JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

Compression of Molecular Interaction Fields Using Wavelet Thumbnails: Application to Molecular Alignment

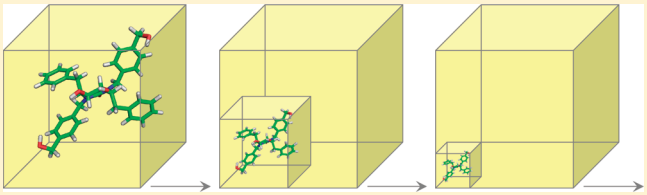
Richard L. Martin,[†,‡,§] Eleanor J. Gardiner,[†] Stefan Senger,[‡] and Valerie J. Gillet*[,†]

[†]Information School, University of Sheffield, Regent Court, 211 Portobello Street, Sheffield S1 4DP, United Kingdom

[‡]Computational and Structural Chemistry, GlaxoSmithKline Research & Development, Gunnels Wood Road, Stevenage, Hertfordshire SG1 2NY, United Kingdom

Supporting Information

ABSTRACT: Molecular interaction fields provide a useful description of ligand binding propensity and have found widespread use in computer-aided drug design, for example, to characterize protein binding sites and in small molecular applications, such as three-dimensional quantitative structure−activity relationships, physicochemical property prediction, and virtual screening. However, the grids on which the field data are stored are typically very large, consisting of thousands of data points, which make them cumbersome to store and manipulate. The wavelet transform is a commonly used data compression technique, for example, in signal processing and image compression. Here we use the wavelet transform to encode molecular interaction fields as wavelet thumbnails, which represent the original grid data in significantly reduced volumes. We describe a method for aligning wavelet thumbnails based on extracting extrema from the thumbnails and subsequently use them for virtual screening. We demonstrate that wavelet thumbnails provide an effective method of capturing the three-dimensional information encoded in a molecular interaction field.



INTRODUCTION

Molecular interaction fields (MIFs) describe the interaction energies between a molecule and a probe group distributed on a three-dimensional (3D) grid. Negative regions on the grid indicate energetically favorable positions for the probe to reside, and positive regions indicate positions where interaction between the probe and the molecule is energetically unfavorable. By varying the probe group, the propensity for the molecule to form different types of interactions, such as hydrogen bonds and hydrophobic interactions, can be characterized. The GRID program[1,2] is the most well established method for generating MIFs. GRID was initially developed to characterize protein binding sites[3] and has since been applied to small molecules for a variety of applications, including quantitative structure−activity relationship (QSAR) modeling,[4−6] adsorption, distribution, metabolism, excretion (ADME) property prediction,[7−9] and virtual screening.[10,11]

The grids used to encode MIFs are typically very large so that several thousand data points are required to store and manipulate them. Furthermore, the large number of variables, many of which are highly correlated, can lead to misleading results in some applications. Several of the applications based on MIFs use some form of data extraction in which the full grids are reduced to a much smaller number of data points. For example, in the GRid-INdependent Descriptors (GRIND) descriptors, extreme values calculated by GRID are extracted into a 1D descriptor that is similar in concept to the pharmacophore key.[12] The GRIND descriptors were developed for QSAR modeling and have also been used in similarity searching as well as to predict various properties, such as human ether-a-go-go-related gene (hERG) inhibition and metabolism. The related fingerprints for ligands and proteins (FLAP) descriptors were developed to characterize both proteins and ligands and have been used for various virtual screening applications, including protein−ligand docking.[13] The FieldScreen program is also based on extracting the key information contained within a field into a small number of points, although here, the spatial relationships between the extrema are retained as a 3D graph to enable molecules to be aligned.[14]

In previous work, we investigated the use of wavelet techniques for reducing the data present in MIFs. Wavelet transforms are commonly used in signal processing and are related to the Fourier transform. We demonstrated that wavelet transforms can be used to achieve very high degrees of compression of GRID fields without loss of information and that they are better suited to the irregular data contained within a GRID than the Fourier transform.[15] When used as input to 3D QSAR model building for sets of prealigned ligands, the wavelet compressed fields resulted in models that were at least as accurate as those generated using the whole (uncompressed) fields, and in some cases, the predictive performance of the models was superior. It was postulated that the latter was due to the removal of noise in the GRIDs and a significant reduction in the number of correlated variables. We examined

Received: July 26, 2011
Published: February 13, 2012

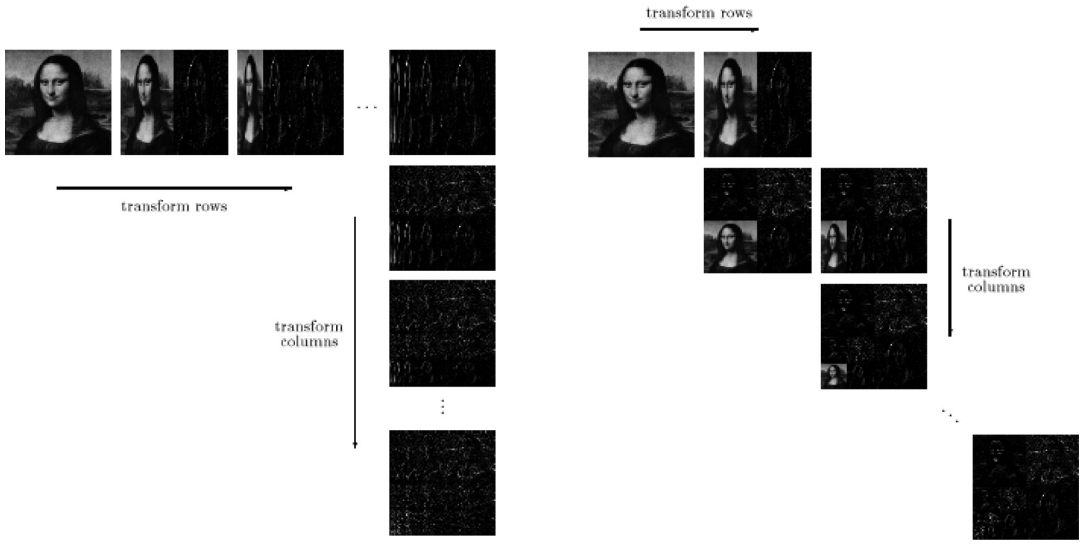


Figure 1. The standard wavelet transform is shown on the left and involves iteratively transforming the rows followed by the columns. The nonstandard wavelet transform is shown on the right in which the rows and columns are transformed alternately with the main signal concentrated increasingly toward the bottom left-hand portion of the image, with the remaining space containing the details. The square thumbnail images which arise after each pair of row and column transformations are readily visible. Images published in Stollnitz.[17] Copyright Elsevier (1996).

two types of data compression: a truncation method, which is commonly used in image compression, and what we have referred to as wavelet thumbnails, where a 3D volume is iteratively reduced in size to smaller and smaller volumes. Both of these are described in more detail below. The advantage of the wavelet thumbnail is that the spatial relationships between the various data points are retained, albeit in a smaller volume. This allows easy interpretation of the reduced data and also enables thumbnail comparisons to be carried out directly, which is exploited in this paper.

Here, we further investigate the effectiveness of wavelets for compressing MIFs by analyzing the performance of wavelet approximated fields in virtual screening. We begin with a brief introduction to wavelets and the wavelet approximation techniques used. We then describe two experiments that investigate the effectiveness of the wavelets at capturing information relevant to 3D virtual screening. First, we examine the ability of the wavelet compressed fields to maintain the order of similarity-ranked lists of molecules as the data is progressively compressed. In this case, the molecules are prealigned using an established method. In the general case, however, molecules in a database may not be in a meaningful alignment. We then present a methodology for aligning a pair of molecules in arbitrary orientation based on their wavelet thumbnail representation. The resulting alignment can then be used to calculate the similarity of the pair of molecules and thereby enables the wavelet thumbnails to be used for virtual screening. Performance of the wavelets is compared with the more established 3D virtual screening program, ROCS.

Wavelet Transform. The wavelet transform is a technique for signal decomposition that is related to the well-known Fourier transform; the main difference between the transforms is the nature of the basis functions used in the decomposition process.[16,17] The Fourier transform decomposes a signal into sine waves of different frequencies with the smooth character of sine waves making the technique suitable for representing periodic signals of long duration. The wavelet transform, on the other hand, is based on functions that have irregular shape and decay over time, which makes them more suitable for representing complex, nonperiodic signals.

For simplicity, the wavelet transform is first described here for a 1D signal. The wavelet transform recursively splits a signal into two parts: smoothed low-frequency components and high-frequency differences or details, with each part half the size of the input data. Thus, if the original signal consists of 256 elements, the first pass of the transform produces 128 smoothed components and 128 detail coefficients. A wavelet consists of a scaling function, which is used to produce the smoothed components that approximate the original signal, and a wavelet function, which is used to produce the detail coefficients that represent the components which are not captured by the scaling function. As an example, in the well-known Haar wavelet, a window of size two is moved over the input data using a step size of two. For each step, the scaling function is the average of the two elements in the window, while the wavelet function takes the difference between the two elements. Under Mallat's pyramid algorithm,[18] used in this work, the smoothed low-frequency components form the input to the next pass of the transform with the process continuing until there is one smoothed component and $2^N - 1$ detail coefficients, for a signal of length 2^N. Thus a signal with 256 elements will be represented by one smoothed component and 255 detail coefficients.

No loss of information has occurred during the wavelet transform, and the original signal can be reconstructed by applying the inverse of the transform recursively. However, the effect of the wavelet transform is to concentrate the signal into a small number of coefficients, since a large number of the detail coefficients tend to be of very small magnitude. Removal of a large number of these prior to reconstruction results in small errors only in the signal and allows for efficient data compression and denoising of the signal.

When applying the wavelet transform to 2D images and 3D volumes, it is necessary to process the data as a series of 1D signals. Thus, a 2D $2^N \times 2^N$ square image is considered as a set of 2^N rows and 2^N columns, each of length 2^N. There are two common approaches to applying the wavelet transform, which

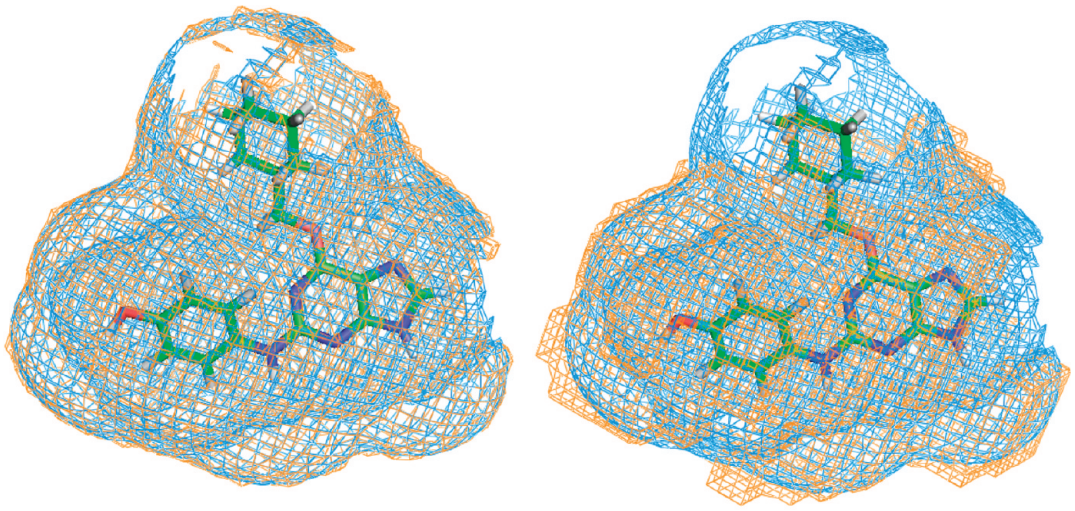


Figure 2. Wavelet compression. Haar compression applied to the GRID O^- probe map of the CDK2 ligand extracted from Protein Data Bank (PDB) entry 1oi9.[29] The uncompressed GRID map is shown in blue, and the reconstructed surface generated following wavelet compression is in orange at (a) 90% and (b) 99% compression. The overall shape of the field is encoded successfully in both cases, however at 99% compressions, the image is noticeably less smooth than the original due to the discontinuous character of the Haar wavelet (see Figure S1, Supporting Information), although the shape of the original field is still discernible.

are referred to as the standard and the nonstandard approaches. In the standard approach, the wavelet transform is applied to each row iteratively to give a single average value and a set of $2^N - 1$ detail coefficients for each row. Next the transformed rows are treated as an image, and the wavelet transform is applied iteratively to each column. The resulting values are all detail coefficients except for a single averaged value. In the nonstandard approach, the wavelet transform is applied to the rows and columns alternately so that one iteration of the transform consists of first approximating the signal in half the width of the image followed by halving its length. Thus the smoothed components are contained in one-fourth of the original image, with the rest of the image containing the details. If N iterations are performed, then, as for the standard approach, the result is all detail coefficients except for a single averaged value. However if the nonstandard approach is halted after m iterations (where $m < N$), then the resulting smoothed image constitutes a 2^{N-m} square representation of the original data. Thus following each pass of the transform, the image is concentrated in the bottom left corner with the rest of the image consisting of the details. Figure 1 illustrates the application of both the standard and the nonstandard approaches to the wavelet transform to a 2D image.

Note that there are many different wavelet functions. Those investigated in this work include Haar and Daubechies 4-tap (D4) following our previous paper, and these functions are illustrated in the Supporting Information.

Wavelet transforms form the basis of the JPEG 2000 standard for 2D image compression.[19] They are also widely used in medical imaging for 2D image compression and feature preservation.[20] The first use of wavelets for describing 3D data was by Muraki[21] who used them to compress volumetric data determined from magnetic resonance images.

One of the first applications of wavelets to molecular data was for the visualization of protein structure, rather than data compression, as described by Carson[22] who focused on the multiresolution aspect of the wavelet representation to allow protein folds to be viewed in greater or lesser detail. Since then wavelets have been applied to a number of applications in bioinformatics,[23] for example, they have been used for fast compression/decompression to allow interactive visualization of macromolecular maps in a molecular graphics application.[24] In small molecule applications, wavelets have been used to represent various 1D spectra and property distributions, for example, electron densities, with the resulting wavelet coefficients used as descriptors in applications such as QSAR.[25] More recently, wavelets have been applied to the representation of 3D molecule field data generated from prealigned ligands to generate descriptors for 3D QSAR, as described in our own work[15] and that of Beck and Schindler.[26]

Wavelet Approximations of GRID Fields. The wavelet transform in three dimensions is a simple extension of the 2D approach above, since 3D volumetric data can also be considered as a set of 1D signals in each of the three axes. Two different approaches to approximating the GRID fields are investigated here, referred to as wavelet compression and wavelet thumbnails. In wavelet compression, the data are wavelet transformed, and a large percentage of the lowest magnitude coefficients in the post-transform representation are set to zero; setting these points to zero is known as truncation and causes data in the signal to be lost. To complete the compression process, only the nonzero entries (and their locations) are retained.[17,27,28] The volume can then be reconstructed for visualization or comparison by applying the reverse transform to the truncated data; however, since some data points were truncated, some signal is lost. The resulting volume is of the same size as the input, allowing a wavelet approximated GRID to be compared directly with the original GRID from which it was derived. In this work we investigate the degree of compression that can be achieved without significant loss of information. A variety of compression percentages are investigated, where the percentage is relative to the number of nonzero coefficients in the original field, i.e., a n% reduction is with respect to the nonzero data in the original field.

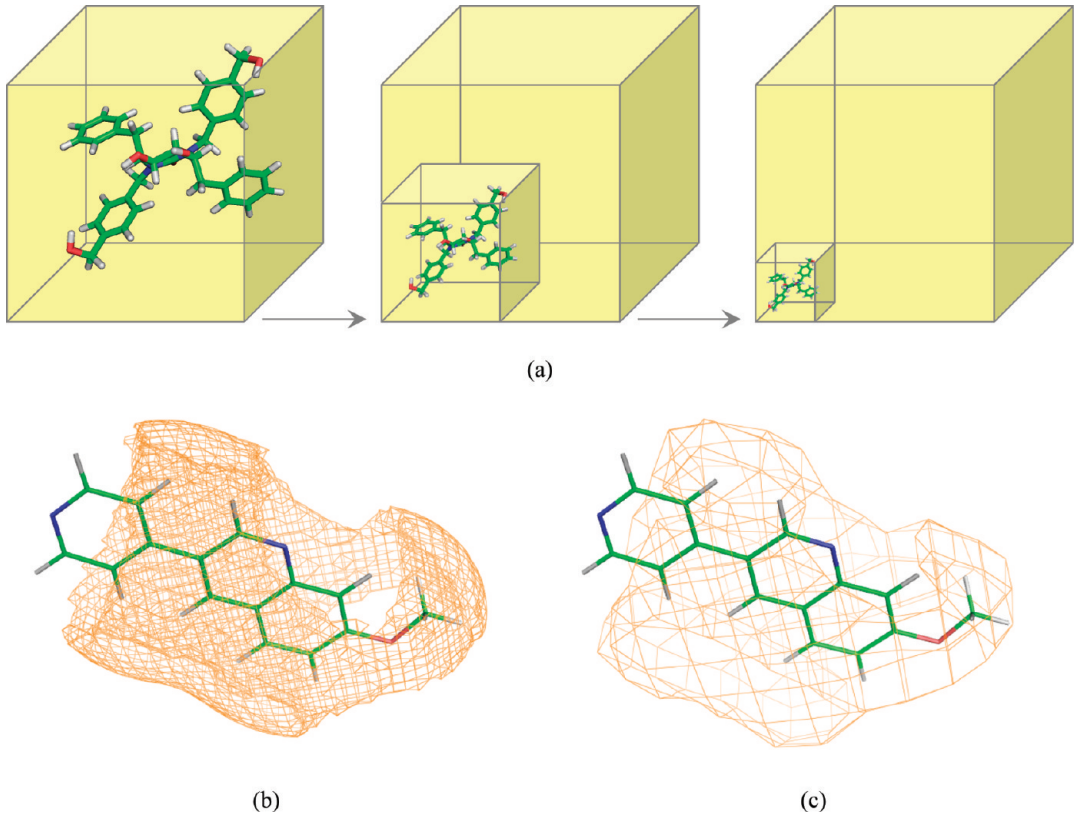


Figure 3. Wavelet thumbnails. (a) Schematic of the thumbnail approach which shows how the signal is increasingly concentrated in the bottom left corner of the grid with each pass of the transform. Note that for clarity the effect is illustrated using an atom and bond representation of the molecule, rather than the actual field data. (b) The reduction in field data is illustrated in the bottom half of the figure: (b) shows the uncompressed 64^3 dry probe field for a ligand extracted from the Zinc database with code ZINC03832261 (left) and (c) shows its 16^3 D4 thumbnail (right), scaled up to the same size for comparison.

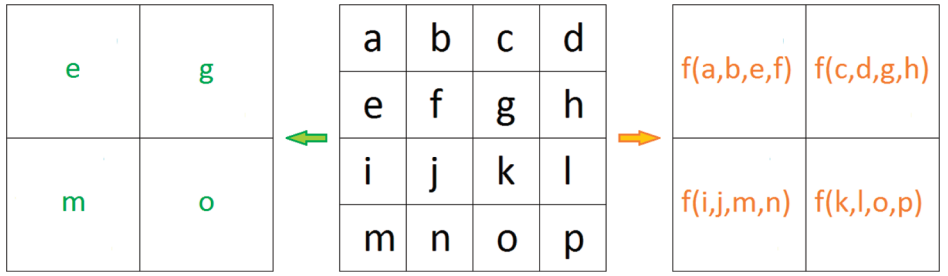


Figure 4. An example 2D image consisting of a 4 × 4 grid of data points (center). Coarse sampling of this image yields a 2 × 2 image comprising a single representative point from each distinct 2 × 2 subgrid (green, left). A single pass of the Haar wavelet transform also produces a 2 × 2 image, but where each point is a weighted average of the four constituent points of each subgrid (orange, right).

In the wavelet thumbnail, compression is achieved by representing the data with fewer data points, rather than by using fewer wavelet coefficients in data reconstruction. Indeed, no reconstruction takes place, since the thumbnail acts as a standalone image, just of a smaller size than the original. Note that here the term thumbnail is used to describe a reduced 3D volume representation of the original data. The wavelet thumbnail is generated by simply retaining the smoothed components and ignoring the details following each pass of the wavelet transform. The nonstandard transform is such that the smoothed components of a 3D grid are concentrated in a smaller cubic volume which approximates the full grid. Each pass results in a progressively smaller thumbnail, thus for a GRID field of size 64^3, the thumbnail sizes that are possible are 32^3, 16^3, 8^3, and 4^3. While the degree of data reduction in this approach is not as great as that achieved in wavelet compression, the thumbnails can save time in field comparison and manipulation relative to the reconstructed volumes. For example, the comparison of two 8^3 thumbnails involves 512 times fewer comparisons between data points than are required for 64^3 fields.

The two different approaches to approximating GRID fields using wavelets are illustrated in Figures 2 and 3, respectively.

A much simpler way of reducing the data in a GRID field is to simply sample the field at a lower resolution, without the need for wavelet techniques. For example, if a grid spacing of 0.5 Å is used to generate a GRID field of size 64^3, then reducing the grid spacing to 1 Å will result in a GRID of size 32^3; a resolution of 2 Å will give a GRID of size 16^3 and so on. In practice, since each data point is calculated independently, the

dx.doi.org/10.1021/ci200348h | J. Chem. Inf. Model. 2012, 52, 757−769

same result can be achieved by considering alternate points at each level of reduction. We refer to grids produced using this method of data reduction as coarse GRIDs. Coarse GRIDs can be generated at the same sizes as the wavelet thumbnails and allow a direct comparison to be made of the two data reduction techniques. The expectation is that wavelet thumbnails will be more effective at retaining the information in a GRID field, since each data point in a thumbnail is some function of a region of data points in the large grid (with the function depending on the particular wavelet transform), whereas the same is not true for coarse GRIDs. This difference is illustrated for a 2D image in Figure 4.

DATA SETS

Nine activity classes were selected from the filtered DUD database for the main experiments.[30] The active compounds therein are a subset of those in the original DUD database chosen to remove the effect of analogue bias; achieved by clustering the actives and choosing a single representative compound from each cluster.[31] The decoys for each activity class in DUD were selected to have similar physicochemical properties but different chemical topologies to the actives. For each class, all of the actives were taken and were augmented by exactly eight times as many decoys chosen at random from the full set of decoys. A further data set was compiled based on 18 hand-picked thrombin inhibitors. Thrombin inhibitors that exhibit four recurrent hydrogen-bond interactions with the protein backbone and a similar large quantity of hydrogen-bond donors and acceptors were obtained from Relibase+.[32] The ligands were protonated using MOE.[33] The data sets are shown in Table 1. In contrast to the DUD actives, the thrombin ligands are all similar with respect to topology: For example, the average pairwise similarity of the active compounds measured using ECFP4 is 0.118 for CDK2 and 0.227 for thrombin and 0.475 and 0.653, respectively, using MDL Public Keys. The thrombin actives were also augmented by exactly eight times as many inactives (144), chosen at random from the subset of DUD thrombin decoys.

Table 1. Data Sets Extracted From the DUD Database[a]

data set	number of actives	mean pairwise similarity using MDL Public Keys	mean pairwise similarity using ECFP4
CDK2	32	0.475	0.118
COX2	44	0.461	0.176
EGFR	40	0.524	0.186
INHA	23	0.455	0.195
P38	20	0.554	0.276
PDE5	22	0.533	0.142
PDGFRB	22	0.466	0.184
SRC	21	0.482	0.125
thrombin	18	0.653	0.227
VEGFR2	31	0.476	0.126

[a]All of the experiments reported here are based on a single conformation of each ligand.

We also used a set of 21 ligands to parametrize our alignment method. We identified the 21 activity classes in the filtered DUD that contain at least 10 ligands and simply selected the first ligand in each download file.

METHODS

We have examined the effectiveness of wavelet techniques for approximating the data in GRID fields using two experiments. First, we investigated whether the information loss that occurs during wavelet compression affects the ability to separate prealigned active and inactive compounds in a similarity-based virtual screening experiments. Second, we present a method for aligning pairs of molecules based on the wavelet thumbnail representations of their molecular interaction fields. The resulting alignments were used to calculate pairwise similarities of molecules and thus enabled virtual screening to be applied to molecules in arbitrary orientation.

Prealigned Ligands. Seven actives were selected at random from each of the DUD data sets and used as queries to generate alignments using Openeye's well-established ROCS program and the ComboScore similarity measure.[34,35] Note that this resulted in seven sets of alignments for each data set, each based on a different query. For each ROCS-aligned set, GRID fields at 0.5 Å resolution and size 64^3 were generated. Four GRID probes were investigated independently: the dry, $N1^+$, O^-, and water probes to give a total of 280 ranked lists (7 queries × 10 data sets × 4 probes). The four probes were chosen to model common interaction types: the dry probe models hydrophobic interactions; $N1^+$ models hydrogen-bond acceptance; O^- models hydrogen-bond donation; and the water probe models both hydrogen-bond donation and acceptance. The GRID field of each compound was compared to the GRID field of the query in their ROCS alignment, and each compound list was ordered on increasing Euclidean distance to the query. The area under the receiver operator characteristic curve (AUC) and enrichment factor at five percent (EF5) were calculated for each list. (Note that although it is customary to report EFs at lower percentage values, such as 1 or 2%, this is not appropriate here due to the comparatively small sizes of the data sets, e.g., a 1% EF would consider only the top three compounds in the CDK2 set). The AUC and EF values were averaged over the 7 queries in each activity class and over the 10 activity classes.

The GRID fields were then compressed using the truncation method to produce wavelet compressed GRIDs at compression percentages of 90, 99, 99.9, and 99.99% for each of the four probes and for each wavelet. For each level of compression, the compressed GRIDs were reranked relative to each query compound and new AUC and EF values calculated and averaged as before. The aim was to investigate the extent to which the wavelet compressed GRIDs retain the same enrichment as that obtained with the uncompressed GRIDs.

Alignment Using Wavelet Thumbnails. The method developed to align thumbnails is based on graph theory techniques. The wavelet thumbnails are represented as fully connected 3D graphs by identifying extrema in the thumbnails which become nodes in a graph. This approach is similar to that used in FieldScreen and in the earlier Field-Graph program, although these are based on different field representations (FieldScreen[14] is based on the XED field, whereas Field-Graph is based on the molecular electrostatic potential),[36] and in both cases the extrema are selected from the full, uncompressed fields. Each node in a graph is labeled according to the value of the extremum it represents, and the edges are labeled according to distances between the extrema in thumbnail coordinate space. Potential mappings between the graph representations of two thumbnails are identified using the Bron–Kerbosch clique

dx.doi.org/10.1021/ci200348h | J. Chem. Inf. Model. 2012, 52, 757–769

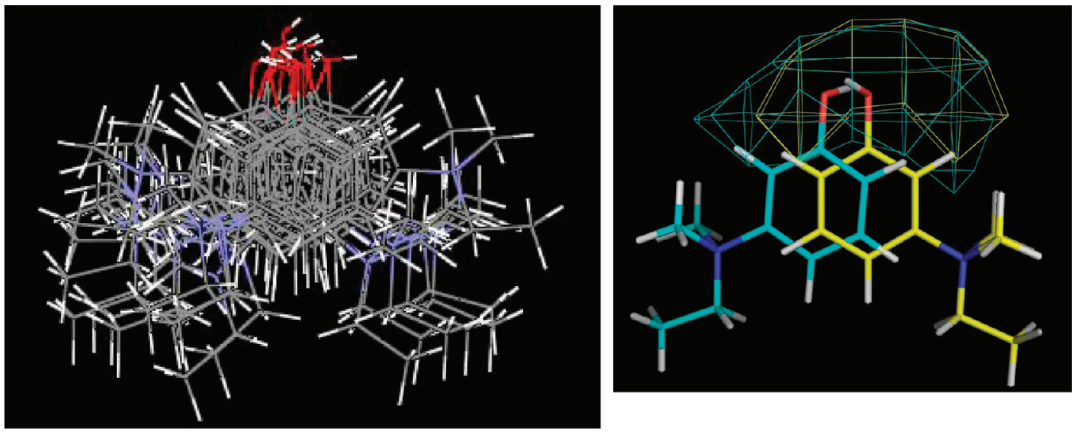


Figure 5. An example of a poor self-alignment. (a) Various poses of the AChE representative ligand are aligned using $N1^+$ Haar 16^3 thumbnails and (b) two poorly overlaid ligands. By contouring the $N1^+$ field data to only show large magnitude areas, it is clear that only the region around the sole oxygen atom is considered in the alignment procedure (image generated in PyMOL).[39]

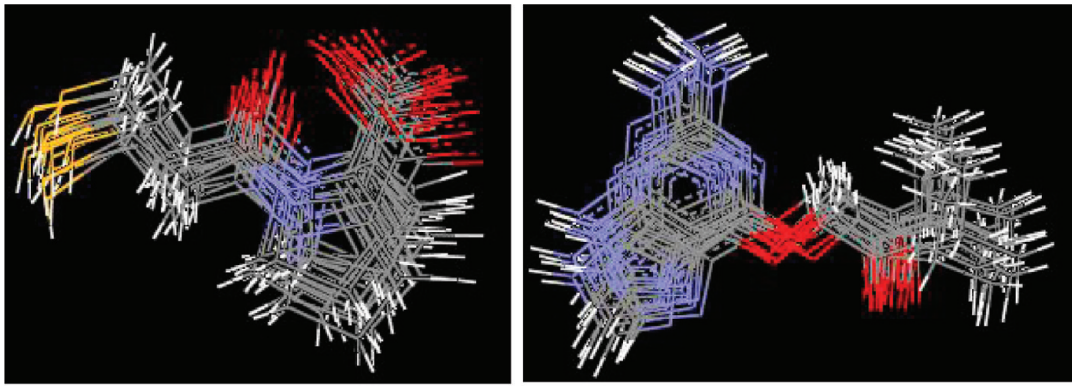


Figure 6. Alignments generated using the water probe and the Haar wavelet: (a) ACE and (b) CDK2 ligands.

detection algorithm.[37] This step involves the construction of a correspondence graph which encodes relationships between pairs of nodes, one from each graph. The resulting mappings form the input to a Kabsch least-squares fitting procedure.[38] Each resulting alignment is then scored using the continuous Tanimoto coefficient applied to the entire thumbnails.

The thumbnail alignment procedure has three parameters for which optimum values were determined: (i) the threshold value used to define which data points are extracted as nodes in the graph representations, referred to as the node threshold; (ii) the tolerance on node magnitude used to determine valid node mappings in the correspondence graph, referred to as the magnitude ratio; and (iii) the tolerance on edge distance in the thumbnail graphs used to generate edges in the correspondence graph, referred to as the distance tolerance.

The range of values present in a thumbnail is dependent on the particular probe used, the molecules in the data set, and the wavelet technique that is applied. The threshold value was, therefore, chosen for each data set and method so that each thumbnail in the data set is represented by some minimum number of nodes.

In contrast, the magnitude ratio and distance tolerance are independent of data set and probe. The magnitude ratio, M_r, is defined as follows: A pair of nodes, one from each thumbnail represented as n_a and n_b, forms a valid mapping if

$$\frac{\max(n_a, n_b)}{\min(n_a, n_b)} \leq M_r$$

The following values were investigated: 1.1, 1.2, 1.3, 1.4, and 1.5.

The distance tolerance values investigated were 0.2, 0.4, 0.6, 0.8, and 1.0. Note that the distance values relate to distances in thumbnail coordinate space so that it is inappropriate to specify units of distance, however, since the width, height, and breadth of a 16^3 thumbnail is one-fourth of those of a 64^3 GRID, a distance of 1.0 in a 16^3 thumbnail could be said to correspond to a GRID distance of 4.0, i.e., 2 Å.

Preliminary experiments based on thumbnails generated from individual probes revealed poor results for the $N1^+$ and O^- probes. Figure 5 shows several incorrect alignments generated when trying to align an AChE ligand with itself in different starting poses; the alignments are shown on the left with the thumbnail representations of an example alignment shown on the right with the extreme regions highlighted. The $N1^+$ probe models hydrogen-bond donation only, and it is clear that the poor alignment is due to the extrema being concentrated around the single oxygen atom only; the oxygen atoms are overlaid, but there is little with which to orientate the

dx.doi.org/10.1021/ci200348h | J. Chem. Inf. Model. 2012, 52, 757−769

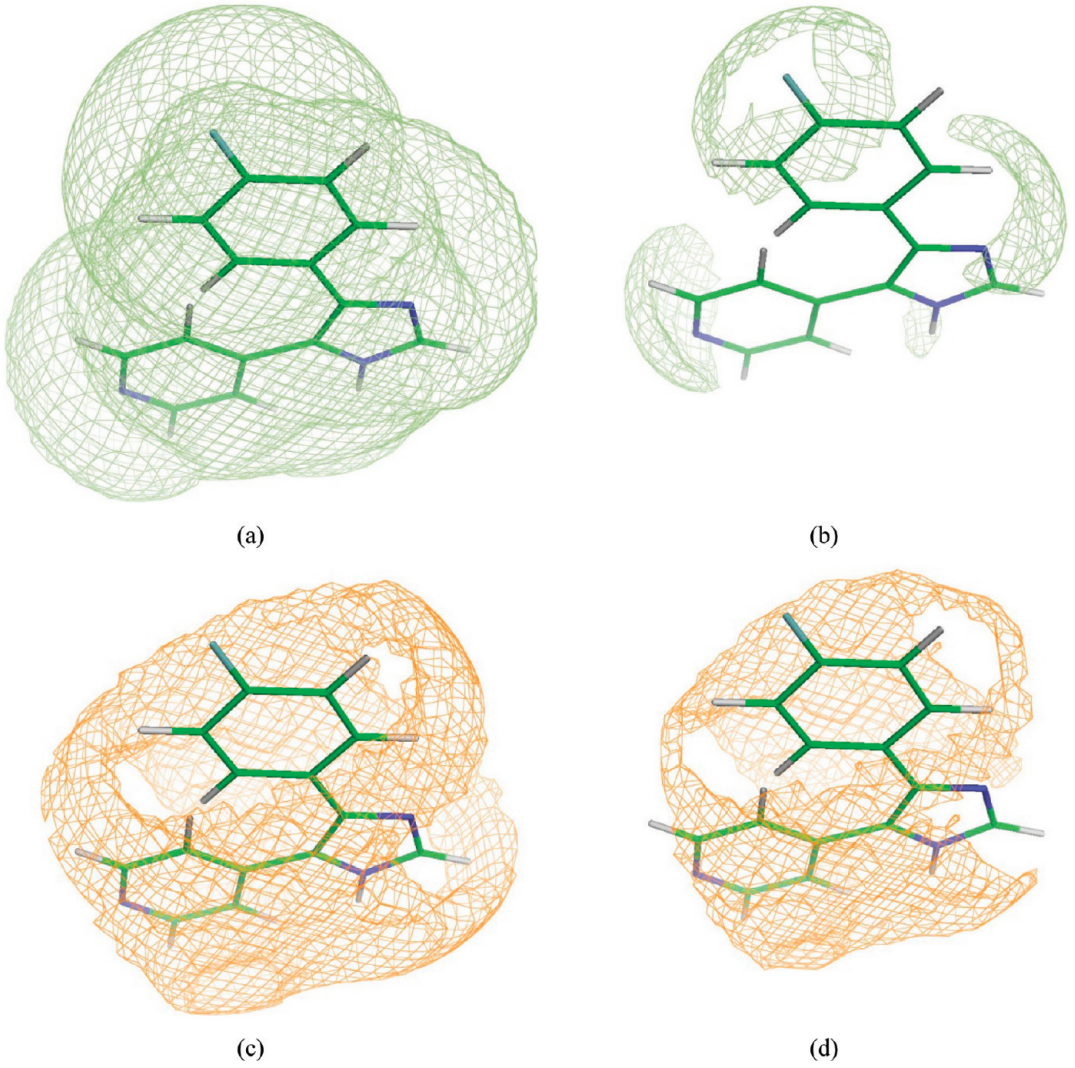


Figure 7. P38 active ligand with code ZINC00007301. Water probe field contoured at (a) −1 and (b) −5 kJ mol^{-1}. Dry probe field contoured at (c) −1 and (d) −5 kJ mol^{-1}. Surfaces contoured using PyMOL.[39]

rest of the molecules. Similar results were also seen for the O^- probe which models hydrogen-bond acceptors only. Thus it was decided not to attempt to generate alignments using these probes in isolation.

In contrast, Figure 6 demonstrates two good alignments using the water probe for ACE and CDK2 ligands, respectively. The water probe models both hydrogen-bond donors and acceptors and so can be considered a good approximation of the important parts of both $N1^+$ and O^-, thus encodes more of the important features of molecules enabling good alignments to be found.

The dry probe was also examined as it is complementary to the water probe, giving a less localized representation of the ligand with large regions of similar magnitude data points. Figure 7 illustrates this complementary character with respect to the P38 active ligand with code ZINC00007301.

The five values examined for each of the three alignment parameters resulted in a total of 125 parameter combinations. Optimum values were chosen by carrying out a self-similarity experiment. Each of the 21 ligands in the parametrization set (see Data Sets Section) was centered on the origin, with this pose referred to as the static pose, and 20 alternative poses were obtained by performing random (known) rotations upon the static pose. GRID fields and 16^3 Haar and D4 thumbnails were generated for each pose using the dry and water probes. Each probe and each wavelet was considered in turn. One ligand was considered at a time, and the thumbnail for the static pose and one of the rotated poses formed the input to the clique detection algorithm using one set of clique detection parameters. The root-mean-square deviation (rmsd) between the extrema represented by the clique following alignment was calculated. This process is illustrated in Figure 8.

The alignment which produced the smallest rmsd was chosen as the best solution for that pose. The process was repeated for all 20 poses, and the mean rmsd was recorded. This was repeated for all 21 ligands, and the rmsd values averaged to give an overall measure of performance for that set of clique detection parameters, probe, and wavelet combination. The whole process was then repeated for all 125 parameter combinations. Finally, the 125 parameters were ranked according to decreasing mean rmsd, and a mean rank was obtained for each parameter combination.

Overall, it was found that the parameters which place the fewest restrictions on the detection of cliques (i.e., larger node threshold, magnitude ratio, and distance threshold) tended to give the best performance, which is likely to be due to the large

dx.doi.org/10.1021/ci200348h | J. Chem. Inf. Model. 2012, 52, 757−769

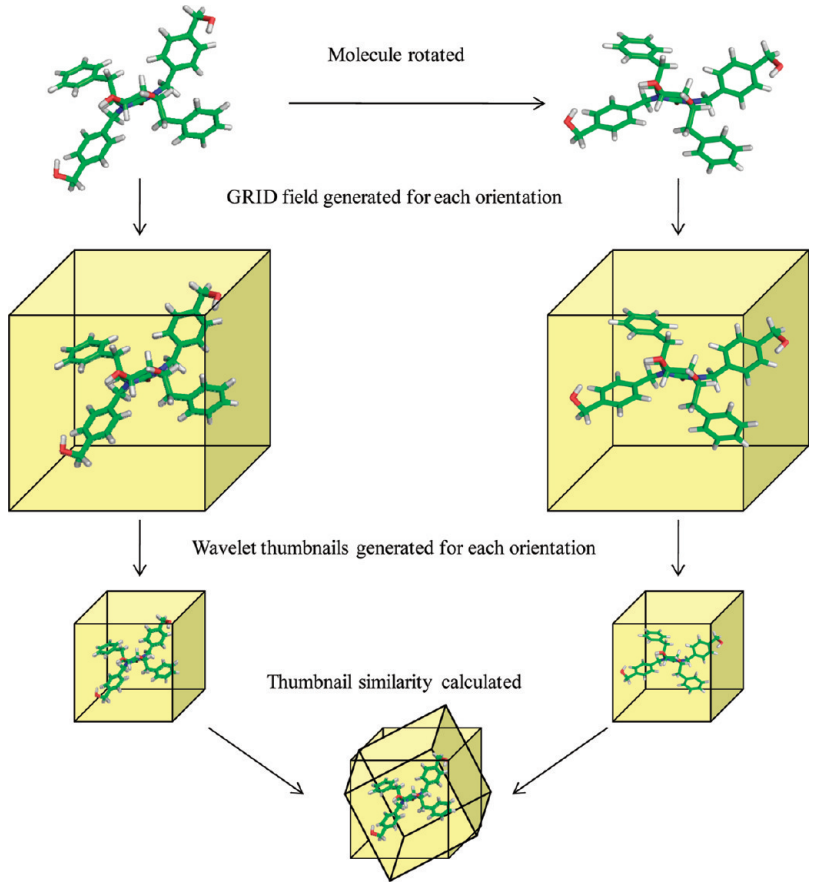


Figure 8. A schematic of the how alignment is performed between compounds, illustrated by a self-similarity example. Atom and bond representations of the molecules are used rather than the actual field data, for clarity.

Table 2. Mean AUC Values (and Standard Deviations) Across 10 Activity Classes for Haar and D4 Wavelets at Different Degrees of Wavelet Compression

		no compression	90%	99%	99.9%	99.99%
dry	Haar	0.802 (0.052)	0.797 (0.053)	0.734 (0.063)	0.668(0.064)	0.615 (0.071)
	D4		0.798 (0.052)	0.777 (0.057)	0.690 (0.060)	0.627 (0.063)
$N1^+$	Haar	0.714 (0.067)	0.709 (0.067)	0.662 (0.057)	0.681 (0.064)	0.665 (0.070)
	D4		0.687 (0.069)	0.667 (0.066)	0.678 (0.066)	0.652 (0.058)
O^-	Haar	0.702 (0.041)	0.701 (0.041)	0.677 (0.043)	0.648 (0.052)	0.653 (0.056)
	D4		0.697 (0.040)	0.680 (0.042)	0.647 (0.043)	0.618 (0.068)
water	Haar	0.730 (0.056)	0.726 (0.056)	0.698 (0.054)	0.671 (0.053)	0.644 (0.065)
	D4		0.724 (0.054)	0.692 (0.052)	0.677 (0.062)	0.647 (0.061)

size and high connectivity of the correspondence graphs, which in turn leads to a large number of cliques being found. The likelihood of finding a good mapping is thus increased; however, this also leads to a significant increase in the computational requirements. Conversely, parameters which place greater restrictions reduce the size and connectivity of the correspondence graph, with a consequent reduction in the number of cliques found for a particular pair of thumbnails. The parameter combination 20−1.5−0.6 (signifying a node threshold of 20, a magnitude ratio of 1.5, and a distance tolerance of 0.6) was top-ranking overall, however, it is one of the most computationally expensive options due to the large number of nodes, and so the combination chosen for the virtual screening experiments was 10−1.5−0.6 (reducing the node threshold to 10 gave the best compromise between overall performance and computational speed).

The 10 data sets and their query compounds discussed earlier were used in the alignment experiments. For each data set and for each query, all compounds in the data set were aligned to the query using the thumbnail representations, and the compounds were ranked relative to the query. To avoid any bias in the starting orientations, each ligand was first centered using the mean of its atomic coordinates and then subjected to a random rotation prior to field generation. Fields of size 64^3 were generated using the dry and water probes and reduced to 16^3 thumbnails using the Haar and D4 wavelets. Two constraints were applied to reduce the number of cliques output during the alignment phase. First, edges were not generated in the correspondence graphs for distances less than two units in thumbnail space (equivalent to 4 Å in full 64^3 GRID space) to avoid mappings consisting of very local features. Second, the clique detection program was terminated after 60 s. In the majority of cases, this time limit is sufficiently

dx.doi.org/10.1021/ci200348h | J. Chem. Inf. Model. 2012, 52, 757−769

Table 3. Mean AUC Values (and Standard Deviations) Across 10 Activity Classes for Wavelet Thumbnails of Different Sizes Generated Using the Haar and D4 Wavelets[a]

		no compression	32	16	8	4
dry	Haar	0.802 (0.052)	0.803 (0.054)	0.799 (0.058)	0.780 (0.063)	0.697 (0.069)
	D4		0.806 (0.054)	0.802 (0.058)	0.787 (0.064)	0.713 (0.068)
	Coarse		0.796 (0.053)	0.816 (0.048)	0.757 (0.054)	0.615 (0.052)
$N1^+$	Haar	0.714 (0.067)	0.708 (0.066)	0.695 (0.066)	0.674 (0.062)	0.630 (0.058)
	D4		0.701 (0.067)	0.697 (0.067)	0.672 (0.068)	0.624 (0.049)
	Coarse		0.711 (0.068)	0.717 (0.067)	0.702 (0.068)	0.609 (0.064)
O^-	Haar	0.702 (0.041)	0.699 (0.041)	0.693 (0.040)	0.678 (0.040)	0.604 (0.047)
	D4		0.701 (0.041)	0.696 (0.039)	0.670 (0.040)	0.609 (0.044)
	Coarse		0.703 (0.040)	0.695 (0.045)	0.692 (0.042)	0.598 (0.049)
water	Haar	0.730 (0.056)	0.726 (0.055)	0.719 (0.053)	0.705 (0.056)	0.650 (0.068)
	D4		0.727 (0.055)	0.720 (0.054)	0.706 (0.051)	0.639 (0.056)
	Coarse		0.730 (0.057)	0.732 (0.056)	0.725 (0.066)	0.610 (0.059)

[a]Results are also shown for coarse GRIDs with the same number of data points.

long to have no effect, but for some ligand combinations, the correspondence graph was such that cliques could be continuously detected for several minutes, leading to an impractical number of cliques and a prohibitively expensive procedure overall. Furthermore, once a clique of size n had been discovered for a pair of thumbnails, subsequently detected cliques of smaller size were not output. It should also be noted that the Bron−Kerbosch algorithm favors the early detection of large cliques, further reducing the volume of output. Even with these constraints, several alignments were typically generated for each pair of thumbnails. Each alignment was scored using the continuous Tanimoto similarity applied to all points in each thumbnail. The alignment with the greatest similarity for a pair of thumbnails was retained.

RESULTS

Prealigned Ligands. The AUC values for the uncompressed GRID fields for each of the four probes are reported in the column headed "no compression" in Tables 2 and 3 (EF5s are given in Table S1 in the Supporting Information). These values form base-lines against which to compare the wavelet results. Results for wavelet compression for the four different probes are reported in Table 2. The same process was repeated for thumbnails at 32^3, 16^3, 8^3, and 4^3 and also for coarse GRIDs of the same sizes, reported in Table 3 (EF5s are in Table S2 in the Supporting Information).

The performance of the wavelet compressed grids at 90% compression is comparable with that of the original GRIDs. However, at higher degrees of compression the ability to distinguish between actives and decoys begins to deteriorate. For thumbnails, the performance at 32^3 and 16^3 volumes is comparable to the original GRIDs. For the dry probe, the performance of the coarse GRIDs deteriorates more rapidly than for the wavelet thumbnails, especially at sizes 8^3 and 4^3. The same is not true of the other probes, however, where the coarse GRID's performance is comparable to that of the wavelet thumbnails at all resolutions. No appreciable differences were seen between the two different wavelets. These findings for the wavelet approximated GRIDs are consistent with those reported earlier in the 3D QSAR studies.[15] However, the apparent good performance of the coarse GRIDs seen here was unexpected and is inconsistent with the QSAR experiments where the performance was reduced relative to the wavelet thumbnails. It is postulated that this apparent good performance is due to the molecules already being in alignment prior to field generation. Since the process of coarse sampling results in the removal of signal elements based only on their position in space, if the compounds in the data set are already aligned prior to field generation, then it is likely that the features which are removed (or not) in the coarse sampling of field A will also be removed (or not) for field B, such that the similarity between the structures will be largely uncompromised. However, when the compounds are not in alignment prior to field generation, as will be the general case, this coincidence of features is less likely. For example, a feature which occurs in both A and B may be removed in the sampling of structure A, but since structure B is in a different orientation relative to A, it is not guaranteed that the same feature within B will also be in a position in space such that it is removed during sampling. These differences which arise in the coarse representations of similar structures may lead to difficulties in performing alignment when using coarse GRIDs, which would not be expected for wavelet thumbnails; this is investigated in the following section.

Although these results demonstrate that high degrees of compression can be achieved for the wavelet compression method without significant loss of information, it should be noted that analyzing compressed GRIDs involves reconstructing the post-transform data to be the same size as the original, albeit, after very large percentages of the detail coefficients are set to zero (truncated). Although the small quantity of post-truncation coefficients contain the same information as the reconstructed GRIDs, it would not be trivial to devise a method for comparing compounds based on these coefficients alone, thereby bypassing the requirement for reconstruction. This is because there is no correspondence in location between coefficients that are retained following compression of one GRID and those arising from the compression of the GRID representation of a different molecule or even from the GRID representation of the same molecule in a different orientation. For wavelet thumbnails however, although the degree of compression is not as great, the data points provide a uniform representation of the original GRID, with the additional benefit of reduced size of the representation.

Alignment using Wavelet Thunmbnails. The performance of the alignment method, applied to each data set, was measured using enrichments as above. AUC values averaged over the seven queries for each data set are shown in Tables 4 and 5 for the Haar and D4 wavelets for the dry and water probes, respectively. (EF5 values are provided in Tables S3 and

Table 4. AUC Results Following Wavelet Thumbnail and Coarse GRID Alignments Using the Dry Probe Are Compared with ROCS Shape and ROCS ComboScore[a]

	Haar	D4	coarse GRID	coarse* GRID	ROCS Shape	ROCS ComboScore
CDK2	0.711	0.667	0.547	0.575	0.655	*0.730*
COX2	*0.890*	0.884	0.590	0.607	0.702	0.745
EGFR	0.587	0.580	0.602	0.604	0.678	*0.808*
INHA	0.685	*0.728*	0.566	0.594	0.696	0.588
P38	*0.827*	0.820	0.596	0.660	0.762	0.792
PDE5	0.748	0.734	0.625	0.617	*0.941*	0.877
PDGFRB	*0.852*	0.851	0.620	0.681	0.733	0.764
SRC	0.898	0.911	0.598	0.662	*0.940*	0.883
thrombin	0.892	0.810	0.615	0.678	*1.000*	0.990
VEGFR2	0.832	*0.842*	0.539	0.588	0.812	0.798
mean	0.792	0.783	0.590	0.627	0.792	*0.798*

[a]The best performing alignment technique for each data set is italics. The column headed coarse GRID presents the results using the same parameters as used for the wavelet methods. The column headed coarse* GRID refers to a parameter set optimized on the coarse GRIDs (see text for details).

Table 5. AUC Results Following Wavelet Thumbnail and Coarse GRID Alignments Using the Water Probe Are Compared with ROCS Shape and ROCS ComboScore[a]

	Haar	D4	coarse GRID	coarse* GRID	ROCS Shape	ROCS ComboScore
CDK2	0.532	0.556	0.533	0.560	0.655	*0.730*
COX2	0.598	0.592	0.621	0.602	0.702	*0.745*
EGFR	0.643	0.638	0.555	0.589	0.678	*0.808*
INHA	0.572	0.591	0.591	0.594	*0.696*	0.588
P38	0.759	0.764	0.733	0.741	0.762	*0.792*
PDE5	0.654	0.623	0.591	0.654	*0.941*	0.877
PDGFRB	0.563	0.585	0.554	0.554	0.733	*0.764*
SRC	0.651	0.639	0.647	0.633	*0.940*	0.883
thrombin	0.831	0.861	0.737	0.783	*1.000*	0.990
VEGFR2	0.546	0.556	0.548	0.569	*0.812*	0.798
mean	0.635	0.641	0.611	0.628	0.792	*0.798*

[a]The best performing alignment technique for each data set is italics. The column headed coarse GRID presents the results using the same parameters as used for the wavelet methods. The column headed coarse* GRID refers to a parameter set optimized on the coarse GRIDs (see text for details).

S4 in the Supporting Information.) Results for coarse GRIDs are also presented for comparison alongside the ROCS Shape (a shape only score) and ROCS ComboScore. It is clear that the dry probe tends to give much better performance than the water probe, which is consistent with the results shown previously.

When compared to ROCS Shape and ROCS ComboScore, the thumbnails for the dry probes demonstrate comparable virtual screening performance, although the relative performance varies by data set. However, ROCS is faster by 2 orders of magnitude (e.g., to align 395 structures to a query takes approximately 3 s in ROCS, and approximately 600 s using wavelet thumbnails, excluding the time taken to generate the thumbnails). It should be noted here that we have not attempted to optimize the alignment method, and in addition to code optimization, there is also considerable scope for parallelism in our method. Since each alignment can be examined independently, these tasks could be distributed across multiple CPU cores. Furthermore, both the alignment and the comparison steps involve calculations performed on hundreds (or possibly thousands, depending on the size of the volumes) of independent grid points. These types of processes are well suited to a GPU implementation which would result in significant speed-ups for these steps in the algorithm.[40]

The coarse GRID results are considerably worse than the thumbnail scores, especially for the dry probe. However, the clique detection parameters were not optimized for coarse GRIDs, therefore an additional experiment was carried out in which the optimum parameters for coarse GRIDs were identified based on the parametrization experiment described previously. The best parameter combination for coarse GRIDs was 25−1.5−0.6, i.e., the same magnitude ratio and distance tolerance as for thumbnails, however, for a considerably larger number of nodes. The increase in the size of the graphs resulted in much longer computation times, and although the results were improved (column headed coarse GRIDs*), they were still considerably worse that the results for wavelet thumbnails. Thus we conclude that the thumbnails provide a significantly more effective method of data reduction than simply sampling a field at lower resolutions, with this due to each point in a thumbnail encoding data from multiple points in the uncompressed field.

The degree to which the actives retrieved by the thumbnails are complementary to those retrieved by ROCS ComboScore and ROCS Shape was also investigated. This was quantified

Table 6. Tanimoto Similarity (Tan) between Actives and Normalized Sum of Unique Actives (norm sum) Retrieved in the top 5% by Dry Probe Thumbnail Alignment Compared To ROCS ComboScore and ROCS Shape

	ROCS ComboScore				ROCS Shape			
	Haar Tan	Haar norm sum	D4 Tan	D4 norm sum	Haar Tan	Haar norm sum	D4 Tan	D4 Norm Sum
CDK2	0.292	0.162	0.197	0.229	0.199	0.210	0.152	0.257
COX2	0.253	0.479	0.244	0.500	0.264	0.500	0.231	0.529
EGFR	0.334	0.127	0.384	0.119	0.296	0.159	0.337	0.159
INHA	0.134	0.299	0.101	0.390	0.313	0.143	0.228	0.260
P38	0.464	0.238	0.350	0.333	0.299	0.397	0.279	0.429
PDE5	0.335	0.286	0.337	0.300	0.320	0.243	0.305	0.286
PDGFRB	0.362	0.486	0.467	0.386	0.308	0.471	0.371	0.386
SRC	0.385	0.386	0.374	0.400	0.402	0.314	0.461	0.271
thrombin	0.575	0.190	0.574	0.111	0.531	0.222	0.530	0.143
VEGFR2	0.396	0.357	0.395	0.357	0.540	0.194	0.524	0.204
mean	0.353	0.301	0.342	0.312	0.347	0.285	0.342	0.292

dx.doi.org/10.1021/ci200348h | J. Chem. Inf. Model. 2012, 52, 757−769

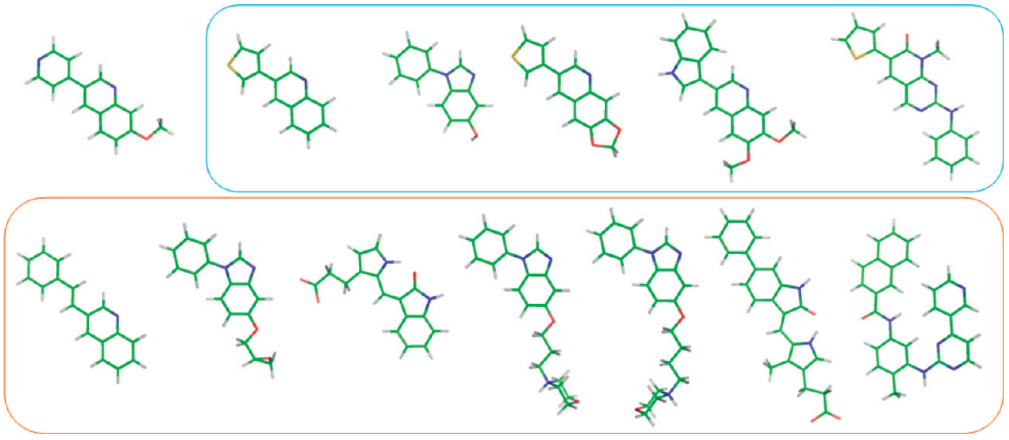


Figure 9. The unique actives retrieved by ROCS ComboScore (blue) each exhibit significant structural similarity to the query. The unique actives retrieved by dry probe thumbnail alignment (orange) demonstrate greater structural diversity.

using the Tanimoto similarity coefficient applied to the top 5% of ligands retrieved by each method and by the normalized sum of actives uniquely retrieved by the wavelet thumbnails. The latter was given as the number of unique actives retrieved using thumbnail alignment divided by the number of compounds in the top 5%, which varies by data set. These results are provided in Table 6. To demonstrate complementary behavior, the normalized sum of unique actives should be large, while the Tanimoto similarity between actives retrieved by both techniques should be small. On average, the number of unique actives retrieved through thumbnail alignment is indeed large. The Tanimoto similarity is also found to be high, indicating that in addition to the retrieval of a large number of unique actives, a large number of common actives are also retrieved.

The dry probe thumbnail alignment is superior in some cases when compared to ROCS for heterogeneous data sets (see mean pairwise similarities using MDL public Keys in Table 1) both in terms of the numbers of actives retrieved (Table 4) and in the complementarity of the hits (Table 6). Since it could be said that discriminating between heterogeneous actives and decoys is a more difficult task than for homogeneous actives, this is an encouraging result concerning the utility of dry probe thumbnail alignment. Furthermore this difference in behavior between data sets of a particular character emphasizes the complementary nature of these two approaches. Comparing the actives retrieved using thumbnails to those retrieved using ROCS shape only, it can be seen that for the dry probe, the normalized sum of actives unique to thumbnail alignment is lower than when compared with ROCS ComboScore, but the Tanimoto similarity scores are very similar. This indicates that ROCS Shape retrieves more actives than ROCS ComboScore, of which more are common to thumbnail alignment. Thus, as expected, dry probe thumbnail alignment cannot be said to be as complementary to ROCS shape only, as it is to ROCS ComboScore.

Figure 9 provides an example of the complementary behavior of the wavelet thumbnails and ROCS ComboScore. The PDGFRB data set provides the greatest normalized sum of unique actives using dry probe Haar thumbnails; within this data set, the query with code ZINC03832261 provides the greatest number of unique actives retrieved using wavelet thumbnails. The query is illustrated along with the actives uniquely retrieved using ROCS ComboScore, and the actives uniquely retrieved using Haar 16^3 thumbnails. The unique actives retrieved using ROCS ComboScore are similar in size to the query, whereas the unique actives retrieved using thumbnail alignment tend to be larger and indicate that the combination of thumbnail alignment, and the dry probe has been more effective in identifying compounds that share a similar skeleton to the query but which have quite different substituents attached to the skeleton. This can be explained by the use of clique detection in the thumbnail alignment method, which enables similar subshapes to be identified and used to generate an alignment.

CONCLUSIONS

We have shown that wavelet thumbnails provide an effective way of compressing the data stored in molecular interaction fields. In particular, they are significantly more effective than the much simpler method of extracting the data at a lower resolution, in what we have termed coarse GRIDs. Their superior performance compared to coarse GRIDs is due to each point in the thumbnail being some function of a larger region of points in the original grid; this is not the case for coarse GRIDs. We have demonstrated their effectiveness by developing an alignment procedure for thumbnail representations of GRIDs which enables the thumbnails to be used in virtual screening experiments. The average performance of the wavelet thumbnails over a number of queries and several different activity classes is comparable to the well-known ROCS program in terms of enrichment factors. The virtual screening results presented here are based on a single probe, with the dry probe providing the best results. Our current work is focused on combining the fields from different probes in order to generate alignments that take account of electrostatic propensities in addition to shape. In terms of speed, however, despite the significant reduction in size of the thumbnails compared to the original GRID fields, our alignment method does not compete with ROCS at present, being 2 orders of magnitudes slower.

Molecular interaction fields, such as those computed by GRID, are recognized as providing a very good description of ligand binding propensity and are increasingly being used in the drug discovery process in applications, such as docking and 3D QSAR. However they are very large and difficult to use in a virtual screening context. We have shown that the wavelet thumbnails provide an effective way of compressing this data into a reduced volume that retains the spatial relationships between the important parts of the field. The effectiveness of

the thumbnails in retaining the information content of the original fields, as demonstrated here, also supports our earlier findings where wavelet thumbnails were shown to be effective in 3D QSAR applications. If the main issue is simply compression, then it may be that the truncation method is most useful, rather than thumbnails, since greater degrees of data compression can be achieved with full-sized GRIDs reconstructed as required.

■ ASSOCIATED CONTENT

Supporting Information

Graphical illustrations of the Haar and Daubechies 4-tap wavelets and enrichment factors corresponding to the results presented in Tables 2−5. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: v.gillet@sheffield.ac.uk. Telephone:+44-1142-222652.

Present Address

[§]Computational Research Division, Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, California 94720, United States.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank Molecular Discovery for provision of the GRID program, Openeye for ROCS, the Chemical Computing Group for MOE, and the Cambridge Crystallographic Data Centre for Relibase+. The work was funded by BBSRC and GlaxoSmithKline via an industrial CASE studentship awarded to R.L.M.

■ REFERENCES

(1) Goodford, P. J. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J. Med. Chem.* **1985**, *28*, 849−857.

(2) Cruciani, G. *Molecular Interaction Fields: Applications in Drug Discovery and ADME Prediction*; Wiley-VCH: Weinheim, Germany, 2006.

(3) Vonitzstein, M.; Wu, W. Y.; Kok, G. B.; Pegg, M. S.; Dyason, J. C.; Jin, B.; Phan, T. V.; Smythe, M. L.; White, H. F.; Oliver, S. W.; Colman, P. M.; Varghese, J. N.; Ryan, D. M.; Woods, J. M.; Bethell, R. C.; Hotham, V. J.; Cameron, J. M.; Penn, C. R. Rational design of potent sialidase-based inhibitors of influenza-virus replication. *Nature* **1993**, *363*, 418−423.

(4) Pastor, M.; Cruciani, G.; Watsons, K. A. A strategy for the incorporation of water molecules present in a ligand binding site into a three-dimensional quantitative structure activity relationship analysis. *J. Med. Chem.* **1997**, *40*, 4089−4102.

(5) Green, S. M.; Marshall, G. R. 3D-QSAR: a current perspective. *Trends Pharmacol. Sci.* **1995**, *16*, 285−291.

(6) Goodford, P. Multivariate characterization of molecules for QSAR analysis. *J. Chemometr.* **1996**, *10*, 107−117.

(7) Cruciani, G.; Carosati, E.; De Boeck, B.; Ethirajulu, K.; Mackie, C.; Howe, T.; Vianello, R. MetaSite: understanding metabolism in human cytochromes from the perspective of the chemist. *J. Med. Chem.* **2005**, *48*, 6970−6979.

(8) Zamora, I.; Afzelius, L.; Crucianis, G. Predicting drug metabolism: a site of metabolism prediction tool applied to the cytochrome P450 2C9. *J. Med. Chem.* **2003**, *46*, 2313−2324.

(9) Cruciani, G.; Crivori, P.; Carrupt, P. A.; Testa, B. Molecular fields in quantitative structure−permeation relationships: the VolSurf approach. *J. Mol. Struct.* **2000**, *503*, 17−30.

(10) Baroni, M.; Cruciani, G.; Sciabola, S.; Perruccio, F.; Mason, J. S. A common reference framework for analyzing/comparing proteins and ligands. Fingerprints for ligands and proteins (FLAP): Theory and application. *J. Chem. Inf. Model.* **2007**, *47*, 279−294.

(11) Sciabola, S.; Stanton, R. V.; Mills, J. E.; Flocco, M. M.; Baroni, M.; Cruciani, G.; Perruccio, F.; Mason, J. S. High-throughput virtual screening of proteins using GRID molecular interaction fields. *J. Chem. Inf. Model.* **2010**, *50*, 155−169.

(12) Pastor, M.; Cruciani, G.; McLay, I.; Pickett, S.; Clementi, S. GRid-INdependent Descriptors (GRIND): A novel class of alignment-independent three-dimensional molecular descriptors. *J. Med. Chem.* **2000**, *43*, 3233−3243.

(13) Perruccio, F.; Mason, J. S.; Sciabola, S.; Baroni, M. FLAP: 4-point pharmacophore fingerprints from GRID. In *Molecular Interaction Fields: Applications in Drug Discovery and ADME Prediction*; Cruciani, G., Ed.; Wiley-VCH: Weinheim, Germany, 2006; pp 83−102.

(14) Cheeseright, T. J.; Mackey, M. D.; Melville, J. L.; Vinter, J. G. FieldScreen: Virtual screening using molecular fields. Application to the DUD data set. *J. Chem. Inf. Model.* **2008**, *48*, 2108−2117.

(15) Martin, R. L.; Gardiner, E.; Gillet, V. J.; Munoz-Muriedas, J.; Senger, S. Wavelet approximation of GRID fields: Application to quantitative structure-activity relationships. *Mol. Inf.* **2010**, *29*, 603−620.

(16) Burrus, C. S.; Gopinath, R. A.; Guo, H. *Introduction to Wavelets and Wavelet Transforms*; Prentice-Hall Inc.: Upper Saddle River, NJ, 1998.

(17) Stollnitz, E. J.; DeRose, T. D.; Salesin, D. H. *Wavelets for Computer Graphics*; Morgan Kaufmann Publishers, Inc.: San Francisco, CA, 1996.

(18) Mallat, S. G. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Trans. Patt. Anal. Mach. Intell.* **1989**, *11*, 674−693.

(19) Skodras, A.; Christopoulos, C.; Ebrahimi, T. The JPEG2000 Still Image Compression Standard. *IEEE Signal Process. Mag.* **2001**, *18*, 36−58.

(20) Nath, S. K.; Vasu, R. M.; Pandit, M. Wavelet based compression and denoising of optical tomography data. *Opt. Commun.* **1999**, *167*, 37−46.

(21) Muraki, S. Volume data and wavelet transforms. *IEEE Comput. Graphics Appl.* **1993**, *13*, 50−56.

(22) Carson, M. Wavelets and molecular structure. *J. Comput-Aided. Mol. Des.* **1996**, *10*, 273−283.

(23) Lio, P. Wavelets in bioinformatics and computational biology: state of art and perspectives. *Bioinformatics* **2003**, *19*, 2−9.

(24) Bajaj, C.; Castrillon-Candas, J.; Siddavanahalli, V.; Xu, Z. Q. Compressed representations of macromolecular structures and properties. *Structure* **2005**, *13*, 463−471.

(25) Sundling, C. M.; Sukumar, M.; Zhang, H.; Embrechts, M. J.; Breneman, C. M. Wavelets in chemistry and cheminformatics. In *Reviews in Computational Chemistry*, Lipkowitz, K. B., Cundari, T. R., Gillet, V. J., Eds.; Wiley: Hoboken, NJ, USA, 2006; Vol. *22*, pp 295−329.

(26) Beck, M. E.; Schindler, M. Quantitative structure-activity relations based on quantum theory and wavelet transformations. *Chem. Phys.* **2009**, *356*, 121−130.

(27) Bajaj, C.; Ihm, I.; Park, S. 3D RGB image compression for interactive applications. *ACM Trans. Graphics* **2001**, *20*, 10−38.

(28) Rodler, F. F. Wavelet Based 3D Compression for Very Large Volume Data Supporting Fast Random Access. *BRICS*, Computer Science Department, University of Aarhus: Aarhus, Denmark, 1999.

(29) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235−242.

(30) Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking sets for molecular docking. *J. Med. Chem.* **2006**, *49*, 6789−6801.

(31) Good, A. C.; Oprea, T. I. Optimization of CAMD techniques 3. Virtual screening enrichment studies: a help or hindrance in tool selection? *J. Comput-Aided. Mol. Des.* **2008**, *22*, 169−178.

(32) Hendlich, M.; Bergner, A.; Günther, J.; Klebe, G. Relibase: Design and development of a database for comprehensive analysis of protein-ligand interactions. *J. Mol. Biol.* **2003**, *326*, 607−620.
(33) *Molecular Operating Environment (MOE)*; The Chemical Computing Group: Montreal, Canada, 2000.
(34) Grant, J. A.; Gallardo, M. A.; Pickup, B. T. A fast method of molecular shape comparison: A simple application of a gaussian description of molecular shape. *J. Comput. Chem.* **1996**, *17*, 1653−1666.
(35) Rush, T. S.; Grant, J. A.; 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*, 1489−1495.
(36) Thorner, D. A.; Willett, P.; Wright, P. M.; Taylor, R. Similarity searching in files of three-dimensional chemical structures: Representation and searching of molecular electrostatic potentials using field-graphs. *J. Comput-Aided. Mol. Des.* **1997**, *11*, 163−174.
(37) Bron, C.; Kerbosch, J. Finding all cliques of an undirected graph. *Commun. ACM* **1973**, *16*, 575−577.
(38) Kabsch, W. Solution for best rotation to relate 2 sets of vectors. *Acta Crystallogr., Sect. A* **1976**, *32*, 922−923.
(39) DeLano, W. L. *The PyMOL Molecular Graphics System*; DeLano Scientific LLC: San Carlos, CA, 2006.
(40) Liao, Q.; Wang, J.; Watson, I. A. Accelerating two algorithms for large-scale compound selection on GPUs. *J. Chem. Inf. Model.* **2011**, *51*, 1017−1024.