Molecular Diversity, **5:** 13–24, 2000.
KLUWER/ESCOM

Comments on the design of chemical libraries for screening

Hugo O. Villar* & Ryan T. Koehler
Telik, Inc., Discovery Technologies Division, 750 Gateway Blvd., South San Francisco, CA 94080, U.S.A.

Received 7 May 2000; Accepted 16 June 2000

Key words: conformational flexibility, diversity analysis, diversity measure, library design, molecular descriptors, molecular representation, pharmacophore representation, small molecule libraries

Summary

Different representations of molecules, based on distinct sets of properties can yield different perspectives of the issues involved in library design. In particular, different chemical representations can give rise to very different estimates of required library sizes. We provide a preliminary mathematical framework that examines the size of libraries required to adequately sample the spaces corresponding to some commonly used property sets. Introduction of conformational flexibility is also discussed as a means of increasing coverage of chemical libraries, while at the same time considering the thermodynamic consequences of flexibility upon detectable activity. Our theoretical analysis reveals that the property spaces currently in use are extremely large and unlikely to provide adequate discrimination among compounds.

Introduction

The design of chemical libraries for high throughput screening has become an important problem in modern drug discovery with the advent of high and ultra-high throughput screening techniques [1]. The ultimate goal of library design is the selection of compounds for screening that maximize the chances of identifying ligands for any given target, that could possibly be developed into drugs [2].

The chances of identifying such ligands can be increased when knowledge about the characteristics of the target, or the ligands that bind to it, is exploited to bias the selection process. However, many times little is known about the target, or in other cases the same library is to be adopted repeatedly for a number of unrelated targets. Under those circumstances, the mind set in the field of library design has been that no assumptions should be made regarding the nature of the targets. Compound libraries should, thus, be designed without preconceived notions about what properties are desirable. These unbiased libraries increase the chance that at least one of the compounds contained in it will complement the properties of an arbitrary target. However, if the compounds are ultimately to be used as drugs, certain biases are necessary and helpful to limit the ranges and types of properties to those relevant to pharmaceuticals [2]. Even within these constraints maximal variability of the properties is important to maximize the chances of success.

Methods that use the structural or physicochemical properties of the molecules in the library are the most commonly used for diversity analysis [3–10]. These computational methods have been very valuable in removing improper biases or redundancies as libraries are assembled. Each property type or set of molecular descriptors studied defines a distinct type of chemical representation, and provides a different interpretation of chemical diversity. Compounds that appear diverse in one representation may well be considered similar in another. Among the most common representations [3–7], discrete-valued descriptors derived directly from features of chemical structure, such as structural keys, have been frequently applied [11]. Continuous-valued global properties [6,7] derived from chemical structure have also been used, with examples that include physicochemical parameters such as log P and topological indices. To encode

* To whom corespondence should be addressed. E-mail: hugo@telik.com

the spatial relationship among atoms, or properties that are a function of such arrangement, features of three-dimensional (3D) structure models are used [3,9]. A 'property space' is then formed from the union of molecular descriptors chosen for study. Ideally, a well designed library should saturate the chosen property space, without overly populated regions, but perhaps with vacant regions tailored in to avoid combinations of properties incompatible with drugs [2].

Diversity analysis has been limited, with a few exceptions [12,13], to the properties of small molecules in isolation. Nevertheless, the ability of the molecules to interact with a macromolecular target depends on multiple factors, some of which are beyond the scope of the isolated molecular descriptors. These factors, which are missed by simple structural or physicochemical representations of the molecules in isolation, have been considered extensively in the context of biophysical studies and structure based drug design but are seldomly invoked in relation to library design [14,15]. Identification of ligands for a given target will be limited by the free energy of the interaction between the molecules, but this is only partly encoded in the simpler properties describing a ligand in isolation.

The goal of library design could be taken beyond avoiding redundancy, to increase the chances of identifying a ligand. Because the goal is to identify potentially interesting new molecules, other constraints in addition to those imposed by drug compatibility should be considered. In particular, compounds likely to suffer significant energetic penalties upon interaction with a macromolecule should be disfavored during the process of designing a library.

A key factor affecting the ability of a ligand to interact productively with a macromolecule is its flexibility. Library design literature has only paid peripheral attention to this issue [16]. Most steric and electronic properties of small molecules are dependent on the conformation. Flexible compounds may therefore be expected to sample a significantly larger number of properties and pharmacophore arrangements than rigid ones, and consequently should be more likely to accommodate the requirements of any target. This argument is compelling and it is tempting to conclude that flexible ligands should be preferred to rigid ones when selecting compounds for a generic screening library. However, flexible molecules need to overcome significant energetic penalties [17,20] in order to achieve the conformation(s) required for target binding, which may effectively render them inactive at a preset concentration.

This manuscript has two purposes. We want to analyze how different property spaces alter our perception of how complete or close to saturation a given library is. Limits on the number of molecules required to completely cover the property space corresponding to different chemical representations are discussed. The study is complemented with some comments on the influence that conformational flexibility has on the ability to cover a selected property space.

Coverage of the property space

The importance attributed to molecular diversity as a tool for library design stems from our belief that the larger the fraction of a relevant property space that is covered, the greater the chance that the library will contain hits capable of complementing any random target. Different sets of descriptors and properties may be employed to describe molecules for use in the design of chemical libraries and each of these different sets will lead to different impressions of how complete a given library is. Hence, the number of molecules that is required to sufficiently cover a property space is an open question. It will depend on the descriptors and similarity metrics chosen as well as the desired compound density, among other factors.

Two-dimensional binary representation

Descriptors encoding structural features of two-dimensional (2D) chemical graphs, such as structural keys, have been used with some success in different aspects of library design and the selection of compounds for screening [6,21]. Structural keys [11] typically take the form of a binary array, where each array element represents the presence or absence of a specific 2D fragment in a given molecule. For a molecule M this is represented by:

$$M = (m1,m2,m3,m4,...,mn), \qquad (1)$$

where mi = 1 if the structural feature associated with the i-th position in the array is present and 0 otherwise. The number of structures that are required to completely cover chemical space in this representation is given by:

$$\sum_{j=1}^{n} \frac{n!}{j![n-j]!}, \qquad (2)$$

where n is the dimension of the array. The assumption is that each of the elements of the array is independent. If all the possibilities are enumerated, then the number of structures required for complete coverage is 2^n. In the case of the MDL MOLSKEYS [22], a widely used set that have given good results in the past [6], there are 166 distinct feature bits. For the case with n = 166 the number of different possibilities is approximately 10^{50} and even larger key sets have been analyzed.

In general, small molecules cannot possibly contain all of the different features encoded in structural key bits simultaneously. Indeed, only a fraction of discriminated features are actually present in most molecules, and consequently only a fraction of the corresponding keys are set in a small molecule. If only k elements could be set at any one time then for an array of dimension n, the total number of possibilities is given by:

$$\sum_{j=1}^{k} \frac{n!}{j![n-j]!} \qquad (3)$$

As a practical example let's analyze the number of features encoded by the MDL MOLSKEYS [22] that occur in a sample of 4000 compounds from the comprehensive medicinal chemistry (CMC) database (MDL, Inc., San Leandro, CA) or 10 000 chemicals from the Maybridge library (Maybridge Chemical Co, Ltd, Cornwall, U.K.). Figure 1 shows a distribution of the count of bits set for records for these different databases. The average number of encoded features, i.e. bits, encountered per compound in the CMC set is 44, out of 166 possible. Less than 1% of the chemicals sampled have more than 78 or less than 16 bits on. Results are similar for the Maybridge collection. Restricting ourselves to only those possibilities with 16 to 78 bits on (i.e. j = 16, k = 78 in Equation 3), the total number of possibilities is still on the order of 10^{49}.

Therefore, even when a substantial restriction on the number of elements that may be set simultaneously is taken into account, the number of compounds that may be discriminated is still a large fraction of all possible representations. The number of different possibilities is truly enormous and consequently the space will inevitably be sparsely populated, regardless of library size. In any case, an analysis of this type is useful to help avoid marked redundancies in the libraries.

The dimensionality of the problem is the issue, and vectors of 166 elements will yield corresponding spaces that are extremely large, regardless of other restrictions. Attempts to focus on only a fraction of the properties, e.g. by considering a subset of the bits, have been made, but mostly to assess similarity [21,23]. Reductions in the number of features could provide more realistic property spaces, but care must be taken to ensure that encoded features are selected to maximize the information relevant to a given task, such as drug discovery. However, at this stage our knowledge about chemical diversity as it relates to the drug discovery problem is in its infancy and significant work needs to be done before such tasks may be expected to succeed.

Global molecular representation

Global molecular properties have a long tradition as descriptors in QSAR and have been employed for library design [6,24]. Contrary to structural key representations, where features are represented as discrete bits, these properties generally take on continuous values. Some global molecular properties such as octanol–water partition coefficient (logP), pKa or topological indices are routinely used to identify similar compounds. Because of the continuous nature of the properties, quantifying the number of compounds required to saturate a corresponding property space requires some subjective judgment about the required density of sampling for each property. One way to choose an appropriate density is to assume that compounds with properties falling within certain value ranges are similar enough to be considered equivalent [3]. The specific ranges of property values in which compounds can be considered equivalent will, of course, depend on the property considered. Value ranges chosen for each property determine how densely that dimension of property space will be populated.

Each property may warrant sampling at different resolutions. In some cases, presence or absence of a certain property may be considered sufficient to discriminate molecules, while in other cases, the property should be more finely sampled. In the past, global properties have been normalized and used in combination with principal component analysis as a means to reduce the dimensionality of the problem [6]. In such cases, all resulting dimensions should in principle be equally populated.

Ascribing property values to discrete ranges permits simplification of the counting process because this reduces each property dimension to a finite number of elements. This process of 'binning' continuous properties has been put forward in the past to analyze chemical diversity [3]. In such a representation,

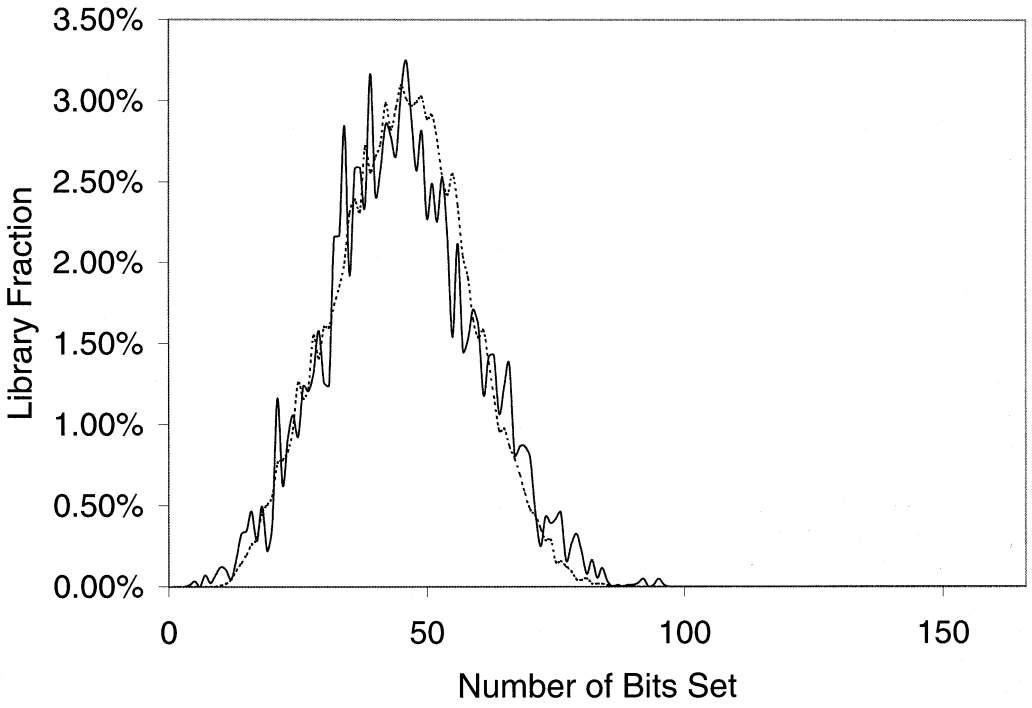

Figure 1. Percentage of molecules as a function of the number of bits set in the MOLSKEYS representation. The solid line shows 4000 compounds from the cmC database, while the dashed line shows the values for 10 000 compounds from the Maybridge collection.

a library with compounds in all bins provides an even sampling of the property space. We will adopt the binning concept of continuous valued properties, as it simplifies the counting process.

As an example, a scalar molecular property p with values that can range between p0 and pt can be divided into n segments. If these all have the same length, then $\Delta p = (p0–pt)/n$. Once discretized, the property may then be represented by a binary array:

$$P = (b0, b1, b2, ..., bn), \quad (4)$$

where bi = 1 if $p \in [p0 + (i–1)\ \Delta n, p0 + i\Delta n)$, otherwise bi = 0. Thus, the vector that represents that property for that molecule has a single 1 in a position in the array that corresponds to the range within which the value of the property p falls. Assuming there are m properties P, a description of the molecule M results from a combination of the property vectors.

$$M= P1 \cup P2 \cup P3 \cup ...\cup Pm \quad (5)$$

M is a sparse vector with exactly m elements equal to 1 and the remaining null.

Because each property Pj can be divided into different numbers of segments bj, the total number of molecules necessary to cover all property space, using the density desired for each property is given by:

$$\prod_{j=1}^{m} bj, \quad (6)$$

where m is the total number of properties being considered. For example, suppose 12 independent properties (or the first 12 principal components of a larger property space [6]) are each equally divided into 10 bins. This corresponds to dimension of m = 12, dimension of n = 10 (for each), dimension of m = 120, and each representation having exactly 12 bits set. In this case the total number of molecules that is required to completely cover the space is 10^{12}. Typically, the number of properties and ranges considered for these spaces are still extremely large, compared to the size of current chemical collections, even if orders of magnitude smaller than the structural keys.

Pharmacophore-based representation

Molecules are, of course, 3D entities where shape and conformation generally relate to the possibility of having a productive interaction with the macromolecular target. Accordingly, numerous approaches based on

this relation have been devised for the definition of property spaces to be used for diversity assessment. Pharmacophore-based diversity measures, where 3D arrangements of chemical groups in space are used to derive descriptors, have been implemented [3]. In fact, several groups have used the number of three or four point pharmacophores represented in a collection of compounds as a diversity criterion [25]. Pharmacophore centers are typically those associated with properties involved in intermolecular interactions. Commonly, hydrogen bond donor and acceptor sites, positively or negatively charged centers and lipophilic rings are singled out.

In general, distances among pharmacophoric centers will vary in a continuous fashion among compounds (and conformations). When the variations between two different pharmacophore distances are small, the two may be considered equivalent, as done in the case of global properties. The equivalency of structures leads to a finite number of distinct pharmacophore arrangements, and allows us to parallel the analysis directly from that of continuous global properties. Binning of pharmacophore patterns is used in the major commercial software for diversity analysis based on pharmacophores [25,26].

The arrangement of pharmacophore centers in 3D space is normally expressed in terms of the distances among those centers. With the exception of mirror image effects, the overall arrangement of points may be defined by specifying, for each unique pair of centers, both the type of interaction and the corresponding distance. If P kinds of pharmacophore properties (H-bond donor, H-bond acceptor, etc.) are considered, then the number of unique *types* of pairwise interactions that can be described is given by:

$$n_P = P(P+1)/2 \qquad (7)$$

Ranges of distances for the same pair of properties can be considered equivalent, and therefore the pairwise distances can be binned. Figure 2 illustrates such a pharmacophore key. Two hydrogen donor centers (H), a hydrogen bond acceptor (N), and a lipophilic center represent a scheme of a pharmacophore contained in a hypothetical molecule. All possible distances are binned into five groups. If P1 is hydrogen bond acceptors, P2 is hydrogen bond donors and P3 is lipophilic centers, the first 15 elements of the pharmacophoric key are shown (corresponding to P11, P12, P13); the later 15 are not shown (corresponding to P22, P23 and P33).

The dimension of the key required to distinguish the different types of interactions is, for any given number of properties P, binned into b groups:

$$N_{key} = b \cdot P(P+1)/2 \qquad (8)$$

This is the number of unique types of interactions considered times the number of equivalent bins.

For simplicity, this formulation merely discriminates between the presence or absence (1 or 0) of a specific type of pairwise arrangement and will not keep track of the number of occurrences of such arrangements. The conclusions would be magnified if the number of occurrences were taken into account.

Generally speaking, a pharmacophore may involve any number of centers, each center is identified with any one of the P pharmacophore properties. Thus, the term 'A-point pharmacophore' is used to denote an arrangement of A pharmacophore centers in 3D space. Every triangle in Figure 2 shows a 3-point pharmacophore, while the entire figure represents a 4-point pharmacophore. 3-Point pharmacophores, or pharmacophore triangles, have been used routinely for a number of years for library design [3,26], while 4-point pharmacophores have more recently been implemented [25].

If we restrict ourselves to A-point pharmacophores, then the number of pairwise arrangements that define the pharmacophore key is given by:

$$n_A = A(A-1)/2 \qquad (9)$$

Each 3-point pharmacophore (A = 3) has 3 distances that define it ($n_3 = 3$), while a 4-point pharmacophore has 6 pairwise distances ($n_4 = 6$). Note that Equation 9 differs in form from Equation 7 because pairwise counts of pharmacophore centers to themselves *are not* considered, whereas arrangements between two distinct centers of the same pharmacophore type *are* counted. The n_A pairs of centers may or may not give rise to n_A *unique* types of pairwise arrangements, if there is redundancy. For example, if every center in the pharmacophore is a hydrogen bond donor, then there is only one type of pairwise arrangement accounted for in the pharmacophore key. Assuming that $A \leq P$, then an A-point pharmacophore may give rise *at most* to n_A unique types of pairwise interactions.

When figuring the number of distinct bits to set in a pharmacophore key, distance must also be considered. Each arrangement is associated with one of b different distance categories, so even if every center in the A-point pharmacophore is a hydrogen bond

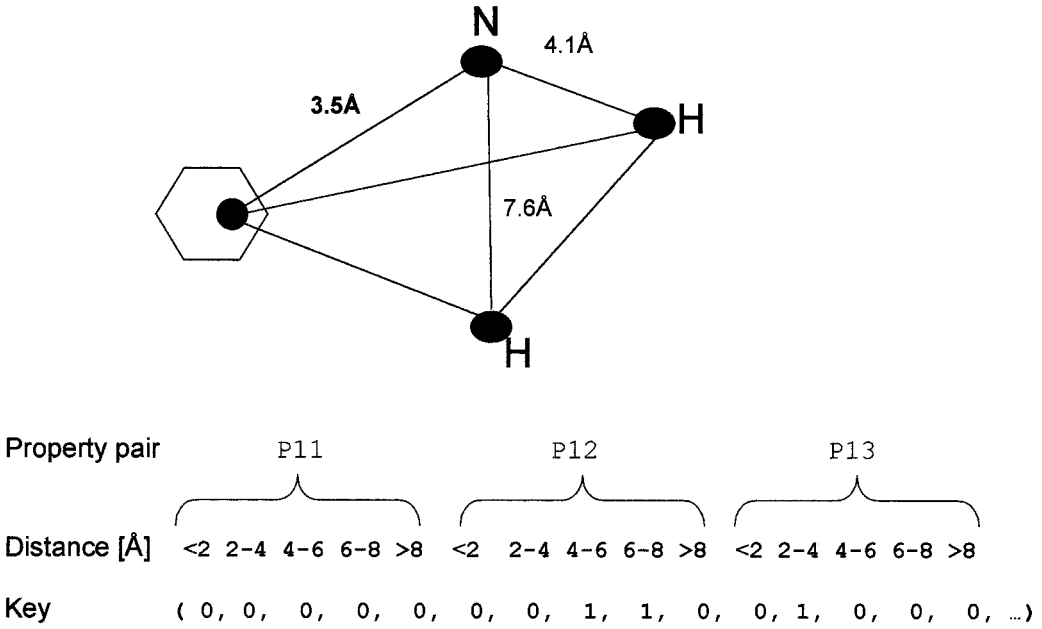

Figure 2. Scheme of a pharmacophore, and its pharmacophore key. N represents a hydrogen bond acceptor (P1), H a hydrogen bond donor (P2), the hexagon a lipophylic center (P3), and the pairwise distances are grouped into 5 bins (b). Only the first terms of the pharmacophore key for such a scheme are shown.

donor, multiple bits may still be set if more than one distance category is encountered. However, if a molecule contains only one distance bin for all equivalent centers (all hydrogen bond donors), only one element of the pharmacophore key will be set on. These are two extreme cases and all intermediate situations are also possible. Therefore, n_A is the upper bound for the number of distinct bits that may be set for an A-point pharmacophore. A-point pharmacophores can have any number of bits set between 1 and n_A but no more.

If we know that an A-point pharmacophore sets exactly j distinct bits, then the number of different bit patterns that could be generated is given by the binomial coefficient:

$$m_j = \frac{N_{key}!}{j!(N_{key}-j)!} = \frac{b{\cdot}P(P+1)/2]!}{j![b{\cdot}P(P+1)/2-j]!} \quad (10)$$

Knowing that j can take on values between 1 and n_A, the number of different A-point pharmacophores that can be distinguished in this representation is:

$$N_{A-point} = \sum_{j=1}^{n_A} mj = \sum_{j=1}^{A(A-1)/2} \frac{[b{\cdot}P(P+1)/2]!}{j![b{\cdot}P(P+1)/2-j]!} \quad (11)$$

The expression corresponds to the number of possible combinations of bits that could be set on for an A-point pharmacophore, when P properties are used and binned into b-pairwise bins. The expression gives the number of compounds that could be differentiated in pharmacophore-based representations.

Equation 11 has a stronger dependence on the number of pharmacophore properties (P) than on the number of pharmacophore points (A) involved. Figure 3 shows how the log of $N_{A-point}$ varies as a function of both P and A when the number of distance bins is 10 for typical values of P and A.

While Equation 11 provides the number of distinct A-point pharmacophores in this representation, it does not indicate the number of compounds that would normally be required to cover all of these possibilities or saturate the property space. A molecule could contain multiple A-point pharmacophores, as in Figure 2, the scheme represents a molecule with 4 possible 3-point pharmacophores. If a molecule contains a total of C centers which can be ascribed to any of the P pharmacophore properties, then the number of A-point pharmacophores found in these structures (assuming $C \geq A$) is given by

$$N_{C,A} = \frac{C!}{A!(C-A)!} \quad (12)$$

Table 1. Average properties of 4000 compounds selected from the CMC database

Property	Average
Molecular weight	357
Number of rings	2.6
Number of H-bond donors	2.0
Number of H-bond acceptors	5.9
Number of nitrogens	2.1
Number of oxygens	3.3
Number of carbons	18.5

An analysis of a selected list of compounds from the Comprehensive Medicinal Chemistry database (MDL Inc., San Leandro, CA) was done in order to determine typical values for C (Table 1). Considering just hydrogen bond donor and acceptor centers, as well as rings that may comprise lipophilic centers, C is found to be approximately 10 for compounds in the CMC database. An average molecule would thus contain 120, 3-point and 210, 4-point pharmacophores.

From a different angle, some general empirical rules have been put forward in the literature regarding the structural characteristics of compounds with optimal physicochemical properties for drug development [27]. The rules impose limits on the physicochemical properties of the molecules that could be viable for drug development. The most common rules [27] impose four constraints on compounds, as they should have less than 5 hydrogen bond donors, less than 10 hydrogen bond accepting centers, molecular weights of less than 500, and a cLogP (partition coefficient) of less than 5. Limits on the total numbers of pharmacophores result from their application. For example, since the sum of hydrogen bond donors and acceptors may not be larger than 15, the maximum number of 4-point pharmacophores is 1365, according to Equation 12. Because the total molecular weight is also constrained to 500, it is hard to envision compounds that satisfy the rules with 15 centers in total, including lipophilic areas. Thus, the overall number of pharmacophores that can be represented in pharmacologically viable molecules is relatively small to make a significant impact in the size of library.

In general, the pharmacophore properties of C centers in a given molecule will not all be distinct, and so the associated A-point pharmacophores may be indistinguishable. A given 3D arrangement of pharmacophore centers may occur more than once in a single molecule. As Equation 11 before, Equation 12 provides only an upper bound for the number of unique A-point pharmacophores that could come from a particular compound. Nevertheless, Equation 12 would be exact if each pharmacophore found is unique, and if we further assume that each compound examined provides a collection of pharmacophores that is distinct from all of the other compounds in the library, then an estimate of the *minimum* number of molecules required to cover all A-point pharmacophores is afforded:

$$\frac{N_{A-point}}{N_{C,A}} = \frac{A!(C-A)!}{C!} \sum_{j=1}^{A(A-1)/2} \frac{[b{\cdot}P(P+1)/2]!}{j![b{\cdot}P(P+1)/2-j]!} \quad (13)$$

This function shows a stronger dependence on the number of properties (P) that are considered than on the number of pharmacophore centers (A). Considering 3-point pharmacophores and four property types leads to a number of molecules on the order of half a million compounds (see also [3]). This is the size of many current libraries. As more complex pharmacophores, or additional properties are considered, the number of potential compounds required to sample the entire space defined grows extremely fast. Since the term that premultiplies Equation 13 is small compared to the summation, the rate of growth is similar to that shown in Figures 3a and 3b for the summations alone. Even when a reasonable number of properties and pharmacophoric points are considered, the numbers of possibilities continue to be overwhelming. In part this is the reason why chemical libraries have increased but the potency of the compounds does not parallel the increase in library size. An increase in potency would require an increase in the number of pharmacophoric points matched, which is not linear with the total number of compounds required to sample the space. The order of magnitude of these numbers also reveals the enormity of the task faced by medicinal chemistry. If medicinal chemistry were carried out completely at random, it would be very hard to increase the potency of compounds. By experience, medicinal chemists have generated a process that follows an implicit approach to optimizing the variables involved.

Consideration of conformational flexibility on the part of the ligand can alter our estimate of the minimum number of molecules needed for coverage provided by Equation 13, as it will increase the

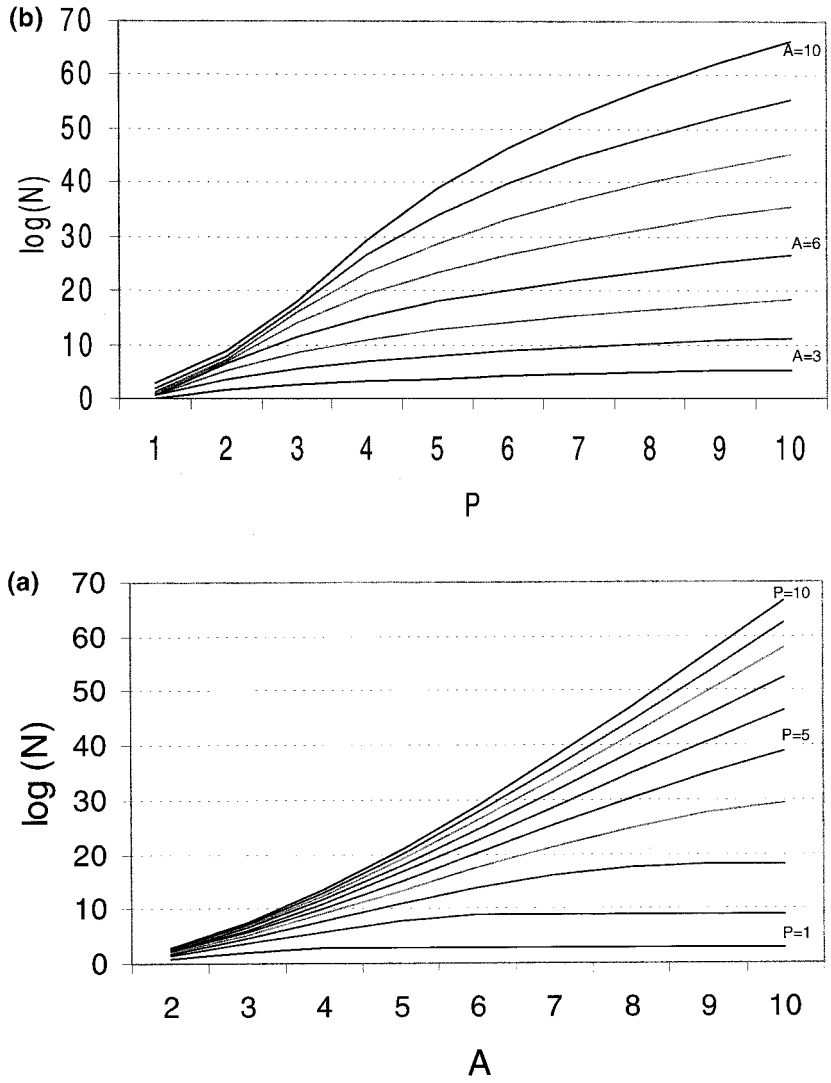

Figure 3. Decimal logarithm of the estimate of the minimum number of molecules required to cover all A-point pharmacophores considering P properties, according to Equation 13. In all cases C = 10 and b = 10. (a) as a function of the number of points considered for the pharmacophore, when P = 4; (b) as a function of the number of classes of atomic centers considered, when A = 4. The equation shows strong dependence with both variables, slightly more pronounced with the number of properties (P) considered.

average number of different pharmacophores per molecule, effectively increasing coverage of the chemical space for the same number of compounds.

Effects of conformational freedom

Conformational flexibility on the part of the ligand should increase the coverage of a property space that includes descriptors that encode 3D features of molecules. This is because flexible molecules will be able to sample a larger number of different pharmacophores than rigid ones, as every unique placement of atoms in space may yield a new set of arrangements of pharmacophore centers. Flexibility enhances the chances that a given molecule may adopt a favorable spatial positioning of atoms to complement a target. It also changes the ranges of values that any conformationally dependent 3D property may adopt. Flexibility will increase the coverage of an A-point pharmacophore space afforded by a given number of molecules when compared to the case where flexibility is absent or ignored.

The increase in property space coverage afforded by flexibility is a function of the total number of conformers present in each molecule. If each conformer were to yield a completely unique arrangement of pharmacophore centers, Equation 12 could simply be multiplied by the number of conformers per molecule. This situation is seldom the case, because subsets of pharmacophoric centers can remain in the same relative orientations in different conformations, resulting in equivalent pharmacophores being obtained. Thus, the total number of pharmacophores sampled by a molecule will generally be smaller than the number of conformers present multiplied by Equation 12, but this quantity provides again an upper limit, which in turn bounds the minimum number of molecules that are necessary to completely cover property space. The lower boundary provided by Equation 13 can be rewritten as:

$$\frac{N_{A-point}}{N_{C,A}} = \frac{1}{n_{conf}} \frac{A!(C-A)!}{C!} \sum_{j=1}^{A(A-1)/2} \frac{[b \cdot P(P+1)/2]!}{j![b \cdot P(P+1)/2 - j]!}, \quad (14)$$

where n_{conf} is the average number of conformers generated by each molecule.

Still the number of molecules required is reduced, but in addition conformational flexibility can be a problem in the design of libraries.

Free energy penalty and conformational flexibility

The chance of finding a hit or a lead in a library is determined by a preselected cut-off value for accepting a compound as a hit in the assay. Depending on the targets and other assay specific criteria, those cut-offs may typically be set between 20 μM and submicromolar. These different potency criteria translate into different minimum free energy (ΔG) values that need to be achieved for a compound to be detected as a hit (ΔG_{det}), according to [28]:

$$\Delta G = - RT \ln K_i \quad (15)$$

If the inhibition constant K_i corresponds to the cut-off set for that assay, then ΔG corresponds to ΔG_{det}.

For every 10-fold increase in the cut-off, ΔG_{det} is increased by 1.5 kcal/mol. This energy is on the order of approximately an additional interaction between ligand and target when the decrease in entropy and changes in solvation are taken into consideration [27]. If covering an A-point pharmacophore space is required to ensure that a hit will be found at the 10 μM level, increasing the cut-off value for K_i by 10-fold to 1 μM implies that an A+1 pharmacophore space should be covered to ensure the existence of a hit. This is because an additional interaction between the receptor and the ligand should be fixed.

When small molecule libraries are composed of equivalent compounds in terms of average number of pharmacophore centers present, additional energy terms allow us to compare the free energy of binding for flexible ligands to otherwise equivalent rigid ones. One term is the internal energy difference between bound and free states (ΔG_{conf}), and the other is the difference due to loss of conformational entropy (ΔG_{loss}) upon binding.

For a significant number of ligands, the conformational energy penalty, ΔG_{conf}, has been found to be less than 3 kcal/mol [20], but it may be larger in some cases. The loss of conformational entropy ($\Delta G_{loss} = -T\ \Delta S_{loss}$) is also significant and has been found to be, on average, 0.5 to 0.7 kcal/mol per degree of freedom restrained upon binding [17]. For any molecule to be detected in a screening assay the detection limit becomes [17]:

$$\Delta G_{det} \geq \Delta G_{int} + \Delta G_{conf} + \Delta G_{loss} \quad (16)$$

For a rigid molecule, the last two terms are null. Therefore, if two molecules were able to interact with the target in such a way that they generate a similar free energy of interaction (ΔG_{int}), the more flexible molecule would not be detected because it would have to pay an additional penalty to reach the detection level.

As an example, let's assume that the threshold is set to be equivalent to a K_i of 10 μM, then, ΔG_{det} must be lower than –7.1 kcal/mol. For a rigid molecule, a ΔG_{int} of –7.1 kcal/mol would be sufficient for detection. But for an equivalent flexible molecule each degree of conformational freedom that is lost upon binding requires that ΔG_{int} increases by at least 0.7 kcal/mol in order for the molecule to be detected. If the molecule loses two degrees of freedom, ΔG_{int} should be –8.5 kcal/mol. Molecules that on average lose two degrees of freedom upon binding incur a penalty similar to increasing the detection cut-off by one order of magnitude in rigid molecules.

The magnitude of energies we have been discussing is, coarsely speaking, that typically associated with a single hydrogen bond (1.4 kcal/mol) [29]. One way to compensate for the penalty imposed by the loss of conformational freedom is by picking up at least one additional point of ligand-target interaction. The penalty is even larger when ΔG_{conf} is not null, which is

the most common occurrence, and may account for 3 kcal/mol in a majority of the interactions [20]. Certainly, changes in the free energy of solvation and other factors could compensate, but then the molecules could not be regarded as equivalent.

One way to compensate for the energy penalty stemming from greater degrees of freedom associated with flexible ligands is to increase the number of interactions to reach the preset ΔG_{det}. For instance, if the level of detection for a rigid molecule could be attained with four interactions (4-point pharmacophore), an otherwise equivalent flexible molecule that lost two degrees of freedom upon binding to the target would require five interactions to reach the detection level. The penalty imposed by two degrees of conformational flexibility makes it necessary to have a larger number of matching interactions to ensure detection. Therefore, a larger number of pharmacophore points are necessary to ensure that flexible compounds will be detected in a library.

The difference between A-point pharmacophores and the (A+1), i.e. the amount necessary to approximately offset the loss of two degrees of freedom can be derived from Equation 14:

$$\frac{1}{N_{conf}} = \frac{A!(C-A)!}{C!} \sum_{j=A(A-1)/2+1}^{A(A+1)/2} \frac{[b{\cdot}P(P+1)/2]!}{j![b{\cdot}P(P+1)/2-j]!} \quad (17)$$

With typical values of P = 5, b = 10 and all triangle pharmacophores (A = 3), the additional number of pharmacophore configurations that are needed to ensure a saturated library would be on the order of 10^6. The situation becomes worse when larger numbers of degrees of conformational flexibility are considered.

At the same time, the maximum contribution a flexible molecule could make to the total number of pharmacophores is proportional to the number of unique conformers, provided that each of these contributes a new pharmacophore. *The number of pharmacophore configurations that are required to ensure the additional interaction that has to be picked up because of the entropy loss, cannot be compensated by the number of conformers added by a flexible molecule.* Two degrees of conformational freedom provide a smaller fraction of the 4-point pharmacophore possibilities than a library of the same number of rigid compounds provides for a 3-point pharmacophore, both of which are able to reach an equivalent level of detection. Since the coverage is less, it is also less likely that the right compound will be found for an arbitrary target when a flexible library is used.

Conclusions

The different representations that are used to study chemical diversity generate a larger number of possibilities than the size of the most ambitious real libraries. Typically used triangle pharmacophores provide the possibility to discriminate over half a million compounds, and libraries of this size are in use for high throughput screening [3]. Other representations provide vast spaces that can seldom be covered, and could bring into question even their effectiveness in removing redundancy at the level of today's screening libraries.

Conformational flexibility is not the solution to the problem of increasing the coverage of property space for a chemical library. Indeed, in those cases, the loss of conformational flexibility of the ligands upon binding imposes a penalty on the free energy of interaction that is, on average, equivalent to having an additional point of interaction for every two degrees of freedom lost upon binding. If a three-center pharmacophore is sufficient to produce a productive interaction between a rigid small molecule and a macromolecule, a library with compounds of about two degrees of freedom will require 4-point pharmacophore interactions. With two additional degrees of freedom, a library provides significantly smaller coverage of the 4-point pharmacophore space than a library of the same number of rigid compounds provides for a 3-point pharmacophore.

The complementary idea that making molecules rigid increases the affinity of a compound for the target, provided the elements of the pharmacophore are kept intact, is widely used in medicinal chemistry [30]. Numerous examples exist of enhancement of the affinity when the conformational flexibility is decreased. Significant conformational flexibility may hinder our ability to detect a productive interaction.

Our analysis is of a coarse nature, when the pharmacophore representations are considered. The most significant shortcoming of the analysis is that it assumes that only one type of molecule or pharmacophore is able to generate an appropriate interaction with the binding site. Indeed, the same sites can be inhibited by many different molecules in radically different ways, even when they are structurally related.

The large numbers of molecules resulting from our estimates do not reflect the ability of the binding sites to alter their conformations and adapt to a potential partner via induced fit. Certain sites are clearly unique in their ability to bind promiscuously, as is the case for most metabolizing enzymes, reflecting their ability to recognize multiple pharmacophores. Therefore, Equation 14 should also be divided by the average number of A-point pharmacophores that a binding site contains, even when the conformational flexibility of the site is considered. However, evaluating such numbers is not a trivial task and requires a systematic evaluation of the characteristics of binding sites. Even if the magnitude of the numbers produced with the pharmacophore analysis is not precise, the trends and relative values should be accurate.

Even with those limitations, the analysis reveals some of the problems with our approach to diversity analysis. Defining exactly which properties are most relevant to the problem at hand is still a critical aspect of the analysis. Significantly different representations can be obtained using different types of molecular properties. The large numbers that follow from all of these descriptions are to some extent a reflection of the inadequacy of the properties widely in use.

Finally, it is clear that when designing libraries, rigid compounds should be preferred. Flexible compounds have to pay an additional energy penalty that is not required from rigid ligands, which would make it harder to reach a pre-set detection cut-off, as usually done in high throughput screening. The concept is not new, but the magnitude of the penalty imposed in terms of compounds necessary to achieve a similar probability of success can be surprising.

References

1. Gibbon, J.A., Taylor, E W. and Braeckman, R.A., In Gordon, E.M. and Kerwin, J.F. (Eds.), Combinatorial Chemistry and Molecular Diversity in Drug Discovery, Wiley-Liss, New York, NY, 1998, p. 453–474.
2. Martin, E.J. and Critchlow, R.E., *Beyond mere diversity: Tayloring combinatorial libraries for drug discovery*, J. Combinat. Chem., 1 (1999) 32–45.
3. Martin, Y.C., Brown, R.D. and Bure, M.G., *Quantifying diversity*, In Gordon, E.M. and Kerwin Jr., J.F. (Eds.), Combinatorial Chemistry and Molecular Diversity in Drug Discovery, Wiley-Liss, New York, NY, 1998, pp. 369–385.
4. Martin, Y.C., *Challenges and prospects for computational aids to molecular diversity*, Perspect. Drug Discov. Design, 7/8 (1997) 159–172.
5. Bures, M.G. and Martin, Y.C., *Computational methods in molecular diversity and combinatorial chemistry*, Curr. Opin. Chem. Biol., 2 (1998) 376–380.
6. Brown, R.D. and Martin, Y.C., *The information content of 2D and 3D structural descriptors relevant to ligand-receptor binding*, J. Chem. Inf. Comput. Sci., 37 (1997) 1–9.
7. Brown, R.D., *Descriptors for diversity analysis*, Perspect. Drug Discov. Design, 7/8 (1997) 31–49.
8. Willett, P., *Computational tools for the analysis of molecular diversity*, Perspect. Drug Discov. Design, 7/8 (1997) 1–11.
9. Chapman, D., *The measurement of molecular diversity: A three-dimensional approach*, J. Comput.-Aided Mol. Design, 10 (1996) 501–512.
10. Lajiness, M.S., *Dissimilarity-based compound selection techniques*, Perspect. Drug Discov. Design, 7/8 (1997) 65–84.
11. Martin, E.J., Blaney, J.M., Siani, M.A., Spellmeyer, D.C., Wong, A.K. and Moos, W.H., *Measuring diversity: experimental design of combinatorial libraries for drug discovery*, J. Med. Chem., 38 (1995) 1431–1436.
12. Briem, H. and Kuntz I.D., *Molecular similarity based on DOCK-generated fingerprints*, J. Med. Chem., 39 (1996) 3401–3408.
13. Dixon, S.L. and Villar, H.O., *Bioactive diversity and screening library selection via affinity fingerprinting*, J. Chem. Inf. Comput. Sci., 38 (1998) 1192–1203.
14. Eaton, B.E., Gold, L. and Zichi, D.A., *Let's get specific: The relationship between specificity and affinity*, Chem. Biol., 2 (1995) 633–638.
15. Böhm, H.J. and Klebe, G., *What can we learn from molecular recognition in protein-ligand complexesfor the design of new drugs?*, Angew. Chem. Int. Ed. Engl., 35 (1996) 2588–2614.
16. Newton, C.G., *Molecular diversity in drug design. Application to high speed synthesis and high throughput screening*, In Dean, P.M. and Lewis, R.A. (Eds.), Molecular Diversity in Drug Design, Kluwer, Dordrecht, 1999, pp. 23–42.
17. Vajda, S., Weng, Z., Rosenfeld, R. and DeLisi, C., *Effect of conformational flexibility and solvation on receptor-ligand binding free energies*, Biochemistry, 33 (1994) 13977–13988.
18. Wang, J., Szewczuk, Z., Yue, S.Y., Tsuda, Y., Konishi, Y. and Purissima, E.O., *Calculation of relative binding free energies and configurational entropies: A structural and thermodynamic analysis of the nature of non-polar binding of thrombin inhibitors based on hirudin55-65*, J. Mol. Biol., 253 (1995) 473–493.
19. Vieth, M., Hirst, J.D. and Brooks III, C.L., *Do active site conformations of small ligands correspond to low free-energy solution structures?*, J. Comput. Aided Mol. Des., 12 (1998) 563–572.
20. Bostrom, J., Norrby, P.O., and Liljefors, T., *Conformational energy penalties of protein bound ligands*, J. Comput.-Aided Mol. Design, 12 (1998) 383–396.
21. Dixon, S.L. and Villar, H.O., *Investigation of classification methods for the prediction of activity in diverse chemical libraries*, J. Comput.-Aided Mol. Design, 13 (1999) 533–45.
22. MACCS-II Menu Reference, Version 2.2, MDL Information Systems, San Leandro, CA, 1994.
23. Johnson, M.A. and Maggiora, G.M., Concepts and Applications of Molecular Similarity, Wiley, New York, NY, 1990.
24. Lewis, R.A., Mason, J.S. and McLay, I., *Similarity measures for rational set selection and analysis of combinatorial libraries: The Diverse Property Derived approach*, J. Chem. Inf. Comput. Sci., 37 (1997) 599–614.
25. Mason, J.S., Morize, I., Menard, P.R., Cheney, D.L., Hulme, C. and Labaudiniere, R.F., *New 4-point pharmacophore method for molecular similarity and diversity applications, including a novel approach to the design of combinatorial libraries containing privileged substructures*, J. Med. Chem.,

42 (1999) 3251–3264.
26. Matter, H. and Potter, T., *Comparing 3D pharmacophore triplets and 2D fingerprints for selecting diverse compound subsets*, J. Chem. Inf. Comput. Sci., 39 (1999) 1211–1225.
27. Lipinski, C.A., Lombardo, F., Dominy, B.W. and Feeny, P.J., *Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings*, Adv. Drug Delivery Res., 23 (1997) 3–25.
28. Nogrady, T., Medicinal Chemistry: A Biochemical Approach, Oxford University Press, New York, NY, 1988.
29. Andrews, P.R., Craik, D.J. and Martin, J.L., *Functional group contributions to drug receptor interactions*, J. Med. Chem., 27 (1984) 1648–1657.
30. Kollman, P.A., *Drug-target binding forces*, In Wolff, M.E. (Ed.), Burger's Medicinal Chemistry and Drug Discovery, Vol. 1, Wiley, New York, NY, 1995, pp. 309–412.