



# Shape Signatures: speeding up computer aided drug discovery

Peter J. Meek<sup>1</sup>, ZhiWei Liu<sup>1</sup>, LiFeng Tian<sup>1</sup>, Ching Y. Wang<sup>2</sup>, William J. Welsh<sup>2</sup> and Randy J. Zauhar<sup>1</sup>

<sup>1</sup> Department of Chemistry & Biochemistry, University of the Sciences in Philadelphia, 6005. 43rd Street, Philadelphia, PA 19104, USA

<sup>2</sup> Department of Pharmacology, Robert Wood Johnson Medical School, University of New Jersey, 675 Hoes Lane, Piscataway, NJ 08854, USA

Identifying potential lead molecules is becoming a more automated process. We review Shape Signatures, a tool that is effective and easy to use compared with most computer aided drug design techniques. Laboratory researchers can apply this *in silico* technique cost-effectively without the need for specialized computer backgrounds. Identifying a potential lead molecule requires database screening, and this becomes rate-limiting once the database becomes too large. The use of Shape Signatures eliminates this concern and offers molecule screening rates that are in advance of any currently available method. Shape Signatures provides a conduit for researchers to conduct rapid identification of potential active molecules, and studies with this tool can be initiated with only one bioactive lead or receptor site.

## Background

Identification of small molecules with selective bioactivity, whether intended as potential therapeutics or as tools for experimental research, is central to progress in medicine and the life sciences. One approach to generating active compounds is to synthesize vast numbers of molecules (perhaps using combinatorial synthesis) that conform to Lipinski's Rule of Five [1], a powerful set of criteria obeyed by the majority of druglike compounds. It is estimated that  $10^{50}$ – $10^{60}$  compounds [2–4] could be synthesized while adhering to these rules.

Lipinski's Rules provide an important filter for selecting compounds with druglike characteristics [5–8], although they say nothing about a compound's potential for activity against a particular target (nor, for that matter, the toxicological or metabolic profile). This leaves a vast chemical space to be explored, and has prompted the development of numerous tools to further refine the selection of candidate compounds from a given library. With current drug development costs to get a drug to market fast approaching US\$1000 million, the need for reliable screening methods has become paramount. Computer aided drug design (CADD) represents a potentially useful tool for this purpose, and

has made great inroads into improving the odds of finding bioactive leads [9,10]. However, for CADD to be truly successful its tools should be provided in an easily available and usable format, for the benefit of the wider scientific community.

Four factors exist with respect to the successful implementation and dissemination of CADD techniques. These include: the availability and utility of databases; cost; time; and dependence upon the knowledge and expertise of the investigator. It is of extreme importance to address these issues to bring CADD to life science researchers. The tools we are implementing go a long way toward addressing the above issues. In fact, our goal is that the user submits their query, consisting of a bioactive molecule or a receptor site, and receives a list of hits with no further need for specialized chemical or computing expertise.

Making use of known scientific knowledge and data is imperative for the application of a CADD investigative process [9,11–13]. There are countless examples of online resources that are available [14] (<http://cactus.nci.nih.gov/ncidb2/download.html>, <http://chembank.broad.harvard.edu>), but these tend to be in an unsuitable format for immediate use. These resources often require manipulation by personnel with significant computer programming expertise before the information can be applied to the investigation of scientific data

Corresponding author: Meek, P.J. ([p.meek@usip.edu](mailto:p.meek@usip.edu))

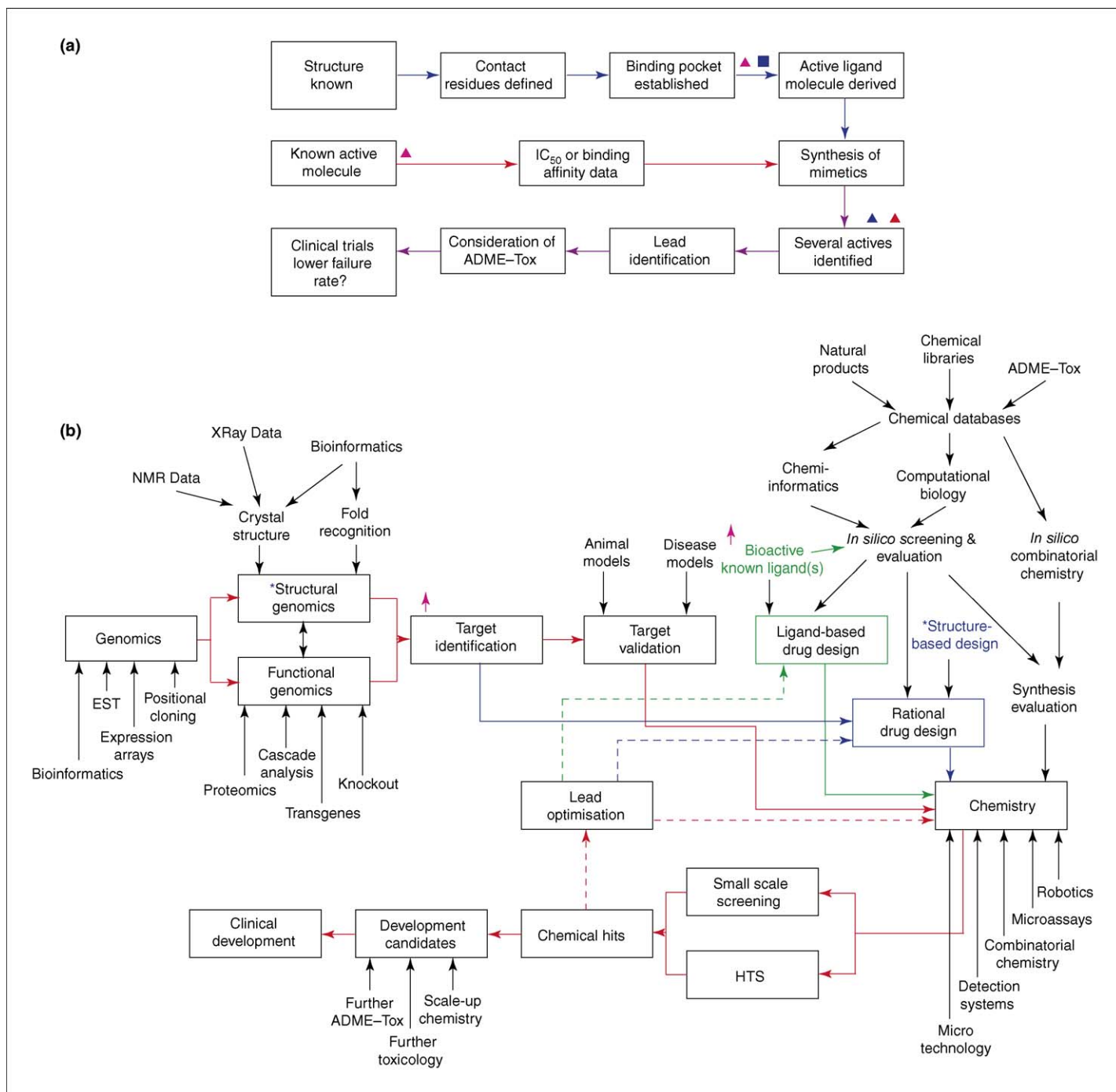


FIGURE 1

### Flow charts depicting the role of Shape Signatures and other computational techniques used in the drug discovery process.

**(a)** A flow diagram of the initial phases of drug discovery, the diagram indicates where computer aided drug design (CADD) techniques are used at each phase. The blue arrow depicts structure-based enquiries, red arrows are ligand-based enquiries and purple arrows depict either or both approaches. The specified computational techniques can be found online via most search engines. Applying Shape Signatures (pink triangle) in a structure-based or ligand-based manner requires one receptor binding pocket or one biologically active molecule, respectively. Shape Signatures identifies possible cross reactivity with other ligands and/or targets. Further, Shape Signatures could offer the potential to consider ADME-Tox even before actual chemical synthesis and testing, this is a very active area of our research. Structure-based (blue square): if a binding site is determined use Combiflexx, FAST and/or SPROUT to build fragments to fill the binding pocket. Use CAESA to determine feasibility of chemical synthesis. Test molecules experimentally to evaluate *in silico* procedure(s). Structure-based (blue triangle): if active molecules >30–40 and binding affinity or IC<sub>50</sub> data spans 3–4 orders of magnitude use QSAR or CoMFA, with GOLD and/or database screening UNITY. Establish a pharmacophore based on QSAR or CoMFA output. If active molecules <30–40 use GOLD to define structural features and establish a structure-based pharmacophore in UNITY. Ligand-based (red triangle): if active molecules >15 and binding affinity or IC<sub>50</sub> data spans 3 orders of magnitude use HQSAR. If active molecules >30–40 use DISCotech, QSAR, CoMFA or Catalyst to aid in establishing a ligand-based pharmacophore in UNITY and/or Catalyst. **(b)** The flow diagram depicts the orchestration between different disciplines and techniques that amalgamate into the process of drug discovery. To track the general outline of the process effectively, the unbroken red line should be followed from box to box in the direction of the arrows. Black arrows highlight areas of science that feed

and hypotheses. By contrast, tailor-made resources (<http://www.gykbio.com/informatics/dbprod.htm>; <http://www.wiley.com/WileyCDA/WileyTitle/productCd-0471748927,miniSiteCd-STMDB.html>; <http://www.wiley.com/WileyCDA/WileyTitle/productCd-0471743925,miniSiteCd-STMDB.html>; <http://www.scientific.thomson.com/products/wdi>) incur significant financial outlay.

Databases of small molecules with measured physicochemical properties are the foundation of CADD. Especially important are catalogues of commercially available compounds and known drugs [14] (<http://chembank.broad.harvard.edu>), but not to be overlooked are libraries of natural products [15–23]. A National Cancer Institute (NCI) survey quoted that, between 1981 and 2002, 61% of 877 small-molecule new chemical entities introduced as drugs could be traced to, or inspired by, natural compounds. This included 78% of antibacterial and 74% of anticancer drugs in this period [24]. Also to be considered is the complementary approach of taking the structure of the protein receptor and using it to screen databases for compatible molecules. Although not as well-developed as small-molecule methods, receptor-based techniques are bolstered by the ever-expanding collection of potential drug targets found in the Protein Data Bank (<http://pd-beta.rcsb.org/pdb/Welcome.do>), and successes derived thereof [25].

### Current CADD approaches

There are many techniques for conducting CADD [10,13], all of which aim to assist the discovery of new and potent molecules that could eventually be useful as biological tools or ideally go on to become drugs [26]. Initially, the goal is to accomplish ‘rapid elimination of swill’ [2]; in other words, removal of inactive molecules while retaining all potentially active molecules within a database. CADD techniques fall into one of two broad categories: ligand-based methods that focus on the structures of known active molecules and scan databases for compounds that are similar on the basis of structure, shape or physicochemical properties; and structure(receptor)-based methods that use the structure of the target active site as a template, seeking small molecules that will fill the available void, while fulfilling hydrogen-bonding opportunities. These two approaches can be combined as required and as the data available warrants. Full, well-documented reviews of many such CADD techniques, including descriptions of theory and practical details, are readily available [10,13]. When conducting CADD it is of great importance to use the techniques in conjunction with biologically derived data [9,11–13]. Although there have been successes when biological data have not been used with CADD [27–30], there have also been many failures [31–34].

### Shape Signatures

The Shape Signatures method [35] has several advantages over other CADD techniques. The entire focus of the method is on molecular shape and polarity and does not rely on additional computed or measured physicochemical properties, nor does it require training data concerning activity – it is therefore applicable

at the earliest stages of drug discovery (Figure 1). Most important, from the perspective of the end user, is the simplicity of applying the method; there is no need to worry about constructing multiple chemical queries or building complex pharmacophore models, all that is needed is the structure of a single, mildly active compound. The method is extremely fast, capable of screening  $\sim 10^9$  molecules per day against a single query on a single processor (Table 1), and it is trivial to extend the method to multiple processors.

If we cut away all the elaborate intricacies – complex motion and dynamics of molecular interaction between ligand and receptor – what is left? The answer, first proposed by Emil Fischer (in 1890) [36], is shape. How and why does a certain small molecule interact with a comparatively huge biological target and induce a response? The answer, again, is shape [37]. Using Shape Signatures deals specifically with the fundamental commonality between ligand shape and receptor shape [35]. The method finds other molecules within a database that have a shape close to a query molecule, or a shape similar to a target binding site and, therefore, aids identification of potential lead compounds. As mentioned previously, the principle objective is to screen through a molecular database rapidly and to enrich for molecules similar to the query (i.e. rapid elimination of swill [2]). This is rapidly and effectively [38] accomplished with the Shape Signatures method (Table 1). It is crucial to note that this apparent simplicity of shape plus polarity encompasses more-precise and widely used descriptors, such as hydrogen bonding, hydrophobicity and electrostatics. One should also be aware that a protein is merely a coiled spring awaiting a substrate of the right shape and polarity to induce a response; flexibility of such a target can be approximated by observing the shapes of several structures of the same target with different ligands bound.

The basis of the Shape Signatures method [35] is ray-tracing, a technique adapted from computer graphics [39–41] for a radically different application. The first step is to encapsulate the molecule in a solvent-accessible surface (Figure 2a), giving the appearance of a rubber sheet wrapped around the atoms. This surface is triangulated, meaning that it has been broken into small triangular elements and is effectively an irregular polyhedron. The Shape Signatures algorithm selects one of the triangles at random, and initiates a ray (Figure 2b) with a random direction, which is directed into the interior of the polyhedron. This ray then propagates, without attenuation, by the rules of optical reflection (Figure 2c and 2d) – think of the molecule as a very oddly shaped and inside-out disco ball! The reflecting ray explores the geometry of the molecule and in an interesting way; because the ray cannot pass through the surface it specifically samples those pairs of surface points that can be connected by an uninterrupted line of sight – a feature that is intimately connected to the overall shape of the molecule. This exploration is carried out very efficiently because of the stochastic nature of the algorithm.

Once a ray-trace is prepared for the molecule, probability distributions are derived from the trace (Figure 2d); it is precisely these distributions, accumulated and stored as histograms, that

information into the central drug discovery process. The blue arrows indicate computational structure-based design, whereas green arrows indicate computational ligand-based drug design. Broken arrows indicate feedback from initial discoveries and additional data to improve the candidate molecule for downstream application. The two small pink arrows demonstrate where Shape Signatures can be applied within the discovery process (adapted from [http://www.cfes.com/documents/pharma/03-Technology\\_Trends.PDF](http://www.cfes.com/documents/pharma/03-Technology_Trends.PDF)).

TABLE 1

## Comparison of 'between speeds' for database construction and screening

Technique	Number of molecules per unit of time					(CPU + RAM)
	Minute	Hour	Day	Month	Year	Computer specifications
<u>Database construction</u>						
Shape Signatures (ray-tracing)						
	5.6	$3.3 \times 10^2$	$8.0 \times 10^3$	$2.4 \times 10^5$	$2.9 \times 10^6$	1 × 3.5 GHz Intel Pentium 4, 2 GB
	$1.1 \times 10^4$	$2.1 \times 10^4$	$5.1 \times 10^5$	$1.5 \times 10^7$	$1.9 \times 10^8$	32 × Opteron 2.6 GHz, 0.5 GB
RAPTOR model (generation)						
	4.5	$2.7 \times 10^2$	$6.5 \times 10^3$	$1.9 \times 10^5$	$2.4 \times 10^6$	1 × 1.8 GHz Pentium 4
<u>Database screening</u>						
Shape Signatures (screening)						
	$1 \times 10^6$	$6.0 \times 10^7$	$1.4 \times 10^9$	$4.3 \times 10^{10}$	$5.3 \times 10^{11}$	1 × 3.5 GHz Intel Pentium 4, 2 GB
	$6.4 \times 10^7$	$3.8 \times 10^9$	$9.2 \times 10^{10}$	$2.8 \times 10^{12}$	$3.4 \times 10^{13}$	32 × Opteron 2.6 GHz, 0.5 GB
Rapid overlay of chemical structures (ROCS)						
	$7.6 \times 10^2$	$4.6 \times 10^4$	$1.1 \times 10^6$	$3.3 \times 10^7$	$4.0 \times 10^8$	1 × 3.5 GHz Intel Pentium 4, 2 GB
UNITY (pharmacophore)						
3pt	81.6	$4.9 \times 10^3$	$1.2 \times 10^5$	$3.5 \times 10^6$	$4.3 \times 10^7$	1 × RG14000 MIPS SGI 1 × 500 mHz, 1 GB
4pt	68.9	$4.1 \times 10^3$	$9.9 \times 10^4$	$3.0 \times 10^6$	$3.6 \times 10^7$	
5pt	56.6	$3.4 \times 10^3$	$8.2 \times 10^4$	$2.4 \times 10^6$	$3.0 \times 10^7$	
Docking GOLD 2.2						
	0.6	34.7	$8.3 \times 10^2$	$2.5 \times 10^4$	$3.0 \times 10^5$	2 × 3.0 GHz Intel Xeon, 4 GB
RAPTOR (molecule evaluation)						
	8.0	$4.8 \times 10^2$	$1.2 \times 10^4$	$3.5 \times 10^5$	$4.2 \times 10^6$	1 × 1.8 GHz Pentium 4

RAPTOR data was obtained from Marcus Lill [50]. ROCS was conducted in-house and is available from openeye (<http://www.eyesopen.com/products/applications/rocs.html>).

have been named Shape Signatures. The simplest derivation is the distribution of ray-trace segment lengths, where a segment is just the portion of the trace between two successive reflections. Because the domain of this signature is 1D (length only), it is termed a 1D signature (Figure 2e). Molecular properties are coupled to shape by considering the total length of the segments on either side of a reflection point, together with some surface property measured at the reflection. For example, the molecular electrostatic potential (MEP) computed over the surface. In this case, a joint probability distribution is constructed that depends upon geometry and molecular polarity. Because the domain of this distribution is 2D, and the property in this case is the MEP, this is termed a 2D-MEP signature (Figure 2f). Signatures of even higher dimension can be contemplated. The ray-trace for indinavir, the Merck HIV protease inhibitor, is shown in Figure 2d. The 1D and 2D-MEP signatures (Figure 2e and 2f) computed for this molecule are also illustrated.

Evaluation of Shape Signatures comparisons between the query and database entries is a process distinct from that of generating the Shape Signatures. The comparison between query molecule and each database entry is achieved by using a separate comparison algorithm. The comparison is achieved by integrating the total absolute difference between query and target signature, taking into account the dimension of the domain (1D, 2D or higher dimension). Because distributions are approximated as histo-

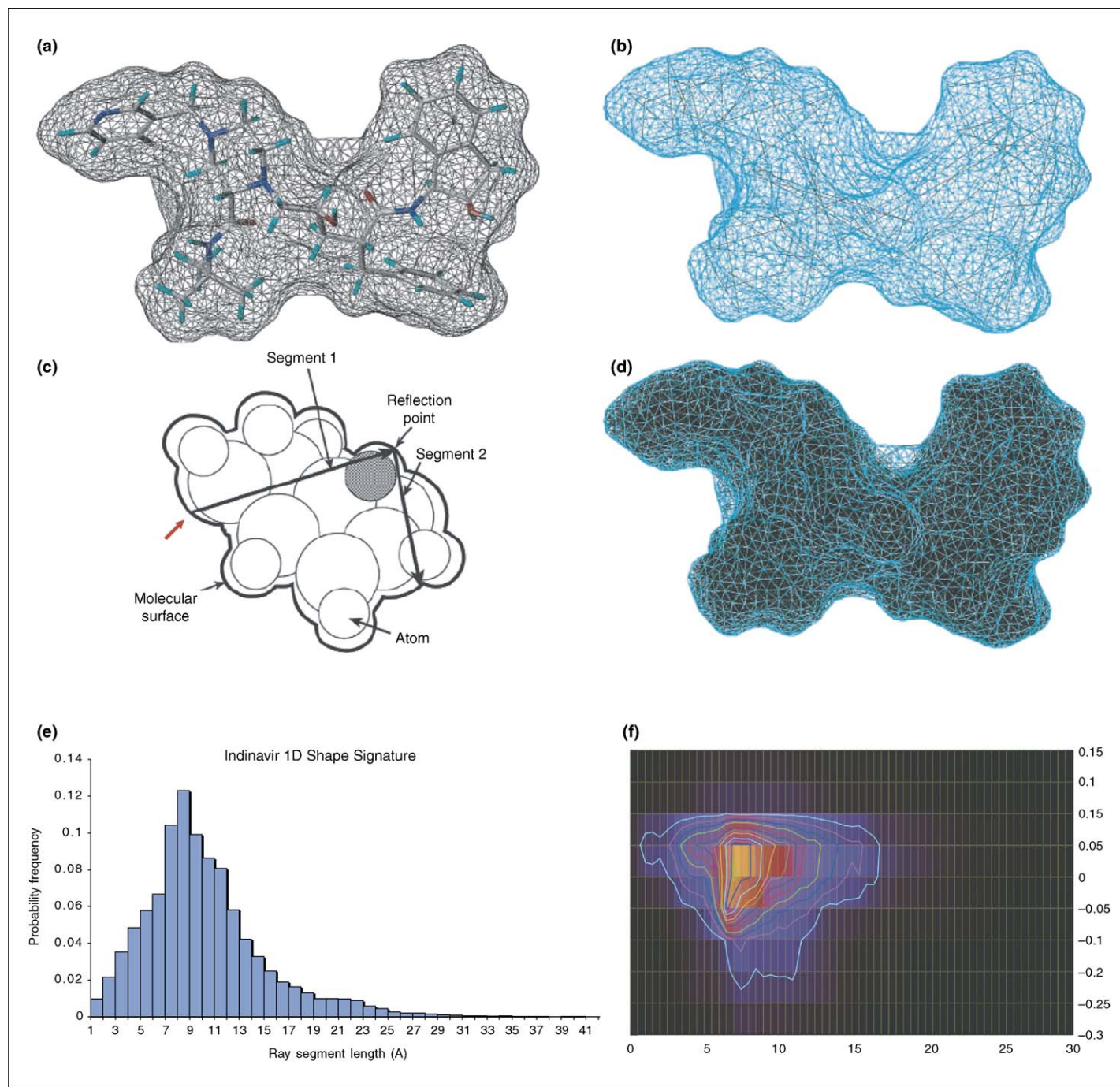
## BOX 1

## Shape Signatures equations for 1D and 2D-MEP comparisons

$$L_1^{1D} = \sum_i |H_i^1 - H_i^2|$$

$$L_1^{2D} = \sum_i \sum_j |H_{ij}^1 - H_{ij}^2|$$

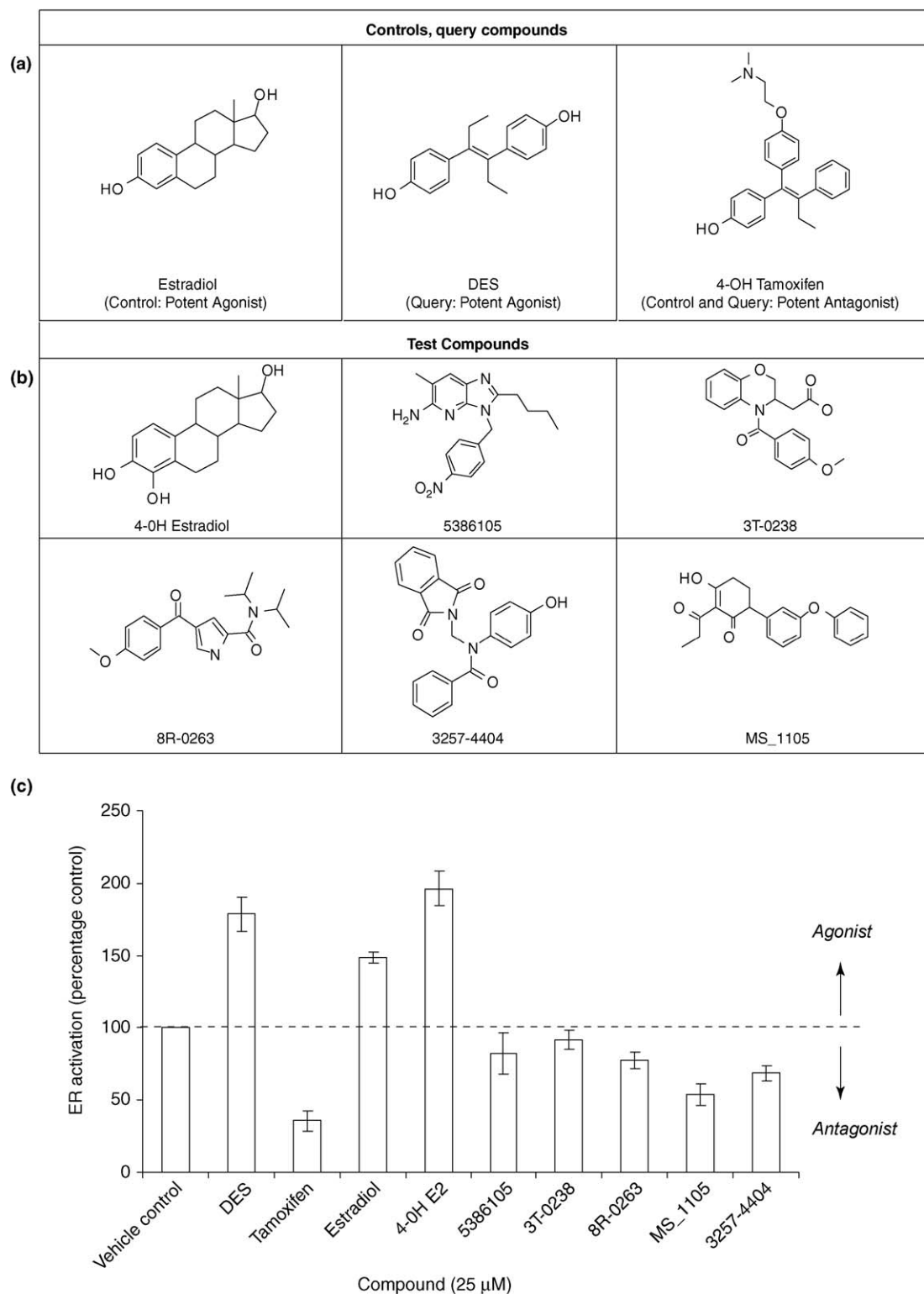
These equations are used for computing distances between Shape Signatures histograms for 1D and 2D signatures. Here, for 2D signatures,  $H_{i,j}^1$  represents the bin count or probability for the first Shape Signature,  $H_{i,j}^2$  the second, and where the bin with length index  $i$  and MEP index  $j$  is indicated. For the 1D signature only the length index is used (Figure 2e compared to itself (1D) would yield a Shape Signature score of 0.0, likewise Figure 2f (2DMEP) to itself would yield 0.0. The reason is that there are no differences in the probability distributions. Differences will only be observed between different molecular structures).

**FIGURE 2**

**Process of Shape Signatures generation.** The structure of indinavir enclosed in the triangulated solvent accessible surface (a) generated using SMART [35]. Propagation of a ray-trace around the inside of the triangulated molecular surface schematic (b), the red arrow indicates the start of the ray-trace. Propagation of the ray-trace around the indinavir molecule with 100 ray-trace segments (c) and 10,000 ray-trace segments (d). On completion of a Shape Signature (d) the generated trace is illustrated by (e), a histogram denoting probability distribution (ordinate) of ray-trace segment lengths (abscissa), a 1D Shape Signature. A Shape Signature trace defined by ray-trace segment lengths and a mean electrostatic potential (MEP) 'descriptor' is shown by (f) a contour plot, a 2D Shape Signature. It is these data stored as text files that are compared when performing a database screen (i.e. a molecular comparison).

grams, this numerical integration really means finding differences between corresponding bins in the signatures being compared, and requires little computing time (Box 1). The smaller the overall difference (score) the more similar the signatures are and the greater the molecular similarity we infer. The end result of a Shape Signatures database search is a collection of hits that are presented in order of increasing score (decreasing similarity).

Although the method is in its infancy, comparisons have already been shown to be very effective for biological applications [35,38,42]. A current area of intensive research in our laboratory is the development of further ways of applying Shape Signatures for multiple queries and catering for structural and ligand-based perspectives. It is not unreasonable to expect to be able to carry out well over a thousand million comparisons a day on a single processor

**FIGURE 3**

**Examples of molecules selected by Shape Signatures demonstrating estrogenic activity on human estrogen receptor (ER) based on known controls.** Three known estrogenic compounds **(a)** estradiol, diethylbesterol and tamoxifen are all known to interact with human ER control. Using these as query molecules to screen an in-house database ( $1.2 \times 10^6$  molecules), the returned molecules **(b)** were tested by assay [50]. Taking each of the selected 'hits' from the search in turn, a 25  $\mu\text{M}$  sample was tested using the nuclear receptor (NR) peptide ER $\alpha$  ELISA kit (Active Motif, Carlsbad, CA, USA) according to manufacturer's instructions. 17 $\beta$ -Estradiol and tamoxifen were included as part of the kit. Briefly, a pre-coated 96-well-plate was supplied with an optimized peptide containing the consensus binding motif of ER $\alpha$  co-activator SRC-1. Each compound was incubated for 1 hour with MCF-7 nuclear extract (Active Motif) and the co-activator peptide in each well. The ligand-activated

(Table 1) – numbers that, until now, were not attainable unless using 2D and/or 3D fingerprints (in addition, these fingerprints take time to generate before they can be used). The method is trivial to parallelize for database generation and comparison, hence submission of queries via the internet could be implemented and accomplished in the next few years. The data returned dramatically reduces the initial database size, a vital part of any successful CADD strategy [2]. Further, Shape Signatures also achieves the second important quality of CADD: enrichment for molecules from the same class as the query molecule [38,42]. Initial successes with Shape Signatures and estrogen-related molecules (Figure 3) were supported with experimental findings [42]. A potential novel molecule involved in analgesia has been discovered using Shape Signatures and offers considerable benefits over current compounds in this class. Many groups [39–41] have previously used similar techniques for virtual-spatial recognition, but they have not applied them for use in biological applications [with the exception of Ankerst and colleagues (<http://www.dbs.informatik.uni-muenchen.de/Publikationen/Papers/ISMB99.final.pdf>)]. The Shape Signatures method is distinct from these other previous attempts that use histogram comparisons for recognition of objects [39–41] or for lead discovery (<http://www.dbs.informatik.uni-muenchen.de/Publikationen/Papers/ISMB99.final.pdf>). The Shape Signatures technique is rotationally invariant, meaning that it does not require, nor depend on, orientation of the ligand or receptor site to obtain a reproducible histogram profile [35,39]. This is imperative for users submitting queries or engaging in regular use and systematic comparisons. The Shape Signatures technique will execute in the same manner regardless of whether the molecule sits on the side, head, bottom or any combination thereof.

### Computational evaluation of Shape Signatures

The most time-consuming phase of generating Shape Signatures is ray-tracing, typically consuming ~10 s of central processing unit (CPU) time for a single molecule [although this has diminished significantly (~20-fold) with implementation of recent experimental code and more up-to-date processors]. Fortunately this penalty must be paid only once, when the database is first constructed. Although significant, this computational demand has not proved to be an obstacle and we have readily generated databases for hundreds of thousands (up to millions) of compounds on a modest cluster (nonetheless, we are exploring the possibility of using the same specialized hardware that is employed for computer-generated image effects in Hollywood films to the very different problem of computing Shape Signatures libraries). Again, we stress that once the database has been generated it is ready to be used in any number of studies, thus in its application (query, comparison of molecules) the Shape Signatures approach is very fast indeed. This is ideal for the current way *in silico* database screening applications are utilized and conducted.

### Validating the Shape Signatures approach to CADD

Much effort and time is invested to orchestrate the completion of a well-kept molecular database; forethought, standardization,

adaptability, portability and unique identification are all essential for robust functionality. The important caveat is simply one of search speed, because, despite all the time invested in database construction, the amount of queries and questions posed (additively) to a database far exceeds the amount of time for construction. Hence, subsequent to the creation of the Shape Signatures database, there will be considerable net gain by reducing the time needed to perform *in silico* screening. This is further augmented and readily accomplished by searches carried out on parallel processors providing a linear increase in the already noteworthy performance of the method.

Shape Signatures can be performed to model biological interaction even when there is relatively little known biological data and, as a result, they can be positioned into the very early stages of drug discovery (Figure 1). The technique is very fast (Table 1) and the molecules in the databases we are developing [14] can be obtained from vendors. One of the greatest advantages of Shape Signatures, however, is that, unlike many other CADD methods, it will be free of charge for academic researchers. Via collaborations (and once Shape Signatures is online) feedback from life science experts will aid development of the algorithm and iron out any interface teething problems. Use of Shape Signatures has demonstrated significant reduction in the number of molecules that require testing to produce bioactive molecules [42,43], around one out of every three molecules reported active (Figure 3).

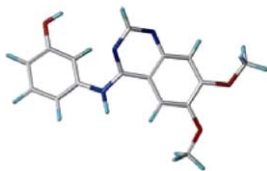
As an example of the application of the Shape Signatures method, the query molecule WHI-P131 [44,45], a tyrosine kinase inhibitor of interest as a therapeutic against several diseases (including leukaemia and amyotrophic lateral sclerosis), is shown along with the top Shape Signatures hits located in the NCI database (Table 2). Note the clear shape similarity between query and hits that nonetheless differ significantly in details of chemical structure. Finding these matches by other CADD methods would entail either generating a large set of structural queries based on the molecule of interest or in generating a pharmacophore model for the inhibitor. In either case there would be the presumption of specialized knowledge to construct the queries actually used to scan the database and a small omission in constructing that query could mean missing important hits. By contrast, the ligand-based Shape Signatures search is very easy to carry out; the investigator need only present the single compound of interest as the query. By contrast, the technique Topomers is an effective means to achieve lead hopping [46], without the need to generate lots of structural queries, but on the downside it requires significant financial investment and computer expertise.

Shape Signatures is not overtly sensitive to small changes in conformation. However, it is vital to point out that the method is sensitive to stereoisomerism, especially when multiple chiral centers are present or when *cis* and *trans* states need to be distinguished. Using the program STERGEN (<http://www.mol-net.de/software/stergen/index.html>) to generate the stereoisomers compatible with the chemical formulae of the compounds in our databases, we are presently following this up with conformational analysis using the Molecular Operating Environment (MOE) pack-

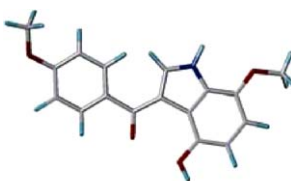
ER $\alpha$  was first detected using a primary antibody specific for ER $\alpha$  and further with horseradish peroxidase (HRP)-conjugated secondary antibody. (c) One molecule was strongly agonistic (4-OH E $_2$ ), and two were strongly antagonistic MS\_1105 and 3257-4404. The other three molecules were only marginally antagonistic.

TABLE 2

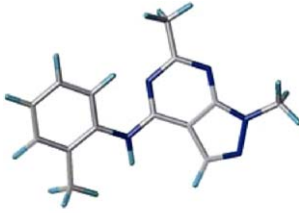
## Shape Signatures search of the NCI database with WHI-P131



WHI-P131

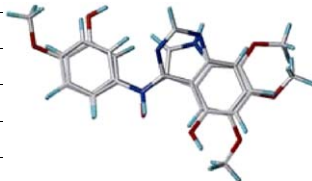


NCI\_HIT1




NCI\_HIT2

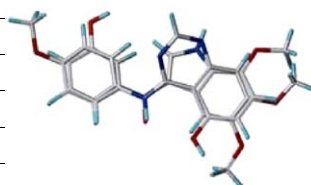
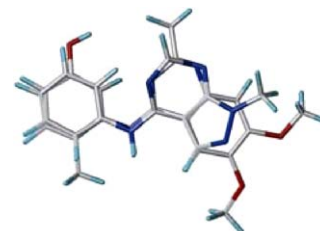
Rank	Molecule	Score
<b>1D</b>		
1	WHI-P131	0.000000
2	NCI_HIT1	0.050280
3	NCI_HIT2	0.081440
<b>2D</b>		
1	WHI-P131	0.000000
2	NCI_HIT2	0.255917
3	NCI_HIT1	0.273748



WHI-P131  
+  
NCI\_HIT1



WHI-P131 + NCI\_HIT2

WHI-P131  
+  
NCI\_HIT1

WHI-P131 + NCI\_HIT2

age (<http://www.chemcomp.com/software.htm>). The intent is that the production databases derived by Shape Signatures will bundle multiple stereoisomers and conformers with each chemical formula, each with its own collection of Shape Signatures descriptors. This will provide maximal flexibility in those cases where the chiral form of a compound, or the differences between bound and unbound conformations, are unknown. This represents a huge leap in CADD technology offering a myriad of applications for database screening. Determining a predicted bound state of a ligand without the requirement of a large biologically derived dataset is a particularly enticing prospect. Although expecting the Shape Signatures approach to become an important and widely used CADD tool, it is not claimed to be a panacea for all of the many issues and complications that plague CADD, nor a substitute for established known techniques [10,13]. What Shape Signatures will offer is the availability of an effective CADD technology to a much larger audience.

There is one potential drawback of Shape Signatures that must be addressed up front. Because the method collapses a great deal of chemical space onto very compact descriptors, it is inevitable that a Shape Signatures search will turn up false-positives in the hit list. To test how significant this issue might be, an independent assessment of the quality of the hits produced by a Shape Signatures search [with the angiotensin-converting enzyme (ACE) inhibitor enalapril] was established. First, docking of known ACE inhibitors that have measured  $IC_{50}$  data [47] into ACE (PDB code 1O86) [48] was carried out using GOLD [49] producing a significant correlation between inhibitor  $pIC_{50}$  and GOLD score, with scores for the known actives ranging from 50 to 87. In the second phase of our study we identified the top 250 hits produced by a Shape Signatures search for the single query, Enalapril. This was conducted using an in-house database of 423 drugs and the NCI database (<http://cactus.nci.nih.gov/ncidb2/download.html>) of

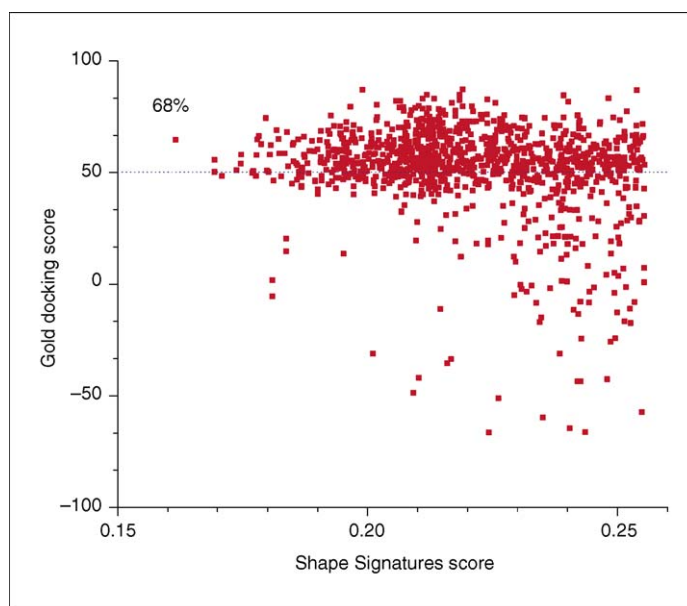


FIGURE 4

**Evaluation of Shape Signatures with the GOLD algorithm.** The plot indicates Shape Signatures scores (abscissa) versus GOLD (ordinate), with the dashed line at 50 for GOLD indicating a good docking hit. Angiotensin-converting enzyme (ACE) inhibitors with measured  $IC_{50}$  data [47] were docked into ACE (PDB code 1O86) [48] using GOLD [49]. This produced significant correlation between the inhibitor  $pIC_{50}$  and GOLD score (the scores for the known active ACE inhibitors ranged from 50 to 87). Subsequently, the top 250 hits produced by a Shape Signatures search of the query molecule enalapril (from the in-house database of 423 drugs and the NCI database [15] of >250,000 compounds) were evaluated with GOLD. The search identified 11 of the possible 20 known ACE inhibitors, and the fitness scores by GOLD of all 250 hits gave >75 for 4% and >50 for almost 70% of the hits. Assuming that the correlation between experimental  $IC_{50}$  and GOLD score holds, the Shape Signatures method is shown to readily identify known actives and interesting lead molecules, and with a modest proportion of false-positives.

>250,000 compounds. The hit compounds were extracted from the database and docked into the active site of ACE using GOLD. Not only did our search, conducted using a single query, immediately identify 11 of the 20 well-characterized ACE inhibitors but the fitness scores by GOLD evaluation of all 250 hits gave >75 for 4% and >50 for 70% of the hits (Figure 4). Assuming that the correlation between experimental  $IC_{50}$  and GOLD score holds, the Shape Signatures method has been shown to readily identify known actives and interesting lead molecules, and with a modest proportion of false-positives.

## Summary

Shape Signatures is a new and exciting CADD technique that is unlike other known and established tools. It produces compact descriptors of molecular shape that can easily be coupled with other properties of interest. Using this representation, comparisons between molecules can be made extremely quickly using readily available computing hardware (Table 1). It is evident that Shape Signatures has the capacity to identify and suggest similar molecules to a query by comparing shape, or shape with an associated descriptor (e.g. MEP). In the study we reviewed, Shape Signatures was shown to quickly identify molecules likely to bind to a selected target, as validated by an independent method (GOLD). The synergy in the results between Shape Signatures and the well-known molecular docking technique GOLD endorses the technique for identifying potential lead molecules. Detection of active molecules with high efficiency (Figure 3) and novel action provided further evidence that Shape Signatures truly functions in a real biological context. In summary, the speed and reproducibility of results

offered by Shape Signatures, coupled with a high incidence of identified bioactive molecules and novel leads, demonstrates the desired qualities for a new and emerging CADD technique.

The Shape Signatures method has huge advantages over other CADD techniques and can be used in the absence of a receptor with as little as one weakly bioactive molecule or ligand. Thus, the Shape Signatures method offers the user an excellent means of searching a large diverse collection of compounds with very little information. Before the advent of Shape Signatures, CADD techniques required considerably more input data before they could be implemented. Hence, Shape Signatures can be applied to the very earliest stages of the drug discovery process (Figure 1). Our goal is to develop Shape Signatures using databases of molecules that are already available from vendors, so that the researcher will simply be able to order them from the respective provider for assay. Promising compounds can then be used in further investigations and be used as starting points for medicinal chemistry campaigns to optimize their activity.

The overarching goal is to unite life science experts with powerful CADD technology, to conduct research rapidly and effectively without the need to read huge tomes of user manuals. We live in an age of plug and play, point and click and turn on and go, and it is about time CADD followed suit.

## Acknowledgements

We kindly acknowledge Kim Sweeny for the drug discovery scheme (Figure 1) that was adapted from his report for the Centre for Strategic Economic Studies at the Victoria University of Technology in Australia (<http://www.cses.com>).

## References

- Lipinski, C.A. *et al.* (1997) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* 23, 3–25
- Zeman, S.P. (2004) Charting chemical space: finding new tools to explore biology. *Nature* 1–3. The 4th Horizon Symposium, Black Point Inn, Maine, U. S. A., on 20th–22nd May
- Dean, P.M. *et al.* (2001) Industrial-scale, genomics-based drug design and discovery. *Trends Biochem. Sci.* 19, 288–292
- Oprea, T.I. (2002) Virtual screening in lead discovery: A viewpoint. *Mol.* 7, 51–62
- Myers, P.L. (1997) Will combinatorial chemistry deliver real medicines? *Curr. Opin. Biotechnol.* 8, 701–707
- Lazo, J.S. and Wipf, P.J. (2000) Combinatorial chemistry and contemporary pharmacology. *Pharmacol. Exp. Ther.* 293, 705–709
- Li, J. *et al.* (1998) Targeted molecular diversity in drug discovery: Integration of structure-based design and combinatorial chemistry. *Drug Discov. Today* 3, 105–112
- Lipinski, C.A. (2003) Chris Lipinski discusses life and chemistry after the Rule of Five. *Drug Discov. Today* 8, 876–877
- Burley, S.K. and Park, F. (2005) Meeting the challenges of drug discovery: a multidisciplinary re-evaluation of current practices. *Genome Biol.* 6, 330
- Blundell, T.L. *et al.* (2002) High-throughput crystallography for lead discovery in drug design. *Nat. Rev. Drug Discov.* 1, 45–54
- Marx, V. (2004) Structures speak up. *Chem. Eng. News* 82, 22–30
- Gryzbowski, B.A. *et al.* (2002) Combinatorial computational method gives new picomolar ligands for a known enzyme. *Proc. Natl. Acad. Sci. U. S. A.* 99, 1270–1273
- Bender, A. and Glen, R.C. (2004) Molecular similarity: a key technique in molecular informatics. *Org. Biomol. Chem.* 2, 3204–3218
- Irwin, J.J. and Shoichet, B.K. (2005) ZINC – A free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* 45, 177–182
- Lesney, M.S. (2004) Nature's pharmaceuticals: natural products from plants remain at the core of modern medicinal chemistry. *Today's Chemist at Work* pp. 26–32
- Maclean, D.B. and Luo, L.G. (2004) Increased ATP content/production in the hypothalamus may be a signal for energy-sensing of satiety: studies of the anorectic mechanism of a plant steroidal glycoside. *Brain Res.* 1020, 1–11
- Natarajan, S. *et al.* (2001) Healing of an MRSA-colonized, hydroxyurea-induced leg ulcer with honey. *J. Dermatolog. Treat.* 12, 33–36
- Tonks, A.J. *et al.* (2002) Honey stimulates inflammatory cytokine production from monocytes. *Cytokine* 21, 242–247
- Lusby, P.E. *et al.* (2002) Honey: a potent agent for wound healing? *J. Wound Ostomy Continence Nurs.* 29, 295–303
- Lusby, P.E. *et al.* (2002) Honey: a potent agent for wound healing? Comment in *J. Wound Ostomy Continence Nurs.* 29, 273–274
- Cooper, R.A. *et al.* (2002) The efficacy of honey in inhibiting strains of *Pseudomonas aeruginosa* from infected burns. *J. Burn Care Rehabil.* 23, 366–370
- Knepper, K. *et al.* (2003) Natural product-like and other biologically active heterocyclic libraries using solid-phase techniques in the post-genomic era. *Comb. Chem. High Throughput Screen.* 6, 673–691
- Mineno, T. *et al.* (2002) Solution-phase parallel synthesis of an isoflavone library for the discovery of novel antiangiogenic agents. *Comb. Chem. High Throughput Screen.* 5, 481–487
- Newman, D.J. *et al.* (2003) Natural products as sources of new drugs over the period of 1981–2002. *J. Nat. Prod.* 66, 1022–1037
- Borman, S. (2005) Drugs by design. *Chem. Eng. News* 83, 28–30
- Mohan, V. *et al.* (2005) Docking: successes and challenges. *Curr. Pharm. Des.* 11, 323–333
- Babine, R.E. and Bender, S.L. (1997) Molecular recognition of protein-ligand complexes: applications to drug design. *Chem. Rev.* 97, 1359–1472
- Farber, G.K. (1999) New approaches to rational drug design. *Pharmacol. Ther.* 84, 327–332
- Kubinyi, H. (1997) QSAR and 3D QSAR in drug design Part 1: methodology. *Drug Discov. Today* 2, 457–467
- Kubinyi, H. (1997) QSAR and 3D QSAR in drug design Part 2: applications and problems. *Drug Discov. Today* 2, 538–546
- Martin, Y.C. (1997) Challenges and prospects for computational aids to molecular diversity. *Perspect. Drug Discov.* 7/8, 159–172
- Ajay, V. and Murcko, M.A. (1995) Computational methods to predict binding free energy in ligand-receptor complexes. *J. Med. Chem.* 38, 4953–4967

- 33 Böhm, H.J. and Klebe, G. (1996) Detailed review on ligand-protein interactions, binding modes of ligands, structure-based and computer-aided drug design. *Angew. Chem. Int. Ed. Engl.* 35, 2589–2614
- 34 Böhm, H.J. *et al.* (1999) Combinatorial docking and combinatorial chemistry: Design of potent non-peptide thrombin inhibitors. *J. Comput. Aided Mol. Des.* 13, 51–56
- 35 Zauhar, R.J. *et al.* (2003) Shape Signatures: a new approach to computer-aided ligand- and receptor-based drug design. *J. Med. Chem.* 46, 5674–5690
- 36 Stryer L. *et al.* (2006) Biochemistry (6th edn), pp. 215, 1014, W.H. Freeman and Company, New York
- 37 Connolly, M.L. (1983) Solvent-accessible surfaces of proteins and nucleic acids. *Science* 221, 709–713
- 38 Nagarajan, K. *et al.* (2005) Enrichment of ligands for the serotonin receptor using the Shape Signatures approach. *J. Chem. Inf. Model.* 45, 49–57
- 39 Liu, X. *et al.* (2003) Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'03), pp. 1–8
- 40 Osada, R. *et al.* (2001) Matching 3D Models with Shape Distributions, *Shape Model. Int. May*, 154–166
- 41 Ohbuchi, R. *et al.* (2005) Shape-similarity search of 3D models by using enhanced shape functions. *Int. J. Comp. Appl. Tech.* 23, 70–85
- 42 Nagarajan, K. *et al.* Shape Signatures, a novel tool for virtual screening of chemical databases: detecting estrogenic compounds, *Chem. Res. Toxicol.* (in press)
- 43 Zhang, Q. *et al.* (2006) Discovery of novel triazole-based opioid receptor antagonists. *J. Med. Chem.* 49, 4044–4047
- 44 Amin, H.M. *et al.* (2003) Inhibition of JAK3 induces apoptosis and decreases anaplastic lymphoma kinase activity in anaplastic large cell lymphoma. *Oncogene* 22, 5399–5407
- 45 Jilek, R.J. *et al.* (2004) Lead Hopping Method Based on Topomer Similarity. *Chem. Inf. Comput. Sci.* 44, 1221–1227
- 46 Ghosh, S. *et al.* (1999) Structure-based design of potent inhibitors of EGF-receptor kinase as anti-cancer agents. *Anticancer Drug Des.* 14, 403–410
- 47 Kamenska, V. *et al.* (1999) The COREPA approach to lead generation: an application to ACE-inhibitors. *Eur. J. Med. Chem.* 34, 687–699
- 48 Natesh, R. *et al.* (2003) Crystal structure of the human angiotensin-converting enzyme-lisinopril complex. *Nature* 421, 551–554
- 49 Jones, G. *et al.* (1997) Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* 267, 727–748
- 50 Lill, M. (2004) RAPTOR: combining dual-shell representation, induced-fit simulation, and hydrophobicity scoring in receptor modeling: application toward the simulation of structurally diverse ligand sets. *J. Med. Chem.* 47, 6174–6186

## Five things you might not know about Elsevier

### 1.

Elsevier is a founder member of the WHO's HINARI and AGORA initiatives, which enable the world's poorest countries to gain free access to scientific literature. More than 1000 journals, including the *Trends* and *Current Opinion* collections and *Drug Discovery Today*, are now available free of charge or at significantly reduced prices.

### 2.

The online archive of Elsevier's premier Cell Press journal collection became freely available in January 2005. Free access to the recent archive, including *Cell*, *Neuron*, *Immunity* and *Current Biology*, is available on ScienceDirect and the Cell Press journal sites 12 months after articles are first published.

### 3.

Have you contributed to an Elsevier journal, book or series? Did you know that all our authors are entitled to a 30% discount on books and stand-alone CDs when ordered directly from us? For more information, call our sales offices:

+1 800 782 4927 (USA) or +1 800 460 3110 (Canada, South and Central America)  
or +44 (0)1865 474 010 (all other countries)

### 4.

Elsevier has a long tradition of liberal copyright policies and for many years has permitted both the posting of preprints on public servers and the posting of final articles on internal servers. Now, Elsevier has extended its author posting policy to allow authors to post the final text version of their articles free of charge on their personal websites and institutional repositories or websites.

### 5.

The Elsevier Foundation is a knowledge-centered foundation that makes grants and contributions throughout the world. A reflection of our culturally rich global organization, the Foundation has, for example, funded the setting up of a video library to educate for children in Philadelphia, provided storybooks to children in Cape Town, sponsored the creation of the Stanley L. Robbins Visiting Professorship at Brigham and Women's Hospital, and given funding to the 3rd International Conference on Children's Health and the Environment.