第 37 卷 第 9 期
Vol.37 No.9

计 算 机 工 程
Computer Engineering

2011 年 5 月
May 2011

·人工智能及识别技术· 文章编号：1000—3428(2011)09—0184—03 文献标识码：A 中图分类号：TP391

基于多级滤波器的蛋白质热点区域预测

刘文远，王东伟，王常武，王宝文

(燕山大学信息科学与工程学院，河北 秦皇岛 066004)

摘 要：针对数字滤波技术用于蛋白质热点区域预测时预测结果精度低的问题，提出一种基于多级滤波器的预测方法。依据共振识别模型从一组蛋白质中提取出特征频率，设计一个选通特性良好的多级滤波器对蛋白质序列进行滤波，根据滤波器输出序列的能量谱定位蛋白质的热点区域。实验表明，多级滤波器能提高预测结果的精度。

关键词：蛋白质；热点区域；数字信号处理；多级滤波器；共振识别模型

Hot Spots Prediction in Proteins Based on Multi-stage Filter

LIU Wen-yuan, WANG Dong-wei, WANG Chang-wu, WANG Bao-wen

(College of Information Science and Engineering, Yanshan University, Qinhuangdao 066004, China)

【Abstract】Aiming at the low accuracy of digital filter to identify the locations of hot spots in proteins, this paper presents a new technique of using multi-stage filter. The technique selects characteristic frequency from a set of proteins based on resonant recognition model, and designs a special multi-stage filter to process the protein sequence, and then the energy spectrum of the filter output yields the locations of the hot spots in proteins. Experiments show that this filter has higher accuracy.

【Key words】protein; hot spots; digital signal process; multi-stage filter; resonant recognition model

DOI: 10.3969/j.issn.1000-3428.2011.09.064

1 概述

蛋白质是生命的物质基础。组成蛋白质的氨基酸共有 20 种。蛋白质分子中存在着多种相互作用力，如氢键的相互作用力、离子间相互作用力等。由于这些作用力的影响，蛋白质分子产生了一种折叠现象，形成了非常复杂的 3-D 结构[1]。蛋白质通过这种折叠结构与其他分子绑定到一起，这就决定了它与其他分子之间存在着选择性的相互作用。通常，与蛋白质发生选择性相互作用的目标分子是另一种蛋白质或者其他类型的分子，如核酸等。在蛋白质的分子结构中，通过绑定其他分子发生选择性相互作用的区域被称为活性位点。蛋白质间的选择性相互作用是由活性位点中的氨基酸功能团决定的，这些氨基酸功能团通常被称为热点区域(hot spots)。

准确地预测蛋白质的热点区域能够进一步认识蛋白质间产生选择性相互作用的机制，对生物学家进行实验研究具有重要的意义。蛋白质热点区域预测在药物发现领域应用广泛。人们已成功地开发出类似药物的小分子，这些小分子能够在热点区域形成复杂的结构，约束和抑制蛋白质间的相互作用。热点区域预测被认为是药物设计过程中的第一步，因此需要一种精确的方法预测蛋白质的热点区域。

蛋白质热点区域预测有实验方法和计算方法 2 种。实验方法的周期长，费用昂贵。近年来，一些计算方法已应用于蛋白质热点区域的预测。多数计算方法是通过替换蛋白质中的丙氨酸，估算因替换引起的自由能的变化进行预测。它们的时间复杂度非常高，难以适用于较大规模的预测。最近，数字信号处理技术也被应用于蛋白质热点区域预测。文献[2]首先提出了使用傅里叶变换(DFT)的方法，其缺陷是当对频率域的某一个频率做改动时，进行傅里叶反变换后，整个原序列都会发生变化，会影响输入序列的特性，所以该方法不可靠。文献[3]提出了使用加窗傅里叶变换(STDFT)的方法，尽管 STDFT 能够预测出部分热点区域，但是这种方法的复杂度太高。为了提高算法的运行效率，降低算法的时间复杂度，文献[4]又提出了数字滤波器的方法，该方法能够达到和 STDFT 一样的精度，但是算法复杂度明显低于 STDFT。

数字滤波器虽然能够预测出部分热点区域，但是这种单级滤波器的选通特性较差，预测结果的精度较低，仅能够预测出特定热点区域的大概位置，因此可以进一步设计滤波器的选通特性，提高预测的准确性。

本文设计的多级滤波器允许含有特征频率的信号成分通过，并能抑制其他频率的信号成分，具有较好的频率选择性。

2 共振识别模型

2.1 蛋白质序列的数值表示

蛋白质由氨基酸组成，氨基酸由特定的字符来表示。在进行信号处理之前，必须将字符序列映射成数值序列，映射方法参见文献[5]。

生物学家发现蛋白质间的相互作用是由于蛋白质分子中自由能的周期性分布引起，表示氨基酸中自由能的方法是 EIIP(Electron-Ion Interactive Potential)值，一条蛋白质序列能够通过一系列的 EIIP 值来表示。

例如：W_n=KVFGRCEL，该序列是血色素蛋白中的一段氨基酸序列，映射成数值序列为：

X_n=[0.037 1, 0.005 7, 0.094 6, 0.005 0, 0.095 9, 0.082 9, 0.005 8, 0.000 0]

基金项目：河北省教育厅自然科学研究计划基金资助项目(2009339)

作者简介：刘文远(1968－)，男，教授、博士，主研方向：生物信息学，电子商务；王东伟，硕士；王常武，教授、博士；王宝文，副教授

收稿日期：2010-10-25 **E-mail：**wdwgr@163.com

2.2　共振识别模型的原理

Cosic 发现一组蛋白质的氨基酸序列经过傅里叶变换后，这组蛋白质中具有相同生物功能的氨基酸功能团在傅里叶频谱中有共同的频率，这个共同的频率被称为特征频率。文献[5]根据特征频率的概念提出了一种模型，称为共振识别模型(Resonant Recognition Model, RRM)。

RRM 给出了一个计算特征频率的方法，假设 m 条蛋白质中都具有某种功能团，那么这个功能团的特征频率可以通过计算 m 条蛋白质序列的互谱函数来测定:

$$S(e^{j\omega}) = \left| X_1(e^{j\omega}) X_2(e^{j\omega}) \mathrm{L}\ X_m(e^{j\omega}) \right|$$

其中，$X_1,X_2,\cdots,X_m$ 是 m 条蛋白质表示成数值序列后对应的傅里叶变换。$S(e^{j\omega})$是这个功能团的频谱，这个频谱中存在的唯一一个波峰就是特征频率。图 1 中峰值对应的频率表示的是细胞色素 C 的特征频率。

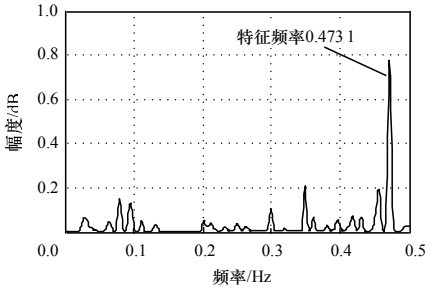
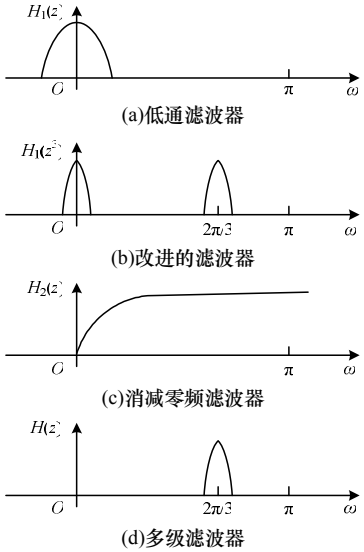


图 1　细胞色素 C 的特征频率

3　蛋白质热点区域的预测

3.1　多级滤波器的设计方法

单级滤波器需要根据数据特性选择滤波器类型。一旦类型确定，就只能通过调整参数对滤波器的选通特性进行优化，这具有一定的局限性。在预测蛋白质的热点区域问题中，反切比雪夫滤波器是几种单级滤波器中效果最好的，但预测结果精度仍不够高。当确定了滤波器的类型后，以多级级联的方式改善滤波器的选通特性，使滤波器对于特征频率具有更强的针对性。首先考虑一个窄带低通滤波器 $H_1(z)$如图 2(a)所示，如果将低通滤波器的的每一个延迟因子 z^{-1} 替换为 z^{-3}，就能够得到替换后的滤波器 $H_1(z^3)$。这样得到的滤波器具有非常好的窄带选通特性，它的频率响应如图 2(b)所示。



(a)低通滤波器

(b)改进的滤波器

(c)消减零频滤波器

(d)多级滤波器

图 2　多级滤波器的设计方法

这个滤波器包括 2 个通带中心，分别在 ω=0 和 ω=2π/3 处，然后在它后面级联一个滤波器 $H_2(z)$。$H_2(z)$能够很好地削减零频率处的通带，从而得到滤波器 $H(z)$:

$$H(z) = H_1(z^3) H_2(z)$$

$H(z)$是一个窄带带通滤波器，它的带通中心在 2π/3 处。由此得到的滤波器 $H(z)$，其阻带衰减明显变得更快，并且具有更好的窄带选通特性，然后通过对该滤波器进行频移可以将滤波器的带通中心移动到特征频率处。

3.2　多级滤波器的实现

椭圆滤波器的幅频响应在通带和阻带内都是等波纹的，对于给定的阶数和给定的波纹要求，它能够获得较其他滤波器更窄的过渡带宽，就这点而言椭圆滤波器是最优的。椭圆滤波器具有阶数低、算法的执行效率高等优点。根据 3.1 节中的方法选择椭圆低通滤波器作为 $H_1(z)$。

设计多级滤波器的步骤如下:

(1)选择椭圆低通滤波器 $H_1(z)$的参数：最大通带衰减为 1 dB，最小阻带衰减为 50 dB，截止频率为 0.1，滤波器的阶数设置为 3 阶就能够达到比较理想的结果，如图 3 所示。

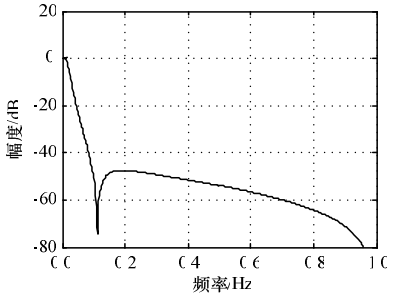


图 3　椭圆低通滤波器 $H_1(z)$

(2)将椭圆滤波器中每一个延迟因子 z^{-1} 替换为 z^{-3}，得到替换后的滤波器 $H_1(z^3)$如图 4 所示。

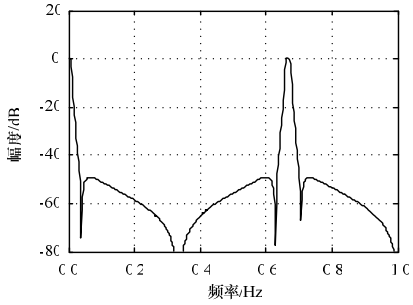


图 4　改进的椭圆滤波器 $H_1(z^3)$

(3)设计滤波器 $H_2(z)$，它在 ω=0 处存在 2 个零点，因此它能够很好地削减 $H_1(z^3)$在零频率处的通带，如图 5 所示。

$$H_2(z) = (1 - z^{-1})$$

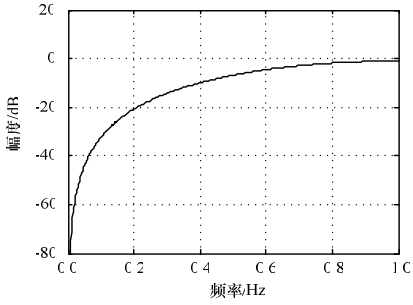


图 5　消减零频率滤波器 $H_2(z)$

(4)将以上 2 个滤波器级联，得到多级滤波器 $H(z)$，该滤

波器的带通中心在 2π/3 处，如图 6 所示。

$$H(z)=H_1(z^3)H_2(z)$$

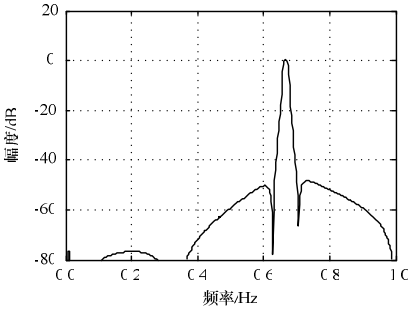


图 6 多级滤波器 $H(z)$

(5)因为 $H(z)$的带通中心在 2π/3 频率处，要实现对特征频率的提取，就必须对 $H(z)$进行频移，把滤波器的带通中心移至前文得到的特征频率处，设频移量为 θ，经过频移后的滤波器为 $H'(z)$，图 7 为平移后滤波器的频率响应，其带通中心为 0.473 1。

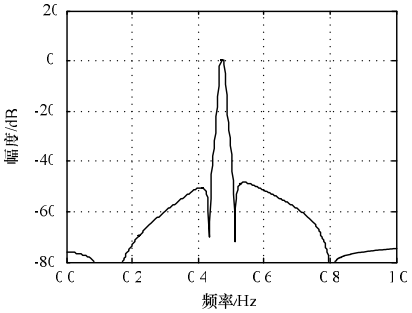


图 7 频移后的多级滤波器 $H'(z)$

由 $H(z)=H(\mathrm{e}^{\mathrm{j}\omega})$，得:

$$H'(z)=H(\mathrm{e}^{\mathrm{j}(\omega+\theta)})$$

将输入序列经过 $H'(z)$后，就能够提取出特征频率的成分。可以根据输出序列的能量谱 $E(y[n])$识别出蛋白质热点区域的位置。如果输出序列表示为 $y[n]$，那么这条序列的能量表示为:

$$E(y[n])=(y[n])^2$$

$E(y[n])$作为特定蛋白质的能量谱，可以根据能量谱中的峰值定位蛋白质的热点区域。能量谱中比较明显的峰值就对应着一个热点区域。

4 仿真实验

数据来自 Uniprot 数据库。实验中选择了 cytochrome C 和 lysozyme 2 组蛋白质(表 1)，分别求出这 2 组蛋白质的特征频率，然后对 Tuna cytochrome C(金枪鱼细胞色素 C)和 Hen egg-white lysozyme(鸡蛋白溶解酶)2 个蛋白质进行了热点区域的预测。例如使用 Human cytochrome C、Bovine cytochrome C、Rabbit cytochrome C 3 个蛋白质求出 cytochrome C 这一族蛋白质的特征频率，根据特征频率对 Tuna cytochrome C 进行了热点区域的预测。

表 1 实验数据

cytochrome C	lysozyme
Tuna cytochrome C	Hen egg-white lysozyme
Human cytochrome C	Human lysozyme
Bovine cytochrome C	Bovine lysozyme
Rabbit cytochrome C	Rabbit lysozyme

图 8、图 9 说明了本文方法对 Tuna cytochrome C 和 Hen egg-white lysozyme 2 个蛋白质的热点区域预测的结果。图中的峰值说明该处存在一个热点区域，图中标注的横坐标表示热点区域在氨基酸序列中的位置。

实验中作为评估热点区域位置准确性的标准数据来自 ASEdb 数据库。表 2 对本文的预测结果与文献[4]的预测结果进行了比较。

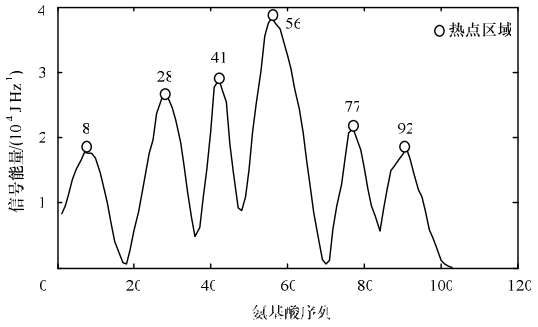


图 8 Tuna cytochrome C 的热点区域

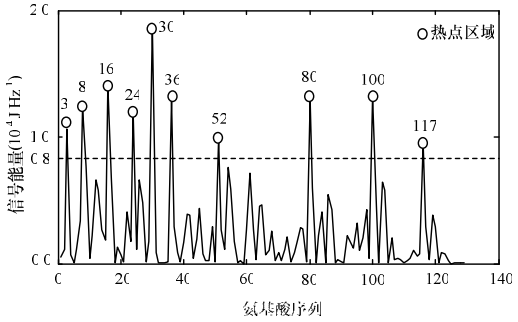


图 9 Hen egg-white lysozyme 的热点区域

表 2 实验结果

序号	蛋白质名称	特征频率	热点区域位置		
			多级滤波器	反切比雪夫滤波器	ASEdb
1	tuna cytochrome C	0.473 1	8, 28, 41, 56, 77, 92	8, 25, 43, 61,78	41, 45, 56, 77
2	hen egg-white lysozyme	0.023 5	3, 8, 16, 24, 30, 36, 52, 80, 100, 117	3, 6, 12, 18, 21, 43, 48, 52, 61, 73, 101, 119	3, 4, 16, 22, 24, 30, 36, 117

从表 2 可见，对于 Tuna cytochrome C 蛋白质多级滤波器能够非常准确地预测出 41、56、77 这 3 个热点区域，而反切比雪夫滤波器仅能预测出这 3 个热点区域的大概位置。另外多级滤波器还预测出 8、28、92 这 3 个潜在的热点区域，潜在的热点区域具有非常重要的生物学意义，可以供生物学家进行更深入的研究。对于 Hen egg-white lysozyme，多级滤波器能够准确地预测出 3、16、24、30、36、117 这 6 个热点区域。而反切比雪夫滤波器的预测结果不够精确，没有预测出 30、36 这 2 个热点区域。使用多级滤波器不但能够准确地预测出这 2 个热点区域，同时也预测出了几个潜在的热点区域 52、80、100。可见多级滤波器比反切比雪夫滤波器的预测结果更加精确。

5 结束语

本文针对单级滤波器用于蛋白质热点区域预测存在的不足，提出了一种基于多级滤波器的预测方法。文中选择 Uniprot 数据库中的 2 组蛋白质进行实验分析，证明了多级滤波器用于蛋白质热点区域预测具有更好的窄带选通特性，较单级滤波器提高了预测结果的精度。由于蛋白质结构的复杂性，使得对蛋白质热点区域的预测比较困难，要获得更加精确的结果需要做进一步深入的研究。

(下转第 189 页)