

IN SILICO DESIGN OF NOVEL BINDING LIGANDS
FOR BIOLOGICAL TARGETS

A Thesis
Presented to
The Academic Faculty

by

C. Denise Enekwa

In Partial Fulfillment
of the Requirements for the Degree
Master of Science in the
School of Chemistry and Biochemistry

Georgia Institute of Technology
August 2010

IN SILICO DESIGN OF NOVEL BINDING LIGANDS
FOR BIOLOGICAL TARGETS

Approved by:

Professor Rigoberto Hernandez, Advisor
School of Chemistry and Biochemistry
Georgia Institute of Technology

Professor Loren Williams
School of Chemistry and Biochemistry
Georgia Institute of Technology

Professor Steve Harvey
School of Biology
Georgia Institute of Technology

Professor Wendy L. Kelly
School of Chemistry and Biochemistry
Georgia Institute of Technology

Date Approved: 3 May 2010

ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. Rigoberto Hernandez, for his support and guidance during these past few years. I have learned a vast amount of information about computational chemistry since I first began as a graduate student, most of which is due to his patience and willingness to teach me. I am forever grateful because this knowledge will always remain with me.

I would also like to thank the members of the Hernandez research group for their help and for many good times. For scientific development, for moral support, and for friendship, I thank Megan Damm, Ashley Tucker, Gungor Ozer, Dr. Alex Popov, Dr. Ashley Ringer, Michael Marshall, and Karen Goeders. I owe special thanks to Dr. Cam Tyson for his support and much appreciated guidance.

I would like to thank Dr. Loren Williams for all his encouragement, and his devotion to my success in graduate school and as a scientist. I thank my other committee members, Dr. Steve Harvey and Dr. Wendy Kelly, for their support, and willingness to take the time to answer questions and offer advice on multiple occasions.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS	iii
LIST OF TABLES	vi
LIST OF FIGURES	vii
LIST OF SYMBOLS OR ABBREVIATIONS	viii
SUMMARY	ix
I INTRODUCTION	1
II THEORETICAL AND PHYSICAL CHEMISTRY FRAMEWORK	4
2.1 Introduction	4
2.2 Molecular Recognition and Docking	5
2.2.1 eHiTS: Electronic High-Throughput Screening	7
2.2.2 AutoDock	8
2.3 Theory	9
2.4 Peptide Conformation	11
2.5 <i>In Silico</i> Design Algorithm	12
2.6 Summary	15
III METHODS AND IMPLEMENTATION	16
3.1 Introduction	16
3.2 Algorithm Details	16
3.3 SubMVL Size	17
3.4 Protein Docking	19
3.5 Pose Ranking	19
3.6 Summary	20
IV RESULTS AND DISCUSSION	21
4.1 Introduction	21
4.2 Pruning Mechanism	21

4.3	SubMVL Size	22
4.4	Pose Ranking	23
4.5	Binding Pocket	24
4.6	Ligand Design with Residue Classes	25
4.7	Discussion	30
V	CONCLUSIONS	34
5.1	Main Conclusions	34
5.1.1	Final Library and Docking Program Invariance	34
5.1.2	Automation	35
5.2	Outlook and Future Work	36
APPENDIX A	DOCKING PARAMETERS	37
REFERENCES	39

LIST OF TABLES

1	MVL size statistics and computation speed details for key rounds. Percent coverage for each of the rounds are shown, and convergence was claimed after 30 iterations. Reprinted from article [1]	18
2	Definition of an 8-res MVL showing residues allowed at each site. One-letter abbreviations for amino acids are used	18
3	Residue classes implemented in MKPEPS; philic = hydrophilic, phobic = hydrophobic, pos = positive, neg = negative	26
4	MVL used for class analysis	27
5	Comparison of weights and entropies for classes Polar and Small, where Polar high weights and low entropies illustrate presence of two dominant residues in the class	28
6	Defined MVLs before and after positive (R,K) residue study, R and K were subsequently removed from all sites but 1 and 2	29
7	Newly defined MVL after class analysis	29
8	Comparison of final predicted libraries using the the different docking programs	29
9	Proportion of good peptides in MVLs from selected rounds	31
10	AutoDock libraries: round 30 library is the converged library obtained by even pruning of the original MVL, round 16 library is obtained by predicting the highest-quality converged library, based on the 200 highest-scoring peptides from round 16	34

LIST OF FIGURES

1	A flowchart of the <i>in silico</i> algorithm for designing the sequence of a peptide that binds to a specified target. Reprinted from article [1] . . .	13
2	Ordered NBIE score versus fraction of the entire library for the two different starting subMVL sizes	22
3	Comparison of the pruning factor $D_i(s)$ for using different subMVL sizes	23
4	CHARMM NBIE scores of selected rounds. Reprinted from article [1]	24
5	DrugScore scores of selected rounds. Docking was performed with AutoDock, round 19 was final round, most favorable peptides have lowest scores.	24
6	SPR analysis of the top scoring peptide of screening rounds 5 ($k_a = 1.2 \times 10^5 M^{-1}s^{-1}$, $k_d = 1.8 \times 10^{-2} s^{-1}$, $K_a = 6.7 \times 10^6 M^{-1}$), 15 ($k_a = 4.3 \times 10^5 M^{-1}s^{-1}$, $k_d = 6.7 \times 10^{-2} s^{-1}$, $K_a = 6.4 \times 10^6 M^{-1}$), and 30 ($k_a = 6.2 \times 10^5 M^{-1}s^{-1}$, $k_d = 1.1 \times 10^{-2} s^{-1}$, $K_a = 5.9 \times 10^7 M^{-1}$) Reprinted from [1]	25
7	Pruning factor ($D_i(s)$) of the 5 highest scoring residues at each site. The two lowest residues at each site are R and K	30
8	NBIE scores of top 200 residues in MVLs where residues R and K have been removed from all other sites and are included only at the indicated site	31
9	Rounds 1, 5 and 10. NBIE scores of MVL improve (<i>i.e.</i> more negative NBIE values) with increasing rounds	32
10	Iterations following class analysis. After round 16, further pruning does not result in better NBIE scores	33

LIST OF SYMBOLS OR ABBREVIATIONS

DNase I	Deoxyribonuclease I.
eHiTS	Electronic High Throughput Screening.
MVL	Massive Virtual Library.
NBIE	Non-Bonded Interactions Energy.
PBP	Partial Binding Probability.
PDB	Protein DataBank.
PMF	Potential of Mean Force.
RMSD	Root Mean Square Deviation.
subMVL	(sub) Massive Virtual Library.

SUMMARY

This thesis presents an *in silico* algorithm that screens massive virtual libraries for peptides that will bind selectively to a given protein target. Recognition is essential for many functional properties of proteins, and is also employed in many different applications. For this reason, the ability to obtain high affinity (nanomolar or lower dissociation constants) binding proteins to specified targets is important. Extensive research goes into generating these binding reagents, either using immunoglobulin, or developing selection rules to identify binding ligands from synthetic libraries.

Existing approaches for generating binding proteins are generally characterized by high throughput screening. It has an inherent limitation: the binding solutions that can be obtained are limited to those that are included in the library being searched. Since this size of this library is necessarily limited by a finite address space (whether constructed experimentally or *in silico*), a ligand might emerge, but there is always the possibility that a better ligand exists that was not defined in the library. The solution to this would be a method which allows implicit access to all of the space available to a peptide binding sequence, as well as effective search techniques to identify high affinity ligands.

Employing principles from protein design and statistical mechanics, a search method is described that defines a virtual library that accesses every possible binding sequence. The binding energy of a candidate ligand is used to ascribe it a probability of binding. A sample of the virtual library of candidate ligands is then used to ascribe implicit weights to all the ligands in the library. These weights are then used to obtain virtual sublibraries which collectively carry a greater probability to bind to

the target.

Several underlying physical assumptions made with this method have been previously validated by the successful identification of small (≈ 8 residue) binding sequences. In this thesis, these assumptions are presented in detail and investigated. Results are presented that demonstrate how different components of the search algorithm affect the binding solutions obtained. The results provide understanding which makes it clear how we can optimize the method for the successful design of high affinity ligands.

CHAPTER I

INTRODUCTION

De novo protein design is the identification of novel protein sequences that will fold into predetermined three-dimensional structures. Experimental protein design techniques (e.g. directed evolution) [2] are limited by the number of sequences that can be screened in a given time, compared to the total available sequence space [3, 4]. A method to overcome this shortfall is computational protein design. *In silico* design methods provide access to larger numbers of sequences than can ever be achieved experimentally [3]. Combinatorial protein libraries are generated [5, 6] computationally and searched.

Computational techniques have proven successful in the design of proteins ‘from scratch’ [7, 8, 9, 10, 11, 12, 13]. While sequences selected evolutionarily exclude non-functional proteins by default, extra care has to be employed with search algorithms for protein design to ensure that identified sequences are active with respect to the desired function [14, 15]. Success in the computational design of functional proteins has been demonstrated in a few instances [16, 15]. With the reported accomplishments using different computational methods to search *in silico* combinatorial libraries for novel proteins with specific properties, it is worthwhile to develop algorithms to achieve these same goals, with even more particular functions in mind. In this work, a method is described to identify novel ligand sequences for protein recognition.

Recognition is a characteristic of biomolecules that is essential for their function. This is seen, for example, in enzyme-inhibitor, receptor-ligand, and antibody-antigen interactions. Recognition is exploited in a wide range of studies and applications.

For instance, immunochemistry (antibody-antigen recognition) is used in several laboratory applications where biomolecular recognition is required e.g. enzyme-linked immunosorbent assays, flow cytometry and immunochemical localization. These immunoglobulins (monoclonal and polyclonal) however have inherent limitations in their use. Some of these limitations include difficulty and costs in manufacturing, and the immunoglobulin (Ig) structure itself [17]. Antibody engineering has been used extensively to study antibody fragments (e.g Fab, Fv, ScFV, Fc) for their therapeutic potential, particularly as an alternative to the standard Ig [18, 19]. Camelid [20] and avian antibodies [21] have also been explored for these purposes.

In addition to the previously described systems, alternative scaffolds have also emerged as a replacement for antibodies. These scaffolds are so named because they are protein frameworks which, by sequence insertion and combinatorial methods, can accommodate novel binding activity [22, 23]. Alternative scaffold molecules have found wide application in many areas where Ig-based structures are limited, including diagnostic and therapeutic uses [24, 25, 26], intracellular applications [27, 28], co-crystallization [29], and chromatography [30, 31].

Common techniques for screening libraries to identify high-affinity binding reagents (such as alternative scaffolds) include phage display (using bacteriophage protein coat fusions) [32], *in vivo* methods based on intracellular expression [33], and *in vitro* methods using ribosomal display [34]. This process is similar to protein design in that the biggest limitation is found in the diversity of the library that can be constructed, and the maximal size that can be searched. Several methods of diversifying combinatorial libraries [35, 36, 37, 38] have been attempted. An achievement of a complexity of 10^{15} at best [39] is known, with the more commonly achievable sizes being between 10^8 and 10^{12} [22, 40]. This work presented herein describes an *in silico* method which circumvents the experimental limits on diversity, to address this problem.

This thesis contains five chapters, including this introductory chapter. Chapter 2

contains the theoretical and physical chemistry framework of the problem and the *in silico* search method which includes the principles of molecular recognition and the theory behind the algorithm. In Chapter 3, the implementation of the method for the design of a peptide to bind to Deoxyribonuclease I is presented. Specific components of the algorithm are considered in detail and studies performed to determine how they affect the predicted sequences. Chapter 4 presents the results of these studies as well as discussions on how the different parts can be optimized to improve the algorithm. Chapter 5 contains the conclusions, detailing the relevance of this work, as well as future considerations.

CHAPTER II

THEORETICAL AND PHYSICAL CHEMISTRY FRAMEWORK

2.1 *Introduction*

The goal of this work is the identification of low-energy binding sequences from a library of non-related peptide sequences. Methods of searching combinatorial libraries for protein design are broadly classified as stochastic and exhaustive. Stochastic methods are considered to be ‘sampling’ algorithms; they move between possible solutions on the energy landscape in a semi-random manner [41]. These methods include Monte Carlo techniques [42, 43] and genetic algorithms [44, 45], which are the two most commonly used in *de novo* protein design. The biggest advantage of these methods is that the probability of becoming trapped in local minima is reduced [46]. On the other hand, a huge disadvantage associated with stochastic methods is that there is no guarantee that solutions near the global minimum will ever be explored. Exhaustive search techniques on the other hand are designed to converge on the global minimum. This is achieved by using rejection criteria to eliminate the majority of the space without having to evaluate them individually [41]. The dead end elimination (DEE) method is one example of an exhaustive algorithm that has been largely validated [47, 48, 49, 50].

In addition to these methods, the use of statistical theories that employ certain criteria to determine the compatibility of certain sequences with desired structure is becoming increasingly popular [51, 14, 52]. Such functions are of the general form:

$$F(x_1, \dots, x_n) = \prod_{i=1}^n f_i(x_i) \quad (1)$$

where the desired structural or functional property, F , of some peptide (x_1, \dots, x_n) is

considered to be composed of the product of the individual properties of the residues in the peptide sequence, $f_i(x_i)$. An advantage of this method is that it is able to yield the composition of sequences that will satisfy given a fold, without a need for explicit enumeration of all the sequences in the library. The application of this method along with principles derived from energy landscape theory of protein folding allows access to the entire sequence space and is not limited by computational sample sizes [52].

A computational method for the identification of binding sequences faces many challenges similar to that of the protein folding problem. Therefore, several of the principles that have been successfully used for *de novo* protein design can be applied to develop an algorithm to search *in silico* peptide libraries to identify recognition sequences [53, 1]. The method described below uses an iterative approach to amplify the information from the screening of a limited *in silico* library, identifying sequence conservation in the best binding peptides. This approach incorporates protein docking as a tool for probing molecular recognition between peptide sequences and a target protein molecule.

2.2 Molecular Recognition and Docking

Molecular recognition between two molecules is characterized by the interactions between them [54, 55, 56, 57, 58, 59]. A change in the entropy and enthalpy of the system describes the affinity with which two molecules can associate non-covalently. The free energy is related to the entropy and enthalpy changes with the thermodynamic equation:

$$\Delta G^0 = \Delta H^0 - T\Delta S^0 \quad (2)$$

The chemical reaction representing a protein-ligand association can be written as:



An equilibrium association constant can be defined for this association:

$$K_a = \frac{k_1}{k_2} = \frac{[PL]}{[P][L]} \quad (4)$$

The inverse of this association constant, $K_d (= 1/K_a)$ has units of moles/liter, (M), which is a practical unit. For this reason, the K_d (equilibrium dissociation constant) is commonly used as a measure of binding affinity, where a small magnitude of K_d corresponds to a low tendency of the complex PL to dissociate. K_d is related to the standard free energy change of binding for the P-L complex, ΔG^0 , by the equation:

$$\Delta G^0 = RT \ln K_d \quad (5)$$

The best recognition sequences identified for binding will be those which possess the lowest free energy of receptor-binding, provided the binding states are accessible.

In order to obtain this free energy value, the ligand binding conformation must be determined. Protein docking is a computational method for predicting protein binding conformations (poses) [60, 61, 62]. An optimal method is used to search the conformational space of the proteins involved. Commonly used search algorithms for protein-ligand docking include molecular dynamics [63], Monte Carlo [64], and genetic algorithms [65, 66]. The difference between one docking program and another consists in the choice of search algorithm, as well as the choice of scoring function used to rank the poses. Previously published results on this work [1] used eHiTS (Electronic High Throughput Screening) [67] for docking, while all studies reported here were done using AutoDock [68].

Free energy of association allows a direct determination of K_d , (Eqn 5), so computational methods for predicting binding conformations of complexes attempt to calculate a $\Delta G_{binding}$ to rank the poses [69, 70]. Computational costs prohibit the calculation of binding free energies from *ab initio* quantum mechanical calculations [71], therefore different molecular mechanics-based methods have been used in an

attempt to obtain this binding free energy. The most successful have been Free Energy Perturbation and Thermodynamic Integration [72, 54]. Both of these are too computationally expensive to be practical for high-throughput ligand screening.

Other approximate methods use a variety of expressions (force-field based, knowledge-based, or empirical functions) to compute free energy. Reduced accuracy is a consequence of these methods, as they cannot fully take into account the flexibility of the proteins involved, and they make use of continuum solvent models. Docking programs are developed with the aim of using a scoring method that balances screening efficiency with as much accuracy as possible [73].

2.2.1 eHiTS: Electronic High-Throughput Screening

eHiTS uses a geometric algorithm to identify clefts on the receptor surface which could be binding pockets [67]. A heuristic selection algorithm then considers the ratio of depth values and volumes to decide which of these clefts is the correct binding site. The ligand is broken up into rigid fragments and connecting flexible chains, and each of these fragments is docked independently in the identified cavities. An exhaustive pose-match algorithm is used to recombine fragments, leading to thousands of pose combinations. After the local energy is optimized in the binding site (using a scoring function), the overall score of the pose set is used to decide which conformations are the best poses.

eHiTS uses a statistically derived empirical scoring function for pose-ranking. Interaction statistics, using 1420 high resolution protein-ligand complexes, were collected for atom pairs (excluding those linked by fewer than 4 covalent bonds) within 5.6Å of each other. These statistics are implemented in a scoring function. The final scoring function also includes terms for steric clash, depth value, receptor surface coverage, ligand conformational strain energy, ligand intramolecular interactions, and entropy loss due to frozen rotatable bonds [67].

eHiTS was validated using 100 protein-ligand complexes obtained from the Protein Databank (PDB). The root mean square deviation (RMSD) of the heavy atom coordinates of the best generated pose from the original PDB structure pose was calculated. The average RMSD was found to be 0.93Å [74]. However in a comparison study to assess the ability of seven different docking programs to predict binding mode and affinity of a protein-inhibitor complex, eHiTS was found to be the least accurate [75].

2.2.2 AutoDock

With AutoDock, the search problem is defined in terms of a global search and a local search. The global search refers to the identification of favorable docking sites on the receptor, corresponding to energetic ‘valleys’. The local search is the optimization of the structure in the current valley (e.g. using a torsional space search) [68]. A genetic algorithm is used to perform the global search. The local search is performed using the energy minimization method of Solis and Wets [76]. The two are used together in a hybrid referred to as the Lamarckian genetic algorithm (LGA) [68].

AutoDock uses an empirical binding free energy function to rank output poses. Using parameters originally derived from the AMBER forcefield and optimized for AutoDock docking [77], the energy term adds entropic terms to the typical functional form of molecular mechanics equations:

$$\Delta G = \Delta G_{vdw} + \Delta G_{hbond} + \Delta G_{elec} + \Delta G_{conform} + \Delta G_{tor} + \Delta G_{sol}$$

(See Appendix A for more details about AutoDock scoring function)

AutoDock is one of the most widely used docking programs in computational binding studies today [78, 79, 80]. It has been highly validated for protein-ligand docking applications [81, 82, 83, 84, 85, 86]. However, in a 2002 study to test the ability of different scoring functions to reproduce experimentally determined structures and binding affinities, six other empirical scoring methods performed better

than the AutoDock scoring function [87]. The scoring function has been updated in later versions of AutoDock, even though similar shortcomings in the scoring function have been found in both AutoDock 3.0 and AutoDock 4.0 [88].

2.3 Theory

The proposed algorithm uses methods similar to those described by Kang *et al* and Floudas *et al* [89, 7] where *de novo* structures are designed by optimizing amino acid sequences for a particular structure. Three critical hypotheses, which will emerge below, are made in the development of this theory. In identifying these sequences, we define an objective function which is to be optimized within the sequence space, with respect to the problem of interest. In this case, recognition is the challenge, and an ideal objective function is the free energy of binding, $\Delta G_{binding}$, which is a measure of binding affinity (Eqn 5).

For each sequence, $y = \{x_1, x_2, \dots, x_n\}$, on the landscape, each x_i is the identity of an amino acid at site i , with a total of n sites (or residue positions) in sequence y . We define the probability of each sequence $P(y)$ as:

$$P(y) = P(x_1, \dots, x_n) \approx \prod_{i=1}^n P_i(x_i) \tag{6}$$

where $P_i(x_i)$ is a partial (site) binding probability (PBP), referring to the probability of residue x_i occupying site i in the binding sequence.

Hypothesis 1: The probability of binding of the sequence, $P(y)$ can be written as a product of partial binding probabilities (PBPs), and we can use these PBPs to refine the space of candidates for the optimum binding peptides. This PBP is an exact probability if only one site is being considered. When multiple sites are considered, correlations are introduced which have to be accounted for. This leads to the second hypothesis:

Hypothesis 2: The correlations between residues at different positions are small in magnitude compared to the PBPs within a given virtual library. Therefore, by

maximizing PBPs, the binding probability of the entire sequence is being optimized. The relationship between the probability of binding, $P(y)$, and the free energy of binding associated with the sequence y is given by the Boltzmann formula:

$$P(y) = P(x_1, \dots, x_n) \propto \exp(-\beta \Delta G_{binding}(x_1, \dots, x_n)) \quad (7)$$

Because the value of $P_i(x_i)$ is necessarily influenced by the residues at all the other sites $j \neq i$, the PBP takes the form:

$$P_i(x_i) = \sum'_{j \neq i} \sum_{x_j} e^{-\beta \Delta G_{binding}(y)} \quad (8)$$

A probability distribution of the residues occupying the different sites in the peptide sequence is used in this theory to approximate the overall binding affinity, with the hypothesis that maximizing $P_i(x_i)$ at each site will minimize the energy of the sequence.

A docking component is incorporated into the algorithm and used to provide these probabilities in the form of a frequency distribution. A small, random sample of the total library is searched and ranked by an energy function and the lowest energy members of this collection (*i.e* the top scorers) are hypothesized to be closely related to the global minimum sequence. These top-ranked peptides are assumed to make the biggest contribution to the minimum structure, so that it is only necessary to analyze this fraction. A frequency distribution is obtained for each residue R at i using the top M scorers:

$$\text{freq}_i(R) = \frac{\sum_{m=1}^M \delta(R_i)}{M} \quad (9)$$

where

$$\delta(R_i) = \begin{cases} 1 & \text{if } x = R_k \\ 0 & \text{if } x \neq R_k \end{cases}$$

Hypothesis 3: $\text{freq}_i(R) \approx P_i(R)$ for optimal binding sequences.

For each site i in the binding sequence, $P_i(R)$ for each residue is approximated by $\text{freq}_i(R)$. By conservatively eliminating residues from different sets based on $P_i(R)$

determined in this manner, we hypothesize that the overall space of the ligands is being effectively optimized.

2.4 Peptide Conformation

The theory above describes the method to optimize primary structure. It is necessary to consider the optimization of three-dimensional structure. The conformations of peptides are specified by dihedral angles. The optimal dihedral angle choices differ between peptides, depending on their amino acid composition. A conformational search would be needed to identify the pose that correlates the best with a given sequence with respect to a given target. Because such a search would dramatically increase the complexity of the design problem, a few options for incorporating conformation specifications exist, and these are described below.

META: with the meta approach, the peptides in the subMVL are ‘linear’, with all the ψ and ϕ angles set to be 180^0 . Side chains on the peptide alternate between being above and below the backbone. Although linear peptides are not biologically practical, this conformation represents the most simplified representation of the peptide, as well as a good starting point for design cases where nothing is previously known about the binding activity of the target protein. The use of flexible docking to obtain binding poses for the peptides (described later) allows this initial specification (of linearity) to not be very crucial, as the final predicted conformation could deviate significantly from the linear starting structure.

PATTERNED: a given protein structure is ‘read’, and the dihedral angles present in this structure are used as a template to construct the subMVL peptides. This approach is useful for designs in which the binding site of some host molecule is to be mutated, because the structure of the mutated fragment is influenced by the fixed positions of the nonsubstituted residues. It also favors the use of non-flexible docking programs, if this ever becomes necessary.

RANDOMIZED: dihedral angles between each pair of residues are chosen according to frequency distributions obtained from the PDB [90]. This statistical approach allows for the consideration that the secondary structure differs with specified primary structure; the first two approaches do not allow this consideration.

In every design case, it is important to use the goal and nature of the ligand to be designed to decide what approach will be used for peptide design. This is demonstrated in the implementation of the design algorithm (Chapter 3).

2.5 *In Silico Design Algorithm*

The sequence of the algorithm is illustrated in the flowchart, Figure 1. A massive virtual library (MVL) is defined that consists of all the possible binding sequences. This library is not literally constructed, so there are no constraints on size. Specifically, the members of the library are obtained by varying sequence composition with respect to the size of a peptide ligand to be designed. For example, $20^N \approx 2 \times 10^{10}$ sequences are possible if the 20 standard amino acids are allowed at each position of an $N = 8$ residue peptide ligand. The MVL represents all of the sequence space available for the structure, and is defined in terms of ‘sets’. A set, S_i , corresponds to the possible residues that are allowed at a given site i . For the design of a ligand with n_{res} sites (i.e. $N = n_{res}$), the size, χ , of the MVL can be represented as follows:

$$\chi(MVL) = \prod_{i=1}^{n_{res}} \chi(S_i) \quad (10)$$

Within each set S_i , a probability $P_i(s)$ is associated with each residue $s \in S_i$, which describes the likelihood of residue s occupying site i . Therefore, for N residues defined in a given set (with normalized probabilities),

$$\sum_{i=1}^N P_i(s) = 1 \quad (11)$$

Initially, equal probabilities are assigned (i.e. for all $s \in S_i$, $P_i(s) = 1/\chi(S_i)$) This definition of sets and probabilities uses **Hypothesis 1** and **Hypothesis 2** (Section

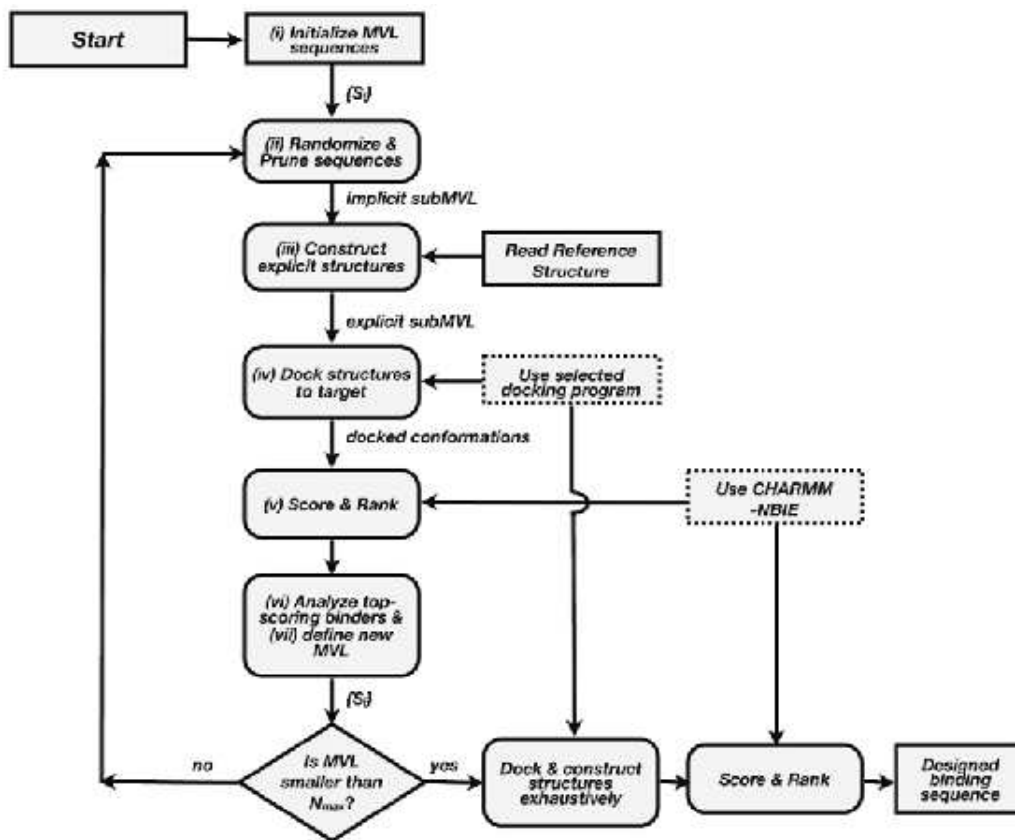


Figure 1: A flowchart of the *in silico* algorithm for designing the sequence of a peptide that binds to a specified target. Reprinted from article [1]

2.3).

A random permutation of the n_{res} sets is performed to construct a sub-library (subMVL) to be screened for high affinity binding sequences. The peptides in the subMVL are constructed *in silico*, with dihedral angles that are specified using one of the three methods described in Section 2.4. The subMVL peptides are then docked to the target using flexible protein docking methods [60, 61, 62]. Binding poses are generated for each member of the subMVL and are scored using the nonbonded interactions energy (NBIE) calculated with the CHARMM forcefield [91]. The sequences are then ranked according to this energy value, and the highest scoring sequences are analyzed further. The relative frequencies, $P_i(s)$, for the residues in this group of high scorers are obtained. Correlations in the identity of the residues at each site are also

analyzed. From this data, criterion 1 below is considered for each site i .

(Let $m = \text{MVL}$, $j = \text{set of high scorers}$ and $d = \text{subMVL}$, so $S_{i,m}$ is the set of residues at site i in the whole MVL, and $S_{i,j}$ is the set of residues at site i in the top scoring fraction, etc).

Criterion 1: At any site i , residue s is excluded from $S_{i,m}$ if $P_i(s_j)$ is less than 0.05. Residues that meet this criterion are excluded from set $S_{i,m}$ for a refinement of the MVL. If no residues at i satisfy criterion 1, criterion 2 is used to eliminate members of a set:

Criterion 2: s such that $P_{i,j}(s) = \min(P_{ij}(s)) \forall s \in S_{i,j}$.

While **Hypothesis 1** and **Hypothesis 2** allow pruning to be performed on a site-by-site basis, **Hypothesis 3** allows the use of $P_{i,j}(s)$ to decide what residues are pruned. The eliminations made in this step allow a reduction of the starting MVL, as one or more of the original sets are inevitably redefined ($\chi(S_i)_{\text{new}} \leq \chi(S_i)_{\text{old}}$). It also allows a refinement, as sequence conservation was applied to the best ligand of the subMVL and used to remove the residues that appeared least likely to make good ligands. The above steps are then iterated (Figure 1), ending with a reduction and refinement of the MVL each time, until the MVL size ($\chi(\text{MVL})$) is sufficiently small that it can be searched exhaustively to obtain the best ligands.

Two definitions of the sets, S_i , are used in the algorithm. In one case, the probability $P_i(s)$ for a residue is either 0 (if $s \ni S_i$), or equal to $1/(\chi(S_i))$ (if $s \in S_i$) This is a ‘black and white’ interpretation where at a given site i in the sequence, a residue is either included or excluded. These sets are referred to as ‘hard sets’.

Using the hard sets, there is the possibility that a residue which is compatible with a given site i in the end may have been excluded in earlier rounds. The solution to avoid this occurrence would be the assignment of nontrivial probabilities $P_i(s)$ to each residue in a given set. A method like this will allow residues to re-emerge in later rounds if they become preferred at those sites, in response to the nature of the

residues at other sites. Sets defined in this way are referred to as ‘soft sets’.

2.6 Summary

The described algorithm begins with a Massive Virtual Library (MVL) and uses information from a limited screen to refine the library for higher binding probability, following the steps shown in Figure 1. The implementation of these individual steps is described in Chapter 3. The different components are investigated separately, and together. Chapter 4 contains the results of these investigations, showing how the overall algorithm is best optimized.

CHAPTER III

METHODS AND IMPLEMENTATION

3.1 Introduction

Deoxyribonuclease I (DNase I) is an endonuclease that degrades DNA. Monomeric G-actin is a known, naturally occurring inhibitor of DNase I [92]. G-actin inhibits DNase I with subnanomolar affinity ($K_a = 5 \times 10^{-8} \text{ M}^{-1}$). This enzyme-inhibitor system was used to test the efficacy of the design algorithm described above. A goal was set to design a peptide to bind to and inhibit DNase I. In addition, as a test for the use of this method as a antibody replacement design strategy, the binding and inhibitory activity of the top-scoring peptide after insertion into an antibody replacement backbone was tested [1]. A family of 8-residue peptides was designed using 30 iterations of the algorithm. These peptides were able to successfully inhibit the activity of the enzyme. *Linum usitatissimum* trypsin inhibitor (LUTI) was used as an alternative scaffold molecule; the top-scoring peptide inserted into LUTI was also able to inhibit DNase I activity. These results are reported in Quirk *et al*, 2009 [1]. Further analyses have been performed on the results in a series of studies aimed at validating the design algorithm and the underlying biophysical considerations. These different questions being addressed and accompanying studies are presented below. The results are given in Chapter 4.

3.2 Algorithm Details

The MKPEPS package allows the definition of the MVL, as well as the construction of the subMVL. The previously described three protein construction methods (META, PATTERNED and RANDOMIZED) are implemented in MKPEPS. This

Fortran/Perl/Csh package has the ability to generate a virtual peptide library (from 2 to 100 amino acids in length), according to a torsional geometry specified by the user. Because the peptide sequence to be designed was only 8 residues long, the META approach was an adequate choice for secondary structure incorporation. In all the described studies, this approach was employed. The PDB structures of the 20 known amino acids are used to define the relative positions of the atoms within each residue. These initial linear structures were relaxed with a short molecular dynamics propagation to reduce unfavorable van der Waals contacts.

Docking was performed using the flexible docking program AutoDock [68]. The use of flexible docking allows the initial specification of peptide conformation to be trivial, as the final pose may deviate significantly from the input conformation. For this reason, it is not detrimental to use the META approach.

AutoDock was selected because it is a tool that has been validated for protein-ligand docking [82, 83, 84, 85, 86]¹ (More details about the docking simulations can be found in Appendix A). In all the docking runs, no initial constraints are placed on the designed ligand. The defined MVLs allow all 20 amino acids to vary at all 8 positions of the peptide. Similarly, no binding site on the molecule was specified in protein docking, neither were any restrictions placed on the search area.

3.3 SubMVL Size

In deciding the size of the subMVL in any given iteration, there is a trade-off between the total percent of sequence space that can be covered and the total screening time that is necessary for the chosen size. This in turn affects the number of rounds required to reach convergence. This point is demonstrated in Table 1 from Ref [1]. 30 iterations of the algorithm were needed to converge to a library size of 432 peptides. Convergence could have been claimed in fewer or more rounds. The choice of $\chi(\text{subMVL})$ at each

¹Previous studies (Quirk *et al*) used eHiTS for docking, this is discussed later.

Table 1: MVL size statistics and computation speed details for key rounds. Percent coverage for each of the rounds are shown, and convergence was claimed after 30 iterations. Reprinted from article [1]

	Screening Round			
	1	5	15	30
χ (MVL)	2.0×10^{10}	5.0×10^7	6.0×10^4	432
SubMVL Size	2.0×10^5	1.0×10^5	1.0×10^4	432
Percent Coverage	$1.0 \times 10^{-3} \%$	0.2 %	16.5 %	100 %
Screen Speed (days)	8.6	4.3	0.9	0.2

Table 2: Definition of an 8-res MVL showing residues allowed at each site. One-letter abbreviations for amino acids are used

S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8
ARNC	ARNCQ	ARNC	ARCQ	ARNQ	ARNC	ARNC	ARNC
QGHI	GHIL	GHIL	GIK	GHIK	GHIL	QHIL	QGHIL
LKPS	KMFP	KFPS	FPS	MFPS	KFPS	KMFP	KMPS
TWYV	STYV	TWYV	TWYV	TWYV	TYV	TWV	TWYV

round and the total number of rounds depends on the preferences of the user.

The larger the subMVL, the bigger the portion of the energy landscape being searched. The advantage that follows with large subMVLs is higher certainty that the prunes at each iteration are eliminating the worst members of the set. However, because of the very iterative nature of the algorithm, in combination with the conservative pruning mechanism, it is not unreasonable to expect that almost any sample size could provide qualitative information about the preferred characteristics of the best ligands.

Parallel runs were performed to test this claim. For a given MVL defined in Table 2, two subMVL sizes: $\chi = 100,000$ (0.0027% coverage) and $\chi = 10,000$ ($2.7 \times 10^{-4}\%$ coverage) were used for an iteration of the algorithm. Results from these parallel runs were compared (presented in Chapter 4).

3.4 Protein Docking

Scores from the docked poses are used in the generation of the PMFs, eventually used to refine the MVL. With respect to the choice of docking program, it is hypothesized that the predicted characteristics (*i.e.* residue identities at different sites) of the best poses predicted will be similar. And as the subMVL scoring is performed using an external energy function (CHARMM forcefield in this case), the ordering of these similar poses will be invariant to the choice of docking program.

This directly depends on the choice of docking method, and its applicability to the problem. Irrespective of the nature of the algorithm employed by the given docking program, it is expected that the ones that have been optimized and validated for protein-ligand docking will yield similar definitions of the final MVL. This is tested by comparing AutoDock generated results to those obtained using the program eHiTS (electronic High Throughput Screening) [67].

3.5 Pose Ranking

The scoring method is used to rank the subMVL peptides, and the use of the CHARMM function (Eqn 12) provides the energy of interaction for the protein-ligand complexes.

$$\begin{aligned}
 U(\vec{R}) = & \sum_{bonds} K_b(b - b_o)^2 + \sum_{angles} K_\theta(\theta - \theta_o)^2 \\
 & + \sum_{torsions} K_\phi(1 + \cos(n\phi - \delta)) \\
 & + \sum_{impropers} K_\varphi(\varphi - \varphi_0)^2 + \sum_{UB} K_{UB}(r_{1,3} - r_{1,3,o})^2 \\
 & + \sum_{nonbonded} \frac{q_i q_j}{4\pi D r_{ij}} + \varepsilon_{ij} \left[\left(\frac{R_{min,ij}}{r_{ij}} \right)^{12} - 2 \left(\frac{R_{min,ij}}{r_{ij}} \right)^6 \right]
 \end{aligned} \tag{12}$$

These energy values also provide information about statistical probabilities. These are implemented in Eqns 6 to 9 above, where the peptides that form the low-energy

complexes are taken to have a higher probability of resembling the global minimum. However, another important question arises: can other non-energy based scoring functions also be used to rank the subMVL poses, eventually leading to successful refinement of the MVL? The scores from non-energy based functions are either potentials of mean force (knowledge based scoring functions), or binding affinity scores (empirical scoring functions) [93, 94]. Can non-energetic measures of binding affinity be applied probabilistically? To address this question, DrugScore, a knowledge-based scoring function [95] was used to rank the poses. Comparing the results in Chapter 4 obtained using DrugScore with those using the CHARMM NBIE scoring method would help to decide if knowledge-based scoring functions can rank the peptide poses as well as force-field based functions.

3.6 Summary

In designing a family of peptides to bind to DNase I, a desired goal was to be able to optimize the different algorithmic components for accuracy and efficiency. The study of the subMVL sizes (Section 3.3) allow us to decide if we need large subMVLs, or if smaller sizes could be used, potentially allowing the iterations to take less time. Comparing the results of different docking programs (Section 3.4) will help to determine whether or not any selected docking method is applicable to this ligand design algorithm. This also applies to the pose ranking method: it can be determined whether or not any chosen scoring function can be used to accurately rank different binding poses (Section 3.5). The results of these studies are presented in Chapter 4.

RESULTS AND DISCUSSION

4.1 Introduction

Results of the studies using the test case of designing an 8-residue ligand for DNase I are reported. By not specifying a binding site for docking or placing constraints on the definition of the binding sequence, it was also possible to test the ability of the algorithm to identify a ligand for a system with no previously known high-affinity binding partner(s).

Results reported by Quirk *et al* [1] used the PATTERNED approach for ligand design. This was a useful approach because the peptides were designed to be inserted into an alternative scaffold. SubMVL peptides were patterned after the binding loop in the LUTI scaffold.

4.2 Pruning Mechanism

According to Criterion 2 in Section 2.5, a residue s at site i is pruned if the frequency of s in the top-scoring set $P_{i,j}(s)$ (recall j = set of high scorers) is lower than the frequencies of all other residues. A modification of this criterion is applied in these results. The frequency $P_{i,j}(s)$ is compared to the frequency of s in the entire subMVL $P_{i,d}(s)$ (recall d = subMVL), and the difference $D_i(s) = P_{i,d}(s) - P_{i,j}(s)$ is obtained. A positive value of $D_i(s)$ (also known as the pruning factor) indicates that residue s is preferred less at site i compared to the other residues in the set. When this pruning factor is obtained for all residues at all sites and they are ordered, the highest positive values represent the residues which are least favored at the given sites, while the lowest values (*i.e.* most negative) represent those that are most favored at the

indicated sites.

Two methods could be used to prune based on this description:

1. Even pruning: the objective with this method would be to reduce the size of all the sets evenly. The worst residue at each site is removed, then the next worst is removed iteratively until the MVL size has been reduced to the desired size.
2. Worst-to-best pruning: with this method, every value of $D_i(s)$ is considered individually and residues are pruned accordingly, even if multiple residues are pruned at a single site before one residue has been removed from every site.

The even pruning method is used in all these studies, allowing the sets to remain roughly the same size over the iterations.

4.3 *SubMVL Size*

The results of using two different starting subMVL sizes are shown in Figure 2.

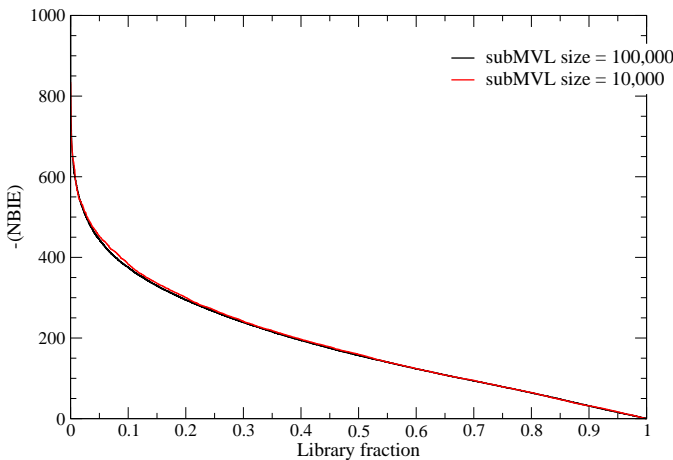


Figure 2: Ordered NBIE score versus fraction of the entire library for the two different starting subMVL sizes

Figure 2 displays the NBIE score trends as a function of the fraction of the entire library. It can be seen that the larger subMVL produces the smoother curve, which

is to be expected. Because one subMVL is larger by a factor of 10, there is increased confidence in pruning from this library over pruning from the smaller one. This is further demonstrated in Figure 3 which shows $D_i(s)$ (recall from Section 4.2) for all the residues in the MVL. Recalling that larger $D_i(s)$ values correspond to residues more eligible for pruning than others, the bigger subMVL appears to show more discrimination.

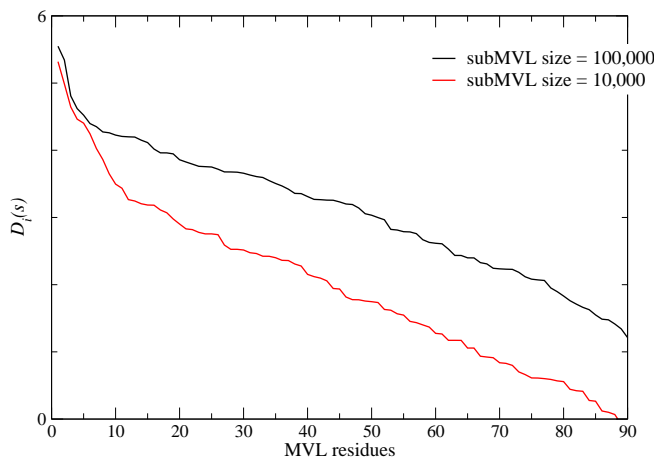


Figure 3: Comparison of the pruning factor $D_i(s)$ for using different subMVL sizes

The two libraries predict different pruning choices and subsequently different next generation MVLs. It is not obvious whether the two different sizes would eventually converge onto the same final library, even if the smaller subMVL-library is pruned with extra conservation. Therefore the results indicate that it is best to, as often as possible, maximize the size of the subMVL with respect to all other factors (time, available computational resources, etc).

4.4 *Pose Ranking*

Figure 5 shows the subMVL ordering predicted for selected rounds using the DrugScore function. This can be compared to Figure 4 which shows CHARMM-scored iterations.

As can be seen, an overall increase in the quality of ligands is predicted with increasing iterations of the algorithm.

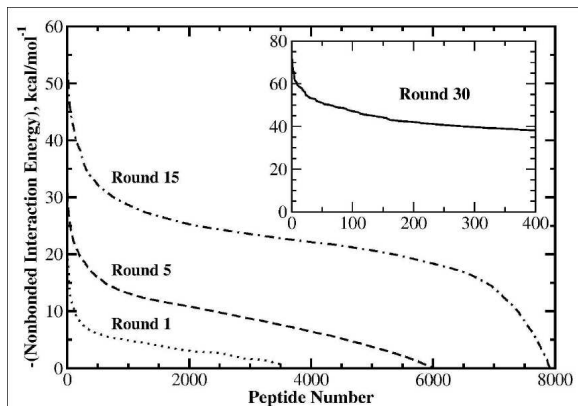


Figure 4: CHARMM NBIE scores of selected rounds. Reprinted from article [1]

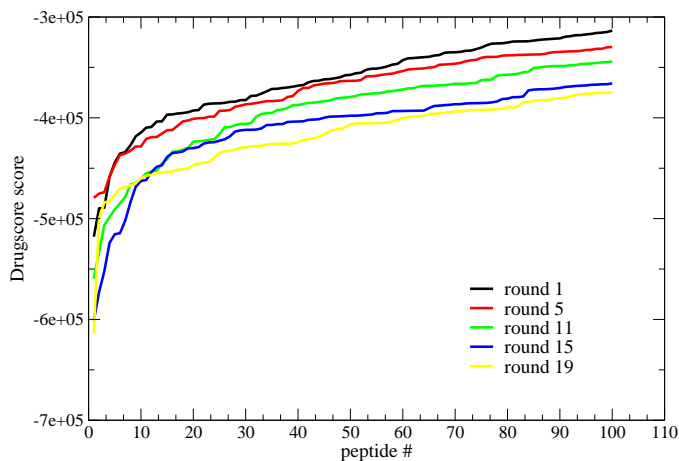


Figure 5: DrugScore scores of selected rounds. Docking was performed with AutoDock, round 19 was final round, most favorable peptides have lowest scores.

4.5 Binding Pocket

In the design of a ligand for DNase I, no binding site on the molecule was specified in protein docking. The previous studies with the algorithm [1] show a positive correlation of the experimental inhibition assay results with increasing iteration (Figure

6). This indicates that the DNase I active site crevice was very likely the preferred binding site identified for the highest scoring peptides. This is a validation for the use of the CHARMM NBIE values to order the subMVL peptides.

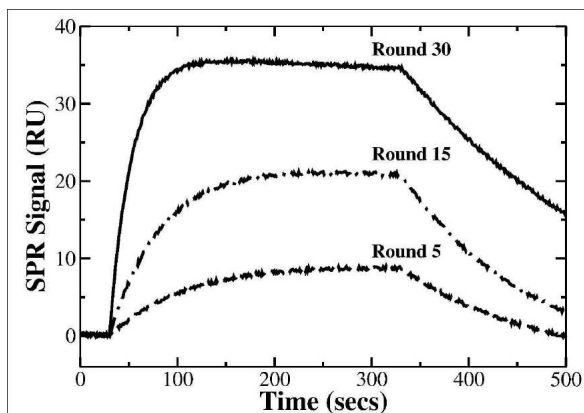


Figure 6: SPR analysis of the top scoring peptide of screening rounds 5 ($k_a = 1.2 \times 10^5 M^{-1}s^{-1}$, $k_d = 1.8 \times 10^{-2} s^{-1}$, $K_a = 6.7 \times 10^6 M^{-1}$), 15 ($k_a = 4.3 \times 10^5 M^{-1}s^{-1}$, $k_d = 6.7 \times 10^{-2} s^{-1}$, $K_a = 6.4 \times 10^6 M^{-1}$), and 30 ($k_a = 6.2 \times 10^5 M^{-1}s^{-1}$, $k_d = 1.1 \times 10^{-2} s^{-1}$, $K_a = 5.9 \times 10^7 M^{-1}$) Reprinted from [1]

However, it is important to realize that an obvious binding site may not exist for every macromolecule to be considered. In these cases, it might be worthwhile to restrict the docking space, so that the peptides are designed for the relevant pocket. The risk with this approach is nonspecific binding, which may or may not be harmful, depending on what recognition application the ligand is to be used for. On the other hand, the entire protein could be searched during docking, and peptides to be analyzed could then be restricted to only those with predicted poses at the correct binding site. The key point here is that just as proteins are very different in composition and properties, it is necessary to consider this individuality when designing binding sequences for the protein of interest.

4.6 *Ligand Design with Residue Classes*

Amino acid residues are commonly classified by their side chain polarity. Three classes arise on this basis: nonpolar, polar/neutral, and polar/charged. Other classification

Table 3: Residue classes implemented in MKPEPS; philic = hydrophilic, phobic = hydrophobic, pos = positive, neg = negative

Classes	Residues
phobic	IVLFCMA
philic	RKNDQEHPYSTG
nonpolar	VLIMFGAPW
polar	RHKNDQESTYW
charged	RHKDE
bulky	FPTYRHW
small	GAVPND CST
pos	RHK
neg	DE
aromatic	FTY
basic	KRH
acidic	DE

methods exist, and it is beneficial to enumerate these possible classes and incorporate them into the algorithm. Outlining the classes will make it easy to identify class preferences at given sites on the designed ligand. Table 3 shows the members of the defined classes for this program.

When the MVL sets had been pruned (using pruning methods described in 4.2) to about a third of their original size, a class analysis was performed. This analysis took the following form:

1. Calculation of class weight: the weight of each class, C , at a given site, i , (C_i) in the top N scoring peptides is given by:

$$W(C_i) = \frac{\sum_{n=1}^N \delta(r_i)}{N} \quad (13)$$

where

$$\delta(r_i) = \begin{cases} 1 & \text{if } r_i \in C \\ 0 & \text{if } r_i \notin C \end{cases}$$

where r_i is a residue at position i , and C is any of the predefined classes.

Table 4: MVL used for class analysis

S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8
ARCG	ARGK	ARGK	RGKS	RNQG	ARGK	RQKM	ARCGK
KPVL	FPSL	PSTF	TYVC	KWYM	PSVN	PTVH	PSTH

2. Calculation of class entropy: the entropy of each class, C , at a given site, i , (C_i) in the top N scoring peptides is given by:

$$S(C_i) = \sum_{r \in C_i} P(r) \ln P(r) \quad (14)$$

where r is a member of class C defined in the set at site i . For each of these residues, $P(r)$ is given by:

$$P(r) = \frac{\chi(r)}{\chi(r_1) + \chi(r_2) + \chi(r_3) + \dots + \chi(r_n)}, r \in C_i \quad (15)$$

where $\chi(r)$ is the frequency at which residue r , member of class C , appears at site i in the fraction being analyzed and r_1, r_2, \dots, r_n are all residues in class C .

3. Class assignment: the class weight shows how well a certain class is represented at a given site in the top fraction of scorers while the class entropy gives a measure of the diversity of the class residues. These two values are taken into consideration and are used to assign classes at different sites.

A class analysis was performed on the MVL in Table 4.

At all sites, a dominance was observed by two amino acids: lysine (K) and arginine (R). As a result of this dominance, the weights of residue classes polar, philic (hydrophilic), charged, pos (positive) and basic were very high. This was accompanied by low polar entropies, as these two residues (R,K) represent less than 20% of the total class (illustrated in Table 5). This dominance is demonstrated in Figure 7, showing the pruning factor, $D_i(s)$, (recall from Section 4.2) of the 5 highest scoring residues at each site.

Table 5: Comparison of weights and entropies for classes Polar and Small, where Polar high weights and low entropies illustrate presence of two dominant residues in the class

Site	Weight		Entropy	
	Polar	Small	Polar	Small
1	72.50 %	26.00 %	0.281	0.715
2	71.00 %	27.00 %	0.378	0.628
3	77.50 %	32.50 %	0.482	0.729
4	80.50 %	36.00 %	0.572	0.718
5	86.50 %	14.50 %	0.557	0.282
6	74.50 %	38.00 %	0.474	0.802
7	81.00 %	20.50 %	0.563	0.498
8	65.50 %	47.00 %	0.563	0.803

In order to forestall a final, nonspecific MVL containing peptides composed only of arginine and lysine residue, 8 parallel studies were performed on all the sites where the two residues (R,K) were in turn excluded from all other sites. This was performed in order to identify what site(s) benefited most from having the positively charged residues at that position.

Figure 8 shows the results. Sites 1 and 2 were identified as the most favorable sites for the positive residues (*i.e.* most negative NBIE scores), and arginine and lysine were removed from the MVL at all other sites.

After the new MVL was docked and scored (see Table 6), another class analysis was performed on the top 200 scorers. At a few of the sites, an almost equal preference was shown for two classes. In these cases, sets were modified to include all other residues which are members of both classes. In the cases where only one class was dominant, sets were modified to encompass all members of that class. The class-modified sets are shown in Table 7.

The results of using two different docking methods to design an 8-residue ligand for DNase I are shown in Table 8, which compares final predicted libraries. The final predicted libraries of the two methods have some overlap at a few sites. These results

Table 6: Defined MVLs before and after positive (R,K) residue study, R and K were subsequently removed from all sites but 1 and 2

Site	Old MVL	New MVL
1	ARCGLKPV	ARCGLKPV
2	ARGLKFPS	ARGLKFPS
3	ARGKFPST	AGFPST
4	RCGKSTYV	CGSTYV
5	RNQGKMWY	NQGMWY
6	ARNGKPSV	ANGPSV
7	RQHKMPTV	QHMPSTV
8	ARCGHKPST	ACGHPST

Table 7: Newly defined MVL after class analysis

Site	MVL	Class(es)	New MVL
1	ARCGLKPV	pos	RHK
2	ARGLKFPS	pos	RHK
3	AGFPST	small, philic	ANGFPST
4	CGSTYV	small, philic	NCGPSTYV
5	NQGMWY	philic, polar	NQHSTY
6	ANGPSV	small	ANCGPSTV
7	QHMPSTV	philic, bulky	HMPTYV
8	ACGHPST	small	ANCGHPSTV

Table 8: Comparison of final predicted libraries using the the different docking programs

	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	total
eHiTS	RK	HQ	N	GV	VI	AMI	PVM	LGA	432
AutoDock	K	RK	AGT	CGP	HS	AGT	PTV	CG	648

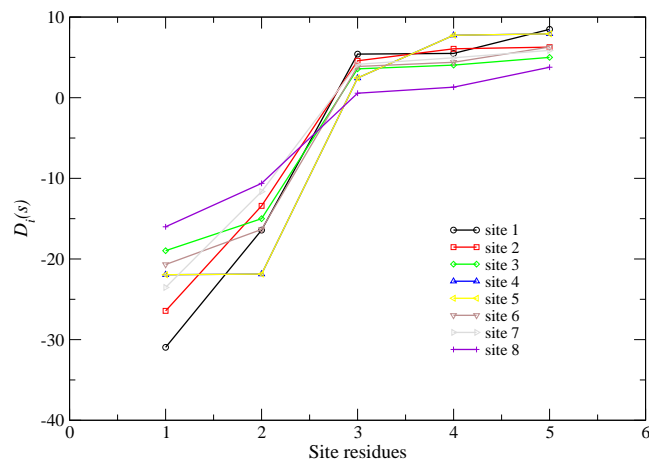


Figure 7: Pruning factor ($D_i(s)$) of the 5 highest scoring residues at each site. The two lowest residues at each site are R and K

appear to invalidate the docking program invariance hypothesis.

The early iterations of the algorithm portray an MVL which is being pruned ideally, improving the quality of the library. Figure 9 shows this trend. After the class analysis (round 13), an overall increase in the NBIE values (*i.e.* lower scores) of the peptides was seen, a result of removing the positive residues which had been contributing greatly to the low energy values. A surprising result was that the overall NBIE scores of the library did not appear to improve after round 16. Even though the library was pruned using the same pruning rules as before round 13 (described in Section 4.2), the library appeared to be optimized at round 16 (Figure 10) with an MVL size of 161,280 peptides.

4.7 Discussion

There are two possible explanations for the conflicting results seen in Section 4.6. One, it is possible that after round 16, the pruning process started to eliminate the residues which would have been desirable for high-affinity sequences. Since the same

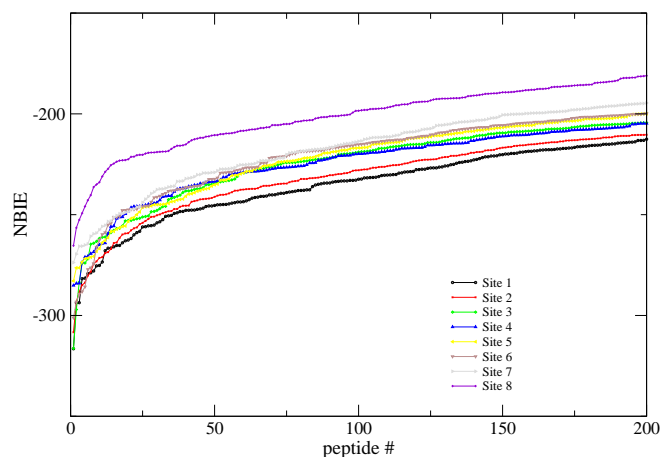


Figure 8: NBIE scores of top 200 residues in MVLs where residues R and K have been removed from all other sites and are included only at the indicated site

Table 9: Proportion of good peptides in MVLs from selected rounds

round	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
ratio ¹	0.087	0.092	0.106	0.107	0.109	0.114	0.106	0.108	0.116	0.110	0.122	0.116	0.131	0.120	0.142

pruning method was used from the beginning, this would imply that from before the class analysis to after, the pruning mechanism stopped working for the system, and needed to be modified in some way to continue accurate pruning.

The second possibility is that after round 16, all the members (residues) of the different sets started to perform equally in the binding sequence. So, it would not matter what was eliminated; as long as the size of the library was reduced, the overall number of high-affinity ligands would be reduced in a proportional manner. Table 9 shows that there is a small variance (0.000182) among this ratio (ratio of peptides with $\text{NBIE} \leq -500$ ($\text{kcal} \cdot \text{mol}^{-1}$)/total no. peptides in subMVL) for MVLs from rounds 16 - 30.

These results could be pointing to a limitation of AutoDock for the purposes of

¹ratio = no. peptides with $\text{NBIE} \leq -500$ ($\text{kcal} \cdot \text{mol}^{-1}$)/total no. peptides in subMVL

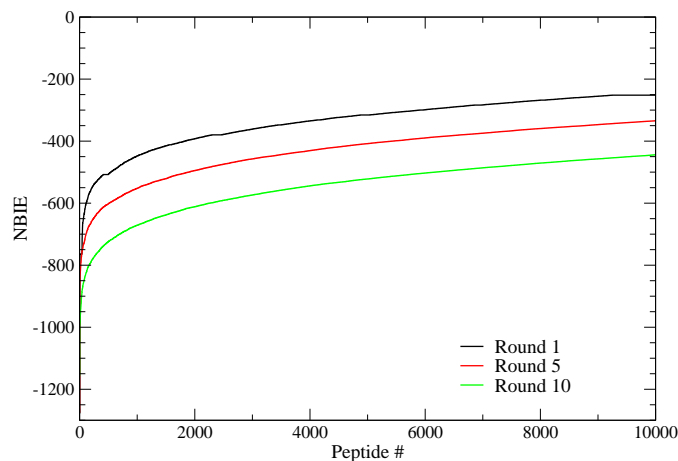


Figure 9: Rounds 1, 5 and 10. NBIE scores of MVL improve (*i.e.* more negative NBIE values) with increasing rounds

this algorithm: removal of the positive residues led to a loss of specificity so that refining the library based on energy no longer appeared to be meaningful between one round and the next. The question arises: were the initial poses that included the positive residues actually good binding conformations, or did they just appear to be reasonable poses because of the electrostatic and hydrogen bond interactions introduced by lysine and arginine. This result shows that the choice of docking program to be used for the algorithm is important. In previous results obtained using HiTS for docking, the preference for the positive residues was not seen, and the class analysis allowed a smooth convergence to the final MVL.

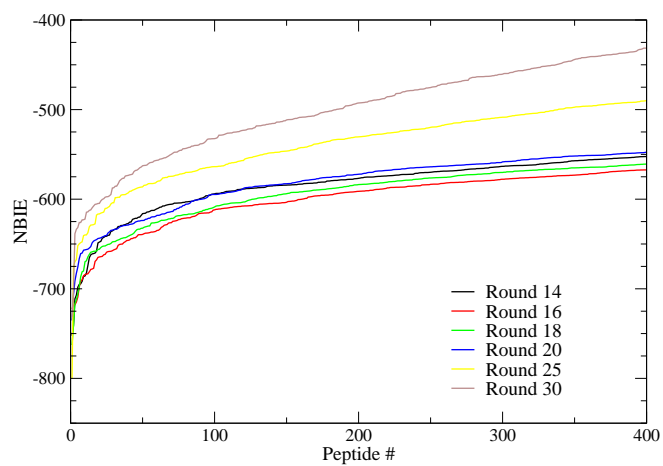


Figure 10: Iterations following class analysis. After round 16, further pruning does not result in better NBIE scores

CHAPTER V

CONCLUSIONS

5.1 *Main Conclusions*

5.1.1 Final Library and Docking Program Invariance

As shown in Section 4.7, the final MVL (Table 8) is different from the one reported in previous studies [1]. Some of the sets (*i.e.* 1, 4, 6, 7 and 8) contain residues in common, but none of the sets proved to be an exact match. The library was considered to be converged at round 30, with 648 peptides. However, because the overall quality of the library did not increase after round 16, it is likely that the best family of peptides were in the MVL from round 16. Analysis of the 200 peptides with the highest NBIE scores from round 16 predict the final MVL shown in Table 10. Both libraries show significant overlap, implying that some slight modification in the algorithmic parameters (e.g. pruning mechanism or $\chi(\text{subMVL})$) might have allowed subsequent iterations to refine the MVL for higher NBIE scores. Optimizing the algorithm for this scenario could be a goal for the future.

The results clearly do not support the docking program invariance hypothesis. This is important because it implies that a certain type of docking program may yield better results for this type of study over another. It is understandable why

Table 10: AutoDock libraries: round 30 library is the converged library obtained by even pruning of the original MVL, round 16 library is obtained by predicting the highest-quality converged library, based on the 200 highest-scoring peptides from round 16

	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	total
Round 30	K	RK	AGT	CGP	HS	AGT	PTV	CG	648
Round 16	K	K	AGST	PG	NSHT	GS	VHTPM	SG	640

a geometric complementarity algorithm like eHiTS would favor residues at different sites regardless of whether they are polar or not, while a force-field based algorithm like AutoDock would seem to show a preference for the polar, positively charged residues over the others.

An experimental binding constant determination of the peptides identified in this study would need to be performed to see how the peptides compare to those previously identified [1] and to the natural inhibitors of DNase I. Until then, we cannot conclude how and if one of the docking methods is better than the other. In the meantime we can argue that in order to design a specific ligand for a receptor, it is better to use a method that does not bias peptides in favor of polar residues which will bind non-specifically.

5.1.2 Automation

The *in silico* design algorithm has been automated to a certain extent. It could not be fully automated for a few reasons. First, the use of AutoDock for docking yielded a different final MVL than that obtained using eHiTS, which suggests that the algorithm may not be invariant to the choice of docking program. Because AutoDock seemed to favor the peptides that were made of predominantly arginine and lysine, a manual adjustment of the library had to be made. This occurrence interfered with the automation.

Second, the inclusion of classes as a determinant to prune the MVL interfered with the automation as well. During the class analysis, two classes were used to prune sets at more than one site. This occurrence was not foreseen, and so was not considered for in the program. Both of these could be incorporated into the code without too much difficulty, but it is hard to say what other possible unforeseen interferences could arise.

5.2 *Outlook and Future Work*

It is important to be able to create binding peptides that optimize affinity and are not necessarily based on the Ig fold. Such molecules can be used in any application where binding is important. The description of the theory and methodology has been provided, and future work will be aimed at upgrading the algorithm so that it is able to design binding loops of any size, and also identify binding peptides for multiple-loop replacement.

It will be important to have side-by-side experimental collaboration in the future, as this will allow a quick determination of binding affinity. This will allow algorithmic modification as necessary.

The challenge with automating the algorithm has to be tackled as well. As has been experienced with this study, the process is not always as clear-cut as eliminating the lowest scoring members of each set. A certain amount of chemical intuition sometimes goes into making prune decisions. It would not be trivial to make the algorithm completely capable of performing these functions.

It would also be necessary to understand the effect of the choice of docking algorithm on the results. There seems to be more of an impact than is immediately obvious. Understanding this effect will allow us to see if and why one type of docking method will work better with the theory and nature of the algorithm to identify optimal binding peptides.

APPENDIX A

DOCKING PARAMETERS

All tested peptides were built with the PDB format, using the MKPEPS program. The structure for the target protein (DNase I) was downloaded from the Protein Data Bank (1ATN), and stripped of all waters and hetero atoms. Docking was performed using AutoDock 4.0. All protein structures were converted into the PDBQT format using AutoDockTools (ADT) [96], as this is the format required by AutoDock.

A grid spacing of 0.5Å and $114 \times 126 \times 122$ points were used. This defined a box that was big enough to contain the receptor. Dimensions for this box were obtained by using the ADT graphical interface to visualize the DNase I molecule. The Genetic Algorithm Local Search with Solis-Wets[76] method was used. Other parameters: population size: 150 individuals, maximum number of energy evaluations: 250000, maximum number of generations: 27000, number of binding conformations generated: 10.

The AutoDock forcefield is made of six pair-wise evaluations (V) and an estimate of the conformational entropy lost upon binding (δS_{conf})[97].

$$\delta G = (V_{L-L}^{bound} - V_{L-L}^{unbound}) + (V_{P-P}^{bound} - V_{P-P}^{unbound}) + (V_{P-L}^{bound} - V_{P-L}^{unbound}) + \delta S_{conf}$$

L refers to the ligand and P refers to the protein. The pair-wise energetic terms includes the terms below:

$$\begin{aligned} V = & W_{vdw} \sum_{i,j} \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^6} \right) + W_{hbond} \sum_{i,j} E(t) \left(\frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^6} \right) \\ & + W_{elec} \sum_{i,j} \frac{q_i q_j}{\epsilon(r_{ij}) r_{ij}} + W_{sol} \sum_{i,j} (S_i V_j + S_j V_i) e^{(-r_{ij}^2/2\sigma^2)} \end{aligned} \tag{16}$$

where these terms are evaluations for dispersion/repulsion, hydrogen bonding, electrostatics and desolvation. W is an empirical weighting constant and $E(t)$ allows for directionality based on the angle t from ideal hydrogen bonding geometry. The fourth term is a desolvation potential based on the volume of atoms (V) that surround a given atom, sheltering it from solvent, weighted by a solvation parameter (S) and exponential term with distance-weighting factor $\sigma = 3.5\text{\AA}$ [97].

REFERENCES

- [1] QUIRK, S., ZHONG, S., and HERNANDEZ, R., De novo identification of binding sequences for antibody replacement molecules, *Proteins: Struct. Func. Bioinfo.*, vol. 76, pp. 693–705, 2009.
- [2] BRYSON, J. W., BETZ, S. F., LU, H. S., SUICH, D. J., ZHOU, H. X., O’NEIL, K. T., and DEGRADO, W. F., Protein design: A hierarchic approach, *Science*, vol. 11, pp. 329–334, 2007.
- [3] VOIGT, C. A., MAYO, S. L., ARNOLD, F. H., and WANG, Z.-G., Computational method to reduce the search space for directed protein evolution, *Proc. Natl. Acad. Sci. U.S.A.*, vol. 98, pp. 3778–3783, 2001.
- [4] FUNG, H. K., WELSH, W. J., and FLOUDAS, C. A., Computational de novo peptide and protein design: rigid template versus flexible templates, *Ind. Eng. Chem. Res.*, vol. 47, pp. 993–1001, 2008.
- [5] SARAF, M. C., MOORE, G. L., GOODNEY, N. M., CAO, V. Y., BENKOVIC, S. J., and MARANAS, C. D., IPRO: An iterative computational protein library redesign and optimization procedure, *Biophys. J.*, vol. 90, pp. 4167–4180, 2006.
- [6] SARAF, M. C., GUPTA, A., and MARANAS, C. D., Design of combinatorial protein libraries of optimal size, *Proteins: Struct. Func. Bioinfo.*, vol. 60, pp. 769–777, 2005.
- [7] KANG, S.-G. and SAVEN, J. G., Computational protein design: structure, function and combinatorial diversity, *Curr. Opin. Chem. Biol.*, vol. 11, pp. 329–334, 2007.
- [8] AMBROGGIO, X. I. and KUHLMAN, B., Computational design of a single amino acid sequence that can switch between two distinct protein folds, *J. Am. Chem. Soc.*, vol. 128, pp. 1154–1161, 2006.
- [9] ALI, M. H., TAYLOR, C. M., GRIGORYAN, G., ALLEN, K. N., IMPERIALI, B., and KEATING, A. E., Design of a heterospecific tetrameric, 21-residue miniprotein with mixed alpha/beta structure, *Structure*, vol. 13, pp. 225–234, 2005.
- [10] NANDA, V., ROSENBLATT, M. M., OSYCZKA, A., KONO, H., GETAHUN, Z., DUTTON, P. L., SAVEN, J. G., and DEGRADO, W. F., De novo design of a redox-active minimal rubredoxin mimic, *J. Am. Chem. Soc.*, vol. 127, pp. 5804–5805, 2005.

- [11] COCHRAN, F. V., WU, S. P., WANG, W., NANDA, V., SAVEN, J. G., THE-RIEN, M. J., and DEGRADO, W. F., Computational de novo design and characterization of a four-helix bundle that selectively binds a nonbiological cofactor, *J. Am. Chem. Soc.*, vol. 127, pp. 1346–1347, 2005.
- [12] DAHIYAT, B. and MAYO, S. L., De novo protein design: fully automated sequence selection, *Science*, vol. 278, pp. 82–87, 2007.
- [13] KUHLMAN, B., DANTAS, G., IRETON, G., VARANI, G., STODDARD, B., and BAKER, D., Design of a novel globular protein fold with atomic-level accuracy, *Science*, vol. 302, pp. 1364–1368, 2003.
- [14] ZOU, J. and SAVEN, J. G., Statistical theory of combinatorial libraries of folding proteins: energetic discrimination of a target structure, *J. Mol. Biol.*, vol. 296, pp. 281–294, 2000.
- [15] KONO, H. and SAVEN, J. G., Statistical theory of protein combinatorial libraries. Packing interactions, backbone flexibility, and the sequence variability of a main-chain structure, *J. Mol. Biol.*, vol. 306, pp. 607–628, 2001.
- [16] PARK, S., YANG, X., and SAVEN, J. G., Advances in computational protein design, *Curr. Opin. Struct. Biol.*, vol. 14, pp. 487–494, 2004.
- [17] BINZ, H. K. and AMSTUTZ, P., Engineering novel binding proteins from non-immunoglobulin domains, *J. Comp. Chem.*, vol. 23, pp. 1257–68, 2005.
- [18] FILPULA, D., Antibody engineering and modification technologies, *Biomed. Eng.*, vol. 24, pp. 201–215, 2007.
- [19] WORN, A. and PLUCKTHUN, A., Stability engineering of antibody single chain Fv fragments, *J. Mol. Biol.*, vol. 305, pp. 989–1010, 2001.
- [20] MUYLDERMANS, S., Single domain camel antibodies: current status, *J. Biotechnol.*, vol. 74, pp. 277–302, 2001.
- [21] WARR, G. W., MAJOR, K. E., and HIGGINS, D. A., IgY: clues to the origins of modern antibodies, *Immunol. Today*, vol. 16, pp. 392–398, 1995.
- [22] NYGREN, P.-A. and SKERRA, A., Binding proteins from alternative scaffolds, *J. Immunol. Meths.*, vol. 290, pp. 3–28, 2004.
- [23] VITA, C., VIZZAVONA, J., DRAKOPOULOU, E., ZINN-JUSTIN, S., GILQUIN, B., and MENEZ, A., Novel miniproteins engineered by the transfer of active sites to small natural scaffolds, *Biopolymers*, vol. 47, pp. 93–100, 1998.
- [24] ABEDI, M. R., CAPONIGRO, G., and KAMB, A., Green fluorescent protein as a scaffold for intracellular presentation of peptides, *Nucleic Acids Res.*, vol. 26, pp. 623–630, 1998.

- [25] RENARD, M., Knowledge-based design of reagentless fluorescent biosensors from recombinant antibodies, *J. Mol. Biol.*, vol. 318, pp. 429–442, 2002.
- [26] ANDERSSON, M., RONNMARK, J., ARESTRÖM, I., NYGREN, P.-A., and AHLBORG, N., Inclusion of a non-immunoglobulin binding protein in two-site ELISA for quantification of human serum proteins without interference by heterophilic serum antibodies, *J. Immunol. Methods*, vol. 283, pp. 225–234, 2003.
- [27] EWERT, S., HONEGGER, A., and PLOCKTHUN, A., Stability improvement of antibodies for extracellular and intracellular applications: CDR grafting to stable frameworks and structure-based framework engineering, *Methods*, vol. 34, pp. 184–199, 2004.
- [28] KOIDE, A., ABBATIELLO, S., ROTHGERY, L., and KOIDE, S., Probing protein conformational changes in living cells by using designer binding protein: application to the estrogen receptor, *Proc. Natl. Acad. Sci. U.S.A.*, vol. 99, pp. 1253–1258, 2002.
- [29] HOGBOM, M., EKLUND, M., NYGREN, P.-A., and NORDLUND, P., Structural basis for recognition by an in vitro evolved affibody, *Proc. Natl. Acad. Sci. U.S.A.*, vol. 100, pp. 3191–3196, 2003.
- [30] NORD, K., GUNNERIUSON, E., UHLEN, M., and NYGREN, P.-A., Ligands selected from combinatorial libraries of protein A for use in affinity capture of apolipoprotein A-1M and TaqDNA polymerase, *J. Biotechnol.*, vol. 80, pp. 45–54, 2000.
- [31] REINA, J., LACROIX, E., HOBSON, S. D., FERNANDEZ-BALLESTER, G., RYBIN, V., SCHWAB, M. S., SERRANO, L., and GONZALEZ, C., Computer-aided design of a PDZ domain to recognize new target sequences, *Nat. Struct. Biol.*, vol. 9, pp. 621–627, 2002.
- [32] SMITH, G. P., Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface, *Science*, vol. 228, pp. 1315–1317, 1985.
- [33] PELLETIER, J. N., ARNDT, K. M., PLUKTHUN, A., and MICHNICK, S. W., An in vivo library-versus-library selection of optimized protein-protein interactions, *Nat. Biotechnol.*, vol. 17, pp. 683–690, 1999.
- [34] MATTHEAKIS, L. C., BHATT, R. R., and DOWER, W. J., An in vitro polysome display system for identifying ligands from very large peptide libraries, *Proc. Natl. Acad. Sci. U.S.A.*, vol. 91, pp. 9022–9026, 1994.
- [35] KOIDE, A., BAILEY, C. W., HUANG, X., and KOIDE, S., The fibronectin type III domain as a scaffold for novel binding proteins, *J. Mol. Biol.*, vol. 284, pp. 1141–1151, 1998.

- [36] NORD, K., GUNNERIUSSON, E., RINGDAHL, J., STAHL, S., UHLEN, M., and NYGREN, P. A., Binding proteins selected from combinatorial libraries of an α -helical bacterial receptor domain, *Nat. Biotechnol.*, vol. 15, pp. 772–777, 1997.
- [37] BINZ, H. K., AMSTUTZ, P., KOHL, A., STUMPP, M. T., BRIAND, C., FORRER, P., GRUTTER, M. G., and PLUCKTHUN, A., High-affinity binders selected from designed ankyrin repeat protein libraries, *Nat. Biotechnol.*, vol. 22, pp. 575–582, 2004.
- [38] BESTE, G., SCHMIDT, F. S., STIBORA, T., and SKERRA, A., Small antibody-like proteins with prescribed ligand specificities derived from the lipocalin fold, *Proc. Natl. Acad. Sci. U.S.A.*, vol. 96, pp. 1898–1903, 1999.
- [39] MERSICH, C. and JUNGBAUER, A., Generation of bioactive peptides by biological libraries, *J. Chromatogr. B.*, vol. 861, pp. 160–170, 2008.
- [40] BARAKAT, N. H. and LOVE, J. J., Molecular diversity in engineered protein libraries, *Curr. Opin. Chem. Biol.*, vol. 11, pp. 335–341, 2007.
- [41] DESJARLAIS, J. and CLARKE, N. D., Computer search algorithms in protein modification and design, *Curr. Opin. Struct. Biol.*, vol. 8, pp. 471–475, 1998.
- [42] LEE, C. and LEVITT, M., Accurate prediction of the stability and activity effects of site-directed mutagenesis on a protein core, *Nature*, vol. 352, pp. 448–451, 1991.
- [43] HELLINGA, H. W. and RICHARDS, F. M., Optimal sequence selection in proteins of known structure by simulated evolution, *Proc. Natl. Acad. Sci. U.S.A.*, vol. 91, pp. 5803–5807, 1994.
- [44] TUFFERY, P., ETCHEBEST, C., HAZOUT, S., and LAVERY, R., A new approach to the rapid determination of protein sidechain conformations, *J. Biomol. Struct. Dyn.*, vol. 8, pp. 1267–1289, 1991.
- [45] DESJARLAIS, J. and HANDEL, T. M., De novo design of the hydrophobic cores of proteins, *Protein Sci.*, vol. 4, pp. 2006–2018, 1995.
- [46] ZOU, J. and SAVEN, J. G., Using self-consistent fields to bias Monte Carlo methods with applications to designing and sampling protein sequences, *J. Chem. Phys.*, vol. 118, pp. 3843–3854, 2003.
- [47] DESMET, J., MAEYER, M. D., HAZES, B., and LASTERS, I., The dead-end elimination theorem and its use in protein side-chain positioning, *Nature*, vol. 356, pp. 539–542, 1992.
- [48] LOOGER, L. and HELLINGA, H. W., Generalized dead-end elimination algorithms make large-scale Protein side-chain structure prediction tractable: implications for protein design and structural genomics, *J. Mol. Biol.*, vol. 307, pp. 429–445, 2001.

- [49] BOLON, D. N. and MAYO, S. L., Enzyme-like proteins by computational design, *Proc. Natl. Acad. Sci. U.S.A.*, vol. 98, pp. 14274–14279, 2001.
- [50] MOOERS, B. H. M., DATTA, D., BAASE, W. A., ZOLLARS, E. S., MAYO, S. L., and MATTHEWS, B. W., Repacking the core of the t4 lysozyme by automated design, *J. Mol. Biol.*, vol. 332, pp. 741–756, 2003.
- [51] BHATTACHERJEE, A. and BISWAS, P., Statistical theory of protein sequence design by random mutation, *J. Phys. Chem. B*, vol. 113, pp. 5520–5527, 2009.
- [52] BISWAS, P., ZOU, J., and SAVEN, J. G., Statistical theory for protein ensembles with designed energy landscapes, *J. Chem. Phys.*, vol. 123, p. 154908, 2005.
- [53] TANG, J., KANG, S.-G., SAVEN, J. G., and GAI, F., Characterization of the cofactor-induced folding mechanism of a zinc-binding peptide using computationally designed mutants, *J. Mol. Biol.*, vol. 389, pp. 90–102, 2009.
- [54] BROOIJMANS, N. and KUNTZ, I. D., Molecular recognition and docking algorithms, *Annu. Rev. Biophys. Biomol. Struct.*, vol. 32, pp. 335–373, 2003.
- [55] JANIN, J., Protein-protein recognition, *Prog. Biophys. Molec. Biol.*, vol. 64, pp. 145–166, 1995.
- [56] WILCHELK, M., BAYER, E. A., and LIVNAH, O., Essentials of biorecognition: the (strept)avidin-biotin system as a model for protein-protein and protein-ligand interaction, *Immunol. Lett.*, vol. 103, pp. 27–32, 2006.
- [57] AJAY and MURCKO, M. A., Computational methods to predict binding free energy in ligand-receptor complexes, *J. Med. Chem.*, vol. 38, pp. 4953–4967, 1995.
- [58] BOHM, H.-J. and KLEBE, G., What can we learn from molecular recognition in protein-ligand complexes for the design of new drugs?, *Angew. Chem. Int. Ed. Engl.*, vol. 35, pp. 2588–2614, 1996.
- [59] CHOTHIA, C. and JANIN, J., Principles of protein-protein recognition, *Nature*, vol. 256, pp. 705–708, 1975.
- [60] PLATZER, K. E. B., MOMANY, F. A., and SCHERAGA, H. A., Conformational energy calculations of enzyme substrate interactions. I. Computation of preferred conformations of some substrates of a-chymotrypsin, *Int. J. Pept. Prot. Res.*, vol. 4, pp. 187–200, 1972.
- [61] WODAK, S. J. and JANIN, J., Computer analysis of protein-protein interaction, *J. Mol. Biol.*, vol. 124, pp. 323–342, 1978.
- [62] GREER, J. and BUSH, B. L., Macromolecular shape and surface maps by solvent exclusion, *Proc. Natl. Acad. Sci. U.S.A.*, vol. 75, pp. 303–307, 1978.

- [63] CARLSON, H. A. and MCCAMMON, J. A., Accomodating protein flexibility in computational drug design, *Mol. Pharm.*, vol. 57, pp. 213–218, 2000.
- [64] GOODSSELL, D. S. and OLSEN, A. J., Automated docking of substrates to proteins by simulated annealing, *Proteins: Struct. Func. Gen.*, vol. 8, pp. 195–202, 1990.
- [65] SANDAK, B., NUSSINOV, R., and WOLFSON, H. J., An automated computer vision and robotics-based technique for 3-D flexible biomolecular docking and matching, *Comput Appl Biosci*, vol. 11, pp. 87–99, 1995.
- [66] CLAR, K. P. and JAIN, A. N., Flexible ligand docking without parameter adjustment across four ligand-receptor complexes, *J. Comp. Chem.*, vol. 16, pp. 1210–1226, 1995.
- [67] ZSOLDOS, Z., REID, D., SIMON, A., SADJAD, S. B., and JOHNSON, A. P., eHiTS: An innovative approach to the docking and scoring function problems, *Curr. Protein Pept. Sci.*, vol. 7, pp. 421–435, 2006.
- [68] MORRIS, G., GOODSSELL, D., HUEY, R., HART, R., HART, W. E., BELEW, R. K., and OLSON, A. J., Automated Docking Using a Lamarckian Genetic Algorithm and an Empirical Binding Free Energy Function, *J. Comp. Chem.*, vol. 19, pp. 1639–1662, 1998.
- [69] DE AZEVEDO JR., W. F. and DIAS, R., Computational methods for calculation of binding affinity, *Curr. Drug Targets*, vol. 9, pp. 1031–1039, 2008.
- [70] DOMINY, B. N., Molecular recognition and binding free energy calculations in drug development, *Curr. Pharm. Biotech.*, vol. 9, pp. 87–95, 2008.
- [71] GILSON, M. K. and ZHOU, H.-X., Calculation of protein-ligand binding affinities, *Annu. Rev. Biophys. Struct.*, vol. 36, pp. 21–42, 2007.
- [72] FOLOPPE, N. and HUBBARD, R., Towards predictive ligand design with free-energy based computational methods, *Curr. Med. Chem.*, vol. 13, pp. 3583–3608, 2006.
- [73] GUVENCH, O. and JR, A. D. M., Computational evaluation of protein-small molecule docking, *Curr. Opin Drug Discov Dev*, vol. 19, pp. 56–61, 2009.
- [74] ZSOLDOS, Z., SZABO, I., SZABO, Z., and JOHNSON, A. P., Software tools for structure based rational drug design, *J. Mol. Struct.*, vol. 666-667, pp. 659–665, 2003.
- [75] ENGLEBIENNE, P., FIAUX, H., KUNTZ, D. A., CORBEIL, C. R., GERBER-LEMAIRE, S., ROSE, D. R., and MOITESSIER, N., Evaluation of docking programs for predicting binding of Golgi α -mannosidase II inhibitors: A comparison with crystallography, *Proteins: Struct. Func. Bioinfo.*, vol. 69, pp. 160–176, 2007.

- [76] SOLIS, F. J. and WETS, R. J.-B., Minimization by random search techniques, *Math. Oper. Res.*, vol. 6, pp. 19–30, 1981.
- [77] GOODSSELL, D. S., MORRIS, G. M., and OLSEN, A. J., Automated docking of substrates to proteins by simulated annealing, *J. Mol. Recognit*, vol. 8, pp. 195–202, 1990.
- [78] ALI, H. I., FUJITA, T., AKAHO, E., and NAGAMATSU, T., A comparative study of AutoDock and PMF scoring performances, and SAR of 2-substituted pyrazolotriazolopyrimidines and 4-substituted pyrazolopyrimidines as potent xanthine oxidase inhibitors, *J. Comput. Aided. Mol. Des.*, vol. 24, pp. 57–75, 2010.
- [79] VOROBEV, Y. N., Blind docking method combining search of low-resolution binding sites with ligand pose refinement by molecular dynamics-based global optimization, *J. Comp. Chem.*, vol. 31, pp. 1080–1092, 2010.
- [80] KAMEL, M. M., ALI, H. I., ANWAR, M. M., MOHAMED, N. A., and SOLIMAN, A. M., Synthesis, antitumor activity and molecular docking study of novel Sulfonamide-Schiff’s bases, thiazolidinones, benzothiazinones and their C-nucleoside derivatives, *Eur. J. Med. Chem.*, vol. 45, pp. 572–580, 2010.
- [81] VAQUE, M., AROLA, A., ALIAGAS, C., and PUJADAS, G., BDT: an easy-to-use front-end application for automation of massive docking tasks and complex docking strategies with AutoDock, *Bioinformatics*, vol. 22, pp. 1803–1804, 2006.
- [82] RAO, M. S. and OLSON, A. J., Modelling of factor Xa-inhibitor complexes: A computational flexible docking approach, *Proteins: Struct. Func. Gen.*, vol. 34, pp. 173–183, 1999.
- [83] BITOMSKY, W. and WADE, R. C., Docking of glycosaminoglycans to heparin-binding proteins: validation for aFGF, bFGF, and antithrombin and application to IL-8, *J. Am. Chem. Soc.*, vol. 121, pp. 3004–3013, 1999.
- [84] SOTRIFFER, C. A., FLADER, W., WINGER, R. H., RODE, B. M., LIEDL, K. R., and VARGA, J. M., Automated docking of ligands to antibodies: methods and applications, *Methods*, vol. 20, pp. 280–291, 2000.
- [85] BARTOLUCCI, C., PEROLA, E., PILGER, C., FELS, G., and LAMBA, D., Three-dimensional structure of a complex of galanthamine (Nivalin) with acetylcholinesterase from torpedo californica: Implications for the design of new anti-Alzheimer drugs, *Proteins: Struct. Func. Gen.*, vol. 42, pp. 182–191, 2001.
- [86] HETENYI, C. and SPOEL, D. V. D., Efficient docking of peptides to proteins without prior knowledge of the binding site, *Prot. Sci.*, vol. 11, pp. 1729–1737, 2002.
- [87] WANG, R., LU, Y., and WANG, S., Comparative Evaluation of 11 scoring functions for molecular docking, *J. Med. Chem.*, vol. 46, pp. 2287–2303, 2003.

- [88] CINCILLA, G., VIDAL, D., and PONS, M., An improved scoring function for sub-optimal polar ligand complexes, *J. Comput. Aided. Mol. Des.*, vol. 23, pp. 143–152, 2009.
- [89] FLOUDAS, C. A., FUNG, H. K., MCALLISTER, S. R., MONNIGMANN, M., and RAJGARIA, R., Advances in protein structure prediction and de novo protein design: A review, *Chem. Eng. Sci.*, vol. 61, pp. 966–988, 2006.
- [90] ZHONG, S., MOIX, J. M., QUIRK, S., and HERNANDEZ, R., Dihedral-angle information entropy as a gauge of secondary structure propensity, *Biophys. J.*, vol. 91, pp. 4014–4023, 2006.
- [91] BROOKS, B. R., BRUCCOLERI, R. E., OLAFSON, B. D., STATES, D. J., SWAMINATHAN, S., and KARPLUS, M., CHARMM: A program for macromolecular energy, minimization and dynamics calculations, *J. Comp. Chem.*, vol. 4, pp. 187–217, 1983.
- [92] LAZARIDES, E. and LINDBERG, U., Actin is the naturally occurring inhibitor of deoxyribonuclease I, *Proc. Natl. Acad. Sci. U.S.A.*, vol. 71, pp. 4742–4746, 1974.
- [93] KROVAT, E. M., STEINDL, T., and LANGER, T., Recent advances in docking and scoring, *Curr. Comput. Aid. Drug Des.*, vol. 1, pp. 93–102, 2005.
- [94] WARREN, G. L., ANDREWS, C. W., CAPELLI, A.-M., CLARKE, B., LALONDE, J., LAMBERT, M. H., LINDVALL, M., NEVINS, N., SEMUS, S. F., SENGER, S., TEDESCO, G., WALL, I. D., WOOLVEN, J. M., PEISHOFF, C. E., and HEAD, M. S., A critical assessment of docking programs and scoring functions, *J. Med. Chem.*, vol. 49, pp. 5912–31, 2006.
- [95] GOHLKE, H., HENDRICH, M., and KLEBE, G., Knowledge-based scoring-function to predict protein-ligand interactions, *J. Mol. Biol.*, vol. 295, pp. 337–356, 2000.
- [96] Molecular Graphics Lab, The Scripps Research Institute, La Jolla, California, *ADT Tools, Version 2.0*, 2005.
- [97] MORRIS, G. M., GOODSSELL, D. S., PIQUE, M. E., LINDSTROM, W., HUEY, R., FORLI, S., HART, W. E., HALLIDAY, S., BELEW, R., and OLSON, A. J., *AutoDock Version 4.2: Automated docking of flexible ligands to flexible receptors*. Molecular Graphics Lab, The Scripps Research Institute, La Jolla, California, May 2009.