

Prediction of Protein Secondary Structures with a Novel Kernel Density Estimator

Yen-Jen Oyang*

Graduate Institute of Biomedical Electronics and
Bioinformatics
National Taiwan University
Taipei, Taiwan, R.O.C.

Darby Tien-Hao Chang

Department of Electrical Engineering
National Cheng Kung University
Tainan, Taiwan, R.O.C.

Yu-Yen Ou

Graduate School of Biotechnology and
Bioinformatics
Yuan-Ze University
Chung-Li, Taiwan, R.O.C.

Hao-Geng Hung

Department of Computer Science and Information
Engineering
National Taiwan University
Taipei, Taiwan, R.O.C.

Chien-Yu Chen

Department of Bio-Industrial Mechatronics
Engineering
National Taiwan University
Taipei, Taiwan, R.O.C.

Abstract - *Though prediction of protein secondary structures has been an active research issue in bioinformatics for quite a few years and many approaches have been proposed, a new challenge emerges as the sizes of contemporary protein structure databases such as the Protein Data Bank (PDB) continue to grow exponentially. The new challenge concerns how to effectively exploit the huge amount of structural information deposited in large protein structure databases and deliver ever-improving accuracy as the sizes of the databases continue to grow. This new challenge is addressed in this article by resorting to a kernel density estimation based approach. The kernel density estimator proposed in this article is distinctive in that the pointwise MSE (mean square error) of its basic form converges at $O(n^{-2/3})$ regardless of the dimension of the vector space, where n is the number of instances in the training dataset. In addition, just like many conventional kernel density estimators, it features average $O(n \log n)$ time complexity for generating the approximation function. The experimental results show that with the novel kernel density estimator the proposed predictor has been able to outperform the state-of-art predictors currently available. Experimental results further reveal that prediction accuracy delivered by the proposed predictor will continue to increase in the future as the size of the protein structure database keeps growing.*

Keywords: data classification, kernel density estimation, protein secondary structure, structural bioinformatics, supervised learning

1 Introduction

In structural biology, protein secondary structures act as the building blocks of the protein tertiary structures [1, 2]. Therefore, analysis of protein secondary structures is an essential intermediate step for obtaining a comprehensive picture of the tertiary structure of a polypeptide. In this respect, one of the main challenges is how to accurately identify the segments in a polypeptide that could fold to form secondary structures. This problem is normally referred to as prediction of protein secondary structures.

Though prediction of protein secondary structures has been an active research issue in bioinformatics for quite a few years and many approaches have been proposed [1, 3-9], a new challenge emerges as the sizes of contemporary protein structure databases such as the Protein Data Bank (PDB) [10] continue to grow exponentially. The new challenge, which has been addressed in several recently completed studies [8, 11], concerns how we can effectively exploit all the structural information deposited in the protein structure databases and deliver ever-improving

* To whom correspondence should be addressed. E-mail: yjoyang@csie.ntu.edu.tw, Tel: +886-2-33664888 ext. 431, Fax: +886-2-23688675.

prediction accuracy. Accordingly, the machine learning algorithm incorporated must feature a low time complexity for constructing a predictor and must be able to deliver superior prediction accuracy with large protein structure databases. Unfortunately, as reported in [3], the state-of-art support vector machine (SVM) algorithm, which prevails in many bioinformatics applications, does not cope well with this new challenge.

In this article, we will propose a novel kernel density estimator designed to address the challenge mentioned above. The major distinction of the proposed kernel density estimator is that the pointwise mean squared error (MSE) of its basic form converges at $O(n^{-2/3})$ regardless of the dimension of the vector space, where n is the number of instances in the training dataset. In addition, just like many conventional kernel density estimators, it features average $O(n \log n)$ time complexity for generating the approximation function [12]. These two favorite characteristics combined implies that a predictor of protein secondary structures designed with the novel kernel density estimator will be able to deliver superior prediction accuracy and will be able to effectively deal with the ever-growing protein structure databases. In this article, we will report the results of the experiments designed to confirm this argument. In the experiments, the training dataset for the proposed kernel density estimation based predictor was derived from the version of PDB at the end of 2005 and the testing dataset was derived from the protein structures deposited into the PDB during March 15 to June 1 in 2006. For comparison, the same testing dataset was used to evaluate the prediction accuracy delivered by the versions of the existing predictors that provide the same function at the end of February in 2006. Experimental results confirm that the proposed predictor is able to outperform the state-of-art predictors in terms of prediction accuracy. Experimental results further confirm the conjecture that the accuracy delivered by the proposed predictor will continue to improve in the future as the size of the protein structure database keeps growing.

2 Proposed Kernel Density Estimator

In this section, we will first elaborate the mathematical basis of the novel kernel density estimator proposed in this article. In particular, we will show that the pointwise mean squared error (MSE) of the basic form of the proposed kernel density estimator converges at $O(n^{-2/3})$, regardless of the dimension of the vector space, where n is the number of instances in the training dataset. Then, we will discuss how the proposed kernel density estimator can be exploited in data classification applications.

Since we can always conduct a translation operation with the coordinate system, without loss of generality, we assume in the following discussion that it is the pointwise

MSE at the origin of the coordinate system that is of concern. Let $f_X(x_1, x_2, \dots, x_m)$ denote the probability density function of the distribution of concern in an m -dimensional vector space. Assume that $f_X(x_1, x_2, \dots, x_m)$ is a smooth function and $f_X(x_1, x_2, \dots, x_m) < \infty$ for all $(x_1, x_2, \dots, x_m) \in \mathbf{R}^m$. Let Z be the random variable that maps a sampling instance s_i taken from the distribution governed by f_X to $\|s_i\|^m$, where $\|s_i\|$ is the distance from the origin to s_i . Accordingly, we have the distribution function $F_Z(z)$ of Z equal to

$$\iint \dots \int_{x_1^2 + x_2^2 + \dots + x_m^2 \leq z^{2/m}} f_X(x_1, x_2, \dots, x_m) dx_1 dx_2 \dots dx_m \quad (1)$$

for $z \geq 0$ and $F_Z(z) = 0$ for $z < 0$.

Theorem 1: Let $f_Z(z) = \lim_{\substack{\varepsilon \rightarrow 0 \\ \varepsilon > 0}} \frac{F_Z(z + \varepsilon) - F_Z(z)}{\varepsilon}$ for $z \geq 0$.

Assume that $f_Z(z)$ is a smooth function in $[0, \infty)$. Then, we have $f_Z(0) = \frac{\pi^{m/2}}{\Gamma(m/2 + 1)} f_X(\mathbf{0})$, where $\Gamma(\cdot)$ is the gamma function [13].

Proof:

Since $F_Z(0) = 0$, we have

$$\begin{aligned} & \lim_{\substack{\varepsilon \rightarrow 0 \\ \varepsilon > 0}} \frac{F_Z(\varepsilon) - F_Z(0)}{\varepsilon} \\ &= \lim_{\substack{\varepsilon \rightarrow 0 \\ \varepsilon > 0}} \frac{\iint \dots \int_{x_1^2 + x_2^2 + \dots + x_m^2 \leq \varepsilon^{2/m}} f_X(x_1, x_2, \dots, x_m) dx_1 dx_2 \dots dx_m}{\varepsilon} \end{aligned}$$

By the Taylor expansion,

$$\begin{aligned} & f_X(x_1, x_2, \dots, x_m) \\ &= f_X(\mathbf{0}) + \frac{\partial f_X(\mathbf{0})}{\partial x_1} x_1 + \dots + \frac{\partial f_X(\mathbf{0})}{\partial x_m} x_m + \text{high-order terms} \end{aligned}$$

Furthermore, in region where $x_1^2 + x_2^2 + \dots + x_m^2 \leq \varepsilon^{2/m}$, we have $x_1 \rightarrow 0$, $x_2 \rightarrow 0$, ..., $x_m \rightarrow 0$ as $\varepsilon \rightarrow 0$. Therefore,

$$\begin{aligned} f_Z(0) &= \lim_{\substack{\varepsilon \rightarrow 0 \\ \varepsilon > 0}} f_X(\mathbf{0}) \cdot \frac{(\sqrt[m]{\varepsilon})^m \pi^{m/2}}{\Gamma(m/2 + 1)} \cdot \frac{1}{\varepsilon} \\ &= \frac{\pi^{m/2}}{\Gamma(m/2 + 1)} \cdot f_X(\mathbf{0}), \end{aligned}$$

where $\frac{(\sqrt[m]{\mathcal{E}})^m \pi^{m/2}}{\Gamma(m/2+1)}$ is the volume of a sphere in an m -

dimensional vector space with radius = $\sqrt[m]{\mathcal{E}}$.

□

Theorem 1 implies that we can obtain an estimate of $f_X(\mathbf{0})$ by first obtaining an estimate of $f_Z(0)$. Since f_Z is a univariate probability density function, if we employ a fixed kernel density estimator [12] to estimate $f_Z(0)$, then, as Theorem 2 shows, we can obtain an estimator of $f_X(\mathbf{0})$ with the pointwise MSE converging at $O(n^{-2/3})$.

Theorem 2: Let $\{s_1, s_2, \dots, s_n\}$ be a set of sampling instances randomly and independently taken from the distribution governed by f_X in the m -dimensional vector space. Then, with $\sigma = \lambda \cdot n^{-1/3}$ and λ being a positive real number,

$$\hat{f}_X(\mathbf{0}) = \frac{\Gamma(m/2+1)}{\pi^{m/2}} \cdot \sum_{i=1}^n \frac{1}{n} \cdot \frac{\sqrt{2}}{\sqrt{\pi}\sigma} \exp\left(-\frac{\|s_i\|^2}{2\sigma^2}\right)$$

is an estimator of $f_X(\mathbf{0})$ with the pointwise MSE converging at $O(n^{-2/3})$.

Proof:

Let $z_i = \|s_i\|^m$ and $\hat{f}_Z(0) = \sum_{i=1}^n \frac{1}{n} \cdot \frac{\sqrt{2}}{\sqrt{\pi}\sigma} \exp\left(-\frac{z_i^2}{2\sigma^2}\right)$ with

$\sigma = \lambda \cdot n^{-1/3}$. We have

$$MSE[\hat{f}_Z(0)] = (E[\hat{f}_Z(0)] - f_Z(0))^2 + Var[\hat{f}_Z(0)] \text{ and}$$

$$E[\hat{f}_Z(0)] = \sum_{i=1}^n \int_0^\infty \frac{1}{n} \cdot \frac{\sqrt{2}}{\sqrt{\pi}\sigma} \exp\left(-\frac{z^2}{2\sigma^2}\right) f_Z(z) dz.$$

As $n \rightarrow \infty$, we have $\sigma \rightarrow 0$ and

$$\begin{aligned} & O(E[\hat{f}_Z(0)] - f_Z(0)) \\ &= O\left[2 \cdot \int_0^\infty \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{z^2}{2\sigma^2}\right) [f_Z(0) + f'_Z(0^+)z] dz - f_Z(0)\right] \\ &= O\left[\sqrt{\frac{2}{\pi}} \cdot f'_Z(0^+) \cdot \sigma\right], \end{aligned}$$

where $f'_Z(0^+) = \lim_{\substack{\mathcal{E} \rightarrow 0 \\ \mathcal{E} > 0}} \frac{f_Z(\mathcal{E}) - f_Z(0)}{\mathcal{E}}$.

Let

$$\hat{f}_{1/n}(0) = \frac{1}{n} \cdot \frac{\sqrt{2}}{\sqrt{\pi}\sigma} \exp\left(-\frac{z_1^2}{2\sigma^2}\right).$$

We have

$$E[\hat{f}_{1/n}^2(0)] = \int_0^\infty \frac{1}{n^2} \cdot \frac{2}{\pi\sigma^2} \exp\left(-\frac{z^2}{\sigma^2}\right) f_Z(z) dz.$$

Due to $\sigma \rightarrow 0$ as $n \rightarrow \infty$,

$$\begin{aligned} O(E[\hat{f}_{1/n}^2(0)]) &= O\left(\int_0^\infty \frac{1}{n^2} \cdot \frac{2}{\pi\sigma^2} \exp\left(-\frac{z^2}{\sigma^2}\right) f_Z(0) dz\right) \\ &= O\left(\frac{f_Z(0)}{n^2 \sigma \sqrt{\pi}}\right). \end{aligned}$$

Therefore, as $n \rightarrow \infty$,

$$\begin{aligned} O(Var[\hat{f}_{1/n}(0)]) &= O(E[\hat{f}_{1/n}^2(0)] - (E[\hat{f}_{1/n}(0)])^2) \\ &= O\left(\frac{f_Z(0)}{n^2 \sigma \sqrt{\pi}} - \frac{1}{n^2} (E[f_Z(0)])^2\right). \end{aligned}$$

Since $\sigma = \lambda \cdot n^{-1/3}$, as $n \rightarrow \infty$, we have

$$O(Var[\hat{f}_{1/n}(0)]) = O(n^{-5/3}).$$

Furthermore, since s_1, s_2, \dots, s_n are taken randomly and independently,

$$Var[\hat{f}_Z(0)] = n \cdot Var[\hat{f}_{1/n}(0)].$$

Therefore, as $n \rightarrow \infty$,

$$\begin{aligned} & O(MSE[\hat{f}_Z(0)]) \\ &= O\left((E[\hat{f}_Z(0)] - f_Z(0))^2 + Var[\hat{f}_Z(0)]\right) \\ &= O(n^{-2/3}). \end{aligned}$$

Let

$$\hat{f}_X(\mathbf{0}) = \frac{\Gamma(m/2+1)}{\pi^{m/2}} \cdot \sum_{i=1}^n \frac{1}{n} \cdot \frac{\sqrt{2}}{\sqrt{\pi}\sigma} \exp\left(-\frac{\|s_i\|^2}{2\sigma^2}\right)$$

with $\sigma = \lambda \cdot n^{-1/3}$. Then, we have

$$\begin{aligned} \text{MSE}[\hat{f}_x(\mathbf{0})] &= E[(\hat{f}_x(\mathbf{0}) - f_x(\mathbf{0}))^2] \\ &= E\left[\left(\frac{\Gamma(m/2+1)}{\pi^{m/2}}\right)^2 \cdot (\hat{f}_z(0) - f_z(0))^2\right] \\ &= \left(\frac{\Gamma(m/2+1)}{\pi^{m/2}}\right)^2 \cdot \text{MSE}[\hat{f}_z(0)] \end{aligned}$$

Since $\text{MSE}[\hat{f}_z(0)]$ converges at $O(n^{-2/3})$ with $\sigma = \lambda \cdot n^{-1/3}$, $\text{MSE}[\hat{f}_x(\mathbf{0})]$ converges at $O(n^{-2/3})$ as well.

□

The estimator presented in Theorem 2 forms the basis of the novel kernel density estimator proposed in this article. Since both Theorem 1 and Theorem 2 address only the pointwise MSE, for real applications we have incorporated the basic idea of variable kernel density estimator [12] to generalize the estimator presented in Theorem 2 and obtain the so-called super-radius based kernel density estimator (SRKDE) as follows:

$$\hat{f}_x^*(\mathbf{v}) = \frac{\Gamma(m/2+1)}{\pi^{m/2}} \cdot \sum_{i=1}^n \frac{1}{n} \cdot \frac{\sqrt{2}}{\sqrt{\pi}\sigma_i} \exp\left(-\frac{\|\mathbf{s}_i - \mathbf{v}\|^{2m}}{2\sigma_i^2}\right),$$

where

- 1) $\sigma_i = \beta \frac{[R_k(\mathbf{s}_i)]^m}{k}$;
- 2) β is the smoothing parameter with order $O(n^{2/3})$;
- 3) $R_k(\mathbf{s}_i)$ is the distance from \mathbf{s}_i to its k -th nearest neighbor;
- 4) k is a parameter to be set.

The proposed kernel density estimator is so named because random variable Z maps a sampling instance \mathbf{s}_i taken from the distribution governed by f_x to $\|\mathbf{s}_i\|^m$ and $\|\mathbf{s}_i\|^m$ is referred to as the super-radius of \mathbf{s}_i in this article. For data classification applications, we will construct one SRKDE to approximate the distribution of one class of training instances in the vector space. Then, a query instance located at \mathbf{v} is predicted to belong to the class that gives the maximum value among the likelihood functions defined in the following:

$$L_j(\mathbf{v}) = \frac{|S_j| \cdot \hat{f}_j^*(\mathbf{v})}{\sum_h |S_h| \cdot \hat{f}_h^*(\mathbf{v})},$$

where $|S_j|$ is the number of class- j training instances and $\hat{f}_j^*(\mathbf{v})$ is the SRKDE corresponding to class- j training

instances. In our current implementation, aiming to improve the execution time of the classifier, we include only a limited number, denoted by k' , of nearest class- j training instances of \mathbf{v} in computing $\hat{f}_j^*(\mathbf{v})$.

As mentioned earlier, one main distinctive property of the kernel density estimation based approach is that the average time taken to construct a classifier is in the order $O(n \log n)$, where n is the total number of training instances. This argument is based on the assumption that the kd-tree structure [14] is employed in the implementation. For detailed analysis of the time complexity, please refer to the discussion presented in [15], which provides the detailed analysis with a similar kernel density estimator. Another advantage enjoyed by the kernel density estimation based approach is that prediction can be made without resorting to the one-against-one or one-against-all mechanism in case there are more than two classes of instances. Concerning the execution time for making prediction with n' incoming objects, it is shown in [15] that the average time complexity is $O(n' \log n)$.

3 Experimental Results

This section reports the experiments conducted to investigate how the SRKDE based predictor actually performs. The experiments were designed to simulate how accurately alternative approaches can exploit the information currently present in the PDB to predict the secondary structures of newly identified protein sequences.

In the following discussion, we will use Prote2S to refer to the SRKDE based predictor that we have recently implemented. In addition, we have created a web server that provides prediction of protein secondary structures with Prote2S (<http://prote2s.csie.ntu.edu.tw>). For Prote2S, the training dataset was derived from the version of PDB at the end of 2005 and the testing dataset was derived from the 1289 protein structures deposited into the PDB during March 15 to June 1 in 2006. In order to reduce redundancy in the training dataset, the CD-HIT clustering algorithm [16] with the similarity threshold set to 0.4 was invoked to remove redundant protein structures in the PDB. After this process, a total of 8163 protein chains remained. For generation of the training dataset, we followed the approach employed in [3]. With this approach, one training instance is created for each residue in the sequences of the 8163 protein chains by associating the residue with the vector computed by the PSI-BLAST software package [17] for the residue and its 14 neighboring residues. As a result, a total of 1,829,733 training instances were generated and each instance was labeled by one of the three types of secondary structure elements, alpha-helix, beta-strand, and loop, determined by DSSP [18]. For generation of the testing dataset, BLAST [17] was invoked to remove redundant p r o t e i n c h a i n s .

Table 1 Parameter settings in Prote2S

Parameter	m	β	k	k'
Value	1	1	7	1500

Table 2 Comparison of prediction accuracy of alternative predictors

(A)			
Predictor	Prote2S	PSIPRED	HYPROSP
Accuracy	85.0%	78.0%	83.7%
(B)			
Predictor	Prote2S	PROTEUS	
Accuracy	83.9%	81.9%	

Table 3 Prediction accuracy of Prote2S vs. size of the training dataset

Portion of dataset	1/16	1/8	1/4	1/2	1
Accuracy	75.9%	77.1%	79.6%	81.9%	85.0%

The criterion guarantees that no two chains remaining have their e-value similarity score computed by BLAST smaller than 0.1. With this criterion applied, the testing dataset then includes a total of 36,443 instances derived from a total of 154 non-redundant protein chains deposited into the PDB during March 15 to June 1 in 2006.

Table 1 shows the parameter settings employed in Prote2S. Parameters m and β have been set to the default values and parameters k and k' have been set through conducting cross validation on the training dataset. One may wonder why the default values of both parameters m and β have been set to 1. According to our experiences with applying the SRKDE based predictor to a variety of problems, there are normally quite a few combinations of parameter settings that yield the same level of accuracy and among them there are always several combinations with $m = 1$ and $\beta = 1$. Therefore, we simply set the default values of both m and β to 1 in order to reduce the number of parameters to be tuned during the cross validation process.

Table 2 shows how the Prote2S performs in comparison with some of the most advanced predictors for analysis of protein secondary structures, including PSIPRED [4], HYPROSP [5, 6], and PROTEUS [7]. HYPROSP and PROTEUS are two most recently released predictors and PSIPRED is probably the most widely used predictor as of today and was incorporated in the design of HYPROSP and PROTEUS.

In the comparison with PSIPRED and HYPROSP presented in Table 2(A), the testing dataset described above has been used. Meanwhile, for comparison with PROTEUS, a separate training dataset and a separate testing dataset have been generated because in the article published by the PROTEUS group VADAR [19], instead of

DSSP, was used to label the classes of the residues in the training dataset. Basically, the same procedures described above for generation of the training and testing datasets have been used but the training and testing datasets generated for the experiments reported in Table 2(B) contain 1,201,565 and 24,649 instances, respectively.

Experimental results presented in Table 2 show that Prote2S is able to outperform PSIPRED, HYPROSP, and PROTEUS in terms of prediction accuracy. Furthermore, since the average execution time to construct a SRKDE based predictor is in the order of $O(n \log n)$, where n is the number of training instances, it is conceivable that the SRKDE based predictor can effectively cope with the fast growth rate of the protein structure database and deliver ever-increasing prediction accuracy through fully exploiting the structural information deposited in the database. Table 3 reports the experiment conducted to confirm this conjecture. In this experiment, we provided the SRKDE based predictor with training datasets derived from randomly selected portions of the 8163 protein chains employed to generate the training dataset in the experiment reported in Table 2. Then, we evaluated the prediction accuracy that the SRKDE based predictor can deliver. The experimental results show that the prediction accuracy delivered by the SRKDE based predictor improves as the number of protein structures exploited to generate the training dataset increases. The experimental results further show that with the current version of PDB we have not yet reached the point where the prediction accuracy saturates. In other words, with the size of PDB continues to grow, it is anticipated that the prediction accuracy delivered by Prote2S will continue to improve.

4 Conclusions

This article proposes the super radius based kernel density estimator (SRKDE) and reports how the SRKDE based predictor of protein secondary structures performs. The major distinction of the proposed kernel density estimator is that the pointwise mean squared error (MSE) of its basic form converges at $O(n^{-2/3})$ regardless of the dimension of the vector space, where n is the number of instances in the training dataset. In addition, the average time complexity for construction of a SRKDE based predictor is in the order of $O(n \log n)$. As a result, the SRKDE based predictor is able to effectively exploit the structural information deposited in the large protein structure database and deliver ever-improving prediction accuracy as the size of the database continues to grow. Experimental results reveal that the SRKDE based predictor has been able to outperform the existing web servers that provide the same function in terms of prediction accuracy. Experimental results further reveal that accuracy delivered by the SRKDE based predictor will continue to increase in the future as the size of PDB keeps growing.

In structural biology, protein secondary structures act as the building blocks of the protein tertiary structures. Therefore, the capability to accurately predict the segments in the polypeptide that could fold to form secondary structures is essential for eventually obtaining a comprehensive picture of the tertiary structure of the polypeptide. In this respect, the SRKDE based predictor can output the values of the likelihood functions to be exploited in subsequent structural analyses of the polypeptide.

References

- [1] I. Eidhammer, I. Jonassen, and W. R. Taylor, *Protein Bioinformatics: An Algorithmic Approach to Sequence and Structure Analysis*. Chichester: John Wiley & Sons Ltd, 2004.
- [2] A. Lesk, *Introduction to Bioinformatics*. Oxford: Oxford University Press, 2005.
- [3] J. J. Ward, L. J. McGuffin, B. F. Buxton, and D. T. Jones, "Secondary structure prediction with support vector machines," *Bioinformatics*, vol. 19, pp. 1650-1655, Sep 1 2003.
- [4] L. J. McGuffin, K. Bryson, and D. T. Jones, "The PSIPRED protein structure prediction server," *Bioinformatics*, vol. 16, pp. 404-405, Apr 2000.
- [5] K. P. Wu, H. N. Lin, J. M. Chang, T. Y. Sung, and W. L. Hsu, "HYPROSP: a hybrid protein secondary structure prediction algorithm - a knowledge-based approach," *Nucleic Acids Research*, vol. 32, pp. 5059-5065, 2004.
- [6] H. N. Lin, J. M. Chang, K. P. Wu, T. Y. Sung, and W. L. Hsu, "HYPROSP II - A knowledge-based hybrid method for protein secondary structure prediction based on local prediction confidence," *Bioinformatics*, vol. 21, pp. 3227-3233, Aug 1 2005.
- [7] S. Montgomerie, S. Sundararaj, W. J. Gallin, and D. S. Wishart, "Improving the accuracy of protein secondary structure prediction using structural alignment," *BMC Bioinformatics*, vol. 7, p. 301, Jun 14 2006.
- [8] O. Dor and Y. Zhou, "Achieving 80% Ten-fold Cross-validated Accuracy for Secondary Structure Prediction by Large-scale Training," *Proteins: Structure, Function, and Bioinformatics*, vol. 66, 2007.
- [9] J. A. Cuff and G. J. Barton, "Application of multiple sequence alignment profiles to improve protein secondary structure prediction," *Proteins-Structure Function and Genetics*, vol. 40, pp. 502-511, Aug 15 2000.
- [10] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The Protein Data Bank," *Nucleic Acids Research*, vol. 28, pp. 235-242, Jan 1 2000.
- [11] S. Montgomerie, S. Sundararaj, W. J. Gallin, and D. S. Wishart, "Improving the accuracy of protein secondary structure prediction using structural alignment," *Bmc Bioinformatics*, vol. 7, pp. -, Jun 14 2006.
- [12] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*. Boca Raton: Chapman & Hall/CRC, 1986.
- [13] E. Artin, *The Gamma Function*. New York: Holt, Rinehart and Winston, 1964.
- [14] F. P. Preparata and M. I. Shamos, *Computational Geometry: An Introduction*: Springer, 1985.
- [15] Y. J. Oyang, S. C. Hwang, Y. Y. Ou, C. Y. Chen, and Z. W. Chen, "Data classification with radial basis function networks based on a novel kernel density estimation algorithm," *IEEE Transactions on Neural Networks*, vol. 16, pp. 225-236, Jan 2005.
- [16] W. Z. Li and A. Godzik, "CD-HIT: a fast program for clustering and comparing large sets of protein or nucleotide sequences," *Bioinformatics*, vol. 22, pp. 1658-1659, Jul 1 2006.
- [17] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. H. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," *Nucleic Acids Research*, vol. 25, pp. 3389-3402, Sep 1 1997.
- [18] W. Kabsch and C. Sander, "Dictionary of Protein Secondary Structure - Pattern-Recognition of Hydrogen-Bonded and Geometrical Features," *Biopolymers*, vol. 22, pp. 2577-2637, 1983.
- [19] L. Willard, A. Ranjan, H. Y. Zhang, H. Monzavi, R. F. Boyko, B. D. Sykes, and D. S. Wishart, "VADAR: a web server for quantitative evaluation of protein structure quality," *Nucleic Acids Research*, vol. 31, pp. 3316-3319, Jul 1 2003.