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

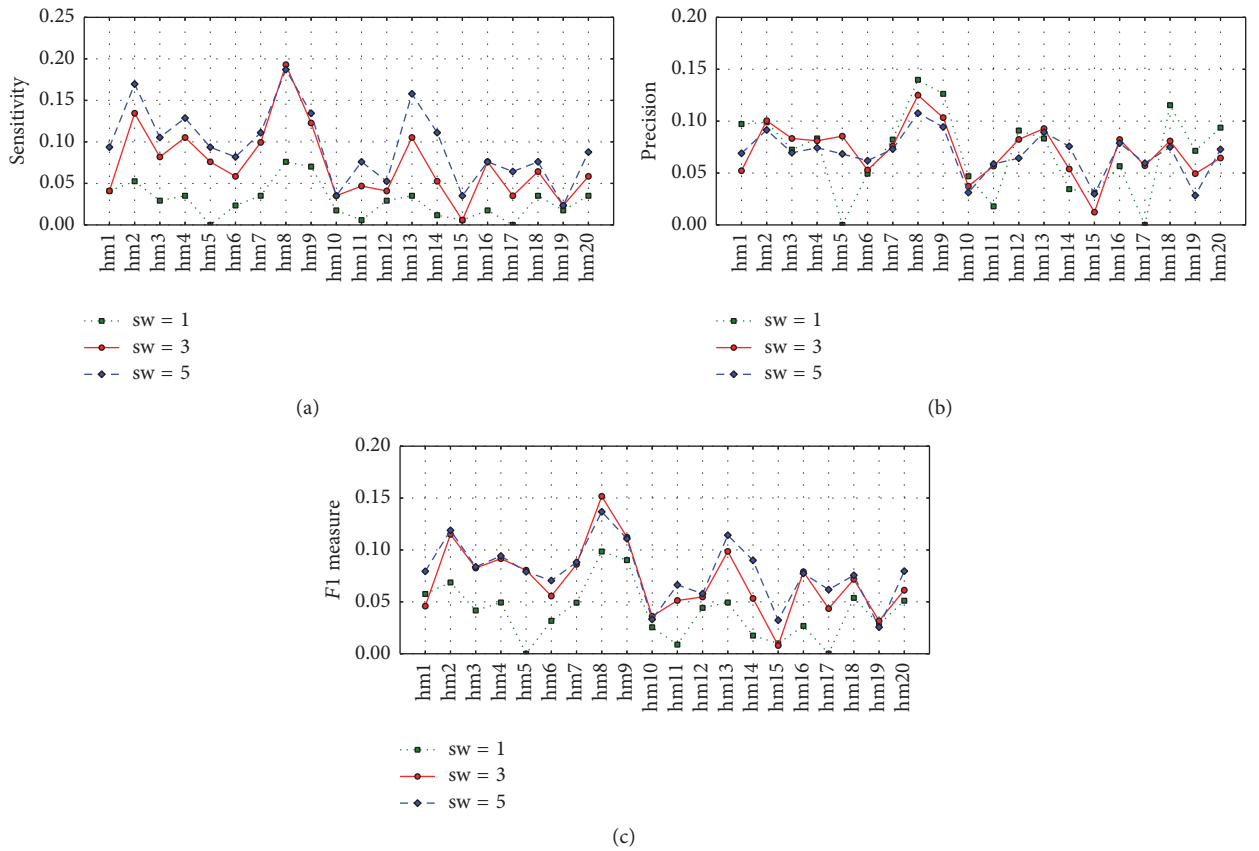

Figure 1: Plots of sensitivity (a), precision (b), and $F1$ values by the single ith highest mode ($i = 1, 2, \ldots, 20$) in three cases of the sliding window sizes (sw) (i.e., sw = 1, 3, 5) for GNB classifiers. The ith highest mode in the figure is denoted as hm$i$.

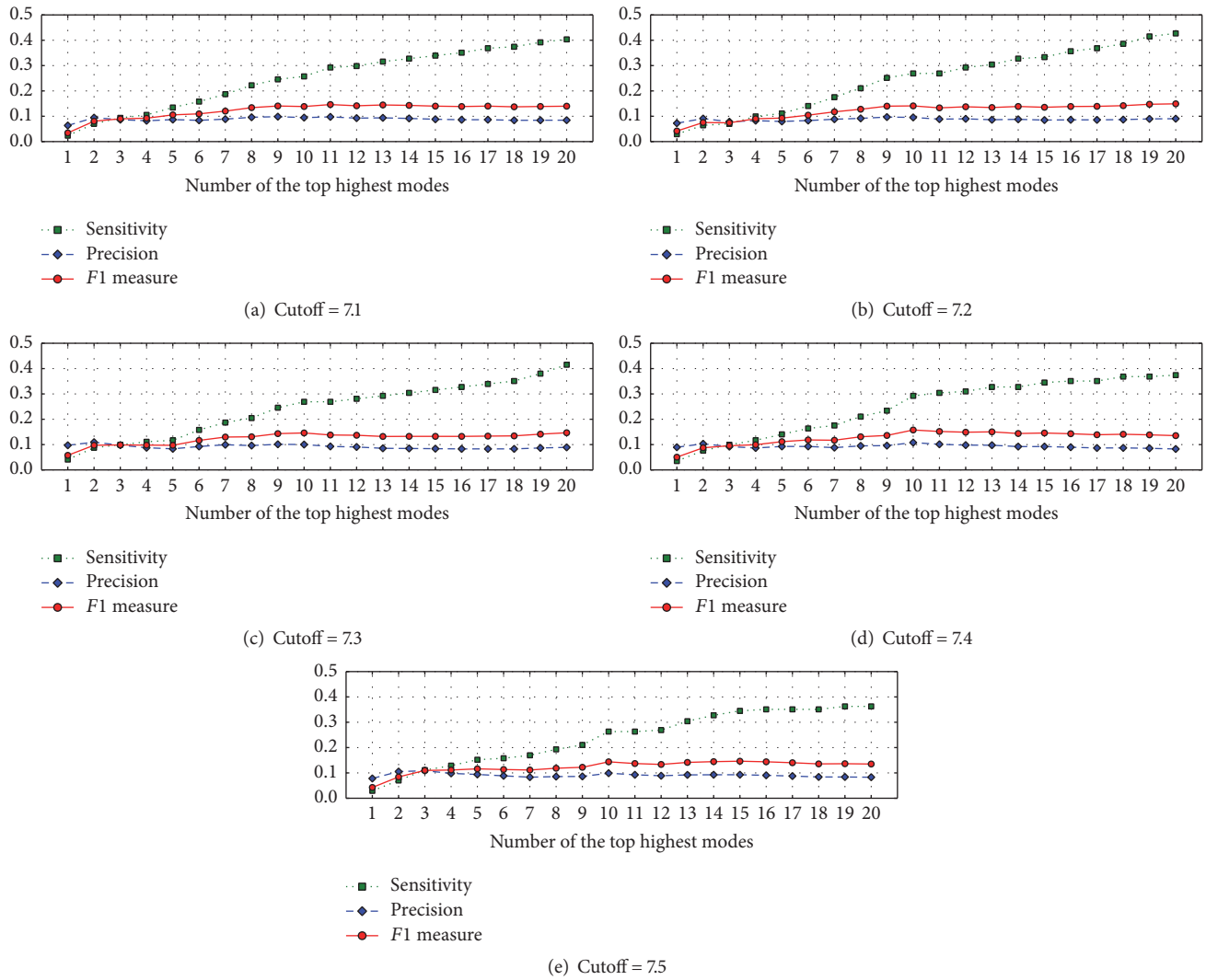

FIGURE 2: Plots of sensitivity, precision, and $F1$ values against m modes with the highest frequency in five cases denoted by the distance cutoffs of 7.1 Å (a), 7.2 Å (b), 7.3 Å (c), 7.4 Å (d), and 7.5 Å (e), respectively, for GNB classifiers.

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.