Feature trees: A new molecular similarity measure based on tree matching

Matthias Rarey\textsuperscript{a,}\textsuperscript{*} & J. Scott Dixon\textsuperscript{b,}\textsuperscript{**}
\textsuperscript{a}German National Research Center for Information Technology (GMD), Institute for Algorithms and Scientific Computing (SCAI), Schloß Birlinghoven, D-53754 Sankt Augustin, Germany; \textsuperscript{b}Smith\&Kline Beecham Pharmaceuticals, Physical and Structural Chemistry, 709 Swedeland Road, King of Prussia, PA 19406, U.S.A.

Received 28 October 1997; Accepted 20 March 1998

Key words: database screening, molecular descriptors, molecular similarity, molecular superposition, structural alignment

Summary
In this paper we present a new method for evaluating molecular similarity between small organic compounds. Instead of a linear representation like fingerprints, a more complex description, a feature tree, is calculated for a molecule. A feature tree represents hydrophobic fragments and functional groups of the molecule and the way these groups are linked together. Each node in the tree is labeled with a set of features representing chemical properties of the part of the molecule corresponding to the node. The comparison of feature trees is based on matching subtrees of two feature trees onto each other. Two algorithms for tackling the matching problem are described throughout this paper. On a dataset of about 1000 molecules, we demonstrate the ability of our approach to identify molecules belonging to the same class of inhibitors. With a second dataset of 58 molecules with known binding modes taken from the Brookhaven Protein Data Bank, we show that the matchings produced by our algorithms are compatible with the relative orientation of the molecules in the active site in 61\% of the test cases. The average computation time for a pair comparison is about 50 ms on a current workstation.

Introduction
In the absence of a three-dimensional structure of a target protein, molecular similarity between small organic compounds is the major concept in the search for new lead structures in drug design. There are several important applications for this concept. First of all, possible new lead structures can be retrieved from a database by searching for molecules similar to a known substrate or inhibitor [1]. Other applications are the clustering of molecules with known activity to identify common ways of binding and the search for a set of dissimilar molecules which may form a good set for screening experiments [2].

Because in most applications the number of molecules is very large, a time-efficient way of comparing them must be chosen. The most common approach is the generation of a linear descriptor, usually a bit- or an integer-string, for each molecule. The strings usually store the absence or occurrence of specific features of the molecule. The features are based either on the two-dimensional structural formula of the molecule (structural keys as used in [3] or hashed fingerprints as used in [4]) or on three-dimensional properties (line segments or triangles between possible pharmacophore points [5–9] or the ability of the molecule to bind to a specific protein [10]). A recent evaluation of a large variety of these methods by Brown and Martin [11] has shown that 2D structural descriptors have the best performance in separating active from inactive compounds.

At the other end of the scale of molecule comparison functions are methods based on a structural alignment of two (or a set of) molecules in three-
dimensional space (see for example [12–15] for algorithms). Because of their lengthy computational time, these methods are used for the identification of pharmacophores and in 3D-QSAR/CoMFA [16] analysis applied to small datasets rather than for the computation of similarity indices. A similarity value can be derived from the structural alignment by comparing the functional groups lying close to each other in three-dimensional space. For the intended application, a similarity measure which is derived by matching groups forming interactions with the same parts of the protein is highly appropriate.

In this paper, we present a new class of descriptors lying somewhere in between the classical descriptors mentioned above and the structural alignment tools. A molecule is described by a tree, called a feature tree. The nodes of the feature tree represent fragments of the molecule. Edges connect nodes which correspond to fragments which are connected in the two-dimensional structure of the molecule. Thus, the feature tree is a rough approximation to the structural formula of the molecule. Each node of the tree contains a set of features which are derived from the molecule fragment corresponding to the node.

To compare two feature trees, a matching between subtrees is computed and a similarity value is derived from this matching. In the Methods section, we describe two different comparison algorithms, called the split-search and the match-search algorithm. We have combined these algorithms with various features describing shape or physico-chemical properties of molecular fragments.

The concept of reducing molecular graphs to simpler forms is already used for representing and matching generic chemical structures for database retrieval [17]. In this work, we use this concept for the detection of molecular similarity. The reduced graph generation, the feature values, and the comparison algorithms are specially developed for this purpose resulting in an algorithmic framework allowing a very high degree of flexibility in the matching process.

In the Results section, we present the application of feature trees to two different datasets. The first set has been assembled by Briem [10] and contains 972 molecules from the MDDR (MACCS Drug Data Report) database [18], 35% of them are from 5 different inhibitor classes and the remainder are randomly selected. With this dataset, we verify the ability of our method to identify inhibitors from the same inhibitor class and compare the results to the Daylight fingerprint approach [4]. The second dataset has been assembled by Klebe and Lemmen (discussed in [15]) and contains 58 molecules taken from the PDB (Brookhaven Protein Data Bank) [19]. Because binding geometries relative to a protein are known, this dataset can be used to evaluate the quality of the matchings produced by the comparison algorithms.

Methods

Feature trees

A feature tree represents a molecule in the following way: a node of the tree describes a set of atoms of the molecule which are connected in the molecular graph. Each atom of the molecule is associated with at least one node. Two nodes which have atoms in common or which contain atoms connected in the molecular graph are connected. Note that in order to have a well-defined tree, the atom sets must be appropriately selected such that no cycles occur. An example of a feature tree is shown in Figure 1. In the following, the terms nodes and subtrees are used for describing feature trees, the term fragment is used to describe a part of the molecule. Subtrees and fragments are always assumed to be connected parts.

Feature trees are able to describe the molecule at various levels of resolution. At the lowest level, the whole molecule can be represented as a single atom set resulting in a feature tree with only one node. At the