Chapter 20

Interactions Between Weighting Scheme and Similarity Coefficient in Similarity-Based Virtual Screening

John D. Holliday
University of Sheffield, UK

Peter Willett
University of Sheffield, UK

Hua Xiang
University of Sheffield, UK

ABSTRACT

Similarity searching is one of the most common methods for ligand-based virtual screening, and is normally carried out using the Tanimoto coefficient with binary fingerprints. However, a recent study has suggested that it may be less appropriate for use with weighted fingerprints in some circumstances. This paper compares the Tanimoto coefficient with other coefficients, and demonstrates that one of these, the cosine coefficient, exhibits a much greater degree of robustness in the face of variations in the nature of the fragment weighting scheme that is being used.

INTRODUCTION

Similarity searching is one of the most common forms of ligand-based virtual screening (e.g., the reviews by Eckert and Bajorath (2007), Geppert et al. (2010), and McGaughey et al. (2007). Given a known bioactive molecule, such as a hit from a high-throughput screening experiment or a compound from the literature, a similarity search involves matching the known molecule (often called the reference structure) against each of the structures in a database, computing the degree of similarity in each case, and then ranking the database structures in order of decreasing simi-
Interactions Between Weighting Scheme and Similarity Coefficient in Similarity-Based Virtual Screening

The similar property principle (Johnson & Maggiora, 1980; Martin et al., 2002) states that molecules that are structurally similar have similar properties, and the top-ranked structures from a similarity search are hence those that are most likely to exhibit the required bioactivity (Sheridan, 2007; Stumpfe & Bajorath, 2011; Willett, 2009).

The effectiveness of similarity searching, i.e., its ability to identify bioactive molecules, is determined by the similarity measure that determines the degree of resemblance between the reference structure and each of the database structures. A similarity measure has three components: the descriptors that are used to represent each of the molecules; the weighting scheme that is used to weight different parts of the representation to reflect their relative degrees of importance; and the similarity coefficient that quantifies the degree of resemblance between two weighted sets of descriptors. Although many types of descriptor have been used in similarity searching, by far the best established is a 2D fingerprint, a binary vector in which bits are set to denote the presence of fragment substructures in a molecule (Willett, 2006, 2009). Binary 2D fingerprints are normally used with the Tanimoto coefficient, a simple association coefficient in which the limiting values of zero and unity denote two fingerprints having no bits (and hence having no substructures) in common and two identical fingerprints, respectively. Many other types of coefficient can be used, but comparative experiments have demonstrated the general effectiveness of the Tanimoto coefficient, and this is the basis for similarity searching facilities in most operational chemoinformatics systems (Leach & Gillet, 2007).

There have been many comparisons of fingerprints and similarity coefficients for similarity searching, e.g., the detailed studies by Bender et al. (2009), Hert et al. (2004), Duan et al. (2010), and Sastry et al. (2010). Despite some limited early work (Willett & Winterman, 1986), there has been less interest in the use of weighted fingerprints, in which the elements of the vector contain not binary values denoting the presence or absence of fragment substructures, but integer or real values denoting the relative importance of the fragments. A fragment with a high weight occurring in both a reference structure and a database structure will then make a greater contribution to the overall degree of inter-molecular similarity than will a fragment in common that has a lesser weight. There are two main sources of frequency information that can be used for fragment weighting: weights based on the number of times that a fragment occurs in an individual molecule; and weights based on the number of times that a fragment occurs in an entire database. Both types of weighting have been studied in recent work by Arif et al. (2009, 2010), who found that the former type of weighting could bring about notable increases in screening effectiveness in some circumstances, but that the latter type was of less general applicability. We hence focus here on the former approach, i.e., on exploiting information on how frequently fragments occur within individual molecules.

Given its widespread usage with binary fingerprints, Arif et al. (2009) used the Tanimoto coefficient in their experiments on frequency-based weighting, but found that problems could arise that were absent when conventional binary fingerprints were being compared. Specifically, they found that even quite small variations in the weighting scheme could affect the magnitudes of the Tanimoto coefficients that are calculated in a similarity search; most notably they found that if there is a large discrepancy in the weights computed for the reference structure and for the database structure then screening effectiveness is likely to be markedly less than if the two weights are of comparable magnitude. This behavior was ascribed to the precise mathematical form of the Tanimoto coefficient, and it was suggested that other types of coefficient might be less affected by changes in the weighting scheme that was being used. Here, we seek to determine whether this is,