Large-Scale Comparison of Four Binding Site Detection Algorithms

Peter Schmidtke,[†,‡] Catherine Souaille,[†] Frédéric Estienne,[†] Nicolas Baurin,[†] and Romano T. Kroemer*[,†]

Sanofi-Aventis VA Research Centre, Structure Design & Informatics, 13 quai Jules Guesde, BP14, 94403 Vitry-sur-Seine, France and Departament de Fisicoquimica and Institut de Biomedicina (IBUB), Facultat de Farmacia, Universitat de Barcelona, 08028, Barcelona, Spain

Received January 18, 2010

A large-scale evaluation and comparison of four cavity detection algorithms was carried out. The algorithms SiteFinder, fpocket, PocketFinder, and SiteMap were evaluated on a protein test set containing 5416 protein−ligand complexes and 9900 apo forms, corresponding to a subset of the set used earlier for benchmarking the PocketFinder algorithm. For the holo structures, all four algorithms correctly identified a similar amount of pockets (around 95%). SiteFinder, using optimized parameters, SiteMap, and fpocket showed similar pocket ranking performance, which was defined by ranking the correct binding site on rank 1 of the predictions or within the first 5 ranks of the predictions. On the apo structures, PocketFinder especially and also SiteFinder (optimized parameters) performed best, identifying 96% and 84% of all binding sites, respectively. The fpocket program predicts binding sites most accurately among the algorithms evaluated here. SiteFinder needed an average calculation time of 1.6 s compared with 2 min for SiteMap and around 2 s for fpocket.

INTRODUCTION

The number of known protein three-dimensional (3D) structures in the public and private domains is constantly on the rise. For drug discovery purposes these structures are of great interest, as they can be exploited in the search for small molecules that bind to them and modulate their function. Of particular importance are the cavities at the protein surface, as they provide the best environment for anchoring small molecules.

In many cases cavities can be identified by the presence of natural substrates, a cofactor or a ligand. However some proteins are crystallized without any partner. Detecting cavities at the surface of these proteins can help in finding the natural substrate binding site, identifying binding pockets or allosteric sites to start the design of small-molecule ligands of therapeutic effect. Moreover, given a cavity in one protein, the detection of similar cavities in other proteins may provide hints to anticipate issues, such as selectivity or toxicity.

A first step toward exploitation and comparison of pockets as well as cavity-based annotation of proteins is therefore the comprehensive scanning of protein 3D structures with the aim of detecting all cavities of interest. These cavities can subsequently be analyzed and compared with novel programs, such as SuMo[1] or FLAP.[2] Given that these programs take surface shape and property into account, they are well suited for the task of cavity/pocket comparison, as opposed to programs that analyze protein backbone and topology, such as SARF2,[3] VAST,[4,5] DALI,[6,7] and FATCAT.[8,9]

Over the past years a number of different approaches have been developed to correctly predict binding sites on the protein surface. One can distinguish two different types of cavity finding algorithms: (i) evolutionary- and (ii) structure-based algorithms. The second category can be subdivided in geometry- and energy-based algorithms.

The most popular example of geometry-based algorithms in the public domain is putative active sites with spheres (PASS).[10] Other well-known geometry-based algorithms are SURFNET[11] and LIGSITE[12,13] (improved version of POCKET).[14] APROPOS[15] and CAST[16] are based on alpha shape analysis.[17,18] The recently published fpocket[19] uses similar properties derived as alpha spheres, already employed by the SiteFinder[20] algorithm.

Energy-based algorithms like PocketFinder,[21] the method introduced by Bliznyuk and Gready,[22] the computational mapping from the Vajda group,[23,24] the multiscale approach from Glick,[25] and the method developed by Ruppert,[26] or SuperStar[27] simulate the interactions of a solvent molecule on the protein surface in order to detect local surface properties of a cavity. Some of these methods still use a geometry-based step in order to measure the extent of the cavity, by tracing rays from grid points in the cavity. Nevertheless approaches like the multiple solvent mapping developed by the Vajda group are fully based on interaction energy calculations.

The accuracy of most of the cavity finding algorithms has not been evaluated on large data sets. Only the PocketFinder algorithm[21] published in 2005 provided a large-scale evaluation using a data set of 17.626 proteins from the Protein Data Bank (PDB). This evaluation included an assessment of the algorithms' capacity to recognize binding sites on apo forms as compared to the corresponding ligated proteins.

In the present study, a large scale evaluation of four cavity finding algorithms has been carried out. Two of them are implemented in two major molecular modeling packages:

* Corresponding author. E-mail: romano.kroemer@sanofi-aventis.com.
[†] Sanofi-Aventis VA Research Centre, In Silico Sciences/Drug Design, 13 quai Jules Guesde, BP14, 94403 Vitry-sur-Seine, France.
[‡] Departament de Fisicoquimica and Institut de Biomedicina (IBUB), Facultat de Farmacia, Universitat de Barcelona, 08028, Barcelona, Spain.

10.1021/ci1000289
Published on Web 09/09/2010

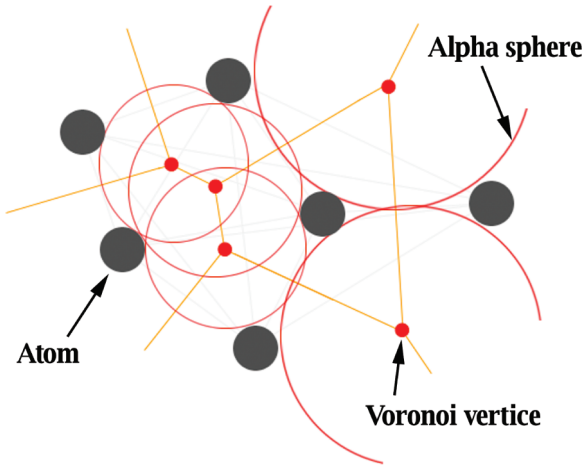

Figure 1. Example of an alpha sphere. The alpha sphere is displayed in red, and the three contacted atoms (2D) are in gray.

SiteFinder, which is implemented in the Molecular Operating Environment (MOE) software provided by the Chemical Computing Group (CCG),[28] and SiteMap,[29,30] provided by Schrödinger[31] and accessible through the Maestro graphical user interface or through the command line. To our knowledge, no evaluation of SiteFinder has been published up to now and a data set of 297 proteins[29,30] has been used for the development and optimization of SiteMap. Furthermore, PocketFinder and fpocket are compared to the previous two. PocketFinder, falling also in the category of energy-based algorithms, was chosen because it had previously been evaluated on the data set used in this study. Fpocket, a geometry-based algorithm, similar to SiteFinder, was included as a sole open source alternative to the previously cited methods.

The main criteria in the evaluation of all algorithms was success rate and accuracy of binding site identification as well as computational performance. Ease-of-use, such as scripting facilities and accessibility of results, were also considered. In the following, the methods evaluated are being described in more detail.

SiteFinder. SiteFinder falls into the category of geometric methods, since no energy models are used. Relative positions and accessibility of the receptor atoms are considered along with an approximate classification of chemical type. The method is based on the identification of regions of tight atom packing on the protein, filtering of exposed regions, hydrophobic/hydrophilic classifications, and the use of a definition of hydrophilicity that is invariant to protonation state.

No grid-based method is used for SiteFinder, this way the method is invariant to rotation of atomic coordinates, and less memory is required for the calculation.

The SiteFinder methodology is based upon alpha shapes which are a generalization of convex hulls developed by Edelsbrunner.[17] A collection of 3D points is triangulated using a modified Delaunay triangulation. For each resulting simplex (collection of four points), there is an associated sphere called alpha sphere (Figure 1). These spheres have different radii.

The collection of alpha spheres is pruned by eliminating those that correspond to inaccessible regions of the receptor as well as those that are too exposed to solvent. In addition, only small alpha spheres are retained since these correspond to locations of tight atom packing in the receptor. Each alpha sphere is classified as either "hydrophobic" or "hydrophilic", depending on whether the sphere is in a good hydrogen-bonding spot on the receptor. Hydrophilic spheres that are far from hydrophobic spheres are eliminated. All alpha spheres are clustered using a single linkage clustering algorithm. A key feature is that each cavity consists of one or more alpha spheres and at least one hydrophobic alpha sphere. Resulting cavities are ranked using the number of contacts with hydrophobic atoms of the receptor.

fpocket. The fpocket algorithm is based on very similar principles as SiteFinder. Using Voronoi tessellation through the free computational geometry library Qhull, fpocket filters Voronoi vertices and their corresponding alpha spheres according to alpha sphere minimum and maximum radii. The next three clustering steps are performed to aggregate nearby alpha spheres to form a pocket (set of alpha spheres). Each cluster of alpha spheres forms a putative pocket. Each pocket is scored based on a knowledge-based SiteScore, and the final pocket list is ranked using this score.

Compared to SiteFinder, fpocket is currently a command line driven open source cavity detection algorithm. Next to basic pocket prediction, fpocket integrates several tools for easy extraction of pocket descriptors and for testing scoring function. A druggability prediction score has recently been integrated as well.[32]

SiteMap. SiteMap is an energy-based cavity finding algorithm. It identifies probable binding sites through three main steps: (i) detection of cavities, (ii) characterization of detected cavities, and (iii) evaluation of characterized cavities.

In the first step, a 1 Å grid of site points is built around the entire protein; points overlapping the protein atoms are deleted. Then the algorithm filters out the site points located too far from the protein or displaying a low degree of "enclosure" within the receptor. The "enclosure" of site points is computed using rays traced in all directions. The number of rays cutting the protein surface at a certain distance is used to estimate the relative "enclosure" of the grid point. The points that fulfill these criteria are clustered into site point groups. Groups of site points are merged if the distance between them is below a predefined threshold and occurs in a solvent-exposed region. The default maximum distance between two grid points to be merged into the same group is 6.5 Å. The ratio of the distance between the centroids of the groups to their effective size (default value is 5) determines whether the groups are considered for merging.

During the second step—the mapping process—various properties of the cavity are calculated using the remaining site points. Hydrophobic and hydrophilic potentials are generated using van der Waals and electric field grids. SiteMap then partitions the accessible space in each site into hydrophobic, hydrophilic, and "neither/nor" regions. The hydrophilic map is further divided into hydrogen-bond donor and acceptor maps. The last step of the SiteMap cavity detection procedure is the site evaluation. Various scores are calculated by SiteMap. The main score (SiteScore) is based on a weighted sum of the following criteria:

- Number of site points: the number of grid points necessary to define the cavity.
- Exposure/enclosure: the property to measure how open the cavity is to the solvent.

- Hydrophobic/hydrophilic character and balance: the measure of the relative hydrophobic/hydrophilic character of the cavity. Balance is the ratio hydrophobic/hydrophilic score.
- Donor/acceptor character: the estimated hydrogen-bond donor intensity of a putative ligand in the cavity.

PocketFinder. The algorithm, published by An et al. in 2005, falls into the category of energy-based pocket detection algorithms. It makes use of a transformed Lennard-Jones potential calculated on the protein structure using an aliphatic carbon atom as a probe placed on a 1.0 Å spaced grid over the protein. Next, the grid is smoothed to emphasize regions with a consistently low Lennard-Jones potential over a given region in space. Out of these isolated regions, envelopes are created and further filtered for envelopes having a volume bigger than 100 $Å^3$. Finally, the resulting binding sites are ranked by volume.

MATERIALS AND METHODS

To carry out the comparison of the algorithms, a two-step procedure was applied. First, a preliminary study was performed on a small calibration data set with a view to eventually optimizing some of the parameters determining the cavity detection. Second, the algorithms with the assessed search parameters were used to check their ability to correctly identify binding sites on a large evaluation data set.

Preliminary Study. The calibration data set consists of 370 protein−ligand complexes obtained by X-ray diffraction with a resolution better than 2.5 Å. These structures are part of a MOE sample database (complex.mdb file shipped with MOE) with structures cleaned from water and crystallization additives, which could alter search results on the protein surfaces. No hydrogen atoms were considered for pocket detection with SiteFinder. Both cavity finding algorithms were applied in the absence of the ligand in the investigated binding site and tested for their performance on the following criteria:

- Percentage of found binding sites ranked as first (according the algorithm score).
- Percentage of found binding sites ranked within the first five positions.

The aim of the calibration step was to minimize: (i) the percentage of not found binding sites; (ii) the percentage of binding sites split up in multiple pockets, and (iii) the mean rank and standard deviation of the rank of a correctly detected binding site.

SiteFinder. Optimization was performed on three parameters of the SiteFinder algorithm (default values in brackets):

- Connect Dist (2.5 Å): connection distance between two alpha spheres, used to cluster alpha spheres into a common group (cavity).
- Minrad (2.0 Å): minimum threshold distance between the alpha spheres of a cluster and the centroid of this cluster.
- Da Dist (3.0 Å): maximum distance between hydrophilic alpha spheres and the nearest hydrophobic alpha sphere.

Connect dist and minrad, tuning the clustering procedure, were optimized with a combinatorial procedure. Da dist was thereafter optimized using the optimum values of the former parameters.

In addition, small cavities can be filtered out with the site minsize (3) parameter, referring to the number of alpha spheres in a site.

SiteMap. This algorithm makes use of 16 parameters that can be modified. The results obtained in the preliminary study with the default values of these parameters were satisfying. Moreover calculations on 370 structures using a single set of parameters took several days on a two-processor Linux workstation, thus it was not feasible performing an exhaustive parameter optimization within a reasonable amount of time. Therefore no parameter optimization was undertaken by us.

The MOE database previously compiled for the evaluation of SiteFinder was used to export receptor and ligand, each into separate files in PDB format. PDB receptor files were further converted into MAE file format (pdbconvert script provided by Schrödinger). Hydrogen atoms were then added to the receptor (applyhtreat script provided by Schrödinger), followed by an atom-typing step required to perform the energy calculation.

SiteMap does not handle structures with missing atoms. In order to overcome this issue, PrimeFill (Schrödinger) was used to build and refine the missing protein regions on the calibration data set. This process was however too long to be used on the evaluation data set. Thus, the evaluation step of SiteMap was performed only on structures prepared automatically using the prepwizard program provided by Schrödinger. Structures that could not be treated by SiteMap were excluded from the analysis for all algorithms.

No evaluation of the influence of different protonation states on the pocket prediction results is performed here, as one of the main objectives of this evaluation is to assess the suitability of SiteMap for high-throughput pocket prediction.

fpocket. No further calibration was carried out for fpocket.

PocketFinder. Also for this algorithm, no calibration was performed, as the results published by An et al.[21] were taken for this comparison.

Evaluation Study. The evaluation data set consisted initially of the 17 126 structures, with a resolution lower than 2.5 Å, generated to evaluate PocketFinder and to compute the so-called Pocketome.[21] When the present study was performed this data set was available on the Web site http://abagyan.ucsd.edu/index.html. It is the largest data set ever used to evaluate cavity finding algorithms. To our knowledge, so far, only PocketFinder has been evaluated with this data set. The set contains 5616 protein−ligand complexes and 11 510 apo structures on which the binding site location is known. At least one holo structure can be found among the 5616 complexes for each apo structure, thus enabling to assess whether the experimentally observed cavities in the complex structures can also be found in the apo structures.

An et al. used data deposited in the PDB from October 30, 2003. Between this date and today, major changes have been made for a multitude of the PDB structures in this data set. For some structures, PDB accession codes simply changed. For others, the known ligand molecule identifier changed. Also, a few PDB structures have been deleted from the PDB since then. For all the previously cited reasons, only 5416 structures out of the initial 5616 holo structures were retained for this study. New chain assignments and the previously cited reasons reduced the apo data set to 9900 structures. This data set reduction is also due to the fact that for example SiteMap could not run successfully (in an

automated manner) on some structures. Thus, to guarantee integrity of the whole data set for the different methods, the data set size was reduced by the structures that SiteMap cannot analyze. The final data set is provided as additional text files (c.f. Supporting Information).

First, a test on the 5416 holo proteins was performed in order to evaluate the ability of the four algorithms to identify the ligand binding site. Next, a second test was performed on the 9900 apo structures.

The criteria to evaluate the performance of both algorithms are:

- Percentage of correctly identified binding sites.
- Accuracy of the prediction.
- Correlation between results obtained on the holo and apo forms of the protein structures.
- Calculation time.
- Ease of use.

In the present analysis, structurally important cofactors, like hemes, biopterins, or chlorophyls, were considered as part of the receptor in order to build a functional unit. Thus these cofactors were included during binding site search. As some of the small molecules in the initial data set by An et al. correspond to this category of molecules, these structures were taken out of the data set. The final data set used on all algorithms is provided as csv files in the Supporting Information.

Definition of a Correctly Identified Ligand Binding Site. The definition of when a binding site is correctly identified is crucial in evaluations, such as the ones carried out here. In order to obtain results comparable to An et al., the same evaluation criterion as in the PocketFinder evaluation was used. This criterion, called relative overlap (RO) is defined as follows: $RO = (A_L \cap A_E)/A_L)$, where A_L is the solvent-accessible area of the receptor atoms within 3.5 Å from a bound ligand, and A_E is the solvent-accessible area of the receptor atoms within 3.5 Å from the predicted pocket envelope. A total mis-prediction would have $RO = 0$, whereas a perfect prediction would have $RO = 1.0$.

Whether the algorithm has detected correctly, a ligand binding site is determined by the overlap between the surface of the cavity atoms and the surface of the actual binding site atoms. If the RO is at least about 0.5, then the cavity will be assessed as "correctly identified".

In apo structures, the ligand binding site is defined by analogy with the holo structure after protein superposition.

Assessing the Accuracy of the Pocket Prediction. An et al. used the RO criterion to assess the accuracy of the pocket prediction. However, the accuracy of a prediction can be defined in various ways. The inherent disadvantage of using a criterion like the RO is that the bigger a predicted pocket gets, the more chance one has to reach a RO close to one, covering completely the known ligand binding site. If the purpose of a pocket prediction algorithm is however to propose reasonably sized binding pockets, then the RO alone is not enough to assess the accuracy of a prediction. Thus, a second criterion, called mutual overlap criterion (MOC) is introduced in this study. It is based on the same principle as the RO but is defined as: $MO = (A_L \cap A_E)/A_E)$. Similarly to RO, if the MO gets close to 1, then it is an indicator the predicted pocket is covering only the surface of the actual overlap between the ligand and predicted binding sites. Associating both RO and MO, one could get an estimate of the accuracy of pocket prediction.

Comparing Geometry-Based Methods with Energy-Based Methods. This evaluation intends to compare two energy-based pocket identification algorithms (SiteMap and PocketFinder) with two geometry-based algorithms (fpocket and SiteFinder). As the pockets are represented in very different ways such a comparison is not straightforward. Both PocketFinder and SiteMap use a grid to delimit the binding site. Thus a binding site can be represented as either an envelope or a set of grid points in the pocket. SiteFinder and fpocket produce more sparsely spaced alpha spheres. In order to assess if a binding site was correctly identified, the solvent accessible surface area of the pocket has to be calculated, and this can be done using a 3.5 Å distance from all grid points of the pocket or all alpha sphere centers. As grid points are more densely packed than alpha sphere centers, this would result in an underestimation of the correctly identified pocket surface for geometry-based methods.

In order to address this, a grid intended to be very similar to the SiteMap grid is packed into the alpha spheres of fpocket and SiteFinder. The general amber force field (GAFF) was used to assign van der Waals parameters to all atoms of the protein. Next, a 0.7 Å spaced grid is placed over the pocket. In these calculations, the radius and the well depth of the Lennard-Jones probe particles are taken to be 1.5 Å and 0.13 kcal/mol, respectively, as used by Halgren in SiteMap.[30] Finally, only those grid points within the alpha spheres were retained that are equal or further than the closest van der Waals equilibrium distance of the probe in the pocket. In all subsequent surface calculations, these retained grid points represent the pocket, on the contrary to the previously used alpha sphere centers (and volume).

PocketFinder uses a very similar representation of pockets and results obtained with SiteMap, and both geometry-based methods using the transformation presented in the previous paragraph are thus considered comparable.

Scripting and Statistical Analysis. Automation of cavity finding was scripted using the programming environment of each molecular modeling package. The Scientific Vector Language (SVL) was used in MOE (version 2006−2008). This scripting language is the proprietary MOE language and provides a flexible platform for users willing to develop their own methods.

In the Maestro molecular modeling suite, the Maestro Command Language is interfaced with Python. Versions 7.5 and 8.0 of Maestro were used in this study (see Results Section).

Analysis was performed using R statistical software version 2.2.1[33] and SpotFire DecisionSite 8.0.

All calculations were performed on biprocessor (2 × 3.6 Ghz) and 2 Gb RAM workstations, running under a RHEL WS release 3 distribution.

Results for pocket prediction for fpocket were taken from a precomputed pocket database (in-house). The transformation from alpha sphere-based pockets to grid-based pockets was performed using several Python based in-house libraries.

RESULTS

SiteFinder Preliminary Study. In order to evaluate SiteFinder with its maximum performance, all search pa-

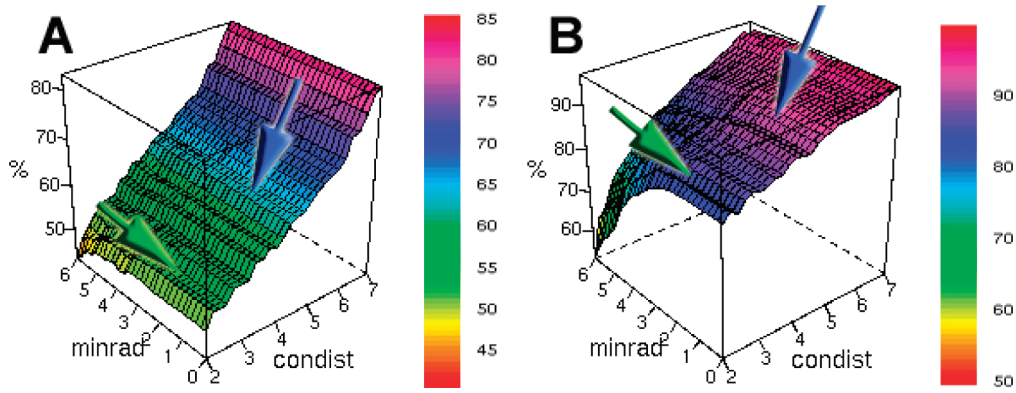

Figure 2. Optimization of SiteFinder parameters. (A) % of correctly identified binding sites ranked on the first rank in function of minrad and connect dist. (B) % of correctly identified binding sites found on the first five ranks in function of minrad and connect dist. Green and blue arrows represent default and optimized parameters, respectively.

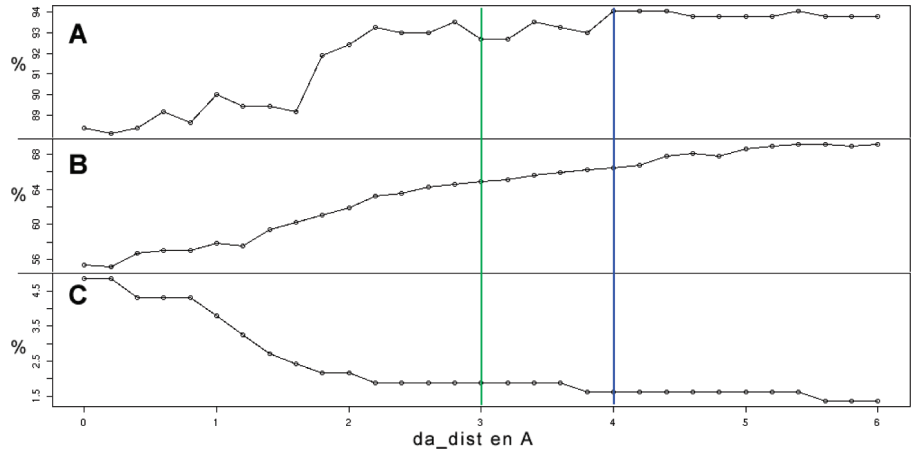

Figure 3. Optimization of SiteFinder parameters. (A) % of correctly identified binding sites ranked within the first five ranks as a function of da dist. (B) % of correctly identified binding sites ranked on the first rank as a function of da dist. (C) % of binding site not found by SiteFinder as a function of da dist.; Green and blue lines represent default and optimized parameters, respectively.

rameters were first optimized during the preliminary study. These calculations led to notable adjustments of minrad, connect dist and da dist values. Minrad was modified from 2.0 (default) to 1.8 Å(optimized), connect dist from 2.5 (default) to 4.6 Å(optimized), and da dist from 3.0 (default) to 4.0 Å(optimized).

In Figure 2A notable increase of correctly identified binding sites ranked on the first rank is observed with increasing connect dist value. Using default parameters (in brackets optimized parameters), SiteFinder ranks about 54 (65%) of identified binding sites on the first rank, 86 (92.7%) within the 5 first ranks. SiteFinder, with default parameters was not able to identify 3.2 (1.6%) of all binding sites during this optimization run on 370 protein structures. This optimization step was performed using a fixed value of 5 for site minsize parameter. Figures 2 and 3 show that cavity detection and ranking performance could be theoretically further enhanced by increasing the connect dist and the da dist values. However, care must be taken with further increasing these values. Considering, for instance, the connect dist parameter with its optimized value fixed at 4.6 Å, a further increase of this value would result in one single cavity at the protein surface. Therefore, connect dist and da dist were not increased further for the purposes of this study.

SiteFinder Full-Scale Evaluation. First, the evaluation step on SiteFinder was performed on the 5416 protein−ligand complexes full data set using default (in brackets optimized) parameters. Figure 4 illustrates that more than 75 (95%) of all binding sites are identified with a RO close to 1. At the threshold of a well-identified binding site, 95 (98%) of all binding sites are found. As shown in Figure 5, 70 (77%) of all found binding sites are ranked as first. Considering all binding sites found on ranks 1−5, the total would amount to 95 (98%) of all found binding sites.

The impressive increase in the percentage of found binding sites with an RO near 100% from SiteFinder with default parameters to the percentage found with optimized parameters shows that the RO alone is not a good enough criterion to evaluate the accuracy of a binding site prediction algorithm. Figure 6 depicts the MO of both parameter sets for SiteFinder, and one can observe a clear shift in accuracy from default to optimized parameters. This important shift toward lower MO values for optimized parameters clearly indicates that, although the RO is very high for most of the found binding sites, this comes at the cost of prediction of far too big binding sites.

Second, SiteFinder using default parameters (optimized parameters in brackets) was evaluated on 9900 apo structures. Here, more contrasted results were obtained. The algorithm was able to retrieve around 40 (65%) of all binding sites with a RO near 1, as shown in Figure 4. This corresponds to a drop of 35 (20%) compared to results obtained on holo structures. Also for ranking performance, a drop in the predictive power can be observed. SiteFinder ranks 42 (62%) of all found binding sites on rank 1 and 83 (98%) within the top five ranks.

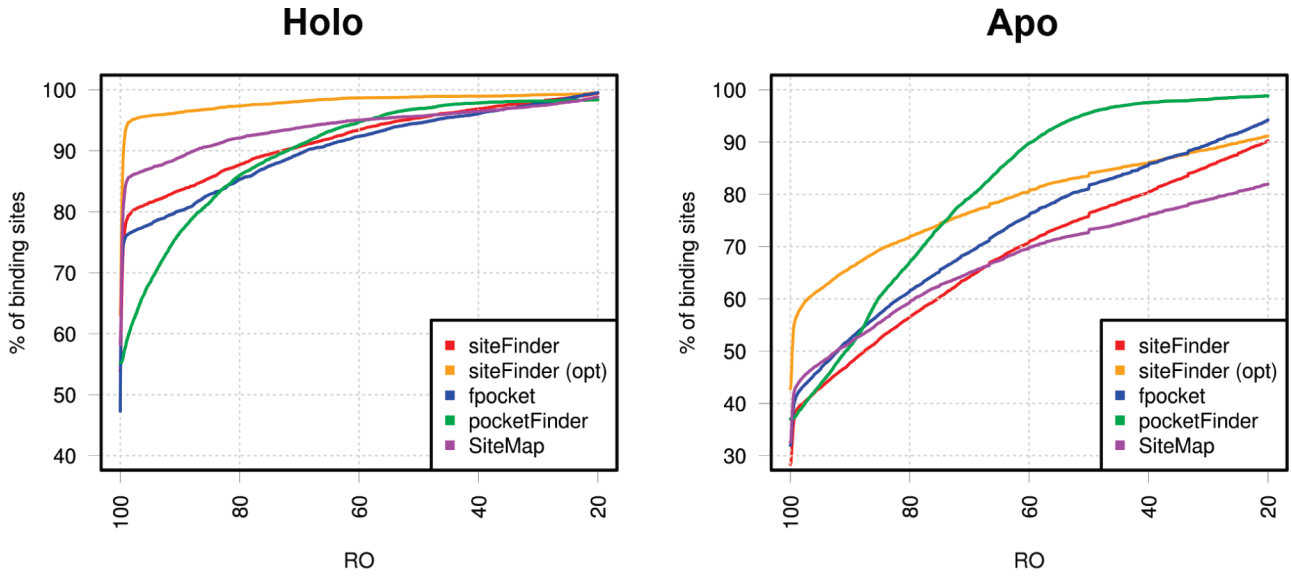

Figure 4. The prediction accuracy measured by the RO between the predicted binding patch A_E, defined as the solvent-accessible surface of the receptor atoms within 3.5 Å from the predicted envelope, and the observed binding patch A_L, defined as the solvent-accessible surface of the receptor atoms within 3.5 Å from the bound ligand. The results of 5416 binding sites from protein−ligand complexes and 9900 binding sites from uncomplexed structures were sorted separately by RO. SiteFinder using default parameters is red; SiteFinder with optimized parameters is orange; fpocket with default parameters is blue; PocketFinder (taken from An et al.)[21] is green; and SiteMap using default parameters is purple.

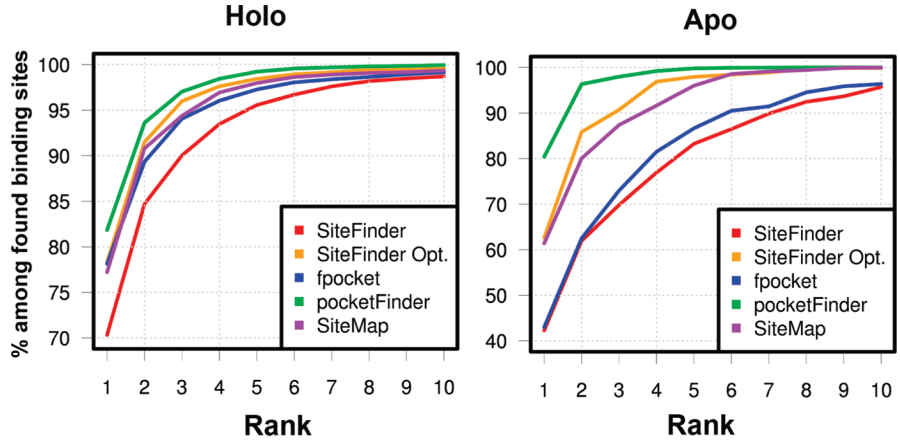

Figure 5. Cumulative percentage of binding sites among found binding sites (RO > 0.5) versus the ranking of those. For all methods, more than 95% of the identified binding sites are found among the first 5 ranks on holo structures.

The average calculation time per protein for SiteFinder was 1.6 s.

SiteMap Preliminary Study. During the preliminary study, the systematic binding site search was performed on 315 proteins out of 370. By default, SiteMap returns only the first five cavities. All binding sites from the 315 proteins were in this set of cavities. A total of 68% of actual binding sites were ranked as first, while 87% of them bind ligands with molecular weights (MW) larger than 250. Thus the SiteMap scoring function used to rank identified cavities performed well for our purpose, as all actual binding sites were retrieved in all cases. Also, the results indicated that no optimization of the search parameters of SiteMap was necessary.

The SiteMap process stopped during the atom-typing step for the remaining 55 proteins, with incomplete coordinates. Missing residues were modeled using Prime, a Schrödinger module that performs homology modeling and side chain and loop prediction. However this process was very time-consuming and did not succeed for all proteins, so it was not applied to the structures used in the full-scale study.

SiteMap Full-Scale Evaluation. As alluded to in the Materials and Methods Section, because SiteMap can handle only complete structures (no missing atoms or residues), structures that cannot be treated by SiteMap in an automated manner (the prepwizard program from Schrödinger was used to prepare the structures) were omitted from the data set. By default, the output of the SiteMap algorithm is limited to the five top-ranked cavities. To enable a relevant comparison with SiteFinder, the SiteMap output set was enlarged to 20 cavities. It should be noted that this modification had an influence on the cavity delimitation, resulting in some cases in the splitting of large cavities.

Figure 4 illustrates that around 85% of all binding sites were found with a relative overlap close to 1. At the threshold of a correctly identified binding site (RO > 0.5) SiteMap gave good predictions for 95% of all binding sites. Considering the ranking performance (Figure 5) of the score implemented in SiteMap, it managed to retrieve around 78% of found binding sites on the very first rank and around 97% within the top five ranks.

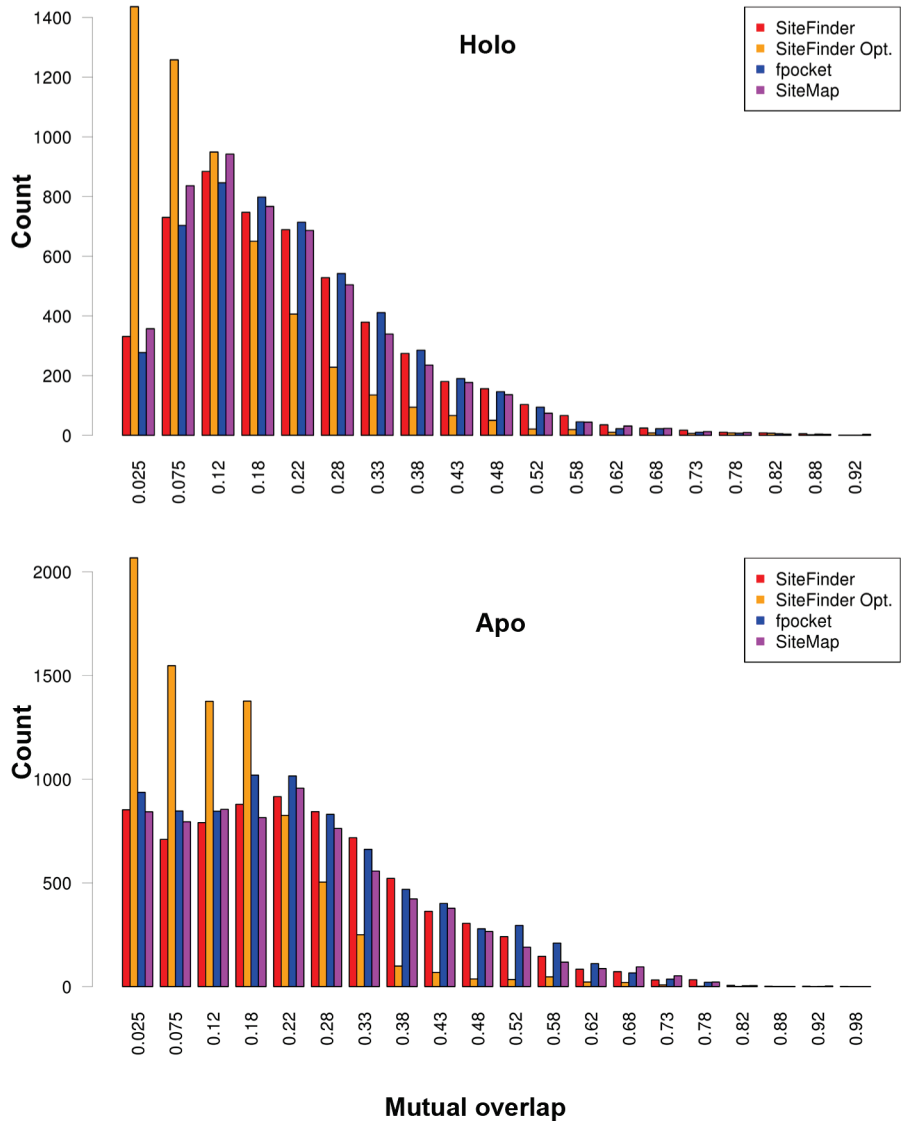

Figure 6. Introduction of mutual overlap, a second measure of prediction accuracy, entitled MO. MO is defined as the ratio between the overlapping ASA between predicted and known pockets and the total predicted pocket ASA. The higher the MO, the more accurately the pocket is predicted. SiteFinder using default parameters is red; SiteFinder using optimized parameters is orange; fpocket using default parameters is blue; and SiteMap using default parameters is purple.

As illustrated in Figures 4 and 5, SiteMap allowed to predict about 73% of all 9900 apo binding sites correctly, while 61% of these were ranked on rank 1 and 96% within the top five ranks.

The preliminary calculations were performed with Maestro 7.5 suite, and the average calculation time was about 13 min per structure, including system preparation. Variations in calculation time were rather large, from 2 min to several hours in a few cases, due to the atom-typing procedure. The large-scale evaluation was performed with Maestro 8.0. No difference in terms of cavity detection performance was observed between both versions. However, the computation time was noticeably improved, down to about 2 min per structure.

Fpocket Full-Scale Evaluation. On the set of 5416 holo structures, fpocket was able to retrieve more than 75% of binding sites with a RO around 1. Similar to the other methods evaluated here, at an RO > 0.5 around 95% of all binding sites could be retrieved. As shown on Figure 5, although being a pure geometry-based method, fpocket has a very similar ranking performance to SiteMap, ranking 78% of all found binding sites on the top rank and 97% within the top 5 ranks.

Considering the set of 9900 apo structures, fpocket managed to correctly identify 82% of all binding sites (Figure 4), ranking 42% of those on rank 1 and 86% among the top 5 ranks (Figure 5). Here again, a drop in ranking performance can be seen compared to the results obtained on holo structures.

Calculation time with fpocket varied between 1 to 3 s per structure.

PocketFinder Evaluation. The results published by An et al. have been taken directly to compare PocketFinder to the other three methods evaluated here. As the initial data set published by An et al. has been slightly modified, these modifications were taken into account in the results presented here.

PocketFinder is able to retrieve around 55% of all holo binding sites with a RO close to 1 and 97% of all binding sites with a RO > 0.5. Although solely volume based, the ranking performance of PocketFinder appears to be generally

better than the other methods, allowing retrieval of 82% of found binding sites on the top rank and 99% on the top five ranks.

Next to the RO for measuring accuracy of pocket prediction, An et al. introduced in their evaluation two other measures extending the assessment of accuracy. The first is the ratio between the binding site and the ligand volumes, and the second is the ratio of the predicted binding patch with respect to the whole protein surface. Neither of these two criteria is used in the present study but rather the MO criterion. As all the results for PocketFinder were directly taken from the publication of An et al., no further calculation of the MO criterion was possible.

On the apo structure data set, PocketFinder identifies about 95% of all binding sites. Also the ranking performance of PocketFinder appears to be satisfying, ranking 80% of all found binding sites on the top rank and nearly 100% within the to 5 ranks.

Comparison between SiteFinder, SiteMap, PocketFinder, and fpocket. Evaluation results for all four algorithms on holo structures are summarized in Figures 4−6. Regarding the capacity of all algorithms to actually find the known binding site, the difference between them is rather small. SiteFinder (default parameters), fpocket, and SiteMap predict around 95% of the known binding sites with a RO above 0.5 (Figure 4). This result is rather interesting given that some of the methods are based on rather different methodologies. A significant improvement in predictive power can be observed for SiteFinder using optimized parameters.

Big differences in performance between SiteFinder (default parameters) and all other algorithms and parameters sets can be seen with respect to ranking of the binding sites. Figure 5 illustrates clearly that the predictiveness of SiteFinder is around 5% lower than the one of the other methods. The results obtained for fpocket, SiteMap, and SiteFinder (optimized parameters) are very similar. PocketFinder shows a slightly better ranking performance, although the ranking is simply based on the pocket volume.

Having observed the big performance increase between the default parameter set and optimized parameter set for SiteFinder as illustrated in Figure 4, one can state nevertheless that the accuracy of the prediction is influenced by changes of these parameters. As shown in Figure 6, the MO for SiteFinder (optimized parameter) is clearly lower than for all other methods. Comparing SiteFinder (default parameters) with fpocket and SiteMap another interesting observation could be made. Although both SiteFinder and fpocket are pure geometry-based approaches, they (especially fpocket) appear to predict binding sites slightly more accurately than SiteMap. This concurs with the smaller RO values obtained for fpocket and SiteFinder around RO = 1, indicating that both geometry algorithms generally produce smaller pockets than SiteMap.

Using all algorithms with the standard parameters, SiteMap clearly outperforms SiteFinder, fpocket, and PocketFinder regarding full coverage of the actual binding sites. Taking a half-covered binding site as correctly identified (as considered here), all methods perform well with a comparable level of predictiveness. Regarding ranking performance, only SiteFinder shows a clearly lower predictive power.

Considering accuracy of prediction, fpocket appears to propose pockets with better MO compared to all other methods.

The evaluation of predictive power of all methods on apo structures allows for the identification of further differences between the four algorithms. First of all, it should be pointed out that PocketFinder results were again taken directly from An et al. As the study on apo structures involves notably a step of structural alignment, it could be a source of variations in pocket definitions using the ligand present in the superimposed holo structure. In the present study, PyMOL's align function was applied on the chains known to hold the apo and holo binding sites. As the structural alignment procedure used by An et al. was not specified in the paper, we simply assumed that the method employed here produced comparable results. However, at least for fpocket, SiteMap, and SiteFinder, the very same protocol was applied, allowing a straightforward comparison between those algorithms.

Bearing these limitations in mind, one can observe that PocketFinder performs better than all other methods regarding accurate prediction of the binding site and the ranking. However, one should bear in mind that the comparison between the other methods and PocketFinder could be skewed. Among the other three methods SiteFinder with optimized parameters performs best, with the caveat that this comes at the cost of reduced accuracy and the prediction of very large binding sites. Among the algorithms with default parameters, fpocket performs best regarding accuracy of binding site prediction, while SiteMap clearly outperforms the other methods regarding ranking of binding sites.

DISCUSSION

A large-scale evaluation of four pocket prediction algorithms, SiteFinder, SiteMap, fpocket, and PocketFinder was performed. All algorithms were able to correctly predict binding sites in almost all proteins for the holo structures. The algorithm with default parameters that allowed the most binding sites with a high RO to be retrieved is SiteMap. By optimizing search parameters, SiteFinder outperforms all other methods at the cost of producing very big binding sites. Although this has no obvious primary sense, such a very comprehensive binding site detection can prove useful in cases where a pocket database is established for comparison of subpockets against other pockets/subpockets. SiteFinder's parameter set allows the construction of a representative collection of cavities containing entire binding sites, that is a cavity database. When exploiting such a database, one must bear in mind that the potential ligands may be smaller than the actual cavity.

Interestingly, SiteFinder using optimized parameters, SiteMap and fpocket show a surprisingly similar performance in ranking binding sites, although all three methods use completely different approaches. Solely SiteFinder (default parameters) appears to exhibit lower ranking performance compared to the latter. Another surprising finding is that both geometry-based methods (using default parameters) tend to produce more accurate binding sites than SiteMap, indicating that SiteMap predicts slightly bigger pockets than SiteFinder and fpocket.

A major performance drop was observed for predictions on apo structures. Assuming that a straightforward compari-

son between PocketFinder and the other algorithms is possible, PocketFinder outperforms all other algorithms. For the other three algorithms, using default parameters, fpocket performs best on accurate binding site prediction, while SiteMap performs best on ranking. Here again, both geometry-based algorithms appear to produce more accurate binding sites (cf., Figure 6). This finding, and the results shown in Figure 4, are at odds with the general idea that geometry-based criteria can only identify well-defined pockets as, for example, mentioned by T. Halgren.[30]

Based on our results it can be postulated that the concave curvature or high degree of burial is a common hallmark of all protein surface patches binding small molecules tightly. Compared to macromolecular interactions, this burial appears to be a necessity to provide sufficient shielding from solvent for a stable interaction to be possible. Furthermore, this will increase the number of contacts that a small molecule can make with the macromolecule and will therefore increase its binding efficiency at this site compared to locations on the macromolecule. Also, one can expect the local water structure to be more ordered within a small, concave, pocket. Release of these molecules into the bulk water upon binding of the ligand provides an entropic gain. Overall, relatively simple geometric rules are sufficient to account for these characteristics, and therefore, a corresponding algorithm can perform well at predicting and ranking binding sites.

Importantly, the herein used data set is not restricted with respect to the characteristics of the ligand molecules. Thus binding of physicochemically different small molecules (sugars, drugs, pro-drugs, etc.) requires generally a concave, solvent-shielded, portion of the protein surface.

Looking at the results published for the Cheng et al. data set[34] regarding the druggability of binding sites, it appears that there is a correlation between size as well as hydrophobicity and druggability (larger binding site, and increased hydrophobicity favoring druggability). Although the algorithm in SiteFinder does not explicitly target druggability, it is apparent that the way the alpha spheres are calculated and used for scoring implies as well that binding sites are ranked with respect to size and hydrophobicity. These characteristics, common to the different algorithms, also correspond to chemical intuition and to the trends very often observed in medicinal chemistry optimization programs, where larger and more hydrophobic molecules tend to have higher affinity. Nevertheless, optimizing molecules solely according to this criterion must be treated with caution because other properties, such as physicochemical and ADMET properties, tend to deteriorate at the same time.

Practical Considerations for Creating a Pocket Database. Both geometry-based algorithms have some inherent advantages over energy-based pocket prediction methods. First, the calculation time is about 90 times faster. Second, both geometry algorithms are robust against structural variations or missing atoms/residues that can occur in PDB files, as no atom-typing step and adding of H-atoms needs to be performed. Care must be taken, however, that the missing atoms or residues do not have an effect on the binding sites that are to be detected, which can be the case with geometry-based methods. In case of a detailed study of a system, where protonation states of all side chains in and around the binding site are known, energy-based pocket predictions can be very useful. However, this type of assignment on a high-throughput level is not realistic. If the task is the creation of a pocket database for the whole PDB or a large in-house databases, then geometry-based methods have a clear advantage.

In terms of userfriendliness, working environment, and informatics skills required, the algorithms cater to different tastes. Regarding the working environment, SiteFinder can be used through the very powerful SVL programming language available in the MOE or within the GUI itself. For SiteMap, the user can interact with the software through the graphical user interface available in Maestro or through the command line. Also, Maestro allows accessing molecular information through a Python-based API. However, for both of these algorithms a certain amount of effort in programming and automation has to be spent to adapt them for the creation of a pocket database, while working with SVL appears to be more straightforward, although it requires some knowledge of the SVL programming language. A very convenient algorithm for creation of a putative pocket database creation is fpocket. In essence it is standalone C code executable, and given that it is command line driven, it makes extraction of pocket scores, ranks, and descriptors very easy via a few command line flags. This information can then be organized using the tools the user prefers and is most comfortable with and not a preimposed working environment that the user has to adapt to.

CONCLUSIONS

In general, the binding site detection algorithms considered in this study exhibit a very good performance. Over 95% of all binding sites are retrieved within the 5 best ranked binding pockets. Considering the trade-off between speed and quality of the results, geometry-based methods like SiteFinder (using optimized parameters) or fpocket appear to be slightly more appropriate for creating a large cavity database for further use by cavity comparison algorithms.

Regarding SiteMap, it would be desirable to improve the treatment of structures with missing atoms or residues, in particular, if it is intended to be used for a systematic study or the preparation of a cavity database. This has been partly accomplished by the prepwizard program provided by Schrödinger. Nevertheless, it adds another intermediate step before pocket prediction using SiteMap.

Given that during the cavity detection process already a number of descriptors for the binding sites are calculated, possible extensions of the binding site algorithms would be the inclusion of a druggability score. Such druggability scores can be based on relatively simple descriptors,[34] as published in recent papers on SiteMap[30] and on fpocket,[32] and provide additional valuable information for the user. In the same vein, the calculated characteristics of a given binding site could be used in order to choose those molecules that should be screened first against the site of interest. Here one could imagine translating the binding site characteristics into a query for shape-based and/or pharmacophore-based screening, in order to identify molecules that are complementary to the binding site characteristics.

Combining the binding site detection algorithms with a binding site comparison tool would allow for the prediction of ligands likely to bind to a new binding site if the ligands for a similar binding site are known.

With the advent of systems biology and pathway- or network-based drug discovery,[35] binding site detection and characterization may gain additional importance, in particular if the idea to interfere at the same time with several targets—representing key players in a pathway or network—gains traction. In this case it might be necessary to find and compare binding sites on several different proteins in a given network with a view to identifying ligands that can bind to these proteins simultaneously, albeit with lower affinity. Binding site detection algorithms that are fast and efficient could prove invaluable for such an undertaking.

Note Added after ASAP Publication. This paper was published ASAP on September 9, 2010 with minor text errors and a corrected version was published on November 12, 2010. The version published ASAP on November 12, 2010 had an error in the optimized and default minrad values. The corrected version was published ASAP on November 17, 2010.

Supporting Information Available: Two files are provided, and both are in csv format and contain the PDB codes of the holo structures (LP_SETFinal.csv) used in this evaluation with corresponding ligand accessions. Another file named UP_SETFinal.csv contains the apo data set used in this study. This information is available free of charge via the Internet at http://pubs.acs.org/.

REFERENCES AND NOTES

(1) Jambon, M.; Imberty, A.; Deleage, G.; Geourjon, C. A new bioinformatic approach to detect common 3D sites in protein structures. *Proteins* **2003**, *2*, 137–145.
(2) 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**, *2*, 279–294.
(3) Alexandrov, N. N. SARFing the PDB. *Protein Eng.* **1996**, *9*, 727–732.
(4) Gibrat, J. F.; Madej, T.; Bryant, S. H. Surprising similarities in structure comparison. *Curr. Opin. Struct. Biol.* **1996**, *3*, 377–385.
(5) Madej, T.; Gibrat, J. F.; Bryant, S. H. Threading a database of protein cores. *Proteins* **1995**, *3*, 356–369.
(6) Holm, L.; Sander, C. Dictionary of recurrent domains in protein structures. *Proteins* **1998**, *1*, 88–96.
(7) Holm, L.; Sander, C. Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.* **1993**, *1*, 123–138.
(8) Ye, Y.; Godzik, A. Flexible structure alignment by chaining aligned fragment pairs allowing twists. *Bioinformatics* **2003**, ii246–ii255.
(9) Ye, Y.; Godzik, A. FATCAT: a web server for flexible structure comparison and structure similarity searching. *Nucleic Acids Res.* **2004**, W582–W585.
(10) Brady, G. P., Jr.; Stouten, P. F. Fast prediction and visualization of protein binding pockets with PASS. *J. Comput.-Aided Mol. Des.* **2000**, *4*, 383–401, 0920−654; 0920−654.
(11) Laskowski, R. A. SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. *J. Mol. Graph.* **1995**, *5*, 323–30, 307−8.
(12) Hendlich, M.; Rippmann, F.; Barnickel, G. LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins. *J. Mol. Graph. Modell.* **1997**, *6*, 359–63, 389.
(13) Huang, B.; Schroeder, M. LIGSITEcsc: predicting ligand binding sites using the Connolly surface and degree of conservation. *BMC Struct. Biol.* **2006**, 19.
(14) Levitt, D. G.; Banaszak, L. J. POCKET: a computer graphics method for identifying and displaying protein cavities and their surrounding amino acids. *J. Mol. Graph.* **1992**, *4*, 229–234.
(15) Peters, K. P.; Fauck, J.; Frommel, C. The automatic search for ligand binding sites in proteins of known three-dimensional structure using only geometric criteria. *J. Mol. Biol.* **1996**, *1*, 201–213.
(16) Binkowski, T. A.; Adamian, L.; Liang, J. Inferring functional relationships of proteins from local sequence and spatial surface patterns. *J. Mol. Biol.* **2003**, *2*, 505–526.
(17) Edelsbrunner, H.; Facello, M.; Fu, P.; Liang, J. Measuring proteins and voids in proteins. Proceedings of 28th Hawaii International Conference on System Science, Hawaii, January 4−7, 1995; IEEE: Piscataway, NJ,1995; pp 256264.
(18) Liang, J.; Edelsbrunner, H.; Woodward, C. Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. *Protein Sci.* **1998**, *9*, 1884–1897.
(19) Le Guilloux, V.; Schmidtke, P.; Tufferry, P. Fpocket: An open source platform for ligand pocket detection. *BMC Bioinformatics* **2009**, *1*, 168–178.
(20) Labute, P.; Santavy, M. Locating Binding Sites in Protein Structures. Chemical Computing Group, Inc.: Montreal, Quebec, Canada, 2001; http://www.chemcomp.com/journal/sitefind.htm. Accessed on June 30, 2010.
(21) An, J.; Totrov, M.; Abagyan, R. Pocketome via comprehensive identification and classification of ligand binding envelopes. *Mol. Cell Proteomics* **2005**, *6*, 752–761.
(22) Cummins, P. L.; Titmuss, S. J.; Jayatilaka, D.; Bliznyuk, A. A.; Rendell, A. P.; Gready, J. E. Comparison of semiempirical and ab initio QM decomposition analyses for the interaction energy between molecules. *Chem. Phys. Lett.* **2002**, *3−4*, 245–251.
(23) Kortvelyesi, T.; Dennis, S.; Silberstein, M.; Brown, L., III; Vajda, S. Algorithms for computational solvent mapping of proteins. *Proteins* **2003**, *3*, 340–351.
(24) Silberstein, M.; Dennis, S.; Brown, L.; Kortvelyesi, T.; Clodfelter, K.; Vajda, S. Identification of Substrate Binding Sites in Enzymes by Computational Solvent Mapping. *J. Mol. Biol.* **2003**, *5*, 1095–1113.
(25) Nettles, J. H.; Jenkins, J. L.; Williams, C.; Clark, A. M.; Bender, A.; Deng, Z.; Davies, J. W.; Glick, M. Flexible 3D pharmacophores as descriptors of dynamic biological space. *J. Mol. Graph. Modell.* **2007**, *3*, 622–633.
(26) Welch, W.; Ruppert, J.; Jain, A. N. Hammerhead: fast, fully automated docking of flexible ligands to protein binding sites. *Chem. Biol.* **1996**, *6*, 449–462.
(27) Verdonk, M. L.; Cole, J. C.; Taylor, R. SuperStar: A Knowledge-based Approach for Identifying Interaction Sites in Proteins. *J. Mol. Biol.* **1999**, *4*, 1093–1108.
(28) Chemical Computing, G. I. Molecular Operating Environment.
(29) Halgren, T. New method for fast and accurate binding-site identification and analysis. *Chem. Biol. Drug Des.* **2007**, *2*, 146–148.
(30) Halgren, T. A. Identifying and Characterizing Binding Sites and Assessing Druggability. *J. Chem. Inf. Model.* **2009**, 1549–960.
(31) *Maestro*, version 8.0; Schrödinger L.L.C.: New York, 2009.
(32) Schmidtke, P.; Barril, X. Understanding and predicting druggability. A high-throughput method for detection of drug binding sites. *J. Med. Chem.* **2010**, *53*, 5858−5867.
(33) *R*, version 2.10.1; R Development Core Team: Vienna, Austria, 2009.
(34) Cheng, A. C.; Coleman, R. G.; Smyth, K. T.; Cao, Q.; Soulard, P.; Caffrey, D. R.; Salzberg, A. C.; Huang, E. S. Structure-based maximal affinity model predicts small-molecule druggability. *Nat. Biotechnol.* **2007**, *1*, 71–75.
(35) Davis, J. C.; Furstenthal, L.; Desai, A. A.; Norris, T.; Sutaria, S.; Fleming, E.; Ma, P. The microeconomics of personalized medicine: today's challenge and tomorrow's promise. *Nat. Rev. Drug Discovery* **2009**, *4*, 279–286.

CI1000289