Interactive evolutionary algorithms and data mining for drug design

Lameijer, E.M.W.

Summary

Designing drugs is hard. For many illnesses there are no effective drugs available (for example, for multiple sclerosis), in other cases existing drugs are not always effective enough (cancer). And even when effective treatment is available (for example for HIV), the drugs may have side effects that investigators and patients alike want to diminish. Chapter 1 introduces this thesis, and explains why it is so hard to design drugs. It also discusses how computers could help in the drug design process, as well as the advantages of effective collaboration between human drug designers and computers. Chapter 1 also explains some of the terms that are used throughout this thesis, such as data mining and docking, and it closes with an overview of the rest of the chapters.

Chapter 2 discusses one of the core methods used in our investigations: evolutionary algorithms. It describes the basic theory of evolutionary algorithms, and gives an overview of the different uses of evolutionary algorithms in drug design, varying from library design to QSAR and docking. It then compares these applications of evolutionary algorithms, and concludes with suggestions for future applications and further research.

The other main computational method we used, data mining, is discussed in Chapter 3. In data mining, one attempts to find patterns or rules in databases. In this chapter, we investigated a database of molecules, the NCI database, which, at the time of our investigation, contained 250,251 molecules. We were especially interested in how diverse this database was, since having a collection of diverse molecular structures leads to a greater chance of finding potential new drug molecules. We split all molecules in the NCI database into ring fragments and non-ring fragments, and determined the frequency of these fragments, as well as the frequency with which fragments occurred together in molecules. The molecules in the NCI database turned out to be remarkably undiverse: while there were over 10,000 different ring structures, the ten most frequent ring systems (<0,1% of the total ring diversity) accounted for 62% of all the ring systems found in the molecules. On the other hand, thousands of ring systems occurred in only one or two molecules. We also found that many substructures occurred much more often with each other than would be expected by chance, forming 'superfragments', and, conversely, that there were also substructures which seemed to avoid each other. The data mining thus yielded lists of rare fragments and fragment combinations, which could be used to design new molecules. Such new molecules, when added to existing libraries, would increase its molecular diversity and thereby increase the chance of finding new drugs.

In Chapter 4 we return to evolutionary algorithms, in particular to the challenge of designing new drug molecules. When we started our study, there had been many investigations using evolutionary algorithms to design candidate drug molecules. However, all these approaches suffered from the problem that it is not yet possible to calculate a suitable fitness function, namely how good a certain molecule would be as a drug. This lack of quality fitness functions made optimization unreliable at best and meaningless at worst. Therefore, we decided to use a different approach, that of an interactive evolutionary algorithm. Interactive evolutionary algorithms use feedback from the user as fitness criterion; so in our case, chemists had to select the molecules they liked (which means, the molecules that seemed novel, drug-like and possible to synthesize without too many problems). In this way we could use the chemists' experience and chemical intuition, which would have been very difficult to 'extract' from their brains and program a computer with. We called the resulting interactive evolutionary algorithm the 'Molecule Evoluator'. Chapter 4 describes the design of the Molecule Evoluator, as well as the coding of the molecules, the different mutations, and the extra features that were needed to make the program usable for chemists and allow optimal use of their creativity.

One of the problems in the first versions of the Molecule Evoluator was that it created many molecules which either were unstable or would be difficult to synthesize. This annoyed chemists, leading them to abandon the optimization process prematurely. To increase the chance that molecules created by the Molecule Evoluator would be chemically sensible, we applied the technique discussed in Chapter 3, data mining. We mined the World Drug Index, a database that, at the time of our investigation, contained the structures of 32,000 drug molecules. In this collection of drug molecules, we determined the relative frequencies of small substructures containing one to four atoms, and used these data to adapt some of the Molecule Evoluator's mutations. We found that this indeed increased the chance that the Molecule Evoluator produced realistic molecules, while at the same time not forbidding substructures which actually occur in drugs. The exact procedures of the data mining, the adaptation of the mutation functions and the methods used to test whether the new mutations improved over the old ones are described in Chapter 5.

Our literature study for Chapter 2 encountered many evolutionary algorithms for designing drug molecules. However, it was difficult to compare the quality and effectiveness of these methods because they had not been compared to each other. This is unfortunate, since proper validation of molecules, either experimentally or by advanced computational methods, costs significant amounts of time and money.

Therefore it is important to make one's evolutionary algorithm as efficient as possible: an evolutionary algorithm which needs to evaluate only 1000 molecules to reach a certain improvement in biological activity would be much preferable over an evolutionary algorithm that needs 100,000 to reach a similar result. But what kind of evolutionary algorithm would be most efficient? Chapter 6 describes our investigation into the importance of one of the main factors in evolutionary algorithms for molecule design, namely whether evolution is atom-based (molecules are changed by adding, changing or removing single atoms), or fragment-based (where mutations add, change or remove multi-atom fragments). We evaluated the molecules by docking them into the reverse transcriptase enzyme of HIV. Docking was chosen as a fitness function here since it approximated the binding of molecules to a real protein with its irregular, three-dimensional cavity. While the experiment did not show clear superiority of either atom-based evolution or fragment-based evolution, a closer study of the optimization processes yielded several insights on how we could improve our evolutionary algorithms. The details of the experiment, as well as the suggestions for future improvement of evolutionary algorithms for drug design, can be found in Chapter 6.

In theory, the Molecule Evoluator we described in Chapter 4 should be able to help chemists design new, biologically active molecules. However, does this also happen in practice? Chapter 7 describes an experiment in which we let the Molecule Evoluator generate 300 structures, of which chemists chose 34 for further investigation. By consulting chemical databases, we determined which of those molecules were novel. Subsequently, the chemists chose a number of the novel molecules and synthesized eight of them. These eight compounds were tested on 83 proteins, and four turned out to be biologically active in some way, amongst others on the α-adrenergic receptor. This result indicated that the collaboration between chemist and computer using the Molecule Evoluator can indeed create novel and biologically active compounds, which may be good leads for new drugs. Chapter 7 gives the detailed experimental protocol, the structures of the molecules synthesized, and the biological test results.

This thesis concludes with Chapter 8. The first part of this chapter contains the general conclusions of the investigations described in this thesis. The second part of Chapter 8 is dedicated to future perspectives: thoughts on the further development of the Molecule Evoluator and on the future role of evolutionary algorithms in drug design. It concludes with my vision on the directions that software for drug design could or should take in the future.