IDENTIFICATION OF PROTEIN-PROTEIN BINDING SITES BY INCORPORATING THE PHYSICOCHEMICAL PROPERTIES AND STATIONARY WAVELET TRANSFORMS INTO PSEUDO AMINO ACID COMPOSITION

[1] Kobilova Guzal Ilhomovna, [2] Rahmatullayeva Madina Jasur kizi
[1] Jizzakh Polytechnic Institute, a great teacher
[2] Jizzakh Polytechnic Institute, student

Annotation: *With the explosive growth of protein sequences entering into protein data banks in the post-genomic era, it is highly demanded to develop automated methods for rapidly and effectively identifying the protein–protein binding sites (PPBS) based on the sequence information alone. To address this problem, we proposed a predictor called iPPBSPseAAC, in which each amino acid residue site of the proteins concerned was treated as a 15-tuple peptide segment generated by sliding a window along the protein chains with its center aligned with the target residue. The working peptide segment is further formulated by a general form of pseudo amino acid composition via the following procedures: (1) it is converted into a numerical series via the physicochemical properties of amino acids; (2) the numerical series is subsequently converted into a 20-D feature vector by means of the stationary wavelet transform technique. Formed by many individual "Random Forest" classifiers, the operation engine to run prediction is a two-layer ensemble classifier, with the 1st-layer voting out the best training data-set from many bootstrap systems and the 2nd-layer voting out the most relevant one from seven physicochemical properties.*

Keywords: *protein–protein binding sites; physicochemical property; stationary wavelet transform; pseudo amino acid composition; random forest; asymmetric bootstrap.*

Introduction

All cellular processes depend on precisely orchestrated interactions between proteins (Chou & Cai, 2006). A critical step in understanding the biological function of a protein is identification of the interface sites on which it interacts with other protein(s). Characterization of protein interactions is important for many problems covering from rational drug design to analysis of various biological networks[1].

The number of experimentally determined structures of protein–protein and protein–ligand complexes is still quite small, as reflected by the fact that the entries in UniprotKB/Swissprot (UniProt, 2013) is much larger than that in the Protein Data Bank

(Berman et al., 2000). The limited availability of structures often restricts the identification of binding sites of proteins and their functional annotation. Furthermore, the chemical or biological experimental methods are expensive, time-consuming and labor-intensive. Therefore, as a complement to the experimental methods, it is highly demanded to develop computational methods for identifying the protein–protein binding sites (PPBSs) according to their sequences information alone (Gallet, Charloteaux, Thomas, & Brasseur, 2000; Valencia & Pazos, 2002) [2].

Given a protein sequence, how can we identify which of its constituent amino acid residues are located in the binding site? Ofran and Rost (2003) and Yan, Dobbs, and Honavar (2004) have reported the following findings: (1) the residues involved in this kind of interactions usually tend to form clusters in sequences within four neighboring residues on either side; and (2) 97–98% of interface residues have at least one additional interface residue and 70–74% have at least four additional interface residues. Their analysis indicates that the neighboring residues of an actual interface residue have higher potential for being the interface residues, suggesting that fragments of protein sequences (referred to as sub-sequences hereafter) may contain useful information or features for discriminating between interaction and non-interaction sites. Several approaches have been proposed for predicting protein–protein interaction sites from amino acid sequence. Kini and Evans (1996), based on their observations on the frequency of proline residues occurring near the interaction sites, proposed a method for predicting the potential PPBSs by detecting the presence of proline bracket. Shortly afterward, using the multiple sequence alignment to detect correlated changes of the interacting protein domains, Pazos, Helmer-Citterich, Ausiello, and Valencia (1997) offered a different method to predict the contacting residue pairs. In 2000, Gallet et al [3]. (2000) introduced an approach to identify the interacting residues by analyzing the sequence hydrophobicity with the method developed by Eisenberg, Schwarz, Komaromy, and Wall (1984) [4].

In 2003, Ofran and Rost (2003) used sub-sequences of nine consecutive residues to develop a neural network-based method with a post-processing filter to predict interface residues. Subsequently, Yan et al. (2004) also used subsequence of nine residues to develop a two-stage classifier by combining support vector machine (SVM) and Bayesian network classifiers, achieving a higher accuracy. Two years later, Wang et al. (2006) also developed a predictor in this regard by using SVM with features extracted from spatial sequence and evolutionary scores based on a phylogenetic tree. Since the three-dimensional (3D) structures are unknown for most of proteins, the sequence-based method plays an important role in protein binding site prediction[5].

Unfortunately, several issues (Chen & Jeong, 2009; Sikic, Tomic, & Vlahovicek, 2009) exist that have made the sequence-based approach particularly difficult. The

main problems are as follows: (i) the effective features common to all the binding sites are hard to extract because the biological properties responsible for protein– protein interacting are not fully understood; (ii) the prediction of binding sites is to deal with a highly imbalanced classification problem because the number of non-binding sites of a protein pair is substantially larger than that of binding ones, and hence prone to cause bias; (iii) there is no good benchmark data-set due to lack of a unique definition for the binding sites, as reflected by the fact that one definition of the binding sites is based on the distance between the carbon atoms concerned, but another on the change of the accessible surface area (ASA) value between the bounded and unbounded status[6].

2. Materials and methods

2.1. Benchmark data-set

Two benchmark data-sets were used for the current study. One is the "surface-residue" data-set and the other is "all-residue" data-set, as elaborated below. The protein–protein interfaces are usually formed by those residues, which are exposed to the solvent after the two counterparts are separated from each other. Given a protein sample with L residues as expressed by

$$\mathbf{P} = R_1R_2R_3R_4R_5R_6R_7 \cdots R_L \qquad (1)$$

where ASA(Ri|P) is the ASA of Ri when it is a part of protein P, ASA(Ri) is the accessible surface area of the free Ri that is actually its maximal ASA as given in Table 1 (Ofran & Rost, 2003) [6,7], and /ð Þ Ri is the ratio of the two. Furthermore, the surface residue Ri is deemed as interfacial residue (Jones & Thornton, 1996) if

$$\phi(R_i) = \frac{ASA(R_i|\mathbf{P})}{ASA(R_i)} > 25\% \qquad (2)$$

where ASA(Ri|P) is the ASA of Ri when it is a part of protein P, ASA(Ri) is the accessible surface area of the free Ri that is actually its maximal ASA as given in Table 1 (Ofran & Rost, 2003), and /ð Þ Ri is the ratio of the two. Furthermore, the surface residue Ri is deemed as interfacial residue (Jones & Thornton, 1996) if

$$ASA(R_i|\mathbf{P}) - ASA(R_i|\mathbf{PP}) > 1\text{Å}^2 \qquad (3)$$

where ASA(Ri|PP) is the accessible surface area of Ri when it is a part of protein–protein complex.

AA	A	B	C	D	E	F	G	H	I	K	L	M
MaxASA	106	160	135	163	194	197	84	184	169	205	164	188
AA	N	P	Q	R	S	T	V	W	X	Y	Z	
MaxASA	157	136	198	248	130	142	142	227	180	222	196	

Table 1. Maximum ASA of different amino acids.a

Note: B stands for D or N; Z for E or Q, and X for an undetermined amino acid.aAmino acids are represented by their one-letter codes[8].

For a given protein, we can use DSSP program (Kabsch & Sander, 1983) to find out all its surface residues based on Equation (2), and use PSAIA program (Mihel, Šikić, Tomić, Jeren, & Vlahovicek, 2008) to find all its interfacial residues based on Equation (3).

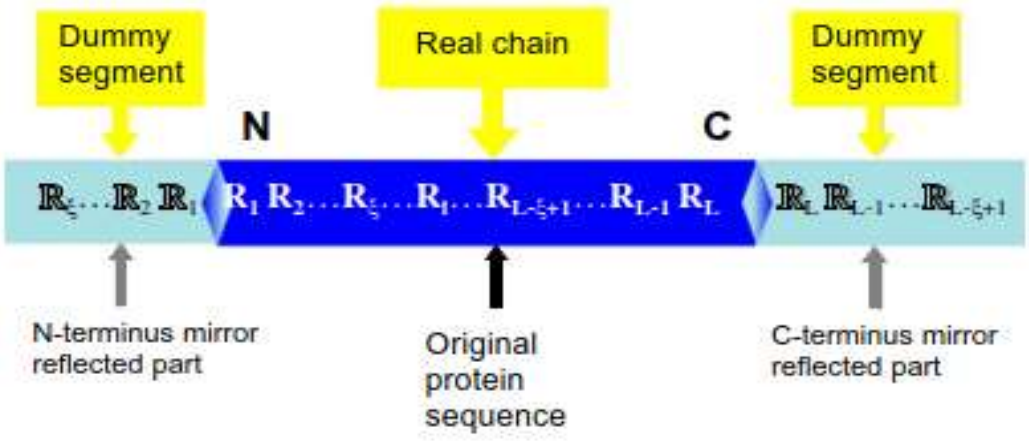

Figure 1. A schematic drawing to show how to use the extended chain of Equation (7) to define the working segments of Equation (6) for those sites when their sequence positions in the protein are less than n or greater L n, where the left dummy segment stands for the mirror image of R1R2 Rn at N-terminus and the right dummy segment for that of RLnþ1 RL1RL at the C-terminus[8,9].

2.2. Flexible sliding window approach

For a protein chain as formulated by Equation (1), the sliding window approach (Chou, 2001a) and flexible sliding window approach (Chou & Shen, 2007b) are often used to investigate its various post-translational modification (PTM) sites protease cleavage sites (Chou, 1996). Here, we also use it to study PPBSs. In the sliding window approach, a scaled window is denoted by ½ n; þn (Chou, 2001a). Its width is 2n þ 1, where n is an integer. When sliding it along a protein chain P (Equation (1)), one can see through the window a series of consecutive peptide segments as formulated by

$$\mathbf{P}_{\xi}(\mathbb{R}_0) = \mathrm{R}_{-\xi}\mathrm{R}_{-(\xi-1)}\cdots\mathrm{R}_{-2}\mathrm{R}_{-1}\mathbb{R}_0\mathrm{R}_{+1}\mathrm{R}_{+2}\cdots\mathrm{R}_{+(\xi-1)}\mathrm{R}_{+\xi}$$

Where Rn represents the n-th upstream amino acid residue from the center, Rþn the n-th downstream amino acid residue, and so forth. The amino acid residue R0 at the center is the targeted residue. When its sequence position in P (cf. Equation (1)) is less than n or greater L n; the corresponding Pnð Þ R0 is defined, instead by P of Equation (1), but by the following dummy protein chain

$$\mathbf{P}(\text{dummy}) = \mathbb{R}_{\xi} \cdots \mathbb{R}_2 \mathbb{R}_1 \\ \Updownarrow \mathrm{R}_1 \mathrm{R}_2 \cdots \mathrm{R}_{\xi} \cdots \mathrm{R}_i \cdots \mathrm{R}_{L-\xi+1} \cdots \mathrm{R}_{L-1} \mathrm{R}_L \\ \Updownarrow \mathbb{R}_L \mathbb{R}_{L-1} \cdots \mathbb{R}_{L-\xi+1}$$

where the symbol ⇳ stands for a mirror, the dummy segment Rn R2R1 stands for the image of R1R2 Rn reflected by the mirror, and the dummy segment RLRL1 RLnþ1 for the mirror image of RLnþ1 RL1RL (Figure 1). Accordingly, P(dummy) of Equation (7) is also called the mirror-extended chain of protein P. [10].

REFERENCES

1. Althaus, I. W., Gonzales, A. J., Diebel, M. R., Kezdy, F. J.,Aristoff, P. A., Tarpley, W. G., & Reusser, F. (1993). Kinetic studies with the nonucleoside HIV-1 reverse transcriptase inhibitor U-88204E.
2. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., … Bourne, P. E. (2000). The protein data bank. Nucleic Acids Research, 28, 235–242.
3. Breiman, L. (2001). Random forests. Machine Learning, 45, 5–32.
4. Cai, Y. D. & Chou, K. C. (2003). Nearest neighbour algorithm for predicting protein subcellular location by combining functional domain composition and pseudo-amino acid composition. Biochemical and Biophysical Research Communications, 305, 407–411.
5. Dehzangi, A., Heffernan, R., Sharma, A., Lyons, J., Paliwal, K., & Sattar, A. (2015). Gram-positive and Gram-negative protein subcellular localization by incorporating evolutionary-based descriptors into Chou′s general PseAAC. Journal Theoretical Biology, 364, 284–294.
6. Eisenberg, D., Schwarz, E., Komaromy, M., & Wall, R. (1984). Analysis of membrane and surface protein sequences with the hydrophobic moment plot. Journal of Molecular Biology, 179, 125–142.
7. Gallet, X., Charloteaux, B., Thomas, A., & Brasseur, R. (2000). A fast method to predict protein interaction sites from sequences. Journal of Molecular Biology, 302, 917–926.
8. Kandaswamy, K. K., Martinetz, T., Moller, S., Sridharan, S., Pugalenthi (2011). AFP-Pred: A random forest approach for predicting antifreeze proteins from sequence-derived properties. Journal of Theoretical Biology, 270, 56–62.
9. Ofran, Y., & Rost, B. (2003). Predicted protein–protein interaction sites from local sequence information. FEBS Letters, 544, 236–239.

10. Schnell, J. R., & Chou, J. J. (2008). Structure and mechanism of the M2 proton channel of influenza A virus. Nature, 451, 591–595.

11. Kobilova G.I. et al. INNOVATSION TEXNOLOGIYALARNI ISHLAB CHIQARISHGA JORIY ETISH //Educational Research in Universal Sciences. – 2023. – Т. 2. – №. 1 SPECIAL. – С. 108-111.

12. Кобилова Г.И. ЭФФЕКТИВНОЕ ИСПОЛЬЗОВАНИЕ ОТХОДОВ ПРИ ПРОИЗВОДСТВЕ КОНСЕРВИРОВАНИЯ. – 2023.