Molecules **1996**, *1*, 57 – 67

Molecular Diversity Assessment: Logarithmic Relations of Information and Species Diversity and Logarithmic Relations of Entropy and Indistinguishability after Rejection of Gibbs Paradox of Entropy of Mixing

Shu-Kun Lin†

Molecular Diversity Preservation International (MDPI), Saengergasse 25, CH-4054 Basel, Switzerland.
Phone +41 79 3223379; Fax +41 61 3028918; (lin@ubaclu.unibas.ch)

† The first version of this paper was prepared at the Pharmaceuticals Division of Ciba-Geigy Ltd., CH-4002 Basel, Switzerland.

Received: 27 May 1996 / Accepted: 7 August 1996 / Published: 24 September 1996

Abstract

It is postulated that the degradation of species diversity results in information loss or entropy increase. To define the molecular diversity of either a single mixture of different compounds and a combinatorial compound library or a collection of pure compounds, we treat them all as molecular assemblages for information registration and consider only the molecular similarity and chemical species numbers of the individual molecules. The entropy of a molecular assemblage is correlated to the chemical species similarity via the von Neumann-Shannon relation and related to the so-defined apparent species indistinguishability number (σ_a) via a logarithmic relation. Information and the apparent species number (M_a) also have a logarithmic relation. M_a is equal to or less than the designated species number M. The diversity index (D) is defined as the ratio of the logarithms of the apparent chemical species number and the designated species number (D=lnM_a/ln$M$). D has a value between 0 and 1 and decreases with the increase in similarity among the species. The decrease in the evenness of the species abundance also results in a decrease in diversity D. Molecular diversity of a combinatorial compound library is determined by the available number of component variants and their similarity. Clearly these concepts and formulae can also be applied to calculate biodiversity.

Keywords: Molecular diversity, species similarity, entropy, information, apparent species number, apparent indistinguishability number, species equitability, diversity index

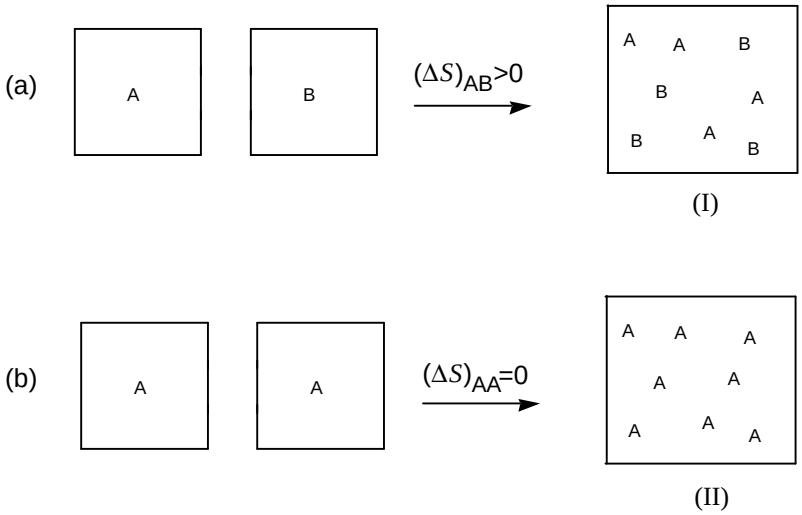


Figure 1. *Suppose compounds A and B have identical entropies when they are in pure and isolated form. The information content difference of a mixture of two different compounds in a container (I) and one containing the same number of compounds with identical properties (II) can be estimated by calculating entropy increments of the two processes (a) and (b). Gibbs paradox statement implies that (I) has a lower information content than (II), because the mixture (I) has an entropy of mixing different compounds. Theoretically [8], we do not agree with Gibbs paradox statement. Practically we observe that a mixture of several different compounds (I) is more interesting than (II), as can be immediately made clear by using (I) and (II) for tests of high throughput screening.*

Introduction

Biodiversity has been a topic of very wide concern for several years. In *The Diversity of Life*, Wilson [1] asserts that biological diversity is priceless and should be protected. There must be some fundamental law of science that stirs our resolve to preserve biodiversity. In all events, diversity appears to be a growing concern. Recently, molecular diversity and its significance in the pharmaceutical sciences have also been the subject of intensive studies [2-6].

Practically, there is a trivial standard about species diversity in chemistry that a collection of many different molecules are more interesting than the collection of the same number but otherwise very similar or indistinguishable compounds. A theoretical explanation for the validity of this standard should be an easy task. However, mathematical relations between entropy [7] and diversity, similarity and indistinguishability were never explicitly established in conventional statistical mechanics texts [8]. What is worse, an informative theoretical explanation is confronted with the famous Gibbs paradox statement of entropy of mixing [8], which says that the mixing (or assembling) of *different* compounds (which should be desirable for a compound library of high-quality molecular diversity or a mixture of several compounds for high throughput screening in drug discovery) has an entropy of mixing, while mixing (or assembling) of indistinguishable molecules has a minimal – which is zero – entropy of mixing (Figure 1). This implies that a mixture (or an assemblage) of many *different* compounds has less information content than a mixture (or an assemblage) comprising many very similar or *identical* compounds, which is ridiculous!

Generally, it appears obvious that a decrease in biodiversity or molecular diversity corresponds to a loss of information and an increase of entropy in the system concerned. To go one step further it would be very interesting to consider the possible mechanism of such an information loss and perform a quantitative calculation. Therefore, a simple theory is developed whereby the informational entropy of an assemblage can be correlated to the similarity of its component individuals. Several logarithmic relations have been set up. Firstly, it is shown that the information (I) and the apparent species number (M_a) have a logarithmic relationship. Then, it is recognized that entropy (S) and the apparent number of indistinguishable microstates (w_a) of the whole system (the assemblage), also have a logarithmic relationship. These and other logarithmic relations of entropy and information are useful: amongst other applications, they can serve to evaluate the quality of species diversity, which lends a rational appeal to the appreciation and the preservation of molecular diversity and biodiversity.

Molecular Assemblages

In order to define the molecular diversity of either a single mixture of different compounds and a combinatorial compound library or a collection of pure compounds, we treat them all as molecular assemblages for information registration at their highest possible capacity and consider only the molecular similarity and chemical species numbers of the individual molecules. Compound mixtures, combinatorial compound libraries and compound collections will all be analysed as static molecular assemblages.

In a previous paper [8], two general cases of information loss in a thermodynamic system of molecules have been discussed. Firstly, the information loss can happen due to spontaneous *dynamic motion* where there are σ interconverting chemical species (e.g., tautomeric equilibrium between enols and ketones or aldehydes, or racemization) or *w* interconverting microstates (e.g., in a fluid), which normally occurs when the temperature is increased. Even though there are different chemical species in a system in consideration, information cannot be registered if there is interconversion between these different chemical species. Secondly, the information loss can occur due to spontaneous formation of *static structures* after phase separation, such as a two-phase structure (one phase is pure A and the other phase is pure B) formed from a mixture of A and B. Different static arrangements of molecules in an assemblage may result in entropy differences [8].

It is not our intention to consider these two information loss mechanisms further here. Therefore, in the following discussion, it is supposed that there is no information loss due to dynamic motion occurring in our system. Direct information registration in a conventional way is impossible when the information registration medium becomes a fluid (a gas or a liquid). Therefore, when the molecular diversity of a mixture, such as a mixture of several samples in DMSO, is considered, the solution is frozen (so that it becomes a static structure) first to form a most heterogeneous molecular assemblage for information registration.

Furthermore, it is assumed that molecules of the separated pure compounds in a collection of compounds are all used to form a most heterogeneous molecular assemblage for registering as much information as possible. Succinictly, only the property differences of the compounds and the relative numbers used to generate the molecular assemblage should be considered.

Information and Species Number

Generally, starting from the von Neumann-Shannon entropy expression [9-11], entropy is

$$S = -\sum_{i=1}^{w} p_i \ln p_i \tag{1}$$

where p_i is the probability of the *i*th microstate [9] with the property that

$$\sum_{i=1}^{w} p_i = 1. \tag{2}$$

Equation (1) is suitable for the distinguishability analysis of one single object by comparing it with *w* designated species. Information and entropy in the whole system of *N* molecules in an assemblage will be considered in the following.

Suppose a system is composed of *N* "unit devices", which can be called the *N* individuals of the system. These individuals appear as *M* attributes, based upon which it is said that the system has *M* species. For fauna, there are *N* animals belonging to *M* species. For a binary system, as used in computer science and communication [10], there are two species: "yes" and "no", or 0 and 1. For faithful registration of information, these species must be truly *distinguishable* or different. As conventionally defined, a binary system of *N* such individuals has a maximum information content of *N* bits [10],

$$I_{\max}(2, N) = \log_2 2^N = N \log_2 2 = N\text{bits} \tag{3}$$

if the logarithmic base of 2 is taken. For a decimal system, we have ten species: 0, 1, 2, 3, 4, 5, 6, 7, 8 and 9, and the maximum information one can register is

$$I_{\max}(10, N) = \log_2 10^N = N \log_2 10 = 3.32N\text{bits} \tag{4}$$

The maximum information content is 3.32 times as much as that of a binary system. As can be seen clearly from equations (3) and (4), it is clear that *the information content increases with the increase in the number of species.* Generally for a system of *M* species, the maximum information is

$$I_{\max}(M, N) = \ln M^N = N \ln M \tag{5}$$

where ln is the natural logarithm, which will be used here.

In equation (5) *M* can be greater than *N* as can be encountered in computer science and communication. However,

$$I_{\max}(N, N) = \ln N^N = N \ln N \tag{6}$$

represents a maximum information value, provided that every individual in the system is a distinguishable species and generally $M \le N$.

As shown in Figure 2, *M*=2 in the assemblage (I), which has a higher information content than that of (II), which has zero information content (M =1, $I_{\max}=N\ln 1=0$).

Because the increase in the species number gives an information increase, it is obvious that the species number *M* is a good indication of maximum species diversity, provided that these species are completely distinguishable.

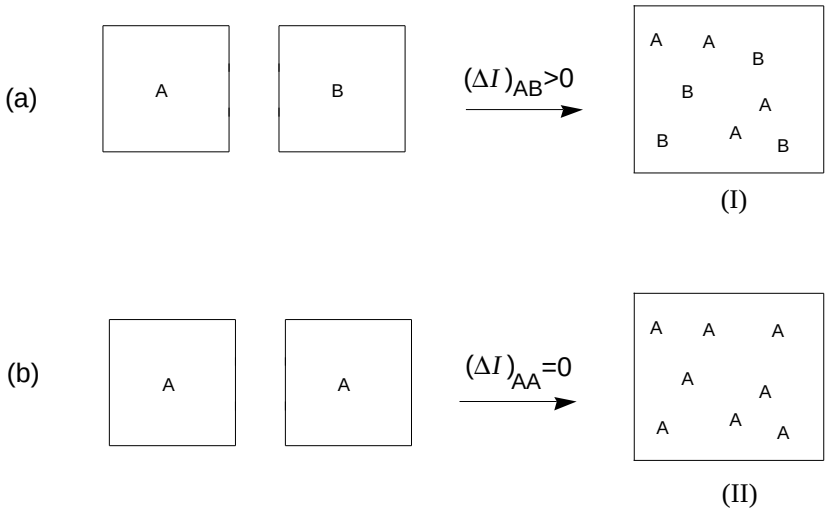


Figure 2. *In contrast to the Gibbs paradox statement (Figure 1), we observe that assembling different molecules to form a molecular assemblage (I) in process (a) is a typical information registration process and gives an information increase. Process (b) has no information increase. Therefore, (I) has a higher information content than (II). The higher molecular diversity in (I) can thus be calculated from its higher information content.*

Apparent Species Number

In order to estimate the entropy of molecular assemblages of molecules with various similarities, the entropy value of an assemblage of M or N completely distinguishable species can be set as the minimum (zero – see state (I) Figure 2):

$$S = 0 \tag{7}$$

corresponding to

$$I = I_{\max}(M, N)$$

or

$$I = I_{\max}(N, N).$$

All the entropy values in the following discussions are relative to this state (equation 7).

More generally, conventional information theory defines that the total amount of information (I) registered by a system is the difference between the system's actual entropy (S) and its maximum entropy ($S_{\max}$),

$$I = S_{\max} - S \tag{8}$$

In a process, if all the information is lost relative to the initial state (equation 5), according to equation (8) the final state has maximum entropy

$$S_{\max}(M, N) = \ln M^N = N \ln M \tag{9}$$

with the indistinguishability number defined as

$$w = M^N$$

which is the number of indistinguishable "microstates" (or, in some cases, the number of the arrangements, or combinations). Again, it is well known that the increase in entropy is synonymous with a loss of information. Clearly,

$$I_{\max} = S_{\max} \tag{10}$$

which means that the maximum information a system can lose equals the maximum information the system may register.

Because, in many cases, the species number M is unknown beforehand or it is only a designated number, it might be realistic to take as a reference an initial state of the highest possible diversity, that is, a state where every component of the system is a distinguishable species (equation 6). Then, a state of total loss of information relative to this state will have

$$S_{\max}(N, N) = \ln N^N = N \ln N \tag{11}$$

with the indistinguishability number of microstates $w = N^N$.

The degradation of diversity means information loss. How is information lost? One most obvious mechanism underlying the degeneration of information is an increase in similarity among the species [8]: *The greater the species similarity, the greater the entropy of an assemblage.* In other words, the greater the similarity, the less becomes the information content of the assemblage.

Similar to equation (1), the entropy for a system of N individuals is

$$S(M,N) = -\sum_{j=1}^{N}\sum_{i=1}^{M}(p_{ij}\ln p_{ij}) \tag{12}$$

where p_{ij} is the probability of finding the *j*th individual of the system as the *i*th species, with the property that

$$\sum_{i=1}^{M} p_{ij} = 1 \tag{13}$$

Similarly, the apparent indistinguishability number of microstates of the whole system is defined as

$$w_a = \exp\left[-\sum_{j=1}^{N}\sum_{i=1}^{M}(p_{ij}\ln p_{ij})\right] \tag{14}$$

and has the property that $w_a \le w$. Or,

$$S(N,N) = -\sum_{j=1}^{N}\sum_{i=1}^{N}(p_{ij}\ln p_{ij}) \tag{15}$$

where p_{ij} is still the probability of finding the *j*th individual of the system as the *i*th species, with the property that

$$\sum_{i=1}^{N} p_{ij} = 1 \tag{16}$$

and,

$$w_a = \exp\left[-\sum_{j=1}^{N}\sum_{i=1}^{N}(p_{ij}\ln p_{ij})\right] \tag{17}$$

There may be many different methods of specifying the p_{ij} values. One obvious possibility in biodiversity assessment, for example, is the comparison of ribosomal RNA sequences, the method frequently used in phylogenetic classification. If comparison indicates that the *M* species are becoming more similar, it will be more than likely that all the p_{ij} values are drawing closer. Several methods of molecular similarity calculation (all have values between 0 and 1) are available [12-14]. If any parameter is given to quantitatively specify the similarity among the species, in principle one can always perform entropy calculation by using equations (12) or (15). The probabilities p_{ij} can always be clearly specified under the restriction imposed by equation (13) or equation (16).

The entropy contents for two extreme cases are obvious: If there are still *M* designated species and these are still distinguishable, it is certain that the *j*th individual of the system is the *i*th species, p_{ij} =1. From equations (8) and (9), the entropy is 0. It is understood that $1ln1 = 0$ and $0ln0 = 0$. Therefore, the information given in equation (5) is conserved. However, *suppose* that there are still *M* species existing in the system (so that one can evaluate the loss of information), but the differences between the individuals are extremely small, or, in other words, they are virtually indistinguishable, all the probabilities will be identical: $p_{ij} = 1/M$; and equation (12) is reduced to equation (9), which is the maximum entropy, a state corresponding to the total loss of the original information.

If species are more similar, p_{ij} values are closer to each other. Clearly the similarity between species determines probability values p_{ij}. Furthermore, one can define

$$Z = \frac{S}{S_{max}} \tag{18}$$

as the general similarity within the system. By inspection, *Z* is within the range 0 to 1. I define similarity index as connected with informational entropy. All the other useful definitions of molecular similarity index have values between 0 and 1 [12-14]. The maximum similarity (*Z* = 1) corresponds to the maximum indistinguishability number. It is worthwhile to emphasizing that the most striking feature of the maximum-entropy state in an assemblage is that all the individual molecules or species are extremely similar (or identical). *The state of maximal entropy is the state of maximal similarity (or indistinguishability).* From equations (8) and (10), the relation of information loss and the similarity can be more explicitly expressed as

$$Z = \frac{I_{max} - I}{I_{max}} \tag{19}$$

Moreover, σ_a is defined as apparent species indistinguishability number among the *M* species,

$$\sigma_a = \exp\left[-\frac{1}{N}\sum_{j=1}^{N}\sum_{i=1}^{M}(p_{ij}\ln p_{ij})\right] \tag{20}$$

If all the supposed *M* species are factually indistinguishable, $\sigma_a = M$. Because *M* is usually unknown, the apparent species indistinguishability number among all the *N* individuals can be calculated by

$$\sigma_a = \exp\left[-\frac{1}{N}\sum_{j=1}^{N}\sum_{i=1}^{N}(p_{ij}\ln p_{ij})\right] \tag{21}$$

Then, if all the N individuals are indistinguishable, $\sigma_a = N$. Obviously, if the apparent-species-indistinguishability number is calculated by equation (21), the similarity formula becomes

$$Z = \frac{\ln \sigma_a}{\ln N} \tag{22}$$

Furthermore, M_a can be defined as the apparent species number. Then, information is

$$I(M, N) = N \ln M_a \tag{23}$$

From equation (8), $N\ln M_a = N\ln M - S(M,N)$. This gives

$$M_a = \exp\left(\ln M + \frac{1}{N}\sum_{j=1}^{N}\sum_{i=1}^{M} p_{ij}\ln p_{ij}\right) = \frac{M}{\sigma_a} \tag{24}$$

Equation (24) clearly indicates that the increase in the apparent species indistinguishability number (σ_a) is equivalent to the reduction in apparent species number (M_a).

Again, if M is unknown, it will be convenient to set $M = N$ as the designated species number. Similarly, one can also calculate M_a from

$$M_a = \exp\left(\ln N + \frac{1}{N}\sum_{j=1}^{N}\sum_{i=1}^{N} p_{ij}\ln p_{ij}\right) = \frac{N}{\sigma_a} \tag{25}$$

where σ_a is calculated from equation (21).

Obviously,

$$\sigma_a \geq 1 \tag{26}$$

and

$$M_a \geq 1 \tag{27}$$

Molecular Diversity Index

Now the molecular diversity can be defined as a diversity index (D):

$$D = \frac{I}{I_{max}} = \frac{\ln M_a}{\ln M} \tag{28}$$

which expresses clearly that information loss is the degradation of diversity. It is a scale of information content related to the state composing completely distinguishable species. The parameter D, as a measure of diversity, has a value limited between 0 and 1. The diversity index value 1 corresponds to the maximum molecular diversity as designated for a molecular assemblage.

It was pointed out earlier that all the calculations can be reduced finally to the similarity analysis. If equation (12) is used, the pairwise comparison to calculate p_{ij} must be carried out between the jth individual with the previously known M specimen. However, if equation (15) is used, the N individuals of the system concerned are directly taken as the N specimen. A general and expedient procedure for calculating these parameters by using equations (15) and (16) may start from the calculations of p_{ij}, which in turn, may be calculated from the pairwise similarity ρ_{ij}. The values of ρ_{ij}, limited between 0 and 1, must be given by direct comparison according to one and only one systematically followed standard of comparison for all the values p_{ij} ($i = 1,\cdots,N;\ j = 1,\cdots,N$). Then a normalization factor c can be calculated:

$$c\sum_{i=1}^{N}\rho_{ij} = 1 \tag{29}$$

or

$$c = \frac{1}{\sum_{i=1}^{N}\rho_{ij}} \tag{30}$$

It follows that

$$p_{ij} = c\rho_{ij} \tag{31}$$

As an example of biodiversity, if a fruit fly of the same species has N=1000 and its species diversity is estimated by using equation (15). Clearly all the values of ρ_{ij} are 1. The normalization factor is $c = 1/1000 = 0.001$. $p_{ij} = 1/1000 = 0.001$. The result will be that $S(1000,1000) \approx 6908$ nats (nat is the natural unit of information or entropy [8]); the apparent species indistinguish-

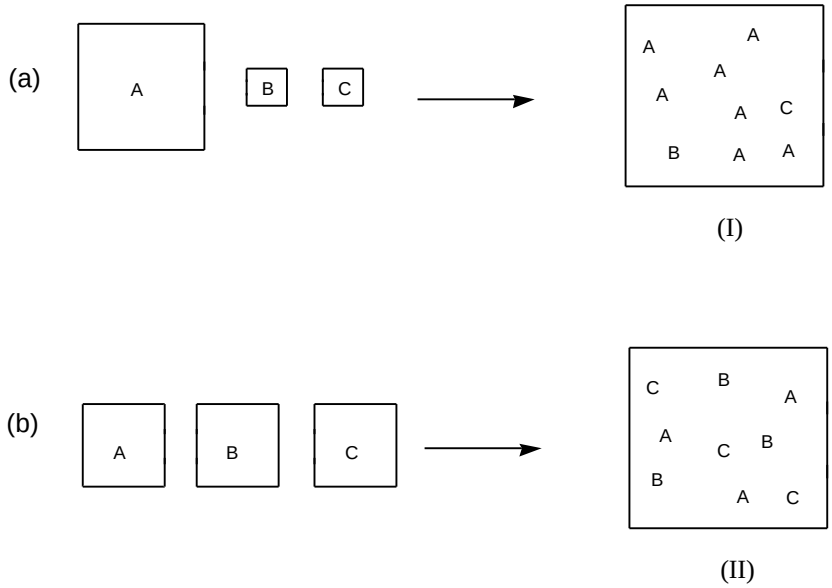


Figure 3. *Schematic representation of the low evenness (or equitability [15]) in (I) and the highest possible evenness (or equitability) in (II) for a mixture of three species A, B, and C.*

ability number is $\sigma_a = 1000$ and the apparent species number becomes $M_a = 1$, exactly as expected from equation (25). Because all the individuals are indistinguishable, the general similarity (*Z*) of the system is 1 according to equation (18) and the diversity index (*D*) is zero according to equation (28).

If a system of two designated species has *N*=1000 and its species diversity is also estimated by using equation (15), the values of ρ_{ij} are either 1 or 0. Suppose there is an equal number of individuals for both species in the system (*vide infra* for the case of non-equal numbers of individuals). Then, the normalization factor will be $c = 1/500 = 0.002$ and p_{ij} will be either 0.002 or 0 so that equation (16) is satisfied. The result will be *S*(1000,1000) = 1000 ln 500 ≈ 6214 nats according to equation (15). The apparent species indistinguishability number is $\sigma_a = 500$ according to equation (21). The apparent species number becomes $M_a = 2$ according to equation (25), exactly as expected. The general similarity $Z \approx 6214/6908 \approx 0.90$, less than 1, but rather large, indicating that a large number of the individuals are indistinguishable (among the 500 individuals for one species and the other 500 individuals for another species). Of course, if equation (12) is applied to this system and *N* = 1000 and *M* = 2, a calculation will finally lead to $Z = 0$ and *D* = 1, showing that the two designated species are truly distinctly different.

For other cases where values of ρ_{ij} are neither 1 nor 0, the calculation can be carried out in a similar way. For a system with a huge number of individuals, sampling of a limited number of individuals from the considered system may be feasible and can give an approximate estimation of the apparent species number M_a and the diversity *D*.

The expressions of M_a and *D* can be directly used to estimate the molecular diversity of a molecular assemblage, such as a compound collection, where a direct similarity comparison of all the composing individual compounds is desirable.

Evenness of Species Abundance

A concept of the evenness of the abundance of species has been used by biologists [15]. Given a fauna of butterflies consisting of 1 million individuals divided into 100 distinguishable species ($M = 100$) (Figure 3b). If all species are equally abundant, and each species has 10000 individuals, from equation (15),

$$\begin{aligned} S\left(10^6,10^6\right) &= -\sum_{j=1}^{10^6}\sum_{i=1}^{10^6}(p_{ij}\ln p_{ij}) \\ &= -\sum_{j=1}^{10^6}\ln\frac{1}{10000} = 10^6\ln 10000 \\ &\approx 9210340\text{nats} \end{aligned} \tag{32}$$

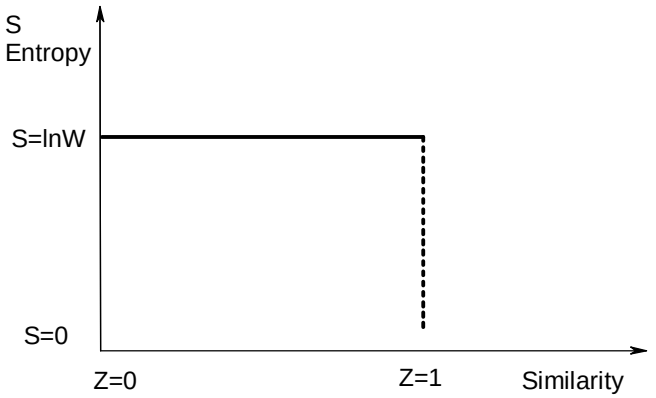
Therefore, $\sigma_a = 10000$ and $M_a = M = 100$, as expected [16], and *D* = 1. This is a fauna of the highest possible evenness (or equitability [15]).

However, if one species is extremely abundant, and has 990100 individuals, and each of the other 99 species comprises only 100 individuals (Figure 3a). Applying equation (15),

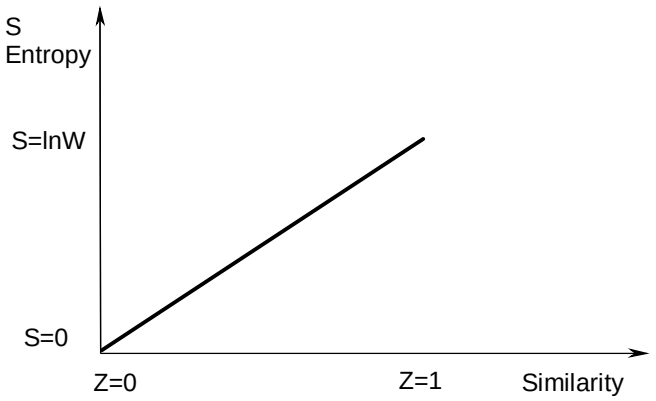
$$S\left(10^6,10^6\right) = -\sum_{j=1}^{10^6}\sum_{i=1}^{10^6}(p_{ij}\ln p_{ij}) = \sum_{j=1}^{990100}\ln 990100 + \sum_{j=990101}^{10^6}\ln 100 \approx 13714477\text{nats} \quad (33)$$

Therefore, comparing with equation (32), entropy is higher, $\sigma_a \approx 903903$, $M_a \approx 1.106$ according to equation (25) [17], and $D \approx 0.022$. The apparent species number (M_a) is much smaller than the species number ($M = 100$). As commented by *Wilson* [15], even though 100 species are present, we encounter the abundant butterfly almost all the time in the concerned system and each of the other 99 species only rarely. This is a system of very low evenness (or equitability [15]). Note equations (12), (20) and (24) are unable to differentiate different situations of evenness of the species abundance [17]. However, equations (15), (21) and (25) can be used to evaluate the influence of the evenness of the species abundance.

Similarly, when a combinatorial compound library is designed, high evenness should be desirable. If in the construction of a combinatorial compound library the reaction yields are different and if there are several reaction steps, some compounds will be much less abundant than others, and the apparent species number will become substantially lower than expected. Consequently, the molecular diversity of this library will be of poor quality (Figure 3a).



(a) Decrease discontinuously (J. W. Gibbs)



(b) Increase continuously (present work)

Figure 4. *Correlation of entropy of mixing with similarity according to conventional statistical physics (a); and the theory of the present author [8, 21,22] (b).*

Diversity of Combinatorial Compound Libraries

The molecular diversity of any assemblage of molecules can be calculated by using the formulas developed in the preceding sections. Because combinatorial compound libraries are prepared by changing the variants of molecular moieties, the calculation of molecular diversity can be substantially simplified.

For example, suppose a combinatorial compound library is constructed by multi-step (to say, *N* steps) reactions to synthesize M^N oligomers from *M* variant monomers, or synthesize M^N molecules composing *M* exchangeable molecular moieties (or *M* exchangeable functional groups or their precursors applied in *N* sites). Obviously the similarity among these M^N chemical species will rely only on the similarity of the *M* monomers. The molecular diversity of the whole combinatorial compound library depends only on the diversity of the *M* monomers, and independent of *N*, if the yields are all identical [18]. We defined $w=M^N$ as the number of microstates of the whole system earlier. In the present case of a combinatorial compound library, M^N is taken as the total number of compounds in the library.

Here the M^N chemical species can be regarded as the composite information registration species [19]. The increase in similarity among these monomers will result in a decrease in M_a from the maximum value *M*, and a decrease in the diversity index *D* from 1. All the formulae for molecular similarity measures can be employed to estimate the diversity of all the available and employed monomers and to calculate M_a and *D* for the available monomers.

Succinctly, *M*, the number of monomers [20], used in the combinatorial synthesis and their similarity, determine the molecular diversity of the combinatorial compounds. The higher the monomer number *M* is and the lower the similarity among the *M* monomers is, the better.

It is easy to see that the apparent distinguishable species number of a combinatorial compound library calculated by directly applying the equations developed in previous sections to all the individual compounds cannot be greater than the total number of the monomer variants [18].

As we have mentioned in the previous section, the abundance of a specific species in a combinatorial compound library relies on the yields of all the N steps of reactions. If the evenness of the abundance of species needs to be considered, equations (15), (21) and (25) should be used. In this case, however, the concentration of all the M^N chemical species should be directly considered.

Discussion

Intuitively, we understand that greater species diversity means greater information content in a system such as an ecosystem. Information is putatively a favorable factor [7a], at least for a compound collection or an ecosystem.

It is concluded that less similarity (Z) of the molecules in a mixture from a combinatorial chemical library implies a lower entropy content in a molecular assemblage and a higher quality of species diversity (D) than a mixture of a large number of indistinguishable compounds. From expressions of D and Z, the diversity and similarity are simply related by

$$D = 1 - Z \tag{34}$$

The logarithmic relations derived here may serve to quantitatively evaluate the quality of any species diversity in general and molecular diversity in particular. Factually, the species diversity assessment discussed here is one of the numerous applications [8, 21,22] of a new theory of entropy of mixing (Figure 4).

Therefore, species such as the millions of compounds generated by combinatorial chemistry and the biological species produced by recombinant DNA technology are cheap species: cheap because they are highly similar and thus of rather high entropy content [23]. For the cheap diversity, even though the prepared compounds may appear as huge numbers (M), the apparent species number M_a can be rather small (equation 24).

The entropy concept, as correlated to similarity, can be applied to differentiate cheap molecular diversity from the priceless diversity of chemical species. Generally, the high-quality diversity is the biodiversity [1] that has evolved and survived throughout millions of years (both the panda and the anaerobic ciliates are such species [24, 25]). For example, the giant panda (*Ailuropoda melanoleuca*) [24] is a precious species, because it is distinctly different from other animals. The ciliated protozoa are likewise precious species [25] in the sense that they are distinctly different from aerobic species and also different from many other anaerobic species. The existence of a large number of such distinctly different species keeps M_a high and is an indication that our biosystem is still far from being the state of maximum similarity and maximum entropy [26].

For molecular diversity, the high quality achieved relies on the distinct differences in both the structures and properties of the collected compounds. These compounds in isolated form are traditionally and still routinely prepared in the research laboratories and isolated from natural sources [27]. Molecules always have certain similarities. For a compound collection, the higher the compound sample number M, the greater the apparent chemical species number (equation 24).

As can be seen from Figure 2b, the information content cannot increase with the increase in the number of indistinguishable compounds. Therefore, in principle, there is no information difference between a container of 1000g of a compound and a container of only 1g of the same compound for one specific sample [28]. In many spectroscopic studies and biological activity tests, the required sample quantities are very small. Sometimes, a compound sample as little as 1 mg can be used for up to dozens of different biological activity tests.

Biodiversity preservation programs, such as a microorganism deposit project [29], are quite successful. Clearly it is worthwhile carrying out a program of coordinated, worldwide collection, deposit, storage and exchange of all the synthetic and natural compounds [27]. The logarithmic relations of entropy and indistinguishability number and the similar relation of information and species number provide a theoretical basis for these activities.

Acknowledgments. The author thanks Dr. E. Felder and many other colleagues in Ciba-Geigy Ltd. for numerous discussions about the molecular diversity preservation programs. He thanks Professor W. Graham Richards (Oxford University) and Professor David Avnir (The Hebrew University, Jerusalem, Israel) for their kind encouragement. The author also thanks one of the reviewers for the helpful comments.

References and Notes

1. Wilson, E. D. *The Diversity of Life*, Cambridge: Balkan Press of Harvard University Press; 1992.
2. Felder, E. R. *Chimia* **1994**, *48*, 531-541.
3. Gallop, M. A.; Barrett, R. W.; Dower, W. J.; Fodor, S. P. A.; Gordon, E. M. *J. Med. Chem.* **1994**, *37*, 1233-1251.
4. Pavia, M. R.; Sawyer, T. K.; Moos, W. H. *Bioorg. Med. Chem. Lett.* **1993**, *3*, 387-396.
5. Lowe, G. *Chem. Soc. Rev.* **1995**, 309-317.
6. (a) "Molecular diversity" has been used as a term in other cases with somewhat different meanings: For example, it is applied to the case where an enzyme or a receptor can bind strongly many structurally rather different ligands. The antonym of the term diversity used there is the specificity of a concerned interaction. A well-known and a very recent example is the binding of the NMDA receptor (e.g. [6b]) found in neuroscience investigations. E.g., the molecular and

functional diversity of the NMDA receptor channel [6c].
(b) Lin, S. -K. *Molecules*, **1996**, *1*, 37-40.
(c) Nishina, M.; Mori, H.; Araki, K.; Kushiya, E.; Meguro, H.; Kutsuwada, T.; Kashiwabuchi, N.; Ikeda, K.; Nagasawa, M.; Yamazaki, M.; Masaki, H.; Yamakura, T.; Morita, T.; Sakimura, K. *Ann. New York Acad. Sci*. **1993**, *707*, 136-152.

7. (a) In this paper, entropy is unambiguously defined as informational entropy. Notably, there were several interesting discussions on the physical and thermodynamic meanings of information. Entropy as well as free energy concepts were related to information registration processes. The discussions on compatibility of thermodynamic entropy and informational entropy can be found in the literature: *Entropy, Information, and Evolution*; Weber, B. H.; Depew, D. J.; Smith, J. D., Eds.; MIT Press: Cambridge, 1988. *Complexity, Entropy and the Physics of Information*; Zurek, W. H., Ed.; Addison-Wesley: Redwood City, California, 1990.
(b) The entropy of mixing ideal gases by opening the walls separating a rigid container has no observable thermodynamic effects such as heat and mechanical work. The Boltzmann constant or a positive constant normally used in entropy calculations of information registration systems is not yet clearly defined [8]. For these reasons entropy is unambiguously defined as informational entropy in the following discussions and this positive constant is taken as 1.
8. Lin, S. -K. *J. Chem. Inf. Comp. Sci.* **1996**, *36*, 367-376.
9. Wehrl, A. *Rev. Mod. Phys.* **1978**, *50*, 221-260.
10. Shannon, C. E.; Weaver, W. *The Mathematical Theory of Communication*, Urbana: University of Illinois Press, 1949.
11. von Neumann, J. *Mathematical Foundations of Quantum Mechanics*, Princeton: Princeton University Press, 1955; p.347.
12. Richards, W. G. *Pure Appl. Chem.* **1994**, *66*, 1589-1596.
13. Johnson, M. A.; Maggiora, G. M. eds. *Concepts and Applications of Molecular Similarity*, New York: Wiley-Interscience Publication, 1990.
14. Kubinyi, H. *QSAR: Hansch Analysis and Related Approaches*, Weinheim: VCH, 1993; 172.
15. The evenness of the abundance of species is defined as equitability, see ref. 1, 151.
16. If equation (20) is used, σ_a=1, because the 100 species are distinguishable.
17. If equation (20) is used, σ_a=1, because the 100 species are still distinguishable. Therefore, equations (12), (20) and (24) cannot differentiate evenness of the species abundance. Equations (15), (21) and (25) should be used.
18. The question why the diversity index (*D*) of a combinatorial compound library is independent of *N* if the reactions all have 100% yields will be discussed in detail elsewhere.
19. For example, ten speices can be represented by 10 binary strings composing only 0 and 1 (*M* = 2). These ten binary strings are examples of composite species. By increasing the length of the string, one can generate millions upon millions of composite species even as a binary string!
20. For an elegant example of monomer selection, see: Martin, E. J.; Blaney, J. M.; Siani, M. A.; Spellmeyer, D. C.; Wong, A. K.; Moos, W. H. *J. Med. Chem.* **1995**, *38*, 1431-1436.
21. (a) Lin, S. -K. *Gibbs paradox of entropy of mixing: Experimental facts, its rejection, and the theoretical consequences*, Papers presented at the Second International Congress on Theoretical Chemical Physics, New Orleans, Louisiana, April 9-13 1996.
(b) Lin, S. -K. *J. Theor. Chem.* **1996**, in press;
22. (a) Lin, S. -K. *The nature of the chemical process. Part 2. Mixing and separation*, Paper presented at The Fourth World Congress of Theoretically Oriented Chemists, Jerusalem, Israel, July 7-12, 1996.
(b) Lin, S. -K. *The nature of the chemical process. Part 3. Self-organization in hierarchical structures*, Paper presented at The Fourth World Congress of Theoretically Oriented Chemists, Jerusalem, Israel, July 7-12, 1996.
(c) Lin, S. -K. *The nature of the chemical process. Part 4. Formation of the chemical bond*, Paper presented at The Fourth World Congress of Theoretically Oriented Chemists, Jerusalem, Israel, July 7-12, 1996.
(d) Lin, S. -K. *The nature of the chemical process. Part 5. Deformation in energy transduction*, Paper presented at the Fourth World Congress of Theoretically Oriented Chemists, Jerusalem, Israel, July 7-12, 1996.
23. Note this is said without any disparagement to the techniques of combinatorial chemical libraries or the relevant biotechnology, which may be valuable for other reasons: the similarity and the gradual variation of structures in the combinatorial libraries permit adjustments designed to optimize structure-activity relations.
24. O'Brien, S. J.; Pan, W. -S.; Lu, Z. *Nature* **1994**, *369*, 179-180.
25. Fenchel, T.; Finlay, B. J. *Am. Scientist* **1994**, *82*, 22-29.
26. According to some definitions entropy is an adverse factor: it is associated with chaos, disorder, and the loss of information [7].
27. (a) Traditional practice of acquiring and distributing only knowledge (chemical information) can be reoriented to acquire and distribute chemical substances also [27b]: (b) Lin, S. -K. *Guide to the De-*

posit and Exchange of Compound Samples, ACS 212th National Meeting, Orlando, Florida, August 25-29, 1996.

28. When collecting a certain number of compound samples within a limited compound acquisition budget, the theoretical results of species abundance suggest that a collection of high evenness has higher molecular diversity. Instead of purchasing several unit amounts of one specific "good" sample, it is more desirable to acquire several different samples, one unit amount each.
29. World Intellectual Property Organization, *Guide to the Deposit of Microorganisms under the Budapest Treaty*, WIPO Publication No. 661 (E), Geneva: WIPO, Reprinted 1994.