A Case Study in Machine Learning for Combinatorial Chemistry

David Page, **Steven Curtis**, **James H. Graham**
Speed Scientific School
University of Louisville
Louisville, KY 40292

Arno F. Spatola
Department of Chemistry
University of Louisville
Louisville, KY 40292

Abstract

Combinatorial chemistry is an advance of this decade that makes it possible to synthesize over a million compounds in a day. These compounds can then be tested for biological activities of various kinds via high-throughput screening, in the search for new drug leads. Unfortunately, the active compounds may have undesirable properties such as metabolic instability or toxicity. Nevertheless, machine learning can be used to determine what the active compounds have in common structurally, in contrast to the inactive compounds; this information can then be used to guide the design of improved structures. This paper describes the challenges raised by this application of machine learning, and it presents a case study of the approach. The paper focuses on the computational issues involved, as opposed to the chemical and biological issues.

1 Introduction

Recent experiments have demonstrated that machine learning can be used successfully for structure-activity prediction [4, 3, 2]. Structure-activity prediction is a key step in the drug design process, described in the next section. In the recent experiments, the examples provided to the machine learning systems were descriptions of molecules; each molecule was labeled as either *active* (a positive example) or *inactive* (a negative example). Due to the widespread use of combinatorial chemistry—also described in the next section—the data used by chemists at academic institutions and pharmaceutical companies sometimes label only *sets* of molecules rather than individual molecules. Such labeling makes the machine learning task much more difficult. This paper describes a case study in the application of machine learning to a data set developed using combinatorial chemistry. Section 2 describes the context of this application, Section 3 describes the case study with an emphasis on the particular challenges raised by this application, and Section 4 presents the conclusions.

2 The Context

The Drug Design Process A drug is a small molecule that works by binding to a target protein in the body, in a way that modifies the protein's behavior. Therefore, the first step in drug design is to identify a target protein and binding site—the target site. If the exact molecular and 3-dimensional structure of the target site is known, then a new molecule can be designed that will fit into it. Unfortunately, the details of the target site typically are unknown. Nevertheless, it is usually possible to design chemical tests—assays—to determine whether any given molecule will bind to the target site. Therefore, researchers often test large libraries of compounds in the hope of finding one or more "lead" compounds capable of binding to the target site. A lead compound may not be acceptable as a drug because of undesirable properties, e.g., it may be toxic, it may not be active or selective enough, it may be metabolically unstable (breaking down when it enters the body), or it may not diffuse from the gut to the bloodstream (so it cannot be taken by mouth). However, given several active compounds, researchers may be able to determine what these compounds have in common structurally that gives rise to their ability to bind to the target. This determination is called *structure-activity prediction*: the attempt to predict the activity of a molecule (its ability to bind to the target site) based on its structure. Once the structural properties are identified that are responsible for the desired activity, it is often the case that new molecules can be synthesized that possess these attributes without the undesirable ones.

Machine Learning for Structure-Activity Prediction In its simplest form, structure-activity prediction is an ideal example of an inductive learning problem. Active molecules constitute positive examples, molecules with little or no activity constitute negative examples, and the task is to learn a structural description that will distinguish the positive examples from the negative ones. Furthermore, the application of machine learning to structure-activity prediction has great potential benefits. Better methods for structure-activity prediction could substantially reduce the time (typically 3-5 years) it takes to move from a lead compound to a potential drug that is ready to undergo intensive testing including clinical trials.

Combinatorial Chemistry For any given target, the drug design process we have described is possible only if the chemical library being tested contains some active compounds, or "hits". The probability of obtaining "hits" for a given target can be raised by increasing the size and diversity of the library of compounds being tested. Combinatorial chemistry is a major advance of this decade that makes it possible to synthesize a very large number of diverse compounds in a short time. While the details of combinatorial chemistry techniques are beyond the scope of this abstract, one aspect of these techniques is central to our discussion: combinatorial chemistry techniques result in the creation of *mixtures* of compounds. If a given mixture is found to be active against a target site of interest, then we know only that at least one of the compounds in the mixture is active, without knowing which one(s). Section 3 discusses the implications of using such mixtures as examples in machine learning.

3 Applying Machine Learning to Data Sets Constructed by Combinatorial Chemistry

When applying machine learning to a chemical library built using combinatorial chemistry, an example (positive or negative) is no longer a single molecule but a set of molecules. For a negative example, we know that every molecule in the set is inactive, but for a positive example, we know only that at least one of the compounds is active. Therefore, the learning problem is a case of the *multiple instance problem*, first defined by Lathrop, Dietterich, and Lozano-Perez [1], and further addressed by Finn, Muggleton, Page, and Srinivasan [2]. The latter work showed that the multiple instance problem can be approached quite naturally using a particular kind of machine learning known as *inductive logic programming*. Nevertheless, while the difficulties posed by combinatorial chemistry are qualitatively similar to those faced by Finn et al., they are quantitatively more difficult. In the earlier work examples were individual molecules, but the multiple instance problem arose because the molecules could take on multiple 3-dimensional *conformations*, and it was not possible to determine which conformation(s) permitted the active molecules to bind to the target site. In the present work, again each molecule can take on multiple conformations (in fact, typically more conformations than in the earlier work), and in addition, each example is a group of eight molecules rather than a single molecule. Thus the present work tests whether the general approach of Finn et al. can be scaled up to meet the demands imposed by combinatorial chemistry.

3.1 Materials and Method

The initial data set consisted of 48 mixtures of pseudopeptides (peptides with amide bond replacements) synthesized by modified solid phase methods and cleaved from the polystyrene matrix [6]. Each mixture had a recorded level of activity, where the activity in question was efficacy against a strain of antibiotic-resistant bacteria. Because of concern over computational complexity, only the seven most highly active mixtures were chosen as positive examples, while four closely-related but completely inactive mixtures were chosen as negative examples. Each mixture consisted of eight compounds. Each compound had a large and flexible molecular structure; the low-energy three-dimensional conformations, or *conformers*, of each molecule were computed using the Sybyl molecular modeling package from Tripos (St. Louis, http://www.tripos.com). Sybyl found 50 to 70 conformers for each molecule, so that in effect each mixture contained a total of roughly 500 distinct structures. The task was then to find the largest 3-dimensional substructure such that at least one member of every active mixture possessed that substructure, while no member of any inactive mixtures possessed the substructure.

The kind of 3-dimensional substructure sought in this work is often referred to as a *pharmacophore*. A pharmacophore is that part of a drug molecule that interacts with the target site and is thereby responsible for the biological response to the drug. Typical components of a pharmacophore are hydrophobic groups and charged atoms or groups. The geometric arrangement of these atoms or groups is also a vital part of the description of a pharmacophore. The task in the present work, therefore, was pharmacophore discovery—specifically, to determine from the ensemble of mixtures the largest number of components, with their geometric arrangement, common to the active subset. For illustration, Figure 1 shows a 6-point pharmacophore within a molecule; the pharmacophore consists of two hydrophobic groups and four positively-charged groups in a particular geometric arrangement.

The machine learning approach was similar to that used by Finn et al. [2]. That earlier work used the inductive logic programming system PROGOL [5]. PROGOL focuses its search on one (arbitrary) positive example, building potential hypotheses that cover or explain that example. PROGOL searches the hypothesis space top-down, building progressively larger and more complex hypotheses. It returns the simplest (textually shortest) hypotheses that correctly label all the examples, that is, that cover all of the positive examples and none of the negative examples. Nevertheless, for the present application, a theoretical analysis revealed that the computation time was likely to be lower for a bottom-up search (considering more complex hypotheses first) than for a top-down search. Therefore, the approach used in the present work was PROGOL-like, in that it focused it search using a single positive example, but the search was bottom-up rather than top-down. The result of the search consisted of the largest hypotheses consistent with the data; the hypotheses represent the largest potential pharmacophores for the activity in question.

3.2 Results

The machine learning algorithm found 12 largest potential pharmacophores for the activity in question. Each potential pharmacophore involves two hydrophobic groups and four positively-charged groups, and they differ from each other in their geometric arrangements. Figure 1 in fact shows one of the molecules in the first mixture, in one of its 3-dimensional conformations, with the points of the first pharmacophore labeled. The run-time for obtaining these results was approximately two weeks real-time on a SUN Ultra.

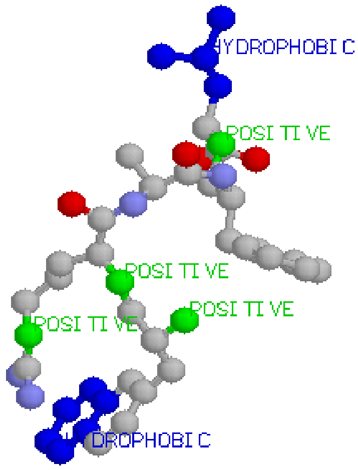

Figure 1: The figure shows one of the molecules in the data set in a low-energy conformation for which it exhibits the first potential pharmacophore found by the machine learning algorithm. The two hydrophobic points and four positively-charged points of the pharmacophore are labeled. Hydrogen atoms have been removed to make the figure easier to view.

In addition, precomputation of conformers using Sybyl also required approximately two weeks real-time on a SUN Ultra.

The use of additional active and inactive mixtures perhaps could have reduced the number of alternative hypotheses consistent with the data. Recall that in the case study, seven highly active mixtures and four inactive mixtures were selected, involving a total of roughly 5000 conformers. Many more inactive mixtures were available for use as data; some additional active mixtures also were available, although their activity levels were less than for the seven selected mixtures. Nevertheless, computation time (worst-case or average-case) for the learning algorithm grows linearly with the total number of conformers used, as does the time to precompute the conformers themselves. Therefore, for example, doubling the number of mixtures used would give approximately 10,000 conformers, which would double the computation time, requiring a month for the precomputation of conformers and a month for the machine learning run.

As an alternative to repeating the machine learning run with additional data, focused experimentation can now be used to determine which of the competing pharmacophores is correct. Each potential pharmacophore predicts which particular molecule(s) from each active mixture will be active. Although the twelve hypotheses largely agree on the active molecules for four of the seven active mixtures, there is wide disagreement on the remaining three mixtures. These points of disagreement can be used as distinguishing experiments. For example, if for a given mixture the first hypothesis (potential pharmacophore) predicts molecule number one of the mixture will be active while the others predict it will be inactive, then molecule number one can be synthesized and tested independently of other molecules. If molecule number one is active, it confirms the first hypothesis, and if it is inactive, it refutes the first hypothesis.

4 Conclusions

On the positive side, the case study has shown that machine learning—in particular, inductive logic programming—can be used with data sets developed through combinatorial chemistry, even though the data points are mixtures of flexible molecules. The case study also has shown that even with a relatively small number of data points, the number of competing hypotheses can be reduced to a manageable number which potentially can be trimmed by experimentation. On the negative side, the case study indicates that computational complexity will prohibit the use of substantially larger data sets or substantially larger mixtures. This is a significant shortcoming, since it would not be surprising to see a data set of 1000 mixtures in which each mixture contains 1000 compounds. The following paragraph suggests two potential approaches to the computational limitations.

The use of parallelism is one potential approach to the complexity of this machine learning task. The algorithm used in the present case study is highly amenable to parallel execution: each hypothesis can be tested against each conformer independently of all other hypotheses and all other conformers. Nevertheless, while parallelism can substantially reduce the time for the machine learning run, the time for the precomputation of conformers remains a practical problem. A second potential approach comes from a relatively recent development in combinatorial chemistry. Combinatorial libraries of *cyclic* peptides and pseudopeptides have been produced and are being tested for various biological activities [7]. Cyclic peptides and pseudopeptides are more rigid than their linear counterparts. Therefore, it may be possible to use effectively only one conformer per molecule. This would allow a machine learning algorithm to handle more molecules per mixture, and it also would eliminate the need for time-consuming precomputation of multiple conformers. Furthermore, the advantages of using cyclic structures can be combined with the benefits of parallelism. It is hoped that this combination might make it possible to process bioassay data obtained from high-throughput screening of combinatorial libraries containing thousands of mixtures, where each mixture has thousands of compounds.

References

[1] T. Dietterich, R. Lathrop, and T. Lozano-Perez. Solving the multiple-instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89(1-2):31–71, 1997.

[2] P. Finn, S. Muggleton, D. Page, and A. Srinivasan. Discovery of pharmacophores using Inductive Logic Programming. *Machine Learning*, 30:241–270, 1998.

[3] A.N. Jain, K. Koile, B. Bauer, and D. Chapman. Compass: Predicting biological activities from molecular surface properties. *Journal of Medicinal Chemistry*, 37:2315–2327, 1994.

[4] R. King, S. Muggleton, R. Lewis, and M. Sternberg. Drug design by machine learning: The use of inductive logic programming to model the structure-activity relationships of trimethoprim analogues binding to dihydrofolate reductase. *Proceedings of the National Academy of Sciences*, 89(23):11322–11326, 1992.

[5] S. Muggleton. Inverse entailment and Progol. *New Generation Computing*, 13:245–286, 1995.

[6] A.F. Spatola. Peptide backbone modifications: A structure-activity analysis of peptides containing amide bond surrogates, conformational constraints, and related backbone modifications. In B. Weinstein, editor, *Chemistry and Biochemistry of Amino Acids, Peptides, and Proteins*, pages 267–357. Marcel Dekker, New York, 1983.

[7] A.F. Spatola, Y. Crozet, D. deWit, and M. Yanagisawa. Rediscovering an endothelin antagonist (bq-123): A self-deconvoluting cyclic pentapeptide library. *Journal of Medicinal Chemistry*, 39:3842–3846, 1996.