. Science, 1983. **220**, 671–680.

[116] Liu, W. & Schmidt, B. *Mapping of Genetic Algorithms for Protein Folding onto Computational Grids*. TENCON 2005 IEEE Region 10, 2005. pages 1–6.

[117] Cox, G. A. *Protein Modelling: Protein Structure Prediction and Energy Landscape Analysis*. Ph.D. thesis, University of Birmingham, Birmingham, UK, 2007.

[118] He, J. & Turpin, E. *(personal communication)*, 2007.

[119] Coutsias, E. A., Seok, C., & Dill, K. A. *Using Quaternions to Calculate RMSD*. J. Comput. Chem., 2003. **25**, 1849–1857.

[120] MacQueen, J. *Some Methods for Classification and Analysis of Multivariate Observations*. In *Proc. Fifth Berkeley Symp. on Math. Statist. and Prob.*. Univ. of Calif. Press, 1967 .

[121] Cox, G. A., Berry, R. S., & Johnston, R. L. *Characterizing Potential Surface Topographies through the Distribution of Saddles and Minima*. J. Phys. Chem., 2006. **110**, 11,543–11,550.

[122] Auer, S., Miller, M. A., Krivov, S. V., Dobson, C. M., Karplus, M., & Vendruscolo, M. *Importance of Metastable States in the Free Energy Landscapes of Polypeptide Chains*. Phys. Rev. Lett., 2007. **99**, 178,104–1–178,104–4.

[123] Wales, D. J. & Bogdan, T. V. *Potential Energy Surfaces and Free Energy Landscapes*. J. Phys. Chem., 2006. **110**, 20,765–20,776.

[124] Ringnér, M. *What is Principal Component Analysis*. Nat. Biotechnol., 2008. **26**, 303–304.

[125] Rylance, G. J., Johnston, R. L., Matsunaga, Y., Li, C., Baba, A., & Komatsuzaki, T. *Topological Complexity of Multidemensional Energy Landscapes*. P. Natl. Acad. Sci. USA, 2006. **103** (49), 18,551–18,555.

[126] Elmaci, N. & Berry, S. *Principal Coordinate Analysis on a Protein Model*. J. Chem. Phys., 1999. **110** (21), 10,606–10,622.

[127] Yongshou, D., Yuanyuan, L., Lei, W., Jungling, W., & Deling, Z. *Adaptive Immune-genetic Algorithm for Global Optimisation to Multivariable Function*. J. Syst. Eng. Electron., 2007. **18** (3), 655–660.