ACTIVE SITE DRIVEN LIGAND DESIGN: AN EVOLUTIONARY APPROACH

SANGHAMITRA BANDYOPADHYAY
Machine Intelligence Unit, Indian Statistical Institute,Kolkata 700 108, India
sanghami@isical.ac.in

ANGSHUMAN BAGCHI
Bioinformatics Center, Bose Institute, Kolkata 700 054,India
angshu@bic.boseinst.ernet.in

UJJWAL MAULIK
Dept. of Comp. Sc. and Engg., Jadavpur University,Kolkata 700 032, India
drumaulik@jdvu.ac.in

Running Title: EVOLUTIONARY APPROACH TO LIGAND DESIGN

Received (day/month/year)
Revised
Accepted

An evolutionary approach for designing a ligand molecule that can bind to the active site of a target protein is described in this article. An earlier attempt in this regard assumed a fixed tree structure of the ligand on both sides of the pharmacophore, and used a genetic algorithm for optimizing the van der Waals energy. However, it is evident that knowledge about the size of the tree is difficult to obtain an *a priori.* Moreover, it will also change from one active site to another. This limitation is overcome in the present article by using variable string length genetic algorithm (VGA) for evolving an appropriate arrangement of the basic functional units of the molecule to be designed, whose size may now vary. The crossover and mutation operators are appropriately redesigned in order to tackle the concept of variable length chromosomes. Once the geometry of the molecule is obtained, the possible three-dimensional structure and its docking energy is determined. Results are demonstrated for five different target proteins both numerically and pictorially. It is found that not only does the molecule designed using variable length representation, in general, have lower energy values, the docking energies are also lower, as compared to the molecule evolved using fixed size representation.

Keywords: Active site; ligand Docking; Target protein receptor; Tree structured representation; Variable string length genetic algorithm;

1. Introduction

The task of drug design has traditionally been dependent on nature, with some of the most effective drugs, like morphine and penicillin, having been obtained from natural sources. Their power stems from their unique structures that have evolved over millions of years of random variation and natural selection. But a potential drawback of the natural processes is that they are extremely slow. Therefore, for developing and commercializing drugs within a reasonable amount of time, pharmaceutical companies have devised proactive drug discovery methods, including a recent innovation called *rational drug design*. In this approach, researchers build and test small drug-like molecules based on prior knowledge about the three-dimensional structures of known drug molecules. Quantitave structure activity relationship (QSAR) [1-2] studies form an integral part of rational drug design. However, this approach has its own limitations, since it's not always clear which variations on known molecules are worth testing. Identification or design of drug molecules, without assuming any similarity with some known structure, that can target proteins crucial for the proliferation of microbial organisms, cancer cells or viruses is one of the important approaches in drug design. Such molecules can disrupt the action of the target protein, that sustain viral proliferation, by binding to its active site; thereby nullifying its activity which can be lethal to

us. Therefore the task of accurately predicting the structure of the potential inhibitors, while utilizing the knowledge about the structure of a target protein, is another important area of research.

Genetic algorithms (GAs) [3-7] are randomized search and optimization techniques guided by the principles of evolution and natural genetics, and have a large amount of implicit parallelism. GAs perform mulitimodal search in complex landscapes and provide near optimal solutions for objective or fitness function of an optimization problem. They have diverse application in the fields as diverse as pattern recognition, image processing, VLSI design, neural networks etc. [8-9]. GAs have also been applied to the domain of bioinformatics [10-13] including that of drug design [14-20]. The approach adopted in [20] is based on the use of genetic algorithms for evolving small molecules represented using a graphical structure composed of atoms as the vertices and the bonds as the edges. The task in [20] is to determine the effectiveness of GAs in evolving a molecule that is similar to a target molecule. Thus knowledge about the target molecule is assumed, which may not be readily available in many situations.

Another approach for ligand design, that is based on the presence of a fixed pharmacophore and that uses the search capabilities of genetic algorithms, was studied by Goh and Foster [16], where the harmful protein human Rhinovirus strain14 was used as the target. This pioneering work assumed a fixed tree structure representation of the molecule on both sides of the pharmacophore. Evidently, an *a priori* knowledge of the size of the tree is difficult to obtain. Moreover, it is known that no unique ligand structure is best for a given active site geometry. Therefore, in this study, we propose to use variable length representation of the trees on both sides of the pharmacophore for designing the ligand molecule. Variable string length genetic algorithm [21] is used for this purpose. Note that, in contrast to [16], here the chromosome, that encodes the ligand tree, can be of any size. This is a more natural representation since the size of the active site will itself be different for different proteins. In this article, we significantly extend the work in [16] by experimenting with a vast suite of proteins with different characteristics of the active site. Moreover, the two-dimensional molecules designed using genetic algorithms are now represented in three dimensions using a tool Insight II (MSI/Accelrys, San Diego, CA, USA). Thereafter the conformations and the energies of the protein-ligand interaction obtained using another software GOLD [22] are computed.

2 Relevance of genetic algorithm for drug design

One of the chief approaches in the field of drug design is the identification of proper ligands so that they can be used to generate the structure of a drug molecule. In general one or more proteins are typically involved in the bio-chemical pathway of a disease. The treatment aims to appropriately reduce the effect of such proteins, by designing a ligand molecule that can bind to its active site. That is, the structure of a ligand molecule is evolved from a set of groups in close proximity to crucial residues of the protein; a molecule is thereby designed that fits the protein target receptor such that a criterion (for example, van der Waals interaction energy) is optimized. Such an approach is adopted in this article.

Assuming that the number of functional groups is g, and the number of positions (in which these functional groups may appear) in the ligand is p, then the number of possible conformations is g^p. This is obviously a combinatorial optimization problem that cannot be solved by traditional methods since the time required would increase exponentially with the value of p. Consequently, application of efficient search and optimization techniques that can provide good (though not necessarily optimal) solutions in reasonably fast becomes imperative for the drug design problem.

Evolutionary computation is a class of search and optimization tools that have proved to be very popular and effective in solving hard and complex problems [12]. The principal components of evolutionary computation are genetic algorithms, evolutionary strategies, evolutionary programming and genetic programming. All these techniques are population based search mechanisms where the parameters of the search space are encoded, and the solutions evolve over a number of generations. Among these, genetic algorithms are widely used by researchers for solving a variety of problems very effectively. Based on this, GAs have been used in this article for solving the problem of ligand design. However, other evolutionary optimization techniques can be used as well for this purpose. We have also implemented evolutionary programming (EP) technique to solve the above problem and have obtained similar results.

In GAs, the parameters of the search space are encoded in the form of strings (called *chromosomes*). A collection of such strings is called a *population*. Initially, a random population is created, which represents different points in the search space. An *objective* and *fitness* function is associated with each string that represents the degree of *goodness* of the string. Based on the principle of survival of the fittest, a few of the strings are selected and each is assigned a number of copies that go into the mating pool. Biologically inspired operators like *crossover* and *mutation* are applied on these strings to yield a new generation of strings. The process of selection, crossover and mutation continues for a fixed number of generations or till a termination condition is satisfied.

Since the work proposed in this article is based on the approach adopted in [16], where GAs are used to design a ligand molecule assuming a fixed tree structure on both sides of the pharmacophore, we describe its basic principles briefly in the next section.

3 Genetic Algorithms for Ligand Design

GAs have been used earlier for different classification and clustering tasks [21], [23]. Recently, an application of GA has been reported in the area of structure based drug design [16]. It attempts to find the structure of a drug molecule for known barrel shaped protein target molecule, the rhinoviral protein, obtained from human rhinovirus strain 14. It uses a two-dimensional model of the system. Here a ligand molecule is assumed to adopt a fixed tree structure on both sides of the pharmacophore, whose location is fixed. Figure 1 shows a diagrammatic representation of the barrel shaped active site of rhinoviral protein, and the fixed tree structure of the ligand assumed in [16]. Note that the left hand side of the tree has seven nodes while the right hand side has ten nodes. Each node of the tree is filled up by a group selected from among a set of seven groups shown in Figure 2.

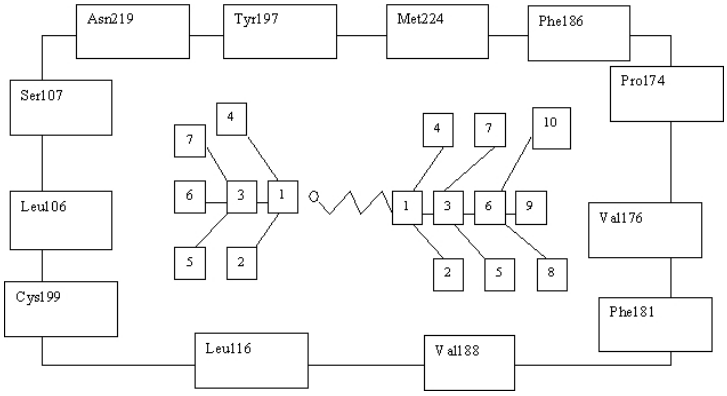


Fig. 1. Active site representation of the target protein along with the tree-structured molecule [16].

A fixed string length representation of the chromosome in the GAs is used that encodes the functional groups that are placed at the different nodes of the tree (shown inside the active site in Figure 1). The van der Waals interaction energy among the functional groups of the ligand and the closest residue of the active site of the target protein is taken as the minimizing criterion. Evidently, the knowledge about size of the ligand tree, which is dependent on the geometry of the active site, is difficult to estimate a priori. Moreover, there is no unique best ligand structure for a given active site geometry. Hence the method of [16], a pioneering work in the field of computer aided evolutionary drug discovery, was bound by design not to generalize. This motivated the work reported in this article, where a variable string length GA based technique is proposed, which will be applicable for protein active sites with differing geometries.

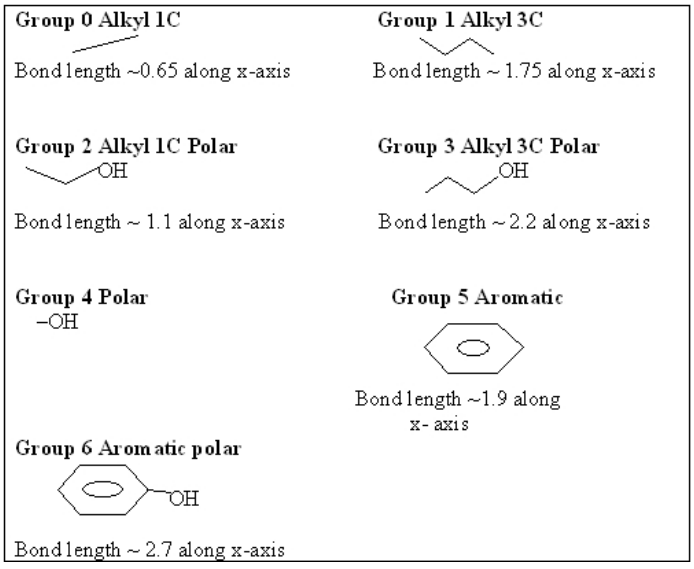


Fig. 2. Representation of the groups [16] (the bond lengths are given in [24]).

4 Proposed Technique

In the present study the size of the tree representing a drug molecule has been made variable, and for that purpose the concept of variable length strings in GAs (VGAs) has been adopted [21], [23]. Unlike conventional GAs, here the length of a string is not fixed. New crossover and mutation operators are accordingly defined in order to take care of the variable length chromosomes.

4.1 *Chromosome representation and population initialization*

As in [16], a chromosome will encode an entire tree structure on one side of the pharmacophore, with the nodes being filled up by the groups shown in Figure 2. However in contrast to [16], the size of the tree (and hence the length of the chromosome) will be variable. For practical purposes, the size of the tree has an upper limit, l_{max}, and a lower limit, l_{min} that is guided by the size of the active site. Moreover, the position of the pharmacophore is also guided by the structure of the active site. Note that in [16], the size of the left tree is seven and that of the right tree is ten. For initializing chromosome i, first a random integer l_i (representing the length of the chromosome) in the range l_{min} to l_{max} is generated. Thereafter each gene position is filled up by a group chosen randomly from the aforementioned 7 groups (see Figure 2). The remaining ($l_{max} - l_i$) positions are filled up by #s (which denote the don't care symbol). For example, suppose that $l_{max} = 8$, the size of the right tree i.e., l_i is five, then a chromosome may look like 01252###. The corresponding right tree is shown in Figure 3.

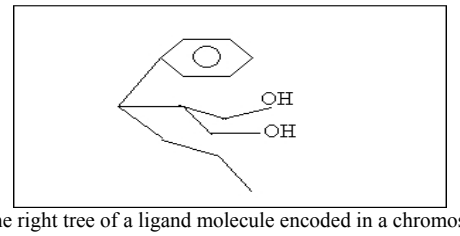


Fig. 3. The representation of the right tree of a ligand molecule encoded in a chromosome mentioned in an example in Section 4.1

4.2 *Fitness computation*

Fitness evaluation is based on the proximity of the residues in the active site to the closest functional groups and the chemical properties of these pairs. Note that the distance between a residue of the target protein receptor and its closest functional group should be at most 2.7Å for the molecules to interact and should not be closer than 0.65 Å for avoiding steric contact. Therefore the chromosomes whose functional groups do not meet these constraints are penalized. If the functional group closest to a particular residue of the receptor target is in a different electronic environment, then again a penalty is imposed. For example, the residue MET 224 is a methionine residue, which is hydrophobic. The group closest to it (present in the drug molecule) should also be a hydrophobic group for the drug molecule to score a good fitness value. The van der Waals energy value is computed which is given by

$$[(C_n / r^6) - (C_m / r^{12})],$$

where n and m are integers and C_n and C_m are constants [25]. The van der Waals energy was computed using the energy expression as given by van der Waals. The total energy of the chromosome is the sum of all the energy terms as obtained by calculating the interaction

energy between the groups to be used in specific positions of the evolved drug molecule and the amino acid residues present at the active site of the protein target receptor. Note that as aforementioned, when a penalty is imposed on particular unfavorable conformations of the molecule being evolved, in effect its energy is increased by adding a large positive integer. The fitness value is taken as the inverse of this energy, the maximization of the fitness by VGAs leads to the minimization of the energy.

4.3 *Genetic operators*

The genetic operators in VGA are selection, crossover, and mutation.

Selection: In order to select the best-fit parents from a mating pool the principle of roulette wheel [3] for implementing the proportional selection strategy is employed. The parents are selected according to their fitness values. The better fitness of the chromosomes, the more is the likelihood of their selection.

Crossover: The crossover operator in VGA approach is modified as follows:
During crossover between two chromosomes of unequal length, the chromosome with smaller length is padded with # (don't care for example) symbols and standard single point crossover is performed. As an example, let there be two chromosomes as:
Chromosome 1: 1024 ====> 1 0 2 4 # # | # #
Chromosome 2: 34651345 ====> 3 4 6 5 1 3 | 4 5
Let the crossover point be as shown above (with a line). After crossover the two offspring are 1024##45 and 346513##

Mutation: Each position of the offspring is considered in turn. It is first checked whether mutation should at all be performed or not according to a mutation probability value. In case the position is to be mutated, then it is checked whether the position contains a valid integer or #. If it contains an integer, then this position is set to # with a probability u/l, where u = number of #s in the chromosome and l = length of the chromosome. Otherwise, conventional mutation is performed. (In conventional mutation, the value at that position is replaced by an integer randomly selected from the range (0-6) corresponding to some other functional group.) If the position contains a #, then it is set to an integer I the range (0-6) (chosen randomly) according to probability $(l-u)/l$. At the end all the #'s are moved to the end of the string, and are subsequently removed. As an example, consider the chromosome 1024##45, which may be converted to 132432#5, and finally to 1324325 after removal of the #.

5 Experimental Results

Five proteins, with known active site conformations, are taken for the purpose of experimentations. These are HIV-I Nef protein [26], HIV-I Integrase [27], Retroviral protease [28], HIV-I capsid protein [29], and Rhinoviral protein obtained from human rhinovirus strain 14 [16].

The coordinates of the proteins were obtained from Protein Data Bank [30]. The human rhinovirus strain 14 is a causative agent of common cold. HIV-I Nef protein is retroviral protease acting on proteins of host cells. HIV-I Integrase accelerates virulent progression of acquired immunodeficiency syndrome (AIDS) by its interaction with specific cellular proteins involved in signal transduction and host cell activation. Retroviral protease is a

protein present in the envelope of the virus that surrounds the retroviral cellular organization. This protein is essential for the structural arrangements of the viral cellular organization. HIV-I capsid protein is the protein present in the outer covering of the virus. The absence of the protein leads to the disorganization of the viral cellular materials. Note that four of the above mentioned proteins (viz., HIV-I Nef protein, HIV-I Integrase, Retroviral protease, HIV-I capsid protein) are known to be involved in the proliferation of the virus causing AIDS inside host cells, though each has a separate function.

The active site geometry of these proteins varies from a barrel shaped structure to an ellipsoidal one. The aim is to find non-peptide molecule(s) to fit into the given active sites. From the active site geometries of the molecules it is quite clear that any appropriate drug molecule must be flexible enough to bend in order to fit into protein target site. This flexible backbone is part of a *pharmacophore*. The essential part of the pharmacophore is found to have more or less the same structural arrangement of groups, i.e., they are made up of alkyl chains, which make them hydrophobic enough in order to fit into the hydrophobic core of the protein target sites.

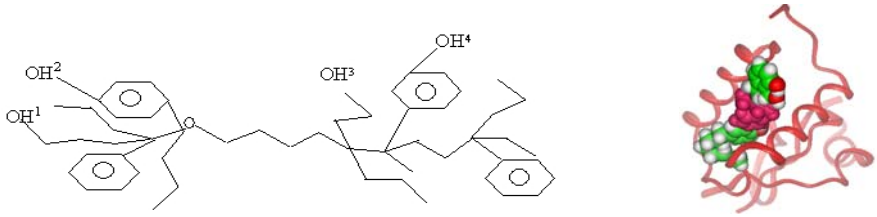
The C programming language on the UNIX platform is used to perform the task. The mutation rate is taken to be 0.01. One point crossover with a rate of 0.8 is used. Population size is taken to be equal to 200 and the algorithm is run for 500 generations. Results are taken for different possible positions of the pharmacophore within the active site, and the evolved ligand having the lowest energy value is taken as the solution.

Figures 4A-8A show the two-dimensional geometries of the ligands designed using the earlier method [16] for the five different proteins HIV-I Nef, HIV-I Integrase, Retroviral proteases, HIV-I capsid, and Rhinoviral protein respectively. Figures 4B-8B show the same molecules in a three-dimensional view (which is obtained using Insight II [MSI/Accelrys, San Diego, CA, USA]) appropriately docked into the active site of the corresponding proteins (where the docking is obtained using GOLD [22]). The target proteins are shown in ribbon form, while the ligand is represented in Corey-Pauling-Koltun (CPK) form, where the balls denote the atoms.

The two dimensional ligand molecules evolved using the proposed VGA based method are presented in Figures 4C-8C for the five proteins. Figures 4D-8D show the three-dimensional geometries of the protein-ligand docking for the corresponding five proteins respectively. As earlier, these are obtained using Insight II (MSI/Accelrys, San Diego, CA, USA) and GOLD. As can be seen from Figures 4B and 4D, the molecule designed for HIV-I Nef protein by VGA based method penetrates more inside the protein target. Note that the bulky phenyl rings present on the ligand obtained by GA based method (shown in Figure 4A) sterically hinder the docking of the ligand, and prevent it from penetrating inside the protein target. In contrast, the ligand designed using VGA (Figure 4C) is found to contain smaller number of such phenyl rings, and the hydrophobic C-skeleton of the ligand helps in its penetration through the hydrophobic core of the protein active site. Similar results were obtained in all other cases except for the ligand designed for HIV-I Integrase.

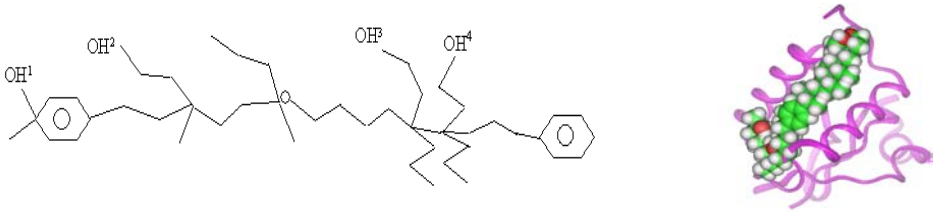
For the sake of comparison, the energy values of the ligands (obtained by the fixed GA based method and the proposed VGA based method) as well as those of the ligand-protein complexes are computed. These values are presented in Tables 1 and 2. The interacting residues of the ligand-protein complexes through the formation of Hydrogen bonds (H-bond) are also identified, along with the corresponding H-bond distances. These values are provided in Tables 3A–7A for the GA based method, and in Tables 3B-7B for the proposed

VGA based method. In these tables the entries are of the form X:Y:Z where X = either the ligand or the target protein (acting as either acceptor or donor depending on the column of the table where it appears), Y = the interacting residue number of X and Z = the atom at Y that is involved in the interaction.



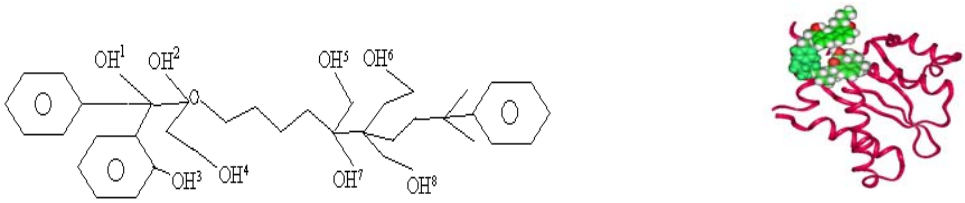
(A) (B)

Fig. 4 A & B. Using the GA based method for HIV-I Nef protein (A) Structure of the molecule (Note: Atoms capable of forming H-bonds are marked with numbers in superscript.) (B) Interaction of the ligand (represented in CPK) with the protein (represented as a red ribbon)



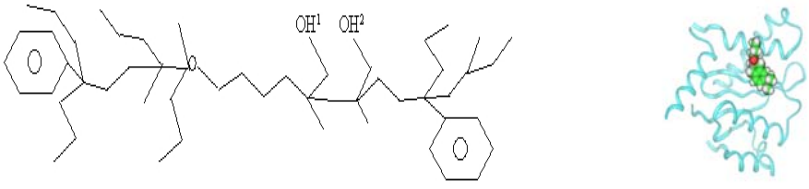
(C) (D)

Fig. 4 C & D. Using the VGA based method for HIV-I Nef protein (C) Structure of the molecule (Note: Atoms capable of forming H-bonds are marked with numbers in superscript.) (D) Interaction of the ligand (represented in CPK) with the protein (represented as a red ribbon)



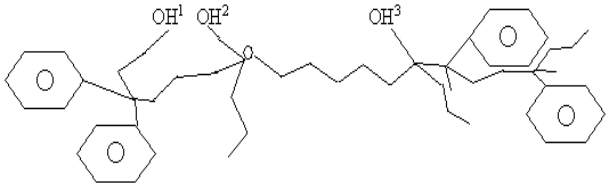
(A) (B)

Fig. 5 A & B. Using the GA based method for HIV-I Integrase (A) Structure of the molecule (Note: Atoms capable of forming H-bonds are marked with numbers in superscript.) (B) Interaction of the ligand (represented in CPK) with the protein (represented as a red ribbon)



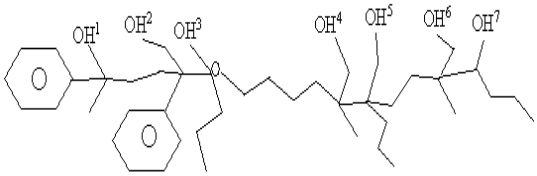
(C) (D)

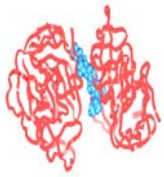
Fig. 5 C & D. Using the **V**GA based method for HIV-I Integrase (C) Structure of the molecule (Note: Atoms capable of forming H-bonds are marked with numbers in superscript.) (D) Interaction of the ligand (represented in CPK) with the protein (represented as a red ribbon)



(A) (B)

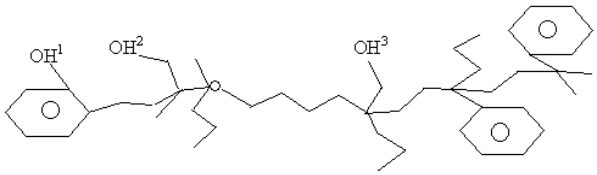
Fig. 6 A & B. Using the GA based method for retroviral protease (A) Structure of the molecule (Note: Atoms capable of forming H-bonds are marked with numbers in superscript.) (B) Interaction of the ligand (represented in CPK) with the protein (represented as a red ribbon)





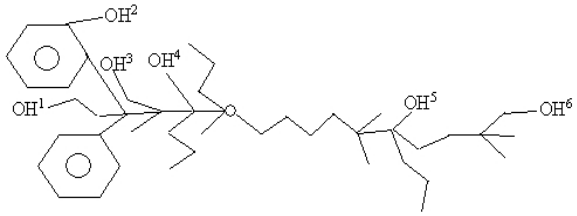
(C) (D)

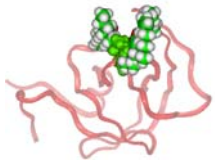
Fig. 6 C & D. Using the VGA based method for retroviral protease (C) Structure of themolecule (Note: Atoms capable of forming H-bonds are marked with numbers in superscript.) (D) Interaction of the ligand (represented in CPK) with the protein (represented as a red ribbon)



(A) (B)

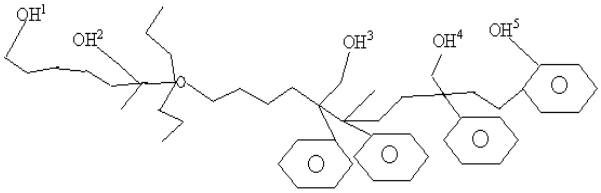
Fig. 7 A & B. Using the GA based method for HIV-I capsid protein (A) Structure of the molecule (Note: Atoms capable of forming H-bonds are marked with numbers in superscript.) (B) Interaction of the ligand (represented in CPK) with the protein (represented as a red ribbon)



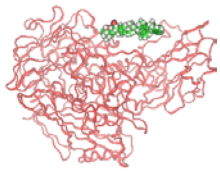


(C) (D)

Fig. 7 C & D. Using the VGA based method for HIV-I capsid protein (C) Structure of the molecule (Note: Atoms capable of forming H-bonds are marked with numbers in superscript.) (D) Interaction of the ligand (represented in CPK) with the protein (represented as a red ribbon)

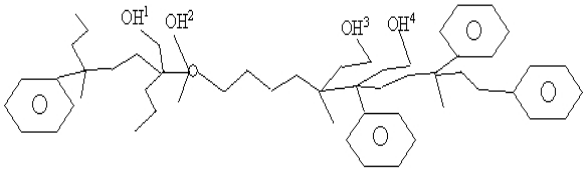


(A)

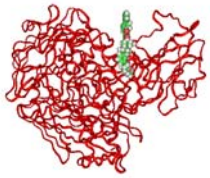


(B)

Fig. 8 A & B. Using the GA based method for Rhinoviral protein (A) Structure of the molecule (Note: Atoms capable of forming H-bonds are marked with numbers in superscript.) (B) Interaction of the ligand (represented in CPK) with the protein (represented as a red ribbon)



(C)



(D)

Fig. 8 C & D. Using the VGA based method for Rhinoviral protein (C) Structure of the molecule (Note: Atoms capable of forming H-bonds are marked with numbers in superscript.) (D) Interaction of the ligand (represented in CPK) with the protein (represented as a red ribbon)

Table 1: Energy Values of the ligands corresponding to the target receptor proteins in Kcal/mol

Process	Nef protein	Integrase	Retroviral Protease	Capsid protein	Rhinoviral protein
VGA	4.04	3.35	2.67	3.51	3.20
GA	4.56	2.45	3.12	4.55	3.64

Table 2: Interaction energies of the ligands with the protein targets in Kcal/mol

Process	Nef protein	Integrase	Retroviral Protease	Capsid protein	Rhinoviral protein
VGA	2.86	3.05	1.33	1.97	2.92
GA	3.92	1.68	2.89	2.61	3.11

Table 3A: H-Bonding interaction of the ligand as obtained by GA with HIV-I Nef protein

H-Bond donor	H-Bond acceptor	H-Bond distance
Protein: 85: N	Ligand: 1:OH	2.71
Protein: 86:N	Ligand: 2:OH	2.55
Protein: 89:ND1	Ligand: 3:OH	2.77
Protein: 89:NE2	Ligand: 4:OH	2.78

Table 3B: H-Bonding interaction of the ligand as obtained by VGA with HIV-I Nef protein

H-Bond donor	H-Bond acceptor	H-Bond distance
Protein: 141:NE1	Ligand: 1:OH	2.61
Protein: 141:NE1	Ligand: 2:OH	2.47
Protein: 142:N	Ligand: 3:OH	2.86
Protein: 184:N	Ligand: 4:OH	2.30
Ligand: 2:HH	Protein: 168:OD1	3.23

Table 4A: H-Bonding interaction of the ligand as obtained by GA with HIV-I Integrase

H-Bond donor	H-Bond acceptor	H-Bond distance
Ligand: 3:HH	Protein: A132:O	3.67
Protein: A136: NZ	Ligand: 1:OH	2.58
Protein: A132:NE1	Ligand: 2:OH	3.04
Protein: A103: N	Ligand: 3:OH	1.95
Ligand: 1:HH	Protein: A125:O	2.18
Protein: A127: N	Ligand: 6:OH	2.78
Protein: A128: N	Ligand: 5:OH	2.64
Protein: A129: N	Ligand: 4:OH	2.69
Protein: A130: N	Ligand: 7:OH	2.91

Table 4B: H-Bonding interaction of the ligand as obtained by VGA with HIV-I Integrase

H-Bond donor	H-Bond acceptor	H-Bond distance
Protein: A107: NE	Ligand: 1:OH	2.88
Ligand: 1:OH	Protein: A116:OD1	2.55
Protein: A107:NH2	Ligand: 2:OH	2.83

Table 5A: H-Bonding interaction of the ligand as obtained by GA with retroviral protease

H-Bond donor	H-Bond acceptor	H-Bond distance
Protein: D1: N	Ligand: 1:OH	2.65
Protein: E99: HN	Ligand: 2:OH	2.85
Ligand: 3:OH	Protein: E99: N	2.88

Table 5B: H-Bonding interaction of the ligand as obtained by VGA with retroviral protease

H-Bond donor	H-Bond acceptor	H-Bond distance
Ligand: 1:OH	Protein: E7:O	3.37
Ligand: 2:HH	Protein: E6:O	4.72
Protein: E6: HN	Ligand: 3:OH	3.94
Ligand: 4:HH	Protein: C4:O	2.56
Ligand: 5:HH	Protein: A48:O	3.02
Protein: A8:HH21	Ligand: 6:OH	2.44
Protein**:** A81: N	Ligand: 7:OH	2.46

Table 6A: H-Bonding interaction of the ligand as obtained by GA with HIV-I capsid protein

H-Bond donor	H-Bond acceptor	H-Bond distance
Ligand: 1:OH	Protein: 24:N	2.39
Ligand: 2:OH	Protein: 81:N	3.02
Ligand: 3:OH	Protein: 95:N	2.14

Table 6B: H-Bonding interaction of the ligand as obtained by VGA with HIV-I capsid protein

H-Bond donor	H-Bond acceptor	H-Bond distance
Ligand: 1:OH	Protein: A118: SD	3.12
Ligand: 2: OH	Protein: A48:OG1	4.32
Protein: A132:HH12	Ligand: 3:OH	2.14
Ligand: 4:OH	Protein: A128:O	3.29
Ligand: 5:OH	Protein: A129: N	2.69
Ligand: 6:OH	Protein: A41:OG	2.28

Table 7A: H-Bonding interaction of the ligand as obtained by GA with Rhinoviral protein

H-Bond donor	H-Bond acceptor	H-Bond distance
Protein : B1439: N	Ligand: 1:OH	2.81
Protein: B1460: N	Ligand: 2:OH	2.86
Protein: B1461: N	Ligand: 3:OH	3.30
Protein: B1440: N	Ligand: 4:OH	3.31

Table 7B: H-Bonding interaction of the ligand as obtained by VGA with Rhinoviral protein

H-Bond donor	H-Bond acceptor	H-Bond distance
Protein : B1381: NZ	Ligand: 1:OH	2.70
Protein: B1434: N	Ligand: 2:OH	2.99
Protein: B1481: N	Ligand: 3:OH	2.94
Protein: B1438: ND1	Ligand: 4:OH	2.83

It is evident from Tables 1 and 2 that for HIV-I Nef protein, the energy values of the ligand as also the ligand-protein interaction energies are lower when VGA is using as compared to the case when GA is utilized as the underlying search tool, thereby indicating better binding in the former case. Tables 3A (GA) and 3B (VGA) show the interacting residues corresponding to this protein, where again we find that the number of such interacting residues is larger for VGA. Note that due to the influence of nearby electron withdrawing -OH group (OH^2 in Figure 4C) and the phenyl ring the reactivity of the -OH group at position 1 (OH^1 in Figure 4C) is enhanced. This makes it function both as H-bond donor as well as H-bond acceptor.

Similar results are obtained for the other proteins where the VGA based method gives better result as compared to the GA based method. The only exception is in the case of HIV–I Integrase for which the GA based method gives a better result that the VGA based method. One possible reason for this may be that the active site geometry of this target protein and the fixed tree structure assumed in [16] is compatible, so that use of VGA does not provide much advantage. In fact, in such a situation, the GA based method is expected to perform better, since its search space is restricted to only the fixed sized trees. In contrast, the search space of VGA is much larger since the size of the tree is kept variable. Therefore, further execution of the VGA is likely to provide better result.

In order to investigate about the synthesizability of the designed molecules, Cambridge Crystallographic Datacentre (CSD) [31] was browsed to identify whether there are any existing crystal structures structurally similar to the evolved molecules. CSD codes of the molecules that are found to be structurally similar to the molecules evolved using GA and VGA are listed in Table 8 along with their corresponding energy values. As evident, except for HIV-I Integrase, in all other cases, the energy values of the molecules obtained from CSD that are similar to the VGA based molecules are significantly lower than those that are similar to the GA based molecules. This indicates that in general the molecules that are structurally similar to those obtained by VGA based method are more stable than those obtained by GA based method.

The algorithms were executed ten times with different initial populations. It was found that the GA based method sometimes got stuck at very poor configurations. In contrast, the VGA based method provided similar sort of results in the different runs. As an illustration, for the HIV-I Nef protein, the GA based method provided the best energy value in 6 out of the ten runs, while in the other four runs, it got stuck a extremely poor values. In contrast, although the VGA based method provided best energy value only once, in nine of the ten runs it

provided energy values that were better than or equal to that provided by the GA based method. In only a single run, it provided a worse value. This indicates that the VGA based method is likely to provide a reasonably good value more often. It was found for the VGA based method that although the different runs provided similar energy values, the structures of the ligands were different. As an illustration consider the right subtree in Figure 4C and Figure 9, obtained in two different runs of the VGA for the HIV-I Nef protein. Both these configurations have similar energy values but different tree depths and structures. This indicates that no unique structure is best for a given active site, thus strengthening the requirement of using variable length representation in GAs.

Table 8: Comparison of the energy values of similar molecules obtained from the CSD.

Name of the protein	Method used	CSD code of the molecule	Energy in Kcal / mole
HIV-I Nef	GA	ACOYUM	-24.29
	VGA	ALITOL	-41.87
HIV-I Integrase	GA	DISWEH	-2.55
	VGA	ACEMAW	1.71
Retroviral protease	GA	ANMDXL	51.03
	VGA	ALEUAC10	-11.9
HIV-I capsid	GA	QERSEL	3.7
	VGA	EDEYUH	-0.99
Rhinoviral protein	GA	ACBUET	-20.56
	VGA	DIQLEU	-83.68

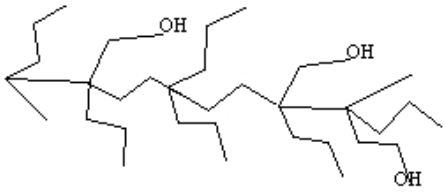


Fig. 9. Right tree of the ligand obtained in a run of the VGA for HIV-I Nef protein.

6 Discussion and Conclusions

An algorithm for the design of a drug like molecule that can bind to the active site of a protein target receptor, there by preventing the proliferation of the microorganism for which the protein is a vital factor, has been proposed in this article. Variable string length GA has been used as the underlying search and optimization tool. Unlike the previous GA based approach [16], the proposed method makes no assumption regarding the size of the tree formed on both sides of the pharmacophore. In order to tackle the problem of variable sized trees, the length of a chromosome in GA is now allowed to vary. Modified crossover and mutation operators are used in this context. Once the two dimensional geometry of the ligand is evolved using this approach, Insight II (MSI/Accelrys, San Diego, CA, USA) and GOLD

are used for determining the possible three-dimensional configuration and the ligand-protein docking respectively.

It is found that the ligand molecules designed using the proposed approach are, in general, associated with lower van der Waals energy values as compared to the fixed string length GA based method. Moreover, it is found that the structure of the evolved molecule is, in general, such that it is amenable to stable configurations because of the presence of hydrogen bonds. The running times of GA and VGA were found to be comparable for the fixed number of iterations.

The actual conformation of a molecule depends not only on the bond lengths and functional groups, but also on non-covalent intermolecular forces, such as electrostatic interaction, hydrogen bond formation, between the drug and the receptor. These factors have not yet been considered in this model. Moreover, here only seven groups are considered. As a scope for future study more groups as well as the energy contributions due to the other interactions can be taken into account. Another area of further research in this regard is to extend the model to three dimensions and use some other representation (e.g., graph representation [20]) to encode the ligand. It is to be noted that in this study we assume that the active site geometries of the receptors are known. Another attempt may be made in the lines of the work done in [32] such that a representation of the active site may be produced using 3DQSAR [1,2] studies as a first step to the ligand design problem.

References:

1. Ghosal N, Mukherjee PK, 3-D QSAR of N-substituted 4-amino-3,3-dialkyl-2(3H)-furanone GABA Receptor Modulators Using Molecular Field Analysis and Receptor Surface Modeling Study, *Bioinorganic & Medical Chemistry Letters* **14**:103-109, 2004
2. Nicolotti O, Gillet VJ, Fleming PJ, Green DVS, Multiobjective Optimisation in Quantitative Structure-Activity Relationships: Deriving Accurate and Interpretable QSARs, *J Med Chem* **45**:5069-5080, 2002
3. Goldberg DE, *Genetic Algorithms in Search, Optimization and Machine Learning*, Addison-Wesley, New York, 1989
4. Davis L, *Handbook of Genetic Algorithms*, New York:Van Nostrand Reinhold, 1991
5. De Jong K, *Learning with Genetic Algorithms: An Overview, Machine Learning 3.* Norwell, MA: Kluwer, pp. 121 – 138, 1988
6. Michalewicz Z, *Genetic Algorithms + Data Structure = Evolution Programs*, Berlin, Germany: Springer-Verlag, 1992
7. Filho JLR, Treleavan PC, Genetic Algorithm Programming Environments*, IEEE Comput*, 28-43, 1994
8. Pal SK, Bhandari D, Selection of Optimum Set of Weights in a Layered Network Using Genetic Algorithm, *Inf. Sci*, **80**: 213-234, 1994
9. Pal SK, Bhandari D, Kundu MK, Genetic Algorithms for Optimal Image Enhancement, *Pattern Recognit. Lett*, **15** : 261-271, 1994
10. Steffen Schulze-Kremer (ed.), *Advances in Molecular Bioinformatics* 258, IOS Press, 1994
11. Leach AR, Gillet VJ, *An Introduction to Chemoinformatic*, Kluwer Academic Publishers, 2003
12. Raidl GR. (ed.), *Applications of Evolutionary Computing: EvoWorkshop 2003: Proceedings of the EvoBIO, EvoCOP, EvoIASP, EvoMUSART, EvoROB, and EvoSTIM*, Essex, UK, April 14-16, 2003, 2611, Lecture Notes in Computer Science Series, 2003
13.Clark DE. (ed), *Evolutionary Algorithms in Molecular Design*, Jown Wiley, 2000.
14. Fogel GB, Corne DW (eds.), *Evolutionary Computation in Bioinformatics*, Morgan Kaufmannm, 2002
15. Nicolotti O, Gillet VJ, Fleming PJ, GreenDVS, Multiobjective Optimisation in Quantitative Structure-Activity Relationships: Deriving Accurate and Interpretable QSARs, *J. Med. Chem* **45**:5069-5080, 2002

16. Goh G, Foster JA, *Evolving Molecules for Drug Design Using Genetic Algorithm, Proc. Int. Conf. on Genetic & Evol. Computing,* Morgan Kaufmann, 27 – 33, 2000
17. Güner OF (ed.), *Pharmacophore Perception, Development and Use in Drug Design*, International University Line:La Jolla, CA. 2000
18. Pegg SC, Haresco JJ, Kuntz ID, A Genetic Algorithm for Structure-based De Novo Design, *J Comput Aided Mol Des***15**: 911-933, 2001
19. Kamphausen S, Holtge N, Wirsching F, Morys-Wortmann C, Riester D, Goetz R, Thurk M, Schwienhorst A, Genetic Algorithm for the Design of Molecules with Desired Properties, *J Comput Aided Mol Des*: **16** 551-567, 2002
20. Globus A, Lawton J, Wipke T, *Automatic Molecular Design Using Evolutionary Techniques, Sixth Foresight Conference on Molecular Nanotechnology*, Sunnyvale, California, November, 1998 and Nanotechnology, Volume 10, Number 3, September 1999, pp. 290-299
21. Bandyopadhyay S, Murthy CA, Pal SK, VGA–Classifier: Design and Application, *IEEE Transaction on Systems, Man, and Cybernetics-Part B: Cybernetics* **30**: 890-895, 2000
22. Jones G, Willett P, Glen RC, Leach AR, Taylor RJ, Development and Validation of a Genetic Algorithm for Flexible Docking, *J Mol Biol* **267**: 727-748, 1997
23. Maulik U, Bandyopadhyay S, Genetic Algorithm-based Clustering Technique, *Pattern Recognition* **33**: 1455-1465, 2000
24. Levine.I, *Physical Chemistry*, McGraw Hill, New York, 1988
25. Leach AR, *Molecular Modeling Principles and Applications*, Pearson, Prentice Hall, 2001
26. Arold.S, Franken P, Strub. MP, Hoh F, Benichou S, Dumas R, The Crystal Structure of HIV-1 Nef Protein Bound to the Fyn Kinase SH3 Domain Suggests a Role for this Complex in Altered T Cell Receptor Signaling, *Structure* **5** 1361-1372, 1997
27. Hinck T, Wang AP, Nicholson YX, Torchia LK, Wingfield DA, Stahl P, Chang SJ, Domaillel CH, Lam PJ,Three-dimensional Solution Structure of the HIV-1 Protease Complexed with DMP323, a Novel Cyclic Urea-type Inhibitor, Determined by Nuclear Magnetic Resonance Spectroscopy, *Protein Sci* **5** : 495-510, 1996
28. Rose RB, Craik CS, Stroud RM, Domain Flexibility in Retroviral Proteases: Structural Implications for Drug Resistant Mutations*, Biochemistry* **37**: 2607-2617, 1998
29. Worthilake DK, Wang H, Soo Y, Sundquist WI, Hill CP, Structures of the HIV-1 Capsid Protein Dimerization Domain at 2.6 A Resolution, *Acta Crystallogr D Biol Crystallogr* **55**: 85-95, 1999
30. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE, The Protein Data Bank, *Nucleic Acids Res*, **28**: 235-242, 2000
31. Allen FH, The Cambridge structural database: a quarter of a million crystal structures and rising, *Acta Crystallogr. B*, **58**: 283-438, 2002
32. Greenidge PA, Merette SA, Beck R, Dodson G, Goodwin CA, Scully MF, Spencer J, Weiser J, Deadman JJ, Generation of Ligand Conformations in Continuum Solvent Consistent with Protein Active Site Topology: Application to Thrombin, *J. Med Chem* **46**: 1293-1305, 2003