Journal of Mathematical Chemistry Vol. 38, No. 1, July 2005 (© 2005)
DOI: 10.1007/s10910-005-4525-6

An information-theoretic characterization of partitioned property spaces

Gerald M. Maggiora*

College of Pharmacy and BIO5 Institute, University of Arizona, Tucson, AZ 85721, USA
E-mail: maggiora@pharamcy.arizona.edu

Veerabahu Shanmugasundaram

Computer-Assisted Drug Discovery, Pfizer Global Research & Development, Ann Arbor, MI 48105, USA

Received 29 December 2004; revised 20 January 2005

A methodology, derived by analogy to Shannon's information-theoretic theory of communication and utilizing the concept of mutual information, has been developed to characterize partitioned property spaces. A family of non-intersecting subsets that cover the "universe" of objects represents a partitioned property space. Each subset is thus an equivalence class. A partition and it's associated equivalence classes can be generated using any one of a number of procedures including hierarchical and non-hierarchical clustering, direct approaches using rough set methods, and cell-based partitioning, to name a few. Thus, partitioned property spaces arise in many instances and represent a very large class of problems. The approach is based on set-valued mappings from equivalence classes in one partition to those in another and provides a coarse-grained means for comparing property spaces. From these mappings it is possible to compute a number of Shannon entropies that afford calculation of mutual information, which represents that amount of information shared by two partitions of a set of objects. Taking the ratio of the mutual information with the maximum possible mutual information yields a quantity that measures the similarity of the two partitions. While the focus in this work is directed towards small sets of objects the approach can be extended to many more classes of problems that can be put into a similar form, which includes many types of cheminformatic and biological problems. A number of scenarios are presented that illustrate the concept and indicate the broader class of problems that can be handled by this method.

KEY WORDS: information theory, mutual information, partitions, property spaces, Shannon entropy

* Corresponding author.

1. Introduction

The notion of a "property space" provides a rich conceptual framework for understanding the relationships of objects to one another, although the general term 'object' in most cases of interest in this paper is synonymous with 'molecule.' The use of property spaces may be less important for handling cases involving relatively small numbers of objects, but they are crucial when dealing with large numbers objects such as found in combinatorial-chemistry libraries and corporate compound collections. The nature of a given property space is related to the way in which objects are "represented" in the space. This generally is a multi-dimensional vector [1] whose components, called descriptors or attributes, characterize the relevant properties or features of the objects under study. Each object can then be viewed as a point in property space, and each descriptor corresponds to a coordinate in the multi-dimensional space. Note that even in coordinate-free cases where, for example, the relationship amongst objects is given by their similarities, appropriate coordinate systems can be constructed using techniques such as multi-dimensional scaling or non-linear mapping [2–5]. Distance between objects is usually defined by some type of distance function (e.g., Euclidean distance, city-block distance, etc.). Thus, property spaces are metric spaces. It is important to remember that different data representations lead to different property spaces and that the relationships among objects in one property space are not necessarily preserved in another property space [6]–a *universal, intrinsic property space does not exist*. Such a view of property space, which is based on point-to-point mappings, can be considered a fine-grained view. This has important consequences regarding the distribution of objects in a property space. It is entirely possible that clusters of objects in one property space may become uniformly spread out in another property space and vice-versa. In addition, nearest–neighbor relationships may also be lost. The outcome of such traumatic changes can lead to significant problems in, for example, dissimilarity-based sampling procedures [7].

Since the objects we are concerned with are discrete entities, and since their number can be very large but finite, property spaces are discrete, even though the coordinates describing an object's position are continuous. Moreover, in general property spaces are sparse, and the distribution of objects within them is generally non-uniform – objects derived from chemical and biological systems tend to occur in clusters much like galaxies within the universe. And like galaxies there is generally considerable empty space even within clusters of objects.

Considerable work has been carried out over the last 15 or so years to elucidate and study the cluster structure of property spaces in cheminformatics [8,9], especially after the introduction of combinatorial-chemistry methods in the early nineties. Clustering is just one way to carve up a property space into a set of non-interacting subsets that *cover* the space [10,11]. In set-theoretical language such a decomposition is called a *partition*. There is a one-to-one correspondence

between the non-intersecting subsets of a partition and set-theoretic equivalence classes, the latter being generated by some form of set-theoretic relation [12]. Thus, a cluster is also an equivalence class [13], which implies that the objects in a cluster can be interpreted as being, in some sense, equivalent. In addition to the usual clustering algorithms, algorithms exist such as those utilized in rough set theory [14] to determine the equivalence classes and thus partitions of a set of objects. Partitioning property spaces can also be accomplished by dividing them into non-overlapping cells, which are generally taken to be hypercubes, but other non-overlapping partitions are also possible although not necessarily easy to construct for hyper-dimensional property spaces [15]. Thus, such cell-based property spaces are also partitions in the strict mathematical sense, and each non-intersecting subset is perforce an equivalence class. Whether the objects in these hypercubic equivalence classes are in any deep chemical sense equivalent is debatable. Nevertheless, in a strict mathematical sense they are.

Another important feature of a partition is that geometric relationship of its non-intersecting subsets to one another is generally lost, except for certain clustering procedures such a k-means clustering [16] where the location of cluster centroids is preserved. However, such clustering methods generally require specification of the number of clusters, which seriously biases the nature of the clusters.

Because property space is representation dependent (*vide supra*) the partitioning of a set of objects in one representation, by whatever means, will in general differ from the partitioning of that same set of objects represented differently but produced by the same procedure. Alternatively, the same set of objects represented identically but partitioned by two different procedures will also generally result in different partitions. This raises the important question as to how similar are the different partitions. The fundamental approach used here to deal with that important question is based by analogy on Shannon's information-theoretic approach to communication [17]. Specifically, the analogy is drawn between symbols transmitted and received and mappings of objects from an equivalence class in one partitioning to an equivalence class in another partitioning – so-called set-valued mappings. Thus, there is a one-to-one correspondence between symbols and equivalent classes – all identical symbols belong, of course, to the same equivalence class. In contrast to the point-to-point mappings discussed earlier, which are the basis for a fine-grained view of property space, the set-valued mappings between the subsets of different partitions dealt with in this work provide a *coarse-grained view*. Nevertheless, as will be seen in the sequel, even such a coarse-grained view can provide a useful means for characterizing partitioned property spaces.

The key concept used here is that of mutual information, which arises naturally out of the information-theoretic framework erected by Shannon and is becoming of growing importance in a number of fields [18–23]. Yockey has described a procedure analogous to that presented here for determining the similarity of gene and protein sequences [24].

A discussion of the basic methodologies needed to develop the theory as applied to partitioned property spaces is presented. These methodologies include set partitions, set-valued mappings, information and its relationship to Shannon entropy, and mutual information, the seminal concept, which underlies this work. Several simple examples are provided that illustrate the various theoretical points discussed and indicate how the information-theoretic methodology can be applied to many problems of interest. The important question of similarity-induced partitions of property space and the relationships among such spaces from a coarse-grained perspective is also addressed, and several simple examples are presented to illustrate the concepts.

2. Methodological considerations

2.1. Partitioned property spaces

There are many ways to generate subsets of objects; the focus here will be on *partitions* defined loosely as families of non-intersecting subsets. Consider the set of n objects

$$\mathbf{M} = \{m_1, m_2, \ldots, m_i, \ldots, m_n\}, \tag{1}$$

where n may be large (e.g., 500,000), and the set of relations

$$\mathbf{R} = \{R_A, R_B, R_C, \ldots\}. \tag{2}$$

The partition of $\mathbf{M}$ generated by the relation R_A is given as, for example,

$$\mathbf{M} \overset{R_A}{\Rightarrow} \mathbf{A} = \{\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_{N_A}\}, \tag{3}$$

where each subset $\mathbf{A}_i$ contains $n_{\mathbf{A}_i}$ objects

$$\mathbf{A}_i = \{m_1^{\mathbf{A}_i}, m_2^{\mathbf{A}_i}, \ldots, m_{n_{\mathbf{A}_i}}^{\mathbf{A}_i}\} \tag{4}$$

and

$$\mathbf{M} = \bigcup_{\mathbf{A}_i \in \mathbf{A}} \mathbf{A}_i \quad \text{and} \quad \mathbf{A}_i \cap \mathbf{A}_j = \emptyset \quad \text{for all } i, j, \tag{5}$$

that is the subsets $\mathbf{A}_i \in \mathbf{A}$ *cover* $\mathbf{M}$ and are non-intersecting. Moreover, each subset constitutes an equivalence class [12]. Thus, a given set of objects can be described by a family of equivalence classes generated by a given equivalence relation. The *measure* or size of, that is the number of elements in, each of the various sets and subsets is given by

$$|\mathbf{M}| = n = \sum_{\mathbf{A}_i \in \mathbf{A}} |\mathbf{A}_i| = \sum_{i=1}^{N_\mathbf{A}} n_{\mathbf{A}_i} \quad \text{and} \quad |\mathbf{A}| = N_\mathbf{A}. \tag{6}$$

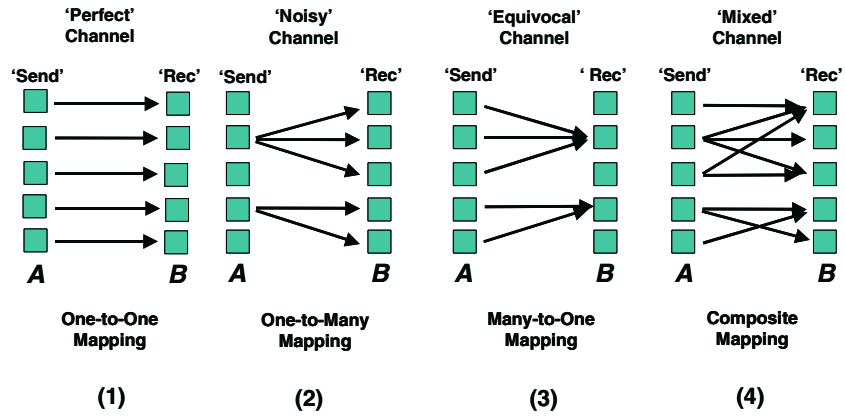

Figure 1. A scheme depicting the analogy between information transmitted and the mapping between the subsets of two partitions.

Other equivalence relations $R_B, R_C, \ldots$ generate different partitions of $\mathbf{M}$, i.e.,

$$\mathbf{M} \overset{R_B}{\Rightarrow} \mathbf{B} = \{\mathbf{B}_1, \mathbf{B}_2, \ldots, \mathbf{B}_{N_\mathbf{B}}\}, \tag{7}$$

$$\mathbf{M} \overset{R_C}{\Rightarrow} \mathbf{C} = \{\mathbf{C}_1, \mathbf{C}_2, \ldots, \mathbf{C}_{N_\mathbf{C}}\}, \tag{8}$$

whose sizes are given by expressions, with appropriate modifications, that are analogous to those given in equation (6).

Partitions can be generated in a number of ways (*vide supra*). In this work the emphasis is on partitions induced by similarity- or distance-based relations, as embodied in hierarchical and non-hierarchical clustering methods [16], by methods such as those employed in rough set theory (RST) to determine equivalence classes (called indiscernablity classes in RST) directly, and in cell-based partitions derived from coordinate-based property spaces such as, for example, those obtained in BCUT property spaces from the program Diverse Solutions (DVS) developed by Pearlman, et al. [25].

2.2. *Set-valued mappings between partitioned property spaces*

Mappings between partitioned property spaces can be viewed as analogous to the transmission of signals between a 'sender' and a 'receiver' as depicted in figure 1. Claude Shannon developed an information-theoretic framework for handling signal transmissions based upon a statistical entropy function [17]. The set-valued mappings indicated in the figure summarize the different possibilities that are typically encountered: (1) one-to-one mappings of the objects in a single subset in one partition to a single subset in another partition correspond to

a 'perfect' communication channel – all symbols sent are faithfully received, (2) one-to-many mappings of the objects in a single subset of one partition to many subsets in another partition correspond to a 'noisy' communication channel – a given symbol that is sent may be received as any one of a number of symbols, (3) many-to-one mappings of objects from many subsets in one partition to a single subset in another partition correspond to an 'equivocal' communication channel – several different symbols sent may be received as the same symbol, and (4) composite or mixed mappings correspond to a 'mixed' communication channel – all of the above may occur, the latter case being the most prevalent. It follows that there is a correspondence between subsets and symbols. Thus, a given partition can in a way be viewed as an alphabet, where each subset of the partition corresponds to an unique symbol of the alphabet [27].

2.3. *Information and Shannon entropy*

Shannon entropy can be derived from the concept of information in the following way. Information, sometimes called 'surprisal,' is defined, in units of 'bits,' as [28,17–21]

$$\mathrm{I}(\mathbf{A}_i) = \log_2 \frac{1}{\mathrm{P}(\mathbf{A}_i)}, \tag{9}$$

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

$$\mathrm{P}(\mathbf{A}_i) = \frac{|\mathbf{A}_i|}{|\mathbf{M}|} = \frac{n_{\mathbf{A}_i}}{n}, \tag{10}$$

and the sizes of the sets are given by equation (6). In addition,

$$\sum_{\mathbf{A}_i \in \mathbf{A}} \mathrm{P}(\mathbf{A}_i) = 1 \quad \text{and} \quad \mathrm{P}(\mathbf{A}_i) \geqslant 0, \quad \text{for } i = 1, 2, \ldots, N_{\mathbf{A}}. \tag{11}$$

Equation (9) 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 [29].

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

$$\begin{aligned} \mathrm{H}(\mathbf{A}) &= \langle \mathrm{I}(\mathbf{A}_i) \rangle_{\mathbf{A}} \\ &= \sum_{\mathbf{A}_i \in \mathbf{A}} \mathrm{P}(\mathbf{A}_i) \log_2 \frac{1}{\mathrm{P}(\mathbf{A}_i)} \\ &= -\sum_{\mathbf{A}_i \in \mathbf{A}} \mathrm{P}(\mathbf{A}_i) \log_2 \mathrm{P}(\mathbf{A}_i). \end{aligned} \tag{12}$$

It can also be shown that the maximum value of $\mathrm{H}(\mathbf{A})$ occurs when all of the equivalence classes are occupied equally, that is $|\mathbf{A}_1| = |\mathbf{A}_2| = \cdots = |\mathbf{A}_{N_{\mathbf{A}}}| =$

$\bar{n}_{\mathbf{A}}$, while the minimum occurs when all of the elements of the set reside in a single subset, $\mathrm{P}(\mathbf{A}_i) = 1 \rightarrow \mathrm{H}(\mathbf{A}) = 0$, so that

$$0 \leqslant \mathrm{H}(\mathbf{A}) \leqslant \mathrm{H}_{\max}(\mathbf{A}) = \log_2 \bar{N}_A, \tag{13}$$

where

$$\bar{N}_{\mathbf{A}} = n/\bar{n}_{\mathbf{A}}. \tag{14}$$

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

Consider the information in the case of *co-occurrences*, that is the *joint information*,

$$\mathrm{I}(\mathbf{A}_i, \mathbf{B}_j) = \log_2 \frac{1}{\mathrm{P}(\mathbf{A}_i, \mathbf{B}_j)}, \tag{15}$$

where $\mathrm{P}(\mathbf{A}_i, \mathbf{B}_j)$ is the probability of observing an object from subset $\mathbf{A}_i \in \mathbf{A}$ [equation (3)] *and* from subset $\mathbf{B}_j \in \mathbf{B}$ [equation (8)]. In an analogy to the above case for a single partition $\mathbf{A}$ [see equation (9)], $\mathrm{I}(\mathbf{A}_i, \mathbf{B}_j)$ is the information gained in observing an object (i.e., object) in $\mathbf{A}_i \in \mathbf{A}$ and in $\mathbf{B}_j \in \mathbf{B}$. Stated mathematically,

$$\mathrm{P}(\mathbf{A}_i, \mathbf{B}_j) = \frac{|\mathbf{A}_i \cap \mathbf{B}_j|}{|\mathbf{A} \cap \mathbf{B}|} = \frac{n_{\mathbf{A}_i, \mathbf{B}_j}}{n}, \tag{16}$$

since [30]

$$|\mathbf{A} \cap \mathbf{B}| = \sum_{\mathbf{A}_i \in \mathbf{A}} \sum_{\mathbf{B}_j \in \mathbf{B}} |\mathbf{A}_i \cap \mathbf{B}_j| = \sum_{i=1}^{N_{\mathbf{A}}} \sum_{j=1}^{N_{\mathbf{B}}} n_{\mathbf{A}_i, \mathbf{B}_j} = n \tag{17}$$

and

$$\mathbf{A} \cap \mathbf{B} = \{\mathbf{A}_i \cap \mathbf{B}_j \,|i = 1, 2, \ldots, N_{\mathbf{A}};\ j = 1, 2, \ldots, N_{\mathbf{B}}\}. \tag{18}$$

The joint Shannon entropy is then given, analogously to equation (12), as the expectation value of the joint information,

$$\begin{aligned} \mathrm{H}(\mathbf{A}, \mathbf{B}) &= \langle \mathrm{I}(\mathbf{A}_i, \mathbf{B}_j) \rangle_{\mathbf{A},\mathbf{B}} \\ &= \sum_{\mathbf{A}_i \in \mathbf{A}} \sum_{\mathbf{B}_j \in \mathbf{B}} \mathrm{P}(\mathbf{A}_i, \mathbf{B}_j) \log_2 \frac{1}{\mathrm{P}(\mathbf{A}_i, \mathbf{B}_j)} \\ &= -\sum_{\mathbf{A}_i \in \mathbf{A}} \sum_{\mathbf{B}_j \in \mathbf{B}} \mathrm{P}(\mathbf{A}_i, \mathbf{B}_j) \log_2 \mathrm{P}(\mathbf{A}_i, \mathbf{B}_j). \end{aligned} \tag{19}$$

In analogy with the case above for the partitioning **A**, the bounds of H(**A**, **B**) are given by [31]

$$0 \leqslant \mathrm{H}(\mathbf{A}, \mathbf{B}) \leqslant \mathrm{H}_{\max}(\mathbf{A}, \mathbf{B}) = \log_2 |\mathbf{A} \cap \mathbf{B}| = \log_2 \bar{N}_{\mathbf{A},\mathbf{B}}, \tag{20}$$

where

$$\left|\mathbf{A}_i \cap \mathbf{B}_j\right| = n_{\mathbf{A}_i,\mathbf{B}_j} = \bar{n}_{\mathbf{A},\mathbf{B}} \quad \text{for all } i = 1, 2, \ldots, N_{\mathbf{A}};\ j = 1, 2, \ldots, N_{\mathbf{B}} \tag{21}$$

and

$$\bar{N}_{\mathbf{A},\mathbf{B}} = \frac{n}{\bar{n}_{\mathbf{A},\mathbf{B}}}. \tag{22}$$

If **A** and **B** are *independent*, i.e., $\mathrm{P}(\mathbf{A}_i, \mathbf{B}_j) = \mathrm{P}(\mathbf{A}_i) \cdot \mathrm{P}(\mathbf{B}_j)$, the maximum joint entropy can be written in terms of the marginal entropies H(**A**) and H(**B**),

$$\mathrm{H}_{\max}(\mathbf{A}, \mathbf{B}) = \mathrm{H}(\mathbf{A}) + \mathrm{H}(\mathbf{B}), \tag{23}$$

and the bounds of H(**A**, **B**), become

$$\max[\mathrm{H}(\mathbf{A}), \mathrm{H}(\mathbf{B})] \leqslant \mathrm{H}(\mathbf{A}, \mathbf{B}) \leqslant \mathrm{H}(\mathbf{A}) + \mathrm{H}(\mathbf{B})\ [32] \tag{24}$$

$$\text{Dependent} \Leftarrow \qquad\qquad \Rightarrow \text{Independent}$$

As also indicated below equation (24), moving towards the left of the inequality leads to dependent, perfect mappings while moving to the right leads to independent, random mappings. It should also be noted that many mappings between the subsets of **A** and those of **B** exist for values of H(**A**) and H(**B**). Joint information can be generalized to more than two set partitions (variables), i.e. **A**, **B**, **C**, **D**, ..., [33,34], but all discussions in this work will be confined to two. A series of *conditional entropies* [17–21] can also be defined but will not be discussed here, as they are not employed in the analyses presented in this paper.

2.4. Mutual information

Mutual information, which is the basis for much of the analysis that follows [35], is defined as the difference between the maximum joint entropy and the observed joint entropy, equations (23) and (19), respectively,

$$\begin{aligned} \mathrm{M}(\mathbf{A}, \mathbf{B}) &= \mathrm{H}(\mathbf{A}) + \mathrm{H}(\mathbf{B}) - \mathrm{H}(\mathbf{A}, \mathbf{B}) \\ &= \mathrm{H}_{\max}(\mathbf{A}, \mathbf{B}) - \mathrm{H}(\mathbf{A}, \mathbf{B}), \end{aligned} \tag{25}$$

which can be viewed as the amount of information "shared" or "transmitted" between the two partitionings.

An alternative but useful expression for mutual information, which can be derived from equation (25), is given by

$$\mathrm{M}(\mathbf{A}, \mathbf{B}) = \sum_{\mathbf{A}_i \in \mathbf{A}} \sum_{\mathbf{B}_j \in \mathbf{B}} \mathrm{P}(\mathbf{A}_i, \mathbf{B}_j) \log_2 \frac{\mathrm{P}(\mathbf{A}_i, \mathbf{B}_j)}{\mathrm{P}(\mathbf{A}_i) \cdot \mathrm{P}(\mathbf{B}_j)}. \tag{26}$$

It is clear from equation (26) that if **A** and **B** are independent, that is if $\mathrm{P}(\mathbf{A}_i, \mathbf{B}_j) = \mathrm{P}(\mathbf{A}_i) \cdot \mathrm{P}(\mathbf{B}_j)$, then $\mathrm{M}(\mathbf{A}, \mathbf{B}) = 0$ since $\log_2 1 = 0$. Thus, mutual information provides a single measure of *statistical independence*, and is superior to the use of correlation coefficients that require $n(n-1)/2$ terms in the case of n variates. Moreover, correlation coefficients only strictly apply if the variates are normally distributed and linearly correlated, while mutual information applies to any distribution and to non-linearly correlated data.

Bounds on mutual information are given by

$$0 \leqslant \mathrm{M}(\mathbf{A}, \mathbf{B}) \leqslant \mathrm{M}(\mathbf{A}, \mathbf{B})_{\max} = \min[\mathrm{H}(\mathbf{A}), \mathrm{H}(\mathbf{B})]. \tag{27}$$

A *similarity metric* [24] can then be defined as the "normalized" mutual information

$$0 \leqslant \mathrm{S}(\mathbf{A}, \mathbf{B}) = \frac{\mathrm{M}(\mathbf{A}, \mathbf{B})}{\mathrm{M}(\mathbf{A}, \mathbf{B})_{\max}} \leqslant 1, \tag{28}$$

more specifically as the fraction of the maximum information that can be shared between two partitionings.

This can also be taken as a "normalized" measure of the dependency between the two partitions, where $\mathrm{S}(\mathbf{A}, \mathbf{B}) = 1$ indicates maximum dependency and $\mathrm{S}(\mathbf{A}, \mathbf{B}) = 0$ indicates complete independence. In the former case, $\mathrm{M}(\mathbf{A}, \mathbf{B}) = \mathrm{M}(\mathbf{A}, \mathbf{B})_{\max} = \min[\mathrm{H}(\mathbf{A}), \mathrm{H}(\mathbf{B})]$ can occur in two ways. The first corresponds to a perfect mapping and obtains when $\mathrm{H}(\mathbf{A}) = \mathrm{H}(\mathbf{B}) = \mathrm{H}(\mathbf{A}, \mathbf{B})$, which from equations (25) and (27) yields $\mathrm{M}(\mathbf{A}, \mathbf{B}) = \mathrm{M}(\mathbf{A}, \mathbf{B})_{\max}$. The situation also obtains when $\mathrm{H}(\mathbf{A}) < \mathrm{H}(\mathbf{B}) = \mathrm{H}(\mathbf{A}, \mathbf{B})$, which does not correspond to a perfect mapping, and thus is called a psuedo-perfect mapping and will be discussed further in section 3.

3. Geometric interpretation of information-theoretic quantities

Additional insights can be obtained by analyzing mutual information, $\mathrm{M}(\mathbf{A}, \mathbf{B})$, and its associated marginal and joint entropies, $\mathrm{H}(\mathbf{A}), \mathrm{H}(\mathbf{B})$, and $\mathrm{H}(\mathbf{A}, \mathbf{B})$, in geometric terms. The basis for constructing the geometric model shown in figure 2 is the set of inequalities given by equations (23) and (24), which can be summarized as

$$\mathrm{H}(\mathbf{A}) \leqslant \mathrm{H}(\mathbf{B}) \leqslant \mathrm{H}(\mathbf{A}, \mathbf{B}) \leqslant \mathrm{H}_{\max}(\mathbf{A}, \mathbf{B}) = \mathrm{H}(\mathbf{A}) + \mathrm{H}(\mathbf{B}), \tag{29}$$

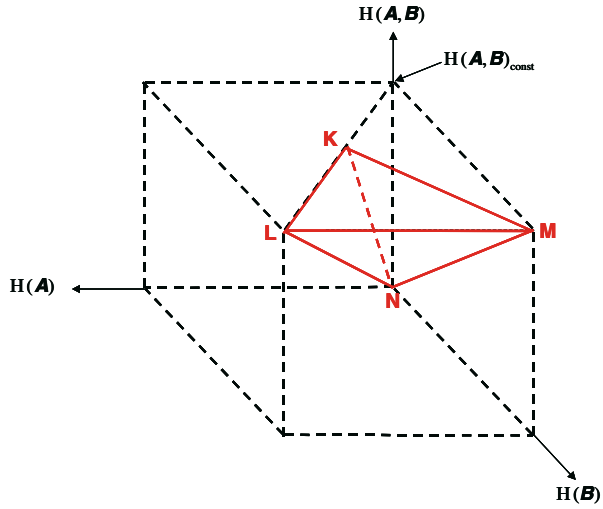

Figure 2. Geometric interpretation of information-theoretic equations.

where taking $H(\mathbf{A}) \leqslant H(\mathbf{B})$ does not affect the generality of the inequality. In the diagram in figure 2 $H(\mathbf{A}, \mathbf{B})$ is plotted against $H(\mathbf{A})$ and $H(\mathbf{B})$. All of the edges of the cube in figure are of equal length, so that at vertex **L**, $H(\mathbf{A}) = H(\mathbf{B}) = H(\mathbf{A}, \mathbf{B})$, at vertex **M**, $H(\mathbf{A}) = 0, H(\mathbf{B}) = H(\mathbf{A}, \mathbf{B})$, at vertex **K**, $H(\mathbf{A}) = H(\mathbf{B}) = 1/2(H(\mathbf{A}, \mathbf{B}))$, and at the origin **N**, $H(\mathbf{A}) = H(\mathbf{B}) = H(\mathbf{A}, \mathbf{B}) = 0$. All points on the top face of the cube satisfy $H(\mathbf{A}, \mathbf{B}) = H(\mathbf{A}, \mathbf{B})_{\text{const}}$. The triangle **KLM** defines the set of allowed values of $H(\mathbf{A})$ and $H(\mathbf{B})$ for a specific value of $H(\mathbf{A}, \mathbf{B})_{\text{const}}$ and thus the set of allowed values of $M(\mathbf{A}, \mathbf{B})$ (see equation (25)). As $H(\mathbf{A}, \mathbf{B})_{\text{const}}$ decreases in value from that shown in the figure 2 triangle **KLM** also decreases in size until it "collapses" to a point at the origin **N**. The infinite set of such triangles "fill in" the tetrahedron **KLMN**. It is important to note that the lines corresponding to edges **NK**, **NL**, and **NM** go to infinity so that the full tetrahedron is actually infinite in extent. Thus, the infinite tetrahedron represents the entire set of allowed values of $H(\mathbf{A})$ and $H(\mathbf{B})$ for all values of $H(\mathbf{A}, \mathbf{B})$.

Figure 3 depicts the $H(\mathbf{A}, \mathbf{B})_{\text{const}}$ plane from above. As indicated in the figure, vertex **L** corresponds to a perfect mapping where $H(\mathbf{A}) = H(\mathbf{B}) = H(\mathbf{A}, \mathbf{B})$ (*vide supra*), so that $M(\mathbf{A}, \mathbf{B}) = H(\mathbf{A}) = M_{\max}(\mathbf{A}, \mathbf{B})$ and thus $S(\mathbf{A}, \mathbf{B}) = 1$. All of the remaining points on edge **LM** correspond to psuedo-perfect mappings where $H(\mathbf{A}) \leqslant H(\mathbf{B}) = H(\mathbf{A}, \mathbf{B})$, which from equations (25) and (27) again give $S(\mathbf{A}, \mathbf{B}) = 1$. The points on Edge **KM** correspond to allowed values of $H(\mathbf{A})$ and $H(\mathbf{B})$ for the case of a random mapping (i.e., independence) where $H(\mathbf{A}) + H(\mathbf{B}) = H_{\max}(\mathbf{A}, \mathbf{B}) = H(\mathbf{A}, \mathbf{B})$, so that $M(\mathbf{A}, \mathbf{B}) = 0$ and thus $S(\mathbf{A}, \mathbf{B}) = 0$. All of the remaining points inside triangle **KLM** plus all points on edge **KL** except for the points at vertices **K** and **L** correspond to allowed values of $H(\mathbf{A})$ and $H(\mathbf{B})$ with respect to $H(\mathbf{A}, \mathbf{B})_{\text{const}}$. In this region, $0 < M(\mathbf{A}, \mathbf{B}) < H(\mathbf{A})$ so that $0 <$

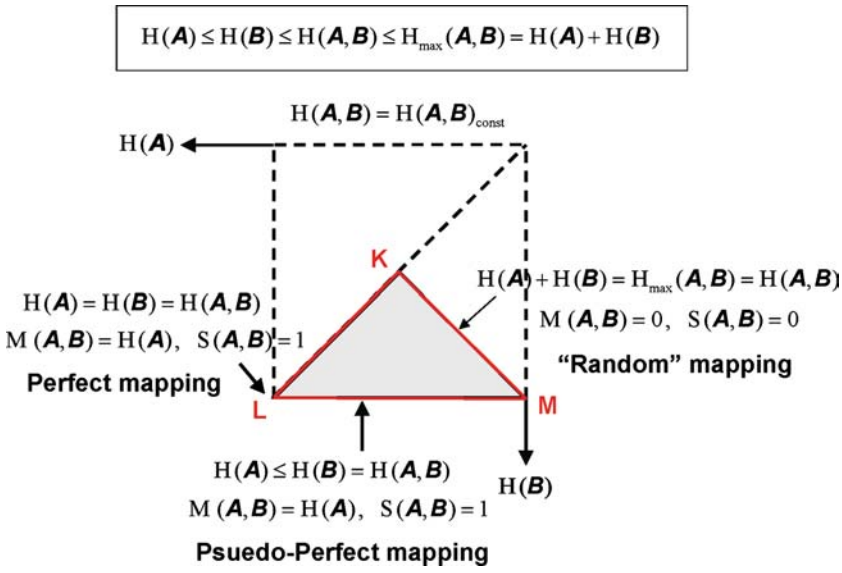

Figure 3. Geometric representation of the plane of mutual information for a fixed value of the joint entropy $H(\mathbf{A}, \mathbf{B})_{const}$.

$S(\mathbf{A}, \mathbf{B}) < 1$ within the entire region. From figure 2 it is clear that as $H(\mathbf{A}, \mathbf{B})_{const}$ decreases in value, the smaller the range of allowed $H(\mathbf{A})$ and $H(\mathbf{B})$ values. Also, for a given value $H(\mathbf{A}, \mathbf{B})_{const}$ of the closer points are to edge **LM** the closer the corresponding mapping is to a perfect/psuedo perfect mapping, while the closer points are to edge **KM** the closer the corresponding mapping is to a random mapping.

4. Illustrative examples

4.1. Perfect mappings

To help clarify the foregoing material simple examples of three mappings are presented – perfect, psuedo-perfect, and mixed (cf. figure 1). First, consider the case of a one-to-one or perfect mapping between two partitions **A** and **B** of the set **M** of 20 objects,

$$\mathbf{M} = \{m_i \mid i = 1, 2, \ldots, 20\}. \tag{30}$$

The first partition (i.e., 'the sender') is given by

$$\mathbf{M} \overset{R_A}{\Rightarrow} \mathbf{A} = \{\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_6\}, \tag{31}$$

Table 1
One-to-one ('perfect') mappings of two property spaces, **A** & **B**: Subset view.[a]

	$\mathbf{B}_1$	$\mathbf{B}_2$	$\mathbf{B}_3$	$\mathbf{B}_4$	$\mathbf{B}_5$	$\mathbf{B}_6$	
$\mathbf{A}_1$	{1,2,3,4}						{1,2,3,4}
$\mathbf{A}_2$		{5,6,7}					{5,6,7}
$\mathbf{A}_3$			{8}				{8}
$\mathbf{A}_4$				{9,10,11,12}			{9,10,11,12}
$\mathbf{A}_5$						{13,14,15,16,17}	{13,14,15,16,17}
$\mathbf{A}_6$					{18,19,20}		{18,19,20}
	{1,2,3,4}	{5,6,7}	{8}	{9,10,11,12}	{18,19,20}	{13,14,15,16,17}	

[a]Note that the numbers in curly brackets correspond to the elements of set M given by equation (32).

where

$$\begin{aligned} \mathbf{A}_1 &= \{m_1, m_2, m_3, m_4\}, & \mathbf{A}_4 &= \{m_9, m_{10}, m_{11}, m_{12}\}, \\ \mathbf{A}_2 &= \{m_5, m_6, m_7\}, & \mathbf{A}_5 &= \{m_{13}, m_{14}, m_{15}, m_{16}, m_{17}\}, \\ \mathbf{A}_3 &= \{m_8\}, & \mathbf{A}_6 &= \{m_{18}, m_{19}, m_{20}\}. \end{aligned} \tag{32}$$

Since a perfect mapping corresponds to the case where all of the objects in a given subset of **A** map to one and only one subset of **B** (i.e., 'receiver'), **B** is given by

$$\mathbf{M} \overset{R_B}{\Rightarrow} \mathbf{B} = \{\mathbf{B}_1, \mathbf{B}_2, \ldots, \mathbf{B}_6\}, \tag{33}$$

where

$$\begin{aligned} \mathbf{B}_1 &= \{m_1, m_2, m_3, m_4\}, & \mathbf{B}_4 &= \{m_9, m_{10}, m_{11}, m_{12}\}, \\ \mathbf{B}_2 &= \{m_5, m_6, m_7\}, & \mathbf{B}_5 &= \{m_{18}, m_{19}, m_{20}\}, \\ \mathbf{B}_3 &= \{m_8\}, & \mathbf{B}_6 &= \{m_{13}, m_{14}, m_{15}, m_{16}, m_{17}\}, \end{aligned} \tag{34}$$

that is $\mathbf{A} = \mathbf{B}$. Table 1 summarizes this data. Note that the far right column corresponds to the subsets of partition **A** (see equation (32)) and the bottom row corresponds to the subsets of partition **B** (see equation (34)). Subsets located within the center of the table correspond to those objects that are common to the subsets of both partitions and are given in set-theoretic language as intersections of subsets. For example, $\mathbf{A}_4 \cap \mathbf{B}_4 = \{9, 10, 11, 12\}$, which describes the mapping of elements in $\mathbf{A}_4$ to $\mathbf{B}_4$ and shows that it is a one-to-one set-valued mapping. In the special case of a one-to-one mapping, $\mathbf{A}_i \cap \mathbf{B}_j = \mathbf{A}_i = \mathbf{B}_j$ [*N.B.* that i need not equal j as seen from equations (32) and (34)] and $\mathbf{A}_i \cap \emptyset = \emptyset \cap \mathbf{B}_j = \emptyset$, and thus each row and column contains only a single subset. All of this is summarized in figure 4. Note that the subsets lying on the 'arrows' in the figure correspond to intersection subsets of $\mathbf{A} \cap \mathbf{B}$. Taking the measure of all of the subsets

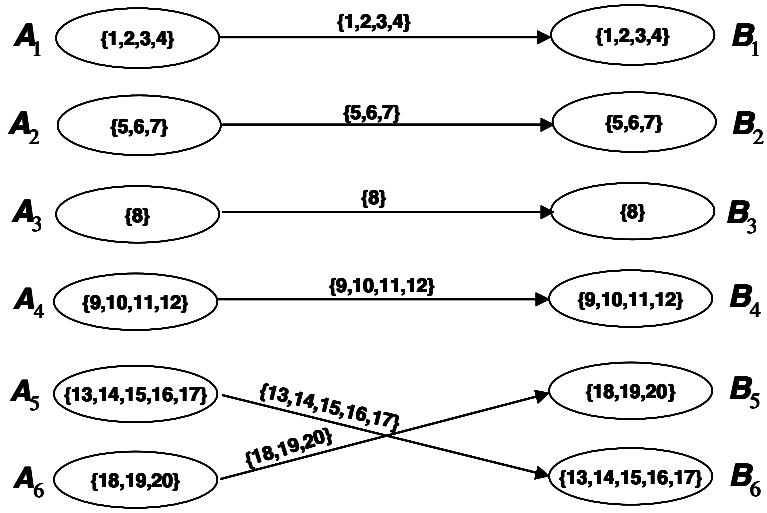

Figure 4. Example of a one-to-one mapping.

Table 2
One-to-one ('perfect') mappings of two property spaces, **A** & **B**: Probabilistic view.

	$\mathbf{B}_1$	$\mathbf{B}_2$	$\mathbf{B}_3$	$\mathbf{B}_4$	$\mathbf{B}_5$	$\mathbf{B}_6$	
$\mathbf{A}_1$	0.20						0.20
$\mathbf{A}_2$		0.15					0.15
$\mathbf{A}_3$			0.05				0.05
$\mathbf{A}_4$				0.20			0.20
$\mathbf{A}_5$						0.25	0.25
$\mathbf{A}_6$					0.15		0.15
	0.20	0.15	0.05	0.20	0.15	0.25	1.00

(see e.g. equation (6)) and converting them into probability estimates (see equations (10) and (16)) yields the values given in table 2, while table 3 contains the values of the various information-theoretic quantities. As is seen in table 3

$$\mathrm{H}(\mathbf{A}) = \mathrm{H}(\mathbf{B}) = \mathrm{H}(\mathbf{A}, \mathbf{B}) \Rightarrow \mathrm{M}(\mathbf{A}, \mathbf{B}) \Rightarrow \mathrm{S}(\mathbf{A}, \mathbf{B}) = 1, \tag{35}$$

which is expected since the two partitions are identical as shown in table 1. This corresponds to point L in figures 2 and 3.

4.2. *One-to-many or psuedo-perfect ('noisy') mappings*

Psuedo-perfect or one-to-many ('noisy') mappings represent a generalization of the previous case, where there is a subsethood relationship between the subsets of the 'sender' $\mathbf{A} = \{\mathbf{A}_1, \ldots, \mathbf{A}_6\}$ and 'receiver' $\mathbf{C} = \{\mathbf{C}_1, \ldots, \mathbf{C}_{10}\}$. Note

Table 3
One-to-one ('perfect') mappings of two property spaces, **A** & **B**: Information-theoretic quantities.

H(**A**) = 2.46596
H(**B**) = 2.46596
H(**A**, **B**) = 2.46596
M(**A**, **B**) = 2.46596
S(**A**, **B**) = 1.00000

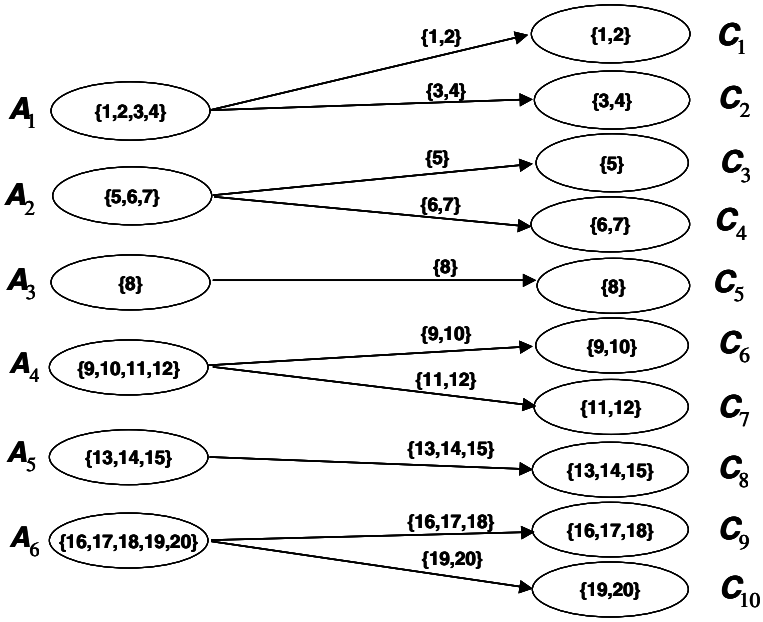

Figure 5. Example of a one-to-many pseudo-perfect mapping.

that the same applies for many-to-one mappings as well as one-to-many mappings, only the roles of 'sender' and 'receiver' are reversed. More explicitly, **C** is partitioned into the following subsets:

$$\begin{aligned}
\mathbf{C}_1 &= \{m_1, m_2\}, & \mathbf{C}_6 &= \{m_9, m_{10}\},\\
\mathbf{C}_2 &= \{m_3, m_4\}, & \mathbf{C}_7 &= \{m_{11}, m_{12}\},\\
\mathbf{C}_3 &= \{m_5\}, & \mathbf{C}_8 &= \{m_{13}, m_{14}, m_{15}\},\\
\mathbf{C}_4 &= \{m_6, m_7\}, & \mathbf{C}_9 &= \{m_{16}, m_{17}, m_{18}\},\\
\mathbf{C}_5 &= \{m_8\}, & \mathbf{C}_{10} &= \{m_{19}, m_{20}\},
\end{aligned} \tag{36}$$

which are all subsets of the subsets of **A** (compare equations (32) and (36)). For example, $\mathbf{A}_4 \supseteq \mathbf{C}_6$ and $\mathbf{A}_4 \supseteq \mathbf{C}_7$. Figure 5 summarizes this information in a form analogous to that given in figure 4. It is clear from the figure that $H(\mathbf{A}) < H(\mathbf{C}) = H(\mathbf{A}, \mathbf{C})$ since the subsets of **C** and $\mathbf{A} \cap \mathbf{C}$ are identical, that is $\mathbf{C} = \mathbf{A} \cap \mathbf{C}$. Thus,

Table 4
Composite ('mixed') mappings of two property spaces **B** & **D**: Subset view.[a]

	$\mathbf{D_1}$	$\mathbf{D_2}$	$\mathbf{D_3}$	$\mathbf{D_4}$	$\mathbf{D_5}$	$\mathbf{D_6}$	
$\mathbf{B_1}$	{1,2}			{3}		{4}	{1,2,3,4}
$\mathbf{B_2}$	{5}		{6}	{7}			{5,6,7}
$\mathbf{B_3}$			{8}				{8}
$\mathbf{B_4}$	{9}	{10,11}		{12}			{9,10,11,12}
$\mathbf{B_5}$						{18,19,20}	{18,19,20}
$\mathbf{B_6}$	{13}	{14}	{15,16}	{17}			{13,14,15,16,17}
	{1,2,5,9,13}	{10,11,14}	{6,8,15,16}	{3,7,12,17}		{4,18,19,20}	

[a]Note that the numbers in curly brackets corresponds to the elements of set given by equation (36).

from equation (25) $M(\mathbf{A}, \mathbf{C}) = H(\mathbf{A})$, which implies from equations (27) and (28) that $S(\mathbf{A}, \mathbf{C}) = 1$. This leads to the somewhat quirky behavior of the similarity index, namely that a 'noisy' mapping between two sets of subsets, **A** and **C**, is similar to a perfect mapping. But this only applies in the limited case where $\mathbf{C} = \mathbf{A} \cap \mathbf{C}$. As will be seen in the sequel, such behavior is observed in real datasets (Shanmugasundaram et al., in preparation). Although their similarities are both equal to unity, a psuedo-perfect mapping can be distinguished from a perfect mapping, since in the latter case $\mathbf{A} = \mathbf{B} = \mathbf{A} \cap \mathbf{B}$, and thus the number of subsets in **A** and **B** is equal, that is $|\mathbf{A}| = |\mathbf{B}|$. As discussed earlier, all psuedo-perfect mappings lie on the line **LM** in figures 2 and 3.

4.3. *Composite ('mixed') mappings*

One-to-one and one-to-many mappings are extremely unlikely to occur in practice, as are many-to-one mappings. Essentially all situations of interest in this work are described by composite ('mixed') mappings, which are illustrated by the following example based upon the same set of 20 objects that was used in the previous examples (see equation (29)). In this case, however, **B** is repartitioned as **D**, i.e.,

$$\begin{aligned} \mathbf{D}_1 &= \{m_1, m_2, m_5, m_9, m_{13}\}, & \mathbf{D}_4 &= \{m_3, m_7, m_{12}, m_{17}\}, \\ \mathbf{D}_2 &= \{m_{10}, m_{11}, m_{14}\}, & \mathbf{D}_5 &= \emptyset, \\ \mathbf{D}_3 &= \{m_6, m_8, m_{15}, m_{16}\}, & \mathbf{D}_6 &= \{m_4, m_{18}, m_{19}, m_{20}\}, \end{aligned} \tag{37}$$

which results in a different mapping. Table 4 summarizes the relevant set-data. As before, subsets located in the center of the table correspond to those objects that are common to subsets in both partitions, but in distinction to the previous case each row and column may contain multiple subsets. Taking the union of all subsets in a given row or column yields the subset in the right most column or bottom row, respectively, in the table. This is graphically depicted in figure 6.

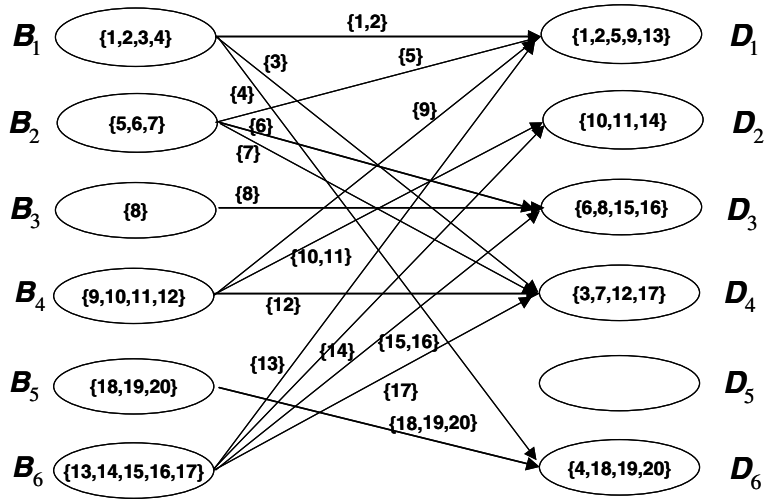

Figure 6. Example of a composite or mixed mapping.

Table 5
Composite ('mixed') mappings of two property spaces **B** & **D**: Probablistic view.

	D_1	D_2	D_3	D_4	D_5	D_6	
B_1	0.1			0.05		0.05	0.20
B_2	0.05		0.05	0.05			0.15
B_3			0.05				0.05
B_4	0.05	0.10		0.05			0.20
B_5						0.15	0.15
B_6	0.05	0.05	0.10	0.05			0.25
	0.25	0.15	0.20	0.20		0.20	1.00

As before, taking the measure of the subsets and converting them to probability estimates gives the values in table 5, while table 6 contains the computed values of the various information-theoretic quantities of interest. As expected $H(\mathbf{B}) \geqslant H(\mathbf{D})$ since **D** is more "concentrated" (or less dispersed) than **B** as seen in tables 1–6. The same holds true for the joint entropies, i.e., $H(\mathbf{A}, \mathbf{B}) \geqslant H(\mathbf{B}, \mathbf{D})$ for exactly the same reason. Based upon this data, it is clear from equation (25) that $M(\mathbf{A}, \mathbf{B}) = M(\mathbf{A}, \mathbf{C}) \geqslant M(\mathbf{B}, \mathbf{D})$. This follows from the fact that maximum dependency obtains in a perfect mapping, and thus the amount of information shared between the two partitions is maximal. As mappings deviate from perfect mappings, dependencies between the two partitions are reduced, reducing the mutual information, until in the case of complete independence, i.e., when $P(\mathbf{B}_i, \mathbf{D}_j) = P(\mathbf{B}_i) \cdot P(\mathbf{D}_j)$ for all $\mathbf{B}_i \in \mathbf{B}$ and $\mathbf{D}_j \in \mathbf{D}$, the mutual information goes to zero. As expected, $S(\mathbf{A}, \mathbf{B}) \geqslant S(\mathbf{B}, \mathbf{D}) \Rightarrow 0$.

Table 6
Composite ('mixed') mappings of two property spaces **B** & **D**: Information-theoretic quantities.

$H(\mathbf{B}) = 2.46596$
$H(\mathbf{D}) = 2.3037$
$H(\mathbf{B}, \mathbf{D}) = 3.78418$
$M(\mathbf{B}, \mathbf{D}) = 0.985475$
$S(\mathbf{B}, \mathbf{D}) = 0.42778$

All values of the mutual information not associated with perfect or psuedo-perfect mappings lie within triangle KLM in figures 2 and 3, and include edges **KL** and **KM**, except for the endpoints **L** and **M**. The edge **KM** is, however, special since the mutual information is zero for all points on this edge except for vertex **M**.

5. Discussion

A methodology, derived by analogy to Shannon's information-theoretic theory of communication and utilizing the concept of mutual information, has been developed to characterize partitioned property spaces. As described in section 2.1 partitioned property space is represented by a family of non-intersecting subsets that cover the "universe" of objects, each subset is thus an equivalence class. A partition and it's associated equivalence classes can be generated using any one of a number of procedures including hierarchical and non-hierarchical clustering, direct approaches using rough set methods, and cell-based partitioning, to name a few. The approach is based on set-valued mappings from equivalence classes in one partition to those in another and provides a coarse-grained means for comparing property spaces as discussed in section 2.2.

As described above, any class of problems that can be formulated as set-valued mappings between two partitionings can, thus, be treated using the formalism described here. And part of the power of the approach is that it can be applied to virtually all types of partitionings as long as the mapping between the equivalence classes can be determined. For example, how similar are cell-based partitionings of property space generated by say 3-D and 2-D BCUTs [25], or how similar are two property spaces where one is a cell-based property space that has been generated by say 3-D BCUTs and the other by c-means clustering generated by a given similarity metric can be analyzed by this approach. In bioinformatics, for example, comparing protein sequences can be cast into the framework described here. In this case the two partitionings are identical and each equivalence class of either partition represents one of the 20 character amino acid symbols of the 'protein alphabet.' The set-valued mappings are

then determined from the sequence alignment. Suppose that the alignment shows that, say, in five cases *alanine* residues, 'A', in protein-1 are replaced by *valine* residues, 'V,' in protein-2. The intersection set of the subset containing 'A' in protein-1 and the subset containing 'V' in protein-2 has five elements, and thus has measure five. This process can be carried for all of the amino acids in protein-1 and those in protein-2, from which the information-theoretic-based similarity (see equation (28)) can be computed. Yockey has dealt with this problem in some detail, although he has not formulated it in terms of set-valued mappings [24].

Another interesting bioinformatics problem is the grouping of proteins into appropriate functional families. This can be accomplished in several steps. If only sequence information is available, the similarity matrix for all of the proteins being considered is computed using any one of a number of sequence-based methods. The similarities are then used as a basis for clustering the proteins using either a hierarchical or non-hierarchical procedure. It is then assumed that those proteins located within the same cluster are functionally related. However, as is well known from numerous studies of small data-sets in property space, the results of these clusterings are highly dependent on both the sequence similarity methodology and the clustering procedure used. Thus, the approach described here can be used to characterize "protein-function space" in terms of the different representations and clustering procedures employed (*vide infra*).

Since the tertiary structure of proteins tends to be conserved to a much greater extent than primary structure, it is expected that similarities determined from 3-D structure would more faithfully represent the conserved functional elements of proteins within a given family or subfamily. Thus, structural similarity should provide a better basis for elucidating the functional relationships among a large set of proteins with diverse functions. Applying the same clustering methodologies as those applied to sequence-based similarities, it is not surprising that similar issues would arise regarding the consistency or lack thereof of results based upon the alignment-based similarity method as well as the clustering methodology used. Again, the current methodology can be applied. In addition, protein clustering produced by sequence based alignments can be compared to those produced by structure-based alignments.

The information-theoretic methodology described in this work also affords the opportunity to carry out a *meta-analysis* of the methodologies used (*vide supra*). For example, it is possible to determine the similarity of different approaches with respect to each other. The similarity matrix can then be used as a basis for clustering the "methodologies" in a fashion that is totally analogous that the used to cluster the proteins into families (*vide supra*). Alternatively, the similarity matrix can be employed as a basis for generating a low-dimensional (2-D or 3-D) coordinate system that can be then be used to portray the "methodological similarity space." [6] Now the clusters emerge quite naturally as groups of points in the space. Points located near each other in the space thus represent methods that generate similar partitions the protein space.

These two examples represent just the "tip of the iceberg." For example, metabolic pathways, which are representable by mathematical graphs, can be analyzed in analogous fashion since it is possible to determine the distance and similarity between two such metabolic pathway graphs [36]. Thus, pathway graphs among different species can be treated in essentially the same way as chemical graphs, and the partitioning of "pathway-graph space" can then be assessed in the same manner as applied in the examples cited earlier. Many other examples can be conceived of, those presented here are meant only to suggest the types of problems the can be addressed.

Acknowledgments

The authors would like to thank Professor Tack Kuntz (University of California, San Francisco) for helpful comments and discussions.

References

[1] Note that while these entities are usually referred to as vectors some of them do not satisfy all of the mathematical axioms of vectors (e.g., with respect to vector addition and scalar multiplication). Thus, property space is not always, strictly speaking, a vector space, although the concept of distance remains valid and thus property space is a metric space with a well-defined geometry.
[2] I. Borg and P. Groenen, *Modern Multi-Dimensional Scaling* (Springer-Verlag, Heidelberg, 1997).
[3] D. Domine, J. Devillers, M. Chastrette, and W. Karcher. Non-linear mapping for structure-activity and structure-property modeling. J. Chemom. 7 (1993) 227–242.
[4] D.K. Agrafiotis and V.S. Lobanov, Nonlinear mapping networks, J. Chem. Info. Comput. Sci. 40 (2000) 1356–1362.
[5] D.N. Rassokhin, V.S. Lobanov, and D.K. Agrafiotis, Non-linear mapping of massive data sets by fuzzy clustering and neural networks, J. Comput. Chem. 22 (2001) 373–386.
[6] G.M. Maggiora and V. Shanmugasundaram, Molecular Similarity Measures, in: *Chemoinformatics: Methods and Protocols*, 1-50, ed. J. Bajorath, (Humana Press, Totowa, New Jersey, 2004).
[7] M.S. Lajiness, Dissimilarity-based compound selection techniques, Perspec. Drug Disc. Design 7/8 (1997) 65–84.
[8] L. Kaufman and P.J. Rousseeuw, *Finding Groups in Data*, (Wiley-Interscience, Wiley, New York, 1990).
[9] P. Arabie, L.J. Hubert, and G. De Spoete (eds.), *Clustering and Classification*, (World Scientific, Singapore, (1996)).
[10] In set theoretic language, a cover of a set X is the union of the set of subsets that is equal to $X : X = \cup_{i=1}^{n} S_i$; if $S_i \cap S_j = \emptyset$. The cover is also a partition.
[11] J. Zupan. Algorithms for Chemists. (Wiley, New York, 1989).
[12] R. Rosen. *Fundamentals of Measurement and Representation of Natural Systems*, (North-Holland, New York, 1978).
[13] Note that this only applies in the case where the clusters are non-interacting subsets.
[14] Z. Pawlak, *Rough Sets: Theoretical Aspects of Reasoning About Data*, (Kluwer Academic Publishers, Dordrecht, The Neatherlands, 1991).

[15] See, for example, "Cell-Based Methods for Sampling in High-Dimensional Spaces," in: *Rational Drug Design*, eds. D.G. Truhlar and W.J. Howe et al., (Springer-Verlag, New York, 1991) pp. 73–79.
[16] A.K. Jain and R.C. Dubes, *Algorithms for Clustering Data*, (Prentice Hall, Englewood Cliffs, New Jersey, 1988).
[17] N.J. Sloane and A.D. Wyner (eds.), *Claude Elwood Shannon: Collected Papers*. (IEEE Press, Piscataway, New Jersey, 1993).
[18] T.M. Cover and J.A. Thomas, *Elements of Information Theory*, (Wiley, New York, 1991).
[19] K. Krippendorff, *Information Theory – Structural Models for Qualitative Data* (Sage Publications, Newbury Park, 1986).
[20] G.J. Klir and M.J Wierman, *Uncertainty-Based Information*, 2nd edition, (Physica-Verlag, Heidelberg, 1999).
[21] R.C.-T. Lee, Application of information theory to select relevant variables, Math. Biosci. 11 (1971) 153–161.
[22] X. Zhou, X. Wang, and E.R. Dougherty. Construction of genomic networks using mutual-information clustering and reversible jump Markov-chain-Monte-Carlo predictor design, Signal Process. 83 (2003) 745–761.
[23] G. Tononi, O. Sporns, and G.M. Edelman, Measures of degeneracy and redundancy in biological networks, Proc. Nat. Acad. Sci. (USA) 96 (1999) 3257–3262.
[24] H.P. Yockey, *Information Theory and Molecular Biology*, (Cambridge University Press, Cambridge, 1992).
[25] R.S. Pearlman and K.M. Smith, Novel software tools for chemical diversity. Perspec. Drug Disc. Design 9–11 (1998) 339–353.
[26] Note that binning or cellularizing property spaces induces equivalence classes, i.e., all objects that belong to a single bin or cell constitute a single equivalence class. Although objects located within the same bin are not necessarily equivalent by "chemical standards" they are equivalent mathematically.
[27] Note that in nucleic acid and protein sequence matching the symbols correspond to the four bases or the 20 amino acids, respectively. In this case, the mappings are determined by first aligning the sequences and then examining the correspondences between the two sequences.
[28] The units used depend upon the base of the logarithm – for base 2 the unit is a 'bit.' Other unit systems are also used but will not be discussed further here.
[29] For example, if an urn is filled with 90 red balls and 10 green balls there is a 90% chance that any ball drawn from the urn will be red compared to a 10% chance for a green ball. Thus, actually drawing a red ball is much less surprising and hence carries less information than drawing a green one.
[30] Note that $\sum_{i=1}^{N_A} \sum_{j=1}^{N_B} n_{A_i,B_j} = \sum_{i=1}^{N_A} n_{A_i} = \sum_{j=1}^{N_B} n_{B_j} = n$.
[31] If all of the equivalence classes are singly occupied $\bar{N}_{A,B} = n$.
[32] Note that $\mathrm{H}_m(\mathbf{A}, \mathbf{B}) \leqslant \mathrm{H}_m(\mathbf{A}) + \mathrm{H}_m(\mathbf{B})$ is called subadditivity and is discussed in Klir and Wierman book (see Reference 20).
[33] W.R. McGill, Multivariate information transmission, Psycometrica 19 (1954) 97–116.
[34] W.R. Ashby, Two tables of identities governing information flows within large systems, Amer. Soc. Cybernetics Comm. 1, 2 (1969) 3–8.
[35] For an interesting discussion of information measures see M. Gell-Mann and S. Lloyd. Information Measures, Effective Complexity, and Total Information, Complexity 2(1) (1996) 44–52.
[36] M. Kinoehisa, *Post-Genome Informatics*, (Oxford University Press, Oxford, 2000).