Journal of Molecular Graphics and Modelling 31 (2011) 28–34

Contents lists available at ScienceDirect

Journal of Molecular Graphics and Modelling

journal homepage: www.elsevier.com/locate/JMGM

A novel method for quantitatively predicting non-covalent interactions from protein and nucleic acid sequence

Jiansheng Wu[a], Dong Hu[a], Xin Xu[a], Yan Ding[b], Shancheng Yan[a], Xiao Sun[b,*]

[a] School of Geography and Biological Information, Nanjing University of Posts and Telecommunications, Nanjing 210046, PR China
[b] State Key Laboratory of Bioelectronics, School of Biological Science and Medical Engineering, Southeast University, Nanjing 210096, PR China

ARTICLE INFO

Article history:
Received 27 June 2011
Received in revised form 5 August 2011
Accepted 5 August 2011
Available online 11 August 2011

Keywords:
Non-covalent interactions
Support vector machine regression models
Conjoint triad
H-VDW

ABSTRACT

Biochemical interactions between proteins and biological macromolecules are dominated by non-covalent interactions. A novel method is presented for quantitatively predicting the number of two most dominant non-covalent interactions, i.e., hydrogen bonds and van der Waals contacts, potentially forming in a hypothetical protein–nucleic acid complex from sequences using support vector machine regression models in conjunction with a hybrid feature. The hybrid feature consists of the sequence-length fraction information, conjoint triad for protein sequences and the gapped dinucleotide composition. The SVR-based models achieved excellent performance. The polarity of amino acids was also found to play a vital role in the formation of hydrogen bonds and van der Waals contacts. We have constructed a web server H-VDW (http://www.cbi.seu.edu.cn/H-VDW/H-VDW.htm) for public access to the SVR models.

1. Introduction

Biochemical interactions between a protein and a drug, or a catalyst and its substrate, and even protein–protein and protein–nucleic acid reactions, are dominated by non-covalent interactions. The class of interaction covers a wide range of binding energies, and contains hydrogen bonding, dipole–dipole interactions, steric repulsion [1]. Molecular structure is governed by covalent bonds, non-covalent bonds, and electrostatic interactions, the latter two of which are the driving force in most biochemical processes [2]. Covalent bonds can be defined in three-dimensional molecular structures, however, non-covalent interactions are hidden within voids in the bonding network [2]. An approach was developed to map and analyze non-covalent interactions, requiring only knowledge of the atomic coordinates, based on the electron density and its derivatives, and the approach is efficient and applicable to large systems, such as proteins or DNA [2,3]. However, high-resolution three-dimensional structures are only available for a small proportion of the known biological macromolecule complexes. Moreover, the process to solve their structure through experimental methods is greatly expensive and time-consuming. Therefore, a method to predict the number of non-covalent interactions potentially occurring in a putative biological macromolecule complex only requiring sequence information would be beneficial to aid understanding of the complex interactions between biological macromolecules.

In the present work, we aim to design an optimal predictor for quantitatively modeling the number of two most dominant non-covalent interactions, i.e., hydrogen bonds (H-bonds) and van der Waals contacts, potentially forming in a hypothetical protein–nucleic acid complex from sequences. In designing models, the following points are taken into consideration. (1) Predicting new data instances based only on sequence information without any structure information is an ideal approach and much more universal, but till now no methods have been proposed for this task only using sequence information. (2) All available data sets of protein–nucleic acid complex structures are employed to build predictors in order to avoid the over-fitting of training data and poor generalization performance for new data. (3) The support vector machine regression (SVR) is a powerful machine-learning algorithm developed from statistical learning theory [4] that has been successfully applied in many fields for data regression. (4) To our knowledge, incorporating effective features is the most important way of achieving excellent performance of predictors. (5) To construct a web server for free access of the models for academic usage.

For the above purposes, we propose a novel method to quantitatively estimate the number of H-bonds and van der Waals contacts from protein and nucleic acid sequences using a SVR model in conjunction with a hybrid feature. The hybrid feature is composed of

Abbreviations: H-bonds, hydrogen bonds; SVR, support vector machine regression; PDB, Protein Data Bank; *RMSE*, root mean square error; *r*, Pearson correlation coefficient; RBF, radial basis function; *DevAp*, the RMS error between the predicted and observed values.

* Corresponding author. Tel.: +86 25 83795174.
E-mail address: xsun@seu.edu.cn (X. Sun).

1093-3263/$ – see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmgm.2011.08.001

the sequence-length fractions, protein sequence attributes named conjoint triad and nucleic acid sequence attributes called the gapped dinucleotide composition. The protein sequence attributes consider the properties of one amino acid and its vicinal amino acids, and also two physicochemical properties (i.e. dipoles and side chain volumes) of the 20 amino acids. The nucleic acid sequence attributes present the frequency of gapped dinucleotides composed of pyrimidine and purine nucleotides. The predictors present good performance, for example, the SVR-based model achieved a Pearson correlation coefficient (r) of 0.660 ± 0.056 for modeling the number of H-bonds formed in protein–DNA complexes. This result showed that the prediction of the regression model closely matches the observed values. We have developed a web server H-VDW (http://www.cbi.seu.edu.cn/H-VDW/H-VDW.htm) for public use of the SVR predictors.

2. Materials and methods

2.1. Datasets

All 1615 protein–DNA and 522 protein–RNA complexes collected from the PDB [5] (release date 2011.04.23) were determined by X-ray crystallography with a resolution <3.0 Å. To evaluate the SVR predictors, two separate test datasets of protein–DNA complexes (PDNA-70) and protein–RNA complexes (PRNA-21) were created by randomly selecting 70 protein–DNA and 21 protein–RNA complex structures, and others as the training datasets.

Hydrogen bonds (H-bonds) and van der Waals contacts in the complexes were calculated using HBPLUS [6]. The H-bonds (i.e. D–H···A) were identified by finding all proximal donor (D) and acceptor (A) atom pairs that satisfy specified geometrical criteria for H-bond formation. The criteria used were: the H–A distance was <2.7 Å, the D–A distance was <3.35 Å, the D–H–A angle was >90° and the H–A–AA angle was >90°, where AA is the atom attached to the acceptor. NUCPLOT [7] used the list of H-bonds generated by to plot all H-bonds between the protein and nucleic acid, between water and nucleic acid, and between protein and nucleic acid via a bridging water molecule. All atoms not involved in H-bonds but separated by <3.9 Å were considered to be interacting through van der Waals contacts [7]. NUCPLOT is an automatic program which can be used for any protein–DNA complex and only for certain protein–RNA structures [7].

In a complex, there may be several amino acid sequences and nucleic acid sequences, such as the complex 1A0A contains two amino acid sequences (A chain and B chain) and two DNA sequences (C chain and D chain). Therefore, four pairs (AB, AC, BC and BD) can be reached for 1A0A as the candidate training samples (each pair consists of one amino acid sequence and one nucleic acid sequence). Meanwhile, it is observed that there are some amino acid sequences or nucleic acid sequences with the same primary structure as each other, such as in the complex 1A0A, the A chain is the same as the B chain, and the C chain and the D chain are exactly identical. In this study, a non-redundant pair was randomly picked from the same candidate training samples. Finally, the protein–DNA training dataset contained 1533 non-redundant pairs, and the protein–RNA one had 258 non-redundant pairs. The number of H-bonds (including H-bonds via bridging water molecules) and van der Waals contacts in each protein–nucleic acid pair were calculated, and the normalized number of H-bonds (θ_h) and van der Waals contacts (θ_v) in each pair were determined by:

$$\theta_h = \frac{N_h}{l_p + l_n}; \quad \theta_v = \frac{N_v}{l_p + l_n} \tag{1}$$

where, N_h and N_v are the number of H-bonds and van der Waals contacts in each pair. l_p and l_n are the length of the protein and nucleic acid chains, respectively. At the first step, we predicted the normalized number (θ_h and θ_v) because it achieved better performance than the prediction of N_h and N_v. At the second step, we recovered the number of H-bonds and van der Waals contacts by:

$$N_h = \text{round}(\theta_h(l_p + l_n)); \quad N_v = \text{round}(\theta_v(l_p + l_n)) \tag{2}$$

Table 1
Classification of amino acids.

No.	Dipole scale[a]	Volume scale[b]	Class
a	–	–	Ala, Gly, Val
a	–	–	Ala, Gly, Val
b	–	+	Ile, Leu, Phe, Pro
c	+	+	Tyr, Met, Thr, Ser, Cys
d	++	+	His, Asn, Gln, Tpr
e	+++	+	Arg, Lys
f	+′+′+′	+	Asp, Glu

[a] Dipole scale (Debye): –, dipole < 1.0; +, 1.0 < dipole < 2.0; ++, 2.0 < dipole < 3.0; +++, dipole > 3.0; +′+′+′, dipole > 3.0 with opposite orientation.
[b] Volume scale (Å^3): –, volume < 50; +, volume > 50.

2.2. Support vector machine regression

The basic idea of support vector machine regression (SVR) was formulated by Vapnik [4]. The SVR algorithm was implemented by the e1071 (version 1.5-16) R package [8]. In ensure that model generation of SVRs are completely independent of the test data, 1/5 of the original samples are randomly selected as the test data for assessment, and the remaining 4/5 of the samples are for model construction where ten-fold cross-validation are used and the parameters C and γ for the RBF kernel are optimized by the standard grid search. The process will be repeated ten times, and the mean and the standard deviation of the prediction results will be reached. However, the SVR models used by the H-VDW server were constructed using all samples in the training datasets.

2.3. Features

Electrostatic (including H-bonding) and hydrophobic interactions usually dominate in protein–DNA complexes, and are reflected by the dipoles and volumes of the side chains of amino acids, respectively [9]. Based on the dipoles and volumes of the side chains, the 20 amino acids can be clustered into seven classes [10]. However, cysteine which represents the seventh class was grouped into the third class in this study. This was because disulfide bonds have no special role in protein–nucleic acid interactions. Consequently, the 20 amino acids were grouped into six classes: Class *a*: Ala, Gly, Val; Class *b*: Ile, Leu, Phe, Pro; Class *c*: Tyr, Met, Thr, Ser, Cys; Class *d*: His, Asn, Gln, Tpr; Class *e*: Arg, Lys; and Class *f*: Asp, Glu [9]. The details are listed in Table 1. In protein–DNA complexes, the pyrimidine–purine dinucleotide steps facilitate the movement of the DNA chain, and such steps may act as long-range signals for protein binding [11]. Nucleic acids were clustered into two classes: purine (p) and pyrimidine (m). Using Shen's method [10], a descriptor named conjoint triad was proposed to represent the protein chain section, whereas another descriptor called gapped dinucleotide composition was used for the nucleic acid chain section. The process of generating descriptors is described as follows.

(1) Conjoint triad

A vector V was used for the protein sequence and each element (v_i) represented a triad type. F_v is the frequency vector and the ith value of F_v (f_{vi}) is the frequency of type v_i within a protein sequence. For the 20 amino acids that were grouped into six classes, the size of F_v was 216 ($6 \times 6 \times 6$). The detailed definition and description for

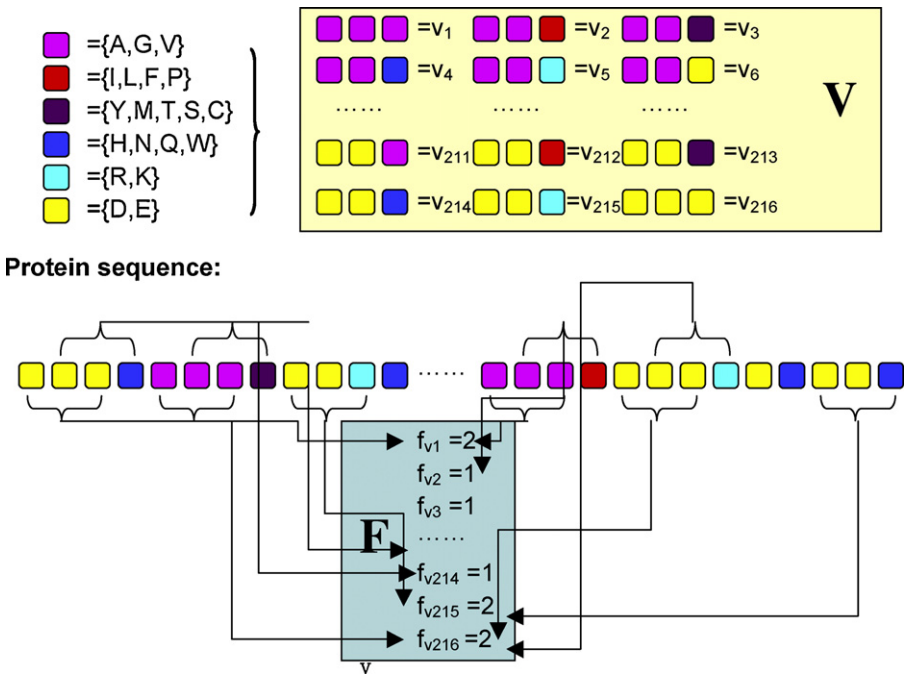


Fig. 1. Schematic diagram for constructing the vector space (V, F_v) of protein sequence. V is the vector space of the sequence features; each feature (v_i) represents a triad composed of three consecutive amino acids; F_v is the frequency vector corresponding to V, and the value of the ith dimension of $F(f_{vi})$ is the frequency that v_i triad appeared in the protein sequence.

(V, F_v) are described in Fig. 1. The elements of F_v were scaled to fall within the range 0–1 by:

$$f_{vi} = \frac{f_{vi} - \min\{f_{v1}, f_{v2}, \ldots, f_{v216}\}}{\max\{f_{v1}, f_{v2}, \ldots, f_{v216}\}} \quad (3)$$

(2) Gapped dinucleotide composition

The gapped dinucleotide means two nucleotides with intervening bases within a sequence [12]. Here, O_k^j is defined as the observed total number of jth two nucleotides with k intervening bases appearing in the nucleic acid sequence [12], and $k=0$ and 1 were used in this paper. Because nucleic acids were clustered into two classes (purines and pyrimidines), $j=1$–4 (2×2). Vectors A and B were used for the nucleic acid sequence. Each element a_j of vector A represents a dinucleotide type without an intervening base ($k=0$) and b_j of vector B means a gapped dinucleotide type with one intervening base ($k=1$). F_A and F_B are the frequency vectors. The jth element of F_A (F_{aj}) and F_B (F_{bj}) are the frequencies of a_j and b_j appearing in a sequence, respectively. As nucleic acids were grouped into two classes, the size of both F_A and F_B are four (2×2). Therefore, an 8-dimensional vector descriptor (4+4) is proposed for each nucleic acid sequence. Fig. 2 presents the details of the definition and description for (A, F_A) and (B, F_B). Each element x_j of vector F_A or F_B was normalized to fall within the range from 0 to 1 by:

$$x_j = \frac{x_j - \min\{x_1, x_2, x_3, x_4\}}{max\{x_1, x_2, x_3, x_4\}} \quad (4)$$

(3) The sequence-length fraction of protein sequences (β) and nucleic acid sequences (η) were defined as:

$$\beta = \frac{l_p}{l_p + l_n}, \quad \eta = \frac{l_n}{l_p + l_n} \quad (5)$$

where l_p is the length of the protein sequences and l_n is the length of the nucleic acid sequences.

In this study, the input vector for each data instance contained 226 feature values, including sequence-length fraction values (β and η, two dimensions), protein sequence descriptor vectors (216 dimensions) and nucleic acid sequence descriptor vectors (eight dimensions).

2.4. Measurement of model performance

The root mean square errors (*RMSE*), the *RMS* error between the predicted and observed values ($DevA_p$) and Pearson correlation coefficient (r) for assessment of the regression models are defined as [13]:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\left|P_i - O_i\right|\right)^2} \quad (6)$$

$$DevA_p = \frac{\sqrt{\sum_{i=1}^{n}(O_i - P_i)^2}}{\sqrt{\sum_{i=1}^{n}(O_i - \bar{O})^2}} \quad (7)$$

$$r = \frac{\sum_{i=1}^{n}(P_i - \bar{P})(O_i - \bar{O})}{\sqrt{\sum_{i=1}^{n}(P_i - \bar{P})^2}\sqrt{\sum_{i=1}^{n}(O_i - \bar{O})^2}} \quad (8)$$

where n is the number of data instances, and P and O are the predicted and observed values for H-bonds or van der Waals contacts, respectively. $\bar{P}$ and $\bar{O}$ are the mean of P and O, respectively.

2.5. Interaction propensity

To analyze the relative importance of different interaction patterns for formation of H-bonds or van der Waals contacts, the interaction propensity (P) was defined for each of the six classes of amino acids (a, b, c, d, e, f) binding each of the two classes of nucleotides (m, p). The surface of protein monomers participates in interactions with nucleic acid molecules. In this study, DSSP

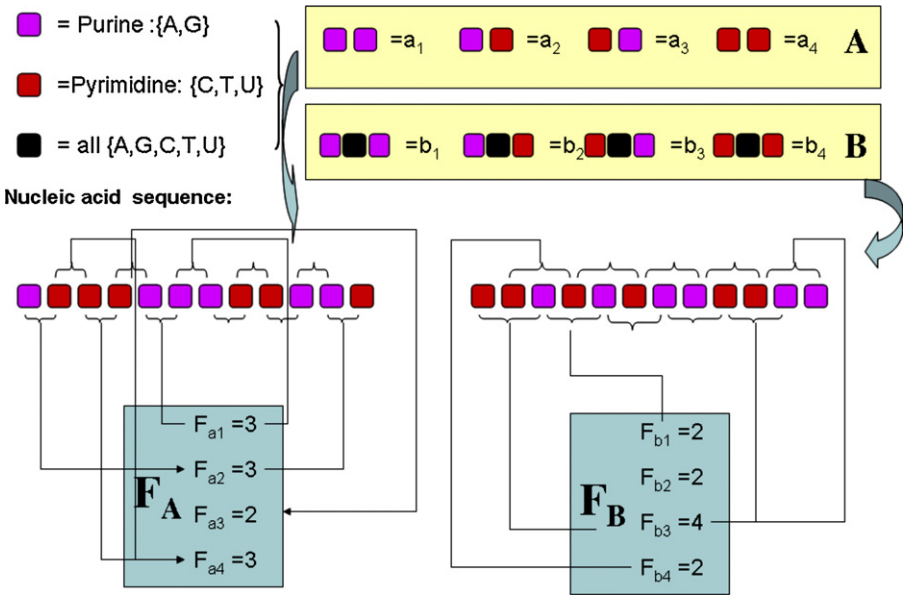


Fig. 2. Schematic diagram for constructing the vector space (A, F_A) and (B, F_B) of nucleic acid sequences. A is the vector space of the sequence features; each feature (a_i) represents a dinucleotide composed of purine and pyrimidine without intervening base; F_A is the frequency vector corresponding to A, and the value of the ith dimension of F_A(F_{ai}) is the frequency that a_i dinucleotide appeared in the nucleic acid sequence. On the other hand, B is another vector space of the sequence features; each feature (b_i) describes a gapped dinucleotide composed of purine and pyrimidine with one intervening base; F_B is the frequency vector corresponding to B, and the value of the ith dimension of F_B(F_{bi}) is the frequency that b_i diad appeared in the nucleic acid sequence.

program [14] is used to calculate the accessible surface area (ASA) values of each residue of protein monomers in complex structures, and a residue with the ASA value greater than the 10% of its surface area is considered as exposed, otherwise as buried [15]. The interaction propensity value P_{ij} between the ith class of amino acids and the jth class of nucleotides was calculated by [16]:

$$P_{ij} = \frac{B_{ij}/\sum B_{ij}}{(N_i/N_x)(N_j/N_y)} \quad (9)$$

Here, B_{ij} is the size of pairs of the ith class amino acid binding the jth class nucleotide, and $\sum B_{ij}$ is the total size of all binding pair, and N_i is the number of the ith class amino acid on the surface of the protein monomers, N_x is the total number of all amino acids on the surface of the protein monomers, and N_j is the size of the jth class nucleotide, and N_y is the sum of all nucleotides.

3. Results and discussion

3.1. Performance of the SVR-based methods

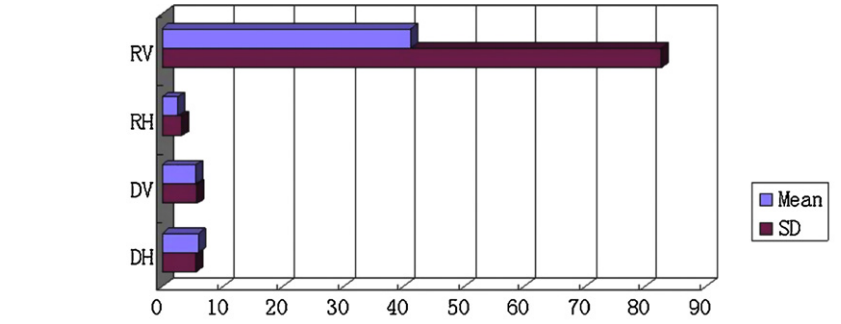
Table 2 shows the performance of the SVR-based models. The results have been obtained using the training parameters, $C = 2$, $\gamma = 0.0078125$ for the protein–DNA training data, and $C = 16$, $\gamma = 0.0078125$ for the protein–RNA training data, which give better results than other values for prediction. For example, on average, the predictor for modeling the number of H-bonds formed in protein–DNA complexes reached an r of 0.660 ± 0.056 (Table 2), indicating that the prediction of the regression model closely matches the observed values. As shown in Table 2, a satisfied r of 0.623 was achieved by the predictor for modeling the number of van der Waals contacts formed in protein–RNA complexes, but its *RMSE* value is much larger than that of other predictors. It is mainly caused by van der Waals contacts in protein–RNA complexes occur much more frequently than others interactions (Fig. 3). We estimated and ranked the relative importance of features used in the SVR prediction models (Table 3). This was evaluated by the relative difference (*RD*) of the r (*RD*-r) values when the feature was included versus when the feature was excluded in the construction of SVR prediction models both following the same strategy indicated in Section 2.2. The *RD*-r values of all 226 features were calculated and the top fifteen features are listed in Table 3. For example, *eba*, *dcf*, *daf*, *fea*, *add*, *bcb*, *fae*, *afb*, *cac*, *ccd*, *ebe*, *abd*, *cfd*, *daa* and *mp1* are the top fifteen most important features in modeling the number of H-bonds in protein–RNA complexes.

3.2. Interaction propensities

The interaction propensities of H-bonds and van der Waals contacts between each of the classes of amino acids and each of the classes of nucleic acids were calculated for all data instances in the training datasets. This included 1533 non-redundant pairs in protein–DNA complexes and 258 pairs in protein–RNA complexes. Each of these pairs consisted of one amino acid sequence and one nucleic acid sequence. Table 4 shows the interaction propensity values. The propensity value represents the frequency of the co-occurrences of each of the classes of amino acids and each of

Table 2
The performance of the SVR models in quantificational modeling of the number of H-bonds or van der Waals contacts.

Predictors	r (mean ± SD)	*RMSE* (mean ± SD)	$DevA_p$ (mean ± SD)
Protein–DNA complexes ($C = 2$, $\gamma = 0.0078125$)			
Hydrogen bonds	0.660 ± 0.056	4.354 ± 0.329	0.755 ± 0.052
Van der Waals contacts	0.573 ± 0.063	4.792 ± 0.288	0.832 ± 0.055
Protein–RNA complexes ($C = 16$, $\gamma = 0.0078125$)			
Hydrogen bonds	0.466 ± 0.077	3.611 ± 0.467	1.100 ± 0.102
Van der Waals contacts	0.623 ± 0.113	64.28 ± 23.19	0.841 ± 0.114



	DH	DV	RH	RV
Mean	6.054794521	5.693411611	2.57751938	41.24418605
SD	5.626157975	5.872664441	3.311899083	82.85089547

Fig. 3. The mean and standard deviation (SD) of the number of H-bonds and van der Waals contacts in protein–nucleic acid complexes.

Table 3
Estimation and ranking of the relative importance of features used in the SVR prediction models.

Rank (importance)	DH		DV		RH		RV[a]	
	Features[b]	*RD-r* [c]	Features[b]	*RD-r*[c]	Features[b]	*RD-r*[c]	Features[b]	*RD-r*[c]
1	*eef*	0.0394	*ace*	0.1330	*eba*	0.5541	*ffc*	0.2314
2	*cdb*	0.0233	*eda*	0.0960	*dcf*	0.5286	*bdb*	0.2198
3	*pm*	0.0231	*abd*	0.0889	*daf*	0.5057	*cbb*	0.1950
4	*faf*	0.0225	*ccc*	0.0864	*fea*	0.5023	*β*	0.1878
5	*ddc*	0.0195	*edd*	0.0831	*add*	0.4984	*fdb*	0.1808
6	*bde*	0.0092	*fba*	0.0805	*bcb*	0.4933	*ddd*	0.1800
7	*cce*	0.0088	*bfd*	0.0795	*fae*	0.4761	*ffe*	0.1791
8	*cae*	0.0062	*fcc*	0.0781	*afb*	0.4732	*η*	0.1566
9	*cfd*	0.0036	*ece*	0.0779	*cac*	0.4686	*bae*	0.1490
10	*ccf*	0.0014	*pp*	0.0767	*ccd*	0.4540	*eda*	0.1482
11	*abf*	0.0014	*afa*	0.0764	*ebe*	0.4516	*bfd*	0.1467
12	*efa*	−0.0004	*edf*	0.0757	*abd*	0.4480	*cbe*	0.1399
13	*aef*	−0.0015	*cbb*	0.0752	*cfd*	0.4463	*adb*	0.1394
14	*bce*	−0.0017	*fad*	0.0734	*daa*	0.4437	*fde*	0.1354
15	*bac*	−0.0020	*cce*	0.0723	*mp1*	0.4303	*edb*	0.1274

[a] DH: prediction for H-bonds in protein–DNA complexes; DV: prediction for van der Waals contacts in protein–DNA complexes; RH: prediction for H-bonds in protein–RNA complexes; RV: prediction for van der Waals contacts in protein–RNA complexes.

[b] *aaa*, *aab*, *aac*, *aad*, *aae*, *aaf*, *aba*, . . ., *fef*, *ffa*, *ffb*, *ffc*, *ffd*, *ffe*, *fff:* the scaled frequency values of all 216 triad types for protein sequences; *mm*, *mp*, *pm*, *pp:* the scaled frequency values of all four dinucleotide types for nucleic acid sequences without intervening bases; *mm1*, *mp1*, *pm1*, *pp1*: the scaled frequency values of all four gapped dinucleotide types for nucleic acid sequences with one intervening base; *η*, *β*: the sequence-length fractions of nucleic acid sequences and protein sequences.

[c] *RD-r*: The relative importance of a feature was estimated by the relative difference of the Pearson correlation coefficient (*r*) values when the feature was included versus when the feature was excluded in the SVR prediction models.

Table 4
Interaction propensities of H-bonds and van der Waals contacts between the amino acid and the nucleic acid classes in protein–nucleic acid complexes.

Classes of nucleic acids	Classes of amino acids						Average
	a	*b*	*c*	*d*	*e*	*f*	
Hydrogen bonds							
Protein–DNA complexes							
m	0.447	0.183	1.320	1.342	2.271	0.456	1.003
p	0.404	0.220	1.457	1.642	3.133	0.542	1.233
Average	0.426	0.202	1.389	1.492	2.702	0.499	
Protein–RNA complexes							
m	0.361	0.276	1.188	1.973	1.624	1.033	1.076
p	0.420	0.223	2.068	1.811	1.363	1.312	1.120
Average	0.391	0.250	1.628	1.892	1.494	1.173	
van der Waals contacts							
Protein–DNA complexes							
m	0.719	0.578	1.209	1.523	1.883	0.425	1.056
p	0.545	0.523	1.043	1.657	2.276	0.563	1.101
Average	0.632	0.551	1.126	1.590	2.080	0.494	
Protein–RNA complexes							
m	0.560	0.465	0.991	1.276	2.563	0.532	1.065
p	0.545	0.419	0.934	1.599	2.550	0.424	1.079
Average	0.553	0.442	0.963	1.438	2.557	0.478	

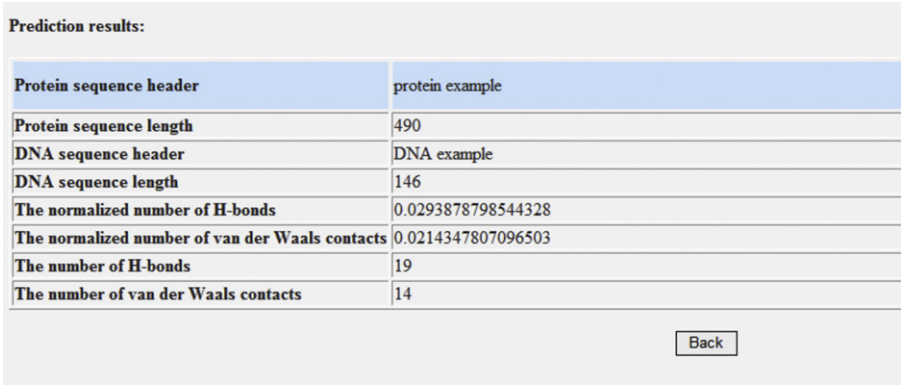


Fig. 4. An output example of the H-VDW server.

the classes of nucleic acids in protein–nucleic acid complexes. A propensity value ≥1 indicates that the combination of a given class of amino acids and a certain class of nucleic acids is a favored interaction pattern for the formation of H-bonds or van der Waals contacts. A propensity value <1 indicates that a given combination is unpopular [17]. The higher the propensity value the more frequently a given combination occurred. The classes of amino acids showed diverse averages of propensity values (ranging from 0.220 to 2.702), whereas there were minor differences in the averages of propensity values for the classes of nucleic acids (ranging from 1.003 to 1.233) (Table 4).

Hydrogen bonds, van der Waals contacts, electrostatic and hydrophobic interactions dominate protein–nucleic acid interactions, and can be reflected by the dipoles and volumes of the side chains of amino acids [10]. For a molecule, the dipole can be characterized by its dipole moment. The dipole moment is a vector quantity, where the positive charge center points towards the negative charge center of a molecule. The dipole moment of a molecule can be used to measure the size of its polarity. As indicated in Table 1, the rank of dipole moments of the six classes of amino acids is: $(a, b) < c < d < (e, f)$ [10]. The increase of the dipole moments of the six classes of amino acids correlated with rising interaction propensity values of H-bonds and van der Waals contacts between classes of amino acids and classes of nucleic acids (Table 4). The results suggest that the polarity of amino acids probably plays a key role in protein–nucleic acid recognition. The results also demonstrated that our SVR models were rational in the effective predication of the number of H-bonds and van der Waals contacts. At the same time we also noted that: (1) Classes *e* and *f* were presented on the same dipole scale (Table 1), but Class *e* had a much higher interaction propensity value than Class *f*. The dipole moments of Class *f* are in the opposite orientation relative to those of Class *e*, indicating that there may be electrostatic exclusive forces between amino acid molecules in Class *f* and nucleic acid molecules. (2) Classes *a* and *b* were also at the same dipole stage (Table 1), whereas the propensities of Class *b* were clearly observed to be lower than those of Class *a*. Volumes of the side chains of amino acids within Class *b* (>50 Å^3) are larger than these of amino acids within Class *a* (<50 Å^3) (Table 1). The results suggest that there are probably steric hindrances between the side chains of amino acids within Class *b* and nucleic acid molecules. Such steric hindrances appear to influence residues–nucleic acid interactions.

3.3. Assessment on independent test sets

The two separate test datasets of protein–DNA complexes (PDNA-70) and protein–RNA complexes (PRNA-21) were processed in the same way as the training dataset by retaining only the non-redundant pairs in complexes. The putative numbers of H-bonds or van der Waals contacts in the PDNA-70 and PRNA-21 datasets were predicted using the H-VDW server. As shown in Table 5, the predictors for protein–RNA complexes (PRNA-21) achieved a better performance than for the protein–DNA complexes (PDNA-70). For example, the predictor achieved an *r* of 0.567 with an *RMSE* of 3.273 and a $DevA_p$ of 1.157 for modeling the number of H-bonds in the PRNA-21 dataset (Table 5). The results suggest that the H-VDW server is reliable in predicting the number of H-bonds and van der Waals contacts formed in protein–nucleic acid complexes.

Table 5
Prediction results on independent test sets.

Test set	*r*	*RMSE*	$DevA_p$
PDNA-70			
Hydrogen bonds	0.563	5.483	1.167
van der Waals contacts	0.479	10.06	1.183
PRNA-21			
Hydrogen bonds	0.567	3.273	1.157
van der Waals contacts	0.580	15.29	1.043

3.4. Web server

H-VDW is available at http://www.cbi.seu.edu.cn/H-VDW/H-VDW.htm. All the CGI scripts of models were written in perl 5.8.4 and the interface was designed using HTML. The SVR algorithm was implemented by the e1071 (version 1.5-16) R package [8]. On the web page, users can choose the complex type to be predicted (either protein–DNA or RNA complexes), and subsequently enter the protein and nucleic acid sequences in FASTA format. H-VDW only allows the prediction for 10 protein and nucleic acid sequences in one round of prediction. In addition, if the DNA or RNA-binding domains are known, users may input the domain sequences rather than the full-length sequences.

For *m* pieces of amino acid sequences and *n* pieces of nucleic acid sequences, $m \times n$ protein–nucleic acid sequence pairs will be implemented to predict the number of H-bonds and van der Waals contacts. As presented in Fig. 4, the prediction returned from H-VDW server includes the following sections: the header and length of protein sequences, the header and length of nucleic acid sequences, the normalized number of H-bonds (θ_h) and van der Waals contacts (θ_v), and the number of H-bonds and van der Waals contacts.

4. Conclusions

A novel method using support vector machine regression (SVR) models in conjunction with a hybrid feature for the quantitative prediction of the number for H-bond or van der Waals contacts in

protein–nucleic acid complexes is presented. The hybrid feature is composed of the sequence-length fraction information, the amino acid sequence attributes termed conjoint triad and the nucleic acid sequence attributes called gapped dinucleotide composition. The SVR-based models reach an excellent performance level. For example, a Pearson correlation coefficient (r) of 0.660 ± 0.056 is reached for modeling the number of H-bonds in protein–DNA complexes. This suggests that the prediction of the regression model closely matches the observed values. The analysis of interaction propensities suggests that amino acid polarity probably plays a key role in the protein–nucleic acid recognition. The analysis also demonstrates that our SVR models are rational in the effective predication of the number of H-bonds and van der Waals contacts. A web server H-VDW (http://www.cbi.seu.edu.cn/H-VDW/H-VDW.htm) has been constructed for public access of the SVR models. Numerous protein–nucleic acid interactions are well characterized. Data from such studies show that nucleic acid molecules interact with particular protein domain architectures. In these cases where the nucleic acid-binding domains are known, users require only to input the domain sequences rather than the full protein sequences to the H-VDW server. More importantly, it is a universal method which can be easily extended to quantitatively predict the number of non-covalent interactions occurring in other biological macromolecule complexes, such as protein–protein interactions.

Conflict of interest statement

None declared.

Acknowledgements

This work is supported by the Research Start-up Funding by Nanjing University of Posts and Telecommunications (Nos. NY209027 and NY210083) and the National Natural Science Foundation of China (No. 61073141).

References

[1] P.A. Kollman, Noncovalent interactions, Chem. Rev. 10 (1977) 365–371.
[2] E.R. Johnson, S. Keinan, P.M. Nchez, J.C. Garca, A.J. Cohen, W.T. Yang, Revealing noncovalent interactions, J. Am. Chem. Soc. 132 (2010) 6498–6506.
[3] J. Contreras-García, E.R. Johnson, S. Keinan, R. Chaudret, J.P. Piquemal, D.N. Beratan, W. Yang, NCIPLOT: a program for plotting non-covalent interaction regions, Chem. Theory Comput. 7 (2011) 625–632.
[4] V.N. Vapnik, Statistical Learning Theory, Wiley, New York, 1998.
[5] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, The Protein Data Bank, Nucleic Acids Res. 28 (2000) 235–242.
[6] I.K. McDonald, J.M. Thornton, Satisfying hydrogen bonding potential in proteins, J. Mol. Biol. 238 (1994) 777–793.
[7] N.M. Luscombe, R.A. Laskowski, J.M. Thornton, NUCPLOT: a program to generate schematic diagrams of protein–nucleic acid interactions, Nucleic Acids Res. 25 (1997) 4940–4945.
[8] E. Dimitriadou, K. Hornik, F. Leisch, D. Meyer, A. Weingessel, e1071: Misc functions of the department of statistics, TU Wien, R Package, Version 1.5-16, 2007, e1071. Available from http://cran.R-project.org.
[9] J. Wu, H. Liu, X. Duan, Y. Ding, H. Wu, Y. Bai, X. Sun, Prediction of DNA-binding residues in proteins from amino acid sequences using a random forest model with a hybrid feature, Bioinformatics 25 (2009) 30–35.
[10] J. Shen, J. Zhang, X. Luo, W. Zhu, K. Yu, K. Chen, Y. Li, H. Jiang, Predicting protein–protein interactions based only on sequences information, Proc. Natl. Acad. Sci. U. S. A. 104 (2007) 4337–4341.
[11] S.A. Coulocheri, D.G. Pigis, K.A. Papavassiliou, A.G. Papavassiliou, Hydrogen bonds in protein–DNA complexes: where geometry meets plasticity, Biochimie 89 (2007) 1291–1303.
[12] P. Jiang, H. Wu, J. Wei, F. Sang, X. Sun, Z. Lu, RF-DYMHC: detecting the yeast meiotic recombination hotspots and coldspots by random forest model using gapped dinucleotide composition features, Nucleic Acids Res. 35 (2007) W47–W51.
[13] J. Song, K. Burrage, Predicting residue-wise contact orders in proteins by support vector regression, BMC Bioinform. 7 (2006) 425.
[14] W. Kabsch, C. Sander, Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features, Biopolymers 22 (1983) 2577–2637.
[15] H. Zhou, Y. Shan, Prediction of protein interaction sites from sequence profile and residue neighbor list, Proteins 44 (2001) 336–343.
[16] J. Wu, S. Yan, L. Tang, D. Hu, Computational analysis of propensities of amino acids and nucleotides usage at protein–nucleic acid interfaces, in: 2011 International Conference on Information Science and Technology, March 26–28, 2011, Nanjing, Jiangsu, China, 2011, pp. 1342–1349.
[17] H. Kim, E. Jeong, S.W. Lee, K. Han, Computational analysis of hydrogen bonds in protein–RNA complexes for interaction patterns, FEBS Lett. 552 (2003) 231–239.