Predicting Residue-Wise Contact Orders in Proteins by Support Vector Regression with Parametric-Insensitive Model

Pei-Yi Hao and Lung-Biao Tsai

Abstract. A major challenge in structural bioinformatics is the prediction of protein structure and function from primary amino acid sequences. The residue-wise contact order (RWCO) describes the sequence separations between the residues of interest and its contacting residues in a protein sequence. RWCO provides comprehensive and indispensable important information to reconstructing the protein three-dimensional structure from a set of one-dimensional structural properties. Accurately predicting RWCO values could have many important applications in protein three-dimensional structure prediction and protein folding rate prediction, and give deep insights into protein sequence-structure relationships. In this paper, we developed a novel approach to predict residue-wise contact order values in proteins based on support vector regression with parametric insensitive model.

1 Introduction

A major challenge in structural bioinformatics is the prediction of protein structure and function from primary amino acid sequences. This problem is becoming more pressing now as the protein sequence-structure gap is widening rapidly as a result of the completion of large-scale genome sequencing projects [1,2]. As an intermediate but useful step, predicting a number of key properties of proteins including secondary structure, solvent accessibility, contact numbers and contact order is a possible and promising strategy, which simplifies the prediction task by projecting the three-dimensional structures onto one dimension, i.e. strings of residue-wise structural assignments [6].

Residue-wise contact order (RWCO) is a new kind of one dimensional protein structure representing the extent of long-range contacts, which is a sum of sequence separations between the given residue and all the other contacting residues [8,9,11]. RWCO provides comprehensive and indispensable important information

Pei-Yi Hao and Lung-Biao Tsai
Department of Information Management, National Kaohsiung University of Applied Sciences, Kaohsiung, Taiwan
e-mail: haupy@cc.kuas.edu.tw

B.-C. Chien and T.-P. Hong (Eds.): Opportunities and Challenges, SCI 214, pp. 7–12.
springerlink.com

to reconstructing the protein three-dimensional structure from a set of one-dimensional structural properties. Kinjo et al. first proposed a simple linear regression method to predict RWCO values and the local sequence information with multiple sequence alignments in the form of PSI-BLAST profiles was extracted using a sliding window scheme centered on the target residue [8]. Song et al. first adopt a support vector regression (SVR) algorithm to predict residue-wise contact order values in proteins, starting from primary amino acid sequences [11]. In this paper, a new SV regression algorithm, called *par-v*-SVR, is proposed by using a parametric insensitive loss function such that the corresponding insensitive zone of *par-v*-SVR can have arbitrary shape. This can be useful in situations where the noise is heteroscedastic, that is, where it depends on **x**.

2 SV Regression with Parametric Insensitive Model

Support Vector (SV) machines comprise a class of learning algorithms, motivated by results of statistical learning theory [12]. Originally developed for pattern recognition, they represent the decision boundary in terms of a typically small subset of all training examples, called the Support Vectors. In order for this property to carry over to the case of SV Regression, Vapnik devised the so-called ε-insensitive loss function [12]:

$$\left|y - f(\mathbf{x})\right|_{\varepsilon} = \max\left\{0, \left|y - f(\mathbf{x})\right| - \varepsilon\right\} \tag{2.1}$$

which does not penalize errors below some ε>0, chosen a priori. To motivate the new algorithm that shall be proposed, note that the parameter ε in original support vector regression (SVR) algorithm can be useful if the desired accuracy of the approximation can be specified beforehand. Besides, the ε-insensitive zone in the SVR is assumed to has a tube (or slab) shape. Namely, the radius of the insensitive zone is a user-predefined constant, and we do not care about the errors as long as they are inside the ε-insensitive zone. The selection of a parameter ε may seriously affect the modeling performance. In this paper, an new SV algorithm, called *par-v*-SVR, is derived to evaluate the interval regression model by using a new parametric-insensitive loss function, which automatically adjusts the interval to include all data [4]. The parametric-insensitive loss function is defined by

$$\left|y - f(\mathbf{x})\right|_{g} := \max\left\{0, \left|y - f(\mathbf{x})\right| - g(\mathbf{x})\right\} \tag{2.2}$$

where f and g are real-valued functions on the a domain R^n, $\mathbf{x} \in R^n$ and $y \in R$. The basic idea of SV regression is that a nonlinear regression function is achieved by simply mapping the input patterns $\mathbf{x}_i$ by Φ: $R^n \rightarrow$ F into a high-dimensional feature space F. Hence, the proposed *par-v*-SVR seeks to estimate the following two functions:

$$f(\mathbf{x}) = \langle \mathbf{w} \cdot \Phi(\mathbf{x}) \rangle + b \text{ , where } \mathbf{w} \in F \text{ , } \mathbf{x} \in R^n \text{ , } b \in R \text{ ,}$$

$$g(\mathbf{x}) = \langle \mathbf{c} \cdot \Phi(\mathbf{x}) \rangle + d \text{ , where } \mathbf{c} \in F \text{ , } \mathbf{x} \in R^n \text{ , } d \in R \text{ .}$$

The problem of finding the **w**, **c**, b, and d that minimize the empirical risk $R_{emp}^{g}[f] = \frac{1}{N}\sum_{i=1}^{N} |y_i - f(\mathbf{x}_i)|_g$ is equivalent to the following optimization problem:

$$\underset{\mathbf{w},\mathbf{c},b,d,\xi_i,\xi_i^*}{\text{minimize}} \ \frac{1}{2}\|\mathbf{w}\|^2 + C\left(v \cdot \left(\frac{1}{2}\|\mathbf{c}\|^2 + d \right) + \frac{1}{N}\sum_{i=1}^{N} (\xi_i + \xi_i^*) \right)$$

subject to

$$\left(\langle \mathbf{w} \cdot \Phi(\mathbf{x}_i) \rangle + b \right) + \left(\langle \mathbf{c} \cdot \Phi(\mathbf{x}_i) \rangle + d \right) \geq y_i - \xi_i \tag{2.3}$$

$$\left(\langle \mathbf{w} \cdot \Phi(\mathbf{x}_i) \rangle + b \right) - \left(\langle \mathbf{c} \cdot \Phi(\mathbf{x}_i) \rangle + d \right) \leq y_i + \xi_i^* \quad \text{and} \quad \xi_i, \xi_i^* \geq 0 \quad \text{for } i=1,\ldots,N.$$

Using the Lagrangian theorem, the dual problem can be formulated as

$$\text{maximize} \begin{cases} \dfrac{-1}{2}\displaystyle\sum_{i=1}^{N}\sum_{j=1}^{N} (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) \langle \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j) \rangle \\ -\dfrac{1}{2Cv}\displaystyle\sum_{i=1}^{N}\sum_{j=1}^{N} (\alpha_i + \alpha_i^*)(\alpha_j + \alpha_j^*) \langle \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j) \rangle + \sum_{i=1}^{N} (\alpha_i - \alpha_i^*) y_i \end{cases}$$

subject to (2.4)

$$\sum_{i=1}^{N} (\alpha_i - \alpha_i^*) = 0, \quad \sum_{i=1}^{N} (\alpha_i + \alpha_i^*) = C \cdot v, \quad \alpha_i, \alpha_i^* \in \left[0, \frac{C}{N} \right]. \quad \text{for } i=1,\ldots,N.$$

From the Karush-Kuhn-Tucker (KKT) conditions, parameters b and d can be computed as follows:

$$b = \frac{-1}{2}\left(\langle \mathbf{w} \cdot \Phi(\mathbf{x}_i) \rangle + \langle \mathbf{w} \cdot \Phi(\mathbf{x}_j) \rangle + \langle \mathbf{c} \cdot \Phi(\mathbf{x}_i) \rangle - \langle \mathbf{c} \cdot \Phi(\mathbf{x}_j) \rangle - y_i - y_j \right) \tag{2.5}$$

$$d = \frac{-1}{2}\left(\langle \mathbf{w} \cdot \Phi(\mathbf{x}_i) \rangle - \langle \mathbf{w} \cdot \Phi(\mathbf{x}_j) \rangle + \langle \mathbf{c} \cdot \Phi(\mathbf{x}_i) \rangle + \langle \mathbf{c} \cdot \Phi(\mathbf{x}_j) \rangle - y_i + y_j \right) \tag{2.6}$$

for some $\alpha_i, \alpha_j^* \in (0, C/N)$.

3 The Concept of Residue-Wise Contact Orders

Prediction of protein three-dimensional structure from primary sequence is the central problem in structural bioinformatics. One protein structural feature is of particular interest here, namely, residue-wise contact order (RWCO) which can be used to enhance protein fold recognition. The concept of residue-wise contact order (RWCO) was first introduced by Kinjo and Nishikawa [8,9]. The discrete RWCO values of the *i*-th residue in a protein sequence with M residues is defined by

$$RWCO_i = \frac{1}{M} \sum_{j:|j-i|>2}^{M} |i-j| \sigma(r_{i,j}) \quad \begin{cases} \sigma(r_{i,j}) = 1, & if\ r_{i,j} < r_d \\ \sigma(r_{i,j}) = 0, & if\ r_{i,j} \geq r_d \end{cases} \tag{3.1}$$

where $r_{i,j}$ is the distance between the C$_\alpha$ atoms of the *i*-th and *j*-th residues (C$_\alpha$ atoms for glycine) in the protein sequence. Two residues are considered to be in contact if their C$_\alpha$ atoms locate within a sphere of the threshold radius r_d. Note that the trivial contacts between the nearest and second-nearest residues are excluded. In order to smooth the discrete RWCO values, Kinjo et al. proposed a particular sigmoid function [8,9], which is given by

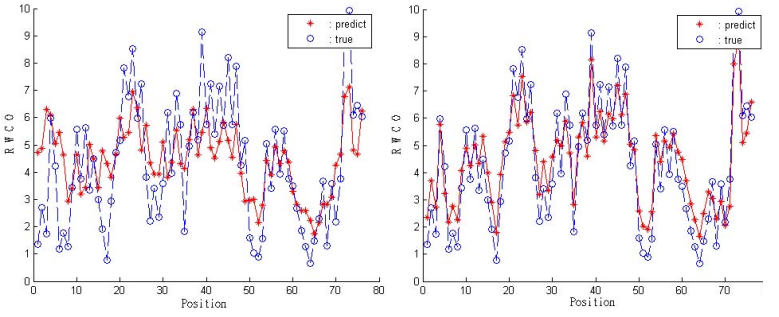
$$\sigma(r_{i,j}) = 1 \Big/ \left\{ 1 + \exp[w(r_{i,j} - r_d)] \right\} \tag{3.2}$$

where w is a parameter that determines the sharpness of the sigmoid function. In the present study, for the sake of comparison, we set rd = 12 Å and w = 3, which was adopted by Kinjo et al. [8,9].

4 Experiments

In this experiment, we apply the proposed *par-v*-SVR to predict residue-wise contact order values in proteins, starting from primary amino acid sequences. The Gaussian kernel $k(\mathbf{x}, \mathbf{y}) = \exp\left(-\|\mathbf{x}-\mathbf{y}\|^2 / 2\sigma^2\right)$ is used here. The optimal choice of parameters C, v and σ was tuned using a grid search mechanism. We used the same dataset previously prepared by Kinjo and Nishikawa [8,9], which included 680 protein sequences and was originally extracted from ASTRAL database version 1.65 [3]. There are a total of 120421 residues in this dataset. The protein chain names and their corresponding amino acid sequences, and the detailed RWCO information with a radius cutoff of 12Å can be found in [11]. To measure the performance of *par-v*-SVR methods in this application, we calculated the Pearson's correlation coefficients (CC) between the predicted and observed RWCO values in a protein sequence as given by

$$CC = \sum_{i=1}^{N} (x_i - \bar{x})(r_i - \bar{r}) \Bigg/ \sqrt{\left[\sum_{i=1}^{N} (x_i - \bar{x})^2 \right] \left[\sum_{i=1}^{N} (r_i - \bar{r})^2 \right]} \tag{4.1}$$



(a) (b)

Fig. 4.1 The predicted RWCO for protein d1n7oa2 obtained by (a) the original SVR and (b) the proposed *par-v*-SVR, respectively. RWCO values are used with a radius cutoff of 12 A. Observed and predicted RWCO are represented by solid and dashed lines, respectively

where x_i and r_i are the observed and predicted normalized RWCO values of the *i*-th residue, and $\bar{x}$ and $\bar{r}$ are their corresponding means. Here *N* is the total residue number in a protein. The root mean square error (RMSE) is also given by

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(r_i - x_i)^2} \tag{4.2}$$

The predicted performances were evaluated on the whole datasets by 5-fold cross-validation. In original SVR, the correlation coefficient (CC) between predicted and observed residue-wise contact order (RWCO) can reach 0.695, with normalized root mean square error (RMSE) less than 2.0545. In our approach, the use of parametric insensitive model can increase the correlation coefficient to 0.734 and decrease the root mean square error to 1.9719. To illustrate the performance in this study, fig 4.1 shows the predicted and observed RWCO values in protein d1n7oa2 obtained by the original SVR and the proposed *par-v*-SVR, respectively.

5 Conclusion

In the present study, we proposed a novel method to predict the RWCO profiles from amino acid sequences based on support vector regression with parametric insensitive model (*par-v*-SVR). Different from the linear regression approach, our method uses the non-linear radial basis kernel function (RBF) to approximate and determine the sequence-RWCO relationship. We compared our prediction accuracy with original SVR. The experimental results show that the proposed *par-v*-SVR is slightly better than the original SVR in predicting protein structural profile values and describing sequence-structure relationships.

References

1. Bairoch, A., Apweiler, R.: The SWISS-PROT protein sequence data bank and its supplement TrEMBL. Nucleic Acids Res. 28, 45–48 (2000)
2. Berman, H.M., et al.: The Protein Data Bank. Nucleic Acids Res. 28, 235–242 (2000)
3. Chandonia, J.M., et al.: The ASTRAL Compendium in 2004. Nucleic Acids Res. 32, D189–D192 (2004)
4. Hao, P.Y.: Shrinking the Tube: A New Support Vector Regression Algorithm with Parametric Insensitive Model. In: The 6th International Conference on Machine Learning and Cybernetics, Hong Kong, China (2007)
5. Jones, D.T.: Protein secondary structure prediction based on position-specific scoring matrices. J. Mol. Biol. 292, 195–202 (1999)
6. Kinjo, A.R., Nishikawa, K.: Recoverable one-dimensional encoding of three-dimensional protein structures. Bioinformatics 21, 2167–2170 (2005)
7. Kihara, D.: The effect of long-range interactions on the secondary structure formation of proteins. Protein Sci. 14, 1955–1963 (2005)
8. Kinjo, A.R., Nishikawa, K.: Predicting Residue-wise Contact Orders of Native Protein Structure from Amino Acid Sequence. arXiv.org. q-bio.BM/0501015 (2005)
9. Kinjo, A.R., Nishikawa, K.: Predicting secondary structures, contact numbers, and residue-wise contact orders of native protein structure from amino acid sequence using critical random networks. Biophysics 1, 67–74 (2005)
10. Plaxco, K.W., et al.: Contact order, transition state placement and the refolding rates of single domain proteins. J. Mol. Biol. 277, 985–994 (1998)
11. Song, J., Burrage, K.: Predicting residue-wise contact orders in proteins by support vector regression. BMC Bioinformatics 7, 425 (2006)
12. Vapnik, V.: The Nature of Statistical Learning Theory. Springer, New York (2000)
13. Yuan, Z., Huang, B.: Prediction of protein accessible surface areas by support vector regression. Proteins 57, 558–564 (2004)
14. Yuan, Z.: Better prediction of protein contact number using a support vector regression analysis of amino acid sequence. BMC Bioinformatics 6, 248 (2005)