Computational screening of combinatorial libraries via multicopy sampling

Qiang Zheng and Donald J. Kyle

Traditionally, computer-aided drug design generates, optimizes and evaluates ligands for binding to a target molecule, one ligand at a time. The advent of high-throughput technologies and the current trend of miniaturization and integration of automated synthesis and screening have stimulated widespread interest in the development of new computational tools for estimating molecular diversity, searching chemical databases and designing ligands. The biggest challenge has always been to improve the accuracy and reliability of such tools. The authors describe a new computational method, which allows simultaneous processing of a mixture of ligands, and may offer new solutions in computer-aided drug design.

Combinatorial chemistry and high-throughput screening have stimulated widespread interest in the development of new computational tools for estimating molecular diversity, searching chemical databases and designing ligands. Because the development of most of these tools has been recently reviewed in this journal[1,2], we have chosen to discuss a relatively new and less known computational method for ligand design and screening. Specifically, we review the locally enhanced sampling (LES)[3] and multiple copy simultaneous sampling[4] methods, which for simplicity are uniformly referred to here as the multicopy sampling method. Our focus is on the various applications of the method, instead of a detailed comparison between this and other methods.

Computer-aided drug design often requires evaluating the binding of many ligands to a given molecular target, such as a protein. The binding affinity of each ligand–protein complex is largely determined by its lowest free energy conformations which, in principle, can be identified by searching through all conformations. In practice, however, the conformational flexibility of the ligand and protein usually yields too many conformations for an exhaustive search. The inherent uncertainty of an energy function may lead to an inaccurate and unreliable estimation of free energy. Moreover, these two complexities are coupled: the free energy calculation relies on extensive conformational sampling, which can be meaningless without an accurate energetic evaluation.

A widely practised and successful approach to improve sampling efficiency and energetic evaluation is to use empirical rules to guide conformational sampling and to simplify energetic evaluation. Multicopy sampling represents a different approach to achieving these goals without resorting to empirical rules. It relies on the fundamentals of molecular mechanics; it is thus the most generic, and yet offers substantial speed-up over the traditional single-copy sampling approach.

Multicopy sampling was first introduced by Elber and Karplus for studying ligand diffusion in myoglobin[3,5–10]. Molecular dynamics was used to simultaneously simulate the diffusion of 60 randomly placed conformations (copies)

Qiang Zheng* and **Donald J. Kyle**, Scios Inc., 820 West Maude Avenue, Sunnyvale, CA 94086, USA. *tel: +1 408 523 7215, fax: +1 408 481 9188, e-mail: qiang@netcom.com

Copyright © Elsevier Science Ltd. All rights reserved. 1359-6446/97/$17.00. PII: S1359-6446(97)01046-5

of carbon monoxide. The time required for the simulation was comparable to what is required for the simulation of a single ligand. Subsequently, multicopy sampling has been applied to binding site mapping[4,11–14], ligand docking[12,13], free energy calculation[15], protein side chain placement[16–18], modeling of protein loops[19–21] and small ligands[22–24], and screening of combinatorial libraries[25] (Box 1). Moreover, the multicopy sampling method has been continuously improved[26–29]. Recently, an experimental complement to multicopy sampling was reported, wherein heterogeneous organic solvents were cocrystallized with the enzyme elastase for simultaneous identification of the solvent binding sites[30].

A multicopy simulation involves a fictitious complex, consisting of a protein and replicated copies of one or several different ligands. A ligand can also be a small segment of the protein. The copies are energetically transparent to one another, while each interacts normally with the protein. The protein experiences the average force from all copies. The efficiency of multicopy sampling stems from the fact that the protein self-interaction is evaluated once for all the copies. In contrast, traditional single-copy sampling requires repeated evaluation of the protein self-interaction for all the copies, even if the protein only undergoes minor conformational change upon binding to a ligand. It has been shown that the efficiency gain is $N/2n$-fold, where N and n denote the number of atoms in the protein and ligand respectively[19,28]. Since a protein is usually much larger than a ligand, the efficiency gain by using multicopy sampling is substantial. Perhaps the most unique feature of multicopy sampling is the partially smoothed energy function[16,17], which tends to be more effective for both structural evaluation and conformational sampling.

In this review, we examine three applications of the multicopy sampling method to the design and screening of combinatorial libraries. The concept of combinatorial libraries adopted is an extension of the traditional one of a mixture of ligands; here, a library consists of a mixture of conformations of one or several different ligands. This extension is useful, since a ligand is usually modeled via its conformations. We end the review by discussing some of the limitations of the method, together with several issues that are critical to its future development.

Mapping of ligand binding sites

A simple fact of ligand binding is that a high-affinity ligand tends to consist of high-affinity fragments that bind specifically to the key areas in a binding site, and vice versa. A small set of high-affinity fragments can lead to a large number of high-affinity ligands because there may be numerous ways to connect the fragments to form a larger ligand. Conformational sampling of small fragments is more effective, because the small fragments are easier to move around than larger ligands during energy minimization or dynamic

Box 1. Development of the multicopy sampling method

1990 • Locally enhanced sampling (LES) for simulation of ligand diffusion in proteins[3]
1991 • Multiple copy simultaneous sampling (MCSS) for mapping ligand binding site on a protein[4]
• Observation of the violation of energy equipartition in the LES dynamics[26]
• Modeling of protein side chains[16]
1992 • Free energy calculation of a point mutation[15]
1993 • Collision-corrected LES (cLES) for restoring energy equipartition and collision correlation[27]
• Docking of flexible peptides to a rigid protein via binding site mapping and segment building[11]
• Docking of flexible peptides to a flexible protein via binding site mapping, loop closure and clustering of copies[12]
• Theoretical analysis of the accuracy and interpretation of the multicopy approximation[28]
1994 • Modeling of free hexapeptides in solution[22]
• Examination of the efficiency of multicopy sampling for modeling protein loops[19]
• Self-consistent, iterative, multicopy sampling for protein side chain and homology modeling[18]
1995 • Self-consistent, Boltzmann-weighted, iterative multicopy sampling for peptide modeling[24]
1996 • $D_{\text{clustering}}$ as a quantitative measure of the clustering of copies for structural evaluation[21]
• Method of multiple solvent crystal structures for simultaneous mapping of the binding sites of heterogeneous organic solvents[30]
• *De novo* optimization and screening of combinatorial libraries[25]
1997 • Analytical derivation of an exact LES dynamics, its intrinsic interpretation and truncation errors[29]

simulation. For these reasons, it is convenient to take a two-step approach towards ligand design by first using small fragments as functional probes to map the binding site, and then connecting the mapped fragments to form large ligands. While the use of small probes to map a binding site had been practised for years, most notably with the grid method[31], protein flexibility was often neglected before the introduction of multicopy sampling.

Multicopy sampling was first applied to map the sialic acid binding site of the influenza coat protein, using water, methane, methanol, acetate and methyl ammonium as small functional probes[4]. One hundred copies of each group were randomly distributed inside the binding site, and then subjected to multicopy energy minimization with the protein held fixed. The position and orientation of the minimized functional groups were found to correspond with various fragments of the sialic acid in the cocrystal structure. This mapping method was subsequently extended by using larger and flexible functional groups to map the binding site of the human immunodeficiency virus 1 (HIV-1) proteinase, for which the cocrystal structure with a hexapeptide, *N*-acetyl-Thr-Ile-Nle-Ψ[CH_2-NH]-Nle-Gln-Arg-NH_2 (Nle, norleucine) was known[11]. The larger functional groups included *N*-methylacetamide, acetamide, ethyl guanidinum, propane and isobutane, which were chosen to represent various fragments of the hexapeptide. For each functional group, its optimal positions and orientations were searched via multicopy minimization of 500–3,000 copies of the group. The optimal *N*-methylacetamide moieties were connected to form the peptide backbone by using a pseudo energy function, followed by side chain placement. The resulting hexapeptides were subjected to normal energy minimization with Monte Carlo sampling. The smallest backbone root-mean-square deviation (RMSD) between the generated and crystal structure of the hexapeptide was 1.9 Å. Similarly, functional groups have also been used to map the binding site of FK506 binding protein and the active site cleft of human thrombin[14].

In these applications, the correspondence between the position and orientation of the optimal functional groups found by using multicopy sampling and those in the crystal structure of a ligand is usually incomplete. In general, the crystal structure represents a subset of all the optimal functional groups. This may indicate that individually optimized functional groups are not necessarily optimal when connected to form a ligand. Alternatively, the crystal structure can be viewed as one, although not necessarily the optimal one, of the many putative ligands containing the critical functional groups for binding. In this view, the optimal function groups provide an opportunity for designing ligands with higher binding affinity.

There are, however, exceptions where the optimal or near-optimal functional groups fail to cover all the ligand binding modes in the crystal structure. For example, multicopy sampling was unable to position benzene and phenol into a specific ligand-binding pocket of thrombin[14]. This was partially attributed to the inaccuracy of using a reduced hydrogen representation of the functional groups[14]. It is also possible that the rigid thrombin conformation used for mapping may prevent the pocket from adjusting itself to accommodate the functional groups[4,11,14].

Docking of flexible ligands to a flexible protein

Full-flexible ligand docking with multicopy sampling was introduced to predict the conformation of bound nonapeptides to class I major histocompatibility complex (MHC) receptors by using only the information that the amino- and carboxyl-terminal residues of a peptide bind to the conserved pockets at the ends of the binding groove on an MHC receptor[12,13]. Docking was performed by randomly distributing 20 copies of a terminal residue within a 5–7 Å × 5 Å × 5 Å box surrounding a pocket, followed by multicopy energy minimization. Once the location and orientation of two terminal residues were found, the bound conformation of the entire nonapeptide was determined by using the scaling-relaxation loop-closure method[32,33] in conjunction with multicopy sampling. In this process, ten random initial copies of the peptide were generated by scaling down all the bond lengths to meet the end-to-end distance constraint between the two mapped terminal residues. The copies were subsequently relaxed by gradually restoring the bonds to their equilibrium lengths via multicopy energy minimization, during which both the copies and the receptor remained conformationally flexible. The scaling-relaxation process was repeated several times with different random initial copies. Finally, a predicted conformation was chosen based on a combination of energy and the clustering of copies during minimization.

The average backbone RMSD of the two predicted nonapeptides and their respective crystal structures was 1.8 Å. Note that the 1.8 Å prediction of the nonapeptides represents significant progress from the previous modeling of a hexapeptide inhibitor (with a 1.9 Å backbone RMSD) of the HIV-1 proteinase, because a nonapeptide is 50% longer than a hexapeptide. Moreover, 1.8 Å is for the predicted nonapeptide conformation, whereas 1.9 Å is for the best generated hexapeptide conformation. The backbone RMSD of the best generated nonapeptide was 0.9 Å.

Computational screening of combinatorial libraries

The multicopy sampling applications discussed in the previous sections only involve homogeneous copies that are conformations of a single functional group or peptide segment. The same is true in applications of protein side chain placement and homology modeling that involve simultaneous modeling of multiple side chains or loops of a unique protein sequence[18–20]. However, multicopy sampling is not limited to homogeneous copies. In practice, it is often necessary to evaluate the binding of many ligands against a given target molecule; it would therefore be convenient to sample multiple conformations of many different ligands simultaneously. Such a heterogeneous multicopy sampling, together with a novel criterion for structural evaluation, has recently been introduced for studying site-directed amino acid substitutions in a protein[25]. In that study, multiple conformations of 19 natural amino acids were simultaneously substituted and optimized at position 14 of the first zinc finger domain of protein Zif268 to determine the most favorable amino acid at that position.

An initial combinatorial library of 19 amino acids (excluding Pro and Gly, but including both neutral and protonated His) was constructed. The library consisted of 190 random copies, ten for each amino acid (Figure 1a). All 190 copies were patched at position 14 of the protein, followed by energy minimization of the entire library–protein complex. Both the copies and protein were conformationally flexible during minimization. The scoring function for evaluation was the degree of clustering of copies, $D_{\mathrm{clustering}}$, defined for each amino acid substitution as:

$$D_{\mathrm{clustering}} = (\mathrm{RMSD}_0 - \mathrm{RMSD})/\mathrm{RMSD}_0$$

where RMSD_0 and RMSD are the average root-mean-square deviations of all possible pairs (e.g. a total of $10(10 - 1)/2 = 45$ pairs for ten copies) of the copies before and after minimization, respectively. The normal range of $D_{\mathrm{clustering}}$ is between 0 and 1. Substitutions with larger $D_{\mathrm{clustering}}$ were considered as more favorable. The optimal conformation of each substitution was identified by the largest cluster of its copies after minimization. To increase statistical significance, ten such random library–protein complexes were individually minimized and their average $D_{\mathrm{clustering}}$ was used as the scoring function for evaluation.

Screening of the initial libraries identified nine favorable substitutions: Phe, His, His+, Leu, Glu, Tyr, Gln, Asp and Asn

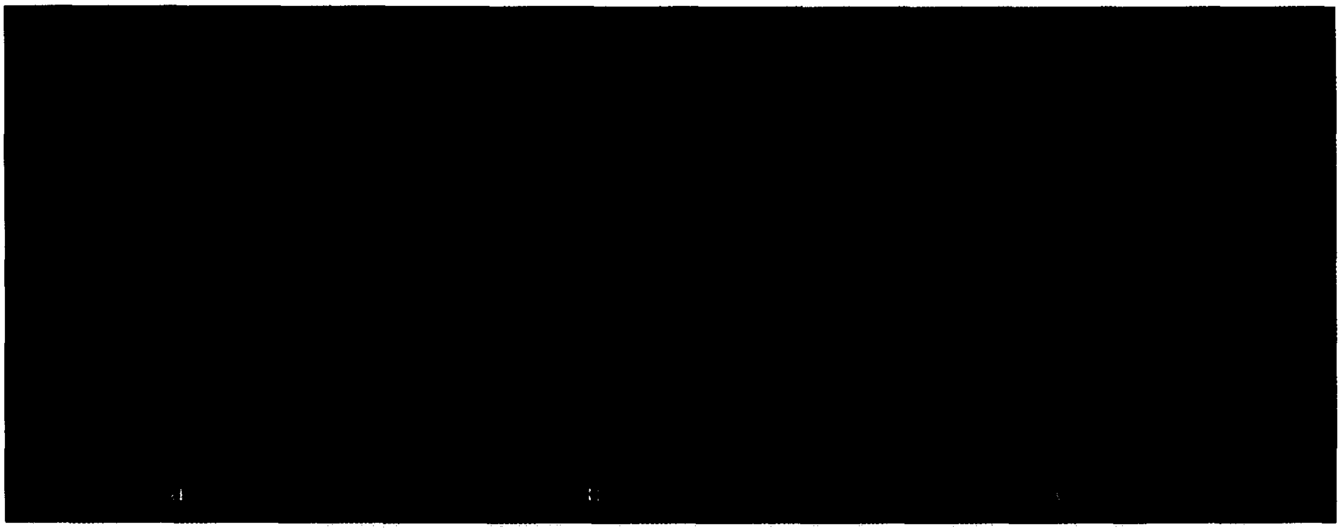

Figure 1. Combinatorial libraries constructed at position 14 of the zinc finger protein, whose α-carbon trace is shown as a blue ribbon, together with the zinc ion (blue sphere). (a) An initial library consists of 190 random initial copies, ten for each amino acid. (b) A sublibrary of nine amino acid substitutions (F, Y, H, H+, E, L, D, Q, and N – single-letter amino acid codes), ten for each amino acid. (c) Ten copies (white) of Phe after the multicopy energy minimization of the sublibrary, together with the native conformation of the Phe side chain (red).

(Figure 2a). Next, ten sublibraries were generated, each sublibrary consisting of 90 random copies, ten for each identified substitution (Figure 1b). The screening of these sublibraries revealed the rank order of scores as Phe, Tyr, His, His+, Glu, Leu, Asp, Gln and Asn, in which Phe was clearly the most favorable substitution (Figure 2b). In addition, the optimal substituted conformations were identified from their respective largest clusters of the copies after minimization (Figure 1c). This result was in qualitative agreement with the published experimental data that Phe is highly conserved and most stable, followed by Tyr and His.

To examine the accuracy and reliability of heterogeneous multicopy sampling, homogeneous multicopy sampling was applied to ten individual libraries of each of the 19 amino acids. Each individual library consisted of ten random copies for a unique amino acid. The rank order of scores for the top eight substitutions was similar to that of the sublibrary screening (Figure 2c). Both screenings suggested that a ring-containing side chain is favored at position 14 of the protein, with Phe being the most favorable one. These results demonstrate that the screening of 19 individual substitutions can be effectively replaced by two rounds of screening of heterogeneous libraries.

The ability to identify ring-containing side chains as favorite substitutions from a structurally diverse library of 19 amino acids suggests that multicopy sampling is an efficient method for the design and screening of structurally biased combinatorial libraries. The most important difference between this and other computational screening methods is that it simultaneously processes an entire library of ligands, whereas other methods process the individual ligands in the library one at a time.

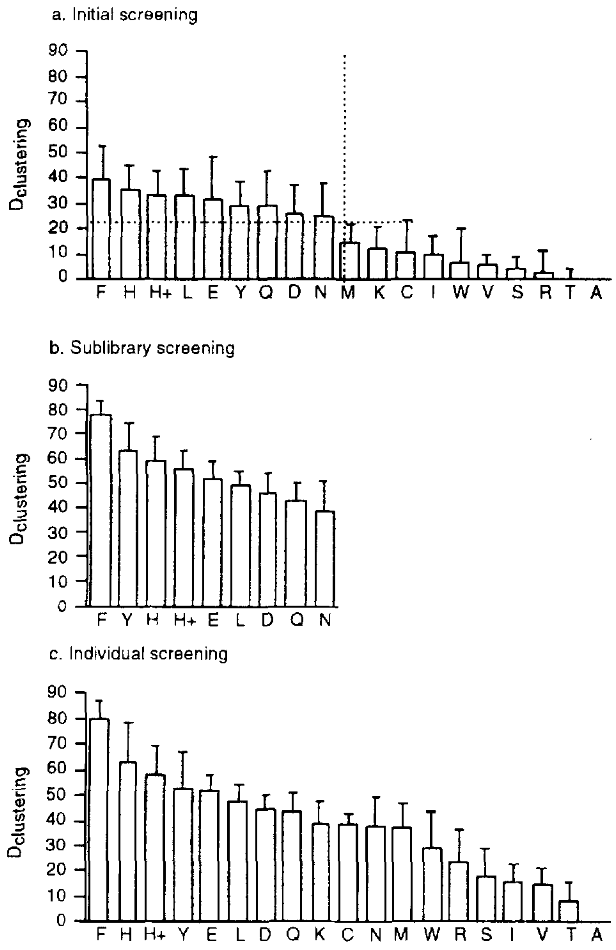

Figure 2. *The* $D_{clustering}$ *profile obtained from the screening of (a) the original 190-copy libraries, (b) the 90-copy sublibraries and (c) the ten-copy individual libraries. In each case, the average and standard deviation of* $D_{clustering}$ *of each amino acid substitution is calculated from the screening of ten different libraries with random initial copies. Substitutions with larger* $D_{clustering}$ *are considered to be more favorable.*

Current and future developments

The past few years have been marked by the development of fast computer programs for assisting combinatorial chemistry and high-throughput screening, which has been further motivated by the current trend of miniaturization and integration of automated synthesis and screening. While speed will always be a challenge, the biggest challenge has been in improving the accuracy and reliability of computer-aided drug design. The multicopy sampling technique represents a unique approach to meeting these challenges. It increases the speed and improves the accuracy and reliability of the traditional single-copy approach in molecular modeling by relying on the fundamentals of molecular mechanics instead of empirical rules[12,13,17]. It is generally applicable to various energy functions and molecular systems, and is compatible with various sampling methods, including Monte Carlo, simulated annealing and genetic algorithms.

While the multicopy sampling method is more efficient than the single-copy methods with an all-atom energy function and a flexible protein, it is slower than methods with empirical rules and a rigid protein. Our experience is that it takes approximately 1 h of CPU time on a Silicon Graphics 150 MHz processor to optimize and screen a 50-copy library of single amino acids with a protein of 100 amino acids.

Assuming two ligands per library, this translates to a throughput of 48 ligands per day, which is far less, for example, than the throughput of 10^5 ligands per day using DOCK (Ref. 34), although a full force-field-based multicopy sampling approach can yield extra information on induced fit, entropy and dynamics. Moreover, the multicopy sampling method shares similar limitations with the traditional single-copy sampling methods (e.g. uncertainties in energy functions, inability for an exhaustive conformational sampling and lack of efficient solvation treatment). Finally, the implementation of multicopy sampling to an existing molecular modeling program requires major modification of the program, thus costing time and effort.

To realize the full potential of multicopy sampling, a number of fundamental and practical issues need to be addressed:

- Multicopy sampling is based on a mean field approximation to the protein, which truncates off certain energetic details from a normal energy function as a trade-off for the gain of sampling efficiency. Because of the loss of these details, the protein becomes less responsive to individual ligand copies. It is important to study the nature of this truncation and its effects on the outcome of a multicopy simulation. Initial progress has been made by the introduction of cLES (collision-corrected LES)[8,27] and a recent derivation of an analytical form of the truncation[29].
- All reported mapping and docking simulations with multicopy sampling involve a single copy of the protein. Such a multicopy implementation is sufficient if the protein conformation changes little upon ligand binding. However, this is not always the case. Suppose that a ligand can bind to a protein in two different conformations, each of which requires a different packing of the protein side chains in the binding site. In order to realize both packings in a single simulation, these side chains must be multicopied along with the ligand. Sometimes, it is also necessary to copy the protein backbone atoms and loops near the binding site, because such a multicopied site is 'softer' and therefore easier to adopt to ligand-specific conformations (Kyle and Zheng, manuscript in preparation).
- Multicopy sampling has two attractive features that are intuitive to understand but difficult to characterize, namely its ability to smooth energy function and to explore conformational entropy. While lowered energy barriers were observed in several simulations, it remains to be demonstrated how multicopy sampling can be used to systematically reduce spurious energy minima. The entropy content of a multicopy simulation is poorly understood. Consider clusters formed by randomly distributed ligand copies during energy minimization. The larger clusters tend to correspond to lower energy minima with a broader attraction basin (entropy). It is important to explore the nature and extent of this correspondence, especially since clustering is an inherent feature of a multicopy simulation. The introduction of $D_{\text{clustering}}$ represents an initial step in this direction, but much remains to be done.

REFERENCES

1 Ashton, M.J., Jaye, M.C. and Mason, J.S. (1996) *Drug Discovery Today* 1, 71–78

2 Manallack, D.T. (1996) *Drug Discovery Today* 1, 231–238

3 Elber, R. and Karplus, M. (1990) *J. Am. Chem. Soc.* 112, 9161–9175

4 Miranker, A. and Karplus, M. (1991) *Proteins* 11, 29–34

5 Czerminski. R. and Elber, R. (1991) *Proteins* 10, 70–80

6 Nowak, W., Czerminski, R. and Elber, R. (1991) *J. Am. Chem. Soc.* 113, 5627–5637

7 Verkhivker, G., Elber, R. and Gibson, Q.H. (1992) *J. Am. Chem. Soc.* 114, 7866–7878

8 Ulitsky. A. and Elber, R. (1994) *J. Phys. Chem.* 98, 1034–1043

9 Carlson, M.L. *et al.* (1994) *Biochemistry* 33, 10597–10606

10 Quillin, M.Q. *et al.* (1995) *J. Mol. Biol.* 245, 416–436

11 Calflish, A., Miranker, A. and Karplus, M. (1993) *J. Med. Chem.* 36, 2142–2167

12 Rosenfeld, R. *et al.* (1993) *J. Mol. Biol* 234, 515–521

13 Rosenfeld, R. *et al.* (1995) *Genet. Anal.* 12, 1–21

14 Grootenhuis, P.D.J. and Karplus, M. (1996) *J. Comput.-Aided Mol. Design* 10, 1–10

15 Verkhivker, G., Elber, R. and Nowak, W. (1992) *J. Chem. Phys.* 97, 7838–7841

16 Roitberg, A. and Elber, R. (1991) *J. Chem. Phys.* 95, 9277–9287

17 Zheng, Q. and Kyle, D.J. (1994) *Proteins* 19, 324–329

18 Koehl. P. and DeLarue, M. (1994) *J. Mol. Biol.* 239, 249–275

19 Zheng, Q. *et al.* (1994) *Protein Sci.* 3, 493–506

20 Koehl. P. and DeLarue. M. (1995) *Struct. Biol.* 2, 163–170

21 Zheng, Q. and Kyle, D.J. (1996) *Proteins* 24, 209–217

22 Simmerling, C.L. and Elber, R. (1994) *J. Am. Chem. Soc.* 116, 2534–2547

23 Simmerling, C.L. and Elber, R. (1995) *Proc. Natl. Acad. Sci. U. S. A.* 92, 3190–3193

24 Huber, T., Torda, A.E. and van Gunsteren, W.F. (1996) *Biopolymers* 39, 103–114

25 Zheng, Q. and Kyle. D.J. (1996) *Bioorg. Med. Chem.* 4, 631–638

26 Straub, J.E. and Karplus, M. (1991) *J. Chem. Phys.* 94, 6737–6739

27 Ulitsky, A. and Elber, R. (1993) *J. Chem. Phys.* 98, 3380–3388

28 Zheng, Q., Rosenfeld, R. and Kyle, D.J. (1993) *J. Chem. Phys.* 99, 8892–9986

29 Zheng, W. and Zheng, Q. (1997) *J. Chem. Phys.* 106, 1191–1194

30 Allen, K.N. *et al.* (1996) *J. Phys. Chem.* 100, 2605–2611

31 Goodford, P.J. (1985) *J. Med. Chem.* 28, 849–857

32 Zheng, Q. *et al.* (1993) *J. Comput. Chem.* 14, 556–565

33 Zheng, Q. *et al.* (1993) *Protein Sci.* 2, 1242–1248

34 Gschwend, D.A. and Kuntz, I.D. (1996) *J. Comput.-Aided Mol. Design* 10, 123–132