Journal of Molecular Graphics and Modelling 66 (2016) 143–154

Contents lists available at ScienceDirect

Journal of Molecular Graphics and Modelling

journal homepage: www.elsevier.com/locate/JMGM

Fragment oriented molecular shapes

Ethan Hain [b], Carlos J. Camacho [a], David Ryan Koes [a,*]

[a] Department of Computational and Systems Biology, University of Pittsburgh, United States
[b] Department of Chemistry, St. Mary's College of Maryland, United States

ARTICLE INFO

Article history:
Received 27 January 2016
Received in revised form 23 March 2016
Accepted 29 March 2016
Available online 2 April 2016

Keywords:
Virtual screening
Molecular shape
Molecular similarity

ABSTRACT

Molecular shape is an important concept in drug design and virtual screening. Shape similarity typically uses either alignment methods, which dynamically optimize molecular poses with respect to the query molecular shape, or feature vector methods, which are computationally less demanding but less accurate. The computational cost of alignment can be reduced by pre-aligning shapes, as is done with the Volumetric-Aligned Molecular Shapes (VAMS) method. Here, we introduce and evaluate fragment oriented molecular shapes (FOMS), where shapes are aligned based on molecular fragments. FOMS enables the use of *shape constraints*, a novel method for precisely specifying molecular shape queries that provides the ability to perform partial shape matching and supports search algorithms that function on an interactive time scale. When evaluated using the challenging Maximum Unbiased Validation dataset, shape constraints were able to extract significantly enriched subsets of compounds for the majority of targets, and FOMS matched or exceeded the performance of both VAMS and an optimizing alignment method of shape similarity search.

1. Introduction

Molecular shape is a fundamental concept in medicinal chemistry [1], and shape-based virtual screens have successfully identified novel inhibitors [2–5]. Shape is used both to assess the similarity of a candidate molecule to a set of known actives and to evaluate the complementarity of a molecular shape to the shape of the binding site on the target receptor. Here, we describe a novel method for *exactly* specifying shape constraints for querying a large library of molecular shapes. The shape constraints are derived from both the receptor, which describes where a molecule cannot be, and from the shape of a known binder, which describes where the molecule should be. Our method is unique in relying on a manually positioned ligand *anchor fragment*. This anchor fragment requirement makes our approach particularly applicable to a fragment-based drug discovery workflow [6,7] as shape constraints are a natural way to search for compounds that extend an identified fragment structure while remaining complementary to the receptor. Additionally, the anchor fragment, by defining a fixed coordinate system, enables the indexing of large libraries of molecular shapes. This indexing allows search times to scale sub-linearly with the size of the library, resulting in search performance that is on an interactive time scale.

Shape-based virtual screening typically attempts to identify the most similar molecules in a virtual library to known active molecules or to a pseudo-ligand that is derived from the desired binding site [8]. Shape similarity is usually assessed either through alignment methods, which seek to maximize the three dimensional overlap of two shapes, or through feature vector methods, which transform shapes into a low-dimension vector of features that can be efficiently compared. As part of the similarity calculation, molecular shapes may be further annotated with electrostatic or pharmacophore features [9–14].

Alignment methods try to find the optimal overlay of two molecules to either maximize the overlapping volume or the correspondence between feature points, such as molecular field extrema [9,10]. The predominant method of maximizing volume overlap is to represent the molecular shapes as a collection of Gaussians [15,16], sample several starting points, and use numerical optimization to find a local maximum [17]. An alternative method is to represent the molecular shape by discrete features and use point correspondence algorithms to generate the alignment. Potential features include pharmacophore features [14], field points [9–11], or hyperbolical paraboloid representations of patches of molecular surface [18]. A number of performance optimizations to the alignment process have been described [2,14,16,19], but alignment methods remain computationally expensive relative to feature vector methods, unless shapes can be pre-aligned to a canonical coordinate system, as is done with Volumetric Aligned Molecular Shapes (VAMS) [20].

* Corresponding author.
E-mail address: dkoes@pitt.edu (D.R. Koes).

http://dx.doi.org/10.1016/j.jmgm.2016.03.017
1093-3263/

Feature vector methods reduce molecular shapes to a simple vector of Boolean or numerical features. Shape similarity is then determined by comparing these vectors using a metric such as Tanimoto or Euclidean distance. The numerical features can be computed using geometric moments [21,22], ray-tracing histograms [23], or a small set of reference shapes [24,25]. Feature vectors enable computationally efficient screening (millions of shape comparisons per a second) [21], but lack the accuracy and interpret-ability of alignment methods [1]. Critically, a feature vector similarity does not generate a molecular overlay suitable for visual inspection and analysis.

In our approach, fragment oriented molecular shapes (FOMS), we eliminate the computational burden of alignment by requiring the presence of a common anchor fragment. Molecules are trivially aligned by a direct superposition of anchor fragments, and the fragment defines a standard coordinate system for describing the shape of the molecule. Prepositioned molecular fragments are a common component of *de novo* drug design [26] where ligands are 'grown' from a prepositioned fragment to fit the binding site. Prepositioned fragments have also been successfully used in structure-based design to identify high-affinity inhibitors for a multitude of targets [27–30]. Notably, our AnchorQuery web service [31] provides interactive pharmacophore virtual screening of billions of custom compounds for protein–protein interaction (PPI) inhibitors by pre-aligning compounds to amino acid side-chain motifs corresponding to anchor residues [32,33] in the PPI interface. Fragment-based drug discovery workflows [6,7] can provide a physical basis for the selection and positioning of an appropriate anchor fragment. Alternatively, virtual docking methods may be used [34].

Anchor fragments present a different modality for shape-based screening: the user is required to identify a fragment structure with a meaningful binding mode and the search space is limited to compounds that contain the specified fragment. These requirements enable a new type of search language that supports explicit shape constraints. In essence, a *partial similarity search* [35] can be performed, where instead of optimizing similarity with the entirety of a query shape, the shape constraints specify only part of the shape in detail (e.g., within the binding site) while leaving other parts unspecified (e.g., interactions with solvent). In addition, the use of anchor fragments enables a new mechanism of search. Instead of evaluating the query against every molecule in the virtual screening library, the molecular shapes of the library can be *indexed* so that searches need only evaluate a fraction of the library. This allows large libraries of millions of shapes to be searched on an interactive time scale of a few seconds.

Here, we describe the retrospective virtual screening performance of FOMS and explore the potential for our explicit shape constraints to define highly specific filters for the creation of significantly enriched subsets of large virtual libraries.

2. Methods

We describe our representation of molecular shapes, how we define and construct shape constraints for searching, and briefly discuss how these shapes can be indexed to support efficient searches. We also describe our benchmarking methodology for performing a retrospective evaluation of FOMS.

2.1. Shape representation

A molecular shape in FOMS is a discretized solvent-excluded volume that is calculated from the heavy atoms of a molecular conformation using a water probe with radius 1.4 Å. The volume is discretized into 0.5 $Å^3$ voxels (three dimensional pixels) and stored as an oct-tree, an efficient data structure for representing volumetric data [37]. Most oct-tree operations take time proportional to the surface area of a shape ($\approx n^2$) instead of the volume ($\approx n^3$). An example of a voxelized shape is shown in Fig. 1. The grid coordinates of a molecular shape are all computed relative to the identified anchor fragment, which is aligned to the axes at the origin. If a molecule contains multiple instances of an anchor fragment or if there are inherent symmetries in the fragment, then a distinct voxelized shape is generated for every valid alignment of the fragment(s). For example, a fragment consisting of ring with a single exit vector will map to four distinct alignments on a ligand that contains a ring with two exit vectors. We grow/shrink molecular shapes by adding/removing the appropriate number of surface voxels.

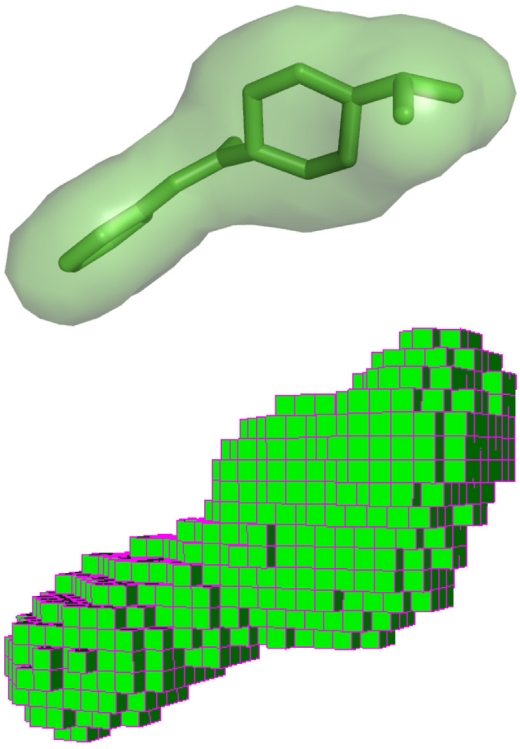

Fig. 1. The Rho-Kinase inhibitor from PDB 2H9V and the voxelization of its molecular, solvent-excluded volume. Voxlized image generated with Sproxel [36].

2.2. Shape constraints

Since we assume all shapes are registered to a common coordinate system defined by the anchor fragment, it is possible to exactly specify regions of space within this coordinate system that a molecule should and should not occupy. Following the nomenclature of VAMS [20], we refer to these constraints as minimum and maximum shape constraints. These constraints combine to form an expressive and exacting specification for a desired target shape. A *minimum shape constraint* sets a strict lower bound on the volumetric shape of a target shape. Every voxel within the minimum shape constraint must be contained within the target shape. A minimum shape constraint can be used to require that a target shape has a specific binding mode and minimum bulkiness. A *maximum shape constraint* sets a strict upper bound on the volumetric shape of a target shape. Every voxel of the target shape must be contained within the maximum shape constraint. The maximum shape can be used to constrain the total volume of the target shape and prevent the target shape from overlapping undesirable areas, such as space filled by a receptor. Shape constraints are distinct from shape similarity. Unlike shape similarity, which produces a continuous ranking