ISSN:0975-9646

A Genetic Algorithm for Conformation Search Optimization in Molecular Docking

Sudha Ramachandra[1] , Vinay Chavan[2]

[1]Department Of Computer Science,
Santaji Mahavidyalaya, Nagpur, Maharashtra, India

[2]Department Of Computer Science
Seth Kesarimal Porwal College , Kamptee,
Nagpur , Maharashtra, India

Abstract: Genetic Algorithm is an indispensible tool in molecular docking for studying binding interactions between target protein and ligand (drug) in drug discovery process. Among the various molecular docking tools available, the most widely used is AutoDock which uses several search algorithms to generate various poses of the ligand in the active site of the protein. Most of the search algorithms suffer from the problem of premature convergence. To overcome the problem of premature convergence and to further enhance the performance of AutoDock interms of finding the lowest binding energy , the researchers in the present work have designed and developed an algorithm called Hybrid-ALPS. Hybrid-ALPS a conformation search optimization algorithm developed in the present investigation has been tested with CYP2B6 and its various polymorphs as receptors and an anti cancer drug Cyclophosphamide as ligand. It has significantly performed better than its precursor Generational-ALPS implemented in AutoDock elsewhere by overcoming premature convergence, in getting lowest binding energy and more number of poses with negative binding energy values. By changing the objective function of Hybrid-ALPS, it can also be used for any search optimization problems in other areas .

Keywords: receptor, ,ligand, , conformation, molecular docking, binding energy, binding affinity, pose, individual, population, selection, crossover, replacement, age, premature convergence

I. INTRODUCTION

Use of computational methods is a cost effective strategy for speeding up the process of drug discovery and development process. The drugs that enter the human body tend to stimulate certain receptors. Most receptors are made up of proteins. The interaction of drug (ligand) with the amino acids of protein as receptor is called as drug binding or docking . Drug binding occurs only at a specific place in the receptor called active site. When a drug binds to a receptor, it results in the change of conformation of the three dimensional structure of the receptor leading to change in the receptor or protein functions in the body. Hence understanding binding interactions between receptor and ligand is very essential for drug discovery scientists.

Molecular docking, a computational method of studying binding interactions in terms of binding energies is immensely used in the process of drug discovery to save on cost and time. In this method , computer generated representation of a small molecule or ligand is placed into the active site of the target or protein's computational structure in a variety of positions, conformations and orientations. The position, orientation and conformation of the ligand in the active site of protein is called as a 'pose'. In order to identify the energetically most favorable pose, each pose of the ligand is evaluated for binding energy computationally. The main objective of molecular docking method is to find a pose which has the lowest binding energy [1,2].

The search algorithm for generating poses of the ligand in the active site of protein and scoring function to score poses which are near to the experimentally determined pose are the two important components of any molecular docking tool. AutoDock is one of the widely used molecular docking tool [3,4,5]. It uses several Evolutionary Algorithms (EA) or Genetic Algorithms (GA) as search algorithms [6]. However, the docking performance of these algorithms is limited by the local optima issues or premature convergence. Various approaches have been tried over the years for overcoming the problem of premature convergence. Earlier research has tried to address this problem through an approach called Age Layered Population Structure (ALPS) [7]. ALPS uses a special notion of age which indicates how long the genotypic material has existed in the population and the evolving population is segregated into age layers. Age is assigned to an individual by different methods. Each of the method set an upper bound on the age of an individual in a particular layer. The EA acts on each age layer and each age layer undergoes selection, crossover and replacement . Age of an individual is incremented after each generation of selection , crossover and replacement. If the age of the individual exceeds the upper limit of the age layer, the individual is moved to the next upper layer. Additionally, to maintain diversity throughout, random individuals are periodically inserted into the youngest population layer. This version of ALPS called Generational-APLS is already implemented as search algorithm elsewhere in AutoDock [7].

An improved version of ALPS called Steady State ALPS (SS-ALPS) which requires a revised measure of age is also reported [8]. In Generational-ALPS, the age of all

individuals is increamented at the end of a generation and SS-ALPS keeps track of the number of energy evaluations [7,8,9].

Incorporating the advantages of both Generational-ALPS and SS-ALPS into a single algorithm to circumvent the problem of premature convergence and to further enhance the performance interms of finding the lowest binding energy of widely used "AutoDock" molecular docking tool was the main aim of the present research work. Hence the authors in the present work have devised and developed an algorithm called Hybrid-ALPS, a conformation search optimization algorithm for ligand - protein docking program in virtual screening of structure-based drug design. Hybrid-ALPS has combined the concept of both Generational-ALPS and SS-ALPS and also novel concepts have been adopted.

II. METHODS

Hybrid-ALPS, a GA incorporating the novel ideas has synergized the concept of both Generational-ALPS and SS-ALPS. In this algorithm, initial population is a set of individuals or chromosomes which constitute solution to the problem. Population is segregated into various age layers. In Hybrid-ALPS, fibonacci aging scheme is used to set upper limit on the age of an individual in a layer. Maximum age limit for each layer according to this scheme are as shown in the following table.

Layer no	Max age limit
0	1
1	2
2	3
3	5
4	8
5	13

Table 1:Age limits for each layer

Every individual has three translational genes representing x, y, z coordinates of the ligand in the grid, four orientational genes constituting the orientation of the ligand and torsional genes depending upon the torsional angles of the ligand. The translational genes are given the random values within the boundary of the grid, orientational genes are also given random values and torsional genes are given random values between -180 degree to +180 degree. The fitness of each individual given by binding energy is evaluated by an Objective Function. Binding energy is computed as the sum of intermolecular interaction energy between receptor and ligand and intra molecular torsional energy of the ligand. AutoGrid tool of AutoDock is used to calculate the intermolecular energy of various atom types of the ligand with the protein. AutoGrid generates the grid map file for each of the atom type of the ligand which are used as look up table to calculate the inter molecular energy between the protein and a particular ligand pose by Trilinear Interpolation. Intra molecular energy is modelled using Fourier Series. The sum of intermolecular energy and intramolecular torsional energy constitutes the total binding energy [10,11,12,13].

Like any other GA, to evolve better and better individuals, Hybrid-ALPS goes through the process of selection, crossover and replacement. Constraint is placed on selecting individuals for crossover.

- The individuals can breed only with individuals from their own layer or from the layer younger to it.
  Example, for layer 0, parents are selected from individuals only in layer 0; for layer 1 parents are selected from individuals in layers 0 and 1; for layer 2 parents are selected from individuals in layers 1 and 2; Mainitaining a crossover probability of 0.85, a dual method of selection of parents for crossover has been proposed in the present work. According to this,
- Individuals are sorted according to energy values and apportioned into two groups. The first group is called as high fit group comprising of lower energy values and the second group is named as low fit group consisting of high energy values.
- In the first method one parent is chosen randomly from high fit group and another randomly from low fit group and in the second method both the parents will be chosen from the high fit group of low energy values.

After selecting parents , heuristic crossover technique [14] is applied genewise to mate parents . Specifically, crossover is performed on three translational, four orientational and torsional genes. Before mating, the values of a gene from both the parents are tested for equality. If the values of a gene from both the parents are same, then the gene is mutated and the modified gene is used for crossover. Age of the parents which undergo crossover are updated according to the formula

$$\mathbf{age = 1 + (evalscurrent - evalscreated) / currentpopulation}$$

where

- evalscurrent : number of evaluations done so far
- evalscreate : number of evaluations done at the time of creation of the individual
- currentpopulation: As there is variable population in Hybrid-ALPS , the value of current population is used to calculate the age of a randomly generated individual.

Offsprings get the age of their youngest parents. The total interaction energy or binding energy for each of the offspring generated from crossover is calculated as the sum of intermolecular and intramolecular energy. Offsprings generated from heuristic crossover replaces the parents according to replacement rule of "deterministic crowding" [15].

The entire method of crossover and replacement is repeated for both the methods of parents selection in sequence. After completion of selection, crossover and replacement of the current layer, 0.05*number of individuals which took part in reproduction are added to the current layer which participated in reproduction.

Hybrid-ALPS adopts the concept of Generational-ALPS by doing 'n' crossovers and replacements and follows the concept of SS-ALPS by not eliminating the individuals

which did not take part in the reproduction and also calculating the age on the basis of energy evaluations.

After reproduction, the individuals exceeding the maximum age limit of a layer are moved to the next higher layer as follows.

- First the least fit individual in the next higher layer is located.
- The fitness of the individual which is to be moved and the fitness of the least fit individual in the next higher layer are compared.
- If the fitness of the individual which is to be moved is better than the least fit individual in the next higher layer then the individual replaces the least fit individual in the next higher layer provided the least fit individual was not moved 'n ' evaluations ago.
- If the least fit individual of the next higher layer was moved 'n' evaluations ago, then the individual from the layer which has exceeded the maximum age limit is just moved to the next higher layer.

An individual gets replaced and moves out of the population as better and better individuals are evolved. This accounts for variable population in Hybrid-ALPS. The entire process of movement is repeated for all the individuals in each layer.

To maintain diversity throughout the algorithm, new randomly generated individuals replace the individuals in the zeroth layer after certain number of energy evaluations. The number of individuals which are added are the same as number of individuals of initial population. After completion of given number of energy evaluations, Hybrid-ALPS outputs

- The lowest binding energy from among the binding energies of the poses generated
- lists and counts the poses which have negative binding energies.
- Counts the poses which have same binding energy values.

Further, to arrive at the actual lowest binding energy of entire docking process, the poses generated by the search algorithm Hybrid-ALPS have to be scored by the scoring function. Implementation of scoring function is beyond the scope of this work.

III. RESULTS AND DISCUSSIONS

Generational-ALPS which is a precursor to Hybrid-ALPS has also been developed and tested on the same platform as that of Hybrid-ALPS for comparative analysis. CYP2B6 is most important but less understood drug-metabolising enzyme and is also polymorphic in nature [16,17,18,19]. Hence both the algorithms have been tested on CYP2B6 and its Single Nucleotide Polymorphs (SNPs) namely SNP 99, SNP 259, SNP 336, SNP 423, SNP 487 as receptor protein and anti-cancer drug Cyclophosphamide as ligand [20] with a population size of 150 and a crossover probability of 0.85.

Both the algorithms generate various poses of the ligand within the grid generated by the AutoGrid tool of AutoDock. They also calculate the lowest binding energy from among the binding energies calculated for various poses generated by the search algorithm

To arrive at the final binding affinity value, the pose which is having the lowest binding energy value have to be scored by a scoring function. This pose may or may not be near to the experimentally determined ones. In such events it is better to have a choice of poses with lower energy values. Hence both the algorithms also outputs number of poses with negative energy values.

Since lowest energy of a pose from among the poses finally present in the population is selected, it is better to have a large universe. Accordingly, both the algorithms also output the final population size which helps in determining how large the universe is.

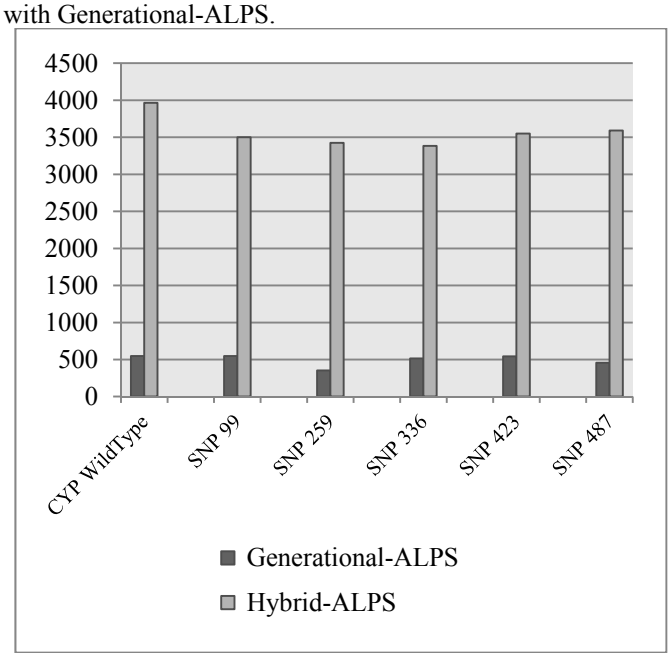
To test for local optima issues, both the algorithms display number of poses which have same energy values.

Also, time taken to complete all the activities in both the algorithms starting from initial population generation, selection of parents, crossovers, mutation and bottom layer replacement is also recorded.

The detailed comparative analysis between Generational-ALPS and Hybrid-ALPS for all the vital parameters namely "Final Population Size" , " Lowest Binding Energy", "No of Poses with Same Negative Binding Energy " and "Time" along with the salient features of the algorithm contributing to the variation of the parameters are enumerated below.

A. Final Population Size

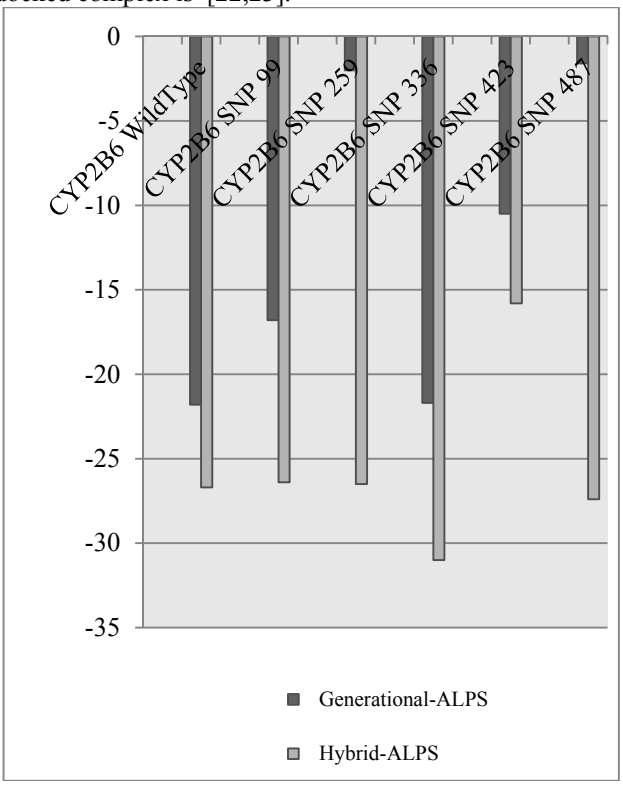
The final population size is very large in Hybrid-ALPS as compared to Generational-ALPS (Graph 1) . Therefore minimum energy found from such a large population is indicative of global minima [21]. The resulting final large population from Hybrid-ALPS is due to the fact that (0.05*size of the layer which undergoes reproduction) new individuals are added each time to the older layer which undergoes reproduction. This is a novel concept built into Hybrid-ALPS and this feature is not available with Generational-ALPS.



Graph 1: Comparison of Final Population Size

B. Lowest Binding Energy

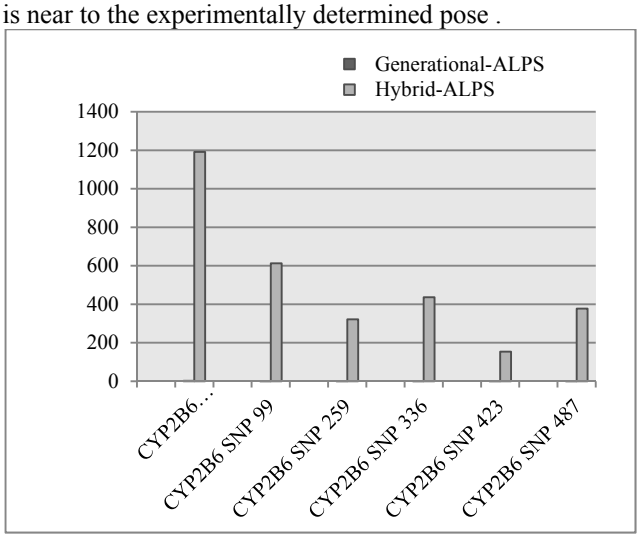
The Lowest Binding Energy found from Hybrid-ALPS with each of the receptor is significantly lower than the Lowest Binding Energy found from Generational-ALPS (Graph 2). Since weak individuals are also taking part in reproduction in Hybrid-ALPS, some of the good genes of weak individuals helps in progressing the search. Selected individuals undergo heuristic crossover in Hybrid-ALPS and "deterministic crowding" replacement policy. In deterministic crowing, each child competes against one of the parents in a tournament, obtaining the individual that survives for the next generation. From this strategy the worst individuals are replaced by the best ones. Furthermore, by doing "n" crossovers and replacements each time, individuals are replaced by fitter and fitter ones in each generation. Besides this, in Hybrid-ALPS individuals which did not take part in reproduction are not eliminated and this helps in the survival of some of the good individuals and these individuals may later partake in reproduction. All these activities coupled with large population size have facilitated Hybrid-ALPS in achieving more lowest binding energy as compared to Generational-ALPS. More lower the binding energy, more stable the docked complex is [22,23].



Graph 2: Comparison of lowest Binding Energy

C. Number of Poses With –ve Binding Energy

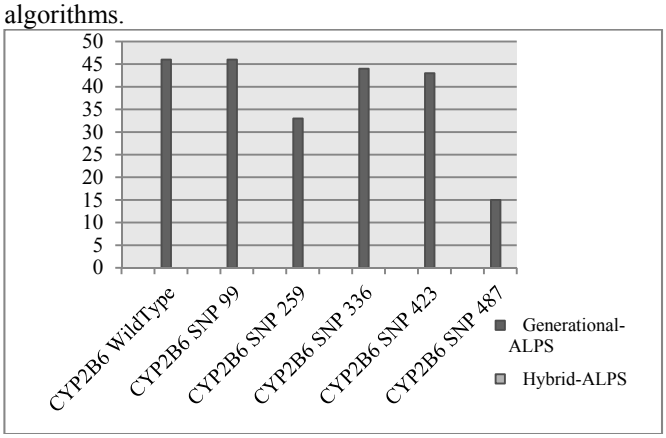
To arrive at the final binding affinity score of docking, all the poses generated by search algorithms such as Hybrid-ALPS and Generational-ALPS have to be scored by a scoring function. Scoring function, an essential component of molecular docking programs ranks the poses generated by the search algorithm and identifies the most stable binding pose of a ligand, when bound to a receptor protein, from among a large set of candidate poses. There are chances that pose with the lowest binding energy found by the search algorithm may not be the preferred binding pose by the scoring function. In such cases, if large number of poses with negative binding energy are available then it helps scoring function to score a most stable binding pose which is near to experimental one. Large number of poses with negative binding energy are available in Hybrid-ALPS with each of the receptor as compared to Generational- ALPS (Graph 3) indicating that Hybrid-ALPS has explored relevant conformational space. This facilitates the scoring function to score a pose which is near to the experimentally determined pose .



Graph 3: Comparison of No of Poses with –ve Energy Values

D. Number of Poses With Same Energy Values

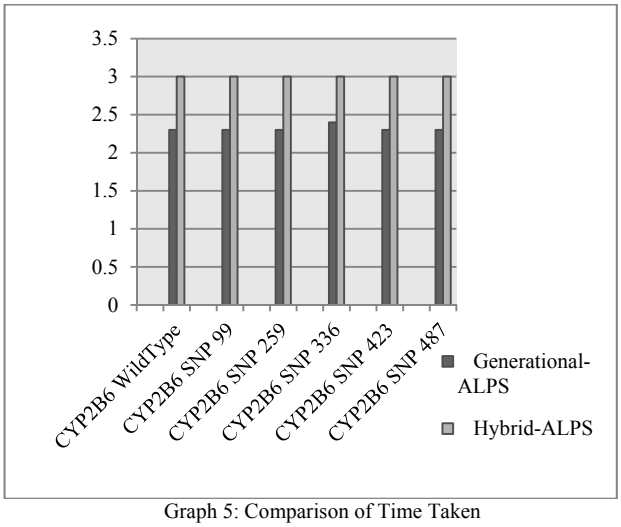
Generational-ALPS has obtained a large number of poses with duplicate energy values with each of the receptor (Graph 4). This is because the selection method may not explore the full search space and during crossovers the duplicate gene values may contribute to the large number of poses with identical energy values. But in Hybrid - ALPS for each of the receptor tested no poses are found with identical binding energy values. It is achieved because of selection method which helps in exploring the larger search space and mutation before crossover has definitely avoided getting the identical energy values. This corroborates that Hybrid-ALPS has surmounted the problem of premature convergence faced by other search algorithms.



Graph 4: Comparison of Number of poses With Same Energy Values

E. Time

As Hybrid-ALPS explores larger search space because of selection methods, mutation before crossover and addition of new individuals to each layer after reproduction, the time taken for execution is slightly more than its precursor with each of the receptor (Graph 5). But, the additional advantages offered by the algorithm overrides the time factor.



Graph 5: Comparison of Time Taken

IV. CONCLUSIONS

Hybrid-ALPS, a conformation search optimization algorithm developed in the present investigation has performed significantly better than its precursor Generational-ALPS by overcoming premature convergence, in getting lowest binding energy and more number of poses with negative binding energy. The search algorithm and scoring function are two important components of any molecular docking tool. Since the output of the search algorithm goes as input to the scoring function, having a good search algorithm in molecular docking tool is very essential in improving the overall performance. Hence, the drug discovery and development process can be enormously benefited by integrating Hybid-ALPS, a better performing conformation search optimization algorithm with AutoDock which is one of the widely used molecular docking tool. Also, in Hybrid-ALPS objective function calculates the total binding energy as the sum of intermolecular and intra molecular energy. By changing the objective function of Hybrid-ALPS, it can be used for any search optimization problems in other areas.

ACKNOWLEDGMENT

Our sincere thanks are due to Dr. P.K.Butey and Dr. Satish Sharma for all the valuable suggestions , support and encouragements made throughout the course of this research work.

REFERENCES

[1]. Raquel Dias , Raquel Dias. 'Molecular Docking Algorithms' Current Drug Targets, 2008.
[2]. B. Vijayakumar and P. Dheen Kumar ,'MOLECULAR DOCKING STUDIES – A REVIEW', IJMCA, Vol 2, Issue 2, 2012 .
[3]. http://autodock.scripps.edu/
[4]. Bachwani Mukesh, Kumar Rakesh ,' Molecular Docking- A review', IRJP ,2(6), 2011.
[5]. iitb.vlab.co.in,. (2011)., "Experiment-12 : Molecular Docking. available: iitb.vlab.co.in/?sub=41&brch=118&sim=698&cnt=1
[6]. David E.Godberg , 'Genetic algorithms', Pearson Education India, 2006.
[7]. Emrah Atilgan , Jianjun Hu , 'Improving Protein Docking Using Sustainable Genetic Algorithms' International Journal of Computer Information Systems and Industrial Management Applications , Volume 3, 2011
[8]. Gregory S. Hornby, ' Steady-State ALPS for Real-Valued Problems ', GECCO'09, July 8–12, 2009.
[9]. Gregory S. Hornby , 'ALPS: The Age Layered Population Structure for Reducing the Problem of Premature Convergence', GECCO'06, Seattle, WA, USA, July 2006.
[10]. http://autodock.scripps.edu/faqs-help/tutorial/using-autodock-4-with-12. autodocktools/2012_ADTtut.pdf
[11]. http://www.csb.yale.edu/userguides/datamanip/autodock/html/Using_AutoDock_305.9.html
[12]. https://en.wikipedia.org/wiki/Trilinear_interpolation
[13]. http://cmt.dur.ac.uk/sjc/thesis_dlc/node74.html
[14]. Yılmaz KAYA Murat UYAR Ramazan TEKĐN ,'A Novel Crossover Operator for Genetic Algorithms:Ring Crossover', http://arxiv.org/ftp/arxiv/papers/1105/1105.0355.pdf
[15]. Manuel Lozano, Francisco Herrera, Jos´e Ram´on Cano,' Replacement Strategies to Maintain Useful Diversity in Steady-State Genetic Algorithms', Information Sciences 178 , 2008.
Ulrich M. Zanger , Matthias Schwab, 'Cytochrome P450 enzymes in drug metabolism: Regulation of gene expression, enzyme activities, and impact of genetic variation', Pharmacology and Therapeutics,Volume 138, Issue 1 ,April 2013.
[16]. https://www.broadinstitute.org/education/glossary/snp
[17]. Sarah C. Preissner ,Michael F. Hoffmann ,Robert Preissner,Mathias Dunkel, Andreas Gewiess, Saskia Preissner ,' Polymorphic Cytochrome P450 Enzymes (CYPs) and Their Role in Personalized Therapy', PLOS, December 10, 2013.
[18]. Magnus Ingelman 'Drug-Metabolising Enzymes: Genetic Polymorphisms',els, 15 November 2011.
[19]. https://pubchem.ncbi.nlm.nih.gov/compound/cyclophosphamide
[20]. PRIGOGINE and S.A. Rice,' Advances in Chemical Physics, Computational Methods for Protein Folding', Wiley,2002.
[21]. https://en.wikibooks.org/wiki/Structural_Biochemistry/Molecular_Moldeling/Molecular_Docking
[22]. Arif Malik, Mahmood Rasool, Abdul Manan, Aamer Qazi, and Mahmood Husain Qazi,'Advances in Protein Thermodynamics',ebook.