CN43-1258/TP
ISSN 1007-130X

计算机工程与科学
COMPUTER ENGINEERING & SCIENCE

2005 年第 27 卷第 8 期
Vol. 27, No. 8, 2005

文章编号:1007-130X(2005)08-0062-03

# 蛋白质结构预测算法的评估*

# Performance Assess of the Algorithms for Protein Structure Prediction

王勇献,王正华,张振慧
WANG Yong-xian, WANG Zheng-hua, ZHANG Zhen-hui
(并行与分布处理国家重点实验室,湖南 长沙 410073)
(National Laboratory for Parallel and Distributed Processing, Changsha 410073, China)

**摘 要**:选取合适的蛋白质结构预测算法的性能评估指标,是直接影响到衡量和比较各种蛋白质结构预测算法优劣的重要问题。本文对目前各种评估指标进行了剖析比较,总结对比了各种评估指标的优缺点,分析了其相互之间的联系与区别,并结合神经网络建模,提出各种评估指标的适用范围与使用原则。

**Abstract**: Accessing the algorithm performance plays an important role in the protein structure prediction. This paper provides a unified overview of performance measurements widely used in the performance assess of the algorithms for protein structure prediction. We briefly discuss the advantages and disadvantages of each approach, and analyze the correlation and comparison among them. Furthermore, we discuss the considerations of some applications based on neural network modeling, and finally give the applicable area and workable principle of these measurements.

**关键词**:蛋白质结构预测;性能评估;生物信息学

**Key words**: protein structure prediction; performance measurement; bioinformatics

**中图分类号**:Q811;TP391　　**文献标识码**:A

## 1 引言

从构成蛋白质的氨基酸序列出发预测其空间结构(即蛋白质结构预测问题),是当前生物信息学研究中的热点问题之一[1,2]。先前的研究已经提出了许多种预测算法。然而,如何客观公平地比较这些算法性能的优劣,从而挑选出一种或几种"最佳"的算法,是研究者经常面临的一个难题[3]。蛋白质结构预测问题中各种分类预测算法的建模主要是通过对训练数据集的学习来完成,而算法的性能则主要在测试数据集上进行。一旦选定了适当的测试数据集后,就要考虑怎样选择适当的性能评估指标,以便对算法做出公正合理的性能评估。

本文主要讨论了蛋白质分类预测算法的各种性能评估指标,并结合神经网络建模的实例,分析总结了各种指标的适用范围及相互联系,为蛋白质结构预测建模中选取合适的性能评估指标提供一个有意义的参考。

## 2 蛋白质结构预测问题描述

考虑测试数据集中一条长为 $N$ 的氨基酸残基序列,已知其真实的结构数据 $D=(d_1,\cdots,d_N)$,以及预测算法产生的输出 $M=(m_1,\cdots,m_N)$,其中,$d_i$, $m_i\in U$, $i=1,2,\cdots,N$。我们的基本问题是:如何评估预测结果 $M$ 的优劣,或者说,如何评价 $M$ 对 $D$ 的近似、拟合能力?

根据实际预测问题中对 $U$ 的不同表示,可将 $D$ 和 $M$ 分成以下几种变量类型:(1)区间标度变量(Interval-Scaled Variable):$U$ 为连续的区间范围(例如表示氨基酸主链中相继两个残基形成的转角范围)。(2)二元变量(Binary Variable):$U$ 只有两个对称的状态分别表示对立的两种意义(例如表示某个残基的结构类为"$\alpha$ 螺旋"或"非 $\alpha$ 螺旋"),一般表示为 $U=\{0,1\}$。(3)标称变量(Nominal Variable):$U$ 有 $K(>2)$ 个无顺序关系的状态值(例如表示某个残基处于 $\alpha$ 螺旋、$\alpha$ 片层和不规则卷曲三种结构区域之一内),为

* 收稿日期:2003-10-28;修订日期:2004-02-04
基金项目:国家自然科学基金资助项目(60003001)
作者简介:王勇献(1975-),男,河南安阳人,博士生,研究方向为生物信息学;王正华,教授,研究方向分布与生物信息学;张振慧,博士生,研究方向为生物信息学。
通讯地址:410073 湖南省长沙市砚瓦池正街 47 号并行与分布处理国家重点实验室;Tel:(0731)4573666;E-mail:wang_yongxian@163.com
Address: National Laboratory for Parallel and Distributed Processing, 47 Yanwachi St, Changsha, Hunan 410073, P. R. China

了方便处理，常用$U=\{1,2,\cdots,K\}$的表示法。

如果利用阈值对区间标度变量的输出作截断处理，则输出就转化成二元变量或标称变量的形式。在实际蛋白质结构预测中，二分类问题一般使用二元变量表示，多分类问题一般使用标称变量表示。对于标称变量和二元变量情形，我们记 $X_{i,j,n}=I_{\{d_n=i,m_n=j\}}$，$i,j\in U$，$n=1,2,\cdots,N$，其中 $I_D$ 表示事件 $D$ 的示性变量，即若 $D$ 成立，$I_D=1$；否则，$I_D=0$。令 $a_{ij}=\sum_n X_{i,j,n}$，则矩阵 $\boldsymbol{A}=(a_{ij})_{K\times K}$ 为预测结果与真实结果的一致性矩阵。显然，有 $\sum_{i,j\in U} a_{ij}=N$。当考虑二元变量时，$U=\{1,0\}$，则矩阵 $\boldsymbol{A}$ 可记为：

$$\boldsymbol{A}=\begin{pmatrix} TP & FN \\ FP & TN \end{pmatrix}$$

其中，$TP$(True Positive，简称 $TP$)表示$\{d_i=1,\ m_i=1\}$的次数，$TN$(True Negative，简称 $TN$)表示$\{d_i=0,\ m_i=0\}$的次数，$FP$(False Positive，简称 $FP$)表示$\{d_i=0,m_i=1\}$的次数，$FN$(False Negative，简称 $FN$)表示$\{d_i=1,\ m_i=0\}$的次数。

# 3 蛋白质结构预测算法性能评估指标

对于蛋白质结构预测问题，研究者提出了各种不同的预测性能评估指标(或称性能度量指标)，用以比较、度量 $M$ 与 $D$ 的“吻合”程度。根据评估指标的来源及特征，可将它们分成百分比类、距离类以及相似性类三大类。

## 3.1 百分比类度量指标

对于二分类问题，最直观的性能评估是考虑从一致性矩阵 $\boldsymbol{A}$ 的四个元素中浓缩出一个指标。百分比类指标最适合于二分类预测问题的性能评估。常见的百分比类度量指标如表 1 所示。

表 1 常见的百分比类度量指标

| 评估指标 | 定义 | 描述及特点 |
|---|---|---|
| 敏感性 | $SE(D,M)=TP/(TP+FN)$ | 敏感性，又称测全率、命中率(recall)，刻画的是在结构类 $c$ 中被正确预测的部分所占的比例。 |
| 特异性 | $SP(D,M)=TP/(TP+FP)$ | 特异性，又称测准率、精度(precision)，主要刻画在预测结果是结构类 $c$ 的数据中正确预测所占的比例。 |
| 总体正确率 | $Q_a(D,M)=\frac{1}{N}\sum_i a_{ii}$ | 刻画预测结果的总体正确性[4]。应用广泛，未充分利用矩阵 $\boldsymbol{A}$ 的所有元素，综合能力略差。 |
| $Q_8$ | $Q_8(D,M)=1-\sqrt{\frac{FN^2+FP^2}{(FP+FN)^2+(TP+FP)^2}}$ | 刻画预测结果的总体正确性[5]。将 $\boldsymbol{A}$ 中元素变换成为高维空间中的点坐标，综合能力较好，但计算公式复杂。 |
| $SOV$ | | SOV (Segment Overlap Measurement，简称 SOV)主要评估局部片段的整体预测准确性，较好解决了区域模糊边界问题[6]。 |

## 3.2 距离类度量指标

若用数值向量表示 $D$ 和 $M$，则向量 $D$ 和 $M$ 之间的 $L^p$ 距离定义为 $L^p(D,M)=(\sum_i |d_i-m_i|)^{1/p}$。应用中常用到的距离有 $L^1$、$L^2$ 和 $L^\infty$，有时也直接使用 $L^2$ 的平方形式 $Q^2(D,M)$，并称作平方距离。

## 3.3 相似性度量指标

若将 $D$ 和 $M$ 视为概率随机变量(向量)，则可以用概率与信息论中的各种相似性来度量 $D$ 与 $M$ 的关系。

### 3.3.1 相关系数和近似相关系数

相关系数也称为 Pearson 相关系数、Matthews 相关系数。向量 $D$ 和 $M$ 的相关系数定义为其标准化向量 $D*$ 与 $M*$ 的内积：$C(D,M)=(D*,M*)$。特别地，在二元变量情形下：

$$C(D,M)=\frac{TP\times TN-FP\times FN}{N\sqrt{(TP+FN)(TP+FP)(FN+FP)(TN+FN)}} \quad (1)$$

为避免被零除的问题，Burset 和 Guigo 定义了一个近似相关系数的指标[7](其中约定 $a/0=0$)：

$$AC=2\times\left(\frac{TP}{TP+FN}+\frac{TP}{TP+FP}+\frac{TN}{TN+FP}+\frac{TN}{TN+FN}\right)-1 \quad (2)$$

通常，近似相关系数 $AC$ 值与相关系数 $C$ 在取值上非常相近。

### 3.3.2 相对熵

相对熵又称交叉熵、$KL$ 距离，来源于信息论[8,9]。对于蛋白质二级结构预测问题，相对熵可以写为：

$$H(D,M)=\sum_{i=1}^{N}\left[d_i\log\frac{d_i}{m_i}+(1-d_i)\log\frac{1-d_i}{1-m_i}\right] \quad (3)$$

在标准化处理后，相对熵适用于区间标度变量及标称变量的场合，但无法直接用于二元变量。

### 3.3.3 互信息

记 $X$ 和 $Y$ 为两个离散型概率分布，$Z$ 是其联合分布，则 $X$ 与 $Y$ 的互信息定义为 $I(X,Y)=H(Z,XY)$。在二元变量场合中，互信息可以写为[10]：

$$I(D,M)=\overline{TP}\log\overline{TP}+\overline{TN}\log\overline{TN}+\overline{FP}\log\overline{FP}+\overline{FN}\log\overline{FN}-\overline{FP}\log[(\overline{TP}+\overline{FP})(\overline{TP}+\overline{FN})]-\overline{FN}\log[(\overline{FN}+\overline{TP})(\overline{FN}+\overline{TN})]-\overline{FP}\log[(\overline{FP}+\overline{TP})(\overline{FP}+\overline{TN})]-\overline{TN}\log[(\overline{TN}+\overline{FN})(\overline{TN}+\overline{FP})] \quad (4)$$

其中，$\overline{TP}=TP/N$，$\overline{TN}$、$\overline{FP}$、$\overline{FN}$的定义类似。若令 $H(D)=-(\overline{TP}+\overline{FN})\log(\overline{TN}+\overline{FP})-(\overline{TN}+\overline{FP})\log(\overline{TN}+\overline{FP})$，则可以定义标准化的互信息系数[11]：$IC(D,M)=I(D,M)/H(D)$。易知，$IC(D,M)\in[0,1]$。

注 1：由于距离类的度量指标都与长度 $N$ 有关，因此在使用距离作为算法性能的评估指标时，一般采用标准化后的距离(即距离值乘以一个因子 $1/N$)的形式。

注 2：三类性能度量指标之间有密切联系。例如：(1)二元变量情形中，$L^1=Q^2$；(2)标准化的 $L1$ 实际上也是一个百分比度量指标；(3)相关系数与 $Q^2$ 距离满足关系式：$C(D,M)=2-Q^2(D^*,M^*)$。

# 4 数值测试与讨论

## 4.1 性能评估指标对神经网络建模的影响

神经网络由于其技术成熟、简单高效等优点而经常用于蛋白质的结构预测研究。神经网络建模时需要从样本数

据出发，通过不断地调整与修正模型参数，最终使目标(误差)函数达到最优。由于目标误差函数选择的不同，最终模型的预测效果也不同[9]。我们使用了不同的优化准则分别建立神经网络模型，并测试其对蛋白质二级结构类(α螺旋、β片层和不规则卷曲)的预测性能。所有试验中均采用了标准的三层 BP 网络结构，各结点使用 S 型激励函数；建模所用蛋白质数据集为 CB396[12]，并通过交叉验证技术对每个模型各测得 20 组性能数值。表 2 示出了分别使用"后验概率最大化"(MAP)和"均方误差最小化"(MSE)两个优化准则建模时所获得的平均预测性能。由此可见，利用 MAP 优化准则的建模具有较好的 $H$ 值，而使用 MSE 优化准则的建模则在 LMS 度量上表现较佳。分析其原因，是因为 MAP 优化准则本质上是优化 $D$ 和 $M$ 的相对熵(不计一个常数因子的影响)[10]，而 MSE 优化准则形式上就是优化 $D$ 和 $M$ 的 $L^2$ 距离。

表 2　不同优化准则下建模的神经网络的预测性能

| | $Q_a$ | *LMS* | *I* | *H* |
|---|---|---|---|---|
| NN-MAP | 0.712 7 | 31.58 | 0.461 3 | 0.910 4 |
| NN-MSE | 0.684 4 | 10.82 | 0.501 6 | 0.749 6 |

表 2 表明，不同的优化准则将导致不同预测性能的模型；同时也说明，利用性能评估指标可以有效地指导建模过程。例如，为了获得最佳的相对熵，就要采用 MAP 优化准则进行建模。此外，使用 LMS 与相对熵作为度量指标时，目标误差函数的全导数可以写成各分量局部误差的线性和，从而实现在线(联机)学习训练；而使用相对系数、互信息系数等度量指标时则没有这种特点，因而只能采用批量(脱机)学习方式。

### 4.2　各种性能评估指标的特点及使用原则

从表 2 还可以看出，使用不同的度量指标对多个算法或模型进行性能评估，结果有很大差异。具体应用时，应当综合各种性能评估指标的特性与适用范围后，合理选择一个或数个指标。

一般来说，百分比类指标问题最适合于二元变量的场合，其中敏感性 $SE$ 与特异性 $SP$ 是使用得最为广泛的两个，而且一般用敏感-特异曲线的表现形式。但是，它们与正确率 $Q_a$ 一样都仅使用了一致性矩阵 $\boldsymbol{A}$ 的部分信息，比较之下 $Q_3$ 则更全面一些。距离类指标刻画了 $D$ 和 $M$ 之间在空间上的相近性，其数值越小，表明预测效果越好；距离类指标适合于所有三种变量类型(区间标度变量、二元变量以及标称变量)，当应用于标称变量时，应当根据具体预测问题选取合适的数值编码表示，在二元变量场合，距离类指标也只涵盖了矩阵 $\boldsymbol{A}$ 中的部分信息；在使用时，为消除氨基酸链长度的影响，距离类指标一般都使用其标准化形式(参见第 3 节注 1)。基于信息论的各种相似性指标主要刻画 $D$ 和 $M$ 之间的相关性，相关系数取值 1、相对熵与互信息取值为 0，都表明 $M$ 对 $D$ 的拟合最好。除了相对熵不能应用于二元变量、近似相关系数仅能应用于二元变量之外，各种指标均适用于所有三种变量类型，当应用于二元变量时，相似性指标都使用了一致性矩阵 $\boldsymbol{A}$ 的所有信息，因而较为全面地反映了 $D$ 和 $M$ 之间的类似程度。

在蛋白质结构预测建模中选择性能评估指标时，总体应遵从以下原则：(1)如果是二分类预测问题，$M$ 和 $D$ 均使用了二元变量，原则上所有前述指标均可使用(除相对熵外)，考虑到表达式的简洁及反映信息的全面性，一般常用敏感-特异曲线及相关系数等指标；(2)若考虑多分类问题，$M$ 和 $D$ 均为标称变量，多使用总体正确率 $Q_a$ 及各种相似性指标，当选用了恰当的数值表示法时，有时也使用距离类指标(主要是欧氏距离)；(3)若 $M$ 和 $D$ 均为区间标度变量，则距离类指标与相似性指标是最佳的侯选，二者各有优缺点：距离类指标计算简便，但易受编码表示法的影响；而相似性指标信息全面，但计算却较为复杂繁冗。

## 5　结束语

蛋白质结构预测算法的评估准则的选取，是直接影响到衡量、比较各种蛋白质结构预测算法优劣的重要问题。每届 CASP 国际比赛都会对参赛的各种预测算法进行性能评估排名，采用不同的性能评估指标可能会导致完全不同的性能评估排名。因此，他们综合使用了数种指标进行评估，尽可能保证结果的客观公正。本文主要就目前各种性能评估指标进行综合剖析，比较了各种评估指标的优缺点，分析了其相互之间的联系与区别，并结合神经网络建模，提出了各种评估准则的适用范围与使用原则。

### 参考文献：


[1] A D Baxevanis, B F F Ouellette. Bioinformatics : A Practical Guide to the Analysis of Genes and Proteins[M]. New York: Wiley-Interscience, 2001.

[2] R Durbin, S R Eddy, A Krogh, et al. Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids [M]. Cambridge University Press, 1998.

[3] D Baker, A Sali. Protein Structure Prediction and Structural Genomics[J]. Science, 2001, 294(5540): 93-96.

[4] J Moult, K Fidelis, A Zemla, et al. Critical Assessment of Methods of Protein Structure Prediction (Casp): Round Iv [J]. Proteins, 2001, 45(Suppl 5): 2-7.

[5] C T Zhang, R Zhang. A Refined Accuracy Index to Evaluate Algorithms of Protein Secondary Structure Prediction[J]. Proteins, 2001, 43(4): 520-522.

[6] A Zemla, C Venclovas, K Fidelis, et al. A Modified Definition of Sov, a Segment-Based Measure for Protein Secondary Structure Prediction Assessment[J]. Proteins, 1999, 34(2): 220-223.

[7] M Burset, R Guigo. Evaluation of Gene Structure Prediction Programs[J]. Genomics, 1996, 34(3): 353-367.

[8] S Kullback, R A Leibler. On Information and Sufficiency[J]. The Annals of Mathematical Statistics, 1951, 22: 79-86.

[9] P Baldi, S Brunak. Bioinformatics : The Machine Learning Approach[M]. Cambridge, Mass. : MIT Press, 2001.

[10] P Baldi, S Brunak, Y Chauvin, et al. Assessing the Accuracy of Prediction Algorithms for Classification: An Overview[J]. Bioinformatics, 2000, 16(5): 412-424.

[11] B Rost. Review: Protein Secondary Structure Prediction Continues to Rise[J]. J Struct Biol, 2001, 134(2-3): 204-218.

[12] J A Cuff, G J Barton. Evaluation and Improvement of Multiple Sequence Methods for Protein Secondary Structure Prediction[J]. Proteins, 1999, 34(4): 508-519.