Shape-based classification of bound ligands

Thomas Funkhouser*[1,2], Fabian Glaser[1], Roman Laskowski[1], Richard Morris[3], Rafael Najmanovich[1], Gareth Stockwell[1], and Janet Thornton[1]

[1] European Bioinformatics Institute
[2] Princeton University
[3] John Innes Centre

1 Introduction

Shape-based classification of molecular structures is important for protein function prediction and virtual ligand screening. The ultimate goal is to predict which ligands will bind to a given protein based on the shapes and properties of protein binding sites.

Two steps are usually required to compute the similarities between protein binding sites: 1) modeling the geometric/chemical/physical properties of the bound ligands, and 2) computing a similarity measure for pairs of those models. In this paper, we focus on the second step – we assume that the geometry and element types of bound ligands are known, taking data from the crystallized structures in the PDB, and study the effectiveness of algorithms that match those structures. Our goal is to determine whether a simple algorithm for matching rigid molecular structures can be effective at classifying ligand types.

Our study is based on a 3D data set of ligands with atoms in the conformations in which they are found when bound to non-homologous proteins in the PDB. We used a simple geometric matching algorithm to compute the similarity between all pairs of these ligands and a nearest neighbor classifier to predict their molecule types (NAD, ATP, ADP, etc.) from the computed similarity values in a "leave-one-out" classification experiment. The results show that even a rigid shape matching algorithm is able to predict the correct type for 92% of the ligands. While this classification performance is not perfect, it is far better than the 3.5% that would be achieved by random chance, and it is remarkably good considering that the data set contained several very similar and flexible molecule types (e.g., ATP, ACP, ANP, GTP, ADP, GDP, AMP).

These results suggest two conclusions. First, the conformational variation of ligands bound to proteins in the PDB is not so great that it completely thwarts a rigid shape matching algorithm. Quite to the contrary, it seems that most ligands appear in a small number of different conformations even when bound to non-homologous proteins, and a nearest neighbor classifier is usually able to pick out another example of the same type in a similar conformation. Second, current geometric matching algorithms seem to provide fairly good performance (92%) for predicting molecule types from descriptions of ligands comprising atom locations and element types. Thus, methods that derive such descriptions from protein structures without bound ligands may be able to use the geometric matching algorithm described in this paper to achieve good ligand binding prediction results.

2 Geometric matching algorithm

Our geometric matching algorithm is based on the Harmonic Shape Descriptor (HSD), a shape representation recently described for computer graphics by Kazhdan, 2004 (Figure 1). Intuitively, this descriptor represents the shape of a molecule with a concise set of numbers (ampli-

tudes of the spherical harmonic coefficients within every radius and frequency around the center of mass) that can be compared efficiently to produce a shape similarity measure. It was chosen as a representative of shape descriptors suitable for virtual screening of large databases.

The first step in building the Harmonic Shape Descriptor for a given molecule is to compute a 3D field for each element type (C, O, N, P, and S) representing the locations of atoms of that type. Each 3D field used in our study is a Gaussian function of the Euclidean distance from every point in space to the closest atom of the given element type. The distance between the fields for two molecules provides a smoothly decreasing penalty function for misalignment of atoms of the same element type.

The second step is to build a representation of these 3D fields that can be matched efficiently (i.e., without search over all possible alignments). The main idea is to decompose each 3D field into rotationally independent components and then store the magnitude of the field within each of the components. The HSD decomposition first partitions the field into concentric spherical shells (that rotate independently) and then it partitions each shell into spherical harmonic frequencies (that also rotate independently). The amplitudes of the spherical harmonic coefficients for each radius and each frequency (a 2D array of numbers) provide a shape descriptor that is concise, descriptive, rotation-invariant, and quick to compare. We use the L^2 distance between any two HSD descriptors to represent the dissimilarity of the molecules they represent.

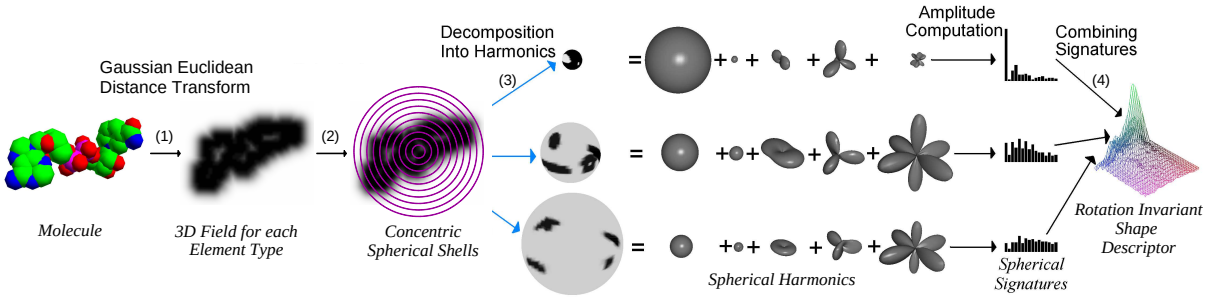

Figure 1: Steps in computing the Harmonic Shape Descriptor: 1) a field is built for each element type (C, O, N, P, S) containing a Gaussian of the Euclidean distance to closest atom of that type, 2) the field is decomposed into a set of concentric spherical shells around the center of mass, 3) the field on each sphere is decomposed into spherical harmonics, 4) the amplitudes of the harmonic coefficients within each frequency of each sphere form a rotation-invariant shape descriptor.

This approach has several advantages over previous methods for molecular shape description, including others based on spherical harmonics (e.g., Duncan *et al.*, 1993; Ritchie *et al.*, 1999; Cai *et al.*, 2002; and Morris *et al.*, 2005). First, it represents any shape, not just star-shaped surfaces. Second, it is invariant to the orientation of the molecule, and thus comparison of the descriptors can be performed without aligning the orientations of molecules and without search over all possible rotations. Third, it provides a provable guarantee that the L^2 distance between two descriptors provides a lower bound on the L^2 distance between the corresponding 3D fields. Finally, it enables efficient indexing based on nearest neighbor search structures, and thus is suitable for use with large databases. In our studies, the time required to compute the descriptor for any given molecule was 2 seconds on average, and the time to find the closest match for any query was less than a millisecond.

3 Data set

Our data set comprises a set of ligands with several examples for each of many different molecule types in the conformations in which they are found when bound to non-homologous proteins in the PDB.

To build this data set, we scanned all crystallized PDB structures with bound ligands (as of October 2004). For simplicity sake, we considered only ligands appearing in a single PDB residue, containing only HETATMS (i.e., no polypeptides), and not covalently bonded to another residue. We then classified each ligand into a molecule type (e.g., ATP, ADP, NAD, ...), verifying that the number, elements, and connectivities of its atoms matched the structure found in the Macromolecular Structure Database, thereby eliminating partial or modified molecules. We also eliminated small molecule types commonly found in crystallization solutions (e.g., SO4, PO4, ...), as they are less likely to be bound in a biologically relevant binding site. Next, in order to avoid evolutionary relationships between ligands in of the same molecule type, we computed all protein-ligand contacts and associated to each ligand the CATH code of the protein domain with the greatest percentage of its contacts, and only one ligands of each molecule type was retained from each CATH superfamily. As a side effect, ligands contacting domains for which no CATH code was available were eliminated as well. Finally, we kept only molecule types containing at least five remaining representatives.

The net result is a set of 545 ligands partitioned into 47 molecule types, with all representatives of every molecule type bound to non-homologous protein domains.

4 Results

We tested whether the geometric matching algorithm can be used to classify the ligands in their bound conformations by using each ligand as a query, matching it to all others in the data set, and reporting how often the nearest neighbor is a member of the same molecule type.

Table 1 summarizes the results. For every molecule type (rows of the table), the columns list from left to right the molecule name, the number of heavy atoms in the molecule, the number of queries of that type, the number correctly classified, the number incorrectly classified (with the types of false matches in parenthesis), and the percentage correctly classified (i.e., the classification rate).

Overall, the classification rate for all ligands in the data set is 92%. Looking in more detail, we observe that the classification rate is usually 100% for rigid molecule types (e.g., HEM), as we would expect. However, it is still quite good for many types of large and flexible ligands. even when there are other ligands in the data set with similar chemical structures. For instance, the classification rate for ATP molecules is 91%, even though the data set also contains ACP, ANP, GTP, ADP, GDP, and AMP. This result suggests that the conformation taken by an ATP molecule bound to a protein is generally closer to at least one other ATP molecule bound to a non-homologous protein than it is to the conformation taken by another (possibly very similar) molecule type.

Of course, the results are not perfect – there are some molecule types for which the classification rate is poor (e.g., NDP). Further investigation is required to characterize these cases, perhaps so that they can be identified quickly as candidates for more expensive matching algorithms.

Ligand Type	# Atoms	# Queries	# Correct	# Incorrect	% Correct)	Ligand Type	# Atoms	# Queries	# Correct	# Incorrect	% Correct)
5GP	24	6	1	5 (AMP,IMP)	17	HEX	6	5	5	0	100
AMP	23	21	20	1 (5GP)	95	HEZ	8	6	5	1 (MPD)	83
ACP	31	8	6	2 (ATP)	75	IMP	23	5	3	2 (5GP)	60
ADN	19	7	7	0	100	IMD	7	18	18	0	100
ADP	27	33	33	0	100	LDA	16	6	6	0	100
ANP	31	11	11	0	100	MAL	23	5	5	0	100
ATP	31	22	20	2 (GTP)	91	MES	12	25	25	0	100
BEZ	9	8	8	0	100	MPD	8	44	44	0	100
BOG	20	14	14	0	100	NAD	44	14	11	3 (COA,FAD,NDP)	79
COA	46	9	5	4 (FAD,NAD,NAP,NDP)	56	NAG	14	11	11	0	100
DIO	6	6	6	0	100	NAP	48	10	4	6 (NAD, NDP)	40
DTT	8	6	6	0	100	NDP	48	5	0	5 (COA, NAP)	0
EPE	15	24	24	0	100	ORN	9	5	5	0	100
EST	20	5	5	0	100	OXL	6	5	6	0	100
F6P	16	5	5	0	100	PG4	13	6	6	0	100
FAD	53	8	3	5 (COA,NAP)	38	PGA	9	5	5	0	100
FLC	13	7	7	0	100	PLM	18	5	5	0	100
FMN	31	10	10	0	100	SAM	27	5	5	0	100
FMT	3	22	22	0	100	SIN	8	7	7	0	100
GAL	11	8	7	1 (GLC)	88	SUC	23	7	7	0	100
GDP	28	7	4	3 (ADP)	57	TAR	10	8	8	0	100
GLC	11	19	19	0	100	TRS	8	38	38	0	100
GTP	32	8	3	5 (ATP)	38	UDP	25	5	4	1 (GTP)	80
HEM	43	21	21	0	100	**Total**	-	**545**	**499**	**46**	**92**

Table 1: Classification results for each molecule type.

5 Conclusion

In this paper, we investigated whether a geometric matching algorithm can provide an effective classifier for rigid ligand structures in their bound conformations. We found that a simple algorithm based on atom positions and element types could correctly classify ligands according to their molecule types 92% of the time. This result suggests that the conformational variation of ligands bound to non-homologous proteins in the PDB is quite small - i.e., the differences between the conformations of ligands of the same type are generally less than the differences between molecules of different types in this study, and also that inexpensive geometric matching algorithms can be effective for protein-ligand binding prediction, as long as the geometry and properties of the bound ligand are modeled accurately. It suggests that good protein-ligand binding prediction is limited mainly by our ability to predict a model of the bound ligand from a binding site, since good classification rates are possible using even fast shape matching algorithms when given a perfect model of the ligand. Future work is required to test whether the geometric matching methods described can be effective in the more difficult case of matching and classifying protein binding site models computed *de novo* without bound ligands.

References

Cai, W., Shao, X., and Maigret, B. (2002). Protein-ligand recognition using spherical harmonic molecular surfaces: towards a fast and efficient filter for large virtual throughput screening, *Journal of Molecular Graphics and Modeling*, **20**, 313-328.

Duncan, B., and Olson, A. (1993). Approximation and characterization of molecular surfaces, *Biopolymers*, **33**, 219-229.

Kazhdan, M. (2004). Shape representations and algorithms for 3D model retrieval, *Ph.D. Thesis*, Department of Computer Science, Princeton University.

Morris, R. J., R. J. Najmanovich, A. Kahraman and J. M. Thornton (2005). Real spherical harmonic expansion coefficients as 3D shape descriptors for protein binding pocket and ligand comparisons, *Bioinformatics*, **21**(10), 2347-2355.

Ritchie D., and Kemp, G. (1999). Fast computation, rotation, and comparison of low resolution spherical harmonic molecular surfaces, *J. Comput. Chem.*, **20**, 383-395.