# Evolving Molecules for Drug Design Using Genetic Algorithms via Molecular Trees

**Gerard Kian-Meng Goh**
Computer Science &Chemistry Depts.
University of Idaho
Moscow, 83844-2343
gerard@uidaho.edu

**James A. Foster**
Computer Science Dept.
University of Idaho
Moscow, 83844-1010
foster@cs.uidaho.edu

## Abstract

We present a new representation for a genetic algorithm to evolve molecular structures representing possible drugs that bind to a given protein target receptor. Our representation is tree-structured with functional groups for leaves, and captures chemically relevant information efficiently. We assume a given target protein structure with known essential residues, and derive the placement of the functional groups in each chromosome from both lengths and the position of a pharmacore in the receptor. Our fitness evaluation takes into consideration both proximities and polarities of the functional groups of the evolved drug structure and the residues. Our evolved structures were intriguingly similar to known active anti-viral drug structures. Our experiments indicate that a tree-structured molecular representation and a simple evolutionary computation can design acceptable molecular structures that are potentially useful for drug design endeavors.

## 1 INTRODUCTION

One major strategy in drug design is to find or build molecules that target proteins crucial to the proliferation of microbes, cancer cells, or viruses. For example, one might design photosensitive compounds to damage such targets (Goh et al. 1997, 1999). Another design strategy, used successfully to design protease inhibitors in HIV research, is to search for compounds that bind to active protein sites which sustain viral proliferation. Much of the challenge involves accurately predicting structures of potential inhibitors, especially if the structure of the protein target is already known. This paper addresses this challenge with genetic algorithms. We use evolutionary computation to seek new molecular structures as possible drugs for a given structurally determined protein target. That is, we evolve molecular structures with their appropriate functional groups in closest proximity to crucial residues, and thereby design molecules that fit the protein receptor perfectly.

Other researchers (Venkatasubramanian et al., 1995, Glen et al., 1994, Globus et al. 1999) have used evolutionary techniques for designing pharmaceutical molecules. However, the recent literature is rather sparse, perhaps due to difficulties with accurate molecular modeling. Our approach simplifies this difficulty with by using a simple tree-structured representation, which brings several advantages. A tree structure arranges crucial data pertaining to the molecular conformations and structures in a an orderly discrete non-linear manner so as to allow easier manipulation by evolutionary operators and by operators used in other artificial intelligence techniques. Given the experimental results of this paper, we argue that the tree structure is more efficient at handling such spatial data, which involved multidimensional information that would not be effectively represented by, for example, a linear chromosome. Also, having trees of functional groups makes it possible to apply further artificial intelligence techniques, such as neural networks, in order to search for similar molecules with similar pharmacophoric properties in current chemical databases. Lastly, we suspect from the somewhat rapid convergence seen in our data that the use of functional groups in tree structure reduces the search space dramatically, since atoms can be covalently arranged together in a vast number of ways.

Our experiments do not employ precise molecular modeling, in the strictest terms of physical chemistry. Rather, our objective was to see if tree structured representations for molecular structures might make genetic algorithms to provide a powerful tool for drug design. As we shall see later, they can.

## 2 EXPERIMENTAL SETUP

### 2.1 THE MODEL

In this drug design project, we assumed that the target protein receptor structure was already known, perhaps by X-ray crystallography or nuclear magnetic resonance (NMR). Our system needed to find non-peptide molecule(s) to fit easily into this given protein structure.

Our specific target was the known antiviral binding site of the human rhinovirus strain 14. This protein site is known as the *VP1 barrel* (see Fig.1), since its structure resembles a barrel. We sought to find molecule(s) that could fit snuggly into this barrel, with the right functional groups of the new molecule in close proximity to each of the crucial amino-acid residues of the barrel.

We downloaded the protein database file of the rhinovirus VP1 protein (protein id: 1RUC) from the RCSB (Research Collaboratory for Structural Bioinformatics) Protein Databank Website, http://www.rcsb.org/pdb/. The important residues used can be seen in the illustration in Fig. 1b (taken with permission from Branden and Tooze, 1991). The crystal structure provides the three-dimensional coordinates of the barrel, which is 25Å deep and 12Å wide with a 3.1 Å resolution, and for the interactive residues.

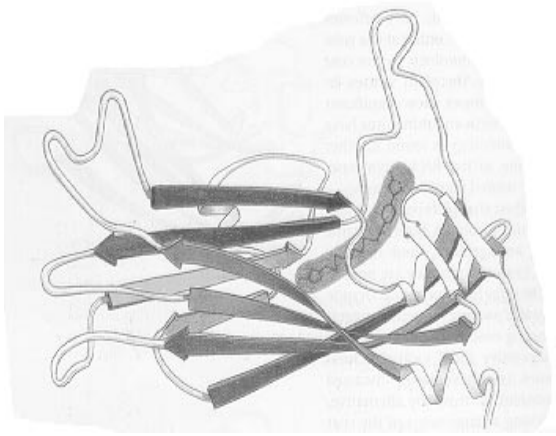


Figure 1a: Protein Structure of the Viral Target. The VP1 barrel (in shaded oval) encapsulates the drug. This protein (VP1) is found in the rhinovirus (Cold Virus) and is responsible for the virion affinity to the host cell via the latter's own protein receptor (adhesion molecule ICAM-D). (Illustration reproduced with permission from Branden and Tooze, 1991)

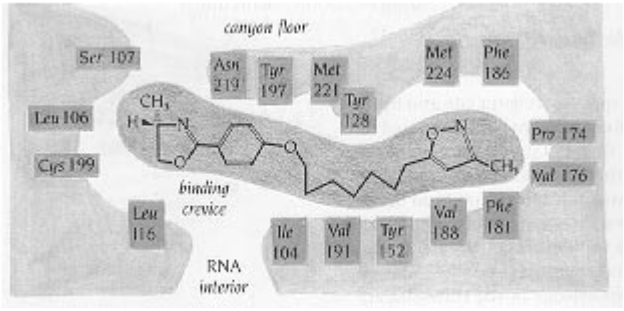



Figure 1b: Anti-viral Drug Encapsulated by the Viral Target**.** The drug shown (in center) binds to sites in the VP1 barrel. Note that this example exhibits poor binding affinity. (Illustration reproduced with permission from Branden and Tooze, 1991).

Looking at the chemical structure of the antiviral drug and the structure of the protein target, one would suspect that a portion of the antiviral compound must be present for the drug to have any activity. We may infer from bends in the barrel that any appropriate drug molecule must be flexible enough to bend in order to fit into the bent barrel. This flexible backbone is part of a *pharmacophore*. See Fig. 2 for a schematic representation.

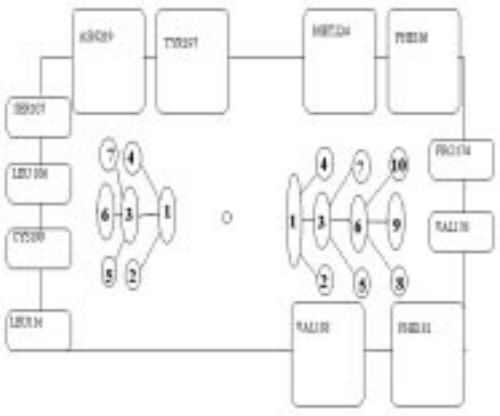


Figure 2: Schematic Representation of the Receptor Target ("Box" or "Barrel"). The residues of this protein target are shown in the rounded boxed, which contain information about the amino-acid, the sequence number and the co-ordinate. The tree structure representing the molecule we will evolve has leaves numbered in a canonical order, and contains a right and left tree.

### 2.2 REPRESENTATION

We represented our evolved molecule with two small bottlebrush shaped tree structures on each side of an

ether O-atom (see Fig. 2). Each labeled node in these trees may be filled by one of the following functional groups (see Fig. 3a): 0) Alkyl-1C (Alkyl chain with only one carbon atom), 1) Alkyl-3C (Alkyl chain with three carbon), 2) Alkyl-1C-Polar (with 1 carbon and a polar group), 3) Alkyl-3C-Polar (Alkyl with 3 carbon and a polar group), 4) Polar, 5) Aromatic (assume a benzene ring), 6) Aromatic-Polar (e.g. Hydroxyl Phenyl), 7) Empty (indicates no functional group at this location). We only allowed "Empty" (No.7) nodes at leaves. In a second experiment, we expanded to include cyclopentanes and cyclopentadiene-like groups.

Fig. 3 also illustrates the lengths of the functional groups, which will be essential to our computation of the coordinates for each of the functional groups in a molecule. Since these trees have a simple, fixed topology, we represented them with a linear chromosome in which each three bit gene encoded a single node in a tree, with the nodes listed in breadth first order and the left tree first.

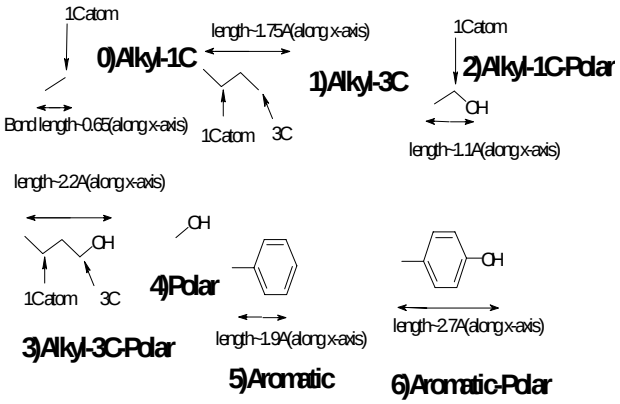



Figure 3: Functional Group Representations.

We computed the co-ordinates for the position of each functional group (the molecule represented by that node of the tree) from the length of the groups' bonds, as given in an elementary physical chemistry text (Levine, 1988). For example, alkyl-1C has a C-C bond length of 0.7Å, whereas a benzene (aromatic) ring has a length of approximately 1.4Å. For computational purposes, the bond lengths and the approximate directions of the bonds were calibrated with the coordinates of the compound in the crystal structure (D. A. Shepard et al, 1993).

## 2.3 FITNESS EVALUATION

Fitness evaluation was based on the proximity of residues (e.g. SER107) to the closest functional groups, and the chemical properties of these pairs. The distance needed to be at least 7.0Å for the molecule to gain any fitness and could not be closer than 2.0Å without being penalized. If the functional group closest to a particular residue of the receptor target was of different polarity, then a penalty was imposed. In addition, the proposed drug molecule had to be within the barrel, so a penalty was given for exceeding cavity limits. We also considered polarity, and penalized matchings with chemically dissimilar polarities. For example, the residue SER107 (See Fig. 1) is a serine group, which is a polar (hydrophilic) group. The group closest to it must also be a polar group for the drug molecule to have a higher fitness level.

To describe our fitness function more formally, let *dd* be the distance from a functional to its closest residue (in angstroms). Let *fd* be (7.0 - *dd*)/7.0. Let *pp* be a penalty value. In particular, let $pp = -|fd|$ if the polarity of the functional group and the corresponding residue are unequal, and let $pp = 10(-|fd|)$ if the position of a functional group of the tree exceeds the boundaries. The fitness of a functional group is defined to be: *pp* if the polarity of the function group is not the same as that of the residue, or if $dd < 2.0$; 0 if $fd < 0$; and *fd* if the polarity of the functional group is approximately the same as the polarity of the residue and $fd > 0$. The total fitness of a molecule is the sum of the fitness of all the functional groups, and this value is to be minimized by the genetic algorithm.

The distance of 2.0Å (in *fd*) is the approximate distance for the distance of minimum potential energy required for hydrogen bonding. The range of between 2.0 Å and 7.0 Å is the approximate distance range required for both effective van der Waal and hydrogen boding attractions (Levin, 1988).

## 2.4 IMPLEMENTATION DETAILS

We used C++ language with object oriented design methodology, and the GNU/Visual C++ compiler in both the UNIX and MS-Windows environments. The chemical drawing package, ISIS/Draw version 2.01 (obtained from Molecular Design Laboratory, Inc., San Leandro, California, http://www.mdli.com) provided measuring rulers for molecules in angstroms, which we used to validate our fitness calculation methods—especially with regards to the length and width of a given molecule.

Originally, we allowed our mutation rate be 10% using one point crossover with a crossover rate of 90%. We later decided that crossovers decrease convergence efficiency. Also, crossover would be theoretically difficult to implement in more complex molecular structure. Therefore, we used all "atomic" mutations, even though other forms of mutations are theoretically still possible. "Atomic" mutations keep track of the number of atoms in each node and randomly mutate the atoms. This also allowed mutations from double to single bonds and vice-versa.

As mentioned, our molecules were divided into two halves, represented by left and right trees. We tested trees

of various heights and found the best results with a height of five for the right side and five on the left side.

## 2.4 RESULTS AND DISCUSSION

The best-evolved right tree of the height five is illustrated in Figure 4a. A possible molecular structure that could satisfy the requirements of Figure 4a is shown in Figure 4b. The structure in Figure 4b is geometrically intriguing. It suggests two layers, one containing an aromatic/phenolic group similar to that of a known active anti-viral molecule (Figure 4b) and an underlying layer that could form a ring-like structure with the upper phenol/aromatic group. In this sense, it structurally resembles the actual known drug. It is particularly intriguing that the molecular tree was apparently able to capture the spatial details of the molecule and of the protein cavity constraints. Our genetic algorithm attempts to place functional groups as close to corresponding residues as possible, which tends to produce molecules that are large, but small enough to fit into the barrel. The "atomic" mutation enabled our algorithm to create many copies of similar structures with slight variations.

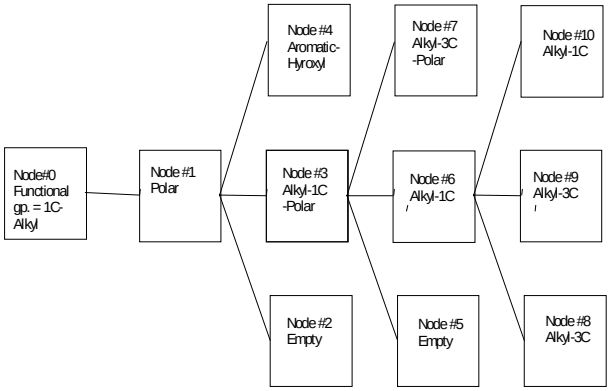



Figure 4a: Tree Representation of the Best Right Half of the Evolved Molecule After 500 Mutations.

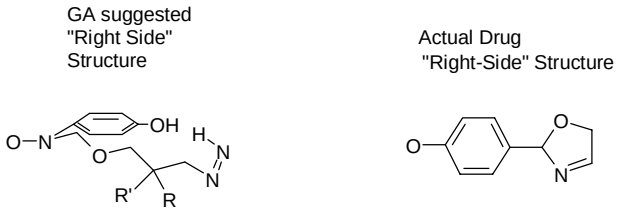



Figure 4b: Possible Structure that Satisfies the Tree Structure Requirements of Fig. 4a, Compared to the Known Anti-viral Drug.

The best tree on the left side is shown in Figure 5a. Notice the similarity of the evolved molecule seen Figure. 5b to that of a known anti-viral drug, even though the information regarding actual structure of the known anti-viral drug was never used as part of the algorithm. We may note that the functional groups at crucial locations in the actual anti-viral compound are similar in polarities to functional groups found in the corresponding positions of the evolved molecules. Furthermore, the sizes and shapes of the evolved molecules do have resemblance to the anti-viral one.

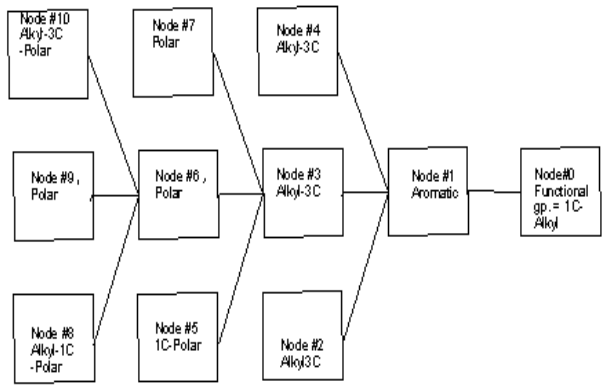



Figure 5a: Tree Representation of the Best Evolved of Left of Molecule after 500 Mutations**.**

Next, we attempted to add cyclopentane-like groups as part of the functional group. The second best right side of the molecule is shown in Fig. 6. Here again, it contains the benzene ring seen in the actual drug molecule (Fig. 6). As mentioned, the tree that describes the molecule in Fig 6 (not shown) was not the best. The best tree was able to reproduce because a single mutation on that tree caused poor fitness on its children. Fortunately, the GA does not restrict replication to just molecular trees of highest fitness, but also allows individuals with high effective fitness—those whose children tend to be fit.

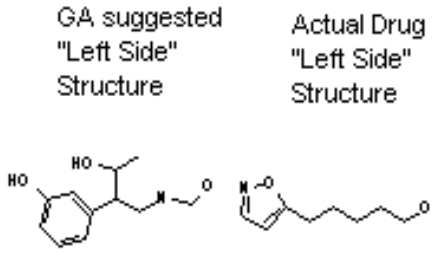



Figure 5b: Possible Structure that Satisfies the Tree Structure Requirements of Fig. 5a, Compared to the Known Anti-viral drug.

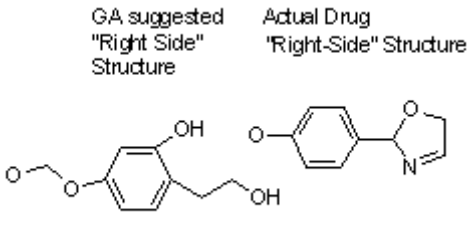



Figure 6: Possible Structure that Satisfies the Tree Structure Requirements Evolved by Adding Cyclo-Pentadiene-like Groups as Compared to Known Anti-viral Drug.

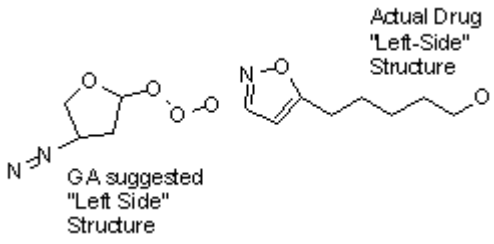
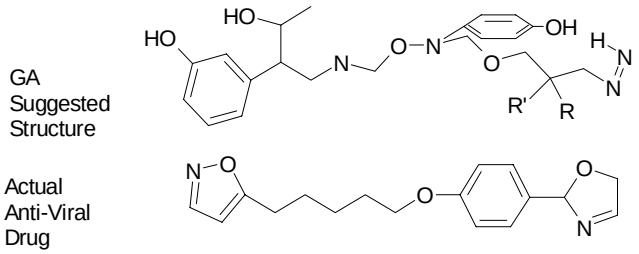



Figure 7: Possible Structure that Satisfies the Tree Structure Requirements, Compared to Known Anti-viral Drug**.**

Figure 7 shows what a molecule that could satisfy the requirement of the output of the molecular tree of the left should look like. It is interesting to note that here again the length and shape of this part of the molecular do resembles that of the antiviral drug. The present of the azide (i.e. N=N) side chain does bear similarity to the N-O atoms of the actual drug.



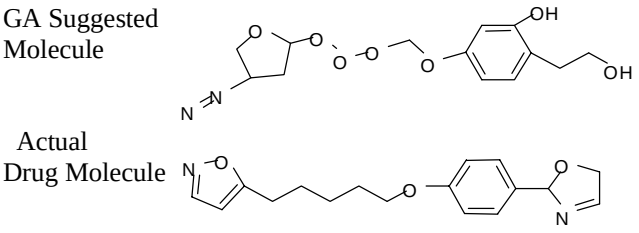




Figure 8: Result Summary**.** This summarizes the results of the our experiments. It is striking that the evolved molecules are geometrically similar in many ways to a known drug compound, despite our several simplifying assumptions, and given the GA used no prior knowledge of the molecule.

## 2.5 CONCLUSION

While our results are intriguing, there are admittedly inherent weaknesses in our model. First, we have not used any energy minimization techniques and we have assumed that the branching of the trees consist of only three children, though it is clear that for our selection of functional groups the tree could be more complicated. Second, we have assumed the bond angles to be constant. Third, we have chosen the PV1 receptor as a case study since the known drug is simple. The reason for this is that this experiment was not meant to be an exact simulation in the strictest sense of physical chemistry, but rather an attempt to evaluate a new approach to using genetic algorithms in drug design.

These simplified assumptions made it possible to test the validity of our approach without requiring extensive computing power or time. Clearly, the computational requirements and the complexities of tree structures can increase dramatically as the simulations become more exact. But this is inherent with all simulations of this scope and application, and is not a particular disadvantage of our tree structured approach.

Nevertheless, clearly our theoretical framework supports for further and more exact simulations, including energy minimization and molecules that are more complex. The conformational search via genetic algorithms of Jones et al. (1996) would be a useful supplement to our approach. There is no reason to constrain molecular trees to structures similar to the ones used in this paper. Structure discovery techniques such as ADFs or cellular encoding, from genetic programming, might help here. Nor is there any reason to limit the functional groups to the ones mentioned above. The flexibility, which allows representation of a wide variety of molecules, of the molecular tree and the convenient reduction of search space arise from the fact that the nodes contain functional groups, as opposed to atoms.

One difficulty with drug design is the complexity of geometric shape of drugs and compounds in general. We propose that molecular trees representations such as ours are more amendable to other AI search techniques such as neural networks, fuzzy logic and the creation of agents that could traverse the branches to analyze the molecular tree. We have preliminary data to confirm this suspicion.

It is also interesting to see how rapidly our population converges: after just 200 mutations. We suspect that this is the result of a streamlined search space, due to our using functional groups.

The fact that the genetic algorithm was able to come up with structures similar to the actual drug is also quite remarkable, since the program had no information about the actual drug except for the calibrations of the bond length and the approximate directions of the bond angle. Apparently, the molecular trees were able to capture three-dimensional spatial details of the molecules, and the algorithm was able to contrast its geometric properties with the protein-receptor structure, despite our simplifying assumptions. This suggests that an algorithm such as ours could be useful when the crystal structure of a potential drug receptor has already been carefully studied, even though no binding compound has yet been found.

An interesting feature of the resulting trees is that many similar trees that were variations of one another arose, though these were not globally optimal. Often, better trees died out because mutations would result in poor fitness, creating an "infertile" tree. The generations of variations of better trees will be useful if we link this program to a chemical database in order to implement drugs high-throughput drug screening.

In conclusion, there are two ways that our drug-design genetic algorithm with tree structures could be used for practical purposes in drug discovery. It could be used to search the database of known compounds for those with molecular structures that match the evolved tree structures, especially if the protein receptor target is already known. It could also provide a powerful exploratory tool for the medicinal synthetic chemist, who could attempt to synthesize the evolved molecular structures.

Given that our current simplified model with a straightforward genetic algorithm has discovered molecules very similar to known antiviral drugs, we are eagerly pursuing improvements. We look forward to using simulated evolution to combat some of the viruses which natural evolution has given us.

**Acknowledgments**

The authors would like to thank Dr. John Tooze of the Imperial Cancer Research Fund, United Kingdom, for his permission to reproduce the mentioned figures that concerned the viral protein. The authors are also grateful to Dr. John Correia, Department of Chemistry, St. Mary's College of California for his helpful comments. The second author (Foster) was support by NIH grant NIH F33 GM20122-01.

**References**

Branden C. and Tooze J. (1991*). Introduction to Protein Structure*, Garland, New York.

Glen R. C. and Payne A. W. R. (1995). A Genetic Algorithm for the Automated Generation of Molecule within Constraints. *Journal of Computer-Aided Molecular Design*, 9, 181-202.

Globus A., Lawtonb J., Wipke T. (1999). *Sixth Foresight Conference on Molecular Nanotechnology.* **10**, 290-299.

Goh G. and Czuchajowski L. (1997). The Synthesis of Isomeric Dihymidyl-phosphorus(V) tetraphenylporphyrins, *Journal of Porphyrins and Phthalocyanines*, **1**, 281-285.

Goh G., Gajewski M., and Czuchajowski L. (2000). Phosphorus(V) Porphyrin Diaxially Substituted with Amino Acids and Crown Ethers, *American Chemical Society National Meeting, San Francisco*.

Jones G., Willett P., and Glen R. C. (1996). Genetic Algorithms for Chemical Structure Handling and Molecular Recognition, Genetic *Algorithms in Molecular Modeling*, 212.

Levine I., (1988). *Physical Chemistry*, McGraw Hill, New York.

Rodgers, D., Lam, P. Y. S. and Erickson-Viitanen, S. (1998). Design and Selection of MP850and DMP 851:The Next Generation of Cyclic Urea HIV Protease Inhibitors, *Biochemistry* **5**, 597.

Shepard, D. A., Heinz, B. A., and Rueckert, R. R. (1993). Win 52035-2 Inhibits Both Attachment and Elipse 1 RUC 32 of Human Rhinovirus 14, *Journal of Virology* **67**, 2245.

Venkatasubramanian, V., Chan, K., and Caruthers, J. M. (1995). Evolutionary Design of Molecules with Desired Properties Using the Genetic Algorithm, *Journal Chemical Information Computer Sci**ence*** **35**, 188-195.