PROTEINS: Structure, Function, and Genetics 42:279–293 (2001)

Docking Molecules by Families to Increase the Diversity of Hits in Database Screens: Computational Strategy and Experimental Evaluation

Andrew I. Su,[1] David M. Lorber,[1] G. Scott Weston,[1] Walter A. Baase,[2] Brian W. Matthews,[2] and Brian K. Shoichet[1]*
[1]*Department of Molecular Pharmacology & Biological Chemistry, Northwestern University, Chicago, Illinois*
[2]*Institute of Molecular Biology and Howard Hughes Medical Institute, University of Oregon, Eugene, Oregon*

ABSTRACT **Molecular docking programs screen chemical databases for novel ligands that fit protein binding sites. When one compound fits the site well, close analogs typically do the same. Therefore, many of the compounds that are found in such screens resemble one another. This reduces the variety and novelty of the compounds suggested. In an attempt to increase the diversity of docking hit lists, the Available Chemicals Directory was grouped into families of related structures. All members of every family were docked and scored, but only the best scoring molecule of a high-ranking family was allowed in the hit list. The identity and scores of the other members of these families were recorded as annotations to the best family member, but they were not independently ranked. This family-based docking method was compared with molecule-by-molecule docking in screens against the structures of thymidylate synthase, dihydrofolate reductase (DHFR), and the cavity site of the mutant T4 lysozyme Leu99 → Ala (L99A). In each case, the diversity of the hit list increased, and more families of known ligands were found. To investigate whether the newly identified hits were sensible, we tested representative examples experimentally for binding to L99A and DHFR. Of the six compounds tested against L99A, five bound to the internal cavity. Of the seven compounds tested against DHFR, six inhibited the enzyme with apparent K_i values between 0.26 and 100 μM. The segregation of potential ligands into families of related molecules is a simple technique to increase the diversity of candidates suggested by database screens. The general approach should be applicable to most docking methods. Proteins 2001;42:279–293.** © 2000 Wiley-Liss, Inc.

Key words: database; screening; diversity; docking; inhibitor discovery; inhibitor design

INTRODUCTION

Molecular docking fits molecules together in favorable configurations.[1–8] Docking programs have predicted the structures of protein–ligand complexes *de novo*[9–13] and discovered novel ligands for proteins of known structure.[10,14–16] In the search for novel ligands, molecular databases are screened for compounds that chemically and sterically complement a binding site. A small number, usually less than 1%, of the best fitting molecules from the database are retained for detailed evaluation. We refer to these as *docking hits*. Some of these docking hits are tested experimentally. The underlying idea is that screening against a protein structure can lead to the discovery of novel chemical scaffolds, dissimilar to known substrates or inhibitors.[17,18]

Ideally, one would like a diverse set of new leads from docking screens. However, docking hit lists are often crowded with molecules that resemble one another. This is because most database molecules have analogs within that database. If a molecule fits a binding site well, its analogs are also likely to fit well. Because the number of docking hits is necessarily small compared with the number of database molecules, having many analogs in a hit list means that disparate, frequently interesting molecules will be absent from that list. This reduces the diversity of the hits, which is a drawback when a major motivation for docking is the discovery of novel ligands.

The problem of increasing diversity in docking calculations resembles that faced by all screening methods, whether computational or experimental. One would like to sample chemical space broadly, discovering as many classes of novel ligands as possible. Methods to do so in experimental screens have received considerable attention.[19–23] A solution often adopted in experimental screening involves clustering. The database is divided into clusters from which representative molecules are chosen, and only these are screened.

The Supplementary Material referred to in this article can be found at http://www.interscience.wiley.com/jpages/0887-3585/suppmat/42_2/v42_2.279.htm

Grant sponsor: National Institutes of Health; Grant numbers: GM59957, GM21967, and T32-GM08382; Grant sponsor: Genetics Institute.

A.I. Su's present address is Department of Chemistry, Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, CA 92037.

G.S. Weston's present address is Computational Design & Informatics Group, ArQule, Incorporated, 19 Presidential Way, Woburn, MA 01801.

*Correspondence to: Brian K. Shoichet, Department of Molecular Pharmacology & Biological Chemistry, Northwestern University, 303 East Chicago Avenue, Chicago, IL 60611-3008. E-mail: b-shoichet@northwestern.edu

Received 18 August 2000; Accepted 12 October 2000

Here we investigate a simple method to increase the diversity of hits in molecular docking screens of chemical databases. Like experimental screens, we group the database into families of similar molecules. Unlike experimental screens, we evaluate every molecule in the database but only assign a rank to the best fitting member of each family (the representative molecule). Lower scoring molecules in that family are written to the hit list as annotations to the representative molecule; these analogs are not independently ranked. This is meant to increase the diversity of the hit list through the elimination of closely related molecules from the same family. At the same time, such clustering might enrich the information for any family of molecules because each high-scoring molecule reports not only its own score but also those of its analogs. We consider whether docking a database in families of related molecules increases the diversity of the docking hits. If so, is the resulting list sensible? That is, have we captured diversity while still finding molecules that will actually bind to the targeted sites? To test this, we experimentally evaluate novel molecules suggested by the new docking algorithm as candidates for binding to two different enzymes.

Overview of the Approach

The Available Chemicals Directory-3D (ACD; MDL Information Systems, San Leandro, CA) was organized into families that shared a common rigid fragment. This common rigid fragment might be, for instance, the largest aromatic ring in a molecule; other molecules that also had such a ring as their largest rigid fragment were put into the same family. We moved the molecules within each family into a common reference frame by superimposing these fragments (Fig. 1). Multiple conformations were generated for every molecule in the family. To dock this ensemble of conformations and molecules into a site, we used a modification of a flexible ligand docking algorithm that we previously described.[24] Orientations of the common rigid fragment for a family, for instance, the largest aromatic ring common to all family members, were calculated in the site. If the rigid fragment could be fit into the site, the rotation–translation matrix used to move it into the site was applied to the various side-chains off this ring in the family of different molecules (Fig. 2). Grouping by rigid fragment was convenient for the docking algorithm used here, but other clustering methods could be used.

Every molecule in a family was explicitly docked and scored in the binding site, typically in hundreds of orientations and hundreds of conformations. Fits were scored for steric[25] and electrostatic[26] complementarity to the binding site. The best conformation and orientation of the best molecule in a family was saved and, if it was among the top-scoring molecules for the database, was included in the hit list. This hit list was composed of several hundred top-scoring, representative molecules written out with coordinates to allow for visual examination in the enzyme site. For any given family, the identity of the representative molecule changed from one receptor to another.

METHODS

Organizing the Database into Families

The molecules of the 95.2 version of the ACD were grouped into families as follows. Beginning with a conformational ensemble database,[24] we calculated chemical ensembles in three steps: identifying the largest rigid fragments, grouping the molecules into families that shared a common fragment, and superimposing related molecules based on this fragment (Fig. 1).

For each molecule in the database, the largest rigid fragment was isolated and expressed in Sybyl Line Notation (SLN) with SYBYL6.4 (Tripos Associates, St. Louis, MO). Scripts to do this are provided in the Supplementary Material. For most molecules, the *rigid fragment* was defined as all the atoms of a ring system plus all directly bonded atoms.[24] Hydrogens were removed from the SLN strings; this ensured that molecules such as toluene and ethyl benzene would have common rigid fragments. Halogen atoms were equated to carbon atoms; this ensured that molecules such as toluene and all of the monohalobenzenes would have common rigid fragments. The list of rigid-fragment SLNs was sorted, and individual clusters were represented by unique SLN strings.

Because of computer memory limitations, several large clusters were divided into subclusters that were docked separately. Because the three-dimensional coordinates of all conformations of all the molecules in one ensemble were stored in memory at once, a maximum of 20,000 conformations per ensemble was imposed at the time of docking. Up to 500 low-energy conformations of each molecule was docked.[24] Molecules in the same chemical ensemble were overlaid with a script written in the SYBYL programming language (see the Supplementary Material). The rigid fragments of each molecule were superimposed through minimization of the root mean square deviation (RMSD) of the paired coordinates.

Preparation of the Test Systems

The ACD database was docked into three enzyme binding sites: dihydrofolate reductase (DHFR; Protein Data Bank (PDB) structure 3dfr[27]), thymidylate synthase (TS; PDB structure 1syn[28]), and Leu99 → Ala (L99A; PDB structure 181L[29]). Each structure was that of an enzyme–ligand complex; the ligands were removed for the docking calculation as were all water molecules, with the exception of the conserved Wat253 for the DHFR calculation.[30] For each protein, excluded and allowed site volumes were calculated to evaluate the shape complementarity of the docked molecules,[25] and electrostatic potentials were calculated, with DelPhi,[31] to evaluate their polar complementarity,[26] both as previously described.[24,30] Interaction energies were corrected for the electrostatic component of ligand desolvation.[30] For fitting the database molecules into the site, sphere positions and chemical labels were calculated as described.[24,30]

Docking Parameters

The ACD was docked into each enzyme site with both family-based and individual docking. The docking parame-

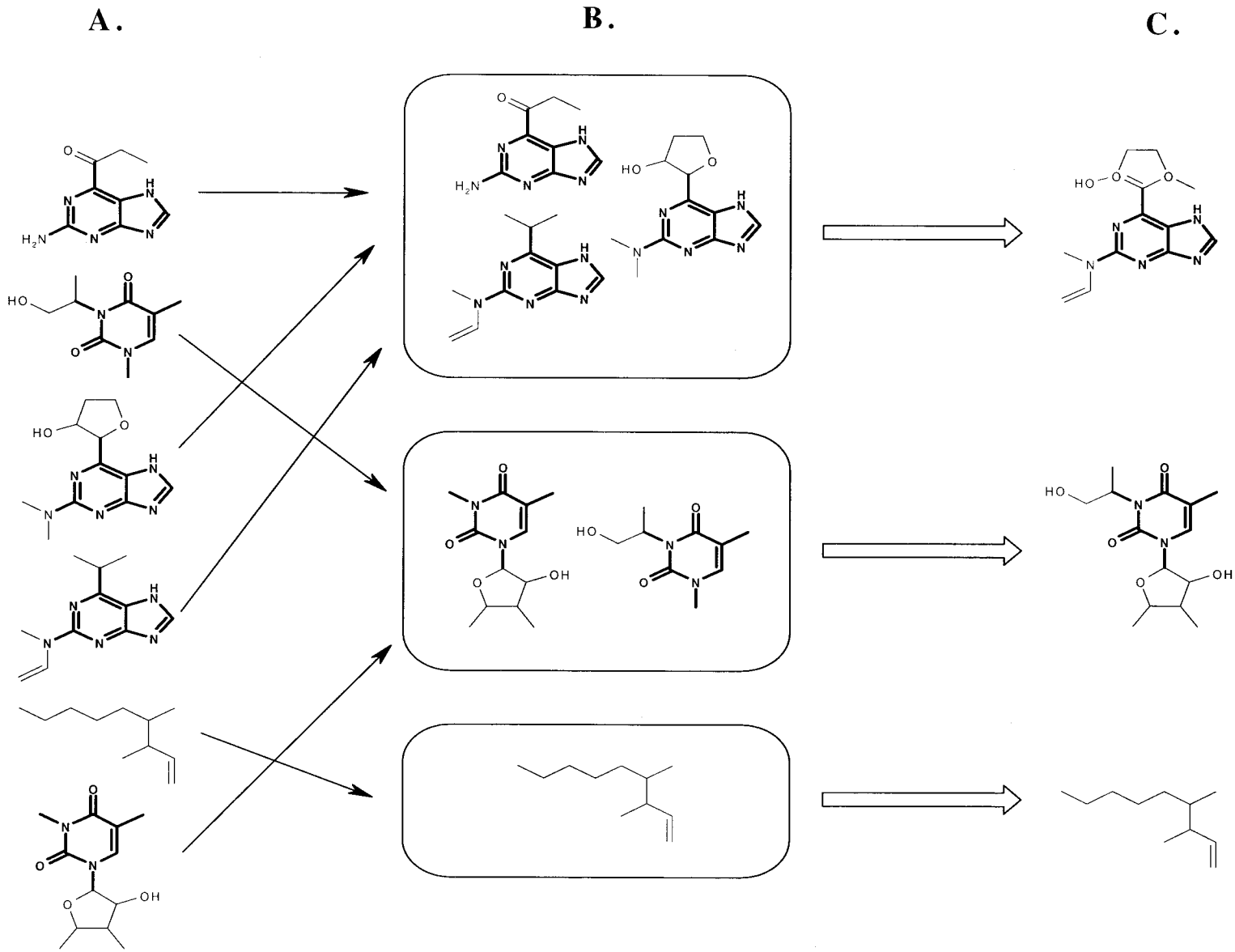


Fig. 1. Clustering a database by rigid fragments into families: (A) the largest rigid fragment in each molecule in the database is identified, (B) the database is grouped into families based on the common rigid fragment, and (C) the molecules are overlaid via the superimposition of their rigid fragments.

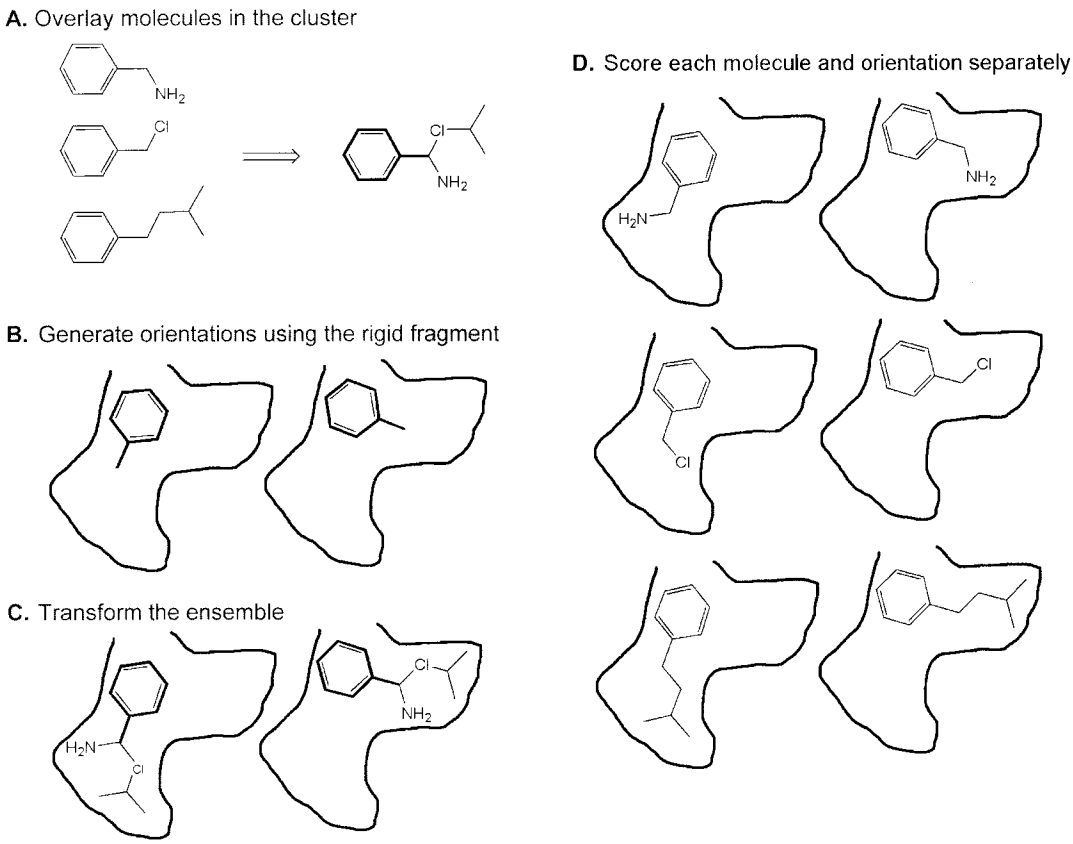


Fig. 2. Schematic drawing illustrating how the superimposed members of a given family are placed in the binding site and scored in multiple orientations.

ters used for each enzyme were identical for both calculations. For DHFR, the polar and nonpolar steric clash limits were 2.4 and 2.6 Å, respectively. The distance tolerance for ligand-atom/receptor-sphere matching was 0.8 Å.[25] The bin size and overlap were 0.3 and 0.2 Å for both the receptor and the ligand, respectively. For TS, the polar and nonpolar steric clash limits were 2.4 and 2.8 Å, respectively. The distance tolerance for ligand-atom/receptor-sphere matching was 1.5 Å. The bin size and overlap for the ligand were both 0.2 Å, and the receptor bin size and overlap were 0.2 and 0.4 Å, respectively. For L99A, the polar and nonpolar steric clash limits were 2.3 and 2.5 Å, respectively. The distance tolerance for ligand-atom/receptor-sphere matching was 0.75 Å. The bin sizes and overlaps for both the ligand and the receptor were 0.2 Å. In each calculation, four ligand-atom/receptor-sphere pairs were used to calculate orientations in the site.

Molecular Diversity

The diversity of the grouped and ungrouped hit lists was measured by pairwise Tanimoto distance comparisons. Each hit list was converted to Smiles format, and Daylight fingerprints were calculated for each set of smiles strings with Daylight toolkit programs (Daylight Chemical Information Software, v. 4.62, Daylight Chemical Information, Inc., Mission Viejo, CA). Tanimoto distances were then calculated for each possible pair of molecules in each of the grouped and ungrouped hit lists with the Daylight toolkit program Simatrix. Pairwise Tanimoto distance values for each hit list were then binned and graphed.

Inhibition and Binding Assays

Compounds were tested for binding to chicken liver DHFR (Sigma). Assays were performed in 50 mM potassium phosphate and 100 mM potassium chloride (pH 6.9) at 23 °C in an HP8453 diode array spectrophotometer with a multicell transporter. Reaction rates were monitored at 340 nm. Nicotine adenine dinucleotide phosphate (NADPH) (Sigma) was present in all assays at 100 μM, and the dihydrofolate (DHF; Sigma) concentration was typically varied between 1 and 100 μM to test for competitive inhibitory behavior. Substrate and cofactor stock solutions were made up fresh daily and stored on ice in 2 or 5 mM dithiothreotol solutions of the reaction buffer for the DHF and NADPH, respectively. Stock solutions of the inhibitors were initially prepared in 20 mM dimethyl sulfoxide (DMSO), except for 3,5,7-triamino-*s*-triazolo(4,3-a)-*s*-triazine and 6-hydroxymethylpteridine, which were made up to 10 mM in DMSO for solubility reasons. For the more active compounds, subsequent stock solutions of 2 mM inhibitor were made in 10% DMSO and 90% reaction buffer. The effect of DMSO on the uninhibited enzyme was

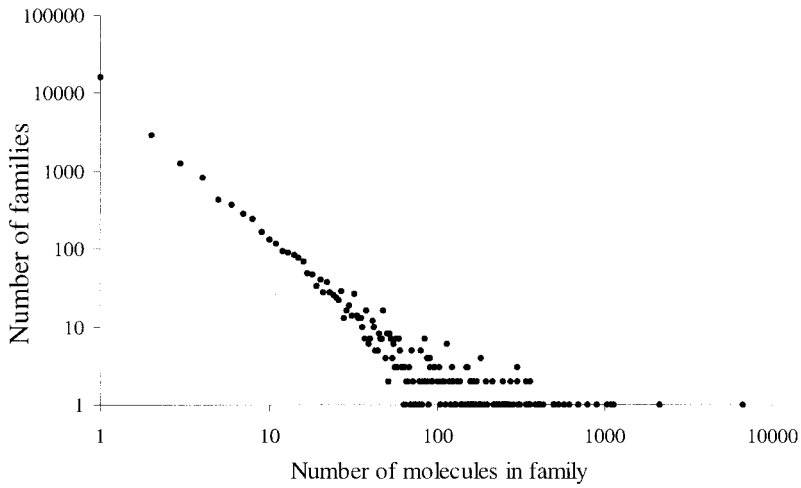


Fig. 3. Distribution of family sizes. A total of 86,619 molecules were grouped into 8,007 families of two or more molecules. In addition, 15,936 molecules were singletons, belonging in families that had only one member.

TABLE I. Identifying Known Ligands Among the Top 500 Hits Obtained by Family-Based and Individual Docking Screens

	Family docking			Individual docking		
Enzyme	Families in the hit list	Hit-list families that include a known ligand[a]	Total ligands in the hit-list families[a]	Families in the hit list	Hit-list families that include a known ligand[a]	Total ligands in the hit list[a]
L99A	500	7 (48)	34 (127)	169	3 (30)	10 (71)
TS	500	5 (6)	8 (16)	171	4 (5)	7 (11)
DHFR	500	5 (12)	14 (36)	221	2 (3)	4 (6)

[a]The numbers in parentheses include the molecules that are known to be ligands plus similar compounds that are presumed to be ligands but have not, to our knowledge, been experimentally tested.[30]

controlled; the percentage of DMSO never rose above 10% by volume. Even at this high concentration, the enzyme retained most of its activity. K_i values were calculated with progress curves.[32] Assays were run near the IC_{50} of the inhibitors, with the DHF substrate between 3 and 10 μM. The K_m of DHF was taken to be 0.15 μM,[33] which is consistent with our own experiments. In all cases, reactions were initiated by enzyme addition.

Compounds were tested for binding to the L99A cavity mutant of T4 lysozyme[34] using the method of thermal upshift.[29] Compounds were tested at 95% of their saturating solubility, which in no case exceeded 1 mM. Measurements were made in 25 mM potassium chloride, 2.95 mM phosphoric acid, and 17 mM potassium dihydrogenphosphate (pH 3.02). Buffers were sparged with wet nitrogen gas overnight to reduce dissolved oxygen. The concentration of protein was 0.5 mg/mL. Melting was conducted in a differential scanning calorimeter with a MicroCal VP-DSC. The rate of temperature increase in the melting studies was 1 °C/min. Melting curves were analyzed with Origin (MicroCal Inc, Northampton, MA). All melts were reversible.

RESULTS

A total of 86,619 molecules from the ACD were grouped into 8,007 families, each of which had at least two members (Fig. 3). Additionally, 15,936 molecules had no rigid fragment in common with any other in the database; these singletons were docked as single molecule conformational ensembles.[24] This produced a database of 102,555 molecules. For families with two or more members, the mean size was 10.8 molecules per family; 50% of the database segregated into clusters of 59 or more molecules. On average, every molecule in the database had 280 conformations, giving rise to a total of 28.7 million distinct conformers in the database. Docking calculations required about 6–8 h of CPU time, depending on the target. Calculations were performed on a single Pentium II 450-MHz processor under Linux.

The docking procedure was tested with three different targets: the cavity mutant of T4 lysozyme resulting from the substitution Leu99→Ala (L99A), the deoxyuridine monophosphate (dUMP) site of TS, and the pterin binding site of DHFR. In each case, the database was evaluated in the conventional way (*individual docking*)[24] and by families (*family docking*). Both algorithms used the same fitting and scoring algorithms, with the exception that family docking calculations calculated the same set of orientations for each member of a family, whereas in individual docking, orientations were calculated separately for each ligand (this typically resulted in the same numbers of orientations being calculated in each case, only more efficiently in the family-

TABLE II. Comparative Rankings of Known L99A Ligands with Family Docking Versus Individual Docking

			Family docking		Individual docking	
MFCD[a]	Compound	ΔT_m (°C[b])	Rank[d]	Score (kcal/mol)	Rank	Score (kcal/mol)
00059193	Phenylacetamide	nd[c]	1	−4.60	2	−4.60
00009502	*n*-Pentylbenzene	1.2	(1)	−1.48	319	−1.49
00009463	*n*-Butylbenzene	1.4	(1)	−1.46	353	−1.47
00009377	*n*-Propylbenzene	6.2	(1)	−1.46	355	−1.47
00008612	Styrene	2.0	(1)	−1.45	404	−1.43
00134644	Toluene	4.4	(1)	−1.44	384	−1.44
00009329	*sec*-Butylbenzene	1.3	(1)	−1.43	374	−1.45
00011647	Ethylbenzene	3.8	(1)	−1.43	405	−1.43
00000280	Fluorobenzene	5.9	(1)	−1.09	817	−1.10
00001029	Iodobenzene	2.6	(1)	−0.40	2,817	−0.29
00008936	Isobutylbenzene	2.9	(1)	−0.34	2,577	−0.34
00009526	*n*-Hexylbenzene	0.5	(1)	−0.28	2,624	−0.33
00007982	*o*-Toluamide	nd[c]	4	−3.92	1,276	−0.83
00008519	*o*-Xylene	2.9	(4)	−1.23	648	−1.22
00001042	2-Iodotoluene	1.9	(4)	−0.74	1,144	−0.88
00009257	2-Ethyltoluene	1.1	(4)	−0.39	2,342	−0.39
00160764	3-Iodobenzaldehyde oxime	nd[c]	8	−3.61	13	−3.24
00008536	*m*-Xylene	1.8	(8)	−2.07	88	−2.07
00001050	3-Iodotoluene	2.0	(8)	−1.98	1,853	−0.55
00009259	3-Ethyltoluene	1.1	(8)	−1.65	385	−1.44
00010611	Methyl-cyclohexanepropionate	nd[c]	32	−2.59	35	−2.59
00001497	Methylcyclohexane	1.1	(32)	−0.56	1,820	−0.56
00160767	4-Iodobenzaldehyde oxime	nd[c]	34	−2.45	101	−2.03
00008556	*p*-Xylene	2.4	(34)	−0.76	1,402	−0.76
00001059	4-Iodotoluene	1.3	(34)	−0.41	2,748	−0.30
00009263	4-Ethyltoluene	1.7	(34)	−0.24	3,086	−0.25
00085154	2-Phenylthiothioacetamide	nd[c]	77	−1.78	1,795	−0.57
00008559	Thioanisole	1.4	(77)	−0.33	2,622	−0.33
00008570	Phenylacetylene	0.7	146	−1.37	485	−1.37
00003777	Indene	1.2	613	−0.52	1,930	−0.52

[a]MDL registry number for the ACD compounds (MDL Inc., San Leandro, CA).
[b]Increase in the temperature of melting of L99A in the presence of ligand compared to the apo-enzyme. Stability upshift indicates ligand binding in this system.[29]
[c]Not determined. This compound has not been tested for binding to our knowledge. It is included here because it is the high-scoring member of a family that has known ligands in it; see the next note.
[d]Phenylacetamide, for example, is in the highest-ranked family and has the best score within that family. Compounds such as *n*-butylbenzene, with ranks given in parentheses, are also in the highest-ranking family but have a lower score.

based calculation). Known ligands and close analogs were found among the top scoring hits for each site by both methods (Tables I–IV). These ligands docked to the enzymes in conformations that resembled those determined by crystallography, typically varying by less than 2-Å RMSD. As expected, identical ligands scored similarly in both calculations. Occasionally, the same ligand scored differently in the two calculations because of differences in the number of ligand orientations in the binding sites that were sampled in the respective calculations. Differences in the numbers of orientations calculated arose when the rigid fragment for the family of molecules differed slightly from the largest rigid fragment for the individual molecules, typically because of differences in bond lengths. For instance, in the family that contained fluro-, chloro-, bromo-, iodo-, and methyl-benzene, the bond distances between the ring and the exocyclic atom were not exactly the same. When the molecules were docked individually, the calculated orientations depended on the atomic distances within each individual molecule. When the molecules were docked as families, a single representative rigid fragment was used to calculate orientations for all the molecules in the family, and this could lead to differences in orientation numbers calculated between single molecule docking and family-based docking. For most families, this discrepancy arose rarely.

More known ligands and close analogs were found in the hit list of the family-based calculations than in the individual calculations (Tables I–IV). Docking in families increased the number of known ligands and analogs between 45% (from 11 to 16 for TS) and 500% (from 6 to 36 for DHFR) for the three binding sites in comparison with docking and ranking each database molecule independently. Similarly, docking in families increased the number of ligand classes between 20% (from 5 to 6 for TS) and 300% (from 3 to 12 for DHFR) in the three sites in comparison with docking one molecule at a time.

To estimate the diversity of the top hits in the family-based calculations versus the individual calculations, Tanimoto coefficients[35] were calculated for all pairs of molecules in each hit list [Fig. 4(A–C)]. For all three enzymes, the distribu-

TABLE III. Comparative Rankings of Known DHFR Ligands with Family Docking Versus Individual Docking

			Family docking		Individual docking	
MFCD[a]	Compound	K_i or IC_{50} (μM)	Rank[b]	Score (kcal/mol)	Rank[b]	Score (kcal/mol)
00012732	4-[*N*-(2,4-Diamino-6-pteridinylmethyl)-*N*-methylamino] benzoic acid	0.3	8	−21.3	44	−20.2
00036692	Aminopterin	0.010	(8)	−17.4	422	−13.6
00006709	4-[*N*-(2,4-Diamino-6-pteridinylmethyl)amino] benzoic acid	1	(8)	−13.6	255	−15.0
00138064	2,4-Diamino-6-methylpteridine	10	(8)	−6.36	6,331	−6.54
nd[c]	2,4-Diaminopteridine	6	101	−15.0	264	−15.0
nd[c]	2,4-Diamino-7-methyl-6-pteridinyl methyl ketone	8	245	−11.9	574	−12.6
00038077	2,4-Diamino-6,7-diisopropylpteridine	0.06	(245)	−7.94	4,104	−7.42
00014658	2,4-Diamino-6,7-dimethylpteridine	10	(245)	−2.60	5,979	−6.65
00038691	2-Amino-4-hydroxy-6-pteridinecarboxaldehyde	300	409	−10.4	1,122	−10.4
00012137	6-Methylpterin	nk[d]	(409)	−8.85	1,975	−9.03
00085369	Monopterin	nk[d]	(409)	−7.68	3,674	−7.66
00036787	Biopterin	nk[d]	(409)	−7.03	4,830	−7.05
00042801	Neopterin	nk[d]	(409)	−2.16	31,143	−3.06
00069326	(6*R*)-5,6,7,8-Tetrahydro-1-biopterin	30	421	−10.3	1,165	−10.3
00066176	2,4-Diamino-6-hydroxypyrimidine	10,000	537	−9.73	1,452	−9.73

[a]MDL registry number for the ACD compounds (MDL Inc., San Leandro, CA).
[b]Compounds with a rank given in parentheses are members of the family with that rank but are not the highest-scoring member. The rank of the best-scoring family member is given without parenthesis.
[c]This compound was in the FCD database, a precursor to the ACD, but is no longer found in the ACD; its structure was known to us from previous work.[40]
[d]Not known. The compound is known to bind to a DHFR, but a formal binding constant was not found in the literature.

TABLE IV. Comparative Rankings of Known TS Ligands with Family Docking Versus Individual Docking

			Family docking		Individual docking	
MFCD[a]	Compound name	K_i or IC_{50} (μM)	Rank[c]	Score (kcal/mol)	Rank[c]	Score (kcal/mol)
00065282	2′-Deoxyuridine 5′-monophosphate	1.6	2	−81.9	3	−81.9
00211208	5-methyluridine-5′-monophosphate	600	21	−72.6	41	−72.6
00057406	5-Bromo-2′-deoxyuridine-5′-monophosphate	4	(21)	−72.2	55	−72.2
00057409	5-fluoro-2′-deoxyuridine 5′-monophosphate	0.014	(21)	−71.4	35	−73.1
00023797	thymidine 5′-monophosphate	15	(21)	−63.5	417	−63.5
00057404	5-bromouridine 5′-monophosphate	nk[b]	(21)	−39.8	1,486	−39.8
00057412	5-iodouridine 5′-monophosphate	nk[b]	(21)	−39.8	73,052	93.1
00044939	phenolphthalein monophosphate	5	32	−71.7	65	−71.7
00038064	5-Methyl-2′-deoxycytidine 5′-monophosphate	100	98	−65.8	310	−65.8
00039043	Pyridoxamine 5-phosphate	nd[d]	261	−48.0	919	−50.8
00149414	Pyridoxal 5-phosphate	1.6	(261)	−26.3	2,395	−26.1

[a]MDL registry number for the ACD compounds (MDL Inc., San Leandro, CA).
[b]Compound that is thought to inhibit TS on the basis of SAR data[41] but for which a binding constant was not found in the literature.
[c]Compounds with a rank given in parentheses are members of the family with that rank but are not the highest-scoring member. The rank of the best-scoring family member is given without parentheses.
[d]Not determined. This compound has not been tested for binding, to our knowledge. It is included here because it is the high-scoring member of the family that includes the known ligand pyridoxal phosphate and resembles this ligand closely.

tion of Tanimoto coefficients are skewed toward lower values, indicating lower similarity and higher diversity in the family calculations versus the individual docking calculations.

We chose to experimentally test molecules that ranked highly when docked as families but not when docked and ranked individually. None of these molecules were known by us to bind to these sites; many of them explored what appeared to be new functionality. Six molecules were chosen to test as binders to the L99A cavity site, and seven molecules were chosen to test as inhibitors of DHFR. Binding to the L99A cavity site was tested on the basis of the ability of the ligands to stabilize the enzyme against thermal denaturation.[29,34] Of the six molecules tested, five were found to bind, with thermal upshift values between 1.1 and 4.5 °C at a ligand concentration of no greater than 1 mM (Table V). In comparison, 1 mM of the characteristic ligand benzene (K_d = 400 μM) stabilized L99A by 2.9 °C.[34] Of the seven molecules tested against DHFR, six inhibited the enzyme with apparent K_i values that ranged from 0.26 to 100 μM (Table VI).

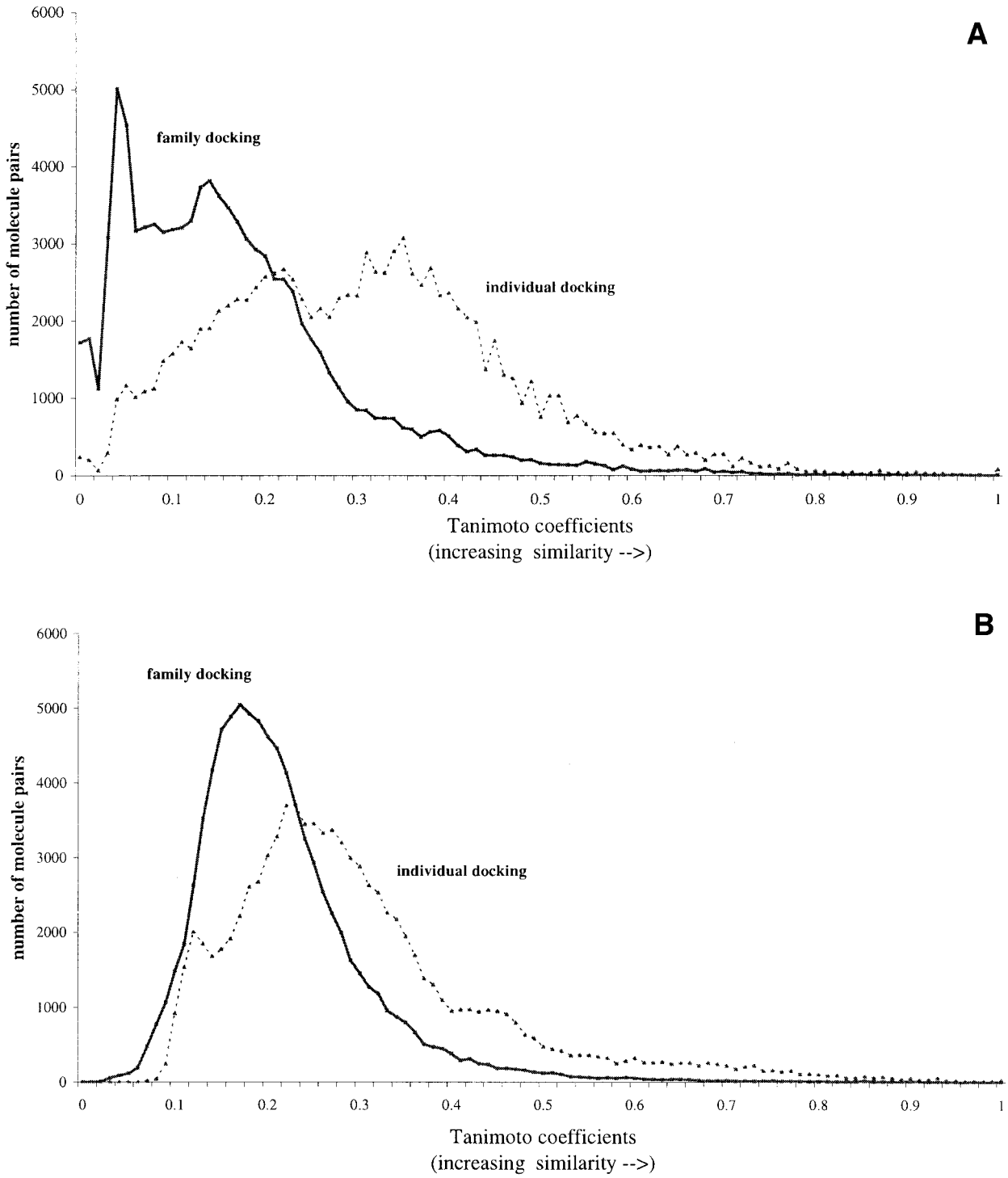


Fig. 4. Pairwise Tanimoto coefficients for the top 500 hits in the family (solid line) and individual (dashed line) docking screens of the ACD database against (A) L99A, (B) TS, and (C) DHFR.

DISCUSSION

Docking programs screen databases of molecules against protein structures in an attempt to discover novel lead ligands. Our objective in docking the database with families of related molecules was to increase the diversity of hits. It is appropriate to ask whether diversity was increased without loss of information, and if so, are the new, more diverse hits sensible. Do they bind to the targeted protein?

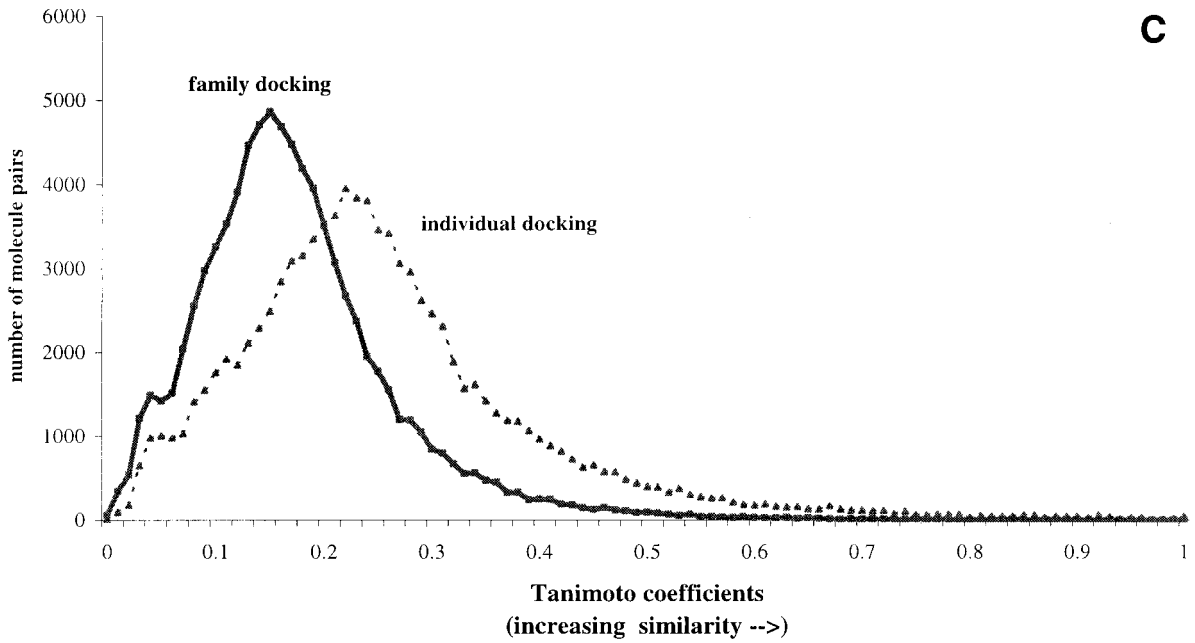


Figure 4. (Continued.)

Docking the ACD database in families increased the diversity of the top-scoring molecules in comparison with docking each molecule independently. This can be seen with two metrics. First, the distributions of Tanimoto coefficients for the top 500 hits on the respective lists are very different [Fig. 4(A–C)]. Tanimoto coefficients are widely used to measure molecular similarity:[35,36] the lower the coefficient, the less similar the two molecules being compared. Looking at all pairs of molecules measures the distribution of similarities in the hit lists. The lower Tanimoto coefficients for the family-based hit lists confirm that these molecules are more diverse and lack the repetition of similar molecules that occurs when the database molecules are docked and ranked individually. A second measure of diversity is to compare the number of known ligand classes that are identified by the respective screens. The family-based docking calculations appear superior by this measure as well (Table I). For example, the 500 hits in the individual docking screen identified only one class of ligands of DHFR: the diaminopteridines (Table III). The family-based hits, however, included not only the diaminopteridines but also the pterin, diaminopyrimidine, and, as is discussed later, 8-azanucleic acid class of inhibitors.

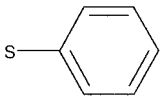
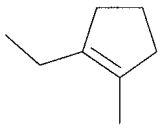
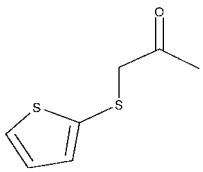
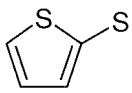
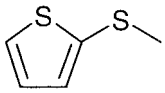
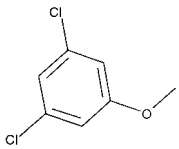
Do we lose information by docking the database molecules in families rather than individually? This is a concern in using clustering in experimental screens because fewer compounds in any given family are actually tested. In family-based docking, this is not an issue; because every molecule in every family is still explicitly docked and scored, one has access to the same amount of information as with individual screening. An advantage of family-based docking is that the scores for related molecules will be grouped together in the hit list. With individual docking, the scores and ranks of similar molecules will be scattered and may not be recorded in the hit list if their scores were not high enough.

The individual docking screen against L99A provides a simple example of how similar molecules receive similar scores, preventing other interesting ligands from appearing as hits. The L99A binding site is a cavity in the core of T4 lysozyme that binds small hydrophobic ligands. When the ACD was docked into the cavity one molecule at a time, the known ligand *n*-pentylbenzene received a score of −1.49 kcal/mol and was ranked 319 out of 102,555 (Fig. 5). The known ligands *n*-butylbenzene, *n*-propylbenzene, and *sec*-butylbenzene received scores of −1.47, −1.47, and −1.45 kcal/mol and were ranked 353, 355, and 374, respectively. Although it is gratifying to find known ligands among the hits, this repetition of similar molecules prevents other dissimilar molecules from appearing in a hit list that is necessarily limited.

When docking is done by families against L99A, *n*-pentylbenzene, *n*-butylbenzene, *n*-propylbenzene, and *sec*-butylbenzene do not receive independent rankings. Although all are listed in the hit list, they are all considered a single hit (Fig. 5). This allows other novel molecules such as thiophenol, ethyl-methyl-cyclopentene, and methylthiothiophene to rank among the top hits in the family-based screen (Fig. 5). These compounds were not found in the hit list when the ACD molecules were docked individually; nevertheless, they appear to be interesting candidates.

To investigate how sensible these novel molecules were, we experimentally tested six of them for binding to L99A (Table V). These six were chosen on the basis of a large differential ranking between the family-based and individual docking calculations and because they seemed to explore new chemistry for this site; to our knowledge, none

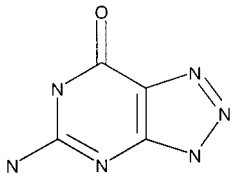
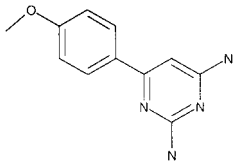
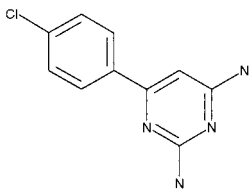
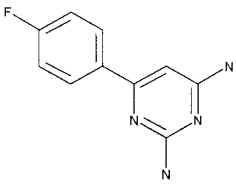
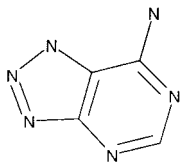
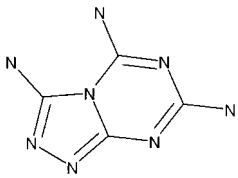
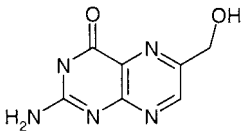
TABLE V. Experimental Testing of Candidate Ligands Identified From the Family-Based Docking Versus the L99A Cavity Site in T4 Lysozyme

Compound	MFCD[a]	Structure	Family docking rank[b]	Individual docking rank	Score (kcal/mol)	T_m (°C)[c]	ΔT_m (°C)	Binder?
None	—	—	—	—	—	38.5	0	—
Thiophenol	00004826		(77)[d]	1,921	−0.52	43.0	4.5	Yes
1-Ethyl-2-methyl-cyclopentene	00036479		(223)[e]	2,956	−0.27	39.9	1.2	Yes
2-(Thioenylthio)-acetone	00067958		302	2,925	−0.94	39.6	1.1	Yes
Thiophene-2-thiol	00051666		(302)	1,320	−0.82	40.6	2.1	Yes
2-(Methylthio)-thiophene	00052382		(302)	2,012	−0.37	41.8	3.3	Yes
3,5-Dichloro-anisole	00000589		362	1,257	−0.84	38.7	0.2	No[f]

[a]MDL registry number for the ACD compounds (MDL Inc., San Leandro, CA).
[b]Compounds with the family docking rank given in parentheses are in the family but are not the highest-scoring member.
[c]Melting temperature.
[d]The high-scoring compound in this family was phenyl vinyl sulfide, which received a docking score of −0.87.
[e]The high-scoring compound in this family was the 1,2-dimethyl analog, which received a docking score of −1.2 kcal/mol. This representative was not readily available to us. The two molecules differ by a methylene.
[f]Binding not detected by thermal upshift.

of the six has been previously tested. For instance, no thiophenes have been previously found to bind to L99A, whereas thiophenol resembles phenol, which was not observed to bind to L99A.[29] Of the six molecules tested, five bound to L99A (Table V). Although binding constants were not determined, these molecules had thermal up-

TABLE VI. Experimental Testing of Candidate Inhibitors Identified From the Family-Based Docking Versus DHFR

Compound	MFCD[a]	Structure	Family docking rank[b]	Individual docking rank	Score (kcal/mol)	K_i (μM)
8-Azaguanine	00056937		427	1,174	−10.3	80
6-(4-Methoxyphenyl)-pyrimidine	00109570		507	1,375	−9.87	0.57
6-(4-Chlorophenyl)-pyrimidine	00068134		(507)	1,480	−9.68	0.26
6-(4-Fluorophenyl)pyrimidine	00052076		(507)	2,588	−8.43	1.9
8-Azaadenine	00005697		547	1,479	−9.68	40
3,5,7-Triamino-*s*-triazolo(4,3-*a*)-*s*-triazine	00039686		555	1,504	−9.64	100
6-Hydroxymethyl-pterin	00038456		672	1,755	−9.15	nd[c]

[a]MDL registry number for the ACD compounds (MDL Inc., San Leandro, CA).
[b]Compounds with the family docking rank given in parentheses are in the family but are not the highest-scoring member.
[c]No inhibition detected.

shifts that compared favorably with well-characterized ligands, such as benzene (thermal upshift occurs because the ligands stabilize L99A against thermal denaturation by binding in the cavity site; many hydrophobic molecules do not have this effect and are considered nonbinders by this technique). Indeed, the thermal upshift for thiophenol

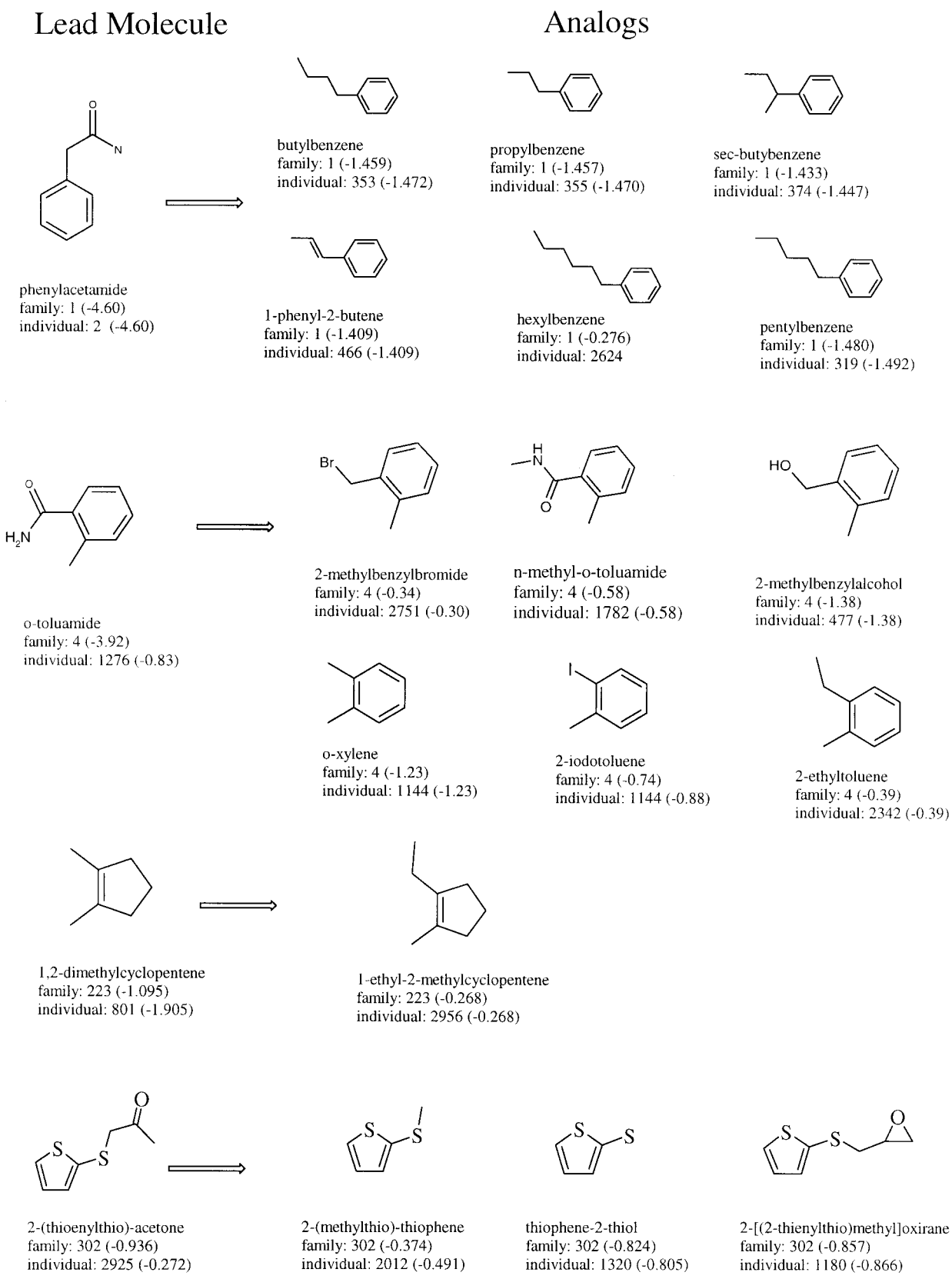


Fig. 5. Rankings and scores of representative family members docked into L99A. On the left are high-scoring molecules from the family-based docking screen, pointing to the other members of the family on the right. The ranks (the scores are in kcal/mol) are given.

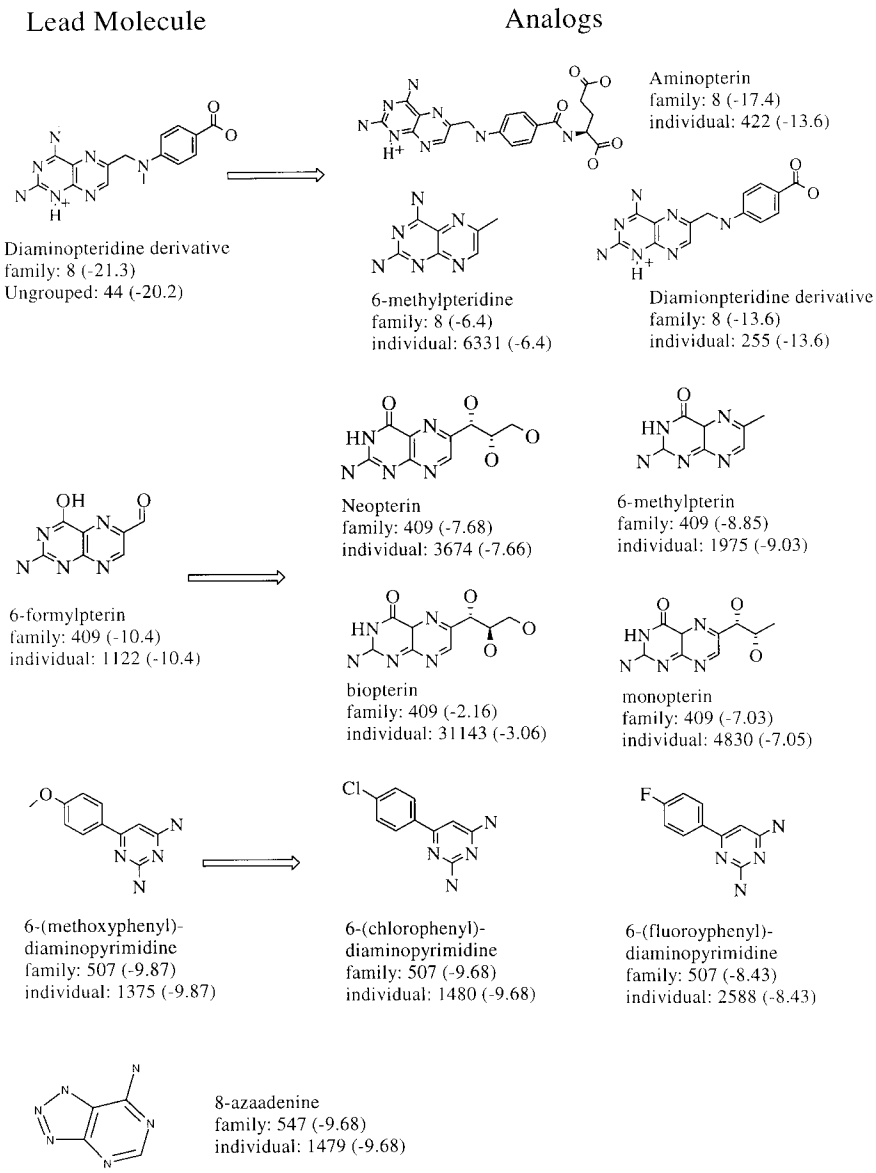


Fig. 6. Rankings and scores of representative family members docked into DHFR. On the left are several high-scoring molecules from the family-based docking screen, pointing to the other members of the cluster on the right. The ranks (the scores are in kcal/mol) are given.

is among the highest measured for an L99A ligand at 1 mM. These experiments suggest that the family-based docking hits for L99A are not only more diverse than the standard, individual docking hits but also that they include bona fide new candidates.

In DHFR too, more inhibitors and inhibitor families are found with family-based docking than with individual docking. For instance, when the database was docked one molecule at a time, the known inhibitors tetrahydrobiopterin and diisopropylpteridine[37] were ranked 1,165 and 4,104 and were well outside the best 500 or best 1% of database molecules that are typically used to define a docking hit list. However, both molecules are ranked among the top 500 hits in the family-based screen (Table III). These inhibitors were not found in the individual docking screen because they were crowded out by a succession of similar molecules in the hit list. Thus, the known DHFR inhibitor[37] 4-[*N*-(2,4-diamino-6-pteridinylmethyl)-*N*-methyl-amino]-benzoic acid (Fig. 6) ranks 44 in the individual docking calculation, and its analogs 4-[*N*-(2,4-diamino-6-pteridinylmethyl)-amino]-benzoic acid and aminopterin (Fig. 6) rank 255 and 422, respectively. In the family-based calculation, these analogs were ranked as a single family and occupied one position, with multiple annotations, in the hit list. This allowed new types of molecules into the hit list, including the diaminopyrimidines, 8-azaguanine, and 8-azaadenine, all of which appeared to be sensible candidates.

To investigate if these novel molecules were, in fact, sensible, seven were tested as inhibitors in an enzyme

assay. As with L99A, the seven were selected on the basis of a large differential ranking between the family-based calculations and the individual docking calculations. These molecules were not known by us to inhibit DHFR, according to the review of Blaney et al.[37] and chemical structure searches of the Chemical Abstracts and Beilstein electronic databases. Also, they appeared to explore new chemistries compared with known DHFR inhibitors. For instance, diaminopyrimidines are well-known DHFR inhibitors, but most of the structure–activity relationships (SARs) in this series seem to have focused on bulky substitutions off of the 5 position; 6-substituted pyrimidines that lack a side-chain off of C5 appeared to be poor inhibitors.[37] Molecules such as 8-azaguanine, 8-azaadenine, and 8-azatriazine fall into classes for which no SAR was known to us. Of the seven molecules tested, six inhibited DHFR with apparent K_i values between 0.26 and 100 μM. As with L99A, the family-based docking against DHFR found molecules that were more diverse than those found with individual docking and were sensible as candidate inhibitors.

Impact on the Hit Rates

A well-known problem with molecular docking is that of false negatives, molecules that would inhibit the target but are not identified as hits. False negatives are often due to limitations or inaccuracies in the docking scoring function.[38,39] This problem is compounded because there is a large number of analogs in most databases, and the differences in scores between docking hits is often small. A trivial difference in scoring energies, one much smaller than the errors in the method, can make the difference between a molecule being included or not included in a hit list. Entire classes of ligands can be missed because a single large class of molecules is overrepresented in a hit list that is necessarily limited in size.

Although the problem of accurate scoring is difficult to solve, we believe that the problem of distinguishing between similar and dissimilar molecules can be addressed at little cost. Docking ligands in families brings a much broader range of sensible molecules into the hit list and reduces dependence on the scoring function by presenting related hits together. This allows the investigator to view the results as a group and to see past small variations in an admittedly inaccurate energy score.

A situation where family-based docking leads to fewer hits might be considered. Imagine that a certain fixed number of top-scoring ligands from a docking screen are chosen for experimental testing. In individual docking, this list might contain several molecules from the same family, only the best scoring of which would be represented explicitly in the hit list of a family docking calculation. If this best scoring molecule did not inhibit the target, but the second best molecule did, then this family of molecules would be found in the individual docking calculation but missed in the family-based docking. The counter argument to this point is that when ligands are docked in families, all the information about a family is readily available, even though the family is represented only once in the hit list. Investigators can choose to screen more or less members from a given family, as guided by the docking scores within that family and their own experience. In family docking, investigators are presented not only with information about more families (more diverse hit lists) but also with more information within each family.

Several limitations to this approach of docking molecules by families deserve mention. Grouping molecules based on common rigid fragments can lead to families of molecules that, although identical in rigid fragment, are otherwise unrelated. We chose to organize the families on the basis of rigid fragments because it allowed us to calculate the scores of related molecules in the configuration of the high-scoring cluster member and it fit well with our docking method. Clustering algorithms have been extensively explored; other ways of organizing the database into related families are certainly conceivable.

This said, it should be clear that clustering of some sort will improve the diversity of docking screens, almost regardless of the type of clustering that is chosen or the type of docking method in which it is implemented. There are some advantages to chemical ensembles, but the most important benefits may be realized by simply clustering the database, ranking only the high-scoring member of the family in the hit list, and recording the scores of related molecules as annotations to this representative family member. Because all the molecules in the database are still docked and scored, no information is lost. The advantages can be considerable: a more diverse hit list, less sensitivity to small differences in docking score, and, consequently, a greater likelihood of finding molecules that both explore new chemistries and bind well to the target.

ACKNOWLEDGMENTS

This work was supported by grants from the National Institutes of Health (GM59957) and the Genetics Institute to B.K. Shoichet and by a grant from the NIH (GM21967) to B.W. Matthews. A.I. Su was partly supported by a Macey summer fellowship, and D.M. Lorber was partly supported by a grant from the NIH (T32-GM08382). The authors thank Paul Charifson, Juan Alvarez, Susan McGovern, Binqing Wei, and Beth Beadle for reading this manuscript, David Hartsough for his assistance and advice with the Daylight toolkit programs, and MDL Information Systems (San Leandro, CA) for the ACD database and the ISIS program.

REFERENCES

1. Goodsell DS, Olson AJ. Automated docking of substrates to proteins by simulated annealing. Proteins 1990;8:195–202.
2. Cherfils J, Duquerroy S, Janin J. Protein–protein recognition analyzed by docking simulation. Proteins 1991;11:271–280.
3. Katchalski-Katzir E, Shariv I, Eisenstein M, Friesem AA, Aflalo C, Vakser IA. Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. Proc Natl Acad Sci U S A 1992;89:2195–2199.
4. Kuntz ID, Meng EC, Shoichet BK. Structure-based molecular design. Acc Chem Res 1994;27:117–123.
5. Totrov M, Abagyan R. Detailed ab initio prediction of lysozyme-antibody complex with 1.6 Å accuracy. Nat Struct Biol 1994;1:259–263.

6. Duncan BS, Olson AJ. Approximation and visualization of large-scale motion of protein surfaces. J Mol Graph 1995;13:250–257.
7. Rarey M, Kramer B, Lengauer T. A fast flexible docking method using an incremental construction algorithm. J Mol Biol 1996;261: 470–489.
8. Schaffer L, Verkhivker GM. Predicting structural effects in HIV-1 protease mutant complexes with flexible ligand docking and protein side-chain optimization. Proteins 1998;33:295–310.
9. Strynadka NCJ, Jensen SE, Alzari PM, James MNG. A potent new mode of β-lactamase inhibition revealed by the 1.7 Å X-ray crystallographic structure of the TEM-1-β-lactamase complex. Nat Struct Biol 1996;3:290–297.
10. Shoichet BK, Kuntz ID. Predicting the structures of protein complexes: a step in the right direction. Chem Biol 1996;3:151–156.
11. Janin J. Protein–protein recognition. Prog Biophys Mol Biol 1995;64:145–166.
12. Hart TN, Ness SR, Read RJ. Critical evaluation of the research docking program for the CASP2 challenge. Proteins 1997; Supplement 1:205–209.
13. Dixon JS. Evaluation of the CASP2 docking section. Proteins 1997;Supplement 1:198–204.
14. Toney JH, Fitzgerald PM, Grover-Sharma N, Olson SH, May WJ, Sundelof JG, Vanderwall DE, Cleary KA, Grant SK, Wu JK, Kozarich JW, Pompliano DL, Hammond GG. Antibiotic sensitization using biphenyl tetrazoles as potent inhibitors of *Bacteroides fragilis* metallo-beta-lactamase. Chem Biol 1998;5:185–196.
15. Burkhard P, Taylor P, Walkinshaw MD. An example of a protein ligand found by database mining: description of the docking method and its verification by a 2.3 Å X-ray structure of a thrombin–ligand complex. J Mol Biol 1998;277:449–466.
16. Tondi D, Slomczynska U, Watterson DM, Costi MP, Ghell S, Shoichet BK. Structure-based discovery and in-parallel elaboration of novel inhibitors of thymidylate synthase. Chem Biol 1999;6:319–331.
17. Kuntz ID. Structure-based strategies for drug design and discovery. Science 1992;257:1078–1082.
18. Aronov AM, Suresh S, Buckner FS, Van Voorhis WC, Verlinde CL, Opperdoes FR, Hol WG, Gelb MH. Structure-based design of submicromolar, biologically active inhibitors of trypanosomatid glyceraldehyde-3-phosphate dehydrogenase. Proc Natl Acad Sci U S A 1999;96:4273–4278.
19. Martin EJ, Blaney JM, Siani MA, Spellmeyer DC, Wong AK, Moos WH. Measuring diversity: experimental design of combinatorial libraries for drug discovery. J Med Chem 1995;38:1431–1436.
20. Brown RD, Martin YC. An evaluation of structural descriptors and clustering methods for use in diversity selection. SAR QSAR Environ Res 1998;8:23–39.
21. Shemetulskis NE, Dunbar JB, Jr, Dunbar BW, Moreland DW, Humblet C. Enhancing the diversity of a corporate database using chemical database clustering and analysis. J Comput Aided Mol Des 1995;9:407–416.
22. Cummins DJ, Andrews CW, Bentley JA, Cory M. Molecular diversity in chemical databases: comparison of medicinal chemistry knowledge bases and databases of commercially available compounds. J Chem Inf Comput Sci 1996;36:750–763.
23. Potter T, Matter H. Random or rational design? Evaluation of diverse compound subsets from chemical structure databases. J Med Chem 1998;41:478–488.
24. Lorber DM, Shoichet BK. Flexible ligand docking using conformational ensembles. Protein Sci 1998;7:938–950.
25. Shoichet B, Bodian DL, Kuntz ID. Molecular docking using sphere descriptors. J Comp Chem 1992;13:380–397.
26. Meng EC, Shoichet B, Kuntz ID. Automated docking with grid-based energy evaluation. J Comp Chem 1992;13:505–524.
27. Bolin JT, Filman DJ, Matthews DA, Hamlin RC, Kraut J. Crystal structures of *E. coli* and *L. casei* DHFR refined to 1.7 Angstrom resolution. 1. General features and binding of methotrexate. J Biol Chem 1982;257:13650–13662.
28. Stout TJ, Stroud RM. The complex of the anti-cancer therapeutic, BW1843U89, with thymidylate synthase at 2.0 Å resolution: implications for a new mode of inhibition. Structure 1996;4:67–77.
29. Morton A, Baase WA, Matthews BW. Energetic origins of specificity of ligand binding in an interior nonpolar cavity of T4 lysozyme. Biochemistry 1995;34:8564–8575.
30. Shoichet BK, Leach AR, Kuntz ID. Ligand solvation in molecular docking. Proteins 1999;34:4–16.
31. Gilson MK, Honig BH. Calculation of electrostatic potentials in an enzyme active site. Nature 1987;330:84–86.
32. Waley SG. A quick method for the determination of inhibition constants. Biochem J 1982;205:631–633.
33. Kaufman BT, Gardiner RC. Studies on dihydrofolic reductase. J Biol Chem 1966;241:1319–1328.
34. Eriksson AE, Baase WA, Wosniak JA, Matthews BW. A cavity-containing mutant of T4 lysozyme is stabilized by buried benzene. Nature 1992;355:371–373.
35. Butina D. Unsupervised data base clustering based on Daylight's fingerprint and Tanimoto similarity: a fast and automated way to cluster small and large data sets. J Chem Inf Comput Sci 1999;39:747–750.
36. Ajay A, Walters WP, Murcko MA. Can we learn to distinguish between "drug-like" and "nondrug-like" molecules? J Med Chem 1998;41:3314–3324.
37. Blaney JM, Hansch C, Silipo C, Vittoria A. Structure–activity relationships of dihydrofolate reductase inhibitors. Chem Rev 1984;84:333–407.
38. Luque I, Gomez J, Semo N, Freire E. Structure-based thermodynamic design of peptide ligands: application to peptide inhibitors of the aspartic protease endothiapepsin. Proteins 1998;30:74–85.
39. Charifson PS, Corkery JJ, Murcko MA, Walters WP. Consensus scoring: a method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. J Med Chem 1999;42:5100–5109.
40. Shoichet BK, Kuntz ID. Matching chemistry and shape in molecular docking. Protein Eng 1993;6:723–732.
41. Santi DV, Danenberg PV. Folates in pyrimidine nucleotide biosynthesis. In Blakely RL, Benkovic SJ, editors. Folates and pterins. New York: Wiley; 1984. p 345–398.