PROTEINS: Structure, Function, and Genetics 44:57–59 (2001)

SHORT COMMUNICATION

Some Insights into Protein Structural Class Prediction

G.P. Zhou* **and Nuria Assa-Munt**
Department of Structural Biology, Burnham Institute, La Jolla, California

ABSTRACT **It has been quite clear that the success rate for predicting protein structural class can be improved significantly by using the algorithms that incorporate the coupling effect among different amino acid components of a protein. However, there is still a lot of confusion in understanding the relationship of these advanced algorithms, such as the least Mahalanobis distance algorithm, the component-coupled algorithm, and the Bayes decision rule. In this communication, a simple, rigorous derivation is provided to prove that the Bayes decision rule introduced recently for protein structural class prediction is completely the same as the earlier component-coupled algorithm. Meanwhile, it is also very clear from the derivative equations that the least Mahalanobis distance algorithm is an approximation of the component-coupled algorithm, also named as the covariant-discriminant algorithm introduced by Chou and Elrod in protein subcellular location prediction (Protein Engineering, 1999; 12:107–118). Clarification of the confusion will help use these powerful algorithms effectively and correctly interpret the results obtained by them, so as to conduce to the further development not only in the structural prediction area, but in some other relevant areas in protein science as well. Proteins 2001;44:57–59.** © **2001 Wiley-Liss, Inc.**

Key words: protein structural class; amino acid composition; Mahalanobis distance; component-coupled algorithm; Bayes decision rule

In a pioneer study, Chou and Zhang[1,2] introduced Mahalanobis distance to reflect the coupling effect among different amino acid components of a protein, significantly improving the success rate of protein structural class prediction. Owing to the normalization of amino acid composition, of the 20 amino acid components, only 19 are independent. Therefore, the Mahalanobis distance based on the 20-dimensional amino acid composition space must be divergent and meaningless. To overcome the divergence difficulty, the Mahalanobis distance originally introduced by Chou and Zhang[1] for protein structural class prediction was defined in a $20-1=19$-dimensional space. Furthermore, to solidify the mathematical frame, an invariance theorem was given by Chou[3] that states that the values of the Mahalanobis distance will remain the same regardless of which one of the 20 components is left out for forming the reduced 19-dimensional space. The introduction of the Mahalanobis distance is a key step in taking into account the coupling effect of different amino acid components for protein structural class prediction. This was further confirmed by Bahar et al.,[4] who also attempted physical analysis trying to understand the recognition of protein structural classes in terms of amino acid composition. However, as shown later, the prediction algorithm based solely on Mahalanobis distance, the so-called least Mahalanobis distance algorithm,[1] does not completely cover the coupling effect. Therefore, for some cases, particularly when the subset sizes in the training data set are substantially different, the least Mahalanobis distance algorithm might yield quite poor results, as elaborated by Chou et al.[5] and Zhou.[6]

In order to cover that part of coupling effect missed in the least Mahalanobis distance algorithm, the component-coupled algorithm was proposed,[5–7] and its mathematical principle, presented by Liu and Chou.[13] It has been indicated from a series of reports in literature that the component-coupled algorithm is much more powerful than the simple geometry algorithms in predicting structural class of proteins based on their amino acid composition. The algorithm also has been successfully used by Chou and Elrod[12] in protein subcellular location prediction.

Recently, Wang and Yuan[9] proposed using Bayes decision rule to predict protein structural class. Claiming that their algorithm was the most powerful one, they used it to determine the upper limit of the prediction rate for structural classes based on amino acid composition. The following questions have been naturally raised. Does their algorithm represent some thing new, or does it just bear a different name? If the least Mahalanobis distance algorithm is an approximation of the Bayes decision rule, what about the component-coupled algorithm to which the least Mahalanobis algorithm is also an approximation? Is there any essential difference between the Bayes decision rule and the component-coupled algorithm, or not at all? It is

*Correspondence to: G.P. Zhou, Department of Structural Biology, Burnham Institute, 10901 N. Torrey Pines Rd., La Jolla, CA 92037. E-mail: gpzhou@burnham-inst.org

Received 17 November 2000; 23 February 2001

important to give a clear-cut answer to these questions so that various confusions caused by the ambiguity can be thoroughly removed. Answering these questions, however, is not quite straightforward and obvious a task. This is because Wang and Yuan[9] used many different mathematical symbols and terms in their equations, which might baffle and puzzle investigators in understanding their real implications, especially those who are not trained in statistical mathematics. To address these questions and reveal the essence hidden in the abstract mathematical symbols, let us consider the following derivation.

The prediction algorithm proposed by Wang and Yuan[9] is based on the following equation (see eq. 13 of their article):

$$d_l(\mathbf{x}) = \ln P(\omega_l) - (19/2)\ln 2\pi - (\tfrac{1}{2})\ln\left|\sum_l\right| - (\tfrac{1}{2})[(\mathbf{x}-\boldsymbol{\mu}_l)^T \sum_l^{-1} (\mathbf{x}-\boldsymbol{\mu}_l)] \quad (1)$$

Multiplying a factor -2 on both side of equation 1, we obtain

$$D_l(\mathbf{x}) = -2d_l(\mathbf{x}) = -2\ln P(\omega_l) + 19\ln 2\pi + \ln\left|\sum_l\right| + [(\mathbf{x}-\boldsymbol{\mu}_l)^T \sum_l^{-1} (\mathbf{x}-\boldsymbol{\mu}_l)] \quad (2)$$

By reorganizing the order of terms, the above equation cab be written as

$$D_l(\mathbf{x}) = [(\mathbf{x}-\boldsymbol{\mu}_l)^T \sum_l^{-1} (\mathbf{x}-\boldsymbol{\mu}_l)] + \ln\left|\sum_l\right| - 2\ln P(\omega_l) + 19\ln 2\pi \quad (3)$$

By contrast, the component-coupled algorithm proposed by the previous investigators is based on the following equation (see eq. 14a of Chou and Maggiora[7]):

$$F_{\mathrm{M}}(\mathbf{X}, \mathbf{X}^{\zeta}) = D^2_{\mathrm{M}}(\mathbf{X}, \mathbf{X}^{\zeta}) + \ln \Pi_{\zeta} - 2\ln \Psi_{\zeta} + \Lambda \ln(2\pi) \quad (4)$$

Both equations 3 and 4 consist of four terms. After carefully reading the mathematical definition of these terms in these two articles, one can immediately find the following. The first term of eq. 3 is equivalent to the first term of eq. 4, both representing the contribution from the Mahalanobis distance to the structural class prediction. The second term of eq. 3 is equivalent to the second term of eq. 4, both representing the contribution from the determinant of the covariance matrix to the structural class prediction. The third term of eq. 3 is equivalent to the third term of eq. 4, both representing the prior probability of the subset concerned. The fourth term of eq. 3 is equivalent to the fourth term of eq. 4, both representing a constant.

Accordingly, the algorithm proposed by Wang and Yuan[9] is essentially the same as the algorithm proposed by the previous investigators. The only difference made by Wang and Yuan[9] was to change the name from the "component-coupled algorithm" to the "Bayes decision rule," as well as using different symbols and term order. Furthermore, following exactly the same procedure as described by the previous investigators,[7] they ignored the constant term and the prior probability term to reduce their eq. 13 to eq. 15, i.e.,

$$d_l(\mathbf{x}) = -(\tfrac{1}{2})\ln\left|\sum_l\right| - (\tfrac{1}{2})[(\mathbf{x}-\boldsymbol{\mu}_l)^T \sum_l^{-1} (\mathbf{x}-\boldsymbol{\mu}_l)] \quad (5)$$

Again letting us multiply a factor -2 on both sides of equation 5 and reorganize the order of terms, we obtain

$$D_l(\mathbf{x}) = -2d_l(\mathbf{x}) = [(\mathbf{x}-\boldsymbol{\mu}_l)^T \sum_l^{-1} (\mathbf{x}-\boldsymbol{\mu}_l)] + \ln\left|\sum_l\right| \quad (6)$$

which is also essentially the same as the eq. 14b of Chou and Maggiora,[7] as given below:

$$F_{\mathrm{M}}(\mathbf{X}, \mathbf{X}^{\zeta}) = D^2_{\mathrm{M}}(\mathbf{X}, \mathbf{X}^{\zeta}) + \ln \Pi_{\zeta} \quad (7)$$

Actually, eq. 6 is also equivalent to eq. 14 of Chou et al.[5] or eq. 8 of Zhou.[6] That is why when applying the component-coupled algorithm[5,6] on the 131 proteins constructed by Nakashima et al.[10] and on the 120 proteins by Chou,[3] the overall self-consistency rates by resubstitution test are 99.2% and 100%, exactly the same as the rates reported by Wang and Yuan[9] in their Table 1. However, if using the least Mahalanobis distance algorithm[1] that only covers part of the coupling effect, the corresponding rates would drop down to 94.7% and 99.2%, respectively. This is because the least Mahalanobis distance algorithm only contains the contribution from the first term of eq. 7 but misses that from the second term. Therefore, the least Mahalanobis distance algorithm is an approximation of the component-coupled algorithm. However, there is no essential difference whatsoever between the Bayes decision rule proposed recently by Wang and Yuan[9] and the component-coupled algorithm proposed by the previous investigators.[5–7]

The success rate in statistical prediction is determined by two factors: one is the power of the prediction algorithm, and the other is the completeness of the training data set. Although the component-coupled algorithm is a quite powerful one, the structural class of some proteins might be incorrectly predicted if they are outside the frame defined by the current very limited training data set, as cautioned by the authors of the component-coupled algorithm. So far the proteins found in nature are about 2 million in number. The glut of genomic information in the form of 3 billion base pairs, assembled into tens of thousands of genes, will need to be translated into proteins. Therefore, the very limited training data set selected by Wang and Yuan[9] for deriving the upper limit of protein structural prediction is far from complete. As a matter of fact, the number of proteins in their training data set is less than 0.05% of the proteins found in nature thus far.

Using such a tiny training data set and an algorithm completely identical to the component-coupled algorithm, Wang and Yuan[9] claimed that they have found the upper limit and resolved a "paradox" in protein structural class prediction. Obviously, this is inappropriate and hard to believe. In addition, the so-called "paradox" targeted by them actually did not exist in the first place for several reasons:

1. As is well known in statistics, the success prediction rate in practical application should be measured by the result of jackknife test, but not by self-consistency test, neither by the limited independent data set test nor by arbitrary subsampling test. (See Mardia et al.,[11] for the mathematical principle and Chou and Zhang[2] for a comprehensive discussion.) However, the argument by Wang and Yuan[9] leading to a paradox was based on the results of the self-consistency test, rather than the jackknife test.
2. The number of protein structural classes is not limited to four, as considered by Wang and Yuan[9] (e.g., in a study performed by Chou and Maggiora,[7] the number of protein structural classes was extended to seven (all-α, all-β, α/β, $\alpha+\beta$, μ, σ, ρ) in order to cover more proteins in nature.
3. The protein structural class prediction is conceptually different from the secondary structure prediction or secondary structural content prediction; it is completely normal and should not be construed as a "paradox," even if the success rate for a given data set in protein structural class prediction is higher than that in secondary structure prediction or secondary structural content prediction, as discussed by Liu and Chou.[8] Accordingly, the "paradox" is actually a delusion caused by incorrectly construing the higher than 90% self-consistency rate as a practical success rate for protein structural class prediction.

In any case, one thing is certain: for a same training data set, the component-coupled algorithm will yield much higher success rate than the simple geometry algorithms in protein structural class prediction. This has been repeatedly demonstrated in many reports in literature. It is anticipated that if we can construct a complete or quasi-complete training data set, the component-coupled algorithm will become a useful tool in proteomics, the next frontier that will take over where genomics leaves off.

REFERENCES

1. Chou KC, Zhang CT. Predicting protein folding types by distance functions that make allowances for amino acid interaction. J Biol Chem 1994;269:22014–22020.
2. Chou KC, Zhang CT. Prediction of protein structural classes. Crit Rev Biochem Mol Biol 1995;30:275–349.
3. Chou KC. A novel approach to predicting protein structural classes in a (20-1)-D amino acid composition space. Proteins 1995;21:319–344.
4. Bahar I, Atilgan AR, Jernigan RL, Erman B. Understanding the recognition of protein structural classes by amino acid composition. Proteins 1997;29:172–185.
5. Chou KC, Liu W, Maggiora GM, Zhang CT. Prediction and classification of domain structural classes. Proteins 1998;31:97–103.
6. Zhou GP. An intriguing controversy over protein structural class prediction. J Protein Chem 1998;17:729–738.
7. Chou KC, Maggiora GM. Domain structural class prediction. Protein Eng 1998;11:523–538.
8. Liu WM, Chou KC. Prediction of protein secondary structure content. Protein Eng 1999; 12:1041–1050.
9. Wang ZX, Yuan Z. How good is prediction of protein structural class by the component-coupled method. Proteins 2000;38:165–175.
10. Nakashima H., Nishikawa K, Ooi T. The folding type of a protein is relevant to the amino acid composition. J Biochem 1986;99:152–162.
11. Mardia KV, Kent JT, Bibby JM. Multivariate analysis. London: Academic Press, 1979; p 322;.381.
12. Chou KC, Elrod DW. Protein subcellular location prediction. Protein Eng 1999;12:107–118.
13. Liu, WM, Chou KC. Prediction of protein structural classes by modified Mahalanobis discriminant algorithm. J Protein Chem 1998;17:209–217.