Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Molecular Graphics and Modelling 22 (2003) 31–40
Journal of Molecular Graphics and Modelling
www.elsevier.com/locate/JMGM

Analysis and optimization of structure-based virtual screening protocols 2. Examination of docked ligand orientation sampling methodology: mapping a pharmacophore for success

Andrew C. Good [a,*], Daniel L. Cheney [c], Doree F. Sitkoff [c], John S. Tokarski [b], Terry R. Stouch [c], Donna A. Bassolino [b], Stanley R. Krystek [c], Yi Li [a], Jonathan S. Mason [b,1], Timothy D.J. Perkins [d,2]

[a] *Bristol-Myers Squibb, P.O. Box 5100, Wallingford, CT 06492, USA*
[b] *Bristol-Myers Squibb, P.O. Box 4000, Princeton, NJ 08543, USA*
[c] *Bristol-Myers Squibb, P.O. Box 5400, Hopewell, NJ 08543, USA*
[d] *Protherics PLC, The West Heath Business and Technical Park, Runcorn, Cheshire W47 4QF, UK*

Accepted 14 February 2003

Abstract

An important element of any structure-based virtual screening (SVS) technique is the method used to orient the ligands in the target active site. This has been a somewhat overlooked issue in recent SVS validation studies, with the assumption being made that the performance of an algorithm for a given set of orientation sampling settings will be representative for the general behavior of said technique. Here, we analyze five different SVS targets using a variety of sampling paradigms within the DOCK, GOLD and PROMETHEUS programs over a data set of ∼10,000 noise compounds, combined with data sets containing multiple active compounds. These sets have been broken down by chemotype, with chemotype hit rate used to provide a measure of enrichment with a potentially improved relevance to real world SVS experiments. The variability in enrichment results produced by different sampling paradigms is illustrated, as is the utility of using pharmacophores to constrain sampling to regions that reflect known structural biology. The difference in results when comparing chemotype with compound hit rates is also highlighted.

Keywords: Ligand sampling; Structure-based virtual screening; Chemotype enrichment tests; Pharmacophore constraints

1. Introduction

A variety of techniques are used to sample potential ligand binding modes in SVS software [1–3]. For techniques such as DOCK [1] where ligand conformations are defined independently of the docking process, binding mode sampling generally occurs as a process independent of scoring and conformation generation. This procedure requires that a series of site points be predefined within the active site. These points are then used as geometric constraints to map ligand atoms into the site. The technique has a long history within SVS [4] and continues to be developed [5]. From the perspective of SVS validation and comparison, however, binding mode sampling has been a somewhat neglected area of research. Indeed, there has been some suggestion that sampling techniques form a significantly lessor role compared to scoring function in terms of docking importance [6]. These studies have generally been based on DOCK results using identical sampling with differing scoring functions, however, with little attempt made to understand the effects of sampling experimentation.

DOCK contains a host of input parameters that can potentially effect ligand sampling. In addition the application of pharmacophore (chemical matching and critical region constraints [7]) can be utilized to impact the sampling paradigm. The differing approaches taken to sampling issues are highlighted by an analysis of two SVS validation papers that include DOCK searches in their experiments [8,9]. Both studies use DOCK4.01, but while one uses geometry searching [8] with core fragment conformational generation, the other [9] applies pharmacophore and critical

* Corresponding author. Present address: Pfizer Central Research, Ramsgate Road, Sandwich, Kent, CT13 9NJ, UK.
E-mail address: andrew.good@bms.com (A.C. Good).
[1] Present address: Pfizer Central Research, Ramsgate Road, Sandwich, Kent. CT13 9NJ, UK.
[2] Present address: AstraZeneca, Charter Way, Silk Road Business Park, Macclesfield, Cheshire, SK10 2NA UK.

1093-3263/03/$ – see front matter © 2003 Elsevier Science Inc. All rights reserved.
doi:10.1016/S1093-3263(03)00124-4

region constraints to the site points in conjunction with pre-generated conformations. If variations in sampling techniques produce a difference in enrichment results, it would be difficult to directly compare the results of these experiments. In this article we study the effect that such sampling alternatives produce, using five different SVS test cases in conjunction with a variety of input parameter and site point combinations. Particular attention has been paid to the construction of pharmacophore constraints and the results they generate. In addition results obtained using DOCK are compared with other SVS paradigms namely PROMETHEUS (a development of the PRO_LEADS SVS search tool [2]) and GOLD [3], to gain further understanding with regard to relative performance and behavior.

2. Methods

2.1. General points

All studies were undertaken using a slightly modified version of DOCK 4.01. Primary code alterations involve changes to the conformation generation algorithm (flex.c in DOCK source) to prevent the program spending too much time generating conformers in heavily branched structures. In addition modifications have been made to the flex.defn file to refine the torsion rules applied during searching.

Table 1 shows the parameters used with the DOCK runs undertaken. Parameters highlighted in bold are those most likely to effect ligand sampling results. The presence of multiple values indicates a parameter that was varied during these studies.

To test the effect of varying DOCK sampling paradigms, 5 SVS target from projects studied within Bristol-Myers Squibb were chosen. Active compounds discovered in-house were combined with competitor leads to construct an active data set for each target.

Much consideration was given to the nature of the targets and active data sets selected for study. In a perfect world these would be taken from literature data to ensure easy reproduction of results by third parties. Unfortunately for studies, such as these such publicly available data sets of diverse active compounds are generally lacking. As a consequence it was felt that this, together with our greater understanding of projects studied internally, would make such targets preferable study material.

2.2. Site preparation

The treatment of protein active sites was dependent on the nature of the target. For highly understood targets (serine protease 1) careful minimization and water placement had been undertaken within the project. For those less well defined sites (e.g. the kinase) the protein was left basically unchanged, save for the manual repositioning of hydroxyl hydrogens. Kollman and co-workers united atom charges [10] were assigned for electrostatic potential calculations.

Table 1
Input parameters used in DOCK

nodes_minimum	**3/4**	gridded_score	yes
nodes_maximum	**3/4**	grid_version	4
distance_tolerance	**0.5/1.0/1.5**	contact_cutoff_distance	4.0
distance_minimum	**1.5/3.0**	contact_clash_overlap	0.75
heavy_atoms_minimum	10	contact_maximum	−10.0
heavy_atoms_maximum	50	energy_cutoff_distance	10
bump_filter	**yes**	distance_dielectric	yes
bump_maximum	**4**	dielectric_factor	4
contact_clash_penalty	25	attractive_exponent	6
conformation_cutoff_factor	**5**	repulsive_exponent	12
flexible_bond_maximum	**15**	atom_model	u
clash_overlap	**0.7**	vdw_scale	1.0
minimize_ligand	**yes**	electrostatic_scale	0.2/1.0
flexible_ligand	**yes**	energy_maximum	−5.0
orient_ligand	yes	contact_minimize	yes
anchor_search	no	energy_minimize	yes
torsion_drive	yes	contact_convergence	0.1
match_receptor_sites	yes	initial_translation	1
automated_matching	no	initial_rotation	0.1
uniform_sampling	no	initial_torsion	10
check_degeneracy	no	maximum_iterations	50
reflect_ligand	no	energy_convergence	0.1
maximum_orientations	**5000**	maximum_cycles	2
random_search	no	cycle_convergence	1.0
intermolecular_score	yes	contact_termination	1
		energy_termination	1

For more information on the nature of each parameter see the online DOCK manual page. URL: http://www.cmpharm.ucsf.edu/kuntz/dock4/html/Manual.19.html#pgfId=8491.

Table 2
Sample definitions for acidic atoms taken from our chemical definitions (chem.defn) file

```
name acid

# optional tyrosine
definition O.3 (H)

# deprotonated carboxyl
definition O.co2 (C)

# tetrazole
definition N.pl3 (H) (N.2 (N.2 (N.2 (C.2))))
definition N.pl3 (H) (N.2 (N.2 (C.2 (N.2))))
definition N.2 (N.2 (N.2 (C.2 (N.pl3 (H)))))
definition N.2 (N.2 (C.2 (N.pl3 (H) (N.2))))
definition N.2 (C.2 (N.2 (N.pl3 (H) (N.2))))
definition N.2 (N.2 (C.2 (N.2 (N.pl3 (H)))))
definition N.2 (N.pl3 (H) (N.2 (N.2 (C.2))))

# acyl sulphonamide
definition N.am (S (2 O.2)) (C.2 (O.2))
definition O.2 (C.2 (N.am (H) (S (2 O.2))))
definition O.2 (S (O.2) (N.am (H) (C.2 (O.2))))
```

The pharmacophoric types we currently use include: donor, acceptor, hydrophobe, aromatic, aromatic_hydrophobic, acidic. Basic, donor_and_acceptor, and special (most commonly metal chelator). For more information see the online DOCK manual. URL: http://www.cmpharm.ucsf.edu/kuntz/dock4/html/Manual.47.html#pgfId=20180.

A crucial part of target preparation in SVS calculation with DOCK is protein site point definition. By default DOCK provides the SPHGEN program as the method for creating such points. This technique essentially places site points across the width of the active site, with the number of points distributed in a given region being roughly proportional to the surface curvature in that portion of the active site. As a consequence points tend to concentrate in the major pockets of the active site, but beyond that there is little that is predictable about the procedure. We are particularly interested in being able to exploit known important interactions (critical salt bridges, hydrogen bonds, etc.). Recent versions of DOCK permit the creation of reasonably complex pharmacophore definitions to exploit such interactions. Table 2 shows a sample of such definitions. In these studies, while we have included SPHGEN-based DOCK runs for reference purposes, our preference is to use alternative procedures for site point generation that produce site points consistent with potential pharmacophore definitions. The first is GRIDDOCK which converts GRID [11] probe maps (COO– (acid), O (acceptor), N2 (donor), N3+ (base), and C3 (hydrophobe) probes used) into mol2 format files, based either on energy minima or allowed energy ranges. To create hydrophobic points all methyl probe points within a user-defined radius of a hydrophilic probe point are removed. By definition interactions near a hydrophilic region are not hydrophobic, and the methyl probe does not differentiate this automatically. An alternative option would be to apply the dry probe, but this has yet to be investigated. The mol2 atom types reflect the chemical (electronic) type of the site point, while the charge field is used to store the associated GRID energy. The second is MAKESITE, which creates site points from CONNOLLY [12] surfaces by back calculating their position a prescribed distance out from each given surface point normal. These points are chemically coded according to the atoms from which the surface point was derived. Again, to permit easy editing, the resulting points are placed in a mol2 file with atom types corresponding to site point chemistry. These can be combined with known ligand points and placed into the active site using the graphical modeling package of choice. At this point comes perhaps the most important and neglected part of site point creation, that of manual analysis and editing. Generally only extensive modification will create a site point set that either reflects the known structural biology of a given target class, or the prevalent binding model hypothesis in the researchers mind. Editing commonly involves extensive site point deletion, together with the assignment of points to critical regions that reflect area of the active site that must be occupied. Modification of site point chemical type is less common, since these are generally already assigned by MAKEPOINT and GRIDDOCK. Once completed the script MOL2SPH converts the resulting mol2 site point file into DOCK SPHGEN format. The atom name field is used to store the critical region assignment, with the atom type being used to assign chemical match type. Table 3 shows the full list of targets used in this study, a basic description of the substructure selection criteria used in chemotype definition, and the primary elements used in site point creation. Fig. 1 highlights the site point definitions used in the kinase target studies. By breaking the site down into

Table 3
Targets used in DOCK calculation, together with primary chemotype and site point assignment definitions

Target	Active chemotype definitions	Defined critical regions (associated pharmacophore type(s)) [critical region total]
Serine protease 1	P1 substituent/P1–P4 linker substituent	S1 sub site (base or hydrophobe) S4 sub site (hydrophobe) [2]
Serine protease 2	P1 substituent/P1–P4 linker substituent	S1 sub site (base) S4 sub site (hydrophobe) [2]
Fatty acid binding protein 1	Core linking acid moiety to remaining substituents	Acid binding sub site (acid) Rear hydrophobic pocket (hydrophobe) [2]
Fatty acid binding protein 2	Core linking acid moiety to remaining substituents	Acid binding sub site (acid) [1]
Kinase	Moiety mimicking adenine/main core of molecules	Adenine hydrogen bonding regions (donor/acceptor) rear hydrophobic pocket (hydrophobe) [3]

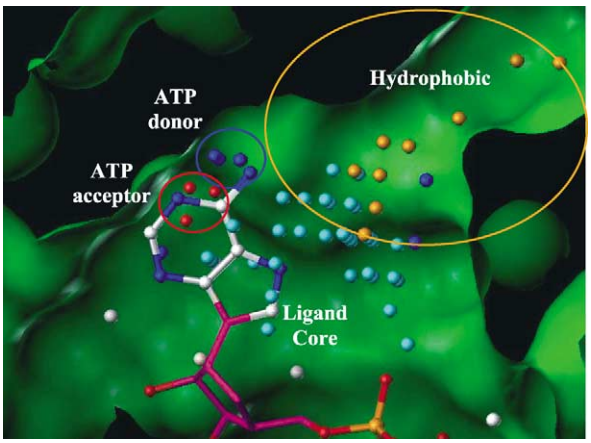


Fig. 1. Site point breakdown for the kinase receptor study. Three primary critical regions were defined: (i) adenine acceptor zone; (ii) adenine donor zone; (iii) ATP binding site rear hydrophobic pocket. Note that the DOCK algorithm always defines those site points not specifically assigned to a critical zone to a region all their own. This means that for n specifically defined critical regions there are $n + 1$ regions that may require specific mapping to by a ligand atom set before docking will occur. The site points in the ligand core portion of the active site constitute such a region.

regions like this, it is possible to simultaneously undertake multiple pharmacophore searches within the confines of the active site. In the case of the kinase ATP site, the hydrogen bonds formed by adenine are known to be important in the majority of known ATP binding inhibitors. By assigning the points where such hydrogen bonds would form as being critical, the formation of such interactions can be enforced when mapping ligand atoms to the site points. It is also possible to assign site points over a large area to a single region (the hydrophobic pocket in the kinase for example), allowing for fine control over the nature of ligand docking while still permitting the exploration of multiple binding modes.

2.3. Data set preparation

Leads retrieved via SVS searches typically undergo a number of 2D similarity and substructure searches in the quest for active analogues. Molecules that cluster together as a result of such analyses can thus be considered to be of a single lead type, since the discovery of one chemotype member could reasonably be expected to result in the unearthing of the remainder. It was therefore decided that each data set would be broken down by lead chemotype based on an examination of their common substructures. While such categorizations inevitably contain an element of subjectivity, enrichments based on chemotype space nevertheless provide a potentially more realistic measure of true SVS search success. This is particularly true in active data sets containing large numbers of closely related analogues (of which there are many).

For each search, the active compound collection was combined with a noise data set comprising 9776 molecules (from 10,000 compounds passed through CONCORD [13] for 3D conversion). All structures were processed to strip/add hydrogens so as to ionize the primary acidic and basic centers. The point charges assigned within CONCORD were left unchanged. These compounds were selected from our in-house inventory in a random fashion, apart from the requirement that the compound flexibility distribution of the set mapped to that of the parent database. The result was a set of compounds that could be analyzed in a reasonable amount of time, but would still provide some reflection of the "real world" searches we are currently undertaking. Active data set compound starting structures were generated in an identical fashion.

2.4. Search variants

The protein site point definitions have been combined with the dock input parameter variations shown in Table 1 to create a host of different DOCK runs. For all targets SPHGEN-based unconstrained, critical region (mapping to important regions of receptor without reference to explicit chemical preferences in said binding pocket) and critical region with chemical matching constraints searches have been undertaken. In addition a variety of sampling conditions have been analyzed (see Table 1) for selected runs. To further benchmark performance the Kinase and fatty acid binding (FAB) protein 2 targets were also searched using PROMETHEUS and GOLD. PROMETHEUS was run in default mode for both targets. In addition for FAB protein 2 an acid critical region was added using a prototypic technique based on smiles definitions to mimic DOCK chemical matching types. This was also attempted for the kinase searches using the three critical region model, however, the resulting level of complexity was found to cause sampling problems in the software version used. GOLD was run in default virtual screen mode. The software is not currently set up to exploit chemical matching or critical regions in virtual screen mode. As a consequence the more generic (i.e. no salt bridges) chemical definitions and critical regions of the kinase were ignored. For the FAB protein search, pharmacophore constraints were simulated by removing all non-acids from the search list. Further details regarding the search criteria applied are given in Section 3.

3. Results

The primary studies undertaken with regard to DOCK sampling variations are highlighted in Fig. 2. Each graph shows the hit rates for a variety of sampling conditions when a specific chemotype enrichment criterion was achieved. For the searches with more generic pharmacophore constraints (kinase and serine protease 1), this was set to 50% of the chemotypes in the data set. For the more constrained searches, i.e. those containing salt bridge critical regions (serine protease 2 and the FAB proteins) the hit rate was set to all known chemotypes or the total number found across

the full search set (when one or more chemotypes were not found). If these primary criteria were not met when the top 500 compounds had been abstracted from the hit list, the results for the top 500 have been given (this was deemed to represent the furthest extent a hit list would likely be mined).

Default conditions for the searches are shown in Table 1. For the SPHGEN-based runs the search defaults were 4 nodes (nodes: number of site points/ligand atoms used in distance matching), with a 0.5 Å distance tolerance (maximum permitted difference for a match in site point/ligand

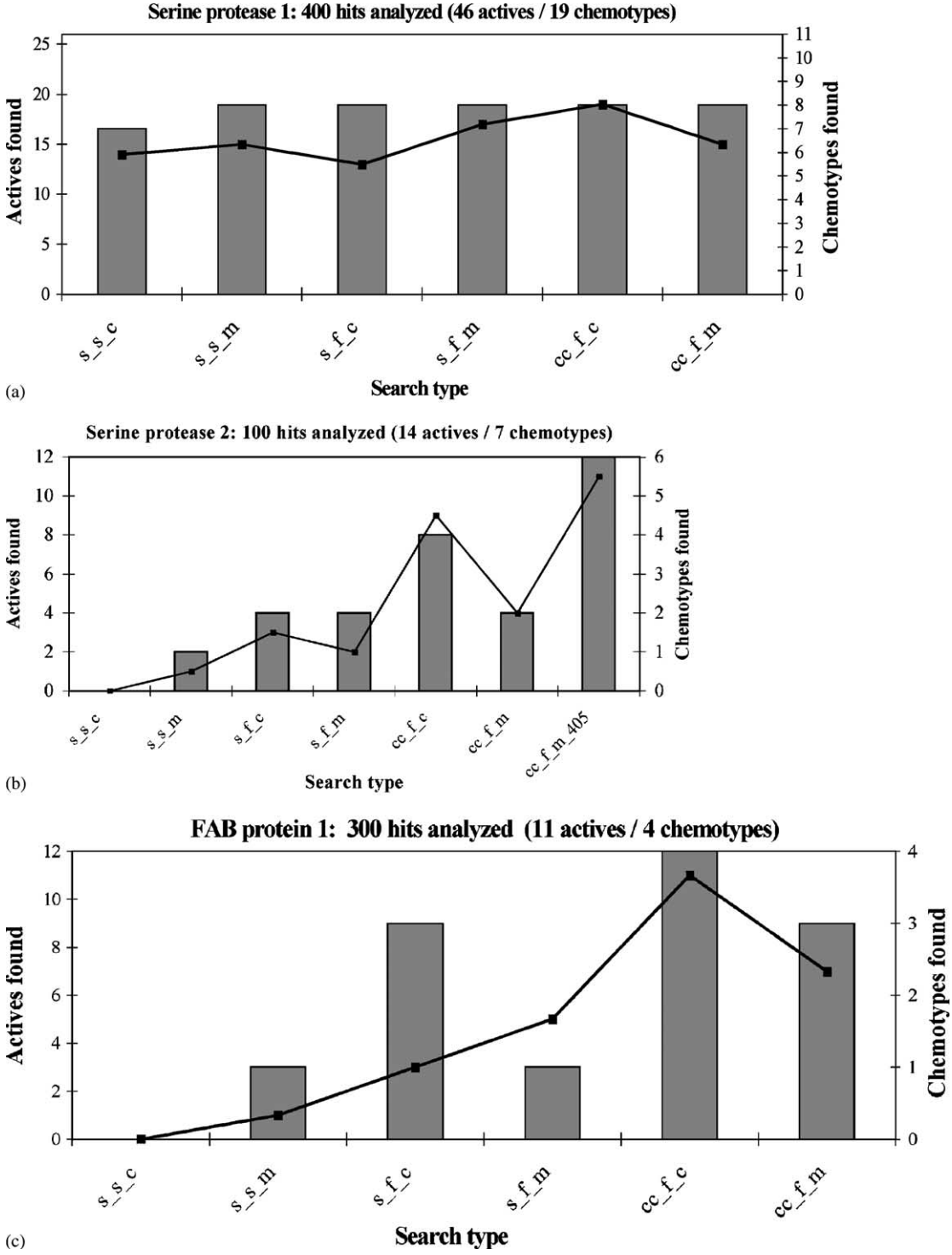
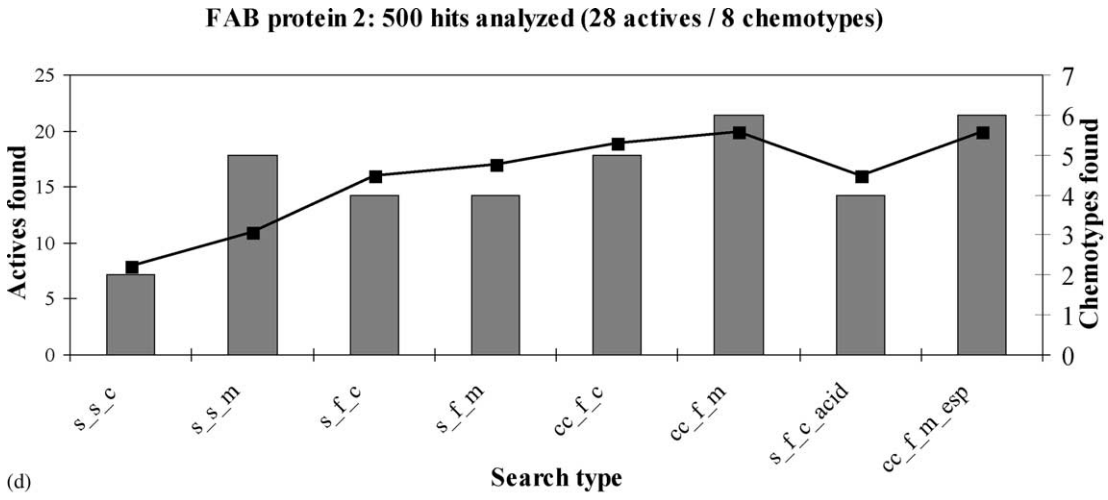


Fig. 2. Comparative hit rate results for the DOCK sampling tests run across the five target proteins. The total number of compounds and chemotypes in each active data set are provided for each target, together with the top ranking hits analyzed to arrive at the results shown. (■) Active compounds (■); active chemotypes.



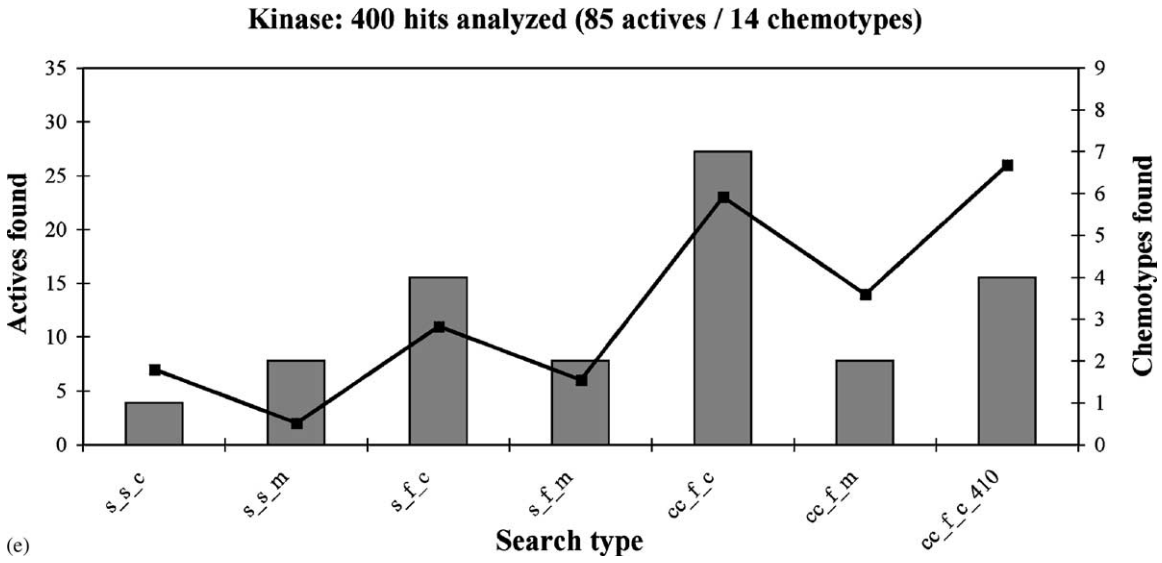


Fig. 2. (*Continued*).

atom pair distances) and a minimum distance (minimum permitted size for site point/ligand atom pair distance match) of 3.0 Å. For the constrained runs 3 nodes and a 0.5 Å distance tolerance were used, except for the kinase search, where the larger number of critical regions required 4 nodes. In addition, since a higher numbers of constraints were being applied which included one short distance constraint (donor–acceptor (see Fig. 1)), a larger distance tolerance (1.5 Å) and smaller minimum distance (1.5 Å) were also used.

SVS docking modes are often of limited accuracy due to the well known issues of rigid receptor approximation, limited ligand conformer accuracy and inadequate binding mode sampling. Taken together with the approximate nature of the partial charges typically used in such searches, the accuracy level of the resultant electrostatic potential interactions are such that their utility is primarily limited to the detection of salt bridge interactions. This being the case, the electrostatic weighting applied to the searches was set to 0.0 unless noted otherwise.

The search keys presented in Fig. 2 are of type a_b_c(_d) e.g. *cc_f_m_410*.

(a) Macro site point definition used, s: generic SPHGEN, cc: site points designed with chemical matching and critical region constraints.
(b) Conformational flexibility paradigm employed s: single conformer only, f: conformational flexibility enabled.
(c) Scoring type used m: force field score (electrostatics ignored), c: contact score.
(d) contains a number of possible parameter variations relating primarily to the constraints applied. These include 405 and 410, which refer to 4 node searches with 0.5 and 1.0 Å distance tolerances, respectively. In addition *acid* denotes the removal of all non acids from search lists, and *esp* highlights calculations where the electrostatic term weighting was set to 1.0.

Averaged enrichment results achieved across all five targets are summarized in Fig. 3.

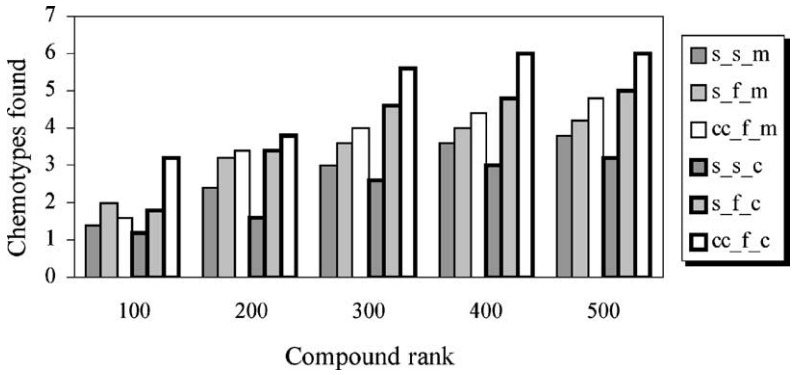


Fig. 3. Averaged chemotype enrichment results for the DOCK searches across the top 500 scoring compounds for all five search sets.

Comparisons between SVS techniques are shown in Fig. 4. For these searches the DOCK results were take from flexible contact searches (s_f_c and cc_f_c). The GOLD and PROMETHEUS searches are described in the methodology section.

4. Discussion

The DOCK results for each target are shown in Fig. 2. For serine protease 1 (Fig. 2a), the performance is pretty uniform across the different search criteria. This is somewhat expected given the generic nature of the pharmacophore constraints employed (primarily hydrophobic (Table 3)). For serine protease 2 (Fig. 2b) inclusion of pharmacophore constraints clearly leads to a major improvement performance (cc_f_c and cc_f_m_405 searches). Again this is not unexpected since constraints are now formulated around docking a base in the S1 sub-site. The top performing search variant shown is cc_f_m_405. This was included since it highlights the complex interplay between sampling and scoring. Running a 4 node search with 0.5 Å distance tolerance significantly reduces amount of sampling undertaken relative to the default 3 node search. In this case the reduced sampling significantly reduces the number of noise molecules that are able to dock successfully. The result is a perceived increase in the enrichment rate of significant proportions. When only 3 nodes are used in conjunction with the same (force field) scoring function, the increased sampling produces somewhat inferior results as more noise molecules make it into the hit list. For the FAB proteins (Fig. 2c and d) once more the inclusion of constraints produces a significant improvement in performance (compare s_f_c with cc_f_m and s_f_c with cc_f_c), with the contact score again outperforming those of the force field. Given the requirement that an acid be present for inclusion in screening, it might reasonably be asked whether simply removing non-acids from the deck and searching would yield similar results. Results for such a search using FAB protein 2 (s_f_c_acid) show an inferior performance relative to the explicitly constrained search (cc_f_c), illustrating the additional descriptive power imparted by 3D pharmacophore constraints. In addition the inclusion of an electrostatic potential term (cc_f_m_esp) illustrates how simple partial charge models often provide little additional information relative to pharmacophore constraints. This is shown by the fact that while inclusion of pharmacophore constraints (s_f_m to cc_f_m) allows the selection of an additional 2 chemotypes, inclusion of electrostatic potential provides no further improvement in performance. For the final kinase search (Fig. 2e), the addition of constraints once more produces a significant improvement in results (cc_f_c). An additional search has been included in this graph (cc_f_c_410), again to highlight the potential vagaries of sampling. For this search the distance tolerance was reduced from 1.5 to 1.0. The consequent decrease in sampling reduces both the number of noise and active molecules present in the hit lists. The result is an increase in the number of active molecules in the top of the list, but a decrease in the number of chemotypes that are able to pass the more stringent distance test. This highlights the potential issues that exist when performance is based on simple active molecule count. The problem is most noticeable for the kinase search, since it has the largest total active molecule to chemotype ratio (85/14, i.e. a higher number of analogues). The issue can be seen in other searches, however, the most prominent example being found in FAB protein 1. Here the s_f_c search finds 3 hits each from a different chemotype, while the s_f_m search locates five hits from a single chemotype.

Fig. 3 highlights the average chemotype enrichment across the top 500 compounds retrieved from each search. Overall it can clearly be seen that the addition of constraints and (as one would expect) conformational flexibility, produce an overall improvement in chemotype enrichment. Further, when both flexibility and constraints are applied, contact score clearly performs better than the force field score for the DOCK searches undertaken. This is likely due in part to the users ability to control the clash penalty in the contact score (the contact_clash_penalty term), which mitigates issues of binding model inaccuracies during docking. The 6–12 potential function used in the force field can on occasion be too sensitive given the sampling approximations inherent in the search procedures. Using a softer

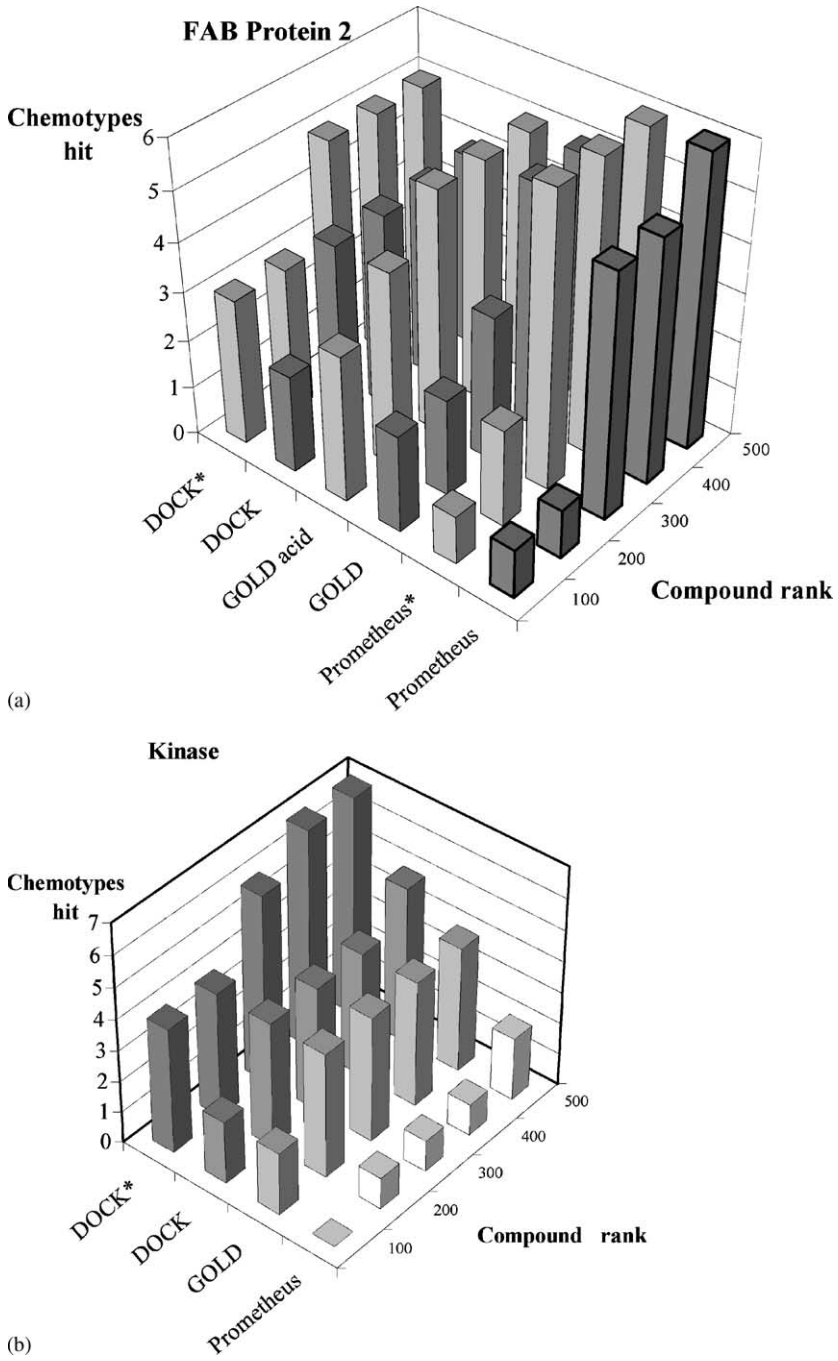


Fig. 4. Chemotype enrichment rates using a variety of SVS algorithms and constraint settings for the (a) FAB protein 2 and (b) kinase targets. The (∗) mark denotes a search including some form of pharmacophoric constraint.

potential (e.g. 4–8 as applied in GOLD) or setting a maximum repulsion cutoff might improve the results for energy, although this has yet to be tested.

Comparisons between SVS techniques over the kinase and FAB 2 targets are shown in Fig. 4. These targets provide interesting counterpoints from the perspective of active site precision. The FAB protein is a closed and relatively rigid active site The kinase on the other hand, is a flexible active site in an inactivated form, creating the potential for extra mobility in the binding domains and possible inaccuracies in the side chain positions, particularly around the phosphate binding pocket. For FAB protein 2 all the search procedures perform well, providing significant enrichments at the top of the hit lists. It is noteworthy, however, that in all cases addition of constraints, whether explicit (DOCK and PROMETHEUS) or implicit (GOLD), provides a measurable improvement in performance. The kinase results are somewhat different. PROMETHEUS, which employs the most complex

scoring function, performs poorly in this search. GOLD performs better, but now lags behind unconstrained DOCK searches, that use the simplest scoring function. While it is difficult to generalize based on two searches, these results suggest that while more exacting scoring functions perform nicely in well-defined active sites, those of a less certain nature benefit from a more fuzzy approach to ligand ranking.

These results highlight the potential utility of pharmacophoric constraints in improving hit list enrichment. This has been shown in a practical context through studies employing LUDI [14] and CATALYST [14] to similar effect in the discovery of novel DNA gyrase inhibitors [15]. There are a number of other advantages to the application of such constraints that are also worthy of mention. For DOCK, the use of such constraints dramatically decreases the number of sub-optimal binding orientations sampled during searching. The result is an equally dramatic increase in search speed, generally in the range of one to two orders of magnitude. For FAB protein 2 and serine protease 2, one chemotype was missed by the constrained search in each case. Analysis of these chemotypes highlights another advantage of using pharmacophore descriptors. The missing serine protease chemotype contained a pyridine basic moiety, while the acid in the FAB protein 2 lead was found to be a 2-hydroxypyridine. Both were absent from the original pharmacophore acid and base definitions, but can be included simply through the addition of two lines in the definitions file (Table 2). Doing the same for electrostatic potential is significantly more involved, however, since the database would need to be rebuilt with (de)protonation of the relevant functionality, followed by recalculation of the charge.

Such flexibility extends to additional areas of SVS searching. For example, if the importance of a group of hydrogen bonds is not known, but the user wishes to ensure their presence in a given search, the relevant group of site points would simply be assigned as a critical region. This forces a match to at least one of the points, thereby ensuring the presence of a matching hydrogen bonding moiety. Should two need to be matched, the site points can be copied and superimposed over their original counterparts. These copies are assigned to a second critical region, with the DOCK *distance_minimum* function ensuring that a site point and its twin can not be matched simultaneously. In this way multiple pharmacophore h bond pair constraints can be searched in a single screen, creating an extremely flexible protocol.

The importance of such user hypotheses have been broached elsewhere. Baxter et al. presented a nice example of VS success for a search targeted at the estrogen receptor (http://www.lib.uchicago.edu/cinf/220nm/slides/220nm16/220nm16.pdf). In it the need for post-screen analysis and filtering criteria are stressed. Pharmacophore constraints permit the user to build such criteria for binding up front, thus saving much time in post-screen hit evaluation.

5. Conclusions

The results presented here highlight the importance of sampling in SVS results, with small changes in sampling parameters producing potentially significant effects on performance. In addition, an alternative method for measuring enrichment rates is presented that provides a potentially more realistic measure of SVS performance. Most importantly, pharmacophore sampling constraints are shown to provide numerous SVS calculation enhancements, potentially improving hit rates, user control and search speed. Indeed, using such constraints at current SVS algorithm sampling levels allows simple scoring functions (DOCK contact score) to easily compete with those of more complex origin. Such constraints rely inherently on a good knowledge of target structural biology. For many target classes this data is available, however, and it is this data that generally guides scientific hypotheses for primary ligand binding modes. Further, given the continuing inability of most scoring functions to differentiate the relative importance of salt bridges and hydrogen bonds, the capacity to exploit such knowledge explicitly takes on added importance. With these advantages in mind, the incorporation of pharmacophore constraint functionality should be considered a high priority for any developer of SVS search algorithms.

References

[1] DOCK, developed and distributed by the Kuntz group, Department of Pharmaceutical Chemistry, 512 Parnassus, University of California, San Francisco, USA. URL: http://www.cmpharm.ucsf.edu/kuntz.
[2] C.A. Baxter, C.W. Murray, B. Waszkowycz, J. Li, R.A. Sykes, R.G.A. Bone, T.D.J. Perkins, W. Wylie, New approach to molecular docking and its application to virtual screening of chemical databases, J. Chem. Inf. Comput. Sci. 40 (2000) 254–262.
[3] GOLD version 1.2, developed and distributed by the CCDC, 12 Union Road, Cambridge, CB2 1EZ, UK. URL: http://www.ccdc.cam.ac.uk/prods/gold/index.html.
[4] A.C. Good, J.S. Mason, S.D. Pickett, Pharmacophore pattern application in virtual screening, library design and QSAR, in: H.J. Böhm, G. Schneider (Eds.), Virtual Screening for Bioactive Molecules: Methods and Principles in Medicinal Chemistry, vol. 10, Wiley, 2000, pp. 131–156.
[5] B.E. Thomas, D. Joseph-McCarthy, J.C. Alvarez, Pharmacophore-based molecular docking, in: O.F. Guner (Ed.), Pharmacophore Perception, Development, and Use in Drug Design, International University line La Jolla, 2000, pp. 351–367.
[6] S. Ha, R. Andreani, A. Robbins, I. Muegge, Evaluation of docking/scoring approaches: a comparative study based on MMP3 inhibitors, J. Comput. Aided. Mol. Des. 14 (2000) 435–448.
[7] B.K. Shoichet, I.D. Kuntz, Matching chemistry and shape in molecular docking, Prot. Eng. 6 (1993) 723–732.
[8] C. Bissantz, G. Folkers, D. Rognan, Protein-based virtual screening of chemical databases. Evaluation of different docking/scoring combinations, J. Med. Chem. 43 (2000) 4759–4767.
[9] P.S. Charifson, J.J. Corkery, M.A. Murcko, W.P. Walters, Consensus scoring: a method for obtaining improved hit rates from docking databases of three dimensional structures into proteins, J. Med. Chem. 42 (1999) 5100–5109.
[10] S.J. Weiner, P.A. Kollman, D.A. Case, C. Singh, C. Ghio, G. Alagona, S. Profeta Jr., P. Weiner, A new force field for molecular mechanical

simulation of nucleic acids and proteins, J. Am. Chem. Soc. 106 (1984) 765–784.
[11] GRID Version 16, Developed and Distributed by Molecular Discovery Limited, West Way House, Elms Parade, Oxford OX2 9LL, UK.
[12] M.L. Connolly, Solvent-accessible surfaces of proteins and nucleic acids, Science 221 (1983) 709–713.
[13] CONCORD 4.0, distributed by Tripos Inc., 1699 South Hanley Road, Suite 303, St. Louis, MO 63144, USA. URL: http://www.tripos.com.
[14] LUDI/CATALYST, Developed and Distributed by Accelrys, 9685 North Scanton Road, San Diego, CA 92121, USA. URL: http://www.accelrys.com.
[15] H.J. Böhm, M. Boehringer, D. Bur, H. Gmuender, W. Huber, W. Klaus, D. Kostrewa, H. Kuehne, T. Luebbers, N. Meunier-Keller, F. Mueller, Novel inhibitors of DNA Gyrase: 3D structure-based biased needle screening, hit validation by biophysical methods, and 3D guided optimization. A promising alternative to random screening, J. Med. Chem. 43 (2000) 2664–2674.