Deriving Knowledge through Data Mining High-Throughput Screening Data

David J. Diller*[,†] and Doug W. Hobbs[‡]

Departments of Molecular Modeling and Chemistry, Pharmacopeia, Inc., CN5350, Princeton, New Jersey 08543-5350

Received February 3, 2004

Deriving general knowledge from high-throughput screening data is made difficult by the significant amount of noise, arising primarily from false positives, in the data. The paradigm established for screening an encoded combinatorial library on polymeric support, an ECLiPS library, has a significant amount of built-in redundancy. Because of this redundancy, the resulting data can be interpreted through a rigorous statistical analysis procedure, thereby significantly reducing the number of false positives. Here, we develop the statistical models used to analyze data from high-throughput screens of ECLiPS libraries to derive unbiased true hit rates. These hit rates can also be calculated on subsets of the collection such as those compounds containing a carboxylic acid or those with molecular weight below 350 Da. The relative value of the hit rate on the subset of the collection can then be compared to the overall hit rate to determine the effect of the substructure or physical property on the likelihood of a molecule having biological activity. Here, we show the effects that various functional groups and the standard physical properties, molecular weight, hydrogen bond donors, hydrogen bond acceptors, log *P*, and rotatable bonds, have on the likelihood of a compound being biologically active. To our knowledge this is the first published account of the use of high-throughput screening data to elucidate the effects of physical properties and substructures on the likelihood of compounds showing biological activity over a broad range of pharmaceutically relevant targets.

1. Introduction

Establishing the relationship between the likelihood of a small molecule being biologically active, i.e., specifically binding to a protein, and its physical−chemical, pharmacophoric, and structural properties is a problem of paramount importance in drug discovery. Despite the abundance of data from sources such as high-throughput screening and medicinal chemistry programs, this problem remains particularly challenging in part because the available data have fundamental limitations. Data from medicinal chemistry programs come from multiple measurements and so are generally of high quality. The difficulty with using data of this type to address general questions about biological activity is the amount of bias introduced by the human decisions made during the programs. Decisions are often based on factors such as availability and cost of reagents, preferences of individual scientists, pharmacokinetic issues, etc. It is nearly impossible to separate the bias toward “druglike” molecules, which is implicit in every medicinal chemist’s goals from the true characteristics of biological activity. For example, one might find that carboxylic acids are underrepresented historically in medicinal chemistry programs. This conclusion leads to the legitimate question: Are they underrepresented because they decrease the likelihood of biological activity or because they are avoided to improve the likelihood of the molecules being bioavailable?

Data from high-throughput screens do not suffer from the same problems as data generated during medicinal chemistry programs. For high-throughput screening data, the compounds screened establish the baseline. To continue with the example from the preceding paragraph, if 10% of all actives from a wide range of high-throughput screens contain a carboxylic acid while only 5% of the compounds screened contain a carboxylic acid, then one can conclude that compounds that contain a carboxylic acid are 2 times more likely to be biologically active than the average compound. Though it is in general less biased than data from medicinal chemistry programs, data from high-throughput screens suffer from problems of its own. High-throughput screening data are generally considered to be of significantly lower quality than data from medicinal chemistry programs and are particularly plagued by false positives. The reason that the false positives cause such a problem is that in a typical screen the inactive compounds vastly outnumber the active compounds. This is best illustrated through an example. Suppose we have a collection of 1 000 000 compounds and that our assays are 99.9% accurate, i.e., 99.9% of active compounds are classified as active and 99.9% of inactive compounds are classified as inactive. Next we assume that for an average target, 1 in a 1000 compounds are truly active. Researchers at Pfizer have found that they must screen approximately 120 000 compounds on average to find one lead series,[1] so the assumption of 1 in a 1000 compounds being truly active is clearly an overestimate. With these assumptions, on average from a single screen we would expect $0.999 \times 1000 = 999$ true positives and $999000 \times 0.001 = 999$ false positives. Thus, even with a nearly perfect assay the false positives are still equal in number to the true positives. Furthermore, if we assume a more realistic true hit rate of 1 in 10 000

* To whom correspondence should be addressed. Phone: (609) 452-3783. Fax: (609) 655-4187. E-mail: ddiller@pharmacop.com.

[†] Department of Molecular Modeling.

[‡] Department of Chemistry.

10.1021/jm049902r CCC: $27.50 © 2004 American Chemical Society
Published on Web 10/20/2004

compounds, we would on average find 10 false positives for every true positive. Any conclusions drawn from data that consist primarily of false positives are suspect. Ultimately, the key to deriving reliable knowledge from high-throughput screening data is minimizing the number of false positives.

ECLiPS[2] (encoded combinatorial library on polymeric support) library synthesis is a form of solid-phase library synthesis that uses a modified version[3] of split and mix as a means to efficiently generate libraries with a large number of compounds, often in excess of 50 000 members. The chief difference between an ECLiPS library and the more traditional forms of combinatorial synthesis is that during each step of the synthesis molecular tags are attached to the solid support in order to encode for the step and reagent used. Once the synthesis of the library is complete, the exact identity of a compound on any particular bead is not known except through detaching and reading the molecular tags. The only information known about the compound at the time of screening is the reagent used in the last synthetic step. The compounds with a common final synthetic step are referred to as a sublibrary. To screen an ECLiPS library, individual beads are placed in wells in a plate. The compound is cleaved from the bead and filtered to a second plate, while the tagged bead remains in the original plate. The second plate is then screened as in any high-throughput screening program. When a well shows sufficient activity, the identity of the molecular tags on the corresponding bead is determined. Because the tags form a binary code for the synthetic history of the compound originally on the bead, the tags contain the necessary information to determine the exact identity of the compound. For this reason, we refer to a compound found in an active well as a decode. Accordingly, we use the terms "active", "decode", and "decoded compound" interchangeably to mean the compound found in a well that has been deemed active by the biological assay.

The data from high-throughput screens of ECLiPS[2] libraries are amenable to a rigorous statistical interpretation that provides a means to eliminate the majority of the false positives for the purposes of data mining. The statistical analysis relies on two features unique to the ECLiPS screening process. First, the compounds being screened are essentially blind to the scientist performing the experiment. As stated above, in these experiments single compounds are put in wells with the only thing known about the compounds being the sublibrary from which they came. Only when a well shows some level of activity is the identity of the compound determined via decoding tags placed on the solid support during the compounds' solid-phase synthesis. Thus, the identities of only the compounds in active wells are determined, and these identities are not determined until after the well is deemed active. The purpose of a blind experiment is to separate the observation from prior expectations, thereby eliminating bias. The concept of a blind experiment is most well-known from clinical trials in which "double blind" trials are routinely run. Clinical trials are run in this fashion to eliminate the bias of those participating in or observing the trial. In the language of statistics, each well is an independent identically distributed experiment from a sublibrary with outcomes being that either the well is inactive or the particular compound in the well is active.

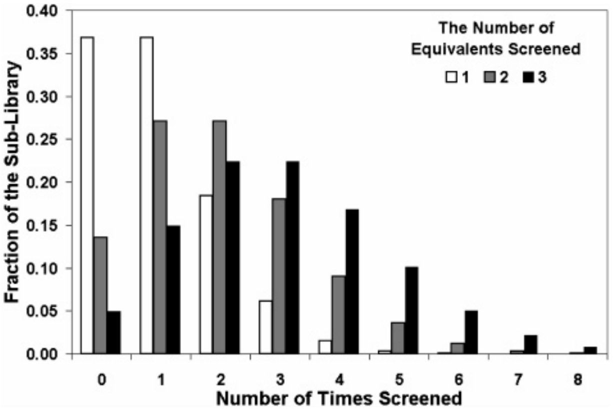

Figure 1. Redundancy in the ECLiPS screening experiment. The white bars represent the distribution of the number of times the compounds are screened if a single equivalent, i.e., the same number of wells as there are compounds, of the sublibrary is screened. The gray bars represent the distribution of the number of times the compounds are screened if two equivalents, i.e., twice the number of wells as there are compounds, of the sublibrary are screened. The black bars represent the distribution of the number of times the compounds are screened if three equivalents, i.e., three times the number of wells as there are compounds, of the sublibrary is screened. Typically, we screen three equivalents of any active sublibrary. From this figure if we screen three equivalents, 5% of the compounds in the sublibrary are not screened at all, 15% are screened one time, 22% are screened two times, 22% are screened three times, etc.

The second and most important feature of an ECLiPS screening experiment is its inherent redundancy. Because the identities of the compounds are not known when the compounds are put in the wells, there is no way to guarantee that each compound in a sublibrary is in fact screened in any particular assay. To offset this uncertainty, more than 1 equivalent of each active sublibrary is screened with a single compound per well. The effect is that while a small fraction of the compounds are not screened the majority will be screened two, three, or more times. Figure 1 shows the distribution of the number of times compounds are screened at 1, 2, and 3 equivalents. Typically, we screen 3 equivalents of every active sublibrary. For example, in a sublibrary with 1000 members 3000 compounds are screened with a single compound per well. The effect is that approximately 5% of the compounds in a sublibrary are not screened, 15% are screened one time, 22% are screened two times, 22% are screened three times, 10% are screened four times, etc. Of course, it is not known which compounds are screened three times, only that a certain fraction of the compounds are. This redundancy, though probabilistic in nature, provides a means to assess whether a set of decodes is consistent with true activity or with simply random compound selection. Random compound selection results in virtually no compounds being decoded more than one time, whereas true activity should result in a reasonable distribution of replication within the decodes; i.e., some compounds appear in single active wells, some in two active wells, some in three active wells, and so on. The statistics around these distributions are developed rigorously in section 2.1.

Since the influential work of Lipinski and co-workers,[4] significant effort has gone into determining the properties and substructures of compounds that make them more likely to become drugs.[4-7] The chief difference between the aim of this large body of work and the present study is that the present study is focused on isolating those factors associated with specific biological activity rather than isolating the factors associated with desirable properties, such as good solubility or high absorption, that are necessary for a compound to become a drug. Gillet and co-workers[8] adopted a different strategy in an effort to delineate the factors associated with biological activity rather than those associated with "druglikeness". In this work the WDI[9] database was used as representative of biologically active molecules and the SPRESI[10] database was used as representative of biologically inactive compounds. The distributions of various molecular properties were calculated for the two databases, and the resulting distributions were compared. A number of descriptors showed statistically significant differences between the two databases. In particular, in decreasing order of significance the number of hydrogen bond donors, the number of hydrogen bond acceptors, molecular weight, the $2\kappa_\alpha$ shape indices, the number of rotatable bonds, and the number of aromatic rings all showed some level of statistically significant separation of the two databases. The WDI contained more compounds with a large number of hydrogen bond donors and with a large number of hydrogen bond acceptors and of high molecular weight, which differs from the majority of the work done on "druglikeness" suggesting that the effects captured in this study are indeed distinct from those captured in the aforementioned "druglikeness" studies. The chief difference between the present study and the study of Gillet and co-workers is the data used with the present study relying on high-throughput screening data and Gillet and co-workers relying on data primarily from medicinal chemistry programs.

In addition to the work on the influence of physical properties, considerable effort has gone into understanding the role of substructures in specific biological activity and "druglikeness". For convenience we split this large body of work into three main areas. The first area is the use of substructure analyses to develop "druglikeness" models[11,12] or models of related phenomena such as bioavailability[13,14] and mutagenicity.[15] The second area where substructure analysis has played a key role is in developing models for the activity of a set of compounds against a single target. There are numerous published examples of the utility of this approach to understanding a structure−activity data set, and we cite only a handful of the earlier examples.[16-19] The general approach in these studies is to look for substructures that are enriched/deriched in the active compounds relative to the inactive compounds. In spirit these analyses are similar to the substructure analysis presented in this work with the chief difference being the scope; in this work we look at high-throughput screening data from a large number of targets. The third main area of substructure analysis is in the identification of "privileged structures"[20-24] where a "privileged structure" is defined as a substructure that occurs frequently in ligands for a variety of targets. Much of this work is focused on "privileged structures" for G-protein-coupled receptors and relies primarily on the intuition and experience of individual scientists.[20,24] Hajduk and co-workers,[21] however, performed a statistical analysis of NMR-derived binding data from 11 protein targets. They found that compounds containing a carboxylic acid were statistically enriched within the active compounds for 6 of the 11 targets and that compounds containing a biphenyl were statistically enriched for 5 of the 11 targets. The next most frequently enriched substructure was enriched in only 3 of the 11 targets.

For this work, the combined results of 100 high-throughput screens were analyzed using the statistical techniques described below. The screens cover a range of pharmaceutically relevant drug targets. The properties and substructures of the actives are compared to those of our overall collection. Because factors such as metabolic stability and absorption are not factors in the assay determinations, any differences between the active set of compounds and our compound collection should be attributable primarily to true biological activity with solubility being the other possible contributing factor. Since these data are less complex than data used in previous studies, these comparisons shed light on the relationship between the likelihood of a compound being biologically active and its physical−chemical and structural properties.

2. Materials and Methods

The materials and methods used for this work are split into three subsections: a rigorous derivation of the statistical interpretation of data produced from a screen of an ECLiPS sublibrary (section 2.1), a concrete example of the process by which the data is analyzed (section 2.2), and a summary of the screens and data that produced the data used (section 2.3).

2.1. Statistical Interpretation of Decode Data. The notation in this section is somewhat involved and as a result is collected and summarized in Table 1. In this section we rigorously develop a statistical model for the screening observation from an ECLiPS sublibrary based on three unknown parameters: (1) the fraction of the library that is active, ϵ; (2) the probability that a well is deemed active (decoded) given the compound in the well is active, P_{11}; (3) the probability that a well is deemed active (decoded) given the compound in the well is inactive, P_{10}. The parameter ϵ is the true hit rate. The parameter P_{10} has the physical interpretation as the well false positive rate; i.e., the probability a well is deemed active when the compound it contains is inactive. The parameter P_{11} is the true positive rate. Therefore $1 - P_{11}$ is the well false negative rate.

For the purposes of this manuscript the observations made from our high-throughput screens are the number of compounds that are decoded one time (Z_1), the number of compounds that are decoded two times (Z_2), the number of compounds that are decoded three times (Z_3), etc. Accordingly, the goal is to derive probability distributions for the screening observations, Z_n, in terms of the unknown parameters ϵ, P_{11}, and P_{10} and the known parameters, the number of distinct compounds in the sublibrary, N**,** and the number of wells screened with single compounds from the sublibrary, M. If the screen is nearly ideal ($P_{11} = 1$ and $P_{10} = 0$), then these

Table 1. Notation and Definitions

symbol	definition	dependencies
ϵ	fraction of the sublibrary that is active	unknown
P_{11}	probability a well is decoded when the compound is active	unknown
P_{10}	probability a well is decoded when the compound is not active	unknown
N	number of compounds in the sublibrary	known
M	number of wells screened from the sublibrary	known
α	number of equivalents of the sublibrary screened	={M}/{N}
k	reference to a single compound	
n	refers to the number of times a compound was decoded	
C_k	state of compound k; active = 1, inactive = 0	
Y_k	probability distribution for the number of times compound k is decoded	$\epsilon, P_{11}, P_{10}, M, N$
p_n	$P(Y_k = n)$	$\epsilon, P_{11}, P_{10}, M, N$
Z_n	probability distribution for the number of compounds decoded n times	$\epsilon, P_{11}, P_{10}, M, N$

distributions should look like those in Figure 1. If the probability distribution for Z_n can be determined as a function of ϵ, P_{11}, and P_{10}, one can in turn estimate these parameters via selection of the values of ϵ, P_{11}, and P_{10} that make the observed Z_n as likely as possible. This is the standard statistical technique of maximum likelihood.[25]

The probability that any single chosen well contains compound k is 1/N, and the probability that this well will be deemed active given that compound k is active is P_{11}. Therefore, given that compound k is active, the probability that any single chosen well contains compound k and is deemed active is P_{11}/N. Since each well is essentially independent from any other well, the probability distribution for the number of times compound k is decoded given it is active will follow a binomial distribution with the number of trials being equal to the number of wells screened and the probability of success being P_{11}/N, i.e.,

$$P(Y_k = n|C_k = 1) = \binom{M}{n}\left(\frac{P_{11}}{N}\right)^n\left(1 - \frac{P_{11}}{N}\right)^{M-n} \quad (1)$$

Note from eq 1 the probability that a compound is not decoded at all given it is active is given by

$$\text{compound false negative rate} = P(Y_k = 0|C_k = 1) = \left(1 - \frac{P_{11}}{N}\right)^M \quad (2)$$

Arguing as in the derivation of eq 1, we find the probability that a compound will be decoded n times given it is inactive is given by the formula

$$P(Y_k = n|C_k = 0) = \binom{M}{n}\left(\frac{P_{10}}{N}\right)^n\left(1 - \frac{P_{10}}{N}\right)^{M-n} \quad (3)$$

Also, from eq 3 the probability that a compound is decoded given it is not active is given by

$$\text{compound false positive rate} = P(Y_k \neq 0|C_k = 0) = 1 - \left(1 - \frac{P_{10}}{N}\right)^M \quad (4)$$

Then through an application of Bayes' theorem,[26] we combine eqs 1 and 3 to derive the probability distribution for the number of times a compound is expected to be decoded:

$$\begin{aligned} p_n &\equiv P(Y_k = n) = P(Y_k = n \cap C_k = 1) + P(Y_k = n \cap C_k = 0) \\ &= P(Y_k = n|C_k = 1)\, P(C_k = 1) + P(Y_k = n|C_k = 0)\, P(C_k = 0) \\ &= \binom{M}{n}\left[\epsilon\left(\frac{P_{11}}{N}\right)^n\left(1 - \frac{P_{11}}{N}\right)^{M-n} + (1-\epsilon)\left(\frac{P_{10}}{N}\right)^n\left(1 - \frac{P_{10}}{N}\right)^{M-n}\right] \end{aligned} \quad (5)$$

Finally, we derive the probability distribution for the number of compounds that will be decoded a given number of times:

$$P(Z_n = r) = \binom{N}{r}{p_n}^r(1 - p_n)^{N-r} \quad (6)$$

where p_n is given in eq 5. A screening observation, which we denote by Θ, consists of the number of compounds decoded one time, the number decoded two times, etc., i.e., an observed value for Z_n for n = 1, 2, 3, ... Thus, the closed form probability for an observation based on the parameters M, N, ϵ, P_{11}, and P_{10} is given by

$$P(\Theta) = \prod_{i=1}^{\infty} P(Z_n = o_n) \quad (7)$$

where o_n is the observed number of compounds that are decoded exactly n times. The exact dependency of $P(\Theta)$ on M, N, ϵ, P_{11}, and P_{10} can be determined by tracing back through eqs 6 and 5. The reader should note that for large n, $o_n = 0$ and $P(Z_n = 0) = 1$. Thus, even though the product in eq 7 is written as an infinite product, it is in reality a finite product.

To this point we have derived an expression for the probability of an observation given the value of the parameters M, N, ϵ, P_{11}, and P_{10}. The parameters M and N are known, whereas the parameters ϵ, P_{11}, and P_{10} are unknown. Ultimately, we wish to estimate the unknown parameters ϵ, P_{11}, and P_{10} from the observed data. The technique we use to perform this estimation is maximum likelihood, which chooses the values of the unknown parameters to maximize the probability of the observation. Thus, we choose the values of ϵ, P_{11}, and P_{10} to maximize $P(\Theta)$ as it is given by eqs 5, 6, and 7. This can be accomplished by a standard gradient based minimization algorithm such as conjugate gradient minimization.[25]

At this point we have derived a means to estimate the free parameters ϵ, P_{11}, and P_{10}, and we would like to derive an estimate for the probability a compound is active given the number of times it was decoded during the screen. This can be achieved from eq 1, eq 5, and two applications of Bayes' theorem:

$$\begin{aligned}&P(C_k = 1|Y_k = n)\\ &= P(C_k = 1 \cap Y_k = n)/P(Y_k = n)\\ &= P(Y_k = n|C_k = 1)P(C_k = 1)/P(Y_k = n)\\ &= \frac{\epsilon\binom{M}{n}\left(\frac{P_{11}}{N}\right)^n\left(1-\frac{P_{11}}{N}\right)^{M-n}}{\binom{M}{n}\left[\epsilon\left(\frac{P_{11}}{N}\right)^n\left(1-\frac{P_{11}}{N}\right)^{M-n}+(1-\epsilon)\left(\frac{P_{10}}{N}\right)^n\left(1-\frac{P_{10}}{N}\right)^{M-n}\right]}\end{aligned} \tag{8}$$

Thus, from the maximum likelihood estimates of ϵ, P_{11}, and P_{10} and eq 8, we can estimate the probability that any compound in the sublibrary is active on the basis of the number of times the compound was decoded. Because the estimated values for ϵ, P_{11}, and P_{10} depend on the entire sublibrary screening observation, the probability that a compound is active depends not only on the number of times it was decoded but also implicitly on the entire observation.

In the following sections we refer to three interpretations of the high-throughput screening data: the random model, the ideal model, and the realistic model. The expected results with all three of these models are determined from eqs 5 and 6 with the differences between the models arising from the differences in the values of the three free parameters ϵ, P_{11}, and P_{10}. The random model is one in which we assume that wells are selected at random and decoded rather than being guided by the biological assay. In this model, $\epsilon = 0$, $P_{11} = 0$, and P_{10} is chosen so that the expected number of decodes is equal to the observed number of decodes. If the observed data are consistent with the expected results from this model, then we conclude that the data are likely not to be true activity and discard the data from further use in this data mining study. The ideal model is one in which we assume the assay classifies inactive compounds as inactive and active compounds as active with no errors. In this model $P_{11} = 1$, $P_{10} = 0$, and ϵ is a free parameter chosen using maximum likelihood. The realistic model is essentially that described above in which all three parameters are treated as unknown and are estimated from the screening observation.

2.2. Example of the Statistical Analysis. The purpose of the preceding section was to develop a rigorous and quantitative means through which we can first assign a significance to the extent of replication found within a set of decodes from a screen of a sublibrary. Should the replication not be sufficiently significant, no further analysis is performed and the decodes are not used further in this study. If the replication is sufficiently significant, then the analysis in the preceding section allows for estimates of the hit rate within the sublibrary and even of probabilities of individual compounds being active depending on the number of times they were decoded. In this section, we show qualitatively how this analysis procedure is applied to screening data from a single active sublibrary. In this particular case, 3 equivalents of a sublibrary with 2268 distinct members were screened, i.e., a total of 6804 wells, each with a single compound from the sublibrary, were screened.

Table 2. Example ECLiPS Screening Observation and the Expected Observations with Different Models for the Data[a]

degree of replication	observed	random model	ideal model	realistic model
1	45	134.3(11.2)	12.1(3.5)	44.9(6.6)
2	12	4.2(2.1)	18.1(4.2)	13.0(3.6)
3	13	0.1(0.3)	18.1(4.2)	9.7(3.1)
4	3		13.6(3.7)	5.5(2.3)
5	2		8.1(2.8)	2.5(1.6)
6	1		4.1(2.0)	0.9(1.0)
7	1		1.7(1.3)	0.3(0.6)
8	0		0.7(0.8)	0.1(0.3)
9	0		0.2(0.5)	0.0(0.1)

[a] The "observed" column indicates the number of compounds that were decoded a single time, two times, etc. In this case the sublibrary contains 2268 members and was screen at 3 equivalents, meaning 6804 compounds were screened in individual wells. Of these, 143 wells were deemed active by the biological assay with 45 compounds appearing in a single active well, 12 compounds in two active wells, etc. Notice that 143 = (45 × 1) + (12 × 2) + (13 × 3) + (3 × 4) + (2 × 5) + (1 × 6) + (1 × 7). The random, ideal, and realistic models are described at the end of section 2.1. The numbers in these columns are the expected number of compounds decoded one times, two times, etc. with the particular model. The numbers in parentheses in the final three columns are the standard deviations. It might seem surprising that there is any uncertainty in the case of an ideal model. This uncertainty arises because of the uncertainty in the number of times each compound is screened as shown in Figure 1 rather than the uncertainty in the assay. As is evident, the realistic model agrees with the observation much better than the ideal model, which agrees with the data significantly better than the random model.

The second column of Table 2 gives an example of an observation from a screen of a single sublibrary. In this case, 143 wells were deemed active, 45 compounds were decoded one time, 12 compounds were decoded two times, 13 compounds were decoded three times, 3 compounds were decoded four times, 2 compounds were decoded five times, 1 compound was decoded six times, and 1 compound was decoded seven times. Notice that the number of active wells,143, is (45 × 1) + (12 × 2) + (13 × 3) + (3 × 4) + (2 × 5) + (1 × 6) + (1 × 7).

Upon receiving the data from an apparent active sublibrary, such as the one described in the preceding paragraph and Table 2, the first question to ask is whether the given observation is consistent with a random selection of compounds from the sublibrary rather than the selection being guided by the biological assay. If the data are consistent with random compound selection, then the hits are unlikely to represent true activity. To address this question, the probability that the observation would have been made had the compounds been chosen at random is calculated. For a highly active sublibrary this number is extremely small ($<10^{-10}$). For this reason we typically deal with the statistical significance of a screening observation that we define as minus the logarithm (base 10) of the probability that the observation would have been made had the compounds been chosen at random.

Returning to the example screening data, the third column of Table 2 shows the expected distribution had the 143 compounds been chosen at random from the sublibrary rather than having been selected via the biological assay. In this case we would expect approximately 134 compounds to be decoded a single time and only a handful of compounds to be decoded twice. Clearly, the discrepancy between this observation and the expected results with random compound selection is so large that there is essentially no chance that these compounds were chosen solely by random means. In fact, for this observation, the significance is 81, which means that the probability of making this observation through random selection is 10^{-81}, which is equivalent to flipping a fair coin and getting heads 269 consecutive times. Clearly, the degree of replication observed in this particular screen could not have happened through random compound selection.

The second step in the analysis process is to compare the data to an ideal screen. An ideal screen is one in which the assay deems a well active when the compound in the well is active, and it deems a well inactive when the compound in the well is inactive. Unlike the case of the random compound selection there is an unknown parameter: the fraction of the sublibrary that is active (the parameter ϵ in the previous section and Table 1). Given any value for the active fraction, the probability of making the given observation can be computed (see the previous section for details). The technique of maximum likelihood[25] is then employed to select the value of the active fraction that best explains the observation. The method of maximum likelihood simply says to choose the value of the parameter that maximizes the probability of having made the particular observation. For the example shown in Table 2 the active fraction that maximizes the likelihood of the observation with the ideal model is 3.6%. Even with this optimal value for the active fraction, the probability of making the given observation from an ideal screen is 10^{-21}. While this is significantly better than the probability 10^{-81}, when a random process was assumed, it demonstrates that this screen is not perfect. The fourth column of Table 2 shows the expected observation if this were an ideal screen with the optimal value for the active fraction. The differences between the expected distribution and the observed distribution are obvious and sufficient to conclude that this is not an ideal screen.

To interpret the data from an ECLiPS screening experiment, we employ a model that we refer to as the realistic model, intermediate between the random process and the ideal screen. In the realistic model there are three free parameters: the active fraction (ϵ), the well false positive rate (P_{10}), and the well false negative rate ($1 - P_{11}$). Here, the well false positive and well false negative rates refer to the probability that the assay incorrectly classifies a well. For any choice of these three parameters, the probability of making an observation can be computed. Thus, maximum likelihood can again be used to select the optimal values for the active fraction, the false positive rate, and the false negative rate. For the particular screen shown in Table 2, the estimated active fraction is 2.1%, the estimated well false positive rate is 0.5%, and the estimated well false negative rate is 24.5%. The expected distribution with these three parameters is shown in the fifth column of Table 2. With these three parameters the model now agrees very well with the observation. The probability of making the observation given this particular model is now 10^{-6}, a vast improvement from either the random or ideal models.

As a side note, one might feel that 1 in a million is in fact very poor odds despite it being a big improvement over previous models. The number is the probability that the exact observation would have been made with this particular model. One would expect that if the experiment were performed many times, the standard deviations of the observation would be comparable to those in Table 2 for the realistic model. If this is the case, there could realistically be 100 000 to 1 000 000 reasonable distinct outcomes of the experiment. Thus, for this particular observation the probability 10^{-6} is nearly as large a probability as one would expect.

The final step in the statistical analysis is to estimate the probability that a compound will be active upon resynthesis. This probability will be dependent on the number of times the compound was decoded and the three parameters (active fraction, false positive rate, and false negative rate) determined in the prior step. This can be accomplished from the probability distributions determined in the prior step and a straightforward application of Bayes' theorem.[26] For the example shown in Table 2, the probability a compound that was decoded exactly one time will be active is 0.25. This probability reflects the fact that the compounds decoded a single time contain the majority of the false positives as well as the less potent actives. For this screen any compound that was decoded multiple times, however, is essentially guaranteed to be active.

2.3. Data Used for the Data Mining. The data from approximately 100 high-throughput screens were analyzed by the above procedure. In these 100 screens approximately 200 000 000 total compounds were screened. Only the decoded compounds from highly significant screens were treated as biologically active compounds in the analysis. The data from all sublibraries for which the significance exceeded 20 were analyzed according to the process outlined in sections 2.1 and 2.2. The actual choice of 20 is somewhat arbitrary, but this left a significant number of active sublibraries and is high enough to guarantee the elimination of most false positives. The decodes from these sublibraries comprise the set of biologically active compounds. The biologically active set comprises 16% GPCR ligands, 17% kinase inhibitors, 27% non-kinase enzyme inhibitors, and 40% from other target families (Table 3).

3. Results

In this section, we consider the effects of the standard physical properties calculated with Cerius2[27] and standard functional groups on hit rates. The general approach will be to compare the fraction of the biologically active set of compounds within a physical property range or having a given substructure to the fraction of

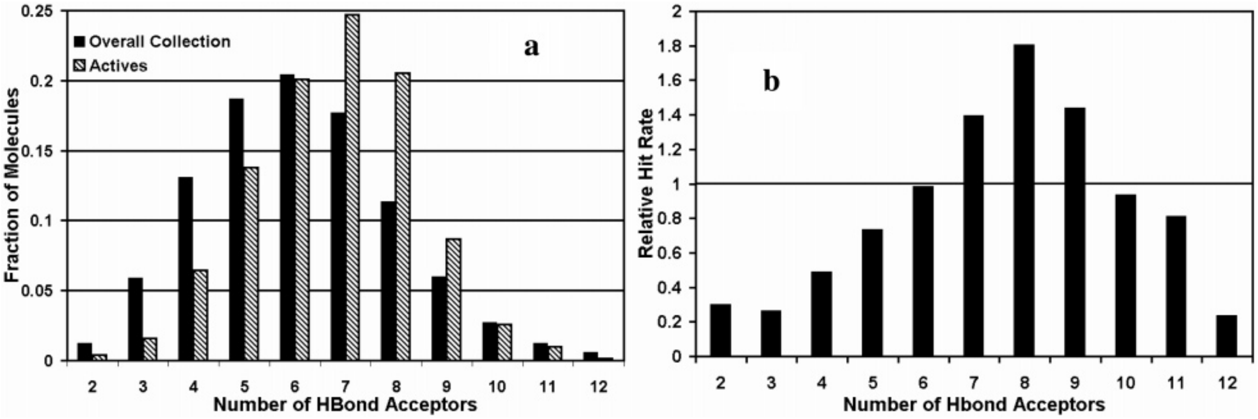

Figure 2. Effects of the number of hydrogen bond acceptors on the likelihood of finding biological activity. (a) Histograms for the number of hydrogen bond acceptors in either the active molecules or the collection. The solid bars show the distribution for the number of hydrogen bond acceptors in our entire collection with the compounds weighted by the number of times their respective library has been screened. The hashed bars show the distribution of the number of hydrogen bond acceptors in the actives found over the course of screening the targets described in Table 3. (b) Relative hit rate for the given number of hydrogen bond acceptors. The height of each bar is the ratio of the height of the hashed bar to that of the solid bar in part a. We refer to these ratios as relative hit rates because they are equivalent to the ratio of the hit rate for those compounds having a fixed number of hydrogen bond acceptors to the hit rate for the entire collection. Note that the maximum number of actives are found in the bin corresponding to seven hydrogen bond acceptors, whereas the maximum hit rate is achieved in the compounds with eight hydrogen bond acceptors.

Table 3. Description of the Targets from Which the Data Used in This Paper Came[a]

target class	fraction of the actives	number of different sublibraries
GPCR	0.16	21
kinase	0.17	16
enzyme	0.27	14
miscellaneous	0.40	30

[a] The "target class" column refers to the target class of the screens. The enzyme class refers to all enzymes other than kinases. The "fraction of the actives" column refers to the fractions of the set of active compounds that were found during a screen with a target of the particular class. The "number of different sublibraries" indicates the number of distinct sublibraries from which the set of active compounds were found separated by target class.

the entire collection within the physical property range or having the given substructure. Typically, we consider the ratio of the first of these fractions to the second of these fractions. If this ratio is greater than 1, we conclude that being within the particular property range or having the particular substructure contributes positively to the chances of finding biological activity. If this ratio is less than 1, then being within the particular property range or having the particular substructure contributes negatively to the chances of finding biological activity. We refer to these ratios as relative hit rates because they are equivalent to the ratio of the hit rate for those compounds within the physical property range or having the particular substructure to the hit rate for the entire collection.

As a first example, we consider the influence of the hydrogen bond acceptor count on hit rates in general. Figure 2a shows the histograms for both the number of hydrogen bond acceptors of our entire collection weighted by the number of times each compound has been screened and the actives we have found over the course of screening the targets described in Table 3 weighted by their probability of being active. As is apparent from Figure 2a, the majority of the active compounds are found in the subset of compounds with five to eight hydrogen bond acceptors. The bin that contains the largest portion of the actives is the bin corresponding to seven hydrogen bond acceptors. Part of the reason the majority of the active compounds are found in these bins is that the majority of the compounds in our collection are found in these bins. More so than the height of any individual bar we are interested in the ratio of the height of the bar for the active compounds to the height of the bar for the entire collection. We refer to this ratio as the relative hit rate because it is equivalent to the hit rate for compounds within the given bin divided by the hit rate for the entire collection. The bin with the largest difference between the fraction of the actives and the fraction of the overall collection is that corresponding to eight hydrogen bond acceptors. In fact, compounds with eight hydrogen bond acceptors have nearly twice the likelihood of showing activity in a typical biological screen than the average compound (see Figure 2b).

The relative hit rates were calculated, in the same manner as for the hydrogen bond acceptors above, for the standard physical properties including the number of hydrogen bond donors, log *P*, molecular weight, and the number of rotatable bonds. The hit rate varies little with the number of hydrogen bond donors (Figure 3a) except for the case when the molecule has five hydrogen bond donors. This result is consistent with the finding of Gillet and co-workers[8] in which they found that the biologically active compounds contained more compounds with a large number of hydrogen bond donors than the nonbiologically active set. It is not clear why the molecules having exactly five hydrogen bond donors are significantly more likely to be active than the rest of the collection, and it is entirely possible that this effect is in part due to the increased solubility of the compounds with a large number of hydrogen bond donors. The increase in the likelihood of finding biological activity in the set of compounds with five hydrogen bond donors is likely not enough to overcome the concern that these compounds would have less than desirable absorption characteristics.[4]

For log *P* (AlogP98[28,29]), the hit rates fall off dramatically for values below 1 and above 6 (see Figure 3b).

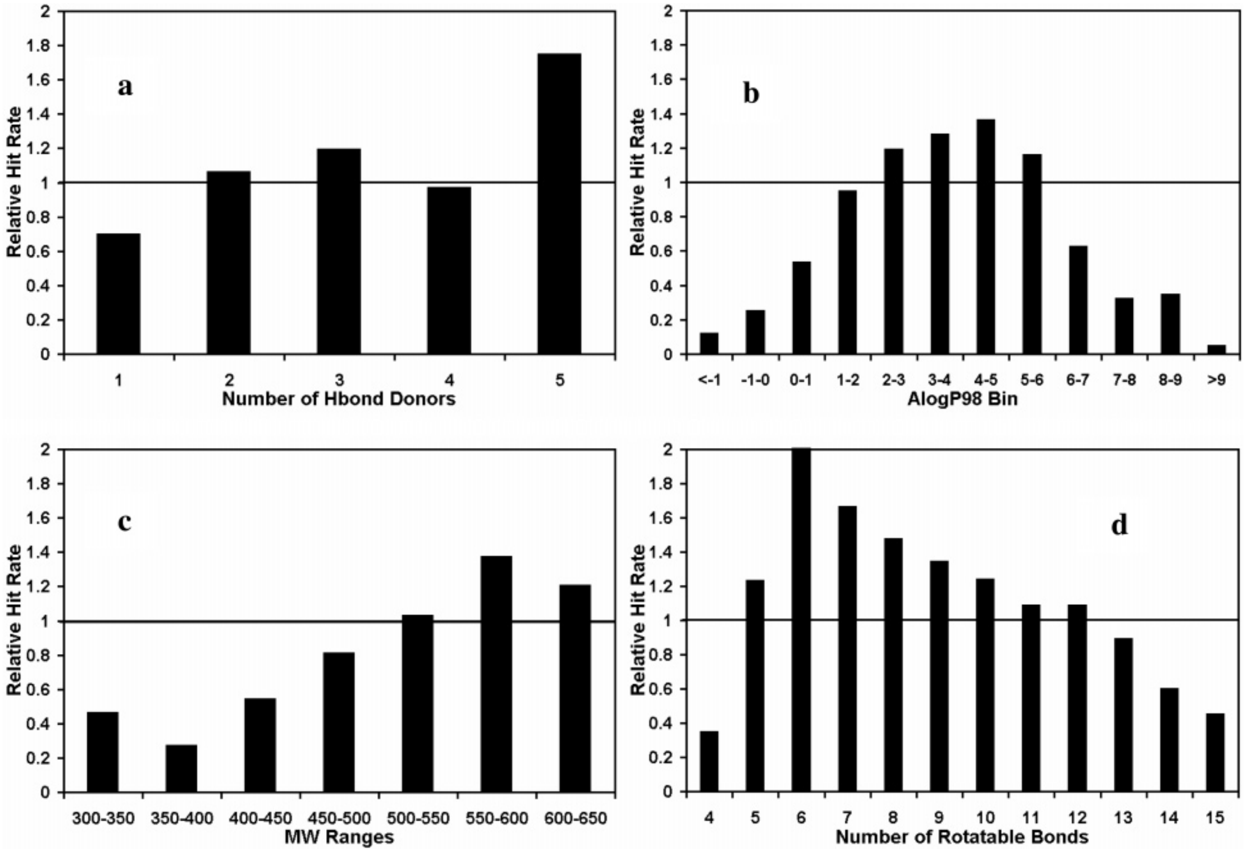

Figure 3. Effects of the standard physical properties on hit rates. Each figure shows the relative hit rate, whose calculation is described in Figure 2, for the particular physical property. (a) Relative hit rates versus the number of hydrogen bond donors. (b) Relative hit rates versus log *P*. The particular log *P* model used was AlogP98.[28,29] (c) Relative hit rates versus molecular weight. (d) Relative hit rate versus the number of rotatable bonds.

The lower hit rates for compounds with AlogP98 < 1.0 can be rationalized as the desolvation penalty being too great to achieve a strong net binding interaction. The lower hit rates for compounds with AlogP98 > 6 are likely the result of lack of solubility more so than the lack of the potential for significant protein−ligand interactions. The range of log *P* between 1 and 6 is consistent with the characteristics of good oral absorption and solubility[4] and provides further reason to restrict the log *P* of screening compounds.

For molecular weight, one sees a steady increase in hit rates from 350 to 600 (see Figure 3c). This is consistent with the observation that more potent compounds tend to be of higher molecular weight.[30] There is a small increase in hit rates in going from above 350 Da to below 350 Da, and a small decrease in the hit rates in going from below 600 Da to above 600 Da. The magnitude of the change at either end of the distribution is not sufficient to be statistically significant.

The physical property with the most impact on the hit rates is the number of rotatable bonds. Here, we use the implementation of the rotatable bond count in Cerius2[27] in which a bond is counted as rotatable if it is an acyclic, nonterminal single bond with hydrogens being excluded except on polar atoms. The count of rotatable bonds used includes partially rigid bonds such as amides and certain bonds that most authors do not count such as a bond to a 0H or a bond to a nitrile group. Thus, the count is an overestimate, but because it is consistent between the active compounds and the overall collection, we felt it unnecessary to correct it. As is apparent from Figure 3d, there is an intermediate number of rotatable bonds at which the hit rate is maximal. For our collection the optimal number of rotatable bonds is 6 where the relative hit rate is 2. Either decreasing or increasing from this number of rotatable bonds results in significant decreases in hit rates.

The number of rotatable bonds of a molecule is correlated with its molecular weight, with larger compounds typically having more rotatable bonds. As a result of this correlation, the exact value for the optimal number of rotatable bonds is likely dependent on the molecular weight profile of the compound collection being studied. As is apparent from Figure 3, molecular weight tends to contribute positively to hit rates whereas hit rate tends to decrease with larger numbers of rotatable bonds. Thus, rotatable bonds and molecular weight are competing factors in the hit rates, and the importance of each is likely partially hidden by the effects of the other.

To deconvolute the effects of rotatable bonds and molecular weight, Figure 4 shows the mean number of rotatable bonds within a small molecular weight range (50 Da) for both the entire collection and the active compounds. It is evident from Figure 4 that the biological assays have preferentially selected for compounds that are less flexible. In fact, regardless of the molecular weight the mean number of rotatable bonds for the active compounds is consistently between 1 and 2 rotatable bonds less than that for the overall collection.

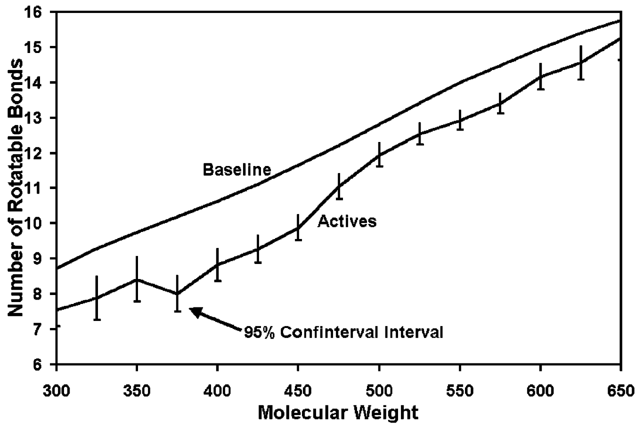

Figure 4. Interplay between molecular weight and rotatable bonds on hit rates. The upper curve shows the mean number of rotatable bonds of compounds in our overall collection within a given molecular weight range (±25 Da). The lower curve is the mean number of rotatable bonds of the active compounds within the given molecular weight range. The error bars on the lower curve are the 95% confidence intervals for the mean, not the standard deviation.

Table 4. Effects of Different Functional Groups on Hit Rates

substructure	relative hit rate	95% confidence interval
cyano	0.13	0.39
ketone	0.23	0.23
thioether	0.48	0.25
amide	0.75	0.04
sulfonamide	0.83	0.22
secondary aniline	0.87	0.12
tertiary amine	0.92	0.09
ether	0.94	0.06
secondary amine	1.00	0.16
phenol	1.14	0.87
alcohol	1.22	0.16
primary amine	1.54	0.58
no amide	1.60	0.20
carboxylic acid	1.79	0.13
urea	1.80	0.16
tertiary aniline	2.21	0.13
carbamate	2.74	0.37

[a] The substructures should for the most part should be self-explanatory. The "no amide" substructure refers to those compounds having 0 amide groups. For all substructures other than "no amide", the results are for those compounds having exactly one instance of the given substructure. The relative hit rate is defined as the fraction of the active compounds containing the substructure divided by the fraction of the overall collection containing the functional group. The 95% confidence interval is that for the relative hit rate.

As a final example of the potential for high-throughput screening data to address the relationship between the likelihood of a small molecule being biologically active and its structural properties, we show the effects of common polar functional groups on hit rates. Table 4 lists several common polar functional groups and their relative hit rates. The trend is that functional groups that are capable of making strong intermolecular interactions generally show larger relative hit rates, whereas those considered to be capable only of weaker hydrogen-bonding interactions show smaller relative hit rates. For example, cyano (0.13), ketone (0.23), and thioether (0.48) all have relative hit rates below one-half whereas carbamate (2.74), tertiary anilines (2.21), urea (1.80), carboxylic acid (1.79), and primary amine (1.54) all have relative hit rates above 1.5. The one group that does not fit this pattern is the tertiary aniline, which has a relative hit rate above 2 but is not capable of hydrogen bonding in any capacity. The only effects this functional group can have is through electronic effects on the nearby substituents, particularly the aromatic ring to which it is bonded. In comparison the secondary aniline, which has the capacity to act as a hydrogen bond donor, has a relative hit rate slightly below 1.0.

Since the physiological role of many proteins is to recognize either a peptide or protein, one might expect that the amide would on average increase the likelihood of a compound showing biological activity. Our data suggest just the opposite. Compounds containing a single amide have a relative hit rate of 0.75, whereas the compounds containing no amide have a relative hit rate of 1.6.

4. Conclusions

While the false positives can make it difficult to derive meaningful knowledge from high-throughput screening data, the process for screening ECLiPS libraries is amenable to a rigorous statistical interpretation. The statistical interpretation stems from both the redundancy in the experiment and the blind nature of the experiment. Most importantly the statistical interpretation allows for a quantitative means to select screens of sublibraries that are of the utmost quality. This filtering allows for significant reduction in the number of false positives. Second, the statistical interpretation allows for the assignment of a consistent probability of activity to every compound from an active sublibrary. Once this assignment is complete for all of the high-quality screens, unbiased hit rates can be calculated. These unbiased hit rates allow for an examination of the relevance for any property or substructure on the general likelihood of a compound showing biological activity.

The statistical data interpretation termed the realistic model estimates an active fraction and false positive and false negative rates for a set of decodes from a single sublibrary. Clearly the assumption that a compound is either active or inactive is not entirely accurate. Some compounds are weakly active, while others are potent. The compounds whose potency is just above or just below the cutoff are much more likely to be classified as false positives or false negatives, respectively. Thus, the assumption that the false positive and false negative rates are independent of the compound is not true. In general, this type of false positive and false negative is not, however, as detrimental to the data as a false positive that occurs because of compound interference, equipment failure, etc. In fact, the net result of weakly potent compounds being sometimes classified as inactive and sometimes classified as active is that the potent compounds have on average a greater impact on the data mining results.

The second shortcoming of the statistical models presented here is their lack of general applicability. Because most pharmaceutical and biotechnology companies screen each compound in their collection a single time, the statistical analysis developed for the analysis of ECLiPS screening results and described here is not directly applicable for most high-throughput screening programs. There are, however, other forms of redundancy that would allow for a similar analysis. Meir and co-workers[31,32] have developed two different approaches to benefit from implicit redundancy in their compound collection even when each compound is screened a single time with multiple compounds per well. In the first of these approaches,[31] they used structural similarity as a form of redundancy. In particular, by prioritizing those compounds for which the most similar hits were found, they showed that they could retain most of the true positives and eliminate many of the false positives. The difficulty in developing a rigorous statistical model using structural similarity as a form of redundancy is that not all highly similar compounds will exhibit the same biological activity. In a study done at Abbott only 30% of compounds with a similarity greater than 85% to a known active had the same biological activity.[33] The second approach used by Meir and co-workers[32] is to represent each compound as a bit string with each bit signifying the presence or absence of a substructure. Each substructure occurs in multiple compounds, and thus, there is redundancy in substructure screening. With the assumption that the impact of each substructure on the likelihood of a compound being active in a screen is independent from the impact of every other substructure, they build a quantitative model from the data arising from a single screen to predict the likelihood that a compound is active in the particular screen. Again, they showed that by using this approach, they were able to maintain the majority of the true positives while eliminating many of the false positives. The advantage of the later approach is that the developed models could be used to prioritize further analogue synthesis, and by development of many models over many screens, trends might be observed in the preference for or against certain substructures.

Most druglikeness models are built from databases such as CMC,[34] MDDR,[35] or WDI.[9] Because these databases comprise almost exclusively of molecules that are biologically active against a protein target, it is possible that to some extent these rules and models also contain information on the features relevant to the likelihood of compounds exhibiting biological activity. The best example in which the relative hit rates we find from our historical high-throughput screens compare favorably to the standard "druglike" criteria is with log *P*. From Figure 3b, it is evident that the optimal range to obtain biological activity for log *P* is between 1 and 6, which is in reasonable agreement with the rule of five for log *P*, which says the optimal range for solubility and oral absorption is between 0 and 5.[4]

The physical property with the most interesting effect on hit rates is the number of rotatable bonds. One could argue that a more flexible compound is more likely to be able to adopt a conformation compatible with a protein binding site, thereby making it more likely to be biologically active. On the other hand, one could argue that a more flexible compound is less likely to be biologically active because it has to overcome a large entropic cost in order to bind. The data presented here suggest that the more rigid compounds are more likely to show biological activity than their more flexible counterparts. In particular, if two collections with identical molecular weight profiles are screened against an identical set of targets, then the more rigid of the collections will have a higher hit rate. This result differs from the result with rotatable bonds from the study of Gillet and co-workers[8] in which they find rotatable bonds to be one of the least significant descriptors and that the biologically active compounds (WDI) have more compounds with many rotatable bonds than the non-biologically active compounds (SPRESI). On the other hand, our finding is similar to that of Veber and co-workers[6] even though they were specifically considering factors that influence bioavailability rather than biological activity.

As a function of the properties studied here, relative hit rates achieve a maximal value of approximately 2. While improvements in hit rates by a factor of 2 may seem to be insignificant, a few factors should be considered. First, because these are general physical properties, it is not clear that there should be any relationship between these properties and the likelihood of a compound being biologically active, and perhaps a factor of 2 is greater than should be expected. It is likely that descriptors with more physical relevance to protein–ligand interactions would show significantly greater impact on the likelihood of a compound being biologically active. Second, any noise in the data will on average cause any result to regress toward the mean, thereby causing any enrichment or derichment in activity to be understated. Because the data used here is from high-throughput screens, there is likely to be a fair amount of noise in the data, suggesting that the true relative hits rates are in fact greater than 2. Third, the set of targets used to produce the data covers a broad spectrum of proteins. If the data were restricted to a smaller family of proteins, it seems reasonable that more significant variations in hit rates would be observed. Fourth, as we saw with molecular weight and rotatable bonds, the effects of differing physical properties on the likelihood of being biologically active are often competing. By uncovering the delicate nonlinear effects of multiple descriptors on biological activity, we may find areas of chemical space that are greater than 2-fold enriched in biologically active molecules. Finally, improving the hit rates of a combinatorial library by merely 25% by fine-tuning the physical properties is well worth the effort to do so.

Acknowledgment. We thank the many chemists and biologists who have assembled and screened our collection over the past 10 years to amass the data used in this manuscript. We also thank the engineers who developed the technology so that this quantity of data could be generated. We thank Vicky Williams for painstakingly organize the data and making this study possible.

References

(1) Spencer, R. W. High-throughput screening of historic collections: observations on file size, biological targets, and file diversity. *Biotechnol. Bioeng.* **1998**, *61*, 61–67.

(2) Ohlmeyer, M. H.; Swanson, R. N.; Dillard, L. W.; Reader, J. C.; Asouline, G.; et al. Complex synthetic chemical libraries indexed with molecular tags. *Proc. Natl. Acad. Sci. U.S.A.* **1993**, *90*, 10922–10926.

(3) Baldwin, J. J.; Horlbeck, E. G. Direct dividing method for synthesis of combinatorial libraries (Pharmacopeia, Inc.). WO9535503, 1995; p 40.

(4) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **1997**, *23*, 3–25.

(5) Oprea, T. I. Property distribution of drug-related chemical databases. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 251−264.

(6) Veber, D. F.; Johnson, S. R.; Cheng, H. Y.; Smith, B. R.; Ward, K. W.; et al. Molecular properties that influence the oral bioavailability of drug candidates. *J. Med. Chem.* **2002**, *45*, 2615−2623.

(7) Vieth, M.; Siegel, M. G.; Higgs, R. E.; Watson, I. A.; Robertson, D. H.; et al. Characteristic physical properties and structural fragments of marketed oral drugs. *J. Med. Chem.* **2004**, *47*, 224−232.

(8) Gillet, V. J.; Willett, P.; Bradshaw, J. Identification of biological activity profiles using substructural analysis and genetic algorithms. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 165−179.

(9) *World Drug Index*; Daylight Chemical Information Systems: Mission Viejo, CA.

(10) *SPRESI*; Daylight Chemical Information Systems: Mission Viejo, CA.

(11) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887−2893.

(12) Bemis, G. W.; Murcko, M. A. Properties of Known Drugs. 2. Side Chains. *J. Med. Chem.* **1999**, *42*, 5095−5099.

(13) Yoshida, F.; Topliss, J. G. QSAR model for drug human oral bioavailability. *J. Med. Chem.* **2000**, *43*, 2575−2585.

(14) Andrews, C. W.; Bennett, L.; Yu, L. X. Predicting human oral bioavailability of a compound: development of a novel quantitative structure−bioavailability relationship. *Pharm. Res.* **2000**, *17*, 639−644.

(15) Bacha, P. A.; Gruver, H. S.; Den Hartog, B. K.; Tamura, S. Y.; Nutt, R. F. Rule extraction from a mutagenicity data set using adaptively grown phylogenetic-like trees. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1104−1111.

(16) Cramer, R. D., 3rd; Redl, G.; Berkoff, C. E. Substructural analysis. A novel approach to the problem of drug design. *J. Med. Chem.* **1974**, *17*, 533−535.

(17) Chu, K. C.; Feldmann, R. J.; Shapiro, M. B.; Hazard, G. F., Jr.; Geran, R. I. Pattern recognitiion and structure−activity relationship studies. Computer-assisted prediction of antitumor activity in structurally diverse drugs in an experimental mouse brain tumor system. *J. Med. Chem.* **1975**, *18*, 539−545.

(18) Adamson, G. W.; Bawden, D. A substructural analysis method for structure−activity correlation of heterocyclic compounds using Wiswesser line notation. *J. Chem. Inf. Comput. Sci.* **1977**, *17*, 164−171.

(19) Hodes, L.; Hazard, G. F.; Geran, R. I.; Richman, S. A statistical-heuristic methods for automated selection of drugs for screening. *J. Med. Chem.* **1977**, *20*, 469−475.

(20) Patchett, A. A.; Nargund, R. P. Privileged structures—an update. *Ann. Rep. Med. Chem.* **2000**, *35*, 289−298.

(21) Hajduk, P. J.; Bures, M.; Praestgaard, J.; Fesik, S. W. Privileged molecules for protein binding identified from NMR-based screening. *J. Med. Chem.* **2000**, *43*, 3443−3447.

(22) Roche, O.; Schneider, P.; Zuegge, J.; Guba, W.; Kansy, M.; et al. Development of a virtual screening method for identification of "frequent hitters" in compound libraries. *J. Med. Chem.* **2002**, *45*, 137−142.

(23) Muller, G. Medicinal chemistry of target family-directed masterkeys. *Drug Discovery Today* **2003**, *8*, 681−691.

(24) Bondensgaard, K.; Ankersen, M.; Thogersen, H.; Hansen, B. S.; Wulff, B. S.; et al. Recognition of privileged structures by G-protein coupled receptors. *J. Med. Chem.* **2004**, *47*, 888−899.

(25) Press, W. H.; Teulkolsky, S. A.; Vetterling, W. T.; Flannery, B. P. *Numerical Recipes in C*, 2nd ed.; Cambridge University Press: Cambridge, 1997; p 994.

(26) Ott, R. L.; Longnecker, M. *Statistical Methods and Data Analysis*, 5th ed.; Duxbury: Pacific Groove, CA, 2001.

(27) *Cerius2-4.8*; Accelrys Inc.: San Diego, CA.

(28) Viswanadhan, V. N.; Ghose, A. K.; Wendoloski, J. J. Estimating aqueous solvation and lipophilicity of small organic molecules: A comparative overview of atom/group contribution methods. *Perspect. Drug Discovery Des.* **2000**, *19*, 85−98.

(29) Ghose, A. K.; Viswanadhan, V. N.; Wendoloski, J. J. Prediction of Hydrophobic (Lipophilic) Properties of Small Organic Molecules Using Fragmental Methods: An Analysis of ALOGP and CLOGP Methods. *J. Phys. Chem. A* **1998**, *102*, 3762−3772.

(30) Kuntz, I. D.; Chen, K.; Sharp, K. A.; Kollman, P. A. The maximal affinity of ligands. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 9997−10002.

(31) Glick, M.; Klon, A. E.; Acklin, P.; Davies, J. W. Prioritization of high throughput screening data of compound mixtures using molecular similarity. *Mol. Phys.* **2003**, *101*, 1325−1328.

(32) Glick, M.; Klon, A. E.; Acklin, P.; Davies, J. W. Enrichment of extremely noisy high-throughput screening data using a naive Bayes classifier. *J. Biomol. Screening* **2004**, *9*, 32−36.

(33) Martin, Y. C.; Kofron, J. L.; Traphagen, L. M. Do structurally similar molecules have similar biological activity? *J. Med. Chem.* **2002**, *45*, 4350−4358.

(34) *Comprehsive Medicinal Chemistry*; MDL Information Systems: San Leandro, CA.

(35) *MDL Drug Data Report*; MDL Information Systems: San Leandro, CA.

JM049902R