

连续小波变换在蛋白质结构预测中的应用*

邱建丁^{1,2}, 梁汝萍^{1,2}, 邹小勇¹, 莫金垣¹

(1. 中山大学化学与化学工程学院, 广东 广州 510275;

2. 萍乡高等专科学校, 江西 萍乡 337055)

摘 要: 将代码为 3pga1 蛋白质的氨基酸序列映射为疏水值序列, 在合适的尺度下, 通过连续小波变换方法可对其非规则二级结构进行预测, 准确率为 83.3%。从 PDBsum 数据库中随机抽取 100 个蛋白质作为测试对象, 经本法处理后, 1 853 个非规则二级结构中有 1 424 个能被准确预测到, 平均预测准确率为 76.8%。结果表明: 该法可较好地预测蛋白质的非规则二级结构, 具有极大的发展前景。

关键词: 连续小波变换; 非规则二级结构; 疏水性

中图分类号: Q71 **文献标识码:** A **文章编号:** 0529-6579 (2003) 02-0059-03

随着人类基因组计划的完成, 生命科学已进入后基因组时代, 人们越来越注重从整体水平上对生物功能的研究。生物功能的主要体现者是蛋白质, 而蛋白质的生物功能很大程度上取决于蛋白质的空间结构, 因此, 蛋白质空间结构的确定是蛋白质科学的重大前沿领域。蛋白质的一级结构 (即蛋白质中氨基酸排列次序) 很大程度上决定其高级结构, 因此根据蛋白质的氨基酸序列提供的信息, 从理论上可以预测其相应的高级结构^[1]。蛋白质分子中氨基酸的疏水性是影响蛋白质结构形成的一个最重要的因素^[2]。不同的氨基酸有着不同的疏水值, 前人^[3,4]研究的工作表明, 蛋白质的氨基酸序列疏水性和它的疏水区域相关, 并且认为疏水值最小值位置的肽链一般是蛋白质的非规则二级结构。非规则二级结构是连接规则二级结构的简单片段, 它在蛋白质折叠、转向和生物功能等方面都起着重要作用^[5,6], 因而预测蛋白质的非规则二级结构具有重要意义。

作为自上世纪 80 年代后期发展起来的信号处理工具—小波变换已广泛应用于各科学领域。小波变换是时间窗和频率窗都可改变的时频局部化分析方法, 即在低频部分具有较高的频率分辨率和较低的时间分辨率, 在高频部分具有较高的时间分辨率和较低的频率分辨率。连续小波变换由于尺度可以连续变化, 比采用离散小波变换提取的信号更显得丰富、全面, 本文采用连续小波变换来预测蛋白质序列的非规则区域。

1 原 理

侧链的不同构成了不同的氨基酸, 从而决定了各种氨基酸有不同的特性, 其中疏水作用是最重要的, 疏水作用很大程度上决定了蛋白质结构的稳定性。而连续小波变换具有“数学显微镜”的功能, 通过拉伸和平移可同时实现时-频的局部化分析, 故可用于揭示疏水性与蛋白质结构之间的关系。但连续小波变换只能处理数字信号, 所以须将蛋白质氨基酸序列转换成数据序列, 即用氨基酸对应的疏水值替换蛋白质中相应的氨基酸, 得到疏水值序列, 此序列代表原有的蛋白质序列, 然后再用连续小波变换处理。表 1 为 20 种氨基酸的疏水值。

表 1 20 种氨基酸的疏水值

Tab.1 The hydrophobic values of 20 amino acids

氨基酸	疏水值	氨基酸	疏水值
A	0.62	M	0.64
C	0.29	S	-0.18
D	-1.05	N	-0.85
E	-0.87	T	-0.05
F	1.19	P	0.12
G	0.48	V	1.08
H	-0.4	Q	-0.78
I	1.38	W	0.81
K	-1.35	R	-1.37
L	1.06	Y	0.26

连续小波变换是一种信号的时间-尺度分析方

* 收稿日期: 2002-08-15

基金项目: 国家自然科学基金资助项目 (29975033); 广东省自然科学基金资助项目 (980340)

作者简介: 邱建丁 (1973 年生), 男, 博士研究生; 通讯联系人: 莫金垣; E-mail: cesmjy@zsu.edu.cn

由表 2 可知, 本法对蛋白质 3pga1 非规则二级结构的预测结果与实际值存在一些偏差, 如连续小波变换位置为 51, 60, 80, 140 的最小值并没有蛋白质非规则二级结构与之对应; 位置为 54, 66, 148-149, 168-170, 200 的 5 个非规则二级结构并没有最小值和它们对应。这是因为虽然疏水性是决定蛋白质折叠的最重要因素, 但并不是唯一因素, 除了疏水性之外, 还有氢键、盐键和范德华力以及肽链内的二硫键, 分子体积和带的电荷等多种因素对蛋白质折叠都起作用, 所以, 我们认为通过把氨基酸序列映射为疏水值序列来预测非规则二级结构, 一定范围内的偏差是允许的。

在作连续小波变换时, 小波尺度 a 对预测的准确率有很大影响。表 3 列出了不同小波尺度对预测准确率的影响。从表 3 可知, 当小波尺度为 9 时, 预测准确率最高。

表 3 小波尺度对预测准确率的影响

Tab.3 The influence of predicted accuracy under the different dilations

小波尺度 a	最小值个数	准确预测个数	准确率/% ¹⁾
6	噪声干扰严重	—	—
8	34	23	76.7
9	30	25	83.3
10	27	21	70.0
11	26	20	66.7
13	24	18	60.0
15	20	13	43.3

1) 准确率 = $\frac{\text{准确预测的个数}}{\text{非规则二级结构总数}} \times 100\%$

为充分验证该方法的可行性, 从 PDBsum 数据库中随机抽取了 100 个蛋白质作为测试对象, 其中全 α 螺旋、全 β 折叠、 α/β 以及 $\alpha + \beta$ 蛋白质各 25 个。在 100 个蛋白质中共有 1 853 个非规则二级结构, 其中 1 424 个能被准确预测到, 平均预测准确率为 76.8%, 比 Hirakawa 等^[8] 利用离散小波变换方法预测的准确率 68.9% 要高。上述结果表明, 连续小波变换可以较好地对蛋白质的非规则二级结构进行预测, 具有很好的发展前景。

参考文献:

- [1] 林治华, 吴玉章. 蛋白质结构预测的方法学评述[J]. 免疫学杂志, 2001, 17(3): 59-62.
- [2] 王克夷. 疏水作用和蛋白质[J]. 生命的化学, 1999, 19(5): 233-235.
- [3] ROSE G D. Prediction of chain turns in globular proteins on a hydrophobic basis[J]. Nature, 1978, 272: 586-591.
- [4] LESZCZYNSKI J F, ROSE G D. Loops in globular proteins: a novel category of secondary structure[J]. Science, 1986, 234: 849-855.
- [5] DILL K A, FIDBIG K M, CHAN H S. Cooperativity in protein-folding kinetics[J]. Proc Natl Acad Sci USA, 1993, 90: 1942-1946.
- [6] CRASTO C J, FENG J. Sequence codes for extended conformation: a neighbor-dependent sequence analysis of loops in proteins[J]. Proteins: Structure, Function, and Genetics, 2001, 42: 299-413.
- [7] 杨福生. 小波变换的工程分析与应用[M]. 北京: 科学出版社, 1999: 1-26.
- [8] HIRAKAWA H, MUTA S, KUHARA S. The hydrophobic cores of proteins predicted by wavelet analysis[J]. Bioinformatics, 1999, 15(2): 141-148.

Prediction of Protein Structure by Continuous Wavelet Transform

QIU Jian-ding^{1,2}, LIANG Ru-ping^{1,2}, ZOU Xiao-yong¹, MO Jin-yuan¹

(1. School of Chemistry and Chemical Engineering, Sun Yat-sen University, Guangzhou 510275, China;

2. Department of Chemical Engineering, Pingxiang College, Pingxiang 337055, China)

Abstract: Non-standard secondary structures can be predicted by using continuous wavelet transform (CWT) under the appropriate dilation after the amino acids of 3pga1 protein are transformed into sequences of hydrophobic values per residue, the predicted accuracy is about 83.3%. One hundred proteins from PDBsum database were chosen randomly as the test objects, 1 424 non-standard secondary structures can be predicted by this method, comparing with that of the sum 1 853, the average accuracy is 76.8%. The result exhibits that CWT is an effective tool to predict the non-standard secondary structures of proteins and has a tremendous development foreground.

Key words: continuous wavelet transform; non-standard secondary structure; hydrophobicity