The Effects of Biasing Torsional Mutations in a Conformational GA

Alex Strizhev,† Edmond J. Abrahamian,* Sun Choi,‡ Joseph M. Leonard,§
Philippa R. N. Wolohan, and Robert D. Clark

Tripos, Inc., 1699 South Hanley Road, St. Louis, Missouri 63144

Received May 27, 2005

This paper describes the effects of incorporating torsional bias into a conformational Genetic Algorithm (GA) such as that found in the GASP program. Several major conclusions can be drawn. Biasing torsional angles toward values associated with local energy minima increases the rate of convergence of the fitness function (consisting of energy, steric, and pharmacophoric compatibility terms) for a set of molecules, but a definite tradeoff exists between total model energy and the steric and pharmacophoric compatibility terms in the fitness score. Biasing torsions in favor of sets of angles drawn from low-energy conformations does not guarantee low total energy, but biased torsional sampling does generally produce less strained models than does the uniform torsional sampling in classical GASP. Overall, torsionally biased sampling produces good models comprised of energetically favorable ligand conformations.

INTRODUCTION

There is currently a great deal of interest in developing automatic pharmacophore elucidation software programs, especially in cases where no crystal structure is available. Several are commercially available programs based on different algorithms.[1−6] Among the most popular is GASP,[5−7] which according to Patel et al.,[8] generally performs best overall. The program utilizes a genetic algorithm[9] to identify feature mappings and bond torsions that maximize an overall score that is the sum of total energy, pharmacophoric overlap, and steric overlap to a template molecule.

To favor incremental over abrupt torsional mutation, a Gray table[10] is used to map between torsional values and their internal representation within the GA (see *detailed methodology*). This mapping explores conformational space uniformly, which means that the GA may spend considerable time exploring high-energy conformations.

The idea behind introducing *torsional bias* into the program is to shift from uniform exploration of conformational space toward a more focused exploration. The regions of focus are determined from torsional angles found in randomized conformations that have been relaxed to their local energetic minima. The work centers on the effect of introducing such a bias.

Although GASP was used for the analysis described here, the idea behind torsional bias is general and can be applied to any GA that encodes torsional angles. In fact, this method has been implemented in Tripos' new pharmacophore elucidation and alignment tool GALAHAD.[11,12]

As noted elsewhere in the literature,[13,14] small ligands generally bind in relatively low-energy conformations similar to their "free" conformations, though often not in their lowest energy configuration. This suggests that one would have a better chance of finding appropriate conformers by preferentially exploring the low-energy conformations of flexible ligands, rather than by exploring conformational space uniformly. One way to identify torsional angles corresponding to low-energy conformations is to randomize the torsions in each molecule in the data set several times and then do a torsional minimization of each to the nearest local minimum. A list of the torsional angle values found in the relaxed conformations can then serve as a basis for exploration by a torsional GA. This procedure has now been implemented as an option in the GASP program in SYBYL 7.1.[15]

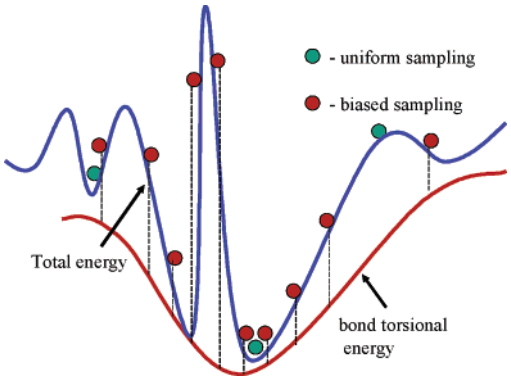

Figure 1. Schematic plot of total energy as a function of torsional angle about one rotatable bond in an otherwise rigid ligand molecule. Green symbols represent uniform sampling, and red symbols represent biased sampling of the torsional profile for this bond.

Figure 1 depicts the kind of energy profile often encountered in a druglike molecule as the torsion around an individual rotatable bond is varied but the parts of the molecule at either end of the bond are kept rigid. The broad underlying curve represents the torsional contribution to the total energy, and the narrow spikes correspond to angles at which long-range intramolecular steric clashes occur. The total number of sampling points across the range of angles

* Corresponding author e-mail: edmond@tripos.com.
† Current address: Amgen, One Amgen Center Drive, Thousand Oaks, CA 91320.
‡ Current address: College of Pharmacy and Division of Molecular Life Sciences, Ewha Womans University, Seoul 120-750, Korea.
§ Current address: Abbott Bioresearch Center, 381 Plantation Street, Worcester, MA 01605.

10.1021/ci0502193 CCC: $33.50
Published on Web 06/10/2006

is the same for uniform sampling (green symbols) and for biased sampling (red symbols), but the sampling frequency is much higher near the local torsional minimum for the latter (red line), where favorable local minima (blue line) are likely to cluster. Thus, the number of sampling points of the biased run falling near the local torsional energy minimum is greater than in the corresponding uniform run, and any solution that lies in that region of torsional space will likely be identified faster—often much faster—when torsional sampling is suitably biased.

The ligand conformation adopted in a binding site represents a balance among many factors.[16] Among the most important factors are torsional energy, hydrogen bonding interactions, intermolecular steric clashes, and steric complementarity to the protein. That conformation is not likely to coincide exactly with the global minimum; its energy may, in fact, lie considerably above the global minimum.[17,18] Nonetheless, every kcal of strain in the bound conformation must be paid for in reduced binding affinity—unless it is compensated for by energy changes in the protein or by entropic factors. Indeed, this is the rationale behind including an energy penalty in the fitness function used in GASP.

Biasing torsions to lie near local minima is therefore a strategy worth considering, since conformations lying near those regions of the internal coordinate space will be relatively low in energy. How far above the global minimum the internal energies of such conformations lie will of course increase with the number of rotatable bonds a ligand contains, just as reported by Perola and Charifson.[17]

One might expect two significant improvements in performance from biased torsional sampling: (1) the GA may converge more quickly, since the GA is already biased in favor of good models; and (2) models involving conformations lying in steep narrow minima of the sort illustrated in Figure 1 are less likely to be missed, since biased sampling of the torsional space necessarily favors consideration of such conformations.

There are, however, risks involved in using biased torsional sampling. As the GA runs, each torsion in a ligand mutates independently of all the others. The combinations of torsions selected for are constrained by overall fitness but not by the combinations of torsions found in the conformations from which the individual torsional profiles were constructed. The individual torsional values found for different rotatable bonds in the GASP models ultimately produced come, in general, from different starting conformations; no attempt is made to maintain the relationship between them. Combining torsional angles that were minimized in different contexts therefore sometimes results in steric clashes and high total energies. This strain can usually be relieved by subsequent local relaxation without greatly affecting overall conformation. Hence, good values for the other fitness terms are minimally affected by such relaxation as well.

Because biased sampling focuses on the angles included in the corresponding torsional profiles, regions of the internal coordinate space lying outside the combinations of those angles are necessarily explored less thoroughly. If a good model exists in such an area, uniform sampling may identify it more quickly.

There is also a risk of overfocusing on the exact torsional angles in the bias profile and skipping values between, which may overlook less obvious but potentially viable solutions that lie nearby. To address this problem, a Gaussian perturbation can be applied to each torsional angle in each model, in such a way that all angles lying between adjacent values in the profile were accessible to the GA. Hence angles lying in torsional ranges corresponding to large gaps in the respective profiles were less thoroughly sampled than those lying between closely spaced values from the profile, but they did have a chance of being considered.

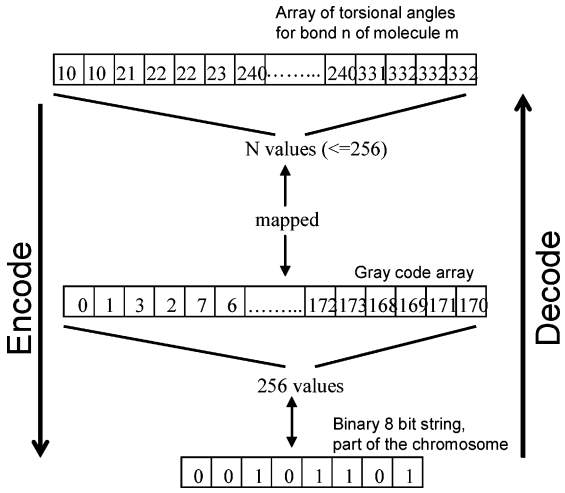

Figure 2. Schematic illustrating how torsional bias works with Gray coding in GASP.

DETAILED METHODOLOGY

Ligand conformations are manipulated as binary chromosomes, with torsion values encoded in chromosomes as 8-bit words. The 8-bit range of integers (0−255) is mapped onto the full 360 degree range of rotation. As the GA progresses, chromosomes exchange torsional information via crossovers.[9] Mutations are carried out by picking one bit in the chromosome at random and "flipping" its value—to 0 if it is 1 and to 1 if it is 0. Since the choice of which bit to flip is random, the most significant bits of an 8-bit word are flipped with the same probability as the least significant ones. The first case would lead to an extreme change in torsion, while the second would be less dramatic. Gray table[10] mapping is used in GASP to alleviate this problem. This approach ensures that most changes are incremental, with the size of larger changes following a roughly Gaussian binomial distribution. In classical GASP, this mapping of 256 binary values to 360 degrees of rotation is uniform, as is the resultant exploration of conformational space.

Torsional bias is introduced into this Gray coding scheme by using the decoded Gray value as an index into a list of favored torsional values—the torsional profile—rather than mapping to an angle directly. The mapping scheme is presented in Figure 2.

The top string consists of the *N* torsional angle values obtained for a specific bond in a specific ligand molecule during conformational preprocessing. Note that some torsional angle values (e.g., 10° and 22° in Figure 2) may appear more than once in a profile and often do. This is to be expected, as two or more different conformations will minimize to the same torsional angle around a given bond if steric clashes are absent. Duplicate values serve as a weighting factor, an angle with more duplicates being more

likely to be present in the final solution. The maximum value of *N* is 100 and the size of the Gray code array is 256, so the angle values are mapped one-to-one onto the Gray code array, wrapping around until the Gray code array is exhausted. This does not disrupt the overall completeness of torsional sampling since the *N* angle values are sorted and cover the range of 0−360°. *N* cannot exceed the size of the Gray code array— 256—, however. The mapping of a binary 8-bit string to the Gray code array is one-to-one, with the decimal translation of the 8-bit string serving as an index into the torsional profile (Figure 2).

The torsional profiles were constructed in three steps: (1) One hundred (100) conformations were generated for each member of the data set using the *DCF* utility in SYBYL, which randomizes the torsion about each rotatable bond in a molecule. No ring flexing takes place. (2) The conformations resulting from step 1 are then relaxed to their local minima using a torsional-space energy minimization based on the Tripos force field.[19] Only torsional changes are allowed during the minimization, and only the torsional and vdW energy components of the force field are taken into consideration. (3) The torsional angles found for each rotatable bond in the various conformers are compiled into torsional profiles, one for each rotatable bond in the data set.

The granularity at which conformational space is explored when torsional sampling is biased is less in some ranges than when sampling is uniform. For example, for the profile illustrated in Figure 2, there is a possibility for a torsional angle changing its value from 23 to 240 degrees due to an incremental mutation in the index. This is because no local energy minima were found between 23 and 240 degrees for that bond. To better cover the continuum of space when the number of entries in the profile is relatively low, selected torsional angles are perturbed by a Gaussian function that is scaled to span the intervals between each angle in the profile and the adjacent ones. This is done after decoding the torsional angle value from the chromosome and before scoring the resulting model. The Gaussian perturbations do not affect the chromosomes themselves but rather express perturbations of the specific values encoded by the genes.

This speeds up the sampling of conformational space overall. If a torsion lying in that range of values is required to get a particular model, however, the likelihood of it being found is lower than with uniform sampling of conformational space.

We used several different criteria to evaluate the quality of the GASP models we obtained: total fitness score of the final model, total energy of the molecules in the final alignment, and the number of iterations required for GASP to converge. These statistics are all produced by GASP in its normal operating mode.

The GASP score is the fitness function used to drive the GA in GASP as well as the scoring function used to rank final models. It is a linear combination of three terms: (1) number and compatibility of overlaid features, (2) volume overlap of other ligands to template molecule (i.e. common volume of aligned molecules), and (3) internal strain (torsional energy and 1−4 van der Waals interactions) of molecular conformations. These terms are described in detail elsewhere in the literature[5,6] and will not be discussed further here. Note that though the score is only defined with respect to a particular data set and template molecule and has no absolute meaning, it can be useful for comparing different alignments created from the same data set.

Two of the criteria used here to compare alignments—fitness score and total energy—are interrelated. Indeed, internal energy is one of the terms used to compute the fitness score. The third criterion—number of iterations required for convergence—requires further explanation. GASP will terminate a run once the number of models evaluated ("operations") exceeds a user-defined maximum, but it will also terminate when the fitness of the best individual in the population has not increased by at least[6] 0.01 in the past 6500 evaluations. Obviously, the effectiveness of the GA can only be measured by the number of iterations required for convergence if a stable fitness value is reached before the maximum allowed number of models has been evaluated. Setting the "maximum number of operations" sufficiently high (here, to 1 million) ensures that the GA only stops because a stable fitness value has been reached. Then the number of operations carried out is a valid measure of how effective the GA is.

RESULTS AND DISCUSSION

We used sets of thrombin, CDK2, thermolysin, DHFR, and HIV-1 reverse transcriptase (RT) inhibitors compiled by Patel et al.[8] to compare the performance of GASP using uniform torsional sampling to that seen when torsional sampling was biased.

The genetic algorithm used in GASP produces a number of alignments (models). GAs are stochastic by nature, so GASP generally produces somewhat different alignments depending on the random number series used in a given run. The union of the alignments produced constitutes a so-called *GASP solution space*. Some seed values may yield much better models than others, in which case the GASP solution space will be narrow and is likely to include the experimentally observed solution. Other seeds may yield poorer models. The number of generations, population size, and mutation rate can also affect the quality of the results obtained. For

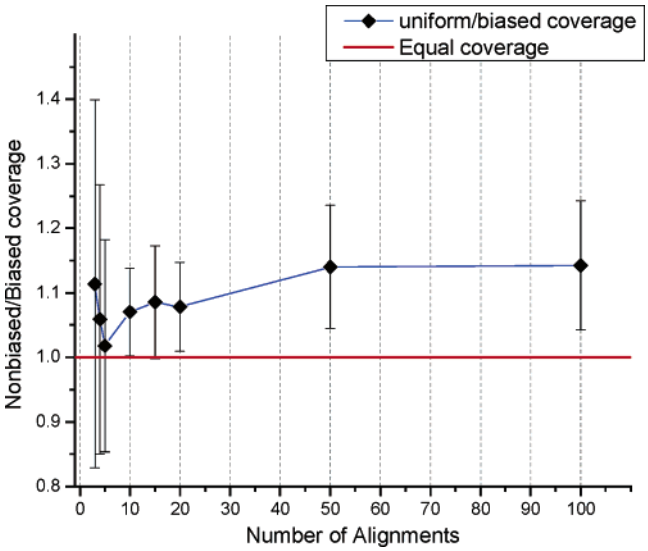

Figure 3. Pharmacophore coverage in terms of the ratio of the number of pharmacophore triplets found in GASP models generated using uniform torsional sampling to the number found in models generated using biased sampling.

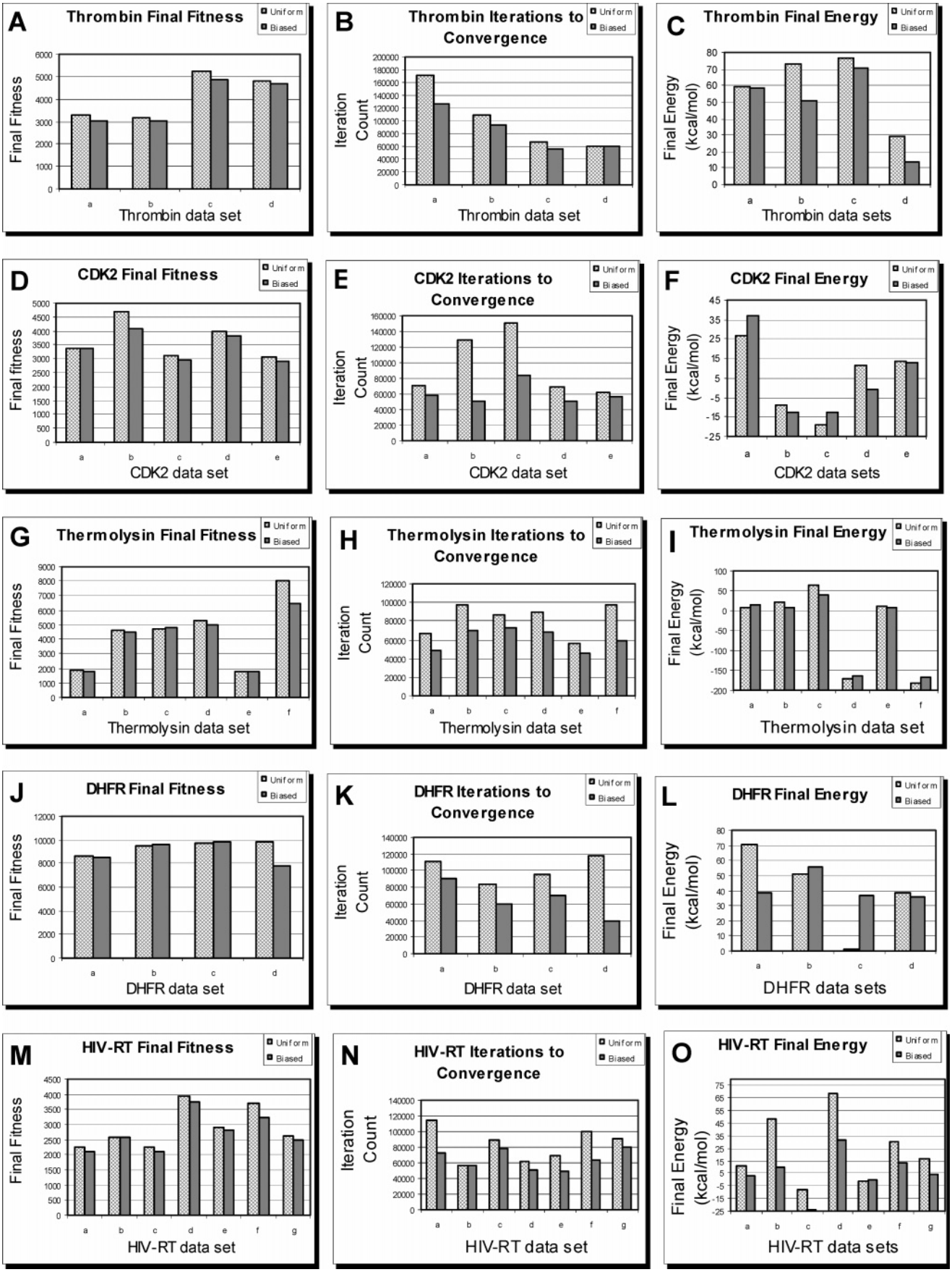

Figure 4. Comparison of GASP results with uniform and biased torsional sampling. The five panels to the left are plots of average final fitness score, central panels show the average number of iterations required for convergence when the maximum number of operations allowed was set to 1 million, and the panels on the right show the average model energy: (**A**−**C**) thrombin inhibitors, (**D**−**F**) CDK2 inhibitors, (**G**−**I**) thermolysin inhibitors, (**J**−**L**) DHFR inhibitors, and (**M**−**O**) inhibitors of HIV-1 reverse transcriptase. The ligands corresponding to each of the data sets used are enumerated in Table 1.

an accurate comparison of the results produced by different runs it is necessary to process a statistically significant number of GASP results, i.e., to generate an adequately large GASP solution space. Determining the number of results required, i.e., how many alignments are required to saturate the GASP solution space, is not straightforward. We started with 100 alignments, i.e., GASP was set to perform 100 consecutive but independent runs.

GASP was first run on a medium-size data set (the RT inhibitors extracted from complexes **1bqm**, **1dtt**, **1ep4**, **1fk9**, **1klm**) in both biased and uniform modes. The conformations of these five ligands were collected in separate hit lists so that each hit list contained different conformations of one ligand. The conformational diversity of these hitlists was then measured using the *dbconform*[20] utility in SYBYL. This program operates on pharmacophore triplet fingerprint metrics[20] and calculates the percentage of pharmacophoric space covered by each hitlist. Next, the ratio of uniform over biased coverage was calculated. The result is shown in Figure 3. Since the biased method is based on a focused and narrower exploration of conformational space, it was expected that such a ratio should be greater than one. The limiting difference observed is gratifyingly small, however, indicating that torsional bias does not unduly limit the range of pharmacophore solutions explored by GASP.

As can be seen from the plot, consistency was achieved after 40 alignments were averaged. Therefore GASP was subsequently set to run 45 times in succession. To randomize the results further, we split those 45 runs into three groups of 15, each with its own initial random number seed.

The results of these analyses (for thrombin inhibitors, CDK2 inhibitors, thermolysin inhibitors, DHFR inhibitors, and inhibitors of HIV-1 reverse transcriptase) are grouped and plotted to show the fitness, energy, and number of iteration differences between uniform and biased runs. Results are presented in Figure 4.

As seen in the right-hand column of panels in Figure 4, the average energy for biased runs was consistently lower than for uniform torsional sampling, as expected. The overall fitness scores from biased runs are somewhat lower as well. Examination of the individual models showed that this reflects the fact that pharmacophorically simpler models are consistently lower (often much lower) in energy than are more complex, higher scoring models. This tradeoff is an inevitable result of including antagonistic criteria in a single fitness function.[16] A second important result is that GASP converged considerably faster when torsional sampling was biased (central panels in Figure 4), with convergence being 28% faster (averaging across all data sets). Per data set average convergence improvement ranged from 15% to 35%.

Incidental Steric Clashes and Proximity to Local Minima. Applying a set of individually favorable torsions to a molecular structure generally yields a low-energy conformation, but there is no guarantee that this will always be the case. On rare occasion, the internal energy may be high—sometimes extremely high—due to steric bumps within the molecule. One illustrative example is discussed in detail here. Thrombin inhibitors from pdb complexes **1d9i**, **1dwd**, and **1d4p** were extracted and aligned in GASP using the conventional uniform torsional sampling method or using torsional bias. Based on similarity between the models and reasonableness of the alignments into consideration, we selected one representative model from each analysis for further comparison. The total energies of ligands from the selected models are listed in Table 2.

Table 1. Ligands Used in Each of the Data Sets that Appear in Figure 4[a]

data set	ligand	data set name in Figure 4
thrombin	1tom; 1d6w; 1c4v; 1fpc	a
	1d9i; 1dwd; 1d4p	b
	1d9i; 1dwd	c
	1c4v; ***1tom***	d
CDK2	1e1v; 1elx; 1fvv	a
	1aq1; 1fin	b
	1di8; 1aq1; 1fin	c
	1di8; 1fvv	d
	1e1v; ***1e1x***; 1aq1	e
DHFR	1drf; 2dhf	a
	1hfp; 1ohk	b
	1boz; 1dlr	c
	1hfp; 1ohk	d
HIV-RT	1bqm; 1fk9; 1klm	a
	1dtt; 1rt5	b
	1rt3; 1vru	c
	1ep4; 1klm	d
	1tvr; 1rt1	e
	1rt5; 1dtt; 1klm	f
	1fk9; 1bqm; 1klm	g
thermolysin	7tln; 1hyt	a
	1qf1; 5tln	b
	4tmn; 5tln	c
	1qf1; 5tmn	d
	7tln; 1hyt	e
	4tmn; 5tmn	f

[a] The ligands are subsets of the thrombin, CDK2, DHFR, HIV-RT, and thermolysin data sets taken from ref 8. The bold, italicized ligands were used as template (rigid) molecules in the GASP runs.

Table 2. Total Energies of Thrombin Inhibitors for Comparative GASP Runs

source of structure	calculated total energy (kcal/mol)	
	uniform torsional sampling	biased torsional sampling
1d9i	161	182
1dwd	73	521
1d4p	70	59

Ligands from the thrombin complexes **1d9i** and **1d4p** do not differ significantly in total energy, while the ligand taken from **1dwd** is much higher in energy for the model produced using torsional bias. Detailed investigation of this ligand in the conformations produced by the two methods revealed a steric clash in the conformation produced by biased torsional sampling (Figure 5).

The clash results from incompatible torsional angles and is readily relieved by energy minimization. The relaxed conformations were then aligned manually to the original, strained conformations with the result depicted in Figure 6. Relaxation sharply reduced the total energy in both cases. This was achieved without departing very much from the original ligand conformation for the model obtained using biased sampling, so the good, nonenergetic components of the corresponding GASP score are still reasonably applicable. In contrast, the model obtained with uniform torsional sampling changed conformation dramatically upon relaxation, enough so that the original GASP score elements are no longer applicable. In fact, the overall shape of the relaxed uniform sampling model is remarkably similar to the model

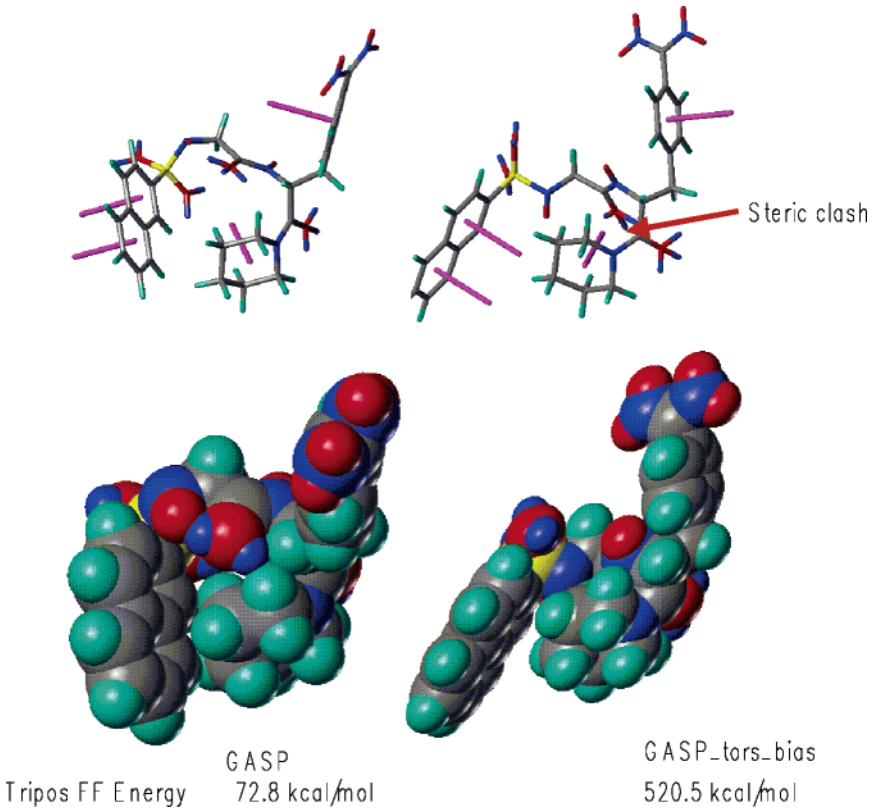

Figure 5. Conformations for the ligand from **1dwd** generated by GASP using uniform torsional sampling (left) and biased torsional sampling (right).

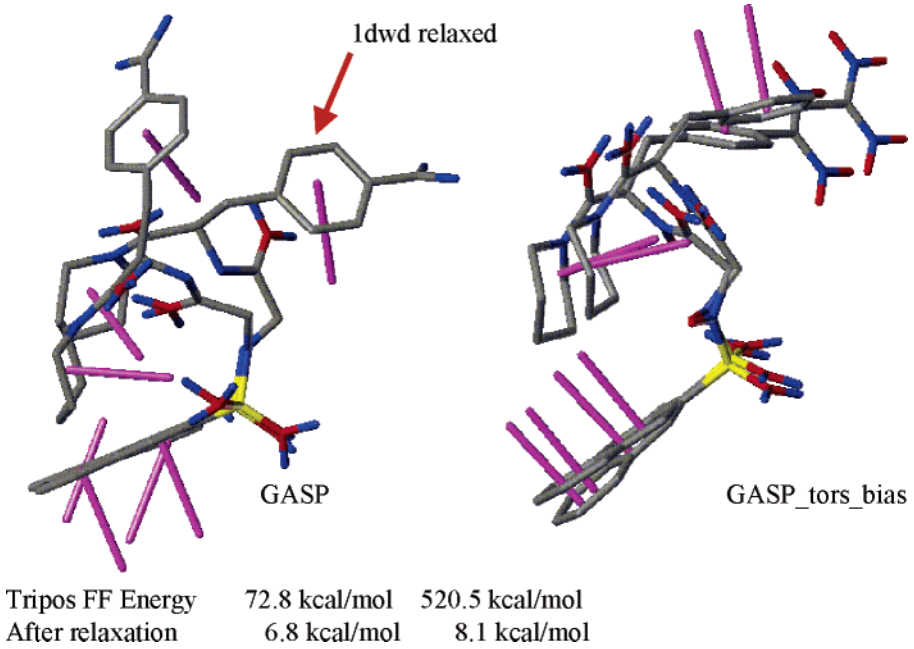

Figure 6. Conformations of the **1dwd** ligand from the models obtained with uniform torsional sampling (left) and using biased sampling (right). The original and torsionally relaxed conformations are shown, along with the respective total energies.

obtained using biased torsional sampling (Figure 6). This example illustrates the point that the models produced from torsionally biased GASP analyses are likely to lie near very similar relatively low-energy models in the solution space even when they are themselves strained due to incidental close contacts.

In fact, such close contacts are not frequently encountered in biased sampling experiments, but they do occur often enough to distort the average energies shown in Figure 4 somewhat. Given that GASP is attempting to generate bound conformations, steric "close calls" may be nearly unavoidable. After all, any molecular regions not involved in interactions between ligand and protein are likely to fold in as tightly as possible to minimize exposed surface area.

Some of the models used to construct Figure 4 were re-examined in light of the example discussed above. Once the GASP runs were complete, the conformations produced were subjected to total energy minimization in SYBYL. The results

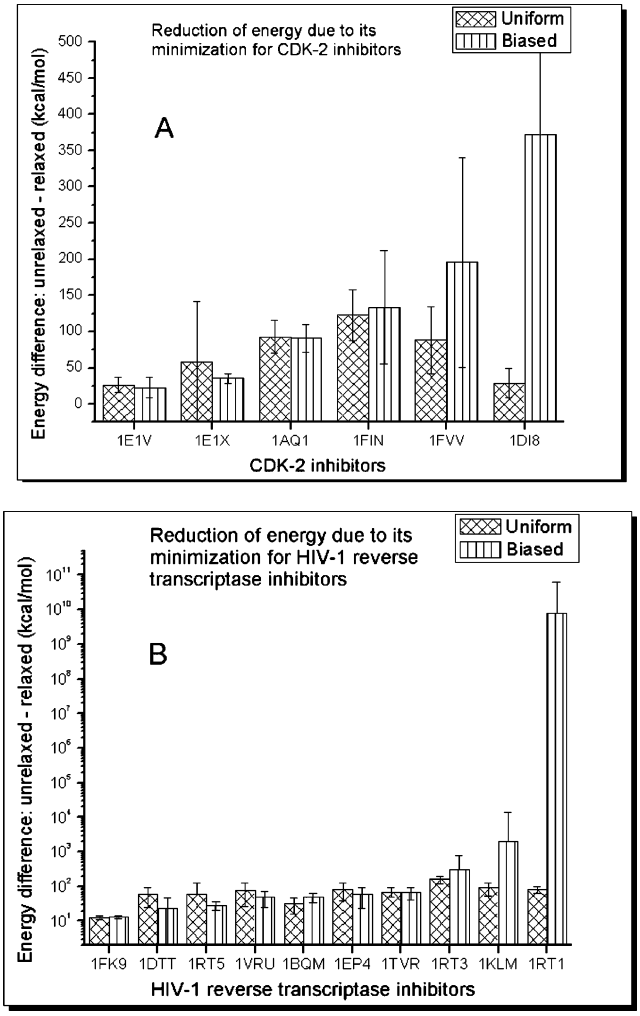

Figure 7. Effect of postrun energy minimization on GASP alignments: (A) CDK-2 inhibitors and (B) HIV-1 reverse transcriptase inhibitors.

are shown in Figure 7 and in Table 3. The histogram presents the reduction of strain energy of CDK-2 and HIV-1 RT ligands achieved by energy minimization using the MaxiMin command in SYBYL using the Tripos force field[19] after the respective GASP runs were complete. Table 3 presents the means along with the root-mean-square deviation (RMSD) between the original and minimized conformations. Note the large energy reduction achieved for CDK2 ligand **1di8** after energy minimization and for HIV-1 RT ligands **1klm** and **1rt1** (Figure 7A) (the energy scales in Figure 7B are logarithmic to accommodate the very large energy differences attributable to van der Waals clashes).

As these examples illustrate, "incompatible" torsional angles applied together can result in severe steric clashes in a conformation. This is generally relieved by the GA itself but can be more efficiently relieved by subsequent energy minimization without unduly disturbing the geometry of the original model. This is reflected in the relatively small changes seen in RMSDs (Table 3).

The quantitative data presented above suggests higher effectiveness of the biased approach over the regular Gray coding approach in most cases. The mean strain energy relief for models obtained using uniform sampling is greater than that for biased sampling for 12 of 19 ligands (63%). Moreover, in most cases where models obtained using

Table 3. Energy Minimization of GASP Models Obtained from Uniform and Torsionally Biased Runs

	uniform torsional sampling				biased torsional sampling			
ligand	ΔE (kcal/mol)		RMSD (Å)		ΔE (kcal/mol)		RMSD (Å)	
source	mean	SD	mean	SD	mean	SD	mean	SD
				Thrombin				
1d9i	71	65	0.33	0.10	36	30	0.25	0.07
1tom	491	27	0.27	0.07	48	46	0.22	0.08
1d6w	731	26	0.36	0.10	73	47	0.33	0.08
1fpc	118	45	0.35	0.08	155	135	0.35	0.10
1d4p	306	40	0.24	0.07	279	15	0.17	0.05
1c4v	336	36	0.25	0.06	332	492	0.23	0.05
1dwd	360	49	0.31	0.07	408	243	0.27	0.09
				DHFR				
1boz	35	23	0.17	0.06	33	12	0.16	0.04
1ohk	78	31	0.29	0.06	77	63	0.30	0.05
1drf	103	62	0.30	0.07	78	57	0.27	0.06
1hfp	95	96	0.30	0.08	103	86	0.28	0.07
2dhf	145	64	0.37	0.08	113	50	0.29	0.07
1dlr	110	77	0.24	0.07	182	67	0.27	0.05
				Thermolysin				
1hyt	13.3	8.3	0.43	0.12	20	17.2	0.33	0.09
5tln	46	33	0.29	0.08	32	22	0.27	0.09
7tln	27.1	11.8	0.26	0.07	37	16.6	0.21	0.05
1qf1	50	34	0.30	0.08	41	35	0.26	0.07
4tmn	136	69	0.40	0.10	122	87	0.32	0.10
5tmn	127	58	0.40	0.09	169	170	0.31	0.08

uniform sampling were less strained on average (e.g., **1c4v** and **1dwd** in Table 3), the standard deviations are much larger for biased sampling, an observation consistent with the high mean being due to a few instances of incidental steric clashes among the models obtained. Such occasional steric bumps in resulting conformations can in general be relieved with a minimal conformational change by total energy minimization. Thus, total postrun energy minimization can be very helpful for a biased run and used together with biased runs produces results superior to the conventional run using uniform torsional sampling.

Pharmacophore Quality. Qualitative comparison of GASP models is necessarily subjective and therefore difficult.[8,16] Illustrative examples are useful; however, flexible alignment of sulfasalazine and folic acid is just such an example. This is a difficult case because of ambiguities in pharmacophore feature mapping and in the high flexibility of the two molecules. We performed both uniform and biased GA runs on this data set, with results presented in Figure 8.

The model produced using torsional bias (the bottom panel in Figure 8) is clearly superior to that from "classical" GASP (top) in terms of pharmacophoric feature overlay, largely due to the inclusion of better torsions for the bonds surrounding the sulfonamide group. The sulfonamide group is not as strained as in the case of the uniform run and donors and acceptors are better aligned. In addition, the distal glutamyl substituent that is not a part of the pharmacophore has an energetically strained configuration position in the uniform run but is nicely extended in the biased run, to yield a low-energy conformation. Finally, the biased run produced lower overall energy conformations than the uniform run.

Tolerated Strain vs Flexibility. Perola and Charifson recently reported that bound ligand conformations may exhibit considerable intramolecular strain[17] and noted that the amount of strain tolerated increases with increasing ligand flexibility. This might be taken to suggest that torsional

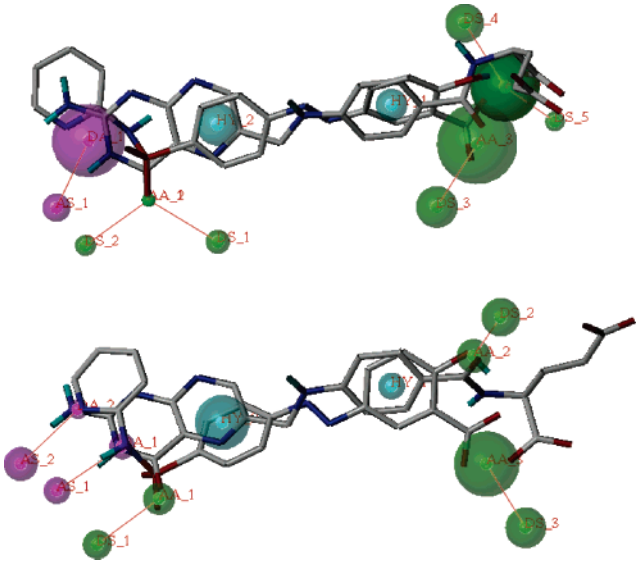

Figure 8. GASP alignments of sulfasalazine and folic acid. Donor atoms and their associated site points are colored magenta, acceptor atoms and associated site points are colored green, and hydrophobic centers are colored green. (Top): model obtained using conventional, uniform torsional sampling. (Bottom): model obtained using biased torsional sampling.

Table 4. Fitness Scores of Final Alignments for GASP

	regular run fitness			biased run fitness		
data set	mean	SDEV	max	mean	SDEV	max
		Flexible ↔ Rigid				
Krystek	6305	1642	8674	5733	1518	8618
dopamine	3285	161	3543	3136	117	3305
angiotensin II	4607	762	5897	4675	624	5742
combretastatin	3580	463	4824	3795	491	5123

biasing would be less advantageous for flexible ligands. That would certainly be so if conformations were being forced to the global minima, but the purpose of biased torsional sampling is to reduce the tendency of classical GASP to produce conformations so torsionally strained that they have no possibility of binding to the target protein.

To address this question, data sets ranging from fairly rigid to rather flexible were analyzed using classical and torsionally biased GASP. The endothelin[21] and dopamine D_4 receptor antagonists[22] data sets are comprised of relatively rigid ligands, whereas the angiotensin II receptor agonists[23] and combretastatin analog[24] data sets were chosen to represent more flexible ligands. Six molecules from the endothelin and dopamine data sets were selected, and GASP was set to run 50 times (producing 50 alignments) in either uniform or biased torsional sampling mode. The criterion used to compare the runs was the fitness score of each final alignment. Table 4 lists the mean and maximum fitness scores obtained, sorted by increasing flexibility. "Flexibility" is used rather subjectively here, since the flexibility of a molecule depends not only on the number of rotatable bonds but also on the topology of the molecule as well. Hence this experiment serves as an illustration, rather than as an exact conclusive study. For this purpose, the "flexibility" of the data set was evaluated in terms of the number of rotatable bonds that determine the position of pharmacophore points in space. Results are presented in Table 4.

These results provide quantitative support for the qualitative conclusion drawn from the folic acid/sulfasalazine example: biased torsional sampling outperforms the uniform torsional sampling in conventional GASP based on fitness value for more flexible data sets as well as in rigid ones.

Note that these results do not contradict the conclusions of Perola and Charifson. Rather, they suggest that strain energy is distributed more or less evenly across all rotatable bonds in bound ligand molecules, with each lying relatively close to a torsional minimum. It follows that more total strain energy is tolerable with a rising number of rotatable bonds. There may also be an entropic effect that reduces the overall free energy of binding, since the vibrational ensemble for torsions near the minimum is expected to have a lower entropic penalty than torsions farther away from the underlying torsional minimum (cf. Figure 1).

CONCLUSIONS

The data presented here demonstrate that the introduction of biased torsional sampling increases the rate of convergence of a torsional GA of the sort used in GASP. Comparison of the number of operations it takes the GA to converge with biased and uniform torsional sampling suggests that this occurs because biased sampling lets the GA skip exploration of large areas of torsional space where the likelihood of finding good solutions is small. Failure to do so is a major disadvantage of uniform torsional sampling: since most torsional mutations result in a small change of torsional angle, once you find yourself in a high-energy area, it takes many instances of torsional angle mutation to find the way out. One may think of biased torsional sampling as fencing around high-energy regions in the GASP solution space. It is important to make sure that the "fence" is not impenetrable and can be escaped if necessary. This is accomplished by way of Gaussian perturbations about the preferred angles in the torsional profile. Thus it is possible to get out of low-energy areas, but you can quickly find your way back if doing so is not productive.

Bias toward low torsional energies does not guarantee that the total energy of all resulting conformations will be low, because the torsional profiles do not take interactions between distal torsions into account directly—the bias is toward torsional energy minima and cannot take full account of topologically long-range steric clashes. Therefore, the total steric energy of a molecule in the alignment may turn out to be quite high in some cases. When that does happen, however, it can generally be relieved by extending the GA run or by relaxing the model conformations to nearby local minima. Such relaxation can generally be accomplished without significantly changing the model as a whole, provided the models have been obtained using biased rather than the uniform sampling conventionally used in torsional genetic algorithms such as GASP.

On the whole, however, biased torsional sampling is a good way to improve the performance of GASP, with speed of convergence and preferential generation of more energetically favorable conformations both benefiting.

GOING FORWARD

Biased torsional sampling appears to be a very promising technique, but the usefulness of its implementation in GASP is somewhat limited by the need to apply an "external" Gaussian perturbation to ensure that angles lying between

those in the bias profile are inadequately explored. This problem can be addressed by using a second 8-bit gene to encode the perturbation instead. Such a bit string would be an integral part of the chromosome and so would be subject to regular GA selection. Just such an approach has been taken in GALAHAD.[11]

Another possible modification to be examined is the generation of bond torsional profiles based on the atomic environments of the corresponding rotatable bonds, rather than by preprocessing using torsional randomization and torsional energy minimization. This would certainly be faster and may provide more thorough sampling as well as greater flexibility in terms of the force field engine used.

In should also be noted that the advantages of biased over uniform torsional sampling are by no means limited to cases where torsional genes are manipulated as binary Gray codes. That is the embodiment chosen for the work presented here, and it is a very efficient one. There is no reason to think that torsions encoded directly as real values would perform any differently, however.

ACKNOWLEDGMENT

The authors wish to thank Evert Homan, Anna-Lena Gustavsson, Peter Brandt, Jerk Vallgårda, and Maria Wirstam of the Biovitrum AB computational chemistry team for valuable input and suggestions.

REFERENCES AND NOTES

(1) Martin, Y. C. DISCO: What We Did Right and What We Missed. In *Pharmacophore Perception, Development, and Use in Drug Design*; Güner, O. F., Ed.; International University Line: La Jolla, CA, 2000; pp 49−68.
(2) Martin, Y. C.; Bures, M. G.; Danaher, E. A.; DeLazzer, J.; Lico, I.; Pavlik, P. A. DISCO: A Fast New Approach to Pharmacophore Mapping and its Application to Dopaminergic and Benzodiazepine Agonists. *J. Comput.-Aided Mol. Des.* **1993**, *7*, 83−102.
(3) Barnum, D.; Greene, J.; Smellie, A.; Sprague, P. Identification of common functional configurations among molecules. *J. Chem. Inf. Comput. Sci*. **1996**, *36* (3), 563−571.
(4) Clement, O. O.; Mehl, A. T. HipHop: Pharmacophores Based on Multiple Common-Feature Alignments. In *Pharmacophore Perception, Development, and Use in Drug Design*; Güner, O. F., Ed.; International University Line: La Jolla, CA, 2000; pp 69−84.
(5) Jones, G.; Willett, P.; Glen, R. C. GASP: Genetic Algorithm Superposition Program. In *Pharmacophore Perception, Development, and Use in Drug Design*; Güner, O. F., Ed.; International University Line: La Jolla, CA, 2000; pp 85−106.
(6) Jones, G.; Willett, P.; Glen, R. C. A genetic algorithm for flexible molecular overlay and pharmacophore elucidation. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 532−549.
(7) GASP (Genetic Algorithm Similarity Program) developed at the University of Sheffield and the Wellcome Research Laboratories. Distributed by Tripos Inc., St. Louis, MO as a module in SYBYL version 7.1, 2005.
(8) Patel, Y.; Gillet, V. J.; Bravi, G.; Leach, A. R. A Comparison of the Pharmacophore Identification Programs: Catalyst, DISCO and GASP. *J. Comput.-Aided Mol. Des*. **2002**, *16*, 653−681.
(9) Holland, J. H. *Adaptation in Natural and Artificial Systems*; University of Michigan Press: Ann Arbor, MI, 1975.
(10) Gray, F. Pulse Code Communications, U.S. Patent No. 2632058, Mar 1953.
(11) Clark, R. D.; Abrahamian, E. J.; Abrams, C.; Brandt, P.; Gustavsson, A. L.; Homan, E.; Richmond, N.; Strizhev, A.; Wirstam, M.; Wolohan, P. Manuscript in preparation.
(12) GALAHAD (Genetic Algorithm with Linear Assignment for Hypermolecular Alignment of Datasets), version 1.2.3 available from Tripos, Inc., St. Louis, MO, 2005.
(13) Moodie, S. L.; Thornton, J. M. A study into the effects of protein binding on nucleotide conformation. *Nucleic Acids Res.* **1993**, *21*, 1369−1380.
(14) Bostrom, J.; Norrby P. O.; Liljefors, T. Conformational energy penalties of protein-bound ligands. *J. Comput.-Aided Mol. Des*. **1998**, *12*, 383−96.
(15) SYBYL molecular modeling software version 7.1 available from Tripos Inc., St. Louis, MO, 2005.
(16) Cottrell, S. J.; Gillet, V. J.; Taylor, R.; Wilton, D. J. Generation of multiple pharmacophore hypotheses using multiobjective optimisation techniques. *J. Comput.-Aided Mol. Des.* **2005**, *18*, 665−682.
(17) Perola, E.; Charifson, P. S. Conformational analysis of drug-like molecules bound to proteins: an extensive study of ligand reorganization upon binding. *J. Med. Chem*. **2004**, *47*, 2499−2510.
(18) Thoma, J. A.; Koshland, D. E. Competitive inhibition by Substrate during enzyme action evidence for the induced-fit theory. *J. Am. Chem. Soc*. **1960**, *82* (13), 3329−3333.
(19) Clark, M.; Cramer, R. D.; Van Opdenbosch, N. Validation of the general purpose Tripos 5.2 force field. *J. Comput. Chem*. **1989**, *10*, 982−1012.
(20) *dbconform*, Unity and SYBYL versions 7.1 are distributed by Tripos Inc., 1699 South Hanley Road, St. Louis, MO 63144, U.S.A., 2005 (http://www.tripos.com).
(21) Krystek, S. R., Jr.; Hunt, J. T.; Stein, P. D.; Stouch, T. R. Three-dimensional quantitative structure−activity relationships of sulfonamide endothelin inhibitors. *J. Med. Chem*, **1995**, *38*, 659−669.
(22) Bostrom, J.; Bohm, M.; Gundertofte, K.; Klebe, G. A 3D QSAR study on a set of dopamine D4 receptor antagonists. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (3), 1020−1027.
(23) Krovat E. M.; Langer T. Non-peptide angiotensin II receptor antagonists: chemical feature based pharmacophore identification. *J. Med. Chem*. **2003**, *46* (5), 716−726.
(24) In H23 data set, http://dtp.nci.nih.gov/docs/cancer/cancer/data.html, accessed August 2005.

CI0502193