De Novo Peptide Design Strategies

by

Evrim Besray Ünal

A Thesis Submitted to the
Graduate School of Sciences and Engineering
in Partial Fulfillment of the Requirements for
the Degree of

Doctor of Philosophy
in
Computational Sciences and Engineering

Koç University

August, 2011

Koç University

Graduate School of Sciences and Engineering

This is to certify that I have examined this copy of a doctoral dissertation by

Evrim Besray Ünal

and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Committee Members:

_____

Prof. Burak Erman

_____

Prof. Attila Gürsoy

_____

Prof. Özlem Keskin

_____

Assist. Prof. Mehmet Sayar

_____

Assist. Prof. Demet Akten

Date:_____

ABSTRACT

Short peptide segments have gained importance as drug candidates. There exist three main problems for peptide design: determining the appropriate sequence with the desired function; properly docking peptide on the protein surface; and the unbound state of the peptide that is to be used as a drug. The 'unbound state' means peptide chains in the denatured state at physiological conditions. The details of the potentials for peptide docking simulations and the statistical features of peptides are defined in the literature.

As a solution to the peptide sequence determination problem, several experimental and *in silico* techniques exist to screen peptides. There is no general computational tool to determine peptide sequences. On the other hand, peptide motifs are crucial for selective and specific binding. There have been successful attempts to discover biological motifs by different research groups. To our knowledge, the efforts in the literature are based on the alignment of evolutionarily conserved motifs from proteins. The evolutionary peptide motif search algorithms/servers/software are available. However, there is no general methodology to discover a binding peptide motif for any protein target. We aim to predict peptide sequences and peptide binding motifs for any given protein using no prior information. Here, four different algorithms are developed for peptide design. The implementation of genetic algorithm, Markov model and hidden Markov model with Viterbi decoding leads to prediction of peptides for different protein targets. The algorithms are successful to determine peptide sequences with good theoretical binding affinities. The peptide motifs for two case-studies are also offered. A web-server, VitAL, is constructed based on Viterbi decoding.

The statistical thermodynamics features of the unbound peptide as a small thermodynamics system in a thermal reservoir is lacking in the literature. A novel statistical thermodynamics approach is applied to the free peptide segments in order to classify them according to their conformational energies and entropies and heat capacities. The conformational partition function, Helmholtz free energy, energy, entropy and heat capacity are obtained. The model is applied to randomly produced peptides and to known peptide inhibitors. Peptides with low energy, low entropy and low heat capacity are determined to be essential for a peptide to be a good candidate inhibitor.

ÖZET

Son yıllarda, kısa peptidler ilaç adayı olarak önem kazanmıştır. Peptid tasarımında üç ana sorun mevcuttur: istenilen fonksiyonlara sahip uygun bir peptid sekansı bulmak; peptidi protein yüzeyine en uygun şekilde yerleştirmek ve ilaç olarak kullanılacak peptidin "bağlanmamış şekli"ni tayin etmek. Peptidin "bağlanmamış şekli" ile kastedilen peptid zincirinin fizyolojik koşullardaki denatüre halidir. Peptidin protein yüzeyine uygun şekilde nasıl yerleştirileceği ve peptidlerin istatistikler özellikleri literatürde tanımlanmıştır.

Peptid sekans tasarımı problemine çözüm olarak, pek çok deneysel ve hesaplamalı teknik mevcuttur. Fakat peptid sekansı belirlemek için genel bir hesaplama yöntemi yoktur. Öte yandan, peptid motifleri, seçici ve spesifik bağlanma için çok önemlidir. Farklı araştırma gruplarının biyolojik motifleri keşfetmek için başarılı girişimleri olmuştur. Bildiğimiz kadarıyla, literatürdeki çalışmalar proteinlerden evrimsel olarak korunmuş motifleri bulmaya dayanmaktadır. Evrimsel peptid motif arama algoritmaları/web-sayfaları/yazılımları mevcuttur. Ancak, herhangi bir protein için selektif olarak bağlanan peptid motifi keşfetmek için genel bir yöntem bulunmamaktadır. Bu çalışmada, hiç bir önbilgi kullanmadan, herhangi bir protein için peptid sekansı ve peptid motifleri bulmak hedeflenmiştir. Dört farklı peptid tasarımı algoritması geliştirilmiştir. Genetik algoritma, Markov modeli ve Viterbi modeli uygulamaları ile farklı protein hedefleri için peptidler tahmin edilmiştir. Algoritmalar farklı proteinler için iyi teorik bağlanma gösteren peptid sekansları tasarlamakta için başarılı olmuştur. Ayrıca çalışmalarımızda kullandığımız algoritmalar ile "VitAL" adlı bir web sunucusu inşa edilmiştir.

Peptidlerin bağlanmamış hallerdeki istatistiksel termodinamik özellikleri literatürde yer almamaktadır. Proteine bağlanmamış peptidleri küçük termodinamik sistem olarak tanımlamak mümkündür. Bu sistemdeki peptidleri konformasyonel enerjilerine, entropilerine ve ısı kapasitelerine göre sınıflandırmak için yeni bir istatistiksel termodinamik yaklaşım uygulanmıştır. Peptidlerin Helmholtz enerjileri, enerjileri, entropileri ve ısı kapasiteleri bu yaklaşım ile elde edilmiştir. Modeli, rasgele üretilen peptidler ve bilinen peptid inhibitörleri uygulanmıştır. Düşük enerji, düşük entropi ve düşük ısı kapasitesine sahip peptidlerin iyi birer inhibitör adayı olduğu tespit edilmiştir.

ACKNOWLEDGEMENT

The easiest and hardest chapter for me to write: naming all the people who helped me will be easy, but I will not be able to thank them sufficiently with words.

I would like to thank, first and foremost, my advisors Prof. Burak Erman and Prof. Attila Gürsoy for their guidance and support throughout my graduate study and during the completion of this thesis. Without their encouragement, knowledge and enthusiastism this thesis would have never been completed.

I am extremely grateful to Prof. Özlem Keskin, Assistant Prof. Mehmet Sayar and Assistant Prof. Demet Akten for their participation in my thesis committee and for the critical reading of my thesis.

I would like to thank Prof. Pınar Çalık, Assistant Prof. Batu Erman and Assistant Prof. Arzu Atalay for collaboration.

Thank you all, who made the four years an enjoyable period of time: Ayşe Küçükyılmaz, Bahar Değirmenci, Billur Engin, Beytullah Özgür, Cahit Dalgıçtır, Cengiz Ulubaş, Ceren Tüzmen, Çiğdem Bayrak, Ece Özbabacan, Ekin Tüzün, Emre Güney, Engin Çukuroğlu, Fatih Toy, Gözde Kar, Güray Kuzu, M.Ali Öztürk, Osman Yoğurtçu, Sefer Baday, Tuğba Özal, Yasemin Oğuz and Zeynep Şimşek. Special thanks to Özge Engin, Nurcan Tunçbağ and Bahar Öndül, who accompanied me in all good and bad times.

Special thanks to Istanbul Lindy Hoppers, who have brightened my days with music, dance and friendship.

Last, I would like to thank my grandparents, my parents Işın & Osman and my sister Eren Naz for their love, affection and support throughout my life, without their encouragement I could not be what I am today. To them I dedicate this thesis.

Words are flying out like endless rain into a paper cup / They slither while they pass / They slip away across the universe / Pools of sorrow waves of joy are drifting through my open mind / Possessing and caressing me

Jai guru deva, om

Nothing's gonna change my world

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

NOMENCLATURE

PDB	Protein Data Bank
GNM	Gaussian Network Model
GA	Genetic algorithm
VitAL	Viterbi algorithm
VitSTR	Viterbi algotihm using structural information from literature

Chapter 1

INTRODUCTION

Computational design of inhibitor molecules for selected target proteins, which play a role in diseases, has gained importance in the last decade. The ability to determine successful inhibitors quickly and with minimal cost with computational design techniques, when compared to the traditional design techniques, has increased interest in computational techniques. In the past years, scientific concentration has mostly been on the design of small organic molecules as protein inhibitors. Some of the small drugs have the required specificity and selectivity; however, most of them cause various side-effects. The toxicity caused by the small drugs can be avoided by using peptide inhibitors instead. Peptide inhibitors are not expected to cause severe side-effects since they are made up of amino acids, which are the main components of proteins. Consequently, peptide inhibitors with high specificity and selectivity with minimum toxicity can be generated.

The focus of computational inhibitor design is currently changing from small molecules to peptides. There are a number of scientific groups working on peptide design, and also some companies' R & D departments are searching peptides for cosmetics. As the research in the area is recent, there is not enough data about the high-quality rational design methods, the superior peptide inhibitors properties, or the limits for the binding energy of peptide inhibitors to their target proteins. The aim of this dissertation is to investigate the common properties of peptide inhibitors and to determine novel strategies for the peptide design problem. First specifically, the known peptide inhibitors were examined and their entropic, energetic properties were determined; then the information of protein data banks was studied in order to determine the thermodynamic properties of the known protein sequences. Subsequently, novel approaches for peptide design was targeted; and this methodologies were applied to selected target proteins. The design of peptides was accomplished by employing probabilistic methods as the hidden Markov models, the Viterbi algorithm and the genetic algorithm.

This dissertation focuses on *de novo* peptide design methods and thermodynamic properties of peptides. The outline of this dissertation is as follows:

Chapter 2 focuses on a recent review on the peptide design strategies, the peptide motifs and the target proteins that studied in this dissertation.

Chapter 3 illustrates a novel approach to determine the thermodynamics of peptides utilizing computational means. Initially, the method is illustrated and then the performance of the method on available peptide dataset is given.

Chapter 4 is based on peptide design via Genetic algorithm. The model is described and the case study results are demonstrated. In Chapter 5, peptide design strategy based on Markov model is described and a case study is illustrated. Chapter 6 explains two peptide design models that are based on Viterbi decoding. At the end of this chapter, the web server VitAL is introduced. The server provides a user-friendly interface to to predict peptide sequences for a given protein target. The proposed algorithms are efficient for *de novo* peptide design.

Chapter 7 demonstrates peptide motifs that are determined by our novel design strategies.

The dissertation is concluded with a short summary of the performed study and future research work.

Chapter 2

LITERATURE REVIEW

2.1. Statistical Thermodynamics Properties of Peptides:

Short peptide segments offer new possibilities as drug candidates as stated in Chapter 1. In view of the fact that the peptides have gained importance in emerging technologies, their structure-function relations need a detailed understanding. The main problem in this area is determining the appropriate sequence that will exhibit the desired function. There exist two main problems for peptide drug design: properly docking a given small molecule on the protein surface, and the unbound state of the peptide that is to be used as a drug. The solutions to the first problem are outlined in the recent excellent review of Kitchen et al. [1]. In the second problem, the term 'unbound state' means peptide chains in the denatured state at physiological conditions [2]. The statistical features of peptides are already well known in the literature, which have been pioneered by Flory and his collaborators [3-6]. Several techniques have been developed to screen probable peptides for use as drugs [7-8] and the details of the potentials for binding of peptides for simulations are already well known [7, 9-13].

The change of free energy ΔF of the protein, peptide/ligand and water system upon binding consists of the following components:

$$\Delta F = \Delta F_{RB}(L) + \Delta F_{conf}(L) + \Delta F_{RB}(P) + \Delta F_{conf}(P) + \Delta F_{binding}(LP) + \Delta F(W) \qquad (1)$$

where, $\Delta F_{RB}(L)$ is the change from rigid body translation and rotation of the ligand, $\Delta F_{conf}(L)$ is the conformational free energy change of the ligand that is the basis of calculations of the present work, $\Delta F_{RB}(P)$ is the change of the rigid body translation and rotation of the protein, $\Delta F_{conf}(P)$ is from the conformational changes of the protein, $\Delta F_{binding}(LP)$ is the change in free energy due to the interaction between the ligand and protein, and $\Delta F(W)$ is the free energy change of water. We assume the terms $\Delta F_{RB}(P)$ and $\Delta F_{conf}(P)$ are small. The contribution of each remaining term is discussed in more detail in order to assess the contribution of $\Delta F_{conf}(L)$ to the total ΔF .

2.1.1. Contributions of $\Delta F_{RB}(L)$ to the change of free energy

The rigid body term for the ligand is mostly entropic in nature, with $\beta\Delta F_{RB} = -\left(\frac{\Delta S_{trans}}{k} + \frac{\Delta S_{rot}}{k}\right)$. Here, $\beta = \frac{1}{kT}$ and $\frac{\Delta S}{k}$ values are dimensionless. We use the dimensionless form instead of the conventional entropy units (e. u.) with 1 e.u. = 1 calorie/mole/Kelvin. The translational entropy loss $\frac{\Delta S_{trans}}{k}$ of a ligand upon binding, with the protein being stationary, is estimated by Amzel to lie in the range 4.5-7 [14]. An equal loss of entropy from the three degrees of freedom for rotation would render the sum $\beta\Delta F_{RB} = -9$ to -14. A value of as large as -25 has also been reported [15]. For the Barnase/Barstar complex, Janin calculated [16] a rotational loss $\frac{\Delta S_{rot}}{k}$ of -11. For the BPTI pancreatic inhibitor in the trypsin-BPTI complex [17] , Finkelstein and Janin obtained for $\frac{\Delta S_{rot}}{k}$ values in the range -24 to -30. The $\frac{\Delta S_{trans}}{k}$ value was determined to be -22.10 and the $\frac{\Delta S_{rot}}{k}$ value to be -24.04 for Ras/Raf complex by Gohlke and Case [18]. The $\frac{\Delta S_{trans}}{k}$ value is determined to be -25 ± 5 on average for protein-protein interactions [19]. Grünberg calculated [20] the $\frac{\Delta S_{rot} + \Delta S_{trans}}{k}$ value in the range -50 to -55 for complexes α-chymotrypsiongen/ Pancreatic secretory trypsin inhibitor; Humanized anti-lysozyme Fv/lysozyme; Anti-lysozyme antibody Hyhel-63/Lysozyme; Barnase/Barstar; HIVB-1 NEF/FYN tyrosin kinase SH3 domain; Uracil-DNA glycosylase/inhibitor; Ras activating domain/Ras. There is a small degree of dependence of translational and rotational entropies on the size of the molecules. But, irrespective of size dependence, the range of values reported is too large, and good experiments seem to be the only way to resolve this issue. It should be noted that there have been criticisms of these values on the grounds that Sakur-Tetrode equation, and/or similar approximations have been used [21-22]. A critical review of the sources of errors in estimates of free energy changes upon binding has been given by Gilson *et al.* [23].

2.1.2. Contributions of $\Delta F_{conf}(L)$ to the change of free energy

Karplus *et al.* [24] calculated the change of the configurational entropy $\frac{\Delta S_{conf}}{k}$ per residue upon denaturation of a native protein into a random state to lie between 2 to 3. The value of unity that we obtained for the binding of a freely jointed peptide is in general agreement with their result. These values may be compared with estimated entropy changes of folding and secondary structure formation in order to gain further insight to the problem. In the crude approximation that folding of a peptide to a single conformational state is similar to the binding of the peptide from many configurations to the single conformation in the bound state on the protein surface, the loss of entropy $\frac{\Delta S_{conf}}{k}$ due to side chains only may be estimated to be around -1.5 per residue [21]. Lim *et al*. calculated [25] entropy of the peptides binding to PABC domain of HYD protein due to the counteracting contributions of $\frac{\Delta S_{conf}}{k}, \frac{\Delta S_{solv}}{k}$ to be in range of 0.52 and -0.7. Frederick *et al.* [26] determined maximum configurational entropy $\frac{\Delta S_{conf}}{k}$ to be maximum -1.51 and minimum 0.03 for good binding peptides of Calmodulin.

2.1.3. Contributions of $\Delta F_{binding}(LP)$ to the change of free energy

We take the binding free energy to be mostly energetic, $\Delta E_{binding}(LP)$ and ignore entropic components. $\Delta E_{binding}(LP)$ includes the van der Waals, electrostatic and hydrophobic, steric, hydrogen bonding, ligand strain and enzyme strain interactions between the peptide and the binding location on the protein [1, 7, 9, 13, 15, 27-28]. Binding interactions seem to be a significant factor and will therefore be discussed here in some detail. An excellent review of the subject is given by Gohlke and Klebe [7]. The upper limit of electrostatic contributions to $\beta\Delta E_{binding}(LP)$ is estimated as -11.2 at 310 K. The contribution of hydrogen bonds and salt bridges to $\beta\Delta E_{binding}(LP)$ varies in the range -1.0 to -3.0. For charge assisted hydrogen bonding and salt bridges, these values increase to -3.5 to -7.7. The contribution of hydrophobic interactions in ligand binding is proportional to the size of the hydrophobic surface buried. Experiments lead to values of $\beta\Delta E$ in the

range of -0.04 $Å^{-1}$ to -0.05 $Å^{-1}$ [28-31] the burial of a methyl group which buries about 25 $Å^2$ of surface, contributes about -1.2 to $\beta\Delta E$ [7].

Desolvation free energy of approximately 0.04 - 0.12 per methyl group is estimated [32]. Kasper described a method [33] to calculate the relative free energies of protein-peptide complexes. The peptides of 7 amino acids bound to DnaK molecular chaperone showed that electrostatic contribution varies in range -2.9 to -6.13; nonpolar binding free energy according to the $Å^2$ of surface of peptides varies in range -3 to -4 [34]. Calorimetric analysis of peptide binding to OppA protein by Sleigh *et al.* [32] indicated that $\beta\Delta E_{binding}(LP)$ varies in the range -8.1 to -10.2.

2.1.4. Contributions of $\Delta F(W)$ to the change of free energy

As a first approximation one may assume that **n** water molecules are localized on the protein surface, which are released upon the binding of the peptide. A single bound water molecule released from the crystalline conformation contributes +1 to $\frac{\Delta S_W}{k}$ [22], where ΔS_w is the change of entropy of water molecules.. A water molecule is transferred from the binding site of a protein upon binding of a small molecule with maximum enthalpy of -6.16 and maximum entropy of -3.47 [32]. Favorable free energy of interaction of water molecule in the binding site of a protein is -2.7. Entropy gain for release of one water molecule to bulk water from the binding site is determined to be 3.35 [35].

2.1.5. The lacking part of the literature

The statistical thermodynamics features of the unbound peptide as a small thermodynamics system in a thermal reservoir is lacking in the literature. In view of the fact that each repeat unit of a peptide can be one of the twenty different amino acid types, and that there are 10^{17} different possible sequences even for a short peptide of seven residues, it is obvious that the calculation of the statistical thermodynamics properties of peptides is a difficult but straight-forward process. Each amino acid residue may be in a large amount of states depending on the values assumed by their two backbone torsion angles; the ϕ and ψ angles, or the Ramachandran variables [6, 8]. Since a peptide may be in any of the possible configurations in the unbound state in the aqueous environment, and since transitions from one conformation to the other takes place rapidly in around the picoseconds range, we need to consider the average conformational energy and the entropy

of the peptide when a free-to-bound state transition is under consideration. The conformational heat capacity of the peptide in solution is needed to be known to understand the stability of the unbound state. In order to calculate these thermodynamic properties, torsion angle data, which may come either from molecular dynamics simulations [36-37] of amino acids in solution or from knowledge based potentials [38-40], should be known. In this dissertation, it is attempted to formulate the statistical thermodynamics features of peptides using knowledge based data relevant to the denatured states of proteins.

2.2. *De novo* Peptide Design:

The determination of a specific peptide sequence with affinity to a particular protein surface is a problem of high degree of complexity arising from the fact that each residue of the peptide could be chosen from a pool of twenty natural amino acids. Even for a peptide with three amino acid residues, there exist $8x10^3$ possible peptide sequences. Screening of such a large number of molecules is complicated with both experimental and computational techniques. A rational methodology for specific and selective peptide sequence prediction is required. The necessary methodology should be time-efficient and be able to design peptides for given locations on a given set of proteins. A fast and global computational tool is desired.

The complete peptide binding problem can be visualized as a three step process: (i) finding a path on the surface of the protein which defines a suitable region for the peptide to bind, (ii) finding the appropriate peptide for this path, and (iii) improving the peptide for a more stable binding required for inhibition. In some cases, the binding surface is known, and a peptide must be designed *de novo*. In other cases, a peptide is given and the best region on the surface that gives the optimum binding conditions is searched. Generally two techniques are used for the peptide design problem: structure-based design and knowledge-based design. In structure-based design, known drugs / peptides are tried to be bind to protein. If the known structures interact with protein, better structures are designed via using this structure. In knowledge-based design the known interaction of amino acids in databases are used to construct a new peptide for a specific protein. The importance, computational and experimental difficulties and the present state of the art of finding new peptides have recently been discussed and reviewed by Petsalaki et al. [41]. The authors

indicate that a short peptide sequence of a protein mediates the interaction with other proteins. The method of Petsalaki et al. [41], based on a knowledge based bioinformatics approach addressed this problem and could successfully find the binding sites for the peptides. A similar 'indirect' design approach has been adopted by Frenkel et al. [42] , using their *de novo* molecular design computational tool Pro_Ligand [42]. The Pro_Ligand methodology is based on molecular fragment placing on target protein sites via a rapid graph-theoretical algorithm and the pre-screening the library fragments. Known peptides were docked to unknown locations on given proteins by Hetenyi and Spoel [43] using AutoDock [44]. There have been successful attempts for computational peptide design that use knowledge-based search strategies and use diverse sets of statistical descriptors, different training databases, hydrophobicity scales, motif regularities, etc. Also, automated peptide binding search techniques [45-47] from known epitopes or protein libraries have been successfully used as bioinformatics tools. There have been applications of bioinformatics computational binding tools such as the sequence moment concept, artificial neural networks, fuzzy neural networks and Hidden Markov Model for checking the suitability of inhibitory peptides for binding on MHC class II proteins [48-57]. Along similar lines, the suitability of a ligand as a drug was tested using Bayesian neural network analysis [58]. Application of genetic algorithms to the design of peptides has been an important line of research, examples of which are: *in silico* peptide screening and application of genetic algorithm to determine inhibitory peptide against Parkinson's-disease-related protein α-Synuclein [59]; peptides as thrombin inhibitors [60-61]; evaluation of energies for peptide binding to a user-defined protein surface patch via Genetic algorithm for p53, oligopeptidase and DNA gyrase [62]; integer linear programming [63]; design of hexapeptides against stromelysin protein by Singh et al. [64], and a peptide buildup approach together with a genetic algorithm [65-67]. The prior knowledge of the structural basis of peptide-protein binding strategies is extremely important.

The recent paper by London et al. [68], that reviews the strategies important for understanding the basis of the peptide interactions and peptide design. London et al. indicates that peptides are stable in their free states and able to stabilize the unstable proteins upon binding. This behavior can be observed in MHC proteins, which are stabilized by peptides. The authors state that the protein site, which a peptide can bind to, is

predefined. The peptides scan protein surfaces for an available binding pocket and prefers largest binding pocket. After finding an appropriate site, the peptide makes interactions with hot-spot residues. For each 3 residues of a peptide 1 hot-spot is observed on the protein binding pocket. Numerous hot-spots are necessary for a long peptide to overcome configurational entropic cost; consequently large peptides are harder to design. The importance of an appropriate binding site is highlighted.

As stated before, the determination of an appropriate peptide sequence for a given protein surface is difficult given that each residue of the peptide can choose any of the twenty amino acids. In such problems, where a solution is hard to discover among many possibilities, Hidden Markov model or heuristic techniques can be applied. Heuristics is the general name for methods that attempt to find the optimum or best possible solution to a problem, using the known properties of the given problem. Hidden Markov model (HMM) is a statistical model in which the system assumed to be a Markov process with unobserved states. Markov process can be defined as a stochastic process with the following properties: given the present state, future states are independent of the past states; at each step the system may change its state from the current state to another state, or remain in the same state, according to a certain probability distribution. In this dissertation the genetic algorithm as heuristics method, a Markov model and Hidden Markov model deciphered by Viterbi algorithm are implemented to the *de novo* peptide design problem.

Once the appropriate peptide sequence is determined, the interaction between the selected target protein and the peptide should be determined via experimental and computational means. The experimental techniques are obligatory to conclude that there is an interaction between selected molecules, but the time-limits for those methods lead to a pre-calculation of interaction with computational tools. Once the computational results are complete, the results may be confirmed by reduced number of experiments. The computational interaction search for possible affinity between a protein and a peptide can be achieved by molecular docking studies. The docking method can be stated as the prediction of the most possible orientation of two interacting molecules [69]. There are various tools that are able to make docking calculations, most of which are academic-free. In this dissertation, AutoDock [44] and GOLD [70] tools are used for docking. AutoDock calculates the hydrogen-bonding, van der Waals, electrostatics, desolvation energies

between protein-peptide complexes and also the peptide intrinsic energies. GOLD can calculate hydrogen-bonding, van der Waals and hydrophobic energies of the protein-peptide complexes.

2.2.1. Genetic Algorithm:

The search space of the peptide design problem is very large, due to the possible sequences and conformations of peptide. Genetic algorithm (GA), which is a population-based metaheuristic optimization algorithm, s ideal for cases that deterministic or analytic methods fail, i.e. problems with underlying mathematical model is not well defined or the search space is too large. Consequently, using GA is very proper for the peptide design problem. GA is inspired by biological evolution: it uses reproduction, mutation, recombination (cross-over) and selection mechanisms. The individuals of the population are represented as chromosomes made up of genes. Darwinistic genetic algorithm is the original genetic algorithm proposed by Holland [71] that utilizes a user-defined population of solutions to determine superior solutions. GA is inspired from the biological evolution: the individuals of the population symbolize the chromosomes; the reproduction, mutation, cross-over and selection mechanisms are employed as operators of the algorithm (Figure 2.1).

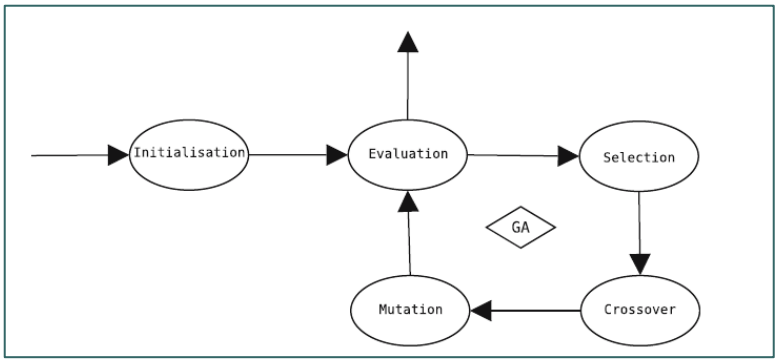


Figure 2.1. Overview of GA.

Other evolutionary algorithms can be listed as: Lamarckian genetic algorithm (LGA), Population-based incremental learning (PBIL) and Bayesian optimization algorithm (BOA) [72]. GA and other evolutionary algorithms are used for peptide design by various groups: Belda et al. [72], Kamphausen et al. [73], Abe et al. [74], Budin et al. [75-77], Schneider et al. [78].

Belda et al. [72] employed different evolutionary algorithms, namely GA, Lamarckian genetic algorithm (LGA), Bayesian optimization algorithm (BOA), and population-based incremental learning (PBIL) to design inhibitor peptides. AutoDock program [44] was used to determine the interaction energy of designed peptide and target proteins. LGA is very similar to GA; the only difference is that it uses a local search method. PBIL is an abstraction of GA; it does not use mutation or cross-over but a statistical vector instead. The genes are represented in binary format (0, 1). The statistical vector consists of the frequency of each gene of the best individuals. Each i^{th} gene of the population is summed and then divided into population-size, a statistical vector is formed. BOA is quite similar to PBIL. Instead of a statistical vector it uses a *Bayesian Network*. A Bayesian network is a directed graph. The nodes are the genes of a chromosome, and the edges indicate the dependencies that exit between those genes [72]. LGA, PBIL and BOA were tested on three different cases, and LGA determined to be the best algorithm. The randomly designed peptides were also seemed to bind protein surfaces with high affinity. This can be due to the limitation of amino acid set. Not the 20 amino acids were used, but a limited set of 6-7 amino acids were used instead. The limitation is inconsistent with the logic of *de novo* design. But the limitation improves the possibility of finding a good peptide candidate in a small period of time. Accurate heuristics seem to be more reliable than rapid heuristics if we compare the results of LGA, PBIL with BOA. The BOA gives fitness values worse than the others. This can be due to time limitation, since less time is used by rapid heuristics for BOA that requires a large population size to infer proper Bayesian network. The disadvantages of their method are that the algorithm uses only limited number of amino acids – not all 20 amino acid types, and the algorithm runs for 2 weeks on 8 processors, which is quite a long period of time.

Kamphausen et al. [73] implemented GA for RNA folding and Thrombin inhibitor peptide design problems. Their work is exciting, since effective peptide inhibitors can be determined in 5 cycles of GA. Known thrombin inhibitors that are highly active were used as the initial population for GA. 30 different building blocks and 16 amino acids were utilized as the fragment library. The kinetics analysis for the peptides was completed and nanomolar binder peptidomimetics were obtained.

Abe et al. [74] also used GA to screen peptide ligands against a Parkinson's-disease-related-protein and achieved peptides with affinity towards target protein after 6 cycles. A

simple GA was implemented with single amino acid mutation, only the superior peptides of each generation were used for new population generation. The binding site of the target protein is highly hydrophobic; consequently *in vitro* screening could not be carried. For this reason authors applied *in silico* screening. Only 10 amino acids we selected 10 amino acids: glycine, alanine, valine, serine, threonine, proline, glutamine, asparagine, phenylalanine and tyrosine were used for GA.

Budin et al. [75-77] utilized GA to build possible binder molecules via combining molecular fragments. An energy function, which takes into account the electrostatic salvation effects, was used to measure the affinity of fragments for target protein. The fragment conformations on binding site were optimized and combined to have a globular ligand structure. The model was tested on thrombin, HIV-1 protease and caspase.

Schneider et al. [78] utilizes a systematic way for peptide design. Initially a seed peptide with ability to bind the target protein is necessary. Variant of the seed peptide was generated by their algorithm, the variants were synthesized and testes on the protein. The quantity sequence-activity relationship was used to train an evolutionary algorithm. E method was applied to determine novel peptides that prevent the positive chronotropic effect of anti-β_1-adrenoreceptor autoantibodies from the serum of patients with dilated cardiomyopathy. The designed peptides with desired activities did not match the seed peptide sequence.

2.2.2. Markov Model and Hidden Markov Model:

Markov chain is a stochastic process, it is a movement from one state to the next state with transition probabilities, over a defined set of states; each move is named as step. Assuming that the current state is $\boldsymbol{S_i}$, at the next step it moves to $\boldsymbol{S_j}$ with a probability $\boldsymbol{P_{ij}}$ [79]. Markov model could be applied to systems with Markov property. The Markov property is defined for stochastic processes with the properties: given the present state, future states are independent of the past states; at each step the system may change its state from the current state to another state, or remain in the same state, according to a certain probability distribution [80].

A specific type of Markov model, Hidden Markov model (HMM) is widely being used in analysis of the biological data and in the bioinformatics area. HMM is a stochastic model in which the system assumed to be a Markov process with unobserved states [80]. HMM is a discrete-time Markov model. When a state is visited by the Markov chain, this state emits

a letter from a pre-defined alphabet. The emissions are time independent. If the sequence of states are denoted by $\mathbf{q_1}$, $\mathbf{q_2}$,… $\mathbf{q_n}$ then the emitted states are denoted by $\mathbf{O_1}$, $\mathbf{O_2}$, … $\mathbf{O_n}$. When a sequence of states is visited a two-step process is observed:

$$\underset{q_1}{\text{initial}} \rightarrow \underset{O_1}{\text{emission}} \rightarrow \underset{\text{to } q_2}{\text{transition}} \rightarrow \underset{O_2}{\text{emission}} \rightarrow \underset{\text{to } q_3}{\text{transition}} \rightarrow \underset{O_3}{\text{emission}} \rightarrow \cdots$$

The **O**'s are known but the **q**'s are hidden, the aim is to dechipher the **q** values given a known **O** sequence. The problem can be defined as

$$\arg\max_{Q} \Pr ob(Q \mid O) \qquad (2)$$

Figure 2.2 represents a frequently used example for HMM: the observed states (**O**) for weather are dry, damp and foggy and the underlying conditions (**Q**) are sunny, cloudy and rainy sky. The most probable sequence of the observed and the hidden states can be determined via listing all possible combinations Pr (Q | O). The computational cost of probability calculation is high for the problems with numerous observed/hidden states.

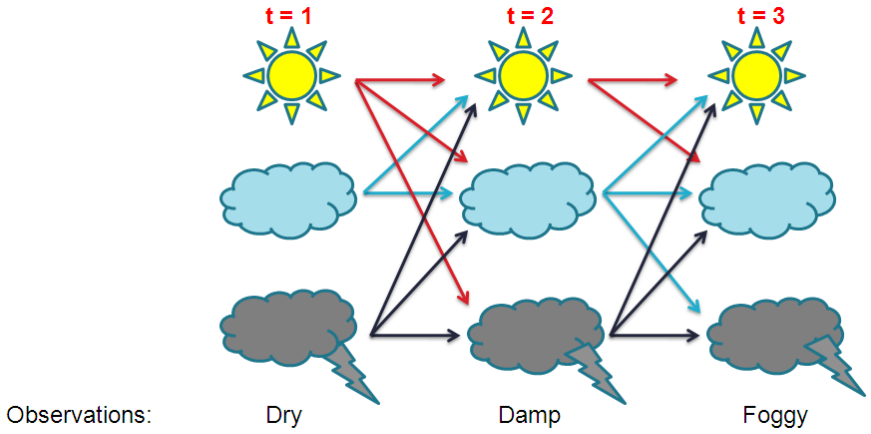


Figure 2.2. The hidden and observed states for the weather problem are depicted. The arrows represent probability of passing through one hidden state to another.

The Viterbi algorithm is an efficient way of determining the best state solution of a hidden Markov model based on a given observation sequence [81-82] and is being used widely in the analysis of biological data and in bioinformatics area [83]. Viterbi decoding reduces complexity via recursively finding the most probable sequence of hidden states given an observation sequence and a HMM. Partial probability **δ** is defined for each particular intermediate state. Each **δ** represents the probability of the most probable path to

a state at time t, and not a total. **δ(i,t)** is the maximum probability of all sequences ending at state i at time t, as shown in Figure 2.3. Partial probabilities **δ**(i,t) are determined for each intermediate and end state. The most probable sequence of states is selected from **δ**'s via maximizing the probabilities. The details, which are described according to our peptide design problem, are given in Chapter 5.

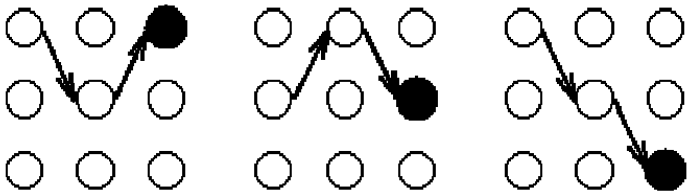


Figure 2.3. δ(i,3) are shown for three different **i** values, **i** = 1,2,3. The scheme on the left indicates the most probable path for **i** = 1, in other words **δ(1,3)**. The middle scheme represents **δ(2, 3)**. The scheme on the right shows **δ(3,3)**.

HMM is utilized in protein structure prediction [84], analysis of protein, RNA or DNA sequences [85], novel peptide hormone prediction via training the HMM algorithm with known receptor protein-peptide hormone interactions [86]. Kobayashi et al. employed HMM on 3 different studies for peptide design against MHC class II proteins: training was achieved by non-binding and binding peptides of target proteins and tests applied to different data-set denoted that method is able to predict binder peptide sequences [52, 55, 87].

The various bioinformatics methodologies that are based on hidden Markov models are summarized by Bystroff et al. [84]: the topology prediction for the transmembrane proteins, the signal peptide determination, and secondary structure prediction methods are summarized. Analysis of protein, RNA or DNA sequences were also achieved by the Viterbi algorithm as illustrated by a study on gene finding from DNA sequence by Sramek et al. [85]: the algorithm enables gene finding, promoter detection, and detection of conserved elements. The prediction of the topology of all-beta membrane proteins combining Viterbi and posterior algorithms was proposed [88] as a modified Viterbi algorithm.

Mirabeau et al. [86] determined two novel peptide hormones by the Viterbi algorithm by training their algorithm with known receptor protein-peptide hormone interactions and testing the model on peptide sequences in databases. The peptide hormone properties are utilized to predict whether a protein/peptide contains a peptide hormone sequence or not. The alignment of known peptide hormones to the top ranking proteins resulted in the identification of two novel candidate peptide hormones.

Noguchi et al. employed the Viterbi algorithm on 3 different studies for peptide design against MHC class II proteins [52, 87, 89]. Training was achieved by non-binding and binding peptides of the target proteins and tests applied to different data-set indicated that the method is able to predict binder peptide sequences.

Viterbi algorithm was adopted by Sonmez et al. [90] to deteremine peptide hormones that interact with G-protein-coupled receptors (GPCRs). Sequence alignment along evolutionary path structure across species was utilized to train their HMM-based computational approach. In other words cross-genomic sequence comparisons are their main methodology. The sequences from Swiss-Prot were the input of the algorithm. The basic molecular biology techniques were employed to verify computationally determined sequences have biological significance. A previously unidentified neuropeptide was discovered, which is mainly localized in brain.

All of the peptide design studies with HMM or Viterbi decoding in literature are based on a training data set. No *de novo* approach using Viterbi decoding has been used for the peptide design problem up to date.

2.3. Peptide Motifs:

The peptides are used as inhibitor / activator of target proteins in order to cure various diseases. Although there has been a wide interest in peptide design, there is no general tool to determine peptide motifs for protein surfaces. The peptide motifs are crucial for the selective and specific binding. There have been successful attempts to discover biological motifs by different research groups. Most of the efforts are to determine specific amino acid or nucleic acid motifs that have been conserved by evolution. Motif search for specific proteins are available in the literature, especially for MHC class proteins and SH3 domains. To our knowledge, the efforts in the literature are to determine the evolutionarily conserved motifs from proteins / DNA / RNA or to discover the binding motifs by means

aligning the binder sequences of the target protein. The evolutionary motif search methods, the peptide motifs discovered for specific proteins and some algorithms / servers / software to determine peptide motifs are reviewed. Neduva et al. provide an excellent review of linear motifs [91]. As staded in ELM server reference [92] the earliest linear motif definition was written by Tim Hunt [93]:

The sequences of many proteins contain short, conserved motifs that are involved in recognition and targeting activities, often separate from other functional properties of the molecule in which they occur. These motifs are linear, in the sense that three-dimensional organization is not required to bring distant segments of the molecule together to make the recognizable unit. The conservation of these motifs varies: some are highly conserved while others, for example, allow substitutions that retain only a certain pattern of charge across the motif.

The short lengths of linear motifs, three to ten amino acids, make them difficult to discover. It has been experimentally shown that the linear motifs do not have strong affinities for their targets, the lowest affinity values are in range of 10 μM [91] . The affinity values indicate that linear motifs make transient interactions with their partners. Same or similar linear motifs are used in diverse species, but conservation over long evolutionary distances is uncommon. Neduva et al. [91] explains the reason for this fuzzy behaviour of nature with plasticity. The point mutations can activate or deactivate linear motifs easily leading to fast adaptation. The loss or gain of a linear motif may not lead to severe influences due to weak affinity of motifs. However, a point mutation on an essential residue of the motif may lead to formation of different biological pathways through the evolution. Saito et al. [94] verified that peptide motifs play role in the plasticity and the evolution of protein network. The group developed an artificial protein library by mixing four peptide motifs of Bcl-2 family proteins. The motifs regulate the apoptosis network in positive or negative fashion. The artificial proteins in the library have pro- and anti-apoptotic properties. Their findings present that common motifs may lead to formation of proteins with divergent attributes.

As stated before, there are several attempts to discover peptide motifs for specific proteins / protein domains. Amy Hin Yan Tong et al. [95] devised a combined strategy to determine peptide recognition modules. The modules recognize and specifically bind to particular peptide motifs. The method described utilizes both experimental and

computational procedures. The phage display library and two-hybrid physical interaction test construct the experimental part. The computational approach is to determine interaction networks for the results of the experimental procedure. The method was successful to determine proline-rich motifs for the SH3 protein domains. Authors imply that their methodology can be used for other peptide recognition modules. Landgraf et al. [96] applied a similar study for SH3 domains with very valuable outcomes. The group combined phage display library and SPOT synthesis. Their aim was to determine all peptides of the yeast proteome with potential affinity for eight SH3 domains. The experimental studies lead to formation of five classes of peptides with overlapping specificity for SH3 domains. Authors imply that the strategy can be applied to the human proteome. Zhao et al. [97] explained the binding of numerous proteins to the catalytic subunit of protein phosphatase-1 with a common motif on the binder proteins. Random peptide library panning was used to determine the binding motif, (2–5 basic residues) VX (F/W) (an acidic residue). Corr et al. [98] identified a binding motif for H-21Dd protein, XGPX(a positively charged residue)XXX(I/L/F)(I/L/F)(a hydrophobic residue). Tan et al. [99] analyzed scorpion toxins by mutation analysis and determined several peptide motifs for ion channels.

There are several web-servers available to predict whether a given sequence is a peptide motif or not. The servers are based on database searches. ELM server [92] presents 146 known motifs, which are experimentally determined. A protein sequence is given as query to server (http://elm.eu.org) and then search for linear motifs on the sequence, complex structure, and related literature are established. Neuwald et al. [100] developed a detection method for motif-encoding regions in sequences. The model is tested on immunoglobulin fold proteins and a set of distantly related 32 bacterial membrane proteins. The method is able to identify very short motif segments that could not be determined by alignment tools available. Thompson et al. [101] described Gibbs Centroid Sampler software that detects conserved regions of biopolymers via centroid alignment. The method is able to predict RNA secondary structure and motif detection (http://bayesweb.wadsworth.org/gibbs/gibbs.html). Rigoutsos and Florafos [102] offered a combinatorial algorithm to determine motifs for a given set of protein sequences. Their methodology assures that the motif is maximal in length and composition. Reiss et al. [103] developed a computational *de novo* method to identify peptides, which bind to

specific peptide recognition modules. The protein sequences and experimentally observed interactions are necessary for this probabilistic method (http://sf.net/projects/netmotsa). The method was tested on SH3 domains and shown to be specific and sensitive. The authors state that their methodology is general for any given peptide recognition module domain. Reche et al. [104] developed a new search algorithm RANKPEP that utilizes a position specific scoring matrix (PSSM) in order to identify binding motifs for MHC class I proteins. The *MHCPEP* database with MHC-binder peptides was used for the methodology. The binding potential of a given peptide sequence for a given MHC protein is obtained from ranked and aligned peptides by PSSM. The prediction for any peptide sequence can be completed at RANKPEP web-server (www.mifoundation.org/Tools/rankpep.html). Reche et al. [105] extended RANPEP for peptide motif prediction of MHC class II proteins. Rajapakse et al. [106] utilized genetic algorithm to determine peptide motifs for medically important protein targets. The method requires experimentally derived peptide motifs and the experimental motifs are used to determine a consensus motif. The method was tested on the MHC Class II molecule I-Ag7 of the non-obese diabetic mouse. The determined motif is better than the experimental motifs and the computationally derived motifs by MEME, RANKPEP and Gibbs Motif Sampler. Dinkel et al. [107] developed a method to identify functional motifs in proteins. The conserved homologous sequences are considered for detection of motifs. The method was tested on experimentally confirmed motifs and found to be successful. The method is able to identify short motifs due to its enhanced filters. Neduva et al. [108] offered a strategy to detect binding motifs for certain proteins. Their methodology utilizes data from genome-scale interaction studies. The over-represented motifs in non-homologous sequences are discovered by the method. The researchers were successful to identify known motifs and new motifs. Savoie et al. [109] described an algorithm to determine peptide motifs for MHC proteins. The computational learning algorithm is named as BONSAI, it utilizes a database of 13000 MHC binding peptides with both positive and negative T cell responses. The authors were successful to reveal motifs.

There is no general methodology to discover a binding peptide motif for any protein target, according to our knowledge. The particular aim of this dissertation is to predict a binding motif for any given protein using no prior information.

Chapter 3

CONFORMATIONAL ENERGIES AND ENTROPIES OF PEPTIDES, AND THE PEPTIDE-PROTEIN BINDING PROBLEM

A novel statistical thermodynamics approach is applied to the free peptide segments in order to classify them according to their conformational energies and entropies and heat capacities. The approach employs the rotational isomeric states (RIS) model in which the states are described by the Ramachandran map of backbone torsion angles. The statistical weight matrices for the pairwise dependent states are derived from the torsion angle probabilities of the consecutive dipeptides in coil library. The partition function is determined for a given sequence via RIS multiplication of the pre-determined matrices. The conformational partition function, Helmholtz free energy, energy, entropy and heat capacity are obtained. The model is applied to randomly produced peptides and also to known peptide inhibitors to analyze their thermodynamic properties.

As stated in Chapter 2, the statistical thermodynamics features of peptides are formulated using knowledge based data relevant to the denatured states of proteins. An efficient algorithm to compute the partition function and the thermodynamics of a free peptide segment, as a function of its sequence type, is proposed. Determining the energy of a peptide is important for the reason that it gives information on the relative magnitudes of the energy of the single conformation that is obtained upon binding and the average configurational energy of the peptide in solution. Knowledge of the entropy of a given sequence in solution is important since it is proportional to the amount of entropy that will be lost when the sequence binds to the surface. The conformational energy and entropy of a peptide in solution are significant contributors to the Gibbs free energy, ΔG, of binding. The conformational heat capacity of a free peptide is also of importance because it is a measure of the stability of the peptide in solution.

The conformational state of a residue is defined by a point on the Ramachandran map. The energy of a given state depends on the following factors [6, 8, 110]: (i) the intrinsic torsion potentials of the peptide backbone bonds, (ii) the attractive and repulsive interactions between atoms of side groups that fall in close proximity upon the choice of the ϕ and ψ pair, (iii) the hydrogen bonds that may form between groups of the backbone

atoms and the backbone-side chains, (iv) interactions between atoms of the residue and the atoms of other residues along the peptide. These are contributions to the energy of a state of a residue coming from intramolecular interactions that arise from the atoms of the peptide. Secondly, interactions between the peptide atoms and atoms of the surrounding medium may contribute to the state of a residue. We focused on intramolecular interactions only. The intramolecular forces that stabilize the state of a residue and determine the values of the ϕ and ψ angles are short range if the interacting atoms are within the residue or between its nearest neighbors along the chain. Forces between atoms of residues not adjacent along the chain are termed long range, and may be significant specifically in or around the native state. Secondary structures such as helices, beta sheets, and tight turns are stabilized by long range interactions.

When the state of a residue depends only on the state of its first neighbors along the chain, its conformations may suitably be studied as a Markov process. The Rotational Isomeric State (RIS) model of polymer statistics is a suitable method for the study of Markov processes [6, 110]. We introduced a novel statistical thermodynamics approach, based on the RIS model that can be used for classifying sequences according to their conformational features such as their entropies and energies and heat capacities.

First of all, various regions on a Ramachandran map as states of residues were defined. Then the probabilities of the states of a pair of neighboring residues were calculated using information from the Protein Data Bank. Inasmuch as we are interested in the denatured conformations of peptides, a coil library was constructed from which we obtain the probabilities [39-40, 111-112]. With the probabilities determined in this manner, the statistical weights for the pair-wise dependent states were obtained, and the partition function for a given sequence was calculated using the RIS matrix multiplication scheme. The conformational partition function, Helmholtz free energy, energy, entropy and the heat capacity of the sequence were obtained accordingly.

3.1. The model and calculations

3.1.1. Identification of isomeric states of a residue:

The ϕ and ψ angles of a residue cannot adopt all values due to backbone intrinsic torsion propensities and attractive and repulsive interactions of atoms that are in close

proximity for certain combinations of these two angles. Among the repulsive interactions, steric hindrances resulting from the side groups are the most pronounced. Hydrogen bonds are the most pronounced favorable interactions. Depending on the type of the residue, these angles show preferences for different states, which we term as 'isomeric states' on the Ramachandran map. The frequency of occurrence of states for the twenty amino acids can be obtained from the Protein Data Bank, PDB. An examination of the frequency of occurrence of the states of neighboring units shows that the probability of the states of two neighboring residues along the chain depends strongly on the type of these residues. Plotting the ϕ, ψ angles on a Ramachandran map from PDB data, irrespective of residue type shows that there are 11 major isomeric states. These are shown in Figure 3.1. We number the regions from 1 to 11 according to the notation shown in Table 3.1.

In obtaining Figure 3.1, the non-redundant PDB Select database of native proteins was used. There are 197,458 points on the Ramachandran map obtained from the database. 96% of these points fell on the eleven regions shown in Figure 3.1. The remaining 4% of points were outside of these regions which are known to correspond to strained states of the residues resulting from long range perturbations. We ignored this set of 4%.

Probabilities of occurrence of a residue in any of the eleven states given above depend on the type of the library used. Since we are interested in the denatured conformations of peptides, we calculated the probabilities over a coil library we constructed. The coil library was downloaded from the website: **http://www.roselab.jhu.edu/coil/**. The library contains fragments selected by Dunbrack's PISCES server according to the following criteria: less than 20% sequence identity, better than 1.6 angstrom resolution, and a refinement factor of 0.25 or better.

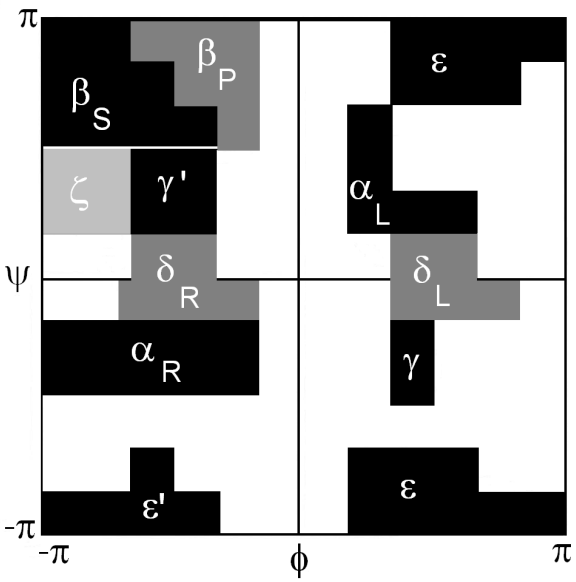


Figure 3.1. Isomeric states on a Ramachandran map.

Table 3.1. Notation for the 11 states.

State	Notation	Description
1	ε'	Mirror image of the extended region ε
2	ε	The extensive regions, $\varphi > 0, \psi \approx \pm 180^o$
3	α_R	Right-handed alpha helix.
4	γ	Tight turn region.
5	δ_R	The right handed bridge region between two β-strands
6	δ_L	Mirror image of the δ_R region
7	ζ	Region observed mostly in residues preceding PRO
8	γ'	Inverse tight turn region
9	α_L	Mirror image of α_R
10	β_S	Extended beta sheet forming region
11	β_P	Region with extended polyproline-like helices

The frequency of occurrence of states was collected in a four dimensional array, $f(i, j, \zeta, \eta)$, which is the normalized frequency that a bond is of residue type i, the succeeding residue is of type j, where the first residue is in state ζ and the succeeding

residue is in state η. When the k-1st residue along the chain is of type i and the kth residue is of type j, the joint probability $p_{k-1,k;\zeta,\eta}$ of having the states ζ and η is

$$p_{k,k+1;\zeta,\eta} = \frac{f(i,j,\zeta,\eta)}{\sum_{\zeta,\eta} f(i,j,\zeta,\eta)} \tag{1}$$

On the left hand side of Eq. 1 we dropped the indices i and j that denote residue types, assuming that the sequence is given and fixed. Thus, it suffices to acknowledge the position dependence with the bond indices k-1 and k. This simplifying notation conforms with previous usage of polymer statistics [6]. For a given k, the matrix $p_{k-1,k;\zeta,\eta}$ is an 11x11 matrix relating the states of two neighboring residues. There are 400 of these matrices for the 20 types of amino acid pairs. In Figure 3.2 we present two sample contour graphs out of the 400. In Figure 3.2a, the k-1st and kth residues are ALA-ALA. In Figure 3.2b, the residue pairs are GLY-GLN. Probability levels are given in four different levels of gray, showing the differences in the conformational propensities of the two pairs.

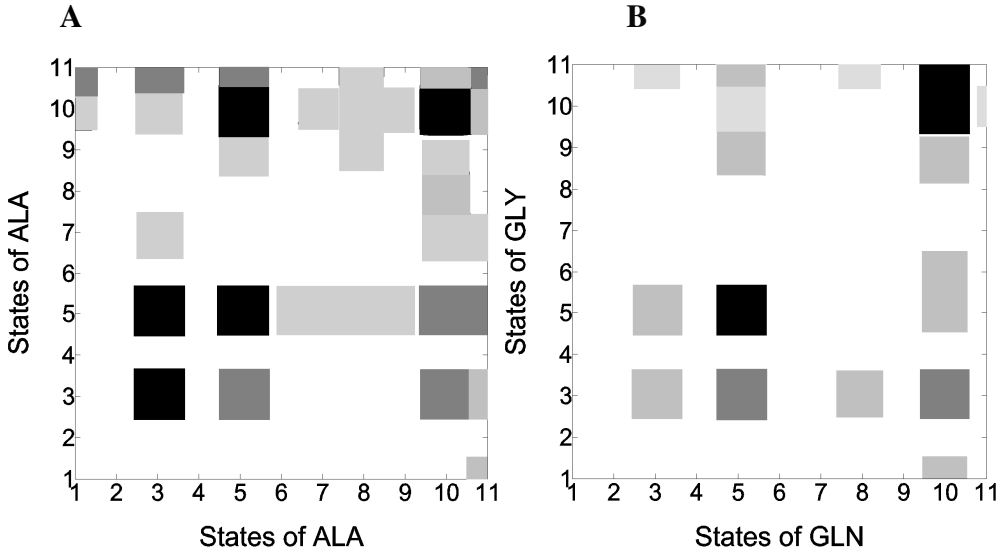


Figure 3.2. State probabilities for ALA-ALA (**A**) and GLY-GLN (**B**). The ordinates denote the states of the previous residue and the abscissa denote the states of the latter residue in the pairs shown. Probability levels are given by four shades of gray.

3.1.2. Matrix multiplication scheme for generating the partition function of a given peptide sequence:

The energy when bond k is in state η and the preceding bond is in state ζ is given by the Boltzmann relation [38]

$$E_{k-1,k;\zeta,\eta} = -\frac{1}{\beta} ln\left(\frac{p_{k-1,k,\zeta,\eta}}{p^0{}_{k-1,\zeta}\, p^0{}_{k,\eta}} \right) \tag{2}$$

where $\beta = 1/kT$, and the probabilities with the superscript zero denote those when the distribution is uniform. For the eleven state model, $p^0{}_{k-1,\zeta} = p^0{}_{k,\eta} = \frac{1}{11}$. The statistical weight matrix U_k for the bond pair k-1, k is

$$U_k = exp\left(-\beta E_{k-1,k;\zeta,\eta}\right) \tag{3}$$

The partition function, Z, of the peptide is obtained according to

$$Z = J^* \left(\prod_{k=2}^{n} U_k \right) J \tag{4}$$

where J^* = [1 1 … 1] ; J = column[1 1 … 1]. (It is to be noted that in the Flory notation, the J* matrix is given as J^* = [1 0 … 0] that assigns the trans state to the first bond. In the present formulation, the choice of the J* matrix allows for the acknowledgement of all of the eleven states to the first residue).

The Helmholtz free energy, $F = -\frac{1}{\beta} ln Z$, energy, $E = -\frac{d}{d\beta} ln Z$, and the entropy [113] $S/k = \beta^2 \frac{dF}{d\beta}$ are obtained in matrix multiplication formalism as

$$F = -\frac{1}{\beta} \ln \left[J^* \left(\prod_{k=2}^{n} U_k \right) J \right] \tag{5}$$

$$E = -\frac{1}{Z} L^* \left(\prod_{k=2}^{n} G_k \right) L \tag{6}$$

and

$$S/k = -\frac{L*\left(\prod_{k=2}^{n} G_k\right)L}{J^*\left(\prod_{k=2}^{n} U_k\right)J} + \frac{1}{\beta} ln\left[J^*\left(\prod_{k=2}^{n} U_k\right)J \right] \tag{7}$$

where

$$G_k = \begin{bmatrix} U_k & U_k' \\ 0 & U_k \end{bmatrix},\ U_k' = \frac{dU_k}{d\beta},\ \text{and L}^* = [\ \ 1\ \ 1\ \ \ldots\ \ 1]\ ;\ \text{L} = \text{column}[\ 0\ 0\ \ldots\ 1\ \ 1\ \ \ldots\ \ 1].$$

The heat capacity, C_v, is given in terms of the partition function as

$$C_v = k\beta^2 \frac{\partial^2 ln Z}{\partial \beta^2} = k\beta^2 \left[-\frac{1}{Z^2}\left(\frac{\partial Z}{\partial \beta}\right)^2 + \frac{1}{Z}\frac{\partial^2 Z}{\partial \beta^2} \right] \tag{8}$$

In terms of matrix notation, C_v takes the following form:

$$C_v = k\beta^2 \left\{ -\frac{L*\left(\prod_{k=2}^{n} G_k\right)L}{\left[J^*\left(\prod_{k=2}^{n} U_k\right)J \right]^2} + \frac{M*\left(\prod_{k=2}^{n} H_k\right)M}{J^*\left(\prod_{k=2}^{n} U_k\right)J} \right\} \tag{9}$$

where,

$$H_k = \begin{bmatrix} U_k & U_k' & U_k' & U_k'' \\ 0 & U_k & 0 & U_k' \\ 0 & 0 & U_k & U_k' \\ 0 & 0 & 0 & U_k \end{bmatrix} \tag{10}$$

and $M* = [1\,1..1 \quad 0...0 \quad 0...0 \quad 0...0 \quad 0...0 \quad 0...0]$

$M = col[0...0 \quad 0...0 \quad 0...0 \quad 0...0 \quad 0...0 \quad 1\,1\,1\,1\,1\,1]$

$$U'' = \frac{\partial^2 U}{\partial \beta^2}$$

3.2. Results

The calculations are performed by randomly generating a sequence of n residues and calculating its partition function, free energy, mean energy, entropy and heat capacity. Peptides of four different lengths with n = 5, 10, 15, and 20 are considered.

We choose the freely jointed model of the peptide as the reference systems for the calculations. For the freely jointed chain all states are equally probable, and the mean energy E may be taken as zero. The entropy is

$$\frac{S}{k} = n\ln(11) = 2.4n \tag{11}$$

In order to refer to the bound state in the discussion of peptides bound on a protein, we also use the firmly bound state, as another reference system. We assume that no conformational transitions can take place and the conformational entropy of the peptide is zero. The conformational energy of a peptide is not unique for this case, and may take a value depending on the peptide type and the configuration of the peptide in the bound state.

Taking the reference state energy of the freely jointed chain as zero leads to negative mean energies for the real peptides with bond energies calculated from the PDB probabilities according to Eq. 2. The energies of the states that are more [114] probable than the random state are negative (positive) according to Eq. 2. Eq. 6 contains $U_k' = \frac{dU_k}{d\beta}$ in each term. The entries of U_k' are of the form $-\beta E_{ij}\, exp(-\beta E_{ij})$ which makes the terms with negative E_{ij} dominate, making the entries positive. The negative sign on the right hand side of Eq. 6 then makes the average energy negative. We let ΔS denote the entropy of the peptide relative to the freely jointed state:

$$\frac{\Delta S}{k} = \frac{S}{k} - 2.4n \tag{12}$$

and ΔE denote the average energy of the peptide relative to the freely jointed state:

$$\Delta E = E - E_{freely\ jointed} = E \tag{13}$$

Here, $E_{freely\ jointed}$ denotes the zero of the energy. Both ΔS and ΔE are negative quantities. We also define the entropy ΔS_b of the peptide relative to the bound state:

$$\frac{\Delta S_b}{k} = \frac{S}{k} - \frac{S_b}{k} = \frac{S}{k} = \frac{\Delta S}{k} + 2.4n \tag{14}$$

where S_b is the entropy of the peptide in the bound state, which we take as zero. Defined in this manner, ΔS_b represents the magnitude of the entropy penalty that will be suffered when a free peptide binds to a protein.

3.2.1. Entropy-energy relations as a function of sequence type

The relationship of $\frac{\Delta S}{k}$ to $\beta \Delta E$ is presented in Figure 3.3 for the four peptide groups, each of length n = 5, 10, 15, and 20. The points are obtained for 10^6 peptides of length n by starting from a random sequence and randomly changing one residue at a time. The points lie on a straight line passing through the origin. The slope $\frac{\Delta S / k}{\Delta E / kT} = \frac{T\Delta S}{\Delta E}$ of the straight line, which is equal to 0.620, is a characteristic property of the coil library and is a result of the energy distribution of the torsional states. If the energies of the eleven states are chosen as random instead of being determined by Eq. 2, the slope of the $\Delta S_b / k$ vs $\beta \Delta E$ line becomes unity. The results of random calculations are shown in Figure 3.4 for the random energies for n = 10. The deviation of the slope of Figure 3.4 from unity is a measure of the non-randomness of the energy surfaces defined by the ϕ and ψ angles.

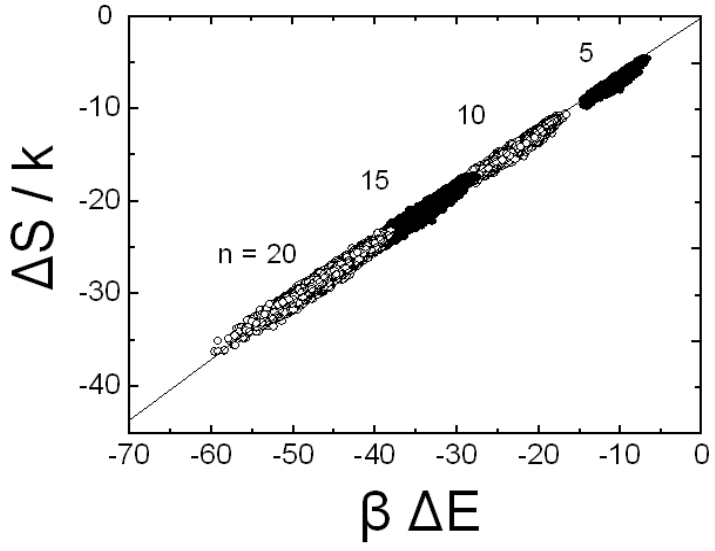


Figure 3.3. Relationship of entropy to energy for different sequence types of four different length peptides.

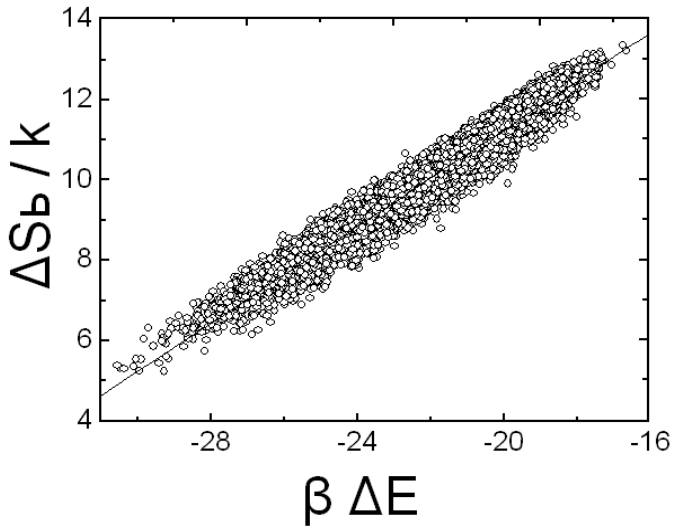


Figure 3.4. Relationship between energy and entropy (relative to the bound state) when the 121 states of the ϕ - φ energy surface are taken randomly. The line through the points with slope = 1.0 is the best fitting straight line through the data points.

In Figure 3.5, the entropy, ΔS_b relative to the bound state is plotted against the average energy relative to the freely jointed state. The points for a given n fall on a straight line. This shows that sequences with low average energies have low entropies. The 'limiting sequence' is represented by the zero entropy intercept. Such a chain has a single conformation that is confined to its minimum energy conformation. When the best fitting lines through the points are extrapolated to zero entropy, the $\beta\Delta E$ intercepts for n = 5, 10, 15 and 20 are obtained as -18.7, -39.2, -59.7, and -80, respectively. The values of the intercepts per residue are approximately constant and are equal to -3.8kT for all n. According to the energies obtained from the PDB data, the sequence that has the lowest energy is all PRO, with an energy of -4.05 per residue and shows a discrepancy of 6% from the extrapolated value of -3.8.

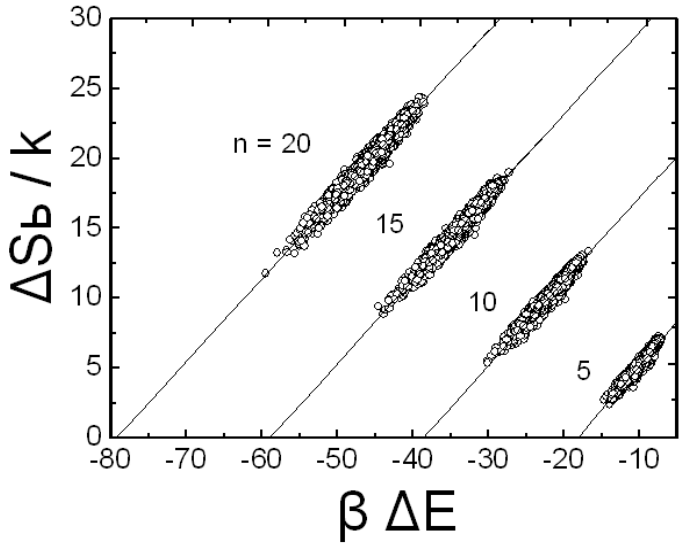


Figure 3.5. Relationship of entropy relative to the bound state to energy for different sequence types of four different length peptides.

Let us take the lowest point in Figure 3.5 on the set n = 10, for example. This is the sequence VAL MET ILE GLU ILE PRO VAL ILE VAL VAL for which $\Delta S_b / k = 5.3$. The entropy penalty per residue to freeze the chain at a single conformation is $T\Delta S = 0.53kT$. The highest point on the n = 10 line, on the other hand corresponds to the sequence GLY SER SER GLY PHE GLN GLY GLY HIS ASP for which the entropy penalty per residue is $T\Delta S = 1.30kT$. Thus the entropy barrier to be overcome for keeping the second sequence at a single conformation is 2.45 times larger than that of the first. As is now known well in the drug design area, and as we show in the discussion part, the entropy barrier plays crucial role in determining the stability of binding. Therefore, choice of the right peptide sequence is of major importance.

3.2.2. Distribution of mean energies, entropies and heat capacities

In Figure 3.6, the distribution of mean energies obtained from 10^6 peptide sequences of n = 10 are presented.

These distributions deviate significantly from Gaussians because they are skewed with longer tails on the low energy and entropy sides. The points are from the results of calculations and the lines are drawn to guide the eye. The mean energy is -24.7 and the peak is observed at -20.52.

In Figure 3.7, the distribution of entropies, relative to the bound state, obtained from 10^6 peptide sequences of n = 10 is presented. The points are from the results of calculations and the lines are drawn to guide the eye. The mean of the graph is 7.6 and the peak is observed at 11.04.

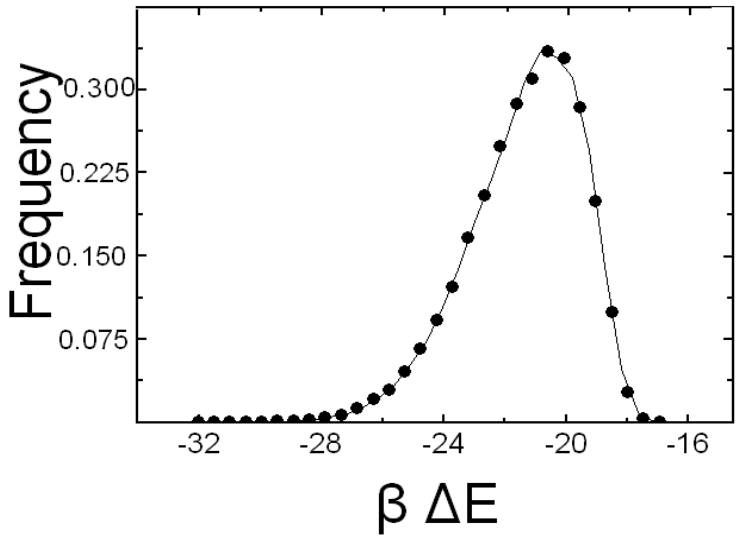


Figure 3.6. Distribution of mean energies for n = 10.

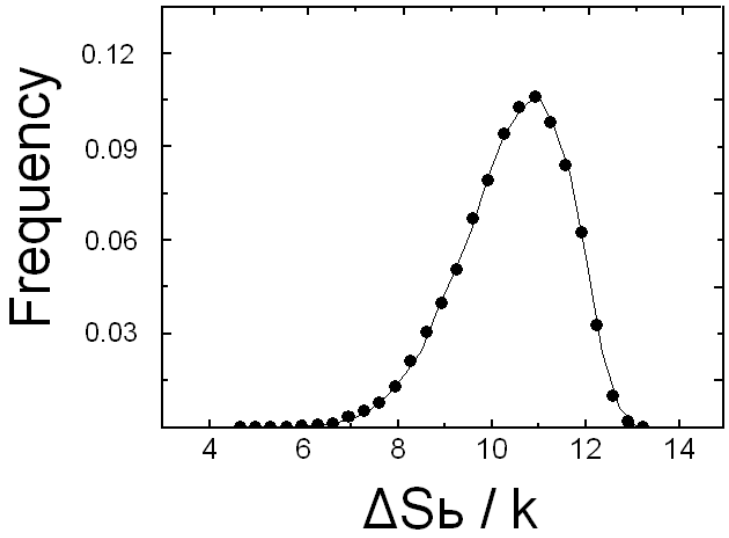


Figure 3.7. Distribution of entropies of peptides for n = 10.

The distributions for both energy and entropy indicate that there is a wide range of these thermodynamic variables depending on the sequence type. The behavior for other sequence length peptides, calculated but not shown in the figures exhibit the same behavior, i.e. the

energies and entropies range over a wide range. The mean energies are approximately -2.5kT per residue which is significantly high when compared to the energy of -3.8kT at the zero configurational entropy intercept discussed above. The mean entropies equate to approximately 0.76k per residue for all sequence lengths.

In Figure 3.8, the distribution of heat capacities for the four different length peptides is presented. The points are from the results of calculations and the lines are drawn to guide the eye. The areas under the curves are normalized to unity. These distributions, unlike the energy and entropy distributions, are skewed with longer tails on the high C_v side, and therefore cannot be represented satisfactorily as Gaussians. The peak per residues number values of the curves are 0.7, 0.69, 0.68, 0.59 for the residue numbers 20, 15, 10 and 5, respectively. The peak of the curves per residue ranges between 0.60 – 0.70. Thus, C_v scales with peptide length.

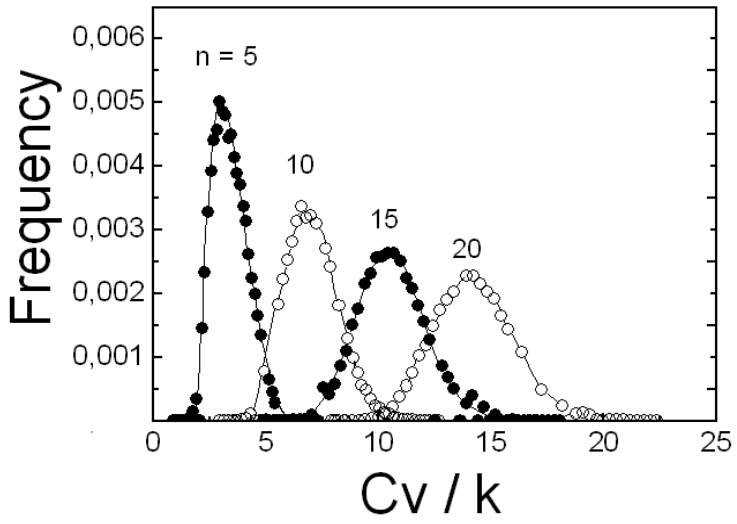


Figure 3.8. Distribution of heat capacities for different length peptides.

At first glance, no functional relationship between Energy and C_v/k, or Entropy and C_v/k exists when $\beta\Delta E$ vs C_v/k and $\Delta S_b/k$ vs C_v/k plots are plotted, as shown in Figures 3.9a and b. In order to see whether sequences with low and high C_v/k values populate different energy and entropy levels than the full set of peptides, we plotted Figures 3.10a and b. In Figure 3.10a, the distribution of $\beta\Delta E$ values for the full set of 10^6 peptides of length n = 10, shown with the open circles, is compared with the distribution of a subset of 11552

peptides whose C_v/k values are less than 4.70, shown with the grey colored circles and also compared with the distribution of a subset of 11944 peptides whose C_v/k values are larger than 9.95, shown with the black colored circles. The low C_v/k peptides have higher energies, meaning that these are closer to the freely-jointed model. On the contrary, high C_v/k peptides have lower energies. Peaks are observed at values -20.52, -20.34 and -23.98 for the all data points, low C_v/k peptides, and high C_v/k peptides, respectively. There is a difference of approximately 4kT between the peak values of the all data points and high C_v/k distributions. In Figure 3.10b, the distributions for $\Delta S_b/k$ for the same sets of 10^6, 11552 and 11944 peptide sequences. Again, the low C_v/k peptides prefer high entropy levels, and therefore are closer to the freely jointed model. The peaks are observed at the values 11, 11.18 and 8.66 for the all data points, low C_v/k peptides, and high C_v/k peptides, respectively.

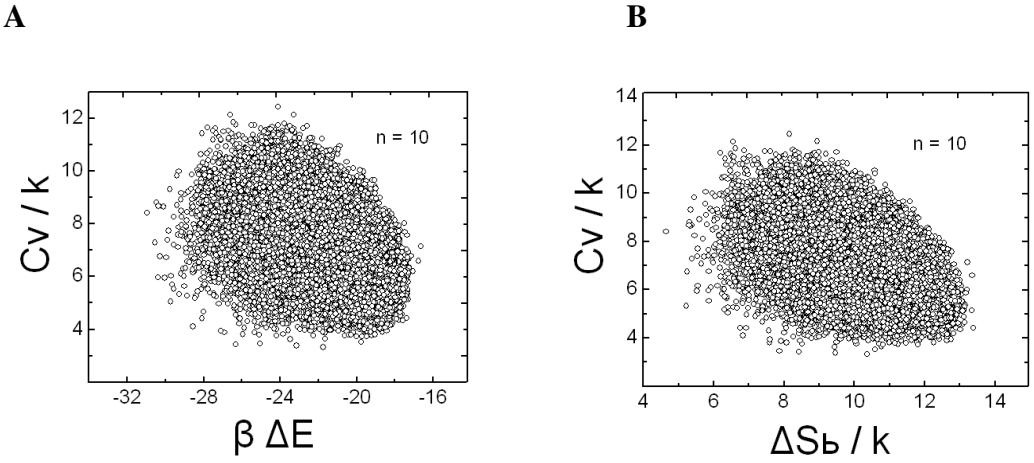
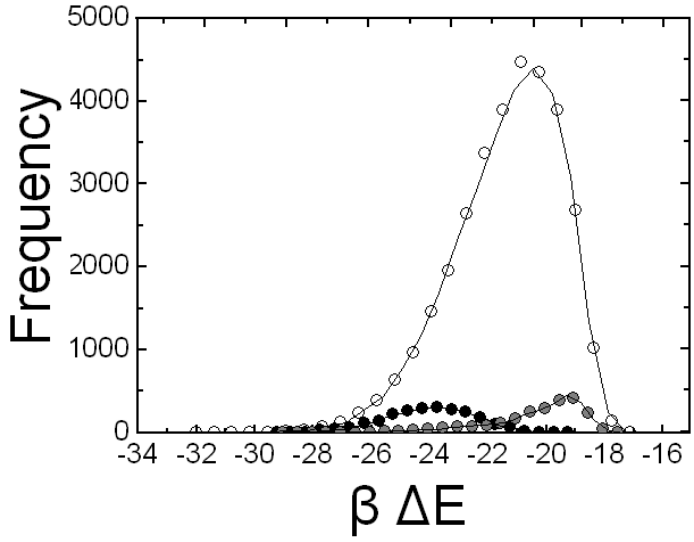


Figure 3.9. Comparison of C_v's with (**A**) $\beta\Delta E$ and (**B**) $\Delta S_b/k$ for n = 10.

A



B

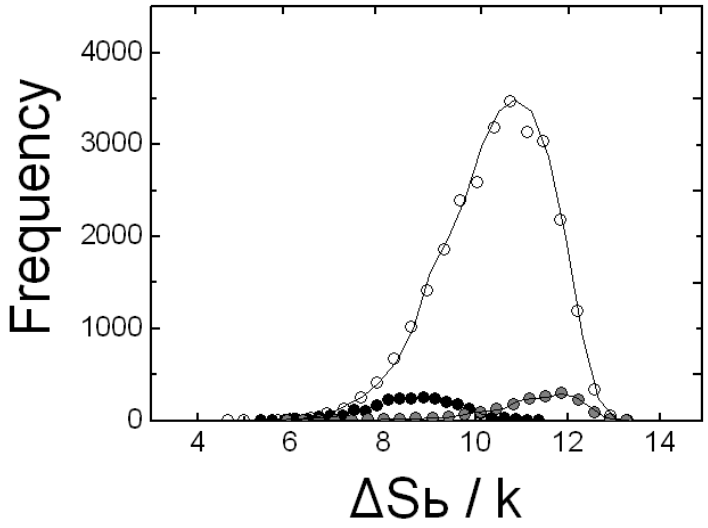


Figure 3.10. (**A**) The distribution of βΔE values for the full set of 10^6 peptides of length n = 10, (open circles), the distribution of a subset of 11552 peptides with low C_v/k values (grey circles), and the distribution of peptides with high C_v/k values (black circles). (**B**) Same comparison for the entropy.

The position of the entropy value of a given peptide in the distribution of entropy values obtained from a large set of peptides for a given value of n may be measured by the ratio r_S defined as

$$r_S = \frac{(\Delta S_b / k) - (\Delta S_b / k)_{min}}{(\Delta S_b / k)_{peak} - (\Delta S_b / k)_{min}} \tag{15}$$

A value of $r_S < 1$ indicated that the entropy is on the low side of the distribution, and vice versa. Similarly, the position of the heat capacity value of a given peptide in the distribution of heat capacity values obtained from a large set of peptides for a given value of n may be measured by the ratio r_C

$$r_C = \frac{(C_v / k) - (C_v / k)_{min}}{(C_v / k)_{peak} - (C_v / k)_{min}} \tag{16}$$

A value of $r_C < 1$ indicated that the heat capacity of the given peptide is on the low side of the distribution, and vice versa. A comparison of r_S and r_C for various peptide drugs shows that in general $r_S < 1$ and $r_C < 1$. In Figure 3.11, the values of r_C and r_S for those drugs are plotted for 31 drugs. The horizontal line indicated the $r_C = 1$ level. One sees all 31 drugs satisfy the $r_S < 1$, $r_C < 1$ criterion.

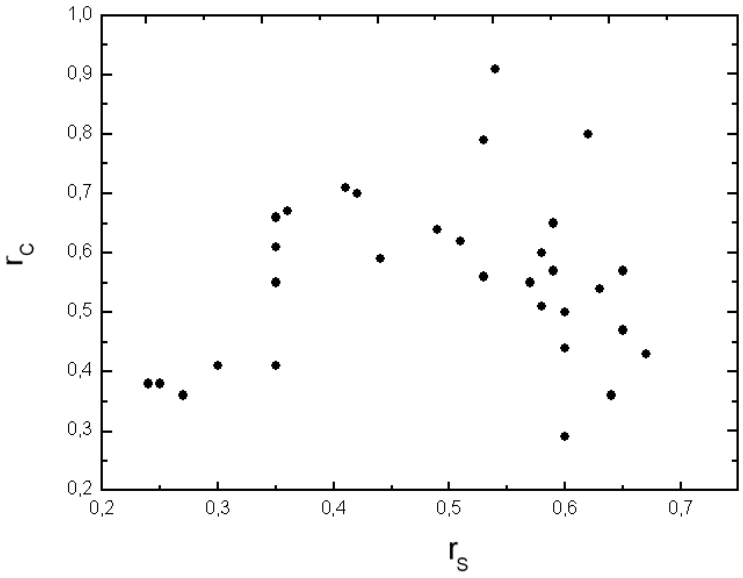


Figure 3.11. The r_C versus r_S values of 31 drug peptides.

The drugs are reported in Table 3.2. These are peptide drugs obtained from three websites:

(1) The KEGG peptide database from GenomeNet, which is a Japanese network of database and computational services for genome research **http://www.genome.jp/en/about.html**

(2) Calbiochem, **http://www.emdbiosciences.com/g.asp?f=CBC/home.html**

(3) The University of Nebraska medical Center Antimicrobial Peptide Database, **http://aps.unmc.edu/AP/main.php**.

The first column shows the number of residues in the peptide. The second column identifies the peptide, the third fourth and fifth columns give the energy, entropy and heat capacity values computed with the present scheme, the last column identifies the sequence. The values of r_S and r_C are indicated following the slashes in columns 4 and 5.

Table 3.2. Drug peptide sequences and their statistical properties.

Peptide Length	**Code**	**βΔE**	**(ΔS/k)/r_s**	**(C_v / k)/r_c**	**Sequence**
8	KEGG C02135	-20.2	6.8 / .41	5.1 / .35	D R V Y I H P F
8	KEGG C16112	-17.1	8.2 / .61	5.1 / .35	F L F Q P Q R F
9	KEGG C07105	-18.6	9.7 / .71	6.4 / .41	C Y F Q N C P R G
9	KEGG C00746	-19.4	9.3 / .66	6.0 / .35	C Y I Q N C P L G
9	KEGG C13662	-18.7	9.6 / .70	6.5 / .42	C Y F Q N C P K G
9	KEGG C00306	-23.4	7.2 / .36	5.4 / .27	R P P G F S P F R
10	KEGG C01505	-25.5	8.3 / .38	5.6 / .25	K R P P G F S P F R
11	KEGG C16008	-27.5	9.2 / .38	6.1 / .24	I S R P P G F S P F R
12	KEGG C16109	-26.9	11.8 / .55	8.1 / .35	M P H S F A N L P L R F
13	KEGG C16046	-28.8	13.2 / .59	8.8 / .44	I L D L W K A I D G L P Y
13	KEGG C02404	-26.0	14.9 / .80	10.5 / .62	G V N D N E E G F F S A R
13	UNMCAP	-26.6	13.3 /	10.1 /	H G V C G H G E H G

	AP00025		.60	.58	V H G
13	UNMCAP AP00159	-30.8	11.7 / .41	7.4 / .30	I L P W K W P W W P W R R
14	CALBIOCHEM 613571	-31.8	13.8 / .62	10.5 / .51	L Q L R D A A P G G A I V S
16	KEGG C16017	-36.0	16.4 / .64	11.9 / .49	Y G G F M T S E K S Q T P L V T
16	KEGG C06006	-35.4	16.5 / .65	12.9 / .59	N D D C E L C V N V A C T G C L
16	KEGG C00952	-33.0	17.9 / .79	12.3 / .53	A D S G E G D F L A E G G G V R
16	UNMCAP AP00218	-38.0	14.4 / .43	13.9 / .67	R G G R L C Y C R R R F C I C V
17	UNMCAP AP00214	-42.2	14.3 / .29	14.3 / .60	K W C F R V C Y R G I C Y R R C R
18	UNMCAP AP00195	-43.2	16.5 / .50	14.6 / .60	R G G R L C Y C R R R F C V C V G R
18	UNMCAP AP00211	-43.8	15.9 / .44	14.6 / .60	R R W C F R V C Y R G F C Y R K C R
18	UNMCAP AP00219	-42.1	17.1 / .55	14.2 / .57	R G G L C C Y C R R R F C V C V G R
18	UNMCAP AP00220	-42.7	16.6 / .51	14.4 / .58	R G G R L C Y C R G W I C F C V G R
18	UNMCAP AP00221	-45.5	15.0 / .36	15.0 / .64	R G G R L C Y C R P R F C V C V G R
18	UNMCAP AP00445	-42.6	17.3 / .57	14.5 / .59	G F C R C L C R R G V C R C I C T R
18	UNMCAP AP00724	-41.9	17.3 / .57	15.1 / .65	R C L C R R G V C R C L C R R G V C
18	UNMCAP AP00725	-42.1	17.2 / .56	13.8 / .53	R C I C T R G F C R C I C T R G F C
20	UNMCAP AP00023	-49.7	17.5 / .54	15.5 / .63	V Y I D K L T P P C G T M Y Y A C E A V
20	UNMCAP	-44.2	20.8 /	14.6 /	I L G P V I S T I G G V

	AP00725		.91	.54	L G G L L K N L
21	KEGG C15876	-46.6	21.8 / .67	14.3 / .36	D M S S D L E R D H R P H V S M P Q N A N
21	UNMCAP AP00028	-49.4	19.6 / .47	17.7 / .65	C L G I G S C N D F A G C G Y A V V C F W

The fact that the peptide drugs cited here all have low entropies indicates that they are relatively stiff and have few conformations that they can have. This is intuitively plausible. However, the physical reason behind the fact that they all have small heat capacities is not immediately apparent. Using the thermodynamic relation and the Onsager fluctuation relation [113], $C_v = T\left(\frac{\partial S}{\partial T}\right) = -\frac{1}{T}\left(\frac{\partial S}{\partial \frac{1}{T}}\right) = \frac{1}{kT}\langle \Delta E \Delta S \rangle$ we see that the heat capacity is related to the cross-correlations of the energy and entropy fluctuations

$$C_v = \beta \langle \Delta E \Delta S \rangle \tag{17}$$

Thus, C_v is determined by the correlation between energy fluctuations of the system, which are large when the energy gaps are large, and entropy fluctuations of the systems between different states, which are small when the energy gaps are large [115]. A small value of C_v means that the peptide takes only a limited number of configurations whose energies are not much separated from each other.

On thermodynamic grounds, the binding constant K_b is obtained from the relative concentrations, $[L]$, $[P]$, and $[LP]$ of the ligand, protein, and their complex, in a bimolecular binding reaction $L + P \Leftrightarrow LP$ as

$$K_b = \frac{[LP]}{[L][P]} \tag{18}$$

The connection of this experimentally measured quantity to the change in free energy in binding is

$$\Delta F \approx \Delta G = -RT \ln K_b \tag{19}$$

where, the change ΔG in the Gibbs free energy is approximated by the change ΔF in the Helmholtz free energy. Here, we discuss the molecular factors that contribute to ΔF and

identify the relative magnitude of conformational free energy changes calculated in this dissertation.

3.2.3. Free energy changes upon binding

As given in Chapter 2, the change of free energy ΔF of the protein, ligand and water system upon binding consists of the following components:

$$\Delta F = \Delta F_{RB}(L) + \Delta F_{conf}(L) + \Delta F_{RB}(P) + \Delta F_{conf}(P) + \Delta F_{binding}(LP) + \Delta F(W) \qquad (20)$$

where, $\Delta F_{RB}(L)$ is the change from rigid body translation and rotation of the ligand, $\Delta F_{conf}(L)$ is the conformational free energy change of the ligand that is the basis of calculations of the present work, $\Delta F_{RB}(P)$ is the change of the rigid body translation and rotation of the protein, $\Delta F_{conf}(P)$ is from the conformational changes of the protein, $\Delta F_{binding}(LP)$ is the change in free energy due to the interaction between the ligand and protein, and $\Delta F(W)$ is the free energy change of water. We assume the terms $\Delta F_{RB}(P)$ and $\Delta F_{conf}(P)$ are small. We discuss the contribution of each remaining term in more detail in order to assess the contribution of $\Delta F_{conf}(L)$ to the total ΔF.

The energies of different conformations of the free peptide will have negative energies when the freely jointed chain is taken as reference. Each energy state will be populated according to the Boltzmann distribution. Upon binding on the surface of the protein, the conformational energy of the peptide will be fixed at one value. This energy may be larger or smaller than the mean energy of the free peptide. The value of this energy depends on the conformation of the bound state, which in turn depends on the shape of the protein surface. A conformational energy in the bound state smaller than the mean energy of the unbound state corresponds to an enthalpically favorable binding. For such a peptide, an estimate of the decrease of conformational energy may be obtained from $-\left\langle (\Delta E)^2 \right\rangle^{1/2}$. Using the definition of the heat capacity, $\frac{C_v}{k} = \beta^2 \left\langle (\Delta E)^2 \right\rangle$, the enthalpic advantage is obtained from $-\beta\left(\frac{Cv}{k}\right)^{1/2}$. From the high tail of the curves in Figure 3.8, where the frequencies of observation fall to less than 1 percent of the peak values, we see (as already

stated in the preceding sections) that $\frac{C_v}{k}$ values are 5, 10, 15, 20 respectively for n = 5, 10, 15, 20. Thus, we have the scaling $\left(\frac{C_v}{k}\right)^{1/2} = n^{1/2}$, and the average conformational energy decreases by $n^{1/2}kT$ upon binding of a good drug. However, the entropy scales with n, and therefore is the dominant factor. The free peptide will have conformational entropy resulting from transitions of bonds from one isomeric minimum to the other. In the bound state, isomeric transitions will vanish. Although the peptide will have some vibrational entropy upon binding [24, 116], the magnitude of this component of the entropy is not yet well characterized. It is mostly regarded within the association entropy of binding, which is $\Delta S_{ass} = S_{res} - S_{trans} - S_{rot}$, where the subscripts are self explanatory [23, 27]. Neglecting this residual entropy, one sees from Figure 3.5 that on the average the entropy decrease is $\frac{\Delta S}{k} = -1.0n$. Thus, in general the conformational free energy of a peptide upon binding, $\beta\Delta F_{conf} = -n^{1/2} + n$. These values are thermodynamic averages however.

In calculating the conformational entropy using the isomeric states model, we did not consider the side chain entropy explicitly. Implicitly, however, the effect is there through the energy maps obtained from PDB. The size of an isomeric state on the Ramachandran map implicitly contains this information. However, explicit consideration of the side chains in addition to the backbone by others [22, 117-118] led to $\frac{\Delta S_{conf}}{k}$ values in the range -4 to -7.5 per residue.

Using the free energy values cited above from different sources that are stated in Chapter 2 [1, 6-7, 14-17, 19-29, 31, 116-119], the various contributions to free energy changes in binding may now be given in Table 3.3. It should be noted however, that these values are only approximate due to the wide range of values reported in the literature. Therefore, the entries of the table should be regarded as to guide the reader. We note here that significant effort is needed to reach a consensus on this important issue.

For n residues, n_{es} electrostatic centers, n_{HB} hydrogen bonds, A^2 of buried hydrophobic area, and n_w water molecules displaced into the bulk water surrounding the protein, the total change of free energy is obtained from Table 3 as:

$$\beta\Delta F = 10 + n - 11n_{es} - 2n_{HB} - 0.05A - n_w \quad (21)$$

The problem of calculating the free energy of binding reduces to the evaluation of n_{es}, n_{HB}, A^2, and n_w , which we do not pursue further. It should be noted that free energy changes due to several effects, notably due protein conformation changes and solvent enthalpy changes, are omitted from Table 1 and Eq. 21, due to uncertainties in their estimation.

Table 3.3. Approximate free energy changes of different components of binding.

$\beta\Delta F_{RB}(L)$	10 [a]
$\beta\Delta F_{conf}(L)$	+1 to +2.5 [b]
$\beta\Delta F_{binding}(LP)$ electrostatic	-11 per pair [c]
$\beta\Delta F_{binding}(LP)$ HB	-2.0 per HB [d]
$\beta\Delta F_{binding}(LP)$ hydrophobic	-.05 Å^{-2} [e]
$\beta\Delta F(W)$	-1 per water molecule [f]

[a] The value of 10 is chosen by taking Amzel's [14] values for translational entropy and multiplying it by 2 to account for rotational entropy.

[b] The lower bound of the range is from the present work, where the contribution from enthalpy is neglected due to $n^{1/2}$ scaling. Entropy loss from side chains is assumed implicit on the RIS formulation. The upper bound is the average of the values given by Karplus [24].

[c,d] From Gohlke and Klebe and the references cited therein [7].

[e] There seems to be a general consensus on this value [29-31, 119].

[f] From Brady and Sharp [22].

3.3. Concluding Remarks

In this chapter, a novel method to determine thermodynamic properties of peptides is presented. Peptides are modeled as sequences of Markov chains where the states defined for each residue are dependenst on the states of the neighbouring residues along the chain. This assumption allows for the application of the Rotational Isomeric States formalism to calculate the thermodynamic averages of conformational properties of the peptides. Here, we used a knowledge based approach to determine the statistical weights of the states of the 20 amino acids and the dependencies of the statistical weights on the neighbouring residues. An alternative approach would be to determine the pairwise statistical weights by

molecular dynamics performed on residue pairs. However, this would require the construction of 400 Ramachandran maps. Choice of the data from Protein Data Bank produces statistical weights that are biased towards the native conformations. We used a coil library data set that is taken from the residues in the coil and loop regions of native proteins. Use of this library is expected to eliminate the helical and beta strand propensities implicit in the full PDB data. Presently, there is no consensus on the choice of the data bank information that would more appropriate to use for the statistical analysis of short peptides. Peptides with low energy, low entropy and low heat capacity are determined to be essential for a peptide to be a good candidate inhibitor. Free energy changes in peptide binding are also discussed.

Chapter 4

DE NOVO PEPTIDE DESIGN WITH GENETIC ALGORITHM

Determination of a specific peptide sequence with affinity to a particular protein surface is a complex problem each residue of the peptide could be chosen from a pool of 20 natural amino acids. Even for a peptide with 3 amino acid residues, there are $8x10^3$ possible peptide sequences. Experimental or computational screening of such a large number of molecules is difficult. A time-efficient rational methodology for specific and selective peptide sequence prediction is necessary. Genetic algorithm (GA) [71] is used for cases that deterministic or analytic methods fail, i.e. problems with underlying mathematical model is not well defined or the search space is too large. Accordingly, employing GA is very proper for our problem, whose search space of peptide design problem is very large. As defined in Chapter 2, GA uses a population of candidate solutions to determine better solutions. Algorithm is inspired by biological evolution: it uses reproduction, mutation, crossover and selection mechanisms. The individuals of the population are represented as chromosomes made up of genes. The algorithm repeats itself for defined number of generations and the given problem evolves toward better solutions in the meantime. For each generation, a new population is formed from the previous population via the selection of parents, mutation and crossover operations on those parents.

In this dissertation, the chromosomes are the peptide sequence and the genes are 20 amino acids. Each chromosome in the problem consists of **7** genes, in other words peptides with 7 amino acids are designed. The genotype is the sequence of a peptide, whereas the phenotype is the three-dimensional structure of that peptide. The input of the algorithm is **n** peptide sequences, which may be selected randomly or selected from literature. The designed algorithm forms new **n** peptides for each generation, and the affinity of the designed **n** peptides for the target protein are determined via AutoDock program. The binding energy of peptides to target protein is regarded as the fitness value. The peptides are aligned according to their binding affinities, from high to low affinity. The choice of parents to form new population is based on tournament selection [120]. Tournament selection randomly chooses pre-defined number individuals from the population and then selects the individuals with best fitness scores to form new individual for next generation.

Also, elitism [121] is applied; the best **4** individuals of each generation will be directly passed to the next generation in order to avoid premature convergence of the population. One- or two-point mutation and one-point / two-point crossover operations are employed. For first generations two-point mutation and two-point crossover are applied to have diversification, while for the final generations the mutation and crossover are applied as one-point. The generation of new population from a previous population is given schematically in Figure 4.1. GA is implemented such that the algorithm works for any given protein target.

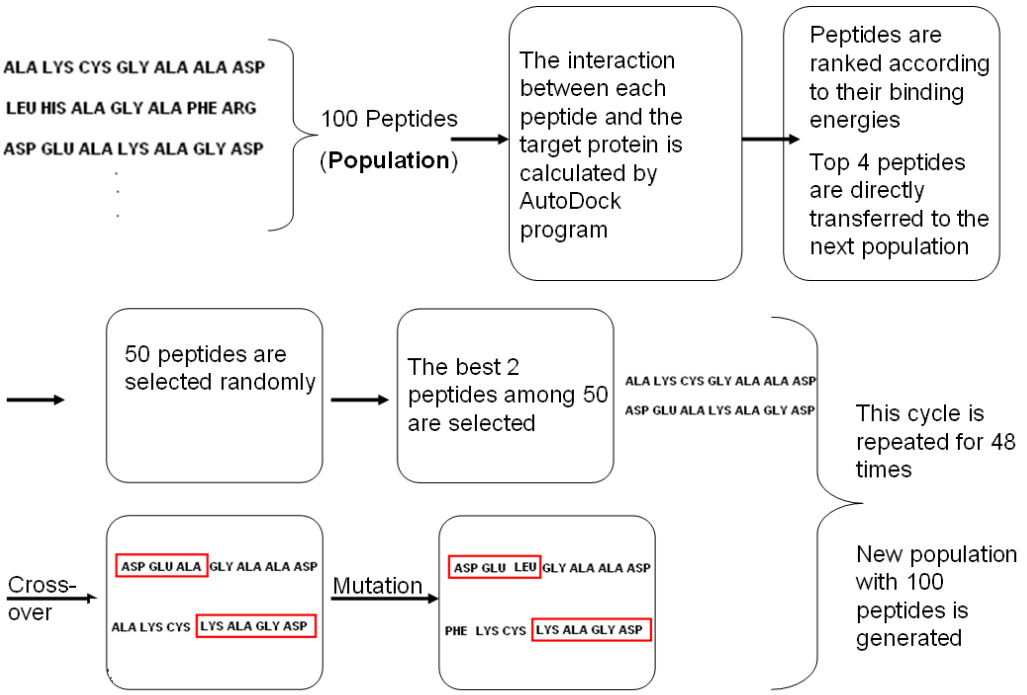


Figure 4.1. Schematic representation of Genetic algorithm.

4.1. The Model and Calculations

Genetic algorithm is implemented to the peptide sequence prediction problem. Firstly the parameters of model are elucidated, and then the model is explained in detail.

4.1.1. Model parameters

The input parameters of our GA are randomly created 100 hepta-peptide sequences, target protein structure file (in *.pdb* or *.mol2* format), a specific binding site on the target protein. The output of GA is 100 peptides with possible affinity to the selected protein target.

The *population* of 100 hepta-peptides is represented by $\mathbf{S_i}$ where *i* indicates the population number. 100 random hepta-peptide sequences form the *initial population*. GA generates a new population using the current population. The formation of one *population* is named as a *generation.* In forming the next population, 2 peptides $\mathbf{P_1}$ and $\mathbf{P_2}$ representing the parent 1 and parent 2, respectively; are selected from the current population. The offspring, i.e. new peptide sequences, formed by $\mathbf{P_1}$ and $\mathbf{P_2}$ are represented by $\mathbf{O_1}$ and $\mathbf{O_2}$.

The cross-over, mutation and elitism operations of GA are denoted by **C**, **M** and **E**, respectively. **C[k]** defines the residues from which the P1 and P2 portions are interchanged; **k** denotes the number of residue (1-6). **M[q, X]** defines which residue to be mutated into which type of amino acid; **q** denotes the residue number (1, 7) and **X** denotes the amino acid type. Elitism operation keeps the best 4 peptides of each population.

4.1.2. GA Implementation

The binding region on the target protein is selected by GNM. 100 hepta-peptides are randomly created by our Python script, without using any priori knowledge about the binding site. Those random 100 peptides form the *initial population*. The 3-dimensional structures, i.e. *.pdb* files, of the peptides are created by *mutate* command of the VMD software [122] using an initial *.pdb* file of hepta-peptide made up of all Alanine residues as input. The peptide torsion angles are set to be flexible by AUTOTORS utility of AutoDock. Polar hydrogen is added to protein and peptide; Gasteiger charges are added by AutoDock Tools (ADT). The grid map is determined by ADT; the binding-site is taken as grid-center. Genetic Algorithm option of AutoDock for docking is selected. Lamarckian Genetic Algorithm is chosen as the docking search parameter. The population size is set to 150; run = 50; maximum number of energy evaluations = 250 000; number of generations = 27 000. The docking parameters are kept at low values to have a fast computation time since we want to rank peptides rapidly. AutoDock program outputs are the docked conformations of the peptide on the protein; the free energy of binding that is summation of the intermolecular energy (van der Waals energy, Hydrogen bonds, desolvation energy, electrostatic energy), final total internal energy, torsional free energy of peptide and unbound system energy; and estimated inhibition constant. The free energy of binding is used as the fitness value of our GA. After each *population* of peptides is docked to the binding site, they are ranked according to their fitness values. The ranking of the *current*

population is the first step of our GA; the *new population* is formed by the ***tournament selection*** [120] procedure with ***elitism*** [121]. The algorithm is defined in detail below. Python is used as the scripting language.

Algorithm:

Input: Random initial population $\mathbf{S_1}$, protein structure file, binding site specified by GNM.

Output: 100 peptides with possible affinity to specified protein binding site.

1 for i = 10 times (10 generations)4

2 Create *.pdb* files for 100 peptides in current population $\mathbf{S_i}$;

3 Prepare AutoDock input files and dock each 100 peptides to target protein binding site.

Rank the peptides according to free energy of binding.

Formation of next population $\mathbf{S_{i+1}}$:

4 E: Elitism, pass the top 4 peptides directly to next population

5 for 48 times

6 Tournament selection part: randomly select 20 peptides; set the top 2 peptides as $\mathbf{P_1}$ & $\mathbf{P_2}$

7 k = random number between 1-6

8 C[k]: single-point cross-over between $\mathbf{P_1}$ and $\mathbf{P_2}$: offspring $\mathbf{O_1}$ and $\mathbf{O_2}$ formed

9 r = random number between 0-1

10 if r > 0.2 **then**

11 q1, q2 = random number between 1-7

12 X1, X2 = any of 20 amino acids selected randomly

13 M[q1, X1]: single-point mutation on $\mathbf{O_1}$ and **M[q2, X2]:** single-point mutation on $\mathbf{O_2}$

4.2. Case Studies

Our GA program runs for 3 days on 30 processors. The processors are 2.4 GHz Intel Xeon with 1 GB RAM memory. 10 generations are completed in 3 days; consequently each generation is evaluated in ~7.2 hours. The algorithm is tested on hGH, CRY2, and NF-κB proteins.

4.2.1. Nuclear factor kappa-light-chain-enhancer of activated B cells

Nuclear factor kappa-light-chain-enhancer of activated B cells (NF-κB) is selected as target in view of the fact that it plays vital role in immune, apoptotic and oncogenic processes [123]. This prominent target is formed by hetero- or homo-dimer of proteins and localized both in cytoplasm and nucleus. In cytoplasm, it is kept inactive by binding of I-κBs (inhibitors of κB) onto the nuclear localization signal (NLS) site. NF-κB becomes active and translocated to nucleus, when I-κB is phosphorylated and degraded by ubiquitin pathway. Degradation of I-κB releases the NLS site of NF-κB; the protein is shuttled to the nucleus. In the nucleus, NF-κB binds to DNA and acts as transcription factor. As a transcription factor, NF-κB mediates production of approximately 180 proteins. The overproduction or constituent production of some proteins by means of NF-κB activity leads to carcinogenesis and metastasis; especially inflammation-associated cancer. Therefore, inhibition of NF-κB is essential for curing inflammatory diseases and NF-κB related cancer types. The system that we aim to design a peptide against is a heterodimer of p65 – p50 subunits. The crystal structure of p50, p65 and I-κB complex could be downloaded from the Protein Data Bank with the accession code: **1IKN** [124]. This complex resides in the cytoplasm as inactive, upon degradation of I-κB the p50-p65 the complex is transferred to the nucleus via the NLS on the p65 subunit. According to mutation experiments, the residues Aspargine-202 and Serine-203 are known to play major role in I-κB binding. The C-terminal of the p65 subunit, which is responsible of p50 and I-κB binding, is chosen as the target region [124]. The site is made up of 2 beta-strands and a cavity forms in between the strands. Also the important residues Asn202 and Ser 203 reside on the chosen region. Figure 4.2 demonstrates the selected cavity: the boundary of the cavity is formed by the amino acids: Glu193, Leu194, Lys195, Ile196, Cys197, Arg198, Val199, Asn200, Arg201, Ser281, Glu282, Pro283, Met284, Glu285, and Phe286. Also the important residues Asn202 and Ser 203 residues [123] on the chosen region.

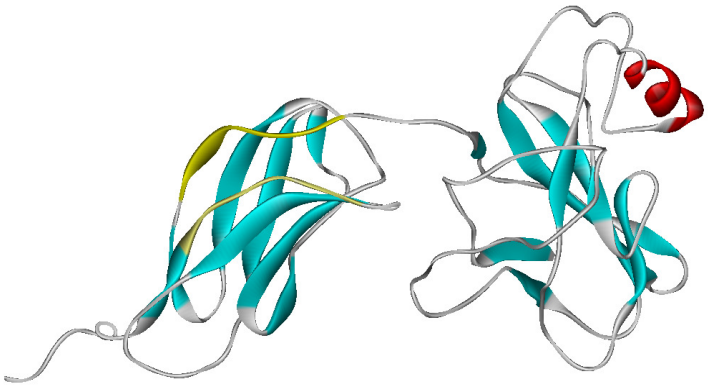


Figure 4.2. The structure of p65 subunit, specified with ribbons diagram representation. The residues that forms the binding site is shown with yellow color.

Best binding peptide sequence of each generation and the corresponding free energy of binding in kcal/mol, which is the fitness value of GA, are given in Table 4.1. The fitness value tends to decrease for each generation. The decline in fitness values indicates that peptides with more affinity to the selected protein surface are determined at each generation. The worse fitness is observed for initial sequence, which is expected since the initial population is created randomly. The program is able to determine best peptide sequence for NF-κB at the **7th** generation. Due to high mutation rate of offspring, the program is able to diversify between good peptide candidates and worse candidates, as could be seen in Table 4. The program moved to *Phe Ala Asn Ala Asn Asn Asp* sequence at the 8th generation after determining best sequence *Cys Ala Asn Ala Asn Asn Asp*; then moved back to the best sequence at the 9th generation. In other words mountain climbing is achieved.

Table 4.1. The first column indicates the GA generation number. The second column indicates the peptide sequence determined to have most affinity to the protein. The last column indicates the best binding energy that is used as the fitness value.

Generation Number	Peptide Sequence	Fitness Value (kcal/mol)
1	Q I N A H H T	-2.63
2	G L N A C N D	-5.56
3	M A N A A N D	-7
4	M A N A H N D	-8.16
5	I A N A N N D	-8.41
6	M A N A N N D	-8.96
7	C A N A N N D	-10.21
8	F A N A N N D	-9.17
9	C A N A N N D	-9.5
10	C A N A N N D	-9.37

The convergence for some amino acid types at specific positions is observed, those repeated and conserved residues may play role in specific binding to the selected protein surface. The *Asn* at the 3rd position and *Ala* at the 4th position are conserved for all generations, those 2 residues are most possibly responsible of initial random peptides' affinity for target protein. At the second generation *Asn* at the 6th position and *Asp* at the 7th position are observed to increase binding affinity and conserved for all next generations. At the 5th generation, the importance of *Asn* at the 5th position is realized and the amino acid is kept for further generations. At the 7th generation the best sequence of 10 generations is determined.

The physical and chemical properties of amino acids are also considered by our GA, without any priori information. For the 1st position, mostly *Cys* and *Met* both having sulfur atom are conserved but the hydrophilic *Cys* is selected. *Cys* also has a more active sulfur atom. The binding surface is known to have Cys197 residue that the *Cys* of peptide may make sulfur-bridge with. For the 2nd position, GA tried only hydrophobic residues: *Leu* and *Ile* with large hydrophobic side-chains seem not to fit in the binding region, consequently hydrophobic residue with smallest side-chain *Ala* conserved. For the 5th and 6th positions some hydrophilic residues are tried and *Asn* seem to be the most appropriate one.

10^3 peptide sequences are analyzed for the selected protein surface. 10^{17} possible hepta-peptide sequences exist for a hepta-peptide. Almost 60% of the tried sequences have free energy of binding that is lower than -7 kcal/mol. This outcome indicates that the searched 10^3 sequences are among the best sequences of the search space.

4.2.2. Fibroblast growth factor receptor 2

Fibroblast growth factor receptor (FGFR2) is a cell-membrane protein made up of an extracellular region; a transmembrane helice region; and a cytoplasmic tyrosine kinase (TK) domain. FGFR has three extracellular domains: D1, D2 and D3. The extracellular portion of the protein interacts with fibroblast growth factors (FGF). There is 4 types of FGF and more than 19 types of FGFR in humans. There exists 1:1 interaction between each FGF:FGFR and two FGF:FGFR complexes bind to each other to form a dimeric complex for activation [125]**.** Two FGFR:FGF complexes form a tetramer that triggers downstream signals into the cell initiating mitogenesis or differentiation. The mutations of FGFR2 cause the protein to bind FGF's strongly and lead to permanent differentiation and growth signals. As a result breast, lung, ovary, stomach cancers are initiated. The FGF proteins bind to the D2 and D3 domains of FGFR proteins, the interactions of D2 domain are almost the same for all FGF:FGFR complexes, while the D3 domain interactions differ for different FGF:FGFR complexes. Our aim is to design a selective peptide for FGFR2, consequently the selective interactions of FGF2:FGFR2 were taken into consideration. In the literature there exists various attempts to design drug against cytoplasmic TK region of the protein - but a variety of proteins also contain TK domain, which became a target to those drugs. Designing a drug / peptide against the extracellular domain of the protein may lead to specificity and will hopefully decrease the side-effects of cancer therapies. The PDB files, with best X-Ray diffraction values, from protein data bank (www.pdb.org) was selected, the PDB ID is 1EV2 [114]. The extracellular part of the FGFR2 is selected as target region (Figure 4.3). The binding site contains the hot-spot amino acids: Arg251, Asp283, Gln285, Tyr281, Arg251, Glu250, and Ala171.

The initial population of GA is randomly generated peptides, which binds to FGFR2 with low affinity or no affinity at all. The GA is set for 13 generations. The results are given in Table 4.2. The fitness value tends to decrease for each generation. The program is able to determine best peptide sequence at the last generation. Due to high mutation rate of offspring, the program is able to diversify between good peptide candidates and worse

candidates, as could be seen in Table 5. The program moved to *Asn Ala Ala Ala Asn Asn Ala* sequence at the 11th generation after determining a better sequence *Asn Ala Ser Asn Asn Asn Ala*; then moved back to the best sequence *Asn Val His Asn Ala Asn Ala* at the 13th generation. In other words mountain climbing is achieved, as in the case of NF-κB.

Table 4.2. Best peptides for FGFR2 generated by each cycle of GA. The first column indicates the generation number. The second column indicates the peptide sequence determined to have most affinity to the protein. The last column indicates the best binding energy that is used as the fitness value.

Generation Number	Peptide Sequence	Fitness Value (kcal/mol)
1	M A H N L H A	-4.97
2	M A H N I N A	-5.28
3	M A H N I N D	-6.5
4	M A H F I N A	-6.59
5	V G H N N N A	-7.18
6	M A V N N N A	-7.28
7	M A A N N A	-7.94
8	N A H N N N A	-7.71
9	F A A N N N A	-7.61
10	N A S N N N A	-8.83
11	N A A A N N A	-7.88
12	N A S N N N A	-8.57
13	N V H N A N A	-8.85

The convergence for some amino acid types at specific positions is observed. *His* at the 3rd position, *Asn* at the 6th and *Ala* at the 7th position are conserved for almost all generations, those 3 residues are most possibly responsible of initial random peptides' affinity for target protein. *Val* is observed at the 5th, 6th and the last generations.

The detailed examination of the interaction between this peptide and the protein reveals that, *Asn Val His Asn Ala Asn Ala* peptide makes 7 hydrogen bonds with Tyr281, Arg251, Glu250, Ala171 residues of FGFR2. The Arg251 residue of FGFR2 is known to interact

with Asn102 – Asn104 of FGF [125]. If Asn104 of FGF2 is mutated to *Ala*, the FGFR2-FGF2 interaction reduces 400-fold [125]. Arg251 is very prominent for FGF binding and since the peptide has 2 *Asn* residues it mimics FGF. There is a hydrophobic interaction between peptide and the hydrophobic pocket formed by Val249 and Ala171 residues of FGFR2 [125]. Furthermore, electrostatic complementarity exists between protein and peptide. The GA seems to work well and it is able to determine peptide sequences that would make important interactions with the target protein. Figure 4.3 represents the interaction between FGFR2 and *Asn Val His Asn Ala Asn Ala* peptide.

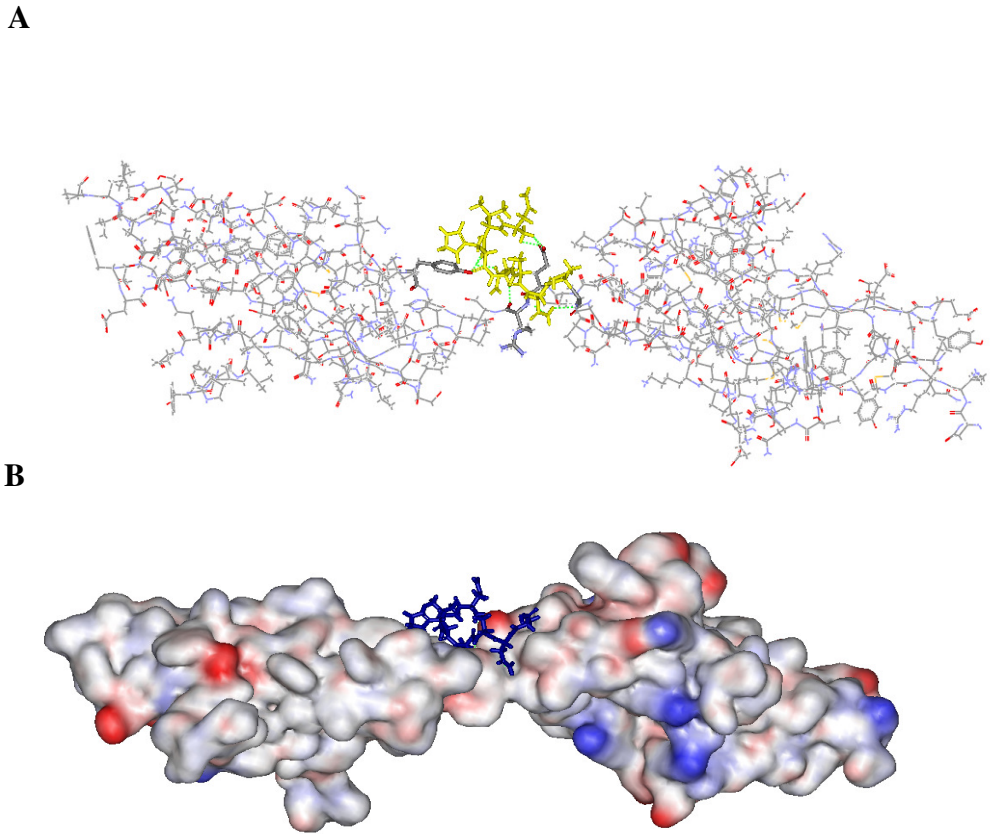


Figure 4.3. The FGFR2 and the best peptide complex. **A.** Both the protein and the peptide are shown in lines representation, the peptide is shown with yellow color and the H-bonds are indicated with green dashed-lines. **B.** The protein is shown in surface representation and the peptide is shown in lines respresentation with blue color.

4.2.3. Cryptochrome 2

Cryptochromes (CRY) are blue-light receptor proteins, which display a high homology to the DNA photolyases. DNA photolyases are DNA repair enzymes, which are activated by blue-light. Apart from sequence similarity, CRY was discovered to have the chromophores

that DNA photolyase has; flavin adenine dinucleotide (FAD) and 5, 10-methenyl tetrahydrofolate (MTHF). The chromophores of CRY have the ability to play role in the electromagnetic transmission of signal from eyes to SCN. Although both CRY s and photolyases are similar in structural level, CRY s do not posses DNA repair activity [126]. The first cryptochrome in human was found when *E. coli* DNA photolyase was aligned against human EST databank [127]. The second cryptochrome of human was found in Prof. Aziz Sancar's lab. [126] was found to have FAD and MTHF as chromophores; but they lack photolyase activity. The hCRY were named as; Cry1 and Cry2 [128]. It is not possible to purify this protein with its chromophores. We aimed to design a peptide that selectively binds to protein without disturbing its contacts with FAD and MTHF.

The binding site of CRY2 contains the residues between Arg341-Ala413. The initial population of GA is randomly generated peptides and the GA is set for 10 generations. The results are given in Table 4.3. The fitness value tends to decrease for each generation. The program is able to determine best peptide sequence at the last generation. The results could be seen in Table 6. The program determined the best sequence *Asp Trp Trp Trp His Ala Cys* at the 10^{th} generation.

Table 4.3. Best peptides for CRY2 generated by each cycle of GA. The first column indicates the generation number. The second column indicates the peptide sequence determined to have most affinity to the protein. The last column indicates the best binding energy that is used as the fitness value.

Generation Number	Peptide Sequence	Fitness Value (kcal/mol)
1	D R W W H A C	-2.15
2	D A W W H A C	-4.05
3	D E W W H A C	-6.12
4	D V V F H A C	-6.51
5	D D V F H A C	-7.16
6	D L W W H A C	-7.35
7	D I W W D A C	-7.54
8	D M V F H A W	-7.57
9	D V V F H A C	-8.84
10	D W W W H A C	-10.34

The convergence for some amino acid types at specific positions is observed. *Asp* at the 1st position, *His* at the 5th position, and *Cys* at the 7th position are conserved for all generations, those 3 residues are most possibly responsible of initial random peptides' affinity for target protein.

The detailed examination of the interaction between this peptide and the protein reveals that, *Asp Trp Trp Trp His Ala Cys* peptide makes 6 hydrogen bonds with Glu344, Pro343, Glu351, Ser395, and Arg398. *Trp* at the 4th position and CRY2 Met363 form S ··· aromatic interaction. The S ··· aromatic interaction provides extra overall stability to the native fold of the peptide and also sets the orientation of the conserved aromatic. Furthermore, electrostatic complementarity exists between protein and peptide. Figure 4.4 represents the interaction between CRY2 and *Asp Trp Trp Trp His Ala Cys* peptide.

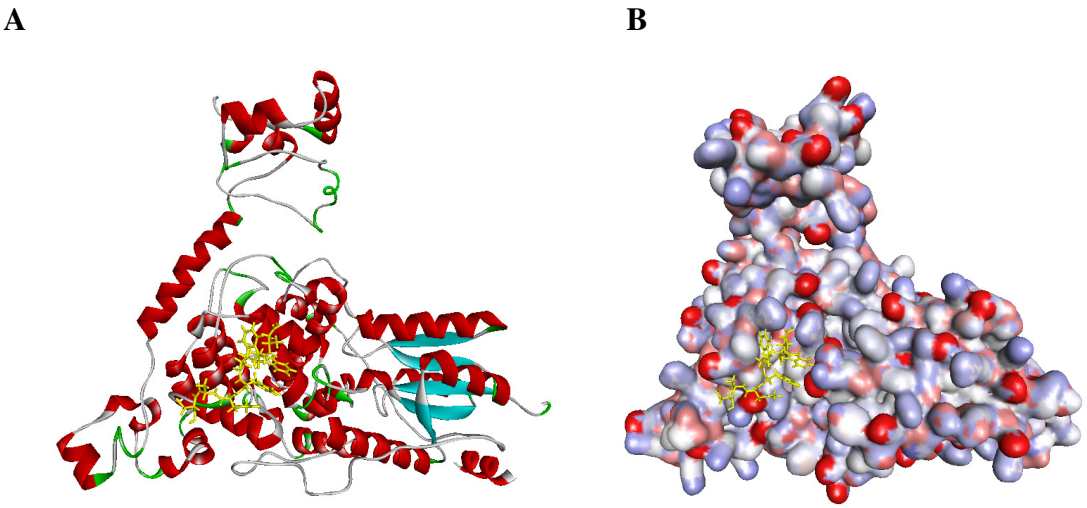


Figure 4.4. The CRY2 and the best peptide complex. **A.** The protein is shown with ribbon diagram and the peptide is shown with lines representation, the peptide is shown with yellow color. **B.** The protein is shown in surface representation and the peptide is shown in lines respresentation with yellow color.

4.3. Concluding Remarks

A metaheuristic optimization method Genetic Algorithm is employed for the peptide design problem. Tournament selection for GA is implemented. The method is global and works well with any random initial population. The performance of predictions is encouraging, and can be improved further via addition of some filters according to the

target protein information or formation of non-random initial population. The filters may also lessen the time of prediction. The conservation of amino acids and amino acid properties at specific positions for each generation is a promising outcome.

The designed peptides have good theoretical binding energies for target proteins. The experimental affinity values of designed peptides are necessary for comparison with the defined theoretical values. The GA offers potential inhibitor peptide sequences for any target protein within three days.

The importance of the program is that no manipulation during the runs or no priori information are needed to determine a good candidate peptide, also the conservation of amino acids at specific positions is outstanding.

Chapter 5

DE NOVO PEPTIDE DESIGN WITH MARKOV MODEL

The Markov model offered in this dissertation to determine peptide sequences with optimum affinity to the selected binding site of a protein. The method offers a novel procedure for *de novo* peptide design that sequentially generates the peptide by docking its residues pair by pair along a chosen path on a protein. The states are amino acid types, in form of dipeptides, forming the peptide of given length. Important points that was taken into consideration in the sequential generation of ligand molecules concerning the placement of fragments on the surface are discussed by Pegg et al. [129]. Two novel approaches are adopted: (i) the first one is the determination of the binding site on a given protein which we determine using the coarse grained Gaussian Network Model (GNM). Recently, it was shown that the GNM identifies the surface residues that are suitable binding sites [130-132]. Once the binding region is determined, the binding path is obtained on it by docking an arbitrarily chosen peptide using AutoDock [44]. (ii) The second novel aspect of the model is the choice of the best fitting peptide to this path. All possible amino acid pairs at each point along the path are generated; the binding energies between these pairs and the specific location on the protein are calculated via AutoDock. The binding energies form the statistical weight of each pair of amino acids. Once all possible pairs are determined for the full path, the Markov based partition function for all possible choices of the peptides are formed using the Ising model matrix multiplication scheme [6]. The transition probabilities are evaluated based on this partition function, and the best peptide for the given surface is selected employing a Markov model. The types of amino acids are the states of the algorithm to be obtained as the solution of the problem.

5.1. The Model and Calculations

Markov model is implemented to peptide sequence prediction problem. Firstly the parameters of model are elucidated, and then the model is explained in detail.

5.1.1. Model Parameters

The input parameters are target protein structure file (in *.pdb* or *.mol2* format), binding site of target protein and a path on this binding site, 20 amino acid *.pdb* files, 400

dipeptide *.pdb* files, and 6 **T** transition probability matrices. The output of is 100 peptides with possible affinity to the selected protein target. The binding path is divided into 7 consecutive grids denoted by $\mathbf{G_m}$ representing the m^{th} grid. $T_{m;\zeta,\eta}$ is the element of the m^{th} transition matrix: the probability of the dipeptide formed by ζ,η types of amino acids to bind to $\mathbf{G_m}$.

5.1.2. Markov Model Implementation

The binding region on the target protein is selected by GNM. To select a path on this binding site, a peptide with 7 residues made up of all Glycine (7-Gly) is used. The affinity of this initial peptide for the selected binding site; and the binding orientation of peptide are determined via the docking tool GOLD [70]. The GoldScore scoring function of GOLD is used with detailed analysis options turned on. The most possible bound conformation of 7-Gly peptide is used for further analysis.

The orientation of the 7-Gly peptide bound to the target protein is divided into 7 grids in space: $\mathbf{G_1}$ contains the conformation of the 1st amino acids, $\mathbf{G_2}$ contains the conformation of the 1st - 2nd amino acids, $\mathbf{G_3}$ contains the conformation of the 2nd - 3rd amino acids, $\mathbf{G_4}$ contains the conformation of the 3rd - 4th amino acids, $\mathbf{G_5}$ contains the conformation of the 4th - 5th amino acids, $\mathbf{G_6}$ contains the conformation of the 5th - 6th amino acids; $\mathbf{G_7}$ contains the conformation of the 6th - 7th amino acids. To put it systematically; the $\mathbf{G_1}$ contains only the conformation of the 1st amino acids, and any of the remaining $\mathbf{G_m}$ piece contains the conformation of the $(m-1)^{th}$ and m^{th} amino acids. The chiral carbon coordinate of the 1st amino acid is used as center for grid $\mathbf{G_1}$; and the arithmetic mean of the chiral carbon coordinates of the $(m-1)^{th}$ and m^{th} amino acids' is used as the center for the grid $\mathbf{G_m}$. The schematic representation can be seen in Figure 5.1.

20 amino acids are docked to $\mathbf{G_1}$. All 400 dipeptides are docked to all m^{th} grid, $\mathbf{G_m}$, $7 \geq m > 1$. The amino acids and dipeptides *.pdb* files are prepared by the HyperChem software [133]. The docking is achieved by AutoDock. The amino acid and dipeptide torsion angles are set to be flexible. Polar hydrogen and Gasteiger charges are added by ADT. The grid map is determined by ADT; the pre-determined spatial coordinates of chiral carbon atoms on the path are used as the AutoDock grid-center. The grid size -with a spacing of 0.375 Angstrom between grid points- is set such that grid-box containing the amino acid / dipeptide could rotate and translate freely. The maximum length of each amino acid and

dipeptide is measured with VMD program; the measured values are used for grid-box size determination. Genetic Algorithm option of AutoDock is selected. Lamarckian Genetic Algorithm is chosen as docking search parameter. The population size is set to 150; runs = 10; maximum number of energy evaluations = 250 000; number of generations = 27 000. The remaining parameters are set as default values. The docking parameters are kept at low values to have fast computation time; since we want to rank amino acids / dipeptides rapidly.

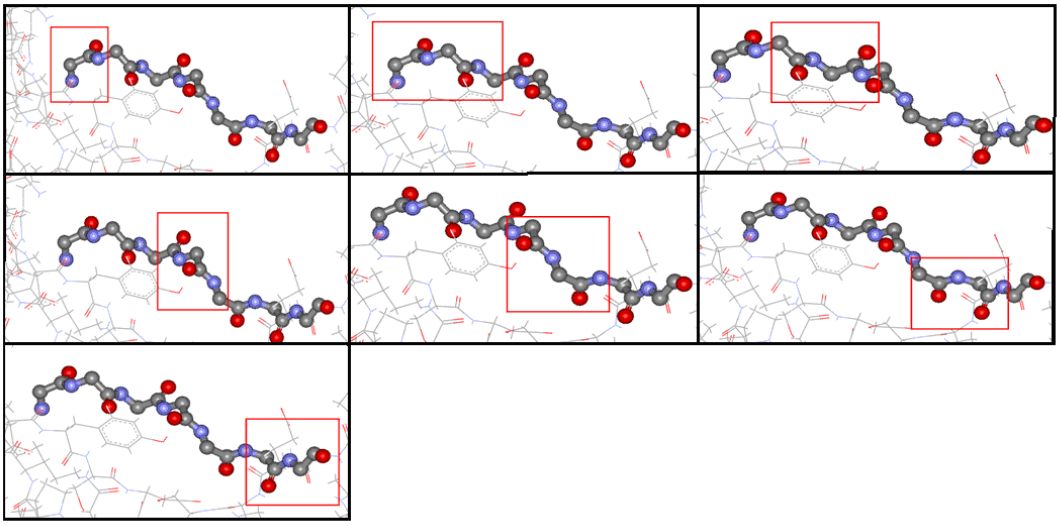


Figure 5.1. Partitioning a path formed by 7-Gly peptide. The upper left illustration indicates the $\mathbf{G_1}$; the lower left illustration shows the $\mathbf{G_7}$ grid box.

The docking results are analyzed and for each grid box $\mathbf{G_m}$ the amino acid / dipeptide affinity in terms of free energy of binding is determined. The determined energy values are used to calculate the probability of binding of each amino acid / dipeptide to the protein binding. The probabilities form the transition matrices.

For $\mathbf{G_1}$, 20 binding energies are available, while for the remaining grids 400 binding energies are available; the corresponding statistical weight matrix formation is given in Eqs. 1.1 and 1.2. The RIS multiplication scheme [6] is used to determine transition probabilities.

The statistical weight matrix U_1 for the amino acid bound to the 1st grid is

$$U_1 = \exp\left(-\beta E_{1;\zeta}\right) \tag{1.1}$$

where ζ stands for the any of 20 amino acids and E indicates the free energy of binding of amino acids to $\mathbf{G_1}$.

The statistical weight matrix U_m for the dipeptide formed residues is

$$U_m = \exp\left(-\beta E_{m,m+1;\zeta,\eta}\right) \tag{1.2}$$

where m represents amino acid position of peptide, $7 \geq m > 1$. The ζ, η values represent any of the 20 amino acids; Ala, Cys, Asp, Glu, Phe, Gly, His, Ile, Lys, Leu, Met, Asn, Pro, Gln, Arg, Ser, Thr, Val, Trp, Tyr. The order of amino acids is given as they are used in the formation of the U matrices. The partition function, Z, of the peptide is obtained according to

$$Z = J^* \left(\prod_{k=2}^{n} U_m \right) J \tag{2}$$

where $J^* = U_1$; J = column [1 1 … 1]. (It is to be noted that in the Flory notation, the J* matrix is given as J^* = [1 0 … 0] that assigns the Alanine to the first residue of peptide. In the present formulation, the choice of the J* matrix allows for the acknowledgement of all of the 20 amino acids to be the first residue with equal probability).

The probability of having residue ζ at the m^{th} position and residue η at the $(m+1)^{th}$ residue is determined by:

$$p_{m,m+1;\zeta,\eta} = \frac{J^* U_2 \ldots U_m U'_{m+1} \ldots U'_7 J}{Z} \tag{3}$$

Equation 3 is used to form probabilities of the transition matrix $\mathbf{T_{m+1}}$ for $\mathbf{G_{m+1}}$; the elements of the matrix are probability values $p_{m,m+1;\boldsymbol{\zeta},\boldsymbol{\eta}} = T_{m+1;\boldsymbol{\zeta},\boldsymbol{\eta}}$.

The calculated probabilities score for the binding affinity of the 20 amino acids to $\mathbf{G_1}$; and for the 400 dipeptides' binding affinity to each of $\mathbf{G_m}$. In order to calculate 100 top binding peptide sequences for the target protein, the calculated transition probabilities are used. The probability of a specific peptide sequence to have affinity for target protein is calculated by the formula given in Eq. 4. The formula basically multiplies the probability of X-type amino acid for $\mathbf{G_1}$; XY-type dipeptide for $\mathbf{G_2}$; YZ-type dipeptide for $\mathbf{G_3}$, etc.

$$\begin{aligned} & prob(Sequence = A_1 A_2 A_3 A_4 A_5 A_6 A_7) \\ & = p_{1;A_1} p_{1,2;A_1,A_2} p_{2,3;A_2,A_3} p_{3,4;A_3,A_4} p_{4,5;A_4,A_5} p_{5,6;A_5,A_6} p_{6,7;A_6,A_7} \\ & = T_{1;A_1} T_{2;A_1,A_2} T_{3;A_2,A_3} T_{4;A_3,A_4} T_{5;A_4,A_5} T_{6;A_5,A_6} T_{7;A_6,A_7} \end{aligned} \tag{4}$$

where A_m represents any of 20 amino acids.

Python used as scripting language. Our script determines the probability for all possible peptides, but keeps track of the top 100 peptides only.

5.2. Case Study: NF-κB

The Markov program runs for 1 day on 10 processors. The processors are 2.4 GHz Intel Xeon with 1 GB RAM memory. 10 processors are used in docking procedure; only 1 processor is used for probability calculations.

Peptide design for NF-κB is studied. To determine the path 7-Gly is used, the bound conformation of 7-Gly peptide to NF-κB is given in Figure 5.2. The peptide makes 6 Hydrogen bonds with the Glu-285, Glu-282, Leu-194, Leu-280, Pro-283, and Ile-196 residues of the protein. The path is used for amino acid / dipeptide docking and determination of most possible 100 peptide sequences as explained in Section 5.1.

37 of 100 peptides bind to NF-κB in nanomolar quantities according to AutoDock results. The docking results imply that the **6** of top **10** peptides, which are predicted to be best binding peptides by our probability calculations, actually show nanomolar binding to NF-κB. From that point of view, the docking results support model and calculations.

Table 5.1 summarizes the properties of the top 5 peptides and NF-κB interaction. All 100 peptide properties could be found in Appendix 1. The 1st column indicates the peptide sequences in one-letter representation, the 2nd column gives the minimum binding energy of 100 runs for each peptide in kcal / mol calculated by AutoDock, the 3rd column demonstrates inhibition constant in molar quantities calculated by AutoDock, the 4th column gives the average binding energy of 100 runs for each peptide

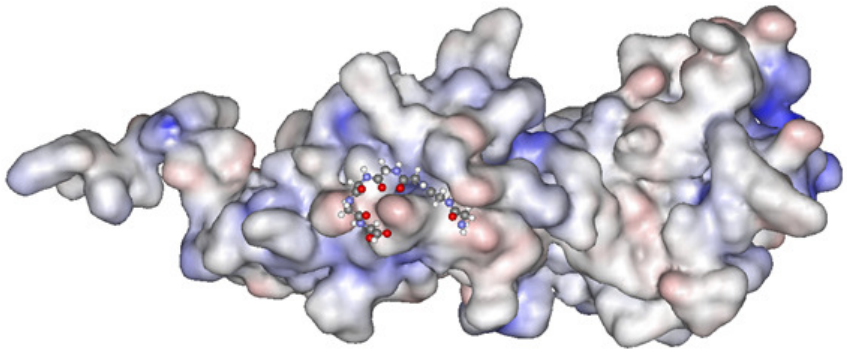


Figure 5.2. 7-GLY peptide bound to Nf-kB. Nf-kB shown with surface representation and peptide with ball-and-stick representation.

Table 5.1. Interaction details of Markov Model peptides.

Peptide Sequence	Binding Energy (kcal/mol)	K_i	Average Binding Energy (kcal/mol)	Probability Order	Number of H-bonds	Number of Short Contacts	Electrostatic Energy (kcal/mol)
FSWFWY	-11.25	5.71	-6.58	**4**	6	15	-0.73
KFWFFY	-11.11	7.16	-6.95	95	5	14	-2.43
FKFEFWY	-10.37	25.14	-4.42	44	3	17	-0.83
FKHKFY	-10.22	32.17	-4.81	**2**	7	7	-3.39
RHTWFY	-10.13	37.78	-5.6	28	10	20	-1.9

in kcal / mol, the 5th column represents the peptides' order in the 100 probability order; the 6th column shows the number of Hydrogen-bonds formed with NF-κB, the 7th column shows the number of short contact formed with NF-κB, the last column indicates the electrostatic affinity of peptide for NF-κB in kcal / mol.

The peptides generally make various short contacts and minimum 3 Hydrogen bonds with the defined binding path. Electrostatic complementarities between protein and peptides exist according to AutoDock results.

There are also peptides that do not show good binding to NF-κB according to docking results, but predicted to have affinity for target protein by our Markov model. For instance, the peptide *Phe Lys His Trp Phe Trp Tyr*, which is predicted as the best binding peptide by our probability calculations, only show -4.37 kcal / mol binding affinity. The main reason

for those overestimations may be explained by the fact that two dipeptides may show high affinities for two consecutive specific sites on the protein; but when those two dipeptides are overlapped to form a tripeptide, any / all of them may not be able to take the conformations in their dipeptide forms. The overall conformation of a protein may be very different from the conformations of dipeptides that form this peptide, the predicted binding affinity and the calculated binding affinity may show variance. In order to overcome this problem, the conformation of dipeptides should also be regarded in the Markov model. We have implemented conformational information to the problem, which is described in Chapter 6.

The detailed analysis of top 5 peptides from AutoDock results with inhibition constant lower than 50 nM and minimum binding energy lower than -10 kcal / mol are given in Figure 4.6. *Phe Ser Trp Phe Trp Trp Tyr* makes hydrogen bond with Ile-196, Arg-198, Val-199, Asn-202, and Glu-285 residues. *Lys Phe Trp Phe Phe Trp Tyr* makes hydrogen bonds with Asn-200, Asn-202, and Glu-285. *Phe Lys Phe Glu Phe Trp Tyr* makes hydrogen bond with Leu-194, Cys-196, and Glu-285. *Phe Lys His Lys Phe Trp Tyr* makes hydrogen bonds with Cys-197, Val-199, Glu-282, Pro-283, and Glu-285. *Arg His Thr Trp Phe Trp Tyr* makes hydrogen bonds with Arg-198, Val-199, Asn-200, Ser-203, Glu-282, Pro-283, and Glu-285. The designed peptides bind to the selected protein binding site, as expected. The Markov model is able to design specific peptides for the selected protein region.

The residues Asn-202 and Ser-203, which are known to play critical role in I-κB binding, are observed to be in close vicinity of top 5 peptides, making hydrogen bonds or short contacts with peptide. Glu-285 residue of target protein also appears to play major role in peptide binding. The possible interactions between protein and peptide may prevent I-κB binding, if the formed protein-peptide complex is stable the NF-κB protein will not be shuttled to nucleus and start gene transcription. The designed peptides may be used as inhibitory drugs for NF-κB related diseases. Experiments to determine peptide affinities for the target protein are necessary.

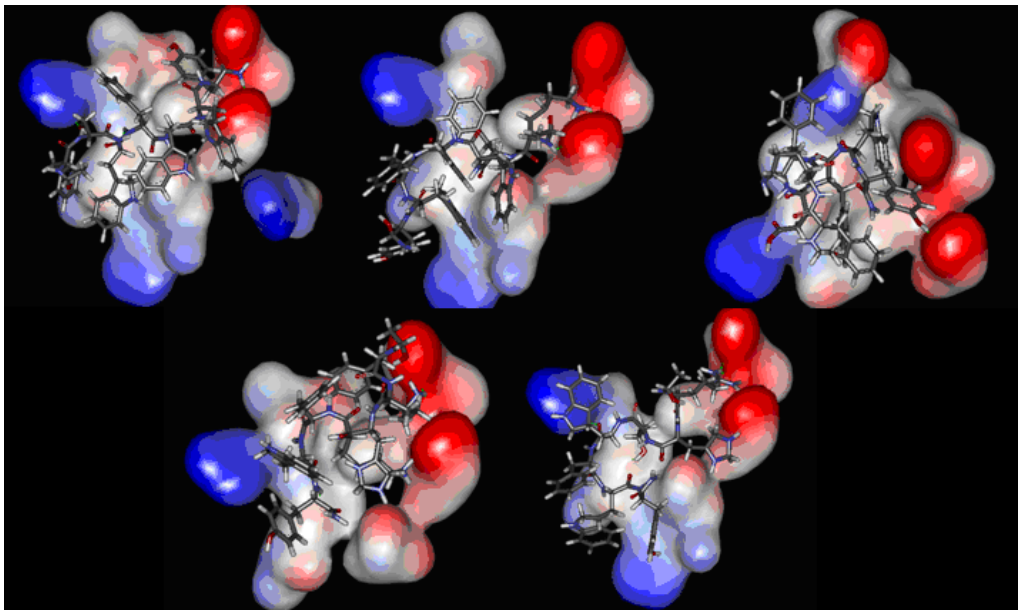


Figure 5.3. The top 5 peptides. The representations are ordered according to the binding affinity; 1st scheme illustrates the **FSWFWWY** bound to NF-κB, the 2nd shows **KFWFFWY**, the 3rd indicates **FKFEFWY**, the 4th indicates **FKHKFWY**, and the 5th shows **RHTWFWY**.

As can be seen from Figure 5.3, all 5 peptides bind to the same region on protein surface, their aromatic side chains interact with the binding cavities of NF-κB. The bulky aromatic side chains of the peptides bind to 5 different grooves of the selected cavity. 4 peptides have *Phe Trp Tyr* at their C-terminal, which bind to the 3 lower grooves. The peptide, with *Trp Trp Tyr* sequence at its C-terminal sequence, binds to only 2 grooves. The *Phe Lys His Lys Phe Trp Tyr* and *Arg His Thr Trp Phe Trp Tyr* peptides show high binding affinity since they have +3 and +2 charges, respectively. Those positively charged N-terminal residues enable the peptides to make electrostatic interactions with the negatively charged part of the selected cavity.

5.3. Concluding Remarks

In this dissertation a model that utilizes Markov property is designed for peptide design problem. This model seems to predict better binding peptide candidates when compared to GA. The time required for prediction is 1/7 of the GA prediction time. The prediction performance could be improved by additional informative attributes: conformation of amino acids on protein surface, chemical and physical features of amino acids, protein binding surface information. Experimental design is necessary to prove interaction between the offered peptide sequences and NF-κB.

Chapter 6

DE NOVO PEPTIDE DESIGN WITH HIDDEN MARKOV MODEL

Two Hidden Markov models are offered in this dissertation as solutions to the *de novo* peptide design problem. The basics and theory of the two methods are the same, while the observable states are different. The first model, namely **VitAL**, utilizes the intrinsic Ramachandran potentials of the ϕ-ψ angles [134] of a coil library [135] as observable states of the HMM. The second model, **VitSTR**, utilizes the intrinsic Ramachandran potentials of the ϕ-ψ angles [134] of a pre-determined, fixed protein structure that is derived from a counterpart of the target protein. The **VitAL** model is used for web server design to predict a possible peptide sequence for any protein target; the designed peptide may act as the inhibitor of the target.

Analogous to our Markov model in Chapter 5, the two offered Hidden Markov models aim to determine peptide sequences with optimum affinity to the selected binding site of a protein. The methods offer a novel procedure for *de novo* peptide design that sequentially generates the peptide by docking its residues pair by pair along a chosen path on a protein. The states are amino acid types, in form of dipeptides, forming the peptide of given length. Three novel approaches are adopted in the proposed design: (i) The first one is the determination of the binding site on a given protein which we determine using the coarse grained GNM [130-132]. Afterwards, the binding path is determined on the target protein by docking an arbitrarily chosen peptide using AutoDock [44]. (ii) The second novel aspect of the model is the choice of the best fitting peptide to this path: all possible amino acid pairs are generated at each point along the path, the binding energies between these pairs and the specific location on the protein are calculated via AutoDock, and the statistical weight of each pair of amino acids are formed. Once all possible pairs are determined for the full path, the Markov based partition function is formed for all possible choices of the peptides using the Ising model matrix multiplication scheme [6]. The transition probabilities based on this partition function are evaluated, and the best peptide for the given surface is selected employing a HMM using Viterbi decoding [80]. The types of amino acids are the hidden variables of the algorithm to be obtained as the solution of

the problem. (iii) The third novel feature of our approach is the consideration of the suitability of the conformations of the peptides that result upon binding on the surface by including the intrinsic Ramachandran potentials of the ϕ-ψ angles [134]. These are the observables of the Viterbi algorithm. As in the choice of the peptides, the Ramachandran torsion angles are assumed to obey Markov statistics according to which a given torsion angle depends on the choice of the preceding torsion angle.

6.1. The Model and Calculations for VitAL

The model consists of five parts: (i) Determining the binding site and the sequence of residues on the surface of the target protein on which the peptide will be docked, (ii) determining the chiral carbon positions of the n residue peptide that will be interacting with the sequence of residues on the target protein surface, (iii) partitioning the path of the n points into a sequence of grids, (iv) docking, by using AutoDock, all 20 peptides to the first grid, and all 400 dipeptides to the succeeding pairs of grids and evaluating their binding affinities, (v) characterizing the ϕ–ψ propensities of the dipeptides. We then use the outcome of these five steps for the implementation of the Viterbi algorithm.

6.1.1. Determining the binding site and the sequence of residues on the target protein

GNM has been shown to predict the protein residues located at specific sites for drug binding [130-132]. We employed GNM to predict 'binding site residues' that play major role in peptide binding. Having identified a site by the GNM, we then need a sequence of residues on the protein that will be in contact with the binding peptide. We choose this sequence of residues using either of the following two approaches: (i) If the protein exists in complex with other proteins, and if the site determined by the GNM lies in an interface in the complex, then the complex is used to determine this sequence of residues on the surface. Recently, two studies indicated that protein-protein interactions, Tuncbag *et al.* [136] and and protein-peptide interactions, Vanhee *et al.* [137] adopt the same structural motifs as monomeric protein folds. The existence of specific folds at the interaction site increases the stability of the formed protein-peptide complex. The counterpart fold of the target protein can be safely used as the fold of designed peptide. The protein crystal structure of the target protein is given as input to the web-server HotPoint [138] which predicts the residues of interest on the surface. In Figure 6.1 an example obtained from the

HotPoint server is given. In this specific example the target protein Human Growth Hormone (HGH), Figure 6.1, is formed by helices and the counterpart protein Human Growth

Figure 6.1. Selection of path by HotPoint server. HGHR (on the left) interacts with HGH (on the right): HotPoint predicts that the regions indicated with green and pink colors interact. The green-colored HGHR residues are selected as binding-sites. The pink-colored HGH residues are selected as binding peptide path.

Hormone Receptor (HGHR) is formed by beta-strands. The binding site of HGH, determined by the GNM, is depicted in pink and the residues that interact with the HGH binding-site, determined by HotPoint, are in green. The region shown in pink will contain the residues to which the peptide will bind. The chiral carbons of the residues shown in green are used as grid-centers for the peptide to be designed, which will be explained in the next section. (ii) If an interacting partner of the protein does not exist, a probe peptide made up of all alanine residues, equal in number to the residues of the peptide to be designed, is docked to the binding site using AutoDock. The specific aim of docking an all alanine peptide is to obtain a path which may well be approximated by the backbone contour. Deviations from this path, due to the forces imposed by placing bulky side groups as the peptide grows on the surface for example, are accounted for as described below. Chiral carbon coordinates of the docked peptide are essential for further steps.

6.1.2. Partitioning the Path into Grids

Our aim is to design a peptide of **n** amino acids. For this purpose, we need **n** points –one for the center of each amino acid of the peptide- that is close to the sequence of residues on the binding surface. We choose these **n** points as the spatial coordinates of the chiral carbons of either the docked probe peptide or the interacting protein portion determined by HotPoint. Once this path of n points is defined, **n** contiguous grid boxes are constructed, each centered around one of the **n** points. The first grid contains the first amino acid. The first and the second grids along the path contain the first and the second amino acids. The t^{th} and $t+1^{st}$ grids contain the t^{th} and $t+1^{st}$ residues. The n chiral carbon atoms of the path define the centers of the n grid boxes.

6.1.3. Docking of amino acids & dipeptides to the binding site

The AutoDock program [44] is used as the docking tool to quantify the binding affinity between the dipeptides and the selected protein surface. Python scripts are written to automate the docking, the probability calculation and the peptide sequence determination procedures. The binding affinity of a given peptide to the surface is determined via AutoDock which gives the affinity in terms of binding energy in kcal / mol.

All 20 amino acids are docked to the 1^{st} grid of the pre-defined path, such that their chiral carbon atoms are forced to coincide with the first grid center. All 400 possible dipeptides are docked to the first and second grid boxes, with the successive chiral carbons located at the grid centers. The pair wise docking of the dipeptides is continued in this way up to the last dipeptide along the path. Pair wise docking rests on the assumption of the Markov property. The amino acids and dipeptides are prepared by the HyperChem software [133]. The dipeptides have Ace-cap on their N-termini.

The grid map is determined by the ADT subprogram of AutoDock. The pre-determined spatial coordinates of the chiral carbon atoms on the path are used as the AutoDock grid box centers. The optimum grid box size is found by trial and error as 2.5 times the length of the distance between the first nitrogen and the last carbon along the backbone of the amino acids. The grids are set such that the chiral carbon atoms of the amino acid coincide with the grid box centers but with a freedom to rotate and translate in the box. This freedom, while keeping the dipeptides to be constrained to the neighborhood of the chosen

grids is necessary to account for the side chain differences of the different amino acids. It also decreases the entropy penalty of constraining a peptide to a certain region.

Docking the dipeptide on a given grid pair leads to a set of 400 binding affinities, one set for each dipeptide, which are used to determine the probability of binding of each dipeptide to the protein binding site as explained in the following section.

6.1.4. Calculation of transition probabilities

The 20 types of amino acids and the peptide length **n** determine the number of states in our model; there exists **20n** states for the problem. For the Markov model adopted here, one needs the transition probabilities, i.e., the probability of an amino acid occupying the $t+1^{st}$ position, given the amino acid at the t^{th} location. The binding energies of dipeptides obtained by AutoDock are used as the statistical weights for determining the transition probabilities. We adopt the Rotational Isomeric States (RIS) approach of polymer physics which is well suited for this purpose [6, 38]. For each grid the probabilities are calculated via the formulation given below by Eqs. 1, 2, 3, and 4. For the 1st grid 20 binding energies are available, while for the remaining grid pairs 400 binding energies are available; the corresponding statistical weight matrix is given in Eq.1 and 2, respectively. The RIS matrix multiplication scheme [6] is used to determine probabilities from the energies.

The statistical weight matrix U_1 for the amino acid bound to the 1st grid box is

$$U_1 = \exp\left(-\beta E_{1;i}\right) \tag{1}$$

where $\beta = 1/kT$, k being the Boltzmann constant and T the temperature, i is any of the 20 amino acids; alanine, cysteine, aspartic acid, glutamic acid, phenylalanine, glycine, histidine, isoleucine, lysine, leucine, methionine, asparagine, proline, glutamine, arginine, serine, threonine, valine, tryptophan, tyrosine.

The statistical weight matrix U_{t+1} for the dipeptide formed by the t^{th} and $t+1^{st}$ residues along the peptide is

$$U_{t+1} = \exp\left(-\beta E_{t,t+1;i,j}\right) \tag{2}$$

where t, $(1 \le t \le n-1)$, represents the amino acid position number (or the grid box number along the path). The *i, j* values represent any of the 20 amino acids.

The partition function, *Z*, of the peptide is obtained according to[6].

$$Z = J^* \prod_{t=2}^{n} U_t \; J \tag{3}$$

where $J^* = U_1$; J = column $\lfloor 1\ 1\ 1 \ldots 1 \rfloor$. (It is to be noted that in the Flory notation, the J* matrix is given as $J^* = \lfloor 1\ 0\ 0\ \ldots 0 \rfloor$ that would assign alanine to the first residue of peptide. In the present formulation, the choice of the J* matrix allows for the acknowledgement of all of the 20 amino acids to be the first residue).

The probability of having residue *i* at the t^{th} position and residue *j* at the $t+1^{st}$ position is determined by:

$$p_{t,t+1;i,j} = \frac{J^* U_2 \ldots U_t U'_{t+1} \ldots U_n J}{Z} \tag{4}$$

Here, U'_{t+1} is the matrix obtained by equating all elements of U_{t+1} to zero except the ij^{th}. The formula given above in *Eq. 4* is used to calculate the probabilities of transition states. Here, we keep the indices t and *t+1*, but they will be dropped in the application to the Viterbi algorithm for simplicity of representation with the understanding that each pair of sites has its own p_{ij} .

6.1.5. Determining the emission probabilities of the ϕ–ψ torsion angles

A fundamental requisite for favorable binding of a peptide to the surface is that the torsion angles of the peptide in the bound conformation should not be forced to have energetically unfavorable ϕ–ψ torsion angle values. The apriori probabilities of the torsion angles are needed for this purpose. In this section we explain the formation of these probabilities, called the emission probabilities, and their incorporation into the Viterbi algorithm as the observable variables.

Two sets of probabilities are needed for specifying the conformation of the peptide. The first set gives the probabilities of the torsion states determined by the angles ϕ_t-ψ_t of the residue at the t^{th} grid. The second set gives the probabilities for the torsion angles ψ_t -ϕ_{t+1} of the dipeptide formed by residues at t^{th} and $t+1^{st}$ grids. The representation of the defined torsion angles are shown in Figure 6.2. The ϕ_t-ψ_t torsion angles of the residue at grid t can select any of eleven regions, as explained below and shown in Figure 6.3. The torsion angles ψ_t -ϕ_{t+1} of the succeeding residue can choose the eleven torsion angles as explained

below and shown in Figure 6.4. Each dipeptide is capped at its N-terminus by an acetyl group in order to define the ϕ_t angle.

The ϕ and ψ angles of a residue cannot adopt all values due to backbone intrinsic torsion propensities and attractive and repulsive interactions of atoms that are in close proximity for certain combinations of these two angles. Among the repulsive interactions, steric hindrances resulting from the side groups are the most pronounced. Hydrogen bonds are the most pronounced favorable interactions. Depending on the type of the residue, these

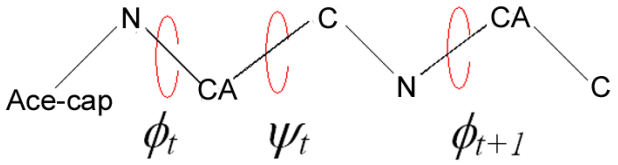


Figure 6.2. Schematic representation of torsion angles of a dipeptide. Only the backbone atoms of dipeptide are given and the acetyl cap on the N-terminal is shown as ACE, for simplicity.

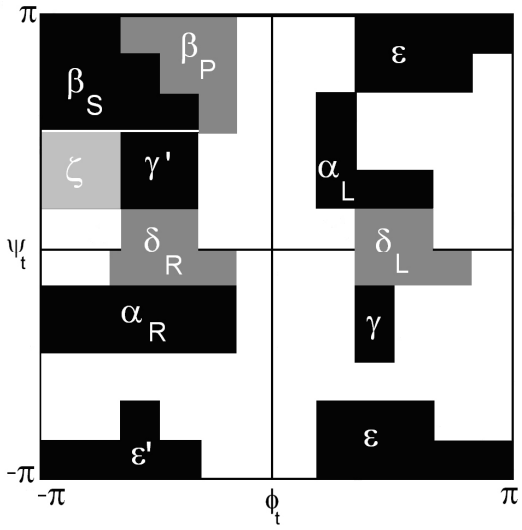


Figure 6.3. The representation of eleven states on Ramachandran map.

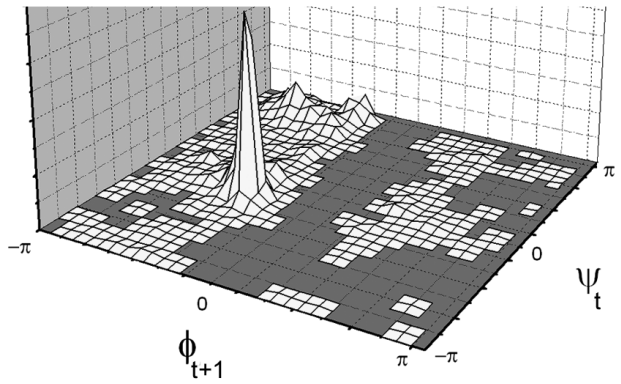


Figure 6.4. The probability distribution of Ψ_t - Φ_{t+1} angles on Ramachandran map.

angles show preferences for different regions on the Ramachandran map. The frequency of occurrence of these regions for the twenty amino acids can be obtained from the Protein Data Bank, PDB. An examination of the frequency of occurrence of the regions for neighboring units shows that there is strong dependence on residue type [38].

Plotting the ϕ_t-ψ_t angles on a Ramachandran map from PDB data, irrespective of residue type shows that there are eleven major isomeric torsion angle states. These are shown in Figure 6.3. We number the regions from 1 to eleven according to the notation of Reference [139] shown in Table 6.1.

Table 6.1. Notation for the eleven states.

State	Notation	Desrcription
1	ε’	Mirror image of the extended region ε
2	ε	The extended regions, $\varphi > 0$, $\psi \sim -+180^o$
3	α_R	Right-handed alpha helix
4	γ	Tight turn region
5	δ_R	The right handed bridge region between two β-strands
6	δ_L	Mirror image of the δ_R region
7	ζ	Region observed mostly in residues preceding PRO
8	γ’	Inverse tight turn region
9	α_L	Mirror image of α_R
10	β_s	Extended beta sheet forming region
11	β_p	Region with extended polyproline-like helices

In obtaining Figure 6.3, the non-redundant PDB Select database of native proteins was used. There are 197,458 points on the Ramachandran map obtained from the database. 96% of these points fell on the eleven regions shown in Figure 6.3. The remaining 4% of points were outside of these regions which are known to correspond to strained conformations of the residues resulting from long range perturbations. We ignored this set of 4%.

The probability of occurrence of a residue in any of the eleven regions given above depends on the type of the library used. Since we are interested in the denatured conformations of peptides, we calculated the probabilities over a coil library that we constructed. The coil library was downloaded from the website: http://www.roselab.jhu.edu/coil/. The library contains fragments selected by Dunbrack's PISCES server according to the following criteria: less than 20% sequence identity, better than 1.6 angstrom resolution, and a refinement factor of 0.25 or better.

A given residue may be of type i $(1 \le i \le 20)$ at the tth grid, and may be in torsion-state m, which is one of the eleven regions shown in Figure 6.3. The frequency of occurrence of ϕ_t-ψ_t in torsion-states m for the i^{th} type of residue was collected in a two dimensional array, $f(i,m)$ which is the normalized frequency that an amino acid is of residue type i in torsion-state m. The singlet probability $b_{i,m}$ that the a residue along the peptide is of type i having the torsion-state m is

$$b_{i,m} = \frac{f(i,m)}{\sum_{m'} f(i,m')} \tag{5}$$

For a given i, $b_{i,m}$ is a column vector of eleven entries. In Figure 6.5, $b_{i,m}$ values are given for GLY as an example.

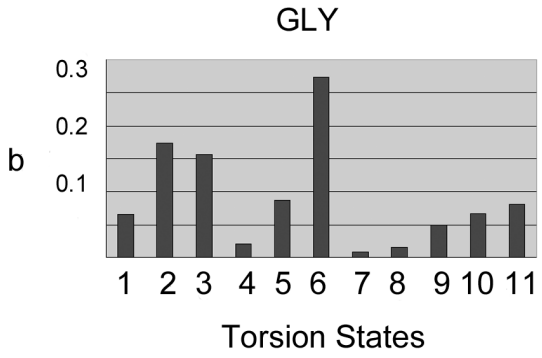


Figure 6.5. The probability distribution of GLY Φ_t - Ψ_t angles, derived from coil library. The dipeptide is mostly observed in the 6th state; and the least observation occurs for state 7.

The choice of the torsion state of the t^{th} residue places restrictions on the choice of the torsion state of residue $t+1$ due to the dependence of the torsion angle ϕ_{t+1} on ψ_t. The extent of this dependence for all pairs of residues in the coil library is depicted in Figure 6.4 where the joint probability distribution of ψ_t-ϕ_{t+1} is presented. For uniformity of representation, we keep the same eleven regions of ψ_t-ϕ_{t+1}. In obtaining Figure 6.4, the non-redundant PDB Select database of native proteins was used. There are 955,679 points on the Ramachandran map obtained from the database. 91% of these points fell on the eleven regions shown in Figure 6.3. The remaining 9% of points were outside of these regions. We ignored this set of 9%. The probabilities are calculated over the coil library of http://www.roselab.jhu.edu/coil/.

The probability $b_{i,j,m}$ that the residue j is in state m of ψ_t-ϕ_{t+1} space when the preceding residue is of type i is

$$b_{i,j,m} = \frac{f(i,j,m)}{\sum_{m'} f(i,j,m')} \tag{6}$$

For a given ij, $b_{i,j,m}$ is a column vector of eleven entries.

For each **n** grid box, the docked conformation of the dipeptides is determined by AutoDock. Knowing the conformation leads to the ϕ_t-ψ_t and ψ_t-ϕ_{t+1} angles. Thus, the

torsion-state of the angles is determined. Equations 5 and 6 give the apriori probability of observing the torsion-states. The product of the two probabilities, $b_{i,m}$ and $b_{i,j,m}$, is called the emission probability for the torsion state of residue *j* when the preceding residue *i* is already prescribed. In the Viterbi algorithm below, this is indicated as $b_{i,j}$. The index *m* will be removed for simplicity.

6.1.6. Implementation of the Viterbi Algorithm

We follow the notation of Reference [83] in our application of the Viterbi algorithm. We need the following definitions:

n : Number of residues of the peptide.

t : Grid number, $1 \le t \le n$

m: Index identifying the torsion state, $1 \le m \le 11$

$S = S_1, S_2, ..., S_{20}$ The 20 natural amino acids set.

$A = \{a_1, a_2, ..., a_{11}\}$: The eleven torsion angle regions.

q_t: The state of the t[th] grid. For example $q_t = S_i$ means that the state of the t[th] grid is the amino acid S_i.

a_t: The torsion-state of the amino acid at the t[th] grid.

$P = (p_{ij})$: The transition probability, $p_{ij} = \Pr(q_{t+1} = S_j \mid q_t = S_i)$.

$b = (b_{i,j})$: The emission probability, $b_{i,j} = \Pr(a_{t+1} = A_j \mid a_t = A_i)$

$\pi = (\pi_i)$: The initial distribution vector, where $\pi_i = \Pr(q_1 = S_i)$.

The sets *S* and *A* describe the structure of the model, *P*, *b* and π describe the parameters. Our intention is to determine $\arg\max_S \Pr(q_1, q_2, ..., q_n \mid a_1, a_2, ..., a_n)$, meaning a peptide sequence made up of preferable torsion angles and with high affinity to protein binding-site. The algorithm is divided into two parts. Firstly, in the forward tracking step, the algorithm finds $\max \Pr(q_1, q_2, ..., q_n \mid a_1, a_2, ..., a_n)$. The forward tracking employs emission and transition probabilities. Then the algorithm backtracks to determine an $\alpha_1, \alpha_2, ..., \alpha_n$ that realizes this maximum.

For an arbitrary position *t* and amino acid type *i*:

$\delta_t(i)$: The maximum probability of all ways to end in state S_i at grid *t* and have observed the torsion states $a_1, a_2, ..., a_t$.

$$\delta_t(i) = \max_{q_1, q_2 \cdots q_{t-1}} \Pr(q_1, q_2, ..., q_{t-1}, q_t = S_i \text{ and } a_1, a_2, ..., a_t) \quad (7)$$

where

$$\delta_1(i) = \Pr(q_1 = S_1 \text{ and } a_1) \quad (8)$$

then,

$$\max_{q_1, q_2, ..., q_n} \Pr(q_1, q_2, ..., q_n \text{ and } a_1, a_2, ..., a_n) \quad (9)$$

is determined.

Initialization step

$$\delta_1(i) = \pi_i b_i(\alpha_1), \ 1 \le i \le 20 \quad (10)$$

$\boldsymbol{\delta}_1$ is 1x20 array, keeping the binding affinity probabilities of each amino acid for the 1st grid box. $\boldsymbol{\pi}_i$ is the initial probability of binding of any 20 amino acids to the 1st grid box and is a 1x20 array. All entries of $\boldsymbol{\pi}_i$ are chosen as unity to give equal chance for selection of any 20 amino acids as the initial residue of the peptide. The binding affinity and the corresponding torsion states are determined by AutoDock. The choice of the first residue will contain the information of its torsion angles from AutoDock, leading to the knowledge of the emission probability $b_i(\alpha_1)$ which is accounted for in Eq. 10.

Induction step

$$\delta_{t+1}(j) = \max_{1 \le i \le 20} \delta_t(i) \, p_{ij} b_{i,j}(a_{t+1} \mid a_t), \ 1 \le t \le n-1, \ 1 \le i, j \le 20 \quad (11)$$

δ_{t+1} is 1x20 array, keeping the binding affinity and torsion angle preference probabilities of each dipeptide for $t+1^{st}$ grid.

Backtracking

For each grid box, the maximum probabilities are kept in $\delta_t(i)$ arrays, as defined by Eq. 11. Backtracking of those arrays leads to the determination of a peptide sequence with a possible affinity to protein binding-site.

Initially, the last residue of the peptide sequence is determined from the 20 entries of δ_n, the array of the last grid box. The maximum entry of δ_n is selected, which determines the amino acid residue q_n.

We let

$$\Xi_n = \arg\max_{1\le i\le 20} \delta_n(i) \tag{12}$$

and choose $q_n = S_{\Xi_n}$. Thus, q_n is the final state of the last residue of the peptide.

The remaining q_i, that is the amino acid types, for $1 \le t \le n-1$ are found recursively by determining:

$$\Xi_t = \arg\max_{1\le i\le 20} \delta_t(i) p_{i\Xi_{t+1}} \tag{13}$$

and then putting $q_t = S_{\Xi_t}$. The backtracking method leads to a peptide sequence with high affinity to the target protein surface.

6.2. The Model and Calculations for VitSTR

VitSTR uses the same model and calculations with VitAL; the only difference of the two models is the emission probability calculation.

The model consists of nine parts: (i) determining the binding site and the sequence of residues on the surface of the target protein on which the peptide will be docked, (ii) determination of the counterpart protein(s) is any exists, (iii) investigation of each protein complex and determination whether the counterpart protein interacts with the pre-determined binding site or not, (iv) selection of **n** residues on the counterpart protein interacting with the selected sequence of residues on the target protein surface, (v) determining the chiral carbon positions of the selected **n** residues, (vi) partitioning the path of the **n** points into a sequence of grids, (vii) docking, by using AutoDock, all 20 peptides to the first grid, and all 400 dipeptides to the succeeding pairs of grids and evaluating their binding affinities, (viii) characterizing the $\phi-\psi$ propensities of the dipeptides forming the selected **n** residues, (ix) characterizing the $\phi-\psi$ propensities of the dipeptides. The outcome of these nine steps is used for the implementation of the Viterbi algorithm.

The target proteins' structure is searched from Protein Data Bank (www.pdb.org) and only the crystal structures formed by complexes are downloaded as PDB files. The protein crystal structures are given as input to the web-server HotPoint [138]. The web-server discovers the interacting residues of protein complexes. The residues of counterpart protein that interacts with the binding site are determined according to the HotPoint results. Based on the HotPoint results, **n** residues are selected on the counterpart binder region. The number **n** is the pre-defined residue number of the peptide to be designed. The chosen **n** residues of counterpart protein (named as *binding path* hereafter) and the target protein are the only necessary inputs for our method. The chiral carbons of the *binding path* are used as grid-centers for the peptide to be designed, as explained below. The conformation of the *binding path* can be safely used to determine the conformation of the designed peptide.

A given residue may be of type i $(1 \le i \le 20)$ at the 1st grid, and may be in torsion-state m, which is one of the eleven regions. If the 1st residue of *binding path* is in torsion-state m, then the singlet probability $b_{i,m}$ that the a residue along the peptide is of type i having the torsion-state m is

$$b_{i,m} = 1 \tag{14}$$

If the 1st residue of *binding path* is in any other torsion-state, then the singlet probability $b_{i,m}$ that the a residue along the peptide is of type i having the torsion-state m is

$$b_{i,m} = 0 \tag{15}$$

For a given i, $b_{i,m}$ is a column vector of eleven entries.

A given residue may be of type i $(1 \le i \le 20)$ at the t^{th} grid, and may be in torsion-state m. A given residue may be of type j $(1 \le j \le 20)$ at the $t+1^{st}$ grid, and may be in torsion-state r, which is one of the eleven regions. If the t^{th} residue of *binding path* is in torsion-state m and $t+1^{st}$ residue of *binding path* is in torsion-state r then the singlet probability $b_{i,j,m,r}$ that the a dipeptide along the peptide is of type i,j having the torsion-states m,r is

$$b_{i,j,m,r} = 2 \tag{16}$$

A given residue may be of type i $(1 \le i \le 20)$ at the t^{th} grid, and may be in torsion-state m. A given residue may be of type j $(1 \le j \le 20)$ at the $t+1^{st}$ grid, and may be in torsion-state r. If the t^{th} residue of *binding path* is in torsion-state m and $t+1^{st}$ residue of *binding path* is not

in torsion-state *r* then the singlet probability $b_{i,j,m,r}$ that the a dipeptide along the peptide is of type *i,j* having the torsion-states *m,r* is

$$b_{i,j,m,r} = 1 \tag{17}$$

Similarly, a given residue may be of type i $(1 \le i \le 20)$ at the t^{th} grid, and may be in torsion-state *m*. A given residue may be of type j $(1 \le j \le 20)$ at the $t+1^{st}$ grid, and may be in torsion-state *r*. If the t^{th} residue of *binding path* is not in torsion-state *m* and $t+1^{st}$ residue of *binding path* is in torsion-state *r* then the singlet probability $b_{i,j,m,r}$ that the a dipeptide along the peptide is of type *i,j* having the torsion-states *m,r* is

$$b_{i,j,m,r} = 1 \tag{18}$$

A given residue may be of type i $(1 \le i \le 20)$ at the t^{th} grid, and may be in torsion-state *m*. A given residue may be of type j $(1 \le j \le 20)$ at the $t+1^{st}$ grid, and may be in torsion-state *r*. If the t^{th} residue of *binding path* is not in torsion-state *m* and $t+1^{st}$ residue of *binding path* is not in torsion-state *r* then the singlet probability $b_{i,j,m,r}$ that the a dipeptide along the peptide is of type *i,j* having the torsion-states *m,r* is

$$b_{i,j,m,r} = 0 \tag{19}$$

For a given *ij*, $b_{i,j,m,r}$ is a column vector of eleven entries.

The choice of the torsion state of the t^{th} residue places restrictions on the choice of the torsion state of residue $t+1$ due to the dependence of the torsion angle ϕ_{t+1} on ψ_t. The probability $b_{i,j,m}$ that the residue *j* is in state *m* of ψ_t-ϕ_{t+1} space when the preceding residue is of type *i* is

$$b_{i,j,m} = 1 \tag{20}$$

if and only if the t^{th} $t+1^{st}$ residues of *binding path* is in state *m*. Otherwise,

$$b_{i,j,m} = 0 \tag{21}$$

For a given *i,j*, $b_{i,j,m}$ is a column vector of eleven entries.

For each **n** grid box, the docked conformation of the dipeptides is determined by AutoDock. Knowing the conformation leads to the determination of ϕ_t-ψ_t and ψ_t-ϕ_{t+1} angles. Thus, the torsion-state of the angles is determined. Equations 14-19 and 20-21 give

the probability of observing the torsion-states. The product of the two probabilities, $b_{i,j,m,r}$ and $b_{i,j,m}$, is called the emission probability for the torsion state of residue *j* when the preceding residue *i* is already prescribed.

6.3. Case Studies

The AutoDock software is used for validating the accuracy of the solutions. The tertiary structure of the designed peptide is prepared using HyperChem. The binding affinity calculation between the target protein and the designed peptide was carried out by AutoDock.

6.3.1. Protein-tripeptide case studies

As a proof of concept, the VitAL is tested on five known protein-tripeptide complexes. The aim is to demonstrate the feasibility and the reliability of the algorithm. The binding energy between the target protein and the peptide designed by Viterbi are determined by the AutoDock program. The inhibition constant K_i is also calculated by AutoDock. The visualization of complexes is achieved by Accelerys Discovery Studio 2.5 program [140].

HIV-1 protease

The first complex is HIV-1 protease interacting with the tripeptide *Glu Asp Leu*. The PDB accession number of this complex is **1A30** [141]. This tripeptide is the smallest analogue of HIV-1 transframe octapeptide (TFP) *Phe Leu Arg Glu Asp Leu Ala Phe*, which is known to be the most potent inhibitor of the target protein. The inhibition of the protein with this peptide is selective and specific.

The VitAL designed the *Trp Tyr Val* tripeptide with high binding affinity for the HIV-1 protease. The binding affinity was calculated as -9.59 kcal/mol by AutoDock; the K_i value was 94.12 nanomolar. The binding affinity of the known inhibitor *Glu Asp Leu* was determined by AutoDock as -7.66 kcal/mol; the K_i value was 2.43 micromolar. The binding region of the peptides is shown in Figure 6.6. Figure 6.6A indicates the complex formed by the HIV-1 protease with *Glu Asp Leu*, while Figure 6.6B indicates the complex formed by the HIV-1 protease with *Trp Tyr Val*. As the figures imply, both of the peptides bind to the same active-site on the protein. The HIV-1 protease is formed by 2 identical chains: chain A and chain B. *Glu Asp Leu* interacts with the chain A residues Asp-25, Gly-27 Ala-28, Asp-29, Asp-30, Met-46, Gly-48; and with the chain B residues Arg-8, Asp-25,

Val-82. The peptide makes five Hydrogen bonds with Gly-27, Asp-29, Asp-30 and Gly-48 shown in Figure 6.7A. A salt bridge is present between Asp-29 of the chain A and Glu residue of the peptide. The sulfide atom of Met-46 from the chain A makes a bond with the oxygen atom of Asp. Trp-Tyr-Val interacts with the chain A residues Asp-25, Gly-27, Ala-28, Asp-29, Asp-30, Gly-48, Pro-81, Val-82, Ile-84; and with the chain B residues Asp-25, Gly-27, Ala-28, Ile-50, Thr-80, Pro-81, Val-82, Ile-84. The peptide makes nine Hydrogen bonds with Asp-25 of the both chains, Ile-50, Thr-80 and Val-82 as shown in Figure 6.7B. There are 2 salt bridges between Asp-25 of the chain A and the Trp residue of the peptide. Another 2 salt bridges are present between Asp-25 of the chain B and Trp residue of the peptide.

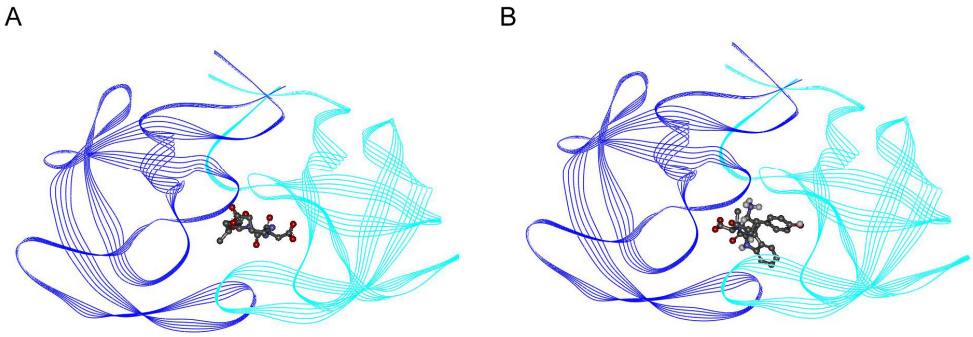


Figure 6.6 HIV-1 protease peptide complexes. (A) Glu-Asp-Leu and HIV-1 protease. (B) Trp-Tyr-Val and HIV-1 protease.

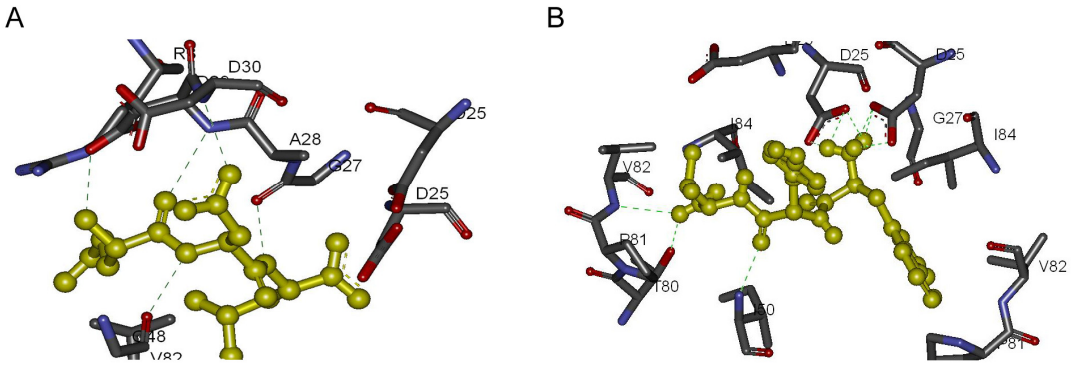


Figure 6.7. Detailed analysis of HIV-protease peptide complexes. Hydrogen bonds are indicated with green lines. (A) Glu-Asp-Leu and HIV-1 protease. (B) Trp-Tyr-Val and HIV-1 protease.

Scytalidocarboxyl Peptidase B

The second complex is scytalidocarboxyl peptidase B protein interacting with the tripeptide *Ala Ile His* (**1S2K**) [142]. The protein is pepstatin-insensitive carboxyl peptidase from the organism *Scytalidium lignicolum*. The crystal structure contains the protein and the cleaved angiotensin II peptide: Ala-Ile-His. The tripeptide is bound to the catalytic residues *Gln-53* and *Glu-136* of the protein.

The VitAL designed the tripeptide *Arg Arg Arg* as potential binder peptide for the scytalidocarboxyl peptidase B protein. The binding affinity of this peptide was calculated as -12.96 kcal/mol; the K_i value was 315.78 picomolar. The known inhibitor *Ala Ile His* binding affinity for the target protein was determined as -5.33 kcal/mol; the K_i value was 124.38 micromolar. The binding region of the peptides is shown in Figure 6.8; Figure 6.8A indicates the complex formed by the scytalidocarboxyl peptidase B with Ala-Ile-His, while Figure 6.8B indicates the complex formed by the scytalidocarboxyl peptidase B with *Arg Arg Arg*. As the figures imply, both of the peptides bind to the same active-site. Ala-Ile-His interacts with the residues Gln-53, Asp-57, Trp-67, Glu-136, Phe-138, Glu-139, Glu-140, Cys-141, and Cys-148. The peptide makes two Hydrogen bonds with Glu-139 as shown in Figure 6.9A. $\pi - \pi$ stacking between Trp-67 and His residues of the peptide is present. There is a salt bridge between Glu-139 and Ala residue of the peptide. Sulfide atom and aromatic ring interaction is present between residues Cys-141 and His; Cys-148 and His. S-O bonding is present between the residues: Cys-141 and Ala; Cys-148 and Ala. Arg-Arg-Arg interacts with the residues Trp-6, Trp-39, Gln-53, Asp-57, Tyr-59, Asp-65, Trp-67, Glu-69, Glu-73, Glu-136, Phe-138, Glu-139, Glu-140, and Cys-141. The peptide makes fifteen Hydrogen bonds with the residues Asp-57, Tyr-59, Glu-69, Glu-73, Glu-136, Glu-139, and Glu-140 as indicated in Figure 6.9B. π - cation interactions are observed between *P*he-138 and Arg-1 residue of the peptide; Trp-39 and Arg-3 residues. There are salt bridges Glu-69 and Arg-1; Glu-140 and Arg-1; Glu-136 and Arg-3; Asp-65 and Arg-3; also an internal salt bridge between Arg-1 and Arg-3 residues of the peptide. Arg-1 interacts with Cys-141 through S-O bonding.

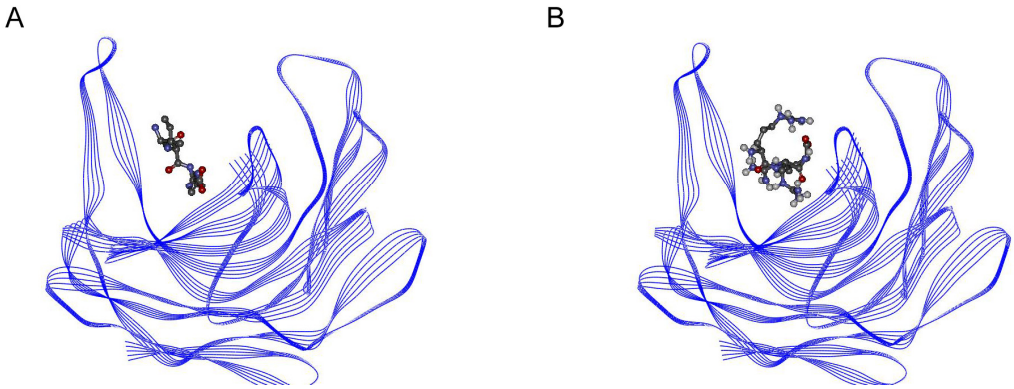


Figure 6.8. Scytalidocarboxyl peptidase B peptide complexes. (A) Ala-Ile-His and scytalidocarboxyl peptidase B. (B) Arg-Arg-Arg and scytalidocarboxyl peptidase B.

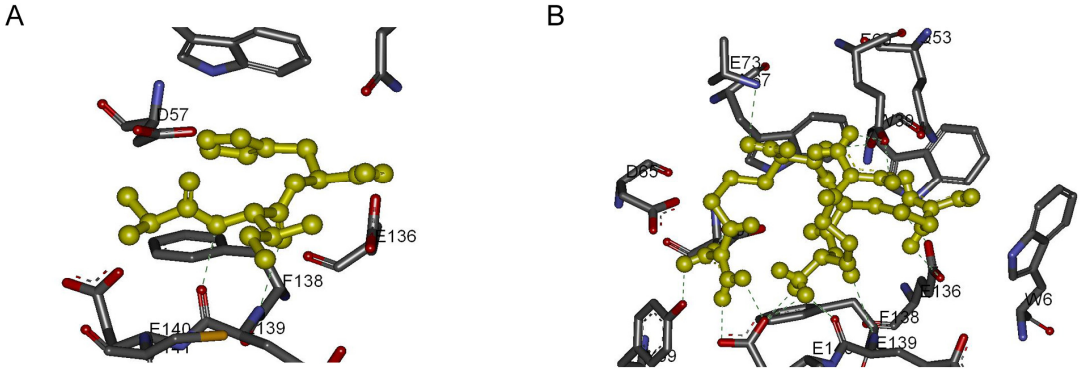


Figure 6.9. Detailed analysis of scytalidocarboxyl peptidase B peptide complexes. Hydrogen bonds are indicated with green lines. (A) Ala-Ile-His and scytalidocarboxyl peptidase B. (B) Arg-Arg-Arg and scytalidocarboxyl peptidase B.

SPG-40

The other complex chosen is the signaling protein from goat mammary gland (SPG-40) and the tripeptide *Trp Pro Trp* (**1ZBW**) [143]. The protein plays role in signaling for reductive remodeling of mammary gland. The remodeling is necessary after cessation of lactation in female mammals. The protein is known to interact with oligosaccharides; the binding enhances protein-protein interactions. The tripeptide Trp-Pro-Trp sits at the active site, to which oligosaccharides bind.

The VitAL designed the tripeptide *Trp Tyr Tyr* as the sequence with possible binding affinity to SPG-40. The binding affinity of this peptide was calculated as -10.97 kcal/mol; the K_i value was 9.04 nanomolar. The known inhibitor *Trp Pro Trp* binding affinity for the target protein was determined by AutoDock as -9.10 kcal/mol; the K_i value was 215.27

nanomolar. The binding region of the peptides is shown in Figure 6.10. Figure 6.10A indicates the complex formed by SPG-40 with *Trp Pro Trp*, while Figure 6.10B indicates the complex formed by SPG-40 with *Trp Tyr Tyr*. As the figures imply, both of the peptides bind to the active-site of the protein. *Trp Pro Trp* interacts with the residues Trp-10, Arg-14, Asn-79, Thr-267, and Glu-269. The peptide makes two Hydrogen bonds with *Asn-79* shown in Figure 6.11A. There is a salt bridge between Glu-139 and Ala residue of the peptide. Trp *Trp Tyr Tyr* interacts with the residues Trp-10, Arg-14, Cys-20, Phe-37, Trp-78, Asn-79, Asp-186, Arg-242, Glu-269, Ile-272, Trp-331, Asp-334, and Leu-335. The peptide makes six Hydrogen bonds with the residues Arg-14, Arg-242, and Glu-269 shown in Figure 6.11B. $\pi - \pi$ stacking is observed between Trp-10 and Trp-1 residue of the peptide and Phe-37 and Trp-1 residues. π - cation interactions are observed between Trp-78 and Tyr-2; Trp-10 and Trp-1; also between Tyr-2 and Trp-1 residues of the peptides; Tyr-2 and Tyr-3 residues of the peptides. There are salt bridges between Glu-269 and Trp-1. There exists aromatic ring and sulfide interaction between Cys-20 and Tyr-2.

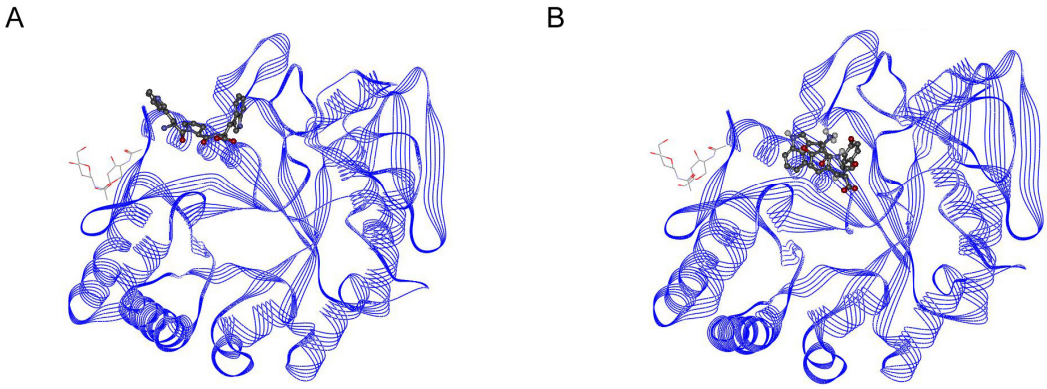


Figure 6.10. SPG-40 peptide complexes. (A) Trp-Pro-Trp and SPG-40. (B) Trp-Tyr-Tyr and SPG-40.

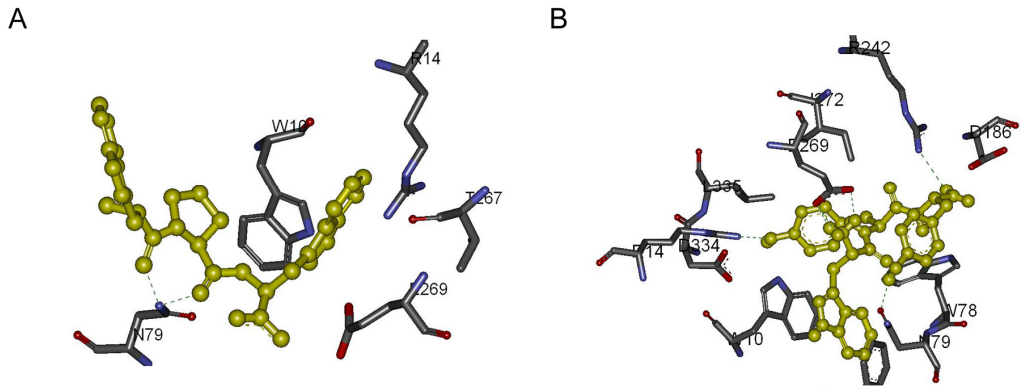


Figure 6.11. Detailed analysis of SPG-40 and peptide complexes. Hydrogen bonds are indicated with green lines. (A) Trp-Pro-Trp and SPG-40. (B) Trp-Tyr-Tyr and SPG-40.

Peptide Deformylase

The peptide deformylase (PDF) from *Enterococcus faecium* organism with Met-Ala-Ser tripeptide is another selected complex (**3G6N**), to which the peptide design algorithm is applied [144]. *E. faecium* are found in normal flora of the intestinal track, but the Vancomycin (an antibiotic) resistant types of bacteria cause infection commonly in hospitals and the strain is resistant to all commercially available antibiotics. The PDF protein is essential for bacterial growth, making it a potential drug target. The *Met Ala Ser* motif is recognized by the active site of the protein.

The Viterbi program designed *Val Trp Trp* as peptide with possible binding affinity to PDF. The binding affinity of this peptide was calculated as -9.52 kcal/mol and the K_i value was 105.71 nanomolar. The known inhibitor Met-Ala-Ser binding affinity for the target protein was determined as -8.07 kcal/mol. The K_i value was 1.21 micromolar. The binding region of the peptides is shown in Figure 6.12. Figure 6.12A indicates the complex formed by the PDF with *Met Ala Ser*, while Figure 6.12B indicates the complex formed by the PDF with *Val Trp Trp*. As the figures imply, both peptides bind to the active-site on the protein. The Fe^{+2} ion in the active site is shown with CPK representation. *Met Ala Ser* interacts with the residues Gly-57, Val-59, Gly-60, His-76, Gly-113, Leu-115, Tyr-150, His-157, His-161, and Phe-167. The peptide makes no Hydrogen bonds with the protein, Figure 6.13A. A sulfide atom and aromatic ring interaction occurs between the residues His-76 and Met; Tyr-150 and Met; His-157 and Met; His-161 and Met; Phe-167 and Met. Val-Trp-Trp interacts with the residues Met-4, Gln-45, Gly-57, Gly-58, Gly-60, Leu-108,

Glu-110, Gly-111, Glu-112, Gly-113, Cys-114, Leu-115, Tyr-150, His-157, Glu-158, Met-166, and Glu-187. The peptide makes four Hydrogen bonds with Gly-111, Gly-113 and Tyr-150 as shown in Figure 6.13B. There are internal salt bridges between Val-1 and Trp-3 residues of the peptide. There is an aromatic ring and sulfide atom interaction between the residue pairs Cys-114 and Trp-3, Met-4 and Trp-3, Met-166 and Trp-3.

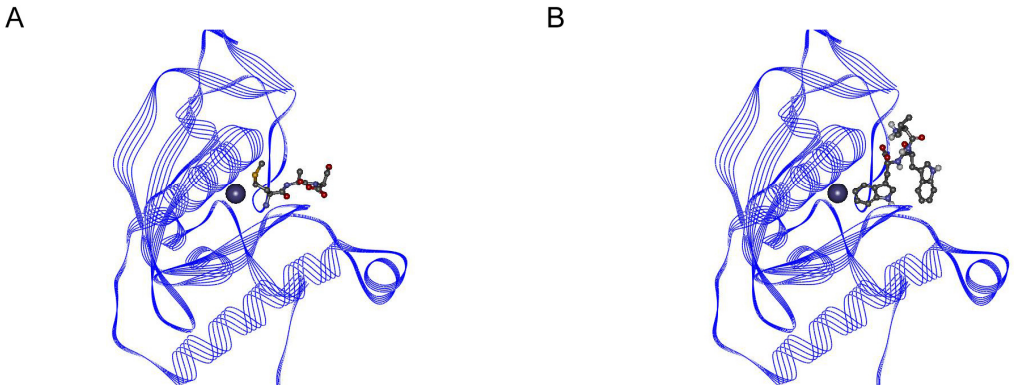


Figure 6.12. PDF peptide complexes. (A) Met-Ala-Ser and PDF. (B) Val-Trp-Trp and PDF.

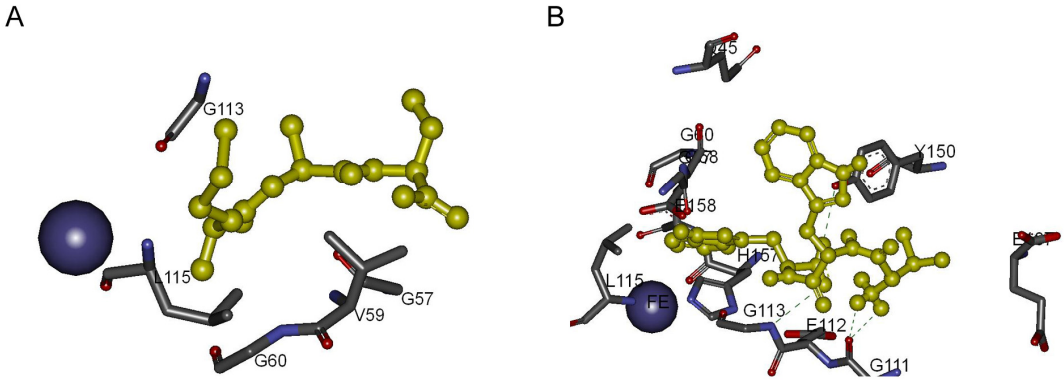


Figure 6.13. Detailed analysis of PDF and peptide complexes. Hydrogen bonds are indicated with green lines. (A) Met-Ala-Ser and PDF. (B) Val-Trp-Trp and PDF.

Concanavalin A

The last test case for tripeptides is concanavalin A (Con A) protein with *Tyr Pro Tyr* peptide (**1HQW**) [145-146]. The peptide is determined to the best binding part out of ~1.4 $x10^9$ octapeptides. A highly diverse phage library procedure is used to determine this sequence. Con A is a carbohydrate-binding protein; the literature indicates the importance of carbohydrate binding in various biological processes. Consequently, inhibition of carbohydrate-specific proteins is important for novel drug development. The chemical synthesis of oligosaccharides is a sophisticated procedure since protection of sugar

monomers is complicated; peptides can be used as ligands for such protein targets. The *Tyr Pro Tyr* peptide is shown to inhibit binding of known monosaccharide ligands to Con A.

The Viterbi program designed the tripeptide *Gly Ala Tyr* for Con A. The binding affinity of this peptide was calculated as -5.70 kcal/mol; the K_i value was 65.83 micromolar. The known inhibitor *Tyr Pro Tyr* binding affinity for the target protein was determined as -5.70 kcal/mol. The K_i value was 144.38 micromolar. The binding region of the peptides is shown in Figure 6.14; Figure 6.14A indicates the complex formed by the Con A with *Tyr Pro Tyr*, while Figure 6.14B indicates the complex formed by the Con A with *Gly Ala Tyr*. As the figures imply, both peptides bind to the active-site of the protein. Tyr-Pro-Tyr interacts with the residues Thr-15, Ser-21. The peptide makes one Hydrogen bond with Thr-15. Gly-Ala-Tyr interacts with the residues Thr-11, Tyr-12, Pro-13, Thr-15, Asp-16, His-205, Pro-206, and Arg-228. The peptide makes seven Hydrogen bonds with Thr-11, Tyr-12, Asp-16, Pro-206 and Arg-228. $\pi - \pi$ stacking is probable between Tyr-12 and Tyr residues.

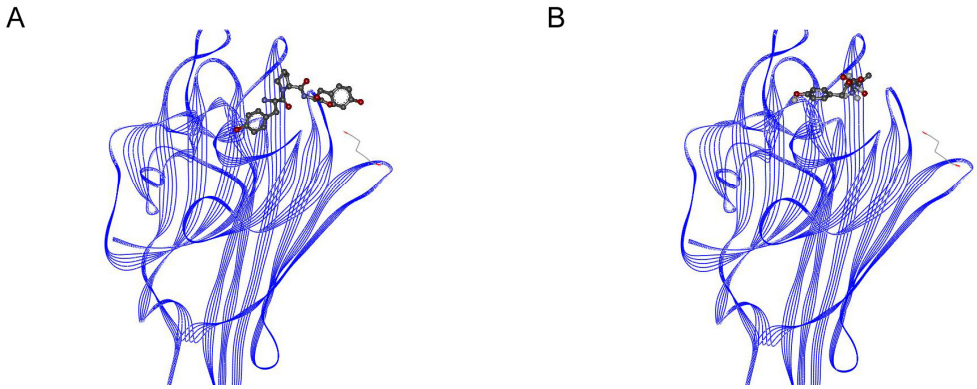


Figure 6.14. Con A peptide complexes. (A) Tyr-Pro-Tyr and Con A. (B) Gly-Ala-Tyr and Con A.

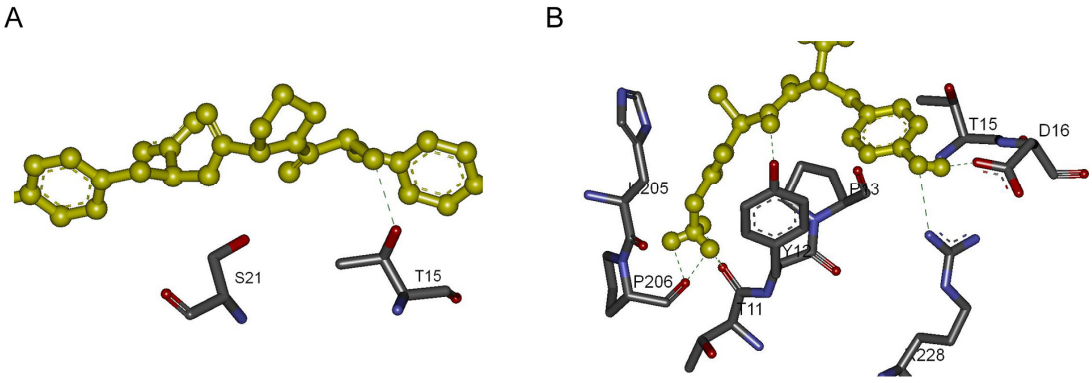


Figure 6.15. Detailed analysis of Con A peptide complexes. Hydrogen bonds are indicated with green lines. (A) Tyr-Pro-Tyr and Con A. (B) Gly-Ala-Tyr and Con A.

6.3.2. Protein- heptapeptide case studies

In the second step of our test, we applied the VitAL method to the design of heptapeptides against Proteinase K, and HLA-B*2705 proteins. Both of these proteins have known peptide ligands in the literature.

Proteinase K

Proteinase K is a serine protease with broad specificity. The PDB accession number of the inhibitor peptide *Pro Ala Pro Phe Ala Ala Ala* in complex with the protein is **1P7V** from the organism *Engyodontium album*. The article about the crystal structure and interaction details of this protein-peptide complex has not yet been published.

The VitAL designed the *Trp Tyr Tyr Tyr Tyr Tyr Tyr* heptapeptide with possible binding affinity to Proteinase K. The binding affinity of this peptide was calculated as -11.59 kcal/mol. The K_i value was 3.22 nanomolar. The known inhibitor *Pro Ala Pro Phe Ala Ala Ala* binding affinity for the target protein was determined as -8.52 kcal/mol; the K_i value was 569.86 nanomolar. Both peptides bind to the active-site on the protein as shown in Figure 6.16. Pro-Ala-Pro-Phe-Ala-Ala-Ala interacts with the residues Asn-67, His-69, Asn-99, Gly-100, Tyr-104, Leu-133, Gly-134, Gly-135, Gly-136, Ala-158, Gly-160, Asn-161, Asn-162, Trp-212, Ile-220, Ser-221, Thr-223, Ser-224, and Met-225. The peptide makes one Hydrogen bond with Gly-102. There exist a π - cation interaction between Tyr-104 and Pro-1. A sulfide atom and aromatic ring interaction is observed between Cys-73, Met-225 residues and Phe-4. The sulfide atom oxygen interaction is present between Cys-73 and Ala-7. The stacking of ring structures is observed for His-69 and Phe-4. Trp-Tyr-Tyr-Tyr-Tyr-Tyr-Tyr interacts with the residues Asn-67, His-69, Asn-99, Ser-101, Gly-102, Gln-103, Tyr-104, Leu-133, Gly-134, Ala-158, Gly-160, Asn-161, Asn-162, Tyr-169, Ser-170, Ala-172, Trp-212, Ile-220, and Ser-224. The peptide makes three Hydrogen bonds with Gln-103, Ser-170 and Asn-161. A sulfide atom and aromatic ring interaction is observed between Cys-73, Met-225 residues and Trp-1, Tyr-2, Tyr3 residues of the peptide. $\pi - \pi$ stacking is observed between His-69, Trp-212 and Trp-1; His-69 and Tyr-2; Tyr-169 and Tyr-3; Phe-192 and Tyr-4; Tyr-104 and Tyr-6; also between Tyr-5 and Tyr-6 of the peptide. A π -cation interaction is present between His-69 and Trp-1.

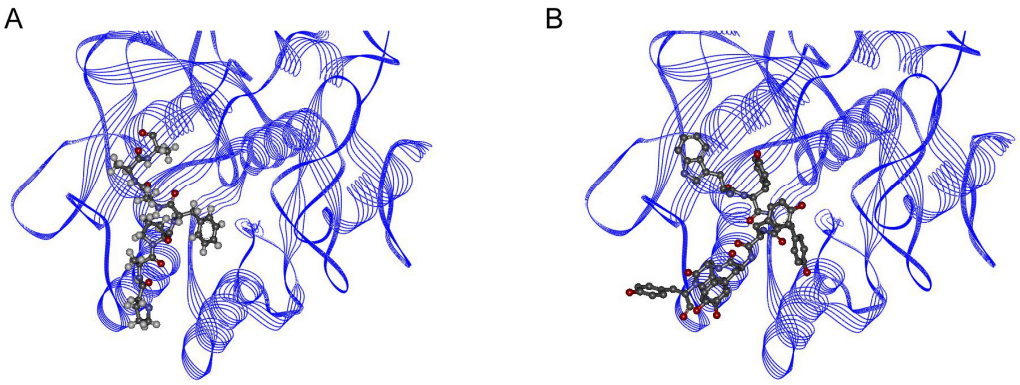


Figure 6.16. Proteinase K peptide complexes. (A) Designed peptide. (B) Literature peptide.

HLA-B*2705

HLA-B*2705 is a disease-associated human MHC class I allele HLA-B27 subtype protein. The protein is the target for nonapeptides. The self peptide sequence of this protein is *Arg Arg Lys Trp Arg Arg Trp His Leu*. The copy number of this peptide in ankylosing spondylitis patients is observed to increase [147]. Viral peptide *Arg Arg Arg Trp Arg Arg Leu Thr Val* derived from Epstein-Barr virus membrane has shown to have affinity to HLA-B*2705 [148]. The glucagon receptor-derived peptide *Arg Arg Arg Trp His Arg Trp Arg Leu* has also proven to interact with HLA-B*2705 [149]. The PDB accession codes of those 3 peptides in complex with HLA-B*2705 are **1OGT**, **1UXS**, **2A83**.

The VitAL designed the *Trp Arg Trp Trp Lys Tyr Tyr* heptapeptide for HLA-B*2705. The binding affinity of this peptide was calculated as -8.97 kcal/mol; the K_i value was 265.82 nanomolar. The known inhibitors *Lys Trp Arg Arg Trp His Leu*, *Arg Trp His Arg Trp Arg Leu*, *Arg Trp Arg Arg Leu Thr Val* binding affinity for the target protein were determined by AutoDock as –6.84, -7.21, and -8.52 kcal/mol, respectively. Both peptides bind to the active-site on the protein as shown in Figure 6.17. Lys-Trp-Arg-Arg-Trp-His-Leu interacts with residues Arg-62, Ile-66, Lys-70, Thr-73, Asp-77, Tyr-99, His-114, Lys-146, Trp-147, Val-152, Gln-155, Lue-156, and Tyr-159. The peptide makes nine Hydrogen bonds with Ile-66, Lys-70, Asp-77, Tyr-84, Tyr-99, Thr-143, Trp-147, and Gln-155. There exist π - cation interactions between Lys-146 and His-6; Lys-1 and Trp-2; Tyr-159 and Lys-1; and Tyr-99 and Lys-1. A salt-bridge is present between Asp-77 and Arg-3. A sulfide atom and aromatic ring interaction is observed between Cys-164, Met-5 residues and Trp-2. Stacking of ring structures is observed for Tyr-99 and Trp-2; Tyr-159 and Trp-

2; Trp-2 and Trp- 7 in the peptide; Trp-147 and Trp-5; Trp-133 and Trp-5; Trp-147 and His-6. Trp-Arg-Trp-Trp-Lys-Tyr-Tyr interacts with the residues Ala-69, Thr-73, Glu-76, Thr-80, Arg-83, His-114, Lys-146, Trp-147, Ala-150, Val-152, Gln-155 and Leu-156. The peptide makes three Hydrogen bonds with Thr-80, Arg-83 and Gln-155. A sulfide atom and aromatic ring interaction is observed between Cys-67 and Trp-4 residue of the peptide. $\pi - \pi$ stacking is observed between Trp-147 and Trp-1; His-114 and Trp-3; Trp-147 and Trp-3; Tyr-99 and Trp-3; Trp-133 and Trp-3; Trp-147 and Tyr-6.

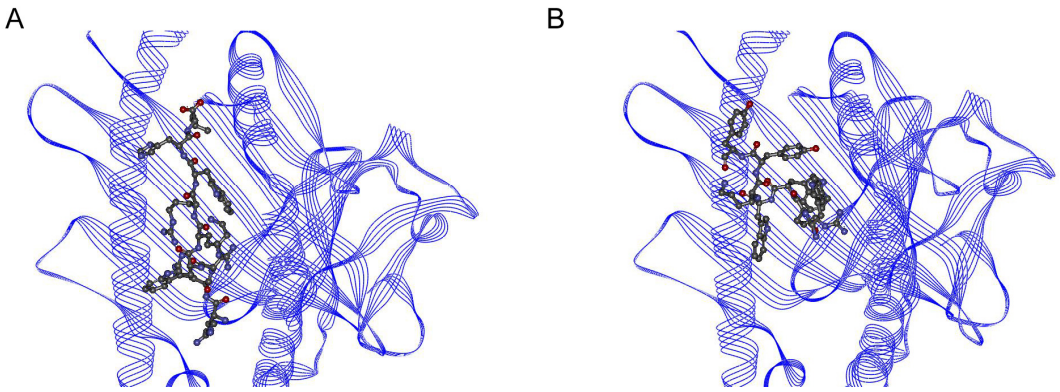


Figure 6.17. HLA-B*2705 peptide complexes. (A) Designed peptide. (B) Literature peptide.

6.3.3. VitAL Case Study: Predicting a peptide for human growth hormone

Growth hormone (GH) is essential for normal linear growth, and have effects on the metabolism [150]. Human Growth Hormone (hGH) is responsible for linear growth in vertebrates via stimulation of skeletal and visceral growth. The protein also plays a role in carbohydrate metabolism and fat mobilization from tissues [151].

It is not possible to purify GH, we aimed to design a peptide bound to the protein may be used to filter out this protein. GH is made up of 4 helix bundles with no symmetry. Human growth hormone (hGH) binds to human growth hormone receptor (hGHR) in 1:1 stoichiometry [152]. Baumgartner et al. [153] indicated that the Ser-226 residue of hGHR is very crucial in hGH binding, Ser interacts with the Glu-174 residue of hGH and offers **Trp X X X Ser** motif for the binding site according to experimental results. Clackson et al. [154] stated that the two nearby Trp residues of hGH-binder protein (hGHbp) play crucial role in the hGH interaction. There is also an Ile residue on hGHbp, which is located next to one of the Trp residues. The other important residues of hGHbp that are in proximity of Trp, Ile and Trp residues are Asn and Arg. Cosman et al. [155] and Bazan et al. [156]

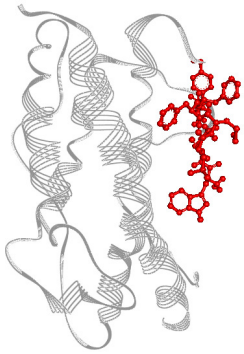
stated that a characteristic Trp Ser X Trp Ser motif is observed in hematopoietic superfamily receptors. According to our knowledge, no hGH binder motif is found with the described conserved motif. Wells [157] studied Alanine scanning of hGH-hGHR complex. Their results indicate that a small set of amino acids are necessary for the two proteins to interact. The most important residues are hydrophobic surrounded by polar and charged interactions of less importance. The Trp-104 and Trp-169 residues of hGHR were found to pack against the Asp-171, Lys-172, and Thr-175 of hGH. Wells indicate that the interaction site for the 2 proteins is large whereas only a few hydrophobic residues are important in binding. Huo et al. [158] devised a computational study to determine the residues playing role in hGH-hGHR binding. The Molecular Mechanics-Poisson-Boltzmann surface area method was applied to achieve computational Alanine scanning and re-computing the binding free energy of the complex. Their studies revealed that Trp-104 of hGHR surrounded by Lys-168, Asp-171, Lys-172, and Thr-175 of hGH play a crucial role for the interaction. The hGHR residues Trp-169, Glu-44, Val-171, Arg-172, Glu-174, Ile-105, Ile-165, Ser-219, and Arg-217 are also very prominent for binding.

The PDB accession code of the protein is **1HGU** [159]. A crystal structure of the protein-peptide complex is not available in PDB. Consequently, the most possible binding site of the protein is determined by GNM.

The VitAL designed *Trp Glu Leu Met Phe Phe Tyr* heptapeptide for HGH. The binding affinity of this peptide was calculated as -8.05 kcal/mol; the K_i value was 1.25 micromolar. The peptide and protein comples is given in Figure 6.18.

Trp Glu Leu Met Phe Phe Tyr interacts with the residues Met-14, His-21, Gln-22, Phe-25, Arg-64, Glu-65, Gln-66, Thr-175, Arg-178, Cys-182, and Cys-189. The peptide makes five Hydrogen bonds with Arg-64 and Arg-178. There exist a π - cation interaction between Arg-178 and Tyr-7. A sulfide atom and aromatic ring interaction is observed between Met-170 and Trp-1; His-18 and Met-4; Met-14, Cys-182, Cys-189 and Phe-5; Met-14, Cys-182, Cys-189 and Phe-6; Cys-182, Cys-189, Met-14 and Tyr-7. Sulfur – oxygen interactions exist between Ser-188 and Met-4; Cys-189 and Met-4; Cys-182, Cys-189 and Phe5; Cys-182, Cys-189 and Phe-6; Cys-182 and Tyr-7. Stacking of ring structures is observed for His-21 and Trp1; Phe-25 and Trp-1; His-18 and Trp-1; His-18 and Phe-5.

A



B

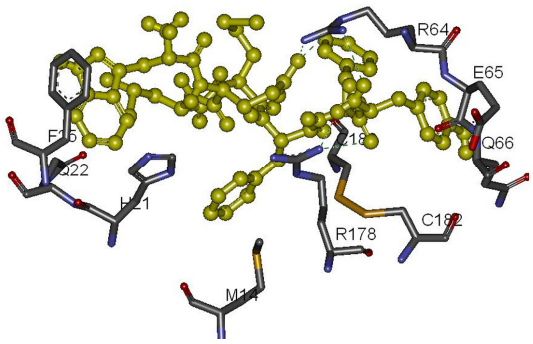


Figure 6.18. hGH peptide complex. (A) Overall structure of the comples. (B) Detailed analysis of peptide bound on hGH, peptide is indicated with yellow color.

The HGH binding region is made up of hydrophobic residues by 18 %. The net charge of the surface is +2 and the net charge of the designed peptide is -1. Consequently, there exist an electrostatic complementarity between the target protein and the designed peptide. There exist two *Cys* and one *Met* residues on protein binding site, which are potential stabilizing residues of peptide binding. The designed peptide also has a Met residue, which makes interactions with the sulfur atom on its side-chain. Numerous sulfur - aromatic ring, sulfur – oxygen interactions are observed due to the *Cys* and *Met* residues of both the

peptide and the protein. The *Trp* and *Phe* residues of the peptide enable stacking of aromatic rings.

6.3.4. VitSTR: Predicting a peptide for human growth hormone

The binding site of hGH is determined to be around His-18 and His-21 residues of hGH by GNM, as shown in Figure 6.19. The counterpart protein (hGHR) residues, which are in contact with the selected binding site, is chosen: *Gln Arg Asn Ser Gly*. The conformation is similar to hairpin structure. Potential penta-peptides with affinity towards hGH are determined.

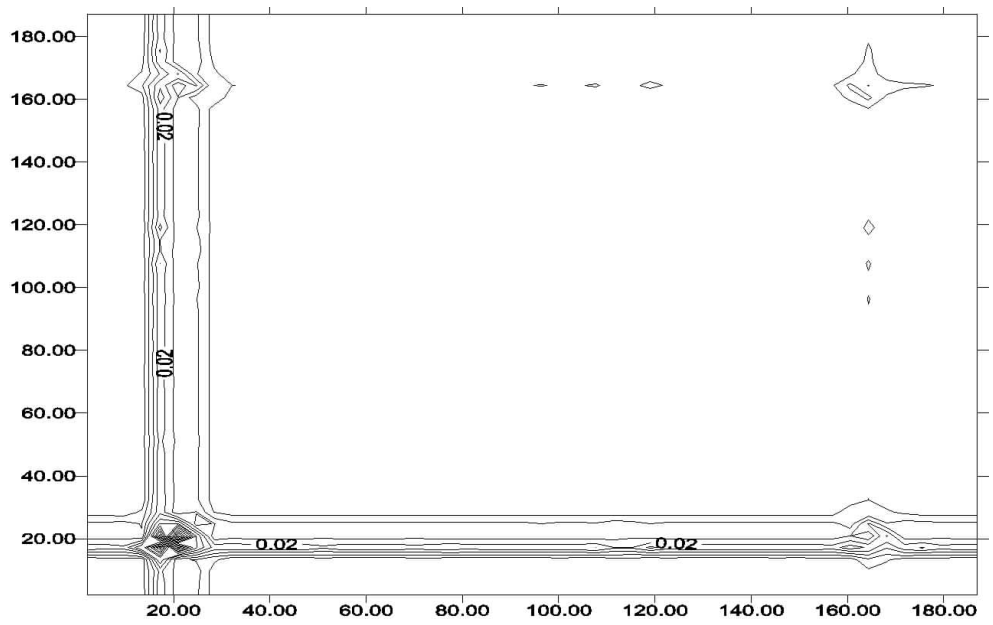


Figure 6.19. The hot-spot residues of hGH protein determined by GNM.

The Viterbi algorithm designed *Trp Trp Ala Ser Tyr* peptide for hGH protein, which is suitable for the **Trp X X X Ser** motif of Baumgartner et al. [153]. The binding affinity is calculated as -9.27 kcal/mol by AutoDock. The inhibition, K_i, is determined to be 161.53 nanomolar. On the other hand, the counterpart protein segment *Gln Arg Asn Ser Gly* peptide binding affinity is determined to be -5.52 kcal/mol; the K_i value of the peptide is 89.23 micromolar. The bound conformations of the peptides are shown in Figure 6.20. This figure indicates the complexes formed by the hGH with *Gln Arg Asn Ser Gly* (green) and *Trp Trp Ala Ser Tyr* (yellow) peptides.

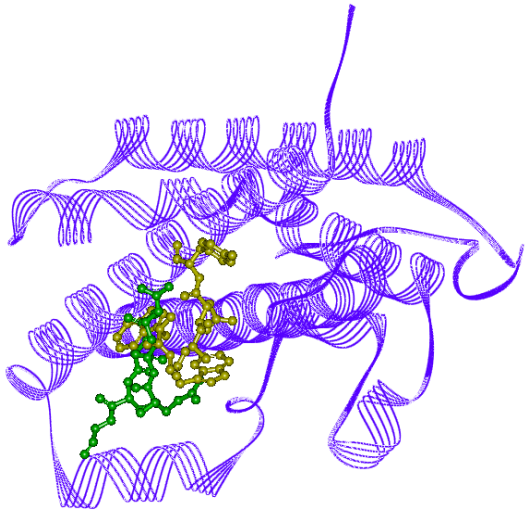


Figure 6.20. hGH with Gln Arg Asn Ser Gly (green) and Trp Trp Ala Ser Tyr (yellow) peptides.

The interactions between protein – peptide complexes are examined in detail with Accelrys Discovery Studio program and our Python scripts. The results imply that the Viterbi peptide has much more interactions with hGH compared to hGHR derived peptide. *Gln Arg Asn Ser Gly* interacts with hGH residues: His 18, Gln 22, Glu 174. There exists a π -cation interaction between Arg 2 of peptide and His 21 of hGH. The π -cation interaction energies are of the same order of magnitude as hydrogen bonds or salt bridges and play an important role in molecular recognition.No hydrogen bond is observed for this complex. *Trp Trp Ala Ser Tyr* interacts with hGH residues: Asp 11, Met 14, Leu 15, His 18, His 21, Gln 22, Phe 25, Asp 171, Glu 174, Thr 175, Arg 178, Gln 181. There exist hydrogen bonds between the residues: Trp 1 – His 21, Ala 3 – Arg 178, Trp 1 – Glu 174. There exists aromatic ring – sulfur interaction between the residues: Trp 1 – Met 14, Trp 1 – Met 170, Trp 2 – Met 14, Trp 2 – Cys 182, Tyr 5 – Met 14, Tyr 5 – Cys 182, and Tyr 5 – Cys 189. There exist S – O atoms interaction between the residues: Ala 3 – Cys 189, Asn 4 – Met 14, Tyr 5- Cys 189. There exist $\pi - \pi$ interactions between the residues: Trp 1 – His 18, Trp 1 – His 21, Trp 1 – Phe 25, Trp 2 – His 21.

As Figure 6.20 implies the hGHR derived peptide and the Viterbi peptides have the same conformation. The binding sites of the two peptides differ slightly. The difference is caused by allowing the grid boxes to be 2.5 times the amino acid sizes for Viterbi

calculations. *Gln Arg Asn Ser Gly* peptide is more exposed to solvent, while the Viterbi peptide is embedded to protein surface. In physiological conditions, the counterpart protein hGHR interacts with hGH protein with various additional residues. The peptide *Gln Arg Asn Ser Gly*, which is a small part of hGHR, is not able to bind to hGH surface as strongly as the whole hGHR protein given that many essential interactions are absent. The designed peptide is able to interact better with hGH protein, albeit it has 5 residues as *Gln Arg Asn Ser Gly* peptide.

The binding energy value of this peptide outperforms the peptide designed by VitAL method using a probe peptide. The new penta-peptide has affinity of -9 kcal/mol for hGH, while the VitAL designed hepta-peptide has affinity of -8 kcal/mol. In other words the binding affinity per amino acid is -1.8 kcal/mol for the described method and -1.14 kcal/mol for the binding path determination by a probe peptide method. Both of the methods are successful to predict potential binder peptides, but it is apparent that using a *binding path* derived from a partner protein is a better choice than docking a probe peptide to determine a *binding path*. The other reason for the improvement may be a result of using the pre-defined torsional angle values instead of employing the probability values derived from coil library. In other words, the use of torsional angles that are specific for the *binding site* surpasses the use of torsional angles derived from a general database. The capability to design a binder peptide increases with the use of properties based on the selected protein binding surface.

6.3.5. VitSTR: Predicting a peptide for caspase-1

Cytokines are secreted proteins/peptides/glycoproteins that play role in signaling for the immune system. Their secretion increases the adhesion factor expression, which enables the leukocyte transmigration to an infection site. Interleukin-1a (IL-1a) and interleukin-1β (IL-1β) are precursor peptides that act as cytokines. Those peptides are cleaved by caspase family proteins to become mature and active [160]. Caspase-1, also named as interleukin-1 converting enzyme (ICE), plays role in inflammatory response, cell proliferation, differentiation, and apoptosis. This protease is a homodimer, and each part is made up of 2 subunits p10 and p20. ICE recognizes Asp-Ala on IL-1β for its cleavage. Various inhibitors of ICE are available, some of which are peptides and peptide mimetics. Shahripour et al. [161] determined diphenyl ether sulfonamide inhibitors for ICE: the

inhibitor leads to a *gauche* to a *trans* orientation of His237 of ICE. The orientation chane of His leads to formation of a large hydrophobic pocket that strengthens the binding. Mialli et al. determined a potetn peptide: Ac-Tyr-Val-Ala-Asp-CO–$(CH2)_5$–Ph that inhibits ICE with K_i=19 nM [162]. Shahripour et al. modified this peptide by replacing Tyr-Val-Ala with a benzyloxycarbonyl group [163]. ICE inhibitor Ac–Tyr–Val–Ala–Asp–H shows potent ICE inhibitory activity with a K_i = 0.76 nM [164], the PDB accession code is **3IBC** for this protein-peptide complex.

There are two binding sites on ICE: the allosteric and active sites [165] (Figure 6.21). Ala-scanning released that Cys331, Arg286, Glu390, Ser332, Asp336, His237, Asn337, Ser333, Thr334, and Thr388 are important residues for protein interactions. The structure-based design and genetic algorithm studies released that peptide sequences *Tyr Val Ala Asp, Asp Glu Val Asp, Ile Glu Thr Asp, Trp Glu Val Asp,* and *Val Glu Thr Asp* have affinity for ICE [76]. The *Asp Glu X Asp* motif was found to bind the conserved pocket with Arg179, Gln283, Ser339 and Arg341 residues [166].

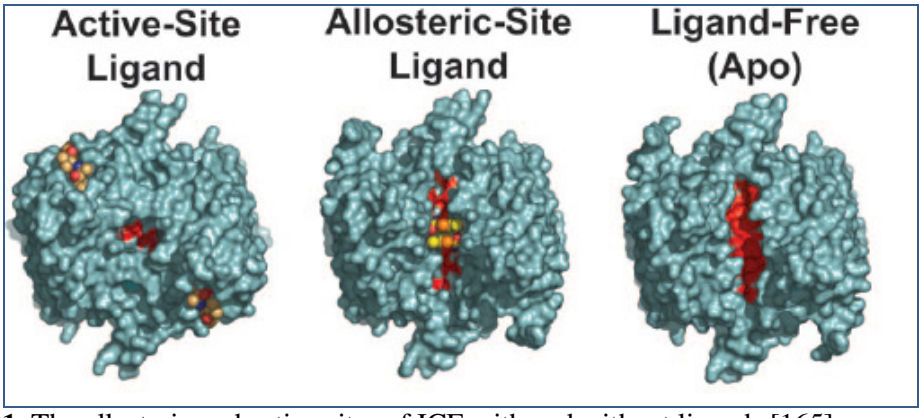


Figure 6.21. The allosteric and active sites of ICE with and without ligands [165].

VitSTR algorithm is applied to ICE with the PDB accession code **1ICE**. The PDB file contains the protein with the peptide *(Tyr Val Ala Asp)-cmk*. The path for VitSTR is selected as the *Tyr Val Ala Asp* peptide chiral carbon atom coordinates. The algorithm output is *Gly Gly Tyr Phe* peptide. This peptide is docked to ICE, and the Arg341 residue is the binding center. The binding energy is determined to be -8.72 kcal/mol, with K_i of 407.06 nM. The binding energy of *(Tyr Val Ala Asp)-cmk* is -7.27 kcal/mol.

The neighboring residues of the *Gly Gly Tyr Phe* peptide are Ser339, Trp340, Arg341, His342, Met345, Gly346, Ser347, Val348, Asp381, Arg383. The peptide makes 5 hydrogen bondes with His342, Arg383, and Asp 381. Figure 6.22 indicates the protein-peptide complex, the protein is shown as surface representation.

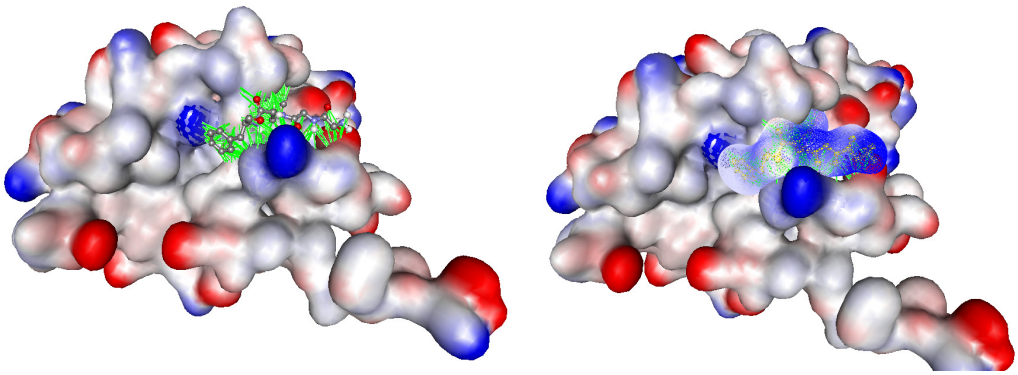


Figure 6.22. The *Gly Gly Tyr Phe* peptide bound on ICE cavity.

The *Tyr* residue of literature peptide *(Tyr Val Ala Asp)-cmk* and the *Tyr* residues of VitSTR peptide *Gly Gly Tyr Phe* bind to the exactly same region on the protein (Figure 6.23). Also the hydrophobic residue *Val* of *(Tyr Val Ala Asp)-cmk* and the *Phe* of *Gly Gly Tyr Phe* bind to the same region. The binding energy of *Gly Gly Tyr Phe* is better than the binding energy of the literature peptide.

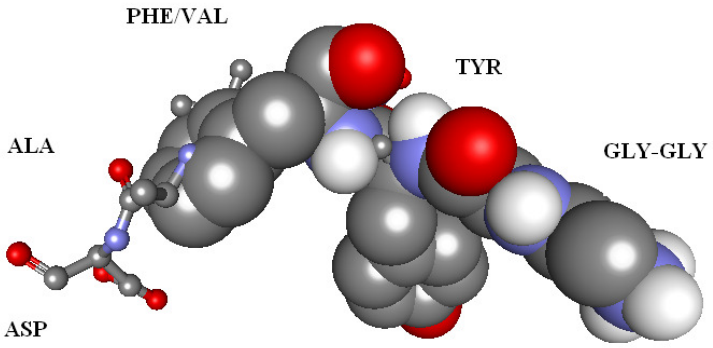


Figure 6.23. The *Gly Gly Tyr Phe* is superimposed with *Tyr Val Ala Asp*. The peptide *Tyr Val Ala Asp* is indicated with ball-and-stick representation, and the peptide *Gly Gly Tyr Phe* is shown with CPK representation.

Since the Ala-Asp residues of the literature peptide is observed to be flanking, it is a good approach to combine *(Ala-asp)-cmk* with the VitSTR peptide *Gly Gly Tyr Phe*.

Similarly other literature peptides can be modified according to the VitSTR peptide sequence. The known inhibitor of ICE is given in Table 6.2 and the modified peptides are given in Table 6.3. According to the Table 6.2 peptide sequences: Ala-Asp, His-Asp, Val-Asp, Ile-Asp, and Thr-Asp dipeptides are added to the C-terminal of VitSTR peptide.

Table 6.2. The inhibitors of ICE from literature.

Val **Ala Asp**
Trp Glu **His Asp**
Val Asp Val **Ala Asp**
Asp Glu **Val Asp**
Tyr Val **Ala Asp**
Val Glu **Ile Asp**
Ile Glu **Thr Asp**
Leu Glu **His Asp**
Ala Glu **Val Asp**

Table 6.3. The modified VitSTR peptides and their affinity for ICE.

Peptide	Min. Binding Energy (kcal/mol)
Gly Gly Tyr Phe Ala Asp-cmk	-6.61
Gly Gly Tyr Phe His Asp- cmk	-7.18
Gly Gly Tyr Phe Ile Asp- cmk	-8.51
Gly Gly Tyr Phe Thr Asp- cmk	-7.36
Gly Gly Tyr Phe Val Asp- cmk	-7.28
Tyr Val Ala Asp -cmk	-7.27

The affinities of the modified peptides for ICE are promising. Experimental verification for each complex is necessary. The conservation of *Val* and the aromatic residue *Phe* instead of *Tyr* is promising. The addition of 2 *Gly* residues at the N-terminal increases the peptide affinity for the protein surface.

6.4. VitAL Server: Viterbi Algorithm Server for *de novo* Peptide Design

Web-server aims to predict a possible peptide sequence for any protein target, the designed peptide may act as the inhibitor of the target. The algorithm is based on VitAL that is described in previous sections. It performs probability calculations based on docking search. The algorithm searches for an appropriate peptide sequence bound to the pre-

defined binding residues of the target protein, which are supplied by the user. The VitAL server uses 3 python scripts and several built-in inputs. The built-in inputs are 20 amino acid and 400 dipeptide structure files, the torsion angle probabilities based on coil-library of Rose Lab. (**http://www.roselab.jhu.edu/coil/**). The e-mail of the user is necessary for the result retrieval and job view during the algorithm continues running. The VitAL web server is available at **http://saros.hpc.eng.ku.edu.tr/vital/**. Server interface is coded in PHP. The codes to predict peptide sequence are written in Python.

6.4.1. Input for the Server

Input data are two files: first one is the protein structure in PDB formatted coordinate file, and the second one is a text file containing the hot-spot residue IDs of the target protein and the desired peptide length. The input text file format is given in Figure 6.24, this example file indicates that residues 151, 152, 153 from chain A are hot-spot residues of the target protein and the desired peptide length is 2 as indicated in the last line. The binding surface is defined by the hot-spot residues supplied by the user. VitAL requires chain identifiers which confine to a protein interface. Server does not work for PDB files containing without any chain identifiers and returns an error. If there is any missing data, the server informs the users of what is missing and terminates the job. The algorithms works for natural 20 amino acid types by default, but user can restrict the amino acid types by selecting the desired types from the submission page (Figure 6.25). The VitAL is free and open to all users and there are no login requirements.

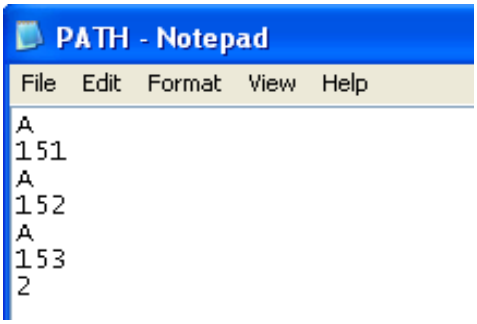


Figure 6.24. An example input file for VitAL server. The residues 151, 152, 153 from chain A are hot-spot residues of the target protein. The desired peptide length is 2 as indicated at the last line.

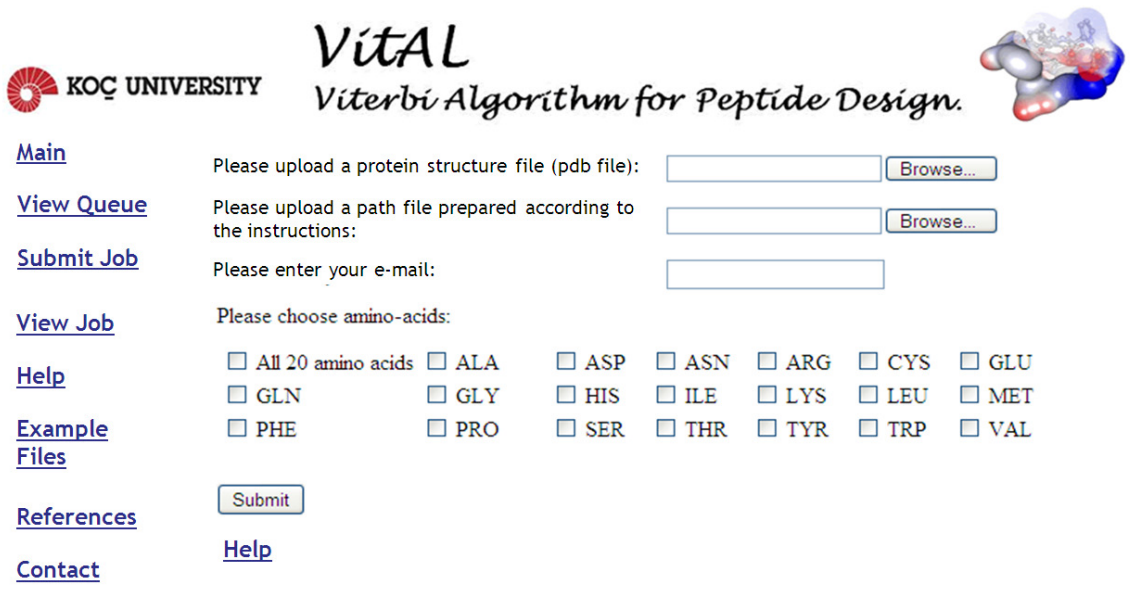


Figure 6.25. The job submission page of the VitAL server.

6.4.2. Output of the Server

After the necessary input files are submitted, a job ID is given to each submitted job enabling the user to track his job (Figure 6.26A). Once the job ID is entered by the user, the details of the given job can be tracked as indicated in Figure 6.26B.

A

B

Figure 6.26. A. The job view page of the VitAL server. **B.** The result for a given job ID, 100642.

VitAL induces the following steps:

1. determines a binding path for the given target protein
2. docks amino acids and dipeptides to the binding site
3. calculates the emission and translation probabilities; determines the peptide sequence
4. docks the peptide to the binding site and calculates the binding energy.

While the job is running, the server informs users about the steps it is performing. The steps can be viewed from the window shown in Figure 6.26B, the “details” link directs the user to a new page (Figure 6.27). If the job is finished, the results are also visualized from the same page: the peptide sequence and the binding energy of this peptide are given. The protein-peptide 3D complex can be downloaded in PDB format, from the “view job” page (Figure 6.26B).

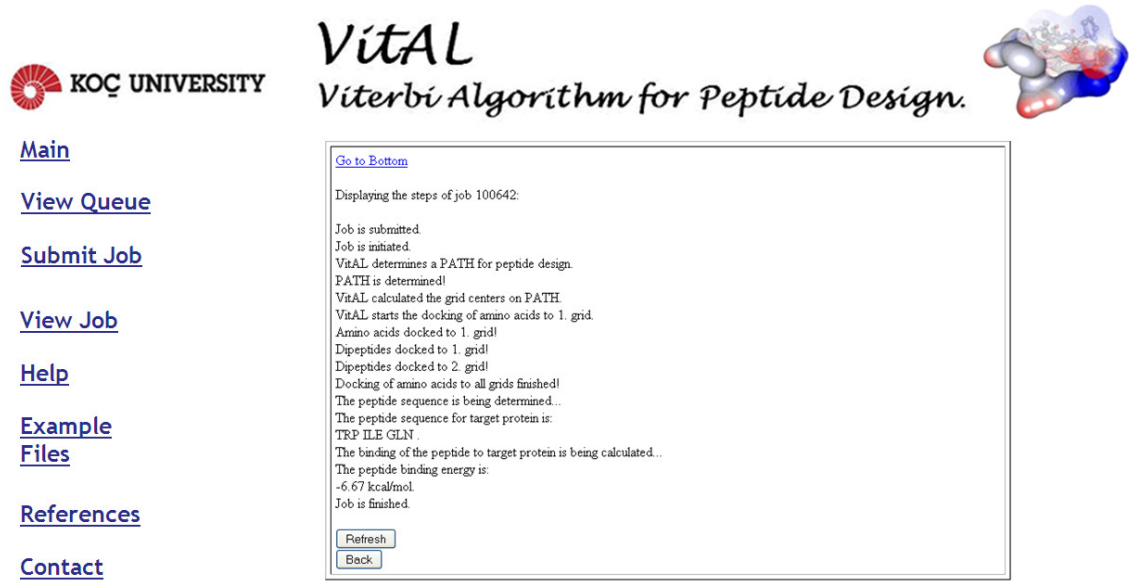


Figure 6.27. The details and the results for a given job, 100642.

The benchmark for VitAL server is given in Appendix 2. Pentapeptides from PEP_X database [167] are employed for the benchmark. The Ramachandran map is divided into 21 areas (as explained in Appendix 2) and this alteration has improved the results compared to 11-state Ramachandran map. The results are exciting since the physic-chemical protperties of the peptides are conserved. The AutoDock favors Trp / Tyr / Phe residues due to its force-field. This problem can be solved via applying other force-fields.

6.5. Concluding Remarks

Novel *de novo* peptide design approaches are developed to block activities of disease related protein targets. No prior training, based on known peptides, is necessary. The methods sequentially generate the peptide by docking its residues pair by pair along a chosen path on a protein. The binding site on the protein is determined via the coarse grained Gaussian Network Model. A binding path is determined. The best fitting peptide is constructed by generating all possible peptide pairs at each point along the path and determining the binding energies between these pairs and the specific location on the protein using AutoDock. The Markov based partition function for all possible choices of the peptides along the path is generated by a matrix multiplication scheme. The best fitting peptide for the given surface is obtained by a Hidden Markov model using Viterbi

decoding. The suitability of the conformations of the peptides that result upon binding on the surface are included in the algorithm by considering the intrinsic Ramachandran potentials.

The model is tested on known protein-peptide inhibitor complexes. The detailed analysis of the designed peptide interactions and comparison of those peptides with the known peptides indicate that our method is able to detect the requirements of binding surface; such as hydrophilicity, electrostatic compatibility, aromatic interactions, sulfur atom and its interactions.

The residues Trp, Tyr and Val are highly favored on the designed peptides. The binding energy of dipeptides containing Trp and Tyr are observed to be superior to other dipeptides. Consequently the transition probabilities of those dipeptides are favored. This may be the reason for high occurrence of those residues in the designed peptide sequences. The conservation of Trp may also be due to the formation of stacking and consequent stabilization. When the protein – designed peptide complexes are analyzed in detail, it is obvious that Trp residues form high number of stacking due to its aromatic ring; this stabilizes the formed complex. The same observation applies to Val for hydrophobic regions; since it has a smaller side-chain compared to other hydrophobic residues - except Ala -; the amino acid is able to interact with small hydrophobic cavities. Nature seems to protect interactions containing amino acids with aromatic side-chains; our methodology also conserves those residues.

The only case that the method failed to determine an outstanding binder peptide is for the Con A protein case. The surface of this protein is highly exposed to water. Consequently, we may state that our algorithm works best for cavities on protein surfaces, with both hydrophilic and hydrophobic residues present. The importance of the binding surface selection is highlighted: a binding path with no cavity and made up of all hydrophilic residues was shown to be not very suitable for determination of a candidate binder peptide. The method is shown to be successful to determine peptide according to the specific properties of the binding surface.

The algorithm requires $O(mn)$ memory and $O(mn^2)$ time to run, where n is the peptide length and m is 20 - the number of states. The time-consuming part for our methodology is the docking process. The Viterbi algorithm works efficiently to determine a single peptide is superior to other possible peptides, i.e. the most possible peptide sequence.

Chapter 7

PEPTIDE MOTIF SEARCH FOR PROTEIN SURFACES: Viterbi and Genetic Algorithms Combined

Our aim is to determine peptide motifs without any prior information the binder sequences of target protein. The combined GA and VitAL/VitSTR algorithms determined peptide motifs for the selected protein binding surfaces. A binding motif is achieved by analysis of the designed peptides with the coarse grained GNM [130-132]. A binding path on the selected binding site is necessary for our calculations. The binding path is selected from a consensus path formed by the partner proteins of the target protein, the HotPoint server [138] is utilized. The best fitting peptide to the selected path is determined via generating all possible amino acid pairs at each point along the path, and calculating the binding energies between these pairs and the specific location on the protein via AutoDock [44] and forming the statistical weight of each pair of amino acids. A Markov based partition function is formed for all possible preferences of the peptides using the Ising model matrix multiplication scheme [6] . The transition probabilities based on this partition function are evaluated, and the best peptide for the pre-defined path is selected via a HMM using Viterbi decoding [81] . The types of amino acids are the hidden variables, and the intrinsic Ramachandran potentials of the ϕ-ψ angles [134] are the observables of the algorithm. The peptide sequence determined by Viterbi algorithm is optimized via a genetic algorithm [71].

7.1. The Model and Calculations

The genetic algorithm is applied as defined in Chapter 5. The algorithm requires target protein structure, a specific binding site, and 50 peptide sequences (initial population) as input. The algorithm creates the structure files of given peptide sequences and then evaluate the binding affinity of each peptide for the pre-defined binding site. The binding affinity is calculated by AutoDock. The best 4 peptides are directly passed through the next population, remaining 46 peptides are formed by crossover and mutation on the pre-existing peptides.

The binding site is a pre-defined GNM site. The input peptide sequences of GA need not to be binder molecules. The output of the Viterbi algorithm is passed to GA, as an input. This peptide is set as the best binding peptide of the initial population. Randomly created 49 peptides are the other input variables of the algorithm. The random peptides are created by our Python script, without using any prior knowledge about the binding site. Each peptide is made up of **n** amino acids. Those 50 peptides form the *initial population*. The 3-dimensional structures, i.e. *.pdb* files, of the peptides are created by *mutate* command of the VMD software using an initial *.pdb* file of peptide made up of **n** alanine residues as input. The peptide torsion angles are set to be flexible by AUTOTORS utility of AutoDock. Polar hydrogen is added to protein and peptide, Gasteiger charges are added by ADT. The grid map is determined by ADT, the binding site is taken as grid center. Genetic Algorithm option of AutoDock for docking is selected. Lamarckian Genetic Algorithm is chosen as the docking search parameter. The population size is set to 150, run = 50, maximum number of energy evaluations = 250 000, number of generations = 27 000. The docking parameters are set such that computation time is feasible, since the aim is to rank peptides rapidly. The 3 AutoDock outputs are the docked conformation of the peptide, the free energy of binding, and the estimated inhibition constant. The free energy of binding is the summation of intermolecular energy (van der Waals energy, hydrogen bonds, desolvation energy, electrostatic energy), final total internal energy, torsional free energy of peptide and unbound system energy. The free energy of binding is used as the fitness value of our GA. As each generation is finished, the *population* of peptides is ranked according to their fitness values. The *new population* is formed by the *tournament selection* and *elitism* procedures from the ranked *current population*.

7.2. Results

7.2.1. Human Growth Hormone Peptide Motifs

The VitSTR determined *Trp Trp Ala Ser Tyr* peptide for hGH protein, as stated in Section 6.3.4. This peptide sequence and 49 randomly created peptides are set as the initial population for GA. The generation number for GA is set as 10, total of 450 docking are designed for the target protein. The designed peptides are clustered into 4 groups: low or no affinity, medium-affinity, good-affinity and high-affinity peptides. Upon the 450 peptides: 58 of the them have low or no affinity for the target protein with free energy of

binding larger than -5 kcal/mol, 228 peptides have medium-affinity with energy value larger than -7 kcal/mol, 142 peptides have good-affinity with energy values larger than -8.5 kcal/mol, and 22 peptides with high-affinity with energy values more negative than -8.5 kcal/mol. For the first 3 generations no high-affinity peptides are observed. The 4th and 5th generations only created 2 peptides with high-affinity. The 6th, 7th and 8th generation has 4 high-affinity peptides. The last generation has 6 high-affinity and 25 good-affinity peptides. GA leads to formation of peptides with more affinity towards hGH at each generation. The average fitness values tend to decrease from generation to generation. The decline in fitness values indicates that peptides with much more binding affinity to the selected protein surface are determined at every generation. The top peptides of the last 3 generations all have high-affinity values. The top peptides sequences of last 3 generations and the corresponding free energy of binding in kcal /mol, which is the fitness value of GA, are given in Table 7.1.

Table 7.1. Binding energy values of the top peptides determined by GA.

Generation Number	**Peptide Sequence**	**Fast Docking Results Binding Energy (kcal/mol)**
7	Trp Trp Ala Ile Phe	-8.93
	Trp Trp Asn Ile Phe	-9.06
	Trp Trp Met Ile Phe	-8.54
8	Trp Trp Tyr Ile Tyr	-9.09
	Trp Phe Phe Ile Trp	-8.69
9	Trp Phe Phe Ile Trp	-9.44
	Trp Trp Ala His Phe	-8.98

GA conserved specific residue types at particular positions: *Trp* is observed at the 1st position and *Phe / Trp / Tyr* are observed at the 5th position as can as given in Table 7.1. Similarly, the Viterbi peptide also has *Trp* for the 1st residue and *Tyr* for the 5th residue. Those repeated and conserved residues may play role in specific binding to the selected protein surface. The significance of the program is that no manipulation during the runs and no priori information are needed to determine a potential binder peptide. The physical and chemical properties of amino acids are also considered by GA, without apriori information. For the 1st position, hydrophobic *Trp* residue with aromatic ring is observed, the residue is able to stack itself among aromatic rings and it is also able to make H-bond due to its –NH group. For the 2nd position, GA tried only hydrophobic residues with

aromatic rings: *Trp* and *Phe*. For the 3rd position, *Ala, Phe, Tyr, Asn*, and *Met* is conserved. The 3rd position is the variable among all positions, it may require hydrophobic or uncharged hydrophilic residues. For the 4th position mostly Ile is conserved, but also *His* is set for this position with a low frequency. The *Ile* amino acid is hydrophobic with long, flexible side-chain. *His* residue, which is charged and polar, is also conserved. For the 5th position hydrophobic residues with aromatic rings are conserved: *Trp, Phe,* and *Tyr*.

The conservation of residues at particular positions lead to a motif for penta-peptide of hGH: **Trp – (Phe/Trp)–(Ala/ Asn/ Met /Phe/Tyr) – (His/Ile) – (Phe /Trp/Tyr)**. The 1st position is fixed to Trp and the 3rd position is the variable part of the motif. In other way to represent a motif is: *aromatic side-chain – aromatic and hydrophobic side-chain – variable – hydrophobic, long side-chain or positively charged side-chain – aromatic side-chain*. There exist 1x2x5x2x3 = 60 possible peptides that can be built up from this motif. In order to determine the best peptide among 60 possibilities AutoDock program is used. The binding affinity in terms of kcal/mol is calculated for each of 60 peptides, the results for good- and high-affinity peptides are given in Table 7.2. The interaction energies between hGH and the motif peptides are given in Table 7.2. All of them bind to the pre-determined binding site.

After the docking of 60 peptides, the top 7 peptides are clustered into 2 groups, as indicated in Table 7.2. The first group has average binding affinity of -10.26 kcal/mol, while the second group has average binding affinity of -8.9 kcal/mol. The first group has the motif **Trp – (Phe/Trp) – (Phe/Tyr) – (His/Ile) – (Trp/Tyr)**. The second group has the motif **Trp – (Phe/Trp)– (Ala/Tyr) – (His/Ile) –(Phe/Tyr)**.

Figure 7.1 indicates the bound *Trp Phe Phe His Trp* peptide on hGH surface. The peptide makes hydrogen bond with protein residues Ser62, Asp17, and Thr175. There exists a $\pi - \pi$ interaction between the residues: Trp5 – Phe25. There is an intramolecular $\pi - \pi$ interaction of peptide formed by the residue pairs Trp1 – Phe2. There exists an intramolecular π -cation interaction of peptide formed by the residue pairs Phe3- His4. As can be seen from the figure, there is almost a perfect geometric match between the peptide and the surface. Especially Trp1 and Phe2 residues match exactly to a cavity on the binding surface.

Figure 7.2 indicates the bound *Trp Phe Phe Ile Trp* peptide on hGH surface. The peptide makes hydrogen bond with protein residues Gln22, Ser62, and Asp171. There

exists a π -cation interaction between the residues: Trp1 – Lys172. There is an intramolecular $\pi - \pi$ interaction of peptide formed by the residue pairs Trp1 – Phe2. A π -sigma interaction is observed for residues Trp5 – Phe25.

Figure 7.3 indicates the bound Trp Trp Tyr Ile Tyr peptide on hGH surface. The peptide makes hydrogen bond with protein residues Gln22, and Asp171. There exists a $\pi - \pi$ interaction between the residues: Tyr5 – Phe25. Two intermolecular π -sigma interactions are observed for residues Trp1 – Phe25, and Tyr5 – His21. The binding orientation of this peptide differs from the *Trp Phe Phe His Trp* and *Trp Phe Phe Ile Trp* peptide conformations on the protein surface.

Table 7.2. Binding energy and K_i values of the motif peptides.

Peptide Sequence	**Detailed Docking Results Binding Energy (kcal/mol)**	**Inhibition Constants K_i**
Trp Phe Phe His Trp	-10.31	27.88 nM
Trp Phe Phe Ile Trp	-10.24	31.30 nM
Trp Trp Tyr Ile Tyr	-10.23	31.91 nM
Trp Phe Ala His Tyr	-8.94	279.62 nM
Trp Trp Tyr His Tyr	-8.94	279.62 nM
Trp Phe Tyr His Tyr	-8.93	286.24 nM
Trp Trp Ala Ile Phe	-8.93	282.88 nM

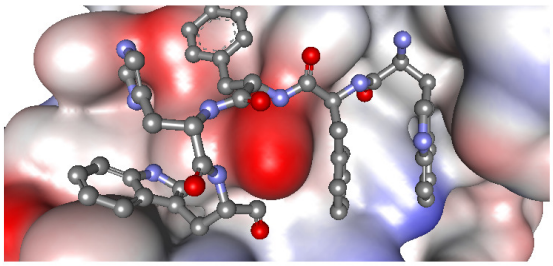


Figure 7.1. *Trp Phe Phe His Trp* peptide bound to hGH surface. The protein is represented as surface; the peptide is represented as ball-and-stick.

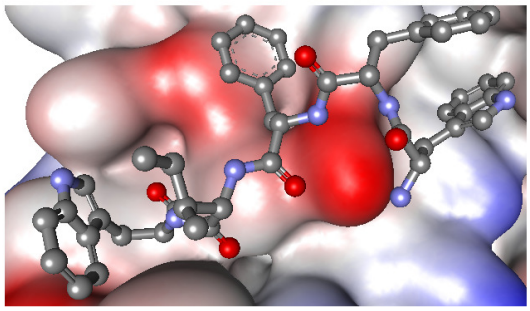


Figure 7.2. *Trp Phe Phe Ile Trp* peptide bound to hGH surface. The protein is represented as surface; the peptide is represented as ball-and-stick.

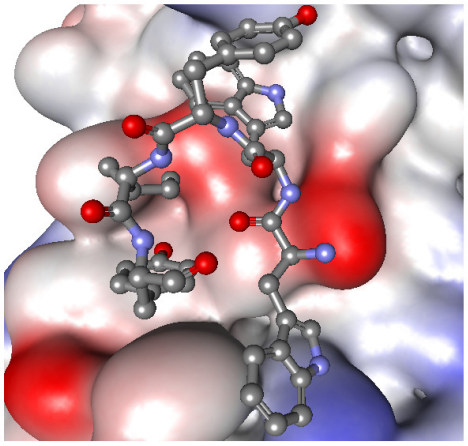


Figure 7.3. *Trp Trp Tyr Ile Tyr* peptide bound to hGH surface. The protein is represented as surface; the peptide is represented as ball-and-stick.

Although no conserved peptide motif has been discovered for hGH binding, there have been successful attempts to determine the binder residues of the protein. The crucial residues of hGHR that play role in hGH binding are studied by many different groups. Baumgartner *et al.* [153] determined that the Ser226 residue of hGHR is very essential in hGH binding, the residue interacts with the Glu174 residue of hGH. Baumgartner indicates that Tyr222 residue of hGHR is surrounded by a pocket formed with Glu175, Pro191 and Arg215 of hGH. The author offers Trp-X-X-X-Ser motif for hGH according to experimental results. The Viterbi algorithm determined Ser residue for the 4th position of the peptide and furthermore the Glu174 of hGH is found in neighboring residues of the designed peptide. The algorithm sets Tyr residue for the 5th position, and Glu175, Pro191 and Arg215 residues oh hGH is available in the binding site. Cosman et al. [168] and

Bazan et al. [169] acknowledged that a typical Trp-Ser-X-Trp-Ser motif is observed in hematopoietic superfamily receptors, of which hGHR is a member of.

The alanine scanning studies of Wells [157] on the hGH-hGHR complex indicates that a small set of amino acids are necessary for the two proteins to interact: Trp104 and Trp169 residues of hGHR pack against the Asp171, Lys172, and Thr175 of hGH. As stated before, those hGH residues are found in our binding region and also 2 *Trp* residues are found both in Viterbi and GA designed peptides. Wells indicate that the interaction site for hGHR-hGH complex is large while only a few hydrophobic residues are essential in binding.

Experimental design is necessary to prove interaction between the offered peptide sequences and hGH. **Trp-X-X-Ser** and **(Phe/Trp)-X-Ile-(Trp/Tyr)** motifs can be offered for hGH as an outcome of our algorithms. The model can be applied to any given protein target for design of any peptide length.

7.2.2. Erythropoietin Peptide Motifs

Erythropoietin (EPO) is a glycoprotein hormone produced mainly in mammalian kidney and liver. EPO regulates the erythropoiesis and the red cell proliferation and differentiation [170].It is not possible to purify EPO directly from blood serum; we aimed to design a selective peptide design to filter the protein. EPO interacts with the membrane bound erythropoietin receptor (EpoR) located on erythroblasts in the bone marrow [171]. EPOR, similar to hGHR, includes a 4 cysteines and **Trp Ser X Trp Ser** motif, which is presumed to be involved in protein-protein interaction. Human EPO was first purified in 1977, from anemia patients. Accordingly, human EPO is the major therapeutic agent to treat anemia of chronic renal failure and other diseases [172]. There exist two binding sites on EPO: the high-affinity receptor binding site and the low affinity receptor binding site. The high-affinity receptor binding site involves residues Asn 147, Arg 150, Gly 151, Thr 41, Lys 52. The Trp 51 and Phe 48 interact with the side chain of Leu 155, suggesting that the residues play a structural rather than functional role in receptor binding. The low affinity receptor binding site comprises the residues Val 11, Arg 14, Tyr 15, Ser 100, Arg 103, Ser 104 and Leu 108 [173].

The binding site of EPO is determined with GNM as indicated in Figure 7.4. The binding site center is set as Leu155. From literature the residues Phe142, Tyr145, Phe148 and Tyr156 are known to be essential for receptor binding [173]. Gly151 and Val46 pack against the methyl groups of Leu155 forming a shallow hydrophobic core [173].

The VitAL determined *Trp Glu Met Met Met Met Met* peptide for EPO protein, the peptide affinity is -4.72 kcal/mol. VitSTR determined *Glu Glu Glu Arg Glu Glu Val* peptide with -10.49 kcal/mol binding energy for EPO (Figure 7.5).

Glu Glu Glu Arg Glu Glu Val forms 5 hydrogen bonds with Phe48, Lys154, Thr163, and Arg166. Salt-bridges are formed between *Glu* residues of peptide and Arg156, Lys154, Arg150.

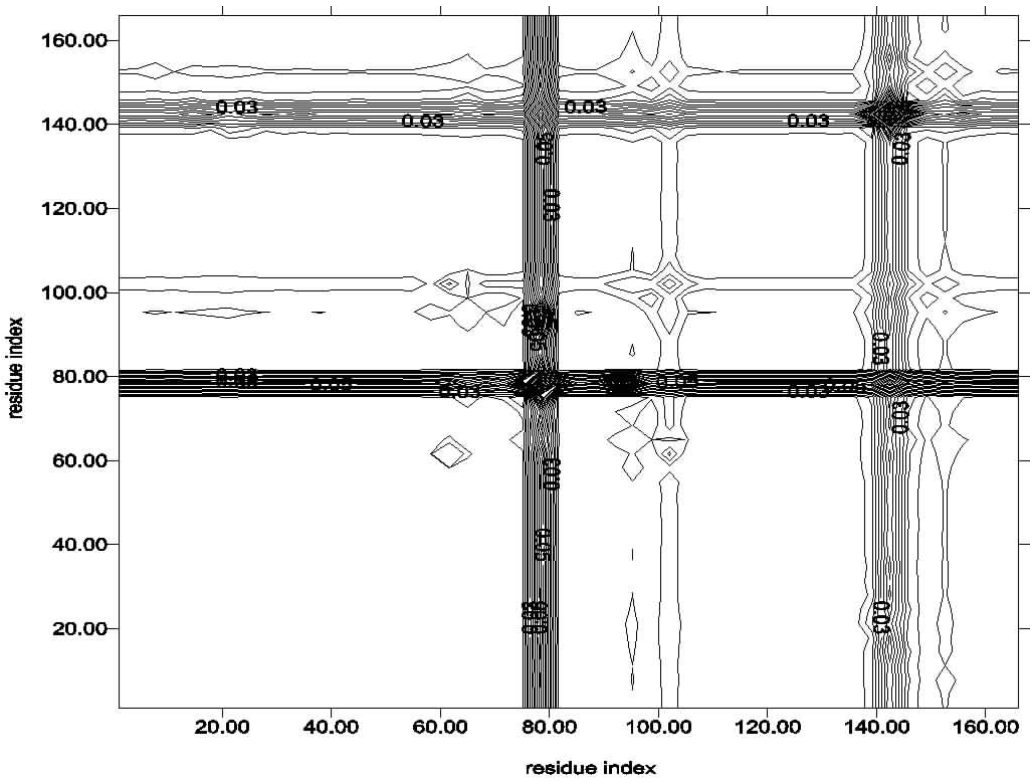


Figure 7.4. The hot-spot residues of EPO protein determined by GNM.

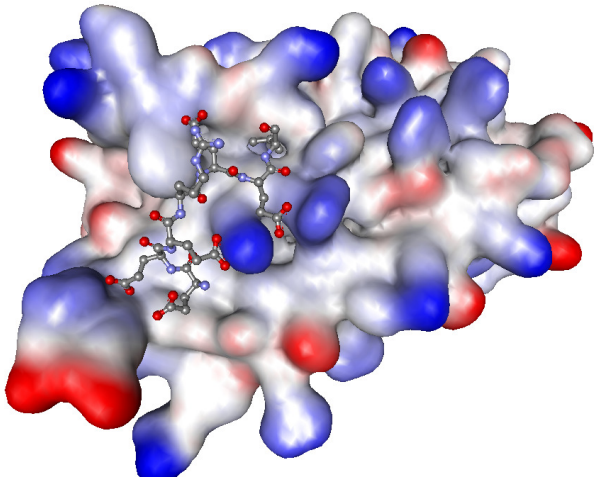


Figure 7.5. *Glu Glu Glu Arg Glu Glu Val* and EPO complex.

The peptides designed by VitAL, VitSTR and 48 random peptides are set as the initial population of GA. After 10 generations, the algorithm determined candidate peptide sequences. The top peptides of the last 4 generations and the corresponding free energy of binding in kcal /mol, which is the fitness value of GA, are given in Table 7.3.

Table 7.3. Binding energy values of the top peptides determined by GA.

Generation Number	Peptide Sequence	Fast Docking Results Binding Energy (kcal/mol)
7	Asn Pro Thr Ala Tyr Phe Ser	-6.5
8	Asp Pro Thr Ser Trp Ala Tyr	-6.24
	Asp Pro Ser Tyr His Ala Val	-6.1
9	Asp Pro Val Ser Trp Ala Tyr	-6.83
	Asp Pro Thr Ala Trp Asp Tyr	-6.3
	Asp Pro Thr Phe Trp Asp Tyr	-6.07
10	Asp Pro Trp Ala Trp Asp Phe	-7.03
	Asp Pro Val Ser Trp His Ser	-6.07

GA conserved specific residue types at particular positions: For the 1st position, negatively charged *Asp* residue with aromatic ring is observed. For the 2nd position the helix-breaker *Pro* residue is conserved as given in Table 7.3. Those repeated and conserved residues may play role in specific binding to the selected protein surface. The significance of the program is that no manipulation during the runs and no priori information are needed to determine a potential binder peptide.

The conservation of residues at particular positions lead to a motif for hepta-peptide of EPO: **Asp–Pro-(Thr/Val/Ser)–(Ala/Ser/Tyr/Phe)–(Tyr/Phe/His)–(Phe/Asp/His/Ala)–**

(Ser/Val/Tyr/Phe). The 1st and the 2nd positions are fixed to Trp and the remaining positions are variable. In other way to represent a motif is: *negative charge – helix breaker – hydrophobic – hydrophobic or aromatic side-chain*. There exist 576 possible peptides that can be built up from this motif. In order to determine the best peptide among 576 possibilities AutoDock program is used. The binding affinity in terms of kcal/mol is calculated for each of all possible peptides and the best peptide sequence is determined to be *Asp Pro Val Tyr Trp His Val* with -9.98 kcal/mol binding energy. The peptide-protein complex is given in Figure 7.6.

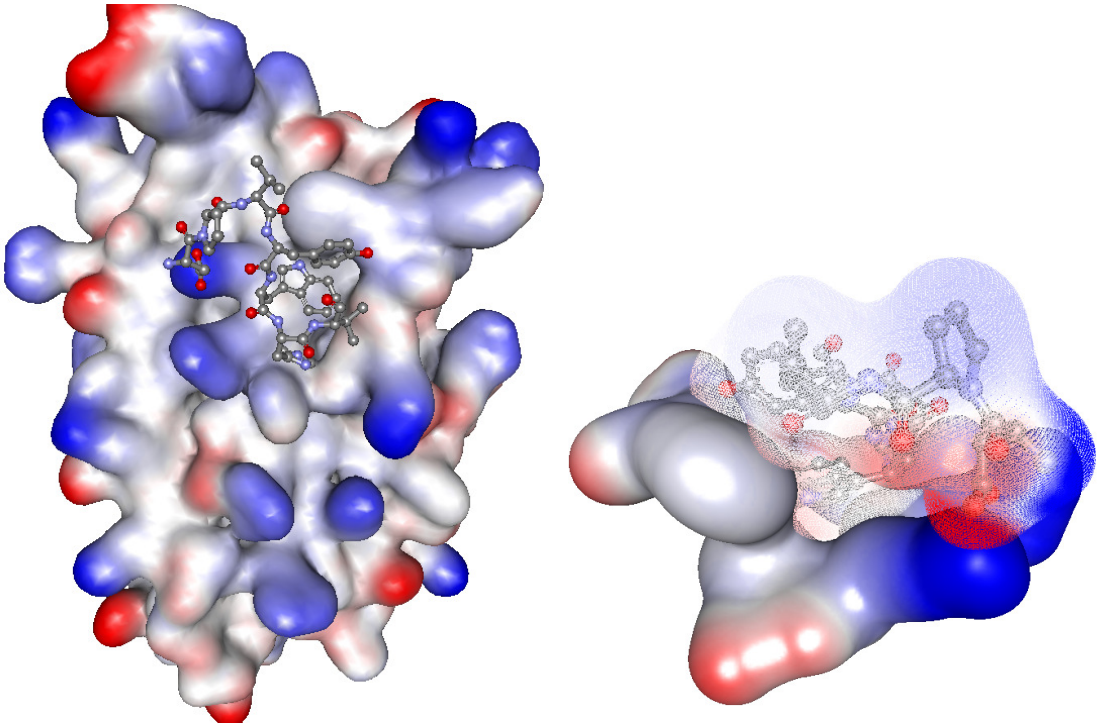


Figure 7.6. *Asp Pro Val Tyr Trp His Val* and EPO complex.

Asp Pro Val Tyr Trp His Val peptide makes specific interactions with EPO: 3 hydrogen bonds with Phe48 and Lys154; π - sigma interaction is observed for *Trp* and Gly151; salt-bridge forms between *Asp* and Lys154; $\pi - \pi$ stacking is available for *Tyr* and Phe48.

7.3. Concluding Remarks

A combined probabilistic approach is offered to define a potential binder peptide motif for a given protein surface. The method is constructed from two parts. The first part is

based on a Hidden Markov model using Viterbi decoding and the second part utilizes genetic algorithm. Initially, a binding site on the selected target protein is determined via the coarse grained Gaussian Network Model. One or more counterpart proteins are determined via employing the HotPoint server if any complex structure is available on Protein Data Bank. The counterpart protein fraction(s) interacting with the pre-determined binding site is analyzed. The interacting part of the counterpart protein is defined as the binding path. The best fitting peptide is constructed by generating all possible peptide pairs at each point along the path and determining the binding energies between these pairs and the specific location on the protein using AutoDock program. The Markov based partition function for all possible choices of the peptides along the path is generated by a matrix multiplication scheme. The best fitting peptide for the given surface is obtained by a Hidden Markov model using Viterbi decoding. The suitability of the conformations of the peptides that result upon binding on the surface are included in the algorithm by considering the pre-determined structure of counterpart protein. The structure consideration is based on the intrinsic Ramachandran potentials. Genetic algorithm with tournament selection is applied to determine motifs for the target protein. The best fitting peptide sequence determined by Viterbi decoding and randomly generated peptide sequences are set as the chromosomes of genetic algorithm. Initial random population leads to determination of 50 possible binding peptides, after 7-9 generations of GA. The conservation of specific amino acids at particular locations leads to formation of binder motifs for the target protein.

Chapter 8

CONCLUSION

The studies in this dissertation shed light to the peptide properties and the peptide design problem. The statistical thermodynamics features of peptides are formulated using knowledge based data relevant to the denatured states of proteins. The method is used to determine the energy of a peptide to gain information about the relative magnitudes of the energy of the single conformation that is obtained upon binding and the average configurational energy of the peptide in solution. Knowledge of the entropy of a peptide can be calculated to understand the amount of entropy that will be lost when the sequence binds to the surface. The conformational heat capacity of a free peptide is also determined to comprehend the stability of the peptide in solution.

The peptide design algorithms provide insights for researchers working on protein purification and binder peptide problems. Three different models are offered to predict inhibitor peptide motifs for any given protein. The methods are global and the performance of predictions is encouraging. The genetic algorithm uses the tournament selection, one-point crossover and mutation operations are implemented. The algorithm converges to certain amino acids for specific positions. Moreover, the method is able to determine better affinity peptides at each generation. The VitAL and VitSTR methods employ Viterbi decoding for HMM, where the hidden states are amino acids and the observables are the torsional angles. The models are able to predict a potential binder peptide sequence. The prediction performance could be improved further by additional informative attributes such as the chemical and physical features of the docked amino acids. The amino acids and their property conservation at specific positions is a promising result. Experimental design is necessary to prove interaction between the offered peptide sequences and the target proteins. The importance of the programs is that no manipulation during the runs or no priori information is needed to determine a good candidate peptide.

The VitAL server would be useful both for the experimentalists and computational scientist working on peptide design problem. The VitAL method is tested on known protein-peptide inhibitor complexes. The detailed analysis of the designed peptide interactions and comparison of those peptides with the known peptides indicate that our

method is able to detect the requirements of binding surface: such as hydrophilicity, electrostatic compatibility, aromatic interactions, sulfur atom and its interactions.

As a future work, improvements for the emission and transition probabilities for Viterbi models are crucial. The genetic algorithm can be improved further via addition of some filters according to the target protein information or formation of non-random initial population. Those filters may also lessen the time of prediction.

The thermodynamics of peptides, *de novo* peptide design and peptide motif determination problem are studied in this dissertation. These studies here will serve to drug design and protein purification.

APPENDIX

A.1. NF-κB peptides designed by Markov Model

Table A.1.1. The top 100 peptides designed by Markov Model. The 1st column indicates the peptide sequence, the 2nd column indicates the AutoDock binding energy of the designed peptide, the 3rd column shows the inhibition constant calculated by AutoDock, and the last column indicates the probability order of the peptide – out of 100. The higher the probability value, the more possible it is that the peptide will bind to target protein according to the Markov model.

Peptide Sequence	AutoDock Binding Energy (kcal/mol)	K_i	Probability Order
FSWFWWY	-11,25	5,71 nM	4
KFWFFWY	-11,11	7,16 nM	95
FKFEFWY	-10,37	25,14 nM	44
FKHKFWY	-10,22	32,17 nM	2
RHTWFWY	-10,13	37,78 nM	28
RHPNFWY	-9,96	49,90 nM	63
FSNFFWY	-9,9	55,72 nM	15
FAQKFWY	-9,83	62,78 nM	79
RSWFWWY	-9,74	73.09 nM	29
FKFFFWY	-9,71	76,32 nM	83
FSQFWWY	-9,67	81,72 nM	49
FKHSWWY	-9,37	135,27 nM	93
FGNFFWY	-9,34	141,91 nM	9
FRVWFWY	-9,25	165,55 nM	53
FKHCFWY	-9,24	167,83 nM	46
FQPKFWY	-9,1	213,51 nM	37
FKFKFWY	-9,07	226,33 nM	3
FRNFWWY	-9,03	239,49 nM	26
FSWFFWY	-9,01	247,73 nM	5
FKHKFWF	-8,9	268,16 nM	33
FAWFFWY	-8,87	314,57 nM	84
FANFFWY	-8,84	332,45 nM	30
FGNWFWY	-8,84	329.84 nM	80

FCNFWWY	-8,81	348,41 nM	97
RHPLFWY	-8,78	366,52 nM	76
FSPKFWY	-8,76	382,05 nM	36
FSWKFWY	-8,75	386,45 nM	62
KFWFWWY	-8,59	504,49 nM	77
RHPSFWY	-8,54	547,86 nM	75
HANFWWY	-8,53	559,60 nM	16
FKFFWWY	-8,41	685,98 nM	74
HAEWFWY	-8,4	696.41 nM	71
FGNFWWY	-8,37	735,51 nM	6
RSWFFWY	-8,3	829,60 nM	42
FSWFWHW	-8,26	889,07 nM	72
RHPCWWY	-8,2	981,01 nM	51
HAVKFWY	-8,19	999,39 nM	38
FSQWFWY	-8,12	1,11 uM	23
FRLWFWY	-8,09	1,17 uM	13
FSFKFWY	-8,03	1,29 uM	66
FHTKFWY	-8,02	1,32 uM	59
FGLWFWY	-7,93	1,55 uM	32
FKHFWWY	-7,91	1,60 uM	8
FSWLFWY	-7,88	1,68 uM	100
FKHCWWY	-7,84	1,80 uM	12
FAVKFWY	-7,53	3,04 uM	60
FRHWFWY	-7,53	3,04 uM	88
FKFWFWY	-7,51	3.11 uM	25
FSQKFWY	-7,37	3,96 uM	14
HANFFWY	-7,36	4,02 uM	21
HAQWFWY	-7,29	4,05 uM	85
FSTKFWY	-7,22	5,07 uM	35
FGRWFWY	-7,15	5,70 uM	82
RHPKFWY	-7,15	5,74 uM	19
FRHKFWY	-7,12	6 uM	86
FKHRFWY	-7,11	6,17 uM	18

HAWFFWY	-7,11	6,13 uM	65
FKFKFWF	-7,09	6,35 uM	34
FKHQWWY	-7,07	6,57 uM	47
FKHWFWF	-6,88	9,10 uM	31
FGRLFWY	-6,85	9,53 uM	87
FFWFFWY	-6,8	10,35 uM	81
FKHEFWY	-6,76	11,17 uM	67
FGFKFWY	-6,66	13,10 uM	61
FSQFFWY	-6,32	23,34 uM	58
RNRLFWY	-6,25	26,18 uM	68
RNFKFWY	-6,12	32,75 uM	22
RHTKFWY	-6,09	34,54 uM	7
FKPKFWY	-5,99	40,34 uM	52
FKFRFWY	-5,79	57,16 uM	91
FKFNFWY	-5,69	67,28 uM	89
HAQKFWY	-5,69	67,68 uM	56
FKHSFWY	-5,63	74,28 uM	41
FKHLFWY	-5,59	80,33 uM	96
FRVKFWY	-5,57	83,19 uM	17
RHWFFWY	-5,34	120,96 uM	40
RHVKFWY	-5,15	168,68 uM	20
FKHIFWY	-4,76	322,51 uM	98
FNFKFWY	-4,72	344,96 uM	55
FSNFWWY	-4,58	441,44 uM	11
FKHWFWY	-4,37	630,09 uM	1
RHVWFWY	-4,26	757,19 uM	64
RHWFWWY	-3,23	4,28 mM	27
RNRWFWY	-3,1	5,33 mM	70
FKFQWWY	-2,9	7,54 mM	43
FFWFWWY	-2,71	10,40 mM	69
FSWFWFW	-2,47	15,35 mM	94
FANFWWY	-2,46	15,61 mM	24
FKHFFWY	-2,45	15,88 mM	10

FSWWFWY	-2,36	18,77 mM	45
FRNFFWY	-2,27	21,60 mM	39
HAWFWWY	-2,27	21,57 mM	48
FAWFWWY	-2,01	33,37 mM	73
FKFHFWY	-2	34,20 mM	90
FSWFFWF	-2	34,44 mM	57
FSWFWWF	-1,93	38,39 mM	50
RNRFWWY	-1,6	67,72 mM	99
FSWFWFR	-1,55	72,69 mM	92
FSQRFWY	0,02	None	54
FGNKFWY	0,57	None	78

A.2. VitAL Benchmark

The Ramachandran map is divided into 11-states as described in Chapter 3. The emission probabilities for dipeptides are calculated according to the probabilities from coil library as defined in Chapter 5. The emission probability values (according to the coil library) for the torsional angles out of the 11-states are small. When the dipeptides are docked to target protein surface by VitAL or VitSTR methods, their torsional angles are mostly observed to be out of the 11-states. Consequently, for suc cases the effect of emission probability is very small compared to the effect of transition probability. When the transition probability dominates, the *Phe* / *Trp* / *Tyr* residues are favored in the designed peptide due to the fact that their AutoDock binding energy is high compared to other amino acid types. The force-filed other than AutoDock can be utilized to overcome the large residue conservation through all peptides.

Since the dipeptide torsional angles are distributed in Ramachandran map, it is necessary to define new states on the map. While implementing the VitAL as a web-server a new peptide library, PEPX [167], is utilized. The usage of the new library is important because the torsional angles are distributed all over the Ramachandran map, not only to the 11-states. The distribution is shown in Table A.2.1. As can be seen from the table 13% of amino acids are observed out of 11-states, and 23% of dipeptides are observed out of 11-states.

Table A.2.1. The observation of 11-states and the other states fot the amino acids (2nd row) and the dipeptides (3rd row) of PEPX library.

States	1	2	3	4	5	6	7	8	9	10	11	Other
Amino Acid Frequeies	185	1563	2659	185	1472	704	161	229	1016	1875	1805	1783
Dipeptide Frequencies	395	1537	2585	357	1262	592	218	399	943	1712	1661	3394

The Ramachandran map is futher divided into 21-states as shown in the Figure A.2.1. The PEPX database is utilized for the new state definitions.

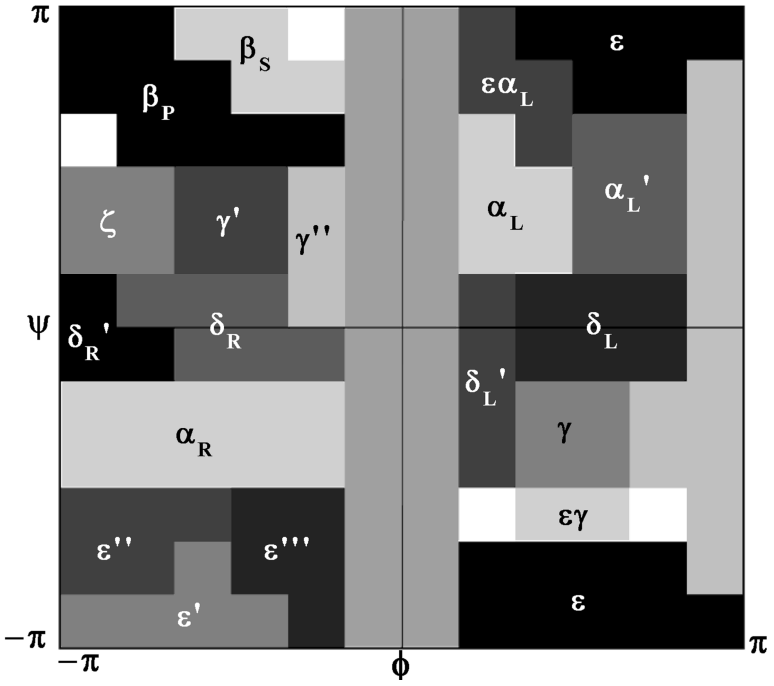


Figure A.2.1.21-states Ramachandran map.

Table A.2.2. Definitions of 21 states.

State	Notation	Definition
1	ε’	Mirror image of the extended region ε.
2	ε	The extended regions, φ >0, ψ ~ -+180°.
3	α_R	Right-handed alpha helix.
4	γ	Tight turn region.
5	δ_R	The right handed bridge region between two β strands.
6	δ_L	Mirror image of the δ_R region.
7	ζ	Region observed mostly in residues preceding PRO.
8	γ’	Inverse tight turn region.
9	α_L	Mirror image of α_R.
10	β_S	Extended beta sheet forming region.
11	β_P	Region with extended polyproline-like helices.
12	δ_R ‘	(-180<=phi<=-120 and -30<=psi<=0) or (-180<=phi<=-150 and 0<=psi<=30)
13	ε’’	(-180<=phi<=-120 and -150<=psi<=-90) or (-120<=phi<=-90 and -120<=psi<=-90)
14	ε’’’	(-90<=phi<=-30 and -150<=psi<=-90) or (-60<=phi<=-30 and -180<=psi<=-150)
15	γ’’	(-60<=phi<=-30 and 0<=psi<=90)
16		(-30<=phi<=30 and -180<=psi<=180)
17	$\varepsilon\alpha_L$	(30<=phi<=60 and 120<=psi<=180) or (60<=phi<=90 and 90<=psi<=150)
18	α_L’	(90<=phi<=150 and 30<=psi<=120)
19		(180<=phi<=150 and -150<=psi<=150) or (120<=phi<=150 and -90<=psi<=-30)
20	δ_L’	(30<=phi<=60 and -90<=psi<=30)
21	Eγ	(20<=phi<=60 and -120<=psi<=-90)

The server is tested on pentapeptides from PEPX database [167]. The PDB accession codes, literature peptide sequences, VitAL server designed sequences and the AutoDock binding energy of designed peptides are given below in Table A.2.3. The conservation of some amino acids and the similar amino acids in VitAL peptides is promising. The location of some amino acids is altered in VitAL peptides when compared to the literature peptides. This is due to the fact that the grid box for each amino acid is not constrained, but it is relaxed to **2.5xamio acid size**. The binding site for the peptide may be shifted 10-15 Angstrom according to the reference size. The most promising results are observed for 1BHX, 1PIP, 1PYO, and 1R9N.

Table A.2.3. The VitAL benchmark

PDB ID	Literature Peptide	VitAL Peptide	Binding Energy of VitAL Peptide (kcal/mol)
1AB9	T P G V Y	W F **Y** M C	-6,67
1BC5	N W E T F	F F P Q H	-4,88
1BE9	K Q T S V	W W V V **V**	-7,65
1BHX	D F E E I	W **E** V **D E**	-6,84
1EE4	K R V K L	W P **L V** A	-5,75
1ELR	M E E V D	Y Y M I E	-4,4
1EVH	F P P P P	W W L L I	-6,19
1FCH	Y Q S K L	W W **Q** C A	-5,85
1GMD	P G A Y D	W W W H P	-8,71
1HTM	T L C L G	W Y T V S	-5,6
1IHJ	T E F C A	W E A A **A**	-4,81
1JQ9	F L S Y K	W Y **Y** G G	-7
1JUQ	D D H L L	W W S D V	-3,75
1JWY	T R G L V	W Y E A E	-5,6
1KL5	S H P Q F	W **P** E C D	-9,16
1KYD	S D P W K	W Q H H S	-6,37
1KZP	K C V I M	W P H **C** C	-8,67
1MF4	V A H R S	W W **R** L I	-4,95
1NRR	K Y E P F	W E H **P** V	-5,3
1NVQ	A S V S A	W Y H H H	-5,39
1NX0	A K A I A	W W V **A A**	-6,26
1OBY	N E F Y A	W F D I L	0,58
1OKV	R R L I F	W W V C R	-6,85

1OYT	F E E I P	W Y A H **P**	-7,09
1PIP	Q V V A A	W Y **Q V V**	-5,12
1PYO	L D E S D	W P **E S** G	-6,83
1R9N	Y P S K P	W F **K** M **P**	-5,8
1SDX	L E Q C A	W **E** Y D D	-4,37
1TG4	F L A Y K	W P C N E	-5,84
1TN8	G C V L S	W W **L** Y A	

BIBLIOGRAPHY

1. Kitchen, D.B., et al., *Docking and scoring in virtual screening for drug discovery: methods and applications.* Nature Reviews Drug Discovery, 2004. 3: p. 935-949.
2. Dill, K.A. and D. Shortle, *Denatured States of Proteins.* Annual Review of Biochemistry, 1991. 60: p. 795-825
3. Brant, D.A. and P.J. Flory, *The Configuration of Random Polypeptide Chains. II. Theory.* J. Am. Chem. Soc., 1965. 87(13): p. 2791–2800.
4. Brant, D.A. and P.J. Flory, *The Role of Dipole Interactions in Determining Polypeptide Configurations.* J. Am. Chem. Soc., 1965. 87(3): p. 663–664.
5. Conrad, J.C. and P.J. Flory, *Moments and distribution functions for polypeptide chains. Poly-L-alanine.* Macromolecules, 1976. 9(1): p. 41-47.
6. Flory, P.J., *Statistical Mechanics of Chain Molecules*. 1969, New York: Wiley.
7. Gohlke, H. and G. Klebe, *Approaches to the Description and Prediction of the Binding Affinity of Small-Molecule Ligands to Macromolecular Receptors.* Angew. Chem. Int., 2002. 41: p. 2644-2676.
8. Mattice, W.L. and U.W. Suter, *Conformational theory of large molecules*. 1994, New York, 1994: Wiley Interscience.
9. Rognan, D., et al., *Predicting binding affinities of protein ligands from three-dimensional models: application to peptide binding to class I major histocompatibility proteins.* J Med Chem., 1999. 42(22): p. 4650-4658.
10. Rognan, D., et al., *Molecular dynamics simulation of MHC-peptide complexes as a tool for predicting potential T cell epitopes.* Biochemistry, 1994. 33(38): p. 11476–11485.
11. Rognan, D., et al., *Modeling the interactions of a peptide-major histocompatibility class I ligand with its receptors. I. Recognition by two alpha beta T cell receptors.* J Comput Aided Mol Des., 2000. 14(1): p. 53-69.
12. Rognan, D., et al., *Molecular dynamics study of a complex between the human histocompatibility antigen HLA-A2 and the IMP58-66 nonapeptide from influenza virus matrix protein.* Eur J Biochem., 1992. 208(1): p. 101-113.
13. Springer, C., et al., *PostDOCK: A Structural, Empirical Approach to Scoring Protein Ligand Complexes.* J. Med. Chem, 2005. 48(22): p. 6821–6831.
14. Amzel, L.M., *Loss of Translational Entropy in Binding, Folding, and Catalysis.* PROTEINS: Structure, Function, and Genetics, 1997. 28: p. 144-9.
15. Chaires, J.B., *Calorimetry andThermodynamics.* Drug Design Annu. Rev. Biophys., 2008. 37: p. 135-51.
16. Janin, J., *The Kinetics of Protein-Protein Recognition.* PROTEINS: Structure, Function, and Genetics, 1997. 28: p. 153-61.
17. Finkelstein, A.V. and J. Janin, *The price of lost freedom: Entropy of bimolecular complex formation.* Protein Eng., 1989. 3: p. 1-3
18. Gohlke, H. and D.A. Case, *Converging Free Energy Estimates: MM-PB(GB)SA Studies on the Protein-Protein Complex Ras-Raf.* J Comput Chem, 2004. 25: p. 238-50.
19. Boniface, J., et al., *Thermodynamics of T cell receptor binding to peptide–MHC: Evidence for a general mechanism of molecular scanning.* Proc. Natl. Acad. Sci. USA, Immunology, 1999. 96: p. 11446-51.
20. Grunberg, R., M. Nilges, and J. Leckner, *Flexibility and Conformational Entropy in Protein-Protein Binding.* Structure, 2006. 14: p. 683-93.
21. Doig, A.J. and M.J.E. Sternberg, *Side-chain conformational entropy in protein-folding.* Protein Scİ., 1995. 4: p. 2247-51.
22. Brady, P. and K.A. Sharp, *Entropy in protein folding and in protein-protein interactions.* Current Opinion in Structural Biology, 1997. 7: p. 215-21.

23. Gilson, M.K., et al., *The Statistical-Thermodynamic Basis for Computation of Binding Affinities: A Critical Review.* Biophysical Journal, 1997. 72: p. 1047-69.
24. Karplus, M., T. Ichiye, and B.M. Petite, *Configurational entropy of native proteins.* Biophys J, 1987. 52: p. 1083-5.
25. Lim, N.S., et al., *Comparative Peptide Binding Studies of the PABC Domains from the Ubiquitin-protein Isopeptide Ligase HYD and Poly(A)-binding Protein.* The Journal of Biological Chemistry, 2006. 281: p. 14376-82.
26. Frederick, K.K., et al., *Conformational entropy in molecular recognition by proteins.* Nature, 2007. 448: p. 325-30.
27. Novotny, J., et al., *Empirical Free Energy Calculations: A Blind Test and Further Improvements to the Method.* J. Mol. Biol., 1997. 268: p. 401-11.
28. Habermann, S.M. and K.P. Murphy, *Energetics of hydrogen bonding in proteins: A model compound study.* Protein Sci., 1996. 5: p. 1229-39.
29. Karplus, P.A., *Hydrophobicity regained.* Protein Sci., 1997. 6: p. 1302-7.
30. Reynolds, J.A., D.B. Gilbert, and C. Tanford, *Empirical correlation between hydrophobic free energy and aqueous cavity surface area.* Proc Nat Acad Sci. USA, 1974. 71: p. 2925-7.
31. Hermann, R.B., *Theory of hydrophobic bonding. 11. The correlation of hydrocarbon solubility in water with solvent cavity surface area.* J Phys Chem, 1972. 76: p. 2754-9.
32. Sleigh, S.H., et al., *Crystallographic and Calorimetric Analysis of Peptide Binding to OppA Protein.* J. Mol. Biol., 1999. 291: p. 393-415.
33. Kasper, P., P. Christen, and H. Gehring, *Empirical Calculation of the Relative Free Energies of Peptide Binding to the Molecular Chaperone DnaK.* PROTEINS: Structure, Function, and Genetics, 2000. 40: p. 185-92.
34. Kasper, P., P. Christen, and H. Gehring, *Empirical Calculation of the Relative Free Energies of Peptide Binding to the Molecular Chaperone DnaK.* PROTEINS: Structure, Function, and Genetics, 2000. 40: p. 185-192.
35. Klebe, G. and H.J. Bohm, *Energetic and Entropic Factors Determining Binding Affinity in Protein-Ligand Complexes.* Journal of Receptor and signal Transduction Research, 1997. 17: p. 459-73.
36. Hu, H., M. Elstner, and J. Hermans, *Comparison of a QM/MM force field and molecular mechanics force fields in simulations of alanine and glycine "dipeptides" (Ace-Ala-Nme and Ace-Gly-Nme) in water in relation to the problem of modeling the unfolded peptide backbone in solution.* Proteins, 2003. 50(3): p. 451-463.
37. Jha, A., et al., *Helix, Sheet, and Polyproline II Frequencies and Strong Nearest Neighbor Effects in a Restricted Coil Library.* Biochemistry, 2005. 44: p. 9691-9702.
38. Keskin, O., et al., *Relationships between amino acid sequence and backbone torsion angle preferences.* Proteins, 2004. 55(4): p. 992-998.
39. Munoz, V. and L. Serrano, *Intrinsic secondary structure propensities of the amino acids, using statistical ϕ - ψ matrices: Comparison with experimental scales.* Proteins: Structure, Function, and Bioinformatics., 1994. 20(4): p. 301-311.
40. Serrano, L., *Comparison between the phi distribution of the amino acids in the protein database and NMR data indicates that amino acids have various phi propensities in the random coil conformation.* J Mol Biol., 1995. 254(2): p. 322-333.
41. Petsalaki, E., et al., *Accurate Prediction of Peptide Binding Sites on Protein Surfaces.* Plos Computational Biology, 2009. 5.
42. Frenkel, D., et al., *Pro-Ligand - an Approach to De-Novo Molecular Design. Application to the Design of Peptides.* J. of Comp.-Aided Mol. Des., 1995. 9: p. 213-225.
43. Hetenyi, C. and D.v.d. Spoel, *Efficient docking of peptides to proteins without prior knowledge of the binding site.* Protein Science, 2002. 11: p. 1729-1737.

44. Morris, G.M., et al., *Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function.* J. of Comp. Chem., 1999. 19(14): p. 1639-1662.
45. Mayrose, I., et al., *Pepitope: epitope mapping from affinity-selected peptides.* Bioinformatics, 2007. 23(23): p. 3244-3246.
46. Moreau, V., et al., *PEPOP: Computational design of immunogenic peptides.* Bmc Bioinformatics, 2008. 9: p. -.
47. Stein, A. and P. Aloy, *A molecular interpretation of genetic interactions in yeast.* Febs Letters, 2008. 582(8): p. 1245-1250.
48. Juretic, D., et al., *Computational Design of Highly Selective Antimicrobial Peptides.* Journal of Chemical Information and Modeling, 2009. 49(12): p. 2873-2882.
49. Brusic, V., et al., *Prediction of MHC class II-binding peptides using an evolutionary algorithm and artificial neural network.* Bioinformatics, 1998. 14(2): p. 121-130.
50. Hammer, J., et al., *Precise Prediction of Major Histocompatibility Complex Class-Ii Peptide Interaction Based on Peptide Side-Chain Scanning.* Journal of Experimental Medicine, 1994. 180(6): p. 2353-2358.
51. Hammer, J., et al., *Peptide Binding-Specificity of Hla-Dr4 Molecules - Correlation with Rheumatoid-Arthritis Association.* Journal of Experimental Medicine, 1995. 181(5): p. 1847-1855.
52. Hanai, T., et al., *Computational Design of Proteinous Drug Employing Hidden Markov Model.* Genome Informatics, 2000. 11: p. 394-395.
53. Harrison, L.C., et al., *Major Histocompatibility Complex (Mhc) Molecules in Insulin-Dependent Diabetes.* Journal of Leukocyte Biology, 1993: p. 105-105.
54. Honeyman, M.C., et al., *Neural network-based prediction of candidate T-cell epitopes.* Nature Biotechnology, 1998. 16(10): p. 966-969.
55. Kato, R., et al., *Hidden Markov model-based approach as the first screening of binding peptides that interact with MHC class II molecules.* Enzyme and Microbial Technology, 2003. 33: p. 472-481.
56. Noguchi, H., et al., *Fuzzy neural network-based prediction of the motif for MHC class II binding peptides.* Journal of Bioscience and Bioengineering, 2001. 92(3): p. 227-231.
57. Noguchi, H., et al., *Hidden Markov Model-Based Prediction of Antigenic Peptides That Interact with MHC Class II Molecules.* Journal of Bioscience and Bioengineering, 2002. 94(3): p. 264-270.
58. Ajay, W.P. Walters, and M.A. Murcko, *Can we learn to distinguish between "drug-like" and "nondrug-like" molecules?* Journal of Medicinal Chemistry, 1998. 41(18): p. 3314-3324.
59. Abe, K., et al., *Peptide ligand screening of a-synuclein aggregation modulators by in silico panning.* BMC Bioinformatics, 2007. 8: p. 451.
60. Kamphausen, S., et al., *Genetic algorithm for the design of molecules with desired properties.* Journal of Computer-Aided Molecular Design, 2002. 16(8-9): p. 551-567.
61. Riester, D., et al., *Thrombin inhibitors identified by computer-assisted multiparameter design.* Proceedings of the National Academy of Sciences of the United States of America, 2005. 102(24): p. 8597-8602.
62. Belda, I., et al., *ENPDA: an evolutionary structure-based de novo peptide design algorithm.* Journal of Computer-Aided Molecular Design, 2005. 19: p. 585-601.
63. Klepeis, J.L., et al., *Design of peptide analogues with improves activity using a novel de novo protein design approach.* Industrial & Engineering Chemistry Research, 2004. 43(14): p. 3817-3826.
64. Singh, J., et al., *Application of Genetic Algorithms to Combinatorial Synthesis: A Computational Approach to Lead Identification and Lead Optimization.* J. Am. Chem. Soc., 1996. 118: p. 1669-1676.

65. Budin, N., et al., *An evolutionary approach for structure-based design of natural and nonnatural peptidic ligands.* Combinatorial Chemistry & High Throughput Screening, 2001. 4(8): p. 661-673.
66. Budin, N., N. Majeux, and A. Caflisch, *Fragment-based flexible ligand docking by evolutionary optimization.* Biological Chemistry, 2001. 382(9): p. 1365-1372.
67. Budin, N., et al., *Structure-based ligand design by a build-up approach and genetic algorithm search in conformational space.* Journal of Computational Chemistry, 2001. 22(16): p. 1956-1970.
68. London, N., D. Movshovitz-Attias, and O. Schueler-Furman, *TheStructural Basis of Peptide-Protein Binding Strategies.* Structure, 2010. 18: p. 188-199.
69. Lengauer, T. and M. Rarey, *Computational methods for biomolecular docking.* Curr Opin Struct Biol., 1996. 6(3): p. 402-406.
70. Jones, G., P. Willett, and R.C. Glen, *Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation.* J Mol Biol., 1995. 245(1): p. 43-53.
71. Holland, J.H., *Adaptation in natural and artificial systems*. 1992, Cambridge, MA: MIT Press.
72. Belda, I., et al., *ENPDA: an evolutionary structure-based de novo peptide design algorithm.* J Comput Aid Mol Des, 2005. 19: p. 585-601.
73. Kamphausen, S., et al., *Genetic algorithm for the design of molecules with desired properties.* J Comput Aid Mol Des, 2002. 16: p. 551-567.
74. Abe, K., et al., *Peptide ligand screening of a-synuclein aggregation modulators by in silico panning.* Bmc Bioinformatics, 2007. 8: p. 451.
75. Budin, N., et al., *An evolutionary approach for structure-based design of natural and nonnatural peptidic ligands.* Comb Chem High T Scr, 2001. 4(8): p. 661-673.
76. Budin, N., N. Majeux, and A. Caflisch, *Fragment-based flexible ligand docking by evolutionary optimization.* Biol Chem, 2001. 382(9): p. 1365-1372.
77. Budin, N., et al., *Structure-based ligand design by a build-up approach and genetic algorithm search in conformational space.* J Comput Chem, 2001. 22(16): p. 1956-1970.
78. Schneider, G., et al., *Peptide design by artificial neural networks and computer-based evolutionary search.* Proc. Natl. Acad. Sci. USA, Biochemistry, 1998. 95: p. 12179–12184.
79. Grinstead, C.M. and J.L. Snell, *Introduction to Probability*. 1997: American Mathematical Society.
80. Rabiner, L.R., *A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition.* Proceedings of the IEEE, 1989. 77(2): p. 257-286.
81. Forney, G.D.J. *The Viterbi Algorithm*. in *Proceedings of IEEE*. 1973.
82. Viterbi, A.J., *Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm.* IEEE Transactions on Information Theory, 1967. 13: p. 260-269.
83. Ewens, W.J. and G.R. Grant, *Statistical methods in bioinformatics: an introduction*. 2001, New York: Springer.
84. Bystroff, C. and A. Krogh, *Hidden Markov Models for Prediction of Protein Features.* Methods in Molecular Biology, 2008. 413: p. 173-198.
85. Sramek, R., B. Brejova, and T. Vinar. *On-line Viterbi Algorithm for Analysis of Long Biological Sequences*. in *7th International Workshop, WABI 2007,* . 2007. Philadelphia, PA, USA: Springer.
86. Mirabeau, O., et al., *Identification of novel peptide hormones in the human proteome by Hidden Markov model screening.* Genome Research, 2007. 17: p. 320-327.
87. Noguchi, H., et al., *Hidden Markov Model-Based Prediction of Antigenic Peptides That Interact with MHC Class II Molecules.* J Biosci Bioeng, 2002. 94(3): p. 264-270.
88. Fariselli, P., P.L. Martelli, and R. Casadio, *A new decoding algorithm for hidden Markov models improves the prediction of the topology of all-beta membrane proteins.* BMC Bioinformatics, 2005(6).

89. Noguchi, H., et al., *Fuzzy neural network-based prediction of the motif for MHC class II binding peptides.* J Biosci Bioeng, 2001. 92(3): p. 227-231.
90. Sonmez, K., et al., *Evolutionary Sequence Modeling for Discovery of Peptide Hormones.* Plos Computational Biology, 2009. 5.
91. Neduva, V. and R.B. Russell, *Linear motifs: Evolutionary interaction switches.* FEBS Letters, 2005. 579: p. 3342–3345.
92. Gould, C.M., et al., *ELM: the status of the 2010 eukaryotic linear motif resource.* Nucleic Acids Res, 2009. 38: p. D167-D180.
93. Hunt, T., *Protein sequence motifs involved in recognition and targeting: a new series.* Trends Biochem Sci, 1990. 15: p. 305.
94. Saito, H., et al., *The role of peptide motifs in the evolution of a protein network.* Nucleic Acids Research, 2007. 35(19): p. 6357–6366.
95. Tong, A.H.Y., et al., *A Combined Experimental and Computational Strategy to Define Protein Interaction Networks for Peptide Recognition Modules.* Science, 2002: p. 295:321.
96. Landgraf, C., et al., *Protein Interaction Networks by Proteome Peptide Scanning.* PLoS Biology, 2004. 2(1): p. 0094-0103.
97. Zhao, S. and E.Y.C. Lee, *A Protein Phosphatase-1-binding Motif Identified by the Panning of a Random Peptide Display Library.* J Biological Chemistry, 1997. 272(45): p. 28368–28372.
98. Corr, M., et al., *H-2Dd Exploits a Four Residue Peptide Binding Motif.* J Experimental Medicine, 1993. 178: p. 1877-1892.
99. Tan, P.T.J., S. Ranganathan, and V. Brusic, *Deduction of functional peptide motifs in scorpion toxins.* J Peptide Sci, 2006. 12: p. 420–427.
100. Neuwald, A.F., J.S. Liu, and C.E. Lawrence, *Gibbs motif sampling: Detection of bacterial outer membrane protein repeats.* Protein Science, 1995. 4: p. 1618-1632.
101. Thompson, W.A., et al., *The Gibbs Centroid Sampler.* Nucleic Acids Research, 2007. 35.
102. Rigoutsos, I. and A. FIorafos, *Motif Discovery without Alignment or Enumeration.* 1998, New York, NY.
103. Reiss, D.J. and B. Schwikowski, *Predicting protein–peptide interactions via a network-based motif sampler.* Bioinformatics, 2004. 20: p. i274–i282.
104. Reche, P.A., P.J. Glutting, and E.L. Reinherz, *Prediction of MHC Class I Binding Peptides Using Profile Motifs.* Human Immunology, 2002. 63: p. 701–709.
105. Reche, P.A., et al., *Enhancement to the RANKPEP resource for the prediction of peptide binding to MHC molecules using profiles.* Immunogenetics, 2004. 56: p. 405–419.
106. Rajapakse, M., et al., *Multi-Objective Evolutionary Algorithm for Discovering Peptide Binding Motifs.* 2006, Heidelberg, Germany: Springer-Verlag Berlin. 149 – 158.
107. Dinkel, H. and H. Sticht, *A computational strategy for the prediction of functional linear peptide motifs in proteins.* Bioinformatics, 2007. 23(24): p. 3297–3303.
108. Neduva, V., et al., *Systematic Discovery of New Recognition Peptides Mediating Protein Interaction Networks.* PLoS Biology, 2005. 3(12).
109. Savoie, C.J., N. Kamikawaji, and T. Sasazuki, *Use of BONSAI decision trees for the identification of potential MHC Class I peptide epitope motifs.* 1999: p. 182-189.
110. Engin, O., M. Sayar, and B. Erman, *The introduction of hydrogen bond and hydrophobicity effects into the rotational isomeric states model for conformational analysis of unfolded peptides.* Physical Biology, 2009. 6: p. 016001.
111. Griffiths-Jones, S.R., et al., *Modulation of intrinsic phi,psi propensities of amino acids by neighbouring residues in the coil regions of protein structures: NMR analysis and dissection of a beta-hairpin peptide.* J Mol Biol., 1998. 284(5): p. 1597-609.
112. Swindells, M.B., M.W. MacArthur, and J.M. Thornton, *Intrinsic phi, psi propensities of amino acids, derived from the coil regions of known structures.* Nat Struct Biol. , 1995. 2(7): p. 596–603.

113. Callen, H.B., ed. *Thermodynamics and an introduction to thermostatistics*. Second ed. 1985, Wiley.
114. Plotnikov, A.N., et al., *Crystal Structures of Two FGF-FGFR Complexes Reveal the Determinants of Ligand-Receptor Specificity.* Cell, 2000. 101: p. 413–424.
115. Prabhu, N.V. and K.A. Sharp, *Heat Capacity.* Proteins Annu. Rev. Phys. Chem, 2005. 56: p. 521-458.
116. Zídek, L., M.V. Novotny, and M.J. Stone, *Increased protein backbone conformational entropy upon hydrophobic ligand binding.* Nature Structural Biology, 1999. 6: p. 1118-21.
117. Boczko, E.M. and C.L. Brooks, *First-principles calculation of the folding free-energy of a 3-helix bundle protein.* Science, 1995. 269: p. 393-6.
118. Makhatadze, G.I. and P.L. Privalov, *On the entropy of protein-folding.* Protein Sci., 1996. 5: p. 507-10.
119. Chothia, C., *Hydrophobic bonding and accessible surface area in proteins.* Nature, 1974. 248: p. 338-9.
120. Goldberg, D.E. and K. Deb, *Foundations of Genetic Algorithms*. 1991, San Mateo, California: Morgan Kaufmann Publishers. 69-93.
121. Michalewicz, Z., *Genetic Algorithms + Data Structures = Evolution Programs*. 1992: Springer.
122. Humphrey, W., A. Dalke, and F. Schulten, *VMD - Visual Molecular Dynamics.* J. Molec. Graphics, 1996. 14: p. 33-38.
123. Huxford, T. and G. Ghosh, *A Structural Guide to Proteins of the NF-kappaB Signaling Module.* Cold Spring Harbor Perspect Biol., 2009. 1(3): p. a000075.
124. Huxford, T., et al., *The crystal structure of the IkappaBalpha/NF-kappaB complex reveals mechanisms of NF-kappaB inactivation.* Cell, 1998. 95(6): p. 759-770.
125. Dickson, C., et al., *Tyrosine kinase signalling in breast cancer. Fibroblast growth factors and their receptors.* Breast Cancer Res, 2000. 2: p. 191–196.
126. Kavakli, I.H. and A. Sancar, *Circadian Photoreception in Humans and Mice.* Molecular Interventions, 2002. 2: p. 484.
127. Todo, T., et al., *Similarity Among the Drosophila (6-4)Photolyase, a Human Photolyase Homolog, and the DNA Photolyase-Blue-Light Photoreceptor Family.* Science, 1996. 272: p. 109.
128. Hsu, D.S., et al., *Putative Human Blue-Light Photoreceptors hCRY1 and hCRY2 Are Flavoproteins.* Biochemistry, 1996. 35: p. 13871.
129. Pegg, S.C.H., J.J. Haresco, and I.D. Kuntz, *A genetic algorithm for structure-based de novo design.* J Comput Aid Mol Des, 2001. 15(10): p. 911-933.
130. Haliloglu, T. and B. Erman, *Analysis of Correlations between Energy and Residue Fluctuations in Native Proteins and Determnation of Specific Sites for Binding.* Physical Review Letters, 2009. 102: p. 088103.
131. Haliloglu, T., A. Gul, and B. Erman, *Predicting the Important Residues and Interaction Pathways in Proteins Using the Gaussian Network Model. Application to Binding and Stability of HLA Proteins.* PLOS computational Biology, 2010.
132. Haliloglu, T., E. Seyrek, and B. Erman, *Prediction of Binding Sites in Receptor-Ligand Complexes with the Gaussian Network Model.* Physical Review Letters, 2008. 100: p. 228102.
133. Inc, H., *HyperChem Professional*: Florida, USA.
134. Ramachandran, G.N., C. Ramakrishnan, and V. Sasisekharan, *Stereochemistry of polypeptide chain configurations.* J Mol Biol, 1963. 7: p. 95-99.
135. Fitzkee, N.C., P.J. Fleming, and G.D. Rose, *The Protein Coil Library: a structural database of nonhelix, nonstrand fragments derived from the PDB.* Proteins, 2005. 4: p. 852-854.

136. Tuncbag, N., et al., *Towards inferring time dimensionality in protein-protein interaction networks by integrating structures: the p53 example.* Molecular Biosystems, 2009. 5: p. 1770-1778.
137. Vanhee, P., et al., *Protein-Peptide Interactions Adopt the Same Structural Motifs as Monomeric Protein Folds.* Structure, 2009. 17: p. 1128-1136.
138. Tuncbag, N., A. Gursoy, and O. Keskin, *Identification of Computational Hot Spots in Protein Interfaces: Combining Solvent Accessibility and Inter-Residue Potentials Improves the Accuracy.* Bioinformatics, 2009. 25: p. 1513-1520.
139. Karplus, P.A., *Experimentally observed conformation-dependent geometry and hidden strain in proteins.* Protein Science, 1996. 5(7): p. 1406-1420.
140. *DS Modeling*, Accelrys Inc.
141. Louis, J.M., et al., *Hydrophilic Peptides Derived from the Transframe Region of Gag-Pol Inhibit the HIV-1 Protease.* Biochemistry, 1998. 37: p. 2105-2110.
142. Fujinaga, M., et al., *The molecular structure and catalytic mechanism of a novel carboxyl peptidase from Scytalidium lignicolum.* PNAS, 2004. 101: p. 3364-3369.
143. Kumar, J., et al., *Crystal structure of the complex formed between signalling protein from goat mammary gland (SPG-40) and a tripeptide Trp-Pro-Trp at 2.8A resolution.* 2009.
144. Nam, K.H., et al., *Crystal structure of an EfPDF complex with Met-Ala-Ser based on crystallographic packing.* Biochemical and Biophysical Research Communications 2009. 381: p. 630-633.
145. Oldenburg, K.R., et al., *Peptide ligands for a sugar-binding protein isolated from a random peptide library.* PNAS, 1992. 89: p. 5393-5397.
146. Zhang, Z., et al., *Crystal Structure of the Complex of Concanavalin A and Tripeptide.* Journal of Protein ChemistrY, 2001. 20: p. 59-65.
147. Hulsmeyer, M., et al., *Dual, HLA-B27 Subtype-dependent Conformation of a Self-peptide.* J Exp Med, 2004. 199: p. 271–281.
148. Fiorillo, M., et al., *Allele-dependent Similarity between Viral and Self-peptide Presentation by HLA-B27 Subtypes.* J Biol Chem, 2005. 280: p. 2962–2971.
149. Rucker, C., et al., *Conformational Dimorphism of Self-peptides and Molecular Mimicry in a Disease-associated HLA-B27 Subtype.* J Biol Chem, 2006. 281: p. 2306–2316.
150. Thorner, M.O. and M.L. Vance, *Growth Hormone.* J. Clin. Invest., 1998. 82: p. 745-747.
151. Lin-Su, K. and M. Wajnrajch, *Growth Hormone Releasing Hormone (GHRH) and the GHRH Receptor.* Reviews in Endocrine & Metabolic Disorders, 2002. 3: p. 313-323.
152. Schiffer, C., et al., *Structure of a Phage Display-derived Variant of Human Growth Hormone Complexed to Two Copies of the Extracellular Domain of its Receptor: Evidence for Strong Structural Coupling between Receptor Binding Sites.* J. Mol. Biol., 2002. 316: p. 277-289.
153. Baumgartner, J.W., et al., *The Role of the WSXWS Equivalent Motif in Growth Hormone Receptor Function.* J Biological Chemistry, 1994. 269(46): p. 29094-29101.
154. Clackson, T. and J.A. Wells, *A Hot Spot of Binding Energy in a Hormone-Receptor Interface.* Science, 2008. 267: p. 383-386.
155. Cosman, D., et al., *A new cytokine receptor family.* Trends Biochem Sci, 1990. 15: p. 265-270.
156. Bazan, F., *Structural Design and Molecular Evolution of a Cytokine Receptor Superfamily.* Proc Natl Acad Sci, 1990. 87: p. 6934-6938.
157. Wells, J.A., *Binding in the Growth Hormone Receptor CoMplex.* Proc Natl Acad Sci, 1996. 93: p. 1-6.
158. Huo, S., I. Massova, and P.A. Kollman, *Computational Alanine Scanning of the 1! Human Growth Hormone-Receptor Complex.* J of Computational Chemistry, 2001. 23(1): p. 15-27.
159. Chantalat, L., et al., *The Crystal-Structure of Wild-Type Growth-Hormone at 2.5 Angstrom Resolution.* Protein Pept Lett., 1995. 2: p. 333.

160. Thornberry, N.A. and S.M. Molineaux, *Interleukin-1P converting enzyme: A novel cysteine protease required for IL-lP production and implicated in programmed cell death* Protein Science, 1995. 4: p. 3-12.
161. Shahripour, A.B., et al., *Structure-Based Design of Caspase-1 Inhibitor Containing a Diphenyl Ether Sulfonamide.* Bioorganic & Medicinal Chemistry Letters, 2001. 11: p. 2779–2782.
162. Mjalli, A.M.M., et al., Bioorg. Med. Chem. Lett., 1993. 3: p. 2689.
163. Shahripour, A.B., et al., *Structure-Based Design of Nonpeptide Inhibitors of Interleukin-1 Converting Enzyme (ICE, Caspase-1).* Bioorganic & Medicinal Chemistry Letters, 2002. 10: p. 31-40.
164. Chapman, K.T., Bioorg. Med. Chem. Lett., 1992. 2: p. 613—618.
165. Romanowski, M.J., et al., *Crystal Structures of a Ligand-free and Malonate-Bound Human Caspase-1: Implications for the Mechanism of Substrate Binding.* Structure, 2004. 12: p. 1361–1371.
166. Nicholson, D.W., *Caspase structure, proteolytic substrates, and function during apoptotic cell death.* Cell Death and Differentiation, 1999. 6: p. 1028 – 1042.
167. Vanhee, P., et al., *Computational design of peptide ligands.* Trends in Biotechnology, 2011. 29(5).
168. Cosman, D., et al., *A new cytokine receptor family.* Trends Biochem. Sci., 1990. 15: p. 265-270.
169. Bazan, F., *Structural Design and Molecular Evolution of a Cytokine Receptor Superfamily.* Proc. Natl. Acad. Sci., 1990. 87: p. 6934-6938.
170. Lee, J.Y., *Purification of Biologically Active Human Erythropoietin-Binding Protein and Detection of its Binding Sites.* Annals of Clinical & Laboratory Science, 2007. 37(1).
171. Khankin, E.V., et al., *Soluble Erythropoietin Receptor Contributes to Erythropoietin Resistance in End-Stage Renal Disease.* PLoS ONE, 2010. 5(2): p. e9246.
172. Wognum, A.W., *Mini-Review. Erythropoietin.*
173. Cheetham, J.C., et al., *NMR structure of human erythropoietin and a comparison with its receptor bound conformation.* Nature structural biology 1998. 5(10).

VITA

Evrim Besray Ünal was born in Istanbul, Turkey, on December 13, 1981. She is an alumnus of Kültür Fen Lisesi, Istanbul. She received her Bachelor of Science degree in molecular biology and genetics from Boğaziçi University, Istanbul, in June 2004. She received M.Sc. Degree in in Computational Science and Engineering from Koc University in 2006. Her research included the interaction between Cryptochrome and Melanopsin proteins; Circadian rhythms. She received the Ph.D. degree from Koc University in Computational Science and Engineering in 2011. From 2007 to 2011, she was a research and teaching assistant in the Computational Science and Engineering Department of Koç University. Her research focuses on peptide design problem and binder peptide properties. She will continue her academic career as a Postdoctoral Associate with Luis Serrano and Johannes Jaeger, at Centre for Genomic Regulation Barcelona, where she will focus on reverse engineering of *Mycoplasma pneumonia*.