On the Use of Information Theory for Assessing Molecular Diversity

Dimitris K. Agrafiotis[†]

3-Dimensional Pharmaceuticals, Inc., 665 Stockton Drive, Suite 104, Exton, Pennsylvania 19341

In a recent article published in *Molecules*, Lin presented a novel approach for assessing molecular diversity based on Shannon's information theory. In this method, a set of compounds is viewed as a static collection of microstates which can register information about their environment at some predetermined capacity. Diversity is directly related to the information conveyed by the population, as quantified by Shannon's classical entropy equation. Despite its intellectual appeal, this method is characterized by a strong tendency to oversample remote areas of the feature space and produce unbalanced designs. This paper demonstrates this limitation with some simple examples and provides a rationale for the failure of the method to produce results that are consistent with other traditional methodologies.

INTRODUCTION

In a recent article published in *Molecules*,[1] Lin proposed a new method for assessing molecular diversity based on the principles of information theory, as it was first formalized by Shannon.[2] In Lin's method, a collection of compounds is viewed as a static molecular assemblage or collection of microstates which can register information about their environment at some predetermined capacity. Molecular diversity is directly related to the information content of the collection, *I*, as given by eq 1

$$I = S_{max} - S \qquad (1)$$

In this equation, S is the "entropy" of the system given by the von Neumann−Shannon expression

$$S = -\sum_{i=1}^{n} p_i \ln p_i \qquad (2)$$

where n is the total number of microstates in the system, and p_i is the probability or frequency of the ith microstate, subject to the constraint:

$$\sum_{i=1}^{n} p_i = 1 \qquad (3)$$

The central concept in Lin's approach is that each compound collection represents a finite number of distinguishable molecular species. In this case, the entropy of the system is given by eq 4

$$S(m,n) = -\sum_{j=1}^{n}\sum_{i=1}^{m} p_{ij} \ln p_{ij} \qquad (4)$$

where m is the number of species, n is the number of individuals in the population, and p_{ij} is the probability of finding the ith individual in the jth species. Again, the probabilities p_{ij} must satisfy the constraint

$$\sum_{i=1}^{m} p_{ij} = 1 \qquad (5)$$

In this case, the maximum entropy of eq 1 is given by eq 6

$$S_{max}(m,n) = n \ln m \qquad (6)$$

As evident from eq 6, the information content of the collection increases as the number of species that are represented by that population increases. The difficulty with this approach stems from the fact that in a typical application m is usually unknown. In this case, each member of the population can be treated as a unique, distinguishable species, and the entropy of the system can be related to the *distinguishability* of these species, rather than their similarity to some *a priori* known set of prototypes. Under this formalism, eq 4 is replaced by eq 7

$$S(n,n) = -\sum_{j=1}^{n}\sum_{i=1}^{n} p_{ij} \ln p_{ij} \qquad (7)$$

subject to the usual constraint

$$\sum_{i=1}^{n} p_{ij} = 1 \qquad (8)$$

and S_{max} becomes

$$S_{max}(n,n) = n \ln n \qquad (9)$$

In eq 7, the values of the p_{ij}'s can be computed directly from a molecular similarity table. Several methods for quantifying molecular similarity have appeared in the literature,[3,4] most of which assign scores in the interval [0, 1]. The method proposed by Lin involves computing the similarity table, ρ_{ij}, using any established method, and applying a normalization factor to derive the actual probabilities. This factor is given by eq 10

$$c = \frac{1}{\sum_{i=1}^{n} \rho_{ij}} \qquad (10)$$

† Tel: (610) 458-6045, Fax: (610) 458-8249, E-mail: dimitris@3dp.com.

Figure 1.

and the actual probabilities by eq 11

$$p_{ij} = c\rho_{ij} \tag{11}$$

Thus, according to Lin, as the probability values draw closer, the species become more indistinguishable, the entropy of the system increases, and the information (and therefore diversity) registered by the population decreases.

DISCUSSION

While the use of information theory to quantify molecular diversity has certain intellectual appeal, the actual implementation has some important limitations. We will first demonstrate these limitations using some simple examples and then discuss the reasons in the context of the underlying theory. We must point out that the following discussion is based on Lin's definitions of entropy and "information"; unless noted otherwise. As we argue toward the end of this paper, it is the exact mathematical definition of these concepts that is responsible for the failures described below.

Consider the following example. Figure 1 shows two different sets of three imaginary compounds plotted against a uniform property scale (in Figure 1b, points 2 and 3 coincide). As usual, the distance between two points, d_{ij}, is taken as a direct measure of their similarity, ρ_{ij}. Two different popular methods are used to perform this transformation

$$\rho_{ij} = \alpha - d_{ij} \tag{12}$$

and

$$\rho_{ij} = \frac{1}{1 + \alpha d_{ij}} \tag{13}$$

where α is a constant (in this case 1.0). Using eq 7, we can compute the following "entropies" for each of these methods and each of the sets (the reader can easily confirm these results)

$$S_a = 2.313$$

and

$$S_b = 1.386$$

using the linear form (eq 12) and

$$S_a = 3.194$$

and

$$S_b = 3.149$$

using the reciprocal form (eq 13), where a and b refer to the sets on the left- and right-hand side of Figure 1, respectively. According to this and eq 1, one would conclude that the diversity of set a is lower than that of set b and that this is

true regardless of the functional form used to derive the similarities. For all practical purposes, this result is suspicious.

Let us take this argument a step further. Figure 2 shows the entropy of the same three-point set shown in Figure 1 as a function of the position of one of these points relative to the other two. The two extremes represent the cases where the third point coincides with one of the two reference points (Figure 1b), while the middle represents the situation depicted in Figure 1a. This profile is based on the reciprocal function shown in eq 13, but similar results are obtained with eq 12 as well. As we can see, the entropy function is at a maximum (and therefore the diversity at a minimum) when the third point is located halfway between the two reference points and monotonically decreases as the point moves away from the center in either direction. This is, again, in sharp contrast to any intuitive interpretation and raises concerns regarding the properties of this metric and its implications in the study of molecular diversity.

The final part of our analysis illustrates the results of Lin's approach when applied to a classical problem in combinatorial library design.[5] The problem can be stated as follows: given an n-membered virtual library and a number k, find the k most diverse compounds in that population. We recently presented a general solution to this problem based on a novel distance-based diversity metric and a stochastic search engine such as simulated annealing or evolutionary programming.[6,7] The power of this method stems from its generality, *i.e.*, that fact that the selection can be carried out based on any desirable selection criterion. In the problem at hand, the selection criterion was to minimize the entropy of the system as measured by eq 7, subject to the constraint that the number of compounds in the final solution should be exactly k. The data set used in the experiment consisted of two- and three-dimensional random vectors, uniformly distributed in the unit square (cube). This property space is densely populated, does not exhibit any significant clustering, and is designed to reveal the structure of the "true" minimum. Euclidean distances were converted to similarities using eq 13. For clarity, our selection was limited to 25 compounds and was performed using simulated annealing. In particular, the simulation was carried out in 30 temperature cycles, using 1000 sampling steps per cycle, a Gaussian cooling schedule, and the Metropolis acceptance criterion ($p = e^{-\Delta E/K_B T}$). Transitions represented single-point mutations in the composition of the current set. Boltzmann's constant, K_B, was adjusted in an adaptive manner, by constantly updating the mean transition energy during the course of the simulation and continuously adjusting the value of K_B so that the acceptance probability for a mean uphill transition at the final temperature was 1%. Details of this algorithm can be found elsewhere.[6,7]

For comparison, Figures 3−5 show the results of three different methods for maximizing diversity. The first is based on the metric described in 1 and 2 (Figure 3), the second on the popular maximin method (Figure 4), and the third on the Lin metric, as implemented in the manner described above (Figure 5). The results are striking. In sharp contrast to the other two methods, Lin's metric does exactly the opposite of what it was designed to do! The points are divided equally among two very tight clusters located at maximal separation along the diagonal. As illustrated in Figure 6, this is also true for the three-dimensional case, and,
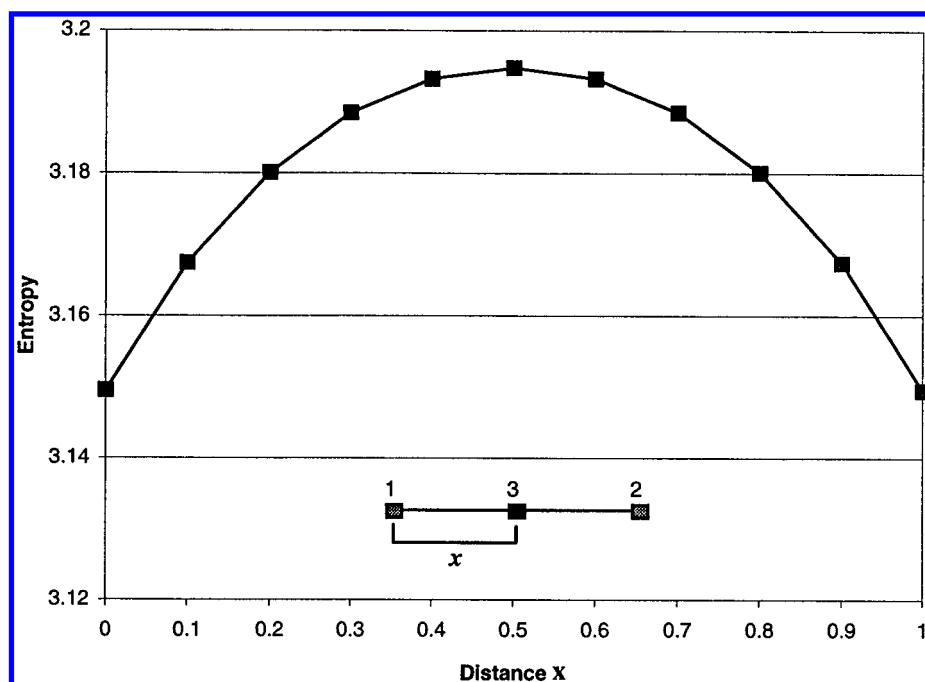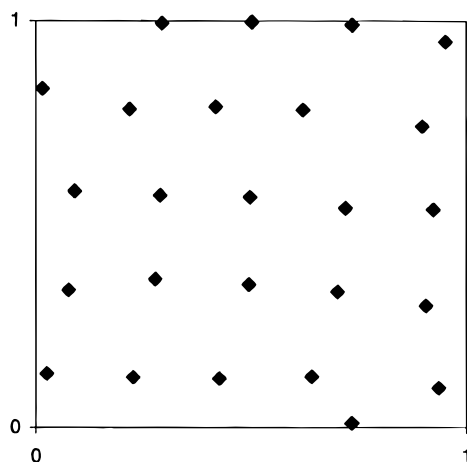
Figure 2. Entropy as a function of *x*.

Figure 3. 2D selection based on Agrafiotis' diversity metric and selection algorithm.

Figure 5. 2D selection based on Lin's diversity metric and Agrafiotis' selection algorithm.
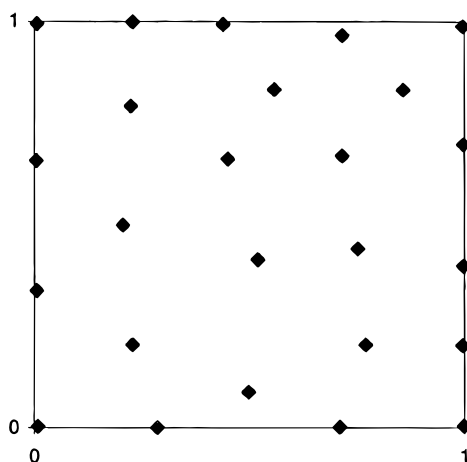

Figure 4. 2D selection based on the maximim algorithm.

in fact, it is the same kind of behavior depicted in Figure 1. Experiments with different sample sizes and density distributions confirm these findings and indicate that this metric has a strong and general tendency for clustering the samples in
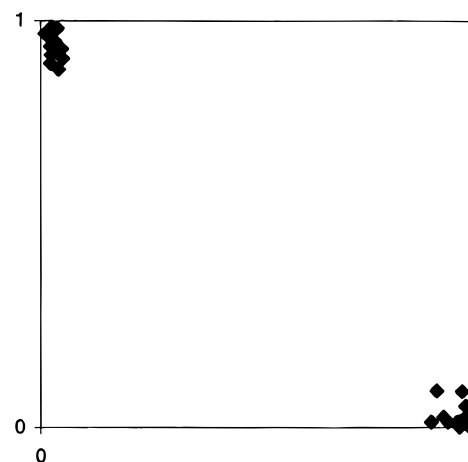
remote areas of the feature space and does an extremely poor job in sampling that property space in an unbiased and uniform way.

The reasons for this behavior can be traced back to eqs 1 and 2, and the definitions of entropy and information. As Lin correctly acknowleges in his paper, the pursuit of molecular diversity aims at maximizing our knowledge or information about the system under investigation. Clearly, we are interested in compound collections that render *more* information about the underlying system or process, not *less*. However, there are many different types of information, and the precise meaning depends on the context in which the term is used. In Shannon's theory,[2] information is used to measure the uncertainty associated with an event that is transmitted across a communication channel. The channel consists of a source that generates signals from some predefined set of possible signals and a receiver that receives these signals with some degree of uncertainty that depends on the characteristics of the source and the communication line. Shannon's own words are as follows: "The fundamental problem in communication is that of reproducing at
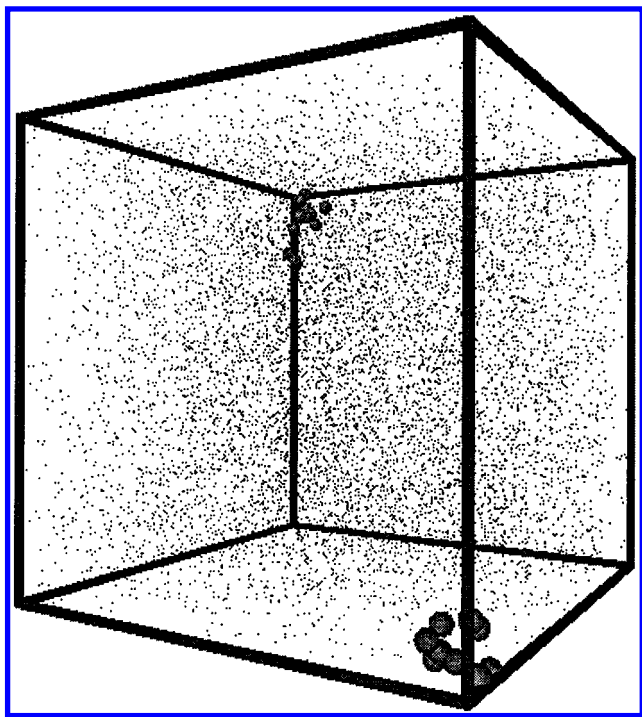
Figure 6. 3D selection based on Lin's diversity metric and Agrafiotis' selection algorithm (dots: uniform data set; spheres: selected points).

one point either exactly or approximately a message selected at another point. Frequently, the messages have meaning; that is, they refer to or are correlated according to some system with certain physical or conceptual entities. These semantic aspects are irrelevant to the engineering problem. The significant aspect is that the actual message is one selected from a set of possible messages.".[2]

Consider, for example, a device (source) that can produce three possible symbols, A, B, or C. As we wait at the receiving end, we are uncertain as to which symbol will appear next. Once the signal is received, our uncertainty decreases, and we remark that we have gained some information. Thus, information (also referred to as "information content" or "information uncertainty") corresponds to a decrease in *uncertainty*. Shannon's entropy equation measures this uncertainty and represents the average *surprisal* for an infinite string of events produced by the source.

The extension of these concepts to the study of molecular systems is straightforward. Indeed, every collection of compounds can be viewed as a string of independent events produced by an imaginary source (selection process) and related to each other by virtue of a similarity metric. Shannon's entropy expression can be used to measure our ability to predict the events (i.e., the structures of the compounds) that are generated by that source. Clearly, if we know that the collection contains very similar compounds, we can predict the overall structure of each compound with sufficient accuracy once we have seen the first few. If, on the other hand, the collection contains diverse compounds, the element of surprise increases, and our ability to predict their structures diminishes. Thus, increase in diversity results in an *increase* in entropy, not a decrease as Lin suggests.

So what is the type of information that one seeks to maximize in a diversity design? Weaver[2] claims that "...the word 'information' in communication theory relates not so much to what you do say, but to what you could say".

Communications engineers are more interested in the *capacity* to store information and have come to associate the term information with potential message variety, freedom of choice, and a large vocabulary. Gatlin,[8] in her book on biological information theory, calls this type of information "potential information" to distinguish it from "stored information", or the "what we do say" in Weaver's statement. Since it varies directly with entropy, high entropy leads to high potential information.

But there are some mathematical errors as well. For example, to satisfy the normality condition in eq 7, the summation in eq 8 (eq 16 in Lin's original paper) should be over both *i* and *j* (the reader may consult any book on information theory). This improper normalization may have serious numerical consequences. Moreover, the coefficient "*c*" in eq 10 (eq 30 in the original paper) should be index-dependent. As pointed out by Maggiora,[9] joint probabilities are not necessarily symmetric, while most similarity indices are. Equation 11 violates this, implying that similarity and probability are both symmetric and are linearly related by a single scalar parameter, which is simply not true.

Would correcting these errors and reversing the sign of entropy lead to the "correct" (i.e., desired) response? The answer is no, and the reason can be found again in eqs 7, 10 and 11. Entropy is maximized when the probability values become identical, and this occurs when all the compounds in the collection become equidistant. This is rarely the case and occurs only for points located at the corners of a triangle, tetrahedron, or a higher order simplex. Actually, this is another interesting consequence of this approach, that is, the fact that based on indistinguishability alone, one cannot distinguish between two different, say, tetrahedra since the diversity (entropy) is the same regardless of the size of the tetrahedron. If one is interested in a uniform sampling of the property space, one needs to consider only the immediate neighborhoods of each compound. This is exactly what maximim and our own algorithm does, and this is why they produce results that are consistent with our expectations and intuition. Although one may argue that some of the properties of Lin's metric could be useful in a different context,[10] it is clear that it cannot produce the kinds of uniform distributions that one typically requires in diversity designs.

One final remark: Lin's concepts of indistinguishability and entropy bear a striking resemblance to similar concepts used in fuzzy clustering.[11] Indeed, if one considers each member of the collection as a singleton cluster, the probabilities in eq 7 become fuzzy cluster membership values, and the entropy expression is identical to the fuzzy entropy of the system. In the context of this discussion, a critical comparison of two different clustering methodologies (fuzzy[11] and possibilistic[12–14]) would be most constructive.

CONCLUSIONS

This paper describes our experience with the use of a new approach for assessing molecular diversity based on the theory of information. As we demonstrate with some simple examples, this method has a strong tendency to oversample remote areas of the feature space and produce unbalanced designs. We believe that this is due to the use of a certain type of information whose mathematical definition is inappropriate for the study of molecular diversity.

REFERENCES AND NOTES

(1) Lin, S. K. Molecular diversity assessment: logarithmic relations of information and species diversity and logarithmic relations of entropy and indistinguishability after rejection of Gibbs paradox of entropy mixing. *Molecules* **1996**, *1*, 57−67.

(2) Shannon, C. E.; Weaver, W. *The mathematical theory of communication*; University of Illinois Press: Urbana, IL, 1949.

(3) *Concepts and applications of molecular similarity*; Johnson, M. A., Maggiora, G. M., Eds.; Wiley-Interscience: New York, 1990.

(4) Kubinyi, H. *QSAR: Hansch analysis and related approaches*; VCH: Weinheim, 1993; p 172.

(5) Agrafiotis, D. K.; Bone, R. F.; Salemme, F. R.; Soll, R. M. System and method of automatically generating chemical compounds with desired properties. United States Patent 5,463,564, 1995.

(6) Agrafiotis, D. K. Stochastic algorithms for maximizing molecular diversity. *Theochem*. **1997**, in press.

(7) Agrafiotis, D. K.; Jaeger, E. P. New metrics and algorithms for assessing molecular diversity. Manuscript in preparation.

(8) Gatlin, L. *Information theory and the living system*; Columbia University Press: New York, 1972.

(9) Maggiora, G. M. Private communication.

(10) For example, one of the referees suggested that Lin's entropy measure may be useful in the analysis of highly clustered data sets.

(11) Bezdek, J. C. *Pattern recognition with fuzzy objective function algorithms*; Plenum: New York, 1981.

(12) Krishnapuram, R.; Keller, J. M. A possibilistic approach to clustering. *IEEE Trans. Fuzzy Syst*. **1993**, *1*, 98−110.

(13) Barni, M.; Cappellini, V.; Mecocci, A. Comments on a possibilistic approach to clustering. *IEEE Trans. Fuzzy Syst*. **1996**, *4*, 393−396.

(14) Krishnapuram, R.; Keller, J. M. The possibilistic c-means algorithm: insights and recommendations. *IEEE Trans. Fuzzy Syst*. **1996**, *4*, 385−393.

CI960156B