Using Extended-Connectivity Fingerprints with Laplacian-Modified Bayesian Analysis in High-Throughput Screening Follow-Up

DAVID ROGERS, ROBERT D. BROWN, and MATHEW HAHN

This article describes the use of a combination of extended-connectivity fingerprints (ECFPs) and Laplacian-modified Bayesian analysis in a study of the inhibition of *Escherichia coli* dihydrofolate reductase. The McMaster High-Throughput Screening Lab at McMaster University proposed a competition to predict the hits in a separate test set of 50,000 compounds. Although the problem seemed best approached with 3D methods, the authors show that 2D methods offer surprisingly competitive results with a low computational cost. (*Journal of Biomolecular Screening* 2005:682-686)

**Key words:** extended-connectivity fingerprints, Laplacian-modified Bayesian analysis, high-throughput screening, computational methods

**INTRODUCTION**

In 2003, SCIENTISTS AT McMASTER UNIVERSITY published a paper describing high-throughput screening (HTS) of a diverse library of 50,000 small molecules against *Escherichia coli* dihydrofolate reductase (DHFR) to detect inhibitors. Thirty-two compounds were identified as having significant inhibitory activity against the enzyme. These screening results were to be used as the training set in a competition cosponsored by the university and the *Journal of Biomolecular Screening*. The participants would use the experimental “training data set” of 50,000 compounds (and any other publicly available data they wished to use) to predict the activity of an additional 50,000-molecule “test library.” Although this problem seems best suited for 3D methods (e.g., docking or pharmacophore models), given that DHFR binding is known to require a precise 3D alignment of the ligand into the active site, we decided to participate using rapid 2D methods available in the SciTegic product Pipeline Pilot™ (version 4.5). These 2D methods, particularly extended-connectivity fingerprints (ECFPs) and Laplacian-modified Bayesian analysis, have been previously used against HTS applications, as reported by Xia et al. It appeared to us that enough information might be available in the 2D information to extract useful information and compete favorably against traditional 3D approaches.

**METHODS**

*Extended-connectivity fingerprints*

ECFPs are a class of 2D fingerprint for molecular characterization. They are based on a process derived using a variant of the Morgan algorithm, which was originally proposed as a method for solving the molecular isomorphism problem (i.e., identify when 2 molecules, with different atom numberings, are the same). In the Morgan algorithm, an iterative process assigns numeric identifiers to each atom, using the identifiers from the previous iteration. The identifiers are independent of the original numbering of the atoms. The process is continued until all atoms have been maximally disambiguated; the intermediate results are discarded, and the final identifiers are used to propose a canonical numbering scheme for the atoms.

Rather than discarding the intermediate atom identifiers, the ECFP algorithm captures them as *features*. (Variants of this type of intermediate feature capture have been seen as early as the DARC program.) Indeed, obtaining these partially disambiguated atom identifiers is the *goal* of the process. This means that the iteration does not have to proceed to completion (i.e., maximum disambiguation), saving the computational effort required when Morgan algorithm variants are used for canonicalization.

The width in bonds of the largest circular substructure in the final iteration of the algorithm is the “size” of the fingerprint and is appended to the name. Thus, ECFP_6 contains all circular substructures, around each atom, up to a maximum width of 6 bonds.

ECFPs have many useful qualities: they can be rapidly calculated; they can represent a very large number of different features...
ECFPs with Laplacian-Modified Bayesian Analysis

(up to 4 billion); features are not predefined and so can represent novel structural variation; features can be interpreted as the presence of particular substructures, allowing easy interpretation of analysis results; and different initial atom identifiers can be used to generate different fingerprints with different uses.

Two different fingerprint variants are discussed in this article, each based on different initial atom identifiers. One fingerprint variant, standard ECFPs, is intended to capture explicit atom-type information and uses specific atom types as initial atom identifiers. The other fingerprint type, termed functional-class fingerprints (FCFPs), is intended to capture general functional roles of atom-and abstract-specific identifiers into 1 of several general atomic classes (e.g., hydrogen-bond donor, acceptor, aromatic, halogen, etc.).

Extended-connectivity fingerprints lead to a high-dimensional representation of the molecular data; rather than “folding” the fingerprints down to a much smaller fixed size (such as 1024 bits), each molecule is annotated with a list of its own features. (Although the lists of features that define ECFPs and FCFPs are treated as unitary descriptors, the features themselves must be “unpack” into separate descriptors for many analysis programs.) Table 1 shows the number of different ECPF and FCPF features within the first 50,000 molecules in a number of different compound libraries.

In all cases, FCPF_6 contains fewer features than ECPF_6, but that is expected, as the atom types were abstracted into a smaller list of functional classes. Combinatorial libraries, because they reuse substructural parts over and over, have the fewest features. Typical vendor starting material libraries contain fewer features than drug combinda, which are the most diverse.

Finally, although extended-connectivity fingerprints were developed for use with our modified Bayesian analysis, they have also proven useful in similarity-based virtual screening. Recent work by Hert et al.3 comparing different fingerprints gave the advantage to either ECFPs or FCFPs over other fingerprints studied.

Laplacian-modified Bayesian analysis

Laplacian-modified Bayesian analysis was developed to take advantage of the high-dimensional representation of molecules provided by extended-connectivity fingerprints, as well as address the problems caused by modeling the large data sets generated by HTS. The modeling process creates a predictive model from the training data that can then be applied to score samples in the test set. The score can be used to prioritize samples for screening.

Early work by Labute4 used Bayesian analysis against a relatively small set of continuous, numeric descriptors. However, the standard naive Bayesian formulation proved difficult to apply to high-dimensional fingerprint data, as the decorrelation step is expensive or impossible for a table that contains possibly millions of columns (1 column for each unique feature in a library). Instead, the Bayesian method is modified by considering only the effect of the presence of a feature and not its absence; to correct for the differing sampling rates of different features (e.g., methyl vs. isocyanate), a Laplacian correction is used. Details of the method are described by Xia et al.2 (Another Bayesian-based method using fingerprints is described in Bender et al.)

Laplacian-modified Bayesian analysis combined with extended-connectivity fingerprints is especially useful for high-throughput data analysis because it is fast, is easily automated, and scales linearly with the number of samples. The low computational effort means it can be used both early in a project and then ongoing as the project develops.

EXPERIMENTAL

The McMaster’s training data consisted of 50,000 molecules with 2 experimentally measured residual activities. An active molecule was determined by consensus scoring: that is, the percent residual activity must be less than 75 in both values for a sample to be labeled a hit. Thirty-two compounds in the data set were active by this measure.

A Bayesian model was built from the training data using the following descriptor set: AlogP, molecular weight, the number of hydrogen-bond donors, the number of hydrogen-bond acceptors, the number of rotatable bonds, and either ECPF_6 or FCPF_6. (The set using ECPF_6 contained the results submitted to the challenge.)

Process validation

Before proceeding with model building on the entire data set, it is important to validate that the process would build a useful model if it were given data “sufficiently similar” to the samples in the training set. This was done using 10-fold cross-validation. In this procedure, one tenth of the samples were left out, and a model was built using the remaining samples; that model was used to predict the scores for the left-out samples. This was repeated until all samples had a prediction. The samples were then sorted by decreasing score, and a receiver operating characteristic (ROC) plot was used to estimate the predictiveness of the modeling process. This validation took about 10 min to perform. The ROC plot is shown in Figure 1.

Table 1. Number of Unique ECPF_6 and FCPF_6 Features in Different Compound Libraries

<table>
<thead>
<tr>
<th>Source of 50,000 Molecules</th>
<th>ECPF_6 Features</th>
<th>FCPF_6 Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple combinatorial library</td>
<td>1292</td>
<td>3738</td>
</tr>
<tr>
<td>MDL Available Chemicals Directory</td>
<td>77,118</td>
<td>176,088</td>
</tr>
<tr>
<td>Asinex</td>
<td>75,955</td>
<td>146,473</td>
</tr>
<tr>
<td>Maybridge</td>
<td>90,490</td>
<td>183,115</td>
</tr>
<tr>
<td>Derwent World Drug Index</td>
<td>152,073</td>
<td>315,782</td>
</tr>
<tr>
<td>MDL Drug Data Report</td>
<td>187,985</td>
<td>394,844</td>
</tr>
</tbody>
</table>

This process was repeated using FCFP_6 instead of ECFP_6, and the resulting ROC score was lower: 0.86. Because ECFPs slightly outperformed FCFPs on the training set, ECFP_6 was chosen for building the model used in the competition.

**Modeling**

Having validated the process, we built a model using the entire data set (this took about 40 s). The model was then used to sort the samples of the data set and generate a ranked list. For the purposes of the competition, only the first 2500 samples in the ranked list were used in scoring. Two scoring functions were used to determine which compounds were hits: *consensus residual activity* (CRA) and *average residual activity* (ARA). A compound is a hit using CRA if *both* of the assays reduced the percent residual activity of DHFR below the cutoff value of 75%. A compound is a hit using ARA if *either* of the assays reduced the percent residual activity of DHFR below the cutoff value of 75%.

When the test set results were released, it was found that we had 2 hits in the top 2500 (using CRA) and 7 hits in the top 2500 (using ARA). For comparison, Tanimoto similarity (with FCFP_4 fingerprints) had 2 hits (either CRA or ARA) in the top 2500 most similar to a training active.

An advantage of 2D fingerprint-based methods such as ECFPs is the speed at which large data sets can be processed. The validation study took under 8 min to run; the model building itself took 40 s; the application of the model to the test data took 30 s. This allows the model to be used against in-house libraries of millions of compounds. Even if a 2D method provides only a percentage of the information that a careful 3D study can provide, its low cost means its use should be considered when speed is important or when the initial results can assist in improving or focusing later 3D efforts.

### Data set comparisons

How much confidence should we have in the applicability of the model against the test data? We previously stated that you can have confidence if the test set is “sufficiently similar” to the training set. More precisely, statistical theory requires the test data and the training data be “drawn from the same distribution of samples” for the validation to be a good estimator of the true enrichment. We used extended-connectivity fingerprints to quantify how good an assumption this is.

Extended-connectivity fingerprints generate a large number of different features; for typical libraries of 50,000 molecules, there are tens or hundreds of thousands of different features, representing the different substructural elements contained in those molecules. More diverse libraries will have a greater number of different features. Furthermore, we can quantify how similar 2 libraries are by calculating the percentage of features they share. (Fingerprint feature comparison operations are simple counting exercises that can be performed in minutes, even on large libraries.)

To quantify the concept of “drawn from the same distribution of samples,” we do exactly that: we randomly split a library into 2 sublibraries (called “A” and “B”), and then we look at the percentage of shared features. Our hypothesis is that a test library “drawn from the same distribution” as the training library will have a similar percentage of shared features. If the percentage of shared features is much smaller, then our confidence in the applicability of the validation study will be reduced. Table 2 shows the counts calculated using the competition’s training and test data sets.

<table>
<thead>
<tr>
<th>Data Set</th>
<th># FCFP_6 Features</th>
<th>Standard Deviation</th>
<th>Shared Features A/B</th>
<th>Standard Deviation</th>
<th>% Shared</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td>82,255</td>
<td></td>
<td>39,440</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Training/2 A</td>
<td>60,732</td>
<td>248</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Training/2 B</td>
<td>60,964</td>
<td>240</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Test</td>
<td>39,006</td>
<td></td>
<td>21,616</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Test/2 A</td>
<td>30,286</td>
<td>153</td>
<td></td>
<td></td>
<td>84</td>
</tr>
<tr>
<td>Test/2 B</td>
<td>30,336</td>
<td>173</td>
<td></td>
<td></td>
<td>71</td>
</tr>
<tr>
<td>Mixture A</td>
<td>77,341</td>
<td>277</td>
<td>13,676</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Training/2 A</td>
<td>60,732</td>
<td>248</td>
<td></td>
<td></td>
<td>23</td>
</tr>
<tr>
<td>Test/2 A</td>
<td>30,286</td>
<td>153</td>
<td></td>
<td></td>
<td>45</td>
</tr>
<tr>
<td>Mixture B</td>
<td>77,737</td>
<td></td>
<td>13,721</td>
<td></td>
<td>88</td>
</tr>
<tr>
<td>Training/2 B</td>
<td>60,964</td>
<td>240</td>
<td></td>
<td></td>
<td>23</td>
</tr>
<tr>
<td>Test/2 B</td>
<td>30,336</td>
<td>173</td>
<td></td>
<td></td>
<td>45</td>
</tr>
</tbody>
</table>

FCFP, functional-class fingerprint.
sublibraries A and B (e.g., Training/2 A). The number of different features using FCFP_6 was counted for each library and sublibrary. The self-similarity of a library was estimated by looking at the percentage of features in each sublibrary shared with the other sublibrary. The results were averaged over 10 runs. (The standard deviations of the percent shared results were all less than 1%.)

One unexpected result is that the count of FCFP_6 features for the training set (82,255) is more than double that of the test set (39,006). The training set is apparently more diverse than the test set. This strongly suggests a violation of the “drawn from the same distribution” requirement.

We can measure the self-similarity of the test and training libraries by looking at the percentage overlap between their respective sublibraries. Both values are high: for the training data, just under two thirds of the features are shared (65%), whereas for the test data, over two thirds of the features are shared (71%). It is likely that the higher overlap for the test data is due to its poorer diversity.

Finally, we can look at mixed libraries, created by mixing sublibraries from each of the training and test sets. Both mixture experiments gave similar results, so we will discuss mixed library A. Only 23% of the features in sublibrary “Training/2 A” are shared with “Test/2 A,” and only 45% of the features in sublibrary “Test/2 A” are shared with “Training/2 A.” This low amount of overlap provides a concrete basis for questioning the applicability of the validation studies to the test predictions.

It might be argued that basing arguments on the “drawn from the same distribution of samples” can be misleading when dealing with data sets containing few actives and many inactives. Note that of the 50,000 compounds in the training set, only 32 were considered to be active; a similar number of actives were present in the test set (95 or 42, depending on the scoring method). Thus, the extended-connectivity fingerprint bits of the inactives will dominate the statistics. This is true but beside the point: when predicting, we want to know if our model was built from molecules with similar structures to the samples we wish to predict. It does not matter if that similarity is to features from an active or an inactive compound. We simply want to know whether we should trust our predictions; considering only the active compounds would leave us blind to all the information in the inactive compounds. An assured prediction of “inactive” is worth more than the complete lack of knowledge when presented with truly novel features.

DISCUSSION

Before discussing the results, it is useful to see how many hits, on average, would be expected on a randomly ordered list. There were a total of 42 hits (CRA) or 9 hits (ARA) over the full data set. In the first 2500, a randomly ordered list would have, on average, 2.1 hits (CRA) or 4.75 hits (ARA). Thus, our method showed no benefit when judged again the CRA benchmark but showed some benefit against the ARA benchmark.

Once the test data results were released, we also wanted to see if there was any difference if we used the more abstract FCFP_6 instead of ECFP_6. The results were better: when the model building and application were repeated using FCFP_6, the procedure discovered 4 hits (using CRA) or 9 hits (using ARA). It turns out that our choice of ECFP_6 rather than FCFP_6 was not optimal. It is likely that the extra abstraction obtained by using FCFP_6 was no benefit when predicting samples similar to those in the training data but became valuable when extrapolating to molecules that were quite different from the training data.

The expected number of hits from a randomly ordered list illustrates the overall weakness of all submissions. Using consensus residual activity as the scoring method, only 1 group found more hits (4) than expected by chance (2.1). Using average residual activity, the situation is slightly improved, with 3 groups (including ours) finding more hits (13, 7, and 6) than expected (4.75). Many groups had zero hits in their lists.

However, one should be careful before concluding that the vast majority of methods failed. Five groups submitted lists greater than 2500 compounds, which was the cutoff used in the judging criteria. Twenty-seven groups submitted lists of less than 2500 compounds or only submitted lists of predicted active compounds. The latter groups would likely have increased the number of hits they predicted simply by “padding” their lists with random choices. More optimistically, if some of the 15 groups that offered 100 or fewer candidates had returned 2500 candidates, there may have been enough data to identify promising methods that happened to not find hits in a small predicted-active set. However, even with the optimistic scenario, the limited nature of the benefit (the best-showing method had an enrichment in 1 case of only 3 times random) deserves explanation.

For those methods that relied on the identity of the initial training hits (versus, for example, pure docking methods that only considered the protein receptor site), 1 difficulty is due to the structural differences between typical training molecules and typical test molecules. As shown by the data set comparison using extended-connectivity fingerprints, the test and training sets have broadly different types of molecules. The violation of the statistical rule that the training and test data be drawn from the same distribution meant that the higher quality enrichments suggested by the cross-validation study were not achieved in real life. This is a serious and common problem with screening efforts that may dramatically change the screening targets by, for example, using a different vendor data source. Statistical counting methods can aid in identifying these problematic situations.

For those methods that relied on docking into a protein receptor site, a combination of factors may be involved in the low prediction results: the docking methods (or their associated scoring metrics) may need continued refinement or tuning for this particular case, the screening process may create noise or artifacts that affect the results but are not directly associated with the ligand-receptor interaction, or the protein itself may have internal flexibility or other macroscale conformational changes that cannot yet be modeled.
One interesting proposal would be to combine 2D and 3D methods into an augmented method. For example, work published by Novartis\textsuperscript{8,9} shows how Laplacian-modified Bayesian analysis can be used to post-process initial 3D screening results, leading to improved hit rates using only preexperimental data. Work by O’Brien and deGroot\textsuperscript{10} suggests that combining different methods (in their case, Laplacian-modified Bayesian and neural networks) can lead to improvements over a single method; such work could be repeated using, for example, Laplacian-modified Bayesian and docking results.

**CONCLUSION**

Competitions such as the McMaster HTS competition offer a real-world opportunity to apply computational methods to a practical task. Such competitions are often, and incorrectly, interpreted primarily as a horserace, with a lucky winning entry and a number of also-rans. That is a mistaken approach; although some methods will, by definition, outperform the others, the element of “luck” is often underappreciated, and without multiple studies with different targets, that element cannot be easily estimated.

More interesting to us was the diversity of methods applied to this problem, which gives a sense of the current state of computational methods (on average). The different methods applied provide a range of possibilities for inclusion into a “toolkit” of diverse techniques. A reasonable computational scientist would likely try multiple approaches to a given problem, with preferred approaches varying from problem to problem. Much as a smart investor will not bet on 1 “best” stock that may be outperformed next year, scientists should resist relying on a single “best” method but instead keep a number of methods available and look for ways to combine them. Although submitting a 2D method to a 3D docking competition might have initially seemed quixotic, the 2D results were competitive with many other methods and at a low computational cost, suggesting that 2D methods still have a place at the 3D table.

**REFERENCES**


Address reprint requests to:
David Rogers
SciTegic, Inc.
9665 Chesapeake Drive #401
San Diego, CA 92123
E-mail: drogers@scitegic.com