DOI:10.16262/j.cnki.1000-8217.2000.05.015

·成果简介·

蛋白质结构分类与结构类预测研究

张春霆*

(天津大学生命科学与生物工程研究院,天津 300072)

[关键词] 蛋白质,结构类,预测,分类,二级结构含量,生物信息学

现代分子生物学研究表明,蛋白质是执行生物功能的大分子,在生命活动中起着极其重要的作用。蛋白质的功能是由其空间结构决定的。结构和功能的一致性促进了对蛋白质结构研究的巨大热忱,结构生物学的发展方兴未艾。虽然空间结构是由一级结构(即氨基酸的排列顺序)决定的这一原理,已经确立了 30 余年,但在一般情况下由一级结构预测其 3 级结构仍然是不现实的。与此同时,用 X 射线衍射和核磁共振技术测定蛋白质的空间结构取得了巨大的进展。迄今为止,已有近万种蛋白质的结构被测定,其中大部分属于同源蛋白质,非同源的仅有一千余种。通过总结这些结构的共性和特性,对于阐明肽链的折叠规律是非常有益的。而其中对蛋白质结构的分类和预测不仅有着巨大的实用价值,而且对于探索肽链折叠的规律(第二套生物学密码)也有重要的理论价值。

蛋白质结构分类的概念是 1976 年由英国学者 Chothia 和 Levitt 提出来的。根据当时已知的 31 种球蛋白质晶体 X 射线衍射所测定的 3 级结构,他们将蛋白质分成 α,β,α+β 和 α/β 4 类,有时往往又把后两类合并成一类,称为 αβ 类。但是这两位学者只给出结构类的定性描述,直到 1986 年日本学者 Ooi 等人根据蛋白质的 2 级结构含量,定量地定义了 α,β 和 αβ 类。此后,又有若干不同的定义发表。但是,这些定义带有相当大的主观任意性。在对大量蛋白质结构进行统计分析的基础上,作者在 1998 年提出了一个客观的定量的分类标准[1],后又做了改进[2]。从蛋白质结构的原子坐标数据出发,经过复杂的计算,可以明确判断该蛋白质属于 α,β 或 αβ 的哪一类。将 αβ 类进一步分类为 α+β 和 α/β 的定量标准,也同时提出[3]。为了简化计算,可直接从蛋白质的 2 级结构出发进行分类,而不必用原子坐标从头算[4]。利用上述定量分类标准对数千个已知结构的蛋白质进行分类,其结果与国际上著名的蛋白质分类数据库 SCOP 高度一致,而 SCOP 是根据进化关系和肽链的折叠原理用手工方法进行蛋白质分类的。作者提出的这一套分类标准是建立在对大量蛋白质结构统计分析的基础之上,因而具有可靠性和先进性,超过了以前已有的一些定量的分类标准,包括 Ooi 等人提出的标准,对于蛋白质结构分类学具有重要的意义。

蛋白质结构分类是在蛋白质的空间结构已知的前提下进行的。如果某蛋白质的空间结构未知,能否根据其一级结构来预测其结构类?这就是结构类预测。1986 年日本学者 Ooi 等人首先用他们提出的分类标准对当时已知结构的 100 余种蛋白质进行了分类,并发现结构类与蛋白质的氨基酸组成有关。利用氨基酸组成,他们提出了结构类预测算法,其准确性已经达到 70%。从 1992 年起,作者及其合作者对结构类预测作了深入、系统和多方面的研究,发表了一系列论文[5—17],对于结构类预测这一课题的本质的了解也日益深入。这些工作引起了国际蛋白质研究界的广泛注意,曾应邀发表了长篇综述文章[18]。但是这些算法和相应软件仍然未能达到实用水平。这主要有两方面的原因:一是训练用蛋白质数据库中蛋白质的数量不够多,缺乏充分的代表性;二是仅仅应用氨基酸组成(对应第一代算法)来表示蛋白质的一级结构是不充分的,因为氨基酸的排列信息被丢失。现在这两方面的问题即将解决。利用氨基酸的排列顺序来预测结构类的第二代算法已开始发表[19],今后将陆续发表。此外,现在已有一千余种已知结构的非同源蛋白质,代表了上万种

* 中国科学院院士.
本文于 2000 年 6 月 18 日收到.

蛋白质。因此,对这一千余种非同源蛋白质的训练参数能够以较高准确度预测数万种蛋白质的结构类。可以预期:有实用水平的算法和软件可望在几年内提出来,并将上网提供服务。届时对于绝大部分任意给定的蛋白质,可望以较高的准确度来预测其结构类。

蛋白质结构类预测的另一途径就是预测其 2 级结构含量,即预测一个蛋白质中有多少个残基采用 α 螺旋构象和 β 折叠构象。一旦 2 级结构含量预测出来了,就可以根据分类标准定出它的结构类。这种途径的一个副产品就是同时预测出其 2 级结构含量,这个量也是生物学家所感兴趣的。经过努力[20—21],终于使 α 螺旋含量的平均预测误差降低到 0.087;β 折叠含量的平均预测误差降低到 0.081。这是迄今为止国际上同类研究的最好结果。预测软件上因特网公布以来,短短数个月就有主要来自国外的上千人次使用这一软件。

这是一项涉及数学、物理学、计算机科学和分子生物学的高度交叉的跨学科研究,是生物信息学的核心课题之一。这也是一项以较少投入获得较多成果的、适合于我国国情的研究项目。

参 考 文 献

[1] Zhang C T, Zhang R. A new criterion to classify globular proteins based on their secondary structure contents. Bioinformatics, 1998, **14**: 857—865.

[2] Zhang C T, Zhang R. A quadratic discriminant analysis for the protein classification based on the helix/strand content. J. Theor. Biol., 1999, **201**:189—199.

[3] Zhang C T, Zhang R. A new quantitative criterion to distinguish between α/β and α + β proteins (domains). FEBS Lett., 1998, **440**: 153—157.

[4] Zhang C T, Zhang R. S Curve, A graphic representation of protein secondary structure sequence and its applications. Biopolymers, 2000, **53**:539—549.

[5] Zhang C T, Chou K C. An optimization approach to predicting the proteinstructural class from amino acid composition. Protein Science, 1992, **1**:401—408.

[6] Zhou G F, Xu X, Zhang C T. A weighting method for predicting protein structural class from amino acid composition. Eur. J. Biochem., 1992, **210**:747—749.

[7] Chou K C, Zhang C T. A correlation-coefficient method to predicting protein-structural classes from amino acid composition. Eur. J. Biochem., 1992, **207**: 429—433.

[8] Zhang C T, Chou K C. Monte Carlo simulation studies on the prediction of protein folding types from amino acid composition. Biophys. J., 1992, **63**:1 523—1 529.

[9] Chou J J, Zhang C T. A joint prediction of the folding types of 1490 human proteins from their genetic codons. J. Theor. Biol., 1993, **161**:251—262.

[10] Mao B, Chou K C, Zhang C T. Protein folding classes: a geometric interpretation of the amino acid composition of globular proteins. Protein Engineering, 1994, **7**:319—330.

[11] Chou K C, Zhang C T. Predicting protein folding types by distance functions that make allowances for amino acid interactions. J. Biol. Chem., 1994, **269**:22 014—22 020.

[12] Zhang C T, Chou K C. An eigenvalue-eigenvector approach to predicting protein folding types. J. Protein Chem., 1995, **14**:309—326.

[13] Zhang C T, Chou K C, Maggiora G M. Predicting protein structural classes from amino acid composition: application of fuzzy clustering. Protein Engineering, 1995, **8**:425—435.

[14] Maggiora G M, Zhang C T, Chou K C et al. Combining fuzzy clustering and neural networks to predict protein structural classes. In: Devillers J ed. Neural Networks in QSAR and *Drug Design*. *London*: Academic Press, 1996, 255—279.

[15] Zhang C T. Relations of the numbers of protein folds, families and sequences. Protein Engineering, 1997, **10**:757—761.

[16] Chou K C, Liu W M, Zhang C T et al. Prediction and classification of domain structural classes. Proteins: Struc., Func. and Genetics, 1998, **31**: 97—103.

[17] Zhang C T, Zhang R. Skewed distribution of protein secondary structure contents over the conformational triangle. Protein Engineering, 1999, **12**:807—810.

[18] Chou K C, Zhang C T. Prediction of protein structural classes. CRC Critical Review in Biochemistry and Molecular Biology, 1995, **30**: 275—349.

[19] Bu W S, Feng Z P, Zhang Z et al. Prediction of protein (domain) structural classes based on amino acid index. Eur. J. Biochem., 1999, **266**:1 043—1 049.

[20] Zhang C T, Zhang Z D, He Z M. Prediction of the secondary structure content of globular proteins based on structural classes. J. Protein Chem., 1996, **15**:775—786.

[21] Zhang C T, Lin Z S, Zhang Z D et al. Prediction of the helix/strand content of globular proteins based on their primary Sequences. Protein Engineering, 1998, **11**:971—979.

STUDY OF STRUCTURAL CLASSIFICATION AND PREDICTION OF STRUCTURAL CLASS OF PROTEINS

Zhang Chunting

(*Institute of Life Science and Biotechnology*, *Tianjin University*, *Tianjin* 300072)

Key words protein, structural class, prediction, classification, secondary structure content, bioinformatics