Study of protein-protein interaction using conformational space annealing. Proteins 60, 257-262

ARTICLE *in* PROTEINS STRUCTURE FUNCTION AND BIOINFORMATICS · AUGUST 2005

Impact Factor: 2.63 · DOI: 10.1002/prot.20567 · Source: PubMed

3 AUTHORS, INCLUDING:

Jooyoung Lee
Korea Institute for Advanced Study

120 PUBLICATIONS **3,890** CITATIONS

SEE PROFILE

Study of Protein–Protein Interaction Using Conformational Space Annealing

Kyoungrim Lee, Jaehyun Sim, and Jooyoung Lee*
School of Computational Sciences, Korea Institute for Advanced Study, Seoul, Korea

ABSTRACT **We apply conformational space annealing (CSA), an efficient global optimization method, to the study of protein–protein interaction. The CSA is incorporated into the Tinker molecular modeling package along with a B-spline method for CAPRI Round 5 experiments. We have used an energy function for the protein–protein interaction that consists of electrostatic interaction, van der Waals interaction, and solvation energy terms represented by the occupancy desolvation method. The parameters of the AMBER94 all-atom empirical force field are used. Each energy term is calculated by precalculated grid potentials and B-spline method approximation. The ligand protein is placed inside a sphere of 50 Å radius centered at an appropriate location, and the CSA rigid docking studies are carried out to find stable complexes. Up to 10 complexes are selected using the K-mean clustering method and biological information when available. These complexes are energy-minimized for further refinement by considering the flexibility of interacting proteins. The results show that the CSA method has a potential for the study of protein–protein interaction. Proteins 2005;60:257–262.**

Key words: protein docking; protein–protein interaction; global optimization; conformational space annealing; B-spline method

INTRODUCTION

One of the goals of the protein–protein docking study is to predict the 3-dimensional (3D) arrangement of a protein–protein complex from the 3D coordinate information of its component molecules. The protein–protein docking study involves 2 orthogonal and yet closely inter-related factors. The first one is how to accurately describe the energy function of a given protein–protein complex, and the second one is how to obtain the global minimum energy structure of the complex using the energy function. The energy function may include geometric and chemical complementarities, as well as electrostatic interaction, hydrogen-bonding interaction, and solvation energy terms. All-atom empirical potentials and/or database-derived score functions can be also used. One of the challenging problems of the docking study is to carry out a rigorous conformational search of a given system considering relative position and orientation of the component molecules as well as their flexibilities. Generally speaking, the protein–protein docking problem can be classified as one of the global optimization problems, since its main purpose is to find the most stable association of protein molecules. Several review articles[1–4] on the current docking methods are available.

Conformational space annealing (CSA)[5–8] is one of the most efficient global optimization methods. The basic idea of the CSA is that it enforces a broad conformational sampling in early stages and then gradually focuses the search into narrow regions populated with low-energy conformations. The most prominent advantage of the CSA is that it can find many distinct families of low-energy conformations. This makes it possible to search the whole intermolecular space of the protein–protein association for a given energy function. The sampling diversity of CSA is maintained by keeping various conformers of local-energy minima as representatives of structurally similar conformations within hyperspheres centered on them. The CSA is achieved by slowly reducing the radius of these hyperspheres. As in the genetic algorithm,[9] CSA evolves the population of solutions through genetic operators to the final population containing *diverse* solutions of a given energy function. The most important feature of CSA is that it can directly control the diversity of the population; consequently, it can generate many distinct low-energy solutions, one of which would correspond to the true solution if the energy function used is reasonably accurate. The CSA method has been successfully applied to various problems including ab initio protein structure prediction,[8,10–12] 3D-structure prediction of multichain homo-oligomer proteins,[13,14] and protein–small-molecule docking.[15]

In this work, we have used the CSA method to study the protein–protein interaction for the CAPRI Round 5 experiment.[16] Energy and its gradient evaluations for local minimization are approximated by cubic B-splines for computational efficiency. The B-spline interpolation was introduced by Oberlin and Scheraga[17] to rapidly evaluate both the potential energy and its gradient in the study of protein–small-molecule docking. Subsequently, the B-spline method combined with a stochastic search method

Grant sponsor: Basic Research Program of the Korea Science and Engineering Foundation; Grant number: R01-2003-000-11595-0.

*Correspondence to: Jooyoung Lee, Department of Computational Science, Korea Institute for Advanced Study, 207-43 Cheongryangri-dong Dongdaemun-gu, Seoul 130-722, Korea. E-mail: jlee@koas.re.kr

Received 16 January 2005; Accepted 9 February 2005

DOI: 10.1002/prot.20567

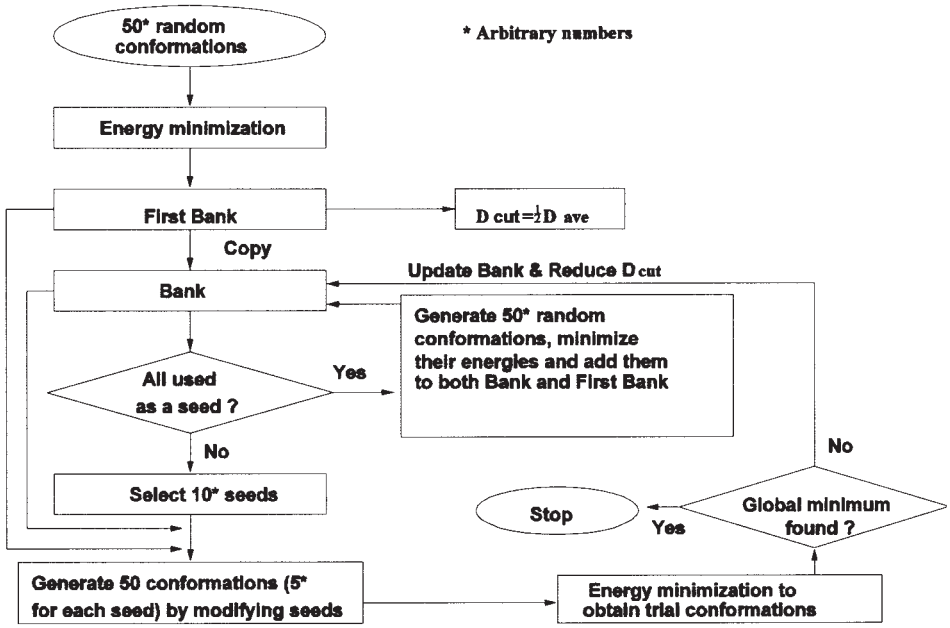

Fig. 1. Flowchart of the CSA algorithm to find the most stable receptor–ligand complex.

is used for gradient-based minimization to study flexible protein–small-molecule docking,[18,19] and its application is further extended to study protein–protein interaction.[20] The energy function used in the current work consists of the AMBER94[21] all-atom van der Waals and electrostatic interactions, and an implicit solvation energy term. The major purpose of this work is to investigate the applicability of the CSA coupled with an all-atom potential for the study of protein–protein interaction and to discuss key findings and modifications necessary for future improvement.

MATERIALS AND METHODS

Conformational Space Annealing

Details of the CSA algorithm and its applications can be found in the references.[5–8,10–12] Here, we provide only a brief description of essential changes of the algorithm for its implementation to the protein–protein docking study with the Tinker program package.[22] The flow chart of the CSA algorithm is shown in Figure 1. It should be noted that, in this work, we have considered only rigid docking. When applying the CSA to docking studies, we first generate a number of random conformations (typically, 50 conformations) whose energies are subsequently minimized. Random conformations are generated by assigning random translational vectors (x,y,z) and rotational Euler angles (ϕ,θ,ψ) to the ligand protein with respect to its receptor protein. We call the set of these minimized complexes the *first bank*. We make a copy of the first bank and call it the *bank*. The conformations in the bank are updated in later stages, whereas those in the first bank are kept unchanged. In addition, the number of conformations in the bank is kept unchanged when the bank is updated. The initial value of D_{cut} is set as $D_{\text{ave}}/2$, where D_{ave} is the average pairwise distance between the conformations in the first bank. New conformations are generated by choosing a certain number of *seed* conformations (e.g., 20 seed conformations in this work) from the bank and by replacing parts of their variables by the corresponding parts of conformations randomly chosen from either the first bank or the bank. Since, in this work, we perform the conformational search in the framework of rigid docking, the total number of degrees of freedom is 6 (3 for the translational vector and the other 3 for Euler angles). New conformations are generated by replacing a subset of variables from a seed conformation by the corresponding one from a conformation in the bank or in the first bank. Energies of these conformations are subsequently minimized, and these minimized conformations become trial conformations.

A newly obtained local minimum-energy conformation (i.e., a trial conformation) α is used to update the bank conformations. One first finds the conformation A in the bank which is the closest to α with the distance $D(\alpha,A)$ defined by

$$D(\alpha,A) = |\Delta T_{\alpha A}(x,y,z)| + \omega_\Theta \Theta_{\alpha A}(\phi,\theta,\psi) \qquad (1)$$

where $\Delta T_{\alpha A}$ (x,y,z) is the translation vector from A to α, and $\Theta_{\alpha A}$ (ϕ,θ,ψ) is the angle between the 2 Euler angle (ϕ,θ,ψ) vectors of A and α. The weight factor ω_Θ is determined dynamically during the docking simulation as follows. After the first bank is generated, we calculate the average values of the 2 terms in the righthand side of Eq. (1) considering all pairs in the first bank. The value of ω_Θ is set so that the 2 terms contribute equally on average.

If $D(\alpha,A) < D_{\text{cut}}$, α is considered as similar to A. In this case, the conformation with a lower energy between α and A is kept in the bank, and the other is discarded. However, if $D(\alpha,A) > D_{\text{cut}}$, α is regarded as distinct from all

conformations in the bank. In this case, the conformation with the highest energy among all bank conformations plus α is discarded, and the rest are kept in the bank so that the total number of the bank conformations remains fixed during this procedure. We perform this operation for all trial conformations.

After the bank is updated using all available trial conformations, the value of D_{cut} is reduced by a fixed ratio, in such a way that it reaches $D_{ave}/5$ after L local minimizations (e.g., $L = 2000$). Then, new seeds are selected from the bank conformations that have not been used as seeds yet, to repeat the aforementioned procedure. The value of D_{cut} is kept constant after it reaches the final value. When all conformations in the bank are used as seeds, one round of iteration is completed. We perform additional iterations of search by first erasing the record of bank conformations having been used as seeds, and by starting a new round of iteration. After 3 iterations are completed, we increase the number of bank conformations by adding 50 randomly generated and subsequently minimized conformations into the bank (and also into the first bank), and reset the value of D_{cut} to its original value of $D_{ave}/2$. The whole procedure stops when the global minimum is found (if its value is known a priori), which is examined immediately after the bank is updated by all trial conformations. For a problem with unknown global minimum (as in the cases of the CAPRI experiment), the conformational search was performed until either the size of the bank becomes larger than the preset maximum number (typically 500), or the preset maximum CPU time elapses. It should be noted that since 1 iteration is completed only after all bank conformations have been used as seeds, and since we add an additional number of conformations whenever our search reaches a deadlock, there is no loss of generality for using particular values for the number of seeds, the number of bank conformations, and so forth.

Energy Function and B-Spline Approximation

We have used the AMBER94 all-atom force field.[21] The energy function contains 3 energy terms: electrostatic (E_{ele}), van der Waals (E_{vdw}), and solvation (E_{solv}) terms:

$$E_{total} = E_{ele} + E_{vdw} + E_{solv} \quad (2)$$

The first two energy terms describe the intermolecular atomic pairwise interactions between a receptor and its ligand molecules, and the third solvation energy term reflects the solvent effect on the complex. For solvation energy, we have used the atomic occupancies model,[23] which is a simple representation of the free energy of interaction of an atom with implicit solvent. This solvation model was applied to protein folding and docking studies[23,24] and yielded quite satisfying results in good agreement with experimental results. Another advantage of this solvation model is that it can be easily implemented along with B-spline approximation,[17] since the solvation term contains only pairwise interactions. Here, we define a grid box large enough to contain the region where the association of the receptor and the ligand proteins may take place. During the rigid docking simulation, the value and the derivative of each energy component are approximately evaluated from its precalculated values[24] at grid points using the cubic B-spline approximation as introduced in Oberlin and Scheraga.[17]

Rigid-Body Docking Simulations

We generate a grid box large enough to contain the search space of the protein–protein interaction and its size is determined based on the size of a target protein–protein complex. The grid spacing is set to 1.0 Å. For each grid point, all components of energy and gradient functions are precalculated. The searches for stable docking complexes are initiated by constructing the first bank containing "*nbank*" (typically, *nbank* = 50) randomly generated docked complexes. The number of seed conformations is set as *nseed* = 20. For each seed conformation, 5 perturbed conformations are generated (3 by replacing the translational vector and 2 by replacing the Euler angles). Therefore, a total of 100 perturbed conformations are generated, and they are energy-minimized to obtain trial conformations. Using these trial conformations, the bank is updated. The rest of the procedure is as explained above. The search for ligand positions to form stable protein complexes is restricted to the inside of a preset sphere (typically of radius 45–60 Å). When no obvious biological information is available (as in Targets 18 and 19), we placed a few spheres on the surface of a receptor molecule to divide the search space into smaller sections, and performed the conformational search separately. However, for Targets 14 and 15, sequence analysis of the component proteins provided us some information regarding the possible locations of their binding sites. For such cases, spheres are located on the potential binding sites. Each sphere is surrounded by a harmonic-soft-wall potential and consequently the ligand is confined inside the sphere.

Filtering and Clustering

The CSA docking procedure has produced between 100 and 400 structurally distinct conformations in the final banks of 4 CAPRI Round 5 targets. Among the final bank conformations, we removed those with the minimum heavy atom distance between the receptor and the ligand proteins larger than 5Å. Then, the K-means clustering method[25] is used to group the remaining conformations into up to 10 structural families. Either the conformer with the lowest energy or the one located at the cluster center is selected as the cluster representative, based on the available biological information. Subsequently, these selected representatives are further energy-minimized by considering the flexibilities of component proteins, as the final models for prediction.

RESULTS

Target 14 (Protein Ser/Thr Phosphatase-1 Complexed With Myosin Phosphatase)

For this target, we have treated the myosin phosphatase, which is the smaller one in size, as a receptor molecule simply because we assumed that its binding

TABLE I. Summary for the Best Prediction

Target	Model	Contact	θ angle	Distance (Å)	L_rmsd (Å)	I_rmsd (Å)
14	1	0	150.8	32.40	20.07	54.38
15	4	0.2	22.8	7.62	8.76	3.25
18	1	0	93.1	28.5	32.44	15.24
19	2	0	81.9	22.14	26.14	14.61

Columns 1 and 2 list the target number and the model number of the submission. Contact (column 3) indicates the fraction of native contact, defined as the number of native residue–residue contact in the model divided by the number of native contact in the experimental structure. θ angle (column 4) is the rotation angle necessary to fit the ligand molecule in the predicted complex to that in the experimental structure. Column 5 lists the distance between geometric centers of predicted and target ligand molecules. L-rmsd and l_rmsd (columns 6 and 7) are the RMSD values of the main-chain of the ligand, and of the main-chain of interface residues in the model versus the experimental structure for the fixed receptor position.

interface would be near the well-conserved region (the inner parallel loop region) from its multiple sequence alignment data. Based on this, the ligand search was limited to the inside of a sphere of radius 40 Å that was centered on the consequently wrong binding site. This, naturally, led to predictions that were far away from the actual crystal structure, and the results were the worst out of the 4 CAPRI Round 5 targets. This illustrates that incorrect or insufficient biological information can easily mislead one toward wrong prediction. The size of the interfacial surface area in the crystal structure is much larger than we expected. The globular protein of Ser/Thr phosphates is well positioned against the concave part of the myosin phosphatase to maximize the interface contact area.

Target 15 (ImmD Immunity Protein Complexed With Colicin D Catalytic Domain)

The prediction for Target 15 is our best overall prediction out of 4 CAPRI Round 5 targets (see Table I). For the best model, the value of ligand RMSA, L_rmsd, which is measured between the predicted ligand backbone structure and the corresponding experimental one for the fixed receptor position, is 8.75 Å. The value of interface RMSD, I_rmsd, which is measured between the predicted residues in the interface and the experimental counterparts, is 3.25 Å. The fraction of native contacts (contact in Table I) is 0.2, which is only a small fraction but still is in the range of being acceptable. The CSA docking overall resulted in 348 structurally distinct complexes in the final bank. They are clustered into 10 structural families using the K-means method. The family containing the best prediction was highest in population with 60 complexes.

Targets 18 and 19

Target 18 is a protein complex consisting of the Triticum aestivum xylanase inhibitor (TAXI) and the *Aspergillus niger* xylanase. Target 19 is an antibody–antigen complex. For these 2 targets, no definite residue conserved regions are detected. These targets are protein complexes of relatively large size, and the CSA searches were terminated in a much premature manner compared to the cases of Targets 14 and 15. The CSA produced 132 and 36 complexes, and 10 and 6 families, for Targets 18 and 19, respectively. For both targets, the predictions were quite different from their experimental structures. The L_rmsd's and the I_rmsd's of the predicted structures are 32.44 and 15.24 Å, respectively, for Target 18 and 26.14 and 14.61 Å, respectively, for Target 19 (see Table I).

DISCUSSION

Modification of Energy Function

The energy function used in this study contained intermolecular van der Waals and electrostatic interactions in terms of AMBER94 parameters and a simple implicit solvation model. The major motivation of this work was to investigate the applicability of CSA with an existing all-atom force field for the study of protein–protein interaction. However, the current form of energy function is shown to be not quite adequate enough to describe accurately the interaction between the receptor and the ligand proteins for the 4 examples in the CAPRI Round 5 experiment. For better description, one might consider using an energy function that includes terms such as the shape complementarities and the coarse-grained residue–residue contact energy between proteins.

Generally, the distances between the 2 interacting proteins in the final bank structures were consistently larger than the experimental value, which led to poor interfacial contacts between proteins. This is partly because we performed conformational searches only in the context of rigid docking, and also because the repulsive part of the van der Waals energy was too strong. Inclusion of the shape complementarity term and the residue–residue contact energy term in the energy function would alleviate this problem. Another possible approach to improve the energy function is to carry out 2-step energy minimization. For example, we can first minimize a structure using an energy function containing only shape complimentary and electrostatic interactions, and then further minimize the resultant structure using a more complete energy function containing both van der Waals and electrostatic energy terms. In this case, appropriate consideration of the side-chain flexibility at the docking interface can lead to the application of the CSA to the soft protein–protein docking.

Conformational Diversity

The investigation to seek possible binding sites based on sequence analysis such as the multiple sequence alignment should be carried out more extensively and more

carefully. Inadequate interpretation of the binding-site information will lead one to incorrect prediction, as demonstrated in the case of Target 14. On the other hand, when interacting proteins are large as in Targets 18 and 19, it was difficult to carry out extensive conformational sampling due to the tremendous amount of CPU time to evaluate the energy function at the atomic level (even with the grid method). In retrospect, even with large targets, we should set CSA parameters so that more diverse sampling is enforced.

Computation Efficiency

The all-atom potential used in this work requires quite extensive CPU time to minimize a large complex such as Targets 18 and 19. The only efforts made were to increase the computational efficiency by precalculated grid potentials and the interpolation of the energy and the gradient values by B-spline approximation. However, this speed-up was not sufficient to finish the calculation in time to meet the deadline for the large targets. For Targets 18 and 19, we had to terminate the CSA docking search prematurely, well before the stopping condition was met. For further development of our method, we need to simplify the energy calculation. One possibility is to consider only the interaction between the surface atoms of the receptor and the ligand proteins for energy evaluation by emphasizing the short-range interaction. Another possiblity is to use a simplified protein model at the coarse-grained level.

CONCLUSION

We have applied the CSA method to an all-atom pairwise force field with a solvation model to predict the protein–protein docking complexes of the 4 CAPRI Round 5 targets. By performing rigid docking calculations, from 50 up to 400 low-energy complexes were obtained by CSA. For targets of reasonable size (i.e., Targets 14 and 15), extensive conformational searches were carried out at an atomic level, and the results from the Target 15 was acceptable. On the other hand, the results of Target 14 were misled by the wrong interpretation of the multiple sequence alignment data. The motivation for this work was to investigate the applicability of CSA with an existing all-atom force field for the study of the protein–protein interaction. However, the current form of energy function is shown to be not quite adequate to describe accurately the interaction between the receptor and the ligand proteins for the four examples in the CAPRI Round 5. For the larger targets, such as Targets 18 and 19, our procedure was hampered by the fact that the current energy function requires an enormous amount of CPU resources to evaluate its value. As beginners in the study of protein–protein docking, we have overlooked a couple of things: selection of an all-atom energy function and the danger of misinterpretation of the binding site information. As a result, only the prediction of Target 15 was acceptable, and the predictions of the other three targets were incorrect. For future improvement, the energy function needs to be modified to include shape complementarity, contact information, and more careful interpretation of the binding site information. From the CAPRI Round 5 experiment, we have learned that the CSA can provide useful information on the possible candidates for a docked complex with an all-atom energy function.

ACKNOWLEDGMENTS

We thank Professor Julian Lee and Dr. Seung-Yeon Kim for their helpful comments on B-spline approximation methods.

REFERENCES

1. Elcock AH, Sept D, McCammon JA. Computer simulation of protein–protein interactions. J Phys Chem B 2001;105:1504–1518.
2. Smith GR, Sternberg MJE. Prediction of protein–protein interactions by docking methods. Curr Opin Struct Biol 2002;12:28–35.
3. Russell RB, Alber F, Aloy P, Davis FP, Korkin D, Pichaud M, Topf M, Sali A. A Structural perspective on protein–protein interactions. Curr Opin Struct Biol 2004;14:313–324.
4. Méndez R, Leplae R, De Maria L, Wodak SJ. Assessment of blind predictions of protein–protein interactions: current status of docking methods. Proteins 2003;52:51–67.
5. Lee J, Scheraga HA, Rackovsky S. New optimization method for conformational energy calculations on polypeptides: conformational space annealing. J Comput Chem 1997;18:1222–1232.
6. Lee J, Scheraga HA. Conformational space annealing by parallel computations: extensive conformational search of met-enkephalin and of the 20-residue membrane bound portion of melittin. Int J Quant Chem 1999;75:255–265.
7. Kim SY, Lee SJ, Lee J. Conformational space annealing and an off-lattice frustrated model protein. J Chem Phys 2003;119:10274–10279.
8. Lee J, Lee IH, Lee J. Unbiased global optimization of Lennard–Jones clusters for $N \geq 201$ using the conformational space annealing method. Phys Rev Lett 2003;91:080201.
9. Goldberg DE. Genetic algorithms in search, optimization and machine learning. Reading, MA: Addison-Wesley; 1989.
10. Lee J, Liwo A, Ripoll DR, Pillardy J, Scheraga HA. Calculation of protein conformation by global optimization of a potential energy function. Proteins 1999;Suppl 3:204–208.
11. Lee J, Liwo A, Ripoll DR, Pillardy J, Saunders JA, Gibson KD, Scheraga HA. Hierarchical energy-based approach to protein–structure prediction: blind-test evaluation with CASP3 targets. Int J Quant Chem 2000;77:90–117.
12. Lee J, Kim SY, Joo K, Kim I, Lee J. Prediction of protein tertiary structure using PROFESY, a novel method based on fragment assembly and conformational space annealing. Proteins 2004;56:704–714.
13. Saunders JA, Scheraga HA. Ab initio structure prediction of two α-helical oligomers with a multiple-chain united-residue force field and global search. Biopolymers 2003;68:300–317.
14. Saunders JA, Scheraga HA. Challenges in structure prediction of oligomeric proteins at the united-residue level: searching the multiple-chain energy landscape with CSA and CFMC. Biopolymers 2003;68:318–332.
15. Lee K, Czaplewski C, Kim SY, Lee J. An efficient molecular docking using conformational space annealing. J Comput Chem 2005;26:78–87.
16. Janin J, Hendrick K, Moult J, Eyck LT, Sternberg MJE, Vajda S, Vakser I, Wodak SJ. CAPRI: A Critical Assessment of PRedicted Interactions. Proteins 2003;52:2–9.
17. Oberlin DJ, Scheraga HA. B-spline method for energy minimization in grid-based molecular mechanics calculations. J Comput Chem 1998;19:71–85.
18. Trosset J-Y, Scheraga HA. PRODOCK: software package for protein modeling and docking. J Comput Chem 1999;20:412–427.
19. Trosset J-Y, Scheraga HA. Flexible docking simulations: scaled collective variable Monte Carlo minimization approach using Bezier splines, and comparison with a standard Monte Carlo algorithm. J Comput Chem 1999;20:244–252.

20. Gillilan RE, Lilien RH. Optimization and dynamics of protein–protein complexes using B-splines. J Comput Chem 2004;25:1630–1646.
21. Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM Jr, Ferguson DM, Spellmeyer DC, Fox T, Caldwell JW, Kollman PA. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. J Am Chem Soc 1995;117:5179–5197.
22. Ponder JW. TINKER 3.9. Available online: http://dasher.wustl.edu/tinker; 2001.
23. Stouten PFW, Frommel C, Nakamura H, Sander C. An effective solvation term based on atomic occupancies for use in protein simulations. Molecular Simulation 1993;10:97–120.
24. Luty BA, Wasserman ZR, Stouten PFW, Hodge CN, Zacharias M, McCammon JA. A molecular mechanics/grid method for evaluation of ligand-receptor interactions. J Comput Chem 1995;16:454–464.
25. Kanungo T, Mount DM, Netanyahu NS, Piatko CD, Silverman R, Wu AY. An efficient K-means clustering algorithm: analysis and implementation. IEEE Trans Pattern Analysis and Machine Intelligence 2002;24:881–892.