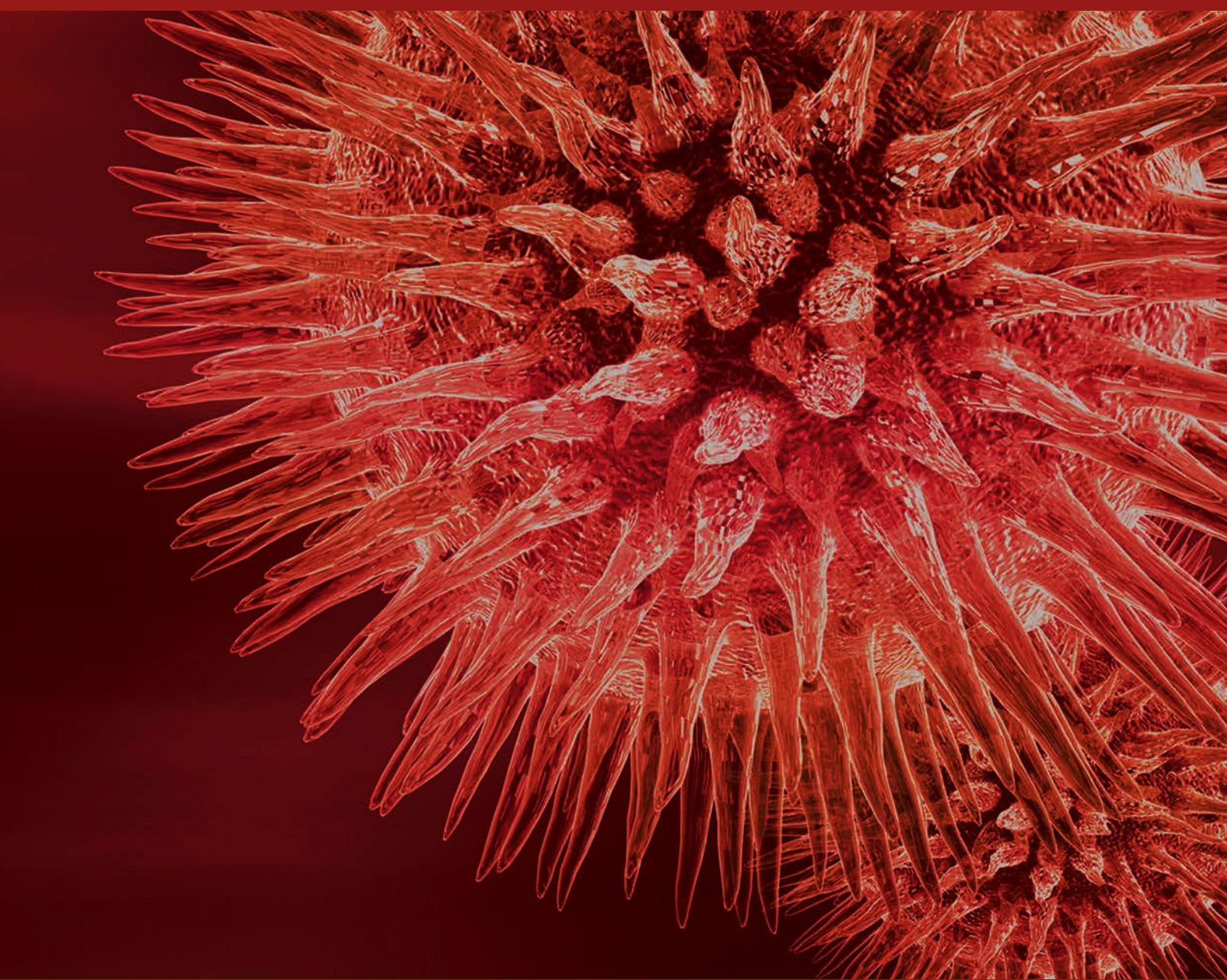
BioMed Research International

Network Proteomics: From Protein Structure to Protein-Protein Interaction

Guest Editors: Guang Hu, Luisa Di Paola, Filippo Pullara, Zhongjie Liang, and Intawat Nookaew



Network Proteomics: From Protein Structure to Protein-Protein Interaction

BioMed Research International

Network Proteomics: From Protein Structure to Protein-Protein Interaction

Guest Editors: Guang Hu, Luisa Di Paola, Filippo Pullara, Zhongjie Liang, and Intawat Nookaew

Contents

Hindawi Publishing Corporation
BioMed Research International
Volume 2017, Article ID 8929613, 1 page
http://dx.doi.org/10.1155/2017/8929613

Editorial

Network Proteomics: From Protein Structure to Protein-Protein Interaction

Guang Hu,[1] Luisa Di Paola,[2] Filippo Pullara,[3] Zhongjie Liang,[1] and Intawat Nookaew[4]

[1]*Center for Systems Biology, School of Electronic and Information Engineering, Soochow University, Suzhou 215006, China*
[2]*Unit of Chemical-Physics Fundamentals in Chemical Engineering, Department of Engineering, Università Campus Bio-Medico, Rome, Italy*
[3]*Department of Computational and Systems Biology, University of Pittsburgh, Pittsburgh, PA, USA*
[4]*Department of Biomedical Informatics, College of Medicine, University of Arkansas for Medical Sciences, Little Rock, AR, USA*

Correspondence should be addressed to Guang Hu; huguang@suda.edu.cn

Received 3 January 2017; Accepted 4 January 2017; Published 16 February 2017

With the emergence of high-throughput "omics" data, network theory is being increasingly used to analyze biomolecular systems. In particular, proteins rarely act alone, which prompted the arrival of proteomics. "Network proteomics" is just defined as the research area that introduces network theory to investigate biological networks ranging from protein structure networks to protein-protein interaction networks. In addition, the application of network proteomics in biomedical fields has increased significantly. This special issue has collected contributions, not only focusing on the state of the art of methodology in "network proteomics" itself but also focusing on the current status and future direction of their applications in translational medical informatics.

This issue starts with discussing the role of protein structure that participates in interactions. N. Raethong et al. developed a strategy to annotate *Aspergillus oryzae*, which would enhance its metabolic network reconstruction. By using the molecular dynamics simulation and principle component analysis, H. Wan et al. investigated the interaction between the last half repeat in TAL effectors and its binding DNA. This work would give a deeper understanding of the recognition mechanism of protein-DNA interactions.

Protein complexes offer detailed structural characteristics of protein-protein interactions. H. Zhang et al. proposed a method by combining the high frequency modes of Gaussian network model and Gaussian Naive Bayes to identify hot spots, which are residues that contribute largely to protein-protein interaction energy. G. Hu et al. performed a comparative study of elastic network model and protein contact network for protein complexes in case of hemoglobin.

Protein-protein interactions not only are limited to binary associations but also employ special topological structures to perform their biological functions. N. Bernabò et al. approached the comparison of two networks models obtained from two different text mining tools to suggest that actin dynamics affect spermatozoa postejaculatory life. J. Xu et al. also compared the topologies of network modes of drug-kinase interactions between established targets and those of clinical trial ones.

Both protein structures and protein-protein interactions are involved in the subject of a growing number of pharmacological studies. F. Ye et al. identified that compounds DC_C11 and DC_C66 are two small inhibitors against protein-protein interactions between CARM1 and its substrates, while Y. Wang et al. identified that CID_70128824, CID_70127147, and CID_70126881 are three potential inhibitors for targeting the androgen receptor to treat prostate cancer. Finally, a review paper completes the issue. X. Li et al. introduced recent advances in the development of network models of identifying synergistic drug combinations.

Altogether, we wish this issue has given a wider development of structural biology and systems biology with the advantage of biological network analysis as well as prospecting the future of this area towards translational bioinformatics and systems pharmacology.

Guang Hu
Luisa Di Paola
Filippo Pullara
Zhongjie Liang
Intawat Nookaew

Hindawi Publishing Corporation
BioMed Research International
Volume 2017, Article ID 3572394, 18 pages
http://dx.doi.org/10.1155/2017/3572394

Research Article

Combined Ligand/Structure-Based Virtual Screening and Molecular Dynamics Simulations of Steroidal Androgen Receptor Antagonists

Yuwei Wang,[1] Rui Han,[1] Huimin Zhang,[1] Hongli Liu,[1] Jiazhong Li,[1] Huanxiang Liu,[1] and Paola Gramatica[2]

[1]*School of Pharmacy, Lanzhou University, 199 West Donggang Rd., Lanzhou 730000, China*
[2]*Department of Theoretical and Applied Sciences, University of Insubria, Via Dunant 3, 21100 Varese, Italy*

Correspondence should be addressed to Jiazhong Li; lijiazhong@lzu.edu.cn

Received 24 August 2016; Revised 11 October 2016; Accepted 16 November 2016; Published 15 February 2017

Academic Editor: Zhongjie Liang

The antiandrogens, such as bicalutamide, targeting the androgen receptor (AR), are the main endocrine therapies for prostate cancer (PCa). But as drug resistance to antiandrogens emerges in advanced PCa, there presents a high medical need for exploitation of novel AR antagonists. In this work, the relationships between the molecular structures and antiandrogenic activities of a series of 7α-substituted dihydrotestosterone derivatives were investigated. The proposed MLR model obtained high predictive ability. The thoroughly validated QSAR model was used to virtually screen new dihydrotestosterones derivatives taken from PubChem, resulting in the finding of novel compounds CID_70128824, CID_70127147, and CID_70126881, whose in silico bioactivities are much higher than the published best one, even higher than bicalutamide. In addition, molecular docking, molecular dynamics (MD) simulations, and MM/GBSA have been employed to analyze and compare the binding modes between the novel compounds and AR. Through the analysis of the binding free energy and residue energy decomposition, we concluded that the newly discovered chemicals can in silico bind to AR with similar position and mechanism to the reported active compound and the van der Waals interaction is the main driving force during the binding process.

1. Introduction

According to the latest World Cancer Report 2014 [1], prostate cancer (PCa) has become the second most common cancer among men in the world. The morbidity rate of PCa has reached 15%, which is merely 1.7% lower than the leading lung cancer. It is reported that about 1100,000 people were diagnosed as new PCa patients in 2012 [2]. Additionally researchers pointed out that prostate cancer is not the privilege of men; women have similar prostate tissue, which also has the risk of cancer [3].

The androgen receptor (AR), a ligand inducible transcription factor in the nuclear hormone receptor super family [4], plays a critical role in the development and progress of PCa. Natural hormone testosterone (T) and dihydrotestosterone (DHT), known as androgens, are the endogenous ligands of AR. When bound to AR, androgens play significant roles in the sexual development, function, and musculoskeletal growth of male. The main mechanism of androgen action is to regulate the gene expression by means of binding to AR, changing the level of specific proteins in cells, and controlling cell behavior [5]. Therefore, a rational approach to cure PCa is the use of antiandrogens to prevent the interaction of T or DHT with AR.

At present, androgen receptor antagonists, such as bicalutamide and flutamide, are used as main hormone therapies for prostate cancer [6]. Although these antiandrogens exhibit good efficacy in many cases and comprise an important part of effective therapeutics [7–10], the emergence of recurrent and metastatic forms of castration-resistant PCa (CRPC)

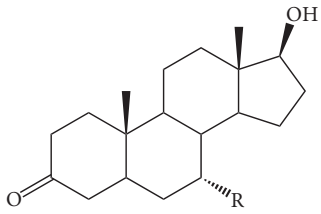


Figure 1: The skeleton structure of 7α-substituted dihydrotestosterones derivatives.

becomes a major challenge, with a median survival of only 1~2 years [11]. A possible reason is that these antiandrogens have partial agonistic activities at high concentration in vitro [12]. Therefore, the discovery of new AR antagonists with high antiandrogen activities is highly expected.

Here, in this study, to aid the research and development of steroidal antiandrogens, we investigated the relationships between a series of 7α-substituted dihydrotestosterone derivatives and corresponding antiandrogen activities. The vital features related to the bioactivities were explored, and a linear quantitative structure-activity relationship (QSAR) model was established according to OECD principles [13], using the QSARINS program [14, 15]. Then the QSAR model, thoroughly and strictly validated by various internal and external validation techniques, is used to virtually screen new dihydrotestosterones, without experimental bioactivities, downloaded from PubChem database [16]. Besides, molecular docking and molecular dynamics (MD) simulations are used to study the possible binding mode of compounds owning high in silico activities with androgen receptor. At last, the most active compounds with good binding affinities to AR, as highlighted by the Insubria graph [17], are proposed for experimental research group to test the antiandrogen activities in the future.

2. Materials and Methods

2.1. Data Set. The success of any QSAR model depends on accurate and clean training data, proper representative descriptor selection methods, suitable statistical methods, and, most critically, both internal and external validation of resulting methods [18, 19]. Here, in this work, a set of 36 7α-substituted dihydrotestosterones derivatives were taken from literatures [20, 21]. The skeleton structure of these derivatives is shown in Figure 1, in which R group represents amine, carboxylic acids, and halogens, and so forth.

These molecules were divided into a training set and a prediction set according to the structure diversity in QSARINS. Finally, 29 compounds were included in the training set and 7 compounds were in the prediction set (prediction set a). The experimental values, half maximal inhibitory concentration (IC_{50}) expressed in nM, were converted into negative logarithmic units marked as pIC_{50}, which was used as dependent variables in the QSAR analyses. The studied molecular structures and corresponding antiandrogen activities were listed in Table 1.

2.2. Descriptors Calculation. To describe a molecule, the molecular structures were firstly sketched in HyperChem program [22] and then were optimized to the minimum energy conformation by using AM1 method. After minimization, we submit the structures to DRAGON 5.5 software [23] to calculate 2914 descriptors including zero-, one-, two-, and three-dimensional (0D, 1D, 2D, and 3D), charge descriptors, and molecular properties. The related theories of the molecular descriptors are provided by DRAGON software, and the calculation procedure is clarified in detail, in the Handbook of Molecular Descriptors [24].

In order to facilitate the successive feature selection process, the constant and near constant descriptors were removed. Besides, if pairwise correlation of two descriptors is larger than 0.98, the one showing the highest pairwise correlation with others will be excluded. Finally, 358 descriptors remained for the next variable selection process.

2.3. QSAR Model Generation. After descriptor calculation, genetic algorithm (GA) implemented in QSARINS software was used to select descriptors. The final model was built by using MLR method based on the selected descriptors, named GA-MLR. The first step of GA is to produce a set of solutions randomly which is called initial population. Each solution, a model based on the contained descriptors by using multiple linear regressions method, is called a chromosome. Subsequently, the fitness function, Friedman LOF, is used to evaluate the fitness of these individuals:

$$\mathrm{LOF} = \left\{ \frac{\mathrm{SSE}}{\left[1 - \left(c + dp/n\right)\right]} \right\}^2. \quad (1)$$

Here, SSE represents the sum of squares of errors, c is the number of basis function, d is the smoothness factor (default 0.5), p is the number of features in the model, and n is the number of samples for model construction. In the successful selection stage, the fittest individuals evolve to the next generation. Then crossover and mutation operators were performed to generate new individuals. Finally a new population is formed consisting of the fittest chromosomes. The above evolution continues until the stop conditions are satisfied. The related parameters that control the GA performance are list as follows: population size (200), maximum generations (10000), and mutation probability (0.05).

2.4. Model Validation. QSARINS is based on GA-MLR method and performed various tools to a rigorous internal and external validation, based on the different validation criteria, as well as for the check of model applicability domain.

The robustness and stability of the built model were validated by several statistical parameters, such as determination coefficient (R^2), leave-one-out (LOO) cross-validation Q^2_{LOO}, and root mean squared error (RMSE). Besides, leave-many-out (LMO) cross-validation method was also performed and Q^2_{LMO} was reported. Randomization technology, by reordering the independent variable, was used to exclude the possibility of the chance correlation. Generally, the correlation coefficient of the built QSAR model should exceed Y randomized generated model. After Y scrambling was

Table 1: The studied compounds and corresponding experimental and predicted pIC_{50} values.

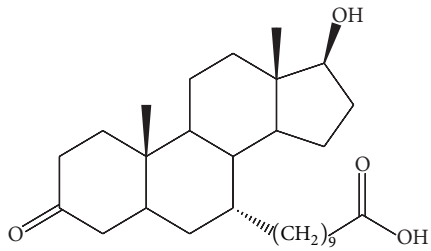
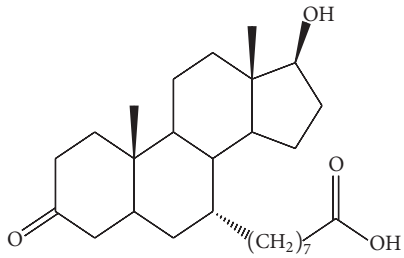
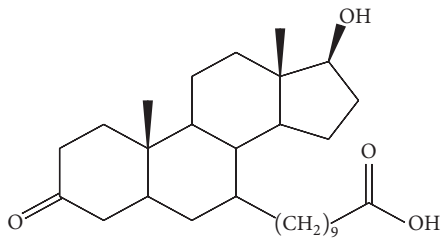
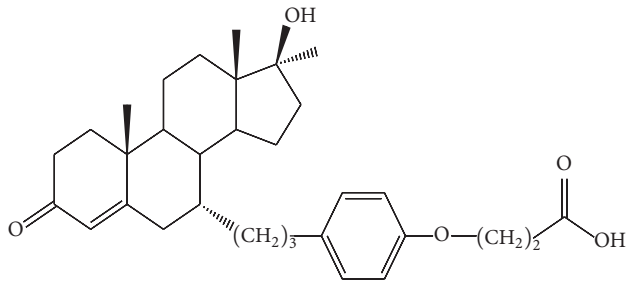
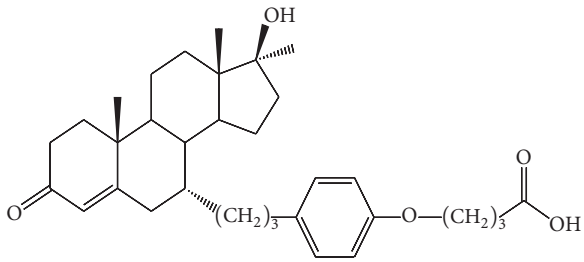
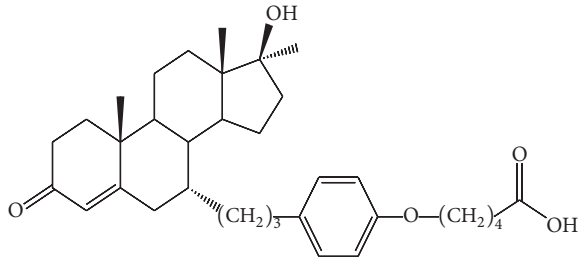
Number	Structure	Experimental pIC_{50}	Predicted pIC_{50}
1[a]		6.04	5.91
2[a]		6.14	6.15
3[a]		5.70	5.81
4		6.77	6.88
5		6.18	6.23
6[a]		6.52	6.71

Table 1: Continued.

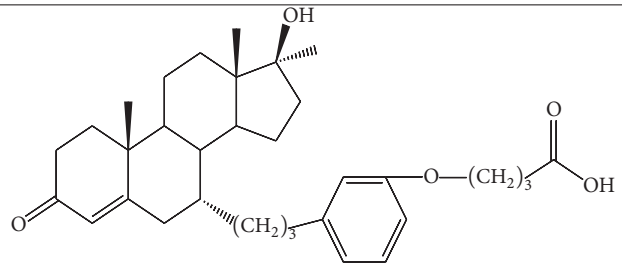
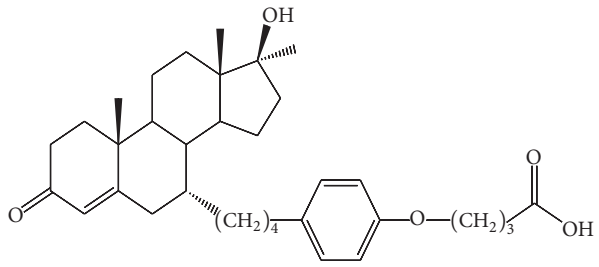
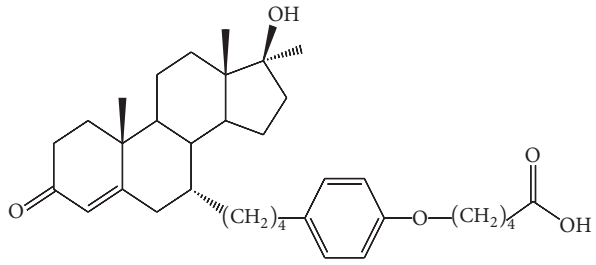
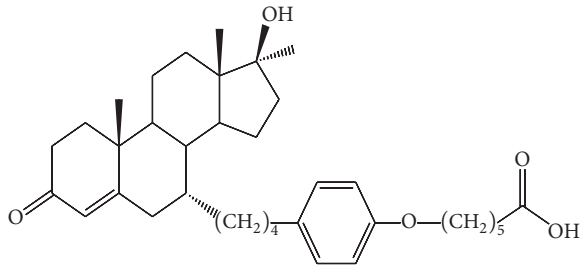
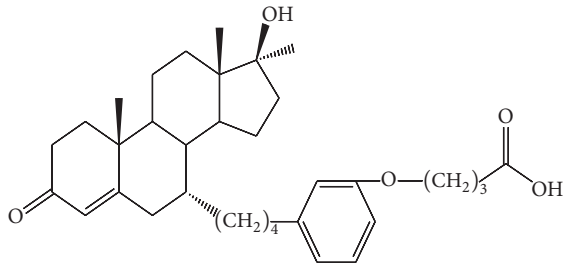
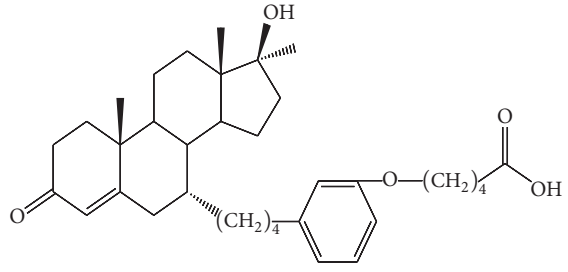
Number	Structure	Experimental pIC_{50}	Predicted pIC_{50}
7		5.24	6.76
8		6.07	5.95
9		6.60	6.53
10		6.38	6.27
11		6.04	6.06
12[a]		6.48	6.48

Table 1: Continued.

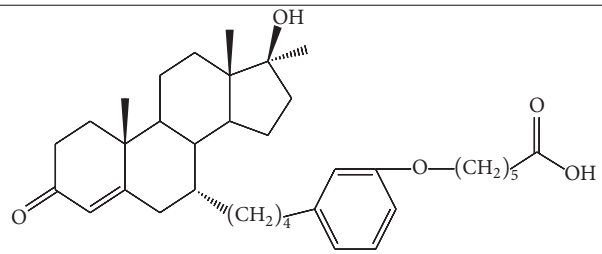
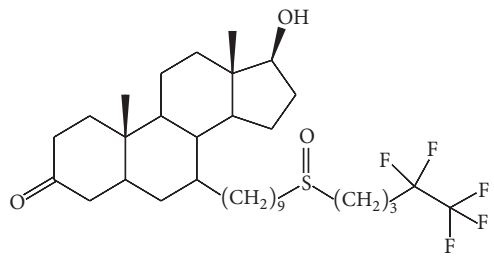
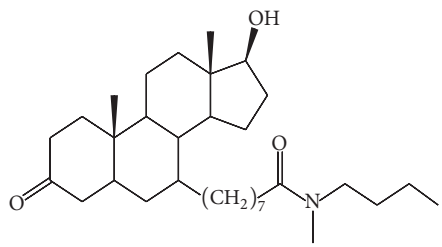
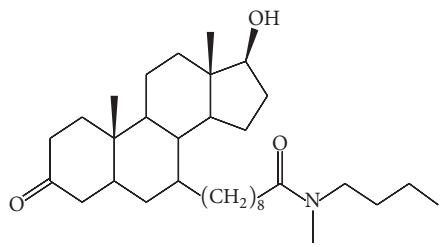
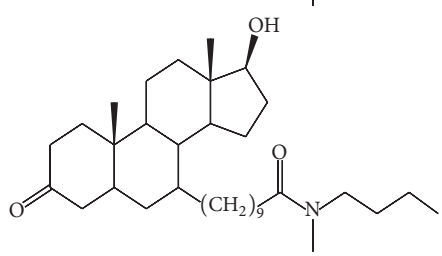
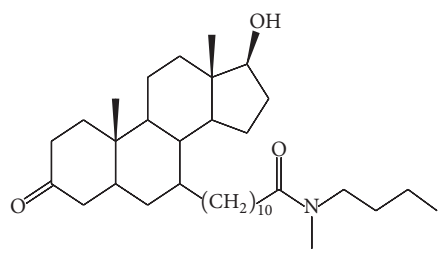
Number	Structure	Experimental pIC_{50}	Predicted pIC_{50}
13		6.28	5.80
14		5.54	5.47
15		6.47	6.17
16		5.47	5.81
17[a]		5.74	5.51
18		5.89	5.80

Table 1: Continued.

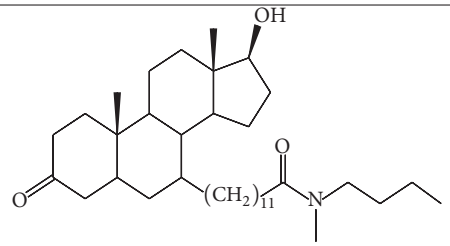
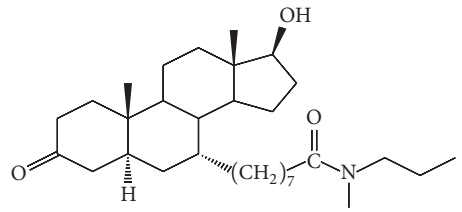
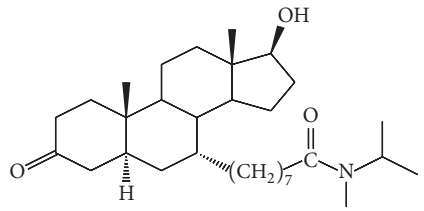
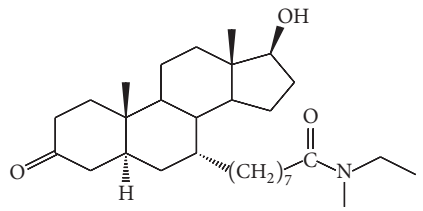
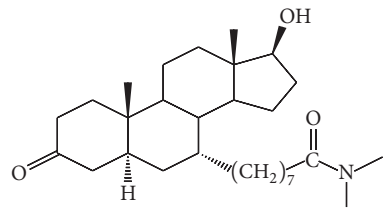
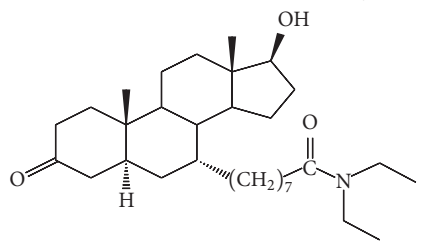
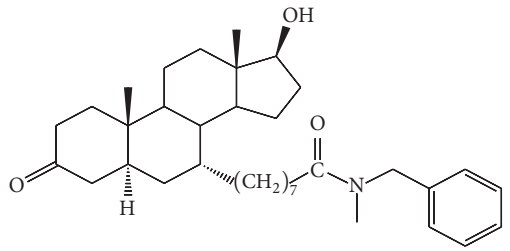
Number	Structure	Experimental pIC_{50}	Predicted pIC_{50}
19		5.46	5.61
20		5.96	6.35
21[a]		6.38	6.12
22		6.38	6.07
23		6.72	6.64
24		6.52	6.46
25		5.80	5.84

Table 1: Continued.

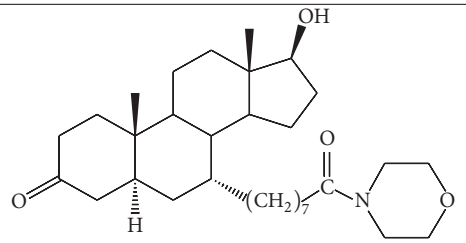
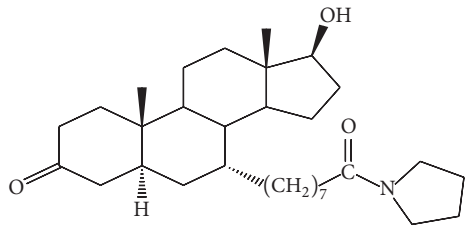
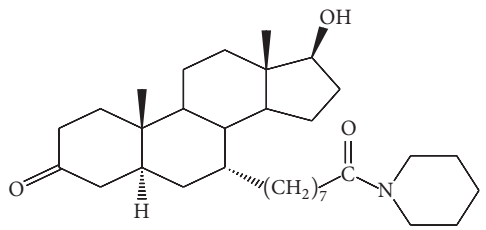
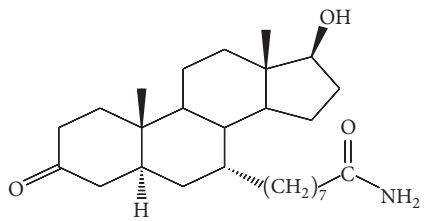
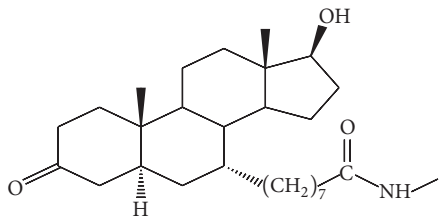
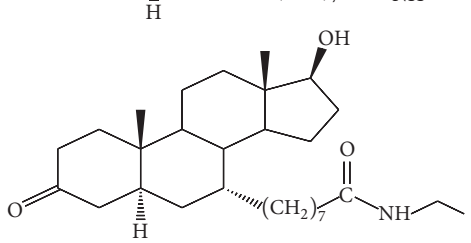
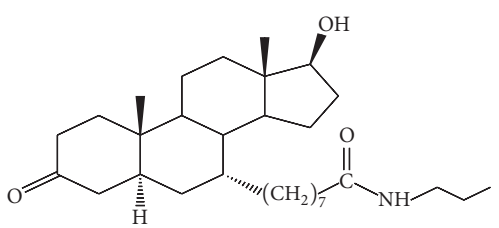
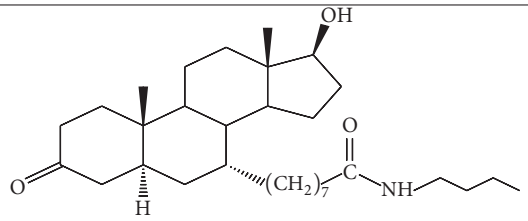
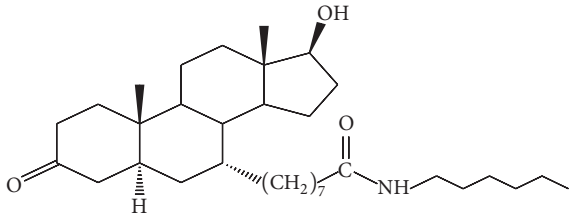
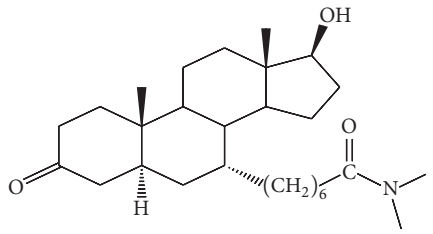
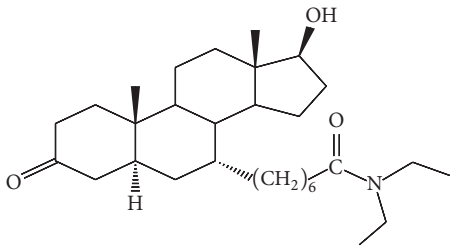
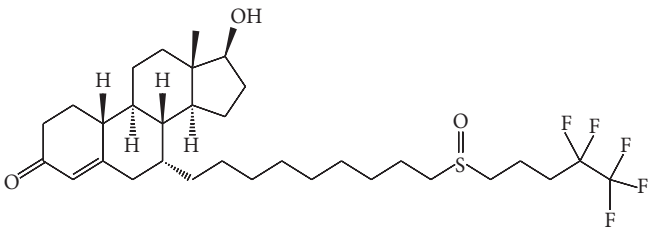
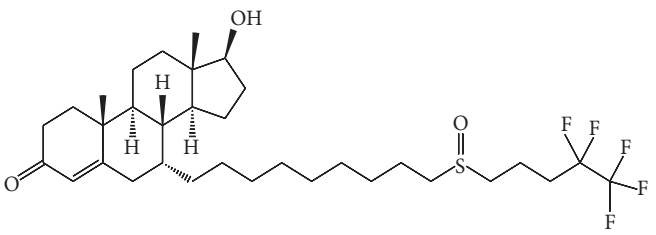
Number	Structure	Experimental pIC_{50}	Predicted pIC_{50}
26		6.14	6.30
27		6.52	6.46
28		6.31	6.48
29		6.35	6.18
30		6.32	6.06
31		6.03	5.80
32		5.74	5.60

Table 1: Continued.

Number	Structure	Experimental pIC_{50}	Predicted pIC_{50}
33		5.20	5.50
34		5.18	5.24
35		6.70	6.59
36		6.28	6.59
37[b]		6.02	6.38
38[b]		5.66	5.49

[a]The prediction set a; [b]the prediction set b (Bradbury et al., 2011).

carried out with iterations of 5,000, the average value of squared correlation coefficient of the randomized models R^2 and Q^2_{LOO} was reported.

A good QSAR model should also have satisfactory predictive ability. The best way to evaluate the predictive ability of a model is its validation by new compounds, called prediction set, which do not participate in the process of model building. After the activities of the prediction set samples were predicted, the agreement between the experimental and predicted values was calculated as a measure of a QSAR

model quality. Here we adopt several ways to calculate this agreement, Q^2_{F1}, Q^2_{F2} [25], Q^2_{F3} [26], and CCC [27–29].

All the above external validation parameters were calculated in the software QSARINS and were combined to evaluate the predictive ability of the proposed model.

2.5. Applicability Domain. To validate the practical applicability of a model to a new chemical, the applicability domain (AD), a theoretical domain which is defined by means of the selected descriptors in the process of modeling, should be defined properly. In this research, the AD was quantitatively assessed by the leverage approach [30, 31]. The leverage (hat, h) was calculated by $h_i = x_i(X^TX)x_i{}^T$ ($i = 1, \ldots, m$), where x_i was the descriptor row-vector of the query compound i and X was the $n * p$ matrix of the training set (p is the number of model descriptors). The limit of model domain was quantitatively defined by the leverage cutoff (h^*), set as $3(p + 1)/n$. A leverage greater than h^* means that the query was outside of the model structural AD, so the predictions were extrapolations of the model and could be less reliable. The AD for chemicals with experimental data was verified by the Williams plot, where the hat values versus the standardized residuals were plotted, while the AD for chemicals without experimental data, which were analyzed in the virtual screening, was verified by the Insubria graph where the hat values were plotted versus the predicted responses [14, 18].

2.6. Virtual Screening of Potent Steroidal Antiandrogens. To explore more 7α-substituted dihydrotestosterones and to find similar derivatives with high antiandrogen activities, the studied skeleton structure was used as a query to search PubChem database for new dihydrotestosterones, without experimental bioactivities. Then the established MLR model, after thoroughly being validated internally and externally, was used to predict the antiandrogenic activities of these new dihydrotestosterones, verifying the AD.

Besides, molecular docking was employed to investigate the possible interaction mechanisms of the samples owning high in silico antiandrogenic activities with AR. Particularly, comprehensively considering the docking speed and accuracy, LigandFit, which is commonly used as a flexible docking method executed in the commercial software Discovery Studio 2.5 [32], was applied into the progress of structure-based virtual screening. The protein structure of AR was firstly downloaded from RCSB Protein Data Bank [33] (PDB entry code: 1T65) and imported in docking process. All ligands and water molecules were removed at first, the charge and polar hydrogen atoms were added, and the incomplete residues were corrected.

2.7. Molecular Dynamics (MD) Simulations. The molecular dynamics (MD) simulations were carried out using the Amber 14 software package [34]. MD is a commonly used methodology in exploring the interaction between ligand and protein. We have investigated the interaction mechanisms of R-bicalutamide/S-1 with WT/W741L AR using molecular dynamics simulations [35]. The docked structures of AR (PDB ID: 1T65) with the reported most active compound number 4 and novel chemicals with high in silico activities were used as the initial structures for MD simulations. During the process of docking, taking into consideration the fact that these residues collide significantly with the compounds, Helix 12 of AR was removed in the model as executed in the literature [36].

All missing hydrogen atoms of the AR were added by the LEaP module of the Amber 14 package. To maintain the electroneutrality of all the studied complexes, the appropriate number of chloride counterions was added. Then each complex was immersed into a cubic periodic box of TIP3P water model [37] with at least 10 Å distance around the complex.

For the ligand, the GAFF parameter assignments [38] were made by using Antechamber program and the partial charges were assigned by using the AM1-BCC method [39].

Amber 14 package and the Amberff03 force field were used for all molecular dynamics simulations. Sander program was carried out for the energy minimization and equilibration protocol. First, energy minimization of four complexes was done through three stages, using the steepest descent method switched to a conjugate gradient every 2500 steps for a total of 5000 steps with a nonbonded cutoff of 10 Å. In the first stage, to enable the added TIP3P water molecules to adjust to their proper orientations, the AR and ligand were restrained with 5.0 kcal mol^{-1} Å^{-2}. In the second stage, to enable the AR to find a better way of accommodating ligand, the protein backbone was restrained with 3.0 kcal mol^{-1} Å^{-2}. In the third stage, the entire solvated system was minimized without any restraint. Additionally, gradual heating, density, and equilibration protocols were performed. During the 100 ps heating procedure, the system was gradually heated from 0 to 310 K, and then the density was at 310 K for 400 ps, and at last the equilibration was at 310 K for 400 ps.

Afterwards, four 20 ns production MD simulations were carried out with the PMEMD program without any restraints in the isothermal isobaric ensemble (NPT, $P = 1$ atm, $T = 310$ K) MD. The time step was set at 2 fs. 10 Å cutoff was applied to treat nonbonding interactions. During the simulations periodic boundary conditions were employed and all electrostatic interactions were calculated using the particle mesh Ewald (PME) method. The SHAKE algorithm was used to restrain all bond lengths containing hydrogen atoms. All of the coordinate trajectories were recorded every 2 ps throughout all MD runs. To analyze the energy and structure, a total of 500 snapshots of the simulated structures stripped in the last 2 ns stable MD production trajectory at 4 ps intervals were extracted.

2.8. Binding Free Energy Calculations. For each protein-ligand complex, the binding free energy was analyzed by the MM/GBSA method [40]. To compare the AR binding free energies with different ligands, MM/GBSA calculation was applied to the 500 snapshots extracted from the final 2 ns of the MD trajectories. The total free energy of binding free energy was composed of the following molecular species (complex):

Table 2: The selected descriptors used to build QSAR model and corresponding meanings.

Descriptor	Meaning	Descriptor type
IC5	Information content index (neighborhood symmetry of 5-order)	Information indices
GATS5e	Geary autocorrelation, lag 5/weighted by atomic Sanderson electronegativities	2D autocorrelations
DISPp	d COMMA2 value/weighted by atomic polarizabilities	Geometrical descriptors
HATS3u	Leverage-weighted autocorrelation of lag 3/unweighted	GETAWAY descriptors

$$\begin{aligned}\Delta G_{\text{bind}} &= G_{\text{complex}} - G_{\text{protein}} - G_{\text{ligand}} \\ &= \Delta E_{\text{MM}} + \Delta G_{\text{sol}} - T\Delta S,\end{aligned} \tag{2}$$

where G_{complex}, G_{protein}, and G_{ligand} are the free energy of complex, receptor, and ligand, respectively. The free energy for each species (complex, ligand, or receptor) can be decomposed into a gas phase energy (ΔE_{MM}), a solvation-free energy (ΔG_{sol}), and an entropy term ($T\Delta S$).

$$\begin{aligned}\Delta E_{\text{MM}} &= \Delta E_{\text{val}} + \Delta E_{\text{ele}} + \Delta E_{\text{vdw}}, \\ \Delta G_{\text{sol}} &= \Delta G_{\text{p}} + \Delta G_{\text{np}}, \\ \Delta G_{\text{np}} &= \gamma\text{SASA} + \beta,\end{aligned} \tag{3}$$

where the ΔE_{MM} is the sum of the internal energy of bonds, angle, and torsion (ΔE_{val}), electrostatic interaction energy (ΔE_{ele}), and van der Waals interaction energy (ΔE_{vdw}). ΔG_{sol} is solvation-free energy and can be divided into two parts, the polar solvation-free energy (ΔG_{p}) and the nonpolar solvation-free energy (ΔG_{np}). The polar solvation-free energy ΔG_{p} is determined by generalized Born (GB) equation. The values of the dielectric constant for solute and solvent were set as 1 and 80. ΔG_{np} is the nonpolar solvation contribution and was calculated with constants 0.0072 kcal mol^{-1} Å^{-2} for surface tension proportionality constant γ and 0.92 kcal mol^{-1} Å^{-2} for the nonpolar free energy for a point solute β. SASA is the solvent accessible surface area and is determined by recursively approximating a sphere around each atom, starting from icosahedra (ICOSA method). $T\Delta S$ is the entropy term, including the translational, rotational, and vibrational terms of the solute molecules.

2.9. Energy Decomposition. Furthermore, to obtain the contribution of each residue to the binding process, we performed binding free energy decomposition. The MM/GBSA approach was used to calculate the per-residue free energy decomposition, which is based on the same 500 snapshots we have extracted from the last 2 ns of the stable MD trajectory.

2.10. Normal Mode Calculation. Entropy was analyzed by normal mode with AMBER14 NMODE module. Due to the high computational cost in the entropy calculation, 50 snapshots were extracted from the last 2 ns trajectory of the simulation with 40 ps time intervals.

3. Results and Discussion

3.1. The Linear MLR Model. The training set samples, 29 compounds as listed in Table 1, were used to build QSAR model by using GA-MLR methods, and the remaining compounds were used to evaluate the predictive ability of the built model. GA provided a series of linear equations containing different descriptor combinations with different performance, but similarly satisfactory. An excellent QSAR model should have high fitting ability, high cross-validated Q^2_{LOO}, high external predictive ability, and little difference between internal and external predictive ability.

Based on the above principles, a four-descriptor model was selected as the final model. The involved descriptors and corresponding physical-chemical meanings were listed in Table 2. The corresponding model equation and statistic parameters are listed as follows:

$$\begin{aligned}\text{pIC}_{50} &= -2.89\text{IC5} + 1.01\text{GATS5e} \\ &\quad - 3.17\text{DISPp} - 12.99\text{HATS3u} \\ &\quad + 27.13, \\ R^2 &= 0.760, \\ Q^2_{\text{LOO}} &= 0.656, \\ Q^2_{\text{LMO}} &= 0.662, \\ Q^2_{F1} &= 0.739, \\ Q^2_{F2} &= 0.731, \\ Q^2_{F3} &= 0.876, \\ \text{CCC} &= 0.891, \\ \text{RMSE}_{\text{training}} &= 0.226, \\ \text{RMSE}_{\text{prediction}} &= 0.270.\end{aligned} \tag{4}$$

From the linear equation and statistic parameters, we could see that the fitting ability of the final model was relatively high with R^2 of 0.760 and the final model was stable with Q^2_{LOO} of 0.656 and Q^2_{LMO} of 0.662. About the predictive ability of the final model, we could find that Q^2_{F1} and Q^2_{F2} have similar high values. Compared with Q^2_{F1} and Q^2_{F2}, the value of Q^2_{F3} was higher. Besides, the value of CCC was as high as 0.891, surpassing the threshold value of 0.85 as suggested in literature [29] for predictive model. Additionally, the RMSE values for the training set and prediction set were similarly very low. All these parameters indicated the higher external prediction ability of the final model. The interrelation coefficients of the selected descriptors were presented in the Table 3. It could

Table 3: The correlation coefficients (K) of the selected descriptors in the model.

	IC5	GATS5e	DISPp	HATS3u
IC5	1			
GATS5e	0.07	1		
DISPp	0.25	0.36	1	
HATS3u	−0.625	0.19	−0.15	1

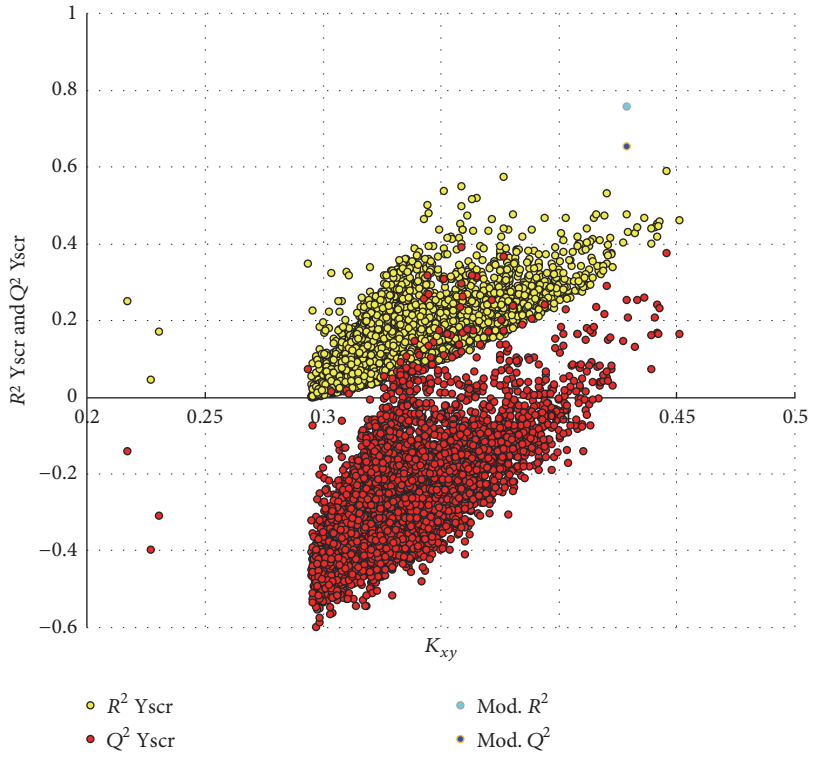


Figure 2: The distribution of R^2 and Q^2 of 5000 iterated Y-scrambled models in comparison to the proposed model performances.

be seen that the highest intercorrelation coefficient K was −0.61 between IC5 and HATS3e, which indicated that the used variables were independent. All results proved that the selected model was reliable, stable, and predictive.

Y randomization technique was carried out with iterations of 5000 in QSARINS. Figure 2 showed the plot of R^2 and Q^2 values versus K_{xy}, automatically obtained in QSARINS. From Figure 2, we could find that the R^2 and Q^2 values of the final model were much higher than the models from scrambled Y-column, because the relationship between molecular structure and response was broken. This result indicated that the relationships between structures of 7α-substituted dihydrotestosterones and corresponding pIC_{50} values did exist in the proposed model, and it was really not obtained by chance.

The predicted pIC_{50} values by MLR model were listed in Table 1. Figure 3 was the scatter plot of the experimental versus the predicted pIC_{50} values. It was obvious that, in Figure 3, all predicted pIC_{50} values were close to the line $y = x$, which indicated that the linear model can accurately predict the antiandrogenic values of these derivatives.

The model applicability domain was evaluated by means of leverage analysis, namely, Williams plot, shown in Figure 4, in which the standardized residuals (σ) and leverage values (h) were plotted. In Figure 4, we could see that all compounds were inside the model structural applicability domain ($h^* = 0.517$) and reasonably well predicted with standard residue smaller than 2.5σ.

After the MLR model was built, we luckily found two new 7α-substituted dihydrotestosterones, showed in Table 1 (marked as "b," prediction set b), with experimental antiandrogen activities from another published literature [41]. These two compounds were additionally used to validate our model. Both of them (numbered in the Williams plot of Figure 4) were located in the applicability domain of the MLR model with h value of 0.332 for compound 37 and 0.237 for compound 38. The predictions on them were quite near to their experimental values. In Figure 3, these two samples (in red) were very near to the line $y = x$. These results further

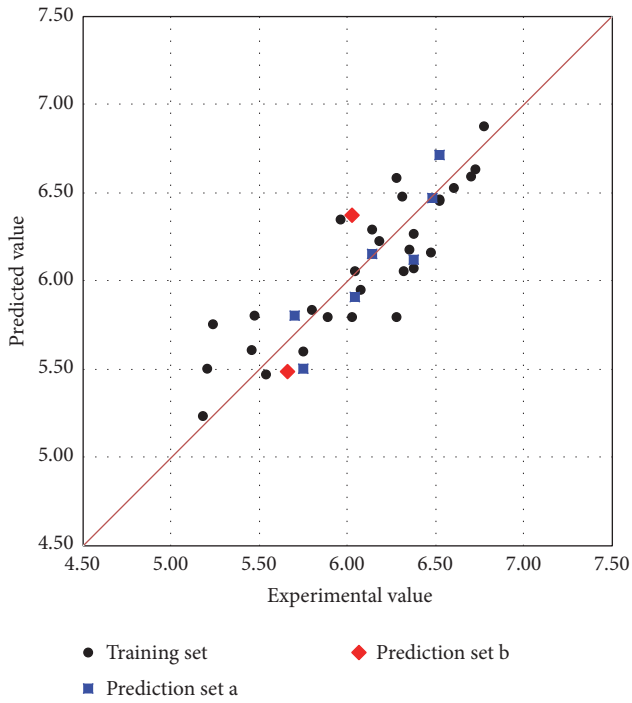


Figure 3: The scatter plot of the experimental versus the predicted pIC_{50} by MLR model.

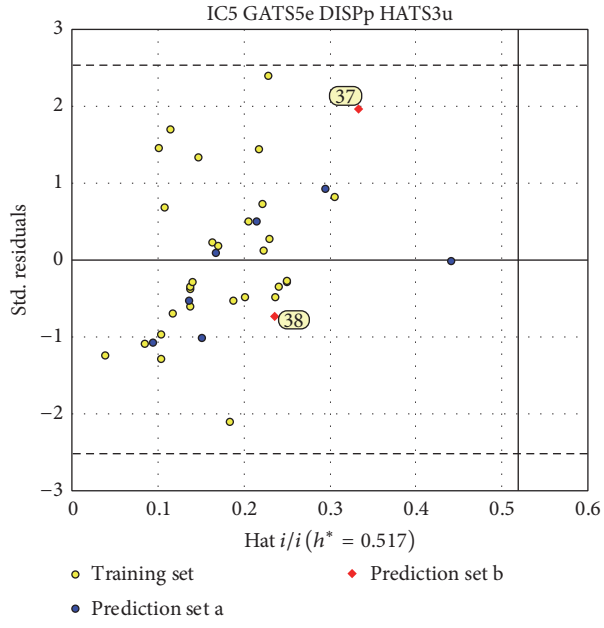


Figure 4: The Williams plot of final MLR model.

indicated the high predictive ability of the proposed MLR model.

By interpreting the meaning of the descriptors used in the model, we could extract vital structural features, to some extent, responsible for the antiandrogenic activities of these steroidal derivatives. IC5 was calculated as the mean

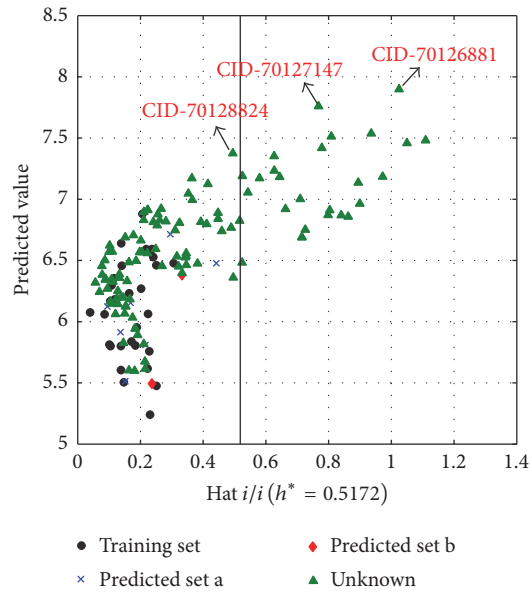


Figure 5: Insubria graph (plot of hat values versus predicted values for the complete compounds).

information content as follows: $IC5 = -\sum_{g-1}^{A_g}(A_g/\mathrm{nAT}) \cdot \log_2(A_g/\mathrm{nAT})$, where g runs over the equivalence classes, A_g was the cardinality of the gth equivalence class, and nAT was the total number of atoms. This index represented a measure of structural complexity per vertex. GATS5e belonged to 2D autocorrelations and was Geary autocorrelation, lag 5/weighted by atomic Sanderson electronegativities. This descriptor was favorable to the antiandrogen activities of these steroidal derivatives. DISPp, geometrical descriptors, indicated the displacement between the geometric center and the center of the polarizability, calculated with respect to the molecular principal axes. HATS3u is a GETAWAY descriptor [42], representing the leverage-weighted autocorrelation of lag 5/weighted by atomic polarizabilities. With the increase of these two descriptors, the bioactivities of the studied compounds decreased.

3.2. Virtual Screening. From PubChem database, we found 110 new 7α-substituted dihydrotestosterone derivatives, without experimental data. By exploring the leverage h values, 77.27% of them were located in the structural applicability domain of the proposed MLR model. Figure 5 is the Insubria graph of these dihydrotestosterones, the plot of leverage values versus predicted pIC_{50}, which was proposed especially for exploring the unknown samples. In Figure 5, most chemicals are in the range of the hat cutoff ($h^* = 0.517$). Inside the model AD, the most active compound is CID_70128824, which has in silico PIC_{50} of 7.37, higher than the reported most active compound number 4 ($pIC_{50} = 6.77$). Outside the model AD, we luckily obtained several samples with higher in silico activities, especially CID_70126881 and CID_70127147, which showed excellent in silico antiandrogen activities as high as 7.90 and 7.76, respectively, even higher than bicalutamide and hydroxyflutamide. The ID of these

Table 4: The 110 new compounds from PubChem database and corresponding predicted activities.

Number	MolID	Pred	AD[a]
1	CID_44421999	6.32	Y
2	CID_44422008	6.49	Y
3	CID_44422014	6.53	Y
4	CID_44422020	6.47	Y
5	CID_44422031	6.34	Y
6	CID_44422034	6.15	Y
7	CID_44422037	6.14	Y
8	CID_44422041	6.07	Y
9	CID_44422043	5.82	Y
10	CID_44422044	5.67	Y
11	CID_44422045	5.62	Y
12	CID_44422047	6.69	N
13	CID_44422053	6.16	Y
14	CID_44422054	6.91	Y
15	CID_44422058	6.90	Y
16	CID_44422064	6.12	Y
17	CID_44422067	5.94	Y
18	CID_44422075	6.25	Y
19	CID_44422080	6.33	Y
20	CID_44433644	6.53	Y
21	CID_44433645	6.20	Y
22	CID_67854257	7.51	N
23	CID_69758112	6.69	Y
24	CID_70126216	7.35	N
25	CID_70126231	7.17	Y
26	CID_70126247	6.91	Y
27	CID_70126297	5.60	Y
28	CID_70126298	5.83	Y
29	CID_70126305	6.56	Y
30	CID_70126327	7.42	N
31	CID_70126491	6.40	Y
32	CID_70126782	6.27	Y
33	CID_70126784	6.24	Y
34	CID_70126798	6.48	N
35	CID_70126802	6.87	N
36	CID_70126837	6.36	Y
37	CID_70126868	6.75	N
38	CID_70126881	7.90	N
39	CID_70126979	6.39	Y
40	CID_70126991	6.89	Y
41	CID_70127144	6.82	Y
42	CID_70127147	7.76	N
43	CID_70127181	6.03	Y
44	CID_70127183	6.35	Y
45	CID_70127185	6.87	N
46	CID_70127188	6.96	N
47	CID_70127192	6.88	Y
48	CID_70127194	6.81	Y
49	CID_70127269	6.35	Y
50	CID_70127287	7.48	N
51	CID_70127296	6.92	N
52	CID_70127297	6.77	Y
53	CID_70127298	6.82	Y
54	CID_70127444	6.80	Y
55	CID_70127446	7.54	N
56	CID_70127567	5.89	Y
57	CID_70127714	7.00	N
58	CID_70127721	6.57	Y
59	CID_70127722	6.82	Y
60	CID_70127760	6.91	N
61	CID_70127771	6.66	Y
62	CID_70127818	7.06	N
63	CID_70127821	7.17	N
64	CID_70128062	6.59	Y
65	CID_70128068	7.19	N
66	CID_70128078	6.06	Y
67	CID_70128083	6.83	Y
68	CID_70128084	6.81	Y
69	CID_70128209	6.75	Y
70	CID_70128238	6.83	Y
71	CID_70128450	6.33	Y
72	CID_70128452	7.23	N
73	CID_70128456	6.71	Y
74	CID_70128462	6.84	Y
75	CID_70128533	6.79	Y
76	CID_70128534	6.92	Y
77	CID_70128574	7.18	N
78	CID_70128608	7.18	N
79	CID_70128824	7.37	Y
80	CID_70128828	6.46	Y
81	CID_70128830	7.00	Y
82	CID_70128847	7.46	N
83	CID_70128902	6.50	Y
84	CID_70128904	6.57	Y
85	CID_70128987	5.60	Y
86	CID_70129065	6.69	N
87	CID_70129161	7.05	Y
88	CID_70129198	6.53	Y
89	CID_70129204	7.14	N
90	CID_70129375	6.18	Y
91	CID_70129377	6.74	Y
92	CID_70129380	7.12	Y
93	CID_9804916	6.38	Y
94	CID_9826776	6.33	Y
95	CID_9827285	6.48	Y
96	CID_9828392	6.12	Y
97	CID_9828587	6.86	N
98	CID_9829426	6.45	Y
99	CID_9891033	6.14	Y
100	CID_9893352	5.94	Y

Table 4: Continued.

Number	MolID	Pred	AD[a]
101	CID_9933226	6.34	Y
102	CID_9935104	6.56	Y
103	CID_9935196	6.57	Y
104	CID_9935788	6.50	Y
105	CID_9936347	6.45	Y
106	CID_9936803	6.36	Y
107	CID_9955874	6.20	Y
108	CID_9957122	6.62	Y
109	CID_9957692	6.57	Y
110	CID_9958161	6.19	Y

[a]AD: model structural applicability domain; Y: compound inside the model structural AD; N: compound outside the model structural AD.

compounds and corresponding predicted pIC_{50} values are listed in Table 4. Though only these three compounds were highlighted here, other samples with high in silico activities, especially those located in the model structural AD, were also worthy of our attention.

To further explore the possible binding mode of the screened compounds, molecular docking was employed to study the interaction between compounds owning high in silico activities (especially CID_70128824, CID_70126881, and CID_70127147), together with the reported most active compound 4 as a comparison, and androgen receptor (PDB ID: 1T65) by using LigandFit module in Discovery Studio 2.5. Firstly, DHT was extracted from crystal structure and redocked into ligand binding pocket to obtain the optimal docking parameters. Secondly, the ligand binding site was defined with the same parameters as DHT. At this point, the radius of SBD_Site_Sphere was set to 10 Å. The other parameters were set by default.

To obtain reasonable conformations of different complex, the top-ranked compounds with lowest RMSD values were extracted. The binding mode of compound 4, CID_70128824, CID_70126881, and CID_70127147 with AR were presented in Figure 6. From this Figure, it could be seen that the docked pose of CID_70128824, CID_70126881, and CID_70127147 located in the same position with similar orientation in the AR ligand binding site to compound 4. All these results indicated that though two of them were outside of the model AD, these three compounds CID_70128824, CID_70126881, and CID_70127147 might have good performance to antagonize androgen receptor and have the potency for further research and development for PCa therapy.

3.3. MD Simulations

3.3.1. System Equilibration. In order to verify whether the studied systems reach equilibrium, the root mean square deviations (RMSDs) of all the backbone atoms of the protein, the C_α atoms for the residues of the active site (residues within 5 Å around ligand), and the heavy atoms of ligand from the initial structure were monitored to examine the dynamic stability of the systems and plotted against time, as shown in Figure 7. The three RMSDs have small fluctuations after 15 ns, implying that the studied systems have reached stability. We used the last 2 ns to analyze the energy and binding modes for the four complexes.

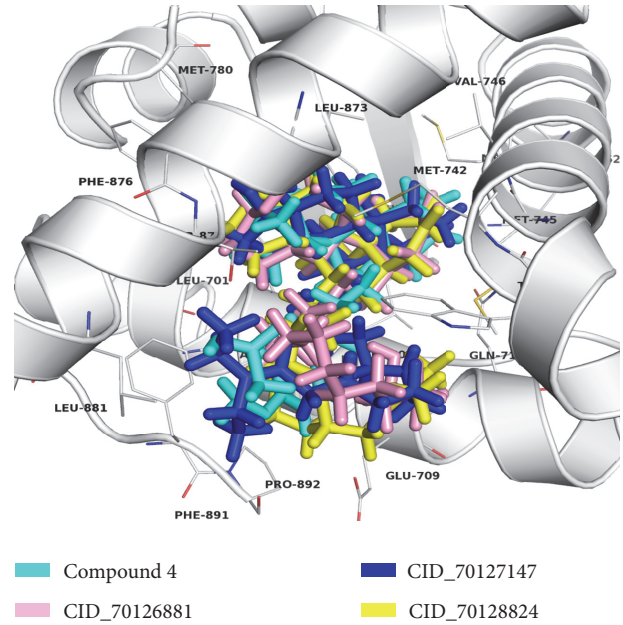


Figure 6: The molecular binding models of compound 4 (cyan), CID_70128824 (yellow), CID_70126881 (pink), and CID_70127147 (blue) in the AR ligand binding site.

3.3.2. Validation of the MD Simulations. We calculated the binding free energy by MM/GBSA method between the four ligands and AR to validate the reliability of the MD simulation. Table 5 lists the binding free energy and all of the energy terms for the four compounds. From Table 5, the four complexes had different binding free energy; the ranking order is CID_70126881 (-41.62 kcal mol^{-1}), CID_70127147 (-33.06 kcal mol^{-1}), CID_70128824 (-31.86 kcal mol^{-1}), and compound 4 (-22.69 kcal mol^{-1}). Different binding free energy means different binding affinity between the four complexes. We have proved that the antiandrogen activity of compounds CID_70126881, CID_70127147, and CID_70128824 are higher than the published best one (compound 4) by MLR model. In addition, the ranking of the calculated binding free energy was consistent with their in silico bioactivities order.

3.3.3. Analysis of the Interaction Mechanism. According to the calculated binding free energy, CID_70126881 holds the strongest binding affinity; on the contrary, compound 4 has the lowest binding affinity. As can be seen from Table 5, the nonpolar interactions ($\Delta G_{nonpolar}$) including van der Waals (E_{vdw}) and nonpolar solvation (ΔG_{np}) terms are the driving force for the binding of the four ligands to AR, and the total polar contributions (ΔG_p) are unfavorable for their binding. In addition, CID_70126881, CID_70127147, and

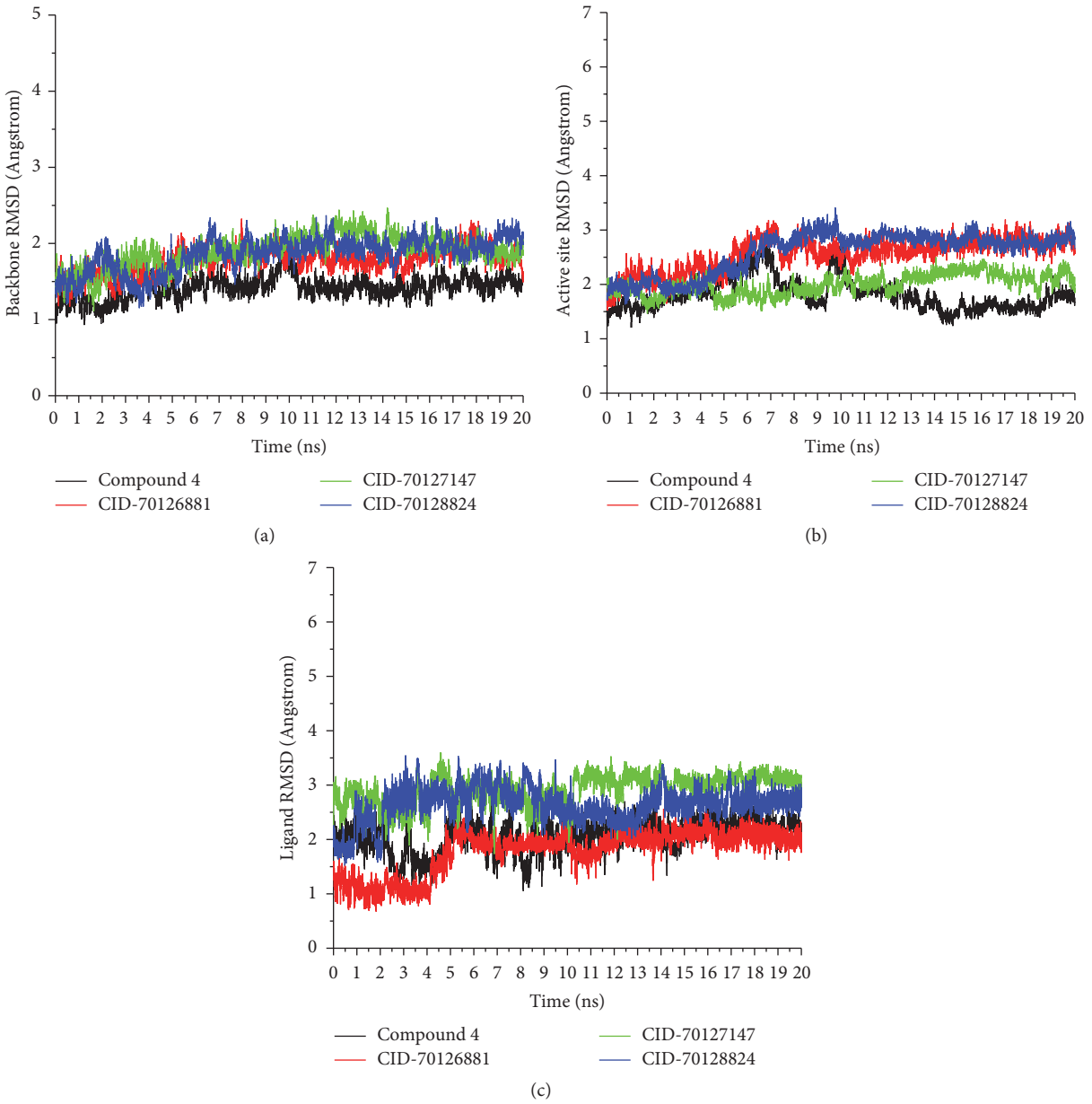


Figure 7: Time series of (a) the RMSDs of backbone atoms of androgen receptor, (b) the RMSD of C_α atoms for the residues around 5 Å of the ligand, and (c) the RMSD of the heavy atoms of ligand.

Table 5: The calculated binding free energies (kcal mol^{-1}) of four systems.

Complex	Contribution								
	ΔE_{ele}	ΔE_{vdw}	ΔG_p	ΔG_{np}	ΔE_{MM}	ΔG_{sol}	ΔE_{bind}	$-T\Delta S$	ΔG_{bind}
Compound 4	−29.58	−58.93	44.95	−6.74	−88.50	38.22	−50.28	27.59	−22.69
CID_70128824	−3.32	−70.56	24.14	−7.25	−73.88	16.89	−56.99	25.13	−31.86
CID_70127147	−16.79	−71.34	33.50	−7.80	−88.13	25.69	−62.44	29.38	−33.06
CID_70126881	−10.01	−69.32	25.14	−8.85	−79.33	16.29	−63.05	21.43	−41.62

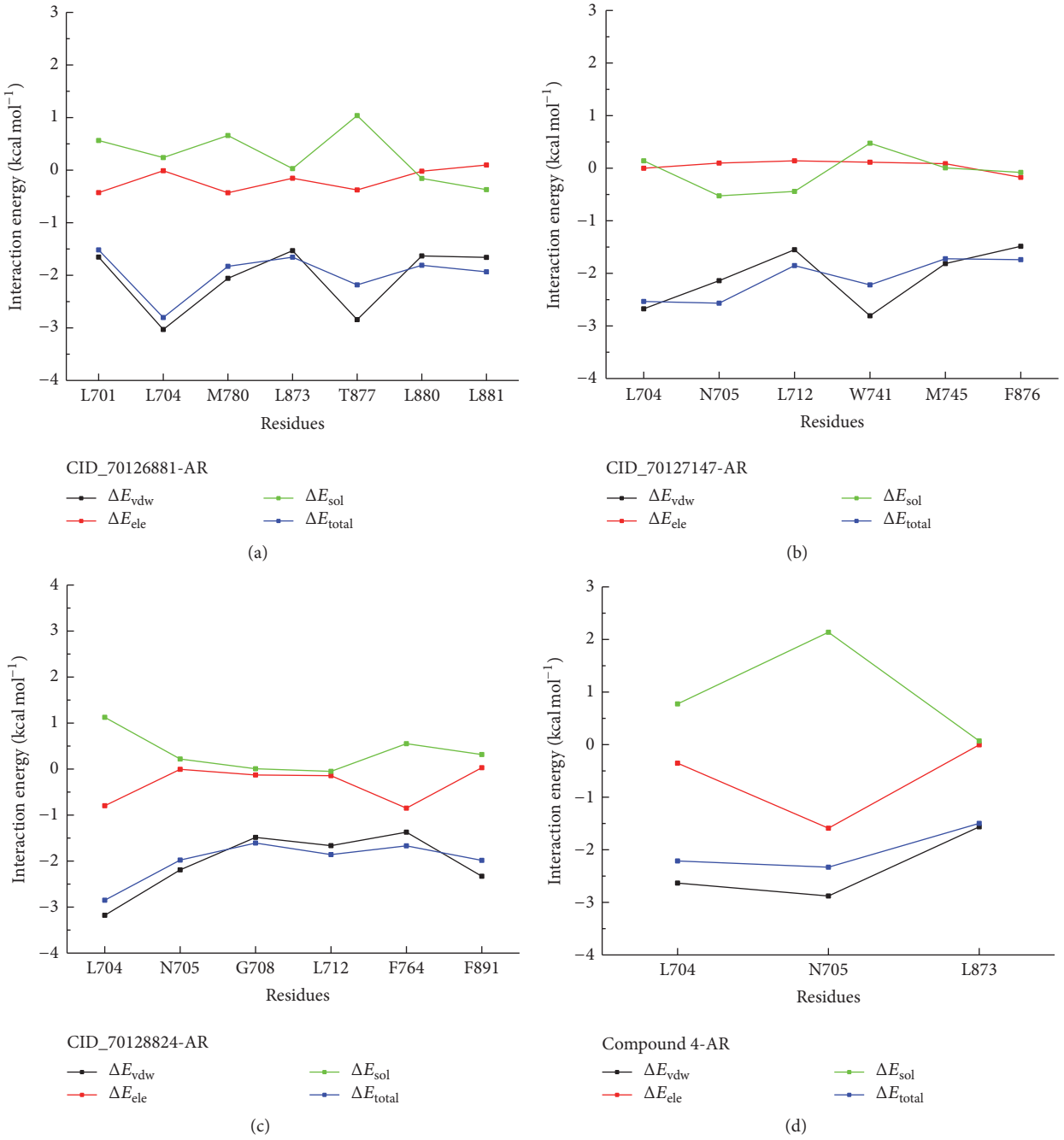


Figure 8: Energy decomposition of key residues in four complexes.

CID_70128824 have almost the same van der Waals interactions toward AR (−69.32 kcal mol^{-1}, −71.34 kcal mol^{-1}, and −70.56 kcal mol^{-1} for CID_70126881-AR, CID_70127147-AR, and CID_70128824-AR), while compound 4 has a low van der Waals value, which may partly explain the reduced binding affinity of compound 4 and prove that the newly discovered chemicals could possess higher antiandrogen activities.

To obtain the detailed interaction between four ligands and AR, the decomposition of binding free energy, which is calculated by MM/GBSA method, was executed to identify key residues during the binding process. The result of energy decomposition contains van der Waals, electrostatic, solvation-free energy, and total energy contribution terms, respectively, for four systems, shown in Figure 8. All the residues with great energy contributions were almost more than 1.5 kcal mol^{-1}. As shown in Figure 8(a), residues L701, L704, M780, L873, T877, L880, and L881 of AR make a significant contribution to the CID_70126881-AR binding, as well as those for the CID_70127147-AR which are L704, N705, L712, W741, M745, and F876 (Figure 8(b)). In Figure 8(c),

residues L704, N705, G708, L712, F764, and F891 of AR make a substantial contribution to the CID_70128824-AR binding. However, only two key residues (L704 and N705) were the major energy contributions to compound 4-AR binding as shown in Figure 8(d). As mentioned previously, the vast majority of key residues of AR were nonpolar; it was reasonable to speculate that these residues can form greater van der Waals interactions with hydrophobic ligand and exhibit more favorable nonpolar interaction contribution to the binding free energy.

The MD simulation, together with the docking results, confirmed that the newly discovered chemicals CID_70126881, CID_70127147, and CID_70128824 share similar binding mode with the reported compound 4, and the in silico antiandrogen activities of them are higher through the calculated binding free energy and decomposition of binding free energy.

4. Conclusions

In this study, the relationships between a series of 7α-substituted dihydrotestosterone derivatives and corresponding antiandrogen activities were explored. A reliable, stable, and robust linear MLR model with four descriptors was built and validated in QSARINS. The predictive ability of the final model, fully evaluated by using two different prediction sets, is excellent enough to be used to virtually screen novel 7α-substituted dihydrotestosterones from PubChem database. After antiandrogenic activity prediction, molecular docking, and molecular dynamic simulations, CID_70126881, CID_70127147, and CID_70128824, as the most potent chemicals with good binding affinities to androgen receptor, were proposed. Of course, bioassay experimental researches are needed to evaluate the virtual screening results. This study provides the theoretical basis and specific chemicals for AR antagonists, which can help the experimental research groups to search for potential antiandrogens.

Competing Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This research is sponsored by the National Natural Science Foundation of China (Grant no. 21205055) and Fundamental Research Funds for the Central Universities (lzujbky-2015-310).

References

[1] http://www.iarc.fr/en/publications/books/wcr/index.php.

[2] http://globocan.iarc.fr/Pages/fact_sheets_cancer.aspx.

[3] W. Dietrich, M. Susani, L. Stifter, and A. Haitel, "The human female prostate—immunohistochemical study with prostate–specific antigen, prostate–specific alkaline phosphatase, and androgen receptor and 3–D remodeling," *Journal of Sexual Medicine*, vol. 8, no. 10, pp. 2816–2821, 2011.

[4] J. Rosen, A. Day, T. K. Jones, E. T. T. Jones, A. M. Nadzan, and R. B. Stein, "Intracellular receptors and signal transducers and activators of transcription superfamilies: novel targets for small-molecule drug discovery," *Journal of Medicinal Chemistry*, vol. 38, no. 25, pp. 4855–4874, 1995.

[5] C. Fix, C. Jordan, P. Cano, and W. H. Walker, "Testosterone activates mitogen-activated protein kinase and the cAMP response element binding protein transcription factor in Sertoli cells," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 30, pp. 10919–10924, 2004.

[6] A. B. Stewart, B. A. Lwaleed, D. A. Douglas, and B. R. Birch, "Current drug therapy for prostate cancer: an overview," *Current Medicinal Chemistry—Anti-Cancer Agents*, vol. 5, pp. 603–612, 2005.

[7] G. T. Kennealey and B. J. A. Furr, "Use of the nonsteroidal antiandrogen casodex in advanced prostatic carcinoma," *Urologic Clinics of North America*, vol. 18, no. 1, pp. 99–110, 1991.

[8] G. Blackledge, G. Kolvenbag, and A. Nash, "Bicalutamide: a new antiandrogen for use in combination with castration for patients with advanced prostate cancer," *Anti-Cancer Drugs*, vol. 7, no. 1, pp. 27–34, 1996.

[9] C. J. Tyrrell, J. E. Altwein, F. Klippel et al., "A multicenter randomized trial comparing the luteinizing hormone-releasing hormone analogue goserelin acetate alone and with flutamide in the treatment of advanced prostate cancer. The International Prostate Cancer Study Group," *Journal of Urology*, vol. 146, no. 5, pp. 1321–1326, 1991.

[10] M. A. Eisenberger, B. A. Blumenstein, E. D. Crawford et al., "Bilateral orchiectomy with or without flutamide for metastatic prostate cancer," *The New England Journal of Medicine*, vol. 339, no. 15, pp. 1036–1042, 1998.

[11] H. I. Scher and C. L. Sawyers, "Biology of progressive, castration-resistant prostate cancer: directed therapies targeting the androgen-receptor signaling axis," *Journal of Clinical Oncology*, vol. 23, no. 32, pp. 8253–8261, 2005.

[12] J. A. Kemppainen and E. M. Wilson, "Agonist and antagonist activities of hydroxyflutamide and casodex relate to androgen receptor stabilization," *Urology*, vol. 48, no. 1, pp. 157–163, 1996.

[13] http://www.oecd.org/chemicalsafety/risk-assessment/37849783.pdf.

[14] P. Gramatica, N. Chirico, E. Papa, S. Cassani, and S. Kovarich, "QSARINS: a new software for the development, analysis, and validation of QSAR MLR models," *Journal of Computational Chemistry*, vol. 34, no. 24, pp. 2121–2132, 2013.

[15] P. Gramatica, S. Cassani, and N. Chirico, "QSARINS-chem: insubria datasets and new QSAR/QSPR models for environmental pollutants in QSARINS," *Journal of Computational Chemistry*, vol. 35, no. 13, pp. 1036–1044, 2014.

[16] http://pubchem.ncbi.nlm.nih.gov.

[17] A. Tropsha, "Best practices for QSAR model development, validation, and exploitation," *Molecular Informatics*, vol. 29, no. 6-7, pp. 476–488, 2010.

[18] P. Gramatica, S. Cassani, P. P. Roy, S. Kovarich, C. W. Yap, and E. Papa, "QSAR modeling is not 'Push a button and find a correlation': a case study of toxicity of (Benzo-)triazoles on Algae," *Molecular Informatics*, vol. 31, no. 11-12, pp. 817–835, 2012.

[19] H. Li, X. Ren, E. Leblanc, K. Frewin, P. S. Rennie, and A. Cherkasov, "Identification of novel androgen receptor antagonists using structure- and ligand-based methods," *Journal of Chemical Information and Modeling*, vol. 53, no. 1, pp. 123–130, 2013.

[20] K. Tachibana, I. Imaoka, H. Yoshino et al., "Discovery of 7α-substituted dihydrotestosterones as androgen receptor pure antagonists and their structure-activity relationships," *Bioorganic and Medicinal Chemistry*, vol. 15, no. 1, pp. 174–185, 2007.

[21] K. Tachibana, I. Imaoka, H. Yoshino et al., "Discovery and structure-activity relationships of new steroidal compounds bearing a carboxy-terminal side chain as androgen receptor pure antagonists," *Bioorganic and Medicinal Chemistry Letters*, vol. 17, no. 20, pp. 5573–5576, 2007.

[22] Hypercube, *HyperChem 7.0*, Hypercube Inc., 2002.

[23] Talete srl, DRAGON for Windows (Software for Molecular Descriptor Calculation), Version 5.5, 2007, http://www.talete.mi.it.

[24] R. Todeschini and V. Consonni, *Handbook of Molecular Descriptors*, Wiley-VCH, Weinheim, Germany, 2009.

[25] G. Schüürmann, R.-U. Ebert, J. Chen, B. Wang, and R. Kühne, "External validation and prediction employing the predictive squared correlation coefficient—test set activity mean vs training set activity mean," *Journal of Chemical Information and Modeling*, vol. 48, no. 11, pp. 2140–2145, 2008.

[26] V. Consonni, D. Ballabio, and R. Todeschini, "Comments on the definition of the Q^2 parameter for QSAR validation," *Journal of Chemical Information and Modeling*, vol. 49, no. 7, pp. 1669–1678, 2009.

[27] L. I.-K. Lin, "A concordance correlation coefficient to evaluate reproducibility," *Biometrics*, vol. 45, no. 1, pp. 255–268, 1989.

[28] N. Chirico and P. Gramatica, "Real external predictivity of QSAR models: how to evaluate It? Comparison of different validation criteria and proposal of using the concordance correlation coefficient," *Journal of Chemical Information and Modeling*, vol. 51, no. 9, pp. 2320–2335, 2011.

[29] N. Chirico and P. Gramatica, "Real external predictivity of QSAR models. Part 2. New intercomparable thresholds for different validation criteria and the need for scatter plot inspection," *Journal of Chemical Information and Modeling*, vol. 52, no. 8, pp. 2044–2058, 2012.

[30] A. Tropsha, P. Gramatica, and V. K. Gombar, "The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models," *QSAR and Combinatorial Science*, vol. 22, no. 1, pp. 69–77, 2003.

[31] P. Gramatica, "Principles of QSAR models validation: internal and external," *QSAR & Combinatorial Science*, vol. 26, pp. 694–701, 2007.

[32] *Discovery Studio Version 2.5*, Accelrys, San Diego, Calif, USA, 2009.

[33] http://www.rcsb.org/.

[34] D. A. Case, V. Babin, J. T. Berryman et al., *AMBER 14*, University of California, San Francisco, Calif, USA, 2014.

[35] H. L. Liu, X. L. An, S. Y. Li, Y. W. Wang, J. Z. Li, and H. X. Liu, "Interaction mechanism exploration of R-bicalutamide/S-1 with WT/W741L AR using molecular dynamics simulations," *Molecular BioSystems*, vol. 11, no. 12, pp. 3347–3354, 2015.

[36] K. Tachibana, I. Imaoka, H. Yoshino et al., "Discovery of 7α-substituted dihydrotestosterones as androgen receptor pure antagonists and their structure–activity relationships," *Bioorganic and Medicinal Chemistry*, vol. 15, no. 1, pp. 174–185, 2007.

[37] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, "Comparison of simple potential functions for simulating liquid water," *The Journal of Chemical Physics*, vol. 79, no. 2, pp. 926–935, 1983.

[38] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, and D. A. Case, "Development and testing of a general Amber force field," *Journal of Computational Chemistry*, vol. 25, no. 9, pp. 1157–1174, 2004.

[39] A. Jakalian, B. L. Bush, D. B. Jack, and C. I. Bayly, "Fast, efficient generation of high-quality atomic charges. AM1-BCC model: I. Method," *Journal of Computational Chemistry*, vol. 21, no. 2, pp. 132–146, 2000.

[40] T. Hou, J. Wang, Y. Li, and W. Wang, "Assessing the performance of the MM/PBSA and MM/GBSA methods. 1. The accuracy of binding free energy calculations based on molecular dynamics simulations," *Journal of Chemical Information and Modeling*, vol. 51, no. 1, pp. 69–82, 2011.

[41] R. H. Bradbury, N. J. Hales, A. A. Rabow et al., "Small-molecule androgen receptor downregulators as an approach to treatment of advanced prostate cancer," *Bioorganic and Medicinal Chemistry Letters*, vol. 21, no. 18, pp. 5442–5445, 2011.

[42] V. Consonni, R. Todeschini, M. Pavan, and P. Gramatica, "Structure/response correlations and similarity/diversity analysis by GETAWAY descriptors. 2. Application of the novel 3D molecular descriptors to QSAR/QSPR studies," *Journal of Chemical Information and Computer Sciences*, vol. 42, no. 3, pp. 693–705, 2002.

Hindawi Publishing Corporation
BioMed Research International
Volume 2017, Article ID 2483264, 15 pages
http://dx.doi.org/10.1155/2017/2483264

Review Article

Comparative Study of Elastic Network Model and Protein Contact Network for Protein Complexes: The Hemoglobin Case

Guang Hu,[1] Luisa Di Paola,[2] Zhongjie Liang,[1] and Alessandro Giuliani[3]

[1]*Center for Systems Biology, School of Electronic and Information Engineering, Soochow University, Suzhou 215006, China*
[2]*Unit of Chemical-Physics Fundamentals in Chemical Engineering, Department of Engineering, Università Campus Bio-Medico di Roma, Rome, Italy*
[3]*Environment and Health Department, Istituto Superiore di Sanità, Viale Regina Elena 299, 00161 Roma, Italy*

Correspondence should be addressed to Guang Hu; huguang@suda.edu.cn

Received 22 June 2016; Revised 17 November 2016; Accepted 20 December 2016; Published 22 January 2017

Academic Editor: Hesham H. Ali

The overall topology and interfacial interactions play key roles in understanding structural and functional principles of protein complexes. Elastic Network Model (ENM) and Protein Contact Network (PCN) are two widely used methods for high throughput investigation of structures and interactions within protein complexes. In this work, the comparative analysis of ENM and PCN relative to hemoglobin (Hb) was taken as case study. We examine four types of structural and dynamical paradigms, namely, conformational change between different states of Hbs, modular analysis, allosteric mechanisms studies, and interface characterization of an Hb. The comparative study shows that ENM has an advantage in studying dynamical properties and protein-protein interfaces, while PCN is better for describing protein structures quantitatively both from local and from global levels. We suggest that the integration of ENM and PCN would give a potential but powerful tool in structural systems biology.

1. Introduction

Proteins rarely act alone: in the great majority of cases they perform a vast array of biological functions by forming functional complexes [1, 2]. The study of protein complexes not only elucidates the molecular mechanism of many diseases [3] but also provides structural information of protein-protein interactions [4]. With the increasing number of structural data, a lot of regularities have been found for protein complexes based on their topological structures [5]. However, the structural and assembly principles underlying protein complexes organization are not yet fully understood, which poses a great challenge in structural systems biology [6]. A well-studied example of protein complex is hemoglobin (Hb) tetramer, which contains two α and two β subunits as a dimer of dimer [7]. Hbs exist in three quaternary conformations: the low-affinity (deoxy, T) state and the high-affinity (oxy, R; carbonmonoxy, $R2$) states. Hbs are never present in cells as monomers. Therefore, Hbs were considered as a sort of 'obliged' allosteric protein complexes and, even thanks to the great amount of both structural and physiological data, attracted a lot of attentions [8–10].

Network theory has become a versatile method to study structures and dynamics of biological systems [11–13]. As a dynamical-based method introduced by Tirion [14], Elastic Network Model (ENM) allows performing normal mode analysis at C_α network level. Two mostly used ENM methods, Gaussian Network Model (GNM) and Anisotropic Network Model (ANM), were further proposed by Bahar and coworkers [15, 16]. ENM is an efficient computational tool to describe the essential vibrational dynamics encoded in the molecular topology [17–20]. It has been proved that the low-frequency modes of ENM are critical of collective motions [21], while the high-frequency modes can identify hot spots for protein-protein interactions [22].

The approach of Protein Contact Network (PCN) was proposed by Kannan and Vishveshwara [23] and now has become a new paradigm in protein ontology [24–28]. In a PCN, nodes correspond to C_α, while edges exist if two amino acid residues (nodes) are close to each other under different

cutoffs [29]. Based on this graphical representation, different topological parameters have been developed to describe protein structures and functions from both the global and the local prospective [30–32].

Both ENM and PCN offer computationally efficient tools to study the structure and function of protein complexes [33, 34], from predicting functionally important residues [35, 36], to characterize protein-protein interactions [37, 38] and allosteric communication paths [39, 40]. Of course, both models have strengths and weaknesses and their comparative study is needed.

In this paper, we have analyzed and compared four applications of ENM and PCN on Hb structures: conformational change characterization, modular analysis, allosteric mechanisms investigation, and interface characterization. Although there are several works reported on the ENM [41–43] and PCN [44, 45] studies of Hb independently, this work revisits Hb as case study and mainly focuses on the methodology comparison of ENM (specifically GNM and ANM) and PCN.

2. Materials and Methods

2.1. Data Sets. Hemoglobins (Hbs) have three states [7]. We select their structures for the ENM and PCN analysis, which are listed as follows: *T*-Hb (PDB code: 2dn2), *R*-Hb (PDB code: 2dn1), and *R*2-Hb (PDB code: 2dn3).

2.2. Gaussian Network Model and Anisotropic Network Model. GNM [15] describes a protein as a network of C_α connected by springs of uniform force constant γ if they are located within a cutoff distance r_c (7 Å in this study). In GNM, the interaction potential for a protein of N residues is [46]

$$V_{\mathrm{GNM}} = -\frac{\gamma}{2}\left[\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\left(R_{ij}-R_{ij}^{0}\right)\cdot\left(R_{ij}-R_{ij}^{0}\right)\Gamma_{ij}\right], \quad (1)$$

where R_{ij} and R_{ij}^0 are the equilibrium and instantaneous distance between residues i and j, and Γ is $N \times N$ *Kirchhoff matrix*, which is written as follows:

$$\Gamma_{ij} = \begin{cases} -1 & i \neq j,\ R_{ij} \leq r_c \\ 0 & i \neq j,\ R_{ij} \geq r_c \\ -\sum_{i,i\neq j}\Gamma_{ij} & i = j. \end{cases} \quad (2)$$

Then, square fluctuations are given by

$$\left\langle \left(\Delta R_i\right)^2 \right\rangle = \left(\frac{3kT}{\gamma}\right)\cdot\left[\Gamma^{-1}\right]_{ii},$$
$$\left\langle \Delta R_i \cdot \Delta R_j \right\rangle = \left(\frac{3kT}{\gamma}\right)\cdot\left[\Gamma^{-1}\right]_{ij}. \quad (3)$$

The normal modes are extracted by eigenvalue decomposition: $\Gamma = U\Lambda U^T$, where U is the orthogonal matrix whose kth column u_k is kth mode eigenvector. Λ is the diagonal matrix of eigenvalues, λ_k. $\langle \Delta R_i \cdot \Delta R_j \rangle$ can be written in terms of the sum of the contribution of each mode as follows:

$$\left\langle \Delta R_i \cdot \Delta R_j \right\rangle = \left(\frac{3kT}{\gamma}\right)\cdot\sum_k\left[\left(U_k\Lambda_k U_k^T\right)^{-1}\right]_{ij}. \quad (4)$$

Thus, the cross-correlation can be calculated by

$$C_{ij} = \frac{\left\langle \Delta R_i \cdot \Delta R_j \right\rangle}{\left[\left\langle \Delta R_i \right\rangle^2 \cdot \left\langle \Delta R_j \right\rangle^2\right]^{1/2}}. \quad (5)$$

The cross-correlation value ranges from −1 to 1: positive values mean that two residues have correlated motions, while the negative values mean that they have anticorrelated motions.

In ANM [16], the interaction potential for a protein of N residues is [46]

$$V_{\mathrm{ANM}} = -\frac{\gamma}{2}\left[\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\left(R_{ij}-R_{ij}^{0}\right)^2\Gamma_{ij}\right]. \quad (6)$$

The motion of the ANM mode of proteins is determined by $3N \times 3N$ *Hessian matrix* H, whose generic element is given as follows:

$$H_{ij} = \begin{bmatrix} \frac{\partial^2 V}{\partial X_i \partial X_j} & \frac{\partial^2 V}{\partial X_i \partial Y_j} & \frac{\partial^2 V}{\partial X_i \partial Z_j} \\ \frac{\partial^2 V}{\partial Y_i \partial X_j} & \frac{\partial^2 V}{\partial Y_i \partial Y_j} & \frac{\partial^2 V}{\partial Y_i \partial Z_j} \\ \frac{\partial^2 V}{\partial Z_i \partial X_j} & \frac{\partial^2 V}{\partial Z_i \partial Y_j} & \frac{\partial^2 V}{\partial Z_i \partial Z_j} \end{bmatrix}, \quad (7)$$

where X_i, Y_i, and Z_i represent the Cartesian components of residues i and V is the potential energy of the system. r_c used here is 13 Å. Accordingly, ANMs provide the information not only about the amplitudes but also about the direction of residue fluctuations.

The similarity between two ANM modes, u_k and v_l, evaluated for proteins with two different conformations can be quantified in terms of inner product of their eigenvectors [39]; that is,

$$O\left(u_k, v_l\right) = u_k \cdot v_l. \quad (8)$$

The degree of overlap between kth ANM modes u_k and the experimentally observed conformation change Δr of Hbs among different states is quantified by $((\Delta r \cdot u_k)/|\Delta r|)$. Therefore, the cumulative overlap CO(m) between Δr and the directions spanned by subsets of m ANM modes is calculated as follows:

$$\mathrm{CO}\left(m\right) = \sqrt{\sum_{k=1}^{m}\left(\Delta r \cdot \frac{u_k}{|\Delta r|}\right)^2}. \quad (9)$$

The Markov model coupled with GNM was used for exploring the signal transductions of perturbations in proteins [47, 48]. The affinity matrix A describes the interactions

between residue pairs connected in GNM; its generic element a_{ij} is defined as follows:

$$a_{ij} = \frac{N_{ij}}{\sqrt{N_i N_j}}, \tag{10}$$

where N_{ij} is the number of atom-atom contacts between residues i and j based on a cutoff distance of 4 Å and N_i is the number of side-chain atoms in residue i. The density of contacts at each node i is given by

$$d_i = \sum_{j=1}^{N} a_{ij}. \tag{11}$$

The Markov transition matrix M, whose element $m_{ij} = d_j^{-1} a_{ij}$, determines the conditional probability of transmitting a signal from residue j to residue i in one time step. Accordingly, the hitting time for the transfer of a signal from residue j to i is given by [47]

$$H(i,j) = \sum_{k=1}^{N} \left\{ \left[\Gamma^{-1}\right]_{kj} - \left[\Gamma^{-1}\right]_{ij} - \left[\Gamma^{-1}\right]_{ki} - \left[\Gamma^{-1}\right]_{ii} \right\} \cdot d_k, \tag{12}$$

where Γ is Kirchhoff matrix obtained by GNM. The average hit time for ith residue $\langle H(i)\rangle$ is the average of $H(i, j)$ over all starting points i. The commute time is defined by the sum of the hitting times in both directions; that is,

$$C(i,j) = H(i,j) + H(j,i). \tag{13}$$

$C(i, j)$ was defined as the corresponding distance, as the weight of the edge between node i and j in the network.

2.3. Protein Contact Networks (PCNs). Protein Contact Networks (PCNs) provide a coarse-grained representation of protein structure [49], based on $C\alpha$ coordinates from PDB files: network nodes are the residues, while links exist between nodes whose Euclidean distance (computed with respect to α-carbons) is within 4 to 8 Å, in order to account only for significant noncovalent intramolecular interactions [24, 50, 51].

After building up the network, it is possible to quantify its features through the adjacency matrix Ad, whose generic element Ad_{ij} is 1 if ith and jth nodes are connected by a link; otherwise it is 0.

The most basic descriptor is the *node degree*, defined for each node as the number of links involving the node itself:

$$k_i = \sum_j \mathrm{Ad}_{ij}. \tag{14}$$

Given a set of vertices V, the *shortest path* $\mathrm{sp}_{u,v}$ between two nodes $u, v \in V$ is the minimum number of edges connecting them (Figure 1). Its role is crucial since it has been demonstrated that the lower the network *average shortest path* (or *characteristic length*, computed as the average value over the whole number of node pairs), the higher the efficiency of signal transmission through the network [52]. In PCNs the average shortest path describes the protein attitude to allosteric regulation.

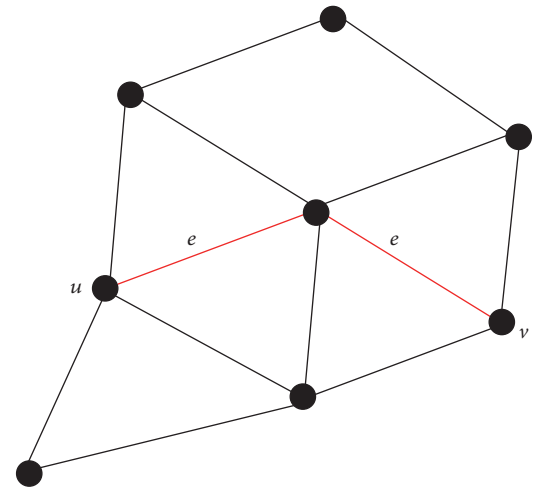


Figure 1: Example of a graph with 8 vertices and 13 edges, while the red line shows a path from vertices u to v.

The *betweenness centrality* of a node describes the number of shortest paths passing by it. Given a set of vertices V, the betweenness centrality of node $s \in V$ is defined as follows:

$$\mathrm{betw}(i) = \sum_{v \in V, v \neq i} \sum_{u \in V, u \neq i} \frac{\sigma_{v,u}(i)}{\sigma_{v,u}}, \tag{15}$$

where $\sigma_{v,u}$ is the total number of the shortest paths connecting two nodes $u, v \in V$, whereas $\sigma_{v,u}(i)$ represents the number of shortest paths connecting the nodes u and v passing on i as well. Therefore, high betweenness centrality nodes take part in many shortest paths, so their removal is likely to be noxious for the whole network connectivity. We computed the betweenness centrality by means of the algorithm described in [53].

Closeness centrality describes the general closeness of a node to all other nodes, in terms of length of shortest paths:

$$\mathrm{close}(i) = \frac{1}{\sum_{u \in V, u \neq i} \mathrm{sp}(u,i)}. \tag{16}$$

Closeness centrality of residues in PCNs has been demonstrated to describe conformational transitions occurring in protein response to environmental stimuli through cooperative processes [54]: residues in the active site of enzymes show both high degree and closeness centrality; however, it does not provide any clue about allosteric regulation in the enzyme-substrate binding.

The Guimerà-Amaral cartography [55] provides a framework to classify nodes according to their topological role in the network. It is based on network clustering into nodes groups (clusters). We applied a spectral clustering procedure, previously demonstrated to catch functional modules in protein structures [56].

The spectral clustering algorithm [57] applies to the Laplacian matrix L defined as the difference between the adjacency matrix Ad and the degree matrix D (a diagonal matrix whose generic element D_{ii} is ith node degree). We applied the eigenvalue decomposition to L: the spectral clustering decomposition refers to the eigenvector v_2 corresponding to the second minor eigenvalue.

The procedure applies iteratively to get the final desired number of clusters (set by defining the number of iterations); nodes are parted according to the sign of corresponding v_2 components. So, for instance, if it is required to part the network into four clusters, the first partition produces two clusters, whose v_2 components have opposite signs and, successively, both clusters undergo the same procedure, applied to single cluster nodes subset.

We represented the clustering partition in two ways: first, we reported on the ribbon representation residues pertaining to different clusters in different colors, to identify at once clusters on the three-dimensional structure representation. Second, we reported the clustering color map, a matrix whose generic element is colored not in blue if residues corresponding to indices pertain to the same cluster and in blue, background, if corresponding residues do not belong to the same cluster. This representation helps understanding the distribution of clusters along sequence.

After clustering partition, it is possible to compute for each node (residue) the *participation coefficient* P, defined as follows:

$$P_i = 1 - \left(\frac{k_{si}}{k_i}\right)^2. \tag{17}$$

k_i is the overall degree of the node, k_{si} is the node degree in its own cluster (number of links the node is involved into with nodes pertaining to its own cluster).

A complementary descriptor is the intramodule connectivity z-score z, defined as follows:

$$z_i = \frac{k_{si} - \overline{k}_{si}}{\mathrm{SD}_{si}}, \tag{18}$$

where $\overline{k}$ and SD are the average value and the standard deviation of the degree k extended to the whole network. The descriptor z catches the attitude of nodes to preferentially connect with nodes in their own clusters; z strongly correlates with node degree, so high z residues are mostly responsible for global protein stability.

The participation coefficient P has been previously demonstrated of a crucial importance in identifying key residues in protein structure with a functional role [43, 56, 58]; residues with P values higher than 0.75 are mostly devoted to the communication between modules (clusters), since they spend more than half of their links with residues pertaining to clusters other than theirs. In other words, signaling pathways between clusters pass by them.

P-z maps show a peculiar shape ("dentist's chair") for PCNs [58]: high P residues show low z values, meaning the role of nodes (communication, high P, and z) are well separated. We previously reported [35] that in protein-ligand binding P shifts from nonnull to null values for residues close to an active site in allosteric proteins.

We computed for each structure P profile and P-z maps. Then, for the two pairs apo-holo forms we report the heat maps of P variation on the ribbon structure, so to highlight regions in the protein structure undergoing changes upon ligand binding.

The analysis was performed by means of a purposed software implemented in Matlab environment v 2014a, including functions from Bioinformatics Toolbox. Heat maps of P variation (comparison between holo and apo forms), Guimerà-Amaral cartography and clusters onto the protein ribbon representation, have been produced by means of a purposed Python script compiled in the embedded Python environment; for further details and application of the method see [37].

Table 1 sums up PCN descriptors, along with their structural and biological relevance.

Table 1: PCN descriptors and their structural and biological relevance.

PCN descriptor	Structural and biological role
Node degree k	Local stability [24]
Betweenness centrality (*betw*)	Signal transmission throughout the structure [26]
Closeness centrality (*close*)	Residues located in the active site of enzymes [26]
Participation Coefficient P	Signal transmission through modules (domains) [27]
Intramodule Connectivity z	Intramodule connectivity and communication [27]

3. Results and Discussion

3.1. ENM Results. GNM and ANM are simple yet effective methods [33]. GNM can only describe the amplitude of residue fluctuations, but ANM can give the direction of the motions. In this section, ANM was used to investigate conformational change between T-Hb and R-Hb, and describe the dynamical properties of protein-protein interfaces. GNM was employed for the modular analysis of Hbs, which was coupled with Markovian stochastic analysis to study the allosteric mechanisms of Hbs. Expecting for the conformational change, we only chose T-Hb to exhibit these investigations. ENM results for other two states of Hbs show similar results, as shown in the supporting information.

3.1.1. Conformational Change. ENM results for the transition of tetrameric Hb between T-state (PDB code: 2dn2) and R-state (PDB code: 2dn1) are shown in Figure 2, while the results of the conformational change between T-state and $R2$-state (PDB code: 2dn3) and R-state and $R2$-state are shown in Supplementary Figure S1 in Supplementary Material available online at https://doi.org/10.1155/2017/2483264. First, the overlap map between the ten ANM slowest modes was calculated to compare the global dynamics of T- and R-Hbs. Along the diagonal in Figure 2(a), only the fifth and sixth modes are

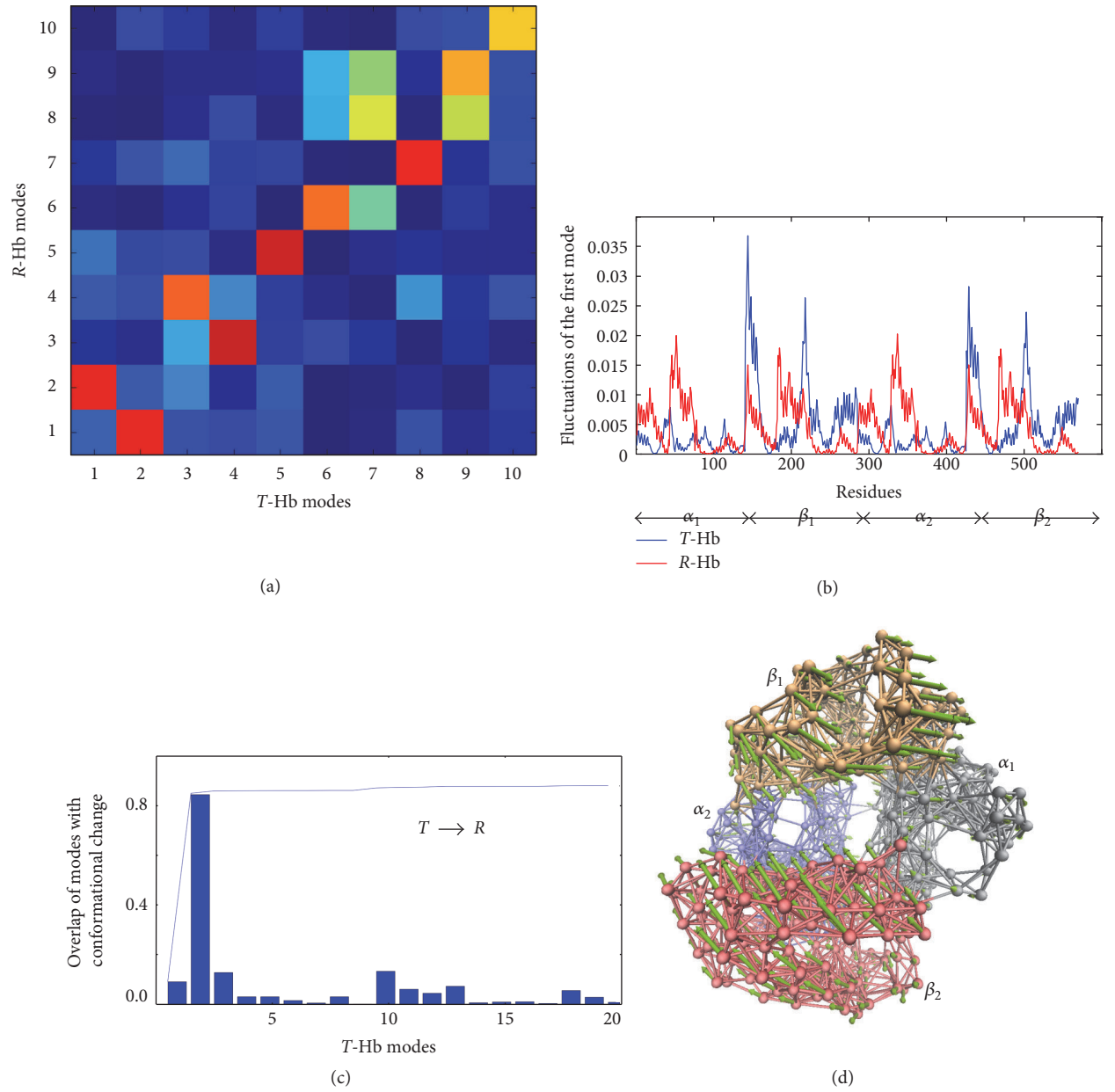


FIGURE 2: ENM results for $T \rightarrow R$ transition of tetrameric Hb. (a) Overlaps between the ten slowest ANM modes of T- and R-Hbs. (b) Distribution of mean-square fluctuations obtained by the first ANM mode of T- and R-Hbs. The residue index of the four chains is 1–140 (α_1), 141–285 (β_1), 286–425 (α_2), and 426–570 (β_2). (c) Overlaps of individual T-Hb ANM modes with the conformational change within $T \rightarrow R$ transition. (d) The motion of the second ANM mode of T-Hb; here the protein is represented as a network.

maintained, with the overlap values of 0.92 and 0.79. For other global modes, there are weaker correlations between two conformations. For example, the reordering of the first two modes was found, which means that the motion of the first mode of T-Hb is similar to the motion of the second mode of R-Hb, while the first mode of R-Hb shifts to the second mode of T-Hb. This result shows that global dynamics greatly changes between the two different states, even for the lowest mode. Then, the difference of two states was further investigated by the distribution of mean-square fluctuations driven by their global ANM modes, as shown in Figure 2(b). For the first mode T-Hb, the two α chains exhibit different dynamical behavior with two β chains, but two dimers of $\alpha_1\beta_1$ and $\alpha_2\beta_2$ show similar global dynamics (the blue line). For the first mode of R-Hb, the mean-square fluctuations profile of α-chain is very similar to that of the β-chain (the red line). Comparing these two structures the fact that α chains are more stable and β chains are less stable in T state than in R state emerges.

Overlaps of each ANM mode with the structural difference between T and R conformations were calculated to detect which individual mode contributes significantly to the

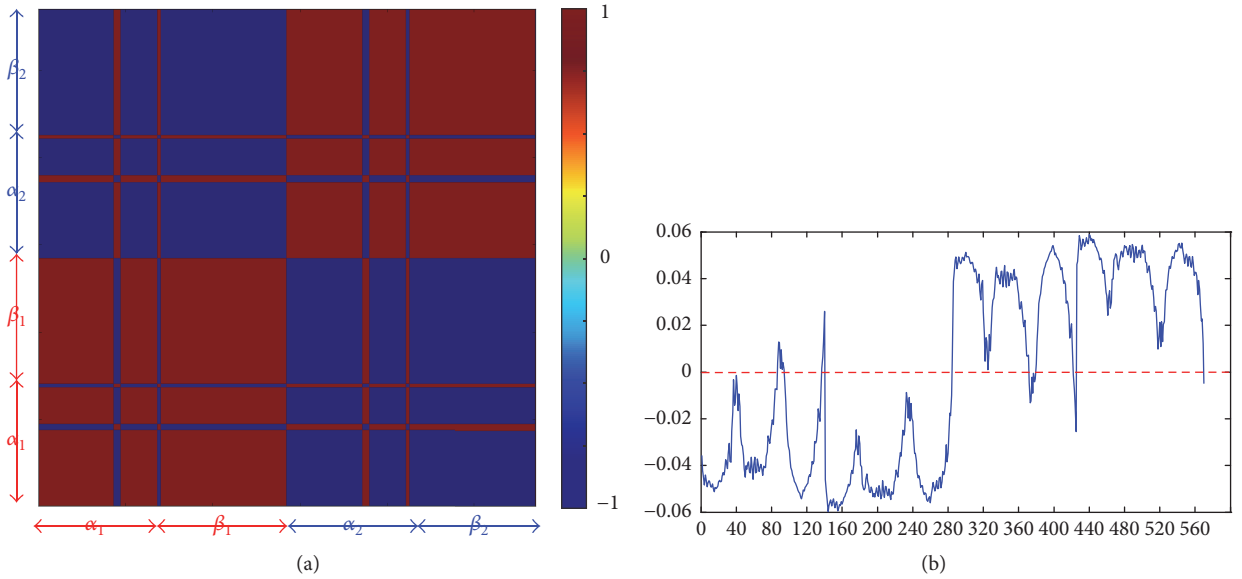


FIGURE 3: Modular analysis of T-Hb based on GNM. (a) The correlation map corresponding to the first mode divides the Hb into two modules. Red regions correspond to collective residue motions and blue-colored regions correspond to uncorrelated motions. (b) The shape of the first mode, which not only shows two modules but also predicts hinge sites.

structural differences between the results from experimental study and are calculated by (9). Figure 2(c) shows that the transformation from T into R is favored by the second mode of T-Hb with the highest overlap (0.84). In this mode, the global motion involves quaternary changes of two dimers, namely, $\alpha_1\beta_1$ dimer, exhibiting a torsional rotation in an opposite direction with $\alpha_2\beta_2$ dimer (Figure 2(d)). Furthermore, this mode is also coordinated by hinge sites at α_1-β_1 and α_2-β_2 interface. Tekpinar and Zheng [42] have previously performed the ENM study of conformation changes from T to R_2 structures, in which they found the first two modes contribute significantly to the conformational change. Our revisiting is in accordance with their results, because mode 2 observed herein seems like the combination motion of their two modes.

3.1.2. Modular Analysis. In their recent work, Li et al. [59] developed a new method based on GNM and ANM for dividing a protein into intrinsic dynamics modular analysis. Here, we adopted a much simpler way, just based on the analysis of the GNM lowest mode. Correlation maps for cross-correlation not only describe collective motion but also reflect the symmetry of proteins [36]. To our aim, the correlation map for the first GNM mode was used for the modular analysis of Hb [60]. In the map, red indicates the highly correlated motions, blue represents the anticorrelated motions, and green is for the uncorrelated motions. As shown in Figure 3(a), the correlation map shows that T-Hb tetramer is divided into two modules, which correspond to $\alpha_1\beta_1$ dimer and $\alpha_2\beta_2$ dimer. Two red blocks indicate that α_1 and α_2 move in the same direction with β_1 and β_2, respectively. Blue blocks indicate that opposite motions are observed between these dimers.

Although the first GNM mode can only generate two modules, it can provide more dynamical information. After diagonalizing the Kirchhoff matrix, the first eigenvector corresponding to the highest eigenvalue can be derived and interpreted to represent the shape of a mode [61]. Figure 3(b) demonstrates that the shape corresponds to GNM mode of the Hb tetramer. It is easy to see that the shape of $\alpha_1\beta_1$ dimer distributes under zero and $\alpha_2\beta_2$ dimer above zero. Thus, the eigenvectors also partition the structure into two modules. In addition, some hinge sites were predicated at near zero positions, which are Thr41, Ala88, and Pro95 in Chain A, His146 in Chain B, Phe98, Leu105, and Ser138 in Chain C, and His2 and His146 in Chain D. Note that these hinge sites always locate at α_1-β_2 and α_2-β_1 interfaces. ENM results for modular analysis of R-Hb and $R2$-Hb are shown in Supplementary Figure S2.

3.1.3. Allosteric Mechanisms. Communication inside protein complexes is implicit in collective motions which are inherent to the network topology [62]. Based on this idea, the signal-processing properties of residues can be investigated by Markovian stochastic analysis coupled with GNM [63, 64]. The commute time, $C(i, j)$, a function of Markov transition probabilities, was used to measure the communication abilities of residue pairs. Figure 4(a) displays the commute time map of T-Hb, while the blue and red regions correspond to short and long hit times. Furthermore, we calculated the average values of each row or column of the commute time map to evaluate the communication abilities of each residue. As shown in Figure 4(b), the minima of the average commute time $\langle C(i)\rangle$ indicate the key residues for T-Hb allostery. The profiles of average commute times for α_1 chain and β_1 chain indicate that Val10, Leu29, Arg31, Thr39, Cys104, Val107,

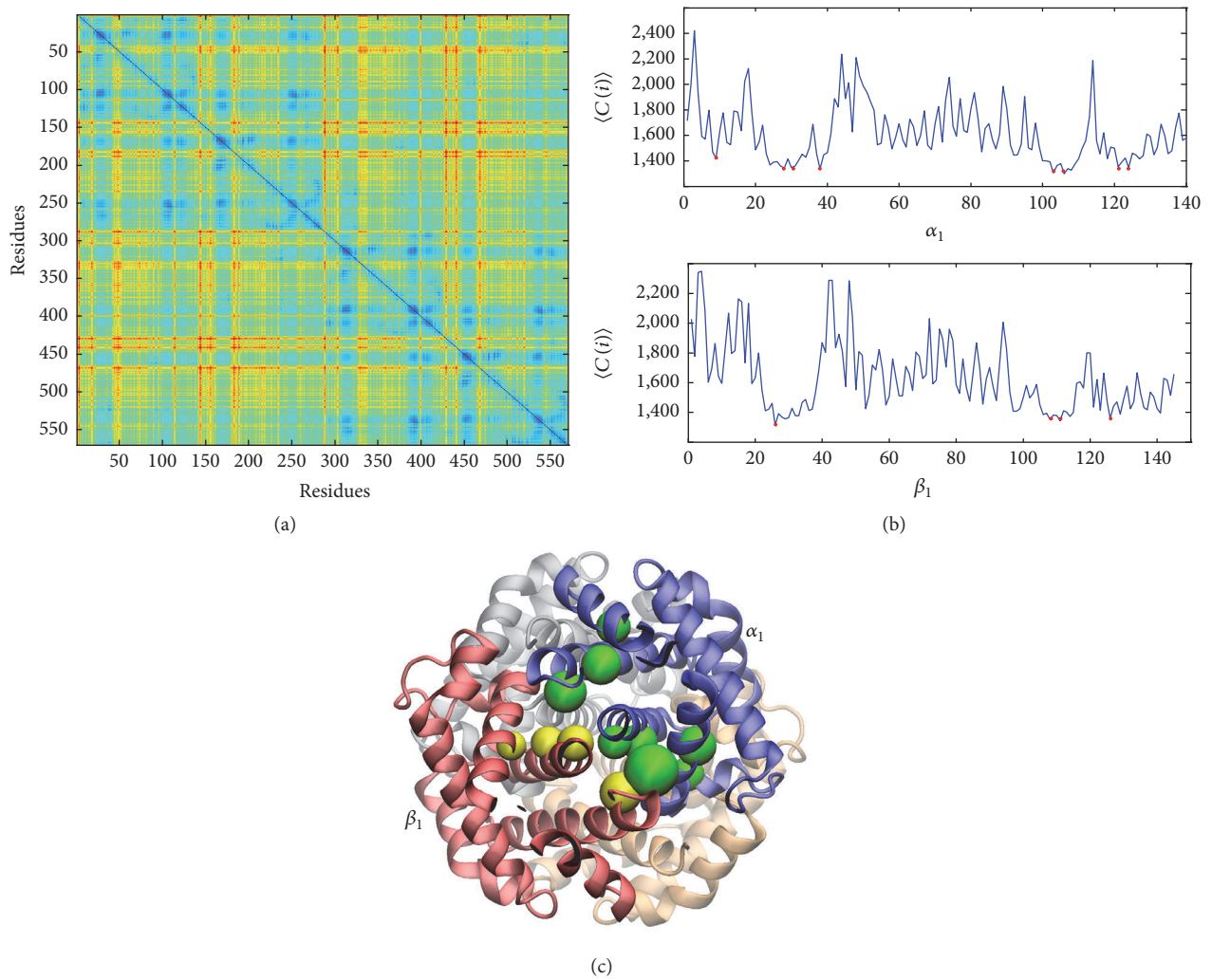


FIGURE 4: Signal propagation of residues for T-Hb. (a) The commute time map of T-Hb. Minimal of average commute time profiles (red circles) for α_1 chain and β_1 chain (b) indicates that most of residues with highest communication abilities (green beads in α_1 chain and yellow beads in β_1 chain) distribute at α_1-β_1 interface (c). Results are presented for 2dn1 (T), and 2dn2 (R) and 2dn3 (R_2) showed similar behavior.

His122, and Leu125 in α_1 chain and Ala27, Val109, Cys112, and Gln127 in β_1 chain are residues with highest communication abilities. It is worth mentioning that the two α chains and two β chains have the same profile shapes. The distributions of these residues in α_1 chain and β_1 chain are also displayed in the three-dimensional representation (Figure 4(c)). It was found that Arg31, Cys104, Val107, and His122 in α_1 chain and Cys112 and Gln127 in β_1 chain are located at α_1-β_1 interface. Likely, the same region was also found at α_2-β_2 interface. ENM results for modular analysis of R-Hb and $R2$-Hb are shown in Supplementary Figure S3.

3.1.4. Interface Characterization. Protein interfaces are the sites where proteins or subunits physically interact. Identification and characterization of protein interfaces are not only important to understand the structures of protein complexes and protein-protein interactions, but also disease phenotypes [65]. Both GNM and ANM have been used to investigate protein-protein interfaces. Kantarci et al. [66] firstly applied GNM to classify interfaces of p53 core domain into the dimerization interface and crystal interface on the base of interfacial dynamics. Zen et al. [67] extended this method to study the interface of 22 representative dimers. More recently, Soner et al. [68] developed a web server to discriminate obligatory and nonobligatory protein complexes. Although GNM is the most used method to study protein-protein interfaces, we have showed here that ANM is also powerful to explore interfacial dynamics of Hbs.

Two kinds of interfaces have been classified in the Hb tetramer: allosteric sites located at α_1-β_1 and α_2-β_2 interfaces, which could be intended as allosteric interfaces. Hinge sites are detected always at α_1-β_2 and α_2-β_1 interfaces, providing structural interfaces. The analyses are in accordance with the results in Tekpinar and Zheng [42], which showed that allosteric interfaces are dynamically variable regions but not necessarily structural interfaces. In this section, square

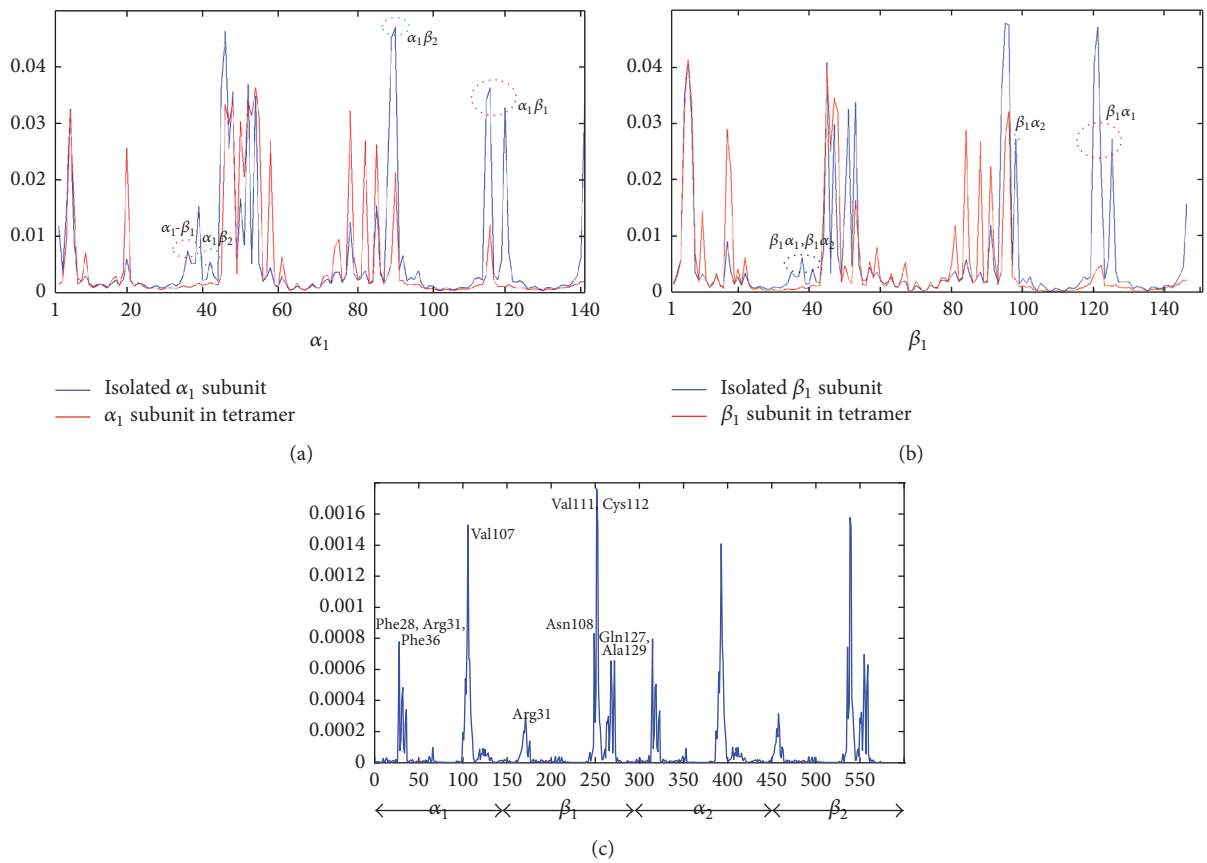


FIGURE 5: Square fluctuations of Hb monomers and tetramers. (a) Square fluctuations of α_1 subunits in isolated and tetrameric states based on the first 20 ANM modes. The differences of mobilities at α_1-β_1 and α_1-β_2 interfaces are indicated by red and green circles. (b) Square fluctuations of β_1 subunits in isolated and tetrameric states based on the first 20 ANM modes. Red, green, and black circles indicate the differences of mobilities at α_1-β_1 and α_2-β_1 interfaces and a common region of these two interfaces. (c) Square fluctuations of T-Hb tetramers based on the highest two modes. Hot spots are predicted by the peaks in the profile, while $\alpha_1\beta_1$ and $\alpha_2\beta_2$ dimer show the same prediction result.

fluctuations of both monomeric and oligomeric proteins based on a large set of slow modes and the highest modes are compared for a deeper analysis of interfaces.

Figures 5(a) and 5(b) show square fluctuations of α_1 and β_1 subunits in isolated and tetrameric states based on the first 20 ANM modes, while α_2 and β_2 subunits show similar behavior. Although α and β subunits are structurally identical, they are different in length and sequence. Accordingly, α_1 and β_1 subunits show different types of fluctuations, which have also been predicted by the previous molecular dynamics (MD) study [69]. The mobility of α_1-β_1 and α_1-β_2 interfacial residues of α_1 subunit is reduced in the tetramer. The same happens for the mobilities of α_1-β_1 and α_2-β_1 interfacial residues of β_1 subunit. Therefore, the flexibilities of residues located at both kinds of interfaces in bound states are lower than in unbound monomers. This kind of dynamical property of interfacial residues has also been detected by the MD simulation [70].

In addition, a similar region (residues 45–57) with high mobility in both isolated states was found, which corresponds to a long loop distributing between two adjacent subunits. The mobility of this region in β_1 subunit was reduced, while no reduction was observed in α_1 subunit. This may suggest that this long loop in β subunits is an allosteric region controlled by interfacial residues.

Among the interface residues, hotspots are defined as residues that have the greatest contribution to the binding energy. The prediction of hotspots is helpful not only to guide drug design but also to understand disease mutations [71]. Based on ENM results, Chennubhotla et al. [72] revealed that hot spots residues show a moderate-high flexibility at global modes. On the other hand, hot spots correlated very well with the residues with high mean-square fluctuations in the highest frequency modes in both GNM [73, 74] and ANM [22]. Ozbek et al. [74] have found that hot spots predictions based on the highest, the second and third highest, and the average three and five highest GNM modes show similar accuracies. Our calculation demonstrates that the square fluctuation based on two highest ANM modes is enough to predict the distribution of of Hb-tetramer hot spots. The result for T-Hb is shown in Figure 5(c). It is surprising that hot spots have been predicted only at α_1-β_1 and α_2-β_2 interfaces: Phe28, Arg31, Phe36, and Val107 in α subunits and Arg31, Asn108, Val111, Cys112, Gln127, and Ala129 in

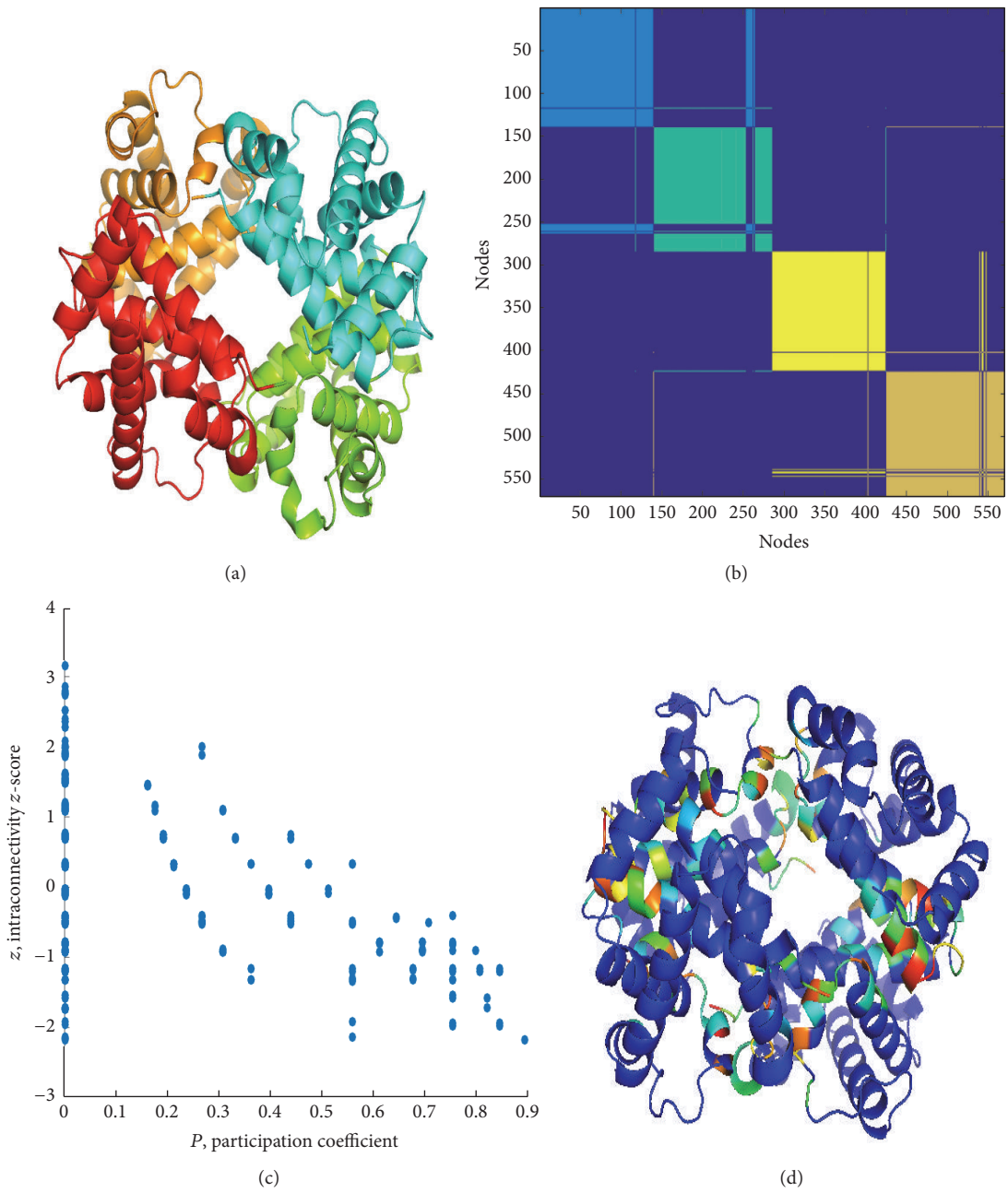


FIGURE 6: PCN results for T-Hb: (a) cluster partition, (b) cluster color map, (c) P-z map, and (d) P over ribbon. Cluster partition satisfactorily matches with chains, and high P residues are mostly located at the chain interfaces.

β subunits. Note that Arg31 and Val107 in α subunits and Arg31, Cys112, and Gln127 in β subunits are overlapped with allosteric sites. It also proved that allosteric interfaces rather than structural interfaces take part in the complex formation. The hotspots predicted for R- and $R2$-Hbs show small differences but still located at the same interfaces (Supplementary Figure S4).

3.2. PCN Results. In this section, results from the application of PCN method and spectral clustering are reported for the three structures under enquiry. Figures 6 and 7 and Supplementary Figure S5 clearly show that cluster partition satisfactorily matches with chains, yet with some divergences (region pertaining to a chain falling in a cluster mainly composed of residues belonging to a second chain, "whiskers" in the clustering color map). In comparison with ENM, PCN results of three states of Hbs exhibit quite high similarity, even between R- and $R2$-states, as emerged mainly from the distribution of P along the ribbon structures (see Figures 8 and 9).

Pz maps show the typical profile for PCNs (and not for other real world networks), with most residues having $P = 0$ (only contacts with residues belonging to their own clusters). Residues with $P > 0$ are mostly interesting for protein functionality, since they account for signaling transmission

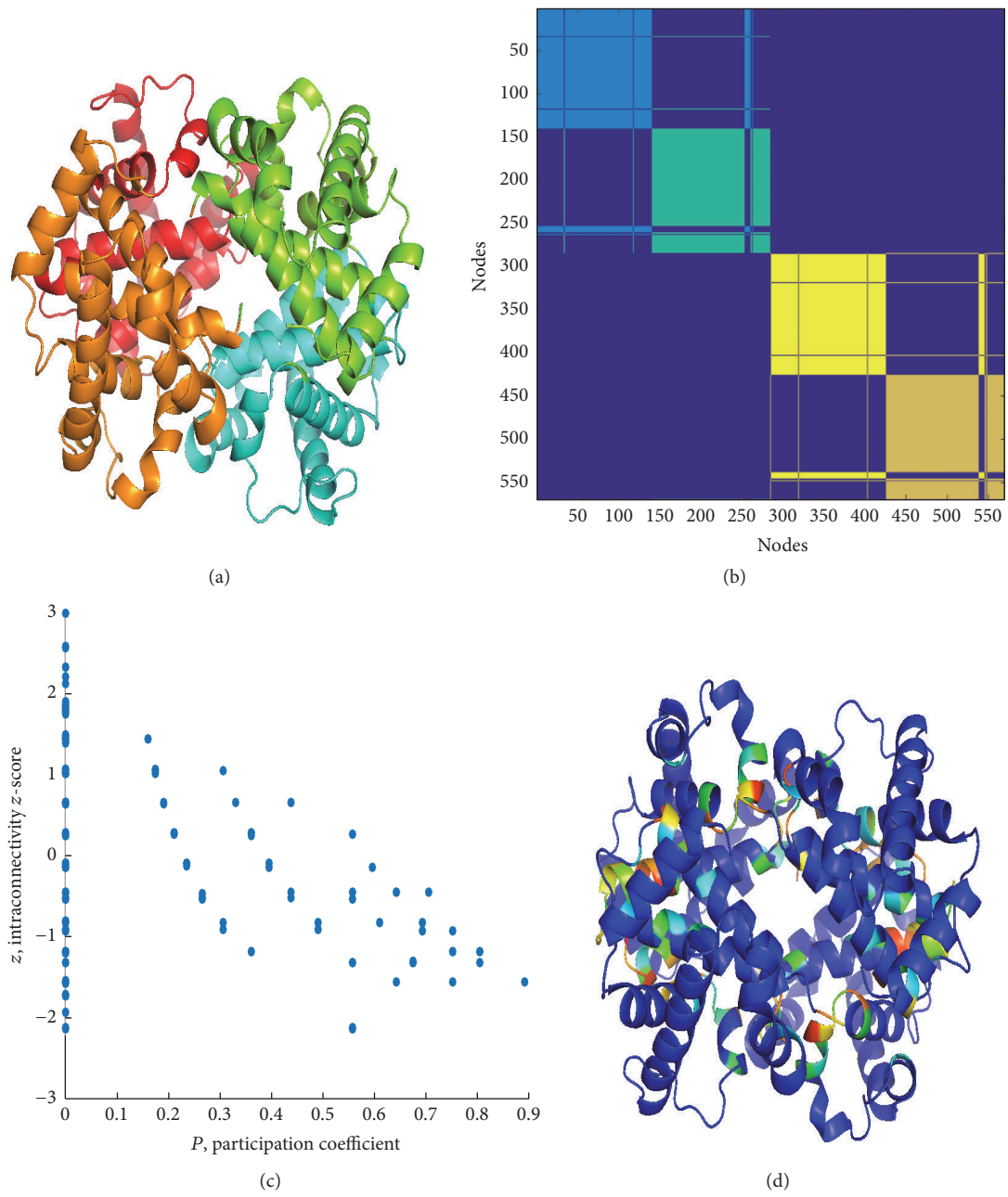


FIGURE 7: PCN results for *R*-Hb: (a) cluster partition, (b) cluster color map, (c) *P*-*z* map, and (d) *P* over ribbon. Again, clusters catch almost perfectly the chains and high *P* residues are located at the interfaces between chains. Similar results are obtained for *R*2-Hb; see Supplementary Figure S5.

through the protein structure (global property of protein structure).

High *P* residues are spotted in the structure and mostly (but not necessarily) placed in the interchain region. In previous works [35, 37, 51, 75], we demonstrated that the participation coefficient *P* addresses the functional role of residues in protein binding and, in general, identifies residues with a key role in protein structural and functional features.

Pz maps instruct a cartography, addressing a specific role to residues, as reported in Table 2. Hubs are nodes with $z > 2.5$, while *P* values address the role of nodes to connect different clusters. The Guimerà-Amaral cartography of the three Hbs is reported in Figures 8 and 9 and Supplementary Figure S6, as original form, on *Pz* maps, and projection on ribbon structures.

Noticeably, in PCNs *R*6 and *R*7 nodes are absent and only few *R*5 nodes are present, all at $P = 0$. In other words, high *z* nodes correspond to residues in charge for protein stability, while nonhub connector nodes are responsible for interdomain (intercluster) communication. Lys60 in α_1, Glu26, His63, Lys66 in β_1, and Leu28, Lys65, Leu68 in β_2 in *T*-Hb, whereas Gly24, Lys61, and Leu141 in two β chains in

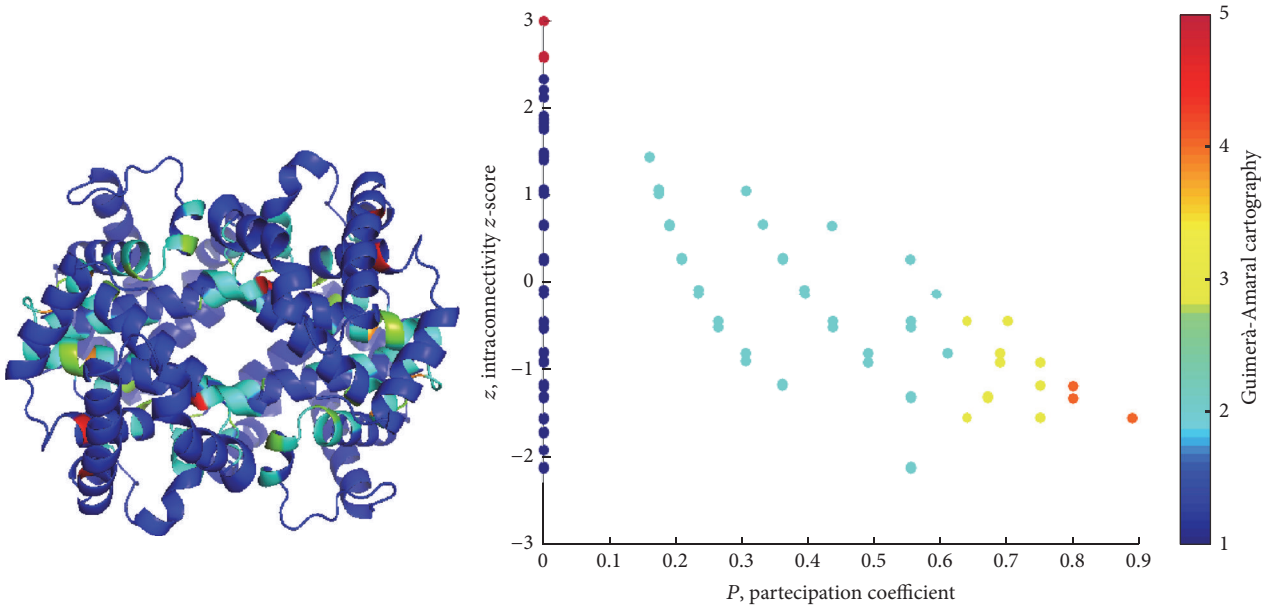


FIGURE 8: Guimerà-Amaral Cartography for *R*-Hb: only very few nodes are classified as hubs (*R*5) and located close to the active site. Non-hub kinless nodes (*R*4) are located in turn on the interface between chains and play a key role in the concerted motion underlying the allosteric regulation of hemoglobin.

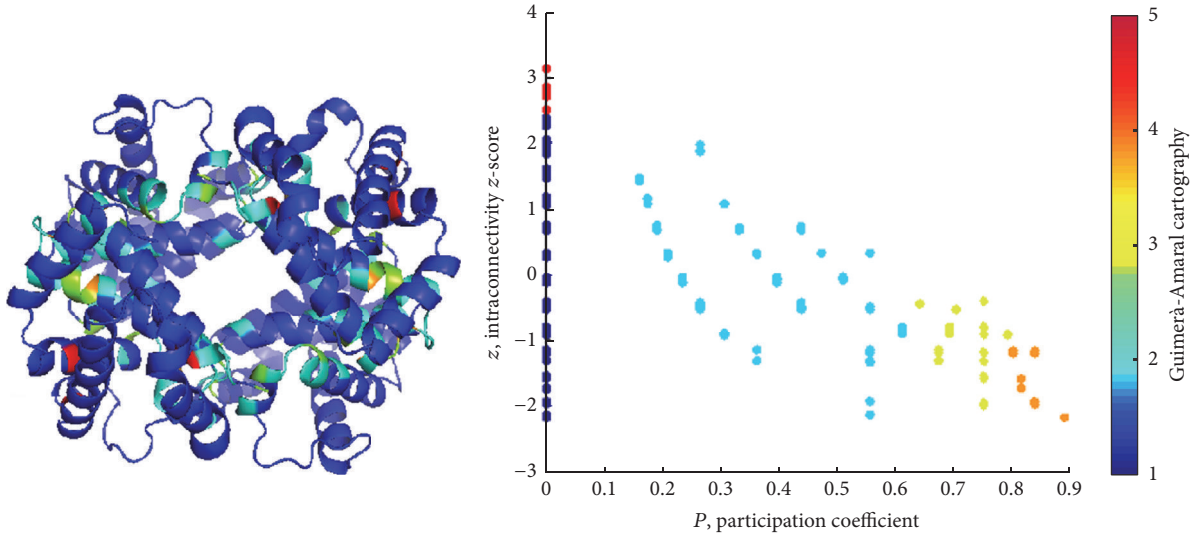


FIGURE 9: Guimerà-Amaral Cartography for *T*-Hb: very few nodes are classified as hubs (*R*5), but more than for R/O_2 complex, again close to the active site. Non-hub kinless nodes (*R*4) lie on the interface between chains. Similar results hold for *R*2-Hb; see Supplementary Figure S6.

TABLE 2: Guimerà-Amaral cartography.

	Regions	z	P
Module nonhubs	*R*1: ultraperipheral node	$z < 2.5$	$P < 0.05$
	*R*2: peripheral node	$z < 2.5$	$0.05 < P < 0.625$
	*R*3: nonhub connectors	$z < 2.5$	$0.625 < P < 0.8$
	*R*4: nonhub kinless nodes	$z < 2.5$	$P > 0.8$
Module hubs	*R*5: provincial hubs	$z > 2.5$	$P < 0.3$
	*R*6: connector hubs	$z > 2.5$	$0.3 < P < 0.75$
	*R*7: kinless hubs	$z > 2.5$	$P > 0.75$

Table 3: Pearson correlation coefficients of network descriptors with mean fluctuations (MF) and average commute times (CT).

	Closeness/MF	Betweenness/MF	P/MF	P/CT
2DN2	−0.3741	−0.1828	−0.0861	−0.2677
2DN1	−0.4320	−0.1226	−0.1878	−0.2904
2DN3	−0.7331	−0.1419	−0.3649	−0.2740

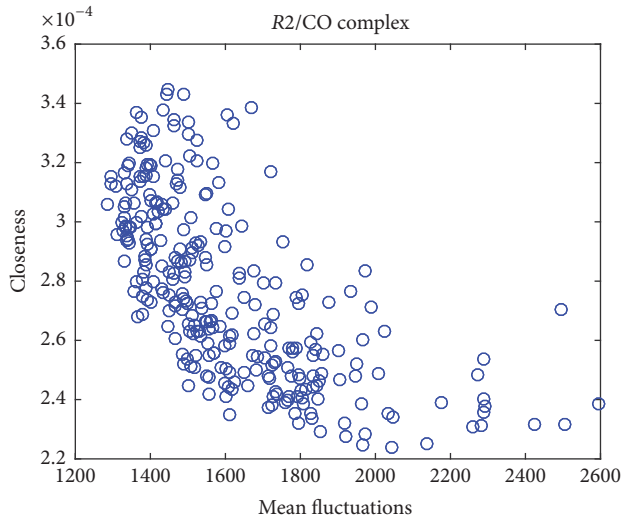


Figure 10: Closeness versus mean fluctuations: in this complex, there is an hyperbolic variation of closeness along with mean fluctuations: the closeness is thus a local rigidity descriptor.

R-Hb belong to *R*5 nodes. It is easy to note that *R*5 nodes distribute at β chains within the protein interior.

As previously stated [58], *R*4 nodes (nonhub kinless nodes) are crucial for the allosteric signal propagation: their kinless nature poses in the gray zone where residues acting at a global level play, so their role in the protein functionality is central. It was found that Leu91 and Arg92 in α_1, His2, His116, and Ala129 in β_1, His89 in α_2, and Thr38, His116, and Phe118 in *T*-Hb and Trp37, Cys112, Phe122, and Pro124 in *R*-Hb belong to *R*5 nodes. Except His2, *R*4 residues were found at all four interfacial regions.

3.3. Comparison between ENM and PCN Results. We finally explicitly superpose the ENMs and PCNs results, in order to better specify key residues and features in allosteric regulation of Hb. Average commute times predict allosteric sites at both protein interior and two interfaces (α_1-β_1 and α_2-β_2 interfaces). In PCN, *R*4 nodes include allosteric sites at interfaces and *R*5 nodes include allosteric sites at protein interior. Combined with modular analysis and hot spots prediction, the use of ENM has advantage to classify protein-protein interfaces.

On the other hand, PCNs analysis relies upon a set of network descriptors to approach the study of protein structures quantitatively. Table 3 reports the Pearson correlation coefficients between mean fluctuations and network descriptors, closeness centrality, betweenness centrality, and participation coefficient.

Betweenness centrality poorly correlates (negatively) with mean fluctuations, while closeness anticorrelates more strongly with mean fluctuations, especially in the more rigid structure (*R*2/complex, Figure 10). The hyperbolic shape of the distribution confirms closeness is a general stiffness descriptor for protein structure. This property may indicate that closeness in PCN could provide an additional evidence to detect hinge sites. There is a relatively poor one-to-one correspondence of functional sites obtained between ENM and PCN, and thus the combination of these two approaches would improve the prediction.

4. Conclusion and Perspectives

ENM and PCN are light yet effective computational methods which simply require the three-dimensional coordinates of atoms in protein structures. In this work, the combination of the ENM and PCN methodologies has provided a plenty of information regarding the dynamic behavior of Hbs. It is noteworthy that the two classes of methods are able to catch the same features without a common, interexchange ground. In comparison with PCN, ENM can find the dominate motion for the conformational change of proteins and detect the dynamics of protein-protein interfaces observed by MD. Except for the topological parameters used in our work, there are more local and global network parameters that can be calculated in PCN to describe protein structures quantitatively. For example, residue centrality as a local network parameter was proposed to identify allosteric sites [76], and coefficient of assortativity as a global network parameter is related to the rates of protein folding [77]. In addition, we have found some correlations between ENM and PCN results. In previous studies [78], the average path lengths are highly correlated with residue fluctuations. Here, we show an additional positive correlation between residue fluctuations predicted by ENM and closeness centrality calculated by PCN. Although the general relationship between dynamical properties and more network parameters is needed to be established, we can conclude that ANM and GNM have advantages in studying dynamical properties and protein-protein interfaces, while PCN is better for describing structures quantitatively from both local and global levels.

In future, the combined description by means of these methods will largely contribute to understanding the dynamic behavior of complexes without heavy computational approaches, such as molecular dynamics (MD). Evidently, MD will anyway provide a very complete and fine description of dynamics, but the combination of lighter methods, such as ENM and PCN, will, for instance, guide MD simulations with well-grounded preliminary results, as preliminary approached in our previous works [79]. On the

other hand, the two methods may help understanding the relationship between local fluctuation of residues and protein stability and functionality, being a primer for identifying key residues, responsible for lethal mutations. For example, the first attempt to combine ENM and PCN has been reported to investigate allosteric communication pathways [80]. In our work, we only indicate that ENM and PCN can be applied to four types of structural and dynamical paradigms. More detailed analysis for each case is needed. Although the integration of these two methods is just at the beginning, it would give a potential but powerful tool in structural systems biology.

Competing Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Authors' Contributions

Guang Hu and Luisa Di Paola are equal contributors.

Acknowledgments

This work was supported by the National Nature Science Foundation of China (21203131 and 81302700), the Natural Science Foundation of the Jiangsu Higher Education Institutions (12KJB180014 and 13KJB520022), and the China Postdoctoral Science Foundation (2013M531406 and 2016M590495). Guang Hu also thanks Professor Chakra Chennubhotla for providing the program of calculating hitting time and commute time.

References

[1] B. J. G. E. Pieters, M. B. Van Eldijk, R. J. M. Nolte, and J. Mecinović, "Natural supramolecular protein assemblies," *Chemical Society Reviews*, vol. 45, no. 1, pp. 24–39, 2016.

[2] T. Perica, J. A. Marsh, F. L. Sousa et al., "The emergence of protein complexes: quaternary structure, dynamics and allostery," *Biochemical Society Transactions*, vol. 40, no. 3, pp. 475–491, 2012.

[3] T. L. Nero, C. J. Morton, J. K. Holien, J. Wielens, and M. W. Parker, "Oncogenic protein interfaces: small molecules, big challenges," *Nature Reviews Cancer*, vol. 14, no. 4, pp. 248–262, 2014.

[4] O. Keskin, N. Tuncbag, and A. Gursoy, "Predicting protein–protein interactions from the molecular to the proteome level," *Chemical Reviews*, vol. 116, no. 8, pp. 4884–4909, 2016.

[5] S. E. Ahnert, J. A. Marsh, H. Hernández, C. V. Robinson, and S. A. Teichmann, "Principles of assembly reveal a periodic table of protein complexes," *Science*, vol. 350, no. 6266, article no. 1331, 2015.

[6] P. Aloy and R. B. Russell, "Structural systems biology: modelling protein interactions," *Nature Reviews Molecular Cell Biology*, vol. 7, no. 3, pp. 188–197, 2006.

[7] S.-Y. Park, T. Yokoyama, N. Shibayama, Y. Shiro, and J. R. H. Tame, "1.25 Å resolution crystal structures of human haemoglobin in the oxy, deoxy and carbonmonoxy forms," *Journal of Molecular Biology*, vol. 360, no. 3, pp. 690–701, 2006.

[8] M. D. Vesper and B. L. de Groot, "Collective dynamics underlying allosteric transitions in hemoglobin," *PLoS Computational Biology*, vol. 9, no. 9, 2013.

[9] Y. Yuan, M. F. Tam, V. Simplaceanu, and C. Ho, "New look at hemoglobin allostery," *Chemical Reviews*, vol. 115, no. 4, pp. 1702–1724, 2015.

[10] J. S. Hub, M. B. Kubitzki, and B. L. de Groot, "Spontaneous quaternary and tertiary T-R transitions of human hemoglobin in molecular dynamics simulation," *PLoS Computational Biology*, vol. 6, no. 5, pp. 1–11, 2010.

[11] A.-L. Barabási and Z. N. Oltvai, "Network biology: understanding the cell's functional organization," *Nature Reviews Genetics*, vol. 5, pp. 101–113, 2004.

[12] P. Csermely, "Creative elements: network-based predictions of active centres in proteins and cellular and social networks," *Trends in Biochemical Sciences*, vol. 33, no. 12, pp. 569–576, 2008.

[13] C. Böde, I. A. Kovács, M. S. Szalay, R. Palotai, T. Korcsmáros, and P. Csermely, "Network analysis of protein dynamics," *FEBS Letters*, vol. 581, no. 15, pp. 2776–2782, 2007.

[14] M. M. Tirion, "Large amplitude elastic motions in proteins from a single-parameter, atomic analysis," *Physical Review Letters*, vol. 77, no. 9, pp. 1905–1908, 1996.

[15] I. Bahar, A. R. Atilgan, and B. Erman, "Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential," *Folding and Design*, vol. 2, no. 3, pp. 173–181, 1997.

[16] A. R. Atilgan, S. R. Durell, R. L. Jernigan, M. C. Demirel, O. Keskin, and I. Bahar, "Anisotropy of fluctuation dynamics of proteins with an elastic network model," *Biophysical Journal*, vol. 80, no. 1, pp. 505–515, 2001.

[17] G. Hu, S. Michielssens, S. L. C. Moors, and A. Ceulemans, "The harmonic analysis of cylindrically symmetric proteins: a comparison of Dronpa and a DNA sliding clamp," *Journal of Molecular Graphics and Modelling*, vol. 34, pp. 28–37, 2012.

[18] S. P. Tiwari and N. Reuter, "Similarity in shape dictates signature intrinsic dynamics despite no functional conservation in TIM barrel enzymes," *PLoS Computational Biology*, vol. 12, no. 3, Article ID e1004834, 2016.

[19] E. Fuglebakk, S. P. Tiwari, and N. Reuter, "Comparing the intrinsic dynamics of multiple protein structures using elastic network models," *Biochimica et Biophysica Acta (BBA)—General Subjects*, vol. 1850, no. 5, pp. 911–922, 2015.

[20] X.-Y. Li, J.-C. Zhang, Y.-Y. Zhu, and J.-G. Su, "Domain motions and functionally-key residues of l-alanine dehydrogenase revealed by an elastic network model," *International Journal of Molecular Sciences*, vol. 16, no. 12, pp. 29383–29397, 2015.

[21] S. Mahajan and Y.-H. Sanejouand, "On the relationship between low-frequency normal modes and the large-scale conformational changes of proteins," *Archives of Biochemistry and Biophysics*, vol. 567, pp. 59–65, 2015.

[22] A. Uyar, O. Kurkcuoglu, L. Nilsson, and P. Doruker, "The elastic network model reveals a consistent picture on intrinsic functional dynamics of type II restriction endonucleases," *Physical Biology*, vol. 8, no. 5, Article ID 056001, 2011.

[23] N. Kannan and S. Vishveshwara, "Identification of side-chain clusters in protein structures by a graph spectral method," *Journal of Molecular Biology*, vol. 292, no. 2, pp. 441–464, 1999.

[24] L. Di Paola, M. De Ruvo, P. Paci, D. Santoni, and A. Giuliani, "Protein contact networks: an emerging paradigm in chemistry," *Chemical Reviews*, vol. 113, no. 3, pp. 1598–1613, 2013.

[25] G. Hu, J. Zhou, W. Yan, J. Chen, and B. Shen, "The topology and dynamics of protein complexes: insights from intra-molecular network theory," *Current Protein and Peptide Science*, vol. 14, no. 2, pp. 121–132, 2013.

[26] W. Yan, J. Zhou, M. Sun, J. Chen, G. Hu, and B. Shen, "The construction of an amino acid network for understanding protein structure and function," *Amino Acids*, vol. 46, no. 6, pp. 1419–1439, 2014.

[27] L. Di Paola and A. Giuliani, "Protein contact network topology: a natural language for allostery," *Current Opinion in Structural Biology*, vol. 31, pp. 43–48, 2015.

[28] L. H. Greene, "Protein structure networks," *Briefings in Functional Genomics*, vol. 11, no. 6, pp. 469–478, 2012.

[29] S. Cheng, H.-L. Fu, and D.-X. Cui, "Characteristics analyses and comparisons of the protein structure networks constructed by different methods," *Interdisciplinary Sciences: Computational Life Sciences*, vol. 8, no. 1, pp. 65–74, 2016.

[30] N. T. Doncheva, K. Klein, F. S. Domingues, and M. Albrecht, "Analyzing and visualizing residue networks of protein structures," *Trends in Biochemical Sciences*, vol. 36, no. 4, pp. 179–182, 2011.

[31] R. K. Grewal and S. Roy, "Modeling proteins as residue interaction networks," *Protein and Peptide Letters*, vol. 22, no. 10, pp. 923–933, 2015.

[32] L. Vuillon and C. Lesieur, "From local to global changes in proteins: a network view," *Current Opinion in Structural Biology*, vol. 31, pp. 1–8, 2015.

[33] I. Bahar, T. R. Lezon, L.-W. Yang, and E. Eyal, "Global dynamics of proteins: bridging between structure and function," *Annual Review of Biophysics*, vol. 39, no. 1, pp. 23–42, 2010.

[34] X. Zhang, T. Perica, and S. A. Teichmann, "Evolution of protein structures and interactions from the perspective of residue contact networks," *Current Opinion in Structural Biology*, vol. 23, no. 6, pp. 954–963, 2013.

[35] M. De Ruvo, A. Giuliani, P. Paci, D. Santoni, and L. Di Paola, "Shedding light on protein-ligand binding by graph theory: the topological nature of allostery," *Biophysical Chemistry*, vol. 165-166, pp. 21–29, 2012.

[36] G. Hu, S. Michielssens, S. L. C. Moors, and A. Ceulemans, "Normal mode analysis of Trp RNA binding attenuation protein: structure and collective motions," *Journal of Chemical Information and Modeling*, vol. 51, no. 9, pp. 2361–2371, 2011.

[37] L. Di Paola, C. B. M. Platania, G. Oliva, R. Setola, F. Pascucci, and A. Giuliani, "Characterization of protein–protein interfaces through a protein contact network approach," *Frontiers in Bioengineering and Biotechnology*, vol. 3, article 170, 2015.

[38] W. I. Karain and N. I. Qaraeen, "Weighted protein residue networks based on joint recurrences between residues," *BMC Bioinformatics*, vol. 16, article 173, 2015.

[39] E. Marcos, R. Crehuet, and I. Bahar, "Changes in dynamics upon oligomerization regulate substrate binding and allostery in amino acid kinase family members," *PLoS Computational Biology*, vol. 7, no. 9, Article ID e1002201, 2011.

[40] O. Schueler-Furman and S. J. Wodak, "Computational approaches to investigating allostery," *Current Opinion in Structural Biology*, vol. 41, pp. 159–171, 2016.

[41] C. Xu, D. Tobi, and I. Bahar, "Allosteric changes in protein structure computed by a simple mechanical model: hemoglobin T ↔ R2 transition," *Journal of Molecular Biology*, vol. 333, no. 1, pp. 153–168, 2003.

[42] M. Tekpinar and W. Zheng, "Coarse-grained and all-atom modeling of structural states and transitions in hemoglobin," *Proteins: Structure, Function and Bioinformatics*, vol. 81, no. 2, pp. 240–252, 2013.

[43] M. Davis and D. Tobi, "Multiple Gaussian network modes alignment reveals dynamically variable regions: the hemoglobin case," *Proteins: Structure, Function and Bioinformatics*, vol. 82, no. 9, pp. 2097–2105, 2014.

[44] A. Giuliani, L. Di Paola, and R. Setola, "Proteins as networks: a mesoscopic approach using haemoglobin molecule as case study," *Current Proteomics*, vol. 6, no. 4, pp. 235–245, 2009.

[45] L. B. Caruso, A. Giuliani, and A. Colosimo, "Allosteric transitions of proteins studied by topological networks: a preliminary investigation on human haemoglobin," *Biophysics and Bioengineering Letters*, vol. 5, no. 1, pp. 1–10, 2012.

[46] L. Meireles, M. Gur, A. Bakan, and I. Bahar, "Pre-existing soft modes of motion uniquely defined by native contact topology facilitate ligand binding to proteins," *Protein Science*, vol. 20, no. 10, pp. 1645–1658, 2011.

[47] C. Chennubhotla and I. Bahar, "Markov propagation of allosteric effects in biomolecular systems: application to GroEL-GroES," *Molecular Systems Biology*, vol. 2, article no. 36, 2006.

[48] C. Chennubhotla, Z. Yang, and I. Bahar, "Coupling between global dynamics and signal transduction pathways: a mechanism of allostery for chaperonin GroEL," *Molecular BioSystems*, vol. 4, no. 4, pp. 287–292, 2008.

[49] K. V. Brinda and S. Vishveshwara, "A network representation of protein structures: implications for protein stability," *Biophysical Journal*, vol. 89, no. 6, pp. 4159–4170, 2005.

[50] A. Giuliani and L. Di Paola, "Protein as networks: will contact maps hold the promise to represent the 'structural-formula' of protein molecules?" *Current Protein and Peptide Science*, vol. 17, no. 1, article 3, 2016.

[51] P. Paci, L. Di Paola, D. Santoni, M. de Ruvo, and A. Giuliani, "Structural and functional analysis of hemoglobin and serum albumin through protein long-range interaction networks," *Current Proteomics*, vol. 9, no. 3, pp. 160–166, 2012.

[52] D. Santoni, P. Paci, L. Di Paola, and A. Giuliani, "Are proteins just coiled cords? Local and global analysis of contact maps reveals the backbone-dependent nature of proteins," *Current Protein and Peptide Science*, vol. 17, no. 1, pp. 26–29, 2016.

[53] J. D. Ullman and M. Yannakakis, "High-probability parallel transitive-closure algorithms," *SIAM Journal on Computing*, vol. 20, no. 1, pp. 100–125, 1991.

[54] G. Amitai, A. Shemesh, E. Sitbon et al., "Network analysis of protein structures identifies functional residues," *Journal of Molecular Biology*, vol. 344, no. 4, pp. 1135–1146, 2004.

[55] R. Guimerà, M. Sales-Pardo, and L. A. N. Amaral, "Classes of complex networks defined by role-to-role connectivity profiles," *Nature Physics*, vol. 3, no. 1, pp. 63–69, 2007.

[56] S. Tasdighian, L. Di Paola, M. De Ruvo et al., "Modules identification in protein structures: the topological and geometrical solutions," *Journal of Chemical Information and Modeling*, vol. 54, no. 1, pp. 159–168, 2014.

[57] F. Cumbo, P. Paci, D. Santoni, L. Di Paola, and A. Giuliani, "GIANT: a cytoscape plugin for modular networks," *PLoS ONE*, vol. 9, no. 10, Article ID e105001, 2014.

[58] A. Krishnan, J. P. Zbilut, M. Tomita, and A. Giuliani, "Proteins as networks: usefulness of graph theory in protein science," *Current Protein and Peptide Science*, vol. 9, no. 1, pp. 28–38, 2008.

[59] H. Li, S. Sakuraba, A. Chandrasekaran, and L.-W. Yang, "Molecular binding sites are located near the interface of intrinsic dynamics domains (IDDs)," *Journal of Chemical Information and Modeling*, vol. 54, no. 8, pp. 2275–2285, 2014.

[60] U. Emekli, D. Schneidman-Duhovny, H. J. Wolfson, R. Nussinov, and T. Haliloglu, "HingeProt: automated prediction of hinges in protein structures," *Proteins: Structure, Function and Genetics*, vol. 70, no. 4, pp. 1219–1227, 2008.

[61] L.-W. Yang and I. Bahar, "Coupling between catalytic site and collective dynamics: a requirement for mechanochemical activity of enzymes," *Structure*, vol. 13, no. 6, pp. 893–904, 2005.

[62] T. L. Rodgers, P. D. Townsend, D. Burnell et al., "Modulation of global low-frequency motions underlies allosteric regulation: demonstration in CRP/FNR family transcription factors," *PLoS Biology*, vol. 11, no. 9, Article ID e1001651, 2013.

[63] C. Chennubhotla and I. Bahar, "Signal propagation in proteins and relation to equilibrium fluctuations," *PLoS Computational Biology*, vol. 3, no. 9, pp. 1716–1726, 2007.

[64] A. Dutta and I. Bahar, "Metal-binding sites are designed to achieve optimal mechanical and signaling properties," *Structure*, vol. 18, no. 9, pp. 1140–1148, 2010.

[65] B. M. Butler, Z. N. Gerek, S. Kumar, and S. B. Ozkan, "Conformational dynamics of nonsynonymous variants at protein interfaces reveals disease association," *Proteins: Structure, Function and Bioinformatics*, vol. 83, no. 3, pp. 428–435, 2015.

[66] N. Kantarci, P. Doruker, and T. Haliloglu, "Cooperative fluctuations point to the dimerization interface of p53 core domain," *Biophysical Journal*, vol. 91, no. 2, pp. 421–432, 2006.

[67] A. Zen, C. Micheletti, O. Keskin, and R. Nussinov, "Comparing interfacial dynamics in protein-protein complexes: an elastic network approach," *BMC Structural Biology*, vol. 10, article 26, 2010.

[68] S. Soner, P. Ozbek, J. I. Garzon, N. Ben-Tal, and T. Haliloglu, "DynaFace: discrimination between obligatory and non-obligatory protein-protein interactions based on the complex's dynamics," *PLoS Computational Biology*, vol. 11, no. 10, Article ID e1004461, 2015.

[69] O. K. Yusuff, J. O. Babalola, G. Bussi, and S. Raugei, "Role of the subunit interactions in the conformational transitions in adult human hemoglobin: an explicit solvent molecular dynamics study," *Journal of Physical Chemistry B*, vol. 116, no. 36, pp. 11004–11009, 2012.

[70] O. N. Yogurtcu, S. B. Erdemli, R. Nussinov, M. Turkay, and O. Keskin, "Restricted mobility of conserved residues in protein-protein interfaces in molecular simulations," *Biophysical Journal*, vol. 94, no. 9, pp. 3475–3485, 2008.

[71] E. Cukuroglu, H. B. Engin, A. Gursoy, and O. Keskin, "Hot spots in protein-protein interfaces: towards drug discovery," *Progress in Biophysics and Molecular Biology*, vol. 116, no. 2-3, pp. 165–173, 2014.

[72] C. Chennubhotla, A. J. Rader, L.-W. Yang, and I. Bahar, "Elastic network models for understanding biomolecular machinery: from enzymes to supramolecular assemblies," *Physical Biology*, vol. 2, no. 4, pp. S173–S180, 2005.

[73] I. Bahar, A. R. Atilgan, M. C. Demirel, and B. Erman, "Vibrational dynamics of folded proteins: significance of slow and fast motions in relation to function and stability," *Physical Review Letters*, vol. 80, no. 12, pp. 2733–2736, 1998.

[74] P. Ozbek, S. Soner, and T. Haliloglu, "Hot spots in a network of functional sites," *PloS ONE*, vol. 8, no. 9, Article ID e74320, 2013.

[75] C. B. M. Platania, L. Di Paola, G. M. Leggio et al., "Molecular features of interaction between VEGFA and anti-angiogenic drugs used in retinal diseases: a computational approach," *Frontiers in Pharmacology*, vol. 6, p. 248, 2015.

[76] A. R. Atilgan, P. Akan, and C. Baysal, "Small-world communication of residues and significance for protein dynamics," *Biophysical Journal*, vol. 86, no. 1, pp. 85–91, 2004.

[77] A. Del Sol, H. Fujihashi, D. Amoros, and R. Nussinov, "Residues crucial for maintaining short paths in network communication mediate signaling in proteins," *Molecular Systems Biology*, vol. 2, no. 1, 2006.

[78] G. Bagler and S. Sinha, "Assortative mixing in protein contact networks and protein folding kinetics," *Bioinformatics*, vol. 23, no. 14, pp. 1760–1767, 2007.

[79] G. Hu, W. Yan, J. Zhou, and B. Shen, "Residue interaction network analysis of Dronpa and a DNA clamp," *Journal of Theoretical Biology*, vol. 348, pp. 55–64, 2014.

[80] F. Raimondi, A. Felline, M. Seeber, S. Mariani, and F. Fanelli, "A mixed protein structure network and elastic network model approach to predict the structural communication in biomolecular systems: The PDZ2 domain from tyrosine phosphatase 1E as a case study," *Journal of Chemical Theory and Computation*, vol. 9, no. 5, pp. 2504–2518, 2013.

Hindawi Publishing Corporation
BioMed Research International
Volume 2016, Article ID 8518945, 11 pages
http://dx.doi.org/10.1155/2016/8518945

Review Article

Biomolecular Network-Based Synergistic Drug Combination Discovery

Xiangyi Li,[1,2] Guangrong Qin,[2] Qingmin Yang,[1,2] Lanming Chen,[1] and Lu Xie[2]

[1]*Key Laboratory of Quality and Safety Risk Assessment for Aquatic Products on Storage and Preservation (Shanghai), China Ministry of Agriculture, College of Food Science and Technology, Shanghai Ocean University, 999 Hu Cheng Huan Road, Shanghai 201306, China*
[2]*Shanghai Center for Bioinformation Technology, Shanghai Academy of Science and Technology, 1278 Keyuan Road, Shanghai 201203, China*

Correspondence should be addressed to Lanming Chen; lmchen@shou.edu.cn and Lu Xie; xielu@scbit.org

Received 24 June 2016; Revised 20 September 2016; Accepted 11 October 2016

Academic Editor: Zhongjie Liang

Drug combination is a powerful and promising approach for complex disease therapy such as cancer and cardiovascular disease. However, the number of synergistic drug combinations approved by the Food and Drug Administration is very small. To bridge the gap between urgent need and low yield, researchers have constructed various models to identify synergistic drug combinations. Among these models, biomolecular network-based model is outstanding because of its ability to reflect and illustrate the relationships among drugs, disease-related genes, therapeutic targets, and disease-specific signaling pathways as a system. In this review, we analyzed and classified models for synergistic drug combination prediction in recent decade according to their respective algorithms. Besides, we collected useful resources including databases and analysis tools for synergistic drug combination prediction. It should provide a quick resource for computational biologists who work with network medicine or synergistic drug combination designing.

1. Introduction

Complex diseases such as cancer are often caused by a collective of abnormalities of correlated genes or biology processes. The traditional paradigm of "one gene, one drug, one disease" has been challenged by the increasing rate of drug failure and the huge costs of time and money in drug development and research [1, 2]. What is worse, the recurrent drug resistance has significantly reduced the efficacy of the existing drugs [3, 4].

To tackle the problem of drug resistance and reduce the great expenses of time and money in drug discovery, researchers have made great efforts to discover synergistic drug combinations. Synergistic drug combination means that the overall therapeutic effect of the combination is larger than the sum of effects independently caused by individual component [5]. Synergistic drug combination can decrease the drug dosage but increase or maintain the same efficacy to avoid toxicity, minimize, or slow down the development of dug resistance [6]. Inspired by the great benefit of synergistic drug combination, both *in silico* methods and *in vitro* methods have been applied to screen synergistic drug combinations. The most straightforward methods for screening synergistic drug combination are *in vitro* methods.

There are three popular reference models *in vitro* methods, namely, the highest single agent (HSA) model [7], the Loewe additivity model [8], and the Bliss independent model [9]. The difference between these models is their definitions of the noninteraction effect of a drug pair which means the expected additive effect of a drug pair. Specifically, HSA defined the noninteraction effect as the highest monotherapy effect among the individual drug in drug combinations. Loewe additivity model defines the noninteraction effect of a drug pair as if single drug is combined with itself. In contrast, Bliss independence model defines the noninteraction effect of a drug pair as if two drugs work independently. In addition, the integration of the HSA model and the Bliss independence model, called zero interaction potency (ZIP), has

been applied to large-scale dose-response matrix experiments [10]. More importantly, Chou proposed a popular algorithm based on the median-effect equation which was encompassed by several complex equations such as Michaelis-Menten, Hill, Henderson-Hasselbalch, and Scatchard equations in biochemistry and biophysics [11, 12]. The core concept of this algorithm, combination index (CI), is an indicator evaluating drug combination interaction effect. CI has been widely used in identification of synergistic drug combination *in vitro*, especially in validation of novel drug combinations from various kinds of computational models. The *in vitro* models mentioned above are all based on drug-treated dose-response curve. Dose-response curve based model can achieve good performance for low-throughput data and be used to validate novel drug combinations. Nevertheless, without involving any molecular level data such as drug-treatment transcriptional expression profile, these models may not help to discover the underlying mechanism of drug synergy [13]. Besides, *in vitro* methods which screen all possible combinations by experimental trials are time and money consuming, and usually only a small number of synergistic drug combinations can be identified.

To address this limitations of *in vitro* methods in identifying synergistic drug combinations, *in silico* methods based on "omics" data have become more and more popular [14–17]. With the explosion of "omics" data in recent years, high-throughput data at various levels related to disease state and drug-treatment have been accumulated rapidly. Also, in the recent decades, protein interactions have been extensively studied, forming comprehensive knowledge background of molecular regulation pathways or networks. In addition, taking the heterogeneity and redundancy of diseases into account, researchers have come to realize that the understanding of mechanism of drug synergistic effect calls for analysis of biology system in a network perspective [1, 18]. Network analysis involves mathematics and computer science into biology and can help to present relationship among molecules in the perspective of network. What is more, network analysis has an advantage of finding newly emerged properties at a network level [19–23]. With the benefit of network analysis, the researchers can make full use of high-throughput data by modeling the interaction among drugs, targets, and diseases, which can definitively promote the discovery of the complex mechanism of drug synergy.

Inspired by the rapid development of biomolecular network-based synergistic drug combination discovery, here we present a brief review on the prediction models of synergistic drug combination. All the models are divided into three classes according to the type of drug pairs used to train the prediction model, namely, unsupervised learning prediction models depending on hypothesis of drug synergy and unlabeled drug pairs, semi-supervised learning prediction models involving few labeled drug pairs, and many unlabeled drug pairs in model training and supervised learning prediction models using labeled drug pairs to train models. Labeled drug pairs denote that effective drug combinations have been approved by FDA or validated by experiments, while unlabeled drug pairs denote drug pairs without synergistic effect evidence. The general work flow of identification of novel synergistic drug combinations based on biomolecular network is shown in Figure 1.

2. Important Public Resources for Network Construction in Predicting Synergistic Drug Combinations

Recently, with the rapid development of next generation sequencing (NGS) technologies and rapidly accumulation of "omics" data [24], there are many useful databases and analysis tools for predicting and screening synergistic drug combinations. Table 1 includes the tools that can be used to identify synergistic drug combinations *in silico*. Table 2 lists important databases containing many instances of drug combinations for cancers as well as other diseases. Table 3 lists the popular biology network-related databases that can aid the construction of molecular interaction network.

The Connectivity Map (CMap) collects a genome-wide transcriptional expression data from cultured human cells treated with various bioactive small molecules, providing a bridge to connect drugs and genes [51]. Drug Combination Database (DCDB) contains 1363 drug combinations, providing an important source for model construction and validation [35]. Here, we collected drug combinations from DCDB with each individual component of these drug combinations in CMap database (see supplementary file in Supplementary Material available online at http://dx.doi.org/10.1155/2016/8518945). With these data, the performance of prediction models based on compound/agent-treated transcriptional expression profile can be evaluated without conducting experiments.

3. Principles of Identifying Synergistic Drug Combinations Based on Biomolecular Network

The therapy strategy of "one drug, one gene" is not always successful to treat disease because cells can often find alternative ways to compensate the function after a gene or protein is perturbed by drug treatment [52]. Thus to treat these complex diseases, it is beneficial to consider the relationships among drugs, disease-related genes, therapeutic targets, and disease-specific signaling pathways as a system. Known drug combinations provide us a useful resource for predicted novel drug combinations. Taking the known drug combinations into consideration, labeled drug combinations can be used for supervised learning methods. The underlying principle in supervised models is that the more similarity to known drug combinations for a novel drug combination, the more likely it can become a synergistic drug combination. On the other hand, unsupervised models use no labeled drug combination to build and train model and mainly analyze the biological networks perturbed by drug combinations. A network presents the relationships (edges) of a set of entities (nodes). These nodes and edges have various important attributes such as degree, betweenness, and eigenvector centrality. Nodes can denote molecules such as genes, proteins, and drugs. Edges

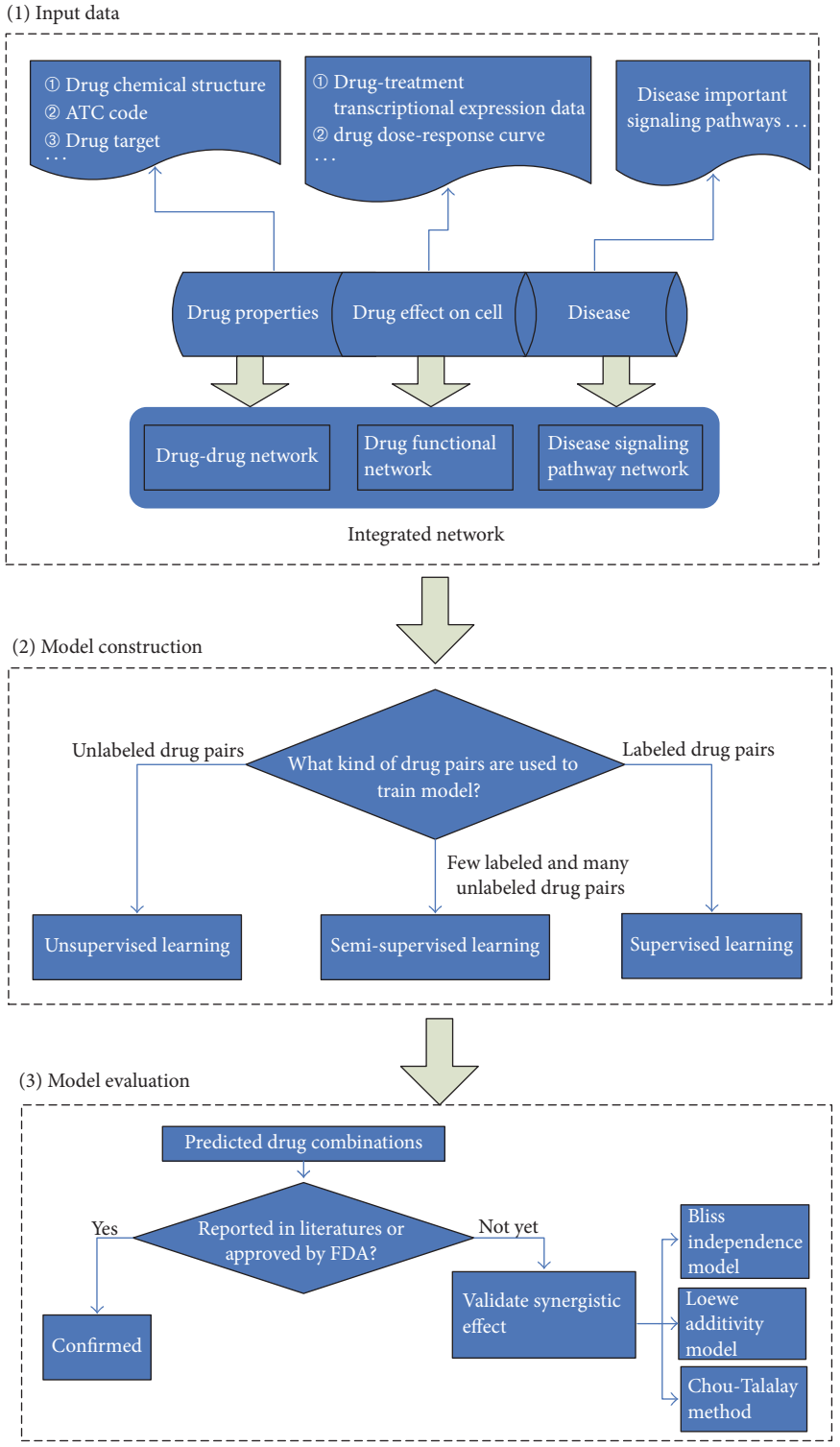


Figure 1: The general work flow of identification of novel synergistic drug combinations based on biomolecular network.

connecting these nodes can represent the interactions such as physical interactions and genetic regulatory interactions [53]. Drug synergy has been reported to be a property largely determined by network topology [54]. Therefore, biomolecular network analysis should provide useful insight into the mechanism of action of drug combinations.

4. Biomolecular Network-Based Unsupervised Learning Models for Synergistic Drug Combination Identification

Unsupervised learning models were mainly based on various features of drugs, targets, drug-treated cellular response data,

Table 1: Tools used to analyze drug combination data.

Tool name	Tool type	Reference model(s)	Input data	Brief description
CompuSyn [25]	Free software	Loewe additivity model	Dose-response data	CompuSyn only allows for manual input of one drug combination at a time
Synergyfinder [26]	R package	HSA, Loewe additivity, Bliss independence	Dose-response data	Synergyfinder is implementations for all the popular synergy scoring models for drug combinations, including HAS, Loewe, Bliss, and ZIP [10]
Mixlow [27]	R package	A nonlinear mixed-effects model	Dose-response data	Mixlow used a nonlinear mixed-effects model to estimate parameters of dose-response curves and required experimental design where the ratio of two drugs in a combination is fixed
COMBIA [28]	R package	Bliss independence, Loewe additivity	Data from wet-lab experimental	Data from wet-lab experimental platforms can be directly used
MacSynergyII [29]	Free software	Bliss independence	Dose-response data	MacSynergy II is essentially an Excel file and it scales the input data to %inhibition using positive and negative controls
Combenefit [30]	Free software	HSA, Loewe additivity, Bliss independence	Dose-response data	Combenefit has advanced graphical capabilities and can be applied to model-based quantification of drug combinations in single and high-throughput settings
Combinatorial Drug Assembler [31] (http://cda.i-pharm.org/)	Free web app implementation	None	Disease-related signaling pathway components	CDA performs expression pattern matching between input gene sets and 6,100 molecule-treated expression profiles of the connectivity map to list up best pattern matching single drugs/combinatorial drug pairs
Synergy Maps [32] (http://richlewis42.github.io/synergy-maps/)	Free web app implementation	None	Drugs or drug combinations in two datasets [25, 26]	Synergy Maps can simultaneously represent individual compound properties and their interactions
DT-Web [33] (http://alpha.dmi.unict.it/dtweb/)	Free web app implementation	None	The name or the accession number of a drug/target	A web-based application for drug-target interaction and drug combination prediction
TIMMA-R [34]	R package	Logic-based network	Drugs' polypharmacological profiles and drug sensitivity profiles from a given cancer cell line	TIMMA-R predicts the effects of drug combinations based on their binary drug target interactions and single-drug sensitivity profiles

Table 2: Integrated drug combination databases.

Database	URL
DCDB [35]	http://www.cls.zju.edu.cn/dcdb/
TTD [36]	http://bidd.nus.edu.sg/group/cjttd/
TCM [37]	http://tcm.cmu.edu.tw/
ASDCD [38]	http://asdcd.amss.ac.cn/

and functional networks following various hypotheses. Drug-treated cellular response data can provide an insight of drug mechanism of action [55]. Transcriptional expression profile is one of the most common cellular response data used to study the mechanism underlying a biological pathway [56] and the biology response of a cell to a certain perturbation [51, 57]. According to different hypotheses, several prediction models have been built.

DrugComboRanker was built based on the hypothesis that effective drug combinations can inhibit major modules of disease signaling networks simultaneously and drugs often have multiple active target genes or proteins [58]. Based on transcriptional expression data from cell lines treated with small molecule compounds in CMap, researchers built a drug functional network and divided the network into numerous drug network communities by a Bayesian nonnegative matrix factorization method [59]. Besides they also constructed a disease-specific signaling network utilizing patients' genomic profile and interactome data. Then they defined a synergy score prioritizing the drug pairs that target on disease-specific signaling network with similar function. Finally, all the drug pairs were ranked in the descend order of the synergy score and the top-ranked drug pairs will be more likely to be synergistic.

Jin et al. built a model called enhanced Petri-net (EPN) to predict the synergistic effect of pairwise drug combinations from genome-wide transcriptional expression data by applying Petri-net to identify drug targeted signaling network [60]. They assumed that there existed at least one molecule that shows the enhanced effect in the pairwise combination compared with the summation of the effect generated by the two drugs individually. They identified synergistic drug combinations by comparing the drug effects from the transcriptional expression data treated by pairwise combination of drugs A and B with those from the corresponding two transcriptional expression data treated by drugs A and B separately.

Wu et al. utilized information of protein interactions, protein-DNA interactions, and signaling pathways to construct a molecular interaction network [61]. Their assumption was that a subnetwork or pathway would be affected in the networked cellular system after a drug was administrated. They built a model to detect the subnetwork perturbed by drug combinations. Based on the molecular interaction network, they defined an interaction score that indicated the gap between drug efficacy effect and side effect. Drug combination whose interaction score for certain subnetwork is higher than that of any individual drug would be recognized as effective drug combination.

Similarly, a model named pathway and pathway interaction (WWI) was based on the assumption that drugs targeting one same pathway or related pathways will be more likely to be synergistic drug combinations [62]. The researchers built two networks, namely, a PPI network based on information from HPRD [44] database and a WWI network based on KEGG database. In WWI network, nodes are the "*Homo sapiens*" pathways, and edges are pathway-pathway interactions. Then for each drug pair they defined a score which indicated the connectivity of pathways perturbed by the individual drug of drug combinations on the WWI network and drug targets on the PPI network. Finally, all the query drug pairs were ranked in descend order of the score and the top-ranked ones would be more likely to be synergetic.

Another group of unsupervised learning prediction models were built according to the DEARM Challenge data. In 2012, the Dialogue on Reverse Engineering Assessment and Methods (DREAM) consortium designed an open competition for researchers all over the world to rank the effect of all 91 pairwise combinations on OCI-LY3 human diffuse large B-cell lymphoma (DLBCL) cell line from the most synergistic to the most antagonistic [63]. This project generated transcriptional expression data only for samples treated by individual drug, dose-response curves for viability of OCI-LY3 cells following perturbation with 14 distinct compounds and baseline genetic profile of the OCI-LY3 cell line. Among all the 31 groups taking part in the project, three of them performed significantly better than random guess. All of the three groups developed their models based on their assumptions about drug synergy, such as assumption that changes in gene expression after drug perturbations could be used to predicted these drug interaction effect [55] or the correlation of differential expression genes (DEGs) after two drugs perturbations would reflect the possibility of drug synergistic effect [64]. Although the final result of this project was modest, the challenge gives us reasons to hope for powerful methods to identify effective drug combinations in the future [65].

Among all of those 31 groups participating in the project, Drug-Induced Genomic Residual Effect (DIGRE) model achieved the best performance on prediction of synergistic drug combinations [66]. DIGRE was developed based on the hypothesis that, for two drugs used sequentially, the first drug would change the transcriptome of the treated cell and thus modulate the effect of the other one. Firstly, the researchers constructed a gene-gene interaction network based on KEGG pathways. Based on the network, DIGRE required transcriptional expression profiles and dose-response curves provided by the DREAM Challenge as the input data. Then drug similarity score of the drug pair was computed by accounting for DEGs of each drug, including common DEGs and upstream and downstream genes in the selected pathways. Finally, all 91 drug pairs were ranked in descend order based on the combinatorial effect score which were computed based on corresponding drug response curve and drug similarity score.

From above, we can see that drug-treated transcriptional expression data is commonly used in unsupervised learning

Table 3: Important biology network-related databases.

Database	URL	Data type
STRING [39]	http://string-db.org/	Protein-protein interactions
Reactome [40]	http://www.reactome.org/	human biological processes
KEGG [41]	http://www.genome.jp/kegg/	Pathway, disease, drug
BioGRID [42]	https://wiki.thebiogrid.org/	PPI/genetic interaction
STITCH [43]	http://stitch.embl.de/	Chemical-protein interaction
HPRD [44]	http://hprd.org/	Protein-protein interaction (PPI)
DIP [45]	http://dip.doe-mbi.ucla.edu/dip/Main.cgi	PPI
IntAct [46]	http://www.ebi.ac.uk/intact/	Molecular interaction
WikiPathways [47]	http://www.wikipathways.org/index.php/WikiPathways	Biological pathways
TRED [48]	https://cb.utdallas.edu/cgi-bin/TRED/tred.cgi?process=home	TF-gene interaction
InterDom [49]	http://interdom.i2r.a-star.edu.sg/	Domain interaction
SignaLink [50]	http://signalink.org/	Signaling pathways

prediction models because of its informative properties, such as reflecting the mechanism of drug action underlying a biological pathway. However, there is no standard rule to process transcriptional expression data. Thus, selection of significant differential expression genes is largely depended on individual researchers and significant gene lists can be quite diverse according to different algorithms. For models based on transcriptional expression data, different networks will be constructed by different researchers depending on diverse processing methods of the same microarray data, which can lead to difficulty in the final interpretation [31].

5. Biomolecular Network-Based Semi-Supervised Learning Models for Synergistic Drug Combination Identification

Traditional classifier uses only labeled data to train the prediction model. However, it is both time and money consuming to collect labeled data by experts. Semi-supervised learning solves this problem by using large amount of unlabeled data together with a limited number of labeled data [67]. To the best of our knowledge, the number of approved drug combinations is still much less than drug combinations without synergistic effect evidence, so semi-supervised learning can solve this problem.

Sun et al. constructed a model called Ranking-system of Anticancer Synergy (RACS) based on semi-supervised learning which was used to rank drug pairs according to their similarity to the labeled samples in a specified multifeature space [68]. Firstly, they performed feature selection to identify significantly different features between labeled samples and the unlabeled samples. Some interesting features had been identified, such as drug target distance in PPI network and the proportion of unrelated pathways regulated by the targets of the two agents. Then all the drug pairs (i.e., labeled and unlabeled samples) were represented by a vector of the selected features (mentioned in the previous step). Finally, they incorporated a manifold ranking algorithm with semi-supervised learning method to enrich the labeled pairs at the top of the drug pair list [69]. To evaluate the performance of RACS in test dataset, the researchers applied RACS to data provided by the above mentioned DREAM Challenge project [49]. It impressively achieved greater progress than that of the best model, DIGRE. However, despite the complexity of the manifold ranking algorithm and some other complex mathematics methods used in RACS, it also largely relied on the known drug targets to calculate the average distance between the target proteins of the two agents in the context of PPI network. So far, a part of compound targets are still unknown; thus this will limit the application of RACS in synergistic drug combination identification.

Chen et al. developed an algorithm termed Network-based Laplacian regularized Least Square Synergistic drug combination prediction (NLLSS) based on their observation that principal drugs which obtain synergistic effect with similar adjuvant drugs are often similar [70], where principal drug means that the drug in synergistic drug combination shows activity in disease treatment and adjuvant drug means drug in synergistic drug combination shows no effect on disease treatment. NLLSS developed a classifier based on the framework of Laplacian Regularized Least Square (LapRLS) which is an popular semi-supervised learning algorithm [71]. Firstly, researchers computed drug similarity for principle drugs and adjuvant drugs, depending on several integrated information such as known synergistic drug combinations, drug combinations without known synergistic evidences, drug target interactions, and drug chemical structures. Finally, a score used to assess synergistic probability of a drug combination can be obtained depending on the result from previous step.

As can be seen from above, RACS and NLLSS share common features. Firstly, they both have a small number of labeled data, such as 26 labeled data compared with 502 unlabeled data for RACS training set and 75 labeled data compared with 4079 unlabeled data for NLLSS training set. Secondly, they were developed based on known and complex machine learning algorithms for manifold regularization which is a technique for using the shape of a dataset to constrain the functions that should be learned on that dataset [71].

6. Biomolecular Network-Based Supervised Learning Models for Synergistic Drug Combination Identification

Supervised learning is a machine learning algorithm of inferring a function from training data. The training data consists of a set of training samples which consist of an input object and a relevant label. Then with the inferred function, new sample can be labeled [72]. Thus with proper amount of known synergistic drug combinations as a training set, researchers can get a learned function by supervised learning which can identify candidate synergistic drug combinations.

Zhao et al. developed a model based on features of US Food and Drug Administration (FDA) approved drug combinations including drug features such as drug target proteins and corresponding downstream pathways, medical indication areas, therapeutic effects as represented in the Anatomical Therapeutic Chemical (ATC) Classification System, and side effects [13]. They performed 5-fold cross-validation on the above mentioned drug combinations to evaluate the performance of these features. Then the learned model was revised after deleting two weakly predictive features. Finally, drug pairs between marketed drugs from FDA orange book were applied to evaluate the performance of the model.

Xu et al. proposed a model called Drug Combination Predictor (DCPred) [73]. With the effective drug combinations collected from DCDB, they built a so-called drug-cocktail network which contained 215 nodes (i.e., unique drug) and 239 edges (i.e., known synergistic effect). Their hypothesis was that two drugs which shared large number of common drugs in drug-cocktail network would be more likely to be effective drug combinations. They found that, compared with drugs in random combination network, drugs in drug-cocktail network tended to have more therapeutic effects and more interaction partners. Based on these two topological features of drug combinations, they built a function which required parameters such as number of common neighbors of two drugs in drug-cocktail network and unique neighbors of individual drug to predict the probability for a drug pair to be an effective drug combination.

Similarly, Li et al. proposed another model called probability ensemble approach (PEA) based on drug-based similarity features including drug chemical structure, ATC code, target side effect, and target-based similarity features include target sequence, target-target interaction in PPI, and Gene Ontology (GO) semantic [74]. The six features for every drug pair were combined using a Bayesian network to calculate a likelihood ratio (LR) which can be used to the estimate of the similarity to known drug pair interaction [75]. A raw score for a query drug pair was defined by summing LRs to all the known drug pairs in each set (i.e., effective drug combinations or undesirable drug-drug interactions), which is further converted to p value based on random distribution. After performance evaluation of PEA using 10-fold cross-validation scheme accompanied with the receiver operating characteristic (ROC) curve analysis, integrative analysis of side-beneficial effects for drug combinations, external literature validation, and experimental validation, tens of effective drug combinations were confirmed.

Chandrasekaran et al. developed a computational model entitled INferring Drug Interactions using chemo-Genomics and Orthology (INDIGO) which used chemogenomics data to predict antibiotic drug combinations [76]. The core of INDIGO is a machine learning algorithm called random forest. To train INDIGO, the researchers firstly performed experimental measurement of 105 interactions ($C_{15}^2 = 105$) among 15 drugs. Then drug interaction data measuring synergistic and antagonistic effect together with chemogenomics data were put into INDIGO as training data. INDIGO only requires chemogenomics data of individual drugs to output novel drug combinations.

The supervised learning models tend to take advantage of drug property information like drug target, ATC code, and chemical structure, while drug synergy is strongly context-dependent, disease type [77] and drug dosages [60] can also modulate the efficacy of drug combinations. Therefore, future supervised learning models should take drug properties as well as drug-treated information into consideration to improve performance of prediction model.

7. Conclusions

With the explosive growth of high-throughput data, *in silico* modeling for synergistic drug combination represents both an opportunity and a challenge for medicine research. Combined with knowledge of mathematics, computer science, and biology, analysis of complex molecule interactions based on biomolecular networks can greatly accelerate the discovery of synergistic drug combinations [78]. To build a model with great prediction performance, intricate mathematics method is necessary to simulate the interactions between molecules in the complex biology system. For validation of the predicted novel drug combinations, *in vitro* methods should be used to get dose-response curves. Finally, either Chou-Talalay method or reference models such as Loewe additivity model and Bliss independent model can be used to determine the effect of drug combination (i.e., synergistic, antagonistic or additive).

Based on the review above, we can see that the number of biomolecular network-based unsupervised learning models is much bigger than that of semi-supervised learning or supervised learning. The possible reason is that only small number of drug combinations has been approved by regulatory agency, which limits the use of machine learning methods such as semi-supervised learning method and supervised learning method. Several significantly different features between synergistic drug combinations and random drug combinations have been identified. The average shortest distance in PPI network of targets between synergistic drug combinations is significantly smaller than that in random drug combinations [68, 73, 79]. Also, dissimilarity of drug chemical structure for individual drug in drug combination is significantly associated with drug synergistic effect [80].

There are several limitations in effectively applying these network-based models. Firstly, synergistic drug combination

modeling utilizing high-throughput data based on biomolecular network is in its infancy and most of these models have not been fully validated in practice. Secondly, despite the complexity of these prediction models, the massive noise of high-throughput data at different levels from different contributors can play an import role in model construction and in turn lower the performance of the prediction model. For instance, prediction models can perform very well in one test set, while poor result may be achieved when the model is applied to another test set. For example, Sun et al. applied the transcriptional expression profiles and target information of drug treatment on the human lung adenocarcinoma cell line A549 to RACS [68]. Results showed that only two drug combinations are synergetic ranked the top 10%; however drug combinations ranked the first and second are both antagonistic. Thirdly, networks such as PPI network, gene-protein interaction network, and drug target interaction network all have been considered in models mentioned above except metabolism network. Drug metabolism processes such as drug absorption and transportation are very important in disease treatment [81]. For example, one of the most important reasons for drug resistance is the overexpression of the P-glycoprotein (P-gp) [82], a drug efflux protein expressed by ABCB1 which is from ATP-binding cassette (ABC) family [83–85], and it has been reported that inhibition of P-gp can enhance the drug efficacy [86–90]. Thus, we deduced that drugs inhibiting efflux genes (e.g., ABC family genes) or activating drug influx genes (e.g., solute carrier transporter genes [91]) can make a contribution to drug synergistic effect. However, few prediction models took drug metabolism processes into consideration.

Future models for synergistic drug combination prediction should pay more attention to incorporating comprehensive information including disease signaling pathways and drug targeting pathways as well as drug metabolism processes such as drug absorption, transportation, metabolism, and clearance.

Competing Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Authors' Contributions

Xiangyi Li, Guangrong Qin, and Qingmin Yang collected the data and tools. Xiangyi Li, Guangrong Qin, Lanming Chen, and Lu Xie wrote and revised the manuscript. Xiangyi Li and Guangrong Qin contributed equally.

Acknowledgments

This work was funded by National Natural Science Foundation of China (31570831), National Key Research and Development Program of China (2016YFC0904100), the innovative pioneer project of Shanghai Industrial Technology Institute (16CXXF001), the National Hi-Tech Program (2015AA020101), and the Chinese Human Proteome Projects (CNHPP: 2014DFB30020 and 2014DFB30030).

References

[1] A. L. Hopkins, "Network pharmacology: the next paradigm in drug discovery," *Nature Chemical Biology*, vol. 4, no. 11, pp. 682–690, 2008.

[2] J. A. DiMasi, H. G. Grabowski, and R. W. Hansen, "Innovation in the pharmaceutical industry: new estimates of R&D costs," *Journal of Health Economics*, vol. 47, pp. 20–33, 2016.

[3] C. Holohan, S. Van Schaeybroeck, D. B. Longley, and P. G. Johnston, "Cancer drug resistance: an evolving paradigm," *Nature Reviews Cancer*, vol. 13, no. 10, pp. 714–726, 2013.

[4] R. Kumar, K. Chaudhary, S. Gupta et al., "CancerDR: cancer drug resistance database," *Scientific Reports*, vol. 3, article 1445, 2013.

[5] N. J. Sucher, "Searching for synergy in silico, in vitro and in vivo," *Synergy*, vol. 1, no. 1, pp. 30–43, 2014.

[6] T.-C. Chou, "Theoretical basis, experimental design, and computerized simulation of synergism and antagonism in drug combination studies," *Pharmacological Reviews*, vol. 58, no. 3, pp. 621–681, 2006.

[7] M. C. Berenbaum, "What is synergy?" *Pharmacological Reviews*, vol. 41, no. 2, pp. 93–141, 1989.

[8] S. Loewe, "The problem of synergism and antagonism of combined drugs," *Arzneimittel-Forschung*, vol. 3, no. 6, pp. 285–290, 1953.

[9] C. I. Bliss, "The toxicity of poisons applied jointly," *Annals of Applied Biology*, vol. 26, no. 3, pp. 585–615, 1939.

[10] B. Yadav, K. Wennerberg, T. Aittokallio, and J. Tang, "Searching for drug synergy in complex dose–response landscapes using an interaction potency model," *Computational and Structural Biotechnology Journal*, vol. 13, pp. 504–513, 2015.

[11] T.-C. Chou, "Derivation and properties of Michaelis-Menten type and Hill type equations for reference ligands," *Journal of Theoretical Biology*, vol. 59, no. 2, pp. 255–276, 1976.

[12] T.-C. Chou, "Drug combination studies and their synergy quantification using the Chou-Talalay method," *Cancer Research*, vol. 70, no. 2, pp. 440–446, 2010.

[13] X.-M. Zhao, M. Iskar, G. Zeller, M. Kuhn, V. Van Noort, and P. Bork, "Prediction of drug combinations by integrating molecular and pharmacological data," *PLoS Computational Biology*, vol. 7, no. 12, Article ID e1002323, 2011.

[14] H. Van De Waterbeemd and E. Gifford, "ADMET in silico modelling: towards prediction paradise?" *Nature Reviews Drug Discovery*, vol. 2, no. 3, pp. 192–204, 2003.

[15] S. Ekins, "Predicting undesirable drug interactions with promiscuous proteins in silico," *Drug Discovery Today*, vol. 9, no. 6, pp. 276–285, 2004.

[16] S. Ekins, A. J. Williams, M. D. Krasowski, and J. S. Freundlich, "In silico repositioning of approved drugs for rare and neglected diseases," *Drug Discovery Today*, vol. 16, no. 7-8, pp. 298–310, 2011.

[17] B. Al-Lazikani, U. Banerji, and P. Workman, "Combinatorial drug therapy for cancer in the post-genomic era," *Nature Biotechnology*, vol. 30, no. 7, pp. 679–692, 2012.

[18] A.-L. Barabási, N. Gulbahce, and J. Loscalzo, "Network medicine: a network-based approach to human disease," *Nature Reviews Genetics*, vol. 12, no. 1, pp. 56–68, 2011.

[19] A.-L. Barabási and Z. N. Oltvai, "Network biology: understanding the cell's functional organization," *Nature Reviews Genetics*, vol. 5, no. 2, pp. 101–113, 2004.

[20] L. Chen, R.-S. Wang, and X.-S. Zhang, *Biomolecular Networks: Methods and Applications in Systems Biology*, vol. 10, John Wiley & Sons, New York, NY, USA, 2009.

[21] Z.-P. Liu, Y. Wang, X.-S. Zhang, and L. Chen, "Network-based analysis of complex diseases," *IET Systems Biology*, vol. 6, no. 1, pp. 22–33, 2012.

[22] L. Chen, R. Wang, C. Li, and K. Aihara, *Modeling Biomolecular Networks in Cells: Structures and Dynamics*, Springer Science & Business Media, 2010.

[23] J. Ruan, A. K. Dean, and W. Zhang, "A general co-expression network-based approach to gene expression analysis: comparison and applications," *BMC Systems Biology*, vol. 4, article 8, 2010.

[24] J. A. Reuter, D. V. Spacek, and M. P. Snyder, "High-throughput sequencing technologies," *Molecular Cell*, vol. 58, no. 4, pp. 586–597, 2015.

[25] T. Chou and N. Martin, *CompuSyn for Drug Combinations: PC Software and User's Guide: A Computer Program for Quantitation of Synergism and Antagonism in Drug Combinations, and the Determination of IC50 and ED50 and LD50 Values*, ComboSyn, Paramus, NJ, USA, 2005.

[26] L. He, E. Kulesskiy, J. Saarela et al., "Methods for high-throughput drug combination screening and synergy scoring," *BioRxiv*, 2016.

[27] J. C. Boik and B. Narasimhan, "An R package for assessing drug synergism/antagonism," *Journal of Statistical Software*, vol. 34, no. 6, pp. 1–18, 2010.

[28] M. Kashif, "Synergy/Antagonism Analyses of Drug Combinations," R package version 104, 2015.

[29] M. Prichard, K. Aseltine, and C. Shipman Jr., *MacSynergy II. Version 1.0. User's Manual*, University of Michigan, Ann Arbor, Mich, USA, 1993.

[30] G. Y. Di Veroli, C. Fornari, D. Wang et al., "Combenefit: an interactive platform for the analysis and visualization of drug combinations," *Bioinformatics*, vol. 32, no. 18, pp. 2866–2868, 2016.

[31] J.-H. Lee, D. G. Kim, T. J. Bae et al., "CDA: combinatorial drug discovery using transcriptional response modules," *PLoS ONE*, vol. 7, no. 8, Article ID e42573, 2012.

[32] R. Lewis, R. Guha, T. Korcsmaros, and A. Bender, "Synergy Maps: exploring compound combinations using network-based visualization," *Journal of Cheminformatics*, vol. 7, no. 1, article 36, 11 pages, 2015.

[33] S. Alaimo, V. Bonnici, D. Cancemi, A. Ferro, R. Giugno, and A. Pulvirenti, "DT-Web: a web-based application for drug-target interaction and drug combination prediction through domain-tuned network-based inference," *BMC Systems Biology*, vol. 9, no. 3, p. 1, 2015.

[34] L. He, K. Wennerberg, T. Aittokallio, and J. Tang, "TIMMA-R: an R package for predicting synergistic multi-targeted drug combinations in cancer cell lines or patient-derived samples," *Bioinformatics*, vol. 31, no. 11, pp. 1866–1868, 2015.

[35] Y. Liu, Q. Wei, G. Yu, W. Gai, Y. Li, and X. Chen, "DCDB 2.0: a major update of the drug combination database," *Database*, vol. 2014, Article ID bau124, 2014.

[36] F. Zhu, B. Han, P. Kumar et al., "Update of TTD: therapeutic target database," *Nucleic Acids Research*, vol. 38, supplement 1, pp. D787–D791, 2010.

[37] C. Y.-C. Chen, "TCM Database@Taiwan: the world's largest traditional Chinese medicine database for drug screening in silico," *PLoS ONE*, vol. 6, no. 1, Article ID e15939, 2011.

[38] X. Chen, B. Ren, M. Chen et al., "ASDCD: antifungal synergistic drug combination database," *PLoS ONE*, vol. 9, no. 1, article e86499, 2014.

[39] D. Szklarczyk, A. Franceschini, M. Kuhn et al., "The STRING database in 2011: Functional interaction networks of proteins, globally integrated and scored," *Nucleic Acids Research*, vol. 39, no. 1, pp. D561–D568, 2011.

[40] G. Joshi-Tope, M. Gillespie, I. Vastrik et al., "Reactome: a knowledgebase of biological pathways," *Nucleic Acids Research*, vol. 33, supplement 1, pp. D428–D432, 2005.

[41] M. Kanehisa and S. Goto, "KEGG: kyoto encyclopedia of genes and genomes," *Nucleic Acids Research*, vol. 28, no. 1, pp. 27–30, 2000.

[42] C. Stark, B.-J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers, "BioGRID: a general repository for interaction datasets," *Nucleic Acids Research*, vol. 34, supplement 1, pp. D535–539, 2006.

[43] M. Kuhn, D. Szklarczyk, A. Franceschini, C. Von Mering, L. J. Jensen, and P. Bork, "STITCH 3: zooming in on protein-chemical interactions," *Nucleic Acids Research*, vol. 40, no. 1, pp. D876–D880, 2012.

[44] L. Baolin and H. Bo, "HPRD: a high performance RDF database," in *Network and Parallel Computing: IFIP International Conference, NPC 2007, Dalian, China, September 18–21, 2007. Proceedings*, vol. 4672 of *Lecture Notes in Computer Science*, pp. 364–374, Springer, Berlin, Germany, 2007.

[45] I. Xenarios, Ł. Salwínski, X. J. Duan, P. Higney, S.-M. Kim, and D. Eisenberg, "DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions," *Nucleic Acids Research*, vol. 30, no. 1, pp. 303–305, 2002.

[46] S. Kerrien, B. Aranda, L. Breuza et al., "The IntAct molecular interaction database in 2012," *Nucleic Acids Research*, vol. 40, no. 1, Article ID gkr1088, pp. D841–D846, 2012.

[47] T. Kelder, M. P. Van Iersel, K. Hanspers et al., "WikiPathways: building research communities on biological pathways," *Nucleic Acids Research*, vol. 40, no. D1, pp. D1301–D1307, 2012.

[48] C. Jiang, Z. Xuan, F. Zhao, and M. Q. Zhang, "TRED: a transcriptional regulatory element database, new entries and other development," *Nucleic Acids Research*, vol. 35, no. 1, pp. D137–D140, 2007.

[49] S.-K. Ng, Z. Zhang, S.-H. Tan, and K. Lin, "InterDom: a database of putative interacting protein domains for validating predicted protein interactions and complexes," *Nucleic Acids Research*, vol. 31, no. 1, pp. 251–254, 2003.

[50] D. Fazekas, M. Koltai, D. Türei et al., "SignaLink 2—a signaling pathway resource with multi-layered regulatory networks," *BMC Systems Biology*, vol. 7, article 7, 2013.

[51] J. Lamb, E. D. Crawford, D. Peck et al., "The connectivity map: using gene-expression signatures to connect small molecules,

genes, and disease," *Science*, vol. 313, no. 5795, pp. 1929–1935, 2006.

[52] S. Frantz, "Drug discovery: playing dirty," *Nature*, vol. 437, no. 7061, pp. 942–943, 2005.

[53] P. Li, Y. Fu, and Y. Wang, "Network based approach to drug discovery: a mini review," *Mini-Reviews in Medicinal Chemistry*, vol. 15, no. 8, pp. 687–695, 2015.

[54] N. Yin, W. Ma, J. Pei, Q. Ouyang, C. Tang, and L. Lai, "Synergistic and antagonistic drug combinations depend on network topology," *PLoS ONE*, vol. 9, no. 4, article e93960, 2014.

[55] C. P. Goswami, L. Cheng, P. S. Alexander, A. Singal, and L. Li, "A new drug combinatory effect prediction algorithm on the cancer cell based on gene expression and dose-response curve," *CPT: Pharmacometrics and Systems Pharmacology*, vol. 4, no. 2, pp. 80–90, 2015.

[56] J. L. DeRisi, V. R. Iyer, and P. O. Brown, "Exploring the metabolic and genetic control of gene expression on a genomic scale," *Science*, vol. 278, no. 5338, pp. 680–686, 1997.

[57] J. Lamb, S. Ramaswamy, H. L. Ford et al., "A mechanism of cyclin D1 action encoded in the patterns of gene expression in human cancer," *Cell*, vol. 114, no. 3, pp. 323–334, 2003.

[58] L. Huang, F. Li, J. Sheng et al., "DrugComboRanker: drug combination discovery based on target network analysis," *Bioinformatics*, vol. 30, no. 12, pp. I228–I236, 2014.

[59] V. Y. Tan and C. Févotte, "Automatic relevance determination in nonnegative matrix factorization with the/spl beta/-divergence," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 7, pp. 1592–1605, 2013.

[60] G. Jin, H. Zhao, X. Zhou, and S. T. C. Wong, "An enhanced Petri-Net model to predict synergistic effects of pairwise drug combinations from gene microarray data," *Bioinformatics*, vol. 27, no. 13, pp. i310–i316, 2011.

[61] Z. Wu, X.-M. Zhao, and L. Chen, "A systems biology approach to identify effective cocktail drugs," *BMC Systems Biology*, vol. 4, no. 2, article 7, 2010.

[62] D. Chen, H. Zhang, P. Lu, X. Liu, and H. Cao, "Synergy evaluation by a pathway–pathway interaction network: a new way to predict drug combination," *Molecular BioSystems*, vol. 12, no. 2, pp. 614–623, 2016.

[63] M. Bansal, J. Yang, C. Karan et al., "A community computational challenge to predict the activity of pairs of compounds," *Nature Biotechnology*, vol. 32, no. 12, pp. 1213–1222, 2014.

[64] J. Zhao, X.-S. Zhang, and S. Zhang, "Predicting cooperative drug effects through the quantitative cellular profiling of response to individual drugs," *CPT: Pharmacometrics & Systems Pharmacology*, vol. 3, no. 2, article 79, pp. 1–7, 2014.

[65] R. B. Altman, "Predicting cancer drug response: advancing the DREAM," *Cancer Discovery*, vol. 5, no. 3, pp. 237–238, 2015.

[66] J. Yang, H. Tang, Y. Li et al., "DIGRE: drug-induced genomic residual effect model for successful prediction of multidrug effects," *CPT: Pharmacometrics and Systems Pharmacology*, vol. 4, no. 2, pp. 91–97, 2015.

[67] X. Zhu, "Semi-supervised learning literature survey," *Computer Science*, vol. 37, no. 1, pp. 63–77, 2008.

[68] Y. Sun, Z. Sheng, C. Ma et al., "Combining genomic and network characteristics for extended capability in predicting synergistic drugs for cancer," *Nature Communications*, vol. 6, article 9481, 2015.

[69] D. Zhou, J. Weston, O. Bousquet, and B. Scholkopf, "Ranking on data manifolds," in *Neural Information Processing Systems*, 2004.

[70] X. Chen, B. Ren, M. Chen et al., "NLLSS: predicting synergistic drug combinations based on semi-supervised learning," *PLOS Computational Biology*, vol. 12, no. 7, Article ID e1004975, 2016.

[71] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: a geometric framework for learning from labeled and unlabeled examples," *Journal of Machine Learning Research*, vol. 7, pp. 2399–2434, 2006.

[72] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of Machine Learning*, MIT Press, Boston, Mass, USA, 2012.

[73] K.-J. Xu, F.-Y. Hu, J. Song, and X.-M. Zhao, "Exploring drug combinations in a drug-cocktail network," in *Proceedings of the 5th IEEE International Conference on Systems Biology (ISB '11)*, pp. 382–387, Zhuhai, China, September 2011.

[74] P. Li, C. Huang, Y. Fu et al., "Large-scale exploration and analysis of drug combinations," *Bioinformatics*, vol. 31, no. 12, pp. 2007–2016, 2015.

[75] R. Jansen, H. Yu, D. Greenbaum et al., "A Bayesian networks approach for predicting protein-protein interactions from genomic data," *Science*, vol. 302, no. 5644, pp. 449–453, 2003.

[76] S. Chandrasekaran, M. Cokol–Cakmak, N. Sahin et al., "Chemogenomics and orthology–based design of antibiotic combination therapies," *Molecular Systems Biology*, vol. 12, no. 5, p. 872, 2016.

[77] J. Lehar, A. Krueger, W. Avery et al., "Synergistic drug combinations tend to improve therapeutically relevant selectivity," *Nature Biotechnology*, vol. 27, no. 7, p. 659, 2009.

[78] Z. Wu, Y. Wang, and L. Chen, "Network-based drug repositioning," *Molecular BioSystems*, vol. 9, no. 6, pp. 1268–1281, 2013.

[79] Y.-Y. Wang, K.-J. Xu, J. Song, and X.-M. Zhao, "Exploring drug combinations in genetic interaction network," *BMC Bioinformatics*, vol. 13, supplement 7, article S7, 2012.

[80] Y. Liu and H. Zhao, "Predicting synergistic effects between compounds through their structural similarity and effects on transcriptomes," *Bioinformatics*, 2016.

[81] W. Hassen, A. Kassambara, T. Reme et al., "Drug metabolism and clearance system in tumor cells of patients with multiple myeloma," *Oncotarget*, vol. 6, no. 8, pp. 6431–6447, 2015.

[82] R. L. Juliano and V. Ling, "A surface glycoprotein modulating drug permeability in Chinese hamster ovary cell mutants," *Biochimica et Biophysica Acta*, vol. 445, no. 1, pp. 152–162, 1976.

[83] M. M. Gottesman, "Mechanisms of cancer drug resistance," *Annual Review of Medicine*, vol. 53, pp. 615–627, 2002.

[84] H. Glavinas, P. Krajcsi, J. Cserepes, and B. Sarkadi, "The role of ABC transporters in drug resistance, metabolism and toxicity," *Current Drug Delivery*, vol. 1, no. 1, pp. 27–42, 2004.

[85] P. Borst, R. Evers, M. Kool, and J. Wijnholds, "A family of drug transporters: the multidrug resistance-associated proteins," *Journal of the National Cancer Institute*, vol. 92, no. 16, pp. 1295–1302, 2000.

[86] G. Szakács, J. K. Paterson, J. A. Ludwig, C. Booth-Genthe, and M. M. Gottesman, "Targeting multidrug resistance in cancer," *Nature Reviews Drug Discovery*, vol. 5, no. 3, pp. 219–234, 2006.

[87] M. Morishita, T. Yamagata, H. Kusuhara, K. Takayama, and Y. Sugiyama, "Use of pharmaceutical excipients in enhancing GI

absorption by inhibiting efflux transporters," in *Proceedings of the Asian Pacific Regional International Society for the Study of Xenobiotics Meeting*, pp. 119–130, 1972.

[88] V. Y. Chua, J. Harvey, and J. Bentel, "Abstract 715: regulation of the ABCG2 drug efflux transporter in breast cancer cells," *Cancer Research*, vol. 75, no. 15, supplement, pp. 715–715, 2015.

[89] G. Pan, T. Li, Q. Zeng, X. Wang, and Y. Zhu, "Alisol F 24 acetate enhances chemosensitivity and apoptosis of MCF-7/DOX Cells by inhibiting P-glycoprotein-mediated drug efflux," *Molecules*, vol. 21, no. 2, 2016.

[90] M. A. Reis, O. B. Ahmed, G. Spengler, J. Molnár, H. Lage, and M. U. Ferreira, "Jatrophane diterpenes and cancer multidrug resistance—ABCB1 efflux modulation and selective cell death induction," *Phytomedicine*, vol. 23, no. 9, pp. 968–978, 2016.

[91] L. Lin, S. W. Yee, R. B. Kim, and K. M. Giacomini, "SLC transporters as therapeutic targets: emerging opportunities," *Nature Reviews Drug Discovery*, vol. 14, no. 8, pp. 543–560, 2015.

Hindawi Publishing Corporation
BioMed Research International
Volume 2016, Article ID 4354901, 9 pages
http://dx.doi.org/10.1155/2016/4354901

Research Article

Identification of Hot Spots in Protein Structures Using Gaussian Network Model and Gaussian Naive Bayes

Hua Zhang,[1] Tao Jiang,[2] and Guogen Shan[3]

[1]*School of Computer and Information Engineering, Zhejiang Gongshang University, Hangzhou, Zhejiang 310018, China*
[2]*School of Statistics and Mathematics, Zhejiang Gongshang University, Hangzhou, Zhejiang 310018, China*
[3]*School of Community Health Sciences, University of Nevada Las Vegas, Las Vegas, NV 89154, USA*

Correspondence should be addressed to Hua Zhang; zerozhua@126.com

Received 21 August 2016; Revised 2 October 2016; Accepted 11 October 2016

Academic Editor: Guang Hu

Residue fluctuations in protein structures have been shown to be highly associated with various protein functions. Gaussian network model (GNM), a simple representative coarse-grained model, was widely adopted to reveal function-related protein dynamics. We directly utilized the high frequency modes generated by GNM and further performed Gaussian Naive Bayes (GNB) to identify hot spot residues. Two coding schemes about the feature vectors were implemented with varying distance cutoffs for GNM and sliding window sizes for GNB based on tenfold cross validations: one by using only a single high mode and the other by combining multiple modes with the highest frequency. Our proposed methods outperformed the previous work that did not directly utilize the high frequency modes generated by GNM, with regard to overall performance evaluated using $F1$ measure. Moreover, we found that inclusion of more high frequency modes for a GNB classifier can significantly improve the sensitivity. The present study provided additional valuable insights into the relation between the hot spots and the residue fluctuations.

1. Introduction

Flexibility and dynamics play key roles for proteins in implementing various biological processes and functions [1, 2]. Residue fluctuations or atomic motions, contributing to large-scale conformational changes of protein structures, are shown to be closely related to functions of native proteins [3–5].

Two methods, molecular dynamic (MD) simulation and normal mode analysis (NMA), are widely used to investigate the dynamic link between protein structures and functions. The main drawback of MD simulations is their computational cost [6, 7]. Coarse-grained NMA, such as elastic network model (ENM) [7], has been increasingly used in recent years as a powerful tool to elucidate the structure-encoded dynamics of biomolecules [2]. The ENMs, including the isotropic Gaussian network model (GNM) [8, 9] and the anisotropic network model [10], define spring-like interactions between residues that are within a certain cutoff distance. They simplify the computationally costly all-atom potentials into a quadratic function in the vicinity of the native state, which allows the decomposition of the motions into vibrational modes with different frequencies that are often known as normal modes. Being simple and efficient, ENM and GNM have been validated in numerous applications that resulted in reasonable agreement with a wealth of experimental data, including prediction of X-ray crystallographic B-factors for amino acids [9, 11], identifications of hot spots [12–14], catalytic sites [15], core amino acids stabilizing rhodopsin [16] and important residues of HLA proteins [17], elucidation of the molecular mechanisms of motor-protein motions [18], and general conformational changes and functions [3, 4, 19–31].

Previous studies have shown in many cases that the normal modes including the high frequency (fast) modes and the low frequency (slow) modes by the GNM are very useful for recognizing several specific types of protein functions. In particular, the highest frequency modes that reflect local events at the residue level can be utilized to identify core residues or binding sites [16, 17, 20, 32], while the lowest frequency modes are usually responsible for the collective functional dynamics of the global protein motions [23, 33].

In area of protein-protein interaction, several studies such as Ozbek et al. [12], Haliloglu et al. [13], and Demirel et al. [14] utilized GNM to identify hot spots that are defined as the residues contributing more than 2 kcal/mol to the binding energy. Their results suggested that hot spots are predefined in the dynamics of protein structures and forming the binding core of interfaces. However, the mean square distance fluctuations of residue pairs and the mean square fluctuations of residues calculated from the highest frequency modes by GNM, rather than the direct usage of the highest frequency modes themselves, were applied to detect the hot spots in the work by Ozbek et al. [12] and by Haliloglu et al. [13] and Demirel et al. [14], respectively.

In addition, several computational methods by utilizing machine learning tools have been developed to predict hot spots from protein sequences and structures [34–37]. The advantage of learning methods is the ability to result in higher quality by sufficiently integrating the extracted feature information from protein structures. In this paper, we follow the work by Ozbek et al. [12] but focus on the direct usage of the highest frequency modes to investigate the relation between the residue fluctuations and the hot spots. The top 20 highest frequency modes by GNM were used as an original feature set inputted into Gaussian Naive Bayes (GNB), as a representative of learning methods, to identify hot spots. The main purpose of this study is to examine whether the raw fast modes can be directly used to differentiate hot spots or non-hot spots and whether the utilization of learning methods can improve the identification quality of hot spots for unbound protein structures.

2. Material and Methods

2.1. Dataset. We used the dataset that was collected by Ozbek et al. [12]. This set was filtered with PISCES culling server [38] at the sequence identity of 25% and was originally composed of 33 unbound protein structures. We had to remove one protein with ID 1lrp from the dataset since its structure cannot be currently found in Protein Data Bank (PDB) [39]. Therefore, the final dataset had 32 unbound protein structures with a total of 4270 residues of which 171 are hot spot residues. The dataset including the detailed information about hot spot residues can be derived from Ozbek et al. [12].

2.2. Gaussian Network Model and Its Applications to Identification of the Hot Spots. GNM describes each protein as an elastic network, where the springs connecting the nodes represent the bonded and nonbonded interactions between the pairs of residues located within a cutoff distance R_C [8, 9]. Assuming that the springs are harmonic and the residue fluctuations are isotropic and Gaussian, the network potential of N nodes (residues) in a protein structure is

$$V_{\mathrm{GNM}} = \frac{\gamma}{2} \sum_{i,j}^{N} \Gamma_{ij} \left(\mathbf{R}_{ij} - \mathbf{R}_{ij}^{0} \right)^2, \tag{1}$$

where $\mathbf{R}_{ij}$ and $\mathbf{R}_{ij}^{0}$ are instantaneous and original distance vectors between residues i and j, respectively, γ is the force constant assumed to be uniform for all network springs, and $\mathbf{\Gamma} = (\Gamma_{ij})$ is the Kirchhoff connectivity matrix defined as

$$\Gamma_{ij} = \begin{cases} -1, & \text{if } i \neq j \text{ and } R_{ij}^{0} \leq R_C \\ 0, & \text{if } i \neq j \text{ and } R_{ij}^{0} \geq R_C \\ -\sum_{j:j\neq i} \Gamma_{ij}, & \text{if } i = j, \end{cases} \tag{2}$$

where R_{ij}^{0} is the distance between residues i and j and R_C is given as a cutoff. Then, the mean correlation between residue fluctuations is calculated as

$$\begin{aligned} \left\langle \Delta\mathbf{R}_i \cdot \Delta\mathbf{R}_j \right\rangle &= \left(\frac{3k_BT}{\gamma} \right) \left[\Gamma^{-1} \right]_{ij} \\ &= \left(\frac{3k_BT}{\gamma} \right) \left[\mathbf{U}\mathbf{\Lambda}^{-1}\mathbf{U}^{T} \right]_{ij}, \end{aligned} \tag{3}$$

where $\mathbf{U}$ is the orthogonal matrix of eigenvectors ($\mathbf{u}_i$), $\mathbf{\Lambda}$ is the diagonal matrix of eigenvalues (λ_i), k_B is the Boltzmann constant, and T is the absolute temperature.

To identify hot spot residues, Ozbek et al. [12] used the mean square distance fluctuations (MSDF), $\langle \Delta\mathbf{R}_{ij}^2 \rangle$, of residues i and j given as

$$\begin{aligned} \left\langle \Delta\mathbf{R}_{ij}^2 \right\rangle &= \left\langle \left(\Delta\mathbf{R}_i - \Delta\mathbf{R}_j \right)^2 \right\rangle \\ &= \left\langle \Delta\mathbf{R}_i^2 \right\rangle + \left\langle \Delta\mathbf{R}_j^2 \right\rangle - 2 \left\langle \Delta\mathbf{R}_i \cdot \Delta\mathbf{R}_j \right\rangle, \end{aligned} \tag{4}$$

which were calculated using high frequency modes of GNM based on a cutoff of 6.5 Å. The residues with relatively high MSDF value were considered functionally probable; see more details in Ozbek et al. [12].

In addition, both Haliloglu et al. [13] and Demirel et al. [14] similarly defined mean square fluctuation (or vibration) (MSF) of residues in the weighted average of several high frequency modes based on a cutoff of 7.0 Å, to identify the hot spot residues. The MSF of residue i weighed by a subset of modes $k_1 \leq k \leq k_2$ is given as

$$\left\langle \Delta\mathbf{R}_i^2 \right\rangle_{k_1-k_2} = \frac{(3k_BT/\gamma) \sum_{k=k_1}^{k_2} \lambda_k^{-1} \left[u_k \right]_i^2}{\sum_{k=k_1}^{k_2} \lambda_k^{-1}}. \tag{5}$$

Then, one residue was predicted as a hot spot if the normalized MSF of the residue (i.e., the measure expressed in (5) divided by $3k_BT/\gamma$) is larger than a given threshold. The main difference between the work by Haliloglu et al. [13] and that by Demirel et al. [14] is the different thresholds adopted. Haliloglu et al. [13] used a constant threshold of 0.005 while it was $6N^{-1}$ given by Demirel et al. [14] where N is the number of residues in a protein sequence.

2.3. Gaussian Naive Bayes. A Naive Bayes (NB) classifier calculates the probability of a given instance (example) belonging to a certain class [40]. Given an instance X described by its feature vector $(x_1, \ldots, x_n)$ and a class target y, the conditional probability $P(y \mid X)$ can be expressed as

a product of simpler probabilities using the Naive independence assumption according to Bayes' theorem:

$$P(y \mid X) = \frac{P(y)\,P(X \mid y)}{P(X)} = \frac{P(y)\prod_{i=1}^{n} P(x_i \mid y)}{P(X)}. \quad (6)$$

Here, the target y may have two values where $y = 1$ means a hot spot residue and $y = 0$ represents non-hot spot residue. X for one residue (one instance) is a feature vector with the same size for describing its characteristic using high frequency modes generated by GNM. For example, X is equal to a vector composed of ith component $\mathbf{u}_{ki}$ for ith residue in a sequence when only one high frequency mode $\mathbf{u}_k$ is used. If three high frequency modes, denoted by $\mathbf{u}_1$, $\mathbf{u}_2$, and $\mathbf{u}_3$, are taken into account, the vector X will be $(\mathbf{u}_{1i}, \mathbf{u}_{2i}, \mathbf{u}_{3i})$ for residue i in a protein sequence. Moreover, if a window size of 3 with respect to the residue i is adopted, X becomes $(\mathbf{u}_{1i-1}, \mathbf{u}_{1i}, \mathbf{u}_{1i+1}, \mathbf{u}_{2i-1}, \mathbf{u}_{2i}, \mathbf{u}_{2i+1}, \mathbf{u}_{3i-1}, \mathbf{u}_{3i}, \mathbf{u}_{3i+1})$.

Since $P(X)$ is constant for a given instance, the following rule is adopted to classify the instance whose class is unknown:

$$\widehat{y} = \arg\max_{y} P(y) \prod_{i=1}^{n} P(x_i \mid y), \quad (7)$$

where "arg" means a value of y so that the above expression is maximized; that is, if $P(y = 1)\prod_i P(x_i \mid y = 1)$ is larger than $P(y = 0)\prod_i P(x_i \mid y = 0)$, $\widehat{y} = 1$; otherwise, $\widehat{y} = 0$.

Moreover, when the likelihood of the features (i.e., $P(x_i \mid y)$) is assumed to be Gaussian, a NB classifier is called Gaussian Naive Bayes (GNB). Due to its simplicity and being computationally fast compared to other more sophisticated methods, GNB has been widely applied to prediction problems in bioinformatics [41, 42]. In this study, GNB was mainly used to train the models by inputting the highest frequency modes to identify hot spot residues.

2.4. Performance Evaluation. In a classification task, the following quality indices, including sensitivity (also known as recall), specificity, precision, and the overall accuracy, were generally used to assess prediction performance:

$$\begin{aligned} \text{Sensitivity: sen} &= \frac{\text{TP}}{\text{TP} + \text{FN}}, \\ \text{Specificity: spe} &= \frac{\text{TN}}{\text{TN} + \text{FP}}, \\ \text{Precision: pre} &= \frac{\text{TP}}{\text{TP} + \text{FP}}, \\ \text{Accuracy: acc} &= \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \end{aligned} \quad (8)$$

where true positives (TP) and true negatives (TN) correspond to correctly predicted hot spot residues and non-hot spot residues, respectively, false positives (FP) denote non-hot spot residues predicted as hot spot residues, and false negatives (FN) denote hot spot residues predicted as non-hot spot residues.

Obviously, the dataset used in this study is extremely unbalanced with a very high proportion of non-hot spot residues. For this reason, the accuracy value is not a good choice to evaluate the overall performance of results. When a dataset includes 95% negative samples but 5% positive samples, a classifier may identify all of them as negative, resulting in 95% overall accuracy and 100% specificity. This is really shown as excellent performance, but it fails to identify the positive samples that we actually need pay close attention to. Moreover, two indices, sensitivity and precision, can both measure the classification correctness for positive samples. It is strongly expected that these two indices can synchronously reach high values, but there exists a trade-off between them in general. Therefore, we used $F1$ measure to evaluate the overall prediction performance:

$$F1 \text{ measure: } F1 = \frac{2 \times \text{sen} \times \text{pre}}{\text{sen} + \text{pre}}, \quad (9)$$

which can balance the sensitivity and the precision in case of the unbalanced dataset. The formula of the $F1$ measure can be changed to be $F1 = 2/((1/\text{sen}) + (1/\text{pre}))$ when both sen and pre are exactly larger than zero. Thus, $F1$ measure can be viewed as an increasing function of sen and pre. The minimum of $F1$ is 0 when sen = 0 or pre = 0, and the maximum of $F1$ is 1 when sen = 1 and pre = 1.

2.5. Identification of Hot Spots Using GNM and GNB. The experimental performance on identification of hot spot residues is tested using n-fold cross validation (nCV) on the dataset composed of 32 unbound protein structures. In the nCV procedure, chains are randomly divided into n subsets with the same numbers of sequences, and the test is repeated n times. In each time, the $n - 1$ subsets are used to build the model, and the remaining one subset is then tested by the prediction model.

In the present study, we performed tenfold cross validation (10CV) based on Gaussian Naive Bayes using the highest modes as features from GNM outputs in different ways. Then, we mainly implemented two schemes concerning feature coding for investigating the relations between the highest modes and the hot spot residues. Firstly, a classifier was modeled by directly using single one of the top 20 high frequency modes (i.e., the eigenvectors ($\mathbf{u}_i$) that correspond to the top 20 largest eigenvalues (λ_i)). Meanwhile, a sliding window of the central residue with sizes ranging from 1 to 21 was utilized to examine the impact of the neighboring residues' fluctuations, and the computation of GNM was carried out by usage of multiple distance cutoffs ranging from 6.0 to 8.0 with a step size of 0.1. Secondly, we combined top m modes with the highest frequency ($m = 1, 2, 3, \ldots, 20$) and utilized similar scheme for the distance cutoff of GNM computation and the sliding window of the central residue to establish the models for identifying hot spot residues.

3. Results and Discussion

3.1. Identification of Hot Spot Residues Using Single One of the Highest Modes. In this work, the overall performance was evaluated by the $F1$ measure in (9), which is able to balance the sensitivity and the precision. Table 1 lists twenty

Table 1: List of top 20 $F1$ measures based on tenfold cross validations of Gaussian Naive Bayes when using single ith highest mode ($i = 1, 2, \ldots, 20$) inputted into the feature vector, where cutoff means the distance threshold for GNM computation that varies from 6.0 to 8.0 with step size of 0.1 and sw represents the size of the sliding window for the central residue that ranges from 1 to 21 with step size of 2.

Top	Cutoff	i	sw	sen	spe	pre	acc	$F1$ measure
1	7.3	8	3	0.1930	0.9436	0.1250	0.9136	**0.1517**
2	7.1	8	9	0.2515	0.9095	0.1039	0.8831	0.1470
3	7.1	8	7	0.2456	0.9119	0.1042	0.8852	0.1463
4	7.1	8	5	0.2164	0.9263	0.1091	0.8979	0.1451
5	7.1	8	3	0.1696	0.9473	0.1184	0.9162	0.1394
6	7.3	8	5	0.1871	0.9354	0.1077	0.9054	0.1368
7	8.0	3	5	0.1930	0.9310	0.1044	0.9014	0.1355
8	7.3	8	7	0.2164	0.9163	0.0974	0.8883	0.1343
9	7.1	8	13	0.2281	0.9090	0.0947	0.8817	0.1338
10	7.1	8	11	0.2281	0.9071	0.0929	0.8799	0.1320
11	7.0	19	17	0.2456	0.8963	0.0899	0.8703	0.1317
12	6.7	13	3	0.1345	0.9619	0.1285	0.9288	0.1314
13	7.8	3	3	0.1520	0.9507	0.1140	0.9187	0.1303
14	7.0	14	21	0.2339	0.901	0.0897	0.8742	0.1297
15	7.0	19	19	0.2456	0.8934	0.0877	0.8674	0.1292
16	7.1	8	15	0.2281	0.9039	0.0901	0.8768	0.1291
17	7.0	4	7	0.2281	0.9022	0.0886	0.8752	0.1277
18	6.6	6	3	0.1520	0.9480	0.1088	0.9162	0.1268
19	6.9	15	21	0.2222	0.9046	0.0886	0.8773	0.1267
20	7.2	14	13	0.2456	0.8897	0.0850	0.8639	0.1263

computational outcomes of the prediction performance that are ordered by $F1$ measure, where the feature vector for a GNB classifier was extracted from single one mode, that is, ith highest mode ($i = 1, 2, \ldots, 20$), the distance cutoff in GNM varied from 6.0 to 8.0 with the step size of 0.1, and the sliding window for one mode ranged from 1 to 21 with a step size of 2. As shown in Table 1, the highest performance was achieved by $F1$ measure of 0.1517 when the distance cutoff is 7.1 Å and the size of the sliding window is 3 in case of the 8th highest mode.

Moreover, top six $F1$ measures shown in Table 1 were from the same 8th highest mode, indicating that the best performance achieved may not belong to the first or second highest frequency mode. Even the 19th and the 13th highest modes can also result in relatively high $F1$ measures. From the aspect of cutoff, it has been shown that majority of the cutoff values shown in Table 1 are in or close to the [7.0, 7.3] interval.

Given the cutoff of 7.3 Å in GNM, we plotted sensitivity, precision, and $F1$ measure for all of the top 20 high modes; see Figure 1. Three cases with sizes of the sliding windows equal to 1, 3, and 5 were examined. It is apparent that the $F1$ measures and the sensitivity values for the majority of the 20 modes can be improved when the size of the sliding window is from 1 to 3. However, there is no sufficient evidence to prove that larger size of the sliding window can further increase the $F1$ measure. On the other hand, the majority of the sensitivity values were improved when the window size was increased from 3 to 5, but no consistent trend can be found for precision values in three cases of the window sizes.

3.2. Identification of Hot Spot Residues by Combining the Highest Modes. Furthermore, top m modes ($m = 1, 2, \ldots, 20$) with the highest frequency were combined to establish the GNB classifier and to investigate whether the prediction performance can be improved. For example, when m is taken to be 10, top ten high modes (i.e., hm1, hm2, ..., hm10) are together inputted into the feature vector of a GNB classifier. Meanwhile, the classification experiments were also performed on various cases in which the distance cutoff is from 6.0 to 8.0 with the step size of 0.1 and the size of the sliding window (sw) ranges from 1 to 21 with the step size of 2. Table 2 lists twenty outcomes of these computational experiments ordered by $F1$ measure. Among these results, the size of the sliding window is almost 1 except the case of the 10th highest $F1$ measure in which 9 high modes and the window size of 3 were used, suggesting that the fluctuation of the central residue may be sufficient to identify hot spot residues by a combination of multiple high frequency modes. Moreover, as shown in Table 2, the distance cutoff often belongs to the [7.1, 7.5] interval, and it seems that a larger m value tends to result in higher sensitivity. For instance, the sensitivity value obtained by a combination of the top 10 high modes with cutoff of 7.4 Å (i.e., the case of top 1 $F1$ measure) is 0.2924, while the sensitivity values in the cases of top 4, 6, and 7 $F1$ measures, which are achieved by the usage of the top 20, 19, and 20 high modes, respectively, are all larger than 0.41.

In Figure 2, we plotted the sensitivity, the precision, and the $F1$ measure against m modes with the highest frequency that were combined as features for five cases denoted by

Table 2: List of the top 20 $F1$ measures based on tenfold cross validations of Gaussian Naive Bayes when using m modes with the highest frequency inputted into the feature vector, where $m = \{1, 2, \ldots, 20\}$, the distance cutoff in GNM varies from 6.0 to 8.0 with step size of 0.1, and the sliding window size (sw) for multiple high modes ranges from 1 to 21 with step size of 2.

Top	Cutoff	m	sw	sen	spe	pre	acc	$F1$ measure
1	7.4	10	1	0.2924	0.8992	0.1080	0.8749	**0.1577**
2	7.4	11	1	0.3041	0.8873	0.1012	0.8639	0.1518
3	7.4	13	1	0.3275	0.8736	0.0976	0.8518	0.1503
4	7.2	20	1	0.4269	0.8207	0.0903	0.8049	0.1491
5	7.4	12	1	0.3099	0.8809	0.0980	0.8581	0.1489
6	7.2	19	1	0.4152	0.8239	0.0895	0.8075	0.1473
7	7.3	20	1	0.4152	0.8229	0.0891	0.8066	0.1467
8	7.1	11	1	0.2924	0.8870	0.0975	0.8632	0.1462
9	7.5	15	1	0.3450	0.8592	0.0928	0.8386	0.1462
10	7.1	9	3	0.3977	0.8312	0.0895	0.8138	0.1461
11	7.3	10	1	0.2690	0.8992	0.1002	0.8740	0.1460
12	7.4	15	1	0.3450	0.8585	0.0923	0.8379	0.1457
13	7.1	13	1	0.3158	0.8727	0.0937	0.8504	0.1446
14	7.5	14	1	0.3275	0.8663	0.0927	0.8447	0.1445
15	7.5	16	1	0.3509	0.8529	0.0905	0.8328	0.1439
16	7.4	14	1	0.3275	0.8653	0.0921	0.8438	0.1438
17	7.6	15	1	0.3333	0.8622	0.0916	0.8410	0.1438
18	7.5	10	1	0.2632	0.9000	0.0989	0.8745	0.1438
19	7.3	9	1	0.2456	0.9090	0.1012	0.8824	0.1433
20	7.1	14	1	0.3275	0.8641	0.0914	0.8426	0.1429

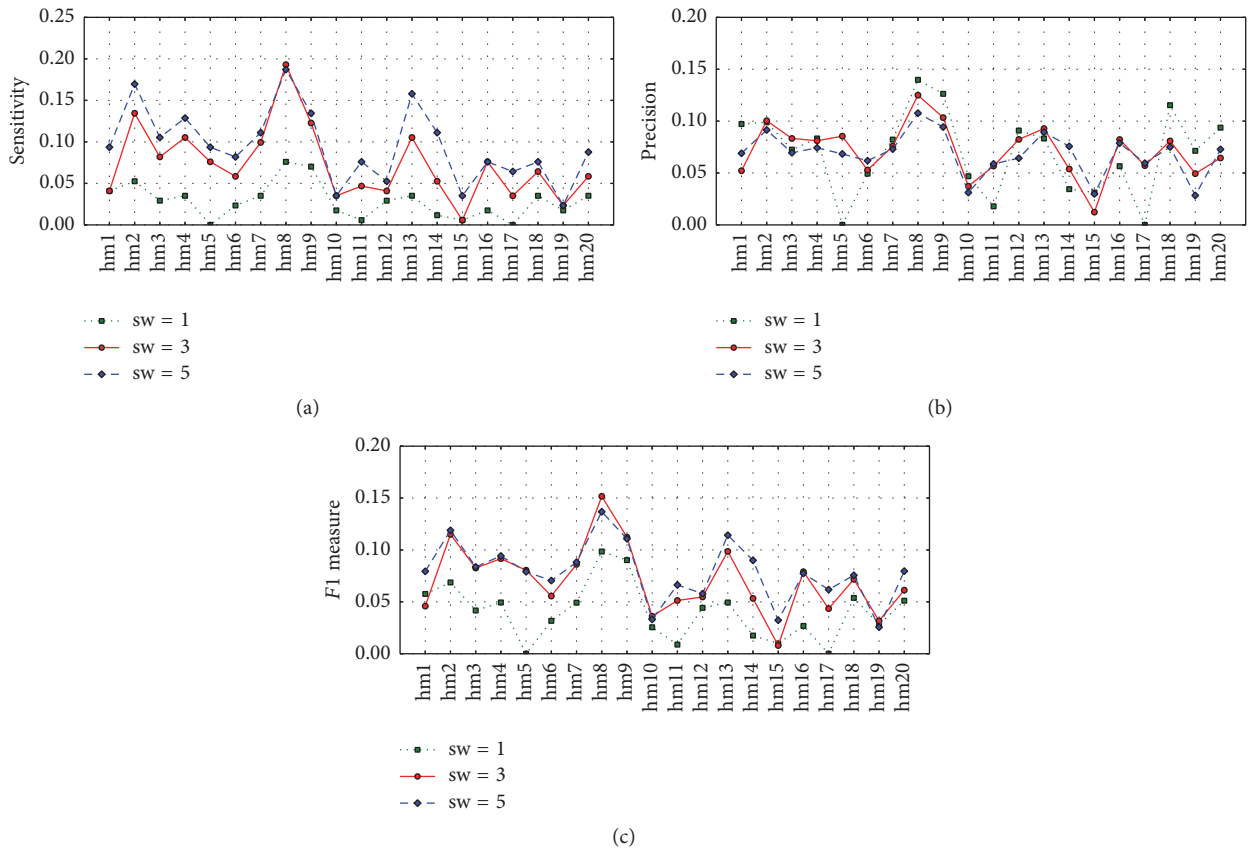


Figure 1: Plots of sensitivity (a), precision (b), and $F1$ values by the single ith highest mode ($i = 1, 2, \ldots, 20$) in three cases of the sliding window sizes (sw) (i.e., sw = 1, 3, 5) for GNB classifiers. The ith highest mode in the figure is denoted as hm$i$.

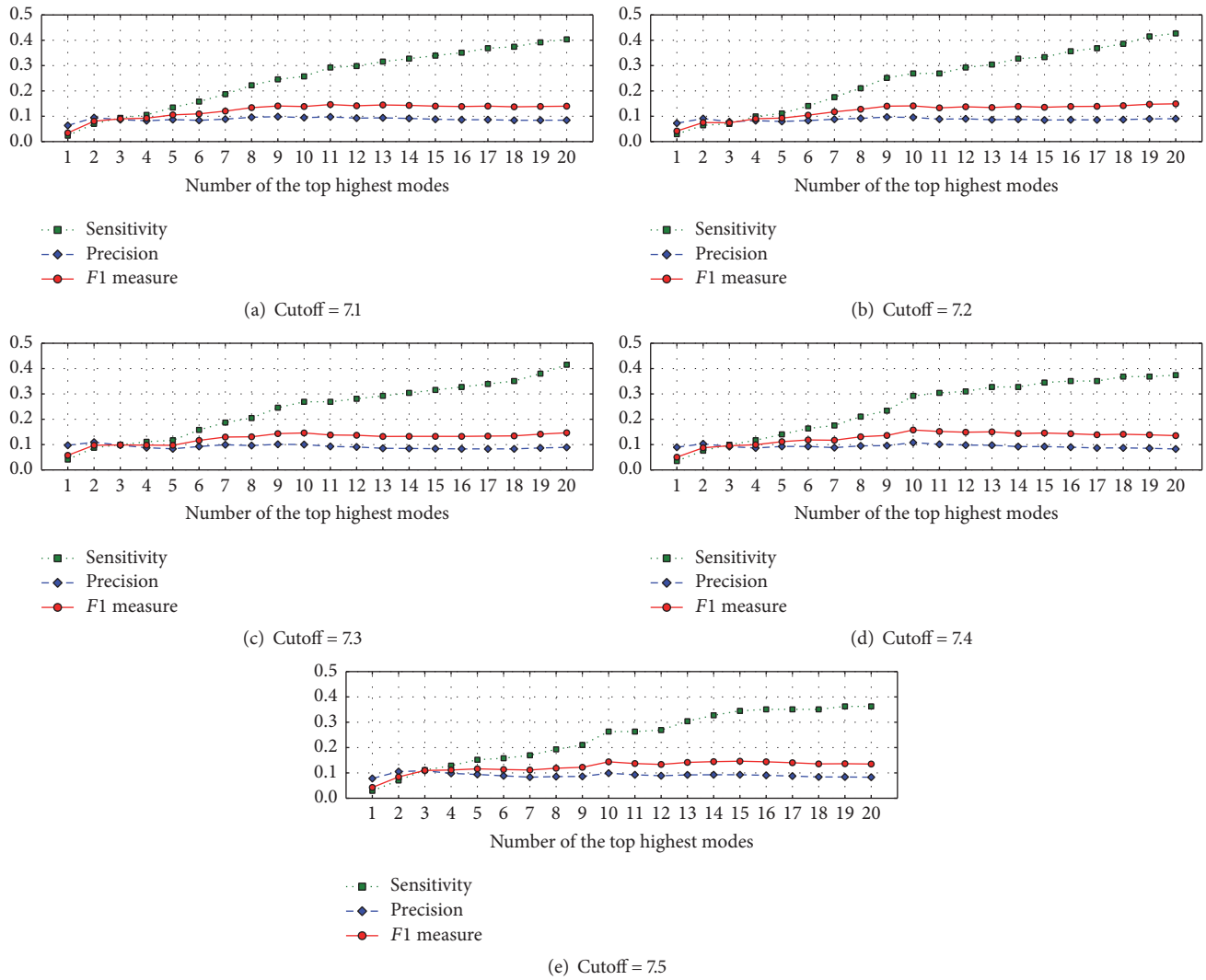


FIGURE 2: Plots of sensitivity, precision, and $F1$ values against m modes with the highest frequency in five cases denoted by the distance cutoffs of 7.1 Å (a), 7.2 Å (b), 7.3 Å (c), 7.4 Å (d), and 7.5 Å (e), respectively, for GNB classifiers.

the distance cutoffs of 7.1 Å, 7.2 Å, 7.3 Å, 7.4 Å, and 7.5 Å, respectively, where the sizes of the sliding window for all cases are 1. It can be seen from the figure that these three indices are consistently improved with the number of top high modes used up to 10. Especially for the case of sensitivity, its value is an increasing function of the number of modes with the highest frequency. It can be concluded that inclusion of more high frequency modes can improve the sensitivity, but the precision values become slightly decreased by adding more high frequency modes when the number of high modes combined is larger than 10. In the meantime, the $F1$ measure tends to be no longer enhanced.

3.3. Performance Comparison with Existing Methods. In the present work, we directly inputted the high frequency modes to a GNB classifier for predicting hot spots when compared with the existing methods proposed by Ozbek et al. [12], Haliloglu et al. [13], and Demirel et al. [14]. Ozbek et al. [12] utilized the mean square distance fluctuations of residue pairs, which were computed at most based on five top high frequency modes, to identify hot spot residues. It may be not appropriate to directly compare our work with the results obtained by Ozbek et al. [12], since the datasets used and the test procedures are both slightly different. However, we reported here again part of outcomes from Table 1 in Ozbek et al. [12] for a comparison. The $F1$ measures were calculated on the reported sensitivity and precision values, as shown in Table 3. In addition, no results concerning the prediction quality of hot spot residues based on a nonredundant dataset were reported in Haliloglu et al. [13] and Demirel et al. [14], where only MSF profiles for a couple of protein cases were depicted and shown as figures. The usage of the number of high frequency modes is not consistent that three, four, or five fast modes may be adopted for different cases. Due to the lack of details and web servers, we here simulated their methods on the dataset in this work by computing the normalized MSF values weighted by one up to five high frequency modes using a cutoff of 7 Å for GNM. A constant 0.005 and a varied value

Table 3: Performance comparison of the proposed models with the work by Ozbek et al. [12] and the simulated methods proposed by Haliloglu et al. [13] and Demirel et al. [14], where hm1–*i* means that a total of *i* high frequency modes (hm1, hm2, ..., hm*i*) are used together.

Reference	GNM modes	Cutoff	sw	sen	spe	pre	acc	*F*1
Ozbek et al. [12]	hm1	6.5 Å	1	0.14	0.89	0.05	0.86	0.0737
	hm2			0.16	0.80	0.05	0.85	0.0762
	hm3			0.24	0.88	0.07	0.85	0.1084
	hm1–3			0.25	0.86	0.07	0.83	0.1094
	hm1–5			0.29	0.84	0.07	0.81	0.1128
Haliloglu et al. [13] (simulated)	hm1	7.0 Å	1	0.1988	0.9019	0.0780	0.8738	0.1120
	hm1-2			0.2690	0.8819	0.0868	0.8574	0.1312
	hm1–3			0.3041	0.8580	0.0820	0.8358	0.1292
	hm1–4			0.3275	0.8429	0.0800	0.8222	0.1286
	hm1–5			0.3450	0.8339	0.0797	0.8143	0.1295
Demirel et al. [14] (simulated)	hm1	7.0 Å	1	0.0468	**0.9773**	0.0792	**0.9400**	0.0588
	hm1-2			0.0526	0.9697	0.0677	0.9330	0.0592
	hm1–3			0.0409	0.9615	0.0424	0.9246	0.0417
	hm1–4			0.0819	0.9573	0.0741	0.9222	0.0778
	hm1–5			0.0936	0.9532	0.0769	0.9187	0.0844
This work	hm8	7.3 Å	3	0.1930	0.9436	**0.1250**	0.9136	0.1517
	hm8	7.1 Å	9	0.2515	0.9095	0.1039	0.8831	0.1470
	hm1–10	7.4 Å	1	0.2924	0.8992	0.1080	0.8749	**0.1577**
	hm1–11	7.4 Å	1	0.3041	0.8873	0.1012	0.8639	0.1518
	hm1–13	7.4 Å	1	0.3275	0.8736	0.0976	0.8518	0.1503
	hm1–20	7.2 Å	1	**0.4269**	0.8207	0.0903	0.8049	0.1491

$6N^{-1}$ with respect to the sequence length N were used to identify hot spot residues for the simulations of the methods by Haliloglu et al. [13] and Demirel et al. [14], respectively. The quality indices including sensitivity, specificity, precision, accuracy, and $F1$ measure for these simulations were then reported in Table 3. We also listed part of the best outcomes from this study in Table 3, two using single high mode and four by a combination of multiple high modes, which have been shown in Tables 1 and 2.

On the whole, if evaluated by $F1$ measure or precision, all of the cases in Table 3 by this work outperformed the results by Ozbek et al. [12] and by the simulated methods of Haliloglu et al. [13] and Demirel et al. [14]. This suggests that the direct usage of the high frequency modes is efficient to identify hot spot residues. Besides, the improvement on $F1$ measure by combining multiple high frequency modes seems to be very slight when compared with the methods only using single high mode, while the sensitivity values in general tend to be improved a lot. This is in good agreement with the work by Ozbek et al. [12] and the simulation results of Haliloglu et al. [13] and Demirel et al. [14] as outlined in Table 3. Additionally, the specificity and accuracy values of the simulated method for Demirel et al. [14] are higher than those of other methods, but on the contrary the values of sensitivity, precision, and $F1$ measure are in general lower. The reason causing worse quality on $F1$ measure achieved by the simulation of Demirel et al. [14] is due to a larger threshold used when compared with the simulated method of Haliloglu et al. [13].

In addition, we also performed computational experiments using several common classifiers, including logistic regression, decision tree, k-nearest neighbor, and support vector machine with default parameters, instead of GNB, where all of the machine learning methods were implemented in scikit-learn [43]. As a consequence, the results (data not shown in this paper) showed that GNB exhibited better performance than other classifiers. This is the reason why we finally adopted GNB as the base classifier for identification of the hot spot residues.

4. Conclusion

In this study, we followed previous work [12–14] focusing on the identifications of hot spots by using GNM but directly used the high frequency modes and further performed GNB classifier. The proposed methods outperformed the outcomes reported in Ozbek et al. [12] and the simulated results of Haliloglu et al. [13] and Demirel et al. [14] based on $F1$ measure to evaluate the overall performance. The results by this work suggested that the high frequency modes can be directly used to identify hot spot residues with reasonable performance. In case of the scheme using only single high frequency mode, the largest $F1$ measure may not be necessarily achieved by one of the top five high frequency modes. In our study, it was surprisingly gained by the 8th highest mode with the distance cutoff of 7.3 and the window size of 3. We further included more modes from total number of 20 high frequency modes when compared with the work

by Ozbek et al. [12] in which at most five frequency modes are used. Of particular interest is the fact that inclusion of more high frequency modes can significantly improve the sensitivity value, but not the $F1$ measure and the precision in general.

The dataset used in this work is obviously unbalanced. There is a trade-off between the sensitivity and the precision. It is not easy for researchers to find a perfect way to determine the proper performance index to evaluate experimental results. Therefore, we finally reported multiple results as listed in Tables 1, 2, and 3, which were considered for choices associated with different purposes in practice. Overall, the present study provided additional valuable insight into the relation between hot spots and residue fluctuations.

Competing Interests

The authors declare that they have no competing interests.

Authors' Contributions

Hua Zhang performed the experiment. Hua Zhang and Guogen Shan designed the research; Hua Zhang and Tao Jiang carried out data analysis. Hua Zhang and Guogen Shan wrote the paper.

Acknowledgments

Hua Zhang was supported by the National Natural Science Foundation of China (Grant nos. 61672459 and 61170099) and the Zhejiang Provincial Natural Science Foundation of China (Grant no. LY15F020001), and Guogen Shan was supported by National Institutes of Health (Grant nos. 5U54GM104944 and P20GM103440).

References

[1] I. Bahar, T. R. Lezon, A. Bakan, and I. H. Shrivastava, "Normal mode analysis of biomolecular structures: functional mechanisms of membrane proteins," *Chemical Reviews*, vol. 110, no. 3, pp. 1463–1497, 2010.

[2] I. Bahar and A. J. Rader, "Coarse-grained normal mode analysis in structural biology," *Current Opinion in Structural Biology*, vol. 15, no. 5, pp. 586–592, 2005.

[3] A. Bakan and I. Bahar, "The intrinsic dynamics of enzymes plays a dominant role in determining the structural changes induced upon inhibitor binding," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 34, pp. 14349–14354, 2009.

[4] T. Haliloglu and B. Erman, "Analysis of correlations between energy and residue fluctuations in native proteins and determination of specific sites for binding," *Physical Review Letters*, vol. 102, no. 8, Article ID 088103, 2009.

[5] S. E. Dobbins, V. I. Lesk, and M. J. E. Sternberg, "Insights into protein flexibility: the relationship between normal modes and conformational change upon protein-protein docking," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 30, pp. 10390–10395, 2008.

[6] M. Rueda, C. Ferrer-Costa, T. Meyer et al., "A consensus view of protein dynamics," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 3, pp. 796–801, 2007.

[7] L. Yang, G. Song, and R. L. Jernigan, "How well can we understand large-scale protein motions using normal modes of elastic network models?" *Biophysical Journal*, vol. 93, no. 3, pp. 920–929, 2007.

[8] I. Bahar, A. R. Atilgan, and B. Erman, "Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential," *Folding and Design*, vol. 2, no. 3, pp. 173–181, 1997.

[9] S. Kundu, J. S. Melton, D. C. Sorensen, and G. N. Phillips Jr., "Dynamics of proteins in crystals: comparison of experiment with simple models," *Biophysical Journal*, vol. 83, no. 2, pp. 723–732, 2002.

[10] A. R. Atilgan, S. R. Durell, R. L. Jernigan, M. C. Demirel, O. Keskin, and I. Bahar, "Anisotropy of fluctuation dynamics of proteins with an elastic network model," *Biophysical Journal*, vol. 80, no. 1, pp. 505–515, 2001.

[11] H. Zhang and L. Kurgan, "Sequence-based Gaussian network model for protein dynamics," *Bioinformatics*, vol. 30, no. 4, pp. 497–505, 2014.

[12] P. Ozbek, S. Soner, and T. Haliloglu, "Hot spots in a network of functional sites," *PloS ONE*, vol. 8, no. 9, Article ID e74320, 2013.

[13] T. Haliloglu, O. Keskin, B. Ma, and R. Nussinov, "How similar are protein folding and protein binding nuclei? Examination of vibrational motions of energy hot spots and conserved residues," *Biophysical Journal*, vol. 88, no. 3, pp. 1552–1559, 2005.

[14] M. C. Demirel, A. R. Atilgan, I. Bahar, R. L. Jernigan, and B. Erman, "Identification of kinetically hot residues in proteins," *Protein Science*, vol. 7, no. 12, pp. 2522–2532, 1998.

[15] L.-W. Yang and I. Bahar, "Coupling between catalytic site and collective dynamics: a requirement for mechanochemical activity of enzymes," *Structure*, vol. 13, no. 6, pp. 893–904, 2005.

[16] A. J. Rader, G. Anderson, B. Isin, H. G. Khorana, I. Bahar, and J. Klein-Seetharaman, "Identification of core amino acids stabilizing rhodopsin," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 19, pp. 7246–7251, 2004.

[17] T. Haliloglu, A. Gul, and B. Erman, "Predicting important residues and interaction pathways in proteins using Gaussian network model: binding and stability of HLA proteins," *PLoS Computational Biology*, vol. 6, no. 7, Article ID e1000845, 2010.

[18] W. Zheng and S. Doniach, "A comparative study of motor-protein motions by using a simple elastic-network model," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 23, pp. 13253–13258, 2003.

[19] W. Zheng and B. R. Brooks, "Normal-modes-based prediction of protein conformational changes guided by distance constraints," *Biophysical Journal*, vol. 88, no. 5, pp. 3109–3117, 2005.

[20] T. Haliloglu, E. Seyrek, and B. Erman, "Prediction of binding sites in receptor-ligand complexes with the Gaussian network model," *Physical Review Letters*, vol. 100, no. 22, Article ID 228102, 2008.

[21] F. Zhu and G. Hummer, "Pore opening and closing of a pentameric ligand-gated ion channel," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, no. 46, pp. 19814–19819, 2010.

[22] O. Kurkcuoglu and P. A. Bates, "Mechanism of cohesin loading onto chromosomes: a conformational dynamics study," *Biophysical Journal*, vol. 99, no. 4, pp. 1212–1220, 2010.

[23] J. Jiang, I. H. Shrivastava, S. D. Watts, I. Bahar, and S. G. Amara, "Large collective motions regulate the functional properties of glutamate transporter trimers," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 108, no. 37, pp. 15141–15146, 2011.

[24] E. Marcos, R. Crehuet, and I. Bahar, "Changes in dynamics upon oligomerization regulate substrate binding and allostery in amino acid kinase family members," *PLoS Computational Biology*, vol. 7, no. 9, Article ID e1002201, 2011.

[25] C. Tuzmen and B. Erman, "Identification of ligand binding sites of proteins using the gaussian network model," *PLoS ONE*, vol. 6, no. 1, article e16474, 2011.

[26] A. Zhuravleva, D. M. Korzhnev, S. B. Nolde et al., "Propagation of dynamic changes in barnase upon binding of barstar: an NMR and computational study," *Journal of Molecular Biology*, vol. 367, no. 4, pp. 1079–1092, 2007.

[27] S. A. Wieninger, E. H. Serpersu, and G. M. Ullmann, "ATP binding enables broad antibiotic selectivity of aminoglycoside phosphotransferase(3′)-IIIa: an elastic network analysis," *Journal of Molecular Biology*, vol. 409, no. 3, pp. 450–465, 2011.

[28] A. Srivastava and R. Granek, "Cooperativity in thermal and force-induced protein unfolding: integration of crack propagation and network elasticity models," *Physical Review Letters*, vol. 110, no. 13, Article ID 138101, 2013.

[29] L. Yang, G. Song, A. Carriquiry, and R. L. Jernigan, "Close correspondence between the motions from principal component analysis of multiple HIV-1 protease structures and elastic network modes," *Structure*, vol. 16, no. 2, pp. 321–330, 2008.

[30] A. Szarecka, Y. Xu, and P. Tang, "Dynamics of firefly luciferase inhibition by general anesthetics: gaussian and anisotropic network analyses," *Biophysical Journal*, vol. 93, no. 6, pp. 1895–1905, 2007.

[31] L.-W. Yang, E. Eyal, C. Chennubhotla, J. Jee, A. M. Gronenborn, and I. Bahar, "Insights into equilibrium dynamics of proteins from comparison of NMR and X-ray data with computational predictions," *Structure*, vol. 15, no. 6, pp. 741–749, 2007.

[32] P. Ozbek, S. Soner, B. Erman, and T. Haliloglu, "DNABINDPROT: fluctuation-based predictor of DNA-binding residues within a network of interacting residues," *Nucleic Acids Research*, vol. 38, no. 2, Article ID gkq396, pp. W417–W423, 2010.

[33] I. Bahar, T. R. Lezon, L.-W. Yang, and E. Eyal, "Global dynamics of proteins: bridging between structure and function," *Annual Review of Biophysics*, vol. 39, no. 1, pp. 23–42, 2010.

[34] Y. Ofran and B. Rost, "Protein-protein interaction hotspots carved into sequences," *PLoS Computational Biology*, vol. 3, no. 7, article e119, 2007.

[35] E. Guney, N. Tuncbag, O. Keskin, and A. Gursoy, "HotSprint: database of computational hot spots in protein interfaces," *Nucleic Acids Research*, vol. 36, no. 1, pp. D662–D666, 2008.

[36] S. J. Darnell, D. Page, and J. C. Mitchell, "An automated decision-tree approach to predicting protein interaction hot spots," *Proteins: Structure, Function and Genetics*, vol. 68, no. 4, pp. 813–823, 2007.

[37] K.-I. Cho, D. Kim, and D. Lee, "A feature-based approach to modeling protein–protein interaction hot spots," *Nucleic Acids Research*, vol. 37, no. 8, pp. 2672–2687, 2009.

[38] G. Wang and R. L. Dunbrack Jr., "PISCES: a protein sequence culling server," *Bioinformatics*, vol. 19, no. 12, pp. 1589–1591, 2003.

[39] H. M. Berman, J. Westbrook, Z. Feng et al., "The protein data bank," *Nucleic Acids Research*, vol. 28, no. 1, pp. 235–242, 2000.

[40] T. M. Mitchell, *Machine Learning*, McGraw-Hill, New York, NY, USA, 1997.

[41] J. Cao, R. Panetta, S. Yue, A. Steyaert, M. Young-Bellido, and S. Ahmad, "A naive Bayes model to predict coupling between seven transmembrane domain receptors and G-proteins," *Bioinformatics*, vol. 19, no. 2, pp. 234–240, 2003.

[42] Y. Murakami and K. Mizuguchi, "Applying the Naïve Bayes classifier with kernel density estimation to the prediction of protein-protein interaction sites," *Bioinformatics*, vol. 26, no. 15, pp. 1841–1848, 2010.

[43] F. Pedregosa, G. Varoquaux, A. Gramfort et al., "Scikit-learn: machine learning in Python," *Journal of Machine Learning Research (JMLR)*, vol. 12, pp. 2825–2830, 2011.

Hindawi Publishing Corporation
BioMed Research International
Volume 2016, Article ID 7086390, 8 pages
http://dx.doi.org/10.1155/2016/7086390

Research Article

Identification of Novel Inhibitors against Coactivator Associated Arginine Methyltransferase 1 Based on Virtual Screening and Biological Assays

Fei Ye,[1,2] Weiyao Zhang,[1] Wenchao Lu,[3,4] Yiqian Xie,[3] Hao Jiang,[3,4] Jia Jin,[1] and Cheng Luo[3]

[1] *College of Life Sciences, Zhejiang Sci-Tech University, Hangzhou, China*
[2] *Key Laboratory of Plant Secondary Metabolism and Regulation of Zhejiang Province, Hangzhou, China*
[3] *Drug Discovery and Design Center, State Key Laboratory of Drug Research, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai, China*
[4] *University of Chinese Academy of Sciences, Beijing, China*

Correspondence should be addressed to Fei Ye; yefei@zstu.edu.cn and Jia Jin; aukaukauk@163.com

Received 2 August 2016; Revised 19 September 2016; Accepted 3 October 2016

Academic Editor: You-Lin Tain

Overexpression of coactivator associated arginine methyltransferase 1 (CARM1), a protein arginine N-methyltransferase (PRMT) family enzyme, is associated with various diseases including cancers. Consequently, the development of small-molecule inhibitors targeting PRMTs has significant value for both research and therapeutic purposes. In this study, together with structure-based virtual screening with biochemical assays, two compounds DC_C11 and DC_C66 were identified as novel inhibitors of CARM1. Cellular studies revealed that the two inhibitors are cell membrane permeable and effectively blocked proliferation of cancer cells including HELA, K562, and MCF7. We further predicted the binding mode of these inhibitors through molecular docking analysis, which indicated that the inhibitors competitively occupied the binding site of the substrate and destroyed the protein-protein interactions between CARM1 and its substrates. Overall, this study has shed light on the development of small-molecule CARM1 inhibitors with novel scaffolds.

1. Introduction

Arginine methylation is an important posttranslational modification catalyzed by protein arginine N-methyltransferases (PRMTs) [1, 2]. During PRMT catalysis, the methyl group of S-adenosyl-L-methionine (AdoMet, SAM) is transferred to the guanidino group of the target arginine, resulting in mono- or dimethylated arginine residues along with S-adenosyl-L-homocysteine (AdoHcy, SAH) as a coproduct [3]. There are nine PRMTs identified so far, which can be classified into three categories: type I (PRMT1, 2, 3, 4, 6, and 8), type II (PRMT5 and 9) and type III (PRM7) [4]. Type I PRMTs catalyze mono- and asymmetric dimethylation of arginine residues, whereas type II PRMTs catalyze mono- and symmetric dimethylation of arginine residues [5]. PRMT7 is the only known type III PRMT, which catalyzes monomethylation of arginine [6].

PRMT4, also known as CARM1 (coactivator associated arginine methyltransferase 1) methylates a wide variety of histone and nonhistone substrates including H3R17, H3R26 [7], SRC-3 [8], CBP/p300 [9], NCOA2 [10], PABP1 [11], and SmB [12]. Consequently, CARM1 participates in many cellular processes by impacting chromatin architecture and transcriptional initiation [9, 13], RNA processing and stability [14], and RNA splicing [12]. Overexpression of CARM1 has been observed in multiple cancer types including myelocytic leukemia [15] and breast [10], prostate [16], lung [17], and colorectal carcinomas [18], making it a potential target for anticancer therapy.

Due to essential roles of CARM1 in the regulation of cellular functions as well as tumorigenesis, discovery of CARM1 inhibitors has recently attracted much attention. To date, a number of CARM1 inhibitors have been reported [19–27]

(see Figure S1 in Supplementary Material available online at http://dx.doi.org/10.1155/2016/7086390). According to the chemical structures, these inhibitors can be divided into several categories: (i) 3,5-bis(bromohydroxybenzylidene) piperidin-4-one inhibitors (compounds 1-2 in Figure S1), (ii) pyrazole inhibitors (compounds 3–10 in Figure S1), (iii) benzo[*d*]imidazole inhibitors (compounds 11–13 in Figure S1), and (iv) other inhibitors (compounds 14-15 in Figure S1) [28]. However, the majority of these inhibitors are lacking selectivity and drug-likeness; thus turning these inhibitors into therapeutically useful compounds is challenging. Therefore, it is still of significant interest to discover selective inhibitors targeting CARM1 with good pharmacological properties.

Virtual screening is an important approach for lead-compound discovery and has been successfully used in multiple projects [29, 30]. Recently, several crystal structures of CARM1 were determined, providing a prerequisite for structure-based virtual screening [26, 31–33]. Herein, due to the convenience and low cost of this approach, docking-based virtual screening was utilized to identify novel inhibitors of CARM1 from the Specs database (http://www.specs.net/). The candidates selected by virtual screening were then tested by biochemical experiments and eventually two novel inhibitors of CARM1 were identified. Among them, the more potent inhibitor DC_C66 displayed selectivity against PRMT1, PRMT6, and PRMT5. Molecular docking was conducted to investigate the binding modes of these inhibitors and molecular basis of selectivity for CARM1. Furthermore, cellular studies revealed that both inhibitors exhibited antiproliferation activity in several CARM1-associated cancer cell lines. Overall, this study has provided chemical probes in exploring biological functions of CARM1 and information for further optimization of potent inhibitors.

2. Materials and Methods

2.1. Virtual Screening Protocol

2.1.1. Protein Preparation. The crystal structure of CARM1 in complex with indole inhibitor (PDB code 2Y1W) was used as a target for subsequent virtual screening [26]. The water molecules and ions were initially removed. The protein status was optimized through the Protein Preparation Wizard Workflow provided in the Maestro [34], with a pH value of 7.0 ± 2.0. Other parameters were set as the default. Residues within a distance of 6 Å around indole inhibitor were defined as binding pocket.

2.1.2. Ligand Database Preparation. The Specs database (http://www.specs.net/), containing ~287,000 compounds, was utilized for the virtual screening. To refine the database, we filtered it by Lipinski's rule of five [35] and removed pan-assay interference compounds (PAINS) [36–38] with Pipeline Pilot, version 7.5 (Accelrys Inc., San Diego, CA, USA) [39], yielding a database of around 180,000 small-molecule compounds. The remaining molecules were treated by LigPrep [40] to generate all stereo isomers and different protonation states with Epik.

2.1.3. Virtual Screening Protocol. The virtual screening protocol is shown in Figure 1. Firstly, the energy scoring function of DOCK4.0 was used to dock the compound library into the defined binding site. The top-ranked 10500 candidates selected by DOCK4.0 were further evaluated and ranked by the AutoDock4.0 program, leading to a list of 1500 compounds. The program Glide 5.5 [41] in XP mode [42] was run to calculate the free binding energy between these 1500 compounds and CARM1 protein. In order to ensure diversity in the candidates, the top 300 compounds from Glide 5.5 were classified to 30 groups by SciTegic functional class fingerprints (FCFP_4) in Pipeline Pilot, version 7.5 (Accelrys Inc., San Diego, CA, USA) [39], and 1–3 compounds were picked from each group. Finally, 57 compounds were selected and purchased for biological evaluation.

2.2. Similarity-Based Analog Searching. According to the results of the biological tests, we used the compound DC_C11 to run a two-dimensional similarity search through the prepared Specs database using Similarity Filter from File in Pipeline Pilot, version 7.5 (Accelrys Inc., San Diego, CA, USA). We purchased 10 compounds and tested their biological activity towards CARM1.

2.3. In Vitro CARM1 Enzyme Inhibition and Selectivity Assay. The enzymatic inhibitory activities of compounds were measured by the AlphaLISA assay provided by Shanghai ChemPartner Co., Ltd. The compounds selected from virtual screening were transferred to the assay plate (white opaque OptiPlate-384, PerkinElmer). 5 μL of enzyme solution (final concentration was 0.1 nM) or pH 8.0 tris-based assay buffer (for Min well) was added to the assay plate and then centrifuged at 1000 rpm for 1 min. Afterwards, the assay plate was incubated for 15 min at room temperature (RT). Then 5 μL of biotinylated H3 peptide/SAM mix (final concentrations were 50 nM and 300 nM, resp.) was added to the assay plate, which was covered with TopSeal-Afilm and incubated for 1 h at RT after centrifuging at 1000 rpm for 1 min (DMSO final concentration 1%). Next, 5 μL of acceptor beads (final concentration was 10 μg/mL) was added to stop the enzymatic reaction. After incubating at room temperature for 60 min, 10 μL of donor beads was added (final concentration was 10 μg/mL) in subdued light and then centrifuged at 1000 rpm for 1 min. Finally, the mixtures were incubated for 30 min at RT, and the signal was read in alpha mode using EnVision readers. The IC_{50} values were calculated by fit inhibition rates under different concentrations into GraphPad Prism 5.0 software.

2.4. Cell Viability Assay. The three cell lines, HELA, K562, and MCF7, were purchased from the American Type Culture Collection (ATCC). HELA, K562, and MCF7 were cultured in DMEM (Life Technologies) supplemented with 10% FBS. All of the cell lines were seeded into 96-well plates at an appropriate density and then treated with compounds

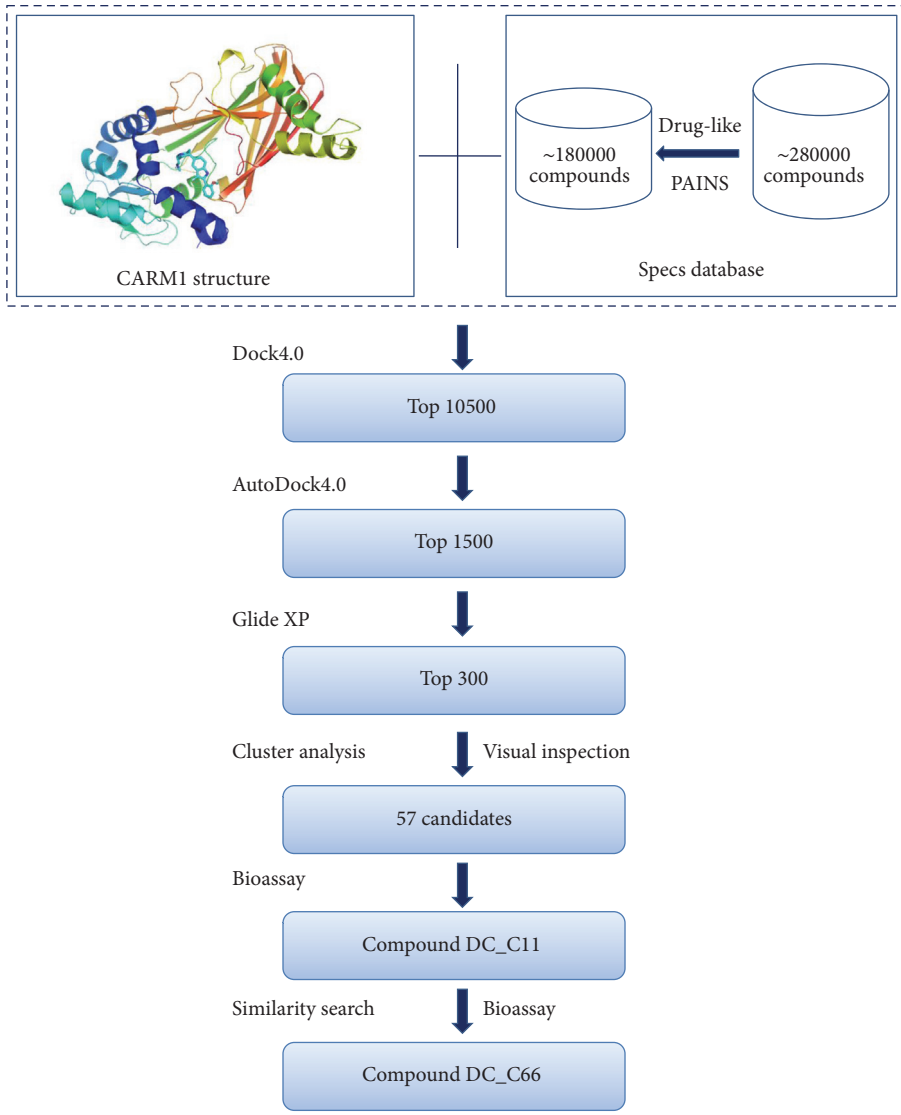


Figure 1: Flowchart of virtual screening procedures for CARM1 inhibitors.

of different concentrations or DMSO control. After 24 hrs, 48 hrs, and 72 hrs, cell viabilities were measured by the MTT assay.

2.5. Binding Energy Calculations. In order to investigate the binding mode of DC_C11 and DC_C66, molecular docking was performed using Glide 5.5 in XP mode. The generated conformations were then used for binding energy calculations by Prime MM-GBSA (Molecular Mechanics/Generalized Born Surface Area method) [43]. The binding energy was calculated as follows:

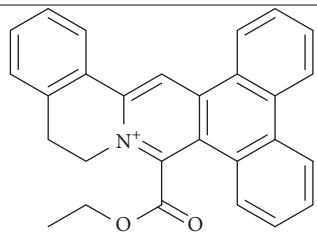
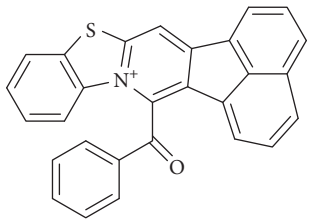
$$\Delta G = E\_\text{complex}\,(\text{minimized}) - \left(E\_\text{ligand}\,(\text{minimized}) + E\_\text{receptor}\right). \quad (1)$$

In the calculations, the protein flexibility was set to 12 Å.

3. Results and Discussion

3.1. Structure-Based Virtual Screening. In this study, docking-based virtual screening was performed to identify CARM1 inhibitors with novel scaffolds, and the flowchart is shown in Figure 1. The crystal structure of CARM1 in complex with indole inhibitor (PDB code 2Y1W) was used as a target for the following *in silico* screening [26]. Residues within a distance of 6 Å around indole inhibitor were defined as binding pocket, which contains the binding site of AdoMet and the arginine substrate. The Specs database (http://www.specs.net/), containing ~287,000 compounds, was utilized for the virtual screening. To refine the database, we filtered it by Lipinski's rule of five and removed pan-assay interference compounds (PAINS) [36–38] with Pipeline Pilot, version 7.5 (Accelrys Inc., San Diego, CA, USA) [39], yielding a database of around 180,000 small-molecule compounds, which were subsequently docked and ranked with

Table 1: Chemical structures and inhibitory activity ([a]IC_{50}, μM) of selected compounds based on virtual screening against CARM1 and several other PRMTs.

Compound ID	Specs ID	Compound structure	IC50 (μM)		
			CARM1	PRMT1	PRMT6
DC_C66	AQ-405/42300312		1.8	21	47
DC_C11	AQ-405/42300392		15	36	41

[a]All assays were conducted in duplicate.

different score functions. The top-ranked 10500 candidates selected using energy scoring function of DOCK4.0 [44] were subsequently evaluated and ranked by the AutoDock4.0 program [45], yielding a list of 1500 compounds. Then, the program Glide 5.5 (XP mode) [42] was chosen to calculate the free energy of binding between these 1500 compounds and CARM1 protein. According to the docking scores, the top-ranked 300 were clustered using Pipeline Pilot to ensure the scaffold diversity in the primary hits. The clustered molecules were cherry-picked by visual inspection based on the following considerations. (1) At least one compound was selected in each clustered group. (2) The binding modes were reasonable and molecules not occupying the SAM or substrate binding pocket were not chosen. (3) Among a group of similar molecules, compounds with lower molecular weight were preferred. Finally, 57 compounds were purchased for further biochemical validation.

3.2. Enzyme Inhibition and Selectivity Assay. All of the selected 57 candidate molecules were tested for CARM1 inhibition to determine their biochemical activities. Here, AlphaLISA assay, which is a powerful and versatile platform, was performed to test the inhibitory activities of the compounds. The enzyme solution and compounds or assay buffer were transferred to assay plates, which was incubated at RT. Then 5 μL of biotinylated H3 peptide/SAM mix was added and incubated for 1 h at RT. Afterwards, acceptor and donor beads were added sequentially. The end point was read in alpha mode using EnVision readers, and IC_{50} values were calculated in GraphPad Prism 5.0 software. Among these candidates, only one compound DC_C11 was found to be active for CARM1 inhibition, which showed an IC_{50} value of 15 μM (Table 1). We used this core structure as a hit to perform a two-dimensional similarity search through the Specs database by Pipeline Pilot, version 7.5 (Accelrys Inc., San Diego, CA, USA) [39], leading to a compound DC_C66 which displayed inhibitory better potency for CARM1 with IC_{50} values of 1.8 μM.

To investigate the selectivity of the compounds, we tested the inhibitory activities of compounds DC_C11 and DC_C66 against several selected members of type I PRMT family, including PRMT1 and PRMT6 (Table 1). It was seen that DC_C66 showed relatively weaker activity against PRMT1 and PRMT6. Moreover, DC_C66 also showed little inhibitory activity of PRMT5, a member of type II PRMT, by <50% inhibition rate at a concentration of 50 μM. Taken together, these results indicated that DC_C66 has a good selectivity for CARM1 against other selected PRMTs.

3.3. Cell-Based Activity. It has been reported that CARM1 was a potential target in many cancers; thus it is well accepted that inhibiting CARM1 could affect cancer cell proliferation. In this study, three human tumor cell lines including HELA (cervical cancer), K562 (myeloid leukemia), and MCF7 (breast cancer) were chosen to evaluate the cellular activity of the two compounds DC_C11 and DC_C66 *in vivo*. Sinefungin, a pan-PRMTs inhibitor which has the same scaffold as the cofactor SAM does, was evaluated for control experiment [46]. As shown in Figure 2, both DC_C11 and DC_C66 could inhibit proliferation of cancer cells in a time-dependent and dose-dependent manner while Sinefungin presented weaker inhibitory activity in cellular level. In the three cell lines, DC_C66 presents better antiproliferative cellular activity, which is consistent with their inhibitory activity *in vitro*. Combined with the biological data *in vitro*, we confirmed that compounds DC_C11 and DC_C66 are cell membrane permeable, which presented promising activity both *in vitro* and in cellular environment.

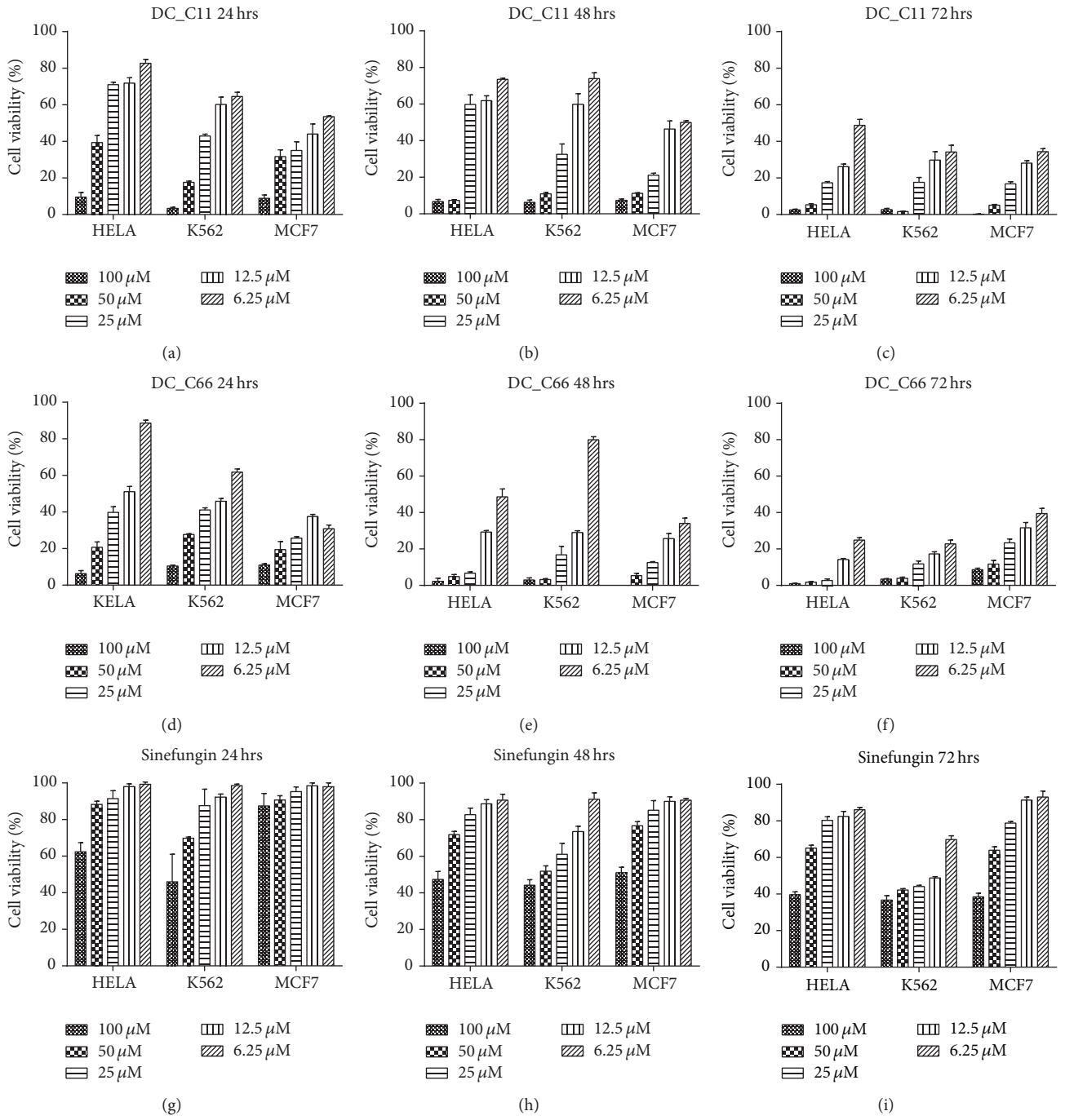


Figure 2: Antiproliferative effect of DC_C11 and DC_C66 on several cancer cell lines. (a–c) Time-dependent and dose-dependent inhibitory effect of DC_C11 on HELA, K562, and MCF7 within 24 hrs, 48 hrs, and 72 hrs, respectively. (d–f) Time-dependent and dose-dependent inhibitory effect of DC_C66 on HELA, K562, and MCF7 within 24 hrs, 48 hrs, and 72 hrs, respectively. (g–i) Time-dependent and dose-dependent inhibitory effect of Sinefungin on HELA, K562, and MCF7 within 24 hrs, 48 hrs, and 72 hrs, respectively.

3.4. Binding-Mode Analysis. To further understand the possible binding mode of DC_C11 and DC_C66 with CARM1, molecular docking study was performed with Glide in XP mode. As shown in Figure 3(a), both of DC_C11 and DC_C66 fit into the negative-charged binding pocket of substrate arginine in H3 peptide [33], implying that the compounds inhibit the activity of CARM1 by destroying the protein-protein interactions between CARM1 and substrate peptide. The phenyl ring bulks of DC_C11 and DC_C66 establish hydrophobic interactions with Y150, F153, Y154, N162, M163, and F475 in active site; the majority of these residues participate in interactions between CARM1 and its substrates (Figure 3) [33]. Besides, DC_C66 forms hydrogen bond with Y262 which probably accounts for its ability

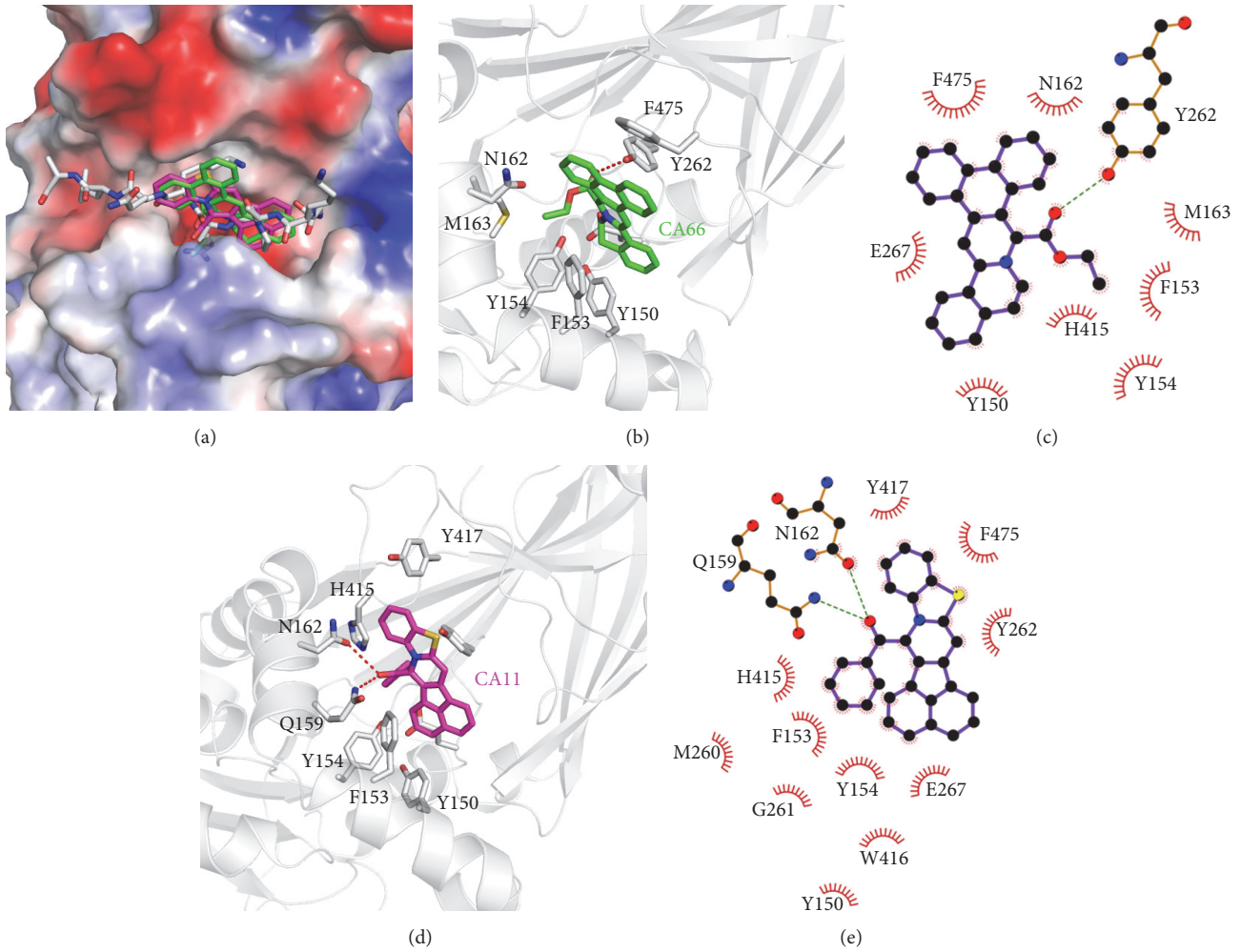


Figure 3: Predicted binding mode of DC_C11 and DC_C66 with CARM1 from docking analysis. (a) Superimposition of the binding modes of the two compounds and substrate H3 peptide (PDB ID: 5DX0). The structure of CARM1 is displayed in vacuum electrostatics. H3 peptide is shown as gray sticks, DC_C11 is shown as magenta sticks, and DC_C66 is displayed as green sticks. (b) A close view of the interactions between DC_C66 and CARM1 in the binding pocket; the key residues are shown as sticks. (c) Schematic diagram showing putative interactions between CARM1 and DC_C66. Residues involved in the hydrophobic interactions are shown as starbursts, and hydrogen-bonding interactions are denoted by dotted green lines. (d) A close view of the interactions between DC_C11 and CARM1 in the binding pocket; the key residues are shown as sticks. (e) Schematic diagram showing putative interactions between CARM1 and DC_C11.

Table 2: Binding energy for compounds DC_C66 and DC_C11.

Compound ID	DC_C66	DC_C11
Binding energy (kcal/mol)	−34.71	−26.72

to inhibit CARM1 activity (Figures 3(b) and 3(c)). Polar interactions between the oxygen in the carbonyl group of DC_C11 and side chain of Q159 as well as N162 also occur (Figures 3(d) and 3(e)). We further calculate binding energies of two compounds using Prime MM-GBSA [47] (Table 2). The results showed that DC_C11 binds to the substrate binding pocket with lower binding energy (−26.72 kcal/mol), followed by DC_C66 with a higher value (−34.71 kcal/mol). The calculated binding energies are in accordance with that of activity, rationalizing our experimental data of bioassays.

The sequence alignment and structural superposition of CARM1, PRMT1, and PRMT6 reveal several differences between these proteins (Figures S2 A-B), which may contribute to selectivity of the CARM1 inhibitors. In the N-terminal helix, which is disordered in the crystal structure of rat PRMT1 and is essential for the enzymatic activity [48], the corresponding residues of F153 in CARM1 are S39 in PRMT1 and C50 in PRMT6 (Figures S2 A-B). Besides, F475 in C-terminal of CARM1 corresponds to R353 in PRMT1 and E374 in PRMT6. Since F153 and F475 are important components of the hydrophobic pocket that accommodates the phenyl ring bulk of DC_C66 (Figure 3), substitutions of the phenylalanine with hydrophilic amino acids may decrease the binding affinity of the CARM1 inhibitors (Figures S2 A-B). These comparisons theoretically explain the selectivity of DC_C66 against CARM1 from the molecular basis.

4. Conclusion

Posttranslational modifications of proteins have been increasingly recognized as essential modulators to their function in cells. In particular, arginine methylation, an important posttranslational modification, is catalyzed by PRMTs. CARM1, a member of PRMTs, has been implicated in a variety of cancers. Thus, the identification of selective inhibitors of CARM1 as probes to investigate CARM1 cellular function and its relevance in disease would be of significant interest in the field of epigenetics. Here in our study, by combining structure-based virtual screening and biochemical assays, we have identified DC_C11 and DC_C66 as novel inhibitors of CARM1, with IC_{50} values of 15 and 1.8 μM, respectively. Notably, DC_C66 displayed good selectivity against PRMT1, PRMT6, and PRMT5. The binding-mode prediction revealed that the two compounds can efficiently bind in the substrate binding site of CARM1 and thus inhibit the enzymatic activity by destruction of protein-protein interactions between CARM1 and its various substrates. Furthermore, the two compounds showed good cell permeability and blocked the proliferation of several cancer cells related to CARM1 overexpression. Overall, this study demonstrated an efficient docking-based virtual screening procedure that can be used to identify novel CARM1 inhibitors. These results paves the way for further development of inhibitors with novel scaffolds and functional probes to target CARM1 on the cellular level for both biological and therapeutic purposes.

Competing Interests

The authors declare no competing interests regarding the publication of this paper.

Authors' Contributions

Fei Ye and Weiyao Zhang contributed equally to this paper.

Acknowledgments

The authors gratefully acknowledge financial support from Zhejiang Province Natural Science Foundation (LQ14H 300003), The National Natural Science Foundation of China (81402849), Public Projects of Zhejiang Province (2015C33159 and 2016C31017), Zhejiang Provincial Top Key Discipline of Biology, Science Foundation of Zhejiang Sci-Tech University (ZSTU) under Grants no. 13042163-Y and no.13042159-Y, and the 521 Talent Cultivation Plan of Zhejiang Sci-Tech University.

References

[1] M. T. Bedford and S. G. Clarke, "Protein arginine methylation in mammals: who, what, and why," *Molecular Cell*, vol. 33, no. 1, pp. 1–13, 2009.

[2] M. T. Bedford and S. Richard, "Arginine methylation: an emerging regulator of protein function," *Molecular Cell*, vol. 18, no. 3, pp. 263–272, 2005.

[3] H. W. Lee, S. Kim, and W. K. Paik, "S-adenosylmethionine: protein-arginine methyltransferase. Purification and mechanism of the enzyme," *Biochemistry*, vol. 16, no. 1, pp. 78–85, 1977.

[4] M. T. Bedford, "Arginine methylation at a glance," *Journal of Cell Science*, vol. 120, no. 24, pp. 4243–4246, 2007.

[5] Y. Yang and M. T. Bedford, "Protein arginine methyltransferases and cancer," *Nature Reviews Cancer*, vol. 13, no. 1, pp. 37–50, 2013.

[6] Y. Feng, R. Maity, J. P. Whitelegge et al., "Mammalian protein arginine methyltransferase 7 (PRMT7) specifically targets RXR sites in lysine- and arginine-rich regions," *Journal of Biological Chemistry*, vol. 288, no. 52, pp. 37010–37025, 2013.

[7] B. T. Schurter, S. S. Koh, D. Chen et al., "Methylation of histone H3 by coactivator-associated arginine methyltransferase 1," *Biochemistry*, vol. 40, no. 19, pp. 5747–5756, 2001.

[8] H. Naeem, D. Cheng, Q. Zhao et al., "The activity and stability of the transcriptional coactivator p/CIP/SRC-3 are regulated by CARM1-dependent methylation," *Molecular and Cellular Biology*, vol. 27, no. 1, pp. 120–134, 2007.

[9] W. Xu, H. Chen, K. Du et al., "A transcriptional switch mediated by cofactor methylation," *Science*, vol. 294, no. 5551, pp. 2507–2511, 2001.

[10] S. Frietze, M. Lupien, P. A. Silver, and M. Brown, "CARM1 regulates estrogen-stimulated breast cancer growth through up-regulation of E2F1," *Cancer Research*, vol. 68, no. 1, pp. 301–306, 2008.

[11] J. Lee and M. T. Bedford, "PABP1 identified as an arginine methyltransferase substrate using high-density protein arrays," *EMBO Reports*, vol. 3, no. 3, pp. 268–273, 2002.

[12] D. Cheng, J. Côté, S. Shaaban, and M. T. Bedford, "The arginine methyltransferase CARM1 regulates the coupling of transcription and mRNA processing," *Molecular Cell*, vol. 25, no. 1, pp. 71–83, 2007.

[13] D. Chen, M. Ma, H. Hong et al., "Regulation of transcription by a protein methyltransferase," *Science*, vol. 284, no. 5423, pp. 2174–2177, 1999.

[14] N. Yadav, J. Lee, J. Kim et al., "Specific protein methylation defects and gene expression perturbations in coactivator-associated arginine methyltransferase 1-deficient mice," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 11, pp. 6464–6468, 2003.

[15] L. P. Vu, F. Perna, L. Wang et al., "PRMT4 blocks myeloid differentiation by assembling a Methyl-RUNX1-dependent repressor complex," *Cell Reports*, vol. 5, no. 6, pp. 1625–1638, 2013.

[16] H. Hong, C. Kao, M.-H. Jeng et al., "Aberrant expression of CARM1, a transcriptional coactivator of androgen receptor, in the development of prostate carcinoma and androgen-independent status," *Cancer*, vol. 101, no. 1, pp. 83–89, 2004.

[17] R. Elakoum, G. Gauchotte, A. Oussalah et al., "CARM1 and PRMT1 are dysregulated in lung cancer without hierarchical features," *Biochimie*, vol. 97, no. 1, pp. 210–218, 2014.

[18] C. Y. Ou, M. J. LaBonte, P. C. Manegold et al., "A coactivator role of CARM1 in the dysregulation of beta-catenin activity in colorectal cancer cell growth and gene expression," *Molecular Cancer Research*, vol. 9, no. 5, pp. 660–670, 2011.

[19] A. Mai, D. Cheng, M. T. Bedford et al., "Epigenetic multiple ligands: mixed histone/protein methyltransferase, acetyltransferase, and class III deacetylase (Sirtuin) inhibitors," *Journal of Medicinal Chemistry*, vol. 51, no. 7, pp. 2279–2290, 2008.

[20] D. Cheng, S. Valente, S. Castellano et al., "Novel 3,5-bis(bromohydroxybenzylidene)piperidin-4-ones as coactivator-associated arginine methyltransferase 1 inhibitors: enzyme selectivity and cellular activity," *Journal of Medicinal Chemistry*, vol. 54, no. 13, pp. 4928–4932, 2011.

[21] A. V. Purandare, Z. Chen, T. Huynh et al., "Pyrazole inhibitors of coactivator associated arginine methyltransferase 1 (CARM1)," *Bioorganic & Medicinal Chemistry Letters*, vol. 18, pp. 4438–4441, 2008.

[22] T. Huynh, Z. Chen, S. Pang et al., "Optimization of pyrazole inhibitors of Coactivator Associated Arginine Methyltransferase 1 (CARM1)," *Bioorganic & Medicinal Chemistry Letters*, vol. 19, no. 11, pp. 2924–2927, 2009.

[23] M. Allan, S. Manku, E. Therrien et al., "*N*-Benzyl-1-heteroaryl-3-(trifluoromethyl)-1*H*-pyrazole-5-carboxamides as inhibitors of co-activator associated arginine methyltransferase 1 (CARM1)," *Bioorganic & Medicinal Chemistry Letters*, vol. 19, no. 4, pp. 1218–1223, 2009.

[24] E. Therrien, G. Larouche, S. Manku et al., "1,2-Diamines as inhibitors of co-activator associated arginine methyltransferase 1 (CARM1)," *Bioorganic and Medicinal Chemistry Letters*, vol. 19, no. 23, pp. 6725–6732, 2009.

[25] H. Wan, T. Huynh, S. Pang et al., "Benzo[d]imidazole inhibitors of Coactivator Associated Arginine Methyltransferase 1 (CARM1)-hit to lead studies," *Bioorganic & Medicinal Chemistry Letters*, vol. 19, no. 17, pp. 5063–5066, 2009.

[26] J. S. Sack, S. Thieffine, T. Bandiera et al., "Structural basis for CARM1 inhibition by indole and pyrazole inhibitors," *The Biochemical Journal*, vol. 436, no. 2, pp. 331–339, 2011.

[27] B. R. Selvi, K. Batta, A. H. Kishore et al., "Identification of a novel inhibitor of Coactivator-associated Arginine Methyltransferase 1 (CARM1)-mediated methylation of histone H3 Arg-17," *The Journal of Biological Chemistry*, vol. 285, no. 10, pp. 7143–7151, 2010.

[28] H. Hu, K. Qian, M.-C. Ho, and Y. G. Zheng, "Small molecule inhibitors of protein arginine methyltransferases," *Expert Opinion on Investigational Drugs*, vol. 25, no. 3, pp. 335–358, 2016.

[29] L. Li, R. Zhou, H. Geng et al., "Discovery of two aminoglycoside antibiotics as inhibitors targeting the menin-mixed lineage leukaemia interface," *Bioorganic & Medicinal Chemistry Letters*, vol. 24, no. 9, pp. 2090–2093, 2014.

[30] S. Chen, Y. Wang, W. Zhou et al., "Identifying novel selective non-nucleoside DNA methyltransferase 1 inhibitors through docking-based virtual screening," *Journal of Medicinal Chemistry*, vol. 57, no. 21, pp. 9028–9041, 2014.

[31] W. W. Yue, M. Hassler, S. M. Roe, V. Thompson-Vale, and L. H. Pearl, "Insights into histone code syntax from structural and biochemical studies of CARM1 methyltransferase," *The EMBO Journal*, vol. 26, no. 20, pp. 4402–4412, 2007.

[32] N. Troffer-Charlier, V. Cura, P. Hassenboehler, D. Moras, and J. Cavarelli, "Functional insights from structures of coactivator-associated arginine methyltransferase 1 domains," *The EMBO Journal*, vol. 26, no. 20, pp. 4391–4401, 2007.

[33] P. A. Boriack-Sjodin, L. Jin, S. L. Jacques et al., "Structural insights into ternary complex formation of human CARM1 with various substrates," *ACS Chemical Biology*, vol. 11, no. 3, pp. 763–771, 2016.

[34] R. T. Sauer, D. N. Bolon, B. M. Burton et al., "Sculpting the proteome with AAA(+) proteases and disassembly machines," *Cell*, vol. 119, no. 1, pp. 9–18, 2004.

[35] C. A. Lipinski, F. Lombardo, B. W. Dominy, and P. J. Feeney, "Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings," *Advanced Drug Delivery Reviews*, vol. 46, no. 1–3, pp. 3–26, 2001.

[36] J. B. Baell and G. A. Holloway, "New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays," *Journal of Medicinal Chemistry*, vol. 53, no. 7, pp. 2719–2740, 2010.

[37] A. Whitty, "Growing PAINS in academic drug discovery," *Future Medicinal Chemistry*, vol. 3, no. 7, pp. 797–801, 2011.

[38] J. Baell and M. A. Walters, "Chemistry: chemical con artists foil drug discovery," *Nature*, vol. 513, no. 7519, pp. 481–483, 2014.

[39] Accelrys Software Inc, *Pipeline Pilot. Version 7.5*, Accelrys Software Inc, San Diego, Calif, USA, 2008.

[40] D. Frees, A. Chastanet, S. Qazi et al., "Clp ATPases are required for stress tolerance, intracellular replication and biofilm formation in Staphylococcus aureus," *Molecular Microbiology*, vol. 54, no. 5, pp. 1445–1462, 2004.

[41] O. Gaillot, S. Bregenholt, F. Jaubert, J. P. Di Santo, and P. Berche, "Stress-induced ClpP serine protease of *Listeria monocytogenes* is essential for induction of listeriolysin O-dependent protective immunity," *Infection and Immunity*, vol. 69, no. 8, pp. 4938–4943, 2001.

[42] R. A. Friesner, R. B. Murphy, M. P. Repasky et al., "Extra precision glide: docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes," *Journal of Medicinal Chemistry*, vol. 49, no. 21, pp. 6177–6196, 2006.

[43] J. Wang, J. A. Hartling, and J. M. Flanagan, "The structure of ClpP at 2.3 Å resolution suggests a model for ATP- dependent proteolysis," *Cell*, vol. 91, no. 4, pp. 447–456, 1997.

[44] T. J. A. Ewing, S. Makino, A. G. Skillman, and I. D. Kuntz, "DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases," *Journal of Computer-Aided Molecular Design*, vol. 15, no. 5, pp. 411–428, 2001.

[45] D. S. Goodsell, G. M. Morris, and A. J. Olson, "Automated docking of flexible ligands: applications of AutoDock," *Journal of Molecular Recognition*, vol. 9, no. 1, pp. 1–5, 1996.

[46] J. Zhang and Y. G. Zheng, "SAM/SAH analogs as versatile tools for SAM-dependent methyltransferases," *ACS Chemical Biology*, vol. 11, no. 3, pp. 583–597, 2016.

[47] R. M. Raju, A. L. Goldberg, and E. J. Rubin, "Bacterial proteolytic complexes as therapeutic targets," *Nature Reviews Drug Discovery*, vol. 11, no. 10, pp. 777–789, 2012.

[48] X. Zhang and X. Cheng, "Structure of the predominant protein arginine methyltransferase PRMT1 and analysis of its binding to substrate peptides," *Structure*, vol. 11, no. 5, pp. 509–520, 2003.

Hindawi Publishing Corporation
BioMed Research International
Volume 2016, Article ID 8036450, 11 pages
http://dx.doi.org/10.1155/2016/8036450

Research Article

Potential Role of the Last Half Repeat in TAL Effectors Revealed by a Molecular Simulation Study

Hua Wan,[1] Shan Chang,[2] Jian-ping Hu,[3] Xu-hong Tian,[1] and Mei-hua Wang[1]

[1]*College of Mathematics and Informatics, South China Agricultural University, Guangzhou, China*
[2]*School of Electrical and Information Engineering, Jiangsu University of Technology, Changzhou, China*
[3]*Faculty of Biotechnology Industry, Chengdu University, Chengdu, China*

Correspondence should be addressed to Mei-hua Wang; wangmeihua@scau.edu.cn

Received 20 May 2016; Revised 16 August 2016; Accepted 24 August 2016

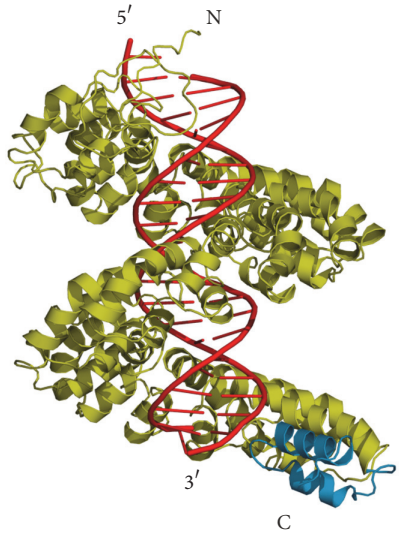
Academic Editor: Zhongjie Liang

TAL effectors (TALEs) contain a modular DNA-binding domain that is composed of tandem repeats. In all naturally occurring TALEs, the end of tandem repeats is invariantly a truncated half repeat. To investigate the potential role of the last half repeat in TALEs, we performed comparative molecular dynamics simulations for the crystal structure of DNA-bound TALE AvrBs3 lacking the last half repeat and its modeled structure having the last half repeat. The structural stability analysis indicates that the modeled system is more stable than the nonmodeled system. Based on the principle component analysis, it is found that the AvrBs3 increases its structural compactness in the presence of the last half repeat. The comparison of DNA groove parameters of the two systems implies that the last half repeat also causes the change of DNA major groove binding efficiency. The following calculation of hydrogen bond reveals that, by stabilizing the phosphate binding with DNA at the C-terminus, the last half repeat helps to adopt a compact conformation at the protein-DNA interface. It further mediates more contacts between TAL repeats and DNA nucleotide bases. Finally, we suggest that the last half repeat is required for the high-efficient recognition of DNA by TALE.

1. Introduction

Transcriptional activator-like effectors (TALEs) are DNA-binding proteins secreted by *Xanthomonas* bacteria [1]. In TALEs, the DNA-binding domain is composed of a repeated highly conserved 33~35 (mostly 34) amino acids' sequence with the exception of the 12th and 13th amino acids. These two residues, known as repeat-variable diresidues (RVDs), are responsible for the specific nucleotide recognition [2, 3]. Both experimental [2] and computational [3] studies found that there is a strong correlation between RVDs and target DNA bases. For example, RVDs Asn/Ile (NI), His/Asp (HD), and Asn/Gly (NG) recognize adenine (A), cytosine (C), and thymine (T), respectively. This simple code allows the design of specific TALE protein by selecting a combination of repeats with appropriate RVDs [4, 5]. The modularity of DNA-binding domain of TALEs has been widely used in biotechnological applications [5, 6], such as genome editing in plants, animals, and human cells, as well as to induce gene expression.

To understand the modular nature of TALE-DNA binding, a series of studies focused on the structural basis for TALE-DNA recognition. In 2010, a nuclear magnetic resonance (NMR) structure of TALE protein PthA was solved by Murakami et al. [7]. The NMR analysis revealed that there are two α antiparallel helices in each repeat. In 2012, researchers led by Shi and Yan crystallized two structures of 11.5-repeat TALE dHax3 in the presence and the absence of DNA at resolutions of 1.8 Å and 2.4 Å, respectively [8]. This study uncovered that amino acid 13 of RVD specifies the identity of a DNA base while amino acid 12 of RVD stabilizes the repeat structure. Separately, researchers led by Stoddard determined the 3.0 Å structure of the naturally occurring TALE PthXo1 bound to DNA [9]. This structure contains over 20 repeats, showing examples of the six most common RVD types. In 2013, Stella et al. reported the crystal structure of TALE AvrBs3 in complex with its target DNA, with the last half repeat being unresolved [10]. This study shows a new interaction mode of the initial thymine T_0 recognition by TALE protein. Additionally, several studies investigated the



(a)

Repeat number	N RVDs	5′ Nucleotide
1	HD	A
2	NG	T
3	NI	A
4	NG	T
5	NI	A
6	NI	A
7	NI	A
8	HD	C
9	HD	C
10	NG	T
11	NI	A
12	NI	A
13	HD	C
14	HD	C
15	HD	C
16	NG	T
17	HD	C
17.5	NG	T
	C	3′

(b)

Figure 1: Complex structure and domain organization of AvrBs3 bound to DNA. (a) The complex structure of AvrBs3-DNA. In the crystal structure, AvrBs3 (yellow) contains a 17.5-repeat domain to mediate the DNA (red) binding. The unresolved last half repeat $R_{17.5}$ was modeled based on the last half repeat in the dHax3-DNA structure (PDB codes: 3V6T) and was colored in blue separately. (b) The 17.5-repeat domain of AvrBs3 conferring DNA sequence. In each repeat, RVD residues are responsible for the specific nucleotide recognition of the DNA sense strand.

specificities and efficiencies of TALE-DNA binding [11–13]. The above biochemical data is important for exploring the TALE-DNA recognition mechanism.

Furthermore, theoretical studies also improved our understanding of TALE-DNA interactions. Moscou and Bogdanove used a computational method to decide the TALE recognition code [3]. Bradley modeled the structure of TALE in complex with DNA based on the Rosetta package and successfully predicted the TALE-DNA interaction [14]. Grau et al. developed a new software platform for predicting TAL effector target sites based on a statistical model [15]. Several molecular simulation studies were applied to investigate the specificities of TALE-DNA binding and conformational changes of TALE [16–19]. Nevertheless, some interesting issues still need to be further probed. In all natural TALEs, surprisingly, the last repeat of tandem repeats is always a truncated half repeat [1]. The previous crystallographic data [8] and our molecular simulation study [17] showed that the last repeat of TALE protein dHax3 forms a stable interaction with DNA. It suggests a necessity of the last half repeat for biological functions. However, the last half repeat was also considered to be dispensable for the function of gene activation by both transient expression assays in *Nicotiana benthamiana* and gene-specific targeting in the rice genome [20]. In order to reduce the complexity and costs, the last half repeat was suggested to be omitted in the design of TALE nucleases [20]. Then, is there the necessity for the last half repeat to occur in TALEs? If yes, how does the last half repeat affect the TALE-DNA binding in detail? What is the difference of the protein-DNA interaction between the two DNA-bound TALE proteins, lacking and having the last half repeat?

In order to answer the above questions, we selected the crystal structure of TALE AvrBs3 (lacking the last half repeat) to perform the comparative molecular dynamics (MD) simulations. The two simulated systems, in the absence and the presence of the last half repeat, were built. By performing MD simulations, we compared the stabilities of the two systems. Principal component analysis (PCA) was applied to probe the functional dynamics in the two systems. The groove deformation of TALE-bound DNA was analyzed at the base pair level. To explain the conformational difference between the two systems, we investigated the specific and nonspecific interactions at the TALE-DNA interface. Finally, we proposed the potential role of the last half repeat in the specific recognition and binding of TALE-DNA.

2. Systems and Methods

2.1. The Structures of AvrBs3-DNA Complex Systems. The crystal structure of the AvrBs3-DNA complex (PDB codes: 2YPF) was obtained from the Protein Data Bank [10]. In the crystal structure, AvrBs3 (yellow) contains a 17.5-repeat TALE domain to confer DNA sequence (red) specificity (Figure 1(a)), with the last half repeat $R_{17.5}$ being unresolved. Then, repeat $R_{17.5}$ (blue) was modeled based on the last half repeat in the TALE dHax3-DNA structure (PDB codes: 3V6T) [8]. A total of 17.5 repeats form a superhelix and bind with the sense strand along the DNA major groove. In each repeat, the RVDs are responsible for recognizing one specific nucleotide (Figure 1(b)). For convenience, the two systems lacking and having repeat $R_{17.5}$ were referred to as the nonmodeled and the modeled systems, respectively.

2.2. Molecular Dynamics Simulation. Two independent simulation systems were prepared using VMD 1.9 [21]. In each system, the complex structure was solvated in a periodic box

filled with TIP3P water molecules. The minimum distance is about 10 Å from the solute unit to the box wall. Each of the two systems was neutralized by adding 49 sodium ions (Na^+) with VMD 1.9. Then, the two MD simulations were performed with the NAMD 2.9 program [22] using the CHARMM27 all-atom additive force field for nucleic acids [23]. The SHAKE algorithm [24] was used to constrain all bonds involving hydrogen atoms, and particle mesh Ewald (PME) method [25] was applied to evaluate electrostatic interactions. Meanwhile, Lennard-Jones potential was truncated at a cut-off distance of 12 Å. Each simulation included two stages. (i) The systems were minimized with 20000-step energy minimization and then slowly were heated from 0 to 310 K over 0.5 ns. To keep the stabilization of systems, all backbone atoms of protein and DNA were restrained with a harmonic constant of 0.1 kcal $\cdot$ mol^{-1} $\cdot$ Å^{-2}. (ii) After the positional constraints were removed, the productive MD simulations were run for 15 ns under constant pressure (1 atm) and temperature (310 K) conditions. The pressure and temperature were kept using the Langevin piston method [26]. The atomic coordinates were stored every 2.0 ps. Hence, 7500 snapshots in each system were collected for further analysis.

2.3. Principal Component Analysis. Principal component analysis (PCA) is a standard method for obtaining a brief picture of motions. This method exacts the highly correlated fluctuations from the MD trajectories through dimensionality reduction. The definition of PCA is based on the construction and diagonalization of the covariance matrix. The element C_{ij} in the matrix is calculated according to [27]

$$C_{ij} = \left\langle \left(x_i - \left\langle x_i \right\rangle\right)\left(x_j - \left\langle x_j \right\rangle\right)\right\rangle, \quad (1)$$

where $x_i(x_j)$ is the coordinate of the ith (jth) atom of the systems and $\langle \cdots \rangle$ represents an ensemble average. The eigenvectors of the matrix give the directions of the concerted motions. The eigenvalues indicate the magnitude of the motions along the direction. The first few principal components (PCs) usually contain the most important conformational changes of a biomolecular system [17, 28, 29]. In this study, PCA was performed with Gromacs 4.5 package [30] to detect the conformational difference between the two systems.

2.4. Conformational Analysis of Nucleic Acids. Curves program is the most widely used in analysis of nucleic acid conformations [31]. This program can provide an entire set of DNA structural parameters. By using the Curves program, we obtain the groove parameters to describe the DNA groove deformation in this paper.

3. Results and Discussion

3.1. MD Results. Two 15 ns MD simulations were carried out for the nonmodeled (lacking the last half repeat) and the modeled (having the last half repeat) systems, respectively. Figure 2(a) compares the root mean square deviation values (RMSDs) of backbone atoms of the AvrBs3-DNA complex from the two systems. The two systems remain relatively stable after 9 ns, and then the last 6 ns MD trajectories are taken as the equilibrium portions for the two systems. Figures 2(b), 2(c), and 2(d) display the distributional probability of RMSD from the equilibrium trajectories. In the nonmodeled system, the RMSDs converge to about 3.07 Å, 3.37 Å, and 2.40 Å for the AvrBs3-DNA complex, AvrBs3, and DNA, respectively. In the modeled system, the RMSDs converge to about 2.38 Å, 2.44 Å, and 2.29 Å for the AvrBs3-DNA complex, AvrBs3, and DNA, respectively. This indicates that the modeled system is more stable than the nonmodeled system. The only difference between the two systems is that the modeled system has an additional repeat, $R_{17.5}$. The previous crystallographic data revealed that the last half repeat contributes to the protein-DNA binding in the structure of DNA-bound TALE dHax3 [17]. All these suggest that the last half repeat increases the structural stability.

We also calculated the root mean square fluctuation values (RMSFs) of the common 17 repeats (from repeat 1 to repeat 17) of AvrBs3 and 20 bases (from position −1 to position 18) of DNA in the two systems from the equilibrium trajectories. The results are given in Figures 2(e) and 2(f), and 17 repeats are labeled as R_1 to R_{17}. In each system, the linker between two adjacent TAL repeats shows higher RMSFs (Figure 2(e)). The RVD loop within each repeat has lower RMSFs because the RVD loop region is the DNA-binding site in a repeat. Of all the repeats, R_{17} undergoes the highest fluctuations. Notably, in the nonmodeled system, the RMSFs of the RVD loop of R_{17} increase markedly relative to the other RVD loops. However, in the modeled system, the RVD loop of R_{17} still maintains relatively lower RMSFs. Meanwhile, the 3′ end of the DNA sense strand is more flexible in the nonmodeled system compared with the modeled system (Figure 2(f)). It indicates that the AvrBs3 of the modeled system is well constrained by DNA. In contrast, the nonmodeled system loses some important protein-DNA contacts. The RMSFs analysis implies that the absence of the last half repeat will partially impair the binding of AvrBs3 to DNA.

3.2. Conformational Change of AvrBs3. Previous studies revealed the conformational plasticity of TALEs bound to DNA [7, 8, 17]. To detect the conformational change of DNA-bound AvrBs3, the PCA was performed for Cα atoms of protein and P atoms of DNA to obtain slow motions based on the equilibrium trajectories of the nonmodeled and the modeled systems. Figure 3 gives the proportion of system's variance accounted for by the first 50 PCs, which was calculated from the diagonalization of the covariance matrix. The proportion rapidly decreases and converges to zero with the increasing of PC index in each system. The first two PCs together account for approximately 47.9% and 45.6% of the total variance in the nonmodeled and the modeled systems, respectively. In an equilibrium system, the motions on the backbone are mainly the localized random motions. Thereby, PC1 and PC2 of the two systems capture higher fraction of the system's variance.

Figure 4 describes the first and the second slowest motion modes. The first slowest motion exhibits some swing motions towards the DNA major groove in the two systems (Figures 4(a) and 4(b)). By observing their average structures, in the

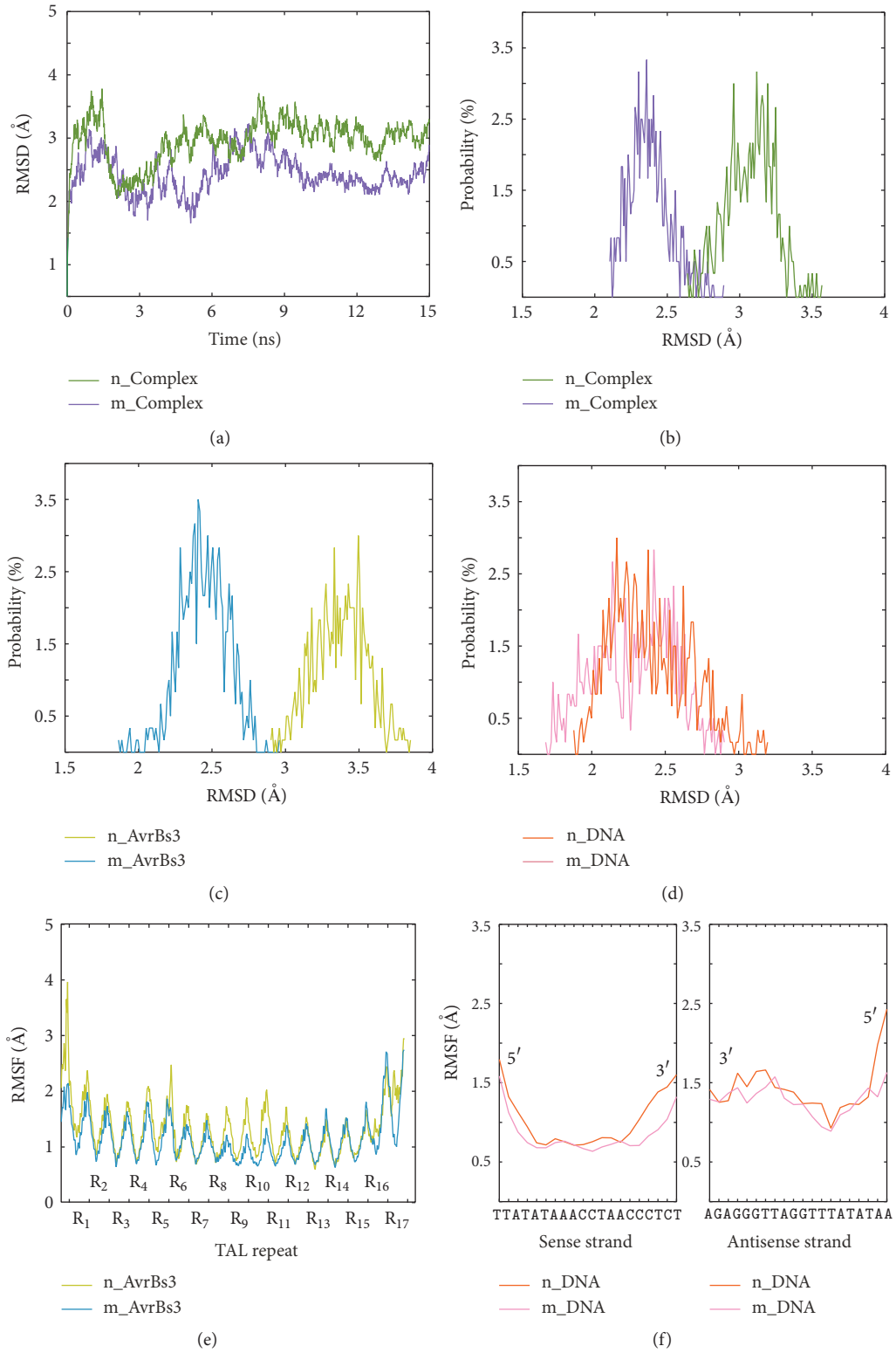


Figure 2: Comparative MD analysis of the nonmodeled system (n_Complex: green; n_AvrBs3: yellow; n_DNA: orange) and the modeled system (m_Complex: purple; m_AvrBs3: blue; m_DNA: pink). (a) The RMSDs of the AvrBs3 backbone atoms versus simulation time. (b~d) The RMSD probability distribution of the AvrBs3-DNA complex (b), AvrBs3 (c), and DNA (d) calculated from the equilibrium trajectories. (e) The RMSFs of the Cα atoms of AvrBs3 calculated from the equilibrium trajectories. (f) The RMSFs of the P atoms in the sense strand (left) and the antisense strand (right) of DNA calculated from the equilibrium trajectories.

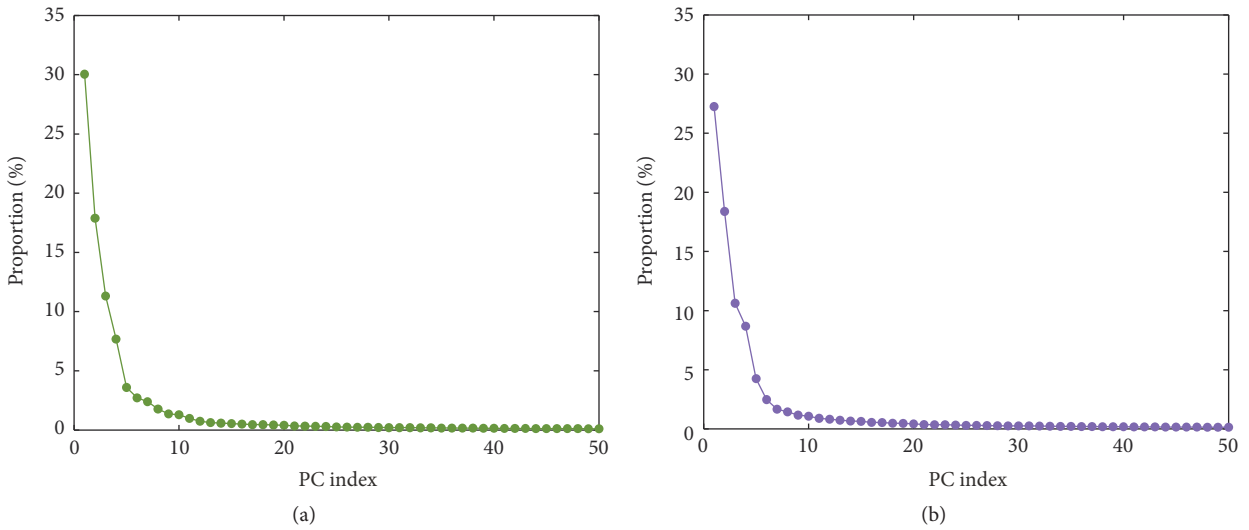


Figure 3: The proportion of system's variance accounted for by the first 50 PCs of the nonmodeled system (a) and the modeled system (b).

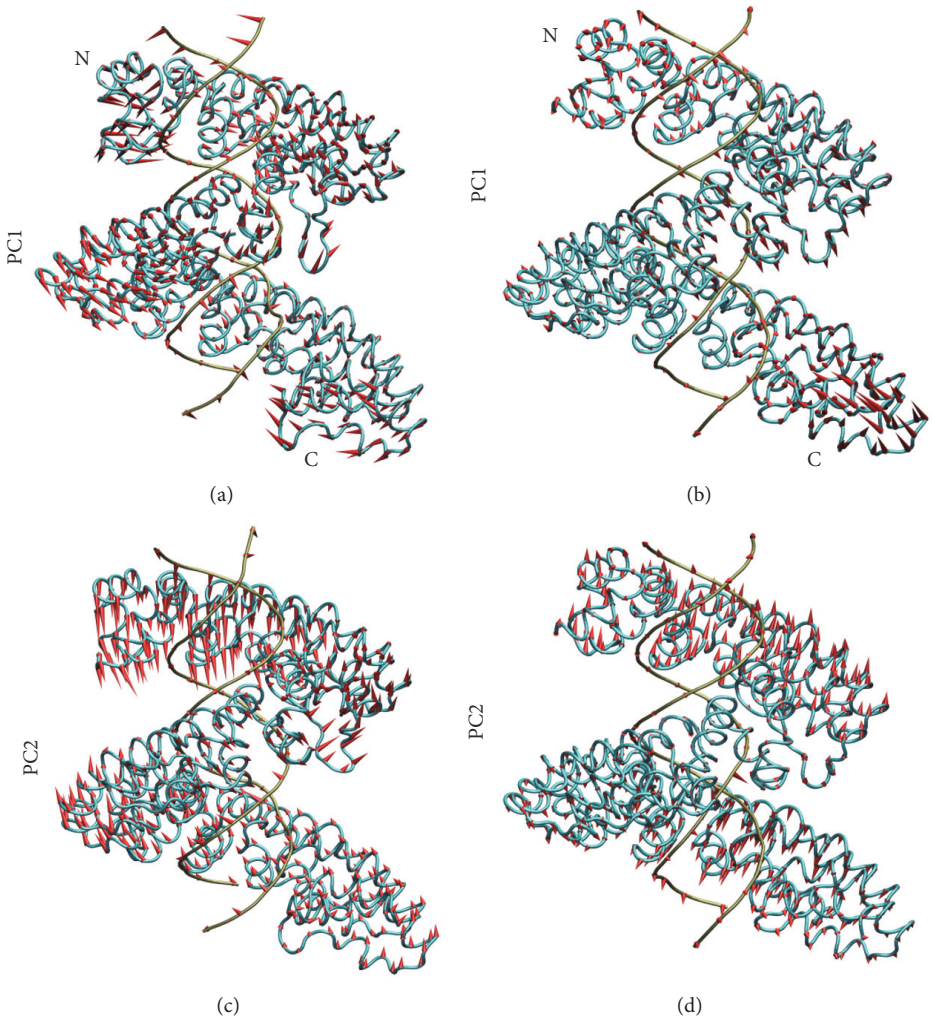


Figure 4: The first and the second slowest motion modes of the nonmodeled system (a and c) and the modeled system (b and d). The average structure is based on the equilibrium trajectories. The length of cone is positively correlated with motive magnitude, and the motive direction is depicted with the orientation of cone.

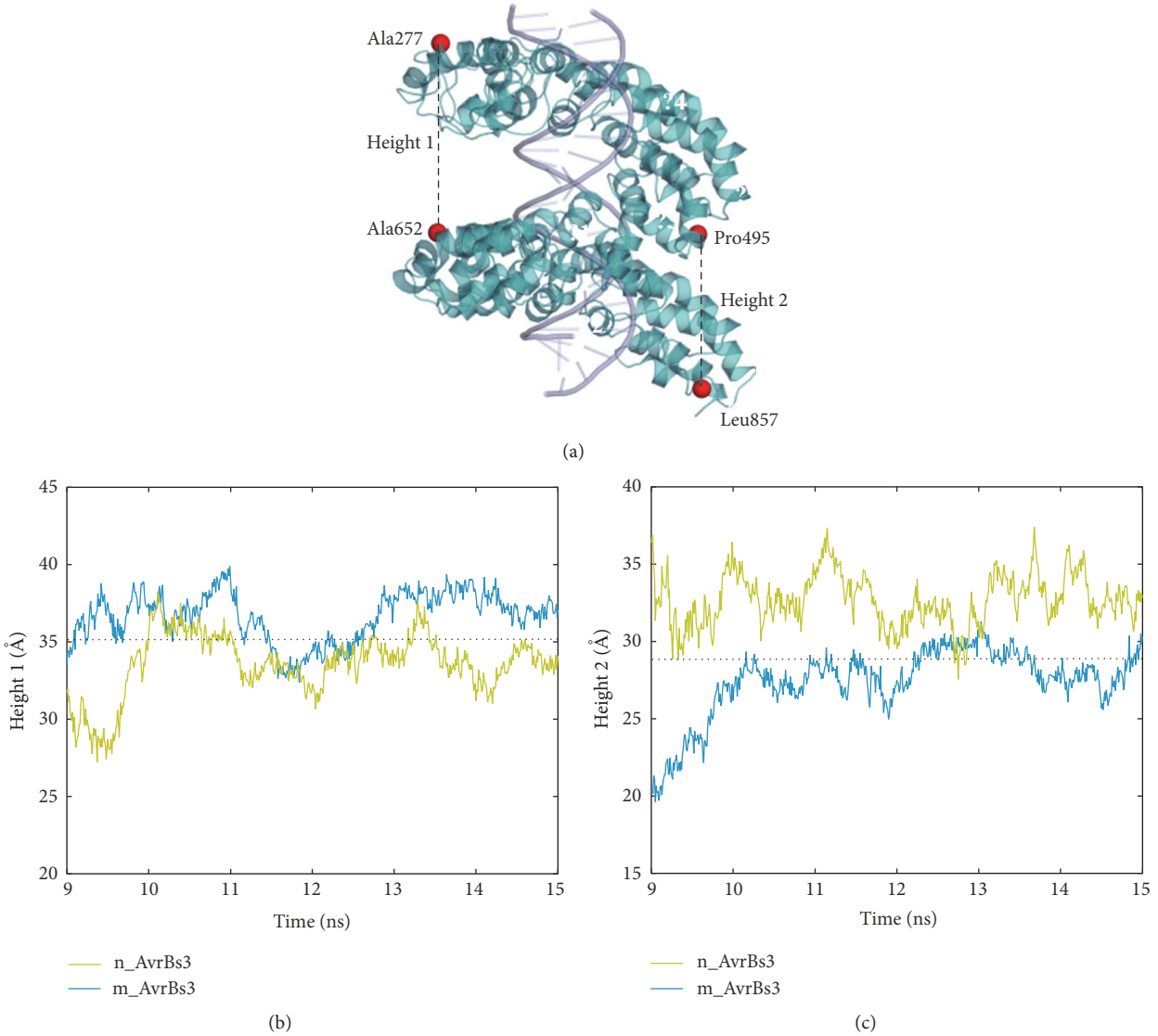


FIGURE 5: The height change of the superhelical structure of AvrBs3. (a) The height of the first half of the superhelical structure is assessed by the distance between the Cα atoms of Ala277 and Ala652 and that of the second by the distance between the Cα atoms of Pro495 and Leu857. (b) The height change of the first half of the superhelical structure versus simulation (solid line) and the value from crystal structure (dotted line). (c) The height change of the second half of the superhelical structure versus simulation (solid line) and the value from crystal structure (dotted line).

nonmodeled system the last few repeats show a conformation far from the DNA major groove (Figure 4(a)). It is presumably because the swing motion breaks the protein-DNA interaction at the binding interface. In contrast, the protein-DNA interface of the modeled system still keeps a compact conformation at the C-terminus (Figure 4(b)). This conformation difference of the C-terminus between the systems is consistent with the above RMSFs analysis.

The second slowest motion mode shows some extension-compression movements of the superhelical structure of AvrBs3 (Figures 4(c) and 4(d)). The previous X-ray scattering (SAXS) data [7] and crystal structure study [8] revealed that TALEs underwent a compressed conformational change upon DNA interaction. This conformational change caused the height change of the superhelical structure of TALE protein [8]. Then, the four atoms, which are Cα atoms of Ala277 (repeat 0), Pro495 (repeat 7), Ala652 (repeat 11), and Leu857 (repeat 17), were selected to measure the height change of the first and the second halves of the superhelical structure (Figure 5(a)). For the first half of the superhelical structure, the average height is 35.1 Å, 33.5 Å, and 36.7 Å for the crystal structure, the nonmodeled system, and the modeled system, respectively (Figure 5(b)). For the second half of the superhelical structure, the average height is 28.9 Å, 32.7 Å, and 27.4 Å for the crystal structure, the nonmodeled system, and the modeled system, respectively (Figure 5(c)). As a whole, the modeled system still maintains a compressed conformation relative to the crystal structure. In the nonmodeled system,

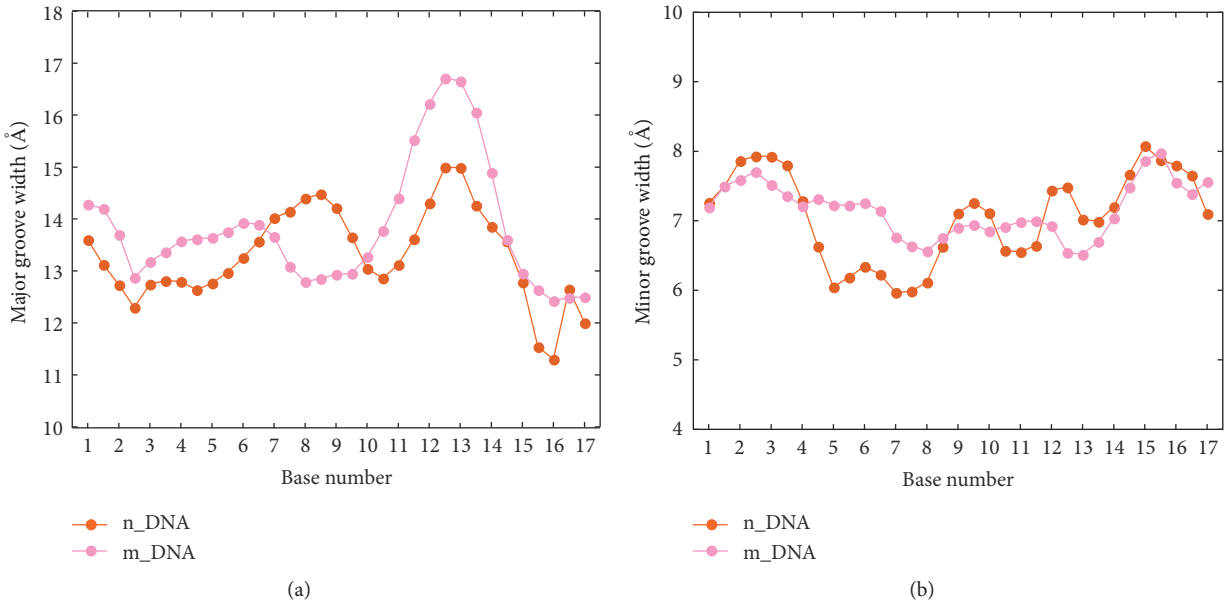


FIGURE 6: Average values of groove widths calculated from the equilibrium trajectories along the target sequence (from position 1 to position 17) in the nonmodeled (orange) and the modeled (pink) systems. (a) Major groove widths. (b) Minor groove widths.

the superhelical structure of AvrBs3 is comparatively more extended. The combined analyses of the first and the second slowest motions clearly show that the AvrBs3-DNA complex structure keeps a more compact conformation in the presence of the last half repeat. Meanwhile, the increase of structural compactness of TALE is associated with the DNA binding [7, 8]. Therefore, the last half repeat makes an important contribution to the TALE-DNA binding.

3.3. Groove Deformation of DNA. DNA groove dimensions are important structural feature in processes involving specific protein-DNA binding [32]. Then, the DNA groove parameters of the two systems were calculated by the Curves program [31] from the equilibrium trajectories. The result is shown in Figure 6. Along the target sequence, except for positions 8 and 9, the major groove of the modeled system is almost always wider than that of the nonmodeled system (Figure 6(a)). The wider major groove makes the side chain of the key amino acid of protein more accessible to nucleotide bases and then can mediate more protein-DNA contacts. It is suggested that the efficiency of DNA major groove binding by AvrBs3 should be relatively higher in the modeled system. The interactions at the protein-DNA interface will be analyzed in the next section.

Notably, the major groove at positions 8 and 9 is markedly narrowed in the modeled system relative to the nonmodeled system. To investigate whether there is some relationship between the groove narrowing of DNA and the structural compression of AvrBs3, we compared the time-dependent fluctuation of groove width at each base pair step with the height change of the superhelical structure of AvrBs3. For the first part of the complex structure (Figure 5(a)), the height change of AvrBs3 (Figure 5(b)) is similar to the fluctuation of minor groove width at position 5 (Figure 7(a)). For the second part of the complex structure (Figure 5(a)), the height change of AvrBs3 (Figure 5(c)) accompanies the deformations of major groove at position 8 and of minor groove at position 13 together (Figure 7(b)). It indicates that the TALE-DNA binding process is associated with some structural adaptation of the DNA as well as the AvrBs3 in order to accommodate each other. The conformational difference between the two systems may reflect the changes of the TALE-DNA binding.

3.4. Interactions at the Interface. To compare the difference of the protein-DNA interaction between the two systems, we examined the hydrogen bonds along the DNA major groove based on the equilibrium trajectories. The hydrogen bond calculation was performed with VMD 1.9 [21] using a distance cut-off value of 3.5 Å and an angle cut-off value of 45°. The result is listed in Table 1 with occupancy over 30%. Relative to the nonmodeled system, the modeled system has four additional specific hydrogen bonds and four additional nonspecific hydrogen bonds. The calculation of hydrogen bond proves that the modeled system has a higher protein-DNA binding efficiency in the DNA major groove. These additional interactions help the modeled system to achieve higher stability, which is consistent with the above analysis of RMSDs.

Compared with the nonmodeled system, the additional specific interactions of the modeled system are mainly formed by the N- and C-terminal repeats, especially by the last few repeats (Table 1). Figure 8 describes the difference of the specific interaction between the two systems. In the nonmodeled system (Figure 8(a)), Asp743 (repeat 14) forms a direct and a water-mediated hydrogen bond with cytosine 14 and cytosine 15 separately. OE2 of Gln781 (repeat 15) interacts with O3′ of cytosine 14. Meanwhile, repeats 16~17 lose the

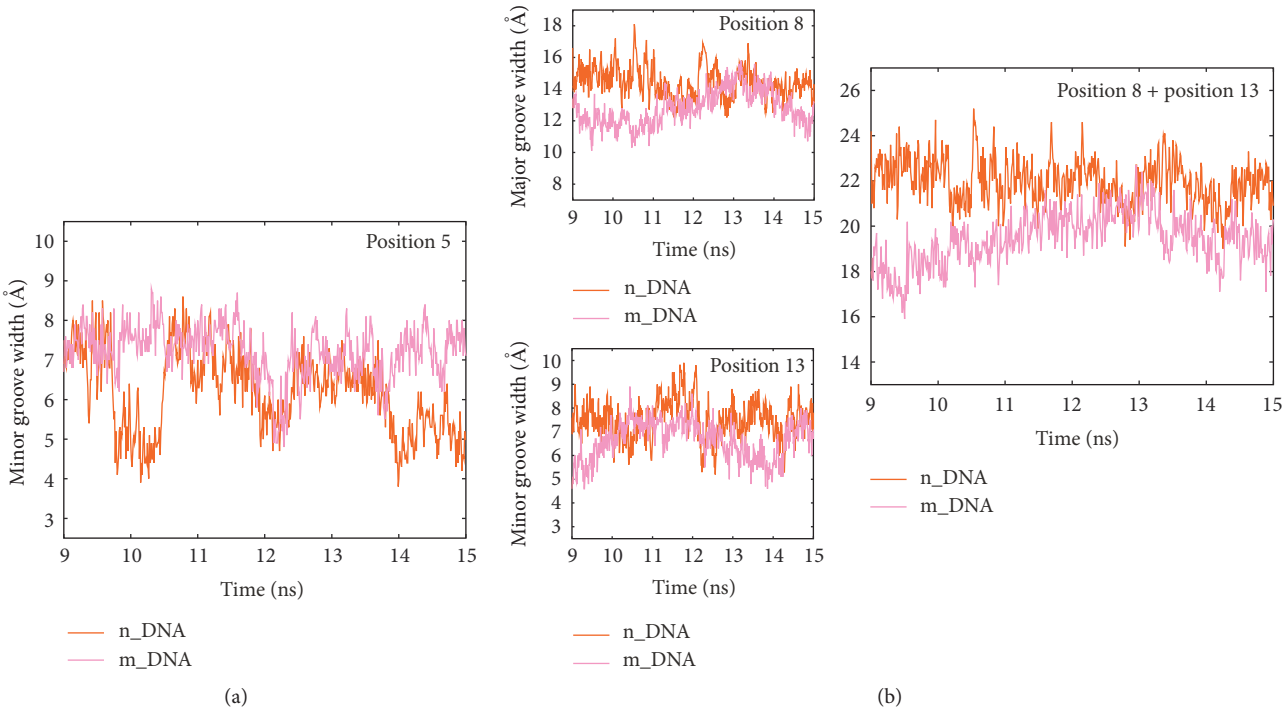


FIGURE 7: Time-dependent fluctuations of DNA groove widths at positions 5 (a) and 8 and 13 (b) calculated from the equilibrium trajectories in the nonmodeled (orange) and the modeled (pink) systems.

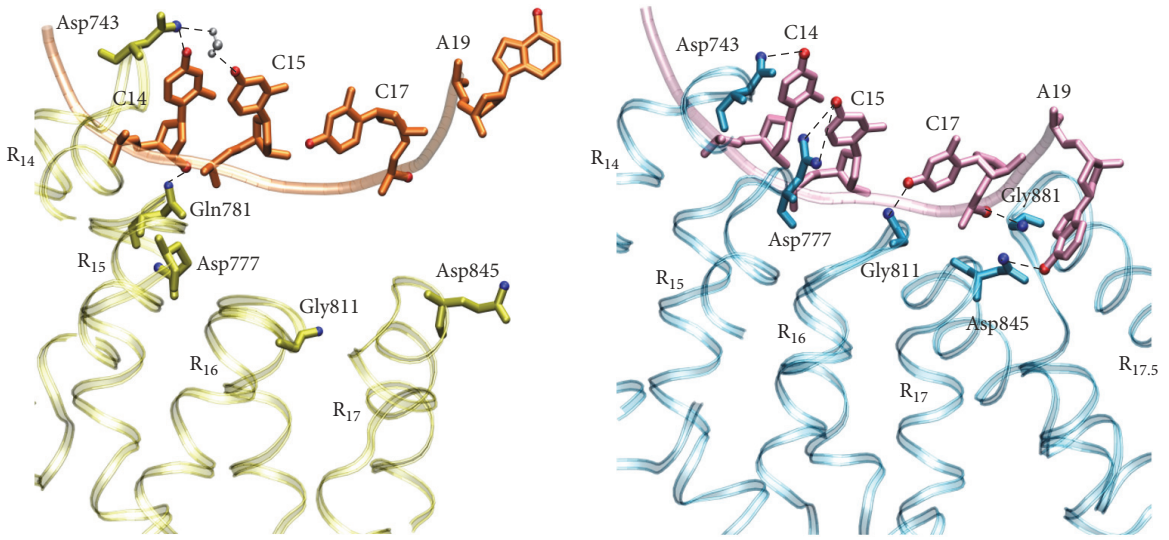


FIGURE 8: The interactions between the last few repeats and DNA from representative structures in the nonmodeled (a) and the modeled (b) systems. The repeats, DNA, and water molecule are depicted with ribbons, tube, and CPK models, respectively. Repeats 14, 15, 16, 17, and 17.5 are labeled as R_{14}, R_{15}, R_{16}, R_{17}, and $R_{17.5}$, respectively. Nucleotide bases cytosine 14, cytosine 15, cytosine 17, and adenine 19 are labeled as C14, C15, C17, and A19, respectively. Thymine 16 and thymine 18 are omitted for clarity.

contact with nucleotide bases. The C-terminal repeats show a conformation far from the backbone of DNA. In the modeled system (Figure 8(b)), Asp743 (repeat 14), Asp777 (repeat 15), Gly811 (repeat 16), and Asp845 (repeat 17) form stable specific hydrogen bonds with cytosine 14, cytosine 15, cytosine 17, and adenine 19, respectively. Notably, N of Gly881 (repeat 17.5) interacts with O1P of cytosine 17. This phosphate binding adopts a compact conformation at the protein-DNA interface and further helps to mediate more base-specific interactions. The previous study revealed that the last repeat is always a

Table 1: The hydrogen bonds with occupancy over 30%.

Base	Nonmodeled system			Modeled system		
Position	Protein$^{(id)}$	DNA	HDO*	Protein$^{(id)}$	DNA	HDO*
0	Gly302-N$^{(1)}$	T0-O2P$^{\alpha}$	74.88%	Thr270-N$^{(0)}$	T0-O2P$^{\alpha}$	83.19%
				Gln305-N$^{(1)}$	T0-O1P$^{\alpha}$	63.73%
1	***Asp301_OD2*$^{(1)}$**	***A1-N6*$^{\alpha}$**	***39.41*%**	***Asp301_OD1*$^{(1)}$**	***A1-N6*$^{\alpha}$**	***60.35*%**
				***Asp301_OD1*$^{(1)}$**	***A1-N7*$^{\alpha}$**	***58.47*%**
	Gln339-NE2$^{(2)}$	A1-O1P$^{\alpha}$	50.89%	Gln339-NE2$^{(2)}$	A1-O1P$^{\alpha}$	88.85%
2				***Asp301_OD1*$^{(1)}$**	***T2-O4*$^{\alpha}$**	***47.65*%**
	Gln373-NE2$^{(3)}$	T2-O1P$^{\alpha}$	35.44%	Gln373-NE2$^{(3)}$	T2-O1P$^{\alpha}$	67.05%
3	Gln407-NE2$^{(4)}$	A3-O2P$^{\alpha}$	53.24%	Gln407-NE2$^{(4)}$	A3-O2P$^{\alpha}$	58.90%
4	Gln441-NE2$^{(5)}$	T4-O1P$^{\alpha}$	63.73%	Gln441-NE2$^{(5)}$	T4-O1P$^{\alpha}$	86.36%
5	Gln475-NE2$^{(6)}$	A5-O2P$^{\alpha}$	30.78%	Gln475-NE2$^{(6)}$	A5-O2P$^{\alpha}$	45.92%
6	Gln509-NE2$^{(7)}$	A6-O2P$^{\alpha}$	63.56%	Gln509-NE2$^{(7)}$	A6-O2P$^{\alpha}$	79.03%
7	Gln543-NE2$^{(8)}$	A7-O2P$^{\alpha}$	96.01%	Gln543-NE2$^{(8)}$	A7-O2P$^{\alpha}$	94.18%
8	**Asp539-OD2$^{(8)}$**	**C8-N4$^{\alpha}$**	**30.23%**	**Asp539-OD2$^{(8)}$**	**C8-N4$^{\alpha}$**	**35.44%**
	Gln577-NE2$^{(9)}$	C8-O1P$^{\alpha}$	92.68%	Gln577-NE2$^{(9)}$	C8-O1P$^{\alpha}$	63.73%
9	**Asp573-OD1$^{(9)}$**	**C9-N4$^{\alpha}$**	**36.77%**	**Asp573-OD1$^{(9)}$**	**C9-N4$^{\alpha}$**	**34.11%**
	Asp573-OD2$^{(9)}$	**C9-N4$^{\alpha}$**	**30.28%**	**Asp573-OD2$^{(9)}$**	**C9-N4$^{\alpha}$**	**60.90%**
	Gln611-NE2$^{(10)}$	C9-O1P$^{\alpha}$	54.08%	Gln611-NE2$^{(10)}$	C9-O1P$^{\alpha}$	90.18%
10	Gln645-NE2$^{(11)}$	T10-O1P$^{\alpha}$	88.19%	Gln645-NE2$^{(11)}$	T10-O1P$^{\alpha}$	88.85%
11	Gln679-NE2$^{(12)}$	A11-O2P$^{\alpha}$	73.21%	Gln679-NE2$^{(12)}$	A11-O2P$^{\alpha}$	88.52%
12	Gln713-NE2$^{(13)}$	A12-O2P$^{\alpha}$	88.69%	Gln713-NE2$^{(13)}$	A12-O2P$^{\alpha}$	81.70%
13				Gln747-NE2$^{(14)}$	C13-O1P$^{\alpha}$	57.90%
14	**Asp743-OD2$^{(14)}$**	**C14-N4$^{\alpha}$**	**64.73%**	**Asp743-OD2$^{(14)}$**	**C14-N4$^{\alpha}$**	**85.69%**
	Gln781-OE2$^{(15)}$	**C14-O3′$^{\alpha}$**	**32.78%**			
15	***Asp743-OD2*$^{(14)}$**	***C15-N4*$^{\alpha}$**	***60.12*%**	**Asp777-OD1$^{(15)}$**	**C15-N4$^{\alpha}$**	**61.40%**
				Asp777-OD2$^{(15)}$	**C15-N4$^{\alpha}$**	**37.27%**
	Lys814-NZ$^{(16)}$	C15-O1P$^{\alpha}$	57.90%	Lys814-NZ$^{(16)}$	C15-O2P$^{\alpha}$	99.83%
16				Lys848-NZ$^{(17)}$	T16-O1P$^{\alpha}$	83.86%
17				**Gly811-O$^{(16)}$**	**C17-N4$^{\alpha}$**	**88.02%**
				Gly881-N$^{(17.5)}$	C17-O1P$^{\alpha}$	35.62%
19				**Asp845-OD1$^{(17)}$**	**A19-N6$^{\alpha}$**	**43.43%**

^{id}The index of a repeat that a residue belongs to.
*HDO is the abbreviation of hydrogen bond occupancy.
$^{\alpha}$DNA base belonging to the sense strand of DNA.
Hydrogen bonds in bold and nonbold reflect the specific and nonspecific interactions, respectively. Bold in italics denotes the specific and water-mediated hydrogen bonds.

truncated half repeat in all natural TALEs [1], but the role of this last half repeat is not clear in the specific binding process of TALE-DNA. Our study indicates that the last half repeat helps to stabilize a compact conformation at the TALE-DNA interface and then indirectly facilitates the specific interactions between TAL repeats and nucleotide bases. Therefore, the last half repeat is required for improving the recognition efficiency of specific DNA sequences by TALE.

4. Conclusions

In this study, MD simulations were performed to investigate the role of the last half repeat in the recognition and binding of TALE-DNA. The simulated result indicated that the stability of the modeled system (having the last half repeat) is higher than that of the nonmodeled system (lacking the last half repeat). The PCA analysis revealed that the AvrBs3 structure of the nonmodeled system is more extended in comparison with the crystallographic data. In contrast, the AvrBs3 of the modeled system still keeps the structural compactness. According to the previous experimental studies, this increase of the structural compactness of TALE is associated with the DNA binding. We also compared DNA groove parameters of the two systems. As a whole, the DNA major groove of the modeled system is relatively wider, which allows the side chain of the key amino acid of protein to be more accessible to nucleotide bases. It was suggested that the protein-DNA binding efficiency of the modeled system may

be relatively higher. Then, we calculated the hydrogen bonds at the protein-DNA interface. Comparatively, the nonmodeled system loses a considerable number of hydrogen bonds. The modeled system still keeps relatively stable protein-DNA binding. These additional interactions are mainly formed by the N- and C-terminal repeats. In particular, the last half repeat stabilizes the phosphate binding with DNA at the C-terminus and then helps to adopt a compact conformation at the protein-DNA interface. This compact conformation improves the specific recognition efficiency between TAL repeats and nucleotide bases. Our study reveals the important role of the last half repeat in high-efficient recognition of the DNA target sequence by TALE. It provides a deeper understanding of the recognition mechanism of TALE-DNA.

Competing Interests

The authors have declared that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (31600591), the Science and Technology Planning Project of Guangdong Province (2016A020210087, 2015A020224038, 2015A020209124, 2015B010131015, 2014A030308008, and 2014A050503057), the Science and Technology Planning Project of Guangzhou (1563000117), and Special Program for Applied Research on Super Computation of the NSFC-Guangdong Joint Fund (the second phase).

References

[1] J. Boch and U. Bonas, "*Xanthomonas* AvrBs3 family-type III effectors: discovery and function," *Annual Review of Phytopathology*, vol. 48, pp. 419–436, 2010.

[2] J. Boch, H. Scholze, S. Schornack et al., "Breaking the code of DNA binding specificity of TAL-type III effectors," *Science*, vol. 326, no. 5959, pp. 1509–1512, 2009.

[3] M. J. Moscou and A. J. Bogdanove, "A simple cipher governs DNA recognition by TAL effectors," *Science*, vol. 326, no. 5959, p. 1501, 2009.

[4] M. Christian, T. Cermak, E. L. Doyle et al., "Targeting DNA double-strand breaks with TAL effector nucleases," *Genetics*, vol. 186, no. 2, pp. 757–761, 2010.

[5] A. J. Bogdanove and D. F. Voytas, "TAL effectors: customizable proteins for DNA targeting," *Science*, vol. 333, no. 6051, pp. 1843–1846, 2011.

[6] J. C. Miller, S. Tan, G. Qiao et al., "A TALE nuclease architecture for efficient genome editing," *Nature Biotechnology*, vol. 29, no. 2, pp. 143–150, 2011.

[7] M. T. Murakami, M. L. Sforça, J. L. Neves et al., "The repeat domain of the type III effector protein PthA shows a TPR-like structure and undergoes conformational changes upon DNA interaction," *Proteins: Structure, Function, and Bioinformatics*, vol. 78, no. 16, pp. 3386–3395, 2010.

[8] D. Deng, C. Yan, X. Pan et al., "Structural basis for sequence-specific recognition of DNA by TAL effectors," *Science*, vol. 335, no. 6069, pp. 720–723, 2012.

[9] A. N.-S. Mak, P. Bradley, R. A. Cernadas, A. J. Bogdanove, and B. L. Stoddard, "The crystal structure of TAL effector PthXo1 bound to its DNA target," *Science*, vol. 335, no. 6069, pp. 716–719, 2012.

[10] S. Stella, R. Molina, I. Yefimenko et al., "Structure of the AvrBs3-DNA complex provides new insights into the initial thymine-recognition mechanism," *Acta Crystallographica Section D: Biological Crystallography*, vol. 69, no. 9, pp. 1707–1716, 2013.

[11] J. Streubel, C. Blücher, A. Landgraf, and J. Boch, "TAL effector RVD specificities and efficiencies," *Nature Biotechnology*, vol. 30, no. 7, pp. 593–595, 2012.

[12] J. F. Meckler, M. S. Bhakta, M.-S. Kim et al., "Quantitative analysis of TALE-DNA interactions suggests polarity effects," *Nucleic Acids Research*, vol. 41, no. 7, pp. 4118–4128, 2013.

[13] A. Richter, J. Streubel, C. Blücher et al., "A TAL effector repeat architecture for frameshift binding," *Nature Communications*, vol. 5, article 3447, 2014.

[14] P. Bradley, "Structural modeling of TAL effector-DNA interactions," *Protein Science*, vol. 21, no. 4, pp. 471–474, 2012.

[15] J. Grau, A. Wolf, M. Reschke, U. Bonas, S. Posch, and J. Boch, "Computational predictions provide insights into the biology of TAL effector target sites," *PLoS Computational Biology*, vol. 9, no. 3, Article ID e1002962, 20 pages, 2013.

[16] L. Cong, R. H. Zhou, Y.-C. Kuo, M. Cunniff, and F. Zhang, "Comprehensive interrogation of natural TALE DNA-binding modules and transcriptional repressor domains," *Nature Communications*, vol. 3, article 968, 2012.

[17] H. Wan, J.-P. Hu, K.-S. Li, X.-H. Tian, and S. Chang, "Molecular dynamics simulations of DNA-free and DNA-bound TAL effectors," *PLoS ONE*, vol. 8, no. 10, Article ID e76045, 2013.

[18] B. I. M. Wicky, M. Stenta, and M. Dal Peraro, "TAL effectors specificity stems from negative discrimination," *PLoS ONE*, vol. 8, no. 11, Article ID e80261, 9 pages, 2013.

[19] H. Flechsig, "TALEs from a spring-superelasticity of Tal effector protein structures," *PLoS ONE*, vol. 9, no. 10, Article ID e109919, 2014.

[20] C.-K. Zheng, C.-L. Wang, X.-P. Zhang, F.-J. Wang, T.-F. Qin, and K.-J. Zhao, "The last half-repeat of transcription activator-like effector (TALE) is dispensable and thereby TALE-based technology can be simplified," *Molecular Plant Pathology*, vol. 15, no. 7, pp. 690–697, 2014.

[21] W. Humphrey, A. Dalke, and K. Schulten, "VMD: visual molecular dynamics," *Journal of Molecular Graphics*, vol. 14, no. 1, pp. 27–38, 1996.

[22] J. C. Phillips, R. Braun, W. Wang et al., "Scalable molecular dynamics with NAMD," *Journal of Computational Chemistry*, vol. 26, no. 16, pp. 1781–1802, 2005.

[23] K. Vanommeslaeghe, E. Hatcher, C. Acharya et al., "CHARMM general force field: a force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields," *Journal of Computational Chemistry*, vol. 31, no. 4, pp. 671–690, 2010.

[24] J.-P. Ryckaert, G. Ciccotti, and H. J. C. Berendsen, "Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes," *Journal of Computational Physics*, vol. 23, no. 3, pp. 327–341, 1977.

[25] T. Darden, D. York, and L. Pedersen, "Particle mesh Ewald: an $N\cdot\log(N)$ method for Ewald sums in large systems," *Journal of Chemical Physics*, vol. 98, pp. 10089–10092, 1993.

[26] T. Hatano and S. Sasa, "Steady-state thermodynamics of Langevin systems," *Physical Review Letters*, vol. 86, no. 16, pp. 3463–3466, 2001.

[27] G. G. Maisuradze, A. Liwo, and H. A. Scheraga, "Relation between free energy landscapes of proteins and dynamics," *Journal of Chemical Theory and Computation*, vol. 6, no. 2, pp. 583–595, 2010.

[28] H. Wan, J.-P. Hu, X.-H. Tian, and S. Chang, "Molecular dynamics simulations of wild type and mutants of human complement receptor 2 complexed with C3d," *Physical Chemistry Chemical Physics*, vol. 15, no. 4, pp. 1241–1251, 2013.

[29] H. Wan, S. Chang, J.-P. Hu, Y.-X. Tian, and X.-H. Tian, "Molecular dynamics simulations of ternary complexes: comparisons of LEAFY protein binding to different DNA motifs," *Journal of Chemical Information and Modeling*, vol. 55, no. 4, pp. 784–794, 2015.

[30] D. Van Der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, and H. J. C. Berendsen, "GROMACS: fast, flexible, and free," *Journal of Computational Chemistry*, vol. 26, no. 16, pp. 1701–1718, 2005.

[31] R. Lavery, M. Moakher, J. H. Maddocks, D. Petkeviciute, and K. Zakrzewska, "Conformational analysis of nucleic acids revisited: curves+," *Nucleic Acids Research*, vol. 37, no. 17, pp. 5917–5929, 2009.

[32] C. Oguey, N. Foloppe, and B. Hartmann, "Understanding the sequence-dependence of DNA groove dimensions: implications for DNA interactions," *PLoS ONE*, vol. 5, no. 12, Article ID e15931, 2010.

Hindawi Publishing Corporation
BioMed Research International
Volume 2016, Article ID 9795409, 8 pages
http://dx.doi.org/10.1155/2016/9795409

Research Article

Networks Models of Actin Dynamics during Spermatozoa Postejaculatory Life: A Comparison among Human-Made and Text Mining-Based Models

Nicola Bernabò,[1] Alessandra Ordinelli,[1] Marina Ramal Sanchez,[1] Mauro Mattioli,[1,2] and Barbara Barboni[1]

[1] *Faculty of Veterinary Medicine, University of Teramo, Via Renato Balzarini 1, 64100 Teramo, Italy*
[2] *Istituto Zooprofilattico Sperimentale dell'Abruzzo e del Molise "G. Caporale", Campo Boario, 64100 Teramo, Italy*

Correspondence should be addressed to Nicola Bernabò; nbernabo@unite.it

Received 27 May 2016; Revised 26 July 2016; Accepted 27 July 2016

Academic Editor: Guang Hu

Here we realized a networks-based model representing the process of actin remodelling that occurs during the acquisition of fertilizing ability of human spermatozoa (HumanMade_ActinSpermNetwork, HM_ASN). Then, we compared it with the networks provided by two different text mining tools: Agilent Literature Search (ALS) and PESCADOR. As a reference, we used the data from the online repository Kyoto Encyclopaedia of Genes and Genomes (KEGG), referred to the actin dynamics in a more general biological context. We found that HM_ALS and the networks from KEGG data shared the same scale-free topology following the Barabasi-Albert model, thus suggesting that the information is spread within the network quickly and efficiently. On the contrary, the networks obtained by ALS and PESCADOR have a scale-free hierarchical architecture, which implies a different pattern of information transmission. Also, the hubs identified within the networks are different: HM_ALS and KEGG networks contain as hubs several molecules known to be involved in actin signalling; ALS was unable to find other hubs than "actin," whereas PESCADOR gave some nonspecific result. This seems to suggest that the human-made information retrieval in the case of a specific event, such as actin dynamics in human spermatozoa, could be a reliable strategy.

1. Introduction

Postgenomic era offers to researchers amazing opportunities in approaching a myriad of biological problems. One of the most interesting issues is the use of computational models for representing and analysing complex biological systems. They make researchers able to face important problems, such as those arising from the availability of a huge amount of data to be analysed (the so-called big data challenge) and from the creation of new information from the already available published data. This last issue, on one hand, is very timely and offers fascinating horizons, whereas, on the other one hand, it requires further studies to verify the reproducibility and the reliability of the obtained data. In this context, here we focused our attention on a biological event, which has a great importance in spermatology and in applied andrology: the dynamics of actin during the postejaculatory life of male gametes. Indeed, immediately after ejaculation, mammalian spermatozoa are virtually unable to fertilize the homologous oocyte. They become fully fertile only after they reside for hours to days within the female genital tract, where they complete a complex process of functional maturation known as capacitation. During capacitation spermatozoa biochemical machinery changes its function as a result of the dialogue between male gametes and female environment (tubal epithelium, tubal fluid, and female endocrine axis). The ionic intracellular concentration of ions changes, the protein phosphorylation is modified, sperm motility becomes hyperactivated, and plasma membrane (PM) and outer acrosome membrane (OAM) became gradually more fluid and tend to fuse each other. In this context, to date, it is believed that immediately after ejaculation the actin present in sperm head is mainly in globular unpolymerized form (G-actin). As the capacitation progresses, the actin undergoes polymerization,

forming a network of F-actin that interposes between outer acrosome membrane (OAM) and plasma membrane (PM), thus avoiding their premature fusion. When the physiological stimulus of acrosome reaction, the zona pellucida proteins, is met, this diaphragm is destroyed and the two membranes can fuse. Recently it has been suggested that the role of actin dynamic in this context could go beyond the merely mechanical function, but that this protein could be involved in the pathway as an active signal transducer [1].

From this point of view, it will be very interesting to have available a computational model of actin dynamics during the postejaculatory life of spermatozoa. At the present, a specific model devoted to the representation of actin dynamics during capacitation life is not already available; thus we carried out a study comparing a new model based on the manual compilation of a database, analogously to other database that we have already realized [2, 3] with ones obtained by a text mining-based approach. We paid our attention to text mining because it represents a new, important, and fascinating resource for information retrieval [4] and for constructing interaction network from biomedical texts [5]. Recently, this approach has been adopted to explore the biology of different phenomena, such as the prostate cancer protein interaction network, by using a reinforcement learning-based algorithm [5], or in studying other types of tumours [6–9] and physiological [10–12] and pathological events [13–15]. Here, in detail, we realize a model, starting from the analysis of published literature on this topic and we compared it with models realized by two different text mining tools, able to produce networks: Agilent Literature Search and PESCADOR. As a reference, we used the data from the online repository KEGG (Kyoto Encyclopaedia of Genes and Genomes), which are referred to the actin polymerization and depolymerisation in a wide variety of cells and not specifically to the spermatozoa.

2. Materials and Methods

2.1. Data Collection, Network Creation, and Analysis

2.1.1. Human-Made Spermatozoa Actin Network (HM_SAN). In this work, we used different networks. The first was realized by considering the scientific literature published in peer-reviewed international papers indexed in PubMed archive (http://www.ncbi.nlm.nih.gov/pubmed/) in the last 15 years [2, 3]. As reference, we used the data referred to human species. Following an already validated protocol [16], two researchers expert on spermatozoa biology carried out an independent literature analysis on papers using the following key words: "Actin polymerization", "Actin depolymerisation", "Actin dynamics", and "Actin remodelling". Then, the two databases have been compared, and a third researcher verified the correctness of the record inserted and resolved eventual conflicts. The freely available and diffusible molecules such as H_2O, CO_2, P_i, H^+, and O_2 were omitted, when not necessary, and in some cases the record did not represent a single molecule but a complex event, such as "protein tyrosine phosphorylation" because all the single molecular determinants of the phenomenon are still unknown [10, 17].

This database (interaction database), was realized in Microsoft Excel 2013 and contained the following fields:

(i) *Source molecule*: here are reported the molecules source of the interaction.

(ii) *Interaction*: here is described what kind of interaction the molecules carry out.

(iii) *Target molecule:* here are reported the molecules that are target of the interaction.

(iv) *Alias*: eventual aliases are described.

(v) *Role*: the physiological and/or pathological role of the molecule in epididymis is reported.

(vi) *Reference*: it represents the paper reporting the above mentioned data.

(vii) *Notes*: any further information that could be useful in the study is mentioned here.

2.1.2. Agilent Literature Search-Spermatozoa Actin Network (ALS_SAN). This network was realized by using Agilent Literature Search Software, a metasearch tool for automatically querying multiple text-based search engines that can be used in conjunction with Cytoscape, thus generating a network view of protein associations. In particular, we used the Cytoscape 3.3.0. App Agilent Literature Search 3.1.1 beta (LitSearch version 2.69), using as data source the papers contained in PubMed database. As key words, we used the same key words used to build HM_ASN, using as context "spermatozoa". Max Engine Matches was set at 1.000 (which always was higher than the number of articles found; thus in all the cases all the available information was processed); the "Use Aliases," the "Use Context," and the "Concept Lexicon Restrict Search" options were set. As Concept Lexicon "Homo sapiens" we used. The data have been accessed until April 15, 2016. We created ALS_SAN by merging all the obtained networks and removing self-loops and the duplicated edges [10].

2.1.3. PESCADOR-Spermatozoa Actin Network (P_SAN). This network was created by using PESCADOR (Platform for Exploration of Significant Concepts AssociateD to co-Occurrence Relationships), which is a platform independent web resource (http://cbdm.mdc-berlin.de/tools/pescador/) [18]. It analyses a query composed of a list of PMIDs to be scanned for gene/protein cooccurrences and, optionally, of a list of words (ideally, biological concepts related to protein interactions, such as "aggregation" or "phosphorylation") to be found in the cooccurrence analysis, as text mining engine to extract sentences with cooccurring bioentities from the text of the PubMed abstracts requested that it uses LAITOR (Barbosa altro). P_SAN was created by using the list of PMIDs of the papers we have manually selected for the realization of HM_SAN.

Table 1: Main topological parameters assessed in this study.

Parameter	Definition
Connected components	It is the number of networks in which any two vertices are connected to each other by links and which is connected to no additional vertices in the network.
Number of nodes	It is the total number of molecules involved.
Number of edges	It is the total number of interactions found.
Clustering coefficient	It is calculated as $CI = 2nI/kI(kI - 1)$, where nI is the number of links connecting the kI neighbors of node I to each other. It is a measure of how the nodes tend to form clusters.
Network diameter	It is the longest of all the calculated shortest paths in a network.
Shortest paths	The length of the shortest path between two nodes n and m is $L(n, m)$. The shortest path length distribution gives the number of node pairs (n, m) with $L(n, m) = k$ for $k = 1, 2, \ldots$.
Characteristic path length	It is the expected distance between two connected nodes.
Averaged number of neighbors	It is the mean number of connections of each node.
Node degree	It is the number of interactions of each node.
Node degree distribution	It represents the probability that a selected node has k links.
γ	Exponent of node degree equation.
R^2	Coefficient of determination of node degree versus number of nodes, on logarithmized data.

2.2. KEGG_AN. This network, used as reference, has been created by importing the data from KEGG (Kyoto Encyclopaedia of Genes and Genomes), a database resource for understanding high-level functions and utilities of the biological system, and from molecular-level information, especially large-scale molecular datasets generated by genome sequencing and other high-throughput experimental technologies (http://www.genome.jp/kegg/). We analysed the data from the pathway: map04810—regulation of actin cytoskeleton. This network is not specifically designed to represent the actin dynamics occurring during sperm capacitation, but it is generically referred to the actin cytoskeleton rearrangement. It was used to compare the other networks with a network representing a strongly related biological event and certified by a rigorous quality control [19, 20].

2.3. Networks Visualization and Analysis. All these networks have been realized, visualized, and analysed using Cytoscape 3.1.2 [21]. The analysis was carried out considering the networks as undirected and assessing the topological parameters listed and described in Table 1.

To represent the nodes as Venn's diagram, we used Venny, a specific tool, available at http://bioinfogp.cnb.csic.es/tools/venny/.

2.4. Network Randomization. To compare our networks with a computer-generated network following the Barabasi-Albert model, we used the Cytoscape plug-in Network Randomizer 1.1 (http://apps.cytoscape.org/apps/networkrandomizer). We used the Barabasi-Albert model and we set the parameters $N = 128$ and $m = 2$. We obtained the Barabasi_Albet random network (BA_RN) constituted by 2 connected components of, respectively, 125 (main component, BA_RN, and MC_BA_RN) and 3 nodes.

3. Results

We obtained five different networks: HM_SAN, P_SAN, ALS_SAN, KEGG_AN, and BA_RN. The results of their topological analyses are shown in Table 2, where the values of main topological parameters are listed. In the case of the network obtained by using PESCADOR, we found that it contained several nonspecific nodes (such as "acrosome", "spermatozoa", "membrane", and "in vitro"). After their removal, we obtained P_ASN and a second network, its main connected component, MC_P_ASN. Also in the case of ASL_SAN, KEGG_AN, and BA_RN we extracted the main connected components (MC_ALS_SAN, MC_KEGG_SAN, and MC_BA_RN). In Table 3 are reported the results of the fitting of node degree versus the number of nodes. In Table 4 are shown the results of the correlation analysis between the node degree and the clustering coefficient of all the networks. In Table 5 are listed the hubs of the networks. In Supplementary Material (available online at http://dx.doi.org/10.1155/2016/9795409) are listed the articles we used to build our database and those used by ALS, highlighting the common ones.

4. Discussion

Here, we realized a network representing actin dynamics during sperm capacitation (HM_ASN); then we compared it with two networks generated by two text mining software, able to directly provide networks models (P_ASN and ALS_SAN). As reference we used a peer-reviewed and quality controlled network (KEGG_AN) related to the same biological event, but it referred to a more general context and a Barabasi-Albert scale-free network generated by the computer (BA_RN). See Figure 1. From our analysis, it is clear that HM_SAN has a scale-free topology, in keeping

Table 2: Results of topological analyses of networks.

Parameter	P_ASN	MC_P_ASN	ALS_ASN	MC_ALS_ASN	KEGG_AN	MC_KEGG_AN	HM_ASN	BA_RN	MC_BA_RN
Connected components	10	1	7	1	23	1	1	2	1
Number of nodes	136	109	86	66	84	60	128	128	125
Number of edges	283	234	161	141	75	73	187	196	194
Clustering coefficient	0.577	0.552	0.637	0.656	0.017	0.024	0.073	0.024	0.025
Network diameter	9	9	5	5	9	9	10	10	10
Shortest paths	11624 (63%)	8230 (69%)	4344 (59%)	4290 (100%)	3546 (50%)	3540 (100%)	16256 (100%)	13812 (95%)	13806 (100%)
Characteristic path length	3.799	3.876	2.332	2.346	4.294	4.299	4.064	4.440	4.441
Avg. number of neighbors	4.162	4.294	3.744	4.273	1.786	2.433	2.921	2.975	3.017

Table 3: Results of node degree analysis of networks.

	P_ASN	MC_P_ASN	ALS_ASN	MC_ALS_ASN	KEGG_AN	MC_KEGG_AN	HM_ASN	BA_RN	MC_BA_RN
γ	−1.348	−0.941	−0.881	−0.741	−1.596	−1.540	−1.459	−1.317	−1.299
r	0.741	0.825	0.862	0.790	0.530	0.494	0.979	0.895	0.860
R^2	0.736	0.546	0.671	0.617	0.799	0.778	0.860	0.805	0.780

Table 4: Results of fitting on node degree versus clustering coefficient.

	P_ASN	MC_P_ASN	ALS_ASN	MC_ALS_ASN	KEGG_AN	MC_KEGG_AN	HM_ASN	BA_RN	MC_BA_RN
γ	−0.763	−0.479	−0.915	−0.921	−0.178	−0.198	−0.490	−0.640	−0.663
r	0.737	0.662	0.704	0.708	0.121	0.128	0.647	0.414	0.438
R^2	0.477	0.342	0.810	0.815	0.030	0.038	0.505	0.482	0.391

Table 5: Hubs of the networks.

HM_ASN	MC_P_ASN	MC_ALS_ASN	KEGG_ASN
PKA	RHOA	ACTIN	RAC1
Actin polymerization	MSP		ROCK1
Tyrosine phosphorylation	EGFR		PAK4
$[Ca^{2+}]_i$	LIMK		RHOA
cAMP	CDC42		CDC42
ROS	GNA13		ACTIN
Actin depolymerisation	ROCK2		ARHGEF7
F-actin	LIMK2		MYL12B
PLD	ACE		RRAS2
Rho GTPase	AKAP4		
H_2O_2	AKAP3		
PIP2 cleavage	PRKAR2		
Arp2/3 complex	ROPN1		
ADF/cofilin			
EGFR			
HCO_3^-			
PKC			

with the Barabasi-Albert model. Indeed, it is very close to BA_RN and it has the node degree (i.e., the number of nodes per link) probability distribution following an exponential law with a negative exponent and uncorrelated with the clustering coefficient (which represents the network tendency to develop clusters). In addition, the network has a small world topology, as evident from the values of shortest paths (100%), characteristic path length, and averaged number of neighbors (4.064 and 2.921, resp.). These measures suggest that the information is spread within the network in a very fast and efficient way and that the network is able to quickly adapt to the external perturbations. In particular, the low value of clustering coefficient indicates that loop or clusters, that could interfere and slow the propagation of messages, are virtually absent in HM_ASN. KEGG_SN has virtually the same topology of HM_ASN, thus suggesting that the network we created could be representative of a similar biological event, and that this pattern could be typical of signalling pathways. This finding is in accordance with those we have found when analysing several other networks referred either to sperm signalling or to other biologically relevant events. Indeed, recently, we compared the networks representing the biochemical machinery involved in spermatozoa in sea urchin, *Caenorhabditis elegans*, and human male gametes, with networks representing ten pathways of relevant physiopathological importance and with a computer-generated network [22]. As a result, we have found that all the networks studied are characterized by robustness against random failure, controllability, and efficiency in signal transmission. In all the cases, the clustering coefficient had values near zero [22]. Interestingly, the two networks generated by the text mining software have a different topology. Both of them are characterized by a lower absolute value of exponent of node degree distribution (see Figure 2) and by a higher value of clustering coefficient, whose distribution correlated with the node degree, as shown in Table 4. Then they could be considered hierarchical networks. This finding highlights that ALS and PESCADOR seem to give results not completely comparable with those from manual compilation of databases. This idea is also highly strengthened by the analysis of networks hubs. Indeed, the scale-free topology of all the networks allows one to identify the nodes exerting the higher level of control within the network, the hubs, calculated as the nodes with a node degree with a degree at least one standard deviation above the network mean [23]. As it is reported in Table 5 we found great differences either in number or in identity among the hubs from the different networks. Interestingly the only hub shared by all the networks is "actin" (see Figure 3).

The hubs of HM_ASN are F-actin and complex events related to the signal transduction pathway involved in actin remodelling occurring during the process of spermatozoa acquisition of fertilizing ability such as "Actin polymerization" and "Actin depolymerization", or proteins "Tyrosine phosphorylation". In addition we have identified as hubs several molecules involved in input of control messages (EGFR, H_2O_2, and HCO_3^-), second messengers ($[Ca^{2+}]_i$, cAMP, ROS, and PIP2 cleavage), and effector molecules (PKC, PLD, Rho GTPase, Arp2/3 complex, and ADF/cofilin). This finding

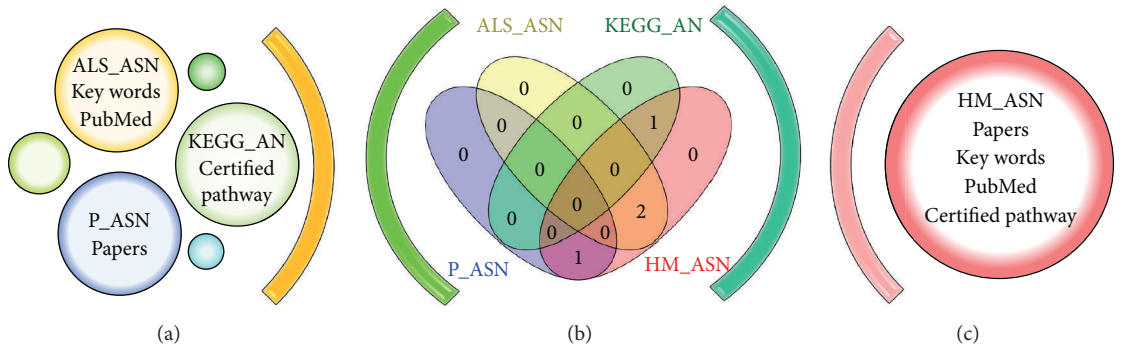


Figure 1: (a) In yellow are represented the search parameters that our database, HM_ASN, has in common with the ALS_ASN software, in blue with P_ASN software (two text mining tools), and in green with the KEGG database. (b) Venn diagram representation, which allowed the identification of the intersections between different databases other than HM_ASN. (c) In red are represented the elements used to create the database HM_ASN.

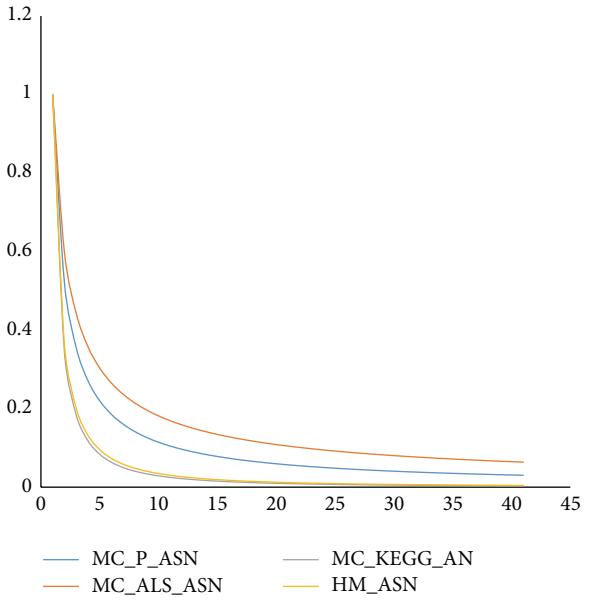


Figure 2: Curves that represent the node degree distribution in HM_SAN, MC_P_ASN, MC_ALS_ASN, and MC_KEGG_AN.

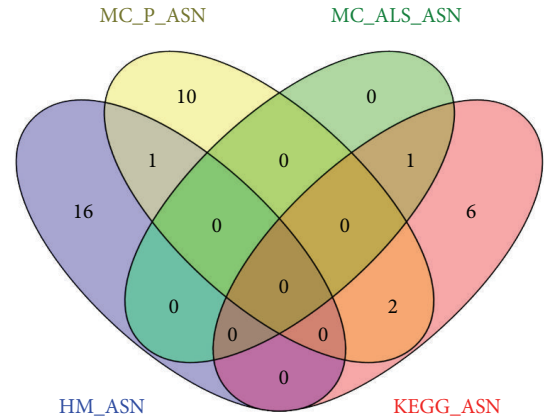


Figure 3: Venn's diagram showing the common hubs in HM_SAN, MC_P_ASN, MC_ALS_ASN, and MC_KEGG_AN.

is in agreement with the currently proposed model of actin signalling transduction pathway active in human and mammalian spermatozoa, based on experimental data. Indeed, the actin dynamics occurring during capacitation and acrosome reaction are under the control of several activating factors. The most important extracellular activating messenger is thought to be the bicarbonate [24–27], which is able to enter the cells and to stimulate the production of cAMP, via the activation of a specific soluble adenylyl cyclase (sAC). The rise in intracellular level of cAMP leads to the increase in membrane scrambling and directly or indirectly causes the increase in cytosolic concentration of the other second messengers: Ca^{2+} [28, 29], cAMP [30], ROS [17, 31, 32], and DAG and IP3 (resulting from PIP2) [33, 34]. This promotes the activation of a myriad of cellular effectors that directly and indirectly control the actin polymerization status [35, 36].

In particular, it has been demonstrated that PKA, PKC, and PLD1 play a key role in modulating the actin polymerization/depolymerisation status [35, 37, 38]. KEGG_ASN contains several proteins involved in cell signalling, such as RAC1, ROCK1, PAK4, RHOA, CDC42, ARHGEF7, MYL12B, and RRAS2, and virtually all those involved in Rho signalling and, of course, it is known to participate in actin cytoskeleton remodelling (see for reference [39]).

More interestingly, ALS_SAN contains only one hub: actin. This could be explained with the search logic of ALS that, likely, is able to consider only the molecules directly interacting with actin, thus excluding from the results indirect relationships, which were instead took into account by human database compilers. This reason will explain also the hierarchical structure network. We examined also the papers identified by human manual compilers of database and those identified by ALS. We have found 26 papers related to the used key words and published in last 15 years suitable to be used to gather information about actin dynamics. ALS identified 31 papers, 4 of which have been published before this range of time; see Supplementary Material. Twelve papers have been identified by both the systems; the others differ.

This difference could be, in our opinion, explained with two hypothesis:

(i) Human compilers discarded similar papers (mainly reviews) from the same group, using only the most recent ones.

(ii) Human compilers included also papers, which did not have "actin" in key words, expanding the selection criteria.

PESCADOR gives a high number of hubs, actually corresponding to proteins involved in actin signalling. Curiously, it considers also MSP, the Major Sperm Protein, which is involved in spermatozoa cytoskeleton signalling, but in Nematoda that lack actin [40].

5. Conclusions

In conclusion, we could affirm that

(i) HM_ASN and KEGG_AN are very similar, in terms of topology; this could suggest that the human information retrieval in the case of a specific event, such as actin dynamics during mammalian spermatozoa, could be a reliable strategy;

(ii) PESCADOR seems to give nonspecific results that need to be manually removed from the model; thus the reliability of their results needs to be improved;

(iii) ALS tends to be less "elastic" than human retrieval; indeed it collects only the data strictly related to the actin, leaving out the molecules indirectly interacting with actin.

It is possible to hypothesize that when searching for a very specific query the human bases research could offer more reliable data, in comparison with text mining tools. Likely, these systems could be needed when the number of papers to be checked is larger.

Competing Interests

The authors declare that there are no competing interests regarding the publication of this paper

Acknowledgments

Marina Ramal Sanchez is granted by MarieSkłodowska-Curie ITN REP-BIOTECH 675526, European Joint Doctorate in Biology and Technology of the Reproductive Health.

References

[1] N. Bernabò, P. Berardinelli, A. Mauro et al., "The role of actin in capacitation-related signaling: an in silico and in vitro study," *BMC Systems Biology*, vol. 5, article 47, 2011.

[2] N. Bernabò, M. Mattioli, and B. Barboni, "The spermatozoa caught in the net: the biological networks to study the male gametes post-ejaculatory life," *BMC Systems Biology*, vol. 4, article 87, 2010.

[3] N. Bernabò, B. Barboni, and M. Maccarrone, "Systems biology analysis of the endocannabinoid system reveals a scale-free network with distinct roles for anandamide and 2-arachidonoylglycerol," *OMICS: A Journal of Integrative Biology*, vol. 17, no. 12, pp. 646–654, 2013.

[4] N. Bernabò, A. Ordinelli, R. Di Agostino, M. Mattioli, and B. Barboni, "Network analyses of sperm-egg recognition and binding: ready to rethink fertility mechanisms?" *OMICS*, vol. 18, no. 12, pp. 740–753, 2014.

[5] F. Zhu, Q. Liu, X. Zhang, and B. Shen, "Protein interaction network constructing based on text mining and reinforcement learning with application to prostate cancer," *IET Systems Biology*, vol. 9, no. 4, pp. 106–112, 2015.

[6] P. F. Cheng, R. Dummer, and M. P. Levesque, "Data mining The Cancer Genome Atlas in the era of precision cancer medicine," *Swiss Medical Weekly*, vol. 145, Article ID w14183, 2015.

[7] H. Ye, X. Zhang, Y. Chen, Q. Liu, and J. Wei, "Ranking novel cancer driving synthetic lethal gene pairs using TCGA data," *Oncotarget*, 2016.

[8] L. Uttley, B. L. Whiteman, H. B. Woods, S. Harnan, S. T. Philips, and I. A. Cree, "Building the evidence base of blood-based biomarkers for early detection of cancer: a rapid systematic mapping review," *EBioMedicine*, 2016.

[9] J. Coates, L. Souhami, and I. El Naqa, "Big data analytics for prostate radiotherapy," *Frontiers in Oncology*, vol. 6, p. 149, 2016.

[10] N. Bernabò, A. Ordinelli, R. Di Agostino, M. Mattioli, and B. Barboni, "Network analyses of sperm-egg recognition and binding: ready to rethink fertility mechanisms?" *OMICS: A Journal of Integrative Biology*, vol. 18, no. 12, pp. 740–753, 2014.

[11] K. Seenprachawong, P. Nuchnoi, C. Nantasenamat, V. Prachayasittikul, and A. Supokawej, "Computational identification of miRNAs that modulate the differentiation of mesenchymal stem cells to osteoblasts," *PeerJ*, vol. 4, Article ID e1976, 2016.

[12] A. Mohammadnia, M. Yaqubi, F. Pourasgari, E. Neely, H. Fallahi, and M. Massumi, "Signaling and gene regulatory networks governing definitive endoderm derivation from pluripotent stem cells," *Journal of Cellular Physiology*, vol. 231, no. 9, pp. 1994–2006, 2016.

[13] D. E. Jones, H. Ghandehari, and J. C. Facelli, "A review of the applications of data mining and machine learning for the prediction of biomedical properties of nanoparticles," *Computer Methods and Programs in Biomedicine*, vol. 132, pp. 93–103, 2016.

[14] T. B. Ho, L. Le, D. T. Thai, and S. Taewijit, "Data-driven approach to detect and predict adverse drug reactions," *Current Pharmaceutical Design*, vol. 22, no. 23, pp. 3498–3526, 2016.

[15] D. Svenstrup, H. L. Jørgensen, and O. Winther, "Rare disease diagnosis: a review of web search, social media and large-scale data-mining approaches," *Rare Diseases*, vol. 3, no. 1, Article ID e1083145, 2015.

[16] B. B. Nicola Bernabò, R. Di Agostino, A. Ordinelli, and M. Mattioli, "The maturation of murine spermatozoa membranes within the epididymis, a computational biology perspective," *Systems Biology in Reproductive Medicine*, In press.

[17] N. Bernabò, L. Greco, A. Ordinelli, M. Mattioli, and B. Barboni, "Capacitation-related lipid remodeling of mammalian spermatozoa membrane determines the final fate of male gametes: A Computational Biology Study," *OMICS*, vol. 19, no. 11, pp. 712–721, 2015.

[18] A. Barbosa-Silva, J.-F. Fontaine, E. R. Donnard, F. Stussi, J. M. Ortega, and M. A. Andrade-Navarro, "PESCADOR, a web-based tool to assist text-mining of biointeractions extracted

from PubMed queries," *BMC Bioinformatics*, vol. 12, article 435, 2011.

[19] M. Kanehisa, S. Goto, Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe, "Data, information, knowledge and principle: back to metabolism in KEGG," *Nucleic Acids Research*, vol. 42, no. 1, pp. D199–D205, 2014.

[20] M. Kanehisa, Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe, "KEGG as a reference resource for gene and protein annotation," *Nucleic Acids Research*, vol. 44, no. D1, pp. D457–D462, 2016.

[21] P. Shannon, A. Markiel, O. Ozier et al., "Cytoscape: a software environment for integrated models of biomolecular interaction networks," *Genome Research*, vol. 13, no. 11, pp. 2498–2504, 2003.

[22] N. Bernabò, M. Mattioli, and B. Barboni, "Signal transduction in the activation of spermatozoa compared to other signalling pathways: a Biological Networks Study," *International Journal of Data Mining and Bioinformatics*, vol. 12, no. 1, pp. 59–69, 2015.

[23] O. Sporns, C. J. Honey, and R. Kötter, "Identification and classification of hubs in brain networks," *PLoS ONE*, vol. 2, no. 10, Article ID e1049, 2007.

[24] B. M. Gadella and R. A. P. Harrison, "Capacitation induces cyclic adenosine 3′,5′-monophosphate-dependent, but apoptosis-unrelated, exposure of aminophospholipids at the apical head plasma membrane of boar sperm cell," *Biology of Reproduction*, vol. 67, no. 1, pp. 340–350, 2002.

[25] T. Leahy and B. M. Gadella, "New insights into the regulation of cholesterol efflux from the sperm membrane," *Asian Journal of Andrology*, vol. 17, no. 4, pp. 561–567, 2000.

[26] L. Botto, N. Bernabò, P. Palestini, and B. Barboni, "Bicarbonate induces membrane reorganization and CBR1 and TRPV1 endocannabinoid receptor migration in lipid microdomains in capacitating boar spermatozoa," *Journal of Membrane Biology*, vol. 238, no. 1–3, pp. 33–41, 2010.

[27] B. Barboni, N. Bernabò, P. Palestini et al., "Type-1 cannabinoid receptors reduce membrane fluidity of capacitated boar sperm by impairing their activation by bicarbonate," *PLoS ONE*, vol. 6, no. 8, Article ID e23038, 2011.

[28] F. A. Navarrete, F. A. García-Vázquez, A. Alvau et al., "Biphasic role of calcium in mouse sperm capacitation signaling pathways," *Journal of Cellular Physiology*, vol. 230, no. 8, pp. 1758–1769, 2015.

[29] M. S. Rahman, W.-S. Kwon, and M.-G. Pang, "Calcium influx and male fertility in the context of the sperm proteome: an update," *BioMed Research International*, vol. 2014, Article ID 841615, 13 pages, 2014.

[30] B. M. Gadella and C. Luna, "Cell biology and functional dynamics of the mammalian sperm surface," *Theriogenology*, vol. 81, no. 1, pp. 74–84, 2014.

[31] R. J. Aitken, "The capacitation-apoptosis highway: oxysterols and mammalian sperm function," *Biology of Reproduction*, vol. 85, no. 1, pp. 9–12, 2011.

[32] R. J. Aitken, M. A. Baker, and B. Nixon, "Are sperm capacitation and apoptosis the opposite ends of a continuum driven by oxidative stress?" *Asian Journal of Andrology*, vol. 17, no. 4, pp. 633–639, 2015.

[33] N. Etkovitz, S. Rubinstein, L. Daniel, and H. Breitbart, "Role of PI3-kinase and PI4-kinase in actin polymerization during bovine sperm capacitation," *Biology of Reproduction*, vol. 77, no. 2, pp. 263–273, 2007.

[34] D. Ickowicz, M. Finkelstein, and H. Breitbart, "Mechanism of sperm capacitation and the acrosome reaction: role of protein kinases," *Asian Journal of Andrology*, vol. 14, no. 6, pp. 816–821, 2012.

[35] H. Breitbart, G. Cohen, and S. Rubinstein, "Role of actin cytoskeleton in mammalian sperm capacitation and the acrosome reaction," *Reproduction*, vol. 129, no. 3, pp. 263–268, 2005.

[36] L. Daniel, N. Etkovitz, S. R. Weiss, S. Rubinstein, D. Ickowicz, and H. Breitbart, "Regulation of the sperm EGF receptor by ouabain leads to initiation of the acrosome reaction," *Developmental Biology*, vol. 344, no. 2, pp. 650–657, 2010.

[37] G. Cohen, S. Rubinstein, Y. Gur, and H. Breitbart, "Crosstalk between protein kinase A and C regulates phospholipase D and F-actin formation during sperm capacitation," *Developmental Biology*, vol. 267, no. 1, pp. 230–241, 2004.

[38] H. Breitbart, T. Rotman, S. Rubinstein, and N. Etkovitz, "Role and regulation of PI3K in sperm capacitation and the acrosome reaction," *Molecular and Cellular Endocrinology*, vol. 314, no. 2, pp. 234–238, 2010.

[39] C. Guilluy, R. Garcia-Mata, and K. Burridge, "Rho protein crosstalk: another social network?" *Trends in Cell Biology*, vol. 21, no. 12, pp. 718–726, 2011.

[40] L. L. LeClaire III, M. Stewart, and T. M. Roberts, "A 48 kDa integral membrane phosphoprotein orchestrates the cytoskeletal dynamics that generate amoeboid cell motility in *Ascaris* sperm," *Journal of Cell Science*, vol. 116, no. 13, pp. 2655–2663, 2003.

Hindawi Publishing Corporation
BioMed Research International
Volume 2016, Article ID 2509385, 9 pages
http://dx.doi.org/10.1155/2016/2509385

Research Article

Comparison of FDA Approved Kinase Targets to Clinical Trial Ones: Insights from Their System Profiles and Drug-Target Interaction Networks

Jingyu Xu,[1,2] Panpan Wang,[1] Hong Yang,[1] Jin Zhou,[1] Yinghong Li,[1] Xiaoxu Li,[1] Weiwei Xue,[1] Chunyan Yu,[1] Yubin Tian,[2] and Feng Zhu[1]

[1]*Innovative Drug Research and Bioinformatics Group, School of Pharmaceutical Sciences and Innovative Drug Research Centre, Chongqing University, Chongqing 401331, China*
[2]*School of Mathematics and Statistics, Beijing Institute of Technology, Beijing 100081, China*

Correspondence should be addressed to Weiwei Xue; xueww@cqu.edu.cn and Feng Zhu; zhufeng@cqu.edu.cn

Received 31 January 2016; Revised 14 June 2016; Accepted 28 June 2016

Academic Editor: Filippo Pullara

Kinase is one of the most productive classes of established targets, but the majority of approved drugs against kinase were developed only for cancer. Intensive efforts were therefore exerted for releasing its therapeutic potential by discovering new therapeutic area. Kinases in clinical trial could provide great opportunities for treating various diseases. However, no systematic comparison between system profiles of established targets and those of clinical trial ones was conducted. The reveal of probable difference or shift of trend would help to identify key factors defining druggability of established targets. In this study, a comparative analysis of system profiles of both types of targets was conducted. Consequently, the systems profiles of the majority of clinical trial kinases were identified to be very similar to those of established ones, but percentages of established targets obeying the system profiles appeared to be slightly but consistently higher than those of clinical trial targets. Moreover, a shift of trend in the system profiles from the clinical trial to the established targets was identified, and popular kinase targets were discovered. In sum, this comparative study may help to facilitate the identification of the druggability of established drug targets by their system profiles and drug-target interaction networks.

1. Introduction

The human kinome (defined as the protein kinase complement of the human genome) provided a starting point for full-scale understanding of protein phosphorylation in normal and disease states and for a comprehensive discovery of the kinase target [1]. Phylogenetic tree of the human kinome revealed that kinase was one of the most productive classes of established therapeutic targets [2]. According to the latest reports [3, 4], 46 drugs targeting the human kinome have received approval by the US Food and Drug Administration (FDA), which include 35 small molecular drugs, 6 monoclonal antibodies, and 5 biologics. The targets of these 46 drugs had attracted extensive attentions from many pharmaceutical companies owing to their pivotal roles in not only cancers [5–8] but also other disease indications, such as central nervous system disorder, inflammation, and ophthalmology [4]. However, the majority (37 out of 46) of approved drugs against kinase were developed for treating cancer with only a few exceptions like metformin for diabetes and tofacitinib for rheumatoid arthritis [9, 10]. Intensive efforts were thus exerted for releasing the therapeutic potential of the human kinome by discovering new therapeutic area of established targets [11] or by identifying novel target from those undiscovered kinase families [4].

As an effective new way to reveal the multifactorial nature of disease, network medicine was proposed to discover new therapeutic area for the established targets [12]. Particularly, kinase was found to be capable of regulating diverse disease indications other than cancer by pathway affiliation and network analysis of drug-kinase interactions [13]. Moreover, the accelerated identification of novel drug targets, especially the clinical trial ones, provided more opportunities

for treating a variety of diseases [14, 15]. The clinical trial targets defined here refer to kinases that have not yet been utilized by FDA approved drugs but are under investigation in clinical trials. As reported, intensive efforts in the exploration of clinical trial target have dramatically extended the coverage of druggable families in the human kinome from the tyrosine kinase family to several other families like the calmodulin/calcium-regulated kinase, the glycogen synthase kinase (GSK), the cGMP-dependent protein kinase (PKG), the cAMP-dependent protein kinase (PKA), the CDC-like kinase (CLK), and the protein kinase C (PKC) [4, 10].

Although proteins in the human kinome demonstrated much closer homology relation to each other than to protein outside of kinase family, their sequence, structure, physicochemical properties, and many other characteristics vary significantly. As one of the most important properties reflecting the druggability of target, the system profile was frequently analyzed to evaluate the likelihood of a target to achieve therapeutic effects [16–18]. In particular, typical system profiles of a therapeutic target include the following: target affiliated signaling pathways, target subcellular locations, similarity proteins outside target's biochemical family, and level of sequence and structure similarities to the established drug targets [16–18]. Based on the system profiles of established drug targets, systems-level druggability rules were derived [16–18], which could be generalized as follows: targets similar to fewer human proteins outside of target family and associated with fewer human pathways tend to target drugs with reduced side-effects; efficacy drugs are more readily achieved by working on targets expressed in fewer tissues. In order to understand and evaluate the current trends in clinical trial development, it is of great interest to identify shift of trend between established targets and clinical trial ones from the system profiles' point of view. However, the system profiles of clinical trial kinase targets have not yet been analyzed, and no study of systematic comparison between the system profiles of established targets and that of clinical trial ones was conducted. Therefore, a comparison of system profiles would help to discover key factors defining the druggabilities of established targets [19–22].

In this study, a comparative analysis on the system profiles between established and clinical trial targets was conducted. Firstly, system profiles of these two types of targets were compared on 3 aspects: (1) the number of human proteins outside of the target families; (2) the number of target affiliated pathways; (3) the number of tissues the target is expressed in. Secondly, a reported combinational method predicting the promising targets by integrating multiple profiles (these system profiles, drug binding domain structural conformations, and protein physicochemical properties) of the target was further evaluated and discussed. Thirdly, the drug-target interaction networks were used to identify popular established and clinical trial kinase targets by both approved and clinical trial drugs.

2. Materials and Methods

2.1. Collection of FDA Approved and Clinical Trial Drugs Together with Their Kinase Targets. Firstly, 1,767 approved drugs were collected from the FDA official website (Drugs@FDA), and their corresponding primary therapeutic targets were matched from the Therapeutic Target Database (TTD) [3] or identified through extensive literature review (find more details in Sections 2 and 2.3), which resulted in 1,521 FDA approved drugs with 361 identifiable primary targets. Secondly, to make a comprehensive collection of clinical trial drugs, multiple resources were searched to collect more than 10,000 clinical trial drugs, which include the TTD [3], the PhRMA (http://www.phrma.org/) medicines in development, the drug pipeline reports from the websites, and annual reports of more than 150 pharmaceutical and biotechnology companies, as well as additional literature search. Thirdly, the clinical status of those clinical trial drugs was identified by the US National Institutes of Health's (NIH) ClinicalTrials.gov website (https://clinicaltrials.gov/) and the public announcements by the drug developers. As a result, ~6,000 clinical trial drugs with available clinical trial information against ~800 primary therapeutic targets were identified. Among these targets, ~500 were clinical trial targets that have not yet been utilized by approved drugs but are under investigation in clinical trials. Fourthly, the biochemical classes of established and clinical trial targets were collected from the UniProt database [23, 24]. Only drugs targeting protein kinase were analyzed in this study, which included 46 approved drugs against 25 established targets and 149 clinical trial drugs against 39 clinical trial targets.

2.2. System Profiles of Established and Clinical Trial Kinase Targets. Sequences of studied targets were downloaded by mapping their name to the UniProt database [23, 24]; pathway information was collected from the KEGG database [25] by crossmatching IDs of the UniProt database; tissue distribution information was collected from the TissueDistributionDBs [26] by querying using gene name of the targets. Moreover, similarity level among protein sequences were calculated by the tool of BLAST [27] which was provided by the US National Center for Biotechnology Information. Statistical comparison of system profiles were conducted by R software [28] and all figures were drawn in Microsoft Excel. In particular, the boxplot function in the basic package of R was applied in this study to draw the boxplot of system profiles among established and clinical trial (phase 3, phase 2, and phase 1) targets.

2.3. Identification of the Primary Therapeutic Targets of Approved and Clinical Trial Drugs. The primary therapeutic targets of approved and clinical trial drugs were identified by a well-established target validation process, which demands several key criteria [29, 30]. Firstly, targets of interest should be expressed in the disease-relevant cells or tissues. Secondly, the targets should be effectively modulated by a drug or drug-like molecule with adequate biochemical activity. Thirdly, the modulation of target in cell or animal models should ameliorate the relevant disease phenotype. Last but not least, manual literature search in PubMed [31] was used to guarantee the data quality. Only when three types of validation data were collected could the target of interest be validated

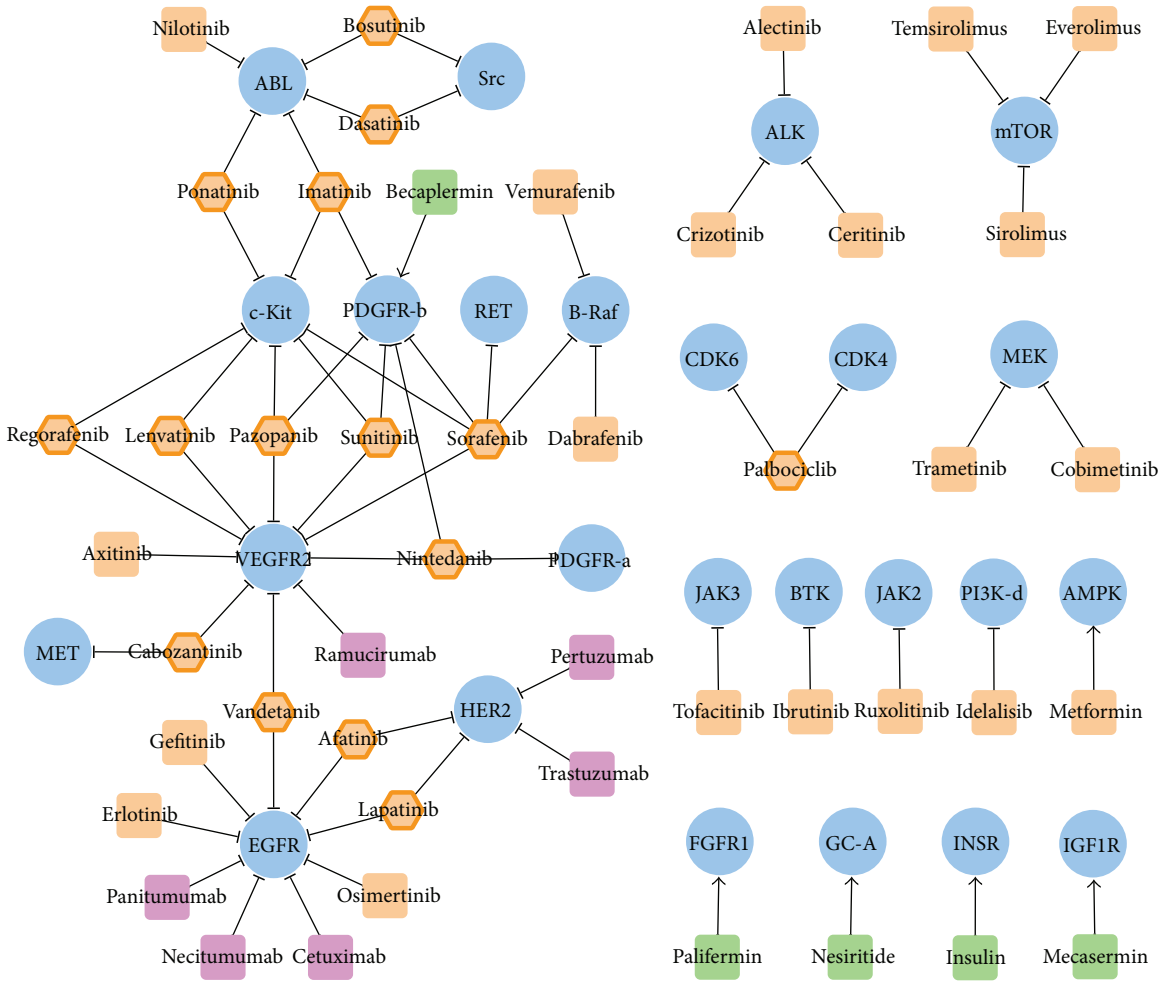


FIGURE 1: Drug-target interaction network of FDA approved drugs targeting kinase. Single target drugs were represented by round rectangle (small molecular drugs in orange, monoclonal antibodies in magenta, and biologics in green), while multitarget drugs were represented by orange hexagon and highlighted by additional orange hexagon line. All kinase targets were shown by blue ellipse. Interactions between drug and target were displayed by edges with shapes of arrow and "T" representing activation and inhibition, respectively.

as a primary one. Those three validation data types include the following: experimentally determined potency of drugs against their primary targets, observed potency of drugs against disease models linked to their corresponding targets, and the observed effects of target knockout, transgenetic, RNA interference, antibody, and antisense *in vivo* models.

3. Results and Discussions

3.1. Construction of Drug-Target Interaction Networks and Subnetworks. Drug-target interaction networks of approved and clinical trial drugs were constructed and displayed by Cytoscape 3.3.0 [32], which is a stand-alone platform for visualizing molecular interaction networks. 46 FDA approved drugs together with their corresponding 25 targets were uploaded to and displayed in Cytoscape. As shown in Figure 1, single target drugs were shown as a round rectangle (small molecular drugs in orange, monoclonal antibodies in magenta, and biologics in green), while the multitarget drugs were displayed by orange hexagon and highlighted by additional orange hexagon line. All kinase targets were shown by blue ellipse. Interactions between drug and target were displayed by edges with shapes of arrow and "T" representing activation and inhibition, respectively. Moreover, 149 clinical trial drugs along with their 39 targets were inputted and shown in Cytoscape. The network representing drug-target interaction was provided in Figure 2 with the representation of target the same as that in Figure 1 (blue ellipse). Due to the huge number of clinical trial drugs and targets, subnetwork of specific disease class according to the International Classification of Diseases (ICD) was generated. The ICD was provided by the World Health Organization as the standard diagnostic tool for epidemiology, health management, and clinical purpose. Firstly, a specific disease class (at level 2 of ICD) named as the "malignant neoplasms of female genital organs" was selected, and clinical trial drugs and targets within this disease class were identified. Consequently, 13 drugs against 9 kinase targets were displayed [32]. As shown in Figure 2, single target drugs were shown as a round rectangle, while the multitarget drugs were displayed by

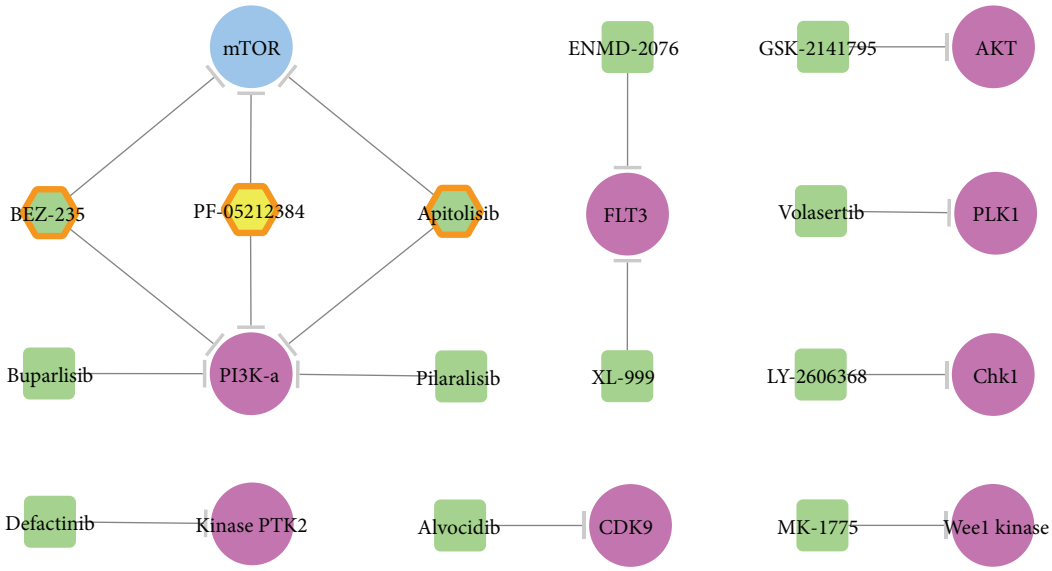


Figure 2: Drug-target interaction network of kinase inhibitors in clinical trial—a subnetwork of the malignant neoplasms of female genital organs (C51–C58). Single target drugs were shown as a round rectangle, while the multitarget drugs were displayed by hexagon. Colors of the drugs were defined as follows: phase 2 clinical trial drugs are in green and phase 1 clinical trial drugs are in yellow. The multitarget drugs were highlighted by an additional orange hexagon line. Established and clinical trial targets were shown by blue and violet ellipses, respectively.

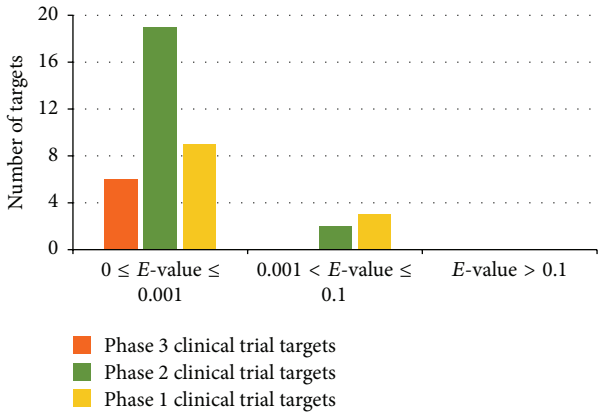


Figure 3: Distribution of phase 1 (yellow), phase 2 (green), and phase 3 (orange) clinical trial targets by level of similarity to established targets. The level of similarity to established targets is classified into very similar, marginally similar, and unsimilar with the BLAST E-values in the range of ≤0.001 and 0.001~0.1 and >0.1, respectively.

hexagon. Colors of drugs were defined as phase 2 clinical trial drugs in green and phase 1 clinical trial drugs in yellow. Multitarget drugs were highlighted by additional orange hexagon lines.

3.2. Comparison of System Profiles between FDA Approved Kinase Targets and Clinical Trial Ones. Comparison of the characteristics of the 39 clinical trial kinase targets with those of established kinase targets provides clues about common and distinguished features and shift of trends in profiles of clinical trial targets that can be retained, enhanced, or improved. Figure 3 illustrated the distribution of phases 1, 2, and 3 clinical trial targets with respect to the level of sequence similarity to the established targets. Based on the BLAST E-value, the levels of similarity were classified into very similar ($E \leq 0.001$), marginally similar ($0.001 \leq E \leq 0.1$), and unsimilar ($E > 0.1$). The majority of the clinical trial kinase targets (100%, 90%, and 75% of phases 3, 2, and 1) were very similar to the established ones. In addition, no clinical trial kinase target was significantly different in sequence to the established ones.

Figure 4 illustrated the distributions of clinical trial kinase targets and established kinase targets with respect to the number of human similarity proteins outside families of the target (Figure 4(a)), the number of target affiliated signaling pathways (Figure 4(b)), and the number of tissues that the target is distributed in (Figure 4(c)). The distribution profiles of clinical trial kinase targets were comparable to those of the established ones [17, 18]. As shown in Figure 4, 88% and 84% of the established and clinical trial targets had <15 human similarity proteins outside their target family. Furthermore, 71% and 68% of the established and clinical trial targets were affiliated to ≤3 human signaling pathways. In addition, 100% and 95% established and clinical trial targets were distributed in ≤5 human tissues. In summary, the systems profiles of vast majority of clinical trial kinase targets appear to be very similar to those of established ones [16], but the percentages of established targets obeying all three system profiles appear to be slightly but consistently higher than those of clinical trial targets.

Figure 5 illustrated the distributions of phase 1, phase 2, and phase 3 clinical trial kinase targets with respect to the number of human similarity proteins outside families of the target (Figure 5(a)), the number of target affiliated signaling

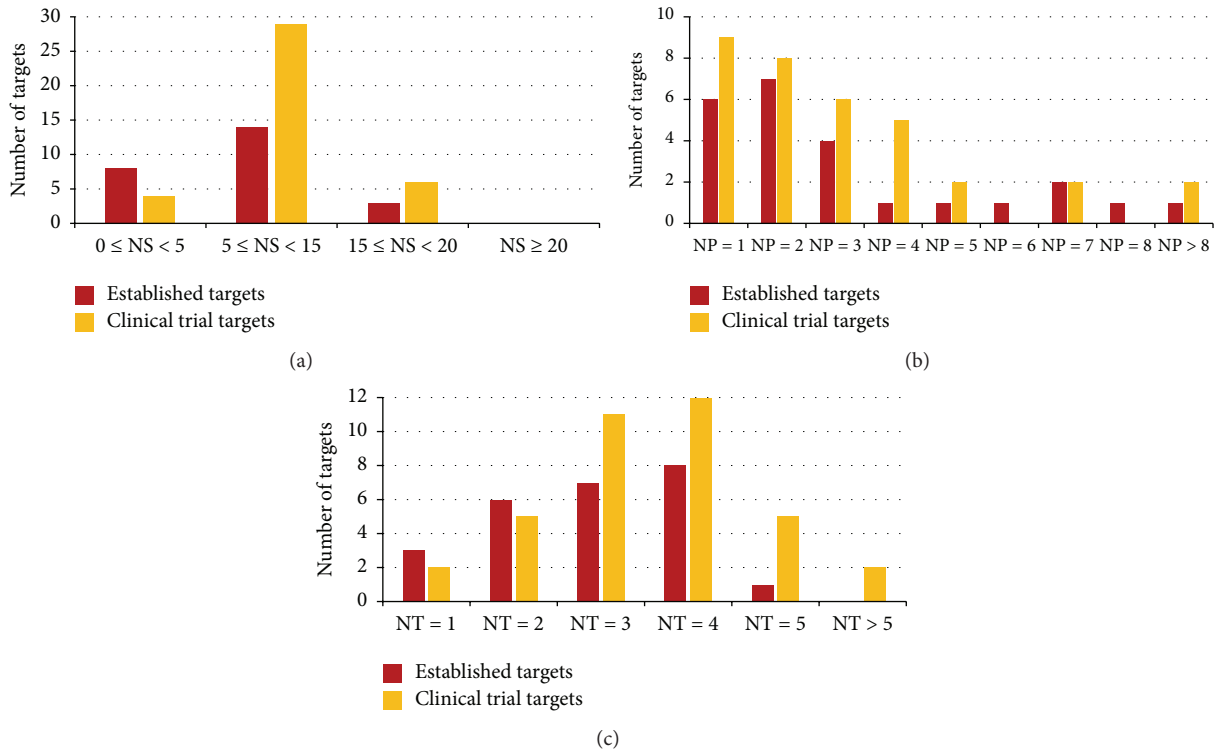


FIGURE 4: Distribution of all clinical trial kinase targets (orange) and established kinase targets (red) with respect to (a) the number of human similarity proteins (NS) outside the target family, (b) the number of human pathways (NP) the target is associated with, and (c) the number of human tissues (NT) the target is distributed in.

pathways (Figure 5(b)), and the number of tissues that the target is distributed in (Figure 5(c)). As shown in figures, 86%, 85%, and 75% of phase 3, phase 2, and phase 1 clinical trial targets had <15 human similarity proteins outside their target family. Furthermore, 83%, 50%, and 82% of phase 3, phase 2, and phase 1 clinical trial targets were associated with ≤3 human pathways. In addition, 100%, 95%, and 91% of phase 3, phase 2, and phase 1 clinical trial targets were distributed in ≤5 human tissues. Consequently, percentages of phases 3, 2, and 1 clinical trial kinase targets obeying two system profiles (number of similarity proteins and tissues) appear to follow a clear descending trend, which indicates more similar profiles between established and phase 3 targets comparing to phases 2 and 1 targets.

In the meantime, the distributions of those three types of system profiles of phase 1, phase 2, and phase 3 clinical trial kinase targets and that of established targets were compared by *boxplot* (Supplementary Figure S1 in Supplementary Material available online at http://dx.doi.org/10.1155/2016/2509385). Although no significant statistical difference was observed between different clinical statuses of three types of system profiles, a shift of trend in 3 system profiles could be identified. In particular, from the established to phases 3, 2, and 1 clinical trial targets, there was a clear ascending trend of the mediums of the number of human similarity proteins outside their target family and the number of tissues that the target is distributed in. Similar ascending trend could also be observed for the number of target affiliated signaling pathways, but the medium of phase 1 targets was lower than that of phase 2 targets. In summary, as shown in Figures 4 and 5 and Supplementary Figure S1, systems profiles of vast majority of clinical trial kinase targets (especially phase 3 targets) appear to be very similar to those of established ones, which indicates that, despite extensive exploration on the innovative therapeutic target, kinases capable of entering clinical trial are those very similar to the established ones in system profiles. However, as shown in Supplementary Figure S1, there is a clear shift of trend in the system profiles from the clinical trial (phase 1 to phase 2 to phase 3) to established targets.

3.3. Evaluating the Performance of the Combinational Method Used for Identifying Promising Target. Majority of clinical trial targets were reported to be similar to established ones in their systems profiles [17, 18, 33–36]; target druggability may be further revealed by two more profiles: drug binding domain structural conformations [37] and protein physicochemical properties [38]. As reported, a combinational method was able to identify 50%, 25%, and 10% of the analyzed phases 3, 2, and 1 targets and 4% of nonclinical trial targets as similar to the established targets in at least 3 of the 4 profiles by systematically analyzing sequence, structural, physicochemical, and system profiles of these targets [16]. It has been 7 years since the publication of that combinational method, and it would be of great interest to evaluate its

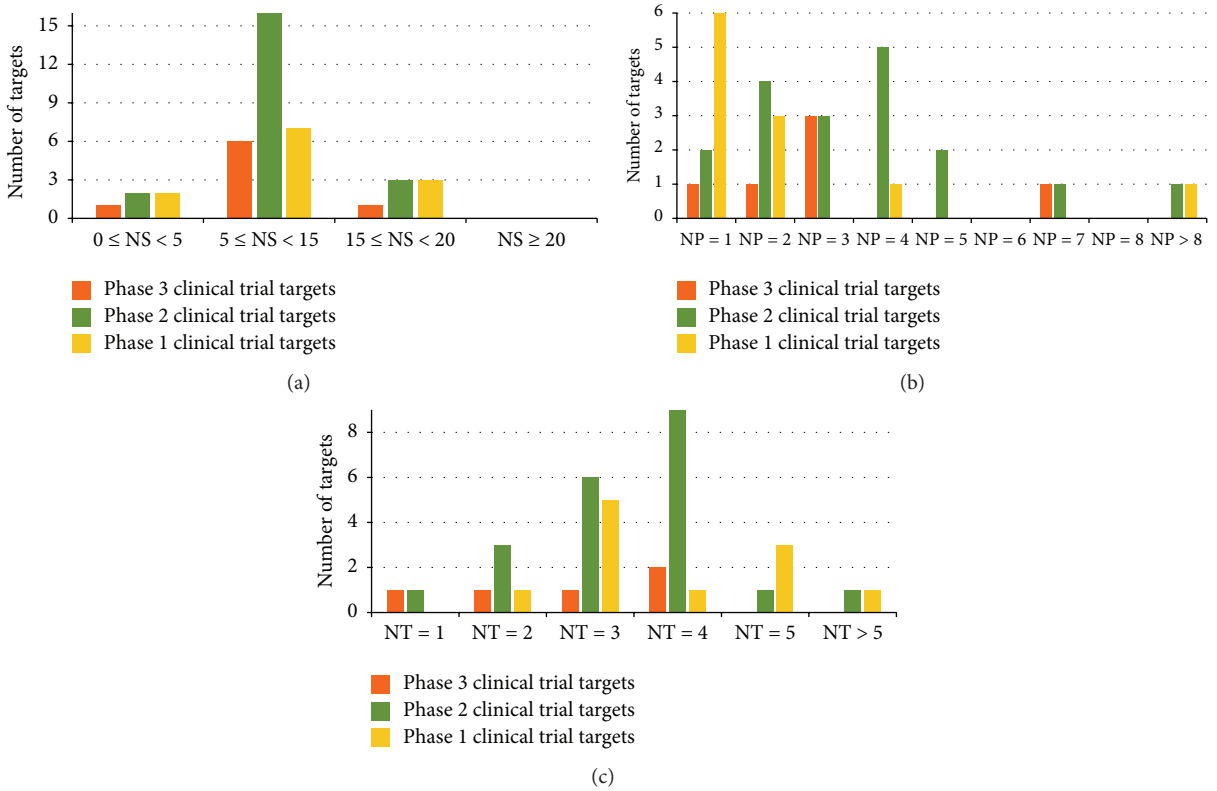


Figure 5: Distribution of phase 1 (yellow), phase 2 (green), and phase 3 (orange) clinical trial kinase targets with respect to (a) the number of human similarity proteins (NS) outside the target family, (b) the number of human pathways (NP) the target is associated with, and (c) the number of human tissues (NT) the target is distributed in.

predictive performance by investigating the current developmental status of those clinical trial targets. As shown in Table 1, of the 16 phase 3 targets similar to the established ones in no less than 3 profiles [16], 5 (31%) have been approved and 6 (37%) have shown positive phase 3 results. Moreover, no positive result has been reported for 13 of the 15 phase 3 targets similar to the established ones in less than 3 profiles (with only one exception (FPTase), whose drug was filed for approval but was deemed not approvable by FDA) [16]. In particular, the corresponding phase 3 drugs of 3 targets (HSP90, squalene synthase, and FLAP) were all discontinued, and those of 5 targets (AKT, MMP-2, MMP-9, MMP-12, and sphingosine kinase) were reported with negative phase 3 results. Because of its strong predictive power reflected by the real world test in this study, the combinational method appeared to be capable of capturing target druggability by the genetic, structural, physicochemical, and system profiles [16]. Moreover, these have in turn led to the exploration of individual [17, 18, 22, 35–40] and combination of profiles [16], perspectives [41–43], and algorithms [44, 45] for *in silico* target analysis and prediction.

3.4. Drug-Target Interaction Networks of FDA Approved and Clinical Trial Drugs Targeting Kinase. To understand drug-target interaction of FDA approved drugs targeting kinase, network of those drugs as well as their corresponding targets was shown in Figure 1. As a widely used statistical concept in network analysis, degree was applied to assess interactions of targets and drugs. Degree of a specific node (drug or target) refers to the number of edges (interaction from other nodes) connected to this node. As shown in Figure 1, the maximum and minimum numbers of degree of approved kinase inhibitors equal 5 and 1, respectively. Particularly, 1, 4, 10, and 31 kinase inhibitors target on 5 kinases, 3 kinases, 2 kinases, and 1 kinase as their primary therapeutic targets. In particular, drug of the highest degree was sorafenib.

The maximum and minimum numbers of degree of targets equal 10 and 1, respectively. Particularly, 1, 1, 1, 1, 1, 1, 3, 2, and 14 targets were targeted by 10, 9, 7, 6, 5, 4, 3, and 2 multitarget drugs and 1 multitarget drug, respectively. The targets of substantially high degree (>8 drugs) were VEGFR2 and EGFR. As reported, VEGF and its receptors were essential in the development of the renal cell carcinoma (RCC) [46], and the inhibition of VEGFR2 could provide substantial influence on RCC's pathogenesis. In the meantime, EGFR was reported to play critical roles in and acted as primary target for non-small cell lung cancer [47], breast cancer [48], and colorectal cancer [49, 50]. Based on the network analysis, VEGFR2 and EGFR were identified as the most popular primary therapeutic kinase targets of all FDA approved drugs. Supplementary Figures S2 and S3 illustrated a comprehensive drug-target interaction network including all 46 FDA approved drugs (together with their corresponding 25 established targets) and 239 drugs in clinical trial (including 81 drugs targeting

Table 1: Latest development status of the previously analyzed phase 3 targets similar to established targets in sequence (A), drug binding domain structural fold (B), physicochemical features (C), and systems (D) profiles.

Target (drug previously reported to be in phase 3 trial)	Similar to established targets in combination of A, B, C, and D profiles	Targeted disease conditions	Latest development status (year of report)
CCK-A receptor (dexloxiglumide)	Combination of A, B, C, and D	Irritable bowel syndrome	Positive results in phase III trial (2007) and a large European phase III trial (2010), in talks with FDA for approval (2010)
Coagulation factor IIa (SR-123781A)	Combination of A, B, C, and D	Venous thromboembolism	Positive results in a large European phase III trial (2008)
NTRK1 (lestaurtinib)	Combination of A, B, C, and D	Acute myeloid leukemia	Lestaurtinib approved by FDA as orphan drug (2006)
5HT 3 receptor (cilansetron)	Combination of A, C, and D	Irritable bowel syndrome	Positive phase III trial results (2004), filed but withdrawn for FDA approval (2005), still in talks with MHRA and EU (2010)
Heparanase (PI-88)	Combination of A, C, and D	Hepatocellular cancer	PI-88 fast tracked by FDA (2008)
MDR 3 (LY335979)	Combination of A, C, and D	Acute myeloid leukemia	
Orexin receptor (almorexant)	Combination of A, C, and D	Sleep disorders	Positive phase III trial result (2010)
Somatostatin receptor 1 (pasireotide)	Combination of A, C, and D	Cushing's disease, renal cell carcinoma	
NK-2 receptor (saredutant)	Combination of A, C, and D	Depression	Positive phase III trial result (2007), trial discontinued (2009)
BK-2 receptor (icatibant)	Combination of A, B, and C	Hereditary angioedema, traumatic brain injuries	Positive phase III trial results (2006), icatibant approved in EU (2008)
Thrombin receptor (SCH-530348)	Combination of A, B, and C	Cardiovascular disorders	
CXCR4 (plerixafor)	Combination of A, B, and D	Non-Hodgkin's lymphoma, late-stage solid tumors	Plerixafor approved by FDA (2008)
C1 esterase (Cinryze)	Combination of A, B, and D	Hereditary angioedema	Cinryze approved by FDA (2008)
Sphingosine 1-phosphate receptor 1 (Gilenia)	Combination of A, B, and D	Multiple sclerosis	Positive phase III trial results (2008). FDA granted priority review (2010)
NPYR5 (CGP71683A)	Combination of A, B, and D	Obesity	
Plasma kallikrein (ecallantide)	Combination of A, B, and D	Hereditary angioedema	Positive phase III trial results (2007), ecallantide approved by FDA (2009)

only on 17 established targets, 140 drugs targeting only on 36 clinical trial targets, and 29 multitarget drugs targeting on 13 established and 13 clinical trial targets). As shown, established targets of substantially high degree of clinical trial drugs (>10 drugs) were EGFR (26 drugs), mTOR (19 drugs), VEGFR2 (14 drugs), and IGF1R (13 drugs). In summary, EGFR and VEGFR2 were identified as the most popular established targets utilized by the highest number of both approved and clinical trial drugs, while mTOR and IGF1R were also the popular established targets with high number of drugs tested in the clinical trial.

Moreover, clinical trial targets of substantially high degree of clinical trial drugs (>10 drugs) were CDK1/2 (13 drugs), Glucokinase (13 drugs), AKT (13 drugs), and Aurora B (12 drugs). Figure 2 illustrated a subnetwork of drug-target interaction of clinical trial drugs used for treating the malignant neoplasms of female genital organs (C51–C58). Typical diseases within this class included the ovarian cancer and cervical cancer. In this disease class, the maximum degree of drugs equals 2, while the minimum is 1. In particular, 3 and 10 drugs worked on 2 primary targets and 1 primary target, respectively. BEZ-235, PF-05212384, and apitolisib were dual PI3K-alpha/mTOR inhibitors currently in phase 2 or 1 clinical trials. Take BEZ-235 as an example; its dual inhibition disturbed the PI3K/AKT/mTOR signaling pathway, leading to cell apoptosis of endometrial cancer overexpressing PI3K and mTOR [51].

The maximum number of degrees of targets equals 5, while the minimum number is 1. In particular, 1, 1, 1, and 6 targets were targeted by 5, 3, and 2 clinical trial kinase inhibitors

and 1 clinical trial kinase inhibitor for treating cancers of the female genital organ, respectively. Target of the highest degree was the PI3K-alpha. The development of endometrial cancer was reported to be closely associated with the disruptions in both Wnt/beta-catenin and Akt/PI3K/mTOR pathways. Particularly, the genetic mutations in the catalytic subunit of PI3K were considered typical for endometrial cancer and were present in 26%~36% of cases [52]. Moreover, target of the second largest degree was mTOR. PI3K/mTOR pathway was frequently activated in the endometrial cancer through various genetic alterations [53], which double confirmed the pivotal roles of both targets in endometrial cancer [51]. Thus, based on network analysis, mTOR and PI3K-alpha were discovered as the most popular targets of kinase inhibitors in clinical trial for cancers of female genital organs.

4. Conclusion

In this study, a comparative analysis on system profiles of both targets was conducted. Moreover, a previously reported combinational method used for predicting the promising targets was discussed and evaluated. Drug-target interaction networks were used to identify popular established and clinical trial kinase targets. As a result, systems profiles of the majority of clinical trial kinase targets were identified to be very similar to those of established ones, but a shift of trend in the system profiles from the clinical trial to the established targets was identified.

Competing Interests

The authors declare that they have no competing interests.

Authors' Contributions

Jingyu Xu, Panpan Wang, and Hong Yang contributed equally to this work.

Acknowledgments

This work was funded by the Chongqing Graduate Student Research Innovation Project (CYB14027); by the research support of National Natural Science Foundation of China (81202459, 21505009); and by Fundamental Research Funds for the Central Universities (CDJZR14468801, CDJKXB14011, and 2015CDJXY).

References

[1] G. Manning, D. B. Whyte, R. Martinez, T. Hunter, and S. Sudarsanam, "The protein kinase complement of the human genome," *Science*, vol. 298, no. 5600, pp. 1912–1934, 2002.

[2] J. Zhang, P. L. Yang, and N. S. Gray, "Targeting cancer with small molecule kinase inhibitors," *Nature Reviews Cancer*, vol. 9, no. 1, pp. 28–39, 2009.

[3] H. Yang, C. Qin, Y. H. Li et al., "Therapeutic target database update 2016: enriched resource for bench to clinical drug target and targeted pathway information," *Nucleic Acids Research*, vol. 44, pp. D1069–D1074, 2016.

[4] M. Rask-Andersen, J. Zhang, D. Fabbro, and H. B. Schiöth, "Advances in kinase targeting: current clinical use and clinical trials," *Trends in Pharmacological Sciences*, vol. 35, no. 11, pp. 604–620, 2014.

[5] M. Bellon, L. Lu, and C. Nicot, "Constitutive activation of Pim1 kinase is a therapeutic target for adult T-cell leukemia," *Blood*, vol. 127, no. 20, pp. 2439–2450, 2016.

[6] F. E. Bleeker, S. Lamba, C. Zanon et al., "Mutational profiling of kinases in glioblastoma," *BMC Cancer*, vol. 14, article 718, 2014.

[7] P. Lahiry, A. Torkamani, N. J. Schork, and R. A. Hegele, "Kinase mutations in human disease: interpreting genotype-phenotype relationships," *Nature Reviews Genetics*, vol. 11, no. 1, pp. 60–74, 2010.

[8] W. Xue, P. Wang, B. Li et al., "Identification of the inhibitory mechanism of FDA approved selective serotonin reuptake inhibitors: an insight from molecular dynamics simulation study," *Physical Chemistry Chemical Physics*, vol. 18, no. 4, pp. 3260–3271, 2016.

[9] P. Wu, T. E. Nielsen, and M. H. Clausen, "Small-molecule kinase inhibitors: an analysis of FDA-approved drugs," *Drug Discovery Today*, vol. 21, no. 1, pp. 5–10, 2015.

[10] P. Wu, T. E. Nielsen, and M. H. Clausen, "FDA-approved small-molecule kinase inhibitors," *Trends in Pharmacological Sciences*, vol. 36, no. 7, pp. 422–439, 2015.

[11] "Efficacy and safety of nintedanib in idiopathic pulmonary fibrosis," *The New England Journal of Medicine*, vol. 373, no. 8, pp. 782–782, 2015.

[12] A.-L. Barabási, N. Gulbahce, and J. Loscalzo, "Network medicine: a network-based approach to human disease," *Nature Reviews Genetics*, vol. 12, no. 1, pp. 56–68, 2011.

[13] K. Strebhardt, "Multifaceted polo-like kinases: drug targets and antitargets for cancer therapy," *Nature Reviews Drug Discovery*, vol. 9, no. 8, pp. 643–660, 2010.

[14] F. Zhu, Z. Shi, C. Qin et al., "Therapeutic target database update 2012: a resource for facilitating target-oriented drug discovery," *Nucleic Acids Research*, vol. 40, no. 1, pp. D1128–D1136, 2012.

[15] F. Zhu, B. Han, P. Kumar et al., "Update of TTD: therapeutic target database," *Nucleic Acids Research*, vol. 38, no. 1, pp. D787–D791, 2009.

[16] F. Zhu, L. Han, C. Zheng et al., "What are next generation innovative therapeutic targets? Clues from genetic, structural, physicochemical, and systems profiles of successful targets," *The Journal of Pharmacology and Experimental Therapeutics*, vol. 330, no. 1, pp. 304–315, 2009.

[17] C. Zheng, L. Han, C. W. Yap, B. Xie, and Y. Chen, "Progress and problems in the exploration of therapeutic targets," *Drug Discovery Today*, vol. 11, no. 9-10, pp. 412–420, 2006.

[18] C. J. Zheng, L. Y. Han, C. W. Yap, Z. L. Ji, Z. W. Cao, and Y. Z. Chen, "Therapeutic targets: progress of their exploration and investigation of their characteristics," *Pharmacological Reviews*, vol. 58, no. 2, pp. 259–279, 2006.

[19] A. C. Cheng, R. G. Coleman, K. T. Smyth et al., "Structure-based maximal affinity model predicts small-molecule druggability," *Nature Biotechnology*, vol. 25, no. 1, pp. 71–75, 2007.

[20] D. Kozakov, D. R. Hall, R. L. Napoleon, C. Yueh, A. Whitty, and S. Vajda, "New frontiers in druggability," *Journal of Medicinal Chemistry*, vol. 58, no. 23, pp. 9063–9088, 2015.

[21] T. Masini, B. S. Kroezen, and A. K. H. Hirsch, "Druggability of the enzymes of the non-mevalonate-pathway," *Drug Discovery Today*, vol. 18, no. 23-24, pp. 1256–1262, 2013.

[22] A. L. Hopkins and C. R. Groom, "The druggable genome," *Nature Reviews Drug Discovery*, vol. 1, no. 9, pp. 727–730, 2002.

[23] E. Boutet, D. Lieberherr, M. Tognolli et al., "UniProtKB/Swiss-Prot, the manually annotated section of the UniProt KnowledgeBase: how to use the entry view," *Methods in Molecular Biology*, vol. 1374, pp. 23–54, 2016.

[24] A. Bairoch, "The ENZYME database in 2000," *Nucleic Acids Research*, vol. 28, no. 1, pp. 304–305, 2000.

[25] M. Kanehisa, Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe, "KEGG as a reference resource for gene and protein annotation," *Nucleic Acids Research*, vol. 44, no. 1, pp. D457–D462, 2016.

[26] S. Kogenaru, C. del Val, A. Hotz-Wagenblatt, and K.-H. Glatting, "TissueDistributionDBs: a repository of organism-specific tissue-distribution profiles," *Theoretical Chemistry Accounts*, vol. 125, no. 3–6, pp. 651–658, 2010.

[27] M. Johnson, I. Zaretskaya, Y. Raytselis, Y. Merezhuk, S. McGinnis, and T. L. Madden, "NCBI BLAST: a better web interface," *Nucleic Acids Research*, vol. 36, supplement 2, pp. W5–W9, 2008.

[28] W. Huber, V. J. Carey, R. Gentleman et al., "Orchestrating high-throughput genomic analysis with Bioconductor," *Nature Methods*, vol. 12, no. 2, pp. 115–121, 2015.

[29] M. A. Lindsay, "Target discovery," *Nature Reviews Drug Discovery*, vol. 2, no. 10, pp. 831–838, 2003.

[30] O. Vidalin, M. Muslmani, C. Estienne, H. Echchakir, and A. M. Abina, "*In vivo* target validation using gene invalidation, RNA interference and protein functional knockout models: it is the time to combine," *Current Opinion in Pharmacology*, vol. 9, no. 5, pp. 669–676, 2009.

[31] E. W. Sayers, T. Barrett, D. A. Benson et al., "Database resources of the national center for biotechnology information," *Nucleic Acids Research*, vol. 39, no. 1, pp. D38–D51, 2011.

[32] P. Shannon, A. Markiel, O. Ozier et al., "Cytoscape: a software environment for integrated models of biomolecular interaction networks," *Genome Research*, vol. 13, no. 11, pp. 2498–2504, 2003.

[33] M. A. Lindsay, "Finding new drug targets in the 21st century," *Drug Discovery Today*, vol. 10, no. 23-24, pp. 1683–1687, 2005.

[34] F. Sams-Dodd, "Target-based drug discovery: is something wrong?" *Drug Discovery Today*, vol. 10, no. 2, pp. 139–147, 2005.

[35] L. Yao and A. Rzhetsky, "Quantitative systems-level determinants of human genes targeted by successful drugs," *Genome Research*, vol. 18, no. 2, pp. 206–213, 2008.

[36] M. K. Sakharkar, P. Li, Z. Zhong, and K. R. Sakharkar, "Quantitative analysis on the characteristics of targets with FDA approved drugs," *International Journal of Biological Sciences*, vol. 4, no. 1, pp. 15–22, 2008.

[37] P. J. Hajduk, J. R. Huth, and S. W. Fesik, "Druggability indices for protein targets derived from NMR-based screening data," *Journal of Medicinal Chemistry*, vol. 48, no. 7, pp. 2518–2525, 2005.

[38] L. Y. Han, C. J. Zheng, B. Xie et al., "Support vector machines approach for predicting druggable proteins: recent progress in its exploration and investigation of its usefulness," *Drug Discovery Today*, vol. 12, no. 7-8, pp. 304–313, 2007.

[39] P. J. Hajduk, J. R. Huth, and C. Tse, "Predicting protein druggability," *Drug Discovery Today*, vol. 10, no. 23-24, pp. 1675–1682, 2005.

[40] H. Xu, H. Xu, M. Lin et al., "Learning the drug target-likeness of a protein," *Proteomics*, vol. 7, no. 23, pp. 4255–4263, 2007.

[41] E. Klipp, R. C. Wade, and U. Kummer, "Biochemical network-based drug-target prediction," *Current Opinion in Biotechnology*, vol. 21, no. 4, pp. 511–516, 2010.

[42] S. Pérot, O. Sperandio, M. A. Miteva, A.-C. Camproux, and B. O. Villoutreix, "Druggable pockets and binding site centric chemical space: a paradigm shift in drug discovery," *Drug Discovery Today*, vol. 15, no. 15-16, pp. 656–667, 2010.

[43] N. Huang and M. P. Jacobson, "Binding-site assessment by virtual fragment screening," *PLoS ONE*, vol. 5, no. 4, Article ID e10109, 2010.

[44] U. Rix and G. Superti-Furga, "Target profiling of small molecules by chemical proteomics," *Nature Chemical Biology*, vol. 5, no. 9, pp. 616–624, 2009.

[45] G. Hu and P. Agarwal, "Human disease-drug network based on genomic expression profiles," *PLoS ONE*, vol. 4, no. 8, Article ID e6536, 2009.

[46] B. I. Rini, "Vascular endothelial growth factor-targeted therapy in metastatic renal cell carcinoma," *Cancer*, vol. 115, no. 10, pp. 2306–2312, 2009.

[47] W. Han and H.-W. Lo, "Landscape of EGFR signaling network in human cancers: biology and therapeutic response in relation to receptor subcellular locations," *Cancer Letters*, vol. 318, no. 2, pp. 124–134, 2012.

[48] S. O. Lim, C. W. Li, W. Xia et al., "EGFR signaling enhances aerobic glycolysis in triple-negative breast cancer cells to promote tumor growth and immune escape," *Cancer Research*, vol. 76, no. 5, pp. 1284–1296, 2016.

[49] A. Bertotti, E. Papp, S. Jones et al., "The genomic landscape of response to EGFR blockade in colorectal cancer," *Nature*, vol. 526, no. 7572, pp. 263–267, 2015.

[50] S. Misale, R. Yaeger, S. Hobor et al., "Emergence of KRAS mutations and acquired resistance to anti-EGFR therapy in colorectal cancer," *Nature*, vol. 486, no. 7404, pp. 532–536, 2012.

[51] K. J. Dedes, D. Wetterskog, A. Ashworth, S. B. Kaye, and J. S. Reis-Filho, "Emerging therapeutic targets in endometrial cancer," *Nature Reviews Clinical Oncology*, vol. 8, no. 5, pp. 261–271, 2011.

[52] D. Llobet, J. Pallares, A. Yeramian et al., "Molecular pathology of endometrial carcinoma: practical aspects from the diagnostic and therapeutic viewpoints," *Journal of Clinical Pathology*, vol. 62, no. 9, pp. 777–785, 2009.

[53] S. Murayama-Hosokawa, K. Oda, S. Nakagawa et al., "Genome-wide single-nucleotide polymorphism arrays in endometrial carcinomas associate extensive chromosomal instability with poor prognosis and unveil frequent chromosomal imbalances involved in the PI3-kinase pathway," *Oncogene*, vol. 29, no. 13, pp. 1897–1908, 2010.

Hindawi Publishing Corporation
BioMed Research International
Volume 2016, Article ID 8124636, 13 pages
http://dx.doi.org/10.1155/2016/8124636

Research Article

Sequence- and Structure-Based Functional Annotation and Assessment of Metabolic Transporters in *Aspergillus oryzae*: A Representative Case Study

Nachon Raethong,[1] Jirasak Wong-ekkabut,[2,3,4] Kobkul Laoteng,[5] and Wanwipa Vongsangnak[1,3]

[1]*Department of Zoology, Faculty of Science, Kasetsart University, Bangkok 10900, Thailand*
[2]*Department of Physics, Faculty of Science, Kasetsart University, Bangkok 10900, Thailand*
[3]*Computational Biomodelling Laboratory for Agricultural Science and Technology (CBLAST), Faculty of Science, Kasetsart University, Bangkok 10900, Thailand*
[4]*Center of Advanced Science in Industrial Technology, Faculty of Science, Kasetsart University, Bangkok 10900, Thailand*
[5]*Food Biotechnology Research Unit, National Center for Genetic Engineering and Biotechnology (BIOTEC), National Science and Technology Development Agency (NSTDA), Pathum Thani 12120, Thailand*

Correspondence should be addressed to Wanwipa Vongsangnak; wanwipa.v@ku.ac.th

Received 25 January 2016; Accepted 6 April 2016

Academic Editor: Luisa Di Paola

Aspergillus oryzae is widely used for the industrial production of enzymes. In *A. oryzae* metabolism, transporters appear to play crucial roles in controlling the flux of molecules for energy generation, nutrients delivery, and waste elimination in the cell. While the *A. oryzae* genome sequence is available, transporter annotation remains limited and thus the connectivity of metabolic networks is incomplete. In this study, we developed a metabolic annotation strategy to understand the relationship between the sequence, structure, and function for annotation of *A. oryzae* metabolic transporters. Sequence-based analysis with manual curation showed that 58 genes of 12,096 total genes in the *A. oryzae* genome encoded metabolic transporters. Under consensus integrative databases, 55 unambiguous metabolic transporter genes were distributed into channels and pores (7 genes), electrochemical potential-driven transporters (33 genes), and primary active transporters (15 genes). To reveal the transporter functional role, a combination of homology modeling and molecular dynamics simulation was implemented to assess the relationship between sequence to structure and structure to function. As in the energy metabolism of *A. oryzae*, the H^+-ATPase encoded by the AO090005000842 gene was selected as a representative case study of multilevel linkage annotation. Our developed strategy can be used for enhancing metabolic network reconstruction.

1. Introduction

Aspergillus oryzae belongs to a group of filamentous fungi that has long been used for the commercial production of different industrial enzymes, such as alpha-amylases [1], proteases [2], glucoamylases [3], xylanases [4], other hydrolytic enzymes [5], and organic acids [6]. Not only does *A. oryzae* produce various biological compounds, but also it has beneficial features, such as acting as a robust host system with high production yields and acclimatization to environmental and nutritional duress [7]. In 2005, the whole genome of *A. oryzae* strain RIB40 was sequenced and annotated [8]. Very recently, the quality of the genome sequence was improved and verified using next-generation sequencing platforms, such as SOLiD [9] and Illumina MiSeq [10]. Moreover, the advancement of multilevel omics integrative analysis (genomics, transcriptomics, and proteomics) has enabled the interpretation of high-throughput data for functional annotation. In addition, the number of annotated genes in *A. oryzae* was enhanced using expressed sequence tags data [11]. Clusters of genes were then identified and annotated by

oligonucleotide microarrays [12, 13] and mRNA sequencing technology [14].

Using a systems biology approach, a genome-scale metabolic network of *A. oryzae* was reconstructed based on annotated genomic data, which contains 1,314 enzyme-encoding genes including 53 metabolic transporter-associated genes [11]. Modeling of the genome-scale metabolic network of *A. oryzae* has been used to evaluate fungal biological processes and cellular physiology. However, the connectivity of metabolic networks remains incomplete because of the poor annotation of transporter genes. Among the 161 unique transport reactions, only 33% of annotated genes were identified and used in the network [11]. In metabolic pathways, transporters appear to play crucial roles in controlling the flux of molecules into and out of cells [6, 15, 16]. Additionally, several transporters regulate metabolic energy generation, delivery of essential nutrients, waste product elimination, and survival under environmental changes [17].

The techniques used for transporter annotation are often performed by sequence-based analysis using pairwise and multiple sequence alignment. Many studies of fungal transporters have relied on similarity searching between orthologous sequences using the BLASTP algorithm [18], such as investigating the gene encoding glucose transporter (hxtB–E) in the genome of *Aspergillus nidulans*. In particular, use of the ClustalW program [19] allowed for the clustering and the identification of conserved sequences and evolutionary relationship among orthologs of fungal transporters. In a study of amino acid uptake in rust fungi (plant pathogenic fungi), 60 genes were identified from rust fungal genomes and then clustered into three different transporter families, including 33 genes in yeast amino acid transporters, 20 genes in amino acid/choline transporters, and 7 genes in L-type amino acid transporters [20]. This study indicated several transporter genes in rust fungal genomes, which may play a role in interactions between plant and rust fungi [20]. However, sequence-based analysis is limited to functional annotation. For example, there is a case of two proteins, which have overall identical protein folds implying their closely related functions, but no statistically significant degree of sequence identity was observed [21]. To address such this case, structural studies through three-dimensional (3D) structure from crystallography have greatly enhanced our understanding of the potential protein function. As an example case presented in yeast, the structure of V-ATPase from *Saccharomyces cerevisiae* was determined using electron cryomicroscopy wherein the conformational changes for three functional states were observed during proton translocation [22]. Recently, the crystal structure of the phosphate transporter from *Piriformospora indica* was determined using X-ray crystallography, suggesting both proton and phosphate exit pathways and the mechanism of phosphate transport [23]. However, the number of molecules with unsolved 3D structures and unknown functions is increasing rapidly because the experimental assays to determine these properties are time-consuming and expensive. Computational approaches enable functional annotation and can be used to overcome these limitations. As observed in *A. nidulans*, the relationship between the structure and function of the subfamily of urea/H^+ membrane transporter for the UreA gene was studied [24]. Homology models of the urea transporter were developed from the crystal structures of other organisms [25, 26] as templates combined with site-directed and classical random mutagenesis. This computational approach can be used to identify critical residues for urea transport and understand the binding, recognition, and translocation of urea [24]. However, the structure-based approaches generally rely on single static structure and do not involve dynamic information. In fact, structural dynamics can enhance functional prediction, in which the homology modeling and molecular dynamics (MD) simulation have already been extensively used as tools to further access possible functions of several specific fungal transporters (e.g., proline permease [27] and purine and pyrimidine transporters [28]). Moreover, dynamic information from MD simulation revealed the molecular mechanism of the proton pump related to conformational changes during proton translocation through H^+-ATPase [29, 30].

As described above, current approaches can only be performed manually and specifically and cannot be used to describe the relationship between sequence, structure, and function for annotating high-throughput data of transporters. Based on experimental data of *A. oryzae*, very few reports involved in metabolic transporters, such as maltose permease [31, 32], sulphate permease [33], malic acid transporter [6], C_4-dicarboxylate transporter [34], and uric acid-xanthine permease [35], existed. Therefore, the advanced annotation approaches can be used to increase the efficiency of transporter annotation. In this study, we developed a metabolic annotation strategy to determine the relationship between sequence, structure, and function to annotate metabolic transporters in the *A. oryzae* genome. Sequence-based analysis is used to predict transporter genes. Next, candidate transporter genes were subjected to functional classification. The transporters involved in metabolic process were manually curated by integrative analysis (i.e., integrative databases, phylogenetics, protein domains, or transporter components). In addition, the combination of homology modeling and MD simulation was used to determine the relationship between sequence to structure and structure to function. This proposed metabolic annotation strategy can be used to improve the genome-scale metabolic network of *A. oryzae* and relevant fungi.

2. Materials and Methods

2.1. Sequence Alignment Analysis for Transporter Gene Prediction. To identify all possible candidate transporter genes, 12,096 protein sequences from *A. oryzae* genome [8] were searched against protein sequences from two different transporter databases that are available that is, transporter classification database (TCDB) [36] and TransportDB [37] using BLASTP (version 2.2.29$^+$) [18] under bidirectional best-hit and sensitivity analysis [38] as shown in Figure 1 (1st panel). For TCDB, it is a curated transporter database of factual information from over 10,000 published references. Unique proteins in TCDB are deposited over 10,000 sequences

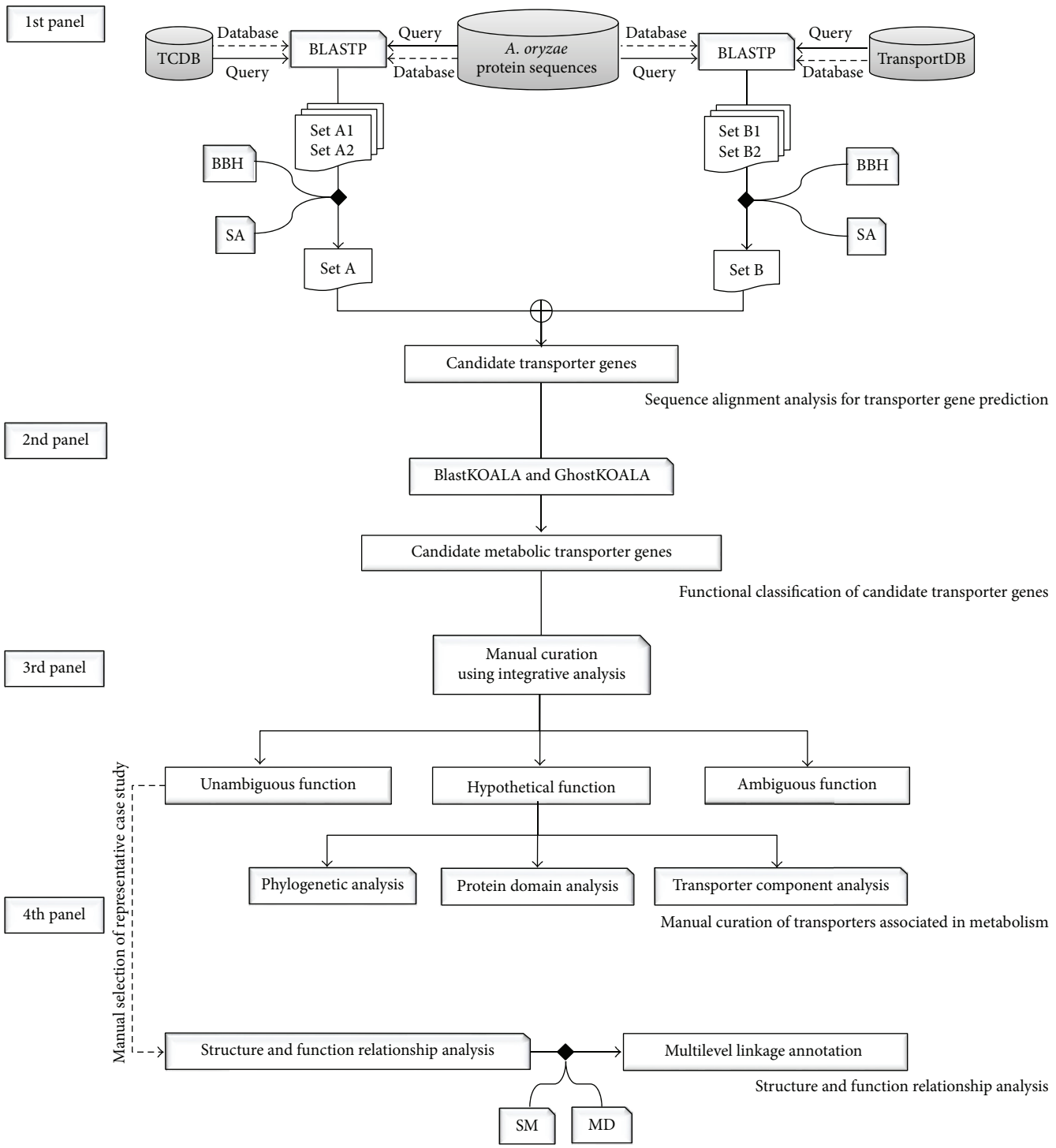


Figure 1: Diagram shows overall framework of a metabolic annotation strategy for linkage between sequence, structure, and function for annotating metabolic transporters in *A. oryzae* genome. In the 1st panel, Sets A and B indicate *A. oryzae* protein sequences searched against TCDB and TransportDB databases, respectively, under bidirectional best-hit analysis (BBH) and sensitivity analysis (SA). In the 4th panel, dash line implies the manual selection of a metabolic transporter from unambiguous function group as a representing case study of multilevel linkage annotation. SM and MD stand for SWISS-MODEL and molecular dynamics simulation, respectively.

which are classified into over 800 transporter families based on the transporter classification (TC) system according to functional and phylogenetic information [39]. In contrast, TransportDB is a relational database describing the predicted transporters based on automated annotation tool for organisms whose complete genome sequences are available [40].

2.2. Functional Classification of Candidate Transporter Genes. For functional classification, the candidate transporter genes obtained were submitted as dataset queries using the BlastKOALA and GhostKOALA annotation tools [41] as shown in Figure 1 (2nd panel). These are KEGG internal annotation tools for assignment of KEGG Orthology (K) number to the query protein sequences by BLAST searching against a nonredundant set of KEGG GENES, which was determined using a 50% identity cut-off [42, 43]. It is noted that GhostKOALA is suitable for annotating a large amount of metagenome sequence data by GHOSTX searching using a cut-off GHOSTX score of 100. After the submission of queries, the annotation data with assigned K numbers was downloaded and used for KEGG Mapper analysis to determine the full details of the assigned K numbers for each candidate transporter gene [41]. The function of candidate transporter gene was then manually classified into two main categories, including (i) metabolic process and (ii) nonmetabolic process. Candidate transporter genes involved in various metabolisms (i.e., energy, lipid, nucleotide, amino acid, glycan, and others) and metabolic transport processes (i.e., solute carrier family, nutrient uptake, and ion channel) were categorized into the metabolic process. Candidate transporter genes related to signaling, cellular, and genetic information were categorized into the nonmetabolic process. Candidate transporter genes with unclassified functions were categorized into the unclassified process. Only candidate transporter genes associated with the metabolic process were subsequently performed by manual curation.

2.3. Manual Curation of Transporters Associated with Metabolism. Candidate transporter genes categorized into the metabolic process were manually curated functions using integrative databases, including TCDB [36], KEGG [42, 43], and PFAM [44], as shown in Figure 1 (3rd panel). If transporters showed the same functions in all the three databases, they were categorized into the unambiguous function group. Otherwise, they were included in the hypothetical function group. These further required additional manual curation for transporter function. Such phylogenetic analysis combined ClustalW [45] with MEGA6 (Molecular Evolutionary Genetics Analysis, version 6.0) [46] and was manually performed to reveal evolutionary relationship of hypothetical metabolic transporter gene based on the maximum likelihood approach [47]. Alternatively, protein domain analysis was performed. Hypothetical metabolic transporter gene was manually submitted to HMMER [48] and MEME [49] and then searched for protein domains using the hidden Markov models [44, 50]. Otherwise, transporter component analysis was done. Hypothetical metabolic transporter gene was manually searched against protein sequences in TCDB based on sequence similarity to identify transporter components. Each component was afterwards curated against several protein databases (e.g., carbohydrate-active enzymes database (CAZy) [51] and Universal Protein Resource (UniProt) database [52]). Transporters showing ambiguity remained in the ambiguous function group.

2.4. Structure and Function Relationship Analysis. Protein structure is more evolutionarily conserved than amino acid sequence. Therefore, the analysis of 3D structures is a promising method for the functional annotation of transporters. Homology modeling was performed as shown in Figure 1 (4th panel). Initially, *A. oryzae* protein sequences belonging to the unambiguous function group were submitted as queries to the SWISS-MODEL [53] for searching the template against the Protein Data Bank (PDB) [54]. Next, a metabolic transporter from unambiguous function group that showed the highest quality with the best-identified structural template (i.e., sequence identity and percent coverage) was manually selected as the representative case study of multilevel linkage annotation. For structure-based sequence alignment of the query and template, the conserved residues between the query and template were retained in the homology model using ProMod II [55]. Remodeling was carried out by substitution of the appropriate amino acids. In order to obtain the homology protein structure, MD simulation was conducted using GROMACS version 4.5.5 [56]. Protein topology was created using the standard GROMOS96 force field parameter set 53a6 [57] and solvated based on the simple point charge water model [58]. To remove steric conflicts between atoms and to avoid high energy interactions, system energy was minimized for 2,000 steps. MD simulation was afterwards run in the NVT (constant particle number, volume, and temperature) ensemble for 100 ns with an integration time step of 1 fs. The temperature was kept constant at 298 K using the V-rescale algorithm with a time constant of 0.1 ps [59–61]. Periodic boundary conditions were applied in all directions. The real-space part of the electrostatic and Lennard-Jones interaction was set at a 1.0 nm cut-off. Long-range electrostatics were calculated using particle-mesh Ewald [62, 63] with a 0.12 nm grid and the cubic interpolation of order four in the reciprocal-space interactions. To avoid physical artifacts, the tested protocol was employed [64–66]. All bond lengths were constrained using the LINCS algorithm [67]. System visualization was performed using Visual Molecular Dynamics software [68]. The structural template was used as the reference, in which the homology model was created and simulated for comparison. At equilibrium, the trajectories were determined as the stability of global protein structure by calculating the root mean square deviation (RMSD) and root mean square fluctuation (RMSF).

3. Results and Discussion

Using our developed metabolic annotation strategy for transporters, we achieved four main results as described in the following. First, we describe the assessment of candidate transporter genes. Next, we present the classified functions of candidate transporter genes. Focusing on metabolic process

Table 1: Number of candidate transporter genes identified by sequence alignment analysis.

Database-based annotation	E-value*	Number of candidate transporter genes
TCDB	$6E-09$	112
TransportDB	$5E-04$	18
		123

*Suitable estimated cut-off values.

category, we describe the manually curated transporters associated in metabolism. To this end, the structure and function relationship assessment of unambiguous metabolic transporter is discussed.

3.1. Assessment of Candidate Transporter Genes. Candidate transporter genes were identified by sequence alignment analysis using 12,096 protein sequences of *A. oryzae* against protein sequences in TCDB and TransportDB. We identified 129 and 23 protein sequences with one-to-one homologous relationship by bidirectional best-hit analysis in TCDB and TransportDB, respectively. These results were subsequently subjected to sensitivity analysis by varying the E-values as cut-offs. The E-values of $6E-09$ and $5E-04$ were selected as the suitable estimated cut-off values. Hereby, we obtained 112 and 18 possible transporter genes from TCDB and TransportDB, respectively. All possible transporter genes under statistical significance were overlapped and removed duplicate data. Consequently, 123 candidate transporter genes of *A. oryzae* were obtained as presented in Table 1. Full list of candidate transporter genes is provided in Table S1 in Supplementary Material available online at http://dx.doi.org/10.1155/2016/8124636.

3.2. Classified Functions of Candidate Transporter Genes. A total of 123 candidate transporter genes were submitted as dataset queries to the BlastKOALA and GhostKOALA annotation tools. Based on the KEGG database results, 87 of the 123 submitted queries were assigned K numbers, which were manually classified into the metabolic process and nonmetabolic process categories (Table S2). As shown in Figure 2, the major category (65 of 123 candidate transporter genes) was in the metabolic process (Table S3), which was divided into seven subcategories, including 41 genes involved in metabolic transport processes, 15 genes involved in energy metabolism, 4 genes involved in glycan metabolism, and 5 genes involved in another four subcategories (Figure 2). In contrast, 17 candidate transporter genes were classified in the nonmetabolic process category, which was divided into two subcategories. These were 8 genes involved in signaling and cellular process and 9 genes involved in genetic information process. It has been reported that transporter genes involved in genetic information and cell signaling process are important in regulation level which can trigger cellular response process by transporting transcription factors, DNA binding protein, mRNA, miRNA, and other related genetic factors across compartments [69]. For candidate transporter genes with unclassified functions (41 genes), they were separated into unclassified process category.

3.3. Manually Curated Transporters Associated with Metabolism. Initially, 65 candidate metabolic transporter genes were manually curated to determine their functions using integrative databases, including TCDB, KEGG, and PFAM (Table S4). The results showed that the transporter functions were classified into three assigned function groups, namely, unambiguous, hypothetical, and ambiguous functions.

For the unambiguous function group, 55 of the 65 transporter genes were manually curated and found to be overlapped among the integration of three databases as summarized in Table 2. The 55 transporter genes were clustered into three classes using the TC system. Seven of the 55 transporter genes were involved in ammonium, magnesium, copper, and water transporters, which belonged to channels and pores (class 1). Most of the unambiguous function group (33 of 55 transporter genes) were involved in electrochemical potential-driven transporters (class 2), such as carbohydrate, amino acid, and nutrient uptake transporters. As example in class 2, AO090009000688 gene was curated as a nucleotide sugar transporter involved in transporting GDP-mannose, which was synthesized in the cytosol and nucleus and transported to the endoplasmic reticulum and the Golgi apparatus for mannosylation process [70]. Dean et al. demonstrated that a mutation in the gene encoding GDP-mannose transporter (*VRG4*) in *S. cerevisiae* caused a loss of mannosylation in *vrg4* mutants, leading to cell death [71]. For gene orthologs of *VRG4* identified in *Aspergillus fumigatus* [72] and *A. nidulans* [73], they were also found to be associated with polysaccharide synthesis during spore germination. In addition, three zinc transporter genes (AO090005000026, AO090011000831, and AO090026000441) corresponded to zinc tolerance and accumulation in *A. oryzae* [74]. Interestingly, large amounts of zinc could be accumulated in mycelial cells of *A. oryzae* [74]. Accordingly, this suggests that zinc transporter can be used to improve the absorption capacity of *A. oryzae* towards pollutant metals. For the other remaining manually curated genes, 15 of 55 transporter genes were functionally assigned for the primary active transporters (class 3). As seen in class 3, observably most of the transporter function utilized energy from ATP hydrolysis to transport ions through cellular membranes against a concentration gradient [29] (Table 2). For instance, AO090102001037 gene encoding proton-translocating transhydrogenase can hydrolyze ATP to transport proton through cellular membrane. Notably, this AO090102001037 gene showed evolutionary relationship among *Aspergillus* species in terms of gene sequence and expression [75].

For the hypothetical function group, 3 of the 65 transporter genes (i.e., AO090001000747, AO090023000801, and AO090005000980) were manually curated for individual transporter function by either phylogenetic, protein domain, or transporter component analysis, respectively.

Performing phylogenetic analysis, the hypothetical metabolic transporter gene, for example, AO090001000747 in *A. oryzae* and oligosaccharyl transferase (OST3) in *S. cerevisiae,* showed a closer evolutionary relationship than magnesium transporter (MAGT1) in *Homo sapiens* as illustrated in Figure 3. As a result, it is promising that AO090001000747 gene

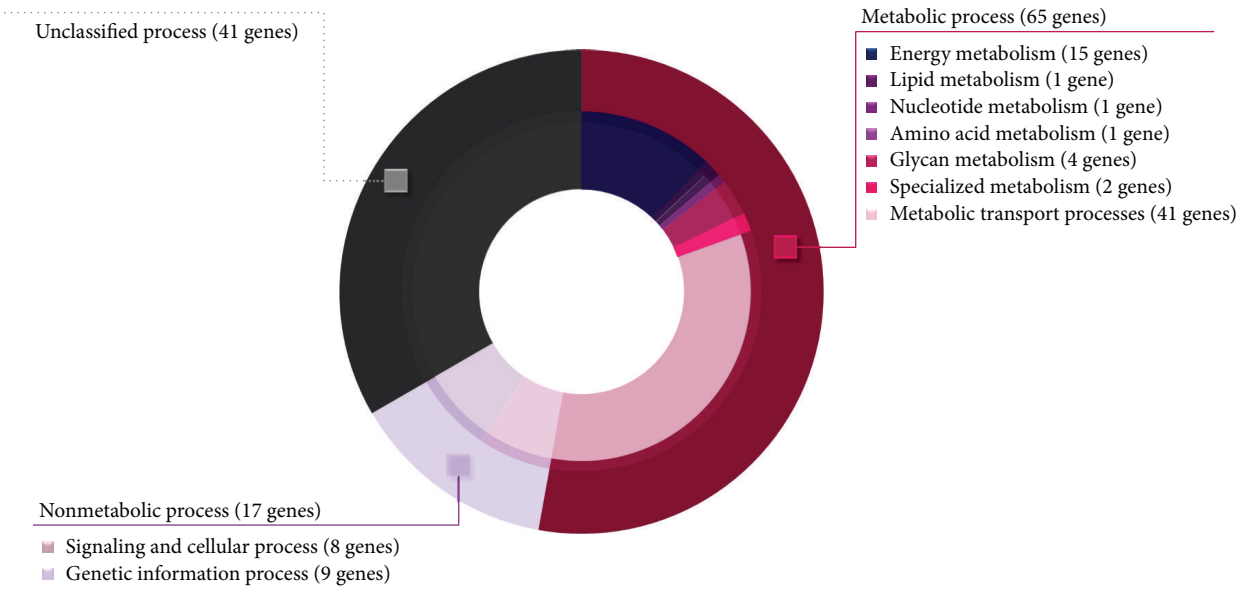


Figure 2: Doughnut chart illustrates different functional categories of *A. oryzae* candidate transporter genes. Outer layer shows three main functional categories (i.e., metabolic, nonmetabolic, and unclassified processes). Inner layer shows seven subcategories distributed into metabolic process and two subcategories distributed into nonmetabolic process. Ring size reflects the relative ratio of genes identified in each category.

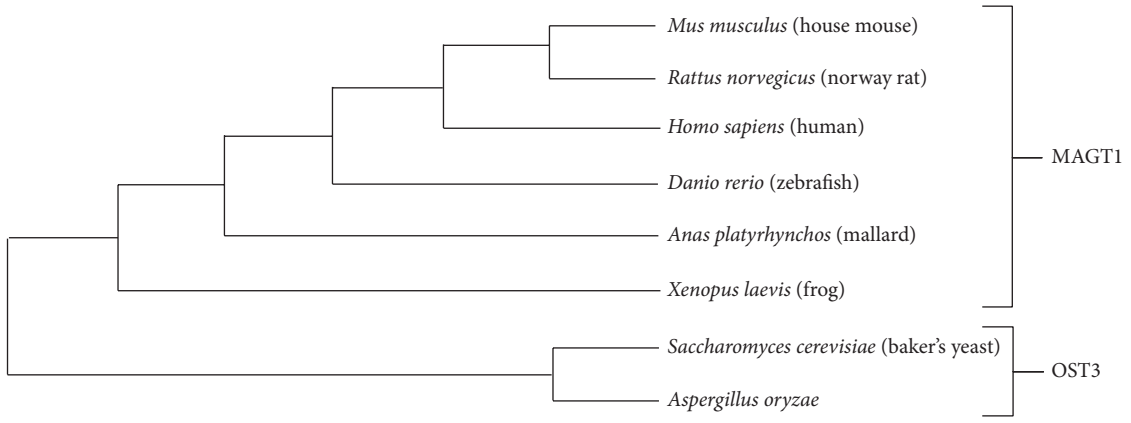


Figure 3: Horizontal cladogram shows an evolutionary relationship of oligosaccharyltransferase (OST3) and magnesium transporter (MAGT1) among *A. oryzae* and 7 different model organisms (i.e., *Mus musculus*, *Rattus norvegicus*, *H. sapiens*, *Danio rerio*, *Anas platyrhynchos*, *Xenopus laevis*, and *S. cerevisiae*). The figure is generated by the MEGA6 [46] and ClustalW [45].

is potentially encoded for the endoplasmic reticulum resident oligosaccharide transporter involved in N-glycosylation according to the function of OST3 in *S. cerevisiae* [76]. Previously, it has been reported that OST3 is a gate keeper for the secretory pathway [77] and it can catalyze the priority step in protein secretion [78]. Therefore, the significant transcriptional upregulation of AO090001000747 gene (OST3 ortholog) was accordingly reported in an *A. oryzae* alpha-amylase overproducing strain [1]. Our finding implies that AO090001000747 gene is contributed for transporting and encompassing secretory proteins, which is favorable for increasing the efficiency of commercial protein secretion in *A. oryzae*. Full details of horizontal cladogram can be seen in Figure S1.

Considering protein domain analysis, it is an alternative way for manual curation of transporter function. Once HMMER [48] and MEME [49] were used for searching the protein domains of hypothetical metabolic transporter gene, for example, AO090023000801, observably this gene contains the conserved carboxylase domain which represents a conserved region in pyruvate carboxylase and oxaloacetate decarboxylase. A report by Knuf et al. supported that AO090023000801 gene encoding pyruvate carboxylase was involved in organic acid production [6]. Besides, a manual sequence searching by TCDB [36] also supported that AO090023000801 gene encoding oxaloacetate decarboxylase was involved in sodium transport. These results thus imply that the AO090023000801 gene may have two transporter

Table 2: List of manually curated transporter genes and functions in unambiguous function group.

Name of transporter gene	TCID	Name of transporter function*
Class 1: channels and pores		
AO090023000569	1.A.1.7.1	Outward-rectifier potassium channel
AO090038000314	1.A.11.3.2	Ammonium transporter
AO090003001402	1.A.35	Magnesium transporter
AO090120000141	1.A.35.5.1	Magnesium transporter
AO090120000214	1.A.56.1.4	Copper transporter
AO090011000329	1.A.8.8.8	Aquaporin
AO090023000895	1.B.8.1.1	Voltage-dependent anion channel porin
Class 2: electrochemical potential-driven transporters		
AO090003000050	2.A.1.7.1	L-fucose permease
AO090012000623	2.A.1.8.5	Nitrate transporter
AO090010000135	2.A.100.1.3	Iron-regulated transporter
AO090010000229	2.A.17.2.2	Proton-dependent oligopeptide transporter
AO090026000828	2.A.19.4.4	Sodium/potassium/calcium exchanger
AO090009000637	2.A.2.6.1	Alpha-glucoside permease
AO090003001404	2.A.20	Phosphate transporter
AO090012000901	2.A.20.2.2	Phosphate transporter
AO090103000274	2.A.22.3.2	Sodium and chloride dependent GABA transporter
AO090009000405	2.A.29.1.3	Mitochondrial adenine nucleotide translocator
AO090005000114	2.A.3.10.2	Amino acid transporter
AO090009000636	2.A.36.1.12	Sodium/hydrogen exchanger
AO090005000019	2.A.39.3.1	Allantoin permease
AO090005000455	2.A.40.5.1	Purine permease
AO090003000443	2.A.41.2.7	H^+/nucleoside cotransporter
AO090003000920	2.A.47.2.2	Phosphate transporter
AO090026000432	2.A.49.1.3	Chloride channel
AO090005000026	2.A.5.1.1	Zinc transporter
AO090011000831	2.A.5.5.1	Zinc transporter
AO090026000441	2.A.5.7.1	Zinc transporter
AO090011000817	2.A.52.1.3	Nickel transporter
AO090003000798	2.A.53.1.2	Sodium-independent sulfate anion transporter
AO090003001119	2.A.55.1.1	High-affinity metal uptake transporter
AO090003001233	2.A.57.3.1	Nucleoside transporter
AO090005001332	2.A.59.1.1	Arsenite transporter
AO090120000217	2.A.6.6.5	Hydroxymethylglutaryl-CoA reductase
AO09M000000016	2.A.63	NADH-ubiquinone oxidoreductase
AO090001000748	2.A.66	Polysaccharide exporter
AO090010000775	2.A.7.10.2	UDP-xylose/UDP-N-acetylglucosamine transporter
AO090009000400	2.A.7.11.1	UDP-galactose transporter
AO090009000688	2.A.7.13.2	GDP-mannose transporter
AO090026000255	2.A.72.3.2	Potassium transporter
AO090005001455	2.A.97.1.4	Potassium and hydrogen ion antiporter
Class 3: primary active transporters		
AO090009000651	3.A.1.201.11	Multidrug resistance protein 1
AO090038000399	3.A.1.31.1	Possible ABC transporter permease for cobalt
AO090003000688	3.A.19.1.1	Arsenite-translocating ATPase
AO090010000482	3.A.2	V-type ATPases
AO09M000000001	3.A.2.1.3	F-type ATPase
AO090012000797	3.A.2.2.3	V-type ATPase

Table 2: Continued.

Name of transporter gene	TCID	Name of transporter function*
AO090038000088	3.A.3.1.7	P-type ATPase
AO090012000773	3.A.3.10.1	P-type ATPase
AO090038000322	3.A.3.2.2	P-type ATPase
AO090005000842	3.A.3.3.6	Plasma membrane proton ATPase
AO09M000000013	3.D.1.2.1	NADH dehydrogenase
AO09M000000015	3.D.1.6.2	NADH-ubiquinone oxidoreductase
AO090102001037	3.D.2.4.1	Proton-translocating transhydrogenase
AO090010000475	3.D.3.2.1	Cytochrome b-c1 complex subunit Rieske
AO09M000000014	3.D.4.8.1	Cytochrome oxidase

*Names of transporter functions are based on KEGG, PFAM, and UniProt databases.

functions related to the conserved region. For the other transporter component analysis, the hypothetical metabolic transporter gene, for example, AO090005000980, was manually searched against protein sequences in TCDB [36] based on sequence similarity to identify transporter components. Accordingly, AO090005000980 gene was identified as potassium transporter (Ktr) containing three different components (i.e., the potassium-translocating protein (KtrB), regulatory protein (KtrA), and Slr1508 protein). Using CAZy [51] and UniProt [52], the protein function of Slr1508 was glycosyl transferase involved in glycosylphosphatidyl inositol anchor formation. After using PFAM [44], the results also supported that the Slr1508 protein has glycosyl transferase function. These suggest that the AO090005000980 gene may have two transporter functions relevant to the transporter components. Transporter genes showing functional ambiguity remained in the ambiguous function group (7 of 65 transporter genes), namely, genes AO090005001300, AO090120000224, AO090011000320, AO090020000415, AO090020000492, AO090010000212, and AO090012000733.

3.4. Structure and Function Relationship Assessment of Unambiguous Metabolic Transporter. To ensure the functional role of the unambiguous metabolic transporter, a combination of homology modeling and MD simulation was used to assess the relationship between sequence to structure and structure to function, which provides stronger evidence for functional conservation and annotation of transporter beyond sequence-based analysis. To do this, a metabolic transporter from unambiguous function group was manually selected based on the central transporter role in metabolism of *A. oryzae* with the highest sequence identity and coverage from sequence alignment analysis between the query (e.g., metabolic transporter gene) and the well-known structure and function of transporter in PDB. Among the unambiguous metabolictransporters, favorably AO090005000842 gene encoding for H^+-ATPase was selected as a representative case study of multilevel linkage annotation due to the highest sequence identity and percent coverage between AO090005000842 gene and the well-known structure and function of the H^+-ATPase of *Neurospora crassa*. To elaborate, AO090005000842 gene was initially submitted as a query onto the SWISS-MODEL for template searching against the PDB. According to the highest quality results among the top 10 identified templates (Table S5), the electron crystallography structure of H^+-ATPase in *N. crassa* (PDB ID: 1MHS) [30] showed the highest sequence identity (77.47%) and percent coverage (94%). Therefore, 1MHS was used as a template for the homology modeling of *A. oryzae* H^+-ATPase. Thus, the model was generated with detailed sequence alignment between H^+-ATPase in *A. oryzae* and *N. crassa* (Figures 4(a) and S2). Overall, 681 residues in the five principal domains were identical in both proteins, as shown in Figure 4(b). The most homologous domain was the phosphorylation (P) domain (92.12%), followed by the cluster of 10 transmembrane helices (M1-2, M3-4, and M5–10) in the membrane domain (80.52%), the nucleotide-binding domain (72.30%), the actuator domain (64.13%), and the regulatory domain (60.53%). Additional details are shown in Table S6.

In addition to the analysis of static structures by homology modeling, MD simulation was carried out in order to evaluate structural stability during dynamics simulation and the changes in the stability of proton-transporting regions compared with template structures. The dynamics systems of both H^+-ATPase models were created under the GROMOS96 force field and solvated in a simple point charge water model without constraints. These systems were then subjected to MD simulation for 100 ns while monitoring equilibration by examining the stability of the geometrical property (RMSD) of the H^+-ATPase models. Subsequently, the RMSD and RMSF were calculated using the trajectories to quantify the stability and the fluctuation of the protein. The RMSD of global structures of the H^+-ATPase in *A. oryzae* and *N. crassa* reached equilibrium after 50 ns using the quantities as shown in Figure S3. Indeed, all five principal domains in the *A. oryzae* and *N. crassa* H^+-ATPases shared the same average RMSD over the equilibrium which indicated that the dynamic behavior of functional domains was conserved among these species (Table S7).

In fact, the proton-transport region (M-domain) of H^+-ATPase is embedded in membrane environment. Therefore, M-domain of *A. oryzae* H^+-ATPase embedding in palmitoyl oleoylphosphatidylcholine (POPC) lipid bilayer was conducted using the MD simulation. The insertion of M-domain into membrane was done as followed by Kandt et al. [79, 80] (Figure S4). The simulation was performed under NPT (constant particle number, pressure, and temperature) ensemble.

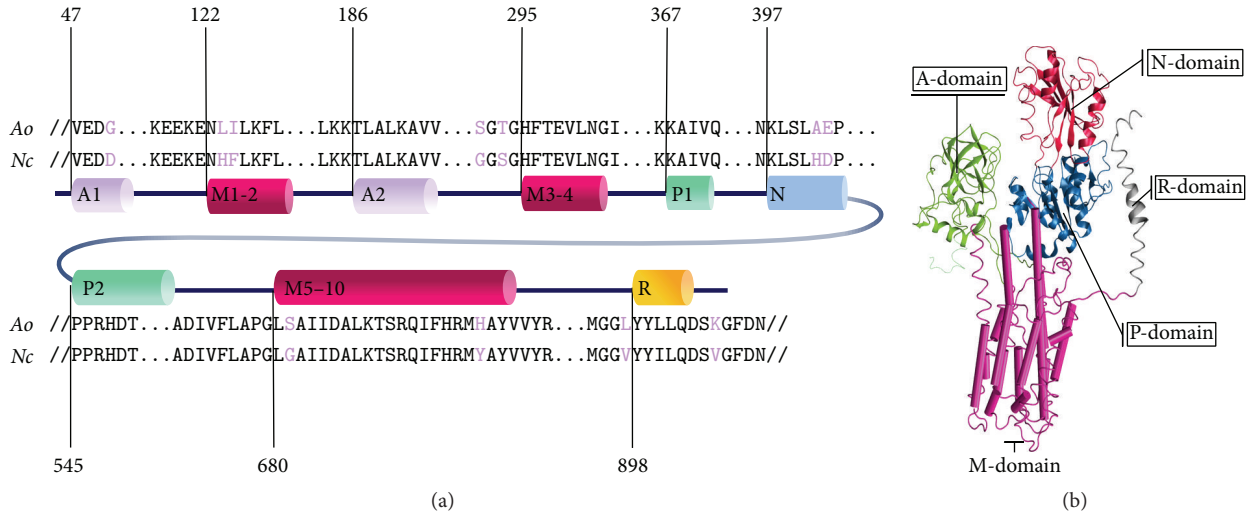


FIGURE 4: Diagram shows sequence alignment between the H^+-ATPase in *A. oryzae* (*Ao*) and *N. crassa* (*Nc*) (PDB ID: 1MHS) [30] in (a) and structural template with five principle domains distinguished with different colors in (b). For both (a) and (b), A1-2 indicates actuator (A) domain shaded in green, P1-2 indicates the phosphorylation (P) domain shaded in blue, N indicates the nucleotide-binding (N) domain shaded in red, M1-2, M3-4, and M5–10 indicate the transmembrane (M) domain shaded in pink, and R indicates the regulatory (R) domain of the H^+-ATPase shaded in grey.

Semi-isotropic pressure was applied by the Berendsen algorithm, at a pressure of 1 bar in both the *xy*-plane and the *z*-direction (bilayer normal) with a time constant of 3.0 ps and a compressibility of 4.5×10^{-5} bar^{-1} [59–61]. The simulation was run for 25 ns and the last 15 ns was used for analysis. The results showed that the average RMSD of the proton-transporting regions, M-domain embedding in POPC of *A. oryzae,* H^+-ATPase was 0.398 ± 0.007 nm (Figure S5). This RMSD result supported that the M-domain embedding in POPC of *A. oryzae* H^+-ATPase was consistently preserved with the corresponding regions in the initial structure of *N. crassa* H^+-ATPase.

In addition, the proton-transporting unit of the H^+-ATPase is defined by the presence critical proton-binding sites along proton translocation path in M-domain [81]. Such mutational H^+-ATPase studies in plants demonstrated that substitution of Asp684 with Asn led to a defect in the conformational change for transporting protons but did not abolish the ability to bind to nucleotides and hydrolyze ATP [82]. Consistently, the substitutions of Asp730 in *N. crassa* H^+-ATPase disrupted a salt bridge between Asp730 and Arg695, preventing the transport of protons along the proton cavity [83]. Similar structural arrangements in the proton-transporting path included positions for each conserved polar and charged residue, which may promote efficient proton transport [81]. Thus, the overall equivalent residues for proton translocation must conserve in identity and position. Therefore, fluctuations in the corresponding proton-binding sites in the *A. oryzae* H^+-ATPase, including basic side chains (Arg705 and His711 on M5), acidic side chains (Asp740 on M6, Glu815 on M8), and polar side chains (Tyr704 and Ser709 on M5, Thr743 on M6), were expected to show RMSF values comparable to those of the *N. crassa* H^+-ATPase (Figure 5). The RMSF of individual equivalent residues in the *A. oryzae* H^+-ATPase also matched with their corresponding sites in the *N. crassa* H^+-ATPase (Table S8). For instance, the acidic side chain Arg705 and the basic side chain Asp740 in the *A. oryzae* H^+-ATPase fluctuated with the RMSF by approximately 0.0737 nm and 0.1530 nm, respectively, which are the corresponding sites in the *N. crassa* H^+-ATPase, Arg695 (0.0770 nm), and Asp730 (0.1118 nm).

In accordance with the overall comparable geometrical properties, the *A. oryzae* and *N. crassa* H^+-ATPase models were substantiated for their structural conservation at the dynamic level. Taken together, the integrative results derived from homology modeling and MD simulation supported that the proton-transporting role along the proton-transporting path in transmembrane domain was structurally conserved between H^+-ATPases in *A. oryzae* and *N. crassa*, where functional conservation for the proton transporter is expected.

4. Conclusion

For the integrative multilevel annotation of metabolic transporters, we propose a metabolic annotation and assessment strategy based on sequence, structure, and function relationship as a platform for increasing the functional efficiency of transporter annotation. Of 12,096 total genes in the *A. oryzae* genome, our strategy could be used to identify 58 metabolic transporter genes. Under consensus integrative databases, 55 unambiguous metabolic transporter genes were distributed into channels and pores (7 genes), electrochemical potential-driven transporters (33 genes), and primary active transporters (15 genes). The remaining 3 hypothetical metabolic transporter genes were manually curated transporter functions by phylogenetic, protein domain, and transporter component analysis. Among the unambiguous metabolic transporter genes, the H^+-ATPase or proton pump encoded by the AO090005000842 gene was selected as a representative case study of multilevel linkage annotation in

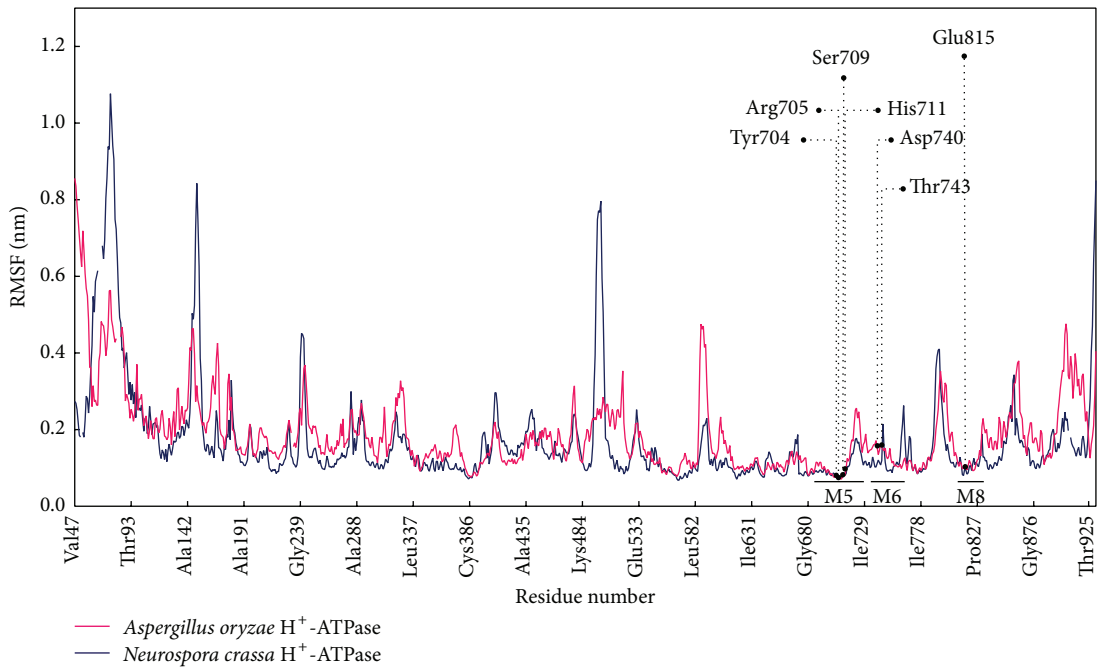


Figure 5: Diagram shows the comparable RMSF between the *A. oryzae* and *N. crassa* H^+-ATPases. This graph is generated using the data in Table S8.

order to reveal the transporter functional role in *A. oryzae* metabolism. Our metabolic annotation strategy can be used for improving functional annotation and enhancing cellular metabolic network and modeling in *A. oryzae* and relevant fungi.

Competing Interests

The authors declare that there is no competing interests regarding the publication of this paper and regarding the funding/grants that they have received.

Acknowledgments

This work was financially supported by Kasetsart University Research and Development Institute (KURDI) at Kasetsart University. Nachon Raethong thanks Science Achievement Scholarship of Thailand (SAST), Department of Zoology, and the Graduate School at Kasetsart University. Wanwipa Vongsangnak and Jirasak Wong-ekkabut gratefully acknowledge financial support from the Faculty of Science at Kasetsart University (Grant nos. PRF4/2558 and PRF-PII/59). The authors also acknowledge Computational Biomodelling Laboratory for Agricultural Science and Technology (CBLAST) at Kasetsart University for computing facilities and resources.

References

[1] L. Liu, A. Feizi, T. Österlund, C. Hjort, and J. Nielsen, "Genome-scale analysis of the high-efficient protein secretion system of *Aspergillus oryzae*," *BMC Systems Biology*, vol. 8, article 73, 2014.

[2] S. O. Budak, M. Zhou, C. Brouwer et al., "A genomic survey of proteases in Aspergilli," *BMC Genomics*, vol. 15, no. 1, article 523, 2014.

[3] H. Hisada, M. Sano, H. Ishida, Y. Hata, and M. Machida, "Identification of regulatory elements in the glucoamylase-encoding gene (glaB) promoter from *Aspergillus oryzae*," *Applied Microbiology and Biotechnology*, vol. 97, no. 11, pp. 4951–4956, 2013.

[4] X. Yin, Y.-Y. Gong, J.-Q. Wang, C.-D. Tang, and M.-C. Wu, "Cloning and expression of a family 10 xylanase gene (Aoxyn10) from *Aspergillus oryzae* in *Pichia pastoris*," *Journal of General and Applied Microbiology*, vol. 59, no. 6, pp. 405–415, 2013.

[5] T. Kobayashi, K. Abe, K. Asai et al., "Genomics of *Aspergillus oryzae*," *Bioscience, Biotechnology and Biochemistry*, vol. 71, no. 3, pp. 646–670, 2007.

[6] C. Knuf, I. Nookaew, I. Remmers et al., "Physiological characterization of the high malic acid-producing *Aspergillus oryzae* strain 2103a-68," *Applied Microbiology and Biotechnology*, vol. 98, no. 8, pp. 3517–3527, 2014.

[7] K. Abe, K. Gomi, F. Hasegawa, and M. Machida, "Impact of *Aspergillus oryzae* genomics on industrial production of metabolites," *Mycopathologia*, vol. 162, no. 3, pp. 143–153, 2006.

[8] M. Machida, K. Asai, M. Sano et al., "Genome sequencing and analysis of *Aspergillus oryzae*," *Nature*, vol. 438, no. 7071, pp. 1157–1161, 2005.

[9] M. Umemura, Y. Koyama, I. Takeda et al., "Fine *de novo* sequencing of a fungal genome using only SOLiD short read data: verification on *Aspergillus oryzae* RIB40," *PLoS ONE*, vol. 8, no. 5, Article ID e63673, 2013.

[10] T. Ikegami, T. Inatsugi, I. Kojima et al., "Hybrid *de novo* genome assembly using MiSeq and SOLiD short read data," *PLoS ONE*, vol. 10, no. 4, Article ID e0126289, 2015.

[11] W. Vongsangnak, P. Olsen, K. Hansen, S. Krogsgaard, and J. Nielsen, "Improved annotation through genome-scale

metabolic modeling of *Aspergillus oryzae*," *BMC Genomics*, vol. 9, article 245, 2008.

[12] K. Tamano, M. Sano, N. Yamane et al., "Transcriptional regulation of genes on the non-syntenic blocks of *Aspergillus oryzae* and its functional relationship to solid-state cultivation," *Fungal Genetics and Biology*, vol. 45, no. 2, pp. 139–151, 2008.

[13] M. R. Andersen, W. Vongsangnak, G. Panagiotou, M. P. Salazar, L. Lehmann, and J. Nielsen, "A trispecies *Aspergillus* microarray: comparative transcriptomics of three *Aspergillus* species," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 11, pp. 4387–4392, 2008.

[14] B. Wang, G. Guo, C. Wang et al., "Survey of the transcriptome of *Aspergillus oryzae* via massively parallel mRNA sequencing," *Nucleic Acids Research*, vol. 38, no. 15, Article ID gkq256, pp. 5075–5087, 2010.

[15] G. Xu, W. Zou, X. Chen, N. Xu, L. Liu, and J. Chen, "Fumaric acid production in *Saccharomyces cerevisiae* by *in silico* aided metabolic engineering," *PLoS ONE*, vol. 7, no. 12, Article ID e52086, 2012.

[16] L. Karaffa and C. P. Kubicek, "*Aspergillus niger* citric acid accumulation: do we understand this well working black box?" *Applied Microbiology and Biotechnology*, vol. 61, no. 3, pp. 189–196, 2003.

[17] S. Sahoo, M. K. Aurich, J. J. Jonsson, and I. Thiele, "Membrane transporters in a human genome-scale metabolic knowledgebase and their implications for disease," *Frontiers in Physiology*, vol. 5, article 91, 2014.

[18] S. F. Altschul, T. L. Madden, A. A. Schäffer et al., "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," *Nucleic Acids Research*, vol. 25, no. 17, pp. 3389–3402, 1997.

[19] F. Sievers and D. G. Higgins, "Clustal omega, accurate alignment of very large numbers of sequences," *Methods in Molecular Biology*, vol. 1079, pp. 105–116, 2014.

[20] C. Struck, "Amino acid uptake in rust fungi," *Frontiers in Plant Science*, vol. 6, article 40, 2015.

[21] G. A. Petsko and D. Ringe, *Protein Structure and Function*, Oxford University Press, Oxford, UK; Sinauer Associates, Sunderland, Mass, USA, 2009.

[22] J. Zhao, S. Benlekbir, and J. L. Rubinstein, "Electron cryomicroscopy observation of rotational states in a eukaryotic V-ATPase," *Nature*, vol. 521, no. 7551, pp. 241–245, 2015.

[23] B. P. Pedersen, H. Kumar, A. B. Waight et al., "Crystal structure of a eukaryotic phosphate transporter," *Nature*, vol. 496, no. 7446, pp. 533–536, 2013.

[24] M. Sanguinetti, S. Amillis, S. Pantano, C. Scazzocchio, and A. Ramón, "Modelling and mutational analysis of *Aspergillus nidulans* UreA, a member of the subfamily of urea/H^+ transporters in fungi and plants," *Open Biology*, vol. 4, no. 6, Article ID 140070, 2014.

[25] S. Faham, A. Watanabe, G. M. Besserer et al., "The crystal structure of a sodium galactose transporter reveals mechanistic insights into Na^+/sugar symport," *Science*, vol. 321, no. 5890, pp. 810–814, 2008.

[26] S. Weyand, T. Shimamura, S. Yajima et al., "Structure and molecular mechanism of a nucleobase-cation-symport-1 family transporter," *Science*, vol. 322, no. 5902, pp. 709–713, 2008.

[27] C. Gournas, T. Evangelidis, A. Athanasopoulos, E. Mikros, and V. Sophianopoulou, "The *Aspergillus nidulans* proline permease as a model for understanding the factors determining substrate binding and specificity of fungal amino acid transporters," *The Journal of Biological Chemistry*, vol. 290, no. 10, pp. 6141–6155, 2015.

[28] E. Krypotou, T. Evangelidis, J. Bobonis et al., "Origin, diversification and substrate specificity in the family of NCS1/FUR transporters," *Molecular Microbiology*, vol. 96, no. 5, pp. 927–950, 2015.

[29] W. Kühlbrandt, "Biology, structure and mechanism of P-type ATPases," *Nature Reviews Molecular Cell Biology*, vol. 5, no. 4, pp. 282–295, 2004.

[30] W. Kühlbrandt, J. Zeelen, and J. Dietrich, "Structure, mechanism, and regulation of the *Neurospora* plasma membrane H^+-ATPase," *Science*, vol. 297, no. 5587, pp. 1692–1696, 2002.

[31] T. Hiramoto, M. Tanaka, T. Ichikawa et al., "Endocytosis of a maltose permease is induced when amylolytic enzyme production is repressed in *Aspergillus oryzae*," *Fungal Genetics and Biology*, vol. 82, pp. 136–144, 2015.

[32] W. Vongsangnak, M. Salazar, K. Hansen, and J. Nielsen, "Genome-wide analysis of maltose utilization and regulation in aspergilli," *Microbiology*, vol. 155, no. 12, pp. 3893–3902, 2009.

[33] Y. Toyoshima, A. Takahashi, H. Tanaka et al., "Lethal and mutagenic effects of ion beams and γ-rays in *Aspergillus oryzae*," *Mutation Research—Fundamental and Molecular Mechanisms of Mutagenesis*, vol. 740, no. 1-2, pp. 43–49, 2012.

[34] S. H. Brown, L. Bashkirova, R. Berka et al., "Metabolic engineering of *Aspergillus oryzae* NRRL 3488 for increased production of L-malic acid," *Applied Microbiology and Biotechnology*, vol. 97, no. 20, pp. 8903–8912, 2013.

[35] Y. Higuchi, T. Nakahama, J.-Y. Shoji, M. Arioka, and K. Kitamoto, "Visualization of the endocytic pathway in the filamentous fungus *Aspergillus oryzae* using an EGFP-fused plasma membrane protein," *Biochemical and Biophysical Research Communications*, vol. 340, no. 3, pp. 784–791, 2006.

[36] M. H. Saier Jr., V. S. Reddy, D. G. Tamang, and Å. Västermark, "The transporter classification database," *Nucleic Acids Research*, vol. 42, no. 1, pp. D251–D258, 2014.

[37] Q. Ren, K. Chen, and I. T. Paulsen, "TransportDB: a comprehensive database resource for cytoplasmic membrane transport systems and outer membrane channels," *Nucleic Acids Research*, vol. 35, supplement 1, pp. D274–D279, 2007.

[38] M. Huynen, B. Snel, W. Lathe III, and P. Bork, "Predicting protein function by genomic context: quantitative evaluation and qualitative inferences," *Genome Research*, vol. 10, no. 8, pp. 1204–1210, 2000.

[39] M. H. Saier Jr., V. S. Reddy, B. V. Tsu, M. S. Ahmed, C. Li, and G. Moreno-Hagelsieb, "The Transporter Classification Database (TCDB): recent advances," *Nucleic Acids Research*, vol. 44, no. 1, pp. D372–D379, 2016.

[40] Q. Ken and J. T. Pauisers, "Comparative analyses of fundamental differences in membrane transport capabilities in prokaryotes and eukaryotes," *PLoS Computational Biology*, vol. 1, no. 3, article e27, 2005.

[41] S. Okuda, T. Yamada, M. Hamajima et al., "KEGG Atlas mapping for global analysis of metabolic pathways," *Nucleic Acids Research*, vol. 36, pp. W423–W426, 2008.

[42] M. Kanehisa, S. Goto, Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe, "Data, information, knowledge and principle: back to metabolism in KEGG," *Nucleic Acids Research*, vol. 42, no. 1, pp. D199–D205, 2014.

[43] M. Kanehisa and S. Goto, "KEGG: kyoto encyclopedia of genes and genomes," *Nucleic Acids Research*, vol. 28, no. 1, pp. 27–30, 2000.

[44] R. D. Finn, J. Mistry, B. Schuster-Böckler et al., "PFAM: clans, web tools and services," *Nucleic Acids Research*, vol. 34, pp. D247–D251, 2006.

[45] M. A. Larkin, G. Blackshields, N. P. Brown et al., "Clustal W and Clustal X version 2.0," *Bioinformatics*, vol. 23, no. 21, pp. 2947–2948, 2007.

[46] K. Tamura, G. Stecher, D. Peterson, A. Filipski, and S. Kumar, "MEGA6: molecular evolutionary genetics analysis version 6.0," *Molecular Biology and Evolution*, vol. 30, no. 12, pp. 2725–2729, 2013.

[47] S. Guindon and O. Gascuel, "A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood," *Systematic Biology*, vol. 52, no. 5, pp. 696–704, 2003.

[48] R. D. Finn, J. Clements, W. Arndt et al., "HMMER web server: 2015 update," *Nucleic Acids Research*, vol. 43, no. 1, pp. W30–W38, 2015.

[49] T. L. Bailey, M. Boden, F. A. Buske et al., "MEME Suite: tools for motif discovery and searching," *Nucleic Acids Research*, vol. 37, no. 2, pp. W202–W208, 2009.

[50] S. R. Eddy, "Hidden Markov models," *Current Opinion in Structural Biology*, vol. 6, no. 3, pp. 361–365, 1996.

[51] V. Lombard, H. Golaconda Ramulu, E. Drula, P. M. Coutinho, and B. Henrissat, "The carbohydrate-active enzymes database (CAZy) in 2013," *Nucleic Acids Research*, vol. 42, no. 1, pp. D490–D495, 2014.

[52] The UniProt Consortium, "Activities at the universal protein resource (UniProt)," *Nucleic Acids Research*, vol. 42, pp. D191–D198, 2014.

[53] L. Bordoli, F. Kiefer, K. Arnold, P. Benkert, J. Battey, and T. Schwede, "Protein structure homology modeling using SWISS-MODEL workspace," *Nature Protocols*, vol. 4, no. 1, pp. 1–13, 2009.

[54] H. Berman, K. Henrick, H. Nakamura, and J. L. Markley, "The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data," *Nucleic Acids Research*, vol. 35, supplement 1, pp. D301–D303, 2007.

[55] N. Guex and M. C. Peitsch, "SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling," *Electrophoresis*, vol. 18, no. 15, pp. 2714–2723, 1997.

[56] B. Hess, C. Kutzner, D. van der Spoel, and E. Lindahl, "GROMACS 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation," *Journal of Chemical Theory and Computation*, vol. 4, no. 3, pp. 435–447, 2008.

[57] C. Oostenbrink, A. Villa, A. E. Mark, and W. F. Van Gunsteren, "A biomolecular force field based on the free enthalpy of hydration and solvation: the GROMOS force-field parameter sets 53A5 and 53A6," *Journal of Computational Chemistry*, vol. 25, no. 13, pp. 1656–1676, 2004.

[58] H. J. C. Berendsen, J. P. M. Postma, W. F. Gunsteren, and J. Hermans, "Interaction models for water in relation to protein hydration," in *Intermolecular Forces: Proceedings of the Fourteenth Jerusalem Symposium on Quantum Chemistry and Biochemistry Held in Jerusalem, Israel, April 13–16, 1981*, B. Pullman, Ed., pp. 331–342, Springer, Dordrecht, The Netherlands, 1981.

[59] G. Bussi, F. L. Gervasio, A. Laio, and M. Parrinello, "Free-energy landscape for β hairpin folding from combined parallel tempering and metadynamics," *Journal of the American Chemical Society*, vol. 128, no. 41, pp. 13435–13441, 2006.

[60] G. Bussi, T. Zykova-Timan, and M. Parrinello, "Isothermal-isobaric molecular dynamics using stochastic velocity rescaling," *The Journal of Chemical Physics*, vol. 130, no. 7, Article ID 074101, 2009.

[61] W. F. van Gunsteren and H. J. C. Berendsen, "Algorithms for macromolecular dynamics and constraint dynamics," *Molecular Physics*, vol. 34, no. 5, pp. 1311–1327, 1977.

[62] U. Essmann, L. Perera, M. L. Berkowitz, T. Darden, H. Lee, and L. G. Pedersen, "A smooth particle mesh Ewald method," *The Journal of Chemical Physics*, vol. 103, no. 19, pp. 8577–8593, 1995.

[63] T. Darden, D. York, and L. Pedersen, "Particle mesh Ewald: an N·log(N) method for Ewald sums in large systems," *The Journal of Chemical Physics*, vol. 98, no. 12, pp. 10089–10092, 1993.

[64] J. Wong-ekkabut and M. Karttunen, "The good, the bad and the user in soft matter simulations," *Biochimica et Biophysica Acta (BBA)—Biomembranes*, 2016.

[65] J. Wong-Ekkabut and M. Karttunen, "Assessment of common simulation protocols for simulations of nanopores, membrane proteins, and channels," *Journal of Chemical Theory and Computation*, vol. 8, no. 8, pp. 2905–2911, 2012.

[66] M. Patra, M. Karttunen, M. T. Hyvönen, E. Falck, and I. Vattulainen, "Lipid bilayers driven to a wrong lane in molecular dynamics simulations by subtle changes in long-range electrostatic interactions," *Journal of Physical Chemistry B*, vol. 108, no. 14, pp. 4485–4494, 2004.

[67] B. Hess, H. Bekker, H. J. C. Berendsen, and J. G. E. M. Fraaije, "LINCS: a linear constraint solver for molecular simulations," *Journal of Computational Chemistry*, vol. 18, no. 12, pp. 1463–1472, 1997.

[68] W. Humphrey, A. Dalke, and K. Schulten, "VMD: visual molecular dynamics," *Journal of Molecular Graphics*, vol. 14, no. 1, pp. 33–38, 1996.

[69] Å. Västermark and M. H. Saier Jr., "The involvement of transport proteins in transcriptional and metabolic regulation," *Current Opinion in Microbiology*, vol. 18, no. 1, pp. 8–15, 2014.

[70] B. Hadley, A. Maggioni, A. Ashikov, C. J. Day, T. Haselhorst, and J. Tiralongo, "Structure and function of nucleotide sugar transporters: current progress," *Computational and Structural Biotechnology Journal*, vol. 10, no. 16, pp. 23–32, 2014.

[71] N. Dean, Y. B. Zhang, and J. B. Poster, "The VRG4 gene is required for GDP-mannose transport into the lumen of the Golgi in the yeast, *Saccharomyces cerevisiae*," *The Journal of Biological Chemistry*, vol. 272, no. 50, pp. 31908–31914, 1997.

[72] J. Engel, P. S. Schmalhorst, and F. H. Routier, "Biosynthesis of the fungal cell wall polysaccharide galactomannan requires intraluminal GDP-mannose," *The Journal of Biological Chemistry*, vol. 287, no. 53, pp. 44418–44424, 2012.

[73] L. Jackson-Hayes, T. W. Hill, D. M. Loprete et al., "GDP-mannose transporter paralogues play distinct roles in polarized growth of *Aspergillus nidulans*," *Mycologia*, vol. 102, no. 2, pp. 305–310, 2010.

[74] A. M. Al Obaid and A. R. Hashem, "Zinc tolerance and accumulation in *Aspergillus oryzae, Penicillium citrinum* and *Rhizopus stolonifer* isolated from Saudi Arabian soil," *Qatar University Science Journal*, vol. 17, no. 1, pp. 103–109, 1997.

[75] M. Salazar, W. Vongsangnak, G. Panagiotou, M. R. Andersen, and J. Nielsen, "Uncovering transcriptional regulation of glycerol metabolism in aspergilli through genome-wide gene expression data analysis," *Molecular Genetics and Genomics*, vol. 282, no. 6, pp. 571–586, 2009.

[76] D. N. Hebert, L. Lamriben, E. T. Powers, and J. W. Kelly, "The intrinsic and extrinsic effects of N-linked glycans on glycoproteostasis," *Nature Chemical Biology*, vol. 10, no. 11, pp. 902–910, 2014.

[77] R. E. Dempski Jr. and B. Imperiali, "Oligosaccharyl transferase: gatekeeper to the secretory pathway," *Current Opinion in Chemical Biology*, vol. 6, no. 6, pp. 844–850, 2002.

[78] A. Feizi, T. Österlund, D. Petranovic, S. Bordel, and J. Nielsen, "Genome-scale modeling of the protein secretory machinery in yeast," *PLoS ONE*, vol. 8, no. 5, Article ID e63284, 2013.

[79] T. H. Schmidt and C. Kandt, "LAMBADA and InflateGRO2: efficient membrane alignment and insertion of membrane proteins for molecular dynamics simulations," *Journal of Chemical Information and Modeling*, vol. 52, no. 10, pp. 2657–2669, 2012.

[80] C. Kandt, W. L. Ash, and D. Peter Tieleman, "Setting up and running molecular dynamics simulations of membrane proteins," *Methods*, vol. 41, no. 4, pp. 475–488, 2007.

[81] M. J. Buch-Pedersen, B. P. Pedersen, B. Veierskov, P. Nissen, and M. G. Palmgren, "Protons and how they are transported by proton pumps," *Pflügers Archiv*, vol. 457, no. 3, pp. 573–579, 2009.

[82] M. J. Buch-Pedersen, K. Venema, R. Serrano, and M. G. Palmgren, "Abolishment of proton pumping and accumulation in the E1P conformational state of a plant plasma membrane H^+-ATPase by substitution of a conserved aspartyl residue in transmembrane segment 6," *The Journal of Biological Chemistry*, vol. 275, no. 50, pp. 39167–39173, 2000.

[83] S. S. Gupta, N. D. DeWitt, K. E. Allen, and C. W. Slayman, "Evidence for a salt bridge between transmembrane segments 5 and 6 of the yeast plasma-membrane H^+-ATPase," *The Journal of Biological Chemistry*, vol. 273, no. 51, pp. 34328–34334, 1998.