Protein Engineering vol.6 no.7 pp.723–732, 1993

Matching chemistry and shape in molecular docking

Brian K.Shoichet[1] and Irwin D.Kuntz[2]

Department of Pharmaceutical Chemistry, University of California, San Francisco, CA 94143-0446, USA

[1]Present address: Institute of Molecular Biology, University of Oregon, Eugene, OR 97403, USA
[2]To whom correspondence should be addressed

We have added a chemical filter to the ligand placement algorithm of the molecular docking program DOCK. DOCK places ligands in receptors using local shape features. Here we label these shape features by chemical type and insist on complementary matches. We find fewer physically unrealistic complexes without reducing the number of complexes resembling the known ligand–receptor configurations. Approximately 10-fold fewer complexes are calculated and the new algorithm is correspondingly 10-fold faster than the previous shape-only matching. We tested the new algorithm's ability to reproduce three known ligand–receptor complexes: methotrexate in dihydrofolate reductase, deoxyuridine monophosphate in thymidylate synthase and pancreatic trypsin inhibitor in trypsin. The program found configurations within 1 Å of the crystallographic mode, with fewer non-native solutions compared with shape-only matching. We also tested the program's ability to retrieve known inhibitors of thymidylate synthase and dihydrofolate reductase by screening molecular databases against the enzyme structures. Both algorithms retrieved many known inhibitors preferentially to other compounds in the database. The chemical matching algorithm generally ranks known inhibitors better than does matching based on shape alone.

Key words: computer-aided drug design/molecular database/molecular docking/shape complementarity/structure-based inhibitor design

Introduction

Molecular docking fits ligands into receptors in favorable configurations. The method is used in studies of the structural basis of biological function (Goodsell and Olson, 1990; Cherfils *et al.*, 1991; Stoichet and Kuntz, 1991; Stoddard and Koshland, 1992) and drug design (Goodford, 1984; DesJarlais *et al.*, 1990; Shoichet *et al.*, 1993).

Docking methods fall into two broad categories: those that use kinetic techniques to smoothly explore molecular potential surfaces (Goodsell and Olson, 1990; Caflisch *et al.*, 1992; Stoddard and Koshland, 1992) and those that sample discrete configurations of the ligand–receptor complex. Discrete sampling can be accomplished by grid search (Wodak *et al.*, 1987; Cherfils *et al.*, 1991; Jiang and Kim, 1991; Wang, 1991) and descriptor matching (Kuntz *et al.*, 1982; Connolly, 1985; Smellie *et al.*, 1991; Bacon and Moult, 1992; Lawrence and Davis, 1992; Shoichet *et al.*, 1992) methods. Grid searches seek complementary ligand configurations at regular intervals of translation and rotation. Descriptor methods predefine regions of likely complementarity on the receptor and the ligand, which they superimpose to fit the two molecules together.

We have previously published a rigid-body docking method that uses molecular descriptors (DOCK program; Kuntz *et al.*, 1982; DesJarlais *et al.*, 1988; Meng *et al.*, 1992; Shoichet *et al.*, 1992). DOCK matches receptor pockets to ligand atoms or bumps in the ligand's surface, to orient a ligand in a receptor. Thousands of orientations are typically sampled. The scoring of complementarity occurs after an orientation has been generated; in DOCK the sampling of configuration space has been distinct from the evaluation of the complex. Until now, *matching* has been guided by the local shape of the docking molecules, but *scoring* has evaluated overall chemical complementarity and steric fit.

Here we introduce a simple modification to our matching algorithm that speeds DOCK by an order of magnitude and improves the program's selectivity. Whereas previously we allowed any receptor descriptor to pair with any ligand atom, we now insist that a descriptor pocket be chemically complementary to the ligand atom with which it is to be matched. We label all descriptors and atoms by chemical category. This modification only affects the generation of ligand orientations. It does not change the DOCK score, which is unmodified from earlier work. We refer to this new algorithm as 'labeled matching' and to the earlier, shape-only algorithms as 'unlabeled matching'.

We test the labeled matching algorithm's ability to reproduce ligand geometries using three crystallographically determined protein–ligand complexes: dihydrofolate reductase (DHFR) with methotrexate, thymidylate synthase (TS) with deoxyuridine monophosphate (dUMP) and trypsin with bovine pancreatic trypsin inhibitor (BPTI). We also test the new algorithm's ability to screen molecular databases for inhibitors of DHFR and TS. All structures are from the Protein Data Bank (PDB) (Bernstein *et al.*, 1977); we use DHFR/methotrexate structure 3drf (Bolin *et al.*, 1982), TS structure 4tms (Finer-Moore *et al.*, 1990), dUMP structure 2tsc (Montfort *et al.*, 1990) and trypsin/BPTI structure 2ptc (Marquart *et al.*, 1983).

Descriptor matching

The computational problem is to describe the features that define the shape of the receptor site and that of the ligand and then match the two sets of features together in favorable ways. There are many ways of describing molecules for this purpose (Kuntz *et al.*, 1982; Connolly, 1985; Leicester *et al.*, 1988; Bacon and Moult, 1992; Lawrence and Davis, 1992). We use spheres that are locally complementary to pockets in the molecular surface of the receptor (Kuntz *et al.*, 1982; Connolly, 1983) and ligand atom centers (DesJarlais *et al.*, 1986) or, for macromolecular ligands, spheres complementary to bumps in the ligand's molecular surface (Shoichet and Kuntz, 1991).

Four matched features suffice to dock a ligand into a receptor (Shoichet *et al.*, 1992). Using more features further constrains the ligand to the site: we typically use set sizes of four or five for our work. Taking the larger number, the number of orientations to explore is bounded by (Shoichet *et al.*, 1992):

$$\text{Number of orientations} = (N_r)^5 \times (N_l)^5 \quad (1)$$

where N_r is the total number of receptor features and N_l is the total number of ligand features.

The features in equation (1) are shape-based, containing no information about ligand chemistry or receptor chemical environment. Thus, many matches will result in high energy complexes. Matching a primary amine into an electrostatically positive region of the receptor, for instance, is usually a bad idea, even if it is sterically acceptable. Forbidding such matches should not diminish the program's accuracy. Eliminating them early in the procedure will speed the docking calculation.

We assign every atom or sphere to one of five chemical categories: positively charged, negatively charged, electron acceptor, electron donor and neutral. The number of categories and the category types are arbitrary and can be extended or modified. Our matching rules permit only positive atoms to match with negative spheres, only electron donor atoms with electron acceptor spheres and only neutral atoms with neutral spheres. Other matching rules can be imagined as the categories are altered or extended. Double-typing an atom or sphere is allowed and will reduce the severity of this restriction. In the DHFR–methotrexate docking calculation (below), for instance, we labeled a sphere of intermediate potential both negative and neutral by listing it twice.

Categorization reduces the number of orientations generated, yielding equation (2):

$$\text{Number of orientations} = \left\{ \sum_i^n (N_{ri} \times N_{li}) \right\}^5 \quad (2)$$

where i indicates the following pairs of classes: neutral atom, neutral sphere; positive atom, negative sphere; negative sphere positive atom; donor sphere, acceptor atom; and acceptor sphere, neutral atom. N_r and N_l are the number of receptor and ligand features in each category.

If we compare equation (2) with equation (1), we see that the number of descriptors in each category in equation (2) is a fraction of the total number of descriptors in equation (1). This will have a dramatic effect on the number of possible orientations. For instance, if all five categories were equally represented in the sphere and ligand sets, i.e. if the fraction of neutral spheres and atoms equalled the fraction of positive spheres and atoms and so on, then each category would contain one-fifth of the total number of descriptors. In this case, the number of possible orientations under categorization (equation 2) would fall to 1/3125 of the number of possible orientations without categorization (equation 1).

To take full advantage of this potential speed-up, we impose the chemical-fit criterion early in the matching process. DOCK builds matching sets by adding one atom–sphere pair to the growing sets at a time. The pairwise internal distances within the new sets are then measured to insure that they can still be overlapped (Kuntz *et al.*, 1982; Shoichet *et al.*, 1992). We now add the condition that the new atom must chemically complement the new sphere, for every atom–sphere pair added to the growing set. We calculate a physical orientation only once the set reaches a certain size (five pairs in equation 2). Since the set will be abandoned due to any non-complementary pair before this size is reached, the chemical complementarity condition eliminates a large number of potential orientations before they are fully produced.

Several other descriptor-based docking algorithms use some form of chemical matching. Kuhl *et al.* (1984) used 'repellent matchings' to prevent ligands from binding in regions where pharmacophore or structural evidence suggests they should not. While this has the same logical sense as chemical matching, the authors did not use it for such a purpose. Lawrence and Davis (1992) have recently published a program, CLIX, that is closer to the work we present here. They chose and categorized receptor 'target sites' (descriptors) by their chemical environment using the GRID algorithm (Goodford, 1985), while an automatic procedure categorized ligand atoms. CLIX orients ligands by picking two pairs of complementary atoms and receptor descriptors and moving the ligand about the resulting line of rotation in the site. We differ from Lawrence and Davis by using five matches to dock a ligand instead of two, which further prunes the search tree (equation 2). Balancing this, Lawrence and Davis (1992) used many more descriptor types, 23 to our five. We do not know how best to categorize ligand atoms and their receptor targets, nor how many categories to use. The best way to choose receptor descriptors is currently a matter of opinion. We have usually chosen to relabel descriptors previously picked for shape definition of the site. It might be useful to choose receptor descriptors from the start using chemical complementarity, as suggested by Lawrence and Davis (1992). These are areas for future research.

Methods

Since DOCK has been described elsewhere (Meng *et al.*, 1992; Shoichet *et al.*, 1992), we outline here only how we type spheres and atoms and how the test systems were run. The classification of atoms and spheres takes place outside the docking program, which acts on whatever categories the user has defined. The user can assign types directly (by hand) or automatically through an atom-typing program—we use both methods in this paper. As currently implemented, up to five different types can be chosen. DOCK reads the atom and sphere types from the input files at the beginning of a run.

To assign sphere types automatically, we used the molecular electrostatic potential at the sphere center. We calculated receptor electrostatic potentials using the program DELPHI (Gilson and Honig, 1987), measuring the potential at sphere centers using the PHITOPDB utility distributed with DELPHI. Sphere centers in regions of potential $> +30$ kT were categorized positive. Sphere centers in regions of potential < -30 kT were categorized negative, except in the generally electropositive TS site, where a cut-off of -20 kT was used. All other sphere centers were categorized neutral.

Receptor descriptors used by DOCK sometimes take the form of inhibitor atom positions, as determined in the structure of an inhibitor–receptor complex (Shoichet *et al.*, 1993), rather than spheres. Using inhibitor atom positions highlights a binding region more exactly than spheres, though one is restricted to a known region of the site in the docking calculations. In the search for DHFR inhibitors, we wished to label the binding region more finely than we had in the TS and trypsin searches, using five categories of labels rather than three. We thus chose to describe the DHFR-binding site using the atomic positions of crystallographically determined inhibitors. We used methotrexate and folate pterin ring atoms from two DHFR complexes (Bolin *et al.*, 1982; Bystroff *et al.*, 1990) to determine the chemical character of the receptor descriptors. These ligand atomic coordinates were used as sphere centers. Pteridine nitrogens were categorized electron acceptors (to complement electron donor ligand atoms)

and the exocyclic oxygen of folate was categorized an electron donor (to complement electron acceptor ligand atoms). The pteridine carbon atoms were categorized neutral.

We assigned chemical types to ligand atoms based on their hybridization, functional group and protonation state. Atoms that were in functional groups carrying a positive charge (sp^3 hybridized nitrogens with four substituents, amidinium and guanidinium nitrogens and carbons) were categorized positive and atoms in functional groups carrying a negative charge (carboxylate oxygens and carbons, phosphate and phosphonate oxygens and phosphorus, sulphate and sulfonate oxygens and sulfurs) were categorized negative. All other ligand atoms were categorized neutral, except for the ligands used in the DHFR database search. In this last case, some atoms were also categorized electron donor or electron acceptor. Electron donor atoms included sp^2-hybridized oxygens and sulfurs and sp^3-hybridized oxygens in ether or ester moieties. Electron acceptor atoms included sp^2 hybridized and aromatic nitrogens, as well as sp^3-hybridized nitrogens possessing a hydrogen and conjugated to a pi orbital system (such as an amide). The ligand atomic classifications were made automatically by a program MOL2DB (a utility program distributed with DOCK) based on their SYBYL (Molecular Modeling System SYBYL, Version 5.4, TRIPOS Associates, Inc., St Louis, MO) atom types and connectivities. We calculated ligand atom partial atomic charges using the Gasteiger option in SYBYL (Gasteiger and Marsili, 1980), except for the PTI structure, where we used the AMBER charges (Weiner *et al.*, 1984).

Test cases: reproducing crystallographically determined complexes

We pre-coded volume elements of the active sites for steric and electrostatic potential using regular lattices. We used the program DISTMAP to score the sites for steric tolerances (Shoichet *et al.*, 1992) and electrostatic grids, calculated with DELPHI, to compute the sites' electrostatic potentials (Meng *et al.*, 1992). The AMBER united atom charge set, distributed with DELPHI, was used for all receptor electrostatic calculations. The EDIT program in AMBER automatically placed all heteroatom hydrogens for the DELPHI calculation.

Trypsin–PTI (2ptc) (Marquart et al., 1983). We calculated trypsin and PTI spheres as previously described (Shoichet and Kuntz, 1991). The molecular electrostatic potential was calculated with DELPHI using a three-step focusing method (Gilson and Honig, 1987) at 20, 60 and 90% containment of the protein using a 65^3 lattice. We did not include any water molecules or counterions in either the electrostatic or steric calculations. We set the dielectric to 2 internally, 78.5 externally and used a 2.0 Å ion exclusion radius. Protein van der Waals radii were taken from Rashin (1990). Eight of 34 receptor spheres had potentials < -30 kT and were categorized negative and three spheres overlapping lysine 15 were categorized positive out of 60 BPTI spheres overall. All other receptors and ligand spheres were categorized neutral. For the docking calculation we used a distance tolerance of 1.5 Å between matching pairs of ligand and receptor nodes, a node limit of 5 and bin sizes and overlaps (Shoichet *et al.*, 1992) of 0.5 Å for the ligand and receptor. We disallowed all bad contacts and used single-step focusing (Shoichet *et al.*, 1992).

DHFR–methotrexate (3dfr). We used DHFR spheres generated in earlier work (Shoichet *et al.*, 1992). The electrostatic potential was calculated as above. Seven of 89 spheres had potentials < -30 kT and were categorized negative. One sphere that had a potential of -24.5 kT was labeled both negative and neutral. Methotrexate atoms N1, N3 and C2 were categorized positive and its carboxylate oxygens and carbons were categorized negative. The docking calculation used the same parameters as the trypsin run, but bin sizes were set to 0.2 and 1.0 Å for the ligand and the receptor respectively, with no overlaps.

TS–dUMP. We generated TS spheres as described previously (Shoichet *et al.*, 1992). Twelve spheres had potentials > 30 kT and were categorized positive, three had potentials < -20 kT and were categorized negative, leaving 52 neutral. The dUMP phosphorus and the four oxygens attached to it were categorized negative, while all 15 other ligand atoms were categorized neutral. The docking calculation used the same parameters as above, except that we used bin sizes of 0.5 and 1.0 Å for the ligand and the receptor respectively, with no overlaps.

Test cases: database searches

DOCK was used to search two molecular databases for compounds complementary to DHFR and TS. Both databases were subsets of the 1990 edition of the *Fine Chemicals Directory* (FCD) (Molecular Design Limited, San Leandro, CA) (Guner *et al.*, 1991). The TS database consisted of 696 compounds, each of which contained either a sulfate, sulfonate, phosphate or phosphonate moiety. The DHFR database was created with the similarity search facility in the program MACCS (Guner *et al.*, 1991), which we used to select all compounds with two fused six membered rings (5909 compounds). We calculated an electrostatic solvation free energy for every compound using the modified Born equation treatment, as implemented by Rashin

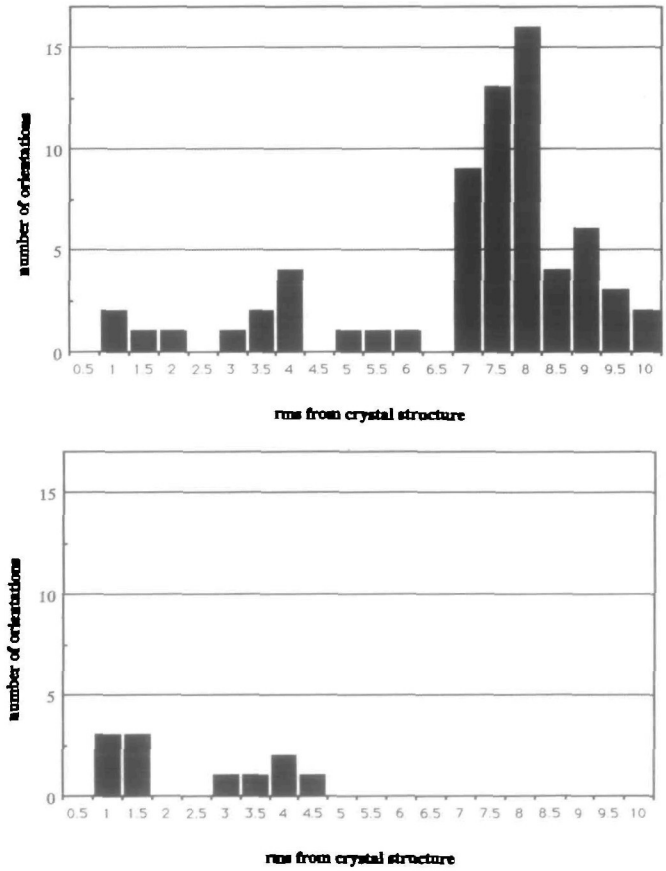


Fig. 1. (a) dUMP docked into TS, distribution of orientations from unlabeled matching calculation (r.m.s. in Å for Figures 1–3). (b) dUMP docked into TS, distribution of orientations from labeled matching calculation.

(1990) and Shoichet *et al.* (1993). The number of low-energy conformations accessible to each compound was calculated using the COBRA program (Leach and Prout, 1990). We calculate an electrostatic interaction energy for every ligand orientation by mapping ligand atoms onto the electrostatic grid potential (P) and multiplying the potential of each atom (i) by its partial charge (C), summing over all atoms (n) (Shoichet *et al.*, 1992). We subtract the solvation energy (ΔG_{solv}) and the natural logarithm of the number of conformations (N) multiplied by RT (the gas constant multiplied by the temperature), from this interaction energy to provide an electrostatic free energy of binding corrected for the loss of ligand conformational entropy to give equation (3):

$$\text{Energy} = \sum_{i}^{n} C_i \times P_i - \Delta G_{solv} - RT \times \ln N \qquad (3)$$

We used SYBYL to add N1 hydrogens to the 47 2,4-diaminopterins and quinazolines in the DHFR database, giving the molecules formal positive charge at this position. The TS database search used the same spheres, steric grid and docking parameters as the dUMP single-molecule run. In the TS and DHFR searches we attempted to allow for the effect of ligand binding on the dielectric in the active site, which can be an important consideration when comparing the interaction energies of chemically different ligands. We included the docking spheres as uncharged receptor atoms of 1.4 Å radius and recalculated the receptor electrostatic potentials with DELPHI (Shoichet *et al.*, 1993). Including the receptor spheres as receptor 'atoms' lowers the active site dielectric, which usually increases the absolute magnitude of the potential in the binding site.

The DHFR database search used the atomic positions of the pterin moieties of methotrexate and folate as receptor 'sphere' centers. We placed folate in the 3dfr site by r.m.s.-fitting the C-alpha atoms from PDB structure 7dfr (Bystroff *et al.*, 1990) onto those of 3dfr using EXIFIT (McLachlan, 1982) and applying the resulting rotation and translation matrix to the folate atoms. We classified all nitrogens 'acceptor', the folate O4 atom 'donor' and all carbons 'neutral'. We also categorized nitrogens N5 and N8 from both methotrexate and folate as 'neutral'—this double-typing allowed either neutral or donor ligand atoms to match with these nitrogens.

The DHFR steric grid was unchanged from the methotrexate docking run. We changed the protein structure in two ways for the DELPHI electrostatic calculation. Structurally conserved water 253 from the 3dfr structure was included as part of the protein—the analog of this water (water 206) ligates the folate O4 in the 7dfr structure (Bystroff *et al.*, 1990). Also, the hydroxyl Oγ to hydrogen dihedral angle of Thr116, which in the AMBER

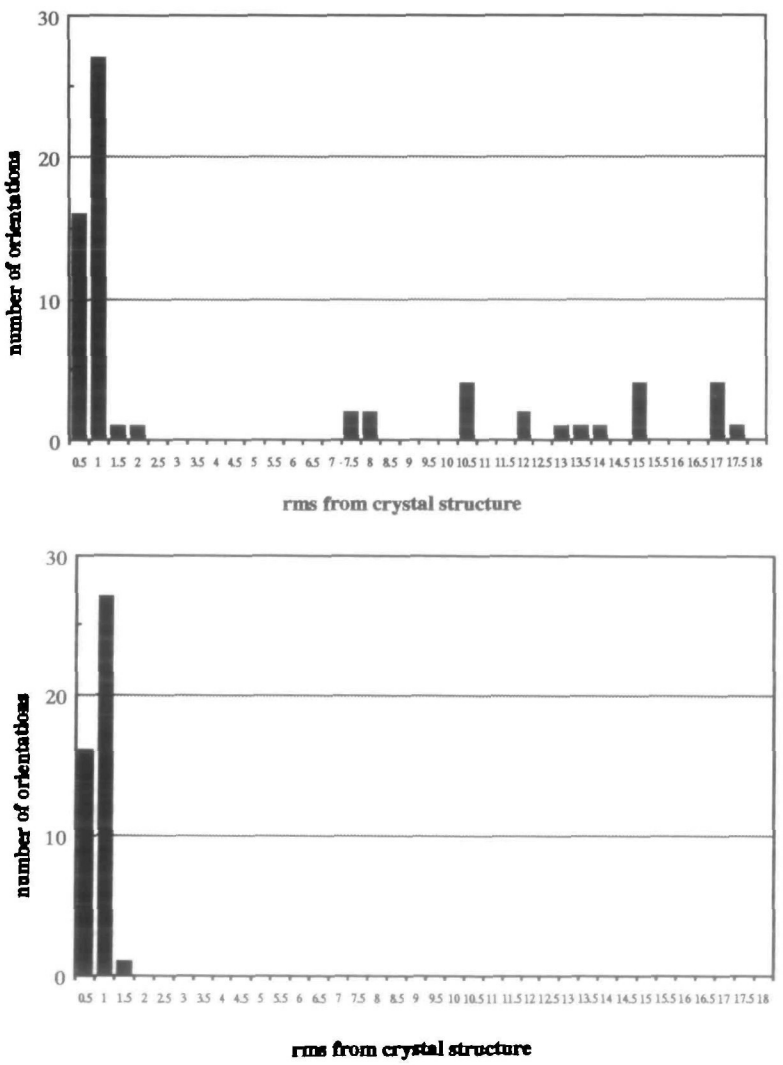


Fig. 2. (a) Methotrexate docked into DHFR, distribution of orientations from unlabeled matching calculation. (b) Methotrexate docked into DHFR, distribution of orientations from labeled matching calculation.

file 'all.in' is set to 180°, was rotated to 120°, allowing it to interact better with methotrexate. No heavy atoms were moved in making this change.

We performed all calculations on Iris PI 35 workstations (Silicon Graphics, Mountain View, CA 94039) with 32 Mb of memory.

Results

Using the labeled matching algorithm for single-ligand docking (DOCK SINGLE mode), we reproduced the crystallographic complexes accurately (Figures 1–3), 2–13 times faster than the unlabeled matching algorithm (Table Ia). Labeled matching was an order of magnitude faster than unlabeled matching in the database searches (DOCK SEARCH mode) (Table Ib).

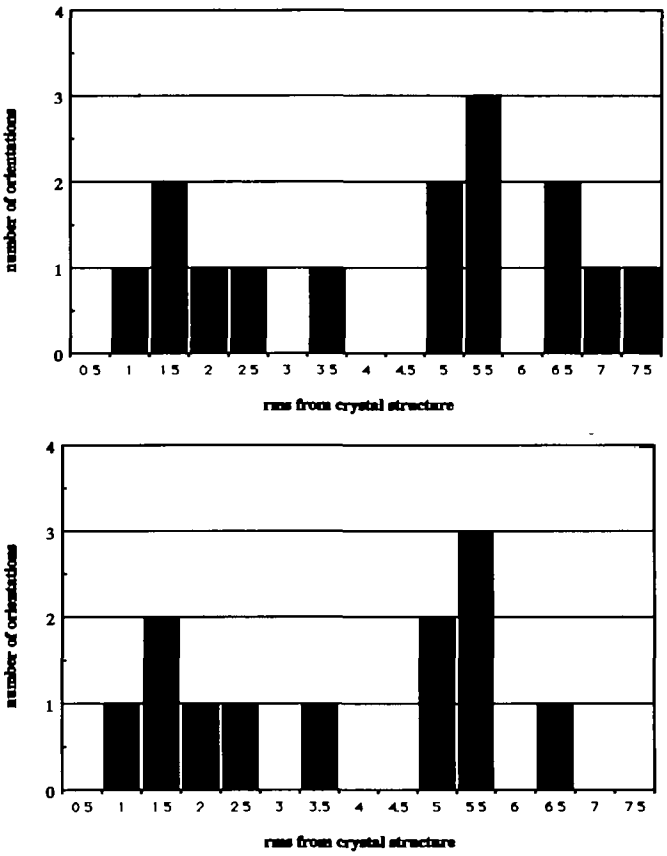


Fig. 3. (a) BPTI docked into trypsin, distribution of orientations from unlabeled matching calculation. (b) BPTI docked into trypsin, distribution of orientations from labeled matching calculation.

With the shorter calculation time of labeled matching came improved selectivity. Labeled matching found a larger percentage of orientations close to the crystallographic complex than did unlabeled matching. In the dUMP–TS and methotrexate–DHFR calculations (Figures 1 and 2) labeled matching found dramatically fewer configurations far from the crystallographic complex than did unlabeled matching. In the BPTI–trypsin calculations (Figure 3) the selectivity improvement is only modest, though the improvement in calculation time is similar. This difference reflects the higher steric specificity of the BPTI–trypsin interface: there are fewer 'wrong' ways to fit BPTI into trypsin, irrespective of the chemistry. In the screens of molecular databases against enzyme structures, the selectivity advantage translated into an improved ability to identify inhibitors. We consider the TS and DHFR results in turn.

TS binds selected pyrimidine monophosphates (PMPs), of which dUMP (Figure 4) is a member, at micromolar concentrations. Of the 696 compounds in the TS database, 47 were PMPs. Labeled matching discriminated in favor of the PMPs compared with other compounds in the database. Labeled matching ranked 27 PMPs higher than they were with unlabeled matching, while only eight had lower ranks (Figure 5a).

Labeled matching excluded two PMPs that are good (micromolar) inhibitors of TS and that unlabeled matching found to have negative (good) scores. These compounds are the naphthyl ester of TMP and the aminophenyl ester of fdUMP (Santi and Danenberg, 1984), ranked 263 and 301 in the unlabeled search. While these ranks are relatively low, the absolute exclusion of these inhibitors from the labeled matching might suggest that the new algorithm achieves some of its improved efficiency at the expense of accuracy. Inspection of the complexes generated by the unlabeled matching algorithm (Figure 4), however, shows that both inhibitors were docked in manners that do not resemble the crystallographic mode of dUMP. The two molecules were docked with their phosphates in the ribose region of the site, 3 Å distant from the crystallographic phosphate and with their pyrimidine rings in the folate region. Neither inhibitor can sterically fit into the dUMP binding pocket in the conformation that they are represented in the FCD. The phosphate moieties

Table I. (a) Calculation times for single-ligand docking runs: labeled and unlabeled. (b) Run times for database searches: labeled and unlabeled

(a) Receptor	Ligand	Algorithm	Best r.m.s. (Å) to crystal structure	Number of orientations	Run time (s)
DHFR	Methotrexate	Labeled	0.286	2,565	18
DHFR	Methotrexate	Unlabeled	0.286	35,912	108
TS	dUMP	Labeled	0.591	676	22
TS	dUMP	Unlabeled	0.591	8,373	40
Trypsin	BPTI	Labeled	0.585	13,233	135
Trypsin	BPTI	Unlabeled	0.585	95,253	1839

(b) Receptor	Number of compounds searched	Algorithm	Mean number of orientations calculated per compound	Run time (s)
DHFR	5909	Labeled	446	3310
DHFR	5909	Unlabeled	9185	43329
TS	696	Labeled	221	399
TS	696	Unlabeled	4076	3391

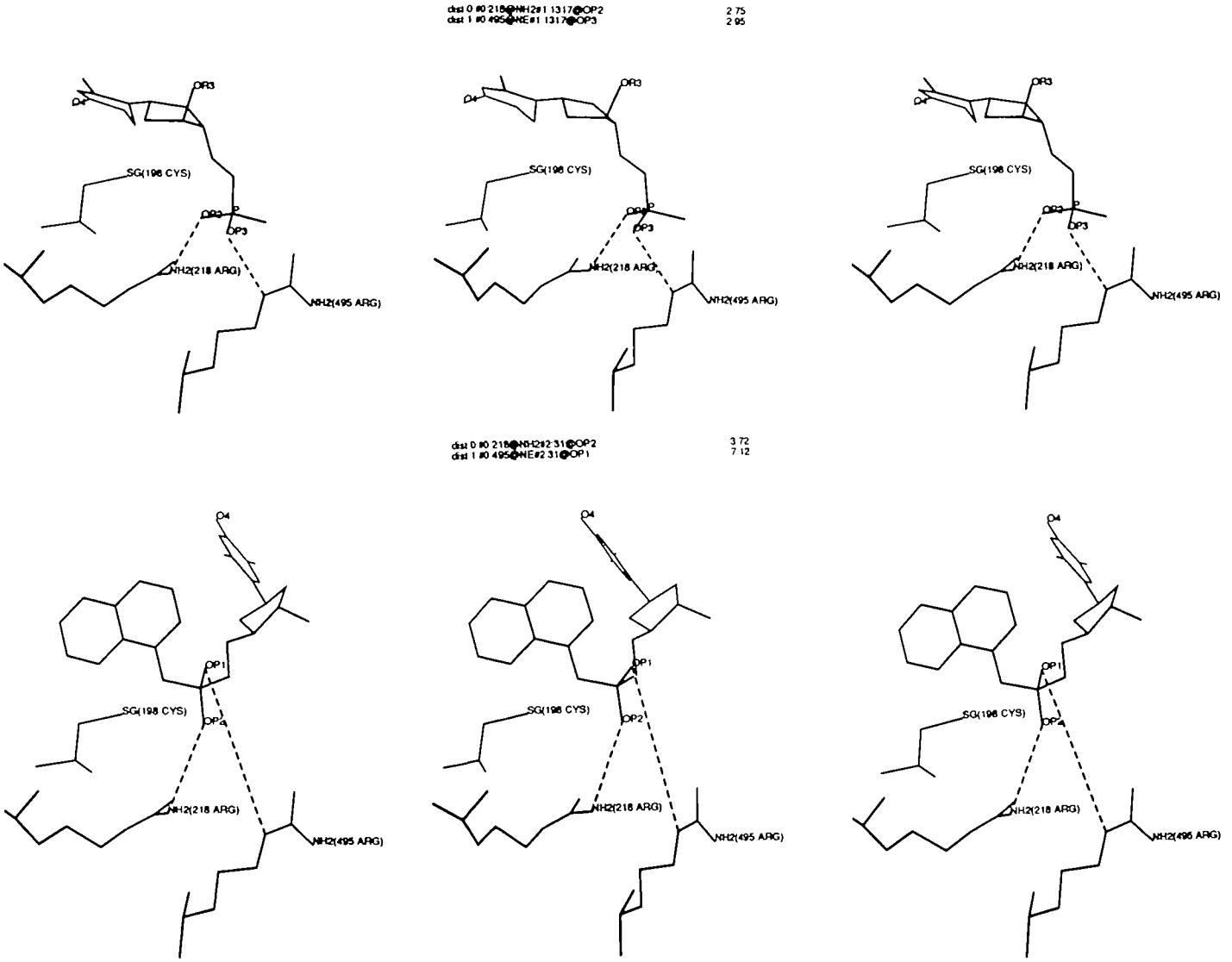


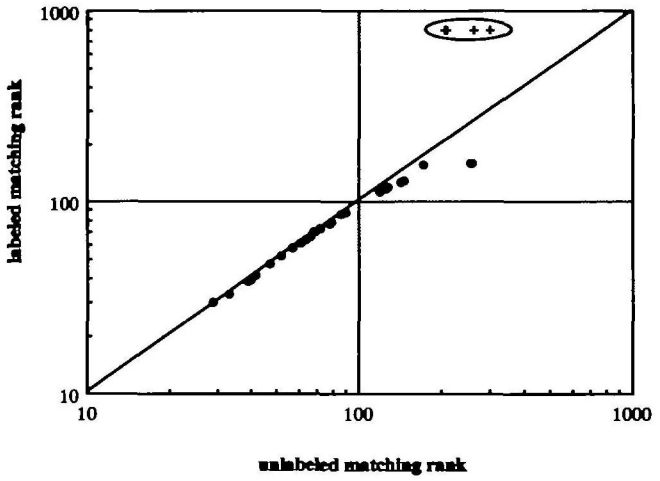
Fig. 4. Stereoimage of the nucleotide binding site of thymidylate synthase, comparing the binding modes of the crystallographic dUMP and naphthyl FdUMP from the unlabeled matching algorithm. Two of the phosphate-recognizing arginines are shown as is the active site cysteine. (a) The crystallographic configuration of dUMP in its Michael adduct conformation (Montfort *et al.*, 1990). Three phosphate moiety atoms are labeled, as is the O3′ (OR3) hydroxyl of the ribose and the O4 of the nucleotide. The dashed lines show hydrogen-bonding interactions between the NH2 of Arg219 and the OP2 oxygen of the dUMP phosphate (2.75 Å) and between the N_ϵ of Arg495 and the OP3 oxygen of the dUMP phosphate (2.95 Å). (b) The unlabeled matching configuration of naphthyl FdUMP. Two phosphate oxygens are labeled, as is the O4 of the nucleotide between Arg219 and Arg495 and the phosphate oxygens are now 3.72 and 7.12 Å respectively, too far for hydrogen-bonding interactions. Note that the nucleotide is pointing in the opposite direction from the crystallographic dUMP.

of these two molecules compelled labeled matching to place them into just this region, explaining why no good scoring configurations were found. Unlabeled matching was not constrained to place these molecules in the dUMP site. The orientations it found (also sampled by the labeled matching) had fortuitously good scores but are nevertheless probably wrong.

DHFR binds selected 2,4-substituted pteridines and quinazolines at micromolar and submicromolar concentrations. Of the 5909 compounds searched, 184 were 2,4-substituted pteridines and quinazolines. As with TS, labeled matching discriminated in favor of the pteridines and quinazolines versus other compounds in the database. Fifty pteridines and quinazolines were ranked higher using chemical matching than they were using unlabeled matching, while 11 had lower ranks (Figure 5b).

Twenty pteridines and quinazolines that had negative (good) scores in the unlabeled matching calculation were excluded or had poor scores by labeled matching. In every case, their presence in the unlabeled matching list is due to the adoption of configurations that do not resemble the binding modes of either folate (in 7dfr) or methotrexate (in 3dfr). Orientations of the pteridines and quinazolines calculated with labeled matching usually overlapped the crystallographic configurations of methotrexate and folate better than did orientations calculated with unlabeled matching (Figure 6).

Two pterins, xanthopterin and tetrahydropterin, had significantly worse ranks in the labeled versus unlabeled matching search (Figure 5b, points 1 and 2). Neither compound is a good inhibitor of DHFR—tetrahydropterin is probably not an inhibitor at all—but the differences between the unlabeled and labeled binding modes illustrates the selectivity advantage of the latter. As with some PMPs in the TS search, unlabeled matching places these two molecules in orientations that do not resemble that of their crystallographically determined analog, in this case folate. Labeled matching, conversely, strongly overlaps these molecules



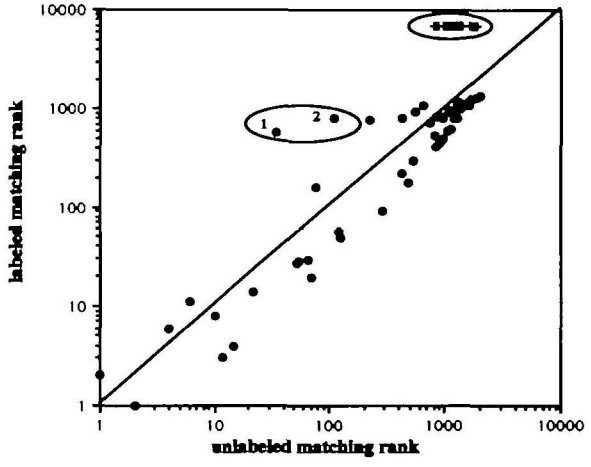


Fig. 5. (a) TS search. Pyrimidine monophosphate ranks (out of 696 total compounds), labeled versus unlabeled matching. Filled circles indicate compounds that were found by both matching algorithms and crosses indicate compounds that were only found by the unlabeled matching. The *y*-axis ranks of the latter are arbitrary and were set to 800 for ease of presentation. (b) DHFR search. Pterin and quinazoline ranks (out of 5909 total compounds), labeled versus unlabeled matching. Filled circles indicate compounds that were found by both matching algorithms and crosses indicate compounds that were only found by the unlabeled matching. The *y*-axis ranks of the latter are arbitrary and were set to 8000 for ease of presentation. Points 1 and 2 are xanthopterin and tetrahydropterin (see text).

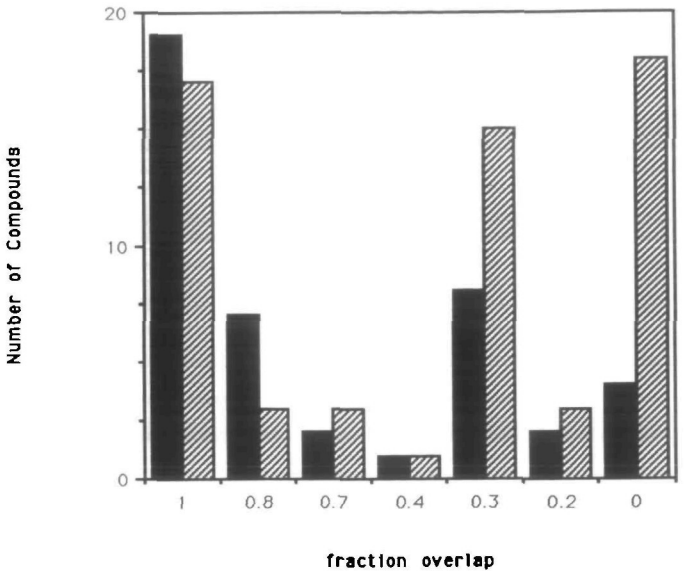


Fig. 6. Correspondence of docked and crystallographic orientations. We measure the overlap of pteridine and quinazoline ring heteroatoms of molecules selected from the DHFR database search with methotrexate pteridine and folate pterin ring heteroatoms. For two atoms to be considered overlapped, the docked atom must be within 0.5 Å of the corresponding methotrexate of the folate atom and the atoms must be of the same type (e.g. nitrogens can only be overlapped on nitrogens). To receive a value of 1, all possible heteroatoms from a docked molecule must overlap with a methotrexate or folate atom of the same type. To receive a value of 0, all possible heteroatoms from a docked molecule must not overlap with any methotrexate or folate atom of the same type. Intermediate values indicate intermediate overlap. Overlap values from the labeled algorithm are shown in the dark columns, overlap values from the unlabeled algorithm are shown in the hatched columns.

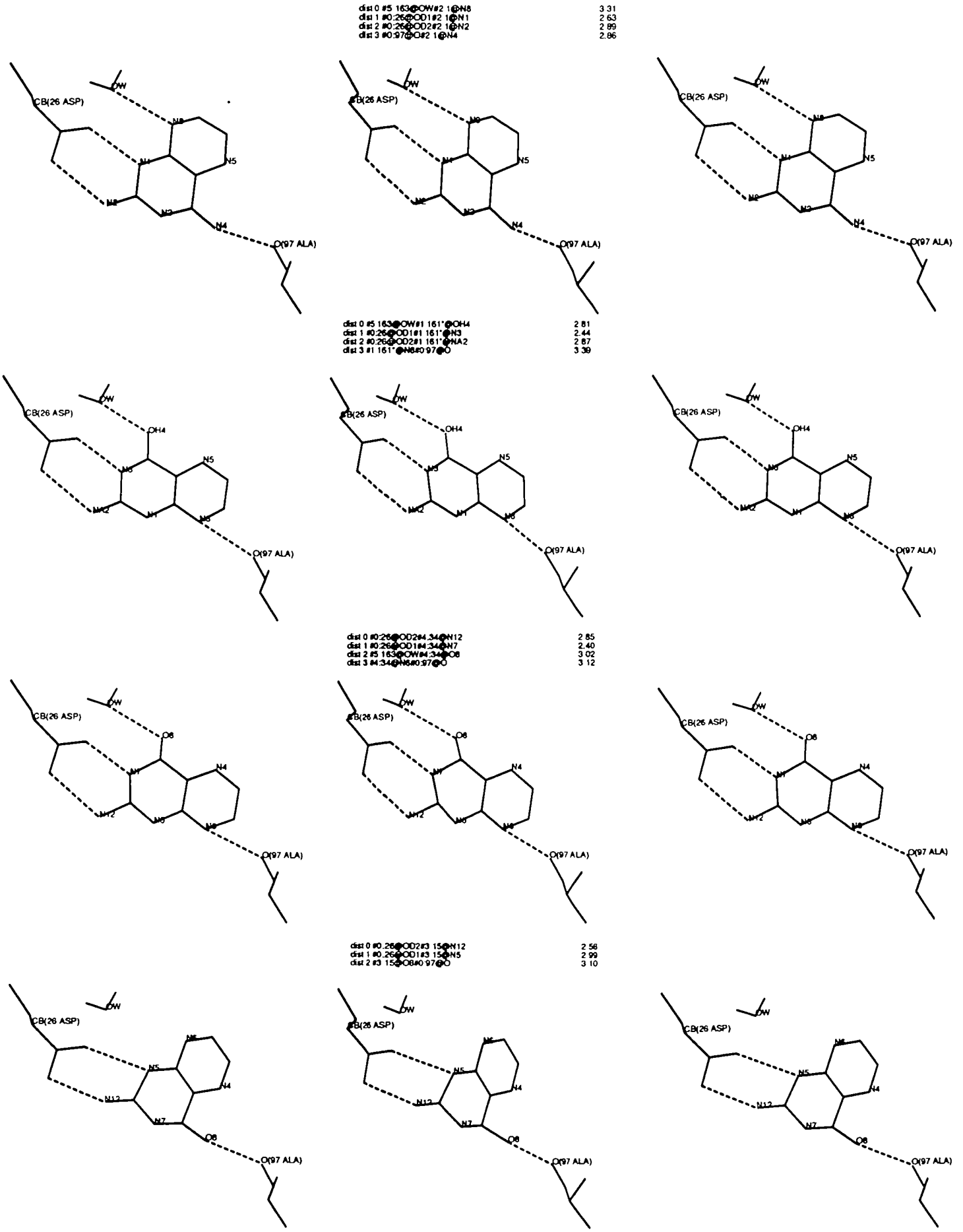
on folate. Because xanthopterin exists in the FCD in its 4-hydroxy tautomer, rather than the 4-keto tautomer expected for pterins and nucleotides, the labeled matching's configuration has a relatively poor interaction energy. The unlabeled matching algorithm, which is not constrained to overlap heteroatoms, finds an orientation of xanthopterin with a better electrostatic interaction energy. Notwithstanding its better energy, the unlabeled orientation is probably wrong. In the unlabeled docking of tetrahydropterin (Figure 7), the molecule makes a close salt bridge to D26. In order to overlap more closely the folate-derived 'spheres', the labeled matching docking moves the exocyclic amine involved in the salt bridge back ~0.2 Å from D26, reducing its electrostatic interaction energy. The salt bridge in the unlabeled docking case is probably too close at 2.5 Å; such a configuration would be penalized with a more sophisticated close contact limit than we used here. For xanthopterin and tetrahydropterin the low resolution chemistry of the labels recovers correct orientations that were missed due to high resolution errors in the ligand structure and the receptor potential function.

Discussion

Molecular docking is increasingly popular for structure-based inhibitor design and discovery. Two aspects of this work may affect such efforts. First, labeled matching improves DOCK's speed by an order of magnitude. Second, labeling improves the method's ability to select inhibitors from large lists of possible candidates.

The improved speed of the algorithm will allow us to treat problems that have been impractical until now. The database search calculations suggest, for instance, that it will be possible to screen databases of a million compounds in a week on a workstation. A faster algorithm is interesting, of course, only if the molecules that it suggests stand some chance of being inhibitors. DOCK has been successfully used for novel inhibitor discovery by searching molecular databases for inhibitors of several different enzymes (DesJarlais *et al.*, 1990; Kuntz, 1992; Shoichet *et al.*, 1993).

The selectivity advantage of labeled matching comes from eliminating many spurious configurations before they are generated. In the single-ligand calculations, fewer non-crystallographic configurations are found (Figures 1–3). In the database searches, eliminating such false positives improves the relative rankings of known inhibitors by displacing compounds that are unlikely to inhibit. *N*-Methyl-isoquinolium (ranked third in the unlabeled search but 942 in the labeled search) is probably not a DHFR inhibitor [it lacks the pattern of



endocyclic and exocyclic nitrogens that are ubiquitous in the pteridine–quinazoline series (Blaney *et al.*, 1984)], just as phosphothreonine (ranked eleventh in the unlabeled search but excluded in the labeled search) is probably not a TS inhibitor [most phosphates that are unable to form Michael adducts with the active-site cysteine inhibit the enzyme at 5 mM or worse (Santi and Danenberg, 1984; B.K.Shoichet, unpublished results)]. Labeled matching also improves the likelihood that the high-scoring docked configuration will resemble the crystallographic mode.

Does the increased speed of the labeled algorithm come at the expense of more false negatives in the database searches? Occasionally, labeled matching eliminates known inhibitors present in the unlabeled matching list. Even here, we believe that labeled matching is wrong for the right reasons and that unlabeled matching is right for the wrong ones. In the case of the two PMPs that were missed in the TS search (Figure 4), for example, unlabeled matching placed these molecules in what are probably unphysical regions of the site. This is also true of the pteridines and quinazolines missed or with reduced ranks in the DHFR search using the labeling algorithm.

DOCK's ability to retrieve known inhibitors from searches of general databases deserves special emphasis. These databases typically have ligands with many different functional groups and charge states. Accurate calculation of the differential binding energy of even two very similar inhibitors is computationally intensive and prone to error (Straatsma and McCammon, 1991). Selecting the few true inhibitors from the large number of possibilities in such a diverse list could conceivably require a calculation beyond our current capabilities. Our results suggest that relatively simple models of protein and ligand interactions are sufficient to identify many known inhibitors and to dock them into their cognate macromolecular receptors in likely binding modes. We cannot predict the binding energy of any single ligand and even our rankings are likely to be wrong in detail. Nevertheless, our algorithm does select many known and novel (Kuntz, 1992; Shoichet *et al.*, 1993) inhibitors by ranking them very highly compared to most other molecules that are found in searches of general databases.

Although we can identify many inhibitors from large lists of dissimilar compounds, it is also true that we miss many others. We did not find tight inhibitors such as methotrexate in the database search of DHFR and we missed several inhibitors in the TS search as well. These compounds were not found because the current algorithm docks molecules in rigid conformations. Methotrexate and the phosphate esters have many possible conformations; in the database they exist in a single one that does not fit into the cognate receptor site. Including conformational flexibility (Goodsell and Olson, 1990; Wilson *et al.*, 1991; Leach and Kuntz, 1992) will improve the algorithm's performance substantially, but will probably add significantly to computation time. Increasing the speed of our docking method by an order of magnitude makes time-consuming computations like flexible docking more tractable.

Conclusion

Adding chemical information to our matching algorithm improves the speed of the docking calculation by an order of magnitude and improves its selectivity. The amount of information that we added for this improvement was modest, involving three to five different categories of atoms or receptor environments. While our energy evaluation techniques are much too crude to predict the free energy of a complex, we have shown that the docking algorithm can select known inhibitors from amongst large lists of dissimilar compounds.

Acknowledgements

We thank Elaine Meng for critically reading this work and for many insightful conversations. Research was supported by NIH grants GM31497 and GM39553 (G.L.Kenyon, PI). The facilities of the Computer Graphics Laboratory (RR-1081 to R.Langridge) were used for part of this work. We thank Molecular Design Ltd for supplying MACCS-3D and the FCD and Tripos Associates for SYBYL. Figures 3 and 6 were produced with MidasPlus (Ferrin *et al.*, 1988).

References

Bacon,D.J. and Moult,J. (1992) *J. Mol. Biol.*, **225**, 849–858.
Bernstein,F.C., Koetzle,T.F., Williams,G.J.B., Meyer,E.F., Jr, Brice,M.D., Rodgers,J.R., Kennard,O., Shimanouchi,T. and Tasumi,M. (1977) *J. Mol. Biol.*, **112**, 535–542.
Blaney,J.M., Hansch,C., Silipo,C. and Vittoria,A. (1984) *Chem. Rev.*, **84**, 333–407.
Bolin,J.T., Filman,D.J., Matthews,D.A., Hamlin,R.C. and Kraut,J. (1982) *J. Biol. Chem.*, **257**, 13650–13662.
Bystroff,C., Oatley,S.J. and Kraut,J. (1990) *Biochemistry*, **29**, 3263.
Caflisch,A., Niederer,P. and Anliker,M. (1992) *Proteins*, **13**, 223–230.
Cherfils,J., Duquerroy,S. and Janin,J. (1991) *Proteins*, **11**, 271–280.
Connolly,M.L. (1983) *Science*, **221**, 709–713.
Connolly,M.L. (1985) *Biopolymers*, **25**, 1229–1247.
DesJarlais,R.L., Sheridan,R.P., Dixon,J.S., Kuntz,I.D. and Venkataraghavan,R. (1986) *J. Med. Chem.*, **29**, 2149–2153.
DesJarlais,R., Sheridan,R.P., Seibel,G.L., Dixon,J.S., Kuntz,I.D. and Venkataraghavan,R. (1988) *J. Med. Chem.*, **31**, 722–729.
DesJarlais,R.L., Seibel,G.L., Kuntz,I.D., Montellano,P.O.D., Furth,D.S., Alvarez,J.C., DeCamp,D.L., Babé,L.M. and Craik,C.S. (1990) *Proc. Natl Acad. Sci. USA*, **87**, 6644–6648.
Ferrin,T.E., Huang,C.C., Jarvis,L.E. and Langridge,R. (1988) *J. Mol. Graphics*, **6**, 13–27.
Finer-Moore,J.S., Montfort,W.R. and Stroud,R.M. (1990) *Biochemistry*, **29**, 6977–6986.
Gasteiger,J. and Marsili,M. (1980) *Tetrahedron Lett.*, **36**, 3219–3222.
Gilson,M.K. and Honig,B.H. (1987) *Nature*, **330**, 84–86.
Goodford,P.J. (1984) *J. Med. Chem.*, **27**, 557–564.
Goodford,P.J. (1985) *J. Med. Chem.*, **28**, 849–857.
Goodsell,D.S. and Olson,A.J. (1990) *Proteins*, **8**, 195–202.
Guner,O.F., Hughes,D.W. and Dumont,L.M. (1991) *J. Chem. Inf. Comput. Sci.*, **31**, 408–414.
Jiang,F. and Kim,S.H. (1991) *J. Mol. Biol.*, **201**, 79–102.
Kuhl,F.S., Crippen,G.M. and Friesen,D.K. (1984) *J. Comput. Chem.*, **5**, 24.
Kuntz,I.D. (1992) *Science*, **257**, 1078–1082.
Kuntz,I.D., Blaney,J.M., Oatley,S.J., Langridge,R. and Ferrin,T.E. (1982) *J. Mol. Biol.*, **161**, 269–288.
Lawrence,M.C. and Davis,P.C. (1992) *Proteins*, **12**, 31–41.
Leach,A.R. and Kuntz,I.D. (1992) *J. Comput. Chem.*, **13**, 730–748.
Leach,A.R. and Prout,K. (1990) *J. Comput. Chem.*, **11**, 1193.
Leicester,S.E., Finney,J.L. and Bywater,R. (1988) *J. Mol. Graphics*, **6**, 104–108.
McLachlan,A.D. (1982) *Acta Crystallogr., Sect. A*, **38**, 871–873.

Fig. 7. Stereoimage of the pteridine binding site of DHFR (Bolin *et al.*, 1992), comparing the binding modes of the crystallographic methotrexate and folate with the binding mode of tetrahydropterin from the labeled and unlabeled matching. DHFR residues Asp26 and Ala97, as well as water 256 are shown. The pteridine and pterin heteroatoms are labeled. The dashed lines indicate hydrogen-bond interactions between the protein and the ligand. (**a**) Crystallographic configuration of methotrexate in DHFR (Bolin *et al.*, 1992). (**b**)Crystallographic configuration of folate (Bystroff *et al.*, 1990) in DHFR. (**c**) The labeled matching configuration of tetrahydropterin. (**d**) The unlabeled matching configuration of tetrahydropterin. The ligand makes a closer salt bridge with Asp26 in its unlabeled matching configuration, but it too far from water 256 to hydrogen bond to it. Note that the exocyclic O8 hydroxyl of tetrahydropterin in its unlabeled matching configuration overlaps the N4 of methotrexate rather than the OH4 of folate.

Marquart,M., Walter,J., Deisenhofer,J., Bode,W. and Huber,R. (1983) *Acta Crystallogr., Sect. B*, **39**, 480.
Meng,E.C., Shoichet,B. and Kuntz,I.D. (1992) *J. Comput. Chem.*, **13**, 505–524.
Montfort,W.R., Perry,K.M., Fauman,E.B., Finer-Moore,J.S., Maley,G.F., Hardy,L., Maley,F. and Stroud,R.M. (1990) *Biochemistry*, **29**, 6964–6976.
Rashin,A.A. (1990) *J. Phys. Chem.*, **94**, 1725–1733.
Santi,D.V. and Danenberg,P.V. (1984) In Blakely,R.L. and Benkovic,S.J. (eds), *Folates and Pterins*. Wiley, New York, pp. 345–398.
Shoichet,B. and Kuntz,I.D. (1991) *J. Mol. Biol.*, **221**, 327–346.
Shoichet,B., Bodian,D.L. and Kuntz,I.D. (1992) *J. Comput. Chem.*, **13**, 380–397.
Shoichet,B.K., Stroud,R.M., Santi,D.V., Kuntz,I.D. and Perry,K.M. (1993) *Science*, **259**, 1445–1448.
Smellie,A.S., Crippen,G.M. and Richards,W.G. (1991) *J. Chem. Inf. Comput. Sci.*, **31**, 386–392.
Stoddard,B.L. and Koshland,D.E. (1992) *Nature*, **358**, 774–776.
Straatsma,T.P. and McCammon,J.A. (1991) *Methods Enzymol.*, **202**, 497–511.
Wang,H. (1991) *J. Comput. Chem.*, **12**, 746–750.
Weiner,S.J., Kollman,P.A., Case,D.A., Singh,U.C., Ghio,C., Alagona,G., Profeta,S. and Weiner,P. (1984) *J. Am. Chem. Soc.*, **106**, 765–784.
Wilson,C., Mace,J.E. and Agard,D.A. (1991) *J. Mol. Biol.*, **220**, 495–506.
Wodak,S.J., De Crombrugghe,M. and Janin,J. (1987) *Prog. Biophys. Molec. Biol.*, **49**, 29–63.

Received on January 13, 1993; revised on April 15, 1993; accepted April 30, 1993