sure of similarity between molecules in some property space and give values in the range of 0 to 1, with 1 being identical. Typical examples are the Tanimoto coefficient and the Cosine coefficient. For real-valued properties the Tanimoto is defined as

$$\text{Tanimoto} = \frac{\sum_{i=1}^{i=N} x_{iA}x_{iB}}{\sum_{i=1}^{i=N} (x_{iA})^2 + \sum_{i=1}^{i=N} (x_{iB})^2 - \sum_{i=1}^{i=N} x_{iA}x_{iB}}$$

where x_{iA} is the value of property i of molecule A. When i can take values of only 0 or 1 as in a bit string, then this reduces to

$$\text{Tanimoto} = ab/(a + b - c)$$

where a is the number of on-bits in A and c is the number of bits in common between A and B. The Cosine coefficient can be defined as

$$\text{Cosine} = \frac{\sum_{i=1}^{i=N} x_{iA}x_{iB}}{\sqrt{\sum_{i=1}^{i=N} (x_{iA})^2 \sum_{i=1}^{i=N} (x_{iB})^2}} \Rightarrow \frac{c}{\sqrt{ab}}$$

For field-based measures and overlap of electron density functions then the Carbo index can be used (49), which is equivalent to the Cosine coefficient.

Distance measures give 0 for identical structures and have an upper bound defined by the property space. The Euclidean and Hamming distances are the most common:

$$\text{Euclidean distance} = \sqrt{\sum_{i=1}^{i=N} (x_{iA} - x_{iB})^2} \Rightarrow \sqrt{a + b - 2c}$$

$$\text{Hamming distance} = \sum_{i=1}^{i=N} |x_{iA} - x_{iB}| \Rightarrow a + b - 2c$$

The fundamental difference between similarity and distance measures is that the latter expressly include the absence of a feature (or low values for real-valued properties) in the measure of similarity. This has led to the suggestion (58) that, in the chemical domain at least, such measures are best for relative similarity; that is, ranking the similarity of two molecules to a target, as opposed to measuring the absolute similarity of molecules for which similarity measures, are preferred.

Similarity and distance measures form the basis for most of the analysis and selection methods described in the next section and the reader is referred to the reviews by Willett et al. (2, and references therein) for a fuller discussion of the characteristics and specific properties of these measures.

2.2 Analysis and Selection Methods

In this section we describe some general methods for analyzing and partitioning large data sets, with particular reference to selecting representative or diverse subsets. Library design also employs many of the strategies described here and is discussed in more detail in Section 4. The methods fall into two broad categories: cell-based or partitioning methods and distance-based methods. Partitioning methods use the population to define the limits for cells into which the compounds are divided. Adding or comparing to other compound sets requires identifying the cells into which the new compounds would fall based on their descriptors. This is very rapid and the partitioning process provides a frame of reference for many design tasks; for example, compounds can be readily identified to fill empty or poorly represented cells. Potential issues are where to place the cell boundaries and the handling of compounds that fall near to a cell boundary. Also, new compounds may fall outside the range of properties of the initial population. Distance-based methods, such as clustering and dissimilarity-based methods, require the calculation of similarity between members of the population and are thus population dependent. Adding new members to the population requires recalculating similarities and could change the distribution of compounds between the clusters. Identifying poorly represented or empty areas of property space is not possible. Each of these methods is further described below with examples of their application.