Molecular Docking Improvement: Coefficient Adaptive Genetic Algorithms for Multiple Scoring Functions

Zhengfu Li[1,2*], Xicheng Wang[3] , Keqiu Li[1] , Junfeng Gu[3] , Ling Kang[2]

[1]*School of Computer Science and Technology*
Dalian University of Technology
2 Linggong Road, Dalian, P.R. China, 116024
E-mails: lizhengfu@hotmail.com, keqiu@dlut.edu.cn

[2]*Department of Computer Science and Technology*
Dalian Neusoft University of Information
8 Software Park Road, Dalian, P.R. China, 116023
E-mail: kangling@neusoft.edu.cn

[3]*Department of Engineering Mechanics*
Dalian University of Technology
2 Linggong Road, Dalian, P.R. China, 116024
E-mails: guixum@dlut.edu.cn, jfgu@dlut.edu.cn

[*]*Corresponding author*

Received: June 12, 2013 **Accepted: March 14, 2014**

Published: March 28, 2014

Abstract: *In this paper, a coefficient adaptive scoring method of molecular docking is presented to improve the docking accuracy with multiple available scoring functions. Based on force-field scoring function, we considered hydrophobic and deformation as well in the proposed method, Instead of simple combination with fixed weights, coefficients of each factor are adaptive in searching procedure. In order to improve the docking accuracy and stability, knowledge-based scoring function is used as another scoring factor. Genetic algorithm with the multi-population evolution and entropy-based searching technique with narrowing down space is used to solve the optimization model for molecular docking. To evaluate the method, we carried out a numerical experiment with 134 protein-ligand complexes of the publicly available GOLD test set. The results validated that it improved the docking accuracy over the individual force-field scoring. In addition, analyses were given to show the disadvantage of individual scoring model. Through the comparison with other popular docking software, the proposed method showed higher accuracy. Among more than 77% of the complexes, the docked results were within 1.0 Å according to Root-Mean-Square Deviation (RMSD) of the X-ray structure. The average computing time obtained here is 563.9 s.*

Keywords: *Genetic algorithms, Coefficient adaptive, Molecular docking, Scoring function, Optimization.*

Introduction

Molecular docking is the prediction of conformation of a ligand within the active site of a receptor and search for the low-energy binding modes [8]. Molecular docking is widely used in virtual screen, and some successful cases have been reported [17]. The docking model and scoring functions have received wide concerns in recent years and a lot of scoring functions have been proposed [12]. As the core of molecular docking, scoring function can help a docking program to efficiently explore the binding space of a ligand. It is also responsible for

evaluating the binding affinity once the correct binding pose is identified [1]. It is an optimization process of finding the best position of a ligand in the binding site of a receptor.

A lot of comparative studies have been done to evaluate the relative performances of these widely used docking programs and scoring methods [3, 9, 10, 13, 14, 16]. However, none of these scoring functions or program is generally applicable for all the situations because the interactions between ligands and receptors are complicated. In addition, it is necessary to simplify docking models to obtain acceptable computing time.

Current scoring functions can be roughly classified into three types: force field-based scoring functions, empirical scoring functions and knowledge-based scoring functions. These models of widespread used docking functions are nearly approximate models. Approximation makes one scoring function inaccurate under some circumstances. Based on force-field scoring function, we also considered hydrophobic and deformation as well in our method. Instead of simple combination of them with fixed weights, coefficients are adaptive in searching procedure. In order to improve accuracy and stability, knowledge-based scoring method was used as another scoring factor with adaptive coefficient. An iteration scheme in conjunction with the multi-population evolution and entropy-based searching technique with narrowing down space was used to solve the optimization model for molecular docking. To evaluate the method, we performed the numerical experiment with 134 protein-ligand complexes from the publicly available GOLD test set (http://www.ccdc.cam.ac.uk/). The results indicated that the scoring function for molecular docking had high accuracy.

Materials and methods

In molecular docking, the process of finding the best conformation is an optimization problem. The problem can be described as follows:

$$\begin{aligned} &\min \left\{F_1(X)+F_2(X)+F_3(X)+F_4(X)\right\} \\ &\text{s.t.} \quad g_i(X)<0,\ i=1,\ 2,\ \ldots,\ n \end{aligned}, \tag{1}$$

where X is a vector of design variables, indicating the orientation and conformation information of a ligand. Due to computational reasons, it is always assumed that the ligand is flexible and that the receptor is rigid. So X can be defined as follows:

$$X=\left\{T_x,\ T_y,\ T_z,\ R_x,\ R_y,\ R_z,\ T_{b1},\ T_{b2},\ \ldots,\ T_{bn},\ C_1,\ \ldots,\ C_4\right\}^{\mathrm{T}}, \tag{2}$$

where T_x, T_y and T_z are the position coordinates of the ligand; R_x, R_y and R_z are the rotational angles of the ligand; T_{b1}, T_{b2}, … , T_{bn} are the torsion angles of the rotatable bonds of the ligand; C_1, … , C_4 are coefficients of each factor. The constraints $g_i(X)$, i = 1, 2, … , n are shown as follows:

$$\begin{cases} \underline{T_x} \le T_x \le \overline{T_x} \\ \underline{T_y} \le T_y \le \overline{T_y} \\ \underline{T_z} \le T_z \le \overline{T_z} \\ -\pi \le R_{x,y,z},\ T_{b_1,\ldots,b_n} \le \pi \\ 0 < C_{1,\ldots,4} < 1 \end{cases}. \tag{3}$$

In Eq. (1), $F_1(X)$ represents the part of Van der Waals; $F_2(X)$ represents hydrophobic; $F_3(X)$ represents deformation and $F_4(X)$ represents knowledge-based scoring. $F_i(X)$ is the product of C_i and force-field factor $U_i(X)$.

$$F_i(X) = C_i U_i(X). \tag{4}$$

The force-field function part of this paper adopts the classical AMBER molecular mechanics energy functions [11, 19]. The objective function is the interaction energy between the ligand and protein, consisting of the Van der Waals and Coulomb terms of force field functions:

$$f_1(X) = \sum_{i=1}^{n_{lig}} \sum_{j=1}^{n_{rec}} \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} + 332.0 \frac{q_i q_j}{D r_{ij}} \right), \tag{5}$$

where each term is a double sum over the ligand atom i and the receptor atom j; n_{lig} and n_{rec} are respectively the number of atoms in the ligand and that in the receptor; A_{ij} and B_{ij} are van der Waals repulsion and attraction parameters; r_{ij} is the distance between atoms i and j; q_i and q_j are the point charges on atoms i and j; D is dielectric function; 332.0 is a conversion factor from the electrostatic energy to kilocalories per mole.

The force-field-based scoring function is widely used in popular docking programs, such as DOCK, AutoDock, GoldScore, etc. To simplify the interactions between ligand and receptor, it cannot provide very accurate results in some cases. Empirical scoring considers the interaction in another way. In most empirical scoring functions, it is assumed that the van der Waals interaction (E_{vdw}), hydrogen-bonding energy (E_{hb}), hydrophobic (E_{hyd}) and deformation (E_{def}) terms are the primary parts of binding energy. Weights of the above factors are fixed and obtained by training set. Fixed weights always mean that empirical scoring will not be very accurate in some cases.

Considering that information of van der Waals and hydrogen-bonding is already taken into account in Eq. (5) (hydrogen-bonding is in the electrostatic term), hydrophobic – $f_2(X)$ and deformation – $f_3(X)$, which from X-Score [18], are kept as the next two elements in Eq. (1). To shorten the computation time is an impetus for doing this.

The knowledge-based scoring function commonly refers to Potential of Mean Force (PMF). Completely different from force-field scoring, knowledge-based scoring considers docking problem from another point of view. PMF helps to improve the accuracy and stability of our method. According to the inverse Boltzmann law, it can be directly derived from the statistical analysis of different types of atom pairs encoded in available crystal complex structures. The scoring function K-Score [20] is considered in this paper and defined as follows:

$$f_4(X) = \sum_{\substack{pl \\ r<r_{cut\text{-}off}}} A_{ij}(r) = \sum_{\substack{pl \\ r<r_{cut\text{-}off}}} -K_B T \ln \left[f^{j}_{vol\text{-}corr}(r) \frac{\rho^{ij}_{seg}(r)}{\rho^{ij}_{bulk}} \right], \tag{6}$$

where K_B is the Boltzmann constant; T is the absolute temperature; $f^{j}_{vol\text{-}corr}$ is the ligand volume correction factor; $\rho^{ij}_{seg}(r)$ is the number density of atom pair ij that occurs in a

spherical shell with a thickness of Δr ranging from r to $r + \Delta r$; ρ_{bulk}^{ij} expresses the number density when no interaction occurs between i and j.

$U_i(X^k)$ is the normalized objective function. In order to improve the stability, the values of the last two generations are used in Eq. (7). Then, the normalized score $U_i(X^k)$ is represented as follows:

$$U_i(X^k) = \frac{f_i(X^k)}{(f_i(X^{k-1}) + f_i(X^{k-2}))/2}, \tag{7}$$

where k is the number of iteration in the optimizing process, and X is the optimal solution of the iteration.

Eq. (1) is a complex single-objective and multi-constraint optimization problem. Because of the huge searching space, it is very difficult to get the best solution. Genetic algorithms (GA) provide such a capability of adaptation and searching in many optimal design problems. In this paper, an improved adaptive GA is adopted [7], in which an entropy-based searching technique with multi-population and the quasi-exactness penalty function is developed to ensure rapid and steady convergence. C_1, … , C_4 are also design variables of GA. During optimization searching, they are approaching a certain value.

For multi-population genetic strategy, the genetic algorithm begins from generating arbitrarily m populations with all the same searching space, i.e. design space. For the improved genetic algorithm with narrowing of the search space, we need only to know efficient narrowing coefficients for the searched space. Shannon's theorem [15] has wide-ranging applications in both communications and data storage applications. This theorem is of foundational importance to the modern field of information theory [2]. There are similarities between the process of optimization and communication of information theory. Information entropy or Shannon entropy H of a discrete set of probabilities p_1, … , p_n is defined by:

$$\begin{aligned} &H = -\sum_{i=1}^{n} p_i \ln p_i \\ &\text{s.t.} \sum_{i=1}^{n} p_i = 1, \quad p_i \in [0, 1]. \end{aligned} \tag{8}$$

Results and discussion

To evaluate the method, we performed the numerical experiment with 134 protein-ligand complexes from the publicly available GOLD test set. This set was originally proposed by Jones [6]. The number of heavy atoms of the ligands ranged from 6 to 55, and the rotatable bond number of ligands ranged from 0 to 22. According to the biological interacting pairs, each complex was divided into a probe molecule and a docking ligand. Protein molecule was obtained by excluding ligands, all structural water molecules, cofactors, and metal ions from the receptor PDB file. Then the mol2 file was generated by adding hydrogen atoms and Kallman charge. Residues around the bound ligand within a radius of 6.5 Å were isolated from the protein to define as the active site. The ligands were then prepared by adding hydrogen atoms and Gasteiger-Marsili atomic charges.

1EAP is a catalytic antibody with a serine protease active site [21]. The hydrophobic surface of active pocket of 1EAP is shown in Fig. 1. The ligand shown in green is the native pose derived from the crystal structure. In the figure, blue part indicates the most hydrophilic surface and orange and red parts indicate the most hydrophobic surface. The active pocket of 1EAP is a cavity and its native pose is almost totally included in the cavity. As shown in Fig. 1, the hydrophobicity is very strong for the majority of the surface is orange. The docking results of 1EAP are shown in Fig. 2. The optimization procedure of 1NCO is provided in Fig. 3. The solid line is binding energy of the force-field part of this paper. The iteration number of the docking procedure of this paper is 74, and that of the force-field score is 185.

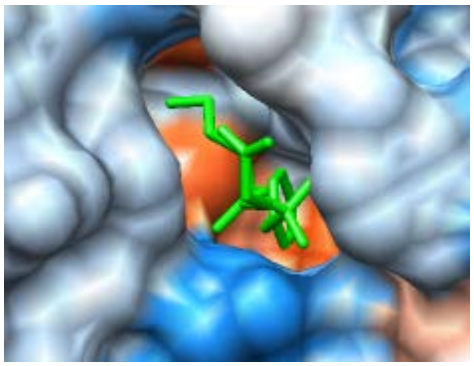


Fig. 1 Active pocket and native pose (green stick) of 1EAP

Docking accuracy is the primary criterion to evaluate docking methods [5]. It is based on the RMSD values of the locations of all of the heavy atoms in the crystal structure. In general, the docking accuracy is acceptable if the RMSD value between the docked pose and X-ray crystal structure is less than 2.0 Å. According to the RMSD values, the accuracy was assigned to four categories. The first category, excellent category, is for those predictions in which the top scoring pose was below 0.5 Å from experimental results. If the RMSD values are between 0.5 Å and 2.0 Å, the results belong to good category. For the third category, close category, the RMSD values are between 2.0 Å and 3.0 Å. For the last category, wrong category, RMSD values are larger than 3.0 Å.

To date, many docking programs are available. Glide [4], GOLD, Surflex [5], FlexX and Dock6 are the commonly used docking programs. The above programs are based on the assumption of rigid receptor for the assumption is conducive to cut off computing time largely. Docking results of flexible receptor are often better than rigid receptor's. In the paper, we selected Dock6 (with flexible default parameters) as a flexible receptor program. Table 1 presents the ratios at different RMSD ranges of these programs.

Table 1. RMSD ratios of this paper and 6 commonly used docking programs

RMSD	Percent (%)						
	This paper	Glide	GOLD	Surflex	FlexX	DOCK6	Dock6-F
≤ 0.5	0.44	0.29	0.08	0.16	0.03	0.15	0.09
> 0.5, ≤ 1.0	0.33	0.19	0.27	0.32	0.18	0.15	0.32
> 1.0, ≤ 2.0	0.06	0.23	0.31	0.29	0.28	0.32	0.39
> 2.0, ≤ 3.0	0.06	0.09	0.05	0.06	0.10	0.12	0.10
≥ 3.0	0.11	0.20	0.28	0.17	0.40	0.27	0.09
Avg. RMSD	1.27	1.98	3.19	2.15	3.69	2.13	1.46

As shown in Fig. 2, the poses obtained in this paper (blue) and native pose (green) coincide with each other very well. Binding score of single force-field function is –28.57 and that obtained in this paper is –20.81. From the point view of force-field, the result of –28.57 should be better (lower score is better). However, the RMSD value of 1EAP obtained in this paper is 0.36, and the RMSD of single force-field is 5.64. That means simple consideration of force-field function fails to provide the best result for 1EAP. The score of hydrophilic part obtained in this paper is –150.70; deformation part is 9.0 and PMF part is –351.55 (scores of different kinds are not comparable with each other). Without considering the hydrophobicity or other factors, single force-field function produces a wrong solution. The deviation of the results of force-field function (yellow) from native pose is very big.

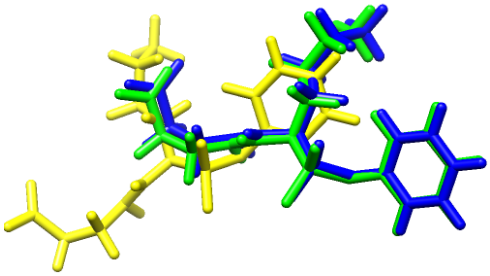


Fig. 2 Results of 1EAP: this paper (blue),
single force-field function (yellow) and native pose (green)

In many cases of test data set, other factors (hydrophobic, deformation and PMF) help force-field function to avoid local optimal solution. As indicated in Fig. 3, the force-field optimization is a continuous decreasing procedure, while multiple factors of this paper can hinder force field from falling into its minimum local energy. The RMSD value of this paper of 1NCO is 0.19, and the RMSD of single force-field is 11.9. Compared with individual force-field function, this paper has the higher capability of finding global optimal solution.

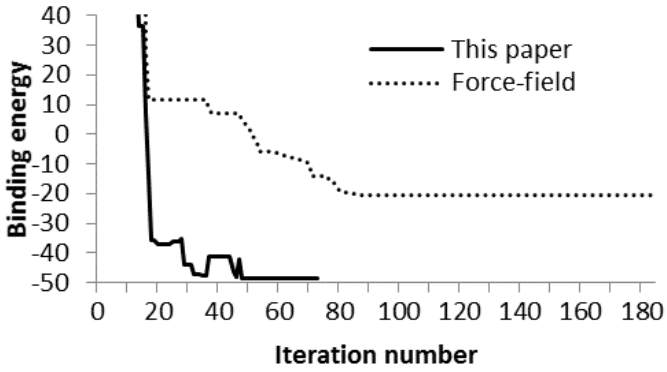


Fig. 3 The optimization procedure of the force-field and this paper of 1NCO

With the method proposed in this paper, we obtained 59 (44%) excellent docking solutions with a RMSD value below 0.5 Å, 52 (39%) good predictions with RMSD between 0.5 and 2.0 Å and only 15 (11%) wrong predictions (RMSD value larger than 3.0 Å). And the average RMSD obtained in this paper is 1.27. In view of RMSD, the method proposed in this paper is excellent among these programs. By considering the conformational situation of the receptor during the docking process, flexible docking can decrease the ratio of wrong prediction.

However, the ratio of excellent results could not be better because it adopted the force-field score function.

Computing time is another important evaluation criterion for a docking method. GA can find the optimum solution under the probability of 1 if the iteration number becomes large enough. The docking accuracy of this paper could be better if it has more computing time. However, simply improving the docking accuracy is pointless, without considering calculation speed. The minimum computing time with the method proposed in this paper is 115.2 s. The maximum computing time is 1895.7 s and the average is 563.9 s. The average computing time of flexible docking is 590.6 s. Considering the high docking accuracy, the computing time with the method proposed in this paper is acceptable.

Conclusion

In this work, we presented a coefficient adaptive method for multiple scoring factors to improve the accuracy of the molecular docking. Based on force-field scoring function, we also considers hydrophobic, deformation and PMF as well in the method. Instead of simple combination with fixed weight, coefficients are adaptive in searching procedure. GA with the multi-population evolution and entropy-based searching technique with narrowing down space is used to solve the optimization model for molecular docking.

The results of the docking experiments on the 134 diverse complexes from the GOLD test data set have shown an obvious improvement of the docking accuracy. This paper has 59 (44%) excellent docking solutions with a RMSD value below 0.5 Å, and 55 (41%) good predictions with RMSD between 0.5 and 2.0 Å. And the average RMSD value (1.22) in this paper is still good. In the view of docking accuracy, our method is better than the other 6 docking programs.

The results indicate that our method can help the force-field function to produce better docking results by introducing other related docking score factors. It is an effective method to improve the docking accuracy. Analyses of failure case indicated that the method proposed in this paper is not helpful in some cases. It is expected to obtain further improvement of this method in our next work.

Acknowledgements

This work was supported by the National Natural Science Funds of China (No. 11202049, No. 11072048, No. 61170168 and No. 61170169).

References

1. Cheng T., Q. Li, Z. Zhou, Y. Wang, S. H. Bryant (2012). Structure-based Virtual Screening for Drug Discovery: A Problem-centric Review, AAPS J, 14, 133-141.
2. Cover T. M., J. A. Thomas (1991). Elements of Information Theory, Wiley, New York.
3. Cross J. B., D. C. Thompson, B. K. Rai, J. C. Baber, K. Y. Fan, Y. Hu, C. Humblet (2009). Comparison of Several Molecular Docking Programs: Pose Prediction and Virtual Screening Accuracy, J Chem Inf Model, 49, 1455-1474.
4. Friesner R. A., J. L. Banks, R. B. Murphy, T. A. Halgren, J. J. Klicic, D. T. Mainz, M. P. Repasky, E. H. Knoll, M. Shelley, J. K. Perry, D. E. Shaw, P. Francis, P. S. Shenkin (2004). Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy, J Med Chem, 47, 1739-1749.
5. Jain A. N. (2003). Surflex: Fully Automatic Flexible Molecular Docking using a Molecular Similarity-based Search Engine, J Med Chem, 46, 499-511.

6. Jones G., P. Willett, R. C. Glen, A. R. Leach, R. Taylor (1997). Development and Validation of a Genetic Algorithm for Flexible Docking, J Mol Biol, 267, 727-748.
7. Kang L, H. L. Li, H. L. Jiang, X. C. Wang (2009). An Improved Adaptive Genetic Algorithm for Protein-ligand Docking, J Comput Aided Mol Des, 23, 1-12.
8. Kang L., Q. Guo, X. C. Wang (2012). A Hierarchical Method for Molecular Docking using Cloud Computing, Bioorganic & Medicinal Chemistry Letters, 22, 6568-6572.
9. Kellenberger E., J. Rodrigo, P. Muller, D. Rognan (2004). Comparative Evaluation of Eight Docking Tools for Docking and Virtual Screening Accuracy, Proteins: Struct Funct Genet, 57, 225-242.
10. Kontoyianni M., G. S. Sokol, L. M. McClellan (2005). Evaluation of Library Ranking Efficacy in Virtual Screening, J Comput Chem, 26, 11-22.
11. Meng E. C., B. K. Shoichet, I. D. Kuntz (1992). Automated Docking with Grid-based Energy Evaluation, J Comput Chem, 13, 505-524.
12. Moitessier N., P. Englebienne, D. Lee, J. Lawandi, C. R. Corbeil (2008). Towards the Development of Universal, Fast and Highly Accurate Docking/Scoring Methods: A Long Way to Go, Brit J Pharmacol, 153, S7-S26.
13. Neves M. A. C., M. Totrov, R. Abagyan (2012). Docking and Scoring with ICM: The Benchmarking Results and Strategies for Improvement, J Comput Aided Mol Des, 26, 675-686.
14. Schulz-Gasch T., M. Stahl (2003). Binding Site Characteristics in Structure-based Virtual Screening: Evaluation of Current Docking Tools, J Mol Mod, 9, 47-57.
15. Shannon C. E. (1948). A Mathematical Theory of Communication, Bell Syst Technical J, 27, 379-423, 623-656.
16. Stahl M., M. Rarey (2001). Detailed Analysis of Scoring Functions for Virtual Screening, J Med Chem, 44, 1035-1042.
17. Villoutreix B. O., R. Eudes, M. A. Miteva (2009). Structure-based Virtual Ligand Screening: Recent Success Stories, Comb Chem High Throughput Screen, 12, 1000-1016.
18. Wang R., L. Lai, S. Wang (2002). Further Development and Validation of Empirical Scoring Functions for Structure-based Binding Affinity Prediction, J Comput Aided Mol Des, 16, 11-26.
19. Weiner S. J., P. A. Kollman, D. T. Nguyen, D. A. Case (1986). An All Atom Force Field for Simulations of Proteins and Nucleic Acids, J Comput Chem, 7, 230-252.
20. Zhao X. Y., X. F. Liu, Y. Y. Wang, Z. Chen, L. Kang, H. L. Zhang, X. M. Luo, W. L. Zhu, K. X. Chen, H. L. Li, X. C. Wang, H. L. Jiang (2008). An Improved PMF Scoring Function for Universally Predicting the Interactions of a Ligand with Protein, DNA, and RNA, J Chem Inf Model, 48, 1438-1447.
21. Zhou G. W., J. Guo, W. Huang, R. J. Fletterick, T. S. Scanlan (1994). Crystal Structure of a Catalytic Antibody with a Serine Protease Active Site, Science, 265, 1059-1064.

Zhengfu Li, M.Sc., Ph.D. Student
E-mail: lizhengfu@hotmail.com

Received his Master degree (2006) from Dalian University of Technology. Now he is a doctor candidate at Dalian University of Technology University. His current research interests include cloud computing, distributed systems and computer-aided drug design.

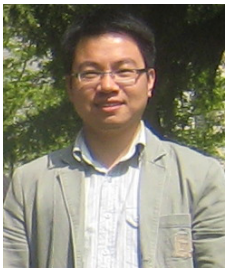
Prof. Xicheng Wang, Ph.D.
E-mail: guixum@dlut.edu.cn

Received his Ph.D. (1989) from Dalian University of Technology. Prof. Wang mainly engages in structural design and optimization, computer-aided pharmaceutical design, parallel algorithms and software development. He has presided over a number of high performance computing and the National Natural Science Foundation funds research work.

Prof. Keqiu Li, Ph.D.
E-mail: keqiu@dlut.edu.cn

Received his Ph.D. (2005) from Japan Advanced Institute of Science and Technology. Now Prof. Li is the vice dean of school of computer science and technology, Dalian University of Technology. His research interests include web technology, grid/cloud computing, mobile computing and network and security.

Junfeng Gu, Ph.D.
E-mail: jfgu@dlut.edu.cn



Received his Ph.D. (2009) from Dalian University of Technology. Now he is a lecturer of department of engineering mechanics, Dalian University of Technology. His research interests include computational mechanics and computational biology.

Ling Kang, Ph.D.
E-mail: kangling@neusoft.edu.cn

Received her Ph.D. (2008) from Dalian University of Technology. Now she is an associate professor of department of computer science and technology, Dalian Neusoft University of Information. Her research interests include cloud computing and computer-aided drug design.