Published in final edited form as:

J Mol Graph Model. 2016 May ; 66: 143–154. doi:10.1016/j.jmgm.2016.03.017.

Fragment Oriented Molecular Shapes

Ethan Hain^b, Carlos J. Camacho^a, and David Ryan Koes^{a,*}

^aDepartment of Computational and Systems Biology, University of Pittsburgh ^bDepartment of Chemistry, St. Mary's College of Maryland

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.

Graphical abstract

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

*Corresponding Author: dkoes@pitt.edu (David Ryan Koes).

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 hyperbolic 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].

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 enable 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.

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.

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 1.4Å. The volume is discretized into 0.5Å³ 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 Figure 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.

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 of similarity with respect to a query shape, shape constraints are binary filters: a shape either matches the constraints or does not.

The structure of a receptor-ligand complex provides a natural starting point for the development of shape constraints. The inverse of the receptor shape provides a starting point for defining a maximum shape constraint. The constraint can be expanded within the binding site as needed to incorporate tolerance for small steric clashes and receptor flexibility while the shape can be shrunk to avoid excessive exposure of the target ligand to solvent. For visualization purposes we display the *inverse* of the maximum shape, which corresponds to the region of space the desired shape is excluded from, to more clearly visualize the minimum shape, which is, by definition, fully subsumed in the maximum shape.

Similarly, the ligand shape can be used to develop a minimum shape constraint. However, constructing a minimum shape constraint directly from the ligand shape, even if the shape is shrunk or skeletonized, results in a highly specific shape query that may be limited in its ability to retrieve novel chemical scaffolds. As an alternative to directly using the shape of a bound ligand, we derive *interaction points* from the ligand-receptor complex. Interaction points are points at the center of clusters of ligand atoms that are interacting with the receptor. Interacting ligand atoms are defined to be those no more than 6Å away from a receptor atom, as measured between atom centers. These atoms are then clustered into groups of three or more that are no more than 4Å across. The center of the cluster is an interaction point. An example of interaction points identified from a ligand-receptor complex is shown with the corresponding minimum and maximum shape constraints in Figure 2. Interaction points provide a more general specification of the binding mode of a ligand that is less dependent on the ligand chemical scaffold than the full ligand shape and ignores non-interacting components of the ligand.

Shape Indexing

As with VAMS[20], we use a matching and packing[38] bulk-loading algorithm to initialize an efficient data structure for volumetric shape constraint searches. Briefly, shapes are stored

in a GSS-tree[39] where each leaf of the tree is a single molecular shape and each internal node includes a maximum included volume (MIV) and minimum surrounding volume (MSV). The MSV is the union of all the molecular shapes beneath the node in the tree while the MIV is the intersection. By appropriately applying minimum and maximum shape constraints to the MIV and MSV, it can be determined if any of the shapes lower in the tree have the potential to match the constraints. If not, the entire subtree of molecular shapes can be eliminated from consideration, resulting in a sub-linear running time.

Shape Similarity

We use the shape Tanimoto [2] to compute the similarity of two voxelized shapes A and B:

$$\delta(A, B) = \frac{A \cap B}{A \cup B}$$

where a larger score indicates a greater degree of similarity. For shape similarity evaluations we consider only similarity with the query ligand.

Dataset

The specific modality of fragment-oriented molecular shapes requires the creation of a custom benchmark for assessing the virtual screening performance of the method. In order to construct the shape constraints, a receptor-ligand structure is required, and, in order to screen a library, all compounds in the library must contain the desired anchor fragment.

We use the Maximum Unbiased Validation (MUV) benchmark [40] as a starting point. MUV includes sets of 30 active and 15000 property-matched decoy compounds for each of 17 targets. Compounds are selected from Pub-Chem bioactivity data using a methodology that both reduces the similarity of actives (to avoid analogue bias [41]) and increases the similarity between actives and decoys (which helps prevent artificial enrichment [42]). MUV is also noteworthy in that the decoys were all assessed to be inactive in the initial high-throughput screen against the target providing some measure of experimental evidence that the negatives are true negatives (as opposed to schemes that generate decoys through random sampling [43]). The construction of the MUV dataset makes it particularly challenging for ligand similarity approaches. In an evaluation of several screening protocols (including molecular shape), poor enrichments were obtained for all protocols for the majority of query molecules [44]. The use of such a challenging benchmark allows us to critically evaluate FOMS and shape constraints in the context of a dataset where traditional approaches often fail.

Of the 17 targets in the MUV dataset, we identified 10 that had a receptor-ligand structure in the Protein Data Bank (PDB) where the ligand had sub- μM affinity. The interaction diagrams of these structures and their PDB codes are shown in Figure 3. For each of these structures we manually identified interacting fragments that could potentially serve as anchor fragments. For each target we selected relatively generic functional groups (at most 7 atoms) that were sufficiently common among both the actives and decoys to yield meaningful results and that were clearly forming interactions with the receptor. An exit

vector atom was included in generic ring fragments to limit the number of innate symmetries in the fragment and reduce the number for aligned poses. Chosen fragments are shown in Table 1 which includes the number of matching actives and decoys for each benchmark. A second fragment is chosen for the PKA target to illustrate the effect of fragment specificity.

Virtual Screening Evaluation

In order to investigate the utility of the FOMS approach we consider both the ability of shape constraint filters to generate enriched subsets and the quality of shape similarity rankings generated using fragment aligned molecules. We compare to VAMS, which aligns all molecules to a canonical reference system based on their moments of inertia, and rdMolAlign, the shape alignment module of RDKit [46] which aligns molecules using Open3DALIGN [47]. Open3DALIGN uses shape and property information to perform an atom-based alignment that attempts to match the largest number of atoms having similar partial charges and MMFF94 atom types. The input conformers for RDKit alignment were the same aligned poses used with VAMS, and the computed similarity score is the Tanimoto coefficient.

For each target, conformers of the active and decoy compounds were generated using RDKit [46]. A maximum of 100 conformers with a minimum RMSD difference of 0.7 Å and a maximum energy difference of 10 kcal/mol were generated for each compound. For each fragment considered, the corresponding conformers were extracted into fragment-specific subsets. The subsets were then preprocessed to create VAMS and FOMS search databases. The VAMS database stores a single pose for each conformation aligned along its moments of inertia. For FOMS, if a compound contains multiple instances of the anchor fragment or the fragment contains symmetries, multiple poses per a conformation are stored to account for the multiple fragments/symmetries.

For each target, a single reference structure was identified by searching BindingDB [48], PDBbind [49], and Binding MOAD [50] for the complex with the best binding affinity. The structures were visually inspected to identify buried anchor fragments that made key contacts with the protein, as shown in Figure 3 and Table 1. All shape queries were constructed using the reference complexes.

We evaluated both shape constraints and shape similarity using the single reference complex. For the shape constraint search we evaluated multiple maximum shape constraints by shrinking the receptor shape by 0, 0.5, 1.0, 1.5, and 2.0 Ångstroms. For the minimum shape constraints we used interaction points with minimal radius (one voxel). We evaluated every possible query of interaction points. That is, if n interaction points were identified, we evaluated 2^n minimum shapes. Selecting subsets of interaction points reduces the dependency of the query on the full shape of the reference molecule and more closely mimics the use case where the shape constraints are manually constructed by an expert to select only the most important interactions. As each shape constraint query is a filter, it does not return a total ranking of database compounds, so we report the true positive and false positive rate of the selected subset. We calculate a p-value for the returned subset (relative to the null hypothesis of random selection) using a hypergeometric test. Reported p-values are

corrected for multiple comparisons using the Bonferroni correction (the calculated p -value is multiplied by the number of comparisons).

Shape similarity results are reported using the area under the curve (AUC) of the receiver operating characteristic (ROC) curve. This metric is not biased by the number of compounds and has well-developed statistical properties [51]. The AUC is equivalent to the probability that a randomly selected active and randomly selected inactive compound will be correctly ranked by a method. An AUC of 0.5 corresponds to random chance while a perfect predictor exhibits an AUC of 1.0.

Results

Consistent with previous studies [44], we find the MUV dataset to be a challenging target for shape-based screening with few targets demonstrating AUCs far from random performance. Overall FOMS either matched or exceeded the virtual screening performance of VAMS and RDKit while retaining most of the benefits of pre-aligned molecules. Specifically, it is orders of magnitude faster than the optimizing alignment of RDKit suggesting, at a minimum, that FOMS is a viable method for rapidly pre-screening large libraries. In general, the Pareto frontier of the shape constraint queries delineates virtual screen performance equivalent or better to a full shape similarity search while performing queries orders of magnitude faster.

For each target, we plot the ROC curves for FOMS, VAMS, and RD-Kit similarity search as well as the results for every interaction point shape constraint query. Shape constraint results along the Pareto frontier are highlighted as solid circles and the most statistically significant result is annotated with its Bonferroni-corrected p -value. We also plot the total time required for the FOMS, VAMS, and RDKit similarity search and provide a box plot of the distribution of times for the shape constraint searches. All time plots are plotted on a logarithmic scale due to the orders of magnitude difference between the methods. Shape constraint search time varies considerably depending on the number of hits retrieved. Most queries are highly selective, return few or no results, and take less than a hundredth of a second. We do not include the time needed to generate the shape constraints in the query time since the receptor-based constraints are generated once and cached. We plot times for only those queries that generate results (queries that were too selective to return any matches were nearly instantaneous). For these queries the median time, as shown in the box plots, remains below a tenth of a second. A few queries are non-selective and return subsets of compounds that are comparable in size to the full database. These queries approach or exceed the running time of a full similarity search. The statistically most significant shape constraint query, annotated with a p -value in the ROC plot, has its running time plotted as a circle. These informative queries typically exceed the median time, but still take less than a second.

Cathepsin G

Virtual screening results for cathepsin G are shown in Figure 4. All three shape similarity methods exhibited poor performance with this query with AUCs near 0.5, but many of the interaction point shape constraint queries were significantly better than random. The

statistically most significant interaction point query is shown in Figure 5. The fragment for the query is deeply buried within the S1 pocket and the interaction points match the S2 pocket and phosphonate moiety of the ligand. However, there are no interaction points selected in the S3/S4 pocket. Interactions in this pocket can increase affinity, but are not necessary for binding [52]. Unlike whole-ligand similarity methods, shape constraint search can ignore a significant part of the geometry of the ligand (the naphthalene in the S3 pocket) and select for only specific parts of the query shape. In this case, this selectivity results in substantially better virtual screening results that is two orders of magnitude faster than they pre-aligned similarity search methods.

Estrogen Receptor α

Virtual screening results for estrogen receptor α agonists and inhibitors are shown in Figures 6 and 7. RDKit was marginally successful at identifying inhibitors with an AUC of 0.59 but the remaining methods were less successful with AUCs ranging from 0.43 to 0.52 and none of the shape-based methods exhibited early enrichment. The shape constraint queries lacked statistical significance and the most significant queries identified a large percentage of the total compounds in the database, resulting in running times that sometimes exceeded a second.

Estrogen Receptor β

Virtual screening results for estrogen receptor β are shown in Figure 8. None of the methods outperform random selection and all methods fail to exhibit early enrichment. Although shape constraints are capable of improving on the performance of the similarity methods, they still lack significance.

Focal Adhesion Kinase

Virtual screening results for focal adhesion kinase (FAK) are shown in Figure 9. Similar to ER β , none of the methods outperform random selection. However, there is at least one shape constraint query capable of generating an enriched subset with statistical significance. Note that unlike the other kinases, Rho and PKA, for FAK a nitrogen containing ring pattern was not chosen for fragment matching since fewer than half of the actives in the benchmark had such a pattern. Unlike the other kinases, using such a pattern for FAK does not improve the virtual screening results (data not shown).

Factor XIa

Virtual screening results for factor XIa (FXIA) are shown in Figure 10. All three shape similarity methods perform worse than random with AUCs ranging from 0.31 (VAMS) to 0.42 (RDKit and FOMS), with both VAMS and FOMS failing to identify any actives early in the screen. Shape constraints do not select enriched subsets and the pre-alignment methods exhibit no early enrichment behavior.

HIV Reverse Transcriptase

Virtual screening results for HIV reverse transcriptase (HIV-RT) are shown in Figure 11. Again, none of the methods outperform random selection. However, FOMS displays the best

early enrichment up to a true positive rate of 17% with RDKit also achieving some early enrichment. The shape constraint queries are statistically significant and can identify enriched subsets with greater proficiency than the shape similarity methods.

Heat Shock Protein 90

Virtual screening results for heat shock protein 90 (HSP90) are shown in Figure 12. All three shape similarity methods perform poorly with AUCs from 0.38 (VAMS) to 0.49 (RDKit). The shape constraint queries can produce statistically significant enrichments and the Pareto frontier matches or exceeds the ROC curve of RDKit.

Protein Kinase A

Virtual screening results for protein kinase A (PKA) are shown in Figure 13. For this kinase FOMS substantially outperforms both RDKit (AUC 0.64) and VAMS (AUC 0.55) with an AUC of 0.90. Shape constraints can perform even better than the full similarity search. Much of this performance is due to the use of a more selective fragment that properly orients the nitrogen in the ring fragment. Figure 14 shows the result when a more generic fragment, an arbitrary aromatic ring (PKA_{alt} in Table 1), is used. The enhanced performance of FOMS disappears, and the ability of shape constraints to generate enriched subsets is reduced. In comparison, the optimizing shape similarity of RDKit performs about the same (AUC 0.62). The use of this fragment pulls in thousands more decoys and introduces many more valid fragment alignments for both actives and decoys. This implies that the good performance of FOMS and shape constraints with this target is likely due to the ability of the fragment to properly position compounds to make conserved interactions with the hinge region of the kinase, as shown in Figure 3. Compounds that contain the specified fragment but cannot position the fragment to make these interactions while maintaining the steric compatibility of the whole ligand with the query ligand and receptor are filtered out.

Rho-Kinase2

Virtual screening results for Rho-Kinase2 are shown in Figure 15 and are similar to those of PKA. FOMS achieves an AUC of 0.94, VAMS an AUC of 0.56, and RDKit an AUC of 0.66. The best shape constraints match or exceed the performance of FOMS. As with PKA, if a less selective fragment is used (a generic aromatic ring), virtual performance is substantially reduced (data not shown).

Discussion

The MUV dataset, with its focus on eliminating analogue bias, is particularly resistant to single-query shape-based virtual screens [44]. This is reflected in our overall results, shown in Figure 16, where only two targets (Rho and PKA) achieve AUCs where the 95% confidence interval does not overlap with 0.5 (random performance). The remaining targets likely lack meaningful whole-molecule shape complementarity between the query ligand and the active compounds of the benchmark. One exception may be HIV-rt, where there is clear early enrichment which indicates that a subset of the actives may be compatible with the query molecule. FOMS dramatically outperformed other methods for the Rho and PKA targets due to correct positioning of a fragment with key, conserved interactions. For

comparison, Figure 16 also shows the performance of a 2D fingerprint, the OpenBabel FP2 [53] path-based fingerprint. 2D information is more successful for three targets (ER β , FXIa, HSP90), and has comparable or worse performance for the remaining targets, illustrating the orthogonality of 2D and 3D approaches.

Our assumption in the design of this study was that, since the evaluated methods utilize rigid conformers, it would be necessary to generate a large sampling of conformers to ensure the biologically relevant conformer was screened and was correctly registered as a hit.

Surprisingly, when we tested this assumption with the Rho Kinase target (results were similar for PKA), we found that the virtual screening performance for all three shape methods was relatively insensitive to the number of conformers used (Figure 17). Interestingly, this is not because the top ranked conformer is the most representative conformer. Instead, as demonstrated in Figure 17, reducing the number of conformers does result in a reduction in similarity scores for active compounds. However, a compensating reduction in scores is observed in the decoy set as the number of conformers sampled is decreased, resulting in similar virtual screening results. This suggests that sampling a large number of conformations will be beneficial in terms of producing meaningful poses with high similarity to the query ligand, a key advantage of shape-based methods, but, counter-intuitively, may not be strictly necessary for virtual screening purposes.

FOMS essentially provides a rapid means of template docking [31, 54, 55] using shape-based scoring. The disadvantage of fragment-oriented approaches is they are critically dependent on the choice of fragment and its proper positioning in defining the query. Provided these requirements can be met, there are several advantages to shape-based fragment alignment search. By enforcing the fragment alignment, key interaction are guaranteed to be conserved. Previous studies have demonstrated the importance of adding pharmacophoric properties (or 'color') to shape similarity [17]. Fragment alignment introduces a hard bias toward matching a key portion of the query molecule without introducing any additional computation, as required by more general methods. In fact, as we have shown, pre-alignment substantially reduces the computational overhead.

Prealignment, whether to fragments (FOMS) or canonical internal coordinates (VAMS) is orders of magnitude faster than methods that dynamically optimize the alignment. This holds true even if the cost to create the search database is taken into account. The time to create the databases scales with the number of molecular shapes (about 10 shapes a second on our system) and compares favorably with RDKit search (2 molecules a second). As the common case is for a fragment oriented database to be re-used to for queries by multiple users investigating multiple targets, in practice the cost of database creation gets amortized into insignificance.

A major advantage of fragment alignment is that it enables the use of shape constraints. Shape constraint search generally tracked or improved upon the performance of FOMS similarity ranking (e.g. Figure 4). As shown in Table 2, shape constraints were able to generate statistically significant ($p < 0.01$) enriched subsets for six of the ten targets. Unlike whole-molecule shape similarity, shape constraints can select for a *subshape* of the query ligand (i.e., perform partial shape matching) and specifically filter out potential clashes with

the receptor. Furthermore, shape constraints support an indexing search algorithm that scales sub-linearly. In our evaluation, shape constraint searches were orders of magnitude faster than even VAMS similarity search and generally took well under a second.

Shape constraints provide a novel and unique method for specifying molecular shape queries. Although we have investigated automatically generated shape constraints, our assumption is that intelligently designed constraints created by human experts will substantially outperform these interaction point based constraints. The speed of shape constraint searches enables interactive analysis and refinement of shape queries where the expert user sculpts their desired molecule guided by the results of searches of full-sized databases. We anticipate that fragment-oriented shape constraints will prove a useful tool in both fragment-based drug discovery workflows, which are naturally centered around a fragment, and lead optimization workflows, where the core scaffold of the lead compound can serve as a fragment.

Acknowledgments

This work was supported by the National Institute of Health [R01GM097082, R01GM108340].

References

1. Nicholls A, McGaughey GB, Sheridan RP, Good AC, Warren G, Mathieu M, Muchmore SW, Brown SP, Grant JA, Haigh JA. Molecular shape and medicinal chemistry: a perspective. *J Med Chem.* 2010; 53(10):3862.10.1021/jm900818s [PubMed: 20158188]
2. Rush TS III, Grant JA, Mosyak L, Nicholls A. A shape-based 3-D scaffold hopping method and its application to a bacterial protein-protein interaction. *J Med Chem.* 2005; 48(5):1489–1495.10.1021/jm040163o [PubMed: 15743191]
3. McMasters DR, Garcia-Calvo M, Maiorov V, McCann ME, Meurer RD, Bull HG, Lisnock JM, Howell KL, DeVita RJ. Spiroimidazolidinone NPC1L1 inhibitors. 1: Discovery by 3D-similarity-based virtual screening. *Bioorganic & medicinal chemistry letters.* 2009; 19(11):2965–2968.10.1016/j.bmcl.2009.04.031 [PubMed: 19410454]
4. Muchmore SW, Souers AJ, Akritopoulou-Zanze I. The Use of Three-Dimensional Shape and Electrostatic Similarity Searching in the Identification of a Melanin Concentrating Hormone Receptor 1 Antagonist. *Chemical biology & drug design.* 2006; 67(2):174–176.10.1111/j.1747-0285.2006.00341.x [PubMed: 16492165]
5. Naylor E, Arredouani A, Vasudevan SR, Lewis AM, Parkesh R, Mizote A, Rosen D, Thomas JM, Izumi M, Ganesan A, Galione A, Churchill GC. Identification of a chemical probe for NAADP by virtual screening. *Nature Chemical Biology.* 2009; 5(4):220–226. URL <http://dx.doi.org/10.1038/nchembio.150>. 10.1038/nchembio.150 [PubMed: 19234453]
6. Rees DC, Congreve M, Murray CW, Carr R. Fragment-based lead discovery. *Nature Reviews Drug Discovery.* 2004; 3(8):660–672.10.1038/nrd1467 [PubMed: 15286733]
7. Congreve M, Chessari G, Tisi D, Woodhead AJ. Recent Developments in Fragment-Based Drug Discovery. *J Med Chem.* 2008; 51(13):3661–3680. [PubMed: 18457385]
8. Ebalunode JO, Ouyang Z, Liang J, Zheng W. Novel Approach to Structure-Based Pharmacophore Search Using Computational Geometry and Shape Matching Techniques. *J Chem Inf Model.* 2008; 48(4):889–901. [PubMed: 18396858]
9. Vainio MJ, Puranen JS, Johnson MS. ShaEP: molecular overlay based on shape and electrostatic potential. *Journal of chemical information and modeling.* 2009; 49(2):492–502.10.1021/ci800315d [PubMed: 19434847]
10. Cheeseright T, Mackey M, Rose S, Vinter A. Molecular field extrema as descriptors of biological activity: definition and validation. *Journal of chemical information and modeling.* 2006; 46(2):665–676.10.1021/ci050357s [PubMed: 16562997]

11. Thorner DA, Wild DJ, Willett P, Wright PM. Similarity searching in files of three-dimensional chemical structures: flexible field-based searching of molecular electrostatic potentials. *J Chem Inf Comput Sci.* 1996; 36(4):900–908.10.1021/ci960002w
12. Tervo AJ, Rönkkö T, Nyrönen TH, Poso A. BRUTUS: optimization of a grid-based similarity function for rigid-body molecular superposition. 1. Alignment and virtual screening applications. *Journal of medicinal chemistry.* 2005; 48(12):4076–4086.10.1021/jm049123a [PubMed: 15943481]
13. Marín RM, Aguirre NF, Daza EE. Graph theoretical similarity approach to compare molecular electrostatic potentials. *Journal of chemical information and modeling.* 2008; 48(1):109–118.10.1021/ci7001878 [PubMed: 18166018]
14. Sastry M, Dixon S, Sherman W. Rapid Shape-Based Ligand Alignment and Virtual Screening Method Based on Atom/Feature-Pair Similarities and Volume Overlap Scoring. *J Chem Inf Model.* 10.1021/ci2002704
15. Good AC, Richards WG. Rapid evaluation of shape similarity using Gaussian functions. *J Chem Inf Model.* 1993; 33(1):112–116.
16. Grant JA, Gallardo MA, Pickup BT. A fast method of molecular shape comparison: A simple application of a Gaussian description of molecular shape. *Journal of Computational Chemistry.* 1996; 17(14):1653–1666.
17. Hawkins PCD, Skillman AG, Nicholls A. Comparison of Shape-Matching and Docking as Virtual Screening Tools. *J Med Chem.* 2007; 50(1):74–82. URL <http://dx.doi.org/10.1021/jm0603365>. 10.1021/jm0603365 [PubMed: 17201411]
18. Proschak E, Rupp M, Derksen S, Schneider G. Shapelets: Possibilities and limitations of shape-based virtual screening. *Journal of Computational Chemistry.* 2008; 29(1):108–114.10.1002/jcc.20770 [PubMed: 17516427]
19. Fontaine F, Bolton E, Borodina Y, Bryant SH. Fast 3D shape screening of large chemical databases through alignment-recycling. *Chemistry Central Journal.* 2007; 1:12.10.1186/1752-153X-1-12 [PubMed: 17880744]
20. Koes DR, Camacho CJ. Shape-based virtual screening with volumetric aligned molecular shapes. *Journal of Computational Chemistry.* 2014; 35(25):1824–1834. URL <http://dx.doi.org/10.1002/jcc.23690>. 10.1002/jcc.23690 [PubMed: 25049193]
21. Ballester PJ, Richards WG. Ultrafast shape recognition to search compound databases for similar molecular shapes. *J Comp Chem.* 2007; 28(10):1711.10.1002/jcc.20681 [PubMed: 17342716]
22. Schreyer AM, Blundell T. USRCAT: real-time ultrafast shape recognition with pharmacophoric constraints. *Journal of Cheminformatics.* 2012; 4(1):27. URL <http://dx.doi.org/10.1186/1758-2946-4-27>. 10.1186/1758-2946-4-27 [PubMed: 23131020]
23. Zauhar RJ, Moyna G, Tian LF, Li ZJ, Welsh WJ. Shape signatures: a new approach to computer-aided ligand-and receptor-based drug design. *J Med Chem.* 2003; 46(26):5674–5690.10.1021/jm030242k [PubMed: 14667221]
24. Haigh JA, Pickup BT, Grant JA, Nicholls A. Small molecule shape-fingerprints. *J Chem Inf Model.* 2005; 45(3):673–684.10.1021/ci049651v [PubMed: 15921457]
25. Putta S, Lemmen C, Beroza P, Greene J. A novel shape-feature based approach to virtual library screening. *J Chem Inf Comput Sci.* 2002; 42(5):1230–1240. [PubMed: 12377013]
26. Schneider G, Fechner U. Computer-based de novo design of drug-like molecules. *Nat Rev Drug Discov.* 2005; 4(8):649–663. URL <http://dx.doi.org/10.1038/nrd1799>. [PubMed: 16056391]
27. Kick EK, Roe DC, Geoffrey Skillman A, Liu G, Ewing TJA, Sun Y, Kuntz ID, Ellman JA. Structure-based design and combinatorial chemistry yield low nanomolar inhibitors of cathepsin D. *Chemistry & Biology.* 1997; 4(4):297–307. [PubMed: 9195867]
28. Murray CW, Clark DE, Auton TR, Firth MA, Li J, Sykes RA, Waszkowycz B, Westhead DR, Young SC. PRO SELECT: combining structure-based drug design and combinatorial chemistry for rapid lead discovery. 1. Technology. *J Comput Aided Mol Des.* 1997; 11(2):193–207. [PubMed: 9089436]
29. Li J, Murray CW, Waszkowycz B, Young SC. Targeted molecular diversity in drug discovery: integration of structure-based design and combinatorial chemistry. *Drug discovery today.* 1998; 3(3):105–112.

30. Liebeschuetz JW, Jones SD, Morgan PJ, Murray CW, Rimmer AD, Roscoe JME, Waszkowycz B, Welsh PM, Wylie WA, Young SC. PRO SELECT: combining structure-based drug design and array-based chemistry for rapid lead discovery. 2. The development of a series of highly potent and selective factor Xa inhibitors. *Journal of medicinal chemistry*. 2002; 45(6):1221–1232. [PubMed: 11881991]
31. Koes D, Khoury K, Huang Y, Wang W, Bista M, Popowicz GM, Wolf S, Holak TA, Dömling A, Camacho CJ. Enabling Large-Scale Design Synthesis and Validation of Small Molecule Protein-Protein Antagonists. *PLoS ONE*. 2012; 7(3):e32839. URL <http://dx.doi.org/10.1371/journal.pone.0032839>. 10.1371/journal.pone.0032839 [PubMed: 22427896]
32. Rajamani D, Thiel S, Vajda S, Camacho CJ. Anchor residues in protein-protein interactions. *Proceedings of the National Academy of Sciences*. 2004; 101(31):11287–11292. URL <http://dx.doi.org/10.1073/pnas.0401942101>. 10.1073/pnas.0401942101
33. Meireles LMC, Domling AS, Camacho CJ. ANCHOR: a web server and database for analysis of protein-protein interaction binding pockets for drug discovery. *Nucleic Acids Research*. 2010; 38(Web Server):W407–W411. URL <http://dx.doi.org/10.1093/nar/gkq502>. 10.1093/nar/gkq502 [PubMed: 20525787]
34. Brenke R, Kozakov D, Chuang GY, Beglov D, Hall D, Landon MR, Mattos C, Vajda S. Fragment-based identification of druggable hot spots of proteins using Fourier domain correlation techniques. *Bioinformatics*. 2009; 25(5):621. 10.1093/bioinformatics/btp036 [PubMed: 19176554]
35. Bronstein AM, Bronstein MM, Bruckstein AM, Kimmel R. Partial similarity of objects, or how to compare a centaur to a horse. *International Journal of Computer Vision*. 2009; 84(2):163–183.
36. proxel, r173, <http://code.google.com/p/sproxel/>.
37. Zhang J, Smith S. Shape Similarity Matching With Octree Representations. *Journal of Computing and Information Science in Engineering*. 2009; 9:034503.
38. Koes DR, Camacho CJ. Indexing volumetric shapes with matching and packing. *Knowledge and Information Systems*. 2014; 43(1):157–180. URL <http://dx.doi.org/10.1007/s10115-014-0729-z>. 10.1007/s10115-014-0729-z [PubMed: 26085707]
39. Keim, DA. Efficient geometry-based similarity search of 3D spatial databases. *Proc of the Intl Conf on Management of Data*; ACM, New York, NY, USA. 1999. p. 419–430. URL <http://doi.acm.org/10.1145/304182.304219>
40. Rohrer SG, Baumann K. Maximum unbiased validation (MUV) data sets for virtual screening based on PubChem bioactivity data. *J Chem Inf Model*. 2009; 49(2):169–84. 10.1021/ci8002649 [PubMed: 19434821]
41. Good AC, Oprea TI. Optimization of CAMD techniques 3. Virtual screening enrichment studies: a help or hindrance in tool selection? *Journal of computer-aided molecular design*. 2008; 22(3):169–178. [PubMed: 18188508]
42. Verdonk ML, Berdini V, Hartshorn MJ, Mooij WT, Murray CW, Taylor RD, Watson P. Virtual screening using protein-ligand docking: avoiding artificial enrichment. *Journal of chemical information and computer sciences*. 2004; 44(3):793. 10.1021/ci034289q [PubMed: 15154744]
43. Mysinger MM, Carchia M, Irwin JJ, Shoichet BK. Directory of Useful Decoys Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking. *J Med Chem*. 2012; 55(14):6582–6594. URL <http://dx.doi.org/10.1021/jm300687e>. 10.1021/jm300687e [PubMed: 22716043]
44. Tiikkainen P, Markt P, Wolber G, Kirchmair J, Distinto S, Poso A, Kallioniemi O. Critical Comparison of Virtual Screening Methods against the MUV Data Set. *Journal of Chemical Information and Modeling*. 2009; 49(10):2168–2178. URL <http://dx.doi.org/10.1021/ci900249b>. 10.1021/ci900249b [PubMed: 19799417]
45. Stierand K, Maass PC, Rarey M. Molecular complexes at a glance: automated generation of two-dimensional complex diagrams. *Bioinformatics*. 2006; 22(14):1710–1716. URL <http://dx.doi.org/10.1093/bioinformatics/btl150>. 10.1093/bioinformatics/btl150 [PubMed: 16632493]
46. Landrum, G. RDKit: Open-source cheminformatics. <http://www.rdkit.org>, [Online; accessed 25-October-2015]

47. Tosco P, Balle T, Shiri F. Open3DALIGN: an open-source software aimed at unsupervised ligand alignment. *Journal of Computer-Aided Molecular Design*. 2011; 25(8):777–783. URL <http://dx.doi.org/10.1007/s10822-011-9462-9>. 10.1007/s10822-011-9462-9 [PubMed: 21792629]
48. Liu T, Lin Y, Wen X, Jorissen RN, Gilson MK. BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Research*. 2007; 35:D198–D201. Database. URL <http://dx.doi.org/10.1093/nar/gkl999>. 10.1093/nar/gkl999 [PubMed: 17145705]
49. Wang R, Fang X, Lu Y, Yang CY, Wang S. The PDBbind Database: Methodologies and Updates. *J Med Chem*. 2005; 48(12):4111–4119. URL <http://dx.doi.org/10.1021/jm048957q>. 10.1021/jm048957q [PubMed: 15943484]
50. Hu L, Benson ML, Smith RD, Lerner MG, Carlson HA. Binding MOAD (Mother Of All Databases). *Proteins*. 2005; 60(3):333–340. URL <http://dx.doi.org/10.1002/prot.20512>. 10.1002/prot.20512 [PubMed: 15971202]
51. Jain AN. Bias reporting, and sharing: computational evaluations of docking methods. *Journal of Computer-Aided Molecular Design*. 2007; 22(3–4):201–212. URL <http://dx.doi.org/10.1007/s10822-007-9151-x>. 10.1007/s10822-007-9151-x [PubMed: 18075713]
52. de Garavilla L, Greco MN, Sukumar N, Chen ZW, Pineda AO, Mathews FS, Cera ED, Giardino EC, Wells GI, Haertlein BJ, Kauffman JA, Corcoran TW, Derian CK, Eckardt AJ, Damiano BP, Andrade-Gordon P, Maryanoff BE. A Novel Potent Dual Inhibitor of the Leukocyte Proteases Cathepsin G and Chymase: MOLECULAR MECHANISMS AND ANTI-INFLAMMATORY ACTIVITY IN VIVO. *Journal of Biological Chemistry*. 2005; 280(18):18001–18007. URL <http://dx.doi.org/10.1074/jbc.M501302200>. 10.1074/jbc.m501302200 [PubMed: 15741158]
53. Boyle NMO, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR. Open Babel: An open chemical toolbox. *Journal of Cheminformatics*. 2011; 3(1):33. URL <http://dx.doi.org/10.1186/1758-2946-3-33>. 10.1186/1758-2946-3-33 [PubMed: 21982300]
54. Ruiz-Carmona S, Alvarez-Garcia D, Foloppe N, Garmendia-Doval AB, Juhos S, Schmidtke P, Barril X, Hubbard RE, Morley SD. rDock: A Fast Versatile and Open Source Program for Docking Ligands to Proteins and Nucleic Acids. *PLoS Comput Biol*. 2014; 10(4):e1003571. URL <http://dx.doi.org/10.1371/journal.pcbi.1003571>. 10.1371/journal.pcbi.1003571 [PubMed: 24722481]
55. Abagyan, R.; Orry, A.; Raush, E.; Totrov, M. ICM Manual, 3.8. MolSoft LLC; La Jolla, CA:

***Highlights (for review)**

- * Fragment Oriented Molecular Shapes (FOMS) pre-aligns molecules to a user-specified fragment
- * FOMS pre-alignment enabled rapid, sub-linear time search of exact shape specifications
- * With the right choice of fragment and target FOMS can outperform alternative shape methods

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

Figure 2.

(Left) The Rho-Kinase inhibitor (sticks) and receptor (surface) from PDB 2H9V shown with the identified interaction points (spheres). (Right) Shape constraints derived from this structure. The interaction points become a minimum shape constraint (yellow) by applying a radius of 1 Å. Here the receptor volume has been shrunk by 1 Å (two layers of surface voxels were removed) to form a maximum shape constraint (inverse shown in cyan). Voxlized image generated with Sproxel [36].

Figure 3.
Graphical depictions of the ten receptor-ligand structures used to define shape constraints.
The fragment of the ligand used for alignment is highlighted. Diagrams generated with
PoseView [45].

Figure 4.
Cathepsin G virtual screening performance and search time for various methods.

Figure 5. Cathepsin G interaction points that defined the shape constraint query with the most significant enrichment. The interaction points (magenta spheres) and fragment (magenta sticks) delineate only half the molecule and are located near a charge-charge interaction and aromatic interaction. Unlike molecular similarity, shape constraints are capable of selecting molecules that sterically match these features while ignoring the remainder of the query ligand shape and avoiding significant clashes with the receptor shape (for this query the receptor was reduced by 1.5Å).

Figure 6.
ERa agonists virtual screening performance and search time for various methods.

Figure 7.
ERA inhibitors virtual screening performance and search time for various methods.

Figure 8.
ER β virtual screening performance and search times for various methods.

Figure 9.
FAK virtual screening performance and search times for various methods.

Figure 10.
Factor XIa virtual screening performance and search times for various methods.

Figure 11. HIV reverse transcriptase virtual screening performance and search times for various methods.

Figure 12.
Heat shock protein 90 virtual screening performance and search times for various methods.

Figure 13.
Protein kinase A virtual screening performance and search times for various methods.

Figure 14.

Protein kinase A virtual screening performance and search times for various methods with a less selective fragment choice.

Figure 15.
Rho-kinase 2 virtual screening performance and search times for various methods.

Figure 16.
AUCs for shape similarity methods across all targets shown with their 95% confidence intervals. Confidence intervals were calculated using 2000 iterations of bootstrapping.

Figure 17.

(Left) The effect of the maximum number of conformers generated for each molecule on screening performance, and (Right) the change in distribution of similarity scores for active and decoy compounds using FOMS for different numbers of conformers.

Table 1

SMARTS expressions for anchor fragments evaluated in each MUV benchmark shown with the number of active and decoy compounds from the benchmark that contain that fragment. Every MUV benchmark starts with 30 active compounds and 15000 decoys.

Target	SMARTS	Confirmatory Assay ID	Actives	Decoys
CathG	<chem>c1ccccc1[!H]</chem>	832	24	12847
ER α agonist	<chem>c1ccccc1[!H]</chem>	737	29	13390
ER α	<chem>c1ccccc1[!H]</chem>	713	22	13063
ER β	<chem>c1ccccc1[!H]</chem>	733	23	12839
FAK	<chem>a1aaaaa1[!H]</chem>	810	29	13891
FXIa	<chem>a1aaaaa1[!H]</chem>	846	24	7691
HIVrt	<chem>a1aaaaa1[!H]</chem>	652	18	9535
HSP90	<chem>c1ccccc1[!H]</chem>	712	29	12915
PKA	<chem>c1[c,n]cccn1</chem>	466	25	5582
PKA _{alt}	<chem>a1aaaaa1[!H]</chem>	466	29	13985
Rho	<chem>c1[c,n]cccn1</chem>	644	18	4593

Table 2

Significance of the best shape constraint queries and the percentage of all evaluated queries with a Bonferroni-corrected p -value below 0.01.

Target	Best p -value	Sig. Queries
CathG	4.23e-06	1%
ER α	0.0505	0%
ER α agonist	0.0106	0%
ER β	0.0328	0%
FAK	0.000289	4%
FXIa	0.387	0%
HIVrt	0.000951	5%
HSP90	0.000755	2%
PKA	1.14e-23	80%
Rho	8.41e-15	80%