LigEvolver: A Tool for ligand formulation using genetic algorithm

Priyadarsini Soundararajan

Follow this and additional works at: https://scholarworks.rit.edu/theses

Recommended Citation

This Thesis is brought to you for free and open access by RIT Scholar Works. It has been accepted for inclusion in Theses by an authorized administrator of RIT Scholar Works. For more information, please contact ritscholarworks@rit.edu.
LigEvolver: A Tool for Ligand Formulation Using Genetic Algorithm

Approved: Paul A. Craig
Thesis Advisor

G. R. Skuse
Director of Bioinformatics or
Head, Department of Biological Sciences

Submitted in partial fulfillment of the requirements for the Master of Science degree in Bioinformatics at the Rochester Institute of Technology.

Priyadarsini Soundararajan
May 2006
Title of thesis or dissertation: LIGEVOLVER: A TOOL FOR LIGAND FORMULATION USING GENETIC ALGORITHM

Name of author: PRIYADARSINI SOUNDIRARATAN
Degree: MASTER OF SCIENCE
Program: BIOINFORMATICS
College: SCIENCE

I understand that I must submit a print copy of my thesis or dissertation to the RIT Archives, per current RIT guidelines for the completion of my degree. I hereby grant to the Rochester Institute of Technology and its agents the non-exclusive license to archive and make accessible my thesis or dissertation in whole or in part in all forms of media in perpetuity. I retain all other ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Print Reproduction Permission Granted:

I, PRIYADARSINI SOUNDIRARATAN, hereby grant permission to the Rochester Institute Technology to reproduce my print thesis or dissertation in whole or in part. Any reproduction will not be for commercial use or profit.

Signature of Author: Priyadarsini Date: 11/24/06

Print Reproduction Permission Denied:

I, __________________________________, hereby deny permission to the RIT Library of the Rochester Institute of Technology to reproduce my print thesis or dissertation in whole or in part.

Signature of Author: __________________ Date: __________

Inclusion in the RIT Digital Media Library Electronic Thesis & Dissertation (ETD) Archive

I, __________________________________, additionally grant to the Rochester Institute of Technology Digital Media Library (RIT DML) the non-exclusive license to archive and provide electronic access to my thesis or dissertation in whole or in part in all forms of media in perpetuity.

I understand that my work, in addition to its bibliographic record and abstract, will be available to the world-wide community of scholars and researchers through the RIT DML. I retain all other ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation. I am aware that the Rochester Institute of Technology does not require registration of copyright for ETDs.

I hereby certify that, if appropriate, I have obtained and attached written permission statements from the owners of each third party copyrighted matter to be included in my thesis or dissertation. I certify that the version I submitted is the same as that approved by my committee.

Signature of Author: __________________ Date: __________
Thesis Advisory Committee

Committee Chair

Dr. Paul Craig
Professor of Chemistry
Department of Chemistry
College of Science
Rochester Institute of Technology

Committee Member

Professor Al Biles
Professor and Undergraduate Program Coordinator
Information Technology Department
Golisano College of Computing and Information Sciences
Rochester Institute of Technology

Committee Member

Dr. Anne Haake
Associate Professor
Information Technology Department
Golisano College of Computing and Information Sciences
Rochester Institute of Technology
To The Immune System
Abstract

Human Immunodeficiency Virus type 1 (HIV – 1) has been found to be the cause of Acquired Immune Deficiency Syndrome (AIDS). This virus caused uproar during the 1980's and started a major drug discovery race among the pharmaceutical companies. It also proved to be a milestone in the establishment of the use of computers in rational drug design.

Evolutionary computation is a commonly used computational approach that has been successfully applied to a variety of fields ranging from engineering to life science. The main reason for its effectiveness is that it is driven by the principles of evolution. A genetic algorithm is an approach to evolutionary computation that allows random combination of data to occur in a series of generations and enables the identification of novel systems that might otherwise have gone undetected.

This work explores the use of genetic algorithms to generate new ligand structures that may be effective in inhibiting HIV – 1 Protease, one of the major drug targets in HIV. One of the computational challenges associated with drug discovery is the conversion of chemical and biological entities into formats that the computer can use. Chemical structures can be represented by linear character strings called SMILES strings. SMILES (Simplified Molecular Input Line Entry Specification) strings are taken as the genetic representation for our approach to drug discovery, and a genetic algorithm has been developed to generate appropriate ligands for HIV-1 protease. Based on a fitness function,
the ligands are evaluated and either kept or removed from the gene pool following the 
"survival of the fittest" pattern found in nature.
Acknowledgements

I would like to thank Dr. Paul Craig, committee chairman and advisor for his unrelenting guidance and encouragement. His patience and enthusiastic supervision has been the driving force behind the successful completion of the project. I am grateful to Prof. Al Biles for his zeal and support. The numerous technical discussions with him enabled me to approach the problem in various angles and his advices towards the computational aspect of the research were useful. I would like to offer my sincere gratitude to Dr. Anne Haake for her assistance and interest. Dr. Haake had been a source of support and encouragement at all times. I would like to thank all the committee members for trusting me and being a part of this thesis.

I would like to extend my thankfulness to Mr. Srinivas Naidu, my uncle and a constant source of moral support, without whom this entire venture would have been impossible.
Table of Contents

1. Introduction ...................................................................................................................... 1
  1.1 Fundamental Concepts ............................................................................................... 1
  1.2 Purpose of Investigation ............................................................................................ 2
  1.3 Problem Description ................................................................................................. 3
  1.4 Rationale .................................................................................................................. 4

2. Methods and Materials .................................................................................................... 5
  2.1 Disease and Enzyme Selection .................................................................................... 5
    2.1.1 HIV Genome ......................................................................................................... 5
    2.1.2 Lifecycle of HIV .................................................................................................. 6
    2.1.3 HIV Protease ....................................................................................................... 8
    2.1.4 Conserved Region ............................................................................................... 9
    2.1.5 Enzymatic Reaction ............................................................................................ 10
  2.2 Genetic Representation in LigEvolver ......................................................................... 12
    2.2.1 Chromosome Representation ............................................................................... 12
    2.2.2 Sample SMILES string and 2D structure of ligand .............................................. 13
    2.2.3 Pharmacophoric Features Considered .................................................................. 14
  2.3 Input .......................................................................................................................... 15
    2.3.1 XML/Perl Parser .................................................................................................. 16
    2.3.2 Protein Data Bank ............................................................................................... 17
  2.4 Output ....................................................................................................................... 17
  2.5 External Packages ...................................................................................................... 18
  2.6 Design of LigEvolver ............................................................................................... 18
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.6.1 Genetic Algorithm</td>
<td>19</td>
</tr>
<tr>
<td>2.6.2 Pharmacophore Elucidator</td>
<td>27</td>
</tr>
<tr>
<td>2.6.3 Molecular Weight Calculator</td>
<td>32</td>
</tr>
<tr>
<td>2.6.4 Van Der Waals Volume</td>
<td>33</td>
</tr>
<tr>
<td>2.6.5 Polar Surface Area</td>
<td>35</td>
</tr>
<tr>
<td>2.6.6 Motif Finder</td>
<td>35</td>
</tr>
<tr>
<td>2.6.7 Fitness Comptor</td>
<td>36</td>
</tr>
</tbody>
</table>

3. Results                                                            41

3.1 Simulation I                                                      41

3.1.1 Significant Results                                            41

3.2 Simulation II                                                     46

3.2.1 Significant Results                                            46

3.3 Simulation III                                                    52

3.3.1 Significant Results                                            52

3.4 Simulation IV                                                     58

3.4.1 Significant Results                                            58

3.5 Simulation V                                                      61

3.5.1 Significant Results                                            61

3.6 Simulation VI                                                     64

3.6.1 Significant Results                                            65

3.7 Fitness Behavior                                                 66

4. Discussion                                                         67
List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1.1.1</td>
<td>HIV Genome</td>
<td>5</td>
</tr>
<tr>
<td>2.1.2.1</td>
<td>HIV Lifecycle</td>
<td>7</td>
</tr>
<tr>
<td>2.1.3.1</td>
<td>HIV Protease and active site</td>
<td>8</td>
</tr>
<tr>
<td>2.1.3.2</td>
<td>HIV Protease inhibited by 1D4Y</td>
<td>8</td>
</tr>
<tr>
<td>2.1.4.1</td>
<td>Asparatic Acid</td>
<td>9</td>
</tr>
<tr>
<td>2.1.4.2</td>
<td>Threonine</td>
<td>9</td>
</tr>
<tr>
<td>2.1.4.3</td>
<td>Glycine</td>
<td>9</td>
</tr>
<tr>
<td>2.1.4.4</td>
<td>Catalytic triad</td>
<td>10</td>
</tr>
<tr>
<td>2.1.5.1</td>
<td>HIV Protease cleaving reaction</td>
<td>11</td>
</tr>
<tr>
<td>2.2.1.1</td>
<td>Transformation from 3D structure to SMILES string</td>
<td>12</td>
</tr>
<tr>
<td>2.2.2.2</td>
<td>TPV Structure</td>
<td>14</td>
</tr>
<tr>
<td>2.2.3.1</td>
<td>TPV Structure with labeled pharmacophoric features</td>
<td>15</td>
</tr>
<tr>
<td>2.3.1</td>
<td>Input Construction</td>
<td>17</td>
</tr>
<tr>
<td>2.6.1.1</td>
<td>Components of LigEvolver</td>
<td>19</td>
</tr>
<tr>
<td>2.6.1.1.1</td>
<td>Flow chart of GA</td>
<td>20</td>
</tr>
<tr>
<td>2.6.1.1.1.1</td>
<td>Example of GenFrag</td>
<td>22</td>
</tr>
<tr>
<td>2.6.1.1.2.1</td>
<td>Example of Crossover</td>
<td>23</td>
</tr>
<tr>
<td>2.6.1.1.3</td>
<td>Mutation rates and genetic adaptability</td>
<td>24</td>
</tr>
<tr>
<td>2.6.1.2.1</td>
<td>Tournament Selection</td>
<td>26</td>
</tr>
<tr>
<td>2.6.2.1.1</td>
<td>Aromatic Structure I – A Benzene ring</td>
<td>27</td>
</tr>
<tr>
<td>2.6.2.2.1</td>
<td>Hydrogen Bond Acceptor Structure I</td>
<td>28</td>
</tr>
<tr>
<td>2.6.2.2.2</td>
<td>Hydrogen Bond Acceptor Structure II</td>
<td>28</td>
</tr>
</tbody>
</table>
Figure 2.6.2.2.3 Hydrogen Bond Acceptor Structure III
Figure 2.6.2.2.4 Hydrogen Bond Acceptor Structure IV
Figure 2.6.2.3.1 Hydrogen Bond A/D Structure I
Figure 2.6.2.3.2 Hydrogen Bond A/D Structure II
Figure 2.6.2.3.3 Hydrogen Bond A/D Structure III
Figure 2.6.2.4.1 Hydrophobic Structure I
Figure 2.6.2.4.2 Hydrophobic Structure II
Figure 2.6.2.4.3 Hydrophobic Structure III
Figure 2.6.2.4.4 Hydrophobic Structure IV
Figure 2.6.2.4.5 Hydrophobic Structure V
Figure 2.6.2.4.6 Hydrophobic Structure VI
Figure 2.6.2.4.7 Hydrophobic Structure VII
Figure 3.1.1.1.1.1 Simulation I – Generation 100 – Compound I
Figure 3.1.1.1.1.2 Simulation I – Generation 100 – Compound II
Figure 3.1.1.2.1.1 Simulation I – Generation 85 – Compound I
Figure 3.1.1.2.2.1 Simulation I – Generation 85 – Compound II
Figure 3.2.1.1.1.1 Simulation II – Generation 40 – Compound I
Figure 3.2.1.1.2.1 Simulation II – Generation 40 – Compound II
Figure 3.2.1.1.3.1 Simulation II – Generation 40 – Compound III
Figure 3.2.1.1.4.1 Simulation II – Generation 40 – Compound IV
Figure 3.2.1.2.1.1 Simulation II – Generation 41 – Compound I
Figure 3.3.1.1.1.1 Simulation III – Generation 100 - Compound I (a)
Figure 3.3.1.1.1.2 Simulation III – Generation 100 - Compound I (b)
List of Tables

Table 2.2.2.1 Ligand Summary 13
Table 2.3.1 List of Databases 16
Table 2.6.1.1.3.1 Amino Acids and their SMILES string 25
Table 2.6.3.1 Molecular weights 32
Table 2.6.4.1 Atoms and their Volumes 34
Table 2.6.6.1 FDA – Approved Protease Inhibitors 35
Table 2.6.7.1.1 Aromatic Ring Count Scoring Scheme 37
Table 2.6.7.2.1 Hydrogen Bond Donor Count Scoring Scheme 38
Table 2.6.7.3.1 Hydrogen Bond Acceptors Count Scoring Scheme 38
Table 2.6.7.4.1 Hydrophobic Regions Count Scoring Scheme 38
Table 2.6.7.5.1 Molecular Weight Scoring Scheme 39
Table 2.6.7.6.1 Polar Surface Area Scoring Scheme 39
Table 2.6.7.7.1 Van Der Waals Volume Scoring Scheme 40
Table 3.1.1 Simulation I Properties 41
Table 3.2.1 Simulation II Properties 46
Table 3.3.1 Simulation III Properties 52
Table 3.4.1 Simulation IV Properties 58
Table 3.5.1 Simulation V Properties 61
Table 3.6.1 Simulation VI Properties 64
Table 4.1 LigEvolver compounds with corresponding input databases 71
1. Introduction

LigEvolver, the tool that produces novel ligand structures, has been implemented for specific reasons. These intentions and the way this tool is different from previous genetic algorithm based tools is presented in this section. The necessary background and the problem are also described.

1.1 Fundamental Concepts

Enzymes.

Enzymes are a specialized class of proteins that act as catalysts to increase the rate of chemical reactions.

Inhibitor

The inhibitors are chemical compounds that serve to block enzymes from functioning.

Ligands

The substrate that is bound to by a protein is also called a ligand. These can be used as effective inhibitors. The terms ligand and inhibitor are used interchangeably in this thesis.

Genetic Algorithm

Genetic Algorithms employ concepts of evolution to search for solutions in domain problem spaces. The individuals in a population are made to crossover and mutate to give rise to new individuals. Depending on the ability of individuals to satisfy a fitness function,
a pre-defined quality criterion, they are either allowed to exist or eliminated from the next generation.

Pharmacophore

This is a common template that depicts the desired spatial arrangement of chemical properties of a potential candidate. It is the minimal structure required for inhibition.

SMILES

This is an acronym for Simplified Molecular Input Line Entry Specification. It is a method of representing a molecule in a linear form.

1.2 Purpose of Investigation

Inhibitor design is a major part of the drug discovery process. Structure-based drug design has been shown to produce effective results and has led to the discovery of a number of inhibitors. There are plenty of features that lead to the design of a reliable drug such as its bioavailability, binding affinity and interaction specificity, among other features.

In the case of enzymes with known inhibitors, it would be worthwhile to explore the possibility of generating ligands based on scalar properties without other added complexities. The focus of this study is directed towards determining if new ligand structures could be derived by using a minimal set of physiochemical properties and comparisons with known inhibitors.
Previous studies such as those by Glen et al., (1994) and Venkatasubramanian et al., (1995) have shown the use of evolutionary techniques for drug design. Douguet et al made use of SMILES line notation for their program LEA (Ligand by Evolutionary Algorithm). LigEvolver differs from the past research as it makes use of the smallest number of constraints and genetic operators for evolving ligands. Its main objective is to see if this limited approach can lead to powerful structures.

1.3 Problem Description

There are many undiscovered therapeutic structures that have the ability to function as inhibitors. The goal of this study is to uncover such structures by constructing a genetic algorithm based computer tool, LigEvolver. The task is to identify new inhibitors for the HIV – 1 Protease enzyme and LigEvolver aides this task by evolving solutions that seem more likely to perform the task. The level of analysis of the chemical structures has been restricted to two dimensions because of the minimalist hypothesis.

Given a set of chemical structures in the SMILES notation, LigEvolver evolves potential inhibitors for the target enzyme. The algorithm conserves the fittest individuals in each generation and iterates until fairly competent candidates have been produced. Each individual is assessed based on pharmaceutically relevant criteria.
1.4 Rationale

Discovering new drugs is an extremely tedious and expensive task. Computational derivation of potential ligand structures would help to reduce the time and the cost of drug development. The use of a genetic algorithm to generate candidate structures could give rise to a variety of combinations that would enable the scientist to pursue lead compounds that have not been identified using traditional approaches.

The intention is to develop a program using a well-characterized system of an enzyme (HIV protease) that has been crystallized in the presence of a large number on inhibitors. In this case, the 3D structures of the drugs and the target protein are known. Once an approach to predicting new ligands based on existing ligands has been developed, it can be extended to open cases, where a number of drugs have been identified, but the structure of the target enzyme and its identity are not known.
2. Methods and Materials

2.1 Disease and Enzyme Selection

2.1.1 HIV Genome

The HIV genome consists of three main genes, namely gag, pol and env. Gag codes for structural proteins that form the viral core. Pol genes code for Reverse Transcriptase, Integrase and Protease, which are three essential enzymes that play a significant role in the formation and maturation of the virus. The env gene codes the envelope for the viral proteins [39]. Figure 2.1.1.1 shows the HIV genome encompassing the three genes.
2.1.2 Lifecycle of HIV

Human immunodeficiency virus (HIV) is a lethal retrovirus that attacks the immune system. This retrovirus follows a specific life – cycle to infect the host cells. The viral activity has been outlined in fig. 2.1.2.1.
BINDING

The HIV binds itself to the CD4 receptor in the cell membrane

FUSION

The virus penetrates the membrane and enters the cell with the help of the co-receptor

TRANSCRIPTION

Reverse Transcriptase transcribes the viral RNA into DNA

INTEGRATION

Integrase integrates the viral DNA with the host cell's DNA

CLEAVING

Protease cleaves the viral polyproteins into functional proteins and the virus exits the cell

Figure 2.1.2.1 HIV Lifecycle


2.1.3 HIV Protease

As can be seen from the lifecycle, HIV Protease is a crucial enzyme that hydrolyzes viral polyproteins into functional protein products. These products then contribute to the further propagate the virus. This work concentrates on identifying new ligands for HIV Protease. The flexibility and mutability of this enzyme makes it a challenging drug design target. HIV protease acts as a dimer with two identical polypeptide chains. It has only one active site that is depicted in the boxed region of fig. 2.1.3.1 [40].

![Figure 2.1.3.1 HIV Protease and active site](image1)

![Figure 2.1.3.2 HIV Protease inhibited by 1D4Y*](image2)
2.1.4 Conserved Region

During the transcription of RNA to DNA, Reverse Transcriptase has an error rate of about 1 in 2000 bases. Due to this the HIV strain undergoes mutation in each generation and the nucleotide sequence of HIV Protease changes from strain to strain. However, a conserved region has been detected. It is the residues, Asp – Thr – Gly, of the catalytic triad.

*Figure generated using DeepView

*Figure 2.1.4.1 Aspartic Acid (Asp)*

*Figure 2.1.4.2 Threonine (Thr)*

*Figure 2.1.4.3 Glysine (Gly)*

*Structures have been taken from Wikipedia [38]*
2.1.5 Enzymatic Reaction

The HIV protease enzyme cleaves the protein at the carbonyl group of the peptide bond. The entire process is shown with respect to the active site (Asp – Thr – Gly) in figure 2.1.5.1 [20]. Step 1 shows the binding of HIV – 1 protease to a peptide substrate; step 2 demonstrates the attack of the carbonyl of the scissile amide by HIV – 1 protease activated water; step 3 illustrates the cleavage of the scissile bond while step 4 depicts the final product.
Figure 2.1.5.1 HIV Protease Catalytic Mechanism
2.2 Genetic Representation in LigEvolver

2.2.1 Chromosome Representation

"Chromosome" is the general term used to address the genotype of the individuals in a population. The SMILES string serves as the input to the genetic program. Its simplified representation supplies the computer with an accurate linear representation of complicated two-dimensional structures, which can be readily manipulated during evolutionary programming [4]. It is also a standard and universal form of representation used by chemists all over the world [1]. These reasons make it a good choice for the depiction of the chemical compounds and a means to represent the chromosome for the genetic program. The mapping of the chemical compound from its 3D structure to the SMILES notation is shown in figure 2.2.1.1.
2.2.2 Sample SMILES string and 2D structure of ligand

An example of the ligand structure is shown below along with the SMILES string. The common name is Tipranavir, which is abbreviated TPV.

<table>
<thead>
<tr>
<th>Name</th>
<th>N-(3-[(1R)-1-[(6R)-4-HYDROXY-2-OXO-6-PHENETHYL-6-PROPYL-5,6-DIHYDRO-2H-PYRAN-3-YL]PROPYL]PHENYL)-5-(TRIFLUOROMETHYL)-2-PYRIDINESULFONAMIDE</th>
</tr>
</thead>
<tbody>
<tr>
<td>HET ID</td>
<td>TPV</td>
</tr>
<tr>
<td>Formula</td>
<td>C_{31}H_{33}N_{2}O_{5}F_{3}S</td>
</tr>
<tr>
<td>SMILES String</td>
<td>CCCC2(CCc1cccc1)CC(O)=C(C(CC)c3cccc(NS(=O)(=O)c4ccc(cn4)C(F)(F)c3)C(=O)O2</td>
</tr>
</tbody>
</table>

Table 2.2.2.1 Ligand Summary
2.2.3 Pharmacophoric Features Considered

For each of the ligand structures the following are the important features considered to form the basic pharmacophore template. These features are thought to be important interaction points between the enzyme and the ligand.

1. Aromatic Rings [Ar]
2. Hydrogen Bond Acceptors [A]
3. Hydrogen Bond Acceptors and Donors [A/D]
4. Hydrophobic center [F]

Figure 2.2.3.1 shows how the pharmacophoric features of ligand TPV are decomposed.
2.3 Input

Before beginning the genetic algorithm, certain pre-processing steps had to be completed for the construction of the input files. SMILES databases, consisting of various chemical compounds represented as SMILES strings, were assembled. The data for these databases were extracted from various public databases. The public databases are listed in the following table.
<table>
<thead>
<tr>
<th>List of Databases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kyoto Encyclopedia of Genes and Genomes (KEGG)</td>
</tr>
<tr>
<td>ChemPDB</td>
</tr>
<tr>
<td>ChemBank</td>
</tr>
<tr>
<td>Macromolecular Structure Database (MSD)</td>
</tr>
<tr>
<td>Anti – HIV National Cancer Institute (NCI)</td>
</tr>
<tr>
<td>Protein Data Bank (PDB)</td>
</tr>
</tbody>
</table>

Table 2.3.1 List of Databases

The above mentioned databases have been downloaded from Ligand.Info which is a collection of various small-molecule databases [27].

2.3.1 XML/Perl Parser

An XML parser was constructed to elucidate the data from the databases and to produce the output in a format easily transferable to the database. In other words, the files from the public databases contain information regarding a ligand such as its molecular weight, geometrical coordinates, etc. The parser scans the dump files for the token “SMILES” and extracts only the SMILES strings for the compilation of the various input files. The Java Parser works the same way. Depending on the type of input data file either of the parsers is used.
2.3.2 Protein Data Bank

A HIV database was constructed using the ligand data of the various HIV - 1 Protease inhibitors. These data were extracted from the Protein Data Bank (PDB) and had to be compiled manually since the dump files from the PDB did not contain SMILES string information of the ligands. This file contained the identifier, molecular formula and the SMILES string information.

![Diagram of data flow](image)

Figure 2.3.1 Input Construction

2.4 Output

Two files are produced as a result of running LigEvolver. One gives the details regarding the number of children produced, maximum, minimum and average fitness for each generation. The other file gives the final set of SMILES strings along with the generation that it was produced.
2.5 External Packages

Certain calculations required the use of PerlMol, an external package comprised of modules for performing basic computational chemistry related functions. This is freely distributed software and a detailed description of PerlMol can be found at (www.perlmol.org). PerlMol provided the following functionalities to LigEvolver:

1. Calculation of Chemical formula from the SMILES string
2. Calculation of bonds from the SMILES string
3. Polar Surface Area (PSA)

A perl script was written to combine the PSA script (available at http://www.perlmol.org/examples/polar_surface_area/) with the chemical formula and bond calculation.

2.6 Design of LigEvolver

This section gives a detailed description of the individual components of LigEvolver. Each of the components has a specific purpose. An overview has been provided in fig. 2.6.1.
2.6.1 Genetic Algorithm

A genetic algorithm (GA) is used to evolve the population in a random manner to generate reasonably good individuals. It is driven by the concept of Darwinian evolution. The use of genetic algorithms in the generation of ligands has been successfully explored [3,4,5]. The main parts of the algorithm fall into two broad categories namely, Genetic operators and Fitness function, which will be discussed in the following sections. The genetic algorithm is the main module. It comprises of functions to initialize the population, mate individuals and generate new population. This is illustrated in fig. 2.6.1.1.
Figure 2.6.1.1 Flow chart of GA
The initial population is seeded in different styles. The ways of initialization are as follows:

1. Initialize with known HIV inhibitors
2. Initialize with random inhibitors
3. Initialize with a mixture of known and random inhibitors

For each generation, the population is rearranged according to the fitness function and only a fixed number of individuals is retained to contribute towards evolution while the rest are discarded.

2.6.1.1 Genetic Operators

The genetic operators are the mechanisms through which the individuals in a population are manipulated and changed. If evolution were viewed as a reaction, then genetic operators can be considered catalysts. Two dominant operators observed in many applications are the Crossover and Mutation operator. Depending on the need of the problem customized operators may also be included as is the case in this study. The operators used in LigEvolver are, GenFrag, Crossover and Mutator.

2.6.1.1.1 GenFrag

The functionality of this operator is to divide the input SMILES string into meaningful fragments. This operator was introduced because random segmentation of the SMILES string may give rise to chemically impossible or meaningless individuals. The numbers in a SMILES string serve as the cut points.
Original String

\[
\text{CC(C)CN(CC(O)C(Cc1ccccc1) NC(=O)OC2CCOC2) (=O)c3ccc(N)cc3}
\]

Fragments after application of GenFrag

The SMILES string is initially cut into fragments.

\[
\text{CC(C)CN(CC(O)C(Cc1ccccc1) NC(=O)OC2CCOC2) (=O)c3ccc(N)cc3}
\]

A random point of concatenation is chosen to give rise to two strings to serve as head and tail. In this case the concatenation point is taken to be two. So the start of the fragment is combined with the second fragment for the head and the rest of the fragments are combined together to form the tail.

Fragment One (Serves as the head for the Crossover operator)

\[
\text{CC(C)CN(CC(O)C(Cc1ccccc1) NC(=O)OC2CCOC2)}
\]

Fragment Two (Serves as the tail for the Crossover operator)

\[
(=O)c3ccc(N)cc3
\]

Figure 2.6.1.1.1 Example of GenFrag

Overlap and duplicity checks are performed within the GenFrag, to avoid redundancy in the gene pool. The fragments to be combined for the head and the tail portion are determined randomly.
2.6.1.1.2 Crossover

The crossover operator is used to mate the parents. The GenFrag operator returns two fragments that are the head and tail of an individual. Crossovers take place 90% of the time.

The crossover operator works as shown in figure 2.6.1.2.1.

![Figure 2.6.1.2.1 Example of Crossover](image)

2.6.1.1.3 Mutator

Mutations have played an important role in the evolution of living organisms. These operators cause variations in the gene pool. They can be both beneficial and harmful, although harmful ones tend to be eliminated by the fitness function. A lack of mutation
leads to no adaptability in the population whereas levels of mutations that are too high lead to genetic breakdown as shown in fig. 2.6.1.3[5].

Figure 2.6.1.3 Mutation rates and genetic adaptability [5]

In LigEvolver, mutations take place 10% of the time.

2.6.1.3.1 Amino Acid Insertion

Random places of insertion are chosen and an amino acid is included. The choice of the amino acid is also done randomly.
<table>
<thead>
<tr>
<th>Amino Acid</th>
<th>SMILES String</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alanine</td>
<td>NC(C)C(O)=O</td>
</tr>
<tr>
<td>Arginine</td>
<td>NC(=N)NCCCC(N)C(=O)O</td>
</tr>
<tr>
<td>Asparagine</td>
<td>NC(=O)CC(N)C(=O)O</td>
</tr>
<tr>
<td>Asparatic Acid</td>
<td>OC(=O)CC(N)C(=O)O</td>
</tr>
<tr>
<td>Cysteine</td>
<td>NC(CS)C(=O)O</td>
</tr>
<tr>
<td>Glutamic Acid</td>
<td>C(CC(=O)N)C(C(=O)O)N</td>
</tr>
<tr>
<td>Glutamine</td>
<td>NC(CCC(=O)N)C(=O)O</td>
</tr>
<tr>
<td>Glycine</td>
<td>C(C(=O)O)N</td>
</tr>
<tr>
<td>Histidine</td>
<td>NC(Cc1nc[nH]1)c(=O)O</td>
</tr>
<tr>
<td>Isoleucine</td>
<td>CCC(C)C(N)C(=O)O</td>
</tr>
<tr>
<td>Leucine</td>
<td>CC(C)CC(N)C(=O)O</td>
</tr>
<tr>
<td>Lysine</td>
<td>NCCCCCC(N)C(=O)O</td>
</tr>
<tr>
<td>Methionine</td>
<td>C(N)(C(O)O)CCSC</td>
</tr>
<tr>
<td>Phenylalanine</td>
<td>NC(Cc1ccccc1)c(=O)O</td>
</tr>
<tr>
<td>Proline</td>
<td>C1CCNC1C(=O)O</td>
</tr>
<tr>
<td>Serine</td>
<td>NC(CO)C(=O)O</td>
</tr>
<tr>
<td>Threonine</td>
<td>CC(O)C(N)C(=O)O</td>
</tr>
<tr>
<td>Tryptophan</td>
<td>NC(Cc1cc2ccccc2[nH]1)c(=O)O</td>
</tr>
<tr>
<td>Tyrosine</td>
<td>NC(Cc1cccc(O)cc1)c(=O)O</td>
</tr>
<tr>
<td>Valine</td>
<td>CC(C)C(N)C(=O)O</td>
</tr>
</tbody>
</table>

Table 2.6.1.3.1 Amino Acids and their SMILES string
2.6.1.2 Parent Selection

Parent selection is another important design factor for a genetic algorithm. There are various selection algorithms and the one employed in this study is tournament selection. In this method two groups of four individuals are selected randomly. The one with the highest fitness score is selected from each group. These two individuals then serve as the parents. In case of a tie with respect to fitness score, the parent is chosen randomly. This selection approach follows the concept of a tournament with teams competing to get selected. Finally the fitness score is the determinant of the winners.

Figure 2.6.1.2.1 Tournament Selection
2.6.2 Pharmacophore Elucidator

This module is used to extract the pharmacophore from a SMILES string. The pharmacophoric features include aromatic rings, hydrogen bond acceptor, hydrogen bond acceptors/donors, hydrophobic regions, positive charge centers and negative charge centers.

2.6.2.1 Aromatic Rings

Aromatic rings are a circular arrangement of atoms that can be represented with alternating single and double – bonds. It is a conjugated system of lone pairs, unsaturated bonds and empty orbitals that put together form a strong stabilization [36]. In chemical terms, aromatic rings contain \((4n+2)\pi\) electrons, where \(n\) is an integer.

The following are the chemical compounds that would be detected by the tool as aromatic rings.

![Aromatic Structure I - A Benzene ring](image)

Figure 2.6.2.1.1 Aromatic Structure I – A Benzene ring*

* Drawing has been generated using Smi2Depict developed by University of California, Irvine [44]
2.6.2.2 Hydrogen Bond Acceptor

Hydrogen atoms that are bound to electronegative atoms such as oxygen or nitrogen can interact with lone pairs to form additional bonding [39]. The lone pairs that interact with the hydrogen atom are termed as hydrogen bond acceptors.

The following are the chemical compounds that would be detected by the tool as hydrogen bond acceptors.

![Figure 2.6.2.2.1 Hydrogen Bond Acceptor Structure I](image1)

\[ \text{O} \]

![Figure 2.6.2.2.2 Hydrogen Bond Acceptor Structure II](image2)

![Figure 2.6.2.2.3 Hydrogen Bond Acceptor Structure III](image3)
2.6.2.3 Hydrogen Bond Acceptor/Donor

A group that can either accept or donate hydrogen bonds is identified as hydrogen bond acceptor/donor.

The following are the chemical compounds that would be detected by the tool as hydrogen bond acceptors/donor. This is not an exhaustive list, but contains a few of the common bonds that occur in HIV inhibitors.

- OH

Figure 2.6.2.3.1 Hydrogen bond A/D Structure I

-NH-

Figure 2.6.2.3.2 Hydrogen bond A/D Structure II
2.6.2.4 Hydrophobic Regions

The regions that repel water are termed hydrophobic regions.

The following are the chemical compounds that would be detected by the tool as hydrophobic regions.

Figure 2.6.2.4.1 Hydrophobic Structure I*

Figure 2.6.2.4.2 Hydrophobic Structure II*
Figure 2.6.2.4.3 Hydrophobic Structure III

Figure 2.6.2.4.4 Hydrophobic Structure IV

Figure 2.6.2.4.5 Hydrophobic Structure V
2.6.3 Molecular Weight Calculator

As the name indicates, the molecular weight calculator computes the molecular weight given the chemical formula. The molecular weight values (round-off weight) used by the tool is shown in the table below.

<table>
<thead>
<tr>
<th>Molecule</th>
<th>Weight</th>
<th>Round-off weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>12.0110369</td>
<td>12.011</td>
</tr>
<tr>
<td>H</td>
<td>1.0079759</td>
<td>1.007</td>
</tr>
<tr>
<td>O</td>
<td>15.9993047</td>
<td>15.999</td>
</tr>
<tr>
<td></td>
<td>Molecular weights</td>
<td></td>
</tr>
<tr>
<td>-------</td>
<td>-------------------</td>
<td>-----</td>
</tr>
<tr>
<td>B</td>
<td>10.8110280</td>
<td>10.811</td>
</tr>
<tr>
<td>N</td>
<td>14.0067231</td>
<td>14.006</td>
</tr>
<tr>
<td>Cl/l</td>
<td>35.4527379</td>
<td>35.452</td>
</tr>
<tr>
<td>Br/r</td>
<td>79.9035266</td>
<td>79.903</td>
</tr>
<tr>
<td>F</td>
<td>18.9984032</td>
<td>18.998</td>
</tr>
<tr>
<td>S</td>
<td>32.0643881</td>
<td>32.064</td>
</tr>
<tr>
<td>P</td>
<td>30.973762</td>
<td>30.973</td>
</tr>
<tr>
<td>I</td>
<td>126.904473</td>
<td>126.904</td>
</tr>
<tr>
<td>Fe/e</td>
<td>55.8451474</td>
<td>55.845</td>
</tr>
<tr>
<td>Na/a</td>
<td>22.9897677</td>
<td>22.989</td>
</tr>
</tbody>
</table>

Table 2.6.3.1 Molecular weights

2.6.4 Van Der Waals Volume

Van der Waals surfaces define the closest contact that atoms in a molecule could make with one another. Hence it acts as an indicator of atom packing in a molecular structure. It also has an important role in establishing the specificity of interactions between protein binding pockets and ligands [36].

Zhao et al have described a method for the calculation of van der Waals volume using the formula, \( V(\text{vdW}) = \text{summation operator all atom contributions} \times 5.92N(B) + 14.7R(A) + 3.8R(NR) \) (\( N(B) \) is the number of bonds, \( R(A) \) is the number of aromatic rings, and \( R(NA) \) is the number of nonaromatic rings).
Table 2.6.4.1 gives the values of the atoms and their volumes.

<table>
<thead>
<tr>
<th>Atoms</th>
<th>$V_{vdw}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>H</td>
<td>7.24</td>
</tr>
<tr>
<td>C</td>
<td>20.58</td>
</tr>
<tr>
<td>N</td>
<td>15.60</td>
</tr>
<tr>
<td>O</td>
<td>14.71</td>
</tr>
<tr>
<td>F</td>
<td>13.31</td>
</tr>
<tr>
<td>Cl</td>
<td>22.45</td>
</tr>
<tr>
<td>Br</td>
<td>26.52</td>
</tr>
<tr>
<td>I</td>
<td>32.52</td>
</tr>
<tr>
<td>P</td>
<td>24.43</td>
</tr>
<tr>
<td>S</td>
<td>24.43</td>
</tr>
<tr>
<td>As</td>
<td>26.52</td>
</tr>
<tr>
<td>B</td>
<td>40.48</td>
</tr>
<tr>
<td>Si</td>
<td>38.79</td>
</tr>
<tr>
<td>Se</td>
<td>28.73</td>
</tr>
<tr>
<td>Te</td>
<td>36.62</td>
</tr>
</tbody>
</table>

Table 2.6.4.1 Atoms and their Volumes
2.6.5 Polar Surface Area

Polar Surface Area (PSA) is a descriptor used to indicate the bioavailability of drugs. The surface belonging to polar atoms correlates with the transport of molecules through membranes. Ertl et al proposed an approach for the calculation of PSA based on the summation of surface contributions of polar fragments. PerlMol uses this method and provides a perl script to calculate the PSA.

2.6.6 Motif Finder

The motif finder is responsible for observing common patterns between a set of strings. It is used to derive structural similarities between compounds through the longest common substring present in their SMILES string.

The structures of the various FDA (U.S. Food and Drug Administration) approved drugs for HIV protease were compiled in the form of SMILES strings. These were then matched with the recently generated ligands to determine if they have a common substring of length greater than 9. Table 2.5.6.1 is the list of drugs that has been taken into consideration.

<table>
<thead>
<tr>
<th>Brand Name</th>
<th>Generic Name</th>
<th>Manufacturer Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agenerase</td>
<td>Amprenavir</td>
<td>GlaxoSmithKline</td>
</tr>
<tr>
<td>Aptivus</td>
<td>Tipranavir</td>
<td>Boehringer Ingelheim</td>
</tr>
<tr>
<td>Crixivan</td>
<td>indinavir, IDV, MK-639</td>
<td>Merck</td>
</tr>
<tr>
<td>Fortovase</td>
<td>Saquinavir</td>
<td>Hoffmann-La Roche</td>
</tr>
<tr>
<td>Invirase</td>
<td>Saquinavir mesylate, SQV</td>
<td>Hoffmann-La Roche</td>
</tr>
<tr>
<td><strong>Kaletra</strong></td>
<td>lopinavir and ritonavir</td>
<td>Abbott Laboratories</td>
</tr>
<tr>
<td><strong>Lexiva</strong></td>
<td>Fosamprenavir Calcium</td>
<td>GlaxoSmithKline</td>
</tr>
<tr>
<td><strong>Norvir</strong></td>
<td>ritonavir, ABT-538</td>
<td>Abbott Laboratories</td>
</tr>
<tr>
<td><strong>Reyataz</strong></td>
<td>Atazanavir sulfate</td>
<td>Bristol-Myers Squibb</td>
</tr>
<tr>
<td><strong>Viracept</strong></td>
<td>nelfinavir mesylate, NFV</td>
<td>Agouron Pharmaceuticals</td>
</tr>
</tbody>
</table>

Table 2.6.6.1 FDA – Approved Protease Inhibitors*

*Data taken from FDA site [42]

2.6.7 Fitness Computor

The fitness function is used to evaluate the strength and robustness of an individual element in the population. The fitness is calculated for every entity and it determines if the individual needs to be retained or eliminated. In a way, the fitness function biases the path of evolution. The following is the fitness function.

\[ F = Ar + Hd + Ha + Hy + Mw + Ps + Vd + Sm \]

Where,

Ar – Aromatic Ring Count Score
Hd – Hydrogen Bond Donor Score
Ha – Hydrogen Bond Acceptor Score
Hy – Hydrophobic Region Count Score
Mw – Molecular Weight Score
Ps – Polar Surface Area Score
Vd – Van Der Waals Volume Score
Sm – Structural Motif Score

For each of the above mentioned features scores were assigned according to their resemblance to the existing known inhibitors with the lowest score being the minimum matching score and the highest being the maximum match. If there are no matches to the known inhibitor feature count then no points are assigned. A list of the known existing inhibitors was made and their feature pattern was analyzed manually to obtain the scoring scheme. This list has been included as appendix II.

2.6.7.1 Aromatic Ring Count

The Aromatic ring is one of the pharmacophoric features that are vital components in determining the way the inhibitors interact with the enzyme. Table 2.6.7.1.1 gives the scoring scheme for the aromatic ring count.

<table>
<thead>
<tr>
<th>Aromatic Ring Count</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt;4 and &lt;7</td>
<td>1</td>
</tr>
<tr>
<td>&gt;1 and &lt;4</td>
<td>2</td>
</tr>
<tr>
<td>== 4</td>
<td>3</td>
</tr>
</tbody>
</table>

Table 2.6.7.1.1 Aromatic Ring Count Scoring Scheme

2.6.7.2 Hydrogen Bond Donor Count

A count of the hydrogen bond donors is taken and scores are assigned according to table 2.6.7.2.1.
### Table 2.6.7.2.1 Hydrogen Bond Donor Count Scoring Scheme

<table>
<thead>
<tr>
<th>Hydrogen Bond Donor Count</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt;1 and &lt;5</td>
<td>1</td>
</tr>
<tr>
<td>&gt;7</td>
<td>2</td>
</tr>
<tr>
<td>&gt;4 and &lt;8</td>
<td>3</td>
</tr>
</tbody>
</table>

### 2.6.7.3 Hydrogen Bond Acceptor Count

A count of the hydrogen bond acceptors is taken and scores are assigned according to table 2.6.7.3.1.

<table>
<thead>
<tr>
<th>Hydrogen Bond Acceptor Count</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt;1 and &lt;5</td>
<td>1</td>
</tr>
<tr>
<td>&gt;7</td>
<td>2</td>
</tr>
<tr>
<td>&gt;4 and &lt;8</td>
<td>3</td>
</tr>
</tbody>
</table>

### Table 2.6.7.3.1 Hydrogen Bond Acceptors Count Scoring Scheme

### 2.6.7.4 Hydrophobic Region Count

A count of the hydrophobic regions is taken and scores are assigned according to table 2.6.7.4.1.

<table>
<thead>
<tr>
<th>Hydrophobic Region Count</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt;1 and &lt;5</td>
<td>1</td>
</tr>
<tr>
<td>&gt;7</td>
<td>2</td>
</tr>
</tbody>
</table>
and
<8

Table 2.6.7.4.1 Hydrophobic Regions Count Scoring Scheme

2.6.7.5 Molecular Weight

Scores are assigned according to the range of molecular weights for each compound. The range and their corresponding score are summarized in table 2.6.7.5.1.

<table>
<thead>
<tr>
<th>Molecular Weight (Daltons)</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;500</td>
<td>1</td>
</tr>
<tr>
<td>&gt;600 and &lt;806</td>
<td>2</td>
</tr>
<tr>
<td>&gt;500 and &lt;600</td>
<td>3</td>
</tr>
</tbody>
</table>

Table 2.6.7.5.1 Molecular Weight Scoring Scheme

2.6.7.6 Polar Surface Area

The ideal polar surface range has been taken from the list of known inhibitors and scores are attributed to a newly generated individual if its polar surface area falls within the range. The details regarding the range and the scores are given in table 2.6.7.6.1.

<table>
<thead>
<tr>
<th>Polar Surface Area (Angstroms)</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt;150 and &lt;200</td>
<td>1</td>
</tr>
<tr>
<td>&gt;100 and &lt;150</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 2.6.7.6.1 Polar Surface Area Scoring Scheme
2.6.7.7 Van Der Waals Volume

Van der waals volume is another descriptor that is used to assign scores to the population units. The volume range and the respective score have been shown in table 2.6.7.7.1.

<table>
<thead>
<tr>
<th>Van Der Waals Volume (Angstroms$^3$)</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt;400 and &lt;750</td>
<td>1</td>
</tr>
<tr>
<td>&gt;800 and &lt;900</td>
<td>2</td>
</tr>
<tr>
<td>&gt;900 and &lt;1000</td>
<td>3</td>
</tr>
</tbody>
</table>

Table 2.6.7.7.1 Van Der Waals Volume Scoring Scheme

2.6.7.8 Structural Motif

The structures of the government approved drugs were compared to the new ligand to see if they include similarities. The presence of a significant resemblance gives a score of one and the absence a score of zero.
3. Results

LigEvolver was run using the data from various public databases. The results obtained are discussed in this section. Unless otherwise specified the entire chemical figures have been generated using CS ChemDraw Pro.

3.1 Simulation I

The inhibitor generation was done using the parameters shown in table 3.1.1.

<table>
<thead>
<tr>
<th>Property</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Database</td>
<td>ChemPDB</td>
</tr>
<tr>
<td>Initial Population Size</td>
<td>4009</td>
</tr>
<tr>
<td>Number of Matings per Generation</td>
<td>150</td>
</tr>
<tr>
<td>Number of Generations</td>
<td>100</td>
</tr>
<tr>
<td>Population size per Generation</td>
<td>500</td>
</tr>
</tbody>
</table>

Table 3.1.1 Simulation I Properties

3.1.1 Significant Results

A few noteworthy compounds that were generated during this simulation are discussed in the following section.
3.1.1.1 Generation #100

3.1.1.1.1 Compound 1

SMILES string

\[
C(=O)NC(Cc3cccc3)C(NC(=O)C(N)c3cccc3)C(=O)N2C1CC1=CN(C2)ccccCC(C)CNc2cccc3CC(C)CNc23
\]

Figure 3.1.1.1.1 Simulation I – Generation 100 - Compound I*

*Figure generated using ChemSketch [45].
### 3.1.1.1.2 Compound II

**SMILES String**

\[
\text{NC(=O)c1c[n+]1(cccc1)(NC(=O)C(N)c3cccccc3)C(=O)N2C1CC1=CN(C2)ccccc3CC(C)C}
\]

\[
\text{Nc2cccc3CC(C)CNc23}
\]

**Reconstruction**

Elements underlined, either in the original SMILES string or in the reconstructed string, signify the portion of the SMILES string that is modified. This convention is followed throughout the remaining sections.

\[
\text{NC(=O)c1c[n+]1(cccc1)(NC(=O)C(N)c3cccccc3)C(=O)N2C1CC1=CN(C2)ccccc3CC(C)C}
\]

\[
\text{c2cccc3CC(C)CNc23}
\]

---

![Figure 3.1.1.1.2 Simulation I - Generation 100 - Compound II*](image-url)
3.1.1.2 Generation #85

3.1.1.2.1 Compound I

SMILES String

```
COC(=O)C(Cc1ccccc1)CCC1=C(C)C(=O)N1C(O)CN(CCC4CCC5OCOC5C4)C(=O)
CCN6C(=O)c7ccccc7C6
```

*Figure generated using ChemSketch [45].

Figure 3.1.1.2.1.1 Simulation I – Generation 85 - Compound I
3.1.1.2.2 Compound II

SMILES String

CC(C)C(NC(C)=O)C(=O)NC(Cc1cccc1)C(O)CN(CCC4CCC5OCOC5C4)C(=O)CCN6
C(=O)c7cccccc7C6
## 3.2 Simulation II

<table>
<thead>
<tr>
<th>Property</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Database</td>
<td>KEGG</td>
</tr>
<tr>
<td>Initial Population Size</td>
<td>10005</td>
</tr>
<tr>
<td>Number of Matings per Generation</td>
<td>100</td>
</tr>
<tr>
<td>Number of Generations</td>
<td>45</td>
</tr>
<tr>
<td>Population size per Generation</td>
<td>500</td>
</tr>
</tbody>
</table>

Table 3.2.1 Simulation II Properties

### 3.2.1 Significant Results

A few noteworthy compounds that were generated during this simulation are discussed in the following section.
3.2.1.1 Generation #40

3.2.1.1.1 Compound 1

SMILES String

CCC(C)C(N(C)C)(=O)NC1C(Oc2ccc(cc2)C(O)CNC(=O)C(Cc3cccc3)NCl)CC(=O)NC1C(O)CC(OC1)

Figure 3.2.1.1.1.1 Simulation II – Generation 40 - Compound I
3.2.1.1.2 Compound II

SMILES String

\[
\text{CC(C)CC(=O)C(=O)NC(=O)NC(=O)NC(Cc1cccc1)C(Oc4ccc(C=CNC(=O)C(Cc5cccc5)NC3c4cccc4)cc5cccc5)}
\]

Reconstruction

\[
\text{CC(C)CC(=O)C(=O)NC(=O)NC(=O)NC(Cc1cccc1)C(Oc4ccc(C=CNC(=O)C(Cc5cccc5)NC3c4cccc4)cc5cccc5)}
\]

Figure 3.2.1.2.1 Simulation II – Generation 40 - Compound II
3.2.1.1.3 Compound III

SMILES String

NC(=O)OC2CCOC2cccc3CC(C)CNe2cccc3CC(C)CNe2cccc3CC(C)CNe2cccc3CC(C)CNe2

Reconstruction

NC(=O)OC2CCOC2cccc3CC(C)CNe2cccc3CC(C)CNe2cccc3CC(C)CNe2cccc3CC(C)CNe2

Figure 3.2.1.1.3.1 Simulation II – Generation 40 - Compound III*

*Figure generated using ChemSketch [45].
3.2.1.4 Compound IV

SMILES String

CC(C)(C)NC(=O)C1CN(CC1)COC(=O)CC(O)(CCC(C)(C)O)C(=O)OC1C2c3cc4OC
Oc4cc3CCN5CCCC25C=C1

Figure 3.2.1.4.1 Simulation II – Generation 40 - Compound IV
3.2.1.2 Generation #41

3.2.1.2.1 Compound I

SMILES String

[nH]c(C=C4NC(=O)C(C)=C4)C(=O)N2CCCC2Oc2ccc(C=CNC(=O)C(Cc3ccccc3)NC(=O)C1NC(=O)C4CCCN4C)cc2

Reconstruction

[nH]c(C=C4NC(=O)C(C)=C4)C(=O)N2CCCC2Oc2ccc(C=CNC(=O)C(Cc3ccccc3)NC(=O)CNC(=O)C4CCCN4C)cc2

Figure 3.2.1.2.1.1 Simulation II – Generation 41 - Compound I*

*Figure generated using ChemSketch [45].
3.3 Simulation III

<table>
<thead>
<tr>
<th>Property</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Database</td>
<td>MSD</td>
</tr>
<tr>
<td>Initial Population Size</td>
<td>6433</td>
</tr>
<tr>
<td>Number of Matings per Generation</td>
<td>100</td>
</tr>
<tr>
<td>Number of Generations</td>
<td>100</td>
</tr>
<tr>
<td>Population size per Generation</td>
<td>500</td>
</tr>
</tbody>
</table>

Table 3.3.1 Simulation III Properties

3.3.1 Significant Results

A few noteworthy compounds that were generated during this simulation are discussed in the following section.
3.3.1.1 Generation #100

3.3.1.1 Compound 1

SMILES String

CN(C)C(C1CCNC1C(=O)Oc1cccc1)C(=O)N2CCCC2C(=O)NC3C(Oc4ccc(C=CNC(=O)C(Cc5ccccc5)NC3

Reconstruction

CN(C)C(C1CCNC1C(=O)Oc1ccccc1)C(=O)N2CCCC2C(=O)NC3COcccc(C=CNC(=O)C(Cc5ccccc5)NC3c4)

Figure 3.3.1.1.1 Simulation III – Generation 100 - Compound I (a)*
Reconstruction

CN(C)(C1CCNC1C(=O)Oc1cccc1)C(=O)N2CCCC2C(=O)NC3C(Oc4ccc(C=CNC(=O)C(Cc5cccc5)N)C34)

Figure 3.3.1.1.2 Simulation III – Generation 100 - Compound I (b)*
Reconstruction

CN(C)C(C1CCNC1C(=O)OC1cccc1)C(=O)N2CCCC2C(=O)NC3C(Oc4cccC=C4iNC(=O)C(Cc5ccccc5)NC3

Figure 3.3.1.1.3 Simulation III – Generation 100 - Compound I (c)*

*Figures generated using ChemSketch [45].
3.3.1.1.2 Compound II

SMILES String

\[
\text{NC(CS)C(=O)OC}_2(O)C_3C(OC)C_1(COC(=O)c_6cccc6)C_45C_2\text{CCC}_8(COC(=O)c_6)C(=O)NCC(=O)N(C)C(Cc_2cccc6)C_23\text{CCCC}(COC(=O)c_1)C(=O)NCC(=O)N(C)C(Cc_2cccc6)
\]

Reconstruction

\[
\text{NC(CS)C(=O)OC}_2(O)C_3C(OC)C_1(COC(=O)c_6cccc6)C_23\text{CCCC}(COC(=O)c_1)C(=O)NCC(=O)N(C)C(Cc_2cccc6)
\]

Figure 3.3.1.1.2.1 Simulation III – Generation 100 - Compound II (a)
Reconstruction

\[ \text{NC(CS)C(=O)OC2(O)C3C(OC)Cl(COC(=O)c6ccc6)c23C(CCC1COC(=O)cC(=O)NCC(=O)N(C)C(Cc2ccc2))} \]

Figure 3.3.1.1.2.2 Simulation III – Generation 100 - Compound II (b)*

*Figure has been generated using Smi2Depict developed by University of California, Irvine [44].

3.3.1.1.3 Compound III

SMILES String

\[ \text{CCC1=C(C)CN(C(=O)NCCc2ccc(cc2)S(=O)(=O)NC(=O)NC3CCC(C)C3)C1OCC1=CNC(=O)NC1C5cccnc5} \]

Figure 3.3.1.1.3.1 Simulation III – Generation 100 - Compound III
3.4 Simulation IV

<table>
<thead>
<tr>
<th>Property</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Database</td>
<td>ChemBank</td>
</tr>
<tr>
<td>Initial Population Size</td>
<td>2344</td>
</tr>
<tr>
<td>Number of Matings per Generation</td>
<td>100</td>
</tr>
<tr>
<td>Number of Generations</td>
<td>100</td>
</tr>
<tr>
<td>Population size per Generation</td>
<td>300</td>
</tr>
</tbody>
</table>

Table 3.4.1 Simulation IV Properties

3.4.1 Significant Results

A few noteworthy compounds that were generated during this simulation are discussed in the following section.
3.4.1.1 Generation #100

3.4.1.1.1 Compound I

SMILES String

NC(=O)C2CCCN2C(=O)C(Cc3cccc3)NC(=O)C(Cc4cccc4)NC1CN1CCN(CC1)CC(C)C1NC(=O)C(Cc2)

Reconstruction

NC(=O)C2CCCN2C(=O)C(Cc3cccc3)NC(=O)C(Cc4cccc4)NC1CN1CCN(CC1)CC(C)C1NC(=O)C(C)

Figure 3.4.1.1.1 Simulation IV – Generation 100 - Compound I
3.4.1.1.2 Compound II

SMILES String

CN1CCN(CC1)CC(C1)NC(=O)C(Cc2ccccc2)NC(=O)C(CCCCCC(=O)NO)NC(=O)C3CCCN3C1c5ccccc5

Reconstruction

CN1CCN(CC1)CC(C1)NC(=O)C(Cc2ccccc2)NC(=O)C(CCCCCC(=O)NO)NC(=O)C3CCCN3C1c5ccccc5

Figure 3.4.1.1.2.1 Simulation IV – Generation 100 - Compound II
3.5 Simulation V

<table>
<thead>
<tr>
<th>Property</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Database</td>
<td>HIV_PDB</td>
</tr>
<tr>
<td>Initial Population Size</td>
<td>71</td>
</tr>
<tr>
<td>Number of Matings per Generation</td>
<td>100</td>
</tr>
<tr>
<td>Number of Generations</td>
<td>100</td>
</tr>
<tr>
<td>Population size per Generation</td>
<td>150</td>
</tr>
</tbody>
</table>

Table 3.5.1 Simulation V Properties

3.5.1 Significant Results

A few noteworthy compounds that were generated during this simulation are discussed in the following section.
3.5.1.1 Generation #12

3.5.1.1.1 Compound I

SMILES String

CC(C)(C)NC(=O)C1CC2CCCCC2CN13=CN[C@](C[C@H](O)[C@H](Cc4cccc4)NC(=O)O[C@H]5CCOC5)(Cc6cccc6)C3

Reconstruction

CC(C)(C)NC(=O)C1CC2CCCCC2CN13=CN[C][C@H](O)[C@H](Cc4cccc4)NC(=O)O[C@H]5CCOC5)(Cc6cccc6)C3

Figure 3.5.1.1.1.1 Simulation V – Generation 12 - Compound I
3.5.1.2 Generation #10

3.5.1.2.1 Compound I

SMILES String

CC1CCC(O)C1CC(C)(C)NC(=O)C1CC2CCCCC2CN1C(O)C(O)C(OCc4ccccc4)C(=O)NCc5c(F)ccccc5

Figure 3.5.1.2.1.1 Simulation V – Generation 10 - Compound I
3.5.1.3 Generation #9

3.5.1.3.1 Compound I

SMILES String

\[(\text{CNC}(=\text{O})\text{c}4\text{cccccc3})\text{CN}(\text{Cc}2\text{cccccc2})\text{C}(=\text{O})\text{CO}3\text{c}(\text{C})\text{ccccc3NC}(=\text{O})\text{C}(\text{CC(N)}=\text{O})\text{NC}(=\text{O})\text{c}4\text{ccc5ccccc5n4})\]

Reconstruction

\[(\text{CNC}(=\text{O})\text{c}4\text{cccccc4})\text{CN}(\text{Cc}2\text{cccccc2})\text{C}(=\text{O})\text{CO}3\text{c}(\text{C})\text{ccccc3NC}(=\text{O})\text{C}(\text{CC(N)}=\text{O})\text{NC}(=\text{O})\text{c}4\text{ccc5ccccc5n4})\]

Figure 3.5.1.3.1.1 Simulation V – Generation 9 - Compound I*

*Figure has been generated using Smi2Depict developed by University of California, Irvine [44].

3.6 Simulation VI

<table>
<thead>
<tr>
<th>Property</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Database</td>
<td>HIVNCl</td>
</tr>
<tr>
<td>Initial Population Size</td>
<td>42689</td>
</tr>
<tr>
<td>Number of Matings per Generation</td>
<td>100</td>
</tr>
<tr>
<td>Number of Generations</td>
<td>200</td>
</tr>
</tbody>
</table>
3.6.1 Significant Results

3.6.1.1 Generation #199

3.6.1.1.1 Compound I

SMILES String

\[
\text{COC(=O)C(O)(CCC(C)C)C(=O)OC1C2c3cc4OCOC4cc3CCN5CCCC25C=C1CCN5CCCC}25\text{C=C1CN=Ne5ccccc5}
\]

Reconstruction

\[
\text{COC(=O)C(O)(CCC(C)C)C(=O)OC1C2c3cc4OCOC4cc3CCN5CCCC25C=C1CCN5CCCC}5\text{C=CCN=Ne5ccccc5}
\]

Figure 3.6.1.1.1.1 Simulation VI - Generation 199 - Compound I
3.7 Fitness Behavior

It was noted that the fitness steadily increased with the number of generations. After a certain number of generations the fitness values approached an upper limit and this value remained constant through the subsequent generations. Figure 3.7.1 shows the fitness behavior chart across two simulations.

Figure 3.7.1 Fitness behavior
4. Discussion

The approach used in this study does not take into account the entire complexities involved in ligand design. It is limited due to the non-incorporation of the 3D features. The main objective of this research was to investigate if scalar properties such as those used in the fitness function would derive promising ligand structures. Another aspect was to experiment with the application of genetic algorithms to the domain of ligand generation.

From the results it can be inferred that the tool generated ligands that were hopeful and novel by nature. It also proved that genetic algorithm is an efficient computational approach towards solving problems of this nature. Certain compounds that were produced by LigEvolver had striking similarities to existing known inhibitors. A few of those are presented here.

![Figure 4.1 LigEvolver Compound I*](image-url)
Figure 4.2 Similarities of Compound I with known inhibitors [43]

Figure 4.3 LigEvolver Compound II*
Figure 4.4 Similarities of compound II with known inhibitor 1D4K [43]

Figure 4.5 LigEvolver Compound III*
A few compounds were generated that had a totally new structure than the ones already proved to be inhibitors. One such compound from this class of result is shown in Figure 4.7.

*Drawings generated using CS ChemDraw Pro.

Table 4.1 shows the various input database for the corresponding compounds.
<table>
<thead>
<tr>
<th>LigEvolver Compound #</th>
<th>Input Database</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>Protein Data Bank (only HIV inhibitors)</td>
</tr>
<tr>
<td>II</td>
<td>KEGG</td>
</tr>
<tr>
<td>III</td>
<td>ChemBank</td>
</tr>
<tr>
<td>IV</td>
<td>KEGG</td>
</tr>
</tbody>
</table>

Table 4.1 LigEvolver compounds with corresponding input databases
5. Conclusion

LigEvolver succeeded in generating new ligands using the minimalist approach. The results have been promising, but given the nature of the problem there is no computational way of verifying the solutions. A worthwhile future enhancement would be to improvise the fitness function by incorporating the 3D properties. Despite some unresolved issues, this tool provides an extensible first step towards generation of inhibitors that possess the potential of becoming a viable drug candidate.
Bibliography

Paper References

1. David Weininger. (1987) SMILES, a chemical language and information system.  


Books


Web References


40. Biology of HIV Protease Inhibitors, \\
    www.unc.edu/~brybar/evolutionmodule/Biology\%20of\%20HIV\%20Protease\%20Inhibitors.doc, accessed as on 16\textsuperscript{th} May 2006.

41. University of Calgary, \\
    http://www.chem.ucalgary.ca/courses/351/Carey5th/Ch27/ch27-4-4.html, \\
    accessed on 10\textsuperscript{th} March 2006.

42. FDA Approved Drugs, http://www.fda.gov/oashi/aids/virals.html, accessed as on 10\textsuperscript{th} March 2006.

43. Protein Data Bank, http://www.rcsb.org/pdb/home/home.do, accessed as on 12\textsuperscript{th} October 2005.

44. ChemDB / Smi2Depict: Generate 2D Images from Molecule Files, \\

45. ACD/ChemSketch 10.0 Freeware, \\
Appendix I

**Smiles Encoding Rules**

1. Non-hydrogen atoms are specified by their atomic symbol enclosed in square brackets. The second letter of the two-letter symbol must be entered in lower case.

2. Elements in the "organic subset", B, C, N, O, P, S, F, Cl, Br and I, may be written without brackets if the number of attached hydrogens conforms to the lowest normal valence consistent with explicit bonds.

3. Elements not in the organic subset must be described in brackets.

4. Attached hydrogens and formal charges are always specified inside brackets. The number of attached hydrogens is shown by the symbol H followed by an optional digit (e.g., [NH4+] – ammonium cation).

5. Similarly, a formal charge is shown by one of the symbols + or -, followed by an optional digit (e.g., [Fe+2] – iron (II) cation). The form [Fe+++] is considered synonymous with the form [Fe+3].

6. If unspecified, the number of attached hydrogens and charges is assumed to be zero for an atom inside the bracket.

7. Atoms in aromatic rings are specified by lower case letters; e.g., normal carbon is represented by the letter C, aromatic carbon by c.

8. Single bond is represented by the symbol –

9. Double bond is represented by the symbol =

10. Triple bond is represented by the symbol #

11. Aromatic bonds are represented by :
12. Single and aromatic bonds may be, and usually are, omitted.

13. Hydrogens can be omitted from the SMILES notation

Structure

CH₂=CH-CH₂-CH=CH-CH₂-OH

Valid SMILES

C=CCC=CCO
C=C-C-C=C-C-O
OCC=CCC=C

14. Branches are enclosed within parentheses

Triethylamine

SMILES
CCN(CC)CC

Isobutyric acid

SMILES
CC(C)C(=O)O

15. Branches can be nested or stacked
3- propyl - 4 - isopropyl - 1 - heptene

\[
\begin{align*}
&\text{CH}_3 \\
&\quad \text{CH}_2 \\
&\quad \text{CH}_2 \\
&\text{H}_2\text{C} = \text{CH} - \text{CH} & \quad \text{CH} - \text{CH}_2 - \text{CH}_2 - \text{CH}_3
\end{align*}
\]

SMILES

\[
\text{C} = \text{CC(CCC)C(C)C} \text{CCC}
\]

16. Cyclic structures are represented by breaking one single (or aromatic) bond in each ring.

17. The bonds are numbered in any order.

18. Ring opening or ring closure bonds are followed by a digit immediately following the atomic symbol at each ring closure.

Cyclohexane
SMILES

C1CCCCC1

19. There are usually many different but equally valid descriptions of the same structure

1-methyl-3-bromo-cyclohexane

\[
\begin{align*}
\text{CH}_3 \\
\text{O} & \text{C-H} \\
\text{C-H} & \text{C-H} \\
\text{C-H} & \text{Br}
\end{align*}
\]

SMILES

a. CCl=CC(Br)CCC1
b. CCl=CC(CCC1)Br

20. A single atom may have more than one ring closure.
21. Digits denoting ring closures can be reused, if desired.

\[
\begin{align*}
\text{SMILES} \\
\text{C12C3C4C1C5C4C3C25}
\end{align*}
\]

22. Disconnected compounds are written as individual structures separated by a period.

\[
\begin{align*}
\text{SMILES} \\
\text{O1CCCCC1N1CCCCC1}
\end{align*}
\]
SMILES

\[[Na^+].[O^+].c1ccccc1\]

Or

c1cc([O^-].[Na^+])ccc1

23. Atoms in aromatic rings are written in lower case letters

\[\begin{align*}
&\text{SMILES} \\
c1ccccc1C(=O)O
\end{align*}\]

24. The rule used in the SMILES system is to eliminate all hydrogen atoms except in the following three cases:
1) Hydrogens connected to other hydrogens

2) Hydrogens connected to zero or more than one other atom

3) In isomeric SMILES, isotopic hydrogen specifications, eg., [2H]

In these cases, hydrogens are retained and are treated like any other atom except that their hydrogen count is always zero.

25. Aromatic nitrogen symbol is ‘n’

26. Tetrahedral chirality can be specified using the "visual mnemonic" `@' character (anticlockwise) or two `@' characters (clockwise).

Reference

# Appendix II

List of HIV inhibitors gathered from Protein Data Bank

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Name</th>
<th>Structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>1D4S</td>
<td>N-(3-{{(1R)-1-[{(6R)-4-HYDROXY-2-OXO-6-PHENETHYL-6-PROPYL-5,6-DIHYDRO-2H-PYRAN-3-YL}PROPYL}PHENYL}-5-(TRIFLUOROMETHYL)-2-PYRIDINESULFONAMIDE</td>
<td><img src="image1.png" alt="Structure of 1D4S" /></td>
</tr>
<tr>
<td>1D4Y</td>
<td>N-(3-{{(1R)-1-[(6R)-4-HYDROXY-2-OXO-6-PHENETHYL-6-PROPYL-5,6-DIHYDRO-2H-PYRAN-3-YL]PROPYL}PHENYL}-5-(TRIFLUOROMETHYL)-2-PYRIDINESULFONAMIDE</td>
<td><img src="image2.png" alt="Structure of 1D4Y" /></td>
</tr>
</tbody>
</table>
1EBY  N,N-[2,5-O-DIBENZYL-GLUCARYL]-DI-[1-AMINO-INDAN-2-OL]

1ECO  N,N-[2,5-O-DI-2-FLUORO-BENZYL-GLUCARYL]-DI-[1-AMINO-INDAN-2-OL]

1G2K  3-(7-BENZYL-4,5-DIHYDROXY-1,1-DIOXO-3,6-BIS-PHENOXYMETHYL-1L6-[1,2,7]THIADIAZEPAN-2-YLMETHYL)-N-METHYL-BENZAMIDE
1HPO  
4-CYANO-N-(3-
CYCLOPROPYL(5,6,7,8,9,10-HEXAHYDRO-
4-HYDROXY-2-OXO-
CYCLOOCTA[B]PYRAN-3-
YL)METHYL)PHENYL
BENZENSULFONAMIDE

1HVH  
[[4-R(-4-ALPHA,5-ALPHA,6-BETA,7-
BETA)]-HEXAHYDRO-5,6-BIS(HYDROXY)-
1,3-BIS(4-HYDROXYMETHYL)METHYL]-
4,7-BIS(PHENYL METHYL)-2H-1,3-
DIAZEPIN-2-YLIDENE]CYANAMIDE

1HVR  
[4R-(4ALPHA,5ALPHA,6BETA,7BETA)]-
HEXAHYDRO-5,6-DIHYDROXY-1,3-BIS[2-
NAPHTHYL-METHYL]-4,7-
BIS(PHENYL METHYL)-2H-1,3-DIAZEPIN-
2-ONE
HYDROXY GROUP

HWR

[4-R-(4-ALPHA,6-BETA,7-BETA]-HEXAHYDRO-5,6-DI(HYDROXY)-1,3-
DI(ALLYL)-4,7-BISPHENYL-METHYL)-2H-
1,3-DIAZEPINONE

1JLD

3-(CARBOXYAMIDE(2-CARBOXYAMIDE-
2-TERTBUTYLETHYL))PENTAN

2-CARBONYLQUINOLINE

PHENYLALANYLMETHANE
1KZK

(R)-N-(2-METHYLBENZYL)-3-[(2S,3S)-2-HYDROXY-3-(3-HYDROXY-2-METHYLBENZOYL)AMINO-4-PHENYL BUTANOYL]-5,5-DIMETHYL-1,3-THIAZOLIDINE-4-CARBOXAMIDE

ETHYLENE GLYCOL

CHLORIDE ION

1QBS

[4-R-(4-ALPHA,5-ALPHA,6-BETA,7-BETA)]-HEXAHYDRO-5,6-BIS(HYDROXY)-[1,3-BIS((4-HYDROXYMETHYL-PHENYL)METHYL)-4,7-BIS(PHENYLMETHYL)]-2H-1,3-DIAZEPINONE

S-HYDROXYCysteine
1T7K

3-((5-BENZYL-6-HYDROXY-2,4-BIS-(4-HYDROXY-BENZYL)-3-OXO-[1,2,4]-TRIAZEPANE-1-SULFONYL)-BENZONITRILE

1UPJ

3-[1-(4-BROMO-PHENYL)-2-METHYL-PROPYL]-4-HYDROXY-CHROMEN-2-ONE

1W5V

N,N-[2,5-O-DI-3-FLUORO-BENZYL-GLUCARYL]-DI-[1-AMINO-INDAN-2-OL]
(2R,3R,4R,5R)-2,5-BIS[(2,4-DIFLUOROBENZYL)OXY]-3,4-DIHYDROXY-N,N'-BIS[(1R,2S)-2-HYDROXY-2,3-DIHYDRO-1H-INDEN-1-YL]HEXANEDIAMIDE

(2R,3R,4R,5R)-2,5-BIS[(2,3-DIFLUOROBENZYL)OXY]-3,4-DIHYDROXY-N,N'-BIS[(1S,2R)-2-HYDROXY-2,3-DIHYDRO-1H-INDEN-1-YL]HEXANEDIAMIDE

(2R,3R,4R,5R)-2,5-BIS[(2,5-DIFLUOROBENZYL)OXY]-3,4-DIHYDROXY-N,N'-BIS[(1S,2R)-2-HYDROXY-2,3-DIHYDRO-1H-INDEN-1-YL]HEXANEDIAMIDE
GLYCEROL

CHLORIDE ION

N-BENZYL-2-(2,6-DIMETHYLPHENOXY)-N-[(3R,4S)-4-[(ISOBUTYL(PHENYLSULFONYL)AMINO)METHYL]PYRROLIDIN-3-YL]METHYL]ACETAMIDE

CHLORIDE ION

N-[(1S)-1-(3-BROMOBENZYL)-4-[(4-BROMOPHENYL)SULFONYL]-6-METHYL-2-OXOHEPTYL]-2-(2,6-DIMETHYLPHENOXY)ACETAMIDE
(S)-N-((2S,3R)-3-HYDROXY-4-(4-((E)-(HYDROXYIMINO)METHYL)-N-ISOBUTYLPHENYL SULFONAMIDO)-1-PHENYL BUTAN-2-YL)-3-METHYL-2-(3-((2-METHYLTIAZOL-4-YL)METHYL)-2-OXOIMIDAZOLIDIN-1-YL)BUTANAMIDE

[1-((1S,2R)-1-BENZYL-2-HYDROXY-3-{ISOBUTYL[(4-METHOXYPHENYL)SULFONYL]AMINO}PROP-1H-1,2,3-TRIAZOL-4-YL)METHYL (1R,2R)-2-HYDROXY-2,3-DIHYDRO-1H-INDEN-1-YLCARBAMATE

TERT-BUTYL 4-[[[1-((1S,2R)-1-BENZYL-2-HYDROXY-3-{ISOBUTYL{(4-METHOXYPHENYL)SULFONYL]AMINO}PROP-1H-1,2,3-TRIAZOL-4-YL]METHYL]AMINO)CARBONYL]BENZYLCARBAMATE
2UPJ

\[2-(3\cdot(6-(1\text{-}\text{BENZYL}\text{-}PROPYL})\text{-}4\text{-}\text{HYDROXY}\text{-}2\text{-}\text{OXO}\text{-}2\text{H}\text{-}\text{PYRAN}\text{-}3\text{-}\text{YL})\text{-}\text{Cyclopropyl}\text{-}\text{Methyl})\text{-}\text{Phenylcarbamoyl})\text{-}\text{Ethyl})\text{-}\text{Carbamic Acid Tert-Butyl Ester}\]

3UPJ

4-HYDROXY-7-METHOXY-3-(1-PHENYL-PROPYL)-CHROMEN-2-ONE

4UPJ

\{(3\cdot(1\text{-}(4\text{-HYDROXY}\text{-}2\text{-}\text{OXO}\text{-}2\text{H}\text{-}\text{CHROMEN}-3\text{-}\text{YL})\text{-}\text{PROPYL})\text{-}\text{Phenylcarbamoyl})\text{-}\text{Methyl})\text{-}\text{Carbamic Acid Tert-Butyl Ester}\}

5UPJ

5,6,7,8,9,10-HEXAHYDRO-4-HYDROXY-3-(1-PHENYLPROPYL)CYCLOOCTA[B]PYRAN-2-ONE
6UPJ

6,7,8,9-TETRAHYDRO-4-HYDROXY-3-(1-PHENYLPROPYL)CYCLOHEPTA[B]PYRAN-2-ONE

7UPJ

N-(3-CYCLOPROPYL(5,6,7,8,9,10-HEXAHYDRO-2-OXO-2H-CYCLOOCTA[B]PYRAN-3-YL)METHYL)PHENYLBENZENSULFONAMIDE

1HEG

METHOXY GROUP

HYDROXYETHYL GROUP

2FDE

POTASSIUM ION
(5R,6R)-5-BENZYL-6-HYDROXY-2,4-BIS(4-HYDROXY-3-METHOXYBENZYL)-1-[3-(4-HYDROXYPHENYL)PROPANOYL]-1,2,4-TRIAZEPAN-3-ONE

N-[2(S)-CYCLOPENTYL-1(R)-HYDROXY-3(R)-METHYL]-5-[(2(S)-TERTIARY-BUTYLAMINO-CARBONYL)-4-(N1-(2)-(N-METHYLPIPRAZINYL)-3-CHLOROPYRAZINYL-5-CARBONYL)-PIPERAZINO]-4(S)-HYDROXY-2(R)-PHENYL METHYL-PENTANAMIDE

N-[2(R)-HYDROXY-1(S)-INDANYL]-5-[(2(S)-TERTIARY BUTYLAMINOCARBONYL)-4(3-PYRIDYL METHYL)PIPRAZINO]-4(S)-HYDROXY-2(R)-PHENYL METHYL-PENTANAMIDE
PYRIDINYL SULFONYL

PIPERIDINE-2-CARBOXYLIC ACID TERT-BUTYLAMIDE

PHENYLALANYLMETHANE

(2,6-DIMETHYL-PHENOXY)-ACETIC ACID
SULFATE ION


ALPHA-AMINOBUTYRIC ACID
CIS-N-TERT-BUTYL-DECAHYDRO-2-{2(R)-HYDROXY-4-PHENYL-3(S)-[[N-2-QUINOLYL-CARBONYL-L-ASPARAGINYL]AMINO]BUTYL}-(4AS)-ISOQUINOLINE-3(S)-CARBOXAMIDE

1QBR

[4R-(4ALPHA,5ALPHA,6BETA,7BETA)]-3,3'-[{TETRAHYDRO-5,6-DIHYDROXY-2-OXO-4,7-BIS(PHENYL METHYL)-1H-1,3-DIAZEPINE-1,3(2H)-DIYL}BIS(METHYLENE)]BIS[N-2-THIAZOLYLBENZAMIDE]

1QBT

[4R-(4ALPHA,5ALPHA,6ALPHA,7ALPHA)]-3,3'-[{TETRAHYDRO-5,6-DIHYDROXY-2-OXO-4,7-BIS(PHENYL METHYL)-1H-1,3-DIAZEPINE-1,3(2H)-DIYL}BIS(METHYLENE)]BIS[N-1H-BENZIMIDAZOL-2-YLBENZAMIDE]
[4R-(1\textalpha,5\textalpha,7\textbeta)]-3-[(\textcyclopropylmethy)l\texthexahydro-5,6-dihydroxy-2-oxo-4,7-bis(phenylmethyl)-1H-1,3-diazepin]-methyl-2-thiazolylbenzamide

\textit{1QBU}

*QUINALDIC ACID*

3-(mercaptomethylene)pyridine

*PIPERIDINE-2-CARBOXYLIC ACID TERT-BUTYLAMIDE*
2-HYDROXY-3-AMINO-4-PHENYL BUTANE

1D4L

(10S,13S,1'R)-13-[1'-HYDROXY-2'-(N-P-AMINO-BENZENESULFONYL-1''-AMINO-3''-METHYLBUTYL)ETHYL]-8,11-DIOXO-10-ISOPROPYL-2-OXA-9,12-DIAZABICYCLO [13.2.2]NONADeca-15,17,18-TRIENE

1CPI

AMINO GROUP

METHYLENE GROUP

DIMETHYLENE CARBONYL GROUP

ALPHA-AMINOBUTYRIC ACID

1HXB

2-CARBONYLQUINOLINE
TERTIARY-BUTYLAMINO GROUP

PHENYLALANINOL GROUP

2-METHYL-DECAHYDRO-ISOQUINOLINE-3-CARBOXYLIC ACID

1NPV

{1-BENZYL-3-[2-BENZYL-3-OXO-4-(1-OXO-1,2,3,4-TETRAHYDROISOQUINOLIN-4-YL)-2,3-DIHYDRO-1H-PYRROL-2-YL]-2-HYDROXY-PROPYL]-CARBAMIC ACID TETRAHYDRO-FURAN-3-YL ESTER
1HIH
ACETYL-NH-VAL-CYCLOHEXYL-
CH2[NCH2CHOH]CH2-BENZYL-VAL-NH-
ACETYL

BETA-MERCAPTOETHANOL

1D4H
2,5-DIBENZYL-OXY-3,4-DIHYDROXY-
HEXANEDIOIC ACID BENZYLAMIDE (2-
HYDROXY-INDAN-1-YL)-AMIDE
1D4I
2,5-DIBENZYLOXY-3-HYDROXY-HEXANEDIOIC ACID BIS-[(2-HYDROXY-INDAN-1-YL)-AMIDE]

1D4J
2,5-DIBENZYLOXY-3,4-DIHYDROXY-HEXANEDIOIC ACID 2-CHLORO-6-FLUORO-BENZYLAMIDE (2-HYDROXY-INDAN-1-YL)-AMIDE
RITONAVIR

1NPW CARBAMIC ACID 1-{5-BENZYL-5-[2-HYDROXY-4-PHENYL-3-(TETRAHYDROFURAN-3-YLOXYCARBONYLAMINO)-BUTYL]-4-OXO-4,5-DIHYDRO-1H-PYRROL-3-YL}-INDAN-2-YL ESTER

1M0B 1-BENZYL-(R)-PROPYLAMINE

AMINO GROUP

NH₃
GLYCEROL

TERT-BUTYLOXYCARBONYL GROUP

SULFATE ION

ACETYL-NH-VAL-CYCLOHEXYL-CH2[NCH2CHOH]CH2-BENZYL-VAL-NH-ACETYL

SODIUM ION

GLYCEROL

CHLORIDE ION
ACETIC ACID

(3R,3S,6R)-HEXAHYDROFURO[2,3-b]FURAN-3-YL(1S,2R)-3-[(4-AMINOPHENYL)SULFONYL](ISOBUTYL)AMINO]-1-BENZYL-2-HYDROXYPROPYLCARBAMATE