J|A|C|S
ARTICLES
Published on Web 03/03/2007

Protein Flexibility and Species Specificity in Structure-Based Drug Discovery: Dihydrofolate Reductase as a Test System

Anna L. Bowman,[†] Michael G. Lerner,[‡] and Heather A. Carlson*[,†,‡]

Contribution from the Department of Medicinal Chemistry and Biophysics Research Division, University of Michigan, Ann Arbor, Michigan 48109-1065

Received November 17, 2006; E-mail: carlsonh@umich.edu

Abstract: In structure-based drug discovery, researchers would like to identify all possible scaffolds for a given target. However, techniques that push the boundaries of chemical space could lead to many false positives or inhibitors that lack specificity for the target. Is it possible to broadly identify the appropriate chemical space for the inhibitors and yet maintain target specificity? To address this question, we have turned to dihydrofolate reductase (DHFR), a well-studied metabolic enzyme of pharmacological relevance. We have extended our multiple protein structure (MPS) method for receptor-based pharmacophore models to use multiple X-ray crystallographic structures. Models were created for DHFR from human and *Pneumocystis carinii*. These models incorporate a fair degree of protein flexibility and are highly selective for known DHFR inhibitors over drug-like non-inhibitors. Despite sharing a highly conserved active site, the pharmacophore models reflect subtle differences between the human and *P. carinii* forms, which identify species-specific, high-affinity inhibitors. We also use structures of DHFR from *Candida albicans* as a counter example. The available crystal structures show little flexibility, and the resulting models give poorer performance in identifying species-specific inhibitors. Therapeutic success for this system may depend on achieving species specificity between the related human host and these key fungal targets. The MPS technique is a promising advance for structure-based drug discovery for DHFR and other proteins of biomedical interest.

Introduction

Candida albicans and *Pneumocystis carinii* are opportunistic fungal pathogens that present a major health problem. The individuals at highest risk for pathogenic infections include those with hematological malignancies, AIDS patients with low CD4+ counts, and transplant recipients. Fungal infections of the blood are an increasingly important cause of morbidity and mortality in immunocompromised patients.[1] *C. albicans* is among the most significant fungal pathogens,[2] and hospital-acquired infections due to *C. albicans* are almost as prevalent as those due to bacteria such as *E. coli*.[3] *P. carinii* is important in the study of *Pneumocystis* pneumonia, one of the major AIDS-defining infections.[4]

Dihydrofolate reductase (DHFR; EC. 1.5.1.3) is a common target for drug action. DHFR is a small enzyme that plays an essential role in various cellular processes including the biosynthesis of DNA. It catalyzes the reduction of dihydrofolate to tetrahydrofolate (and, more slowly, folate to dihydrofolate). These reductions take place via the transfer of a hydride ion from the nicotinamide adenine dinucleotide phosphate (NADPH) cofactor bound to the enzyme.[5] However, problems arise from the use of DHFR as a therapeutic target because the enzyme is found in both microorganisms and humans. To decrease toxic effects in the human host, drugs should selectively target the fungal binding site versus the human.[6]

The active site residues of human DHFR (hDHFR), *P. carinii* DHFR (pcDHFR), and *C. albicans* DHFR (caDHFR) are highly conserved with only three differences, Figure 1: (1) D21 in hDHFR corresponds to S24 in pcDHFR and K24 in caDHFR, (2) F31 in hDHFR corresponds to I33 in the two fungal species, and (3) N64 in hDHFR corresponds to F69 in pcDHFR and F66 in caDHFR. These residues line the edges of the pocket and provide contacts to the "tails" of bound folate molecules. There are differences in the cleft geometry that are apparent in Figure 1. The protein conformation of caDHFR creates a slightly larger active site than that of hDHFR.[7] However, the active site volume of pcDHFR is smaller than that found in hDHFR.[8] Furthermore, studies have shown that distant residues affect catalysis through altered dynamics of the whole protein.[9,10] It

[†] Department of Medicinal Chemistry.
[‡] Biophysics Research Division.

(1) Viudes, A.; Peman, J.; Canton, E.; Ubeda, P.; Lopez-Ribot, J. L.; Gobernado, M. *Eur. J. Clin. Microbiol. Infect. Dis.* **2002**, *21*, 767−774.
(2) Sullivan, D. J.; Moran, G. P.; Pinjon, E.; Al-Mosaid, A.; Stokes, C.; Vaughan, C.; Coleman, D. C. *FEMS Yeast Res.* **2004**, *4*, 369−376.
(3) Edwards, J. E. *N. Engl. J. Med.* **1991**, *324*, 1060−1062.
(4) Morris, A.; Lundgren, J. D.; Masur, H.; Walzer, P. D.; Hanson, D. L.; Frederick, T.; Huang, L.; Beard, C. B.; Kaplan, J. E. *Emerg. Infect. Dis.* **2004**, *10*, 1713−1720.
(5) Schweitzer, B. I.; Dicker, A. P.; Bertino, J. R. *FASEB J.* **1990**, *4*, 2441−2452.
(6) Chan, J. H.; Hong, J. S.; Kuyper, L. F.; Baccanari, D. P.; Joyner, S. S.; Tansik, R. L.; Boytos, C. M.; Rudolph, S. K. *J. Med. Chem.* **1995**, *38*, 3608−3616.
(7) Whitlow, M.; Howard, A. J.; Stewart, D.; Hardman, K. D.; Kuyper, L. F.; Baccanari, D. P.; Fling, M. E.; Tansik, R. L. *J. Biol. Chem.* **1997**, *272*, 30289−30298.
(8) Gschwend, D. A.; Good, A. C.; Kuntz, I. D. *J. Mol. Recognit.* **1996**, *9*, 175−186.

10.1021/ja068256d CCC: $37.00

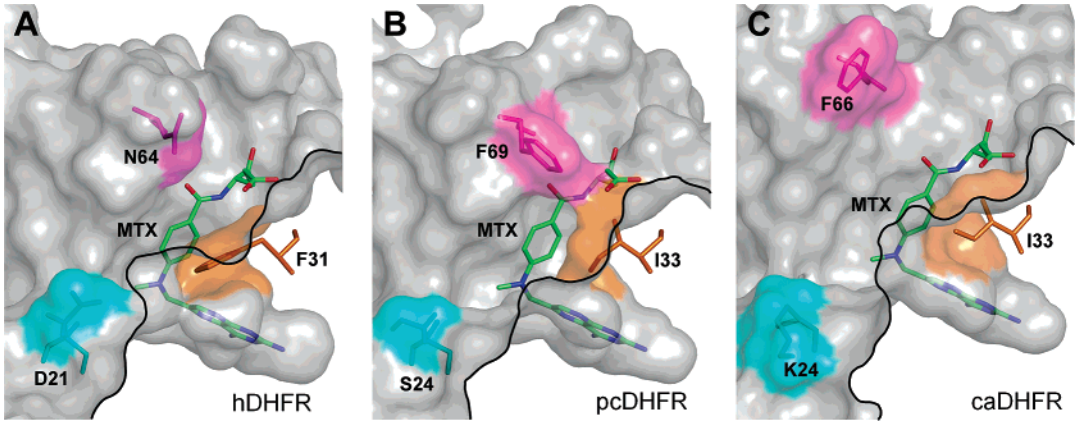

Figure 1. The average surface of the substrate binding pocket is shown in gray. The three residues that are not conserved in the binding site are labeled and shown in stick formation; their contribution to the protein surface is colored accordingly: (a) hDHFR, (b) pcDHFR, and (c) caDHFR. The black line denotes where each protein surface is cut away to reveal the penetration of the binding pocket into the interior of the protein. Methotrexate (MTX, colored by atom with carbons in green) has been manually added to the average structure to aid the viewer in interpreting the potential interactions with substrates.

may be possible for inhibitors to take advantage of these differences in size and conformational sampling. A disparity in activities indicates subtle binding specificities in the active site of the enzymes.[6,11] This implies that there are exploitable differences within the active site that could be probed to develop potent and selective inhibitors.

Currently, the drug of choice for *Pneumocystis* pneumonia is trimethoprim, which is selective for pcDHFR[12] but has only moderate affinity (IC_{50} value $\sim$20 μM).[13] Because of this low potency, trimethoprim must be given in combination with sulfa drugs like sulfamethoxazole or dapsone.[14] However, sulfa drugs cause allergic and toxic side effects in up to 54% of patients, and treatment must be altered or discontinued.[15] A more potent inhibitor, selective for pcDHFR over hDHFR, would be a valuable lead for a drug that is effective without the co-administration of a sulfa drug.[16]

The main drug for treating *C. albicans* infections is Amphotericin B; however, it can cause severe side effects.[17] DHFR is not the target of Amphotericin B, but recent research has discovered inhibitors with significant selectivity for caDHFR over hDHFR.[6] This is a promising avenue for new pharmacological treatment.

Here, we present a computational study of species specificity for inhibitors of DHFR that focuses on incorporating protein flexibility into structure-based drug discovery. Ignoring protein flexibility by using a "static" structure (a single, rigid conformation of a protein) can considerably reduce the computational expense of virtual screening or de novo design, but it can limit the resulting inhibitors to a minute fraction of the appropriate chemical space that could complement that receptor.[18] Our technique uses multiple protein structures (MPS) to represent an ensemble of conformational states of the receptor.[19−21] Our MPS method maps out complementary interactions with the binding sites of each protein conformation. Regions of rigidity and flexibility are identified by the conserved regions where the same complementary interactions are consistently made with most of the MPS.

The method, initially applied to HIV integrase,[20] has seen considerable success with HIV-1 protease.[21,22] However, our previous studies have focused on using molecular dynamics (MD) simulations to generate the ensemble of protein conformations for MPS studies. MD simulations are appropriate for this use, but they are time-consuming. Because of this, they are not typically used during drug discovery in the pharmaceutical industry. Here, we further extend our method to use a collection of crystal structures. By using structures from the PDB (Protein Data Bank),[23] the method will be faster to apply and more accessible to scientists.

Incorporating protein flexibility allows structure-based drug design to expand the conformational and chemical space of the hit list. Of course, a more broad description of the chemical space could identify weaker, more promiscuous inhibitors. For the case of DHFR, a loss of species specificity is detrimental. In this study, we show that the inclusion of flexibility does not lose specificity between hDHFR versus pcDHFR inhibitors. The MPS models are highly selective for known DHFR inhibitors over drug-like non-inhibitors, and, furthermore, they display species specificity in these highly conserved systems. Conversely, we show that the limited conformational flexibility in caDHFR structures leads to moderate performance and a loss of species specificity.

(9) Wong, K. F.; Selzer, T.; Benkovic, S. J.; Hammes-Schiffer, S. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6807−6812.
(10) Rajagopalan, P. T. R.; Lutz, S.; Benkovic, S. J. *Biochemistry* **2002**, *41*, 12618−12628.
(11) Gschwend, D. A.; Sirawaraporn, W.; Santi, D. V.; Kuntz, I. D. *Proteins: Struct., Funct., Genet.* **1997**, *29*, 59−67.
(12) Baccanari, D. P.; Daluge, S.; King, R. W. *Biochemistry* **1982**, *21*, 5068−5075.
(13) Delves, C. J.; Ballantine, S. P.; Tansik, R. L.; Baccanari, D. P.; Stammers, D. K. *Protein Expression Purif.* **1993**, *4*, 16−23.
(14) Medina, I.; Mills, J.; Leoung, G.; Hopewell, P. C.; Lee, B.; Modin, G.; Benowitz, N.; Wofsy, C. B. *N. Engl. J. Med.* **1990**, *323*, 776−782.
(15) Lee, S. A. *J. Pharm. Pract.* **2006**, *19*, 5−9.
(16) Rosowsky, A.; Forsch, R. A.; Queener, S. F. *J. Med. Chem.* **2002**, *45*, 233−241.
(17) Sternberg, S. *Science* **1994**, *266*, 1632−1634.
(18) Carlson, H. A.; McCammon, J. A. *Mol. Pharmacol.* **2000**, *57*, 213−218.
(19) Carlson, H. A.; Masukawa, K. M.; McCammon, J. A. *J. Phys. Chem. A* **1999**, *103*, 10213−10219.
(20) Carlson, H. A.; Masukawa, K. M.; Rubins, K.; Bushman, F. D.; Jorgensen, W. L.; Lins, R. D.; Briggs, J. M.; McCammon, J. A. *J. Med. Chem.* **2000**, *43*, 2100−2114.
(21) Meagher, K. L.; Carlson, H. A. *J. Am. Chem. Soc.* **2004**, *126*, 13276−13281.
(22) Meagher, K. L.; Lerner, M. G.; Carlson, H. A. *J. Med. Chem.* **2006**, *49*, 3478−3484.
(23) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. *Nucleic Acids Res.* **2000**, *28*, 235−242; http://www.rcsb.org/pdb/.

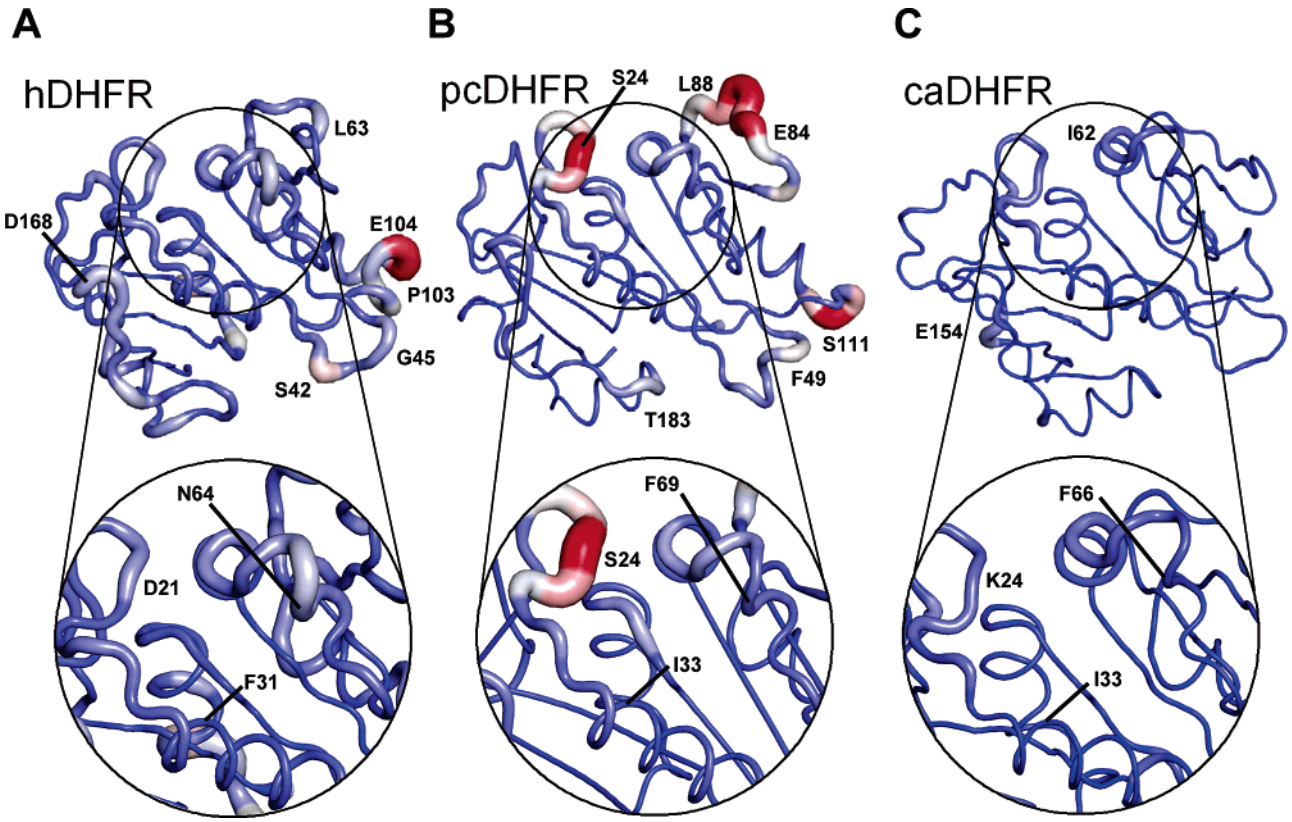

Figure 2. Average backbone structure for (a) hDHFR, (b) pcDHFR, and (c) caDHFR. A red, thicker tube indicates greater rmsd across the ensemble, whereas a blue, narrow tube shows limited flexibility. Residues with the greatest backbone flexibility are labeled in each model. The circle indicates the substrate-binding site region, and a close-up of the region is shown below. Active-site residues that are not conserved between species are labeled.

Methods

Protein Preparation. For each species, X-ray crystallographic structures of the wild-type protein with unique ligands bound were downloaded from the PDB.[23] If, within a species class, a structure with an identical ligand existed, the structure with the better resolution was used. This resulted in seven structures for hDHFR (1HFR,[24] 1OHJ,[25] 1PD8,[26] 1S3W,[27] 1U72,[28] 1KMS,[29] and 1KMV[29]), eight structures for caDHFR (1AOE,[7] 1M78,[7] 1M79,[7] 1M7A,[7] 1IA1,[30] 1IA2,[30] 1IA3,[30] and 1IA4[30]), and nine structures for pcDHFR (1DAJ,[31] 1E26,[32] 2CD2,[32] 3CD2,[32] 1LY3,[33] 1LY4,[33] 1S3Y,[34] 1VJ3,[35] and 1DYR[36]).

All ligands and solvent were removed from each structure; however, the cofactor was retained. Histidine residues were visually inspected to assign the appropriate protonation state. Hydrogen atoms were placed using the LEaP routine[37] in AMBER[38] and optimized to convergence with conjugate gradient energy minimization (convergence criterion: rms gradient <0.0001 kcal/mol Å). A distance-dependent dielectric was used. The force field parameters for the cofactor were provided by Ryde et al.[39]

Structural Comparison. In this work, we utilize available X-ray crystallographic structures to provide the MPS needed for our method. Although using crystal structures is faster, conformational sampling may be more limited. For the hDHFR set, the average Cα rmsd was 0.32 Å with a range of 0.21−0.51 Å. There was a slightly greater degree of variation among the *P. carinii* collection with an average rmsd of 0.34 Å and a range between 0.21 and 0.53 Å. Greater flexibility is seen for the side chains of residues in the binding site; the average rmsd values for side chains within the binding were 1.58 Å for hDHFR and 1.69 Å for pcDHFR. However, there was very little variety among the *C. albicans* ensemble; the average Cα rmsd was 0.13 Å with a range of 0.08−0.16 Å. More significantly, conformational variation in the binding-site side chains was low, with an average rmsd of 1.01 Å. Figure 2 shows that the majority of the conformational variation occurs in the loop regions of the protein. This is analogous to results seen during a 10-ns MD simulation where the loop regions of DHFR were seen to undergo large conformational changes, while the core exhibited more restrained conformational sampling.[40]

(24) Cody, V.; Galitsky, N.; Luft, J. R.; Pangborn, W.; Blakley, R. L.; Gangjee, A. *Anti-Cancer Drug Des.* **1998**, *13*, 307−315.
(25) Cody, V.; Galitsky, N.; Luft, J. R.; Pangborn, W.; Rosowsky, A.; Blakley, R. L. *Biochemistry* **1997**, *36*, 13897−13903.
(26) Cody, V.; Luft, J. R.; Pangborn, W.; Gangjee, A. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2003**, *59*, 1603−1609.
(27) Cody, V.; Luft, J. R.; Pangborn, W.; Gangjee, A.; Queener, S. F. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2004**, *60*, 646−655.
(28) Cody, V.; Luft, J. R.; Pangborn, W. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2005**, *61*, 147−155.
(29) Klon, A. E.; Heroux, A.; Ross, L. J.; Pathak, V.; Johnson, C. A.; Piper, J. R.; Borhani, D. W. *J. Mol. Biol.* **2002**, *320*, 677−693.
(30) Whitlow, M.; Howard, A. J.; Stewart, D.; Hardman, K. D.; Chan, J. H.; Baccanari, D. P.; Tansik, R. L.; Hong, J. S.; Kuyper, L. F. *J. Med. Chem.* **2001**, *44*, 2928−2932.
(31) Cody, V.; Galitsky, N.; Luft, J. R.; Pangborn, W.; Gangjee, A.; Devraj, R.; Queener, S. F.; Blakley, R. L. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **1997**, *53*, 638−649.
(32) Cody, V.; Galitsky, N.; Rak, D.; Luft, J. R.; Pangborn, W.; Queener, S. F. *Biochemistry* **1999**, *38*, 4303−4312.
(33) Cody, V.; Galitsky, N.; Luft, J. R.; Pangborn, W.; Queener, S. F.; Gangjee, A. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2002**, *58*, 1393−1399.
(34) Cody, V.; Luft, J. R.; Pangborn, W.; Gangjee, A.; Queener, S. F. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2004**, *60*, 646−655.
(35) Cody, V.; Chan, D.; Galitsky, N.; Rak, D.; Luft, J. R.; Pangborn, W.; Queener, S. F.; Laughton, C. A.; Stevens, M. F. *Biochemistry* **2000**, *39*, 3556−3564.
(36) Champness, J. N.; Achari, A.; Ballantine, S. P.; Bryant, P. K.; Delves, C. J.; Stammers, D. K. *Structure* **1994**, *2*, 915−924.
(37) Pearlman, D. A.; Case, D. A.; Caldwell, J. W.; Ross, W. S.; Cheatham, T. E., III; Debolt, S.; Ferguson, D. M.; Seibel, G. L.; Kollman, P. A. *Comput. Phys. Commun.* **1995**, *91*, 1−41.
(38) Case, D. A.; et al. *AMBER 6*; University of California, San Francisco, 1999.
(39) Holmberg, N.; Ryde, U.; Bulow, L. *Protein Eng.* **1999**, *12*, 851−856.
(40) Radkiewicz, J. L.; Brooks, C. L., III. *J. Am. Chem. Soc.* **2000**, *122*, 225−231.

Probe Flooding, Minimization, and Clustering. The binding site of each structure was flooded with 1000 small molecule probes using a 12-Å sphere.[41] The probes used were benzene (to identify aromatic and hydrophobic interactions), ethane (to distinguish hydrophobic interactions from aromatic), and methanol (to identify hydrogen-bonding interactions). The flooding sphere was centered at the bottom of the binding pocket. The probes were then optimized through a low-temperature Monte Carlo minimization, the multi-unit search for interacting conformers (MUSIC) routine in BOSS,[42] using the OPLS force field.[43] While the protein is held fixed, the probes undergo multiple, simultaneous gas-phase minimizations and cluster into local minima, which define complementary binding regions.[20] An automated procedure, based on Jarvis−Patrick methodology, was used to systematically define the clustered groups of probe molecules.[41] Each cluster was then represented by the lowest-energy probe in the group, termed its "parent".

The structures were divided into sets by their species of origin. Each set of structures was aligned using a Gaussian-weighted rmsd alignment[44] to the structure in that set with the highest resolution to obtain a common frame of reference. F34 in hDHFR (F36 in pcDHFR and caDHFR) defines the center of the binding pocket. For each independent set of DHFR structures, the parent probes from each structure within 8 Å of Cγ of the corresponding central phenylalanine were combined and clustered to give "consensus clusters". A consensus cluster must contain parents from $\geq$50% of the protein conformations. Each consensus cluster was represented as a spherical pharmacophore element. The center of each element was defined by the average position of the consensus cluster, and the radius was given by the rmsd of the consensus cluster.[21] The conserved regions are favorable across multiple conformations, which is an entropic benefit. No limitations or requirements are made in the flexible region, which allows us to identify many potential ligands with a broad range of sizes, shapes, and chemical scaffolds. To avoid the ligands severely clashing with the small conserved steric regions of the pocket, four excluded volume elements were added to complete the pharmacophore model. These had a radius of 1.5 Å and were centered at the average position of I7 C, A9 Cα, F34 Cγ, and I60 Cα for hDHFR; I10 C, A12 Cα, F36 Cγ, and I65 Cα for pcDHFR; and I9 C, A11 Cα, F36 Cγ, and I62 Cα for caDHFR.

Creation of the Ligand Databases. Each model was screened against three databases: one set of highly specific inhibitors, one set of weaker, less specific inhibitors, and a general set of drug-like non-inhibitors using MOE.[45] There were 493 known hDHFR, pcDHFR, and caDHFR inhibitors in total, each with an IC_{50} value below 1 μM. The full set of structures and affinity data are provided with references in the Supporting Information. hDHFR inhibitors were taken from the literature and from the MDL drug data report.[46] Selected pcDHFR inhibitors were taken from the Sutherland et al. dataset,[47] and caDHFR inhibitors were taken from the literature. A subset of high-affinity inhibitors for each species was compiled. For hDHFR, this comprised the top 50 inhibitors; the IC_{50} values ranged from 0.1 to 4.1 nM. The top 50 pcDHFR inhibitors were selected to be the *P. carinii* high-affinity dataset, with IC_{50} values ranging from 0.035 to 43 nM. However, it was not possible to select 50 equivalently potent caDHFR inhibitors, as there were comparatively few structures in the literature. For this reason, compounds with an IC_{50} value lower than 50 nM were chosen, giving a dataset of 19 high-affinity caDHFR inhibitors. Each species has a slightly different set of less potent DHFR inhibitors where the corresponding high-affinity inhibitors are removed from the full set of 493 DHFR inhibitors. The sizes of these sets were 474 for caDHFR and 443 for both hDHFR and pcDHFR (the compositions were appropriately different because the specific inhibitors are different for each system). An additional set of drug-like molecules was obtained from the CMC.[48] These were filtered to require one hydrogen-bond donor and one aromatic atom. Molecules that had molecular masses less than 100 were removed; this gave a total of 2326 compounds. To produce a dataset of drug-like non-inhibitors of DHFR, a substructure search against 2,4-pyrimidinediamine was performed at 96% identity using the Tanimoto superset/subset similarity metric in MOE.[45] This identified 23 folate-like molecules, which were known to inhibit or were suspected to inhibit DHFR; these included compounds such as methotrexate and trimetrexate. Removal of these compounds yielded a database of 2303 drug-like non-inhibitor compounds. Multiple conformations of each compound were generated using rule-based torsion driving in OMEGA,[49] using an energy cutoff of 14 kcal/mol and a heavy-atom rmsd criterion of 1 Å.

Evaluation of Pharmacophore Models. The predicative quality of each pharmacophore model was evaluated by comparing the percentage of true inhibitors (true positives) identified on the *y*-axis to its percentage of drug-like non-inhibitors found (false positives) on the *x*-axis, on a receiver operator characteristic (ROC) curve. The ideal model would be at the point (0,100) with all of the true positives identified with no false positives found, whereas a random predictor would lie on a line of slope equal to 1.

The number of elements required was varied, for example, requiring seven, six, or five elements from a seven-site pharmacophore model. The radius of each pharmacophore element, given by the rmsd of the consensus cluster, was increased by a multiplication factor of 1−7 in $^1/_3$ increments. In previous studies of HIV-1 protease, the rmsd values for the consensus clusters were in the range of 0.8−1.6 Å, and the optimal models had an element radii multiplication factor of ~2, which gave the radii in the range 1.6−3.2 Å.[21,22] For the current work, the smaller variation in the crystal structures resulted in small, specific pharmacophore elements: 0.2−1.0 Å for hDHFR, 0.3−0.8 Å for pcDHFR, and 0.1−0.5 Å for caDHFR. These small elements required a wider range of multiplication factors to produce reasonable models. The optimal models had radii of 4×rmsd for hDHFR, 3×rmsd for pcDHFR, and $5^1/_3$×rmsd for caDHFR, which led to radii in the range of 0.6−4.1 Å. Models with the largest radii typically perform poorly and are just used to show the limits of the technique.

Results and Discussion

Performance of the MPS Pharmacophore Models. Each model was screened against three databases: one set of potent, species-specific inhibitors for the target DHFR, one set of inhibitors with weaker affinities, and one broad set of 2303 drug-like non-inhibitors. The performance of the pharmacophore models for all three species is given in Figure 3. Compounds identified by the optimal models are listed in the Supporting Information.

As one would expect, the ROC plots show that the models preferentially hit strong and weak inhibitors over non-inhibitors. *What is most striking is that the models for hDHFR and pcDHFR, the models with the most protein flexibility, preferentially hit species-specific inhibitors over other general DHFR inhibitors.* It is a concern that more general models, which incorporate protein flexibility, could lose specificity information and identify weak, promiscuous inhibitors. While this may be

(41) Lerner, M. G.; Meagher, K. L.; Carlson, H. A. Unpublished results.
(42) Jorgensen, W. L. *BOSS*, Version 4.2; Yale University, New Haven, CT, 2000.
(43) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1996**, *118*, 11225−11236.
(44) Damm, K. L.; Carlson, H. A. *Biophys. J.* **2006**, *90*, 4558−4573.
(45) *MOE*; Chemical Computing Group: Montreal, Canada.
(46) *MDDR Drug Data Report*; MDL Information Systems Inc.: San Leandro, CA, 2005.
(47) Sutherland, J.; O'Brien, L.; Weaver, D. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1906−1915.
(48) *Comprehensive Medicinal Chemistry*; MDL Information Systems Inc.: San Leandro, CA, 2005.
(49) *OMEGA*, Version 1.8.b1; OpenEye Scientific Software, Inc.: Santa Fe, NM, 2004; www.eyesopen.com.

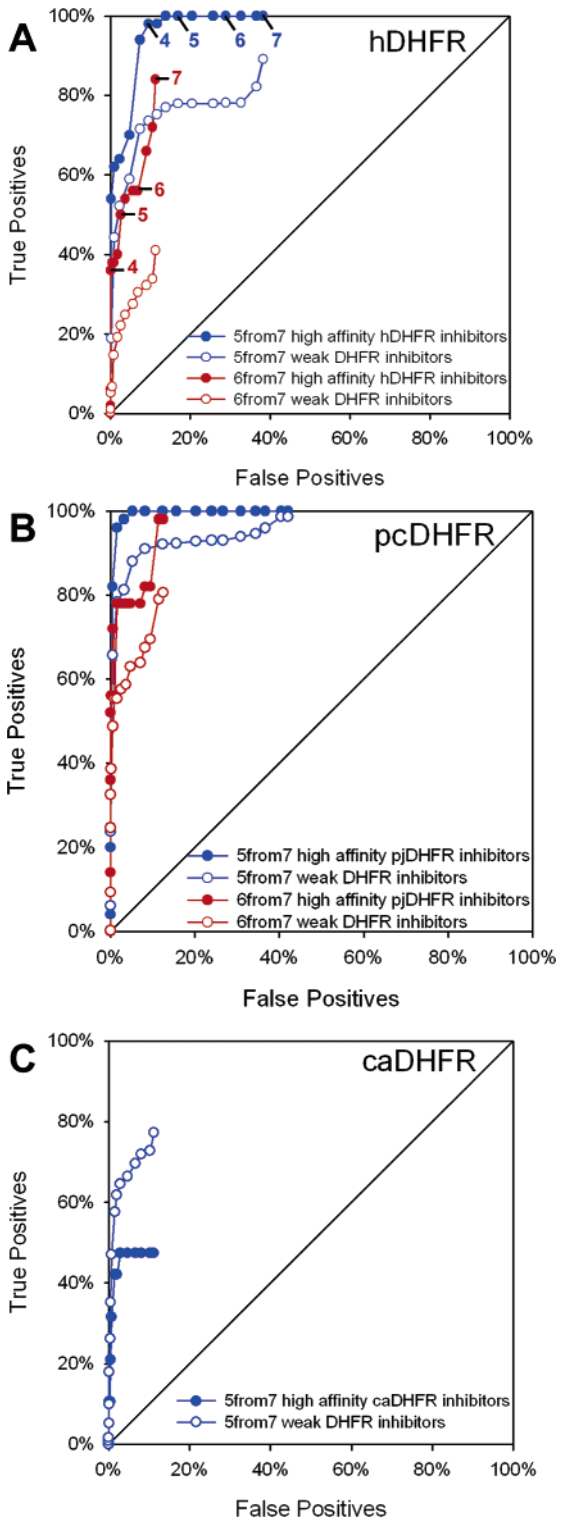

Figure 3. The ROC curves show that strong, specific inhibitors are preferentially identified in virtual database screening results: (a) hDHFR models and (b) pcDHFR models. The limited protein flexibility in (c) the caDHFR models does not capture the species specificity of the system, but strong and weak inhibitors are still preferentially identified over non-inhibitors. Series with filled data points are results from the screening of the related high-affinity database; those with open data points are results from the screening of weaker known DHFR inhibitors. Shown in red are results from models requiring 6 hits from a 7-site pharmacophore model; shown in blue are results from models requiring 5 hits from a 7-site pharmacophore model. Points along each series represent an increase in pharmacophore element radii of 1×rmsd to 7×rmsd. To aid the reader in interpreting the graphs, these scaling factors are labeled in (a) for the plots of hitting high-affinity inhibitors of hDHFR over hitting false positives.

an understandable limitation for a drug discovery technique, it would greatly weaken the applicability of that method. It appears that our MPS technique does not have this limitation!

The MPS models broadly describe the proper chemical space, and yet retain the focus to preferentially identify the most potent and specific inhibitors. Figure 3 shows that high-affinity, specific inhibitors are identified over weaker inhibitors, which are identified over non-inhibitors. This shows that our technique explores the most appropriate chemical space as it pushes structure-based drug design into a broader description of inhibitor space. In fact, the models for caDHFR show significantly less flexibility and show more modest performance. The caDHFR model does identify inhibitors over non-inhibitors, but the species-specific bias is lost.

Pharmacophore Model Comparison. Figure 4 compares the MPS pharmacophore models across the three systems. The pharmacophore elements describe comparable interactions with the protein across the three species, and close resemblance between the models is to be expected due to the conservation of the binding site residues. However, the variation in location of the elements is the result of subtle differences in the conformations of the binding sites, and the different radii reflect the varying degree of conformational flexibility. For instance, the relatively reduced protein flexibility found in the caDHFR ensemble led to much smaller RMSDs for many clusters. The smallest radius of an element in the caDHFR model was 0.1 Å, whereas the smallest radii of the hDHFR and pcDHFR models were 2−3 times larger. These differences should describe the species specificity of each system and lead to selective pharmacophore models.

Each model consists of four excluded volumes and seven pharmacophore elements, six of which are comparable in position and type. The features can be compared back to known motifs of folate ligands. The aromatic features of the pterin rings are consistently represented by an aromatic element at the bottom of the cleft. This tight aromatic element is the result of a highly consistent interaction between benzene probes and F34/F36. The hydrophilic features of the pterin ring are represented by three hydrogen-bond donors with relatively consistent spacing between the excluded volumes representing I7, A9, and F34 (hDHFR numbering). These hydrogen-bond donor elements reflect the consistent interaction of methanol probes with O7 of NADPH and with Oϵ1 and Oϵ2 of E30 hDHFR (E32 caDHFR and pcDHFR).

The models of caDHFR and hDHFR also display a fourth donor element. This reflects an interaction with the backbone oxygen of V115 in hDHFR (I112 in caDHFR). No similar interaction was found for pcDHFR. This may be a result of the slightly rotated orientation of I123, the corresponding residue in pcDHFR, which appears to sterically block interaction with the backbone oxygen from the active site.

In the three models, there is an aromatic element midway in the binding pocket reflecting the interaction between benzene probes and the cofactor (the more general description of aromatic/hydrophobic is seen in hDHFR). Additionally, one hydrogen-bond acceptor is always found on the far side of the excluded volumes of the central phenylalanines in the active site. This element reflects an interaction with R70 hDHFR (R72 caDHFR and R75 pcDHFR) and models the start of the tail region. The location and radius directly reflect the steric

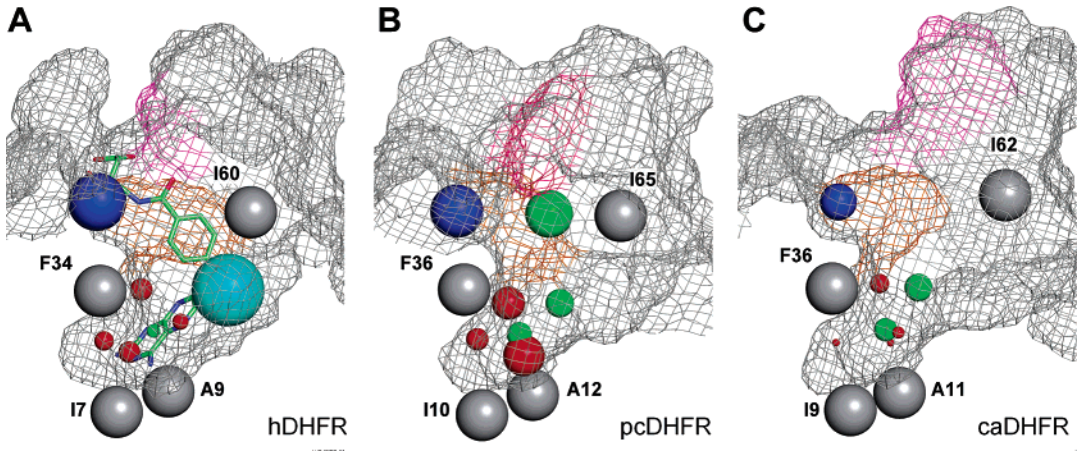

Figure 4. An alternate orientation of the binding pocket is shown that highlights the similarities in the pharmacophore models of (a) hDHFR, (b) pcDHFR, and (c) caDHFR. For clarity, elements of all models are shown with radii of 2×rmsd: green spheres map aromatic moieties, the cyan sphere requires aromatic or hydrophobic groups, blue is a hydrogen-bond acceptor, and red is a hydrogen-bond donor. The gray spheres are excluded volumes, which represent key amino acids that line the pocket. To allow the reader to compare back to Figure 1, the molecular surface for the average structure of each species is shown in gray mesh with the same coloration of nonconserved residues (the mesh does not represent any feature of the pharmacophore models). Also, methotrexate is shown in hDHFR in the same orientation and location as seen in Figure 1.

influence of nearby hydrophobic residues that are not conserved across the species (colored mesh in Figure 4 that corresponds to the same colored surfaces in Figure 1). The close spacing and consistent nature of those unique hydrophobic residues in pcDHFR result in the extra aromatic element, seen in Figure 4B between the hydrogen-bond donor and the excluded volume for I65 in pcDHFR. It should be noted that conformational variations in the twist of F69 in pcDHFR are implicitly accommodated by the relatively large radius of the extra aromatic element. Furthermore, the element does not require a specific orientation of an aromatic group in that location. It simply reflects the favorable interaction created by the smaller and more ordered nature of this portion of the site in pcDHFR. The resulting aromatic element corresponds to the *p*-aminobenzoate linker region of the substrate and known inhibitors. Our models for hDHFR and caDHFR do not restrict the identification of ligands based on this region, so an aromatic feature is perfectly allowed, just not required. In hDHFR, the corresponding residue is N64, which cannot make an aromatic interaction. In caDHFR, the equivalent residue is F66, but the phenyl ring is oriented away from the pocket, contributing to the wider binding site[7] (Figure 1C). Any benzene probes that interacted with F66 in caDHFR fell outside the 8-Å cutoff for the binding site description.

Requiring a molecule to complement all seven sites of any model proved to be too stringent, and inadequate numbers of inhibitors and decoy compounds were identified. This implies that the models are too restrictive, probably due to the limited conformational variance in the structures. For hDHFR and pcDHFR, the models requiring six from seven sites performed well; however, the caDHFR model again had inadequate numbers of hits from any database. This again reflects the limited flexibility in the caDHFR structures as compared to the hDHFR and pcDHFR ensembles. When the criteria were loosened to require only five from the seven sites, all of the models were highly sensitive and highly specific for DHFR inhibitors over drug-like non-inhibitors, as shown in Figure 3. Exceptional performance is seen for hDHFR and pcDHFR; screening with the five-from-seven models consistently identified over 60% of the true positives while hitting less than 2% of the false positives. The performance of the caDHFR models was acceptable, but we address this issue in detail further below. A detailed presentation of true and false positives is included in the Supporting Information.

The optimal hDHFR pharmacophore model (5 from 7, radii = 4×rmsd) identified 98.0% of the high-affinity hDHFR inhibitors and 73.6% of the less potent DHFR inhibitors with 9.2% of the false positives being hit. The optimal pcDHFR model (5 from 7, radii = 3×rmsd) also identified 98.0% of the high-affinity pcDHFR inhibitors. Again, high percentages (81.3%) of the less potent DHFR inhibitors were also identified, and even fewer of the drug-like non-inhibitors were hit (2.9%). It is outstanding that the pharmacophore models for hDHFR and pcDHFR have higher selectivity for their related high-affinity compounds over other DHFR inhibitors. This shows that a degree of species differentiation has been incorporated into these models through the slight variations in functional elements, the location of those elements, and their radii.

The optimal caDHFR pharmacophore model (5 from 7, radii = $5^1/_3$×rmsd) identified 47.4% of the high-affinity inhibitors and 64.6% of the less potent DHFR inhibitors; only 2.7% of the false positives were hit. Although the ability of this model to identify true inhibitors is still quite good, it does not have the same selectivity for the high-affinity compounds that was seen with the hDHFR and pcDHFR pharmacophore models. In previous MPS studies, we have found that more conformational sampling, through longer molecular dynamics simulations, improves performance of the resulting pharmacophore models.[21] The low level of conformational variance seen in the caDHFR structures may restrict the pharmacophore model to closely resemble a static pharmacophore model; certainly, the very small radii (low RMSDs) of the elements are characteristic of models produced from single, rigid structures.[21]

Examination of the 19 inhibitors creating the caDHFR high-affinity dataset (shown in the Supporting Information) revealed that none of the inhibitors had a tail analogous to the glutamate tail of folate. This implies that the acceptor element may not be a crucial part of developing DHFR inhibitors that are selective

for *C. albicans*.[6,50] However, the limited amount of flexibility between each structure in the caDHFR ensemble has enabled this cluster to be present. If the structures represented wider conformational sampling, this site would be more flexible and possibly not conserved. Extraneous sites are another characteristic of static pharmacophore models;[21] for example, a receptor-based pharmacophore model from one static crystal structure of *Lactobacillus casei* DHFR[51] contained 67 site points.[52,53] Some were in good agreement with known inhibitors, but identifying the most important points a priori can be difficult.

The diminished flexibility across the caDHFR structures and the lack of known high-affinity inhibitors make it harder to achieve conclusive results for caDHFR. Additionally, it appears caDHFR has two distinct binding modes: one that is analogous to the binding mode of pcDHFR and hDHFR and another in which more selective inhibitors displaced the NADPH cofactor.[30] Two of the eight X-ray crystallographic structures used in the MPS method displayed the second, more unusual binding mode. As only features that are conserved in ≥50% of protein conformations are preserved, the NADPH displacement binding mode would not be represented in the pharmacophore model.

Conclusion

Using DHFR as a test case, this work has demonstrated that, while expanding chemical space through inclusion of protein flexibility, it is still possible to retain target specificity with our MPS pharmacophore technique. Broadly describing the chemical space of complementary ligands could lead to the identification of weak, promiscuous inhibitors, but that is not the case with the MPS method. Receptor-based pharmacophore models from X-ray crystal structures of hDHFR and pcDHFR were highly selective and identified relevant species-specific, high-affinity inhibitors over other general DHFR inhibitors. The reduced selectivity of the caDHFR model strongly implies that the inclusion of more flexibility improves our pharmacophore models. The species specificity of the model was lost, but the model still retained excellent specificity for DHFR inhibitors over drug-like non-inhibitors. The fact that subtle differences in the binding site can be elucidated by using the MPS method and characterized by a pharmacophore model shows the versatility of the technique. These models may lead to the development of novel, species-specific DHFR inhibitors.

Furthermore, we have extended our technique to use ensembles of protein structures from crystallography. For MPS models based on snapshots from an MD simulation, we previously suggested models use $n - 1$ from n elements and that the radii of the elements be set to 2.3×rmsd.[22] In this study, the more limited conformational sampling requires more relaxed parameters to identify inhibitors. If 6 from 7 elements are used (in keeping with our suggestions from MD models), very large radii are needed (7×rmsd or more). Instead, optimal performance was seen for models requiring 5 from 7 elements and radii of 3×rmsd for pcDHFR and 4×rmsd for hDHFR. Only one less element is needed ($n - 2$ from n), and the moderate increase in radii is reasonable. If collections of crystal structures can be used effectively in the MPS method by these small adjustments, this will make the technique more accessible to scientists who are not experts in dynamic simulations. Also, MD simulations are prohibitively long to be used in the pharmaceutical industry. The extension to use collections of crystal structures makes the MPS method more applicable in the industry's search for new drug leads.

Acknowledgment. This work has been supported by the National Institutes of Health (Grant GM65372) and the Beckman Young Investigator Program. M.G.L. would like to thank the University of Michigan's Molecular Biophysics Training Program for support (NIGMS grant GM008270). We thank OpenEye for their generous donation of the OMEGA software and Allen Bailey for maintaining the computers used in this work.

Note Added in Proof. A predictive, species-specific, docking model of DHFR from *Cryptosporidium hominus* and *Toxoplasma gondii* was recently reported. The study showed that the inclusion of protein flexibility improved the docking scores. Popov, V. M.; Yee, W. A.; Anderson, A. C. *Proteins* **2007**, *66*, 375−387.

Supporting Information Available: Complete ref 38, the data sets of DHFR inhibitors used in this work, the coordinates and rmsd of the pharmacophore elements for all models, lists of true positives identified by each species' optimal model, and examples of false negatives. This material is available free of charge via the Internet at http://pubs.acs.org.

JA068256D

(50) Otzen, T.; Wempe, E. G.; Kunz, B.; Bartels, R.; Lehwark-Yvetot, G.; Hänsel, W.; Schaper, K.-J.; Seydel, J. K. *J. Med. Chem.* **2004**, *47*, 240−253.
(51) Bolin, J. T.; Filman, D. J.; Matthews, D. A.; Hamlin, R. C.; Kraut, J. *J. Biol. Chem.* **1982**, *257*, 13650−13662.
(52) Joseph-McCarthy, D.; Alvarez, J. C. *Proteins: Struct., Funct., Genet.* **2003**, *51*, 189−202.
(53) Joseph-McCarthy, D.; Thomas, B. E.; Belmarsh, M.; Moustakas, D.; Alvarez, J. C. *Proteins: Struct., Funct., Genet.* **2003**, *51*, 172−188.