Analysis and Application of Evolutionary Processes to tackle HIV-1 Entry

Inaugural-Dissertation

zur

Erlangung des Doktorgrades

Dr. rer. nat.

der Fakultät für

Biologie

an der

Universität Duisburg-Essen

vorgelegt von

Reda Rawi

aus Casablanca

Oktober 2013

Die der vorliegenden Arbeit zugrunde liegenden Experimente wurden am Zentrum für Medizinische Biotechnologie (ZMB) in der Abteilung für Bioinformatik der Universität Duisburg-Essen durchgeführt.

1. Gutachter: Prof. Dr. Daniel Hoffmann
2. Gutachter: Prof. Dr. Axel Mosig

Vorsitzender des Prüfungsausschusses: Prof. Dr. Jörg Timm

Tag der mündlichen Prüfung: 12. März 2014

Zusammenfassung

Im Laufe der Jahrmillionen hat die Evolution durch einige einfache Mechanismen wie Mutation, Selektion oder auch Vererbung eine erstaunliche Artenvielfalt hervorgebracht.

Diese Prinzipien können auch beim computergestützten Entwurf von Proteinen und/oder Proteinsequenzen mit gewünschten Eigenschaften, wie z.B. Stabilität oder Funktionalität einer Proteinstruktur, angewandt werden. Da jedoch der mögliche Konformations- und Sequenzraum für bereits kleine Proteine immens groß wird, werden hier vereinfachte Gitterproteinmodelle verwendet.
Im ersten Teil der Promotionsarbeit werden evolutionäre Algorithmen, im Besonderen S Metric Selection - Evolutionary Multi-objective Optimisation Algorithm (SMS-EMOA), implementiert und angewandt um möglichst optimale evolutionäre Parameter zu identifizieren, z.B. Populationsgröße oder Mutationsrate. Interessanterweise spielt die richtige Auswahl der evolutionären Parameter eine entscheidende Rolle bezüglich der Effizienz der Algorithmen.

Im zweiten Teil der Arbeit wird die Evolution von Proteinen beobachtet und analysiert. Ein besonderes Augenmerk wird dabei auf Positionen gelegt, die nicht konserviert sind. Gleichwohl können diese mit kompensatorischen Mutationen an anderen Stellen im Protein strukturell wichtige Funktionen einnehmen. Hierbei werden verschiedene Koevolutionsmethoden, wie z.B. die Mutual Information (MI) oder die Direct Coupling Analysis (DCA), weiterentwickelt und verglichen. Anschließend wird die DCA-Methode mit einer neu verbesserten Gewichtung angewandt um koevolvierende Positionen im Humanen Immundefizienz-Virus (HIV) Hüllprotein-Komplex (Env) vorherzusagen. Bemerkenswerterweise wurden dabei sowohl bereits in der Literatur beschriebene als auch noch unbekannte Po-

sitionen identifiziert, die eine entscheidende Rolle im Eintritt des Viruses in die humane Wirtszelle spielen können. Schließlich wurden die koevolvierenden Positionen bei der Erstellung eines Homologiemodells des Protein-Komplexes verwendet.

Contents

List of Abbreviations

2D	Two-Dimensional
3D	Three-Dimensional
AIDS	Acquired Immunodeficiency Syndrome
AUC	Area Under the Curve
CCR5	C-C Chemokine Receptor 5
CD4	Cluster of Differentiation 4 Receptor
CPU	Central Processing Unit
CXCR4	C-X-C Chemokine Receptor 4
DCA	Direct Coupling Analysis
DI	Direct Information
DNA	Deoxyribonucleic Acid
ECL2	Extracelluar Loop 2
Env	Human Immunodeficiency Virus-1 Envelope
FCC	Face Centred Cubic
FP	False Positive
FRET	Fluorescence Resonance Energy Transfer
GA	Genetic Algorithm

gp120	Glycoprotein 120
gp160	Glycoprotein 160
gp41	Glycoprotein 41
HIV	Human Immunodeficiency Virus
LAV	Lymphadenopathy-Associated Virus
MC	Monte-Carlo
MD	Molecular Dynamics
MI	Mutual Information
MJ	Miyazawa-Jernigan
MOOP	Multi-Objective Optimisation Problem
MSA	Multiple Sequence Alignment
nm	Nanometer
NMR	Nuclear Magnetic Resonance
PDB	Protein Data Bank
ROC	Receiver Operating Characteristic
RNA	Ribonucleic Acid
SIV	Simian Immunodeficiency Virus
SMS-EMOA	S Metric Selection - Evolutionary Multi-objective Optimisation Algorithm
SOOP	Single-Objective Optimisation Problem
TP	True Positive
V1V2	Variable loops 1 and 2
V3	Variable loop 3

List of Figures

List of Tables

1 Introduction

"*Look deep, deep into nature, and then you will understand everything better.*"
Albert Einstein

1.1 Evolution

According to the Oxford Dictionary [37], the concept of Evolution describes "the process by which different kinds of living organism are believed to have developed from earlier forms during the history of earth". Moreover, it is believed that all living organisms originate from one ancestor form, as proposed amongst others by Charles Darwin in the Recapitulation and Conclusion chapter of his book *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life* in 1859: "Therefore I should infer from analogy that probably all the organic beings which have ever lived on this earth have descended from some one primordial form, into which life was first breathed." [32]. Charles Darwin was also the one who introduced the concept of natural selection, which can be summed up by the four main principles variation, competition, adaption and inheritance [32]. The variations between individuals are thereby introduced by mutations. The competition between individuals occurs, because most populations tend to have more offspring than the natural resources are capable of preserving, which leads to the selection of those that are adapted best. The last principle, the inheritance, ensures that the characteristics increasing the probability of survival are inherited to the new offspring generation.
The application of these principles in evolutionary processes generates diversity at all levels of biological organisation, including proteins.

1.1 Evolution

Proteins are one of the most important molecules on earth. They are not only responsible for catalysing nearly all biochemical reactions in cells, but also fulfils functions like gene activity, cell signalling, immune response or structural and mechanical functions in muscles and cytoskeletons. Knowledge of a proteins sequence, structure and function enables thereby scientist to design and implement experiments that can be applied in drug design.
Nevertheless, even for a small protein with a sequence length of 50 amino acids there are 20^{50} possible sequences. In the course of evolution (millions of years) nature explored only a fraction of these possible protein sequences [104], by conserving only biological relevant ones.
One possible approach for the discovery of novel proteins with desired properties or the optimisation of known proteins, is the application and utilisation of evolutionary processes in silico. Basic features of evolution are thereby applied, e.g. reproduction and selection. The fitness, in natural evolution the capability of organisms to reproduce successfully under certain circumstances, is defined according to the designers needs and goals. One typical fitness criterion is the stability of a newly designed protein. Amongst many others [31, 36, 46, 123, 139, 174] Gronwald et al. [63] applied evolutionary Pareto optimisation in order to gain stably folding proteins. However, they performed only a limited number of evolutionary cycles (15 generations), due to the computational very costly fitness calculations of each individual. The evolutionary algorithm required one year Central Processing Unit (CPU) time for 15 generations at eight individuals.

Another interesting research field is the co-evolution within homologous proteins. Positions important for the overall structure and function of proteins are mainly conserved during the course of evolution. Nevertheless, mutations also occur in regions important for the global fold and function, without affecting them. Most of these mutations are linked with compensatory changes in other sequence positions. The identification and analysis of co-evolving positions is of interest for protein contact and de novo structure predictions.

1.2 Research motivation

The first aim of this work was the identification of evolutionary parameters, e.g. mutation rate or population size, that perform best in in silico design of proteins with desired properties. We utilised therefore simplified protein models, which conserve typical protein folding behaviour.
The second aim of this work was the improvement, analysis and comparison of co-evolution detecting methods. Subsequently, the best performing method was applied in order to identify co-evolving and structurally interacting positions in Human Immunodeficiency Virus-1 Envelope (Env), a key protein complex during the entry of Human Immunodeficiency Virus (HIV) into human host cells. Furthermore, we utilised the detected co-evolving positions to predict missing structural regions within the protein complex.

2 Evolutionary Parameter Optimisation

> "*Order and simplification are the first steps toward the mastery of a subject.*"
>
> Thomas Mann

In the following chapter, we apply the S Metric Selection - Evolutionary Multi-objective Optimisation Algorithm (SMS-EMOA) to identify evolutionary parameters that perform best on the Multi-Objective Optimisation Problem (MOOP) protein design. We utilise simplified protein models that conserve typical protein folding behaviour, in order to search through sequence and structure space. First, we give a brief introduction to MOOPs and protein lattice models, followed by an explanation of the applied evolutionary optimisers. Then we present the results followed by a conclusion.

2.1 Introduction

2.1.1 Multi-objective optimisation problems

MOOPs address optimisation problems with more than one objective function. The classical approach solving these kind of problems is to convert the MOOP into a Single-Objective Optimisation Problem (SOOP) by combining the individual objective functions into a single composite one.
The preferred approach solving MOOPs is to obtain the Pareto-optimal solution set or a subset of it. However, we formulate first a minimisation MOOP (without loss of generality) with K objective functions as follows:

2.1 Introduction

"Given an n-dimensional decision variable vector $\mathbf{x} = \{x_1, ..., x_n\}$ in the solution space $\mathbf{X}$, find a vector $\mathbf{x}^*$ that minimizes a given set of K objective functions $z(\mathbf{x}^*) = \{z_1(\mathbf{x}^*), ..., z_K(\mathbf{x}^*)\}$" [85].
Due to the fact that the objectives are in conflict with each other in many optimi-

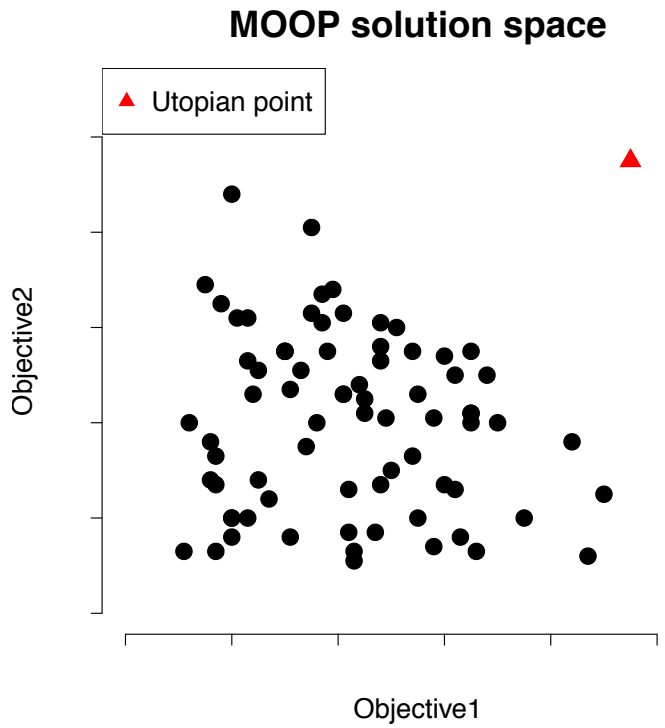

Figure 2.1: Multi-objective solution space including the utopian perfect solution

sation problems, it is rather unrealistic to gain a perfect multi-objective solution as illustrated in Figure 2.1 (red triangle). It is more common that a set of non-dominated solutions represent a more realistic outcome. Dominance is thereby defined as follows:
Solution $\mathbf{x}$ dominates $\mathbf{y}$ if and only if

$$\begin{aligned} &z_i(\mathbf{x}) \geq z_i(\mathbf{y}) \text{ for all } i = 1, ..., K \text{ and} \\ &z_i(\mathbf{x}) > z_i(\mathbf{y}) \text{ for at least one objective.} \end{aligned} \tag{2.1}$$

Furthermore, solution $\mathbf{x}$ is Pareto-optimal, if it is not dominated by any other solution in $\mathbf{X}$. The set of feasible solutions in $\mathbf{X}$ is defined the Pareto-optimal set.

The corresponding objective function values in the objective space are referred to as Pareto front (see Figure 2.2) [33, 85].
As previously mentioned, the ultimate aim for solving a MOOP is the identifi-

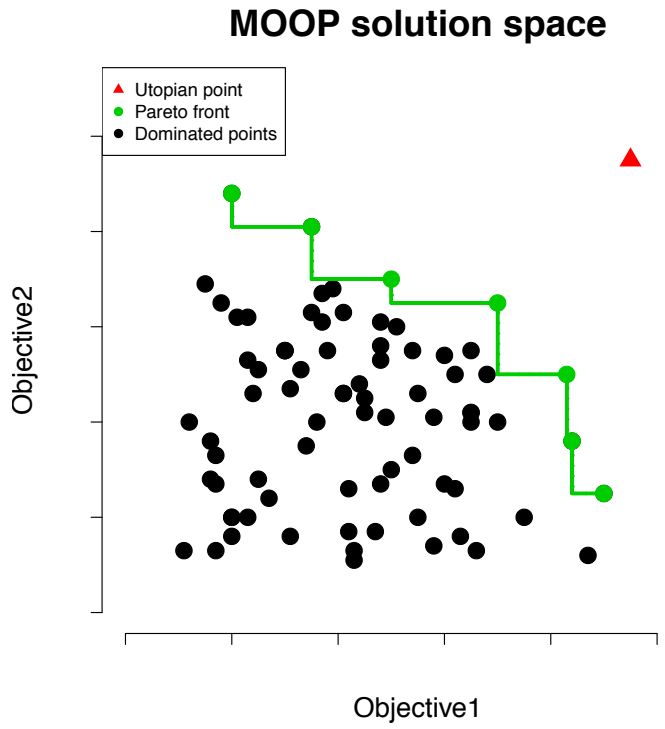

Figure 2.2: Multi-objective solution space including Pareto front

cation of the Pareto-optimal set. However, this is impossible or computationally infeasible for many MOOPs. Hence, a MOOP should attain the following goals:

- The obtained solution set should be as close as possible to the true Pareto front [33].
- The obtained solution set should be as diverse as possible, in order to capture the whole spectrum of the true Pareto front [33].

As already indicated above, most classical multi-objective optimisation algorithms convert MOOPs into SOOPs by applying some sort of a priori knowledge, e.g. weights or constraints/limits. To name a few popular ones, weighted sum

approach [33], ϵ-Constraint method [65], weighted metric methods [33], value function method [33], goal programming methods [19, 74, 94] and interactive methods [7, 15, 18, 76, 112, 162]. However, classical approaches are not practical, since they need to be executed many times in order to find Pareto-optimal solutions and they require a priori knowledge, which might be difficult or even impossible in arbitrary MOOPs.
Later in the thesis (subsection Materials and Methods) we introduce evolutionary algorithms, in particular Genetic Algorithm (GA) and SMS-EMOA, which have proven to be good in order to solve MOOPs.

2.1.2 Protein and protein models

The main building blocks of proteins are amino acids. They are all identical in their backbone formed by an amino group, a central carbon atom (C_α) and a carboxyl group. On the contrary, they differ in their side-chains physiochemical properties, which are bound to the central C_α atom. Some of the side-chains are hydrophobic and prefer other neutral amino acids or non-polar solvent to interact with. Some others are hydrophilic, meaning that they like to interact with other charged or polar molecules and with polar solutions like for instance water.
The primary structure of proteins is defined by their sequence of amino acids, which are covalently linked by peptide bonds in order to form a polypeptide chain. The peptide bond is built by an interaction of one amino acid's backbone carboxyl group with another amino acid's backbone amino group with water as chemical byproduct. The polypeptide chain production is performed during translation[1] of messenger RNA in the ribosomes. It starts with the amino group of the first amino acid (N-terminus) and ends with the carboxyl group of the last amino acid (C-terminus). Immediately after the polypeptide chain production, proteins start to fold into their functional structure, in most cases guided by chaperone molecules, that prevent misfolding of protein parts by interacting with them [177]. Helices and β-sheets are the regular secondary structure of proteins. Helical conformations are stabilised by energetically favourable hydrogen bonding between

[1] Translation is the second central step during gene expression and describes the building of a protein polypeptide chain on the basis of the information coded in the messenger RNA.

2.1 Introduction

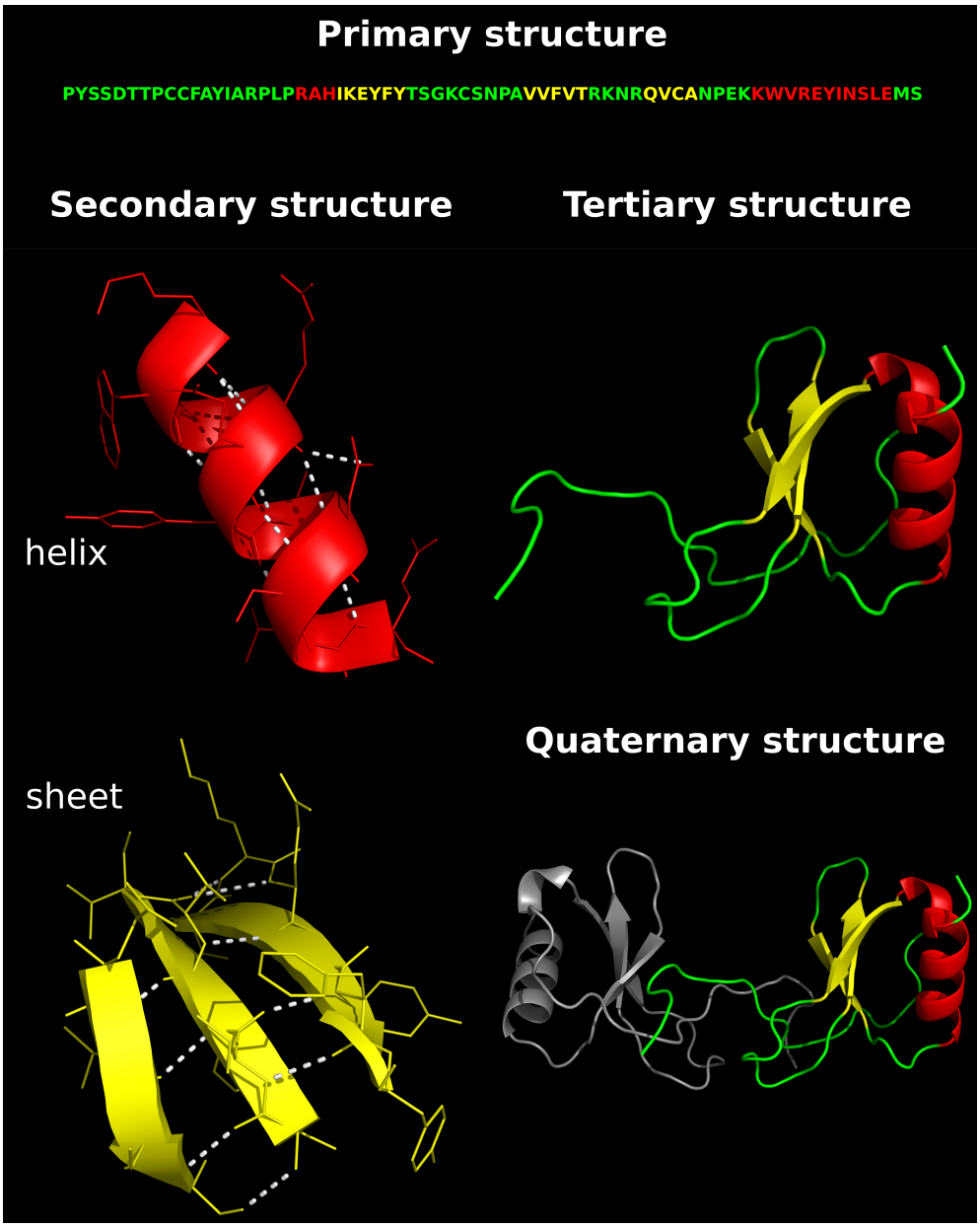

Figure 2.3: Levels of protein structure - We illustrate the four levels of protein structure using the X-ray structure of AOP RANTES [164] (PDB ID: 1B3A): Primary structure - amino acid sequence, Secondary structure - amongst others helices and sheets (hydrogen bonds are indicated by white dashed lines), Tertiary structure - folded conformation of the polypeptide chain, Quaternary structure - complex of two or more polypeptide chains.

backbone oxygen atoms of amino acid's carboxyl group and backbone hydrogen atoms of amino acid's amino group. The most common helical form is the α-helix next to the 3_{10}-helix and the π-helix. They mainly differ in their structural features such as the number of residues per helical turn.
β-sheets are composed of parallel and antiparallel orientated β-strands that can perform hydrogen bonding by their amino group from one side and their carboxyl group from the other side.
Another type of secondary structure are loops, which are less ordered than the previously mentioned helices and sheets. Loops connect helical and sheet secondary elements within a packed folded structure. Moreover, loops are mainly located on the surface of protein folds, where they often fulfil important protein functions, e.g. being part of active or binding sites.
The tertiary structure of proteins is their folded Three-Dimensional (3D)-structure, where the secondary elements are spatially arranged by the following interactions and bonds:

- disulfide bonds
- hydrogen bridges
- van der Vaals interactions
- hydrophobic interactions
- electrostatic interactions.

The hydrophobic forces play an important role, since they almost always pack all hydrophobic amino acids into the core (*hydrophobic core*) in order to avoid the surrounding solution[1]. The charged and polar amino acids are mostly located on the surface of globular proteins, because of their ability to interact with the surrounding solvent. Figure 2.3 illustrates the levels of protein structure using the crystal structure of AOP-RANTES [164] (PDB ID: 1B3A).
According to thermodynamic theory, proteins fold into a conformation with minimal free energy, the native protein state [2]. Nevertheless, depending on the

[1]The surrounding solution of proteins in their natural environment is water.

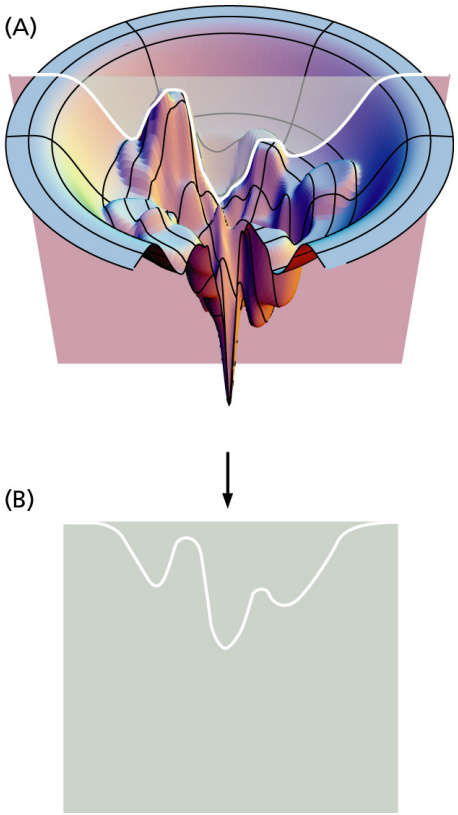

Figure 2.4: Energy landscape - An example energy landscape showing the typical local minima and the native state (centred global minimum). Vertical axis represents the energy in the system. (A) Global view on the energy landscape. (B) Cross-section view. (Copyright 2008 from [183]. Reproduced by permission of Garland Science/Taylor & Francis LLC.)

energy landscape it is possible that proteins fold into other metastable conformations, local minima in the energy landscape, during the folding process. Figure 2.4 illustrates an example energy landscape with several minima and a global minimum representing the native state.

For the purpose of analysis and simulation of proteins and their folding process, protein models are needed that abstract their structure and sequence space as well as the energy functions. Protein structure models can thereby range from detailed all-atom models in 3D applied in Molecular Dynamics (MD) simulations [77] down to coarse grained Two-Dimensional (2D) lattice models [93].

The sequence space is represented by the protein's primary structure, the amino

acids. Models can either apply the set of all 20 naturally occurring amino acids or apply a simplified sequence space, which abstracts the physiochemical properties of amino acids, like for instance the hydrophobicity and polarity model used in the HP-model by Lau and Dill [93].
Protein energy functions are supposed to cover all intra- and inter-protein forces that have an effect on protein folding (see above). Because of the above indicated fact that proteins aspire towards the energetically minimal conformation, protein folding may be defined as minimisation problem for the applied energy function models. Biologically realistic models, like the all-atom energy functions defined in MD simulation force fields, include all intra- and inter-protein forces, but are thus computationally very costly. Simplifications of the energy functions are often realised as distance based potentials or even as contact potentials.
A more detailed introduction into proteins and protein structures can be found in several textbooks, amongst others [12, 96, 138].

2.1.3 Lattice protein models

Lattice protein models are a common simplified representation of proteins. They are mostly applied to perform amongst others folding- [127] and comparative folding-processes [25] or sequence evolution [11]. The proteins are thereby simplified in their sequence and structure space. Furthermore, simplified energy functions are used in order to reduce the complexity during for instance folding experiments. Nevertheless, the abstraction does accompany with loss of accuracy. In the following, we briefly introduce the most common lattices and energy functions.
The classical abstraction of protein sequences is the 'bead on a string' representation, whereupon one amino acid, including backbone and side-chain, is mapped on one bead or sphere. The spheres are thereby connected by rigid edges. Moreover, representations that incorporate side-chains by one additional sphere also exist. Nevertheless, we apply the classical representation for our experiments.

Discrete lattices The abstracted protein strings are in general folded on discrete lattices. The most simple lattice model is the 2D-squared lattice model introduced

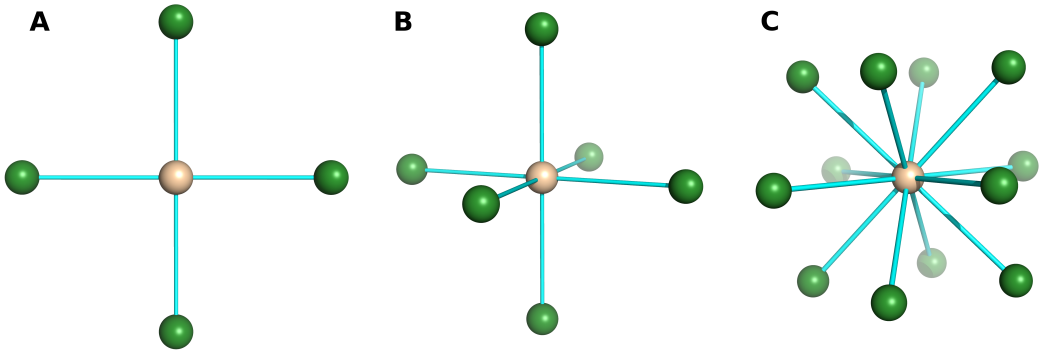

Figure 2.5: Visualisation of lattice's coordination numbers. - The reference sphere is illustrated as wheat coloured sphere, the neighbouring spheres are depicted in green ((A) 2D-squared lattice, (B) 3D-cubic lattice and (C) 3D-face centred cubic lattice).

by Lau and Dill [93]. The coordination number[1] in this lattice is four, which reduces the complexity dramatically at the expense of accuracy. Nonetheless, even on this simple model one can identify protein properties during protein folding, for instance packing of hydrophobic amino acid spheres into the core (*hydrophobic core*).

A more complex model is the 3D-cubic lattice model. It is one of the widely applied protein lattice models, since it is more realistic than 2D ones. The number of nearest neighbours is in this model six, which increases the complexity. On the other hand, more information can be obtained about protein properties and folding behaviour due to the extra dimension.

Another 3D model is the Face Centred Cubic (FCC) lattice model from Raghunathan and Jernigan [130]. The coordination number for this lattice is twelve. Figure 2.5 illustrates the possible nearest neighbours (green spheres) for the central wheat coloured sphere in the introduced lattices. We apply the FCC lattice in our simulations.

The number of possible lattice protein structures for different sequence lengths n in the introduced discrete lattices is given in Table 2.1. It is easily apparent that the number of possible structures in the complex FCC lattice quickly reaches

[1]The number of atoms or ions (spheres) immediately surrounding a central atom in a complex or crystal (lattice). It can be interpreted as a measure of the lattices complexity.

Table 2.1: Number of possible lattice protein conformations

n	2D-square	3D-cubic	3D-FCC
2	1	1	1
3	2	2	4
4	5	6	32
5	13	22	313
6	36	92	3,196
7	98	402	32,835
8	272	1,832	337,056
9	740	8,453	3,452,392
10	2,034	39,640	
$\sim c * b^n$	$0.28 * 2.69^n$	$0.08 * 4.42^n$	$0.05 * 9.57^n$

The number of possible lattice protein structures (self-avoiding chains) for different sequence lengths n. The number already excludes symmetric structures (rotation and reflection products). The last line presents an exponential formula in order to approximate possible conformations for specific sequence lengths n [105].

dimensions, which make a complete enumeration impossible.

Energy functions Due to the structure simplification in lattice protein models, it is not possible to model all forces involved in intra- and inter-protein interactions. Energy functions are therefore mostly approximated by contact potentials, which mimic residue-residue interactions.

The most popular contact potential is the HP-model introduced by Lau and Dill [93]. The HP-model translates the 20 naturally occurring amino acids into two types of residues; H for hydrophobic residues and P for polar residues. The HP-potential is listed in Table 2.2. This simple potential prefers hydrophobic-hydrophobic contacts in contrast to contacts that include polar amino acids. In folding simulations, it is very well suited to discover hydrophobic cores, because it strives for the minimal system energy, which is given by the sum of all contact

potentials.

In the Miyazawa-Jernigan (MJ)-model [113, 114] all 20 amino acids are considered and not split into groups. The MJ-potential is thereby derived from examination of residue-residue interactions within experimentally derived protein structures. The model enables a much more detailed energy calculation than the previously introduced HP one, but has the disadvantage that the complexity is significantly increased. Table A.1 lists the MJ-potentials we applied in our simulations.

Table 2.2: HP-potential according to Lau and Dill [93]

	H	**P**
H	-1	0
P	0	0

2.2 Materials and Methods

2.2.1 Genetic Algorithm

GAs are a popular type of evolutionary algorithms that mimic natural evolutionary processes and parameters, e.g. crossover, mutation or selection, in order to perform optimisation procedures. In contrast to the above-mentioned classical approaches for solving MOOPs, GAs are capable of finding many non-dominated solutions in a single optimisation run. Furthermore, GAs are able to search through different regions of the solution space $\mathbf{X}$, even for difficult MOOPs. The first drafts of GAs were developed by Holland and co-workers in the 1970's [67] and were then extensively applied in many areas, to name only a few, bioinformatics [60, 119, 149, 153], computer science [166], phylogenetics [66] and even engineering and economics. Literature on GAs can be found in quite a few textbooks, for instance Reference [56, 58, 67], while GAs and evolutionary algorithms applied on MOOPs are summarised in detail in [4, 33]. Some of the most commonly used multi-objective GAs are VEGA [136], MOGA [53], SPEA/SPEAII [180, 182],

PAES [84], PESA/PESAII [28, 29] and NSGA/NSGAII [34, 143]. They mainly differ in their approaches to fitness determination, elitism and dispersion along the Pareto front. In most cases, this is realised by different implementation of the selection operator, the most essential parameter during a GA optimisation.
The optimisation during GAs proceeds in cycles, usually referred to as gen-

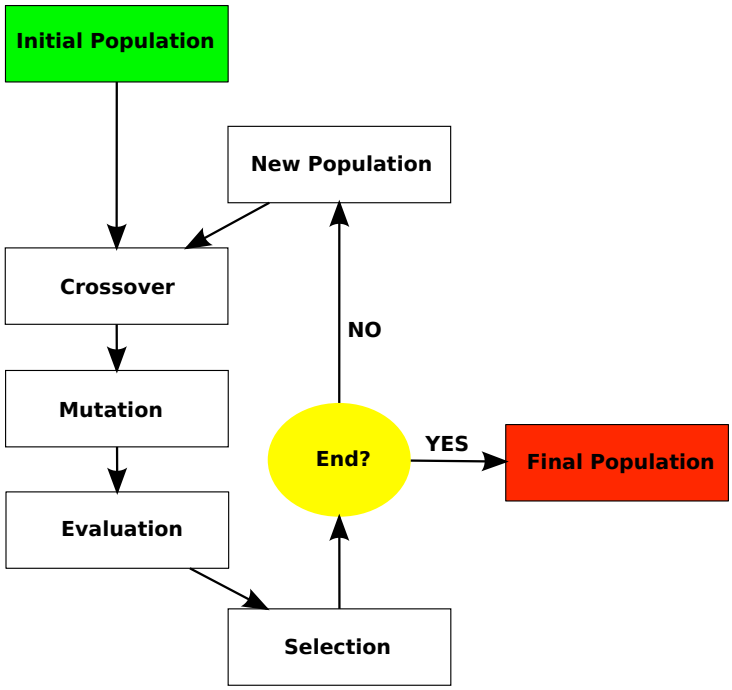

Figure 2.6: GA flowchart - Major steps of a GA.

erations. In every generation, offspring are generated from the parent population. Then the fitness values are determined and the most promising individuals are passed on to the next generation. GA optimisations proceed until a predefined stop criterion is reached. We illustrate the major working steps of GAs as flowchart in Figure 2.6.

Initial population A GA begins the search with a set of chromosomes (set of solutions/individuals) called initial population, which is either randomly chosen or composed of known chromosomes that are supposed to be improved.

Crossover Reproduction operators crossover and mutation are applied on the initial population to produce new offspring. Genes (information) of two or more parent chromosomes are interchanged when using crossover. The most popular crossover operators in GAs are

- the single-point crossover (parent information is split into two halves and interchanged),
- the multi-point crossover (multiple crossover-points are applied)
- and the uniform crossover (parent information is randomly copied).

Mutation The mutation operator is a more local reproduction procedure. The new offspring is changed according to the coding of the chromosomes. For example, in a binary encoding of the offspring, the single bits are mutated with a pre-defined or randomly chosen probability.

Evaluation The next step in the GA cycle is the evaluation process. Depending on the type and number of the objective functions, this step can be the computational most costly one, since sometimes data had to be gained by time-consuming simulations.

Selection The most crucial step for the development of GAs is the selection procedure. As indicated earlier in the text, most GAs differ in the implementation of this operator. According to Darwin's evolution theory, the operator should select the best performing individuals for survival. To mention a few examples,

- roulette wheel selection (the better the individuals are, the higher the probability to be selected),
- Pareto-ranking approaches (individuals are ranked according to the dominance rule and selected on the basis of their Pareto rank) [34, 143, 182],
- and several other selection procedures to maintain diversity in the population by fitness sharing [53, 59], crowding distance [34] or cell-based density

[28, 29, 81, 84, 101, 176].

Stop criterion The last step in the GA cycle is the stop criterion. The optimisation is terminated, if an a priori defined termination condition is satisfied (e.g. number of generations) or a convergence criterion is met, otherwise the GA cycle is run again.

The quality of the performance of GAs (and evolutionary algorithms in gen-

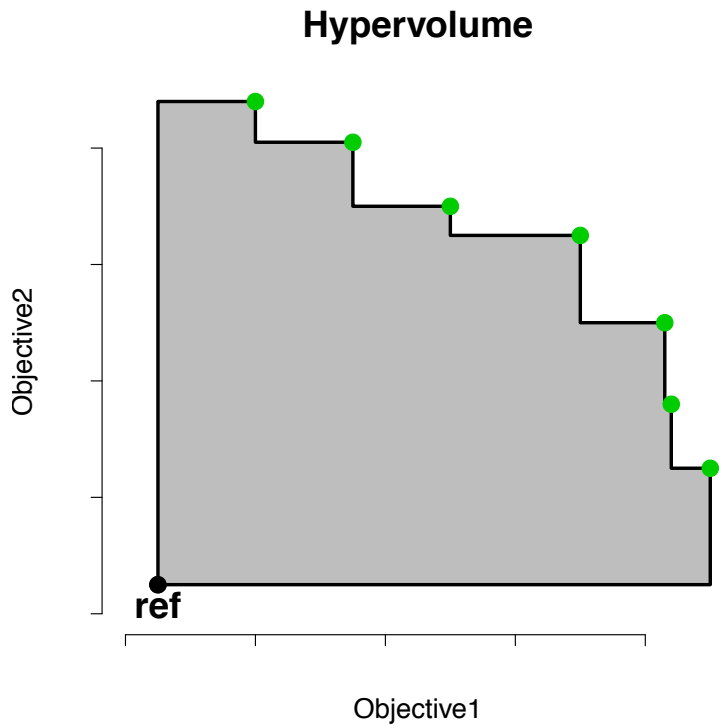

Figure 2.7: Hypervolume - The hypervolume is made up by the grey coloured area between the reference point (black dot) and the actual non-dominated solutions, illustrated as red dots.

eral) is usually determined by the hypervolume [181], a measure that determines the size of the dominated area for a given reference point. The hypervolume is a fair and good performance criterion, since we are always comparing evolutionary optimisers using the same number of iterations/generations. Figure 2.7 illustrates

the hypervolume as grey coloured area between the non-dominated solutions of a 2D maximisation problem in green and the reference point in black. It is desirable to maximise the hypervolume, since it equals the finding of the true Pareto front [136].

2.2.2 SMS-EMOA

The SMS-EMOA applies the hypervolume measure as selection criterion during the optimisation process [9, 44] and incorporates properties from previously developed evolutionary algorithms, e.g. the non-dominated sorting of the NSGAII [34] or archiving strategies from [82, 83]. Furthermore, the algorithm is a steady-state algorithm, i.e. during an evaluation cycle only one offspring is generated by applying reproduction operators and one individual is sorted out by applying selection, which we will explain in detail in the following.
The SMS-EMOA is shown as pseudocode in Algorithm 1 (Page 101) [9]. At the beginning, a population of μ individuals is initialised. Second, a new individual is generated by applying reproduction operators crossover and/or mutation. The now enlarged population of $\mu + 1$ individuals undergoes a two-stage selection process. The primary selection criterion is based on the Pareto dominance. First, all dominated individuals are identified and their dominance number is evaluated. The dominance number specifies the number of individuals that dominate an individual of interest. Figure 2.8 illustrates the removing according to the first selection criterion. Both, the red (**s**) and black (**d**) coloured squared points represent dominated individuals. Point **d** will be removed, since it has the higher dominance number. It is dominated by three other points (black hatched area), while **s** is only dominated by one point (red hatched area). Hence, solution **d** is considered to be more dispensable than **s**, because it is located in a more densely populated area and thus makes less contribution to the diversity. Solution **s** is located in a sparsely populated area and may help the optimiser converging into unexplored solution space areas. If there are multiple individuals having the same dominance number or only non-dominated solutions, the second selection criterion is applied, wherein the individual with the smallest hypervolume contribution is removed. Figure 2.9 illustrates such a situation. Point **s** has the smallest

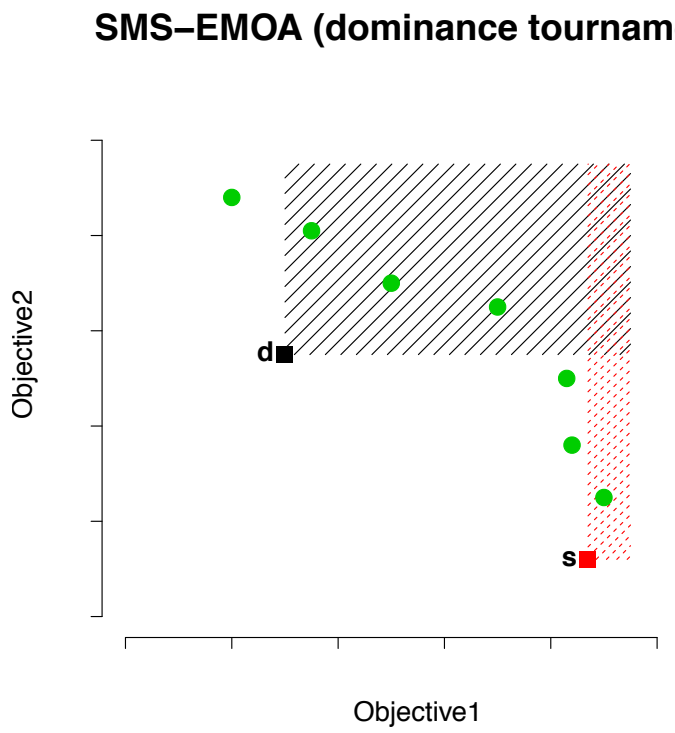

Figure 2.8: SMS-EMOA dominance tournament - The selection of the new population is based on the dominance number. Individuals **d** and **s** are both dominated by green circled points. Subsequently, the individual with the higher dominance number is removed from the population. Individual **d** is dominated by three other individuals (black hatched area), while individual **s** is only dominated by 1 (red hatched area).

hypervolume contribution and is removed from the population.
The SMS-EMOA has numerous advantages. First, the results are well distributed over the Pareto front, since the algorithm focuses the search towards less explored regions near the growing Pareto front. Second, the steady-state approach enables the parallel implementation of function evaluations. For example, we have generated and calculated four individuals in parallel on a computer with four CPUs. We thus reduced the computation time by a factor of four. Furthermore, we introduced a Pareto archive, where we continuously saved the non-dominated solution set for each SMS-EMOA cycle.

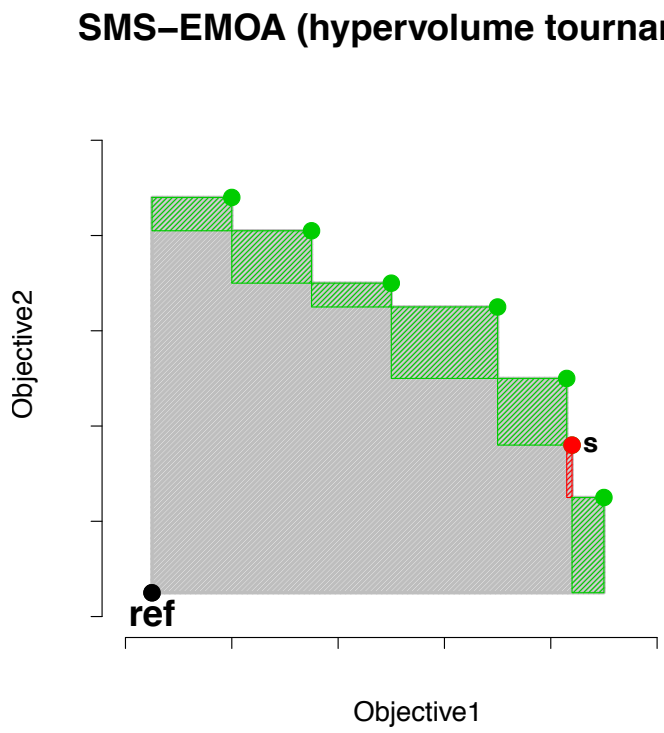

Figure 2.9: SMS-EMOA hypervolume tournament - The selection of the new population is based on the hypervolume contribution. All individuals are not dominated by any other point. Hence, the individual with the smallest hypervolume contribution (**s**) is removed.

2.2.3 Folding simulation via LatFold

We have used the program *LatFold*, embedded in the software package *LatPack* [106], to perform folding simulations of lattice proteins. The folding is applied by Monte-Carlo (MC) simulations using a Metropolis criterion [111].
The Metropolis MC procedure identifies at each step a new random conformation by applying pre-defined move sets. Move sets perform structural changes to a given conformation in order to generate a neighbouring structure in the energy landscape. *LatFold* uses two ergodic move sets. First, the *pull-moves* [95], which perform more local changes, and second *pivot-moves* [103], which in contrast conduct stronger structural changes. Figure 2.10 illustrates the *pull move* implementation on the FCC lattice. The grey coloured current conformation (A)

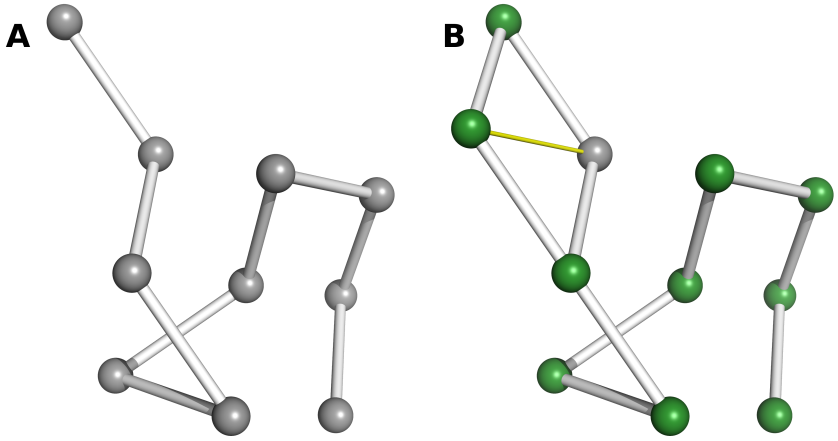

Figure 2.10: Pull move - The actual conformation (A) is transferred by replacement of a single amino acid sphere into a new green coloured (B) conformation.

is transferred by applying a *pull-move* into a new green coloured structure (B). The local change includes only the replacement of one amino acid sphere, indicated by the yellow line. In contrast, a *pivot-move* involves a greater change as shown in Figure 2.11. The grey coloured current conformation (A) is transferred to a new green coloured structure (B) by relocating a whole fraction of the structure, also indicated with a yellow line.
We have illustrated the pseudocode of the Metropolis MC procedure in Algo-

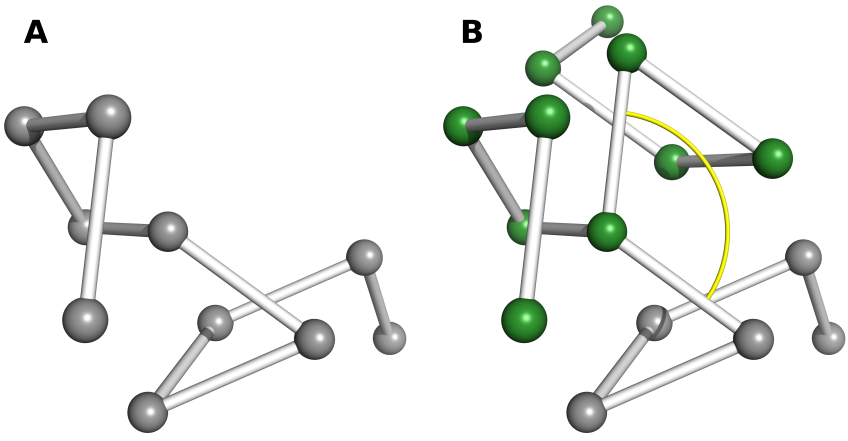

Figure 2.11: Pivot move - The actual conformation (A) is transferred by relocation of a lattice protein's fraction into a new green coloured (B) one.

rithm 2 (Page 102). During the course of the simulation, a new neighbouring conformation is generated at each step by applying the move steps *pull-move* or *pivot-move*. Subsequently, the energy of the new conformation is compared to the energy of the current one. In the case of a lower energy, the new conformation is accepted, while otherwise the Metropolis criterion is applied. The new conformation will then be accepted with probability $e^{-\frac{\Delta E}{k_B T}}$, where ΔE equals the energy difference between the new and the actual conformation ($\Delta E = E_{\mathbf{new}} - E_{\mathbf{actual}}$), k_B represents the Boltzmann constant and T the temperature of the system. The Boltzmann constant k_B cannot be applied, due to the simplified coarse grained energy function [106]. Moreover, the optimal folding temperature T, where the global minimum (native structure) of a lattice protein is adopted and stable, is unknown and has to be approximated for each lattice system independently. We will approximate the product kT for the FCC lattice and the MJ energy function, instead of T and k independently.

2.3 Results

It is computationally very costly to search through all FCC lattice structures, although we are applying simplified protein models (see Table 2.1); e.g. protein Villin Headpiece with a sequence length of 36 amino acids has approximately 9.96^{33} possible conformations. Our main target when applying Metropolis MC simulations for folding is not the identification of native structures for any given sequence, but the sampling of the solution space near native structure conformations in order to test the performance of evolutionary algorithm parameters, like population size, number of generations, mutation or crossover.
However, we have to solve some preliminary issues, before we start the evolutionary optimisations. First, we need to find a native or closely neighboured conformation for our reference structure. The reference structure is the FCC lattice conformation of the arbitrary chosen Villin Headpiece protein (PDB ID: 1VII), a well studied protein due to its short amino acid sequence (36 amino acids) and fast folding kinetics in all-atom protein models applied in MD simulations [6, 63, 80]. We approximate the global minimum energy conformation by application of a

high number of long Metropolis MC simulations at different temperatures. Subsequently, we use the detected native structure and perform a kT-screening in order to identify the relative folding temperature kT. Furthermore, we need to choose the number (seeds) and length (simulation steps per MC run) of MC simulations to perform.

2.3.1 Preliminary experiments

Approximated native conformation We have performed Metropolis MC simulations of Villin Headpiece sequence using 100 seeds and two different simulation lengths:

1. **short** simulations with 1,000,000 Metropolis MC simulation steps.
2. **long** simulations with 10,000,000 Metropolis MC simulation steps.

The *Latpack* developer used in a similar test case only 10,000 steps. Furthermore, we applied the simulations using 10 different kT values ($kT = \{0.05, 0.1, ..., 0.5\}$). In Figure 2.12 we illustrate the **short** (coloured red) and **long** (coloured black) simulations minimum energies (100 different seeds) that have been reached in the course of the Metropolis MC runs. The green horizontal line represents the smallest energy found in all simulations. It is readily apparent that small and large relative folding temperatures do not come within the scope of the minimum energy. It is also striking that the boxes representing small kT values, indicating a higher standard deviation, are apparently bigger than the boxes for large kT values.
Conformations with minimum MJ energies or neighbouring ones are mainly adopted at kT values between 0.15 and 0.3. Hence, we have deepened the analysis in this range of values (see Figure 2.13) and noticed that almost all MC runs assume conformations that are close to the minimum energy one. Astonishingly, none of the MC runs in Figure 2.13 yielded a smaller energy than the highlighted one (green horizontal line), which strengthen the assumption that this is the native or near native minimum energy structure. To be sure, we have to enumerate all structures, which is computational very infeasible.
Moreover, we have plotted the mean energy of each run in Figure 2.14. Noticeable,

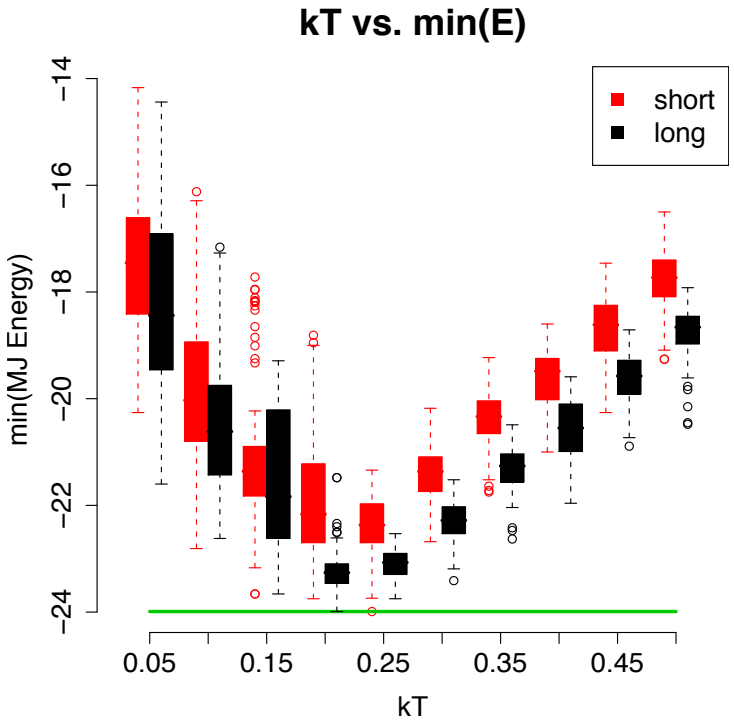

Figure 2.12: Boxplots of minimum MJ energies - Short (red boxes) and **long** (black boxes) simulations minimum MJ energies shown as boxplots for 100 MC seeds. Simulations were performed using 10 different kT values (x-axis). The green line represent the minimum energy found in all simulations.

the mean energies strongly increase with increasing temperature kT, while the standard deviations decrease. The same observations are made when examining the more detailed area between 0.15 and 0.3 kT (see Figure A.1).
We want to explain the previously made observations with the aid of Figures 2.15, 2.16 and 2.17. Figure 2.15 illustrates the energy traces for three different Metropolis MC seeds (coloured black, red and blue) using the relative folding temperature 0.05 kT. The simulations are immediately trapped in a local minimum after a short equilibration time. The relatively low temperature of 0.05 kT has the consequence that energetically inferior conformations are not accepted via the Metropolis criterion, since the probability is getting lower the smaller the kT. This explains also the relatively stable mean energies during the simulations.

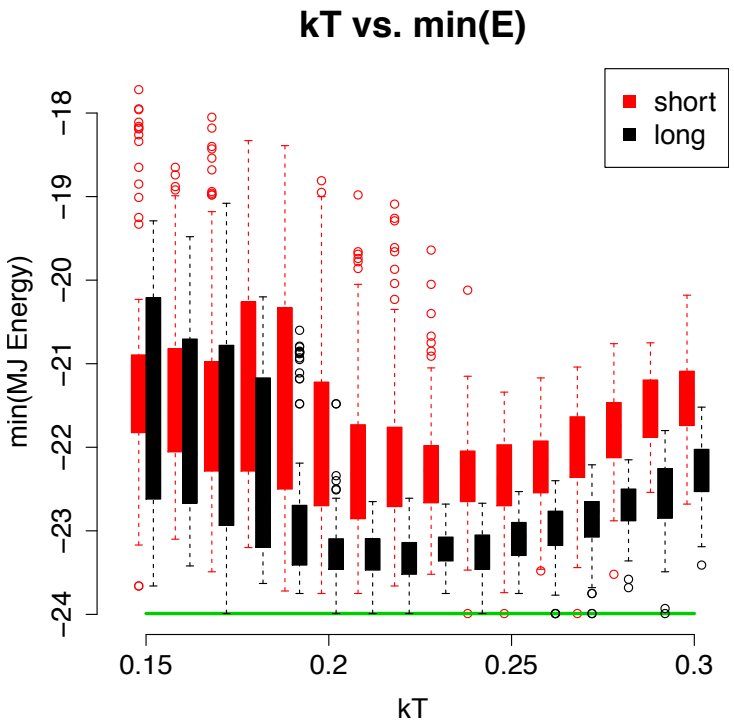

Figure 2.13: Detailed boxplots of minimum MJ energies - Short (red boxes) and **long** (black boxes) simulations minimum MJ energies shown as boxplots for 100 MC seeds. Simulations were performed using kT values ranging from 0.15 to 0.3 (x-axis). The green line represent the minimum energy found in all simulations.

On the other hand, Figure 2.17 illustrates a completely different picture. Due to the relatively high temperature of 0.5 kT, each new conformation is accepted, either because it has a lower energy or by application of the Metropolis criterion, whose probability gets higher the larger the temperature. Thus, a convergence towards the native structure is always prevented.
Figure 2.16 illustrates a more common energy trace behaviour. All three simulations converge towards a MJ energy plateau around -18. Due to the still sufficient high temperature, it is always possible to fall by chance into a deeper energy minimum, as realised in the first simulation (black curve/trajectory).
As already shown in Figure 2.13, several simulations discovered conformations with the same minimal energy. Figures 2.18 and 2.19 depict all three found min-

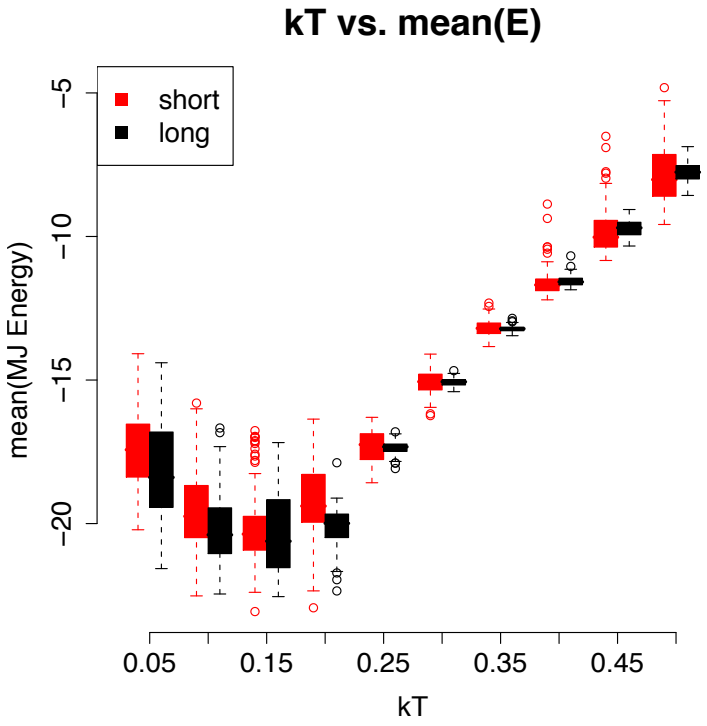

Figure 2.14: Boxplots of mean MJ energies - Short (red boxes) and **long** (black boxes) simulations mean MJ energies shown as boxplots for 100 MC seeds. Simulations were performed using 10 different kT values (x-axis).

imum energy conformations. The conformations differ only in the placement of two amino acid lattice spheres. The differences between conformations A and B and conformations A and C are highlighted by yellow dashed lines.
Moreover, conformation C has another energy contributing interaction than conformation A. Nevertheless, there is no global energy difference, since in both cases the amino acid glutamatic acid (GLU) is involved in the contact.

Relative folding temperature kT In the following, we applied the predicted minimum energy conformation (choose randomly one of the three) in order to perform a kT-screening. The purpose of the kT-screening is the identification of an optimal folding temperature.
The screening procedure is as follows. First, Metropolis MC simulations are per-

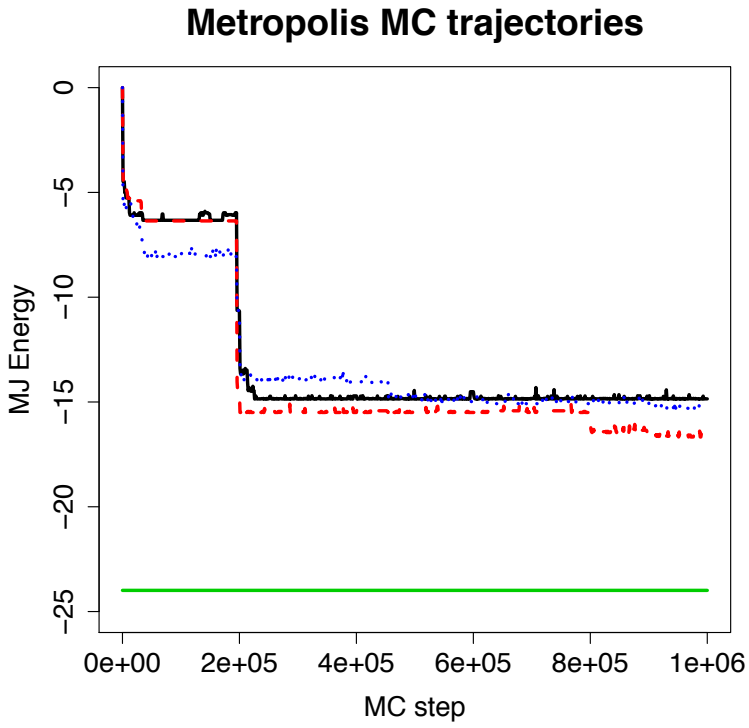

Figure 2.15: Metropolis MC energy trajectory ($kT = 0.05$)

formed using different numbers of seeds and kT values. The MC runs are thereby carried out at a constant kT value. During these simulations the proportion of the unchanged minimum energy conformation is counted. Figure 2.20 illustrates the kT-screening applying **short** MC simulations. Astonishingly, the number of seeds does not have an effect on the proportion of the minimum energy conformation during the MC runs. Furthermore, at 0.25 kT there is no minimum energy conformation present anymore. Hence, we decided to use 0.15 kT, since it maintains a good balance between conserving a well folded lattice conformation and variation by the use of the Metropolis criterion.

2.3.2 SMS-EMOA runs

Workflow In the following, we will perform SMS-EMOA optimisations in order to screen for evolutionary parameters that perform best in the prediction of low energy lattice conformations for given sequences (protein structure prediction).

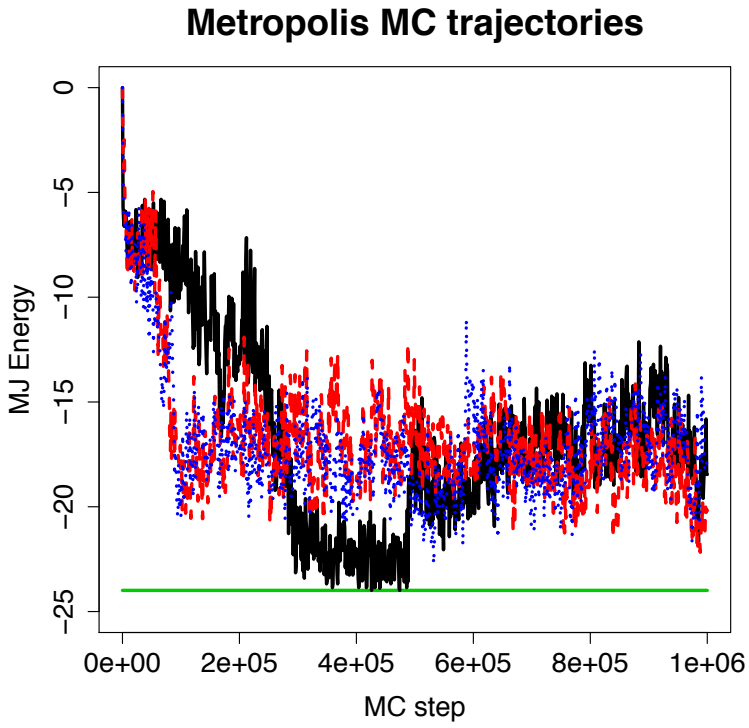

Figure 2.16: Metropolis MC energy trajectory ($kT = 0.25$)

During single SMS-EMOA runs we will apply the following criteria. First, five **short** Metropolis MC simulations are performed in order to evaluate the fitness for each individual, since the number of seeds did not have an effect as shown in the previous subsection. Second, we apply the constant relative folding temperature of 0.15 kT during the fitness evaluations. Furthermore, we split the trajectory into two parts, the equilibration ($\frac{1}{5}$ simulation length) and the production stage ($\frac{4}{5}$ simulation length). The conformations and energies important for the fitness evaluation are exclusively derived from the production stage. Once the simulation has been carried out the following two fitness functions are applied:

1. mean MJ energy.
2. similarity.

The first fitness function ensures that we compare contact potential energies of conformations, which are most frequently adopted during the MC simulations and

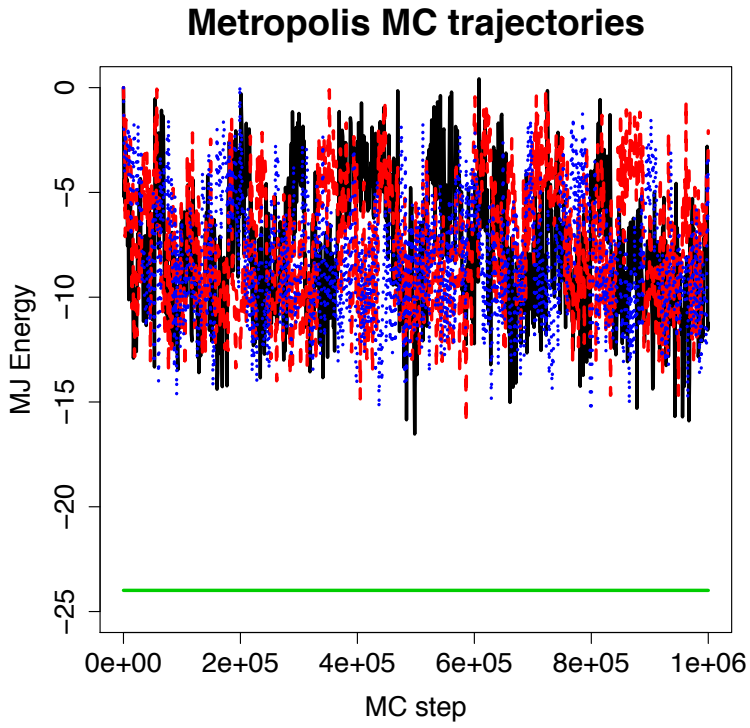

Figure 2.17: Metropolis MC energy trajectory ($kT = 0.5$)

therefore represent the most probable lattice structure for a given individual's sequence.

The second fitness function, the similarity, is defined in order to obtain solutions that are similar to a pre-defined reference structure. The previously predicted minimum energy lattice conformation of Villin Headpiece serves in our case as the reference conformation. The similarity is implemented as follows: A cubic lattice box, with a mesh size of 0.5 Å and an offset of 15 Å, is centred on the reference structure. The box size is chosen quite generously to ensure that all lattice protein structures fit in. Afterwards, each grid point is assigned with a hydrophobicity potential ϕ, which is calculated by

$$\phi = \sum_{i=1}^{n} hydro(res_i) \cdot e^{-dist}, \tag{2.2}$$

with the protein sequence length n, res representing the individuals's amino acid and the distance $dist$ from each amino acid to the current grid point. The function

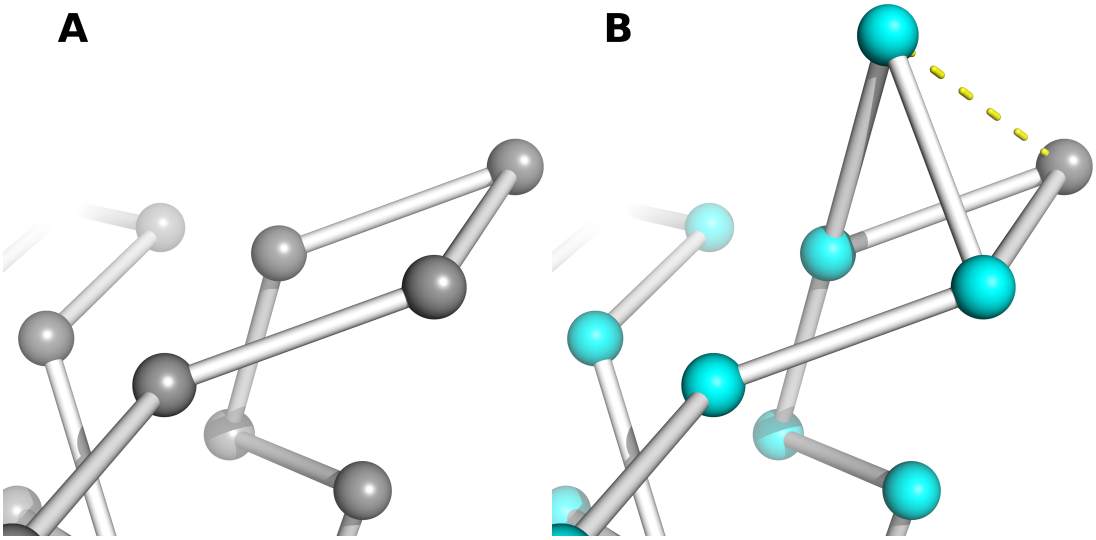

Figure 2.18: Minimum energy conformations A and B - The conformations A and B differ in the placement of one amino acid sphere, indicated by a yellow dashed line. The difference does not have an effect on the overall energy potential.

hydro assigns each residue a hydrophobicity value according to the Kyte-Doolittle scale [90] (listed in Table A.2). Large negative values correspond to strong polar amino acids, while large positive values represent hydrophobic amino acids. The next step is the definition of a hydrophobicity hull around the reference structure. Each amino acid bead of the reference conformation is surrounded by a sphere with a van der Vaals radius of 8 Å. Next, the hull is defined as the surface of these spheres and is the set of grid points, which are considered in the fitness calculation. The hull fulfils the criterion that the same set of points can be derived for every newly generated individual's conformation. This is conducted by the superimposition of every new conformation onto the reference one. The superimposition is performed using the *epitopsy*[1] library coded in Python[2]. Due to the fact that we are considering the same grid points for every newly generated individual's conformation, we are now able to compare the hydrophobicity hulls, by determining the fraction of grid points with the same sign. The resulting fitness function values range thus from zero to one, with one as perfect match of

[1] https://code.google.com/p/epitopsy
[2] http://www.python.org

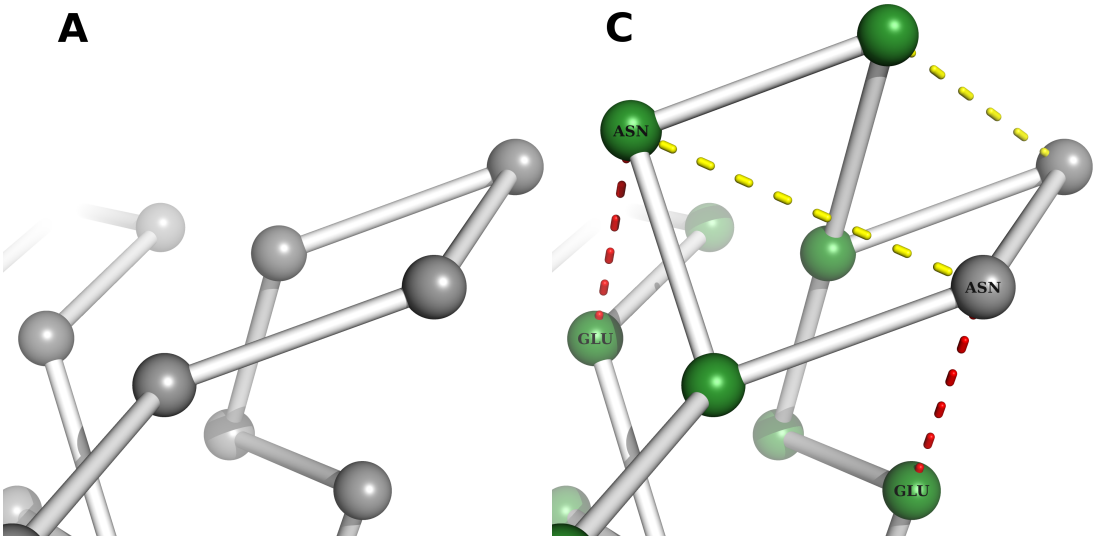

Figure 2.19: Minimum energy conformations A and C - The conformations differ in the placement of two amino acid spheres, indicated by yellow lines. The overall energy potential does not change, since the energy contributing contacts are performed by the same amino acids asparagine and glutamic acid, highlighted by red dashed lines.

the hydrophobicity hulls.

Long SMS-EMOA runs Our main goal is the identification of evolutionary parameters that require a minimum number of computational time as possible to detect new protein sequences and structures having desired properties, e.g. a stable fold or a similarity to a pre-defined protein. We must therefore try to answer the following questions:

- How big should the population size be?
- What is the best performing mutation rate?
- Should we apply crossover?
- How many SMS-EMOA generations should at least be performed?

We have started the analysis with SMS-EMOA optimisations applying the evolutionary parameters listed in Table 2.3.

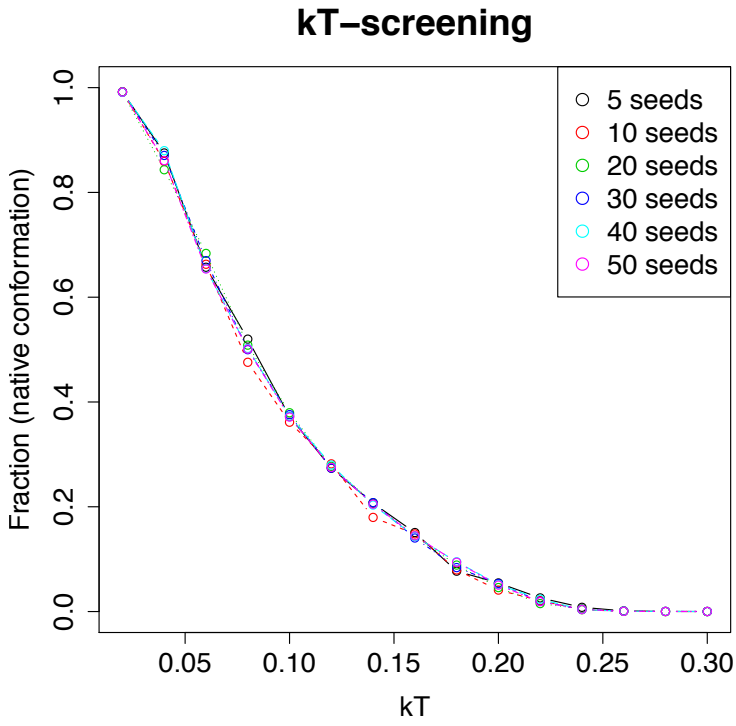

Figure 2.20: kT-screening - Number of seeds does not have an effect on the fraction of global minimum folds. At 0.25 kT no global minimum conformations are present anymore in the trajectory.

Typical illustrations of evolutionary algorithm performances are shown in Figures 2.21 and 2.22.

In Figure 2.21 we plotted the solution space of the SMS-EMOA optimisation using the evolutionary parameters from the second row in Table 2.3. The black circles represent the solution space elements of all individuals generated during the SMS-EMOA run. The red connected filled circles highlight solutions that originate from the last Pareto front. It is readily apparent that the solutions are well distributed over the similarity axis (y-axis), since the axis's range is from zero to one.

The hypervolume development of the same SMS-EMOA run is shown in Figure 2.22. The optimisation converges after around 7000 simulation steps. This explains the accumulation of black circles on the right side of Figure 2.21, because

Table 2.3: Parameters applied in the first SMS-EMOA runs

Simulation steps	Population size	Mutation rate	Crossover
15000	10	$0.014(\frac{1}{2n})$	no
15000	10	$0.028(\frac{1}{n})$	no
15000	10	0.05	no
15000	10	0.075	no
15000	10	0.1	no
15000	10	$0.014(\frac{1}{2n})$	one-point
15000	10	$0.028(\frac{1}{n})$	one-point
15000	10	0.05	one-point
15000	10	0.075	one-point
15000	10	0.1	one-point

it seems that the maximum range of the mean MJ energy is reached, although we do not know where the limit is. A closer examination of the Pareto individuals leads to the following conclusions. The Pareto individuals at the far right side, with high mean MJ energies and low similarities, are highly mutated at sequence level. The sequences consist almost exclusively of polar charged amino acids like aspartic and glutamic acids or lysine, arginine and histidine. Some of the sequences incorporate also a lot of cysteines. The reason for this lies in the fact that the MJ contact potential (see Table A.1) prefers polar and cysteine-cysteine contacts more than others, like hydrophobic ones. The second fitness function, the similarity, prevents the optimisation run to exclusively generate this kind of sequences, by forcing it to search for sequences that have a similar amino acid composition like the reference conformation.

Figure 2.23 shows the hypervolume developments of SMS-EMOA runs applying evolutionary parameters listed in Table 2.3. Red and black lines highlight hypervolume trajectories of simulations applying one-point crossover and no crossover respectively. The different graphical symbols triangle, circle, plus sign, x and diamond differentiate between optimisation runs applying mutation rates $\frac{1}{n}$, $\frac{1}{2n}$ (n =sequence length), 0.05, 0.075 and 0.1.

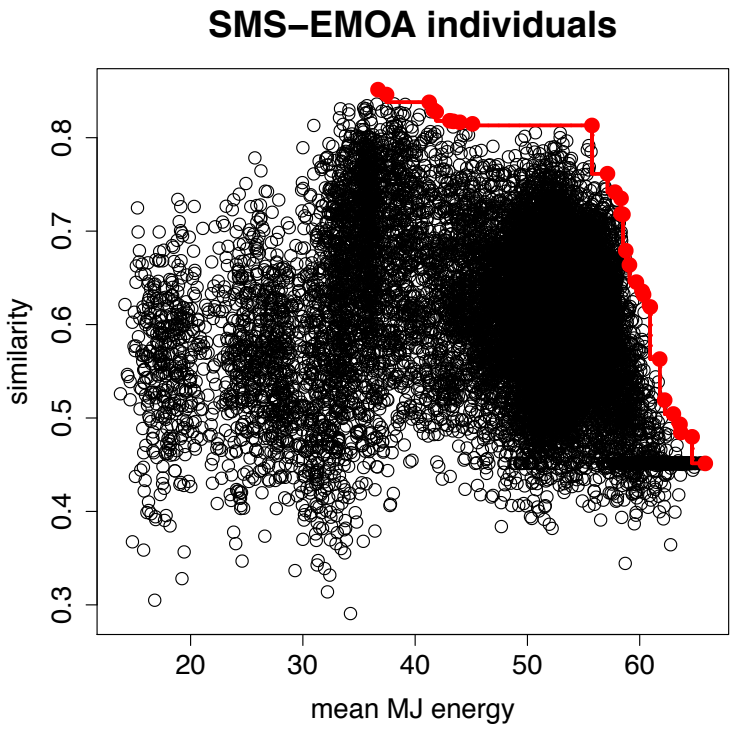

Figure 2.21: SMS-EMOA solution space - Black circles represent solution space entries of all individuals created during the SMS-EMOA run. Red connected circles show the solution space entries for the last Pareto front.

First, it is striking that all trajectories start converging between 5000 and 10000 simulation steps. On closer inspection, we observe the same rapid convergence of the different optimisations with respect to the first fitness function, the mean MJ energy. As shown in the example above, this is due to the highly mutated sequences, which mainly consist of polar charged amino acids. The further relatively small increases in the trajectories derive mainly from the search of better solutions in terms of the second fitness function, the similarity. Depending on the distribution of solutions along the Pareto front, it can be very random and time consuming until better solutions are found. We will describe this by the help of Figure 2.21. There are more Pareto solutions, red circles in Figure 2.21, with a high mean MJ energy (bottom right side) than solutions with low mean MJ energy and high similarity. Due to the fact that the SMS-EMOA only choose

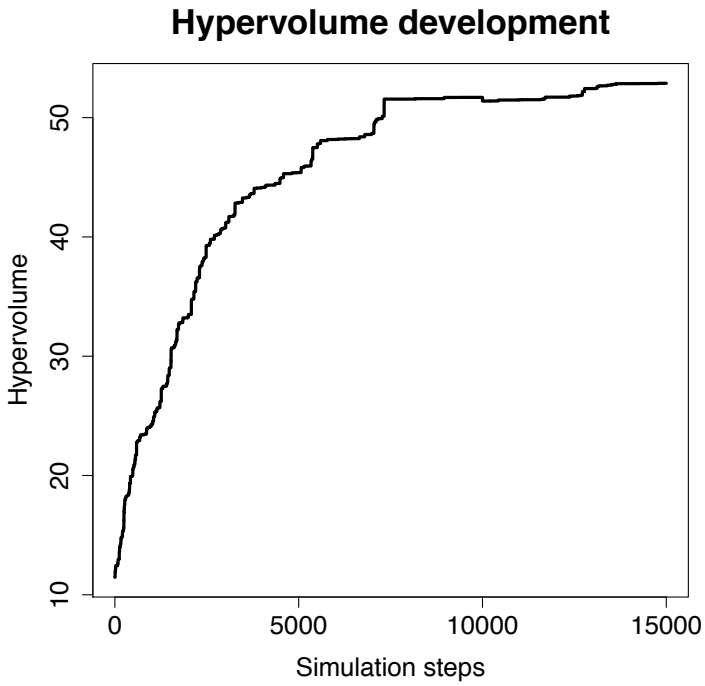

Figure 2.22: Hypervolume development - Typical hypervolume development during a SMS-EMOA optimisation run.

one individual for replication, it occurs less frequently that an individual with high similarity is chosen. Hence, the algorithm gets less chance to improve it's individuals in this area of the solution space.
Moreover, it is noticeable that the trajectories originating from optimisations not using crossover cluster when converging. In contrary, red lines are much more variable, which is due to the properties of crossover operators. Crossover exchanges information, in our case amino acids, from different individuals, resulting in strong reallocation of solution space entries. Hence, the solution space is searched through in larger steps, while mutation only performs local changes.

2.3.3 SMS-EMOA Enumerator

In the following, we carried out a series of experiments (Enumerator) with the assumption that only limited computational resources are available. We performed

2.3 Results

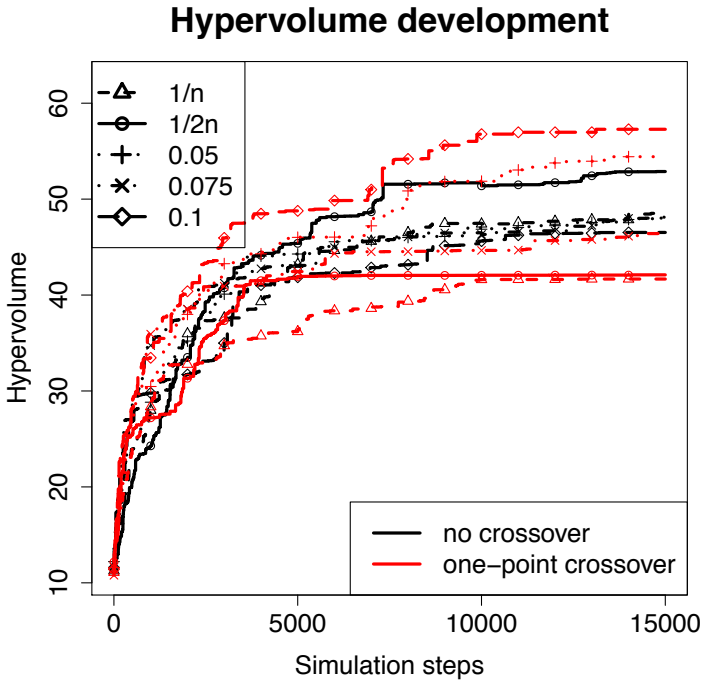

Figure 2.23: Hypervolume development comparison - Comparison of hypervolume developments of SMS-EMOA runs using the evolutionary parameters listed in Table 2.3. Black lines indicate hypervolume developments with no crossover parameter applied. In contrary, red lines represent developments of SMS-EMOA runs including one-point crossover application. The graphical symbols triangle, circle, plus sign, x and diamond refer to SMS-EMOA runs using different mutation rates, in particular $\frac{1}{n}$, $\frac{1}{2n}$ (n =sequence length), 0.05, 0.075 and 0.1 as shown in the legend in the top left corner.

SMS-EMOA optimisations with a maximum of 5000 generations (5000 evaluations steps) in order to identify evolutionary parameters that perform best.
The evolutionary parameters applied in the experiments are the following:

- evaluation step: $e = 5000$.
- population size: $p = \{2, 5, 10, 50, 100\}$.
- mutation rate: $m = \{0.01, 0.02, ..., 0.1, 0.2, ..., 0.5, 0.75, 1\}$.
- crossover: $c = \{0(\text{no crossover}), 1(\text{one-point crossover})\}$.

2.3 Results

For each parameter set we performed three different seeds.
Figures 2.24 illustrates the hypervolume solutions, which have been produced

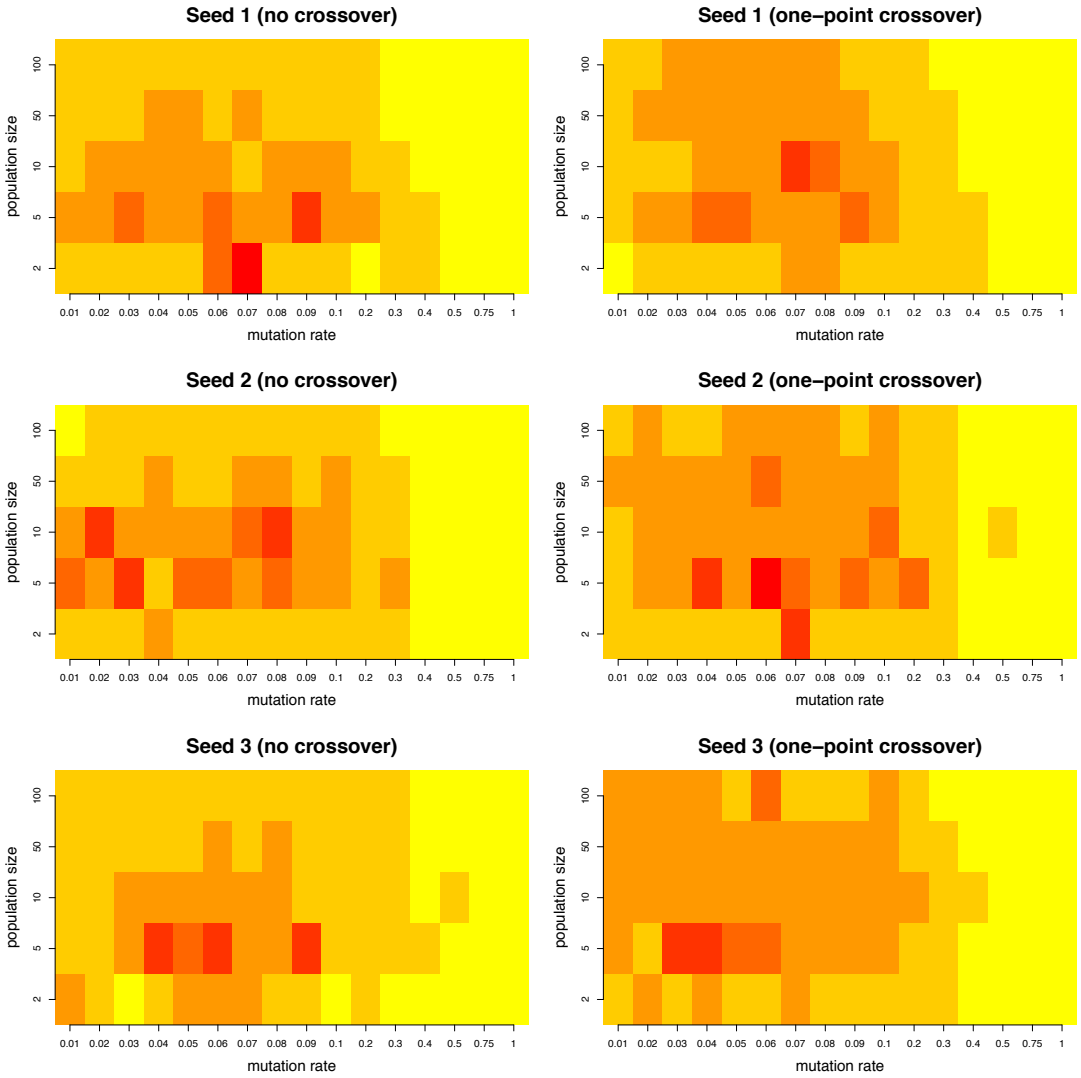

Figure 2.24: Enumerator heat map - The hypervolumes of all parameter set combinations are shown as heat map. In general, the brighter the colour (yellow) the lower the hypervolume and vice versa, the more intensive the red the higher is the hypervolume for the optimisation run applying these parameter sets. The three left-hand figures represent SMS-EMOA runs not applying crossover, while the three right-handed apply one-point crossover. The three different rows indicate the different seeds that have been applied.

during the SMS-EMOA optimisations. The embedded plots are structured as follows. The plots on the left-hand side represent optimisations that did not apply crossover, on the right-hand side those that apply one-point crossover. The three different rows indicate the three distinct seeds that have been applied. The hypervolume size is encoded in heat colours. The brighter the colour (yellow) the lower the hypervolume and vice versa, the more intense the red the higher is the hypervolume. The mutation rates are plotted on the x-axis, the population sizes on the y-axis.

The first thing that strikes is that mutation rates greater 0.1 ($m > 0.1$) lead to poor hypervolume results (yellow squares). Because of this, we made a close up view shown in Figure 2.25, where the x-axis ranges from mutation rates 0.01 to 0.1. The first thing one can notice is that the performances (hypervolumes) in the first three population sizes ($p = \{2, 5, 10\}$) are similar between no and one-point crossover application in all three seeds. However, in optimisations with higher population sizes ($p = \{50, 100\}$), there is a clear advantage (dark orange opposite to bright orange) in the inclusion of crossover, in particular one-point crossover. Furthermore and most interestingly, the best hypervolumes (red squares) are mainly obtained when using population size five in some cases population sizes two and ten. The application of the crossover operator does not have an impact. With respect to the mutation rates, it is quite difficult to name the best mutation rate regarding the performance. Hence, we simplified the results by plotting the mutation rates as a function of the averaged hypervolumes of all seeds (see Figure 2.26). The different population sizes are highlighted by the application of various line types, graphical symbols and colours (see Figure legend). The error bars represent the standard deviation within the three different seeds. It is quickly apparent, that population size five is the superior one with respect to the hypervolume, as already suggested above. In the case of optimisations with no crossover (upper plot), it is difficult to point out one single superior mutation rate, while in optimisation using one-point crossover (bottom plot), mutation rate 0.04 is the best performing one. Moreover, it is astonishing that the choice of the mutation rate does not have a decisive impact when applied larger population sizes ($p = \{10, 50, 100\}$), shown by almost horizontal green, blue and cyan coloured curves.

2.3 Results

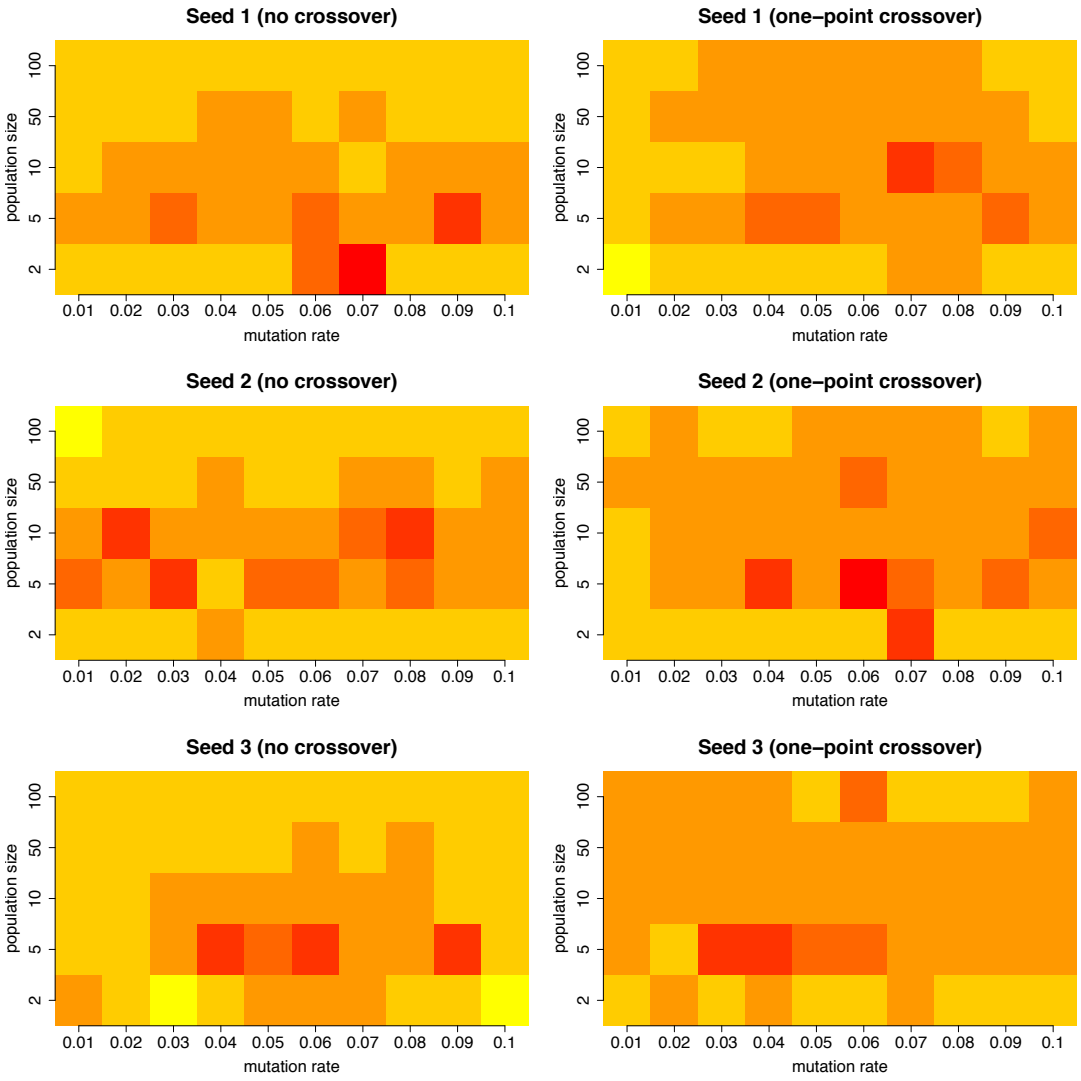

Figure 2.25: Enumerator heat map (close-up view) - The hypervolumes of all parameter set combinations are shown as heat map. In general, the brighter the colour (yellow) the lower the hypervolume and vice versa, the more intensive the red the higher is the hypervolume for the optimisation run applying these parameter sets. The three left-hand figures represent SMS-EMOA runs not applying crossover, while the three right-handed apply one-point crossover. The three different rows indicate the different seeds that have been applied.

In summary, the combination of the correct evolutionary parameters plays a crucial role in the performance of the optimiser (SMS-EMOA).

2.3 Results

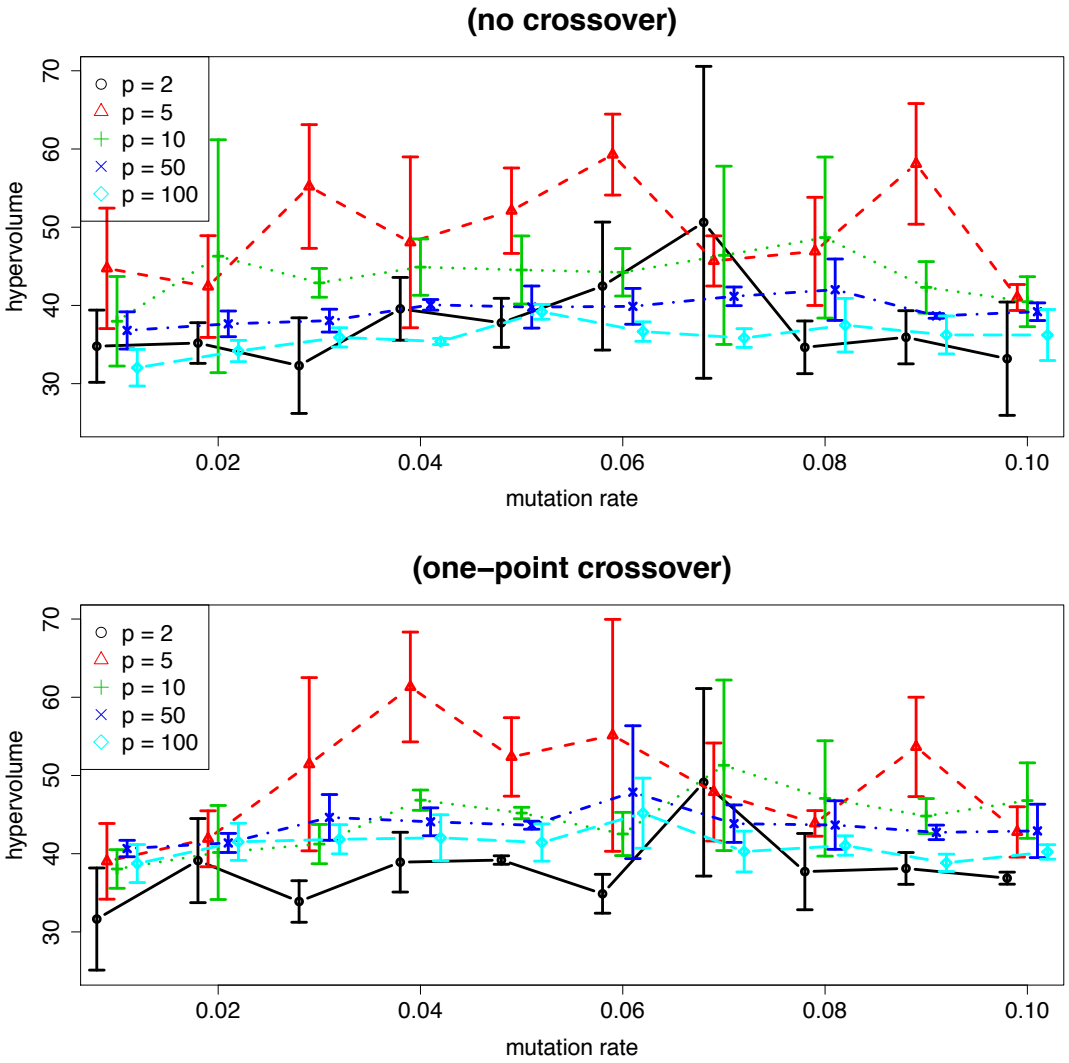

Figure 2.26: Enumerator - mean of three seeds - Averaged hypervolumes of three seed are shown in the plots. The colours, line types and graphical symbols distinguish between optimisations applying different population sizes. The upper plot represents hypervolume output from optimisations not applying crossover, while the bottom plot depicts simulations including crossover operation. The error bars represent the standard deviation of uncertainty.

2.4 Conclusion

We have seen from the extended analysis that the choice of the evolutionary parameters, like population size, mutation rate or crossover, can be decisive in terms of performance.
Starting with the crossover, it seems as if the introduction of crossover (here: one-point crossover) has beneficial effects. Optimisations using high population sizes ($p = \{50, 100\}$) benefit in terms of hypervolume when introducing crossover, whereas the mutation rate is not affected.
Second, the choice of the mutation rate is also essential. Overestimated mutation rates ($m > 0.1$) lead to poor hypervolume results. This is due to the fact that the evolutionary optimiser does not have the chance to converge to an optimum. The mutation rate forces, once a good individual in terms of fitness functions is found, the sequence to mutate strongly. Hence, it is always advisable to choose a more moderate mutation rate, at least less than 0.1 ($m = 0.1$).
The population size is according to our analysis the main determining parameter. The properties of the SMS-EMOA should always be kept in mind when deciding for a proper population size, since only one individual is chosen for reproduction per evaluation step. In contrast, in GAs the complete population is altered by application of crossover and mutation operators. Hence, if we have computationally limited resources and we are interested in quickly finding good results, the population size should be small. In the case of unlimited computational and time resources, one should choose bigger sizes in order to obtain many good and diverse solutions, which capture the whole spectrum of the real Pareto front.
For the MOOP of protein design, we suggest more moderate population sizes like five or ten, since the fitness evaluations are very costly, when applying more detailed protein models, e.g. all-atom models used in MD simulations. Furthermore, we also suggest the introduction of crossover operators, at least on sequence level, since they may enable the optimiser to quickly search in other regions of the solutions space. The mutation rates, however, should be at least lower than 10% of the sequence length, in order to locally search the solution space around an individual. In matters of the mutation rate one can think of an adaptive strategy.

2.4 Conclusion

This can be accomplished by adjusting the mutation rate depending on the convergence condition of the optimiser. For instance, if the SMS-EMOA is already converged with respect to one fitness function, then apply only local search by lowering the mutation rates.

3 Co-evolution in HIV-1 Env

"*In nature we never see anything isolated,*
but everything in connection with
something else which is before it, beside it,
under it and over it."
Johann Wolfgang von Goethe

3.1 Introduction

3.1.1 HIV

HIV/AIDS history The Acquired Immunodeficiency Syndrome (AIDS) was first clinically detected in 1981, where young men exhibit symptoms of Pneumocystis Carinii Pneumonia or of the rare skin cancer Kaposi's Sarcoma, both infections present in people with affected immune systems [54, 62, 73]. The virus causing the disease was then identified by the working group of Luc Montagnier in 1983, called Lymphadenopathy-Associated Virus (LAV) [5] and one year later by Robert Gallo and co-workers, termed HTLV-III [55]. LAV and HTLV-III were renamed in 1986 to HIV, since both groups isolated the same virus. HIV occurs in two types, HIV-1 and HIV-2, where both types most likely evolved from Simian Immunodeficiency Virus (SIV) which infects non-human primates as chimpanzees and gorillas or sooty mangabey respectively [141]. HIV-1 is significantly more common than HIV-2, which is regionally spread in West Africa.

HIV/AIDS epidemic According to the 2012 UNAIDS report on the global

3.1 Introduction

AIDS epidemic [151], 34 million people worldwide were living with HIV at the end of 2011, most of them (69%) in sub-saharan Africa.
The worldwide estimated HIV prevalence is illustrated in Figure 3.1. Regrettably,

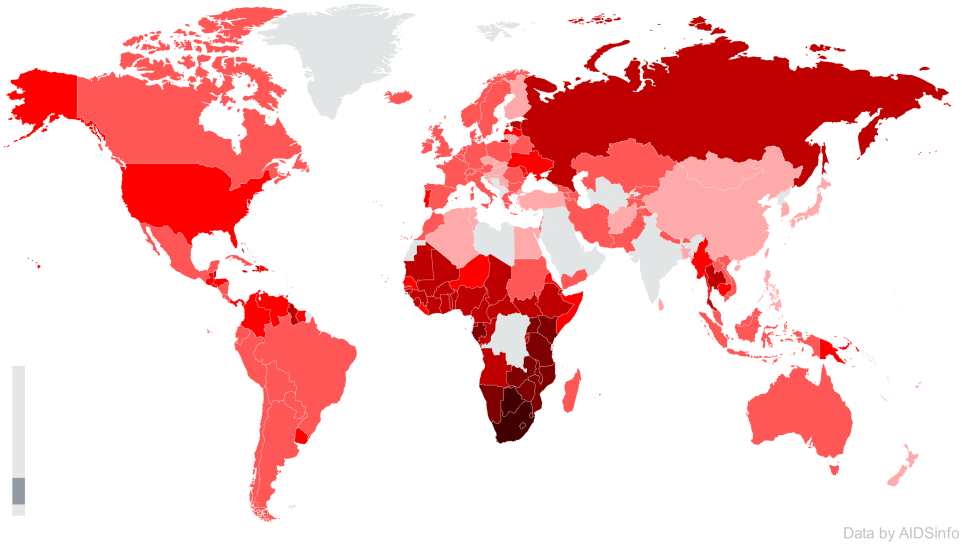

Figure 3.1: HIV prevalence - Estimated HIV prevalence by country of 2011, according to the 2012 UNAIDS report [151]. Source: UNAIDS [152]

the prevalence in southern Africa, e.g. in countries like Swaziland, Botswana, Lesotho or South Africa, reaches values of up to 26%, corresponding to every fourth inhabitant of that countries.
Fortunately, a decline of 20% of the worldwide number of new HIV infections is observed in 2011 compared to 2001, resulting in 2.5 million newly infected people. The decline is mainly based on rising HIV prevention efforts [151].
The number of human deaths from AIDS-related causes (1.7 million) also declined in 2011 compared to 2005 [151].

HIV morphology HIV virions are spherical in shape and have a diameter of around 120 Nanometer (nm) (a schematic morphology of HIV is given in Figure 3.2) [14]. Its outer coat, the viral envelope, is composed of a lipid bilayer, which is extracted from the host cells in the budding step (step six in Figure 3.3) during HIV life cycle. Host cell proteins and envelope protein complexes, key

3.1 Introduction

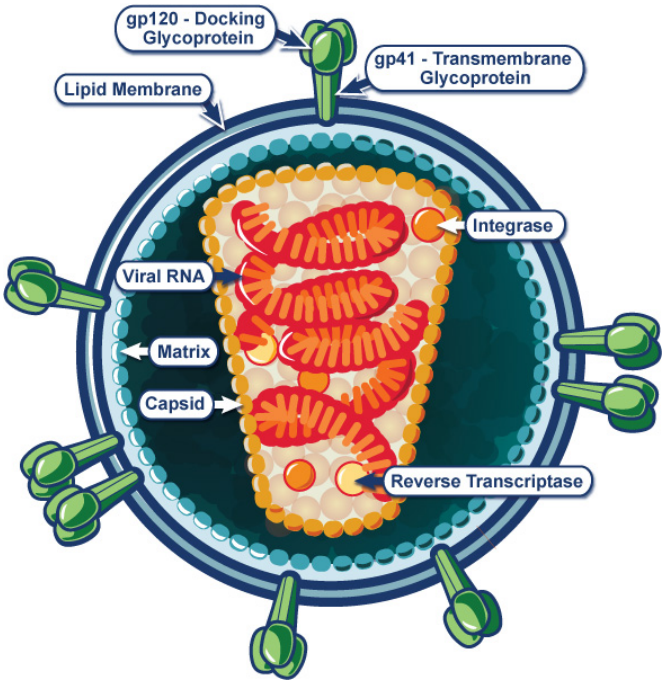

Figure 3.2: HIV virion - Schematic morphology of HIV virion. Source: National Institute of Allergy and Infectious Diseases (NIAID) [118]

players in the initial step of HIV replication cycle, are embedded in the bilayer. The bullet-shaped viral core (also referred to as capsid), embedded in the envelope, is built up by around 2000 capsid-proteins p24 and contains two copies of the viral genes, two single strands of Ribonucleic Acid (RNA), and viral enzymes reverse transcriptase, integrase and protease, which are essential for viral replication.
Between the viral envelope and core, units of matrix proteins (p17) are arranged to assist anchoring the previously mentioned envelope protein complexes.

HIV replication cyle HIV's main targets are Cluster of Differentiation 4 Receptor $(CD4)^+$ T-cells as well as macrophages [24]. The main processes during HIV attack on these cells are illustrated in Figure 3.3 and will be introduced in the following. The initial step, the interaction of the viral envelope protein complex with different host cell membrane proteins, is essential to this thesis and will be described more detailed in the next subsection.

3.1 Introduction

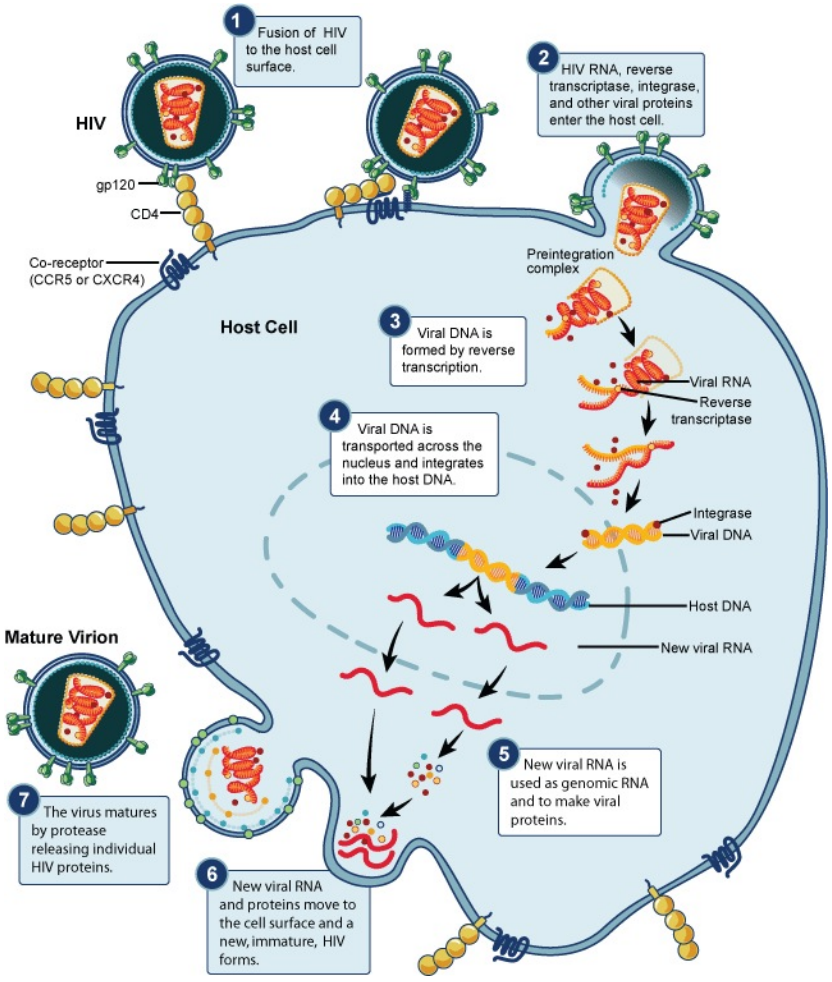

Figure 3.3: HIV replication cycle - Source: National Institute of Allergy and Infectious Diseases (NIAID) [117]

The second step is the fusion of HIV with the host cell membrane followed by the release of the viral capsid content, amongst others the virus RNA and the enzymes reverse transcriptase, integrase and protease, into the intra-cellular space. The next step is the transcription of the viral RNA into viral Deoxyribonucleic Acid (DNA) by the reverse transcriptase. The newly built viral DNA is then transported and integrated by the HIV integrase into the infected cell genome in the nucleus. The new viral RNA, a product of the transcription of the proviral DNA by the cellular machinery, is then used as genomic RNA to make viral proteins Tat and Rev, responsible amongst other things for expression of viral

precursor proteins Gag, Gag-Pol and Env. Subsequently, the new viral RNA and viral precursor-proteins diffuse to the cell membrane forming a new immature HIV virion. The last step of the HIV replication cycle is the cleavage of the pre-cursor proteins into their functional units by the HIV protease resulting in a mature infectious virus particle.

HIV-1 cell entry HIV-1, as well as influenza and syncytial respiratory viruses, enters host cells by applying type I membrane fusion machinery [26, 43]. The type I entry machine in HIV-1 is the membrane spanning Env spike, a trimer consisting of three copies of N-terminal components, responsible for interaction with host cell receptors, and C-terminal parts, positioned in the viral membrane and performing fusion of viral and host membranes. The trimer components are the non-covalently bound exterior Glycoprotein 120 (gp120) (N-terminal) and the transmembrane Glycoprotein 41 (gp41) (C-terminal), both generated by the host cell protease furin by cleaving the pre-cursor polypeptide Glycoprotein 160 (gp160), which is expressed during the HIV replication cycle (step five in Figure 3.3) [173].
The complete entry process can be split into three main parts, attachment, co-receptor binding and fusion, wherein the first two parts are schematically illustrated in Figure 3.4. The first step, the attachment, starts with the recognition and binding of gp120 by the host cell CD4. CD4 interacts with the Phe43 cavity on gp120, a conserved pocket formed by residues located in three different domains of gp120. These domains include helices in the inner domain, the CD4-binding loop in the outer domain and the β20-β21 unit, that becomes one half of the gp120 bridging sheet, formed after CD4 binding and important for co-receptor binding [89, 171]. After CD4 binding, viral and host cell membranes are brought via bending of the flexible region in CD4 into close proximity [134]. Furthermore, substantial conformational changes take place, including the formation of the bridging sheet, spatial approach of inner and outer domain and the detachment of the Variable loop 3 (V3), resulting in formation and exposure of the chemokine co-receptor binding site [20, 89, 131, 135, 168, 172]. The second main step in HIV-1 entry is the binding of gp120 to C-C Chemokine Receptor 5 (CCR5) or C-X-C Chemokine Receptor 4 (CXCR4) [35, 40, 50]. V3 is supposed to interact with the Extracelluar Loop 2 (ECL2), while the bridging sheet interacts with the

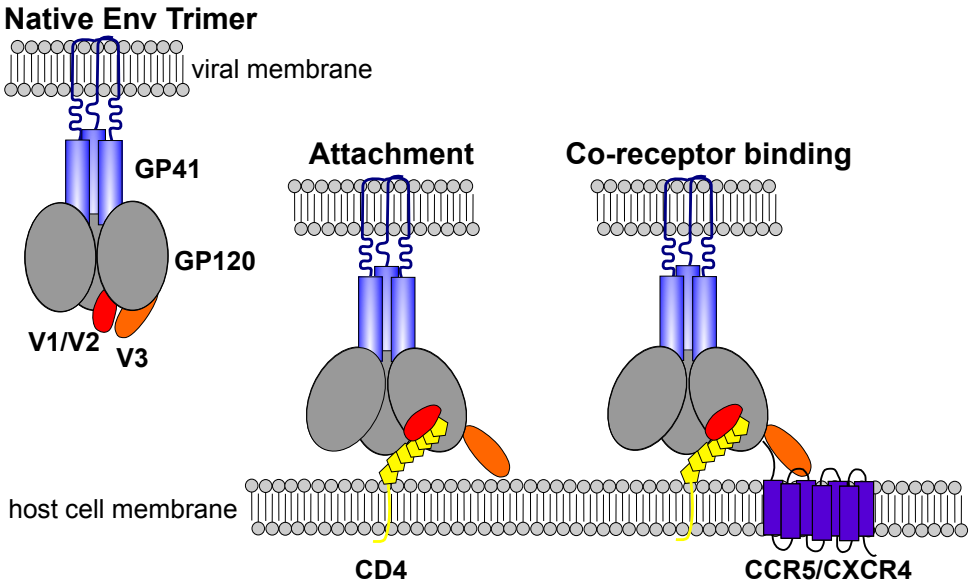

Figure 3.4: HIV cell entry - Schematic illustration of HIV-1 entry steps attachment and co-receptor binding.

N-terminal part of the co-receptor [13, 27, 48, 70, 71]. The co-receptor binding induces further changes in the Env trimer, leading to re-arrangments of the previously inaccessible gp41, that enables in the last step of HIV-1 entry the fusion of the viral and host cell lipid membranes [72].

HIV-1 structural data In order to fully understand the mechanisms and molecular basis of HIV entry, 3D structures of trimeric Env proteins in different conformations and in interaction with involved host cell receptors are needed at the best possible resolutions. Unfortunately, this has not been achieved yet, although an impressive development in terms of this goal is evident, starting with the crystallisation of T-cell CD4 by Wu et al. [168] in 1996. Two years later, the first structure of monomeric HIV-1 gp120, bound to CD4 and a monoclonal antibody 17b, was solved by Kwong et al. [89].
Despite the fact that only a deglycosylated core was crystallised, in total 60% of the polypeptide, gp120 crystal structure revealed the essential molecular sites

3.1 Introduction

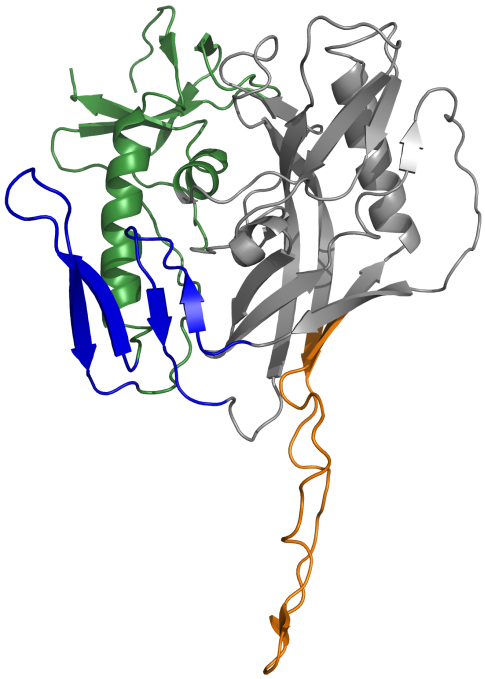

Figure 3.5: Structure of HIV-1 gp120 - HIV-1 gp120 is shown in cartoon illustration with inner domain in green and outer domain in grey. The outer domain structure elements bridging sheet and V3 are shown in blue and respectively orange. Structure taken from PDB entry 2B4C [71].

involved in CD4- and co-receptor binding, which then boosted the development of small molecule CD4 mimics [91, 102] and broadly neutralising antibodies, e.g. VRC01 [169]. Since the publication of Kwong et al. [89], several structures of monomeric gp120 have been crystallised in unliganded state as well as in complex with CD4, CD4 mimics or various antibodies [1, 21, 38, 39, 70, 71, 75, 87, 88, 116, 122, 124, 170, 178, 179]. Nevertheless, the core structure of all these proteins is identical with a packing into an inner and an outer domain, respectively, facing the inner or the outer side of the Env trimer. The most variable parts in gp120, with respect to structural as well as sequential aspects, are the surface-exposed variable loop regions V1-V5. Variable loops 1 and 2 (V1V2) have been crystallised only with a non-HIV scaffold [98, 110], precluding assumptions about intra gp120 domain orientations. In contrast, V3 was successfully solved in com-

plex with CD4 and an antibody by Huang et al. [71] (see Figure 3.5). However, it can vary depending on the conformation and interaction state of gp120. Variable loops 4 and 5 are also disordered in most crystal structures, which seems reasonable, since they are, first, highly variable in sequence and, second, located in the outer domain and thus exposed to immune system attacks. The first insights into trimeric gp120 were given by Liu et al. [99]. They solved the trimeric structure using cryo-electron tomography in unliganded, in complex with the antibody b12 and in complex with CD4 and the 17b antibody.

We superimposed three copies of gp120 structure including V3 and docked CD4 (Protein Data Bank (PDB) [8] ID: 2B4C) onto the gp120 core structures embedded in the trimeric docked coordinates provided by Tran et al. [150] (see Figure 3.6).

Since then, a number of low-resolution electron-microscopy structures of the Env complex were solved [69, 100, 150, 161, 169]. Among these publications, several identified interactions between V1V2 and V3, describing them as a mechanism of HIV to shield the co-receptor binding site (located at the stem of V3) from antibodies [64, 69, 100, 132]. However, there is still controversy whether it is an intra- or inter-gp120 interaction [100, 132].

Even a year before Kwong et al. [89] solved the structure of gp120, multiple groups obtained a six-helix bundle crystal structure of gp41 [17, 145, 160, 163]. Unfortunately, all the reported and following crystal- and Nuclear Magnetic Resonance (NMR)-structures are in the post-fusion conformation. High-resolution structures of gp41 in a critical pre-fusion state, which would give conclusions and hints about overall Env structure, are still unavailable. Recently, Tran et al. [150] determined a cryo-electron microscopy structure of trimeric Env including a gp41 intermediate, providing new insights into conformational changes during HIV entry.

The remaining host cell receptors involved in HIV entry, host cell CCR5 and CXCR4, are unfortunately insufficiently described, at least the complete receptor structure including the tyrosine-sulphated N-terminal regions, crucial for the interaction with gp120. Nevertheless, recent studies reported detailed interaction sites between gp120 and host cell receptors. As already indicated above, Huang

3.1 Introduction

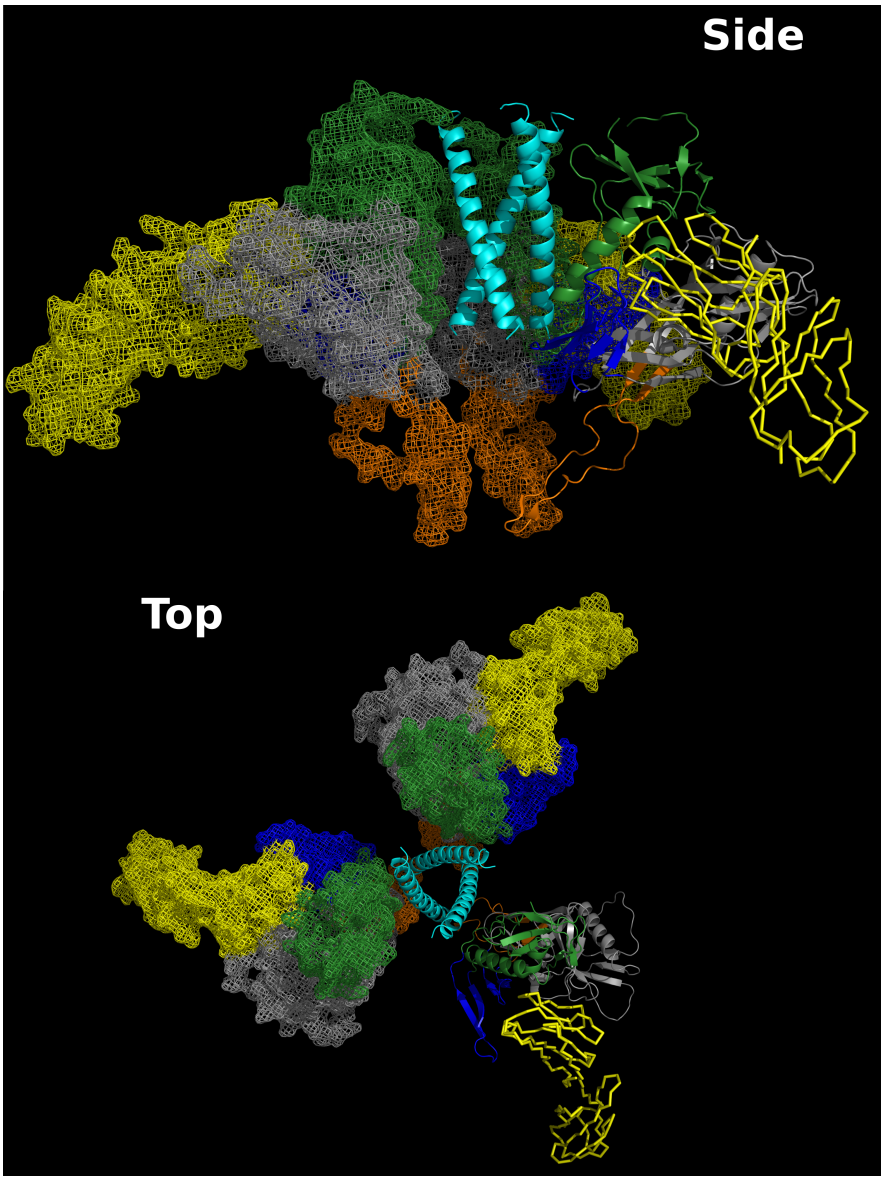

Figure 3.6: Env spike - Structure of Env heterotrimer in a CD4 bound conformation in side and top view. Two protein complexes of the heterotrimer are shown in mesh and one in cartoon illustration. gp120 is coloured in grey with structure elements bridging sheet, inner domain and V3 in blue, green and orange, CD4 in yellow (ribbon illustration) and N-terminal gp41 in cyan. PDB structures 2B4C [71] were superimposed on the trimeric docked coordinates provided by Tran et al. [150].

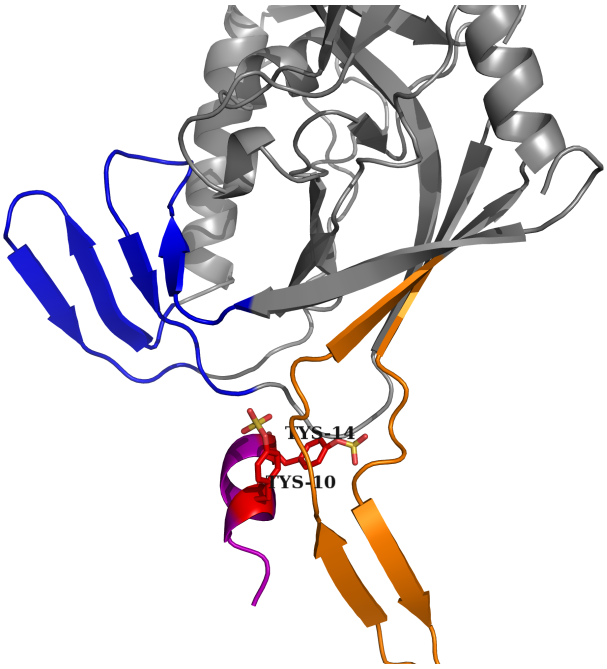

Figure 3.7: Structure of HIV-1 gp120 bound to CD4 and N-terminal CCR5 - HIV-1 gp120 (grey) with V3 (orange), bridging sheet (blue) and N-terminal CCR5 fragment (purple) with sulphated tyrosines as red sticks are shown in cartoon illustration. Docked coordinates provided by Huang et al. [70].

et al. [70] presented a gp120 crystal structure bound to CD4 and a tyrosine-sulphated antibody, which mimics the co-recepor binding. Furthermore, they solved a N-terminal CCR5 fragment (residues 7-15) by NMR and docked it to gp120 crystal structure (see Figure 3.7). Schnur et al. [137] carried out the same experiment by solving a NMR structure of a N-terminal CCR5 peptide. However, their CCR5 fragment was 26 amino acids long and had an oppositely directed orientation after docking in contrast to [70], suggesting other residues as interaction pairs (see Figure 3.8).

3.1.2 Co-evolution

A protein fold can remain almost unchanged during the course of evolution of homologous proteins, while the corresponding amino acid sequences mutate more frequently. Positions that mutate little or are conserved imply thereby func-

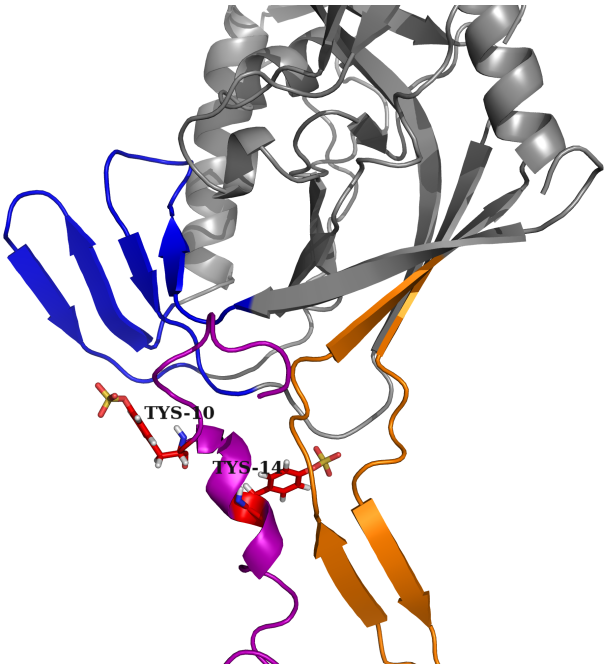

Figure 3.8: Structure of HIV-1 gp120 bound to CD4 and N-terminal CCR5 - HIV-1 gp120 (grey) with V3 (orange), bridging sheet (blue) and N-terminal CCR5 fragment (purple) with sulphated tyrosines as red sticks are shown in cartoon illustration. Docked coordinates provided by Schnur et al. [137].

tional and/or structural importance. However, non-conserved sequence positions may also be important with compensatory mutations in other variable positions [52, 175].

Due to the fact that co-evolution often takes place between residues that are proximal in the folded structure [52, 126, 175], identification of them would prove useful for ab initio structure, interdomain or dynamic prediction of proteins and protein domains [47, 57, 154, 155].

The first step for the identification is the arrangement of the sequences of interest in a Multiple Sequence Alignment (MSA), which helps identifying conserved and co-evolving positions. In Figure 3.9, a cut-out of an example MSA is shown, that includes conserved and co-evolving positions. The first column of the MSA consists exclusively of the nonpolar amino acid proline (P) and thus represents a conserved position. Columns seven and twelve illustrate an example for co-evolving positions. The negatively charged aspartic acid (D) (sequences 1-3) is

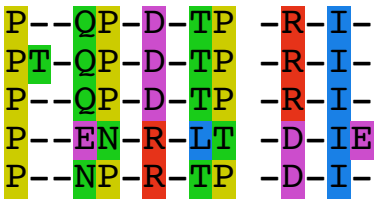

Figure 3.9: MSA - An cut-out of an example MSA illustrating conserved (column 1 and 14) and co-evolving positions (columns 7 and 12).

mutated in column seven into the positively charged arginine (R) (sequences 4-5). Consequently, the positively charged arginine (R) (seqeunces 1-3) is mutated in column twelve into the negatively charged aspartic acid (D) (sequences 4-5). The compensatory mutations are in this case a clear indicator for spatial proximity and ionic interaction.

Many different approaches have been applied to identify co-evolving positions including methods to detect the differences between observed versus expected frequencies of residue pairs (OMES) [78, 92], the McLachlan Based Substitution correlation (McBASC) [57, 120] and Mutual Information (MI) based methods [22, 30, 41, 86, 109, 158, 165]. However, the prediction quality suffers from random noise and phylogenetic relationship of the sequences within a MSA that induce additional indirect couplings [41, 159]. Recently, Direct Information (DI), based on Direct Coupling Analysis (DCA), which disentangles direct from indirect coupling, has been developed by Weigt et al. [159]. DI showed high accuracy in predicting real residue contacts on a variety of protein families [115] and has also been applied to protein ab initio folding with different fold classes [68, 108, 144]. We will focus in the following on the methods MI and DI and apply them together with a new improved sequence re-weighting strategy on the protein dataset provided by Morcos et al. [115]. In addition, we apply DI with the best performing re-weighting scheme to Env sequences in order to identify interacting residues within the Env trimer.

3.2 Materials and Methods

3.2.1 Materials

Bacterial and eukaryote protein sequences For comparing co-evolution methods, bacterial and eukaryote protein sequence alignments with minimum 1000 sequences were extracted from the Pfam database [129]. We retrieved MSAs of all protein families introduced in Morcos et al. [115], except those that have been withdrawn meanwhile, families without known atomic structure or families having too large MSAs for our computational resources. Pfam alignments are generated by applying Hidden Markov models, which can introduce insert states during the alignment of new sequences (lowercase letters). Insertions into the alignment are generally more typical in loop regions.
The MSAs were preprocessed in the same manner as in Morcos et al. [115], where, first, all lowercase amino acids are converted to gaps. Second, all positions that contain non-standard amino acids are also replaced by gaps.
Replacement of all lowercase amino acids by a gap means consequently a non-consideration of loop regions in the co-evolution prediction. Due to this fact, we additionally performed co-evolution prediction including all lowercase amino acids to obtain information about loop regions too.
Furthermore, we retrieved for each family all atomic structures listed on Pfam from the PDB [8]. Table A.3 lists all used protein families.

HIV-1 Env sequences A MSA containing pre-aligned HIV-1 Env sequences was taken from Los Alamos HIV Sequence Database[1]. Apart from the fact that only one sequence per patient is used, we did not consider any other constraints, e.g. HIV subtype or co-receptor usage. Sequences with non-standard amino acid have been deleted resulting in 4844 aligned HIV-1 Env sequences.

[1] http://www.hiv.lanl.gov/

3.2.2 Mutual information

MI is an information theoretical quantity that measures the mutual dependence of two random variables [3, 30]. It is based on Shannon's entropy [140], a measure of uncertainty in a random variable X, which is given by:

$$H(X) = -\sum_{x \in X} p(x) \log p(x). \tag{3.1}$$

The joint entropy is an extension of the entropy to two random variables X and Y with ordered pairs $p(x, y)$. It is defined as follows:

$$H(X, Y) = -\sum_{x \in X} \sum_{y \in Y} p(x, y) \log p(x, y). \tag{3.2}$$

Conditional entropy is in turn given by the difference between the joint and the Shannon entropy:

$$\begin{aligned} H(Y|X) &= \sum_{x \in X} \sum_{y \in Y} p(x, y) \log \frac{p(x)}{p(x, y)} \\ &= -\sum_{x \in X} \sum_{y \in Y} p(x, y) \log p(x, y) + \sum_{x \in X} \sum_{y \in Y} p(x, y) \log p(x) \\ &= H(X, Y) + \sum_{x \in X} p(x) \log p(x) \\ &= H(X, Y) - H(X). \end{aligned} \tag{3.3}$$

MI reduces the uncertainty of random variable X due to the knowledge of Y (and vice versa) and we may therefore specify

$$\begin{aligned} MI(X, Y) &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \\ &= H(X) + H(Y) - H(X, Y). \end{aligned} \tag{3.4}$$

Reorganisation of Equation 3.3 to

$$H(X) = H(X, Y) - H(Y|X) \tag{3.5}$$

and insertion of 3.5 in Equation 3.4 gives

$$\begin{aligned} MI(X, Y) &= H(X, Y) - H(Y|X) + H(X, Y) - H(X|Y) - H(X, Y) \\ &= H(X, Y) - H(X|Y) - H(Y|X). \end{aligned} \tag{3.6}$$

Introduction of joint and conditional entropy definitions (see Equations 3.2 and 3.3) into Equation 3.6 gives the widely used version of MI [30, 51, 115, 159]:

$$\begin{aligned} MI(X,Y) = & -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log p(x,y) \\ & + \sum_{x\in X}\sum_{y\in Y} p(x,y)\log p(x,y) - \sum_{x\in X}\sum_{y\in Y} p(x,y)\log p(y) \\ & + \sum_{x\in X}\sum_{y\in Y} p(x,y)\log p(x,y) - \sum_{x\in X}\sum_{y\in Y} p(x,y)\log p(x) \\ = & \sum_{x\in X}\sum_{y\in Y} p(x,y)\log p(x,y) - \sum_{x\in X}\sum_{y\in Y} p(x,y)\log p(y) \\ & - \sum_{x\in X}\sum_{y\in Y} p(x,y)\log p(x) \\ = & \sum_{x\in X}\sum_{y\in Y} p(x,y)\log \frac{p(x,y)}{p(x)p(y)}. \end{aligned} \tag{3.7}$$

MI can also be applied to MSAs with M as the number of protein sequences (number of MSA rows) and N denoting the protein length (number of MSA columns), by considering the amino acid frequencies $f(A)$ within alignment columns as observations of an alphabet A instead of the probabilities $p(x)$ of a random variable X. The entropy is thus obtained by using the single-site frequencies $f_i(A)$ of MSA columns i,

$$f_i(A) = \frac{1}{M}\sum_{a=1}^{M} \delta_{A,A_i^a} \tag{3.8}$$

with $1 \leq i \leq N$, $1 \leq A \leq q$ ($q = 21$ as the alphabet size, 20 standard amino-acids and gap) and δ denoting the Kronecker symbol equaling one if the two indices (amino acids) match, and zero otherwise. Accordingly, the joint entropy is defined as pair-site frequencies $f_{ij}(A,B)$ of MSA columns i and j,

$$f_{ij}(A,B) = \frac{1}{M}\sum_{a=1}^{M} \delta_{A,A_i^a}\delta_{B,A_j^a} \tag{3.9}$$

with $1 \leq i,j \leq N$, $1 \leq A,B \leq q$. Consequently, MI Equation 3.7 can be adjusted to

$$MI_{i,j} = \sum_{A,B} f_{ij}(A,B)\ln\frac{f_{ij}(A,B)}{f_i(A)f_j(B)}, \tag{3.10}$$

where MI equals zero if MSA columns i and j are uncorrelated and is positive otherwise.

3.2.3 Direct information

A major drawback of covariance analysis methods, like the introduced MI, is that they do not distinguish between correlations arising from direct or indirect interactions that result from substitution patterns of interacting residues. For instance, if residue i is coupled directly with residue j and j with k, then residue i and k are correlated indirectly (see Figure 3.10).
The indirect interaction may even be increased, if there are multiple weak cou-

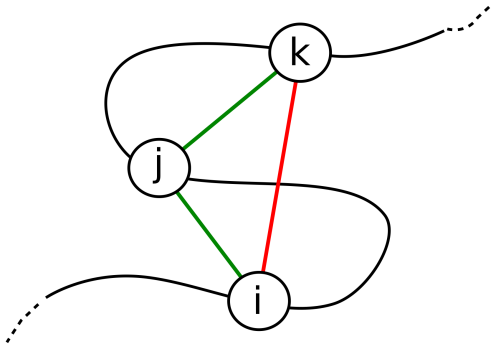

Figure 3.10: Direct and indirect interactions. - Direct interaction, in terms of spatial proximity, of residue i with residue j and j with k illustrated by green dashes. Red dash represents indirect interaction of residue i with residue k.

plings linking i and k, leading to a strong correlation without any direct interaction. Recently, Weigt et al. [159] introduced DCA, a method that disentangles direct from indirect effects. The methods and information presented in the following will however be based on the computationally more efficient implementation of DCA introduced by Morcos et al. [115].
In contrast to the local MI score, because it only considers one residue pair at a time, DCA infers a global statistical model $P(A_1, ..., A_N)$ for all amino acid sequences of the MSA. The model has to be consistent with the empirical data

present in the alignment, i.e. to generate the empirical single- and two-site frequency counts:

$$\begin{aligned} P_i(A_i) &= \sum_{\{A_k|k\neq i\}} P(A_1,...,A_N) = f_i(A_i) \\ P_{ij}(A_i,A_j) &= \sum_{\{A_k|k\neq i,j\}} P(A_1,...,A_N) = f_{ij}(A_i,A_j). \end{aligned} \tag{3.11}$$

The aim is to determine the most general model, i.e. least constrained model $P(A_1,...,A_N)$ and the estimation of the direct couplings $e_{ij}(A,B)$.
A detailed description of the derivation and estimation is given in the supplementary of [115].
Given now the pair couplings $e_{ij}(A,B)$, we can estimate the quantity DI introduced by Weigt et al. [159]. DI measures the MI due to the direct coupling by isolating position pair i, j and introduction of a two-side model

$$P_{ij}^{(dir)}(A,B) = \frac{1}{Z_{ij}} \exp\left\{e_{ij}(A,B) + \tilde{h}_i(A) + \tilde{h}_j(B)\right\}. \tag{3.12}$$

The new fields $\tilde{h}_{i/j}$ are determined by introducing the single-site frequencies as marginal distributions,

$$\begin{aligned} f_i(A) &= \sum_{B=1}^{q} P_{ij}^{(dir)}(A,B) \\ f_i(B) &= \sum_{A=1}^{q} P_{ij}^{(dir)}(A,B), \end{aligned} \tag{3.13}$$

and Z_{ij} follows by normalisation. Consequently, DI is MI associated to $P_{ij}^{(dir)}$:

$$DI_{ij} = \sum_{A,B=1}^{q} P_{ij}^{(dir)}(A,B) \ln \frac{P_{ij}^{(dir)}(A,B)}{f_i(A)f_j(B)}. \tag{3.14}$$

3.2.4 Re-weighting

Re-weighting Biological sequences within a MSA show strong sampling bias due to phylogenetic relationships, especially sequences representing a protein family or sequences originating from only one species. Therefore, Procaccini et al. [128]

introduced a re-weighting scheme that corrects for this sampling bias by weighting each sequence with a factor $\frac{1}{m^a}$, where m^a is the number of sequences with a sequence identity greater than a pre-defined identity threshold x

$$m^a = \sum_{b=1}^{M} c, \quad c = \begin{cases} 1, & seqid(A^a, A^b) \geq xN \\ 0 \end{cases} \tag{3.15}$$

with $0 \leq x \leq 1$. Sequences that have no similar sequences within the given threshold have weight one and sequences with similar sequences are down-weighted with $\frac{1}{m^a}$. The single- and pair-site frequencies are thus re-defined [115, 159],

$$f_i(A) = \frac{1}{\lambda + M_{eff}} \left(\frac{\lambda}{q} + \sum_{a=1}^{M} \frac{1}{m^a} \delta_{A,A_i^a} \right) \tag{3.16}$$

$$f_{ij}(A,B) = \frac{1}{\lambda + M_{eff}} \left(\frac{\lambda}{q^2} + \sum_{a=1}^{M} \frac{1}{m^a} \delta_{A,A_i^a} \delta_{B,A_j^a} \right), \tag{3.17}$$

with pseudo-count λ, to estimate probabilities from counts in biological sequence analysis [42] and $M_{eff} = \sum_{a=1}^{M} \frac{1}{m^a}$ estimating the effective number of independent sequences.

Improved re-weighting The above introduced re-weighting is based on the sequence identity of two sequences extracted from a MSA (see Figure 3.9). The identity (fraction of identical amino acids) is determined from sequences including all gaps (see Figure 3.11), which are treated like regular amino acids (implemented in the code introduced in Morcos et al. [115] and available at `http://evfold.org`). We suspected on the accurateness of this type of sequence identity calculation and extracted sequences from a MSA excluding position where both sequences have a gap (see Figure 3.12), since this is more similar to a pairwise sequence identity calculation. In the following, we will refer to MI and DI using the first option as $\mathbf{MI_{MSA}}$ and $\mathbf{DI_{MSA}}$ and MI and DI applying the second option as $\mathbf{MI_{PW}}$ and $\mathbf{DI_{PW}}$.

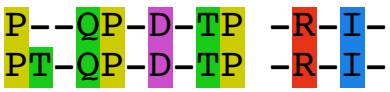

Figure 3.11: Re-weighting$_{MSA}$ - Re-weighting option considering all positions (including gaps).

Figure 3.12: Re-weighting$_{PW}$ - Re-weighting option excluding gapped positions (both sequences).

3.2.5 Homology modelling

Homology modelling, also referred to as comparative or knowledge-based modelling, is a useful technique to construct an atomic-resolution model of a protein with unknown structure (target) from its amino acid sequence and one or more experimentally derived 3D structures of homologous proteins (template(s)) [10, 133]. It makes use of the fact that structures of homologous proteins are more conserved than their corresponding sequences [23]. Figure 3.13 illustrates the main steps during the modelling of a target protein. The first step is the identification of one or more template structures of homologous proteins with a sequence identity of at least 25% - 30% to the target sequence. The second and most critical step is the alignment of the target sequence with the template structures. Subsequent to the alignment is the building of the target model by using the information given in the template structures. The last step is the evaluation of the produced model. In case of insufficient quality of the model, the modelling should be repeated from step two on. More detailed information about homology modelling steps, techniques and programs can be found in literature and in the web, e.g. in chapter 13 of [183].
We applied the homology modelling program MODELLER [133] to generate models of gp120 including V1V2, models of gp41 and models of the gp160.

3.3 Results

We shall first define the requirements necessary for a correct prediction (*contact* or True Positive (TP)). Residue pairs that are predicted by the co-evolution method to be in contact, should have a minimum sequence separation of five and a minimum atomic distance of less than eight Ångström (Å) in one of the

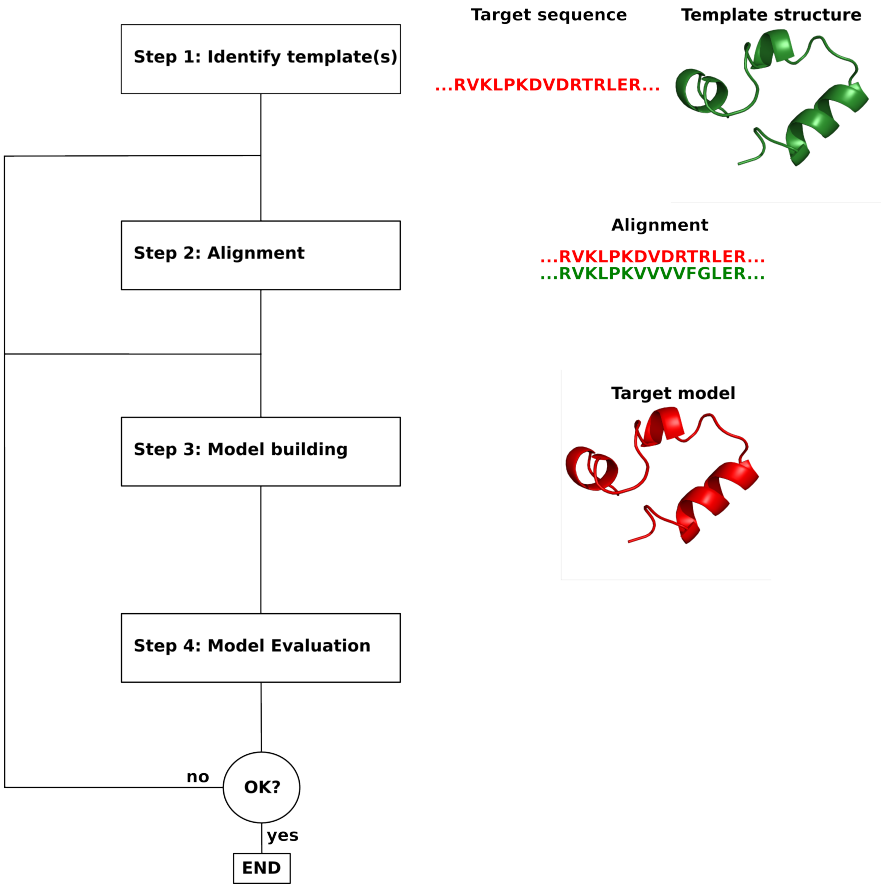

Figure 3.13: Homology modelling steps - A flow diagram illustrating the steps applied in modelling a target sequence based on a template structure. Figure taken from [45].

protein's structures in order to be a TP (as applied in Morcos et al. [115]). These requirements have then been used to test the prediction performance of DI and MI, applying both re-weighting strategies.

3.3.1 Contact prediction in bacterial and eukaryote protein families

Contact prediction excluding lowercase amino acids $\mathbf{DI_{PW}}$ and $\mathbf{DI_{MSA}}$ values have been computed for 124 bacterial protein families, preprocessed in the same manner as in Morcos et al. [115]. The mean TP rate for all families is shown

as a function of the number of predicted residue pairs in Figure 3.14.
The DI predictions are thereby carried out using five different sequence iden-

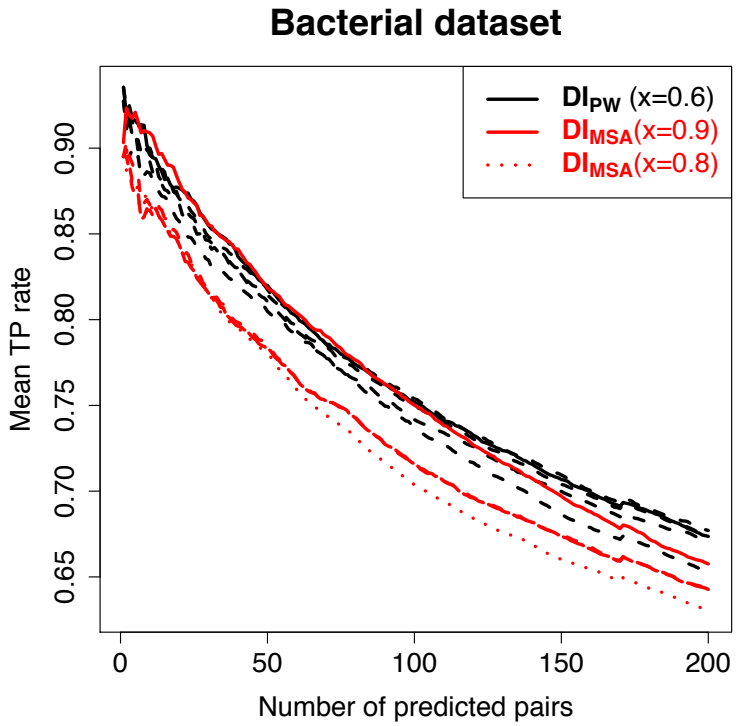

Figure 3.14: $\mathbf{DI_{PW}}$ vs. $\mathbf{DI_{MSA}}$ applied to bacterial protein families - Mean TP rate for 124 bacterial protein families as a function of the number of predicted pairs. The DI predictions were made using five different sequence identity thresholds $x = \{0.5, 0.6, 0.7, 0.8, 0.9\}$ for both, $\mathbf{DI_{PW}}$ and $\mathbf{DI_{MSA}}$. Black and red solid lines represent the sequence identity thresholds that perform best ($\mathbf{DI_{PW}}$: $x = 0.6$, $\mathbf{DI_{MSA}}$: $x = 0.9$). The other thresholds are shown as dashed lines. The dotted line represents the performance of $\mathbf{DI_{MSA}}$ using threshold $x = 0.8$, applied in Morcos et al. [115].

tity thresholds ($x = \{0.5, 0.6, 0.7, 0.8, 0.9\}$). The results of the best performing thresholds[1] are shown as solid black and red lines, for $\mathbf{DI_{PW}}$ and $\mathbf{DI_{MSA}}$ respectively. Dashed lines represent the remaining identity thresholds, except one dotted red line, which displays the performance of $\mathbf{DI_{MSA}}$ using the identity threshold

[1]We determined for each curve the sum of its mean TP rates in order to evaluate the performance between the sequence identity thresholds.

$x = 0.8$ (sequence identity threshold applied in Morcos et al. [115]). The solid lines, representing the best performing thresholds ($\mathbf{DI_{PW}}$: $x = 0.6$ and $\mathbf{DI_{MSA}}$: $x = 0.9$), show a nearly identical progression for the first 200 predicted pairs, while the dotted line is surprisingly the worst preforming one. Strikingly, $\mathbf{DI_{PW}}$ is less dependent on the choice of the sequence identity threshold than $\mathbf{DI_{MSA}}$. This observation is even more pronounced in the eukaryote dataset shown in

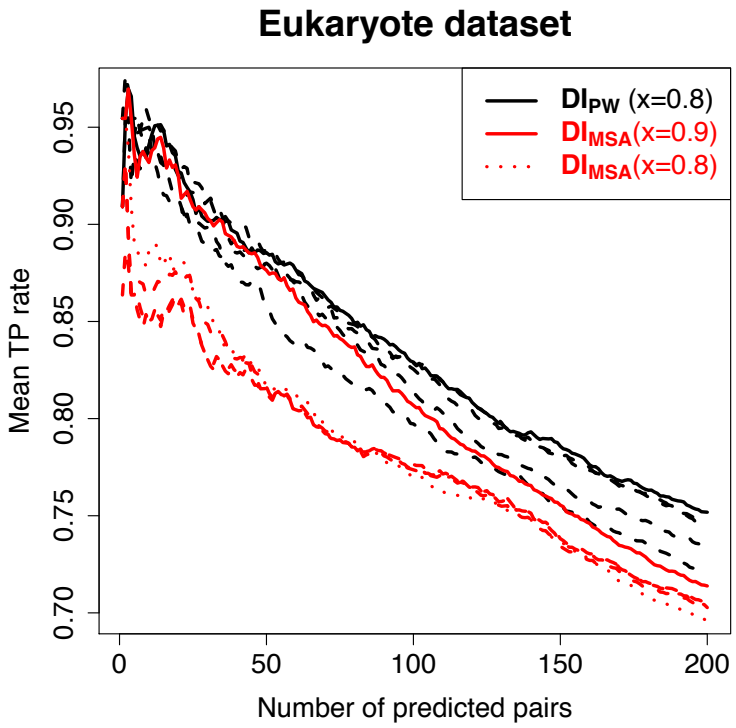

Figure 3.15: $\mathbf{DI_{PW}}$ vs. $\mathbf{DI_{MSA}}$ applied to eukaryote protein families - Mean TP rate for 22 eukaryote protein families as a function of the number of predicted pairs. The DI predictions were made using five different sequence identity thresholds $x = \{0.5, 0.6, 0.7, 0.8, 0.9\}$ for both, $\mathbf{DI_{PW}}$ and $\mathbf{DI_{MSA}}$. Black and red solid lines represent the sequence identity thresholds that perform best ($\mathbf{DI_{PW}}$: $x = 0.6$, $\mathbf{DI_{MSA}}$: $x = 0.9$). The other thresholds are shown as dashed lines. The dotted line represents the performance of $\mathbf{DI_{MSA}}$ using threshold $x = 0.8$, applied in Morcos et al. [115].

Figure 3.15, where all $\mathbf{DI_{MSA}}$ curves, except the curve for the best performing threshold $x = 0.9$, show a lower mean TP rate progression compared to the $\mathbf{DI_{PW}}$

ones. The best performing threshold for $\mathbf{DI_{PW}}$ is in this dataset different from the one in the bacterial one ($x = 0.8$), which reinforces the aforementioned observation that $\mathbf{DI_{PW}}$ is less dependent on the choice of x.
Furthermore, we compared the performance of DI and MI applied on the bacterial dataset, both with and without the re-weighting procedure. Figure 3.16 shows the comparison carried out by averaged Receiver Operating Characteristic (ROC) curves[1] (averaged over 124 protein families).

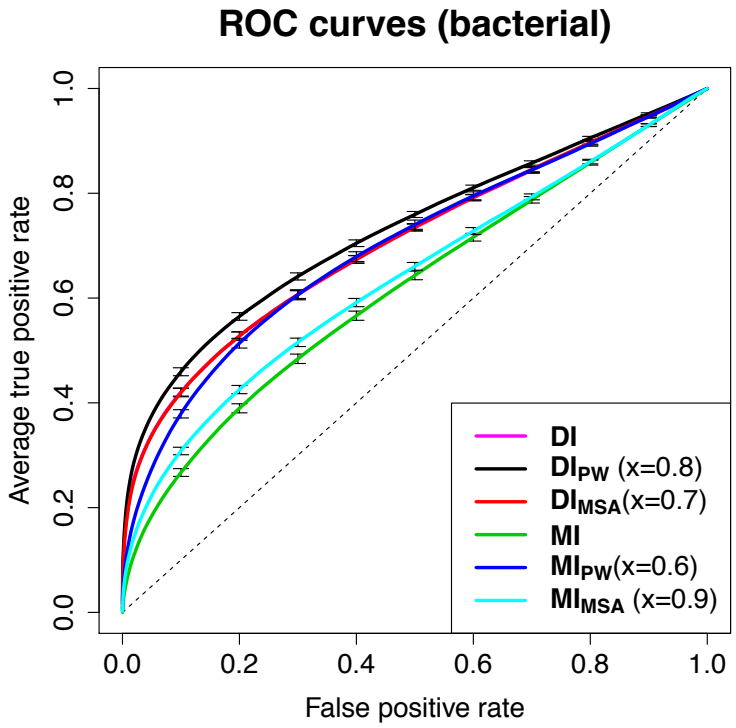

Figure 3.16: Averaged ROC curves for DI and MI prediction methods - The plot shows true positive (y-axis) and false positive (x-axis) rates of DI and MI with no re-weighting and DI and MI applying both re-weighting strategies, while only the best performing sequence thresholds, according to the AUCs, are shown. Strikingly, DI and $\mathbf{DI_{MSA}}$ have a nearly identical performance (magenta coloured line is under the red line), implicating that this type of re-weighting does not supply any prediction improvement.

[1]ROC curves plot the true positive rate as a function of the false positive rate for all predicted pairs and for cutoffs ranging from zero to one.

3.3 Results

DI_{PW} (black line) is the best performing prediction method with a mean Area

Table 3.1: Mean AUC values (bacterial protein families)

Method	Mean	Sd
DI	0.7056	0.0614
DI_{PW}	0.7295	0.0527
DI_{MSA}	0.7065	0.0644
MI	0.6191	0.0693
MI_{PW}	0.6983	0.0718
MI_{MSA}	0.6379	0.0621

Mean AUC values (with standard deviation) of DI and MI applying both re-weigting strategies. **DI_{PW}** outperforms all other methods significantly.

Under the Curve (AUC)[1] of 0.7295 of maximal one and a standard deviation of 0.0527 (see Table 3.1). Moreover, it is significantly[2] better than DI and **DI_{MSA}**, which have an almost identical performance with mean AUCs of 0.7056 and 0.7065 respectively. The averaged ROC curves are so similar that the magenta coloured one, representing DI, is hidden behind the red coloured **DI_{MSA}** curve. Astonishingly, **MI_{MSA}** performs just as well as DI and **DI_{MSA}** with a mean AUC of 0.6983. In a nutshell, MI benefits from both re-weighting strategies, while DI show only an increase when applied with the improved re-weighting strategy.
In summary, the re-weighting has been applied in order to correct for sampling bias by down-weighting the frequency counts depending on the number of similar sequences in the MSAs. The performance of DI is thereby enhanced by the presence of mutually diverse sequences, which tend to introduce many gaps into the MSA. Since **DI_{MSA}** considers also gapped positions for the sequence identity calculation, sequences seem to be more similar than they are in truth, which has

[1] A common application of ROC curves is the integration of the area under the ROC curve, the AUC [49], which is mainly used as measure for comparison of prediction performances. AUC values vary between zero and one, while a random classification results in 0.5, often illustrated as dashed bisecting line in ROC plots.

[2] T-test: p-value=0.0015, with given normal distribution and homogeneity of variance.

the effect of suppressing the information contained by down-weighting the counts. This effect is even more boosted by applying the preprocessing as introduced in Morcos et al. [115], which replaces lowercase amino acids by gaps. This leads to the fact that the choice of the sequence identity threshold is crucial for the performance of the method. $\mathbf{DI_{PW}}$ is in contrast to $\mathbf{DI_{MSA}}$ independent of the choice of the threshold, since it determines the actual identity of the sequences in the MSAs.

Contact prediction including lowercase amino acids We will show in the following the performances of $\mathbf{DI_{PW}}$ and $\mathbf{DI_{MSA}}$ applied on protein families MSAs including lowercase amino acids.
The mean TP rate progressions for theses cases are shown in Figures 3.17 (bacterial families) and 3.18 (eukaryote families) with DI predictions carried out using five different sequence identity thresholds ($x = \{0.5, 0.6, 0.7, 0.8, 0.9\}$).
Both the analysis of the bacterial and the eukaryote protein families yielded similar results compared to the previous subsection. We may therefore sum up as follows:

- The best performing $\mathbf{DI_{PW}}$ and $\mathbf{DI_{MSA}}$ predictions (top 200) have a nearly identical progression of the mean TP curves.
  (The sequence identity thresholds are not known a priori.)
- $\mathbf{DI_{PW}}$ is less dependent on the sequence identity threshold.
- $\mathbf{DI_{MSA}}$ using sequence identity threshold $x = 0.8$ (applied in Morcos et al. [115]) performs by far worst.

However, we must mention that the maximum TP rate for both the bacterial and the eukaryote analysis decreases compared to the previous predictions. The reasons lie above all in the inclusion of lowercase amino acids into the MSAs. The additional information is mainly from loop regions, which are typically highly variable in sequence and thus positioned in the more unaligned regions in the MSA. Furthermore, they are very often located at the solvent-accessible surface of protein folds and thus conformationally very flexible. Accordingly, either the contact predictions may be wrong per se, due to sparsely amino acid information

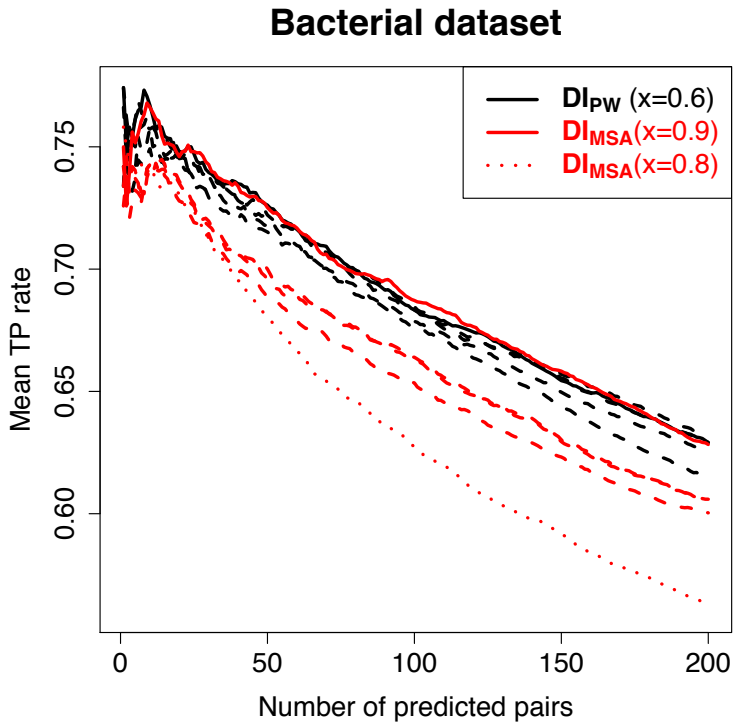

Figure 3.17: DI_{PW} vs. DI_{MSA} applied to bacterial protein families including lowercase amino acids - Mean TP rate (five sequence identity thresholds $x = \{0.5, 0.6, 0.7, 0.8, 0.9\}$) for 124 bacterial protein families, as a function of the number of predicted pairs. Black and red solid lines represent the sequence identity thresholds that perform best ($\mathbf{DI_{PW}}$: $x = 0.6$, $\mathbf{DI_{MSA}}$: $x = 0.9$). The other thresholds are shown as dashed lines. The dotted line represents the performance of $\mathbf{DI_{MSA}}$ using threshold $x = 0.8$, applied in Morcos et al. [115].

in these regions of MSAs, or the validation of the true predictions is wrong due to insufficient information about possible protein fold conformations in the loop regions.

Finally, we want to demonstrate the impact of the two re-weighting strategies with the help of the WD40 repeat domain (Pfam ID: PF00400) as an example. WD40 is a structural motif found in all eukaryotes that carries out functions like signal transduction and transcription regulation to cell cycle control, autophagy and apoptosis [97, 142]. A property of the WD40 family are the diverse sequences resulting in a low average sequence identity of 23% leading to the inclusion of

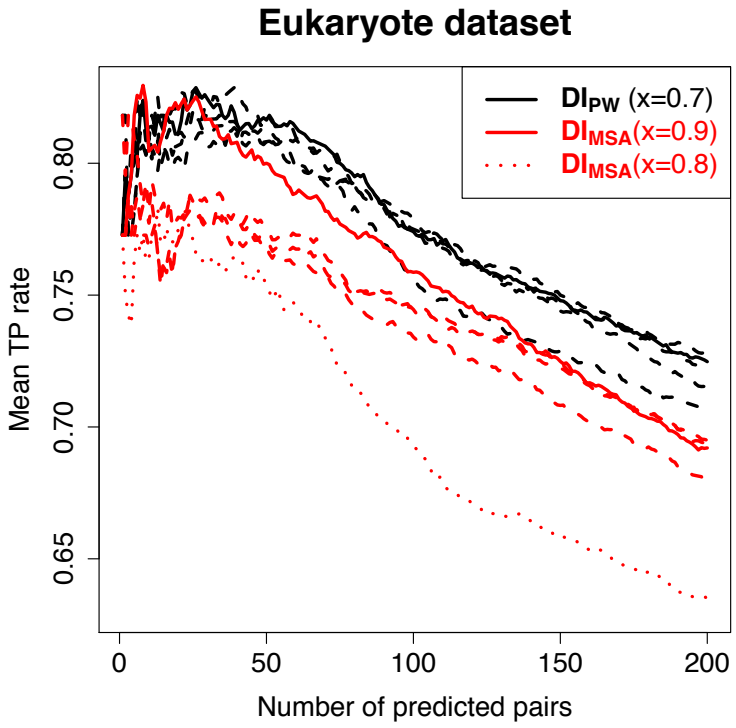

Figure 3.18: $\mathbf{DI_{PW}}$ vs. $\mathbf{DI_{MSA}}$ applied to eukaryote protein families including lowercase amino acids - Mean TP rate (five sequence identity thresholds $x = \{0.5, 0.6, 0.7, 0.8, 0.9\}$) for 22 eukaryote protein families, as a function of the number of predicted pairs. Black and red solid lines represent the sequence identity thresholds that perform best ($\mathbf{DI_{PW}}$: $x = 0.7$, $\mathbf{DI_{MSA}}$: $x = 0.9$). The other thresholds are shown as dashed lines. The dotted line represents the performance of $\mathbf{DI_{MSA}}$ using threshold $x = 0.8$, applied in Morcos et al. [115].

many gaps in the protein families MSA. The residue pairs with the 20 highest DI values and a minimum sequence separation of five are mapped and connected by coloured lines on the WD40 crystal structure (PDB ID: 1YFQ) (see Figure 3.19). Residue pairs with a minimum atomic distance of less than eight Å are TP and shown in green solid lines, whereas dashed orange lines denote TP found in another WD40 protein structure. Red lines highlight residue pairs not in contact. All pairs predicted by $\mathbf{DI_{PW}}$, shown in Figure 3.19 (A), are in contact, indicated by green or orange colouring. In contrast, several of the pairs predicted by $\mathbf{DI_{MSA}}$ are not in contact illustrated by red lines in Figure 3.19 (B). These results clearly

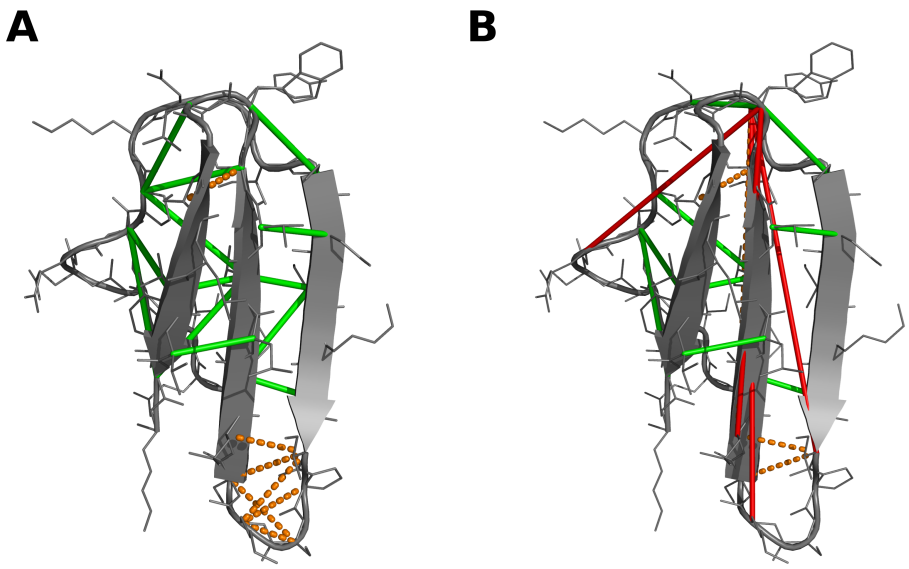

Figure 3.19: DI_{PW} vs. DI_{MSA} applied to the WD40 repeat family - Top 20 predicted contacts of (**A**) $\mathbf{DI_{PW}}$ and (**B**) DI_{MSA} applied to the WD40 repeat family (Pfam ID: PF00400) and mapped on the crystal structure (PDB ID: 1YFQ). Green lines connect residues, crystallised in this structure, with a minimum atomic distance less than eight Å, orange lines represent contacts found in another WD40 repeat domain structure and red lines pairs not in contact.

show that $\mathbf{DI_{PW}}$ outperforms DI_{MSA} on this protein family.

3.3.2 Contact prediction in HIV-1 Env

In the following, we will apply $\mathbf{DI_{PW}}$ to analyse co-evolution within Env. We have previously noted that the choice of the sequence identity threshold has not a major impact with respect to the performance. Nevertheless, we have considered examining the sequence identity distribution within our 4844 HIV-1 sequences, since they originate only from one organism. As expected, the sequences are very similar (see Figure 3.20) with seven out of ten having a sequence identity between 70% and 80%, in contrast to the above applied bacterial and eukaryote datasets, where sequences are more different to each other. Figure A.4 and A.5 show box-plots of the averaged (over 124 bacterial and 22 eukaryote) sequence identities. It is readily apparent that the sequence identity peaks are between 0.1 and 0.3.

The choice of a small sequence threshold (e.g. $x = 0.5$) would in this case cause the down-weighting of all sequences, whereas a big threshold (e.g. $x > 0.9$) results in not weighting any sequence. We choose the threshold $x = 0.8$, since it represents a good balance between no weighting and weighting of all sequences. Marks et al. [108] concluded from their analysis that a minimum of 0.5 to 0.75

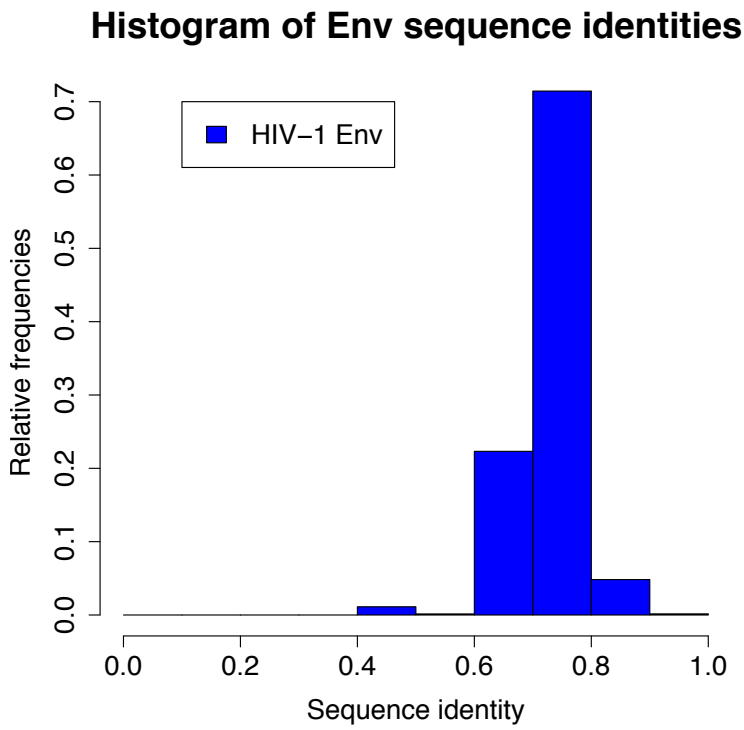

Figure 3.20: Histogram of Env sequence identities - Relative frequencies of HIV-1 Env sequence identities.

predicted constraints per residue is needed to make a reasonable 3D structure prediction. However, they mentioned that this number can vary depending on factors like type of fold and False Positive (FP) rate. We decided to consider a rather conservative number of prediction pairs when analysing Env sequences namely 200, which is around 0.2 of the number of residues in the reference sequence.

Contact prediction in gp120 In Figure 3.21 we utilise the gp120 crystal struc-

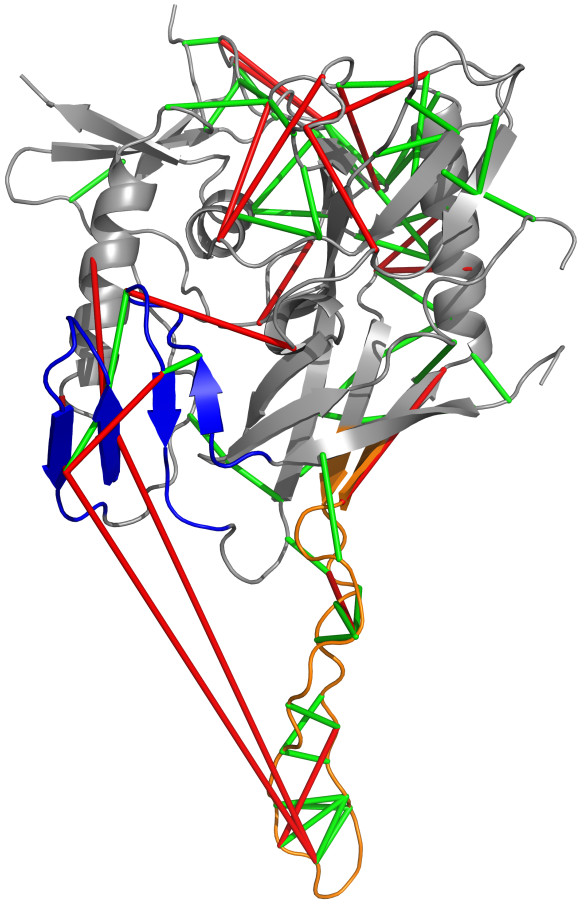

Figure 3.21: DI_{PW} predictions in gp120 - Top 200 predictions for HIV-1 Env. Only contact predictions of crystallised residues are indicated by red and green lines on the crystal structure solved by Huang et al. [70] (PDB ID: 2QAD).

ture, solved by Huang et al. [70], in order to map the contacts predicted by $\mathbf{DI_{PW}}$. Since there is no complete structure of Env (see introduction), we highlight only contacts (as lines), whose residues are crystallised (74 out of 200 residue pairs). Almost 75% of the predictions are TPs, represented by green lines. We even get close to 90%, if we soften our distance threshold from eight Å to ten Å. This may make sense in some cases, since errors, e.g. originating from wrong placement of side chains, may occur during X-ray structure analysis or the protein of interest is very flexible and changes its conformation during lifetime, e.g. Env during host cell entry. Nevertheless, there are some red coloured medium length

dashes (length between 10 Å - 20 Å) such as those slightly above and to the right of the bridging sheet (blue). These three predicted pairs surround the so called Phe43 cavity, which is essential for CD4 binding and may thus be functionally important, for instance to preserve the electrostatics in this region or the cavity per se. Furthermore, there are two medium length dashes ranging from the bridging sheet to the inner domain and two far-reaching dashes. All four predictions are potential TPs, for instance in the unliganded conformation of the Env spike, where it is first, assumed that V1V2 and V3 are in close proximity to shield the co-receptor binding site and second, the bridging sheet is still unformed and thus spatially closer to the inner domain.

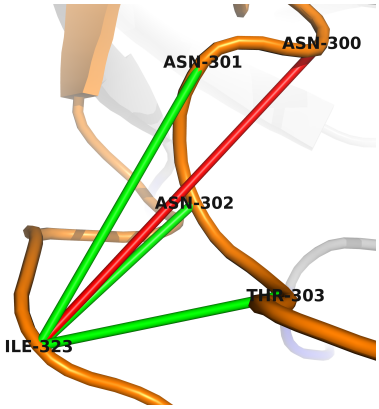

Figure 3.22: DI_{PW} predictions located at the V3 stem

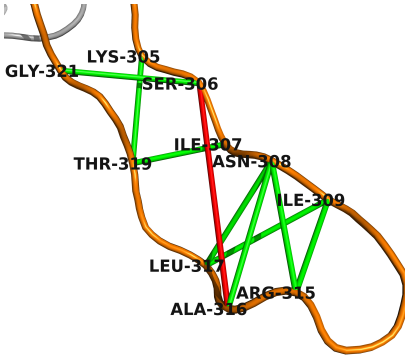

Figure 3.23: DI_{PW} predictions located at the V3 crown

Contact prediction in V3 Figure 3.22 depicts potential interactions of residues located at the stem of V3. Surprisingly, all illustrated residues, except residue Asn 301, are substantial interaction partners of one of the host cells sulphated tyrosines, located at the N-terminal end of one of the co-receptors CCR5 or CXCR4. Position 301 (residue Asn 301) is also an important location, since it is a potential glycosylation site according to the Los Alamos HIV Sequence Database[1]. This region is a perfect example for a functionally motivated direct coupling as the participating residues need to be in close proximity to fulfil the binding between gp120 and the host cell co-receptor.

[1] http://www.hiv.lanl.gov/

Figure 3.23 also shows putative interaction sites between residues within V3, but located more closer to the crown. The crown of V3 is next to the above indicated stem the second essential region responsible for co-receptor binding. It interacts with residues located in the extracellular loops of CCR5 or CXCR4, most likely with residues in the extracellular loop 2 [27, 121, 125, 148]. Moreover, the crown of V3 is supposed to be the one of the main determinants of co-receptor usage [27]. These facts suggest the presence of direct compensatory mutations within this region in order to conserve functional and structural features.
Furthermore, in Figure 3.24 we illustrate potential interactions between residues

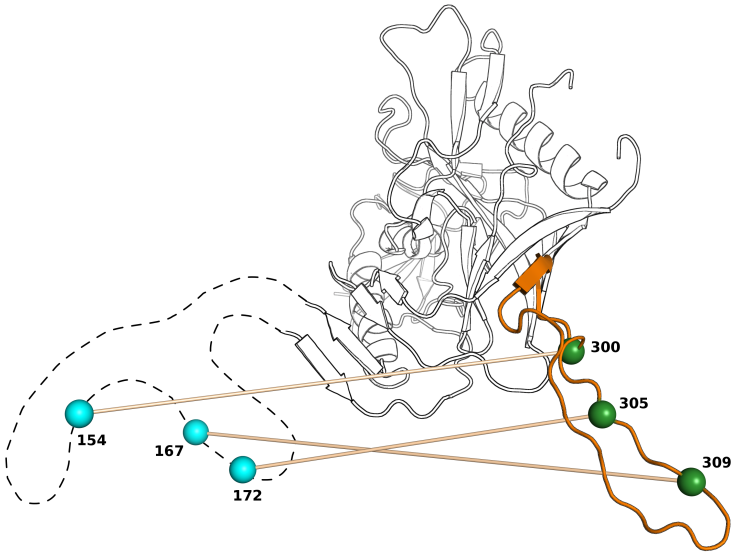

Figure 3.24: Potential interactions between V1V2 and V3 - Structure of gp120 [70] (PDB ID: 2QAD) is shown in cartoon illustration with V3 coloured in orange and V1V2 depicted as dashed line, since it is not crystallised. $\mathbf{DI_{PW}}$ prediction pairs are shown as cyan and green spheres (V1V2 and V3 respectively) and connected by bars.

located in V1V2 and V3. Due to the fact that there is no atomic structure including V1V2 available, we depicted V1V2 as dashed lines. These contact predictions support the above mentioned presumptions that V1V2 and V3 interact with each other in order to shield the co-receptor binding site. Nevertheless, from our $\mathbf{DI_{PW}}$ analysis we cannot distinguish between a monomeric (intra-gp120) [100]

or a neighbouring interaction (inter-gp120) [132].
Moreover, $\mathbf{DI_{PW}}$ predicted V1V2 - V3 interactions helped interpreting signature patterns that distinguishes between Chinese and Non-Chinese HIV-1 genomes [156, 157].

Contact prediction in V1V2 Due to absence of an atomic structure of gp120 including V1V2 domain, we cannot evaluate the performance of intra-V1V2 interactions predicted by $\mathbf{DI_{PW}}$, which are among the top-ranked ones, as can be seen in Table A.5. Nonetheless, the predictions give hints about the conformation of the V1V2 domain and may be applied in structure prediction methods. We will utilise the predicted residue pairs to generate a homology model of gp120 including V1V2 in the next subsection.

Contact prediction between gp120 and gp41 13 out of the top 200 predicted $\mathbf{DI_{PW}}$ pairs are potential interactions between the non-covalently bound proteins gp120 and gp41 (shown in Table 3.2).
We have labelled gp120 positions that have been annotated in literature as gp41 binding [122] with a † and gp120 positions close (maximum sequence distance of two) to annotated ones with a $*$.
Gp120 inner domain one position 114 and signal peptide positions ten, 23, 24 and 25 have not been mentioned in literature previously and may thus be very interesting for future structure prediction analysis of the complete Env trimer. All potential gp41 binding positions are located either in the inner domain, N- or C-terminal gp120 or in the signal peptide and thus face the inner site of the Env trimer. Eight of the 13 predicted gp41 partners are mapped on the gp120 crystal structure solved by Pancera et al. [122] (PDB ID: 3JWD) (shown in Figure 3.25). These findings seem reasonable, since gp41 is located at the inner side of the glycoprotein complex (see Figure 3.6).

Contact prediction in gp41 Table A.6 lists 59 (out of the top 200) putative intra-gp41 interactions predicted by $\mathbf{DI_{PW}}$. An evaluation of the predictions is extremely difficult due to the insufficient structural data for gp41. In fact, the first obstacle is that only a part of gp41 (residues 531-581 and 624-681) is solved

Rank	i (HXB2)	i (domain)	j (HXB2)	j (domain)	$\mathbf{DI_{PW}}$
2	502*	Ct gp120	607	gp41	0.24
73	500†	Ct gp120	619	gp41	0.07
142	114	id1	553	gp41	0.05
149	46*	Nt gp120	632	gp41	0.05
152	25	sp	722	gp41	0.05
153	24	sp	722	gp41	0.05
158	10	sp	809	gp41	0.05
167	244*	id2	629	gp41	0.05
177	236*	id2	792	gp41	0.05
182	499†	Ct gp120	605	gp41	0.05
183	492*	id3	612	gp41	0.05
185	23	sp	853	gp41	0.05
193	92†	id1	633	gp41	0.05

Table 3.2: Potential interactions between the non-covalently bound proteins gp120 and gp41. Positions labelled with a † are annotated as gp41 binding in Pancera et al. [122]. gp120 positions with a $*$ are close to annotated ones (a maximum sequence distance of two). The domain definition are taken from Table S3 in Kwon et al. [87] and are as follows: Nt gp120 - N-terminal gp120, Ct gp120 - C-terminal gp120, sp - signal peptide, id1 - inner domain 1, id2 - inner domain 2 and id3 - inner domain 3.

as crystal or NMR structure (e.g. PDB ID: 2X7R) and only eight of the 59 predicted interactions are within the solved structural region. Second, all available gp41 structures are exclusively in post-fusion state. However, $\mathbf{DI_{PW}}$ contact predictions may be valid in other conformations, e.g. in the unliganded, CD4-bound or co-receptor-bound conformation of the Env spike.

3.3.3 Homology modelling of Env

In the following, we will generate comparative models of gp120, gp41 and a complex of both glycoproteins using the homology modelling program MODELLER. For that purpose, we will utilise the previously introduced $\mathbf{DI_{PW}}$ contact predictions as special distance restraints within the modelling process. The distance

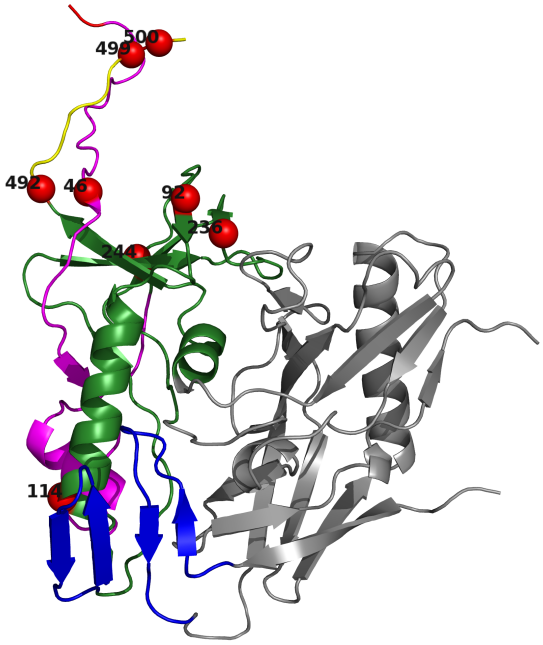

Figure 3.25: Potential gp41 interaction partners - Predicted gp120 positions (gp41 partners) are shown as spheres on the crystal structure solved by Pancera et al. [122] (PDB ID: 3JWD). Signal peptide positions are not shown, since they are unfortunately not crystallised. Inner domain, outer domain, bridging sheet, N-terminal gp120 and C-terminal gp120 are shown in green, grey, blue, magenta and yellow respectively.

constraint in MODELLER is harmonically restrained around a mean distance with an additional standard deviation. Since we do not know a priori which C_α-C_α distance is required, we examine the afore analysed bacterial and eukaryote dataset.

Figure 3.26 illustrates a histogram showing all C_α-C_α distances of the top 200 predicted pairs in each protein family. The distribution exhibits the highest peak at 10.25 Å. Hence, we choose a mean distance of 10.25 Å. Moreover, we mirrored the frequencies left of the peak, which seems to be normal distributed, in order to approximate a normal distribution to obtain a standard deviation, which we can apply in MODELLER. The approximated gaussian is shown as red line in Figure 3.26 with a standard deviation of 2.92 Å.

Furthermore, we included additional information like disulphide bonds and sec-

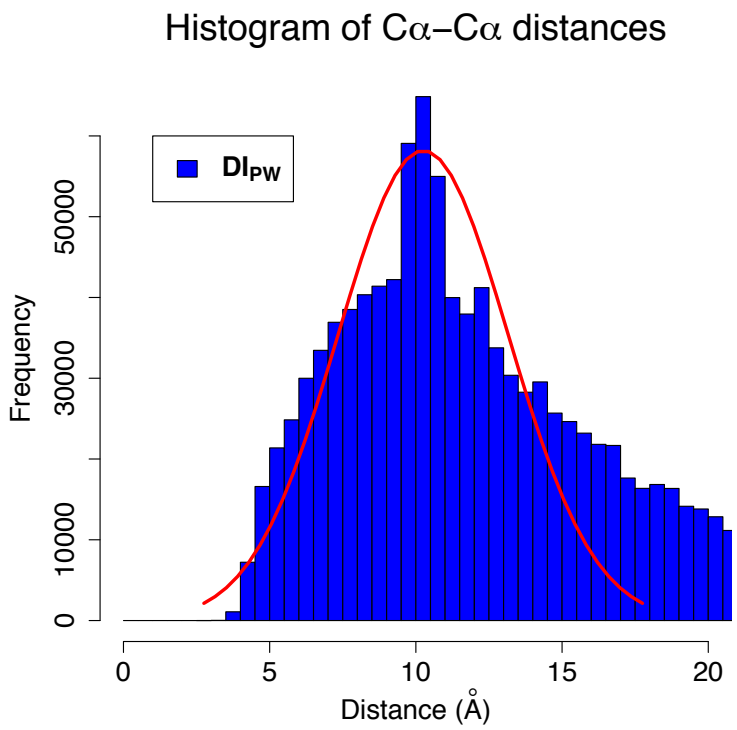

Figure 3.26: Histogram of C_α-C_α distances - Distances of the top 200 predicted pairs from all bacterial and eukaryote protein families.

ondary structure in the modelling process, obtained from Uniprot database[1] [147]. In addition, we applied an extended optimization routine, MD level refinement, with the parameter *very_slow*. We repeated the optimisation ten times.

Homology modelling of gp120 Up to date, there is no complete experimental solved structure of gp120 available, especially V1V2 and the N- respectively C-terminal parts are unavailable.
We identified and applied two template structures. First, the crystal structure solved by Pancera et al. [122] (PDB ID: 3JWD), which includes the N- and C-terminal regions of gp120, and second, the structure solved by Huang et al. [70] (PDB ID: 2QAD), which includes V3. Furthermore, we added 107 distance restraints to the modelling process, based on 107 intra-gp120 $\mathbf{DI_{PW}}$ contact

[1] http://www.uniprot.org/uniprot/P04578

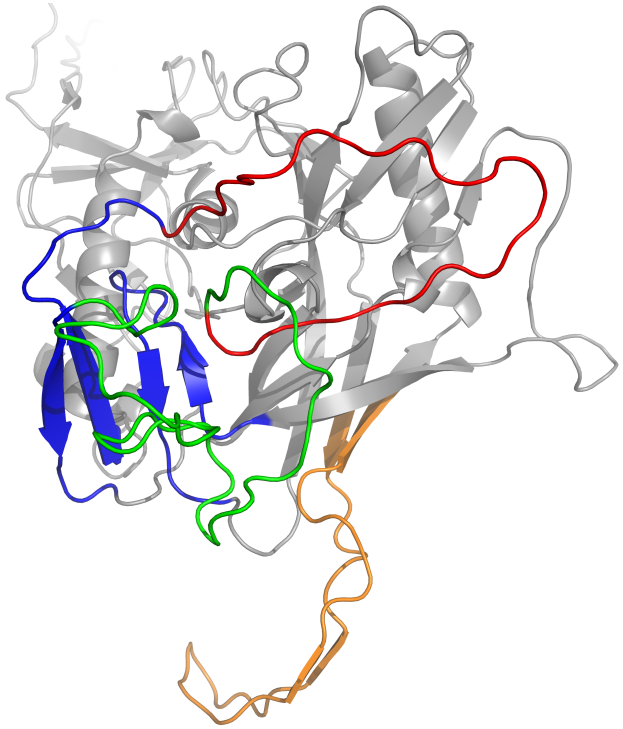

Figure 3.27: Homology model of HIV-1 gp120 including V1V2 - Gp120 is shown in cartoon illustration with V1V2 in red and green. V3 is coloured in orange and the bridging sheet in blue. The top 200 $\mathbf{DI_{PW}}$ contact predictions had been included in the model calculation and cause amongst other things a relocation of V3 close to the bridging sheet and V1V2.

predictions (out of the top 200). Figure 3.27 shows the resulting gp120 homology model including V1V2, which are coloured red and green respectively. Interestingly, according to this model, V3 is folded towards the bridging sheet (blue) and V1V2, which is supported by findings mentioned before in literature [64, 69, 100, 107, 132]. The interactions are supposed to shield the co-receptor binding site located at the stem of V3. V1V2 residues do not form any secondary structure in this model, which could well correspond with the native state, as it takes multiple conformations and locations during the host cell entry. The only structures including V1V2 [98, 110] are solved with a non-HIV scaffold and do not provide possible intra gp120 interactions, as does our comparative model.
However, we must mention that the applied contact predictions may originate

from several conformations during the entry, whereas the model presented here shows only one possible conformation.

Homology modelling of gp41 Due to the fact that all gp41 structures are exclusively solved in post-fusion state, we need to be careful in selecting the templates. We are interested in a more native structure model by applying $\mathbf{DI_{PW}}$ contacts, which can originate from all conformations/states that gp41 adapts. Figure 3.28 shows in (A) the gp41 X-ray structure solved by Weissenhorn et al.

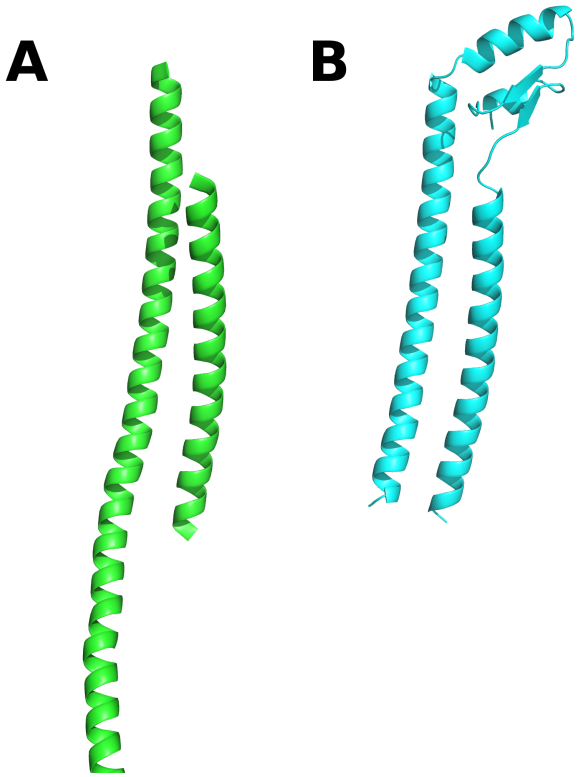

Figure 3.28: Gp41 X-ray structure and homology model - (A) X-ray structure of gp41 (N-terminal part) solved by Weissenhorn et al. [160] (PDB ID: 1ENV). (B) Homology model of N-terminal (residues 540-665) gp41 based on 1ENV template structure and taken from SWISS-MODEL [79] (Repository entry: P04578)

[160] (PDB ID: 1ENV). It is readily apparent that the connection of the two long helices, most probably a loop region, is not solved in the structure. The homology model, illustrated in (B) and obtained from SWISS-MODEL [79] (Repository

entry: P04578), predicts an ordered conformation composed of two helices and a parallel beta sheet. It used the gp41 X-ray structure (PDB ID: 1ENV) as template. However, Tran et al. [150] suggest that gp41 does not adopt the post-fusion conformation of two parallel helices in a pre-fusion intermediate.
Therefore, we mainly depend the protein length on the findings presented in Table 3.2.
If we exclude predicted contacts including the signal peptide (sp) and the con-

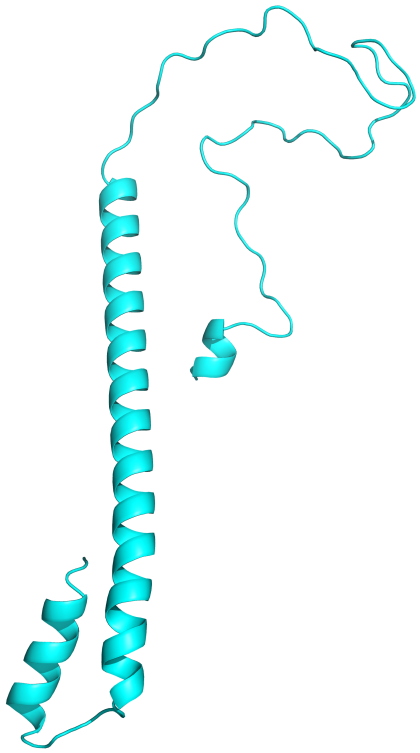

Figure 3.29: Homology model of HIV-1 gp41 - Gp41, residues 512-633 (HXB2 sequence), in cyan coloured cartoon illustration.

tact between the inner domain two (id2) and gp41 (Rank 177), all amino acids have a sequence position ≤ 633. Hence, we generated a model ranging from gp41 N-terminus (HXB2 position 513) up to position 633. As templates, we chose on the one hand the fusion peptide (HXB2 positions 513 - 535) structure solved by Gordon et al. [61] (PDB ID: 1ERF) and on the other hand the N-terminal helix (HXB2 positions 531 - 581) of the structure solved by Buzon et al. [16] (PDB ID:

2X7R). Subsequently, we extracted a subset of Table A.6 that contains predicted interactions of positions ≤ 633 (Table A.7), in order to apply these interacting positions as distance constraints in the modelling process. Furthermore, we applied additional information, a disulphide bridge between position 598 and 604 and secondary structure information, in the modelling process. The information were derived from the Uniprot database[1] [147]. The resulting homology model of gp41 is shown in Figure 3.29. The fusion peptide (small helix on the left side) is folded towards the N-terminal helix as indicated in Figure 4 of the publication by Pancera et al. [122]. According to the gp41 homology model, the fusion peptide folding exists only because of the presence of the interaction between position 518 and 553 (see Table A.7, Rank 93). The looped regions above the long helix are disordered, since these regions are not covered by a template structure or by additional secondary structure prediction.

Homology modelling of gp160 Until today, there is no single atomic structure of gp160 available. For the purpose of producing a comparative model, we generate a template consisting of the above modelled gp120 and gp41 by superimposing both onto the cryo-electron microscopy trimeric Env structure solved by Tran et al. [150] (see Figure 3.6). The gp160 comparative model structure was then solved by adding distance restraints, extracted from Table 3.2, to the modelling process. Figure 3.30 illustrates the gp160 model using the same color code as in Figures 3.27 and 3.29. As already mentioned by Pancera et al. [122], the gp120 termini and the 7-stranded β-sandwich, at the top of gp120, are in close proximity to gp41 and seem to maintain gp120-gp41 interaction (see red spheres in Figure 3.25). Furthermore, gp41 is located at the inner side of the Env trimer and possibly interacts with inner domain residues too. We have to mention nonetheless, that we only consider co-evolving positions as distance constraints during gp160 modelling. Conserved residues in gp120 and gp41 may interact with each other too. However, $\mathbf{DI_{PW}}$ does not detect conserved interacting positions.

.

[1] http://www.uniprot.org/uniprot/P04578

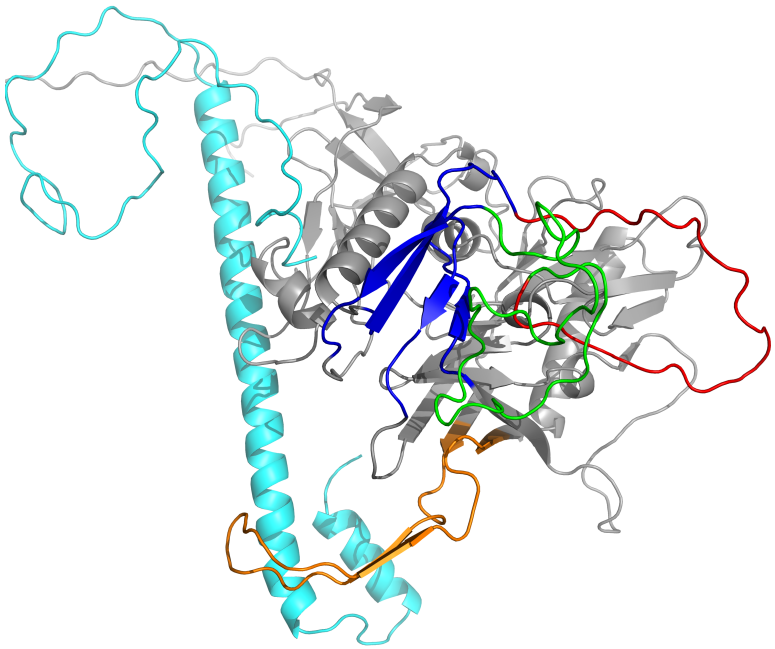

Figure 3.30: Homology model of HIV-1 gp160 - Model has been generated by applying inter gp120-gp41 $\mathbf{DI_{PW}}$ contact predictions as distance restraints. Colour scheme is according to the previous Figures 3.27 and 3.29.

3.4 Conclusion

We have studied the impact of different re-weighting strategies upon the performance of DI and MI, applying the protein family test set introduced in Morcos et al. [115]. In general, we conclude from our analysis that re-weighting has a positive effect with respect to the performance, when applied to both DI and MI. Concerning DI, the new developed re-weighting strategy performs significantly better in predicting real residue contacts than the published one. Furthermore, $\mathbf{DI_{PW}}$ is in contrast to $\mathbf{DI_{MSA}}$ less dependent on the sequence similarity threshold, which has to be chosen a priori.
In addition, we applied the superior method $\mathbf{DI_{PW}}$ in order to identify co-evolving positions close in proximity in HIV-1 gp120 and gp41. Interestingly, the method

predicted a lot of positions already indicated in literature. Nevertheless, due to the lack of atomic structures of different conformations of gp120 and gp41 during the host cell entry, we are not able to validate all predicted positions. These predicted pairs can therefore be at worst FP, TP in other conformations or TP in ligand-mediated conformations (e.g. CD4, co-receptor or glycan shield). However, the predicted pairs are worth an experimental validation, since they may help understanding the molecular basis of the key proteins involved in HIV entry. Beyond that, we applied $\mathbf{DI_{PW}}$ predicted pairs as distance constraints to generate homology models of gp120, gp41 and gp160. We have thereby achieved reasonable models, which provides an insight into the conformations of the involved glycoproteins. A further step would be the inclusion of additional methods in order to validate the produced models and gain more native-like comparative models (e.g. cryo-electron microscopy, Fluorescence Resonance Energy Transfer (FRET) experiments or MD simulations).

4 Summary and Outlook

> "*Prediction is very difficult, especially if it's about the future.*"
>
> Niels Bohr

In this work, we implemented GA and SMS-EMOA codes in Python[1], which are capable of solving different kinds of MOOP. The codes are very flexible and adaptive to the particular needs of the optimisation problems of interest. We used the SMS-EMOA program to test for optimally performing evolutionary parameters in the MOOP protein design. In order to apply several large-scale analyses we utilised simplified protein lattice models to search through sequence and structure space. The results provided new insights in evolutionary parameter selection, especially when applying the hypervolume guided SMS-EMOA to design new protein sequences and structures fulfilling desired tasks. The choice of the right combination of population sizes, mutation rates and crossover operators can save a lot of computation time and yield more promising protein sequence and structure candidates (pareto optimal individuals) at the same time. A further extension of the carried out experiments could on the one hand include an even simpler protein lattice model, for instance the HP model, to enumerate more evolutionary parameters, e.g. various crossover operators, population sizes and evaluation steps. On the other hand one could use more complex and realistic protein lattice models that include solvent-protein contacts. The lack of interaction with the surrounding solvent resulted in our case in the introduction of unrealistic sequence compositions, due to the fact that the applied energy function prefers contacts between charged residues, e.g. some amino acid sequences are built up exclusively

[1] http://www.python.org

by charged residues. A combination of both, a simpler energy function (e.g. HP or HPNX model) and the introduction of solvent interaction may lead to faster and more realistic output.

Furthermore, we modified re-weighting strategies within co-evolution methods DI and MI in order to identify co-evolving positions that are in spatial proximity. The modification made it also possible to be independent of the sequence similarity threshold, which has to be a priori provided for re-weighting. DI already proved its usefulness in the identification of contact positions in several studies [115] and was beyond that applied in structure ab initio folding [68, 108, 144]. The current advances in sequencing methods with a quickly growing number of new sequences make DI even more a promising effective method for co-evolution analysis and structure prediction.
Moreover, we applied the improved DI in order to identify co-evolving contact positions within HIV-1 Env protein complex, which have always been difficult to be solved structurally, due to the flexibility and sequence variability of its variable loops. Our analyses confirms several experimental findings and theories, e.g. the interactions of V1V2 and V3 or V3 and the bridging sheet, which are supposed to be a defence mechanism of HIV in order to shield the co-receptor binding side from immune attacks. Furthermore, we identified interactions between proteins gp120 and gp41, which are also in good agreement with experimental findings, in particular mutagenesis experiments.
In addition, we designed via the help of predicted contacts and cryo-EM density maps a new homology model of HIV-1 protein complex. Previous models, especially gp41 models are only generated in the post-fusion state of HIV entry. However, we performed the modelling process independent of gp41 structures solved in this state. The proposed intra- and inter-domain interactions can be subject to experimental studies in order to confirm and solve HIV bimolecular structures.
Another possibility is the modelling of HIV in combination with one of the co-receptors CXCR4 [167] or CCR5 [146]. The new model could then be further tested by e.g. extensive MD simulations in order to gain knowledge of the binding during the host cell entry.

Another interesting experiment would be the application of SMS-EMOA using the identified evolutionary parameters in order to design peptides/proteins that interact with HIV at critical locations, e.g. the CD4 or co-receptor binding sites. With further development in technology, it may be possible in near future to perform computationally costly MD simulations of big protein complexes including viral and host cell membranes in moderate time frames.

A Appendix

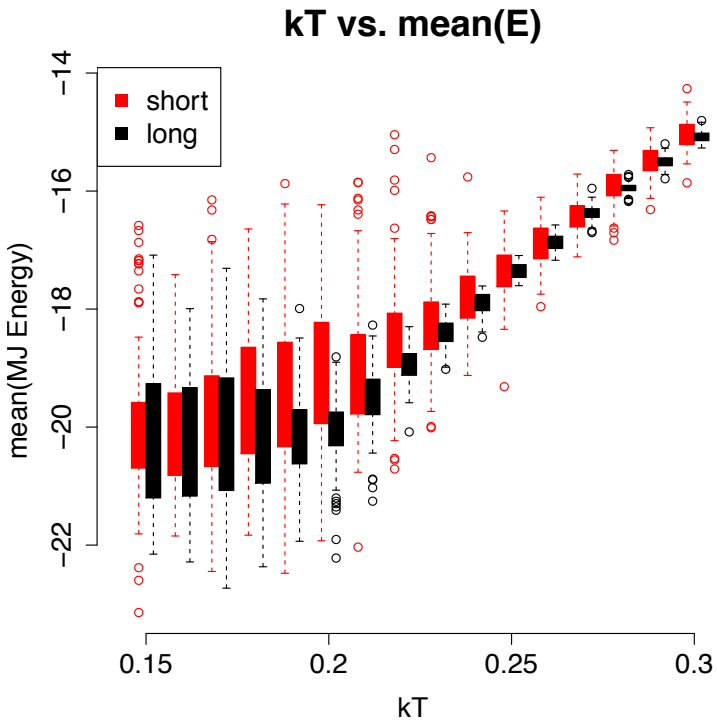

Figure A.1: Detailed boxplots of mean MJ energies - Short (red boxes) and **long** (black boxes) simulations mean MJ energies shown as boxplots for 100 MC seeds. Simulations were performed using kTs ranging from 0.15 to 0.3 (x-axis).

Table A.1: MJ-potential according to Miyazawa and Jernigan [113]

	C	M	F	I	L	V	W	Y	A	G	T	S	Q	N	E	D	H	R	K	P
C	-1.06	0.19	-0.23	0.16	-0.08	0.06	0.08	0.04	0.00	-0.08	0.19	-0.02	0.05	0.13	0.69	0.03	-0.19	0.24	0.71	0.00
M	0.19	0.04	-0.42	-0.28	-0.20	-0.14	-0.67	-0.13	0.25	0.19	0.19	0.14	0.46	0.08	0.44	0.65	0.99	0.31	0.00	-0.34
F	-0.23	-0.42	-0.44	-0.19	-0.30	-0.22	-0.16	0.00	0.03	0.38	0.31	0.29	0.49	0.18	0.27	0.39	-0.16	0.41	0.44	0.20
I	0.16	-0.28	-0.19	-0.22	-0.41	-0.25	0.02	0.11	-0.22	0.25	0.14	0.21	0.36	0.53	0.35	0.59	0.49	0.42	0.36	0.25
L	-0.08	-0.20	-0.30	-0.41	-0.27	-0.29	-0.09	0.24	-0.01	0.23	0.20	0.25	0.26	0.30	0.43	0.67	0.16	0.35	0.19	0.42
V	0.06	-0.14	-0.22	-0.25	-0.29	-0.29	-0.07	0.02	-0.10	0.16	0.25	0.18	0.24	0.50	0.34	0.58	0.19	0.30	0.44	0.09
W	0.08	-0.67	-0.16	0.02	-0.09	-0.07	-0.12	-0.04	-0.09	0.18	0.22	0.34	0.08	0.06	0.29	0.24	-0.12	-0.16	0.22	-0.28
Y	0.04	-0.13	0.00	0.11	0.24	0.02	-0.04	-0.06	0.09	0.14	0.13	0.09	-0.20	-0.20	-0.10	0.00	-0.34	-0.25	-0.21	-0.33
A	0.00	0.25	0.03	-0.22	-0.01	-0.10	-0.09	0.09	-0.13	-0.07	-0.09	-0.06	0.08	0.28	0.26	0.12	0.34	0.43	0.14	0.10
G	-0.08	0.19	0.38	0.25	0.23	0.16	0.18	0.14	-0.07	-0.38	-0.26	-0.16	-0.06	-0.14	0.25	-0.22	0.20	-0.04	0.11	-0.11
T	0.19	0.19	0.31	0.14	0.20	0.25	0.22	0.13	-0.09	-0.26	0.03	-0.08	-0.14	-0.11	0.00	-0.29	-0.19	-0.35	-0.09	-0.07
S	-0.02	0.14	0.29	0.21	0.25	0.18	0.34	0.09	-0.06	-0.16	-0.08	-0.20	-0.14	-0.14	-0.26	-0.31	-0.05	0.17	-0.13	0.01
Q	0.05	0.46	0.49	0.36	0.26	0.24	0.08	-0.20	0.08	-0.06	-0.14	-0.14	0.29	-0.25	-0.17	-0.17	-0.02	-0.52	-0.38	-0.42
N	0.13	0.08	0.18	0.53	0.30	0.50	0.06	-0.20	0.28	-0.14	-0.11	-0.14	-0.25	-0.53	-0.32	-0.30	-0.24	-0.14	-0.33	-0.18
E	0.69	0.44	0.27	0.35	0.43	0.34	0.29	-0.10	0.26	0.25	0.00	-0.26	-0.17	-0.32	-0.03	-0.15	-0.45	-0.74	-0.97	-0.10
D	0.03	0.65	0.39	0.59	0.67	0.58	0.24	0.00	0.12	-0.22	-0.29	-0.31	-0.17	-0.30	-0.15	0.04	-0.39	-0.72	-0.76	0.04
H	-0.19	0.99	-0.16	0.49	0.16	0.19	-0.12	-0.34	0.34	0.20	-0.19	-0.05	-0.02	-0.24	-0.45	-0.39	-0.29	-0.12	0.22	-0.21
R	0.24	0.31	0.41	0.42	0.35	0.30	-0.16	-0.25	0.43	-0.04	-0.35	0.17	-0.52	-0.14	-0.74	-0.72	-0.12	0.11	0.75	-0.38
K	0.71	0.00	0.44	0.36	0.19	0.44	0.22	-0.21	0.14	0.11	-0.09	-0.13	-0.38	-0.33	-0.97	-0.76	0.22	0.75	0.25	0.11
P	0.00	-0.34	0.20	0.25	0.42	0.09	-0.28	-0.33	0.10	-0.11	-0.07	0.01	-0.42	-0.18	-0.10	0.04	-0.21	-0.38	0.11	0.26

Table A.2: Protein amino acid hydrophobicity scale according to Kyte and Doolittle [90]

Residue	Hydrophobicity
Alanine	1.8
Arginine	-4.5
Asparagine	-3.5
Aspartic acid	-3.5
Cysteine	2.5
Glutamine	-3.5
Glutamic acid	-3.5
Glycine	-0.4
Histidine	-3.2
Isoleucine	4.5
Leucine	3.8
Lysine	-3.9
Methionine	1.9
Phenylalanine	2.8
Proline	-1.6
Serine	-0.8
Threonine	-0.7
Tryptophan	-0.9
Tyrosine	-1.3
Valine	4.2

Table A.3: List of Pfam domain families analysed in this study

Pfam Domain Names		
ABC_tran	ABM	AIRS
AIRS_C	AP_endonuc_2	Amidohydro_3
AraC_binding	Arf	ArsA_ATPase
AsnC_trans_reg	B12-binding	BPD_transp_1
Bac_luciferase	CMD	COX1
Continued on next page		

Table A.3 – continued from previous page

Pfam Domain Names		
Cadherin	CbiA	CheW
CoA_transf_3	Cons_hypoth95	Cytochrom_B_C
Cytochrom_B_N	Cytochrom_C	DHH
DHHA1	DNA_gyraseA_C	DegT_DnrJ_EryC1
EAL	FMN_red	Fe-ADH
FecCD	Fer4	Fer4_NifH
Flavin_Reduct	Flavodoxin_2	GGDEF
GTP_EFTU	GerE	Globin
Globin	Glycos_transf_1	Glycos_transf_2
Glyoxalase	GntR	HATPase_c
HD	HTH_1	HTH_3
HTH_5	HTH_8	HTH_IclR
HisKA	HlyD	Homeobox
Hormone_recep	Hpt	HxlR
IclR	IspD	IstB_IS21
Kunitz_BPTI	LacI	Lectin_C
LysR_substrate	MCPsignal	MarR
MerR	MerR-DNA-bind	Methylase_S
MoCF_biosynth	Molydop_binding	Mur_ligase
Mur_ligase_C	Mur_ligase_M	N6_Mtase
N6_N4_Mtase	NMT1	NTP_transferase
Nitroreductase	OEP	OmpA
PAS	PASTA	PAS_3
PD40	PHP	PIN
PQQ	PadR	ParBc
Pentapeptide	Peptidase_M23	Peripla_BP_1
Peripla_BP_2	Phage_integr_N	Phage_integrase
PhoU	PilZ	Plasmid_stabil
Plug	ROK	RRM_1
Radical_SAM	Ras	Resolvase
Response_reg	RibD_C	RimK
Rrf2	RuBisCO_large	SBP_bac_1
SBP_bac_3	SH2	SH3_1
SIS	SLBB	SLT
Serpin	Sigma54_activat	Sigma70_r2
Sigma70_r4	Sigma70_r4_2	Surf_Ag_VNR
Sushi	T2SE	T2SF

Continued on next page

Table A.3 – continued from previous page

Pfam Domain Names		
TOBE	TOBE_2	TP_methylase
TctC	TetR_N	TonB
Toprim	Trans_reg_C	Transpeptidase
Transposase_11	TrkA_N	TrmB
Trypsin	Tubulin	UDPG_MGDP_dh_N
UTRA	Y_phosphatase	YkuD
fn3	zf-C4	

Table A.4: Top 200 interactions predicted by DI_{PW}

Rank	**i** (HXB2)	**i** (domain)	**j** (HXB2)	**j** (domain)
1	283	od	453	od
2	502	gp120_C	607	gp41
3	231	id2	267	od
4	747	gp41	758	gp41
5	97	id1	275	od
6	13	sp	20	sp
7	159	V2	174	V2
8	360	od	465	V5
9	293	od	337	od
10	92	id1	238	id2
11	277	od	352	od
12	308	V3	316	V3
13	816	gp41	824	gp41
14	49	gp120_N	99	id1
15	65	gp120_N	208	id2
16	825	gp41	833	gp41
17	308	V3	315	V3
18	211	id2	379	od
19	232	id2	268	od
20	231	id2	268	od
21	567	gp41	629	gp41
22	219	id2	225	id2
23	114	id1	202	bs

Continued on next page

Table A.4 – continued from previous page

Rank	**i** (HXB2)	**i** (domain)	**j** (HXB2)	**j** (domain)
24	275	od	282	od
25	167	V2	192	V2
26	230	id2	240	id2
27	85	id1	229	id2
28	290	od	340	od
29	178	V2	195	V2
30	106	id1	174	V2
31	801	gp41	825	gp41
32	133	V1	155	V1
33	11	sp	21	sp
34	325	V3	419	od
35	306	V3	321	V3
36	425	bs	432	bs
37	788	gp41	797	gp41
38	770	gp41	783	gp41
39	182	V2	192	V2
40	300	V3	442	od
41	290	od	337	od
42	202	bs	432	bs
43	557	gp41	567	gp41
44	602	gp41	651	gp41
45	845	gp41	851	gp41
46	667	gp41	674	gp41
47	269	od	348	od
48	12	sp	21	sp
49	287	od	481	id3
50	178	V2	194	V2
51	192	V2	426	bs
52	290	od	344	od
53	805	gp41	853	gp41
54	270	od	277	od
55	46	gp120_N	492	id3
56	10	sp	21	sp
57	333	od	389	V4
58	172	V2	305	V3
59	335	od	412	V4
60	12	sp	30	sp

Continued on next page

Table A.4 – continued from previous page

Rank	**i** (HXB2)	**i** (domain)	**j** (HXB2)	**j** (domain)
61	740	gp41	796	gp41
62	309	V3	317	V3
63	360	od	467	V5
64	800	gp41	853	gp41
65	154	V1	300	V3
66	698	gp41	705	gp41
67	121	bs	429	bs
68	700	gp41	758	gp41
69	12	sp	20	sp
70	303	V3	323	V3
71	632	gp41	640	gp41
72	12	sp	23	sp
73	500	gp120_C	619	gp41
74	677	gp41	683	gp41
75	232	id2	269	od
76	273	od	481	id3
77	306	V3	316	V3
78	161	V2	172	V2
79	293	od	446	od
80	720	gp41	727	gp41
81	456	od	466	V5
82	328	V3	334	od
83	726	gp41	736	gp41
84	279	od	474	od
85	721	gp41	732	gp41
86	595	gp41	602	gp41
87	761	gp41	769	gp41
88	164	V2	170	V2
89	353	od	468	V5
90	816	gp41	825	gp41
91	175	V2	194	V2
92	9	sp	21	sp
93	518	gp41	553	gp41
94	13	sp	19	sp
95	725	gp41	731	gp41
96	750	gp41	756	gp41
97	305	V3	319	V3

Continued on next page

Table A.4 – continued from previous page

Rank	**i** (HXB2)	**i** (domain)	**j** (HXB2)	**j** (domain)
98	167	V2	177	V2
99	10	sp	282	od
100	281	od	365	od
101	309	V3	315	V3
102	202	bs	315	V3
103	232	id2	240	id2
104	174	V2	429	bs
105	723	gp41	731	gp41
106	65	gp120_N	379	od
107	10	sp	23	sp
108	95	id1	236	id2
109	784	gp41	800	gp41
110	308	V3	317	V3
111	565	gp41	646	gp41
112	651	gp41	658	gp41
113	346	od	395	V4
114	800	gp41	825	gp41
115	295	od	446	od
116	809	gp41	853	gp41
117	32	sp	500	gp120_C
118	471	V5	477	id3
119	121	bs	202	bs
120	619	gp41	646	gp41
121	665	gp41	677	gp41
122	11	sp	26	sp
123	10	sp	20	sp
124	7	sp	21	sp
125	134	V1	154	V1
126	25	sp	31	sp
127	152	V1	181	V2
128	720	gp41	750	gp41
129	809	gp41	824	gp41
130	289	od	344	od
131	792	gp41	800	gp41
132	106	id1	121	bs
133	283	od	471	V5
134	158	V2	173	V2

Continued on next page

Table A.4 – continued from previous page

Rank	**i** (HXB2)	**i** (domain)	**j** (HXB2)	**j** (domain)
135	671	gp41	683	gp41
136	269	od	351	od
137	602	gp41	654	gp41
138	300	V3	323	V3
139	796	gp41	812	gp41
140	624	gp41	632	gp41
141	725	gp41	743	gp41
142	114	id1	553	gp41
143	792	gp41	798	gp41
144	722	gp41	824	gp41
145	195	V2	432	bs
146	302	V3	323	V3
147	133	V1	152	V1
148	700	gp41	746	gp41
149	46	gp120_N	632	gp41
150	788	gp41	805	gp41
151	162	V2	195	V2
152	25	sp	722	gp41
153	24	sp	722	gp41
154	23	sp	29	sp
155	301	V3	323	V3
156	164	V2	195	V2
157	62	gp120_N	209	id2
158	10	sp	809	gp41
159	746	gp41	758	gp41
160	788	gp41	800	gp41
161	20	sp	26	sp
162	801	gp41	824	gp41
163	7	sp	20	sp
164	183	V2	194	V2
165	167	V2	309	V3
166	121	bs	315	V3
167	244	id2	629	gp41
168	753	gp41	762	gp41
169	369	od	429	bs
170	167	V2	426	bs
171	458	od	466	V5

Continued on next page

Table A.4 – continued from previous page

Rank	i (HXB2)	i (domain)	j (HXB2)	j (domain)
172	22	sp	29	sp
173	749	gp41	758	gp41
174	288	od	341	od
175	270	od	341	od
176	295	od	413	V4
177	236	id2	792	gp41
178	793	gp41	804	gp41
179	746	gp41	756	gp41
180	720	gp41	796	gp41
181	236	id2	275	od
182	499	gp120_C	605	gp41
183	492	id3	612	gp41
184	588	gp41	646	gp41
185	23	sp	853	gp41
186	9	sp	22	sp
187	307	V3	319	V3
188	177	V2	192	V2
189	232	id2	271	od
190	379	od	443	od
191	722	gp41	746	gp41
192	174	V2	333	od
193	92	id1	633	gp41
194	256	od	265	od
195	232	id2	238	id2
196	845	gp41	854	gp41
197	275	od	474	od
198	281	od	353	od
199	750	gp41	758	gp41
200	295	od	444	od

Table A.5: Intra-V1V2 interactions predicted by DI_{PW}

Rank	i (HXB2)	i (domain)	j (HXB2)	j (domain)
7	159	V2	174	V2
25	167	V2	192	V2

Continued on next page

Table A.5 – continued from previous page

Rank	**i** (HXB2)	**i** (domain)	**j** (HXB2)	**j** (domain)
29	178	V2	195	V2
32	133	V1	155	V1
39	182	V2	192	V2
50	178	V2	194	V2
78	161	V2	172	V2
88	164	V2	170	V2
91	175	V2	194	V2
98	167	V2	177	V2
125	134	V1	154	V1
127	152	V1	181	V2
134	158	V2	173	V2
147	133	V1	152	V1
151	162	V2	195	V2
156	164	V2	195	V2
164	183	V2	194	V2
188	177	V2	192	V2

Table A.6: Intra-gp41 interactions predicted by DI_{PW}

Rank	**i** (HXB2)	**i** (domain)	**j** (HXB2)	**j** (domain)
4	747	gp41	758	gp41
13	816	gp41	824	gp41
16	825	gp41	833	gp41
21	567	gp41	629	gp41
31	801	gp41	825	gp41
37	788	gp41	797	gp41
38	770	gp41	783	gp41
43	557	gp41	567	gp41
44	602	gp41	651	gp41
45	845	gp41	851	gp41
46	667	gp41	674	gp41
53	805	gp41	853	gp41
61	740	gp41	796	gp41
64	800	gp41	853	gp41
66	698	gp41	705	gp41

Continued on next page

Table A.6 – continued from previous page

Rank	**i** (HXB2)	**i** (domain)	**j** (HXB2)	**j** (domain)
68	700	gp41	758	gp41
71	632	gp41	640	gp41
74	677	gp41	683	gp41
80	720	gp41	727	gp41
83	726	gp41	736	gp41
85	721	gp41	732	gp41
86	595	gp41	602	gp41
87	761	gp41	769	gp41
90	816	gp41	825	gp41
93	518	gp41	553	gp41
95	725	gp41	731	gp41
96	750	gp41	756	gp41
105	723	gp41	731	gp41
109	784	gp41	800	gp41
111	565	gp41	646	gp41
112	651	gp41	658	gp41
114	800	gp41	825	gp41
116	809	gp41	853	gp41
120	619	gp41	646	gp41
121	665	gp41	677	gp41
128	720	gp41	750	gp41
129	809	gp41	824	gp41
131	792	gp41	800	gp41
135	671	gp41	683	gp41
137	602	gp41	654	gp41
139	796	gp41	812	gp41
140	624	gp41	632	gp41
141	725	gp41	743	gp41
143	792	gp41	798	gp41
144	722	gp41	824	gp41
148	700	gp41	746	gp41
150	788	gp41	805	gp41
159	746	gp41	758	gp41
160	788	gp41	800	gp41
162	801	gp41	824	gp41
168	753	gp41	762	gp41
173	749	gp41	758	gp41

Continued on next page

Table A.6 – continued from previous page

Rank	**i** (HXB2)	**i** (domain)	**j** (HXB2)	**j** (domain)
178	793	gp41	804	gp41
179	746	gp41	756	gp41
180	720	gp41	796	gp41
184	588	gp41	646	gp41
191	722	gp41	746	gp41
196	845	gp41	854	gp41
199	750	gp41	758	gp41

Table A.7: Intra-gp41 (HXB2 position <= 633) interactions predicted by $\mathbf{DI_{PW}}$

Rank	**i** (HXB2)	**i** (domain)	**j** (HXB2)	**j** (domain)
21	567	gp41	629	gp41
43	557	gp41	567	gp41
86	595	gp41	602	gp41
93	518	gp41	553	gp41
140	624	gp41	632	gp41

Algorithm 1: SMS-EMOA pseudocode

$P_0 \leftarrow \text{init}()$ *Initial population with μ individuals*

$t \leftarrow 0$ *Generation index*

repeat

  $o \leftarrow \text{reproduction}(P_t)$ *Generate 1 new offspring by reproduction*

  $D \leftarrow \text{dom}(P_t \cup \{o\})$ *Identify set of dominated individuals*

  if $D \neq \emptyset$ **then**

    $\mathbf{r}^* \leftarrow argmax_{r \in D}[dom(\mathbf{r}, D)]$

    Individual(s) with highest dominance number

    if $\{\mathbf{r}^*\} > 1$ **then**

      More than one individual with the highest dominance number

      $\mathbf{a}^* \leftarrow argmin_{\mathbf{a} \in \{\mathbf{r}^*\}}[hyp(\mathbf{a}, \{\mathbf{r}^*\})]$

      Individual(s) with the lowest hyervolume contribution

      if $\{\mathbf{a}^*\} > 1$ **then**

        More than one individual with the lowest hypervolume contribution

        $\mathbf{s} \leftarrow \text{random}(\{\mathbf{a}^*\})$ *Choose randomly an individual*

      end

    else

      $\mathbf{s} \leftarrow \mathbf{r}^*$ *Choose individual with highest dominance number*

    end

  else

    $\mathbf{r}^* \leftarrow argmin_{\mathbf{r} \in \{P_t \cup \{o\}\}}[hyp(\mathbf{r}, P_t \cup \{o\})]$

    Individual(s) with lowest hypervolume contribution

    if $\{\mathbf{r}^*\} > 1$ **then**

      $\mathbf{s} \leftarrow \text{random}(\{\mathbf{a}^*\})$ *Choose randomly an individual*

    end

  end

  $P_{t+1} \leftarrow \{P_t \cup \{o\}\} \backslash \{\mathbf{s}\}$ *Remove worst individual* $\mathbf{s}$

  $t \leftarrow t + 1$

until *Stop criterion*;

Algorithm 2: Metropolis MC algorithm pseudocode

t_{max} *maximal number of steps*

$conf_{\mathbf{initial}}$ *initial conformation*

$conf_{\mathbf{actual}} \leftarrow conf_{\mathbf{initial}}$ *actual conformation*

$t \leftarrow 0$

while $t < t_{max}$ **do**

  $t \leftarrow t + 1$

  $conf_{\mathbf{new}} \leftarrow random\{N(conf_{\mathbf{actual}})\}$ *choose randomly a neighbourhood conformation*

  read current;

  if $E(conf_{\mathbf{new}}) \leq E(conf_{\mathbf{actual}})$ **then** *new conformation's energy is favourable*

    $conf_{\mathbf{actual}} \leftarrow conf_{\mathbf{new}}$ *accept new conformation*

  else *new conformation's energy is not favourable*

    $r \leftarrow random[0, 1]$ *random number between 0 and 1*

    if $r \leq min(1, e^{-\frac{\Delta E}{kT}})$ **then** *apply Metropolis criterion*

      $conf_{\mathbf{actual}} \leftarrow conf_{\mathbf{new}}$ *accept new conformation*

    else

      $conf_{\mathbf{actual}}$ *reject new and keep actual conformation*

    end

  end

end

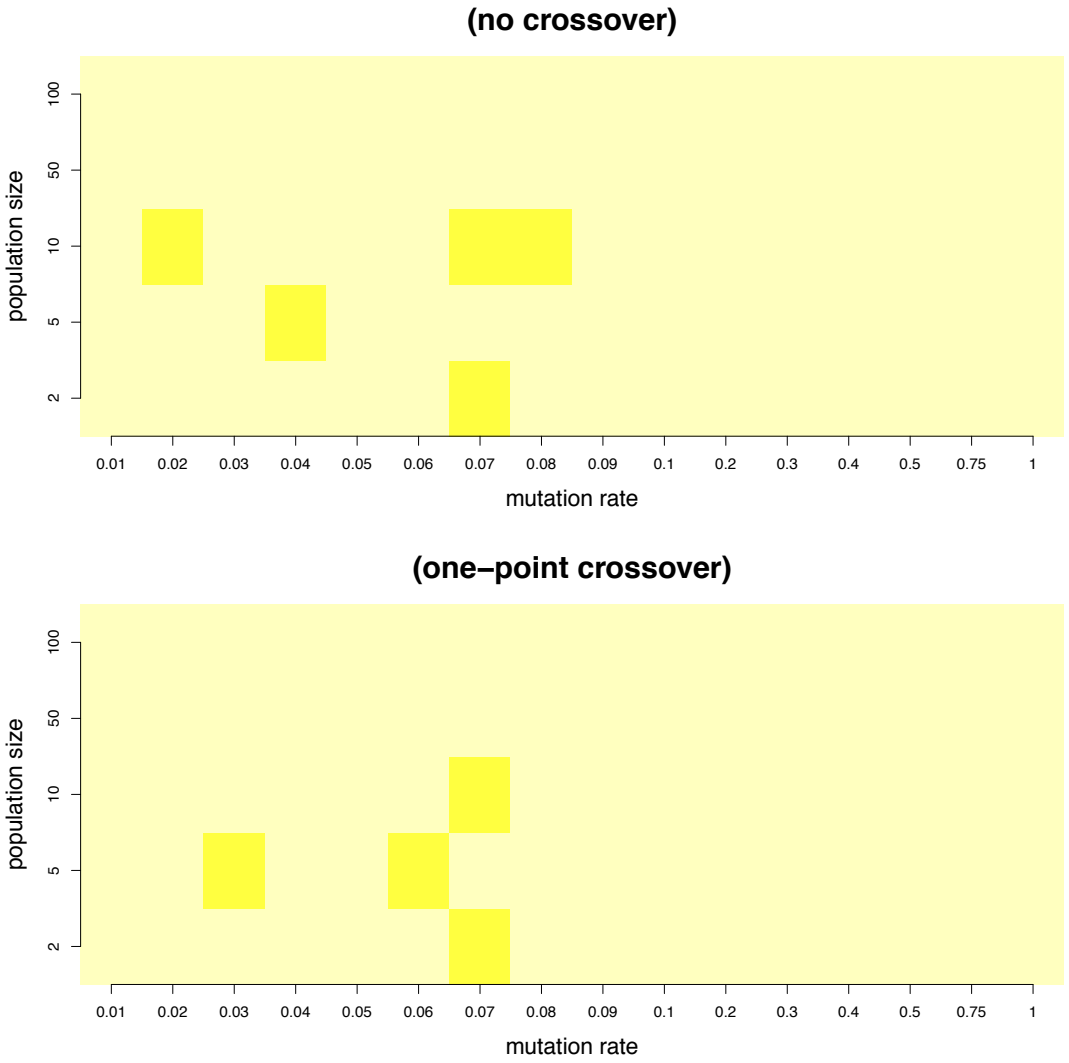

Figure A.2: Enumerator heat map error estimates - The hypervolume standard deviations of three different seeds are shown as heat map for optimisation with and without applying crossover parameter.

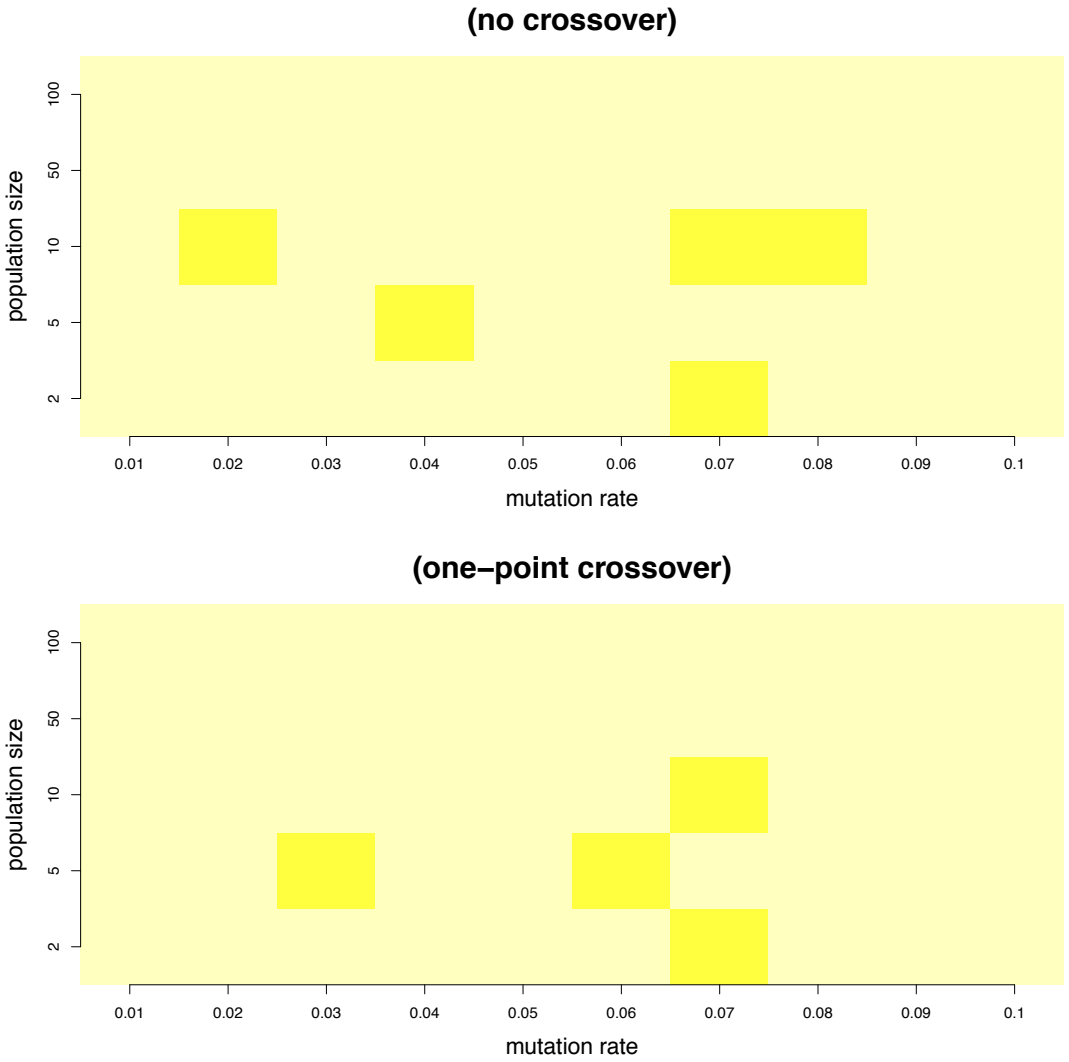

Figure A.3: Enumerator heat map error estimates (close-up view) - The hypervolume standard deviations of three different seeds are shown as heat map for optimisation with and without applying crossover parameter. Only mutation rates between 0 and 0.1 are illustrated.

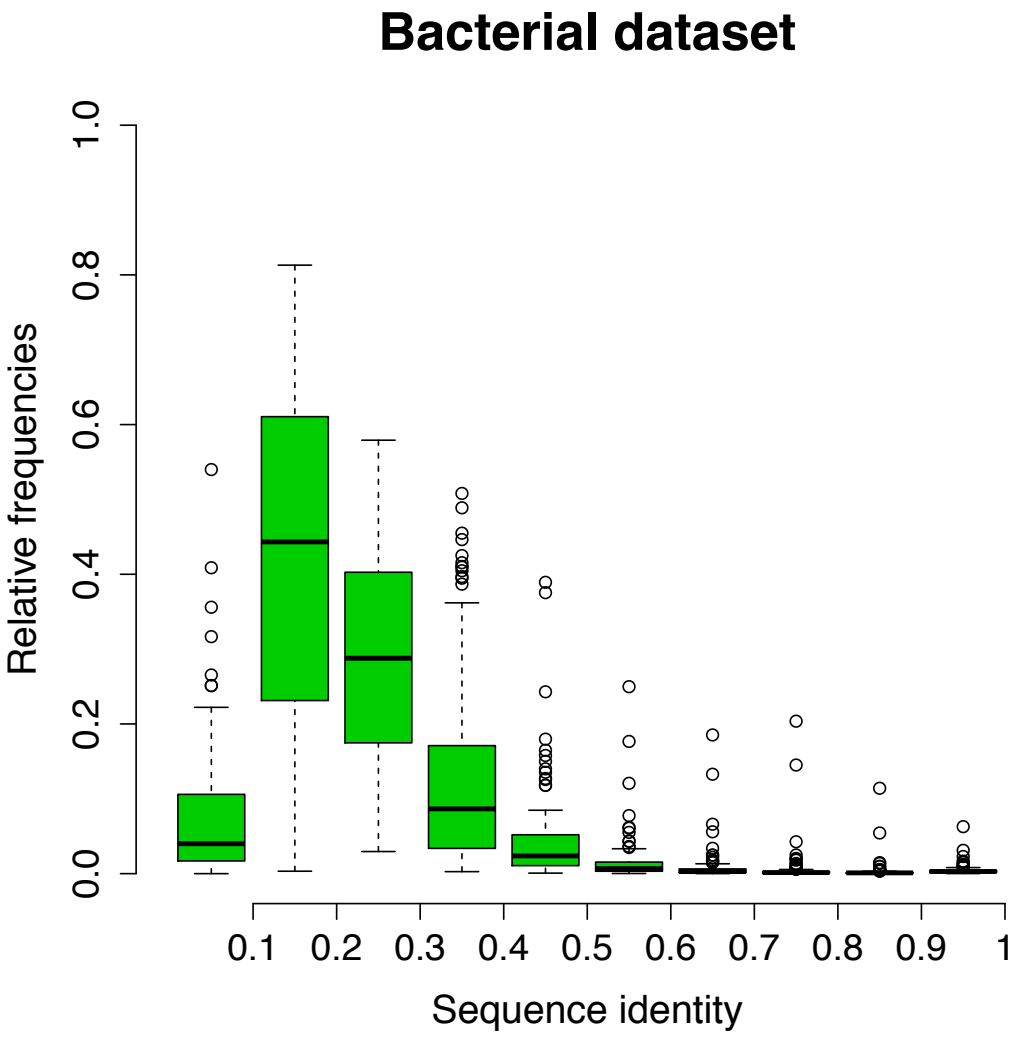

Figure A.4: Boxplot of averaged (over 124 bacterial protein families) sequence identities

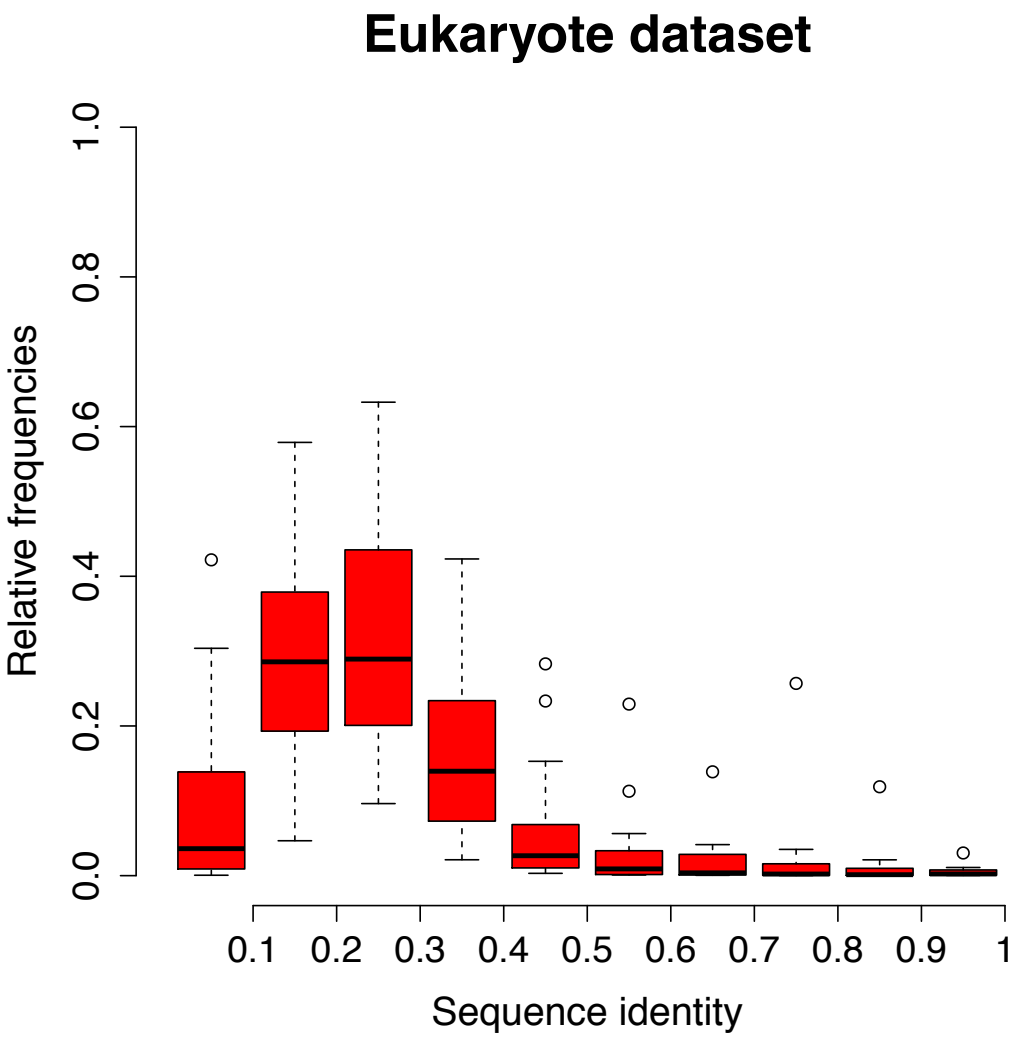

Figure A.5: Boxplot of averaged (over 22 eukaryote protein families) sequence identities

Bibliography

[1] P. Acharya, T. S. Luongo, M. K. Louder, K. McKee, Y. Yang, Y. Do Kwon, J. R. Mascola, P. Kessler, L. Martin, and P. D. Kwong. Structural basis for highly effective HIV-1 neutralization by CD4-mimetic miniproteins revealed by 1.5 Åcocrystal structure of gp120 and M48U1. *Structure (London, England : 1993)*, 21(6):1018–29, June 2013. 49

[2] C. B. Anfinsen. Principles that govern the folding of protein chains. *Science (New York, N.Y.)*, 181(4096):223–30, July 1973. 9

[3] R. Ash. *Information Theory*. John Wiley & Sons, Inc., New York, USA, 1965. 56

[4] T. Baeck, D. B. Fogel, and Z. Michalewicz. *Handbook of Evolutionary Computation*. Oxford University Press, 1997. 14

[5] F. Barre-Sinoussi, J. Chermann, F. Rey, M. Nugeyre, S. Chamaret, J. Gruest, C. Dauguet, C. Axler-Blin, F. Vezinet-Brun, C. Rouzioux, W. Rozenbaum, and L. Montagnier. Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS). *Science*, 220(4599):868–871, May 1983. 43

[6] W. L. Bazari, P. Matsudaira, M. Wallek, T. Smeal, R. Jakes, and Y. Ahmed. Villin sequence and peptide map identify six homologous domains. *Proceedings of the National Academy of Sciences of the United States of America*, 85(14):4986–90, July 1988. 22

[7] R. Benayoun, J. Montgolfier, J. Tergny, and O. Laritchev. Linear programming with multiple objective functions: Step method (stem). *Mathematical Programming*, 1(1):366–375, December 1971. 7

[8] F. C. Bernstein, T. F. Koetzle, G. J. Williams, E. F. Meyer, M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi. The Protein Data Bank: a computer-based archival file for macromolecular structures. *Journal of molecular biology*, 112(3):535–42, May 1977. 50, 55

[9] N. Beume, B. Naujoks, and M. Emmerich. SMS-EMOA: Multiobjective selection based on dominated hypervolume. *European Journal of Operational Research*, 181(3):1653–1669, September 2007. 18

[10] T. L. Blundell, B. L. Sibanda, M. J. Sternberg, and J. M. Thornton. Knowledge-based prediction of protein structures and the design of novel molecules. *Nature*, 326(6111):347–52, 1987. 61

[11] E. Bornberg-Bauer. Randomness, Structural Uniqueness, Modularity and Neutral Evolution in Sequence Space of Model Proteins. *Zeitschrift für Physikalische Chemie*, 216(2/2002), January 2002. 11

[12] C. Branden and J. Tooze. *Introduction to Protein Structure*. Garland Science, New York, 1999. 11

[13] A. Brelot, N. Heveker, K. Adema, M. J. Hosie, B. Willett, and M. Alizon. Effect of mutations in the second extracellular loop of CXCR4 on its utilization by human and feline immunodeficiency viruses. *Journal of virology*, 73(4):2576–86, April 1999. 48

[14] J. A. G. Briggs, T. Wilk, R. Welker, H.-G. Kräusslich, and S. D. Fuller. Structural organization of authentic, mature HIV-1 virions and cores. *The EMBO journal*, 22(7):1707–15, April 2003. 44

[15] J. T. Buchanan. A naïve approach for solving MCDM problems: the GUESS method. *Journal of the Operational Research Society*, 48(2):202–0206, December 1997. 7

[16] V. Buzon, G. Natrajan, D. Schibli, F. Campelo, M. M. Kozlov, and W. Weissenhorn. Crystal structure of HIV-1 gp41 including both fusion peptide and membrane proximal external regions. *PLoS pathogens*, 6(5):e1000880, May 2010. 81

[17] D. C. Chan, D. Fass, J. M. Berger, and P. S. Kim. Core structure of gp41 from the HIV envelope glycoprotein. *Cell*, 89(2):263–73, April 1997. 50

[18] V. Chankong and Y. Y. Haimes. *Multiobjective Decision Making: Theory and Methodology*. North-Holland, New York, USA, 1983. 7

[19] A. Charnes, W. W. Cooper, and R. O. Ferguson. Optimal Estimation of Executive Compensation by Linear Programming. *Management Science*, 1(2):138–151, January 1955. 7

[20] B. Chen, E. M. Vogan, H. Gong, J. J. Skehel, D. C. Wiley, and S. C. Harrison. Determining the structure of an unliganded and fully glycosylated SIV gp120 envelope glycoprotein. *Structure (London, England : 1993)*, 13(2):197–211, February 2005. 47

[21] L. Chen, Y. D. Kwon, T. Zhou, X. Wu, S. O'Dell, L. Cavacini, A. J. Hessell, M. Pancera, M. Tang, L. Xu, Z.-Y. Yang, M.-Y. Zhang, J. Arthos, D. R. Burton, D. S. Dimitrov, G. J. Nabel, M. R. Posner, J. Sodroski, R. Wyatt, J. R. Mascola, and P. D. Kwong. Structural basis of immune evasion at the site of CD4 attachment on HIV-1 gp120. *Science (New York, N.Y.)*, 326(5956):1123–7, November 2009. 49

[22] D. K. Chiu and T. Kolodziejczak. Inferring consensus structure from nucleic acid sequences. *Computer applications in the biosciences : CABIOS*, 7(3):347–52, July 1991. 54

[23] C. Chothia and A. M. Lesk. The relation between the divergence of sequence and structure in proteins. *The EMBO journal*, 5(4):823–6, April 1986. 61

[24] S. K. Choudhary and D. M. Margolis. Curing HIV: Pharmacologic approaches to target HIV-1 latency. *Annual review of pharmacology and toxicology*, 51:397–418, January 2011. 45

[25] M. Citossi and G. Giugliarelli. Lattice protein models: A computatinoal approach to folding and aggregation phenomena. In *Frontiers of Fundamental and Computational Physics*, pages 355–358. Springer Netherlands, 2005. 11

[26] P. M. Colman and M. C. Lawrence. The structural biology of type I viral membrane fusion. *Nature reviews. Molecular cell biology*, 4(4):309–19, April 2003. 47

[27] E. G. Cormier and T. Dragic. The crown and stem of the V3 loop play distinct roles in human immunodeficiency virus type 1 envelope glycoprotein interactions with the CCR5 coreceptor. *Journal of virology*, 76(17):8953–7, September 2002. 48, 74

[28] D. W. Corne, N. R. Jerram, J. D. Knowles, and M. J. Oates. PESA-II: region-based selection in evolutionary multiobjective optimization. In *Proceedings of the Genetic and Evolutionary Computation Conference*, pages 283–290. San Francisco, CA, USA, 2001. 15, 17

[29] D. W. Corne and J. D. Knowles. The Pareto-envelope based selection algorithm for multiobjective optimization. In *Proceedings of the Sixth International Conference on Parallel Problem Solving from Nature*, pages 839–848. Springer, Berlin, 2000. 15, 17

[30] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley Series in Telecommunications. John Wiley & Sons, Inc., New York, USA, 1991. 54, 56, 57

[31] F. L. Custódio, H. J. C. Barbosa, and L. E. Dardenne. Investigation of the three-dimensional lattice HP protein folding model using a genetic algorithm. *Genetics and Molecular Biology*, 27(4):611–615, 2004. 2

[32] C. R. Darwin. *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*. John Murray, 1859. 1

[33] K. Deb. *Multi-Objective Optimization Using Evolutionary Algorithms*. John Wiley & Sons, 2001. 6, 7, 14

[34] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation*, 6(2):182–197, April 2002. 15, 16, 18

[35] H. Deng, R. Liu, W. Ellmeier, S. Choe, D. Unutmaz, M. Burkhart, P. Di Marzio, S. Marmon, R. E. Sutton, C. M. Hill, C. B. Davis, S. C. Peiper, T. J. Schall, D. R. Littman, and N. R. Landau. Identification of a major co-receptor for primary isolates of HIV-1. *Nature*, 381(6584):661–6, June 1996. 47

[36] J. R. Desjarlais and T. M. Handel. De novo design of the hydrophobic cores of proteins. *Protein science : a publication of the Protein Society*, 4(10):2006–18, October 1995. 2

[37] O. Dictionaries. Evolution. 2013. 1

[38] R. Diskin, P. M. Marcovecchio, and P. J. Bjorkman. Structure of a clade C HIV-1 gp120 bound to CD4 and CD4-induced antibody reveals anti-CD4 polyreactivity. *Nature structural & molecular biology*, 17(5):608–13, May 2010. 49

[39] R. Diskin, J. F. Scheid, P. M. Marcovecchio, A. P. West, F. Klein, H. Gao, P. N. P. Gnanapragasam, A. Abadir, M. S. Seaman, M. C. Nussenzweig, and P. J. Bjorkman. Increasing the potency and breadth of an HIV antibody by using structure-based rational design. *Science (New York, N.Y.)*, 334(6060):1289–93, December 2011. 49

[40] T. Dragic, V. Litwin, G. P. Allaway, S. R. Martin, Y. Huang, K. A. Nagashima, C. Cayanan, P. J. Maddon, R. A. Koup, J. P. Moore, and W. A. Paxton. HIV-1 entry into CD4+ cells is mediated by the chemokine receptor CC-CKR-5. *Nature*, 381(6584):667–73, June 1996. 47

[41] S. D. Dunn, L. M. Wahl, and G. B. Gloor. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics (Oxford, England)*, 24(3):333–40, February 2008. 54

[42] R. Durbin, S. R. Eddy, A. Krogh, and G. Mitchison. *Biological sequence analysis: Probabilistic models of proteins and nucleic acids.* Cambridge University Press, 1998. 60

[43] D. M. Eckert and P. S. Kim. Mechanisms of viral membrane fusion and its inhibition. *Annual review of biochemistry*, 70:777–810, January 2001. 47

[44] M. Emmerich, N. Beume, and B. Naujoks. An EMO algorithm using the hypervolume measure as selection criterion. In *EMO'05 Proceedings of the Third international conference on Evolutionary Multi-Criterion Optimization*, pages 62–76. Springer Berlin / Heidelberg, 2005. 18

[45] N. Eswar, B. Webb, M. A. Marti-Renom, M. S. Madhusudhan, D. Eramian, M.-Y. Shen, U. Pieper, and A. Sali. Comparative protein structure modeling using MODELLER. *Current protocols in protein science / editorial board, John E. Coligan ... [et al.]*, Chapter 2:Unit 2.9, November 2007. 62

[46] Q. Fang and D. Shortle. Protein refolding in silico with atom-based statistical potentials and conformational search using a simple genetic algorithm. *Journal of molecular biology*, 359(5):1456–67, June 2006. 2

[47] P. Fariselli, O. Olmea, A. Valencia, and R. Casadio. Progress in predicting inter-residue contacts of proteins with neural networks and correlated mutations. *Proteins*, Suppl 5:157–62, January 2001. 53

[48] M. Farzan, T. Mirzabekov, P. Kolchinsky, R. Wyatt, M. Cayabyab, N. P. Gerard, C. Gerard, J. Sodroski, and H. Choe. Tyrosine sulfation of the amino terminus of CCR5 facilitates HIV-1 entry. *Cell*, 96(5):667–76, March 1999. 48

[49] T. Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, June 2006. 66

[50] Y. Feng, C. C. Broder, P. E. Kennedy, and E. A. Berger. HIV-1 entry cofactor: functional cDNA cloning of a seven-transmembrane, G protein-coupled receptor. *Science (New York, N.Y.)*, 272(5263):872–7, May 1996. 47

[51] A. D. Fernandes and G. B. Gloor. Mutual information is critically dependent on prior assumptions: would the correct estimate of mutual information please identify itself? *Bioinformatics (Oxford, England)*, 26(9):1135–9, May 2010. 57

[52] W. M. Fitch and E. Markowitz. An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. *Biochemical genetics*, 4(5):579–93, October 1970. 53

[53] C. M. Fonseca and P. J. Fleming. Multiobjective genetic algorithms. In *Genetic Algorithms for Control Systems Engineering*. London, 1993. 14, 16

[54] A. E. Friedman-Kien, L. J. Laubenstein, P. Rubinstein, E. Buimovici-Klein, M. Marmor, R. Stahl, I. Spigland, K. S. Kim, and S. Zolla-Pazner. Disseminated Kaposi's sarcoma in homosexual men. *Annals of internal medicine*, 96(6 Pt 1):693–700, June 1982. 43

[55] R. Gallo, S. Salahuddin, M. Popovic, G. Shearer, M. Kaplan, B. Haynes, T. Palker, R. Redfield, J. Oleske, B. Safai, and A. Et. Frequent detection and isolation of cytopathic retroviruses (HTLV-III) from patients with AIDS and at risk for AIDS. *Science*, 224(4648):500–503, May 1984. 43

[56] M. Gen and R. Cheng. *Genetic Algorithms and Engineering Design*. John Wiley & Sons, 1997. 14

[57] U. Göbel, C. Sander, R. Schneider, and A. Valencia. Correlated mutations and residue contacts in proteins. *Proteins*, 18(4):309–17, April 1994. 53, 54

[58] D. E. Goldberg. *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley, 1989. 14

[59] D. E. Goldberg and J. Richardson. Genetic algorithms with sharing for multimodal function optimization. In *Proceedings of the Second International Conference on Genetic Algorithms on Genetic algorithms and their application*, pages 41–49. Cambridge, MA, USA, 1987. 16

[60] C. Gondro and B. P. Kinghorn. A simple genetic algorithm for multiple sequence alignment. *Genetics and molecular research : GMR*, 6(4):964–82, January 2007. 14

[61] L. M. Gordon, P. W. Mobley, W. Lee, S. Eskandari, Y. N. Kaznessis, M. A. Sherman, and A. J. Waring. Conformational mapping of the N-terminal peptide of HIV-1 gp41 in lipid detergent and aqueous environments using 13C-enhanced Fourier transform infrared spectroscopy. *Protein science : a publication of the Protein Society*, 13(4):1012–30, April 2004. 81

[62] M. S. Gottlieb. Pneumocystis PneumoniaLos Angeles. *American Journal of Public Health*, 96(6):980–981, June 2006. 43

[63] W. Gronwald, T. Hohm, and D. Hoffmann. Evolutionary Pareto-optimization of stably folding peptides. *BMC bioinformatics*, 9:109, January 2008. 2, 22

[64] M. Guttman, M. Kahn, N. K. Garcia, S.-L. Hu, and K. K. Lee. Solution structure, conformational dynamics, and CD4-induced activation in full-length, glycosylated, monomeric HIV gp120. *Journal of virology*, 86(16):8750–64, August 2012. 50, 79

[65] Y. Y. Haimes, L. S. Lasdon, and D. A. Wismer. On a Bicriterion Formulation of the Problems of Integrated System Identification and System Optimization. *IEEE Transactions on Systems, Man, and Cybernetics*, 1(3):296–297, 1971. 7

[66] T. Hill, A. Lundgren, R. Fredriksson, and H. B. Schiöth. Genetic algorithm for large-scale maximum parsimony phylogenetic analysis of proteins. *Biochimica et biophysica acta*, 1725(1):19–29, August 2005. 14

[67] J. H. Holland. *Adaptation in Natural and Artificial Systems*. The University of Michigan Press, 1975. 14

[68] T. A. Hopf, L. J. Colwell, R. Sheridan, B. Rost, C. Sander, and D. S. Marks. Three-dimensional structures of membrane proteins from genomic sequencing. *Cell*, 149(7):1607–21, June 2012. 54, 86

[69] G. Hu, J. Liu, K. A. Taylor, and K. H. Roux. Structural comparison of HIV-1 envelope spikes with and without the V1/V2 loop. *Journal of virology*, 85(6):2741–50, March 2011. 50, 79

[70] C.-C. Huang, S. N. Lam, P. Acharya, M. Tang, S.-H. Xiang, S. S.-U. Hussan, R. L. Stanfield, J. Robinson, J. Sodroski, I. A. Wilson, R. Wyatt, C. A. Bewley, and P. D. Kwong. Structures of the CCR5 N terminus and of a tyrosine-sulfated antibody with HIV-1 gp120 and CD4. *Science (New York, N.Y.)*, 317(5846):1930–4, September 2007. 48, 49, 52, 72, 74, 78

[71] C.-c. Huang, M. Tang, M.-Y. Zhang, S. Majeed, E. Montabana, R. L. Stanfield, D. S. Dimitrov, B. Korber, J. Sodroski, I. A. Wilson, R. Wyatt, and P. D. Kwong. Structure of a V3-containing HIV-1 gp120 core. *Science (New York, N.Y.)*, 310(5750):1025–8, November 2005. 48, 49, 50, 51

[72] A. Hughes and M. Nelson. HIV entry: new insights and implications for patient management. *Current opinion in infectious diseases*, 22(1):35–42, February 2009. 48

[73] K. Hymes, J. Greene, A. Marcus, D. William, T. Cheung, N. Prose, H. Ballard, and L. Laubenstein. KAPOSI'S SARCOMA IN HOMOSEXUAL MENA REPORT OF EIGHT CASES. *The Lancet*, 318(8247):598–600, September 1981. 43

[74] J. P. Ignizio. A Review of Goal Programming: A Tool for Multiobjective Analysis. *Journal of the Operational Research Society*, 29(11):1109–1119, November 1978. 7

[75] J. Jardine, J.-P. Julien, S. Menis, T. Ota, O. Kalyuzhniy, A. McGuire, D. Sok, P.-S. Huang, S. MacPherson, M. Jones, T. Nieusma, J. Mathison, D. Baker, A. B. Ward, D. R. Burton, L. Stamatatos, D. Nemazee, I. A. Wilson, and W. R. Schief. Rational HIV immunogen design to target specific germline B cell receptors. *Science (New York, N.Y.)*, 340(6133):711–6, May 2013. 49

[76] A. Jaszkiewicz and R. Slowinsky. The Light Beam Search Over a Non-dominated Surface of a Multiple-objective Programming Problem. In G. H. Tzeng, H. F. Wang, U. P. Wen, and P. L. Yu, editors, *Proceedings of the Tenth International Conference: Expand and Enrich the Domains of Thinking and Application*, pages 87–99. Springer New York, New York, NY, 1994. 7

[77] M. Karplus and J. Kuriyan. Molecular dynamics and protein function. *Proceedings of the National Academy of Sciences of the United States of America*, 102(19):6679–85, May 2005. 10

[78] I. Kass and A. Horovitz. Mapping pathways of allosteric communication in GroEL by analysis of correlated mutations. *Proteins*, 48(4):611–7, September 2002. 54

[79] F. Kiefer, K. Arnold, M. Künzli, L. Bordoli, and T. Schwede. The SWISS-MODEL Repository and associated resources. *Nucleic acids research*, 37(Database issue):D387–92, January 2009. 80

[80] U. Klahre, E. Friederich, B. Kost, D. Louvard, and N. H. Chua. Villin-like actin-binding proteins are expressed ubiquitously in Arabidopsis. *Plant physiology*, 122(1):35–48, January 2000. 22

[81] J. Knowles and D. Corne. The Pareto archived evolution strategy: a new baseline algorithm for Pareto multiobjective optimisation. In *Proceedings of the 1999 Congress on Evolutionary Computation-CEC99 (Cat. No. 99TH8406)*, pages 98–105. IEEE, 1999. 17

[82] J. Knowles and D. Corne. Properties of an adaptive archiving algorithm for storing nondominated vectors. *IEEE Transactions on Evolutionary Computation*, 7(2):100–116, April 2003. 18

[83] J. Knowles, D. Corne, and M. Fleischer. Bounded archiving using the lebesgue measure. In *The 2003 Congress on Evolutionary Computation, 2003. CEC '03.*, volume 4, pages 2490–2497. IEEE, 2003. 18

[84] J. D. Knowles and D. W. Corne. Approximating the nondominated front using the Pareto Archived Evolution Strategy. *Evolutionary computation*, 8(2):149–72, January 2000. 15, 17

[85] A. Konak, D. W. Coit, and A. E. Smith. Multi-objective optimization using genetic algorithms: A tutorial. *Reliability Engineering & System Safety*, 91(9):992–1007, September 2006. 5, 6

[86] B. T. Korber, R. M. Farber, D. H. Wolpert, and A. S. Lapedes. Covariation of mutations in the V3 loop of human immunodeficiency virus type 1 envelope protein: an information theoretic analysis. *Proceedings of the National Academy of Sciences of the United States of America*, 90(15):7176–80, August 1993. 54

[87] Y. D. Kwon, A. Finzi, X. Wu, C. Dogo-Isonagie, L. K. Lee, L. R. Moore, S. D. Schmidt, J. Stuckey, Y. Yang, T. Zhou, J. Zhu, D. A. Vicic, A. K. Debnath, L. Shapiro, C. A. Bewley, J. R. Mascola, J. G. Sodroski, and P. D. Kwong. Unliganded HIV-1 gp120 core structures assume the CD4-bound conformation with regulation by quaternary interactions and variable loops. *Proceedings of the National Academy of Sciences of the United States of America*, 109(15):5663–8, April 2012. 49, 76

[88] P. D. Kwong, R. Wyatt, S. Majeed, J. Robinson, R. W. Sweet, J. Sodroski, and W. A. Hendrickson. Structures of HIV-1 gp120 Envelope Glycoproteins from Laboratory-Adapted and Primary Isolates. *Structure*, 8(12):1329–1339, December 2000. 49

[89] P. D. Kwong, R. Wyatt, J. Robinson, R. W. Sweet, J. Sodroski, and W. A. Hendrickson. Structure of an HIV gp120 envelope glycoprotein in complex with the CD4 receptor and a neutralizing human antibody. *Nature*, 393(6686):648–59, June 1998. 47, 48, 49, 50

[90] J. Kyte and R. F. Doolittle. A simple method for displaying the hydropathic character of a protein. *Journal of molecular biology*, 157(1):105–32, May 1982. 30, 90

[91] J. M. LaLonde, Y. D. Kwon, D. M. Jones, A. W. Sun, J. R. Courter, T. Soeta, T. Kobayashi, A. M. Princiotto, X. Wu, A. Schön, E. Freire, P. D. Kwong, J. R. Mascola, J. Sodroski, N. Madani, and A. B. Smith. Structure-based design, synthesis, and characterization of dual hotspot small-molecule HIV-1 entry inhibitors. *Journal of medicinal chemistry*, 55(9):4382–96, May 2012. 49

[92] S. M. Larson, A. A. Di Nardo, and A. R. Davidson. Analysis of covariation in an SH3 domain sequence alignment: applications in tertiary contact prediction and the design of compensating hydrophobic core substitutions. *Journal of molecular biology*, 303(3):433–46, October 2000. 54

[93] K. F. Lau and K. A. Dill. A lattice statistical mechanics model of the conformational and sequence spaces of proteins. *Macromolecules*, 22(10):3986–3997, October 1989. 10, 11, 12, 13, 14

[94] S. M. Lee. *Goal Programming for Decision Analysis*. Auerbach Publishers, Philadelphia, 1972. 7

[95] N. Lesh, M. Mitzenmacher, and S. Whitesides. A complete and effective move set for simplified protein folding. In *Proceedings of the seventh annual international conference on Computational molecular biology - RECOMB '03*, pages 188–195. ACM Press, New York, New York, USA, 2003. 20

[96] A. M. Lesk. *Introduction to Protein Architecture: The Structural Biology of Proteins*. Oxford University Press, Oxford, 2000. 11

[97] D. Li and R. Roberts. WD-repeat proteins: structure characteristics, biological function, and their involvement in human diseases. *Cellular and molecular life sciences : CMLS*, 58(14):2085–97, December 2001. 68

[98] H.-X. Liao, M. Bonsignori, S. M. Alam, J. S. McLellan, G. D. Tomaras, M. A. Moody, D. M. Kozink, K.-K. Hwang, X. Chen, C.-Y. Tsao, P. Liu, X. Lu, R. J. Parks, D. C. Montefiori, G. Ferrari, J. Pollara, M. Rao, K. K. Peachman, S. Santra, N. L. Letvin, N. Karasavvas, Z.-Y. Yang, K. Dai,

M. Pancera, J. Gorman, K. Wiehe, N. I. Nicely, S. Rerks-Ngarm, S. Nitayaphan, J. Kaewkungwal, P. Pitisuttithum, J. Tartaglia, F. Sinangil, J. H. Kim, N. L. Michael, T. B. Kepler, P. D. Kwong, J. R. Mascola, G. J. Nabel, A. Pinter, S. Zolla-Pazner, and B. F. Haynes. Vaccine induction of antibodies against a structurally heterogeneous site of immune pressure within HIV-1 envelope protein variable regions 1 and 2. *Immunity*, 38(1):176–86, January 2013. 49, 79

[99] J. Liu, A. Bartesaghi, M. J. Borgnia, G. Sapiro, and S. Subramaniam. Molecular architecture of native HIV-1 gp120 trimers. *Nature*, 455(7209):109–13, September 2008. 50

[100] L. Liu, R. Cimbro, P. Lusso, and E. A. Berger. Intraprotomer masking of third variable loop (V3) epitopes by the first and second variable loops (V1V2) within the native HIV-1 envelope glycoprotein trimer. *Proceedings of the National Academy of Sciences of the United States of America*, 108(50):20148–53, December 2011. 50, 74, 79

[101] H. Lu and G. Yen. Rank-density-based multiobjective genetic algorithm and benchmark test function study. *IEEE Transactions on Evolutionary Computation*, 7(4):325–343, August 2003. 17

[102] N. Madani, A. Schön, A. M. Princiotto, J. M. Lalonde, J. R. Courter, T. Soeta, D. Ng, L. Wang, E. T. Brower, S.-H. Xiang, Y. D. Kwon, C.-C. Huang, R. Wyatt, P. D. Kwong, E. Freire, A. B. Smith, and J. Sodroski. Small-molecule CD4 mimics interact with a highly conserved pocket on HIV-1 gp120. *Structure (London, England : 1993)*, 16(11):1689–701, November 2008. 49

[103] N. Madras and A. D. Sokal. The pivot algorithm: A highly efficient Monte Carlo method for the self-avoiding walk. *Journal of Statistical Physics*, 50(1-2):109–186, January 1988. 20

[104] W. Mandecki. The game of chess and searches in protein sequence space. *Trends in Biotechnology*, 16(5):200–202, December 1998. 2

[105] M. Mann. *Computational Methods for Lattice Protein Models*. Ph.D. thesis, University of Freiburg, 2011. 13

[106] M. Mann, D. Maticzka, R. Saunders, and R. Backofen. Classifying proteinlike sequences in arbitrary lattice protein models using LatPack. *HFSP journal*, 2(6):396–404, December 2008. 20, 22

[107] Y. Mao, L. Wang, C. Gu, A. Herschhorn, S.-H. Xiang, H. Haim, X. Yang, and J. Sodroski. Subunit organization of the membrane-bound HIV-1 envelope glycoprotein trimer. *Nature structural & molecular biology*, 19(9):893–9, September 2012. 79

[108] D. S. Marks, L. J. Colwell, R. Sheridan, T. A. Hopf, A. Pagnani, R. Zecchina, and C. Sander. Protein 3D structure computed from evolutionary sequence variation. *PloS one*, 6(12):e28766, January 2011. 54, 71, 86

[109] L. C. Martin, G. B. Gloor, S. D. Dunn, and L. M. Wahl. Using information theory to search for co-evolving residues in proteins. *Bioinformatics (Oxford, England)*, 21(22):4116–24, November 2005. 54

[110] J. S. McLellan, M. Pancera, C. Carrico, J. Gorman, J.-P. Julien, R. Khayat, R. Louder, R. Pejchal, M. Sastry, K. Dai, S. O'Dell, N. Patel, S. Shahzad-ul Hussan, Y. Yang, B. Zhang, T. Zhou, J. Zhu, J. C. Boyington, G.-Y. Chuang, D. Diwanji, I. Georgiev, Y. D. Kwon, D. Lee, M. K. Louder, S. Moquin, S. D. Schmidt, Z.-Y. Yang, M. Bonsignori, J. A. Crump, S. H. Kapiga, N. E. Sam, B. F. Haynes, D. R. Burton, W. C. Koff, L. M. Walker, S. Phogat, R. Wyatt, J. Orwenyo, L.-X. Wang, J. Arthos, C. A. Bewley, J. R. Mascola, G. J. Nabel, W. R. Schief, A. B. Ward, I. A. Wilson, and P. D. Kwong. Structure of HIV-1 gp120 V1/V2 domain with broadly neutralizing antibody PG9. *Nature*, 480(7377):336–43, December 2011. 49, 79

[111] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics*, 21(6):1087, 1953. 20

[112] K. Miettinen and M. Mäkelä. Interactive bundle-based method for nondifferentiable multiobjeective optimization: nimbus . *Optimization*, 34(3):231–246, January 1995. 7

[113] S. Miyazawa and R. L. Jernigan. Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules*, 18(3):534–552, May 1985. 14, 89

[114] S. Miyazawa and R. L. Jernigan. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *Journal of molecular biology*, 256(3):623–44, March 1996. 14

[115] F. Morcos, A. Pagnani, B. Lunt, A. Bertolino, D. S. Marks, C. Sander, R. Zecchina, J. N. Onuchic, T. Hwa, and M. Weigt. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences of the United States of America*, 108(49):E1293–301, December 2011. 54, 55, 57, 58, 59, 60, 62, 63, 64, 67, 68, 69, 83, 86

[116] L. Morellato-Castillo, P. Acharya, O. Combes, J. Michiels, A. Descours, O. H. P. Ramos, Y. Yang, G. Vanham, K. K. Ariën, P. D. Kwong, L. Martin, and P. Kessler. Interfacial Cavity Filling To Optimize CD4-Mimetic Miniprotein Interactions with HIV-1 Surface Glycoprotein. *Journal of medicinal chemistry*, 56(12):5033–47, June 2013. 49

[117] National Institute of Allergy and Infectious Diseases. HIV replication cycle. 2012. 46

[118] National Institute of Allergy and Infectious Diseases. HIV virion. 2012. 45

[119] C. Notredame and D. G. Higgins. SAGA: sequence alignment by genetic algorithm. *Nucleic acids research*, 24(8):1515–24, April 1996. 14

[120] O. Olmea, B. Rost, and A. Valencia. Effective use of sequence correlation and conservation in fold recognition. *Journal of molecular biology*, 293(5):1221–39, November 1999. 54

[121] W. C. Olson, G. E. Rabut, K. A. Nagashima, D. N. Tran, D. J. Anselma, S. P. Monard, J. P. Segal, D. A. Thompson, F. Kajumo, Y. Guo, J. P. Moore, P. J. Maddon, and T. Dragic. Differential inhibition of human immunodeficiency virus type 1 fusion, gp120 binding, and CC-chemokine activity by monoclonal antibodies to CCR5. *Journal of virology*, 73(5):4145–55, May 1999. 74

[122] M. Pancera, S. Majeed, Y.-E. A. Ban, L. Chen, C.-c. Huang, L. Kong, Y. D. Kwon, J. Stuckey, T. Zhou, J. E. Robinson, W. R. Schief, J. Sodroski, R. Wyatt, and P. D. Kwong. Structure of HIV-1 gp120 with gp41-interactive region reveals layered envelope architecture and basis of conformational mobility. *Proceedings of the National Academy of Sciences of the United States of America*, 107(3):1166–71, January 2010. 49, 75, 76, 77, 78, 82

[123] J. T. Pedersen and J. Moult. Genetic algorithms for protein structure prediction. *Current opinion in structural biology*, 6(2):227–31, April 1996. 2

[124] R. Pejchal, K. J. Doores, L. M. Walker, R. Khayat, P.-S. Huang, S.-K. Wang, R. L. Stanfield, J.-P. Julien, A. Ramos, M. Crispin, R. Depetris, U. Katpally, A. Marozsan, A. Cupo, S. Maloveste, Y. Liu, R. McBride, Y. Ito, R. W. Sanders, C. Ogohara, J. C. Paulson, T. Feizi, C. N. Scanlan, C.-H. Wong, J. P. Moore, W. C. Olson, A. B. Ward, P. Poignard, W. R. Schief, D. R. Burton, and I. A. Wilson. A potent and broad neutralizing antibody recognizes and penetrates the HIV glycan shield. *Science (New York, N.Y.)*, 334(6059):1097–103, November 2011. 49

[125] E. J. Platt, S. E. Kuhmann, P. P. Rose, and D. Kabat. Adaptive mutations in the V3 loop of gp120 enhance fusogenicity of human immunodeficiency virus type 1 and enable use of a CCR5 coreceptor that lacks the amino-terminal sulfated region. *Journal of virology*, 75(24):12266–78, December 2001. 74

[126] A. Poon and L. Chao. The rate of compensatory mutation in the DNA bacteriophage phiX174. *Genetics*, 170(3):989–99, July 2005. 53

[127] S. Potzsch, G. Scheuermann, P. Stadler, M. Wolfinger, and C. Flamm. Visualization of Lattice-Based Protein Folding Simulations. In *Tenth International Conference on Information Visualisation (IV'06)*, pages 89–94. IEEE, 2006. 11

[128] A. Procaccini, B. Lunt, H. Szurmant, T. Hwa, and M. Weigt. Dissecting the specificity of protein-protein interaction in bacterial two-component signaling: orphans and crosstalks. *PloS one*, 6(5):e19729, January 2011. 59

[129] M. Punta, P. C. Coggill, R. Y. Eberhardt, J. Mistry, J. Tate, C. Boursnell, N. Pang, K. Forslund, G. Ceric, J. Clements, A. Heger, L. Holm, E. L. L. Sonnhammer, S. R. Eddy, A. Bateman, and R. D. Finn. The Pfam protein families database. *Nucleic acids research*, 40(Database issue):D290–301, January 2012. 55

[130] G. Raghunathan and R. L. Jernigan. Ideal architecture of residue packing and its observation in protein structures. *Protein science : a publication of the Protein Society*, 6(10):2072–83, October 1997. 12

[131] C. D. Rizzuto, R. Wyatt, N. Hernández-Ramos, Y. Sun, P. D. Kwong, W. A. Hendrickson, and J. Sodroski. A conserved HIV gp120 glycoprotein structure involved in chemokine receptor binding. *Science (New York, N.Y.)*, 280(5371):1949–53, June 1998. 47

[132] P. Rusert, A. Krarup, C. Magnus, O. F. Brandenberg, J. Weber, A.-K. Ehlert, R. R. Regoes, H. F. Günthard, and A. Trkola. Interaction of the gp120 V1V2 loop with a neighboring gp120 unit shields the HIV envelope trimer against cross-neutralizing antibodies. *The Journal of experimental medicine*, 208(7):1419–33, July 2011. 50, 75, 79

[133] A. Sali and T. L. Blundell. Comparative protein modelling by satisfaction of spatial restraints. *Journal of molecular biology*, 234(3):779–815, December 1993. 61

[134] Q. J. Sattentau. HIV gp120: double lock strategy foils host defences. *Structure*, 6(8):945–949, August 1998. 47

[135] Q. J. Sattentau and J. P. Moore. Conformational changes induced in the human immunodeficiency virus envelope glycoprotein by soluble CD4 binding. *The Journal of experimental medicine*, 174(2):407–15, August 1991. 47

[136] J. D. Schaffer. Multiple Objective Optimization with Vector Evaluated Genetic Algorithms. In *Proceedings of the 1st International Conference on Genetic Algorithms*. L. Erlbaum Associates Inc., Hillsdale, NJ, US, 1985. 14, 18

[137] E. Schnur, E. Noah, I. Ayzenshtat, H. Sargsyan, T. Inui, F.-X. Ding, B. Arshava, Y. Sagi, N. Kessler, R. Levy, T. Scherf, F. Naider, and J. Anglister. The conformation and orientation of a 27-residue CCR5 peptide in a ternary complex with HIV-1 gp120 and a CD4-mimic peptide. *Journal of molecular biology*, 410(5):778–97, July 2011. 52, 53

[138] G. E. Schulz and R. H. Schirmer. *Principles of Protein Structure*. Springer Verlag, New York, 1984. 11

[139] L. P. Scott, J. Chahine, and J. R. Ruggiero. Using genetic algorithm to design protein sequence. *Applied Mathematics and Computation*, 200(1):1–9, June 2008. 2

[140] C. E. Shannon and W. Weaver. *The Mathematical Theory of Communication*. University of Illinois Press, 1949. 56

[141] P. M. Sharp and B. H. Hahn. Origins of HIV and the AIDS pandemic. *Cold Spring Harbor perspectives in medicine*, 1(1):a006841, September 2011. 43

[142] T. F. Smith, C. Gaitatzes, K. Saxena, and E. J. Neer. The WD repeat: a common architecture for diverse functions. *Trends in biochemical sciences*, 24(5):181–5, May 1999. 68

[143] N. Srinivas and K. Deb. Multiobjective function optimization using nondominated sorting genetic algorithms. *Evolutionary computation*, 2(3):221–48, 1994. 15, 16

[144] J. I. Sulkowska, F. Morcos, M. Weigt, T. Hwa, and J. N. Onuchic. Genomics-aided structure prediction. *Proceedings of the National Academy of Sciences of the United States of America*, 109(26):10340–5, June 2012. 54, 86

[145] K. Tan, J. Liu, J. Wang, S. Shen, and M. Lu. Atomic structure of a thermostable subdomain of HIV-1 gp41. *Proceedings of the National Academy of Sciences of the United States of America*, 94(23):12303–8, November 1997. 50

[146] Q. Tan, Y. Zhu, J. Li, Z. Chen, G. W. Han, I. Kufareva, T. Li, L. Ma, G. Fenalti, J. Li, W. Zhang, X. Xie, H. Yang, H. Jiang, V. Cherezov, H. Liu, R. C. Stevens, Q. Zhao, and B. Wu. Structure of the CCR5 Chemokine Receptor-HIV Entry Inhibitor Maraviroc Complex. *Science (New York, N.Y.)*, September 2013. 86

[147] The UniProt Consortium. Update on activities at the Universal Protein Resource (UniProt) in 2013. *Nucleic acids research*, 41(Database issue):D43–7, January 2013. 78, 82

[148] D. A. D. Thompson, E. G. Cormier, and T. Dragic. CCR5 and CXCR4 usage by non-clade B human immunodeficiency virus type 1 primary isolates. *Journal of virology*, 76(6):3059–64, March 2002. 74

[149] C. C. To and J. Vohradsky. A parallel genetic algorithm for single class pattern classification and its application for gene expression profiling in Streptomyces coelicolor. *BMC genomics*, 8:49, January 2007. 14

[150] E. E. H. Tran, M. J. Borgnia, O. Kuybeda, D. M. Schauder, A. Bartesaghi, G. A. Frank, G. Sapiro, J. L. S. Milne, and S. Subramaniam. Structural mechanism of trimeric HIV-1 envelope glycoprotein activation. *PLoS pathogens*, 8(7):e1002797, January 2012. 50, 51, 81, 82

[151] UNAIDS. *2012 UNAIDS Global Report On the AIDS Epidemic*. UNAIDS, 2012. 44

[152] UNAIDS. HIV prevalence. 2013. 44

[153] F. H. van Batenburg, A. P. Gultyaev, and C. W. Pleij. An APL-programmed genetic algorithm for the prediction of RNA secondary structure. *Journal of theoretical biology*, 174(3):269–80, June 1995. 14

[154] M. Vendruscolo and E. Domany. Protein folding using contact maps. *Vitamins and hormones*, 58:171–212, January 2000. 53

[155] M. Vendruscolo, E. Kussell, and E. Domany. Recovery of protein structure from contact maps. *Folding & design*, 2(5):295–306, January 1997. 53

[156] Y. Wang, R. Rawi, D. Hoffmann, B. Sun, and R. Yang. Inference of global HIV-1 sequence patterns and preliminary feature analysis. *Virologica Sinica*, 28(4):228–38, August 2013. 75

[157] Y. Wang, R. Rawi, C. Wilms, D. Heider, R. Yang, and D. Hoffmann. A small set of succinct signature patterns distinguishes Chinese and non-Chinese HIV-1 genomes. *PloS one*, 8(3):e58804, January 2013. 75

[158] B. Wei, N. Han, H.-z. Liu, A. Rayner, and S. Rayner. Use of mutual information arrays to predict coevolving sites in the full length HIV gp120 protein for subtypes B and C. *Virologica Sinica*, 26(2):95–104, April 2011. 54

[159] M. Weigt, R. A. White, H. Szurmant, J. A. Hoch, and T. Hwa. Identification of direct residue contacts in protein-protein interaction by message passing. *Proceedings of the National Academy of Sciences of the United States of America*, 106(1):67–72, January 2009. 54, 57, 58, 59, 60

[160] W. Weissenhorn, A. Dessen, S. C. Harrison, J. J. Skehel, and D. C. Wiley. Atomic structure of the ectodomain from HIV-1 gp41. *Nature*, 387(6631):426–30, May 1997. 50, 80

[161] T. A. White, A. Bartesaghi, M. J. Borgnia, J. R. Meyerson, M. J. V. de la Cruz, J. W. Bess, R. Nandwani, J. A. Hoxie, J. D. Lifson, J. L. S. Milne, and S. Subramaniam. Molecular architectures of trimeric SIV and HIV-1 envelope glycoproteins on intact viruses: strain-dependent variation in quaternary structure. *PLoS pathogens*, 6(12):e1001249, January 2010. 50

[162] A. P. Wierzbicki. The use of reference objectives in multiobjective optimisation. In G. Fandel and T. Gal, editors, {*MCDM*} *theory and Application, Proceedings*, pages 468–486. Springer Verlag, Hagen, 1980. 7

[163] C. Wild, T. Greenwell, and T. Matthews. A synthetic peptide from HIV-1 gp41 is a potent inhibitor of virus-mediated cell-cell fusion. *AIDS research and human retroviruses*, 9(11):1051–3, November 1993. 50

[164] J. Wilken, D. Hoover, D. A. Thompson, P. N. Barlow, H. McSparron, L. Picard, A. Wlodawer, J. Lubkowski, and S. B. Kent. Total chemical synthesis and high-resolution crystal structure of the potent anti-HIV protein AOP-RANTES. *Chemistry & biology*, 6(1):43–51, January 1999. 8, 9

[165] K. R. Wollenberg and W. R. Atchley. Separation of phylogenetic and functional associations in biological sequences by using the parametric bootstrap. *Proceedings of the National Academy of Sciences of the United States of America*, 97(7):3288–91, March 2000. 54

[166] K.-C. Wong, C.-H. Wu, R. K. Mok, C. Peng, and Z. Zhang. Evolutionary multimodal optimization using the principle of locality. *Information Sciences*, 194:138–170, July 2012. 14

[167] B. Wu, E. Y. T. Chien, C. D. Mol, G. Fenalti, W. Liu, V. Katritch, R. Abagyan, A. Brooun, P. Wells, F. C. Bi, D. J. Hamel, P. Kuhn, T. M. Handel, V. Cherezov, and R. C. Stevens. Structures of the CXCR4 chemokine GPCR with small-molecule and cyclic peptide antagonists. *Science (New York, N.Y.)*, 330(6007):1066–71, November 2010. 86

[168] L. Wu, N. P. Gerard, R. Wyatt, H. Choe, C. Parolin, N. Ruffing, A. Borsetti, A. A. Cardoso, E. Desjardin, W. Newman, C. Gerard, and J. Sodroski. CD4-induced interaction of primary HIV-1 gp120 glycoproteins with the chemokine receptor CCR-5. *Nature*, 384(6605):179–83, November 1996. 47, 48

[169] S.-R. Wu, R. Löving, B. Lindqvist, H. Hebert, P. J. B. Koeck, M. Sjöberg, and H. Garoff. Single-particle cryoelectron microscopy analysis reveals the

HIV-1 spike as a tripod structure. *Proceedings of the National Academy of Sciences of the United States of America*, 107(44):18844–9, November 2010. 49, 50

[170] X. Wu, T. Zhou, J. Zhu, B. Zhang, I. Georgiev, C. Wang, X. Chen, N. S. Longo, M. Louder, K. McKee, S. O'Dell, S. Perfetto, S. D. Schmidt, W. Shi, L. Wu, Y. Yang, Z.-Y. Yang, Z. Yang, Z. Zhang, M. Bonsignori, J. A. Crump, S. H. Kapiga, N. E. Sam, B. F. Haynes, M. Simek, D. R. Burton, W. C. Koff, N. A. Doria-Rose, M. Connors, J. C. Mullikin, G. J. Nabel, M. Roederer, L. Shapiro, P. D. Kwong, and J. R. Mascola. Focused evolution of HIV-1 neutralizing antibodies revealed by structures and deep sequencing. *Science (New York, N.Y.)*, 333(6049):1593–602, September 2011. 49

[171] R. Wyatt, P. D. Kwong, E. Desjardins, R. W. Sweet, J. Robinson, W. A. Hendrickson, and J. G. Sodroski. The antigenic structure of the HIV gp120 envelope glycoprotein. *Nature*, 393(6686):705–11, June 1998. 47

[172] R. Wyatt, J. Moore, M. Accola, E. Desjardin, J. Robinson, and J. Sodroski. Involvement of the V1/V2 variable loop structure in the exposure of human immunodeficiency virus type 1 gp120 epitopes induced by receptor binding. *Journal of virology*, 69(9):5723–33, September 1995. 47

[173] R. Wyatt and J. Sodroski. The HIV-1 envelope glycoproteins: fusogens, antigens, and immunogens. *Science (New York, N.Y.)*, 280(5371):1884–8, June 1998. 47

[174] Y. Yang and H. Liu. Genetic algorithms for protein conformation sampling and optimization in a discrete backbone dihedral angle space. *Journal of computational chemistry*, 27(13):1593–602, October 2006. 2

[175] C. Yanofsky, V. Horn, and D. Thorpe. Protein Structure Relationships Revealed by Mutational Analysis. *Science*, 146(3651):1593–1594, December 1964. 53

[176] G. Yen and H. Lu. Dynamic multiobjective evolutionary algorithm: adaptive cell-based rank and density estimation. *IEEE Transactions on Evolutionary Computation*, 7(3):253–274, June 2003. 17

[177] B.-W. Ying, H. Taguchi, and T. Ueda. Co-translational binding of GroEL to nascent polypeptides is followed by post-translational encapsulation by GroES to mediate protein folding. *The Journal of biological chemistry*, 281(31):21813–9, August 2006. 7

[178] T. Zhou, I. Georgiev, X. Wu, Z.-Y. Yang, K. Dai, A. Finzi, Y. D. Kwon, J. F. Scheid, W. Shi, L. Xu, Y. Yang, J. Zhu, M. C. Nussenzweig, J. Sodroski, L. Shapiro, G. J. Nabel, J. R. Mascola, and P. D. Kwong. Structural basis for broad and potent neutralization of HIV-1 by antibody VRC01. *Science (New York, N.Y.)*, 329(5993):811–7, August 2010. 49

[179] T. Zhou, L. Xu, B. Dey, A. J. Hessell, D. Van Ryk, S.-H. Xiang, X. Yang, M.-Y. Zhang, M. B. Zwick, J. Arthos, D. R. Burton, D. S. Dimitrov, J. Sodroski, R. Wyatt, G. J. Nabel, and P. D. Kwong. Structural definition of a conserved neutralization epitope on HIV-1 gp120. *Nature*, 445(7129):732–7, February 2007. 49

[180] E. Zitzler, M. Laumanns, and L. Thiele. SPEA2: Improving the Strength Pareto Evolutionary Algorithm. Technical report, ETH Zurich, Zuerich, Switzerland, 2001. 14

[181] E. Zitzler and L. Thiele. Multiobjective optimization using evolutionary algorithms A comparative case study. In *Parallel Problem Solving from Nature PPSN V*, pages 292–301. Springer Berlin / Heidelberg, 1998. 17

[182] E. Zitzler and L. Thiele. Multiobjective evolutionary algorithms: a comparative case study and the strength Pareto approach. *IEEE Transactions on Evolutionary Computation*, 3(4):257–271, 1999. 14, 16

[183] M. Zvelebil and J. Baum. *Understanding Bioinformatics*. Garland Science, New York and London, 2007. 10, 61

List of Publications

Journal Articles

Y. Wang, **R. Rawi**, D. Hoffmann, B. Sun, and R. Yang. Inference of global HIV-1 sequence patterns and preliminary feature analysis. *Virologica Sinica*, 28(4):22838, 2013.

Y. Wang*, **R. Rawi***, C. Wilms, D. Heider, R. Yang, and D. Hoffmann. A small set of succinct signature patterns distinguishes Chinese and non-Chinese HIV-1 genomes. PloS one, 8(3):e58804, 2013

R. Rawi, L. Whitmore, M. Topf. CHOYCE: a web server for constrained homology modelling with cryoEM maps. *Bioinformatics*, 26(16):1673-4, 2010

Acknowledgements

> ” *When you practice gratefulness, there is a sense of respect toward others.*”
>
> Dalai Lama

My parents For your everlasting support and love. Allah I-Hfadkum.

My supervisor Daniel Hoffmann For given me the opportunity to be part of your team, but most important for your countless scientific and non-scientific ideas, guidance and help. I really enjoyed and benefited from your knowledge, company and principles. Stay as you are.

My colleagues For sharing ideas, experiences, methods, scripts and a lot of coffee. Cheers.

My love Soumia For your company, love, patience,... . I could name 1001 adjectives and they would still not be sufficient. May you be blessed with happiness and love. I am looking forward to share my life with you.

Curriculum Vitae

For reasons of confidentiality, the curriculum vitae is not included in the online version of this work.

Declarations

Erklärung:
Hiermit erkläre ich, gem. § 6 Abs. (2) f) der Promotionsordnung der Fakultäten für Biologie, Chemie und Mathematik zur Erlangung der Dr. rer. nat., dass ich das Arbeitgebiet, dem das Thema "Analysis and Application of Evolutionary Processes to tackle HIV-1 Entry" zuzuordnen ist, in Forschung und Lehre vertrete und den Antrag von Reda Rawi befürworte und die Betreuung auch im Falle eines Weggangs, wenn nicht wichtige Gründe dem entgegenstehen, weiterführen werde.

Essen, den _____ _____

Unterschrift eines Mitglieds der Universität Duisburg-Essen

Erklärung:
Hiermit erkläre ich, gem. § 7 Abs. (2) c) + e) der Promotionsordnung Fakultäten für Biologie, Chemie und Mathematik zur Erlangung des Dr. rer. nat., dass ich die vorliegende Dissertation selbständig verfasst und mich keiner anderen als der angegebenen Hilfsmittel bedient habe.

Essen, den _____ _____

Unterschrift des/r Doktoranden/in

Erklärung:
Hiermit erkläre ich, gem. § 7 Abs. (2) d) + f) der Promotionsordnung der Fakultäten für Biologie, Chemie und Mathematik zur Erlangung des Dr. rer. nat., dass ich keine anderen Promotionen bzw. Promotionsversuche in der Vergangenheit durchgeführt habe und dass diese Arbeit von keiner anderen Fakultät/Fachbereich abgelehnt worden ist.

Essen, den _____ _____

Unterschrift des Doktoranden