Discovery of Power-Laws in Chemical Space

Ryan W. Benz,[†] S. Joshua Swamidass,[†] and Pierre Baldi*,[†,‡]

Institute for Genomics and Bioinformatics and Department of Biological Chemistry, School of Information and Computer Sciences, University of California, Irvine, Irvine, California 92697-3435

Received September 21, 2007

Power-law distributions have been observed in a wide variety of areas. To our knowledge however, there has been no systematic observation of power-law distributions in chemoinformatics. Here, we present several examples of power-law distributions arising from the features of small, organic molecules. The distributions of rigid segments and ring systems, the distributions of molecular paths and circular substructures, and the sizes of molecular similarity clusters all show linear trends on log−log *rank/frequency* plots, suggesting underlying power-law distributions. The number of unique features also follow Heaps'-like laws. The characteristic exponents of the power-laws lie in the 1.5−3 range, consistently with the exponents observed in other power-law phenomena. The power-law nature of these distributions leads to several applications including the prediction of the growth of available data through Heaps' law and the optimal allocation of experimental or computational resources via the 80/20 rule. More importantly, we also show how the power-laws can be leveraged to efficiently compress chemical fingerprints in a lossless manner, useful for the improved storage and retrieval of molecules in large chemical databases.

1. INTRODUCTION

The discovery and investigation of phenomena following power-law distributions have taken place in a wide variety of areas including physics, geology, biology, computer science, and economics as well as in the structure of the Internet and social networks.[1] Some of the first investigations of power-law distributions were performed by Pareto[2] who observed a power-law distribution of wealth among individuals as well as Estoup[3] and Zipf[4–6] who studied power-law behavior in the frequency of words used in printed texts. Mandelbrot has also studied power-laws and self-similarity arising in natural systems.[7] More recently, the Internet has shown several power-law trends including the distribution of Web site sizes[8] and the distribution of Web link connectivities.[9–11] Additional examples showing power-law distributions can be found in reviews by Mitzenmacher[12] and Newman.[13]

When studying power-law phenomena, two main issues arise. The first involves trying to understand the underlying mechanisms responsible for the power-law behavior. Just as power-laws have been observed in a wide variety of natural and man-made systems, several mechanisms for generating power-law distributions have also been proposed, including the Yule process, models of preferential attachment, simple random models, and distributions that are subject to two exponentials (see the reviews by Mitzenmacher[12] and Newman[13] for further descriptions of these and other mechanisms). No single mechanism has been found to adequately describe all observed power-laws making the process of explaining the existence of a power-law distribution more difficult than simply observing it. The second question of importance to consider is if, and how, the power-law distribution may be of use, either in terms of providing new insight into the phenomenon described by the power-law or through a practical application of the distribution itself (e.g., modeling of disease propagation, computer security).

Though power-laws have been observed in many areas, to the best of our knowledge, no power-law distributions have been reported in the field of chemoinformatics. This can likely be attributed to the lack of, and resistance to, open-access chemistry database systems,[14–16] which have only recently begun to emerge. Open-access databases, such as PubChem (http://pubchem.ncbi.nlm.nih.gov) and ChemDB[17] are, for the first time, enabling large-scale statistical data mining of chemical repositories by the general academic community and have facilitated the discovery of the power-law distributions reported here. In this paper, we present several examples of power-law distributions arising from the features of small molecules including the occurrence frequencies of molecular fragments and molecular fingerprint features and the sizes of clusters determined by clustering molecules according to their Tanimoto similarities with one another. First, a brief review of power-law distributions is presented followed by a description of the various molecular features investigated and how they were obtained. In the next section, the results from the analyses of the power-law distributions are shown including the characteristic exponents as well as further investigation of the distributions, their practical applications, and statistical significance. Finally, a discussion regarding the distributions is given.

2. METHODS

2.1. Review of Power-Law Distributions. When data that distribute according to a power-law are presented as a histogram, the power-law behavior produces a plot that is highly right-skewed. Unlike many other distributions such

* Corresponding author e-mail: @ics.uci.edu.
[†] Institute for Genomics and Bioinformatics.
[‡] Department of Biological Chemistry.

10.1021/ci700353m CCC: $40.75
Published on Web 06/04/2008

as the Gaussian, Poisson, and exponential distributions, power-law distributions decay polynomially instead of exponentially. Power-law distributions are also said to be scale-free, due to the fact that the shape of the distribution looks the same regardless of the scale on which it is observed. This allows one to transfer the behavior observed at one scale to other scales and has given rise to the so-called "80−20" rule, which states that approximately 80% of the distribution can be accounted for by 20% of the values.

Distributions that follow power-law behavior can be written as

$$p(x) = kx^{-\alpha} \tag{1}$$

where k is the normalization constant of the distribution, and $\alpha > 1$ is the exponent of the power-law. In the reported power-law literature, α typically lies in the 1.5−3 range. By taking the log of (1)

$$\log p(x) = -\alpha \log x + \log k \tag{2}$$

the equation takes the form of a line with slope $-\alpha$. As a result, the presence of power-law behavior can be identified by plotting a histogram of a given data set on a log−log scale to determine if the data follow a linear trend. Often times, a data set will not show a linear trend over its entire range but rather above some minimum value, $x_{\min}$. By considering the data above this value, the properties of the power law distribution can be investigated.

In practice, it is usually easiest to identify power-laws by so-called *rank/frequency* plots which are related to $P(x)$, the complementary cumulative distribution function of $p(x)$, which gives the probability of observing a particular value greater than x:

$$P(x) = \int_x^\infty p(x')dx' \tag{3}$$

If $p(x)$ can be written as (1), the complementary cumulative distribution becomes

$$\begin{aligned} P(x) &= k\int_x^\infty x'^{-\alpha}dx' \\ &= \frac{k}{\alpha - 1}x^{-(\alpha-1)} \end{aligned} \tag{4}$$

which is the same form as (1) but with a smaller exponent. An unnormalized complementary cumulative distribution of a data set can be equivalently constructed by simply counting the number of times each value occurs, ranking the data by decreasing frequency, and plotting the rank of each value versus its frequency. If x distribute according to a power-law, a straight line will emerge in this *rank/frequency* plot on a log−log scale with a slope of $-(\alpha-1)$. For many data sets, the quantity of interest is not a frequency but rather a measure expressing a size or an amount. Regardless, a plot analogous to a *rank/frequency* plot can still be produced and is often still referred to by this name despite the fact that a frequency is not directly plotted. While *rank/frequency* plots are typically presented by plotting rank vs frequency, the axes are sometimes switched, which may be a source of confusion. However, regardless of how a *rank/frequency* plot is presented, the underlying data are still the same. It should also be noted that an apparent line on a log−log *rank/frequency* plot does not necessarily imply power-law behavior. For example, data generated from a log-normal distribution with a large width parameter can also produce a seemingly linear trend on a *rank/frequency* plot, making it difficult to differentiate the two models from one another. A recent article by Mitzenmacher[12] provides a concise review of power-law and log-normal distributions.

One of the key values of interest in a power-law distribution is the exponent, α. Unlike other distributions, such as a Gaussian distribution, the convergence of the moments in a power-law distribution depends upon its parameters, namely α. For example, the mean of a power-law distribution is written as

$$\langle \chi \rangle = k\int_{x_{\min}}^\infty x(x^{-\alpha})dx \tag{5}$$

$$= \frac{k}{2-\alpha}[x^{2-\alpha}]_{x_{\min}}^\infty \tag{6}$$

and only converges when $\alpha > 2$. Similarly, the variance diverges for $\alpha \leq 3$. In practice, the nth moment can still be calculated for a data set following a power-law distribution with exponent $\alpha \leq n+1$ by using the maximum observed value, $x_{\max}$, instead of infinity as the upper limit of the integral in calculating the moment. In this case however, the moment will depend on $x_{\max}$ and may not converge to a unique value as more data are observed.

Power-laws have been studied in great detail in the context of word frequencies (i.e., Zipf's law). While the frequencies of real words in many different languages have been shown to distribute via power-laws, it has also been shown that the frequencies of random letter strings follow power-law distributions as well. Specifically, Li has presented a proof for the power-law nature of random word frequency distributions given an alphabet with equal letter probabilities and an associated space probability (used to separate words).[18] For random words generated from alphabets with unequal probabilities, Li also observed power-law distributions empirically, and a more complex proof of this behavior is given by Conrad et al.[19]

In addition to the observation of power-laws in the field linguistics via Zipf's law,[4–6] another "law" know as Heaps' law[20] is also commonly used to investigate the distribution of word frequencies. While Zipf's law describes the power-law distribution of word frequencies in texts, Heaps' law relates the number of unique words (vocabulary size) to the total size of the text via a polynomial

$$V(n) = An^\beta \tag{7}$$

where $V(n)$ is the vocabulary size of a text containing n words, and A and β are parameters of the distribution. Empirically, the value of β has been found to be approximately 0.5.[21,22] It has also been shown that Heaps' law can be derived formally from the Mandelbrot distribution. The Mandelbrot distribution has Zipf's law as a special case.[23]

2.2. Data Sets. To study the power-law trends present in the context of chemoinformatics, data sets consisting of molecules from the ChemDB,[17] which contains approximately 5 million compounds, were used. From the molecules considered, various molecular features, described below and illustrated in Figure 1, were extracted, and their resulting distributions were investigated. Features were generated from a random sample of 50,000 molecules in order to reduce the time required to obtain the features. In the study of similarity cluster sizes, a larger random sample of 1.5 million compounds was considered. To facilitate the parsing and

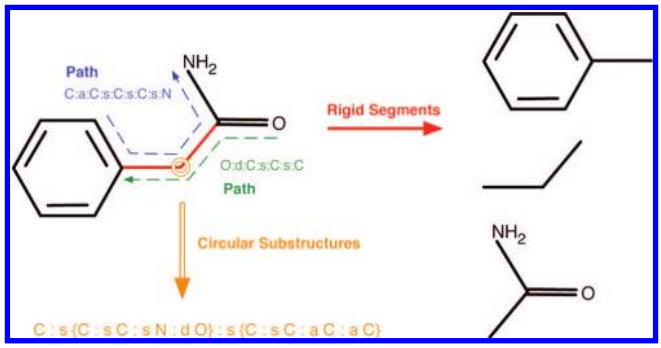

Figure 1. Illustration of the investigated molecular features. A 2D representation of a molecule is shown in black from which several features are extracted. The molecule is fragmented into rigid segments by breaking the molecule apart at its rotatable bonds, colored in red. As the molecule contains $n = 2$ rotatable bonds, $n + 1 = 3$ rigid segments are obtained. Examples of labeled paths are shown in green ($d = 3$) and blue ($d = 4$) accompanied by arrows showing the paths taken to produce them. Here the atoms are labeled according to their element symbols and the bonds are labeled as follows: s = single bond, d = double bond, a = aromatic bond. Finally, an example of a circular substructure label with a depth of 2 is shown for the carbon atom indicated by the orange circle. The label for this carbon is composed of the circular substructure labels for the neighboring atoms connected by one bond. The given carbon has two neighbors whose labels are enclosed in braces, also composed of their neighboring atom and bond labels.

generation of the features, both the OpenBabel[24] (http://openbabel.sourceforge.net) and OpenEye OEChem (http://www.eyesopen.com) libraries were used.

2.3. Rigid Segments and Ring Systems. To determine the distribution of rigid segment occurrences, the individual rigid segments from each molecule were extracted (see Figure 1). Here, rigid segments are defined as the largest fragments containing no rotatable bonds in a molecule. To extract the rigid segments, the rotatable bonds in each molecule were identified using the OEChem toolkit. At each rotatable bond, the molecule was separated into two pieces by breaking the rotor bond, and the adjacent rotor atoms were then reattached to each segment. This process was repeated until all of the separated segments contained no rotor bonds, giving a set of rigid segments for the processed molecule. As such, a molecule with n rotatable bonds will contain $n+1$ rigid segments. After fragmentation, the rigid segments for all of the molecules in the data sets were written to a file in OEChem isomeric SMILES format, and the frequencies of the rigid segments were calculated based upon the SMILES representations to produce *rank/frequency* plots.

Heaps' law for the distribution of rigid segments was also investigated by considering how the number of unique rigid segments varies with the total number of rigid segments observed. In this case, the number of unique rigid segments acts as a "vocabulary" of rigid segments, and the total number of rigid segments represents the length of the rigid segment "text". From the extracted rigid segments, the data were fit to eq 7 to test the applicability of Heaps' law. A similar procedure was also performed for the path and circular substructure features described below.

In addition to fragmenting the molecules into their rigid segments, the ring systems from the molecules were also extracted. Ring systems were identified using the OEChem toolkit, where a ring system is defined as a molecular fragment which is connected through ring bonds only. Like the rigid segments, the ring systems were written to a file as isomeric SMILES strings for frequency and distribution analysis.

2.4. Paths and Circular Substructures. Labeled molecular paths, consisting of alternating atom and bond symbols, were extracted from the investigated molecules as described in ref 25. The frequencies of the paths were then used to study the resulting distributions for power-law behavior. Paths of length $d = 1$ to $d = 8$ were obtained using a depth-first search starting from each atom in the input molecule, where the path length is equal to the number of connections in the path. Path frequencies were investigated by counting the occurrences of the extracted labeled paths. Except where indicated, atoms were represented by their element symbols and bonds were classified into five types (single, double, triple, aromatic, and amide) as parsed by the OpenBabel library.

An important question to consider is if and how the labeling of the paths affects the resulting distributions. At one extreme, all of the atoms and bonds could be labeled uniquely, fully differentiating paths from one another. At the other extreme, the same labels could be used to represent each atom and each bond, providing no differences between the paths. To address this question, a test was performed using four different labeling schemes in this spectrum. The first and most specific scheme labeled the atoms by SYBYL atom type (Tripos Inc., http://www.tripos.com) and used five different bond types (single, double, triple, aromatic, and amide). The second scheme represented atoms by their element symbols and used the same five bond types as the first scheme. The third labeling scheme consisted of three bond types (single, double, triple) and element symbols. In the last and least specific scheme, element symbols were used to represent the atoms, and all bonds were represented simply as connections (single bonds). Using these labeling schemes, labeled paths were produced, and the resulting frequency distributions were compared via *rank/frequency* plots.

To further understand the path distributions and investigate the possibility that simple generative models can be used to generate paths with distributions similar to those observed empirically, Markov models trained on the paths extracted from the investigated molecules were produced. The symbols of the Markov models consisted of individual atom labels (e.g., C, N, O) as well as combined bond + atom labels (e.g., sC: single bond + carbon, dO: double bond + oxygen). The individual atom symbols were needed to initiate the path strings, and subsequent elements were selected from the bond + atom symbols. To prevent the possibility of generating nonphysical paths by violating atom valencies, second-order and higher models were used. To generate paths of a given length, the initial and transition probabilities were calculated from the paths of the same length extracted from the ChemDB molecules. Using these probabilities, paths were then generated in a Monte Carlo manner, and the resulting path frequency distributions were compared to the corresponding ChemDB distributions.

Another class of chemical features, known as circular substructures,[26,27] was also investigated. In this type of representation, each atom in a molecule is labeled based upon its extended neighbors using a modified Morgan algorithm. Instead of using connectivity numbers, the neighboring atom and bond labels are used to construct the circular substructure

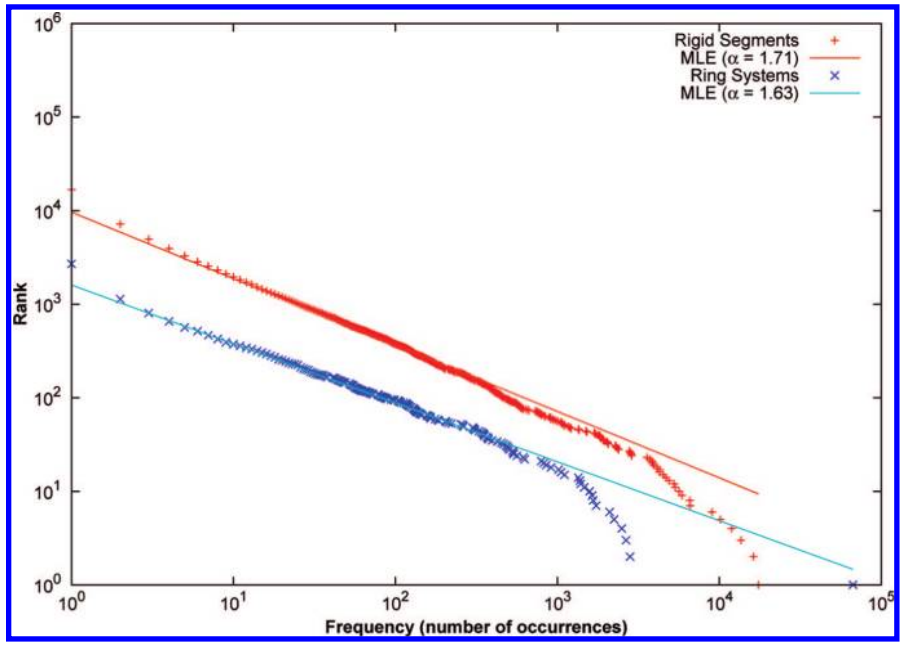

Figure 2. Rank/frequency plots for the rigid segments and ring systems extracted from the ChemDB molecules with their associated power-law exponent MLEs. Linear trends are present on the log−log scale for both sets of molecular fragments, though deviations are seen at high frequencies.

label for each atom. Circular substructures of depth $d = 0-2$ were extracted, where the depth indicates the maximum distance, in number of bonds, to the extended neighbors to consider. For example, a depth of two indicates that the label for a given atom will be composed of the labels for the neighboring atoms which are connected by at most two bonds. The frequency distribution of these labels was then investigated.

2.5. Cluster Sizes. Clustering, a general method of unsupervised data analysis, can be usefully applied in chemoinformatics to investigate the coverage and properties of chemical space represented by a collection of molecules. For example, from a set of molecules clustered according to their similarity to one another, the number of clusters obtained gives an indication of molecular diversity, and the cluster prototypes provide a reduced set of characteristic molecules that represent the molecular properties in the set. Molecules from the ChemDB were separated into clusters using path-based features of lengths up to $d = 8$ with SYBYL labeling and corrected Tanimoto similarity measures.[28] A simple incremental clustering algorithm was used which adds a given molecule to an existing cluster when its Tanimoto similarity to the cluster prototype is greater than a set threshold. If none of the Tanimoto similarities between a molecule and the existing cluster prototypes reaches the threshold, the molecule is assigned to a new cluster. Tanimoto thresholds of $T = 0.7$ and $T = 0.8$ were used in the clustering. Using a fast fingerprint search algorithm,[29] clustering scales as $O(D^{1.6})$, slightly faster than a naive $O(D^2)$ algorithm, where D is the size of the data set.

2.6. Calculation of the Power-Law Parameters. Though power-law distributions manifest themselves as straight lines on log−log distribution plots, extracting the exponent of the distribution using least-squares regression of the log transformed data has been shown to give inaccurate and biased estimates.[13,30,31] Therefore, the power-law exponents were calculated using a maximum likelihood estimator (MLE) of the power-law distribution.[13,31] In the case of continuous data, the power-law MLE can be easily calculated by

$$\alpha = 1 + \frac{N}{\sum_{i=1}^{N} \ln(x_i/x_{\min})} \tag{8}$$

where N is the total number of data points $x_i \geq x_{\min}$. For discrete data, the MLE equation is more complex, and α can be calculated according to

$$\frac{\zeta'(\alpha, x_{\min})}{\zeta(\alpha, x_{\min})} = -\frac{1}{N}\sum_{i=1}^{N} \ln x_i \tag{9}$$

or by numerically maximizing the log-likelihood function

$$\boldsymbol{L}(\alpha) = -N\ln\zeta(\alpha, x_{\min}) - \alpha\sum_{i=1}^{N} \ln x_i \tag{10}$$

Here, $\zeta(\alpha, x_{\min})$ refers to the generalized Reimann zeta function, and $\zeta'(\alpha, x_{\min})$ is the first derivative with respect to α. Though the equations in the discrete case do not lead to an exact closed-form solution for α, an approximate MLE can be written as

$$\alpha = 1 + \frac{N}{\sum_{i=1}^{N} \ln[x_i/(x_{\min} - 0.5)]} \tag{11}$$

which is the same as the MLE in the continuous case, except for the additional -0.5 term in the denominator of the logarithm, and provides an approximate estimate of α when highly accurate results are not needed. See refs 13 and 31 for derivations and additional discussion on the power-law estimators.

Because all of the data considered here are discrete, the power-law exponents were calculated by maximizing the discrete log-likelihood function (eq 10). The $x_{\min}$ value for each data set was calculated using the approach given in ref

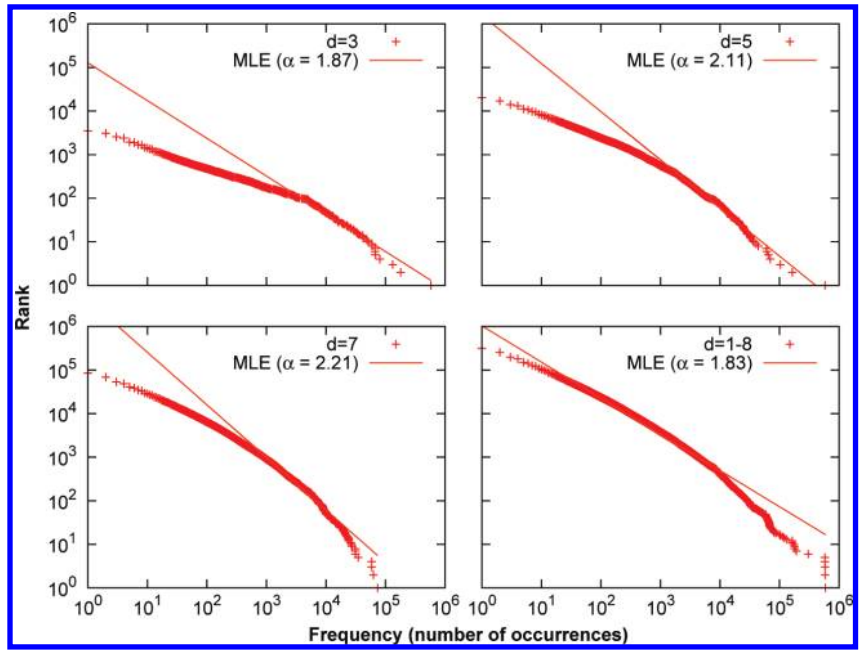

Figure 3. *Rank/frequency* plots for paths of fixed-length $d = 3$, 5, and 7 and combined lengths $d = 1-8$. Compared to the individual path distributions at fixed-lengths, which show power-law behavior mainly in the tails of the distributions, the combined distribution containing all paths, lengths $d = 1-8$, shows linear behavior over a much wider range. For the fixed-length distributions, the x_{min} were selected by hand to capture the linearity in the tails.

31. In this method, the best fit power-law to the data is calculated above x_{min} and compared to the actual data. The difference between the best fit and empirical distributions is then quantified by the Kolmogorov−Smirnov (KS) statistic

$$D = \max_{x \geq x\text{min}} |C(x) - P(x)| \quad (12)$$

where $C(x)$ is the cumulative distribution of the empirical data, and $P(x)$ is the cumulative distribution of the best-fit power-law. The final estimate of x_{min} is chosen such that D is minimized. These calculations were performed using in-house scripts as well as scripts provided by Clauset et al.[31] (http://www.santafe.edu/~aaronc/powerlaws/).

3. IDENTIFICATION OF THE POWER-LAWS

The frequency distributions of the molecular features described above were calculated and investigated for power-law behavior. The details and results are presented for each feature in the following subsections, including the calculated power-law exponents, α.

3.1. Molecular Fragment Distributions. The distributions of the rigid segments and ring systems from the ChemDB molecules are shown in Figure 2 as *rank/frequency* plots. From the data, linear trends for both sets of features are apparent, though deviations exist at large frequencies. These plots also show similar distributions and MLE power-law exponents, calculated as 1.71 for the rigid segments and 1.63 for the ring systems.

3.2. Molecular Path and Circular Substructure Distributions. Figure 3 shows the distributions of the labeled paths extracted from the ChemDB data set for paths of length $d = 3$, 5, and 7. The proposed power-law behavior is weaker and extends over a smaller range compared to the rigid segment and ring system data, located mainly in the tails of the distributions. However, the extent of the linearity as well as α increases with d. These results suggest that the larger diversity of possible paths available at longer lengths may contribute to the power-law nature of the frequency distributions.

By considering the labeled paths as "words", the path distributions were also studied in the context of Zipf's law, originally used to describe the frequencies of word occurrences in texts. As the word frequencies investigated in Zipf's law studies are not separated by word length (i.e., the frequencies of all words of any length, or up to a maximum length, are considered), the path frequencies of length $d = 1$ to $d = 8$ were combined into a single data set producing a "text" of labeled paths. Like the distribution of words in real texts, the resulting *rank/frequency* plot in the last panel of Figure 3 shows linear behavior throughout most of the plot. Compared to the fixed-length path results, which display power-law behavior mainly in the distribution tails, the combined path length results ($d = 1-8$) show linearity over more than 3 orders of magnitude (Figure 3).

In order to test the effect labeling has on the path distributions, *rank/frequency* plots for paths of length $d = 8$, labeled by the four different schemes described in section 2.4, were produced (Figure 4). Except for the labeling scheme using only one bond type, the other schemes produce similar path frequency distributions. These results indicate that labeling does not greatly affect the path distributions, as long as the atoms and bond labels are specified to a reasonable extent. When a large degree of information, important to the representation and differentiation of the paths, is removed like in the least specific scheme with only one bond type, significant changes in the resulting path distributions can be expected.

In Figure 5, the combined distribution of circular substructures is shown for depths $d = 0-2$. Similar to the combined path distribution for $d = 1-8$ (Figure 3), the circular substructure *rank/frequency* plot also shows linearity

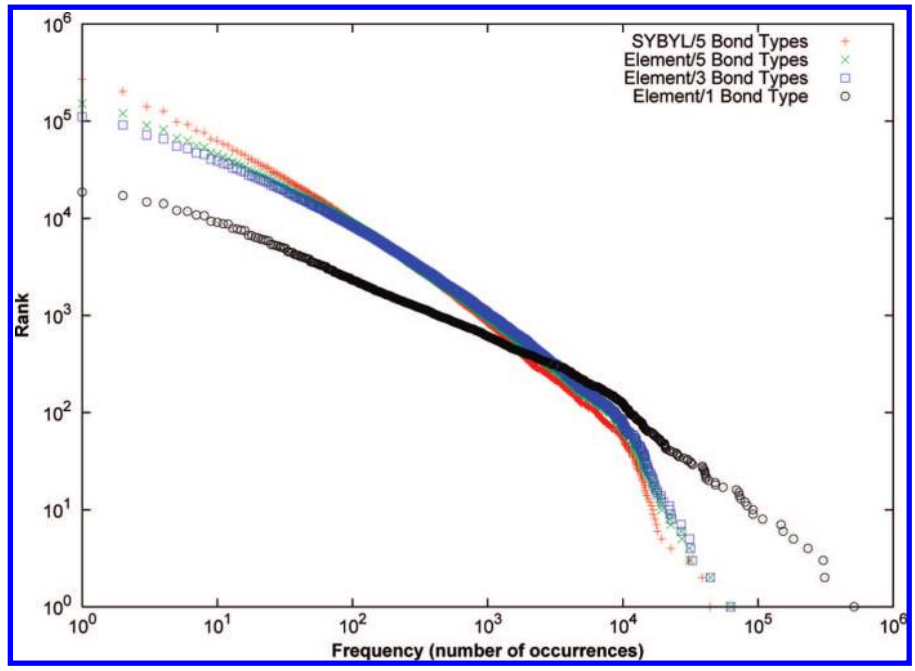

Figure 4. *Rank/frequency* plots for paths of length $d = 8$ specified using the four different labeling schemes described in section 2.4. The results from the most specific scheme are shown in red, followed by green, blue, and finally black for the least specific scheme. The distribution from the least specific scheme shows significant differences compared to the other three distributions.

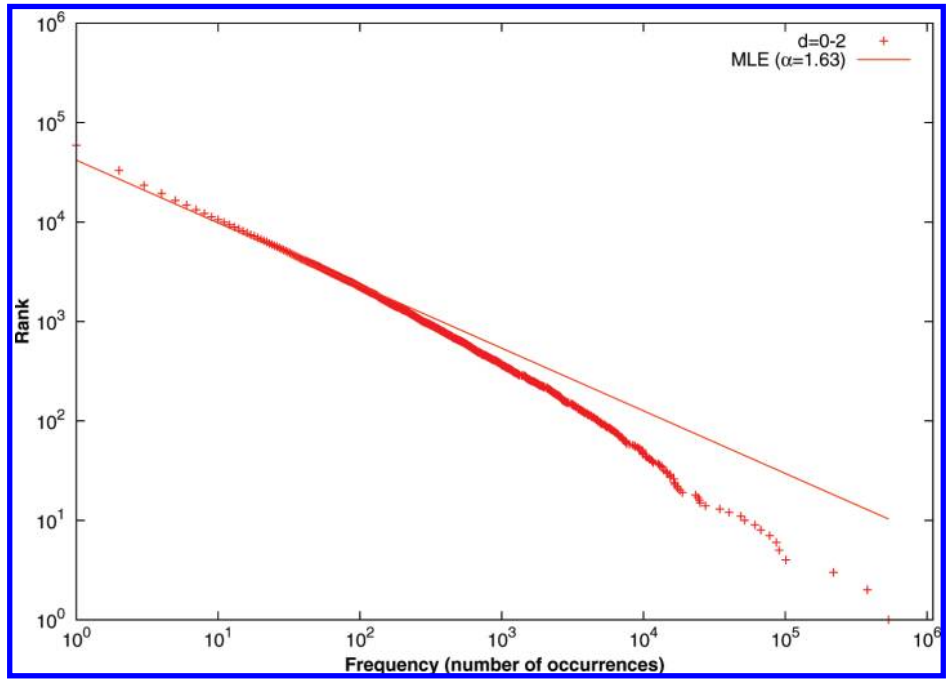

Figure 5. Distribution of circular substructures extracted from the ChemDB molecules for depths $d = 0-2$. Like the combined path length distribution shown in Figure 3 ($d = 1-8$), linearity in the data is seen over nearly 3 orders of magnitude.

in the distribution over ~3 orders of magnitude. Furthermore, when the circular substructure features at fixed-depths are considered individually (data not shown), a decreased extent of linearity in the resulting *rank/frequency* plots is seen, similar to that observed for the path distributions at fixed-length (Figure 3).

3.3. Comparison of Path Frequencies and Random Models. As mentioned in the Introduction, the frequencies of random words have been shown to follow power-law behavior.[18,19] To investigate the possibility that the power-laws observed for the labeled paths can be attributed to these features acting like random "words", Markov models trained on the paths extracted from the ChemDB molecules were used to generate random labeled paths of lengths $d = 4$, 6, and 8. For each length, the number of paths generated was set to the corresponding number of paths extracted from the investigated molecules. In Figure 6, the *rank/frequency* plots of path distributions from ChemDB are compared to those obtained from the trained random models. In the first three panels, paths of length $d = 4$, 6, and 8 are considered with second-order Markov models. Similar distributions are seen between the ChemDB and random path results, though

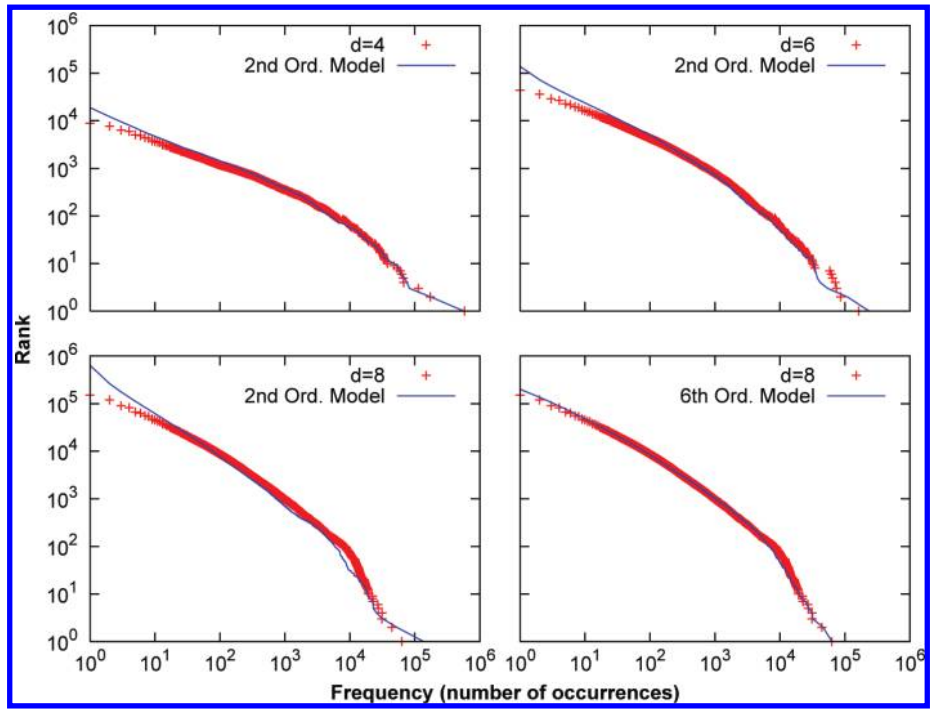

Figure 6. Comparison of path distributions obtained from the ChemDB data set and second- ($d = 4$, 6, and 8) and sixth-order ($d = 8$) Markov models. For the second-order models, shown in the first three panels, the distributions of the random paths approximate the ChemDB distributions well, though differences are seen as the path length increases. The distribution from the sixth-order model for $d = 8$ (last panel), however, is very close to the empirical path distribution.

Table 1. Summary of the Labeled Path Power-Law Exponent MLE and x_{mix} Values

path length	α	x_{min}
$d = 3$	1.87	4000
$d = 5$	2.11	3000
$d = 7$	2.21	2000
$d = 1-8$	1.83	174

deviations increase with the path length. In the last panel, the distribution of paths from a sixth-order model is compared to the corresponding ChemDB distribution for paths of length $d = 8$. Unsurprisingly, as the sixth-order model contains more information about the constitution of the labeled paths from the ChemDB molecules, it produces a path distribution closer to the ChemDB results than the second-order model.

Though the *rank/frequency* plots in Figure 6 may suggest that the empirical and random model distributions are highly similar, differences between them can be masked by the log−log scale. To better understand the differences between the ChemDB and randomly generated paths, the average absolute difference between path frequencies in the two data sets was computed along with the associated standard deviation (Table 2). With the exception of an increase in the standard deviation between the $d = 4$ and $d = 6$ results, both the mean absolute difference and standard deviation decrease with increasing path length for the second-order models. While the increasing differences seen between the *rank/frequency* plots in Figure 6 with increasing path length might suggest that the mean and standard deviation should also increase, the observed decrease can be attributed to the fact that the number of paths in common between the empirical and trained random models decreases with increased path length. Specifically, the second-order models tend to produce a much more diverse set of paths compared to the empirically observed paths, particularly at longer path lengths, resulting in differences between the distributions. This is in contrast to a similar decrease in mean absolute difference and standard deviation values that also occurs as the order of the Markov model increases (from a second-order to a sixth-order model) at fixed path length ($d = 8$). In this case, the fraction of paths in common between the empirical and random model data sets is larger compared to the second-order model results, and the decrease in the mean absolute difference and standard deviation can be attributed to the random path frequencies more closely matching those

Table 2. Comparison of the Frequencies of Paths Extracted from the ChemDB Molecules and Randomly Generated from Markov Models[a]

path length	Markov model order	mean absolute difference	standard deviation	% common paths
$d = 4$	2	124	825	20
$d = 6$	2	92	1381	10
$d = 8$	2	61	613	6
$d = 8$	6	19	211	23

[a] The average absolute difference was calculated as $1/N\sum_i^N|f_{CDB}(path_i) - f_{MM}(path_i)|$ over each path, $path_i$, found in common between the two data sets, where N is the total number of common paths, and $f_{CDB}(path_i)$ and $f_{MM}(path_i)$ are the frequencies that $path_i$ occurs in the ChemDB and Markov model sets, respectively. The standard deviation was calculated according to $\sqrt{1/(N-1)\sum_i^N(d_i-m)^2}$, with $d_i = |f_{CDB}(path_i)-f_{MM}(path_i)|$, and m is the average absolute difference. A general trend of decreasing average absolute difference and standard deviation is seen both as path length increases at fixed Markov model order and for increasing Markov model order at fixed path length. In the latter case, the decrease can be attributed to the random path frequencies more closely matching those from the ChemDB.

Table 3. Summary of the Power-Law Exponent MLE and x_{min} Values for the Molecular Fragments, Labeled Path and Circular Substructure Features, and Similarity Cluster Size Distributions

feature	α	x_{min}
rigid segments	1.71	3
ring systems	1.63	2
paths ($d = 1-8$)	1.83	174
circular substructures ($d = 0-2$)	1.63	1
cluster sizes ($T = 0.7$)	2.80	47
cluster sizes ($T = 0.8$)	3.20	53

from the ChemDB molecules. While not perfect, these results suggest that Markov models can be used to produce realistic labeled paths, to a first-order approximation. Furthermore, by addressing the deficiencies found in these simple random models, more realistic generative models for molecular paths can be developed.

Further exploring the possible connection between the power-law behavior in the frequencies of labeled paths and random words, random sequences of fixed-lengths $k = 5, 6, 7$, and 8 were generated from an alphabet of five symbols with unequal occurrence probabilities (1/15, 2/15, 3/15, 4/15, 5/15). The individual fixed-length and combined sequence distributions were then computed for comparison with the ChemDB path distributions (Figure 7). While the combined random sequence frequencies show linear behavior throughout the entire distribution as expected by Zipf's law[18,19] (Figure 7B), the individual fixed-length random sequence distributions do not (Figure 7A). This is similar to what is observed for the labeled paths, where the individual fixed-length path distributions exhibit power-law behavior mainly in the tails of the distributions (Figure 7C), and the combined path frequencies show linear behavior throughout most of the distribution (Figure 7D). Furthermore, both the individual fixed-length path and random sequence distributions show an increase in the extent of the distribution linearity with increasing length. As the ChemDB path distributions display many of the same features as the random sequence distributions, the power-law behavior for the path frequencies may, in part, be explained by the same arguments describing the existence of power-laws in random sequence distributions.[18,19]

3.4. Cluster Size Distributions. Figure 8 shows the *rank/frequency* plots for the sizes of clusters generated by clustering molecules based upon Tanimoto similarity with threshold values of $T = 0.7$ and $T = 0.8$. The distributions of the cluster sizes show linear trends above cluster sizes of approximately 10 molecules. The larger value of α for the $T = 0.8$ distribution is consistent with the production of more clusters of smaller size, as expected for higher thresholds.

4. APPLICATIONS OF THE POWER-LAWS

4.1. Predicting the Growth of Feature Libraries. By plotting the number of unique rigid segments versus the total number of rigid segments, Heaps' law was investigated for the rigid segment data, shown in Figure 9, along with a fit to eq 7. With an R^2 value of 0.99, the data are well approximated by Heaps' law with an exponent of $\beta = 0.60$. We have also observed that Zipf's and Heaps' laws describe the distribution of rigid segments from molecules extracted from the Cambridge Structural Database (CSD).[32] Based upon a Heaps' law fit using molecules from the 2004 CSD, the number of unique rigid segments from the 2006 CSD was predicted (289, 700) and differed from the actual value of 285, 288 by less than 2%.

Using a similar procedure, the applicability of Heaps' law to the path and circular substructure features was also investigated. The frequencies of the unique ChemDB paths of lengths $d = 1-8$ was fit to eq 7 and were found to be well approximated by Heaps' law with $\beta = 0.46$. Likewise, the circular substructure data are also well approximated by Heaps' law with an exponent of $\beta = 0.48$. These Heaps' laws provide the means to estimate the future growth of feature libraries, useful for the development and design of chemical database systems.

4.2. The 80/20 Rule. The 80/20 rule refers to a commonly observed behavior associated with power-laws whereby 80% of a global effect comes from only 20% of its underlying factors, as in "80% of sales come from 20% of the clientele". Identifying this top 20% is important for maximizing the return on one's investment and to better understand the most important contributions of a particular outcome. Despite the name of the rule, the exact percentages can vary depending upon the situation described. Furthermore, due to the self-similarity property of the underlying power-law, the 80/20 rule applies at any scale. For example, 80% of the remaining 20% of sales not covered by the top 20% of the clientele is covered by the next top 20% of the clientele, and so on.

The consequences of the 80/20 rule for the power-laws described above can be useful for guiding both experimental and computational studies. For example, in the analysis of mass spectrometry data, the fundamental task is to deduce the structure of a molecular fragment from its mass-to-charge ratio. In general however, a mass-to-charge ratio alone is not enough to uniquely identify the corresponding fragment structure, and other sources of information, including fragment databases, are often needed to properly analyze the data. While the immense size of chemical space precludes the storage of all possible molecular species in a fragment database, the power-laws for the rigid segments, ring systems, paths, and circular substructures all indicate that a small fraction of fragments can account for a large majority of overall fragment occurrences. For the rigid segment distribution described in section 2.3, the top 10% of the most frequently occurring rigid segments account for more than 90% of the total rigid segment occurrences. Therefore, the 80/20 rule indicates that a fragment database may only need to contain a relatively small number of entries to effectively cover a large portion of search space. Furthermore, the frequency ranking of the fragments can provide a general indication of how likely a particular fragment match is among a set of fragments with the same or very similar mass-to-charge ratios. This idea could also be applied in reverse when more molecular diversity is desirable by focusing on the least frequently occurring molecular fragments, for example, in X-ray crystallography studies seeking to determine novel chemical structures, or for increasing the coverage of chemical space in small molecule databases.

4.3. Efficient Compression of Chemical Fingerprints. In addition to using the power-law distributions and Heaps' law to predict the growth of the libraries of molecular features or employing the 80/20 rule to help guide experimental or computational studies, the chemical power-law distributions can also be used to losslessly and efficently compress

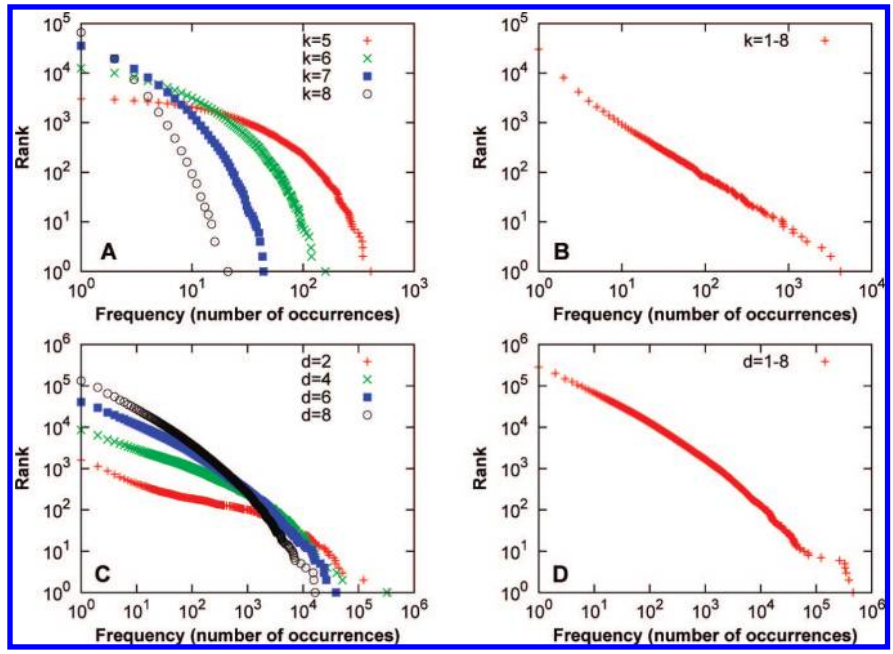

Figure 7. Comparison of the frequency distributions from random letter sequences (Panels A and B) and labeled paths extracted from the ChemDB molecules (Panels C and D). The fixed-length random sequence distributions (Panel A) do not show much power-law behavior, similar to the fixed-length labeled path distributions (Panel C), which show linearity mainly in the tails. However, when the random sequences or labeled paths are combined into a single sets (Panels B and D, respectively), both distributions show increased linearity in the log−log *rank/frequency* plots.

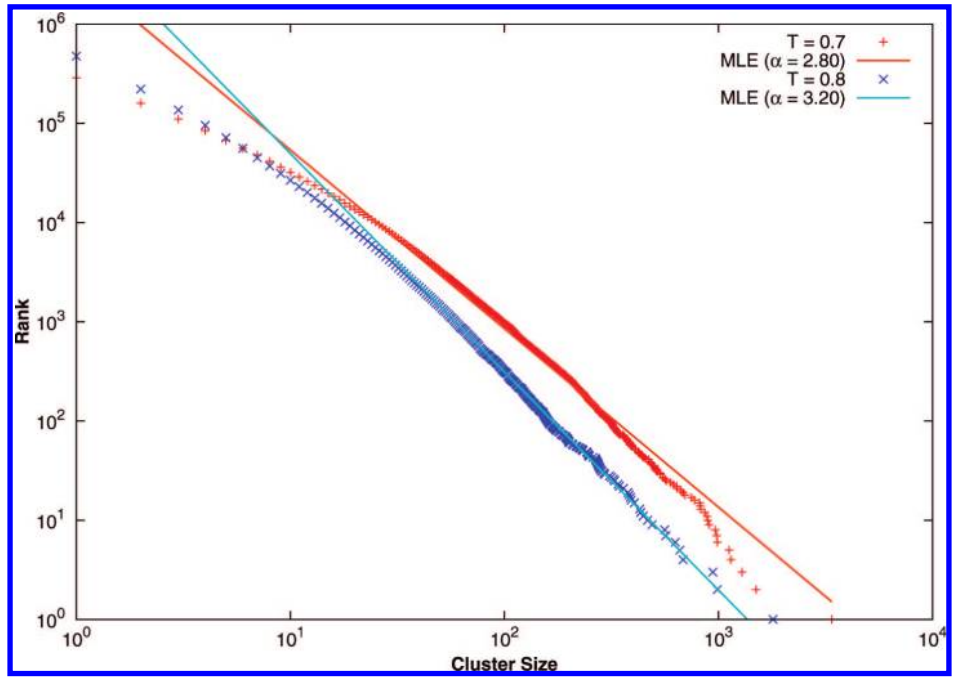

Figure 8. Distributions of similarity cluster sizes for molecules taken from the ChemDB with the associated MLE power-law exponents. The data show linear behavior above cluster sizes of approximately 10 molecules.

chemical fingerprints. Currently in many chemical database systems, fingerprints are based upon features such as labeled paths and trees, which are indexed in bit-vectors containing 10^4-10^6 bits. These long, sparse bit-vectors are then modulo wrapped and combined using the logical OR operator to produce much smaller, compressed vectors, typically 1024-bits in size. Though effective at reducing the size of these vectors, this form of compression is lossy since the original, uncompressed bit-vector cannot be recovered from its compressed form. Alternatively, a new compression scheme based upon the power-law feature distributions can be used to losslessly encode chemical fingerprints in an extremely compact form.

In this compression algorithm, bit-vectors are first converted to index or run length representations, where the indices refer to the locations of the bits set in the uncompressed bit-vector, and the run lengths indicate the number of unset bits between pairs of set bits. Next, the fingerprint bits are sorted by their power-law distributed frequencies, such that more frequent features are mapped to smaller index

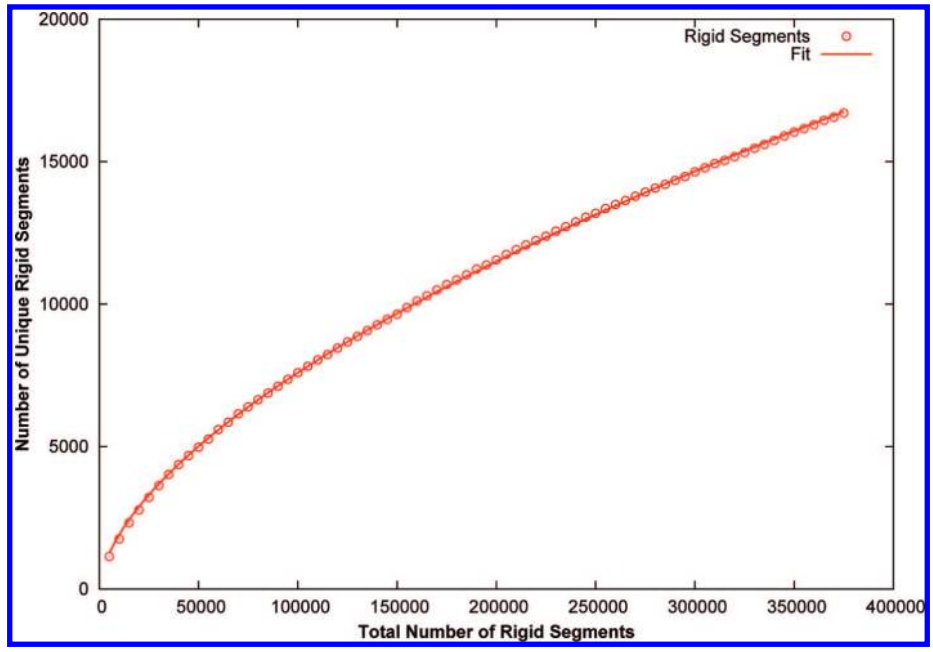

Figure 9. Heaps' law plot for the ChemDB rigid segments (circles: data, line: fit). Equation 7 was used to produce the fit, which describes the observed data well with an exponent of $\beta = 0.60$. Using this Heaps' law, the future growth of the library of rigid segments can be accurately estimated.

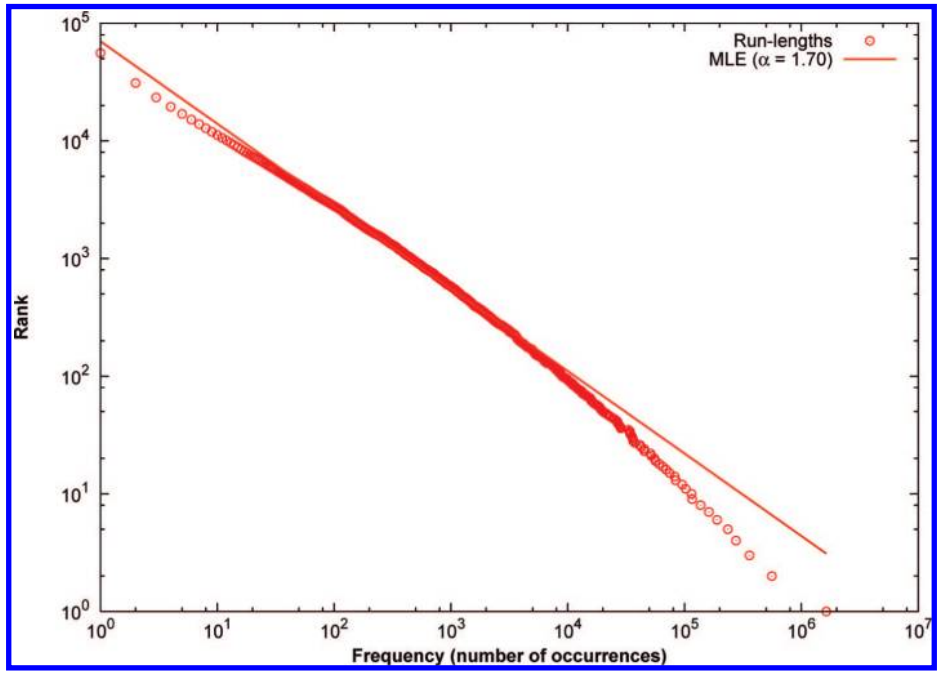

Figure 10. Distributions of run lengths from path-based fingerprints. When the fingerprint bits are sorted by decreasing frequency, the run length distribution becomes power-law and can be used to efficiently compress chemical fingerprints in a lossless manner.

values. Since the distribution of the indices is directly related to the corresponding feature distribution, it follows the same power-law, with and without sorting. However, while the distribution of run lengths is geometric when fingerprint bits are randomly ordered, the sorting process causes the run length distribution to become power-law (Figure 10). These lists of integers can then be compressed using integer entropy encoding schemes.[33–35] In general, an integer n is represented by a bit string $p(n)m(n)$, where $p(n)$ encodes the scale of n, and $m(n)$ is the binary encoding of n. In a list of encoded integers, the $p(n)$ act as delimiters and are necessary for proper decoding. Finally, using the fact that within a particular fingerprint, the list of indices is strictly increasing, and the list of run lengths is quasi-increasing, only changes in $p(n)$ are encoded, as opposed to their full values. An example of this encoding scheme is given in Figure 11. With this compression algorithm applied to circular substructure features represented as run lengths, molecular fingerprints can be losslessly encoded using approximately 300 bits, or 1/3 the size of typical 1024-bit, lossy compressed fingerprints (Figure 12). This compression scheme requires only a slight increase in the computational overhead when computing similarity measures between pairs of molecules, compared to a direct similarity calculation between compressed fin-

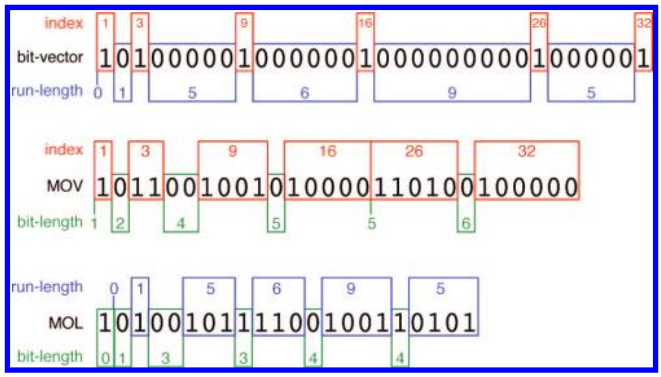

Figure 11. Examples of an uncompressed bit-vector and two losslessly encoded representations using the power-law-based compression method. The bit-vector bits are assumed to be ordered by decreasing frequency such that the first set bit at index 1 corresponds to the most frequently occurring feature. Using the compression scheme described in section 4.3, the uncompressed bit-vector can be efficiently compressed by encoding either a list of index values corresponding to the set bits (MOnotone Value code, MOV) or a list of run lengths corresponding to the number of unset bits between pairs of set bits (MOnotone Length code, MOL). For the MOV and MOL codes, the indices and run lengths are encoded in binary, indicated by the red and the blue boxes, respectively. In the case of the MOV code, the indices are strictly monotonic, and the 0-bits between the encoded indices signal an increase in the bit length for the next index value to decode. In the MOL code, the 0-bits between encoded run lengths also signal an increase in the bit length, while the 1-bits indicate cases where the bit length of the next run length either decreases or stays the same. These 1-bits are needed because run lengths, after bit sorting, are quasi-monotonic (i.e., periodic decreases are possible). Both the MOV and MOL codes provide efficient compression of long bit-vectors without loss of information.

Table 4. Power-Law Distribution *p*-Values for the Investigated Data Sets[a]

distribution	*p*
rigid segments	0.31
ring systems	0.03
cluster sizes ($T = 0.7$)	0.01
cluster sizes ($T = 0.8$)	0.73

[a] Here, a larger *p*-value indicates that there is more evidence for a power-law distribution characterizing the data.

gerprints. Small modifications can also be made to the compression scheme to encode count-based fingerprints or fingerprints for chemical reactions. While the compression scheme has been described in the context of encoding chemical fingerprints, the approach is completely general and can be applied in other domains that employ binary vector representations with power-law distributed features. Additional details on this compression scheme and its performance are given in ref 36.

The lossless nature of the power-law-based encoding scheme also provides better retrieval accuracy using similarity search methods, compared to typical lossy representations. Using six different chemical data sets consisting of inhibitors of several important pharmaceutical drug targets,[37] the recall accuracy of these molecules was evaluated using leave-one-out cross validation. For each data set, molecules were removed one at a time, and the remaining molecules were placed into a large set of random molecules. The extracted molecule was then used as a similarity search query to recall the rest against the background of the random molecules. Receiver operating characteristic (ROC) curves were produced from the results, using both lossy and lossless fingerprints (Figure 13). In each case, the lossless fingerprints outperform the lossy representations, indicated by the higher true positive rate at each false positive rate, and are quantified by an 18% and 11% increase in the ROC curve areas compared to the 512-bit and 1024-bit results, respectively. As expected, for the lossy fingerprints, the amount of compression has a negative effect on the recall performance. This is in contrast to the lossless, power-law-based compression scheme that provides higher fingerprint compression, even compared to 512-bit modulo compression, with dramatically improved recall performance.

5. STATISTICAL SIGNIFICANCE OF POWER-LAWS

While power-law distributions are typically discovered visually through log−log *rank/frequency* plots, it is also possible to assess the statistical significance that a particular data set distributes according to a power-law. Unfortunately, even with statistical tests, it is often difficult to verify or reject the hypothesis that a particular data set is actually described by a power-law and not another heavy-tailed distribution, due to a variety of confounding factors such as small data sets, boundary conditions, and noise in the data. For example, it is not uncommon that a log-normal distribution is a plausible alternative for data believed to distribute by a power-law.[12,31] Generally speaking, the assessment of statistical significance for power-laws in real-world data can be a tricky process.

Despite the above caveats, statistical tests can still provide a general indication about whether or not a power-law is a good model for a particular data set. For the feature distributions described above, the *p*-values, quantifying the hypothesis that the data are power-law distributed, were calculated using the method of Clauset et al.[31] (Table 4). For this particular statistical test, large *p*-values indicate that a power-law is a good model for the data, where a threshold of $p = 0.1$ has been suggested to separate good and bad fits.[31] As shown in Table 4, the rigid segment and cluster size ($T = 0.8$) distributions both have *p*-values larger than 0.1, indicating that the data are plausibly power-law distributed. Compared to the rigid segment distribution, the distribution of ring systems is, visually, very similar (Figure 2), though the *p*-value for the ring system distribution is much smaller, below the threshold of $p = 0.1$. This difference shows how the complex nature of real-world data and the sensitivity of the statistical tests can influence the resulting significance scores. A similar trend is also observed between the cluster size distributions, which appear similar in their *rank/frequency* plots (Figure 8) but differ in their calculated *p*-values.

In the case of the path and circular substructure distributions, the *p*-values are below the threshold, which, strictly speaking, suggest that power-laws are not good models for the data sets. For the distributions at fixed path length, a certain degree of curvature is apparent in the *rank/frequency* plots, though more linearity is seen in the *rank/frequency* plot for combined lengths $d = 1-8$ (Figure 3). For these distributions, it is likely that the data are not ideally described by pure power-laws but are potentially better approximated by a combination of power-law distributions or power-law

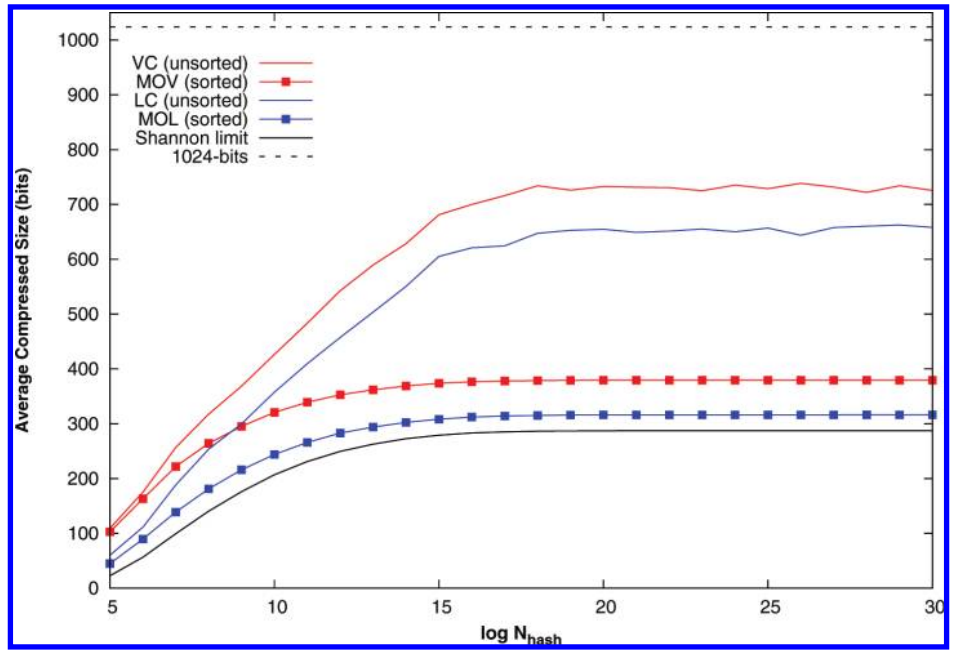

Figure 12. Averaged compressed sizes of circular substructure fingerprints encoded as indices and run lengths. Here, N_{hash} refers to the size of the vector associated with the hash function used in the compression process, with large values of N_{hash} needed for lossless compression. Efficient compression can be achieved using integer encoding, on nonsorted, index (Value Code, VC) and run length (Length Code, LC) representations. However, even greater efficiency can be gained by first sorting the fingerprint bits in decreasing frequency order dictated by the corresponding power-laws (MOnotone Value code, MOV, and MOnotone Length code, MOL). Using MOL encoding, fingerprints can be compressed to just over 300 bits per molecule on average, which is very close to the approximate Shannon limit, estimated by assuming independence between the bits (correlations between bits are small and zero on average). The smaller dynamic range of the run lengths allows for higher compression compared to the index representations.

distributions with exponential cut-offs. However, regardless of whether or not the path and circular substructure distributions are perfect examples of power-laws, the power-law approximations are still useful in practice, as indicated by the applications noted above. Furthermore, as chemical databases grow to encompass a wider variety of molecules, power-laws may become more pronounced in the feature distributions, a plausible trend indicated by the increased linearity of the combined path length distribution compared to the fixed-length distributions.

6. DISCUSSION

We have shown the existence of power-laws in the distributions of several chemical features extracted from a database of small molecules. The calculated power-law exponents for the investigated distributions range from approximately 1.5 to 3, in the same range of power-law exponents previously observed in other contexts (for reviews of power-law phenomena see refs 12 and 13). In the case of the rigid segments and ring systems, linear behavior is seen throughout the frequency distributions, though deviations are present at higher frequencies. For the labeled paths, the distributions at fixed-lengths show linear behavior mainly in the tails of the log−log plots, while the combined distribution incorporating the paths of all investigated lengths displays a greater extent of linearity throughout the distribution. Similar trends are also seen for the circular substructure data. The distributions of cluster sizes also display linearity over several orders a magnitude but show discrepancies at small cluster sizes.

The power-law nature of the molecular fragment distributions helps shed light on the general composition of small molecules in current chemical databases. For the investigated fragments, the frequency distributions are highly skewed toward the tails with exponents of $\alpha = 1.71$ for the rigid segments and $\alpha = 1.63$ for the ring systems. In the case of the rigid segments, the most frequently occurring fragment represents 4.6% of the total number of fragments observed. Furthermore, the top 10% of the set of unique rigid segments accounts for more than 90% of the total rigid segment occurrences. Similar trends are also observed for the ring systems. These numbers illustrate that relatively few rigid segments are used a majority of the time in the composition of the investigated molecules. At the same time, the large numbers of fragments that occur rarely are indicative of the structural diversity also present in the small molecules.

One important issue in the study of power-law phenomena is understanding the underlying mechanism giving rise to the power-law behavior. In the case of labeled paths, the similarities seen between the paths extracted from the ChemDB and the trained random models (Figures 6 and 7) suggest that the observed power-laws in the path distributions may be due to the labeled paths acting, at least in part, like random words, for which power-law frequency distributions have been explained by Li[18] and Conrad et al.[19] However, while the Markov models used to generate paths work to a first level approximation, there are several deficiencies in these simplistic models. In addition to producing realistic labeled paths, more accurate generative models also need to produce paths of varying lengths in the proper proportions observed empirically. By introducing a space symbol into the Markov models studied here, paths of varying length can be generated. However, in such models, shorter paths will, in general, occur more frequently than longer paths. This is

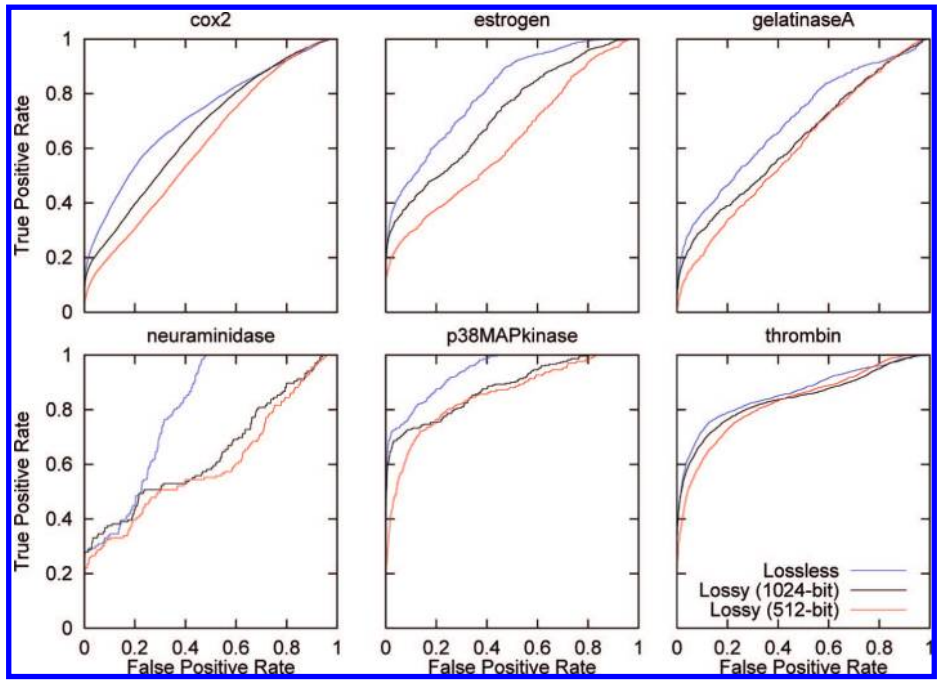

Figure 13. ROC curves illustrating the retrieval performance of druglike molecules from six different data sets against a large background test set of 50, 000 random molecules taken from the ChemDB. In each case, the lossless fingerprints outperform the lossy compressed fingerprints, highlighting the importance of lossless representations for similarity search applications.

different from what is observed for real labeled paths as branching and ring fusion cause molecules to contain a greater number of longer length paths (up to a certain length, dependent upon the size of the molecule) compared to shorter ones. Another inconsistency found in the Markov models is that they tend to produce a more diverse set of paths compared to a similarly sized set of labeled paths from actual molecules, particularly at smaller model orders. This effect decreases as the order of the Markov model increases, indicating the importance of long-range interactions among the path elements that are ignored by lower order models. Finally, the simple Markov models investigated here ignore potential correlations among paths. To properly address these issues, generative models at the full molecule level (i.e., random generation of realistic molecular structures) will likely be needed and may be useful in further understanding the power-law distribution observed here.

We have shown that the power-law models of chemical features have multiple applications, such as predicting the growth of molecular feature libraries and helping to guide experimental and computational studies. Notably, we have leveraged the power-laws to develop new, efficient, lossless compression schemes for chemical fingerprints based upon entropy coding techniques. Given that many chemical database systems currently use lossy compression methods to store molecular representations, lossless compression can help to provide more accurate chemical similarity measures, thereby improving chemical search and retrieval. Further study of the power-law distributions and other statistical properties of small molecules will likely lead to improved chemoinformatics methods as well as to help shed new light on the nature of chemical space.

ACKNOWLEDGMENT

This work was supported by an NIH Biomedical Informatics Training grant (LM-07443-01) to R.B. and P.B., an NSF MRI grant (EIA-0321390), and an NSF grant 0513376 to P.B., by the UCI Medical Scientist Training Program, and by a Harvey Fellowship to S.J.S. We would also like to acknowledge the OpenBabel project, OpenEye Scientific Software, and ChemAxon for their free software academic licenses.

REFERENCES AND NOTES

(1) Baldi, P.; Frasconi, P.; Smyth, P. *Modeling the Internet and the Web*; John Wiley and Sons Ltd.: Chichester, England, 2003.
(2) Pareto, V. *Cours D'Economie Politique*; Geneva, 1896.
(3) Estoup, J. B. *Gammes Stenographiques*; Institut Stenographique: Paris, 1916.
(4) Zipf, G. *Selective Studies and the Priciple of Relative Frequency in Language*; Harvard University Press: Cambridge, MA, 1932.
(5) Zipf, G. *The Psycho-Biology of Language: An Introduction to Dynamic Philology*; Houghton Mifflin Company: Boston, MA, 1935.
(6) Zipf, G. *Human Behavior and the Principle of Least Effort*; Addison-Wesley: Cambridge, MA, 1949.
(7) Mandelbrot, B. B. *The Fractal Geometry of Nature*; W. H. Freeman: New York, NY, 1983.
(8) Huberman, B. A.; Adamic, L. A. Internet - Growth dynamics of the World-Wide Web. *Nature* **1999**, *401*, 131–131.
(9) Barabasi, A. L.; Albert, R. Emergence of scaling in random networks. *Science* **1999**, *286*, 509–512.
(10) Kleinberg, J.; Kumar, R.; Raghavan, P.; Rajagopalan, S.; Tomkins, A. *The Web as a Graph: Measurements, Models, and Methods. In Lecture Notes in Computer Science: Computing and Combinatorics*; Asano, T., Imai, H., Lee, D., Nakano, S., Tokuyama, T. Eds.; Springer: Berlin, Germany, 1999; Vol. 1627, pp 1−17.
(11) Broder, A.; Kumar, R.; Maghoul, F.; Raghavan, P.; Rajagopalan, S.; Stata, R.; Tomkins, A.; Wiener, J. Graph structure in the Web. *Computer Networks-The Int. J. Comput. Telecommun. Networking* **2000**, *33*, 309–320.
(12) Mitzenmacher, M. A Brief History of Generative Models for Power Law and Lognormal Distributions. *Internet Math.* **2003**, *1*, 226–251.
(13) Newman, M. E. J. Power laws, Pareto distributions and Zipf's law. *Contemp. Phys.* **2005**, *46*, 323–351.
(14) Marris, E. American Chemical Society: Chemical reaction. *Nature* **2005**, *437*, 807–809.
(15) Marris, E. Chemistry society goes head to head with NIH in fight over public database. *Nature* **2005**, *435*, 718–719.
(16) Kaiser, J. SCIENCE RESOURCES: Chemists Want NIH to Curtail Database. *Science* **2005**, *308*, 774a.

(17) Chen, J.; Swamidass, S. J.; Bruand, J.; Baldi, P. ChemDB: a public database of small molecules and related chemoinformatics resources. *Bioinformatics* **2005**, *21*, 4133–4139.
(18) Li, W. T. Random Texts Exhibit Zipf's-Law-Like Word Frequency Distribution. *IEEE Trans. Inf. Theory* **1992**, *38*, 1842–1845.
(19) Conrad, B.; Mitzenmacher, M. Power laws for monkeys typing randomly: the case of unequal probabilities. *IEEE Trans. Inf. Theory* **2004**, *50*, 1403–1414.
(20) Heaps, H. S. *Information Retrieval: Computational and Theoretical Aspects*; Academic Press: New York, NY, 1978.
(21) Araujo, M.; Navarro, G.; Ziviani, N. *Large Text Searching Allowing Errors. In Proceedings of the 4th South American Workshop on String Processing*; Baeza-Yates, R. Ed.; Carleton University Press: Valparaiso, Chile, 1997.
(22) de Moura, E.; Navarro, G.; Ziviani, N. *Indexing compressed text. In Proceedings of the Fourth South American Workshop on String Processing, Carleton University Press International Informatics Series, v.8*; Baeza-Yates, R.,. Ed.; 1997.
(23) van Leijenhorst, D. C.; van der Weide, T. P. A formal derivation of Heaps' Law. *Inf. Sci.* **2005**, *170*, 263–272.
(24) Guha, R.; Howard, M. T.; Hutchison, G. R.; Murray-Rust, P.; Rzepa, H.; Steinbeck, C.; Wegner, J.; Willighagen, E. L. The Blue Obelisk-Interoperability in Chemical Informatics. *J. Chem. Inf. Model.* **2006**, *46*, 991–998.
(25) Swamidass, S. J.; Chen, J.; Phung, P.; Ralaivola, L.; Baldi, P. Kernels for small molecules and the prediction of mutagenicity, toxicity and anti-cancer activity. *Bioinformatics* **2005**, *21*, I359–I368.
(26) Dubois, J. E. *Chemical Applications of Graph Theory*; Academic Press: London, England, 1976.
(27) Hassan, M.; Brown, R. D.; Varma-O'Brien, S.; Rogers, D. Cheminformatics analysis and learning in a data pipelining environment. *Mol. Diversity* **2006**, *V10*, 283–299.
(28) Swamidass, S. J.; Baldi, P. Mathematical Correction for Fingerprint Similarity Measures to Improve Chemical Retrieval. *J. Chem. Inf. Model.* **2007**, *47*, 952–964.
(29) Swamidass, S. J.; Baldi, P. Bounds and Algorithms for Fast Exact Searches of Chemical Fingerprints in Linear and Sublinear Time. *J. Chem. Inf. Model.* **2007**, *47*, 302–317.
(30) Goldstein, M. L.; Morris, S. A.; Yen, G. G. Problems with fitting to the power-law distribution. *Eur. Phys. J. B* **2004**, *41*, 255–258.
(31) Clauset, A.; Shalizi, C. R.; Newman, M. E. J. Power-law distributions in empirical data; 2007; arXiv:physics/0706.1062. arXiv.org ePrint archive. http://arxiv.org/abs/0706.1062 (accessed Dec 3, 2007).
(32) Allen, F. H. The Cambridge Structural Database: a quarter of a million crystal structures and rising. *Acta Crystallogr., Sect. B: Struct. Sci.* **2002**, *58*, 380–388.
(33) Elias, P. Universal Codeword Sets and Representations of Integers. *IEEE Trans. Inf. Theory* **1975**, *IT21*, 194–203.
(34) Golomb, S. Run-length Encodings. *IEEE Trans. Inf. Theory* **1965**, *12*, 399–401.
(35) Witten, I.; Moffat, A.; Cell, T. B. *Managing Gigabytes: Compressing and Indexing Documents and Images*; 2nd ed.; Morgan Kauffman: Burlington, MA, 1999.
(36) Baldi, P.; Benz, R. W.; Hirschberg, D. S.; Swamidass, S. J. Lossless Compression of Chemical Fingerprints Using Integer Entropy Codes Improves Storage and Retrieval. *J. Chem. Inf. Model.* **2007**, *47*, 2098–2109.
(37) Stahl, M.; Rarey, M. Detailed Analysis of Scoring Functions for Virtual Screening. *J. Med. Chem.* **2001**, *44*, 1035–1042.

CI700353M