J Math Chem (2011) 49:342–355
DOI 10.1007/s10910-010-9745-8

ORIGINAL PAPER

Application of Shannon-like diversity measures to cell-based chemistry spaces

Veerabahu Shanmugasundaram · Gerald M. Maggiora

Received: 3 December 2009 / Accepted: 27 September 2010 / Published online: 22 October 2010

Abstract The use of multi-dimensional "chemistry spaces" to represent large compound collections has become widespread in pharmaceutical research. In such spaces compounds are treated as points. Points in close proximity represent similar compounds, while distant points represent dissimilar compounds. Assessing the diversity of a compound collection, thus, is tantamount to characterizing the distribution of points in chemistry space. To facilitate many procedures such as selecting subsets of compounds for screening, for compound acquisition and designing combinatorial libraries, chemistry spaces have been partitioned into sets of non-overlapping, multi-dimensional cells, which are generated by dividing each axis into a number of equal-sized bins. This leads to a lattice of $(N_{bins})^{N_{\text{dim}}}$ cells, where N_{bins} is the number of bins on each axis and N_{dim} is the dimensionality of the space. One diversity measure that is typically used in cell-based chemistry spaces is identical in form to Shannon entropy, $D^{cpd}_{N_{cpd}}$ A normalized measure of this Shannon entropy given by, D^{cpd}_{rel} enables comparison between compound collections that occupy different number of occupied cells. Although D^{cpd}_{rel} characterizes the uniformity and "spreadout" of the corresponding compound collection, it treats cells as *positionally independent.* Some of the positional information lost can be recaptured by another diversity measure,

V. Shanmugasundaram (✉)
Anti Bacterials Chemistry/Discovery Technologies, World Wide Medicinal Chemistry,
Pfizer Pharma Therapeutics Research & Development, Groton 06340, USA
e-mail: Veerabahu.Shanmugasundaram@pfizer.com

G. M. Maggiora
Department of Pharmacology & Toxicology, University of Arizona,
College of Pharmacy, Tucson, AZ 85271, USA

V. Shanmugasundaram · G. M. Maggiora
Structural & Computational Chemistry, Pharmacia Corporation,
Kalamazoo, MI 49007, USA

which is also related in form to Shannon entropy. This new measure $D^{cell}_{N_{bin}}(\lambda)$ characterizes the distribution of occupied cells along each axis of chemistry space. The normalized measure $\langle D^{cell}_{rel} \rangle$ over all axes is given then by the average. Examples illustrating the applicability of these two Shannon-like measures to compound collections are presented.

Keywords Diversity measures · Chemistry space · Shannon entropy · Cell-based methods · Diversity assessment · BCUTs · Compound collections · Information theory

1 Introduction

In the early days of building corporate compound collections in the pharmaceutical industry in the 1990s, work focused on acquiring compounds from commercial vendors such that the compounds that were purchased were dissimilar to the compounds that already existed in the corporate collections [1,2]. Molecular diversity analysis provided a range of tools for exploring the extent to which sets of molecules overlapped with the corporate collection and compounds that were chosen for acquisition provided additional coverage of chemistry space that were inaccessible [3]. Later, compound sets were not only proactively selected for inclusion but were also "deselected" to remove those compounds that were considered to be undesirable. Typical deselection criteria contained chemical reactivity descriptors, some molecular property based filters and medicinal chemistry knowledge [4–6]. More recently, in silico filters are used for not only ruling out "undesirable" starting points but also some "unsuitable" starting points (e.g., "leadlike" screening sets). These additional selection filters provide some guiding principles for the type of properties compounds should possess for ensuring successful drug discovery programs [7,8].

However, molecular diversity principles still underlies many approaches to compound selection, screening strategies, comparing compound collections, compound acquisition and design of combinatorial libraries. Many different diversity methods have been described in the literature and new methods continue to appear [9,10]. A wide range of such indices have been reported and discussed in many reviews [11]. Many of the conclusions that were drawn earlier regarding similarity indices are equally applicable to diversity indices, in that the latter involve knowledge of the degree of dissimilarity or distance between pairs of compounds [12].

Finding hits in a drug discovery program is inherently a sequential searching and screening process. Even a large screening campaign where hundreds of thousands or millions of compounds are screened is a modest sampling of the many millions and billions of compounds that are available and/or that could be synthesized. An exhaustive search of the vast chemistry space is not feasible and drug discovery like many other optimization procedures is a sequence of steps in hopefully the right direction. Almost always drug discovery begins with an initial screen or an initial "hit", resulting in further screening or possibly assessing newly synthesized compounds, expanding around the initial hits. A number of companies are pursuing the sequential screening

paradigm where the first step is a small initial screen that is used to mine the remainder of the collection through several iterations of searching and screening [13–15].

There are two distinct scenarios in constructing sets of compounds for screening. The first corresponds to a situation where one does not have any specific target in mind at the time when the set is constructed (ie., a subset of the entire compound collection for any target for a HTS). The second situation arises when constructing a set of compounds for a specific target (or a group of related targets) such as kinases, GPCRs, NHRs etc. In both situations, some common underlying molecular diversity principles are employed based on the molecular structure and pharmacophoric information that is available a priori.

There are several objectives in a sequential screen—some researchers want to find "all active compounds" in their collection; some would like to find "all active chemical classes" in the collection; others have a goal of finding "several active classes" with favorable properties to launch a drug discovery campaign. The follow-up process of drug discovery is very expensive that even large companies cannot afford to follow up on more than two or three chemical classes for a target. Sequential screening therefore typically matches downstream capacity.

The physicochemical properties and the biological activity of a compound are ultimately determined by its molecular structure. The principal aim of molecular diversity analysis is to identify structurally diverse sets of compounds that can be tested for biological activity (with the assumption that evaluating a structurally diverse set of compounds will generate more structure-activity information than a random set of compounds). However, this is subject to the nature of the activity landscape associated with a given assay. For many years it has been assumed that similar molecules tend to have similar activities [16], leading to activity landscapes comparable to the gently rolling hills found on the Kansas prairie [17,18]. Mounting evidence suggests, however, that this picture is not as universal as once thought but is in many cases rather more like the rugged landscapes of Utah's Bryce Canyon. This new topographical metaphor clearly implies that very similar molecules may in some cases possess very different activities leading to what can be called *activity cliffs*. An activity cliff is defined by the ratio of the difference in activity of two compounds to their "distance" of separation in a given chemical space.

The existence of such activity cliffs is not entirely surprising since molecular recognition plays a crucial role in determining activity. The greater prevalence of activity cliffs than was earlier suspected has several important and related implications for sampling chemistry space to infer the underlying biological landscape. First, a uniform sampling of chemistry space might provide adequate coverage for follow-up screening if the underlying biological activity landscape were smooth like the Kansas prairie but might not provide ideal starting points if the underlying activity landscape were rugged like the Bryce canyon. Second, outliers in the data may not be due to statistical fluctuations or due to measurement errors but rather may reflect the presence of activity cliffs. Thus, perfectly valid data points located in cliff regions may *appear* to be outliers. Third, the presence of activity cliffs requires the assay of additional compounds in the neighborhoods around these cliffs to ensure that activity landscapes are adequately represented in these rapidly varying regions. Finally, all of this has

implications to the type of diversity metrics used to characterize molecular diversities of compound collections, sampling strategies and selection of global diversity subsets.

Quantifying diversity in chemistry space is to maximize diversity in biological space where one would ideally like to capture both the extent of coverage of chemistry space (uniformity or "spreadout") as well as the underlying density of the distribution (sampling) of compounds. Here we describe two different indices derived from the formulation of the Shannon entropy that take into account both the density and diversity of coverage of chemistry space. The procedure, which was developed and employed at Pharmacia takes account of the multiplicity of chemistry spaces in a way that the underlying framework can be generalized to any cell-based chemistry space representation. The results reported here, which were gathered over a period of time, tend to indicate that adequate sampling of chemistry space is necessary for different types of problems that reflect the underlying activity landscape. In addition, the diversity assessment using these indices of cell-based chemistry spaces is quite rapid and straightforward even for huge compound collections.

2 Methods

2.1 General

Pre-filtering of databases, if performed, were based on sub-structural filters or property filters in tune with our published compound purchasing strategy [19]. Only the largest molecular component, which effectively eliminated counterions, was considered for each compound in the database. DiverseSolutions and MOE software were run on a Silicon Graphics Origin 200 workstation (R12000 processor), running under the IRIX operating system (version 6.5.5). JMP 5.0 (SAS Institute Inc.) was run on a IBM ThinkPad T23 running under Windows 2000. Computations needed to generate entropies were coded with awk scripting language.

2.2 Compound collections

Four compound collections were analyzed in this study, (1) A filtered subset of ACD (Available Chemicals Directory) database, which contains about 126,000 commercially available compounds (2) A filtered subset of the DSDF (Derwent Standard Drug File), which contains about 25,000 primarily medicinal, pharmacological, and other biologically active compounds (3) and (4) Two proprietary combinatorial libraries (Library A and Library B) containing about 2000 compounds in each library. These were derived from two different chemical templates and were pursued by two different therapeutic area project teams. General molecular property distributions of these four collections are shown in Table 1.

2.3 Chemistry space

BCUTs are a set of molecular descriptors that were developed by Prof. Robert Pearlman at the University of Texas, Austin [20,21]. These descriptors that are computed

Table 1 Summary characteristics of compound collections

Compound collection	Number of compounds	Molecular weight	Number of rotatable bonds	Number of hydrogen bond acceptors	Number of hydrogen bond donors
ACD-F[a]	126,829	276.9(99.3)	4.8(3.6)	3.2(2.1)	1.0(1.2)
DSDF[b]	25,335	351.1(237.9)	6.5(7.7)	4.3(4.7)	2.0(3.0)
LIBRARY-A[c]	1,947	365.1(85.5)	5.6(2.7)	2.9(1.6)	2.1(0.9)
LIBRARY-B[c]	2,249	389.7(82.3)	7.2(2.4)	5.7(1.6)	1.5(0.7)

The figures quoted (to one decimal place) are means and standard deviations (in parenthesis) when averaged over all of the molecules

[a] Filtered library of 126,829 compounds from the Available Chemicals Directory database

[b] Derwent Standard Drug File

[c] Two hypothetical combi-chem libraries

Table 2 The six 3-D BCUTs which best represent the structural diversity of compounds contained in the datasets

BCUT1:	Bcut_gastchrg_S_invdist2_001.250_R_L
BCUT2:	Bcut_gastchrg_S_invdist6_000.600_R_H
BCUT3:	Bcut_haccept_S_invdist_000.600_R_H
BCUT4:	Bcut_hdonor_S_invdist2_001.200_R_H
BCUT5:	Bcut_tabpolar_S_invdist2_001.000_R_L
BCUT6:	Bcut_tabpolar_S_invdist_000.500_R_H

using the structure of the compounds, capture the electrostatic, hydrogen-bonding abilities and polarizabilities (hydrophobic) nature of the molecules. Since these features are critical to the ligand-receptor interactions, BCUT descriptors in a way capture the essential ligand-receptor interactions. Defined in a manner, which incorporates both connectivity information based on actual bonding or inter-atomic distances through space and atomic properties relevant to inter-molecular interactions BCUTs have been repeatedly shown to be useful as descriptors to describe a chemistry space [22,23]. To facilitate many procedures such as selecting subsets of compounds for screening and for compound acquisition, BCUT chemistry spaces have been partitioned into sets of non-overlapping, multi-dimensional cells, which are generated by dividing each axis into a number of equal-sized bins. This leads to a lattice of $(N_{bins})^{N_{\mathrm{dim}}}$ cells, where N_{bins} is the number of bins on each axis and N_{dim} is the dimensionality of the space.

A set of 29 3-D hydrogen suppressed standard BCUT descriptors were computed for all compound collections using the program DiverseSolutions (DVS 5.0.0). 3-D BCUT chemistry space, that best represent the union of the four datasets was generated using the χ^2 algorithm implemented in DVS was used as reference for all molecular diversity analysis. Table 2 describes the 6-dimensional 3D BCUT chemistry space used here.

2.4 Chemistry space statistics

Typically compound collections are not uniformly distributed in BCUT chemistry spaces. Pearlman and Smith recommend that choosing the number of bins/axis that

Table 3 Cell population statistics

Collection	Number of compounds	Number of occupied cells	% Occupancy	Average occupied cell population	Highest cell population
ACD-F	126,829	15,090	12.8	8.40(7.45)	508
DSDF	25,335	7,002	6.0	3.62(2.78)	56
LIBRARY A	1,947	315	0.3	6.18(4.78)	116
LIBRARY B	2,249	228	0.2	9.86(14.65)	524

yields roughly 12–16% occupancy provides an appropriate level of resolution for most applications [20,21]. Accordingly the 3-D BCUT chemistry space coordinates were sliced into 7 bins/coordinate for the union set. This yielded 7^6 cells in the 3-D BCUT chemistry space. Only 12.8% of these cells were occupied by the filtered ACD database with an average number of about 8 compounds in each occupied cell. The highest occupied cell contained 508 ACD compounds. The chemistry space cell population statistics are given in Table 3.

2.5 Information and Shannon entropy

Originally developed for applications in digital communication, Shannon entropy can be derived from the concept of information in the following way [24,25]. Information, sometimes called 'surprisal,' is defined, in units of 'bits,' as

$$\mathcal{I}(\boldsymbol{A}_i) = \log_2 \frac{1}{\mathcal{P}(\boldsymbol{A}_i)}$$

where $\mathcal{P}(\boldsymbol{A}_i)$ is the probability of observing an object from subset $\boldsymbol{A}_i \in \boldsymbol{A}$,

$$\mathcal{P}(\boldsymbol{A}_i) = \frac{|\boldsymbol{A}_i|}{|\boldsymbol{M}|} = \frac{n_{\boldsymbol{A}_i}}{n}$$

In addition,

$$\sum_{\boldsymbol{A}_i \in \boldsymbol{A}} \mathcal{P}(\boldsymbol{A}_i) = 1, \quad \text{and} \quad \mathcal{P}(A_i) \geq 0, \quad \text{for} \quad i = 1, 2, \ldots, N_{\boldsymbol{A}}$$

This equation makes sense from the following point of view, namely, the more likely an event is to be observed the less information will be obtained upon observing it, that is there is less 'surprise' in observing the event.

Shannon entropy is then defined as the expectation value of the information,

$$\begin{aligned}\mathcal{H}(\boldsymbol{A}) &= \langle \mathcal{I}(\boldsymbol{A}_i)\rangle_{\boldsymbol{A}} \\ &= \sum_{\boldsymbol{A}_i \in \boldsymbol{A}} \mathcal{P}(\boldsymbol{A}_i) \log_2 \frac{1}{\mathcal{P}(\boldsymbol{A}_i)} \\ &= -\sum_{\boldsymbol{A}_i \in \boldsymbol{A}} \mathcal{P}(\boldsymbol{A}_i) \log_2 \mathcal{P}(\boldsymbol{A}_i)\end{aligned}$$

It can also be shown that the maximum value of $\mathcal{H}(\boldsymbol{A})$ occurs when all of the equivalence classes are occupied equally, that is $|\boldsymbol{A}_1| = |\boldsymbol{A}_2| = \cdots = \left|\boldsymbol{A}_{N_A}\right| = \bar{n}_{\boldsymbol{A}}$, while the minimum occurs when all of the elements of the set reside in a single subset, $\mathcal{P}(\boldsymbol{A}_i) = 1 \rightarrow \mathcal{H}(\boldsymbol{A}) = 0$, so that

$$0 \leq \mathcal{H}(\boldsymbol{A}) \leq \mathcal{H}_{\max}(\boldsymbol{A}) = \log_2 \bar{N}_{\boldsymbol{A}}$$

where

$$\bar{N}_{\boldsymbol{A}} = n/\bar{n}_{\boldsymbol{A}}$$

If all of the elements are unique $\bar{n}_{\boldsymbol{A}} = 1$ and thus $\bar{N}_{\boldsymbol{A}} = n$ and $\mathcal{H}_{\max}(\boldsymbol{A}) = \log_2 n$.

3 Results and discussion

3.1 Shannon-like diversity index

Let us consider three hypothetical 16 compound libraries A, B and C in a 2-D chemistry space, divided into 4 bins/axis and hence partitioned into a lattice of 16 cells. All three libraries occupy 4 cells. Hence, the coverage of chemistry space (percent of cells occupied) is $(4/16) \times 100 = 25\%$. Libraries A and C occupy 4 cells, with 4 compounds in each cell. However, they differ in the position of occupied cells in chemistry space. Library A and B have the same positional information but differ in the distribution of cell occupancies. Library B has 13 compounds in a single cell and 1 compound in the remaining 3 cells, whereas Library A has 4 compounds in each cell. In addition, Libraries A and B span the entire range of chemistry space along each coordinate (i.e., occupy every bin along each coordinate) whereas Library C covers only a small corner of chemistry space.

A Shannon-like diversity measure could be defined in such a way that it measures the uniformity of the distribution of fractional cell occupancies in chemistry space. Hence, libraries with different numbers of compounds in each cell can be distinguished from each other (e.g., Library A and C from Library B).

Table 4 Distribution of fractional cell occupancies in chemistry space for hypothetical libraries A–C

Library	N_{cpd}	N_{cell}	Average occupied cell population	$D^{cpd}_{N_{cpd}}$
A	16	4	4	2.00
B	16	4	4	0.99
C	16	4	4	2.00

This measure of the uniformity of the distribution of fractional cell occupancies in chemistry space is given in equation (1) below,

$$D^{cpd}_{N_{cpd}} = -\sum_{i=1}^{N_{cell}} f_i \log_2 f_i \tag{1}$$

where $D^{cpd}_{N_{cpd}}$ measures the uniformity of the distribution of fractional cell occupancies in chemistry space, N_{cell} is the number of occupied cells, N_i is the number of compounds in the i-th occupied cell, N_{cpd} is the total number of compounds in the collection given by $N_{cpd} = \sum_{i=1}^{N_{cell}} N_i$ where f_i is the fractional occupancy of the i-th cell, where $f_i = N_i/N_{cpd}$ and $\sum_{i=1}^{N_{cell}} f_i = 1$

Table 4 illustrates the results of computing the Shannon-like diversity measure using equation (1) for the three hypothetical libraries discussed above. Libraries A and C that have a more uniform fractional cell occupancy distribution have measures with greater values ($D^{cpd}_{N_{cpd}}$ equals 2.0 for Libraries A and C compared to a value of 0.99 for Library B).

A second Shannon-like diversity measure can be used to capture some of the positional information that is lost by measuring the uniformity of the marginal distribution of occupied cells along each coordinate that defines the chemistry space. Here libraries that have an even distribution of occupied cells along each coordinate, covering the entire range of chemistry space can be differentiated from libraries that don't (e.g., Libraries A and B from Library C).

This measure of the uniformity of the distribution of occupied cells in chemistry space is given in equation (2) below,

$$D^{cell}_{N_{bin}}(\lambda) = -\sum_{i=1}^{N_{bin}} h_i(\lambda) \log_2 h_i(\lambda) \tag{2}$$

where $D^{cell}_{N_{bin}}$ measures the uniformity of the distribution of occupied cells along each dimension λ, λ refers to a specific dimension in chemistry space, N_{bin} is the number of occupied bins, $h_i(\lambda)$ is the fraction of occupied cells associated with the i-th bin of the λ-th axis, where $h_i(\lambda) = N_i(\lambda)/N_{cell}$ $N_i(\lambda)$ is the number of occupied cells in the i-th bin the λ-th axis, N_{cell} is the total number of occupied cells.

Normalized measures D^{cpd}_{rel} and $\langle D^{cell}_{rel} \rangle$ given in equations (3) and (4) enable comparison between compound collections of different sizes, where N_{dim} refers to the

Table 5 Shannon-like diversity measures for hypothetical libraries A-C

Library	Number of compounds	% Coverage of chemistry space	D_{rel}^{cpd}	$\left\langle D_{rel}^{cell} \right\rangle$
A	16	25	1.0	2.0
B	16	25	0.495	2.0
C	16	25	1.0	1.0

dimensionality of the chemistry space

$$D_{rel}^{cpd} = (\log_2 N_{cell})^{-1} D_{N_{cpd}}^{cpd} \tag{3}$$

$$\left\langle D_{rel}^{cell} \right\rangle = N_{\text{dim}}^{-1} \sum_{\lambda=1}^{N_{\text{dim}}} D_{N_{bin}}^{cell}(\lambda) \tag{4}$$

Consider Library A which contains 16 green compounds that span the entire range of chemistry space and Library C which contains 16 red compounds that does not span the entire range of chemistry space. From the previous exercise we noted that $D_{N_{cpd}}^{cpd}$ equals 2.0 for both libraries. Normalizing this value using equation (3) yields D_{rel}^{cpd} which equals 1.0 for both libraries A and C. This means that both libraries A and C have uniform distribution of fractional cell occupancies. However the libraries differ in their respective values of $\left\langle D_{rel}^{cell} \right\rangle$. Library A has a value of 2.0 and Library C has a value of 1.0. This means that Library A has a more uniform distribution of occupied cells compared to Library C and hence is more diverse (Table 5).

Comparison of D_{rel}^{cpd} and $\left\langle D_{rel}^{cell} \right\rangle$ the normalized Shannon-like diversity measures when used in conjunction, enables us to distinguish between libraries A, B and C. A greater value of D_{rel}^{cpd} indicates a more uniform distribution of fractional cell occupancies and a greater value of $\left\langle D_{rel}^{cell} \right\rangle$ reflects greater coverage of chemistry space and a more uniform distribution of occupied cells in chemistry space.

3.2 Comparing compound collections

Let us now consider the four compound collections (ACD, DSDF, Library A and Library B) and estimate the uniformity of the distribution of fractional cell occupancies and the uniformity of the distribution of occupied cells in BCUT chemistry space. The results are shown in (Table 6; Figs. 1, 2, 3) and images of slices of the 6-dimensional BCUT chemistry space is shown in Fig. 4.

Comparing ACD and DSDF collections, we note that ACD has more compounds (about 3 times the size) than DSDF. ACD also covers a larger fraction of chemistry space than DSDF (12 vs. 6%). However, DSDF is spread more uniformly over the 6% of chemistry space it covers (ie., there is fewer clumping of compounds, indicated by a higher D_{rel}^{cpd} value 0.93 vs. 0.88) and also effectively covers chemistry space better ($\left\langle D_{rel}^{cell} \right\rangle$ equals 0.88 for DSDF versus 0.85 for ACD). This shows that although

Table 6 Shannon-like diversity measures

Collection	Number of compounds	% Coverage of chemistry space	Distribution of cell occupancies		Distribution of occupied cells (Effective coverage)	
			$D^{cpd}_{N_{cpd}}$	D^{cpd}_{rel}	$D^{cell}_{N_{bin}}$	$\langle D^{cell}_{rel} \rangle$
ACD-F	126,829	12.8	12.26	0.88	2.39	0.85
DSDF	25,335	6.0	11.91	0.93	2.47	0.88
LIBRARY A	1,947	0.3	7.04	0.85	1.52	0.54
LIBRARY B	2,249	0.2	4.94	0.63	1.09	0.39

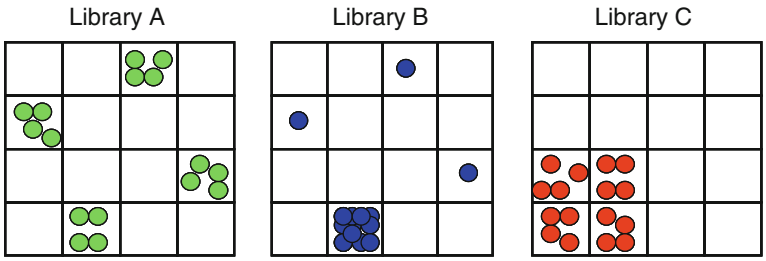

Fig. 1 Hypothetical Libraries—Library A; Library B; Library C (16 compounds each)

DSDF is much smaller in size, it effectively is a more diverse collection than ACD. Depending on the type of underlying biological activity landscape (smooth or rugged) sampling decisions will need to be made for ACD and DSDF. ACD would suit a more rugged activity landscape better than DSDF.

Comparing Libraries A and B we again note that Library A although a bit smaller in size (about 1,900 compounds versus 2,200 compounds) is more diverse by both measures (D^{cpd}_{rel} value of 0.85–0.63 and $\langle D^{cell}_{rel} \rangle$ value of 0.54–0.39)

The Shannon-like diversity measures D^{cpd}_{rel} and $\langle D^{cell}_{rel} \rangle$ for the DSDF collection are greater than all the other three collections. This indicates that the DSDF collection has a more uniform distribution of cell occupancies and occupied cells in BCUT chemistry space. Libraries A and B which are both similar in size and occupy smaller region of chemistry space against the backcloth of ACD and DSDF. However, when compared to each other Libraries A and B seem comparable in their distribution of cell occupancies and occupied cells.

Varying the resolution of the chemistry space from 3 bins/axis to 12 bins/axis does not change the rank ordering of the diversities of the four compound collections. Effects of changes in the resolution of the chemistry space (binning) on the Shannon-like diversity measures can be identified using the sensitivity plots as shown in Fig. 5.

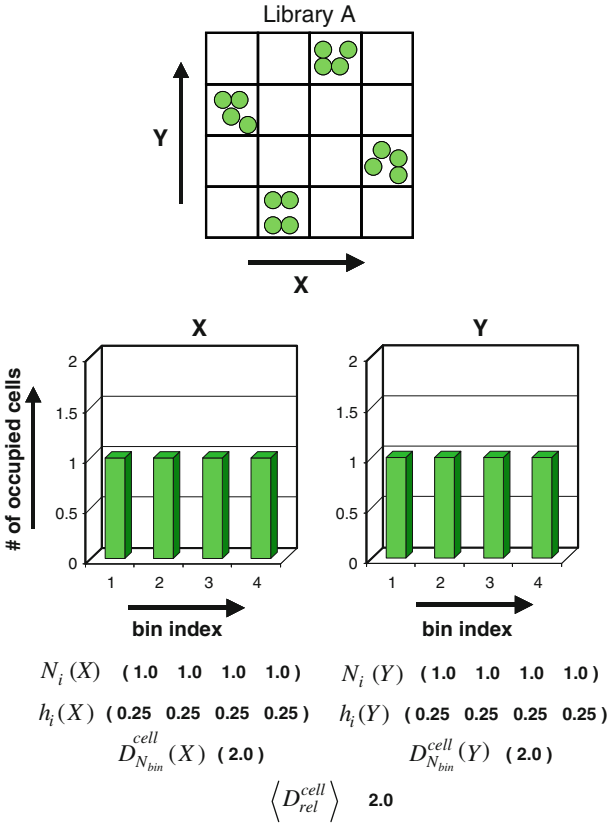

Fig. 2 Distribution of occupied cells: Hypothetical Library A

4 Summary and conclusions

The present work describes applications of Shannon-like diversity measures to cell-based chemistry spaces. A persistent issue associated with chemistry space sampling is the underlying variance of the activity landscape and the lack of invariance or the dependence on the chemistry-space representation used, which leads to a lot of subjectivity in the field. An important consequence of this lack of invariance is that NN relationships are not generally preserved, that is two compounds that are NNs in one chemistry-space representation may not even be close in another. While can be a bit unsettling, several researchers in the field have employed this lack of invariance to advantage through the use of orthogonal HDNN searches of multiple chemistry-space representations of the same compound collection to alleviate the sampling issues in a rugged activity landscape [26]. The results described in this work clearly show that using Shannon-like diversity measures one could measure both the diversity and density of the distribution of compound collections in chemistry space (ie., coverage and sampling could aggregated into two distinct Shannon-like diversity indices). The major advantage of using Shannon entropy as a metric for molecular diversity

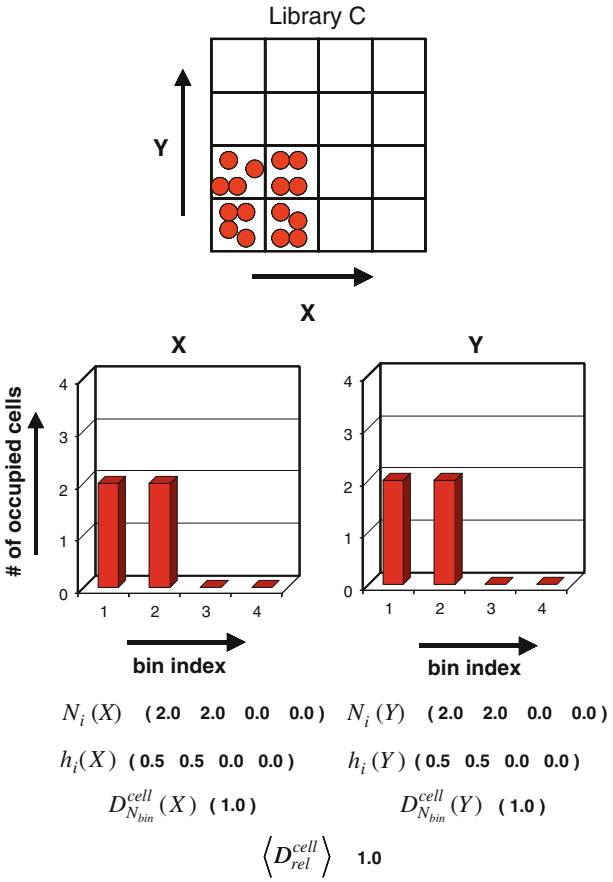

Fig. 3 Distribution of occupied cells: Hypothetical Library C

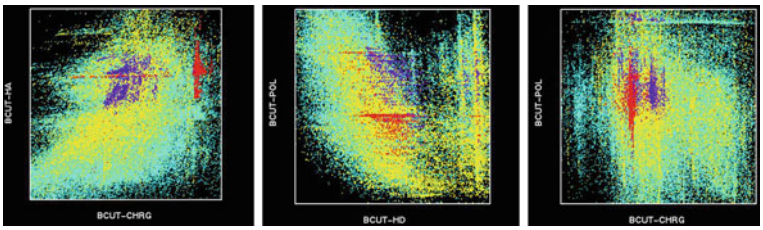

Fig. 4 These pictures capture the overlap of ACD-F (*cyan*), DSDF (*yellow*), LIBRARY-A (*purple*) and LIBRARY-B (*red*) compound collections in the 6-D BCUT chemistry space

over other distance-based metrics is that it is probabilistic, allowing the distribution of chemical properties for a compound collection to be taken into account. Moreover, because the cell based nature of the chemistry space representation, any orthogonal set of descriptors could be used to define the chemistry space. Importantly, there is also a significant lack of sensitivity to number of bins used to partition the chemistry space.

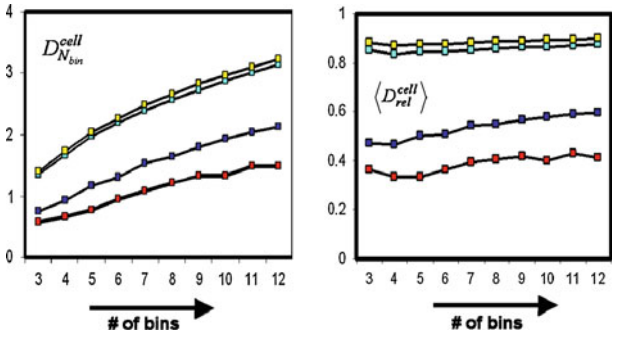

Fig. 5 These pictures illustrate the lack of sensitivity of the Shannon-like diversity measures -$D^{cell}_{N_{bin}l}$ and $\left\langle D^{cell}_{rel} \right\rangle$ to changes in the resolution of the chemistry space (i.e., the number of bins/axis) ACD-F (*cyan*), DSDF (*yellow*), LIBRARY-A (*purple*) and LIBRARY-B (*red*)

The Shannon-like diversity measures are nearly constant and don't change in the rank ordering of the overall diversity of the collections. Thus it appears, at least from a pragmatic viewpoint, that the procedure described here, provides a novel, quick and effective method of measuring and comparing diversities of compound collections.

References

1. Y.C. Martin, J. Comb. Chem. **3**, 231–250 (2001)
2. P. Willett, *Similarity and Clustering in Chemical Information Systems* (Research Studies Press, Letchworth, 1987)
3. M.S. Lajiness, Perspect. Drug Discovery Des. **7/8**, 65–84 (1997)
4. P.S. Charifson, W.P. Walters, J. Comput.-Aided Mol. Design. **16**, 311–323 (2002)
5. G.M. Rishton, Drug. Discov. Today. **2**, 382–384 (1997)
6. M.S. Lajiness, G.M. Maggiora, V. Shanmugasundaram, J. Med. Chem. **47**, 4891–4896 (2004)
7. T.I. Oprea, A.M. Davis, S.J. Teague, P.D. Leeson, J. Chem. Inf. Comput. Sci. **41**, 1308–1315 (2001)
8. M.S. Lajiness, V. Shanmugasundaram, in *Methods in molecular biology, vol. 275*, ed. by J. Bajorath Chemoinformatics: Concepts, Methods and Tools for Drug Discovery (Humana Press, Totawa New Jersey, 2004), pp. 111–129
9. M.F.M. Engels, A.C. Gibbs, E.P. Jaeger, D. Verbinnen, V.S. Lobanov, D.K. Agrafiotis, J. Chem. Inf. Model **46**, 2651–2660 (2006)
10. J.J. Perez, Chem. Soc. Rev. **34**, 143–152 (2005)
11. M. Waldman, H. Li, M. Hassan, J. Mol. Graph. Model. **18**, 412–426 (2000)
12. G.M. Maggiora, V. Shanmugasundaram, in *Chemoinformatics: Concepts, Methods and Tools for Drug Discovery,* Ed. by J. Bajorath. Methods in Molecular Biology, vol. **275** (Humana Press: Totawa, NJ, 2004), pp 1–50.
13. V. Shanmugasundaram, G.M. Maggiora, M.S. Lajiness, J. Med. Chem. **48**, 240–248 (2005)
14. R.P. Sheridan, S.K. Kearsley, Drug Discovery Today **7**, 903–911 (2002)
15. J. Auer, J. Bajorath, J. Chem. Inf. Model. **46**, 2502–2514 (2006)
16. Y.C. Martin, J.L. Kofron, L.M. Traphagen, J. Med. Chem. **45**, 4350–4358 (2002)
17. G.M. Maggiora, J. Chem. Inf. Model. **46**, 1535 (2006)
18. V. Shanmugasundaram, G.M. Maggiora, Abstracts of Papers, 222nd ACS National Meeting, Chicago, IL, United States, August 26-30, 2001 CINF-032
19. G.M. Maggiora, V. Shanmugasundaram, M.S. Lajiness, T.N. Doman, M.W. Schulz, ed. by T. Oprea Cheminformatics Aspects in Drug Discovery (Wiley-VCH, New York, 2004), pp. 317–332
20. R.S. Pearlman, K.M. Smith, Perspect. Drug Discovery Des. **9**, 339–353 (1998)
21. R.S. Pearlman, *Diverse Solutions User's Manual* (University of Texas, Austin, TX, 1995)

22. D.J. Schnur, Chem.Inf. Comput. Sci. **39**, 36–45 (1999)
23. J.S. Mason, S.D. Pickett, Perspect. Drug Discovery Des. **7/8**, 85–114 (1997)
24. C. E. Shannon, W. Weaver, *The Mathematical Theory of Communication* (University of Illinois Press, Urbana, IL, 1963)
25. G.M. Maggiora, V. Shanmugasundaram, J. Math. Chem. **38**, 1–20 (2005)
26. M. Jalaie, V. Shanmugasundaram, Mini. Rev. Med. Chem. **6**, 1159–1167 (2006)