

文章编号:1004-9037(2013)05-0643-07

AAindex 数据库中傅里叶方法的多重氨基酸指数选择

赵 剑^{1,2} 许金涛¹

(1. 南京工业大学理学院, 南京, 210009; 2. 南京工业大学生物与制药工程学院, 南京, 210009)

摘要: 波谱学方法研究蛋白质时, 蛋白质序列的特征频率与蛋白质的位点、功能、空间结构对应。基于时间序列傅里叶变换的交叉谱原理, 本文提出的“一个蛋白质关于多种氨基酸指数的一致谱”能够获取一个或多个蛋白质序列较显著的特征频率, 再由特征频率去筛选与之对应的氨基酸指数。该方法与 Lazović 的筛选方法和筛选意义不同。结果显示, 筛选出的氨基酸指数对应的性质多数与蛋白质的空间结构性质有关。蛋白质特征频率也许是用于特征提取的好方法, 氨基酸指数筛选可被看作蛋白质数字表达的一种优化。

关键词: 氨基酸指数数据库; 傅里叶变换; 特征频率; 氨基酸指数筛选; 信噪比

中图分类号: R857.3

文献标志码: A

Selection of Multi Amino Acid Indices Based on Fourier Transform Analysis in AAindex Database

Zhao Jian^{1,2}, Xu Jintao¹

(1. School of Science, Nanjing University of Technology, Nanjing, 210009, China;

2. Biotechnology and Pharmaceutical Engineering, Nanjing University of Technology, Nanjing, 210009, China)

Abstract: The characteristic frequencies of protein sequences are correlated to protein-target binding sites, function, and structure when the spectroscopy method is applied to protein study. This paper proposes “the consensus spectrum of one protein sequence obtained using a variety of amino acid indices”. It can be used to find significant characteristic frequencies of one or more protein sequences, and then, to select those amino acid indices which refer to the characteristic frequencies. The method and significance of this paper are different from those of Lazović’s paper. The results show that the screened amino acid indices are mostly related to the properties of space structure. The characteristic frequencies derived from its amino acid sequence are probably a good way of feature extraction. Meanwhile, the selection of amino acid index can be regarded as a kind of optimization of protein’s numerical representation.

Key words: amino acid index database; Fourier transform; characteristic frequency; selection of amino acid indices; signal to noise ratio

引 言

AAindex 是氨基酸指数和氨基酸突变矩阵的一个数据库^[1], 该数据库由两部分组成: AAindex1 和 AAindex2。AAindex1 是发表的氨基酸指数的搜集, AAindex2 是发表的氨基酸突变矩阵的搜集。一个氨基酸指数是代表氨基酸的各种物化和生化性质的一套 20 个数值。目前, AAindex 共搜

集了 544 种氨基酸指数。

在蛋白质和蛋白质组的研究中, 蛋白质序列是由 20 个不同氨基酸构成的符号序列, 在使用数值方法研究蛋白质时, 首先需要将符号序列转化为数字序列, 该过程通常称作蛋白质数字表达或者蛋白质数字编码。

通常, 与 AAindex 数据库相关的数字代替方法有以下几类: (1) 从 AAindex 数据库中选择一个或多个氨基酸指数, 将氨基酸序列映射为一维数字

序列或多维向量序列^[2]。这种方法较直接,但常常因为蛋白质长度的差异而导致很多方法无法使用;而且选取氨基酸指数主观性较强。(2)抽取 AAindex 数据库中的全部或部分氨基酸指数,利用主成分分析或因于分析等维数约减方法^[2-9],从大量氨基酸指数中提取其包含的主要信息,得到能够全面反映蛋白质特性的高维数字表达,该类表达,共性太强,反映蛋白质个性方面不足。(3)根据文献或生物学知识,首先从 AAindex 数据库中选取一些属性做数字代替提取氨基酸序列的多种信息,这些信息往往与生物学意义相对应,诸如蛋白质序列的长度,氨基酸在序列当中出现的频率,氨基酸序列的自相关性或者互相关性、位置特异性得分矩阵(Position-specific scoring matrix, PSSM)等,生成与蛋白质序列对应的特征向量或特征信息度量^[10-13]。该类方法有利于使用各种统计学习方法进行处理。

数字表达序列往往包含所要研究蛋白质的特征信息,然后使用统计、信号处理、机器学习等方法处理该数字序列,挖掘蛋白质的各种生物特性。目前该类方法取得了非常多的理论成果,主要有以下几方面^[3]:蛋白质的相似性^[14,15]、蛋白质的二级结构的分类预测^[16]、蛋白质跨膜结构的分类^[17,18]、蛋

白质界面的残基预测^[19]、蛋白质当中的结合位点^[20]、蛋白质中无序区预测^[21]、蛋白质的亚细胞定位^[22]、蛋白质与蛋白质的结合位点、蛋白质与 RNA 和 DNA 相互作用的结合位点识别^[11,12]等。这些研究大量使用了 AAindex 数据库中的氨基酸指数。

不同的蛋白质,其各方面的生物属性不尽相同。无论是一个蛋白质或是一组蛋白质,从 AAindex 中选择一个或多个恰当的氨基酸指数,准确细致地反映蛋白质的性质,尤其是反映蛋白质个性化的性质,这方面的工作较少,但仍是研究蛋白质的一个重要方面。

目前已有的方法中,“伪氨基酸表达”成功地解决了大量的问题,本文将介绍使用傅里叶方法,从 AAindex 数据库中选择氨基酸指数的方法。该方法关键是选择那些能够生成具有较大信噪比特征频率的氨基酸指数,Cosic^[14]对蛋白质组使用 EIIP 做数字表达,发现不同的蛋白质组对应不同的特征频率。具有相同特征频率的蛋白质在功能上具有相似性,Pirogova^[23]对多组蛋白质使用共鸣方法,分别使用 AAindex 中不同的氨基酸指数做数字表达,结果如表 1 所示,表中 Freq 表示特征频率,S/N 表示特征频率处的信噪比。

表 1 选取的氨基酸指数对蛋白质组处理后获取的特征频率和对应的信噪比

参数	EIIP		IC		P001		H085		H371	
	Freq	S/N								
Glucagon	0.087 9	122.7	0.085 9	94.4	0.136 4	166.3	0.152 3	140.1	0.252 0	95.7
Hemoglobin	0.025 4	256.0	0.025 4	256.0	0.136 7	256.0	0.410 2	249.1	0.392 6	256.0
Lysozyme	0.328 1	238.3	0.328 1	249.2	0.470 7	120.5	0.074 2	167.3	0.328 1	217.8
Myoglobin	0.253 9	255.4	0.013 7	191.3	0.025 4	255.2	0.127 0	256.0	0.257 8	256.0
CytochromeC	0.474 6	247.4	0.474 6	208.3	0.474 6	255.5	0.421 9	232.3	0.201 2	150.3
EGF	0.060 5	245.7	0.060 5	255.8	0.062 5	136.8	0.068 4	171.2	0.066 4	143.2

通过表 1 的计算结果不难发现:

(1)对同一蛋白质组,不同氨基酸指数有可能出现相同的特征频率,这说明利用这些氨基酸指数使用共鸣识别方法判别蛋白质相似性结果是等效的,蛋白质关于这些氨基酸指数对应的生物性质具有相似性,可以认为这些氨基酸指数是等效的,如表 1 中 EGF 蛋白组。

(2)对同一蛋白质组,不同氨基酸指数有可能出现不相同的特征频率,这说明利用这些氨基酸指数使用共鸣识别方法判别蛋白质相似性时是有差异的。蛋白质关于这些氨基酸指数对应的生物性质有差异,从而认为这些氨基酸指数体现了蛋白质

之间的差异性。这些差异性的强弱可以通过信噪比来体现,如表 1 中 Myoglobin 蛋白组。

(3)对不同的蛋白质组,分别使用不同的氨基酸指数对不同蛋白质组作共鸣识别,可能出现相同的特征频率,这说明蛋白质关于这些不同的氨基酸指数对应的生物性质是相似的。如表 1 中的 Glucagon 蛋白组, Hemoglobin 蛋白组, Lysozyme 蛋白组,以及 CytochromeC 蛋白组。

蛋白质关于功能、位点、结构等方面的性质,与氨基酸指数做数字表达以后的傅里叶变换谱的特征频率对应^[2,14,23],若蛋白质的不同数字表达对应的傅里叶谱的特征频率相似度较高,则认为蛋白质

之间关于功能、位点、结构等性质较为相似,也即蛋白质之间相似度较高。从而选择能够体现蛋白质相似性和差异性比较明显的多个氨基酸指数研究蛋白质,在频率域上提取蛋白质的特征频率研究蛋白质,符合生物学意义。不同的蛋白质,其功能、位点、结构性质都是有差异的,研究不同的蛋白质就需要选择恰当的氨基酸指数。

Lazović^[24]使用共鸣方法,分别针对不同的蛋白质组,求“同一蛋白质组内的多个蛋白质关于某种氨基酸指数的一致谱”。因为同一蛋白质组内的蛋白具有相似的功能、位点、结构等信息,同一蛋白质组内的蛋白质具有相同的特征频率。从当时发表的 226 个氨基酸指数中筛选出了能够反映生成特征频率的氨基酸指数,文中最后的结论是:在反映蛋白质的相似性方面,使用 EIIP 表达与其他的表达效果是一样的,尽管选用 EIIP 与其他表达生成的特征频率可能不相同,但是在能判断蛋白质的相似性方面,功能是形同的,作者认为 EIIP 是对蛋白质一维数字表达最好的表达,故作者使用 EIIP 对蛋白质做数字表达,研究蛋白质。

本文的研究工作是基于傅里叶变换的 AAindex 中氨基酸指数的选取,但是方法和目的上与 Lazović^[24]不同。在时间序列傅里叶变换^[25,26]中的交叉谱理论基础上,提出“一个蛋白质关于多种氨基酸指数的一致谱”的方法,使用该方法能够计算得到任意一个蛋白质信号特征比较显著的特征频率。本文的研究目的是针对一个蛋白质或一个蛋白质组,或不同组的多个蛋白质,从 544 个氨基酸指数中选择能够生成比较显著“特征频率”的氨基酸指数,利用这些氨基酸指数直接做数字替代。今后,可使用筛选出来的这些氨基酸指数,结合蛋白质的其他生物学信息,对蛋白质做特征提取。

本文在 544 个氨基酸指数模糊聚类分成 7 类的基础上,对每一类氨基酸指数使用时间序列傅里叶变换理论^[25]中的交叉谱理论,建立了蛋白质关于多种氨基酸指数的一致谱,用于查找共有特征频率,然后再进一步查找在特征频率处能够产生较大信噪比的氨基酸指数。目的是利用这些指数,恰当地反映蛋白质的属性,使用这些指数对蛋白质做数字表达,为提取蛋白质序列特征的进一步研究做准备。

1 AAindex 数据库中选择氨基酸指数的方法

在 AAindex 数据库中选择氨基酸指数的方法的具体步骤如图 1 所示。

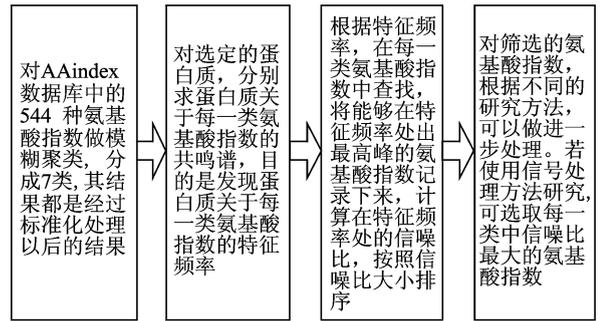


图 1 AAindex 数据库中选择氨基酸指数的流程图

详细解释图 1 中各步骤如下:

(1)Saha 等^[27]使用模糊聚类的方法将 544 种氨基酸指数分成了 7 类。本文使用该结果。

(2)对选定的蛋白质,分别求蛋白质关于每一类氨基酸指数的共鸣谱,得到特征频率。

该过程分两步,首先解决一个蛋白质的氨基酸指数的选择方法。对于多个蛋白质的情形,只需重复一个蛋白质的氨基酸指数的选择方法,将单个蛋白质筛查得到的所有氨基酸指数放到一起,进行二次筛选。如果是同一蛋白质组,其中的氨基酸指数重复出现的程度较高,可选择出现频次高的氨基酸指数。

544 个氨基酸指数可以聚成 7 类,分别就每一类当中的氨基酸指数求其交叉共鸣谱,这样可以得到关于该类性质的显著的共有频率。

假设氨基酸序列为 S ,第 M_i 类 AAindex 当中有 n_i 种氨基酸指数,则氨基酸序列 S 使用第 M_i 类氨基酸指数中第 j 种氨基酸指数所做的傅里叶变换为

$$S_{M_{ij}}(k) = \sum_{n=0}^{m-1} s_{ij}(n) e^{-\frac{i2\pi kn}{m}} \quad (1)$$

式中: m 为氨基酸序列的长度($i=1,2,\dots,7;j=1,2,\dots,n_i;k=0,1,\dots,m-1$); $\{s_{ij}(n)\}$ 表示氨基酸序列 S 使用第 M_i 类氨基酸指数中第 j 种氨基酸指数所做的数字表达序列。 $\{S_{M_{ij}}(k)\}$ 为 $\{s_{ij}(n)\}$ 对应的傅里叶变换。

氨基酸序列 S 关于 AAindex 中第 M_i 类的 n_i 种氨基酸指数的“一致谱”的 $C_{M_i}(k)$ 为

$$C_{M_i}(k) = \prod_{j=1}^{n_i} |S_{M_{ij}}(k)| \quad k=0,1,\dots,m-1 \quad (2)$$

式(2)的作用是求出一个蛋白质序列关于第 M_i 类的 n_i 种多种氨基酸指数的共有特征频率。

(3)根据步骤(2)中计算得到的特征频率,在每

一类氨基酸指数中查找能够在特征频率处出现最高峰的氨基酸指数。能够在特征频率处出现最高峰,表明使用该种氨基酸指数做数字表达能够较好地体现该蛋白质的某种性质。

记录能够在特征频率处出现最高峰的氨基酸指数,大大缩小了选择氨基酸指数的范围。这些氨基酸指数可以反映蛋白质多方面的属性。

对于使用信号处理方法研究蛋白质,目的是要得到较强的信号特征,而且该信号特征又与蛋白质的某种属性相对应,从这个角度可以选择每一类中在特征频率处出现最高峰且信噪比最大的氨基酸指数。

信噪比的定义为

$$\text{SNR}(k) = \frac{|S_{M_j}(k)|}{E} \quad (3)$$

式中: $\text{SNR}(k)$ 表示氨基酸序列 S 使用第 M_i 类氨基酸指数中第 j 种氨基酸指数对应的数字表达序列的傅里叶变换在频率 k 处的信噪比; E 为平均功率谱

$$E = \frac{\sum_{k=0}^{m-1} |S_{M_j}(k)|}{m}$$

式中: m 为氨基酸序列的长度。

(4)下一步进行多个蛋白质的氨基酸指数的选取。对于多个蛋白质,可以是同一蛋白质组中的,也可以是不同蛋白质组中的,如果蛋白质组中的数量较大,可以使用其中的部分蛋白质做训练,利用以上各步骤的方法生成氨基酸指数的候选集,再在此候选集内按照一定的规则筛选。最简单的筛选方法可考虑在此候选集内选择出现频数高的氨基酸指数,出现频数高,表明氨基酸指数对应的属性

是该组蛋白质共有的属性。

2 实验数据及结果

本文选择使用两组蛋白质:肌红蛋白(Myoglobin)和由不同组蛋白质构成的蛋白质组,用于筛选氨基酸指数。

第一组蛋白质:肌红蛋白(Myoglobin)

P02192|MYG_BOVIN;P02189|MYG_PIG;P02144|MYG_HUMAN;P68082|MYG_HORSE;P02185|MYG_PHYMC;P68080|MYG_PHOVI;P04247|MYG_MOUSE;P56208|MYG_CARCR;

P02197|MYG_CHICK;P68276|MYG_DELETE;P02186|MYG_ELEM;G1NJB6|MYG_MELGA;

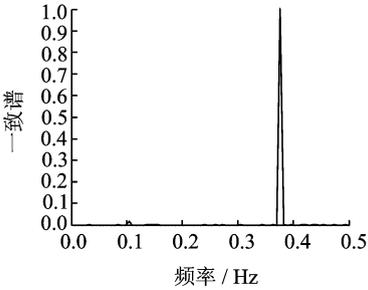
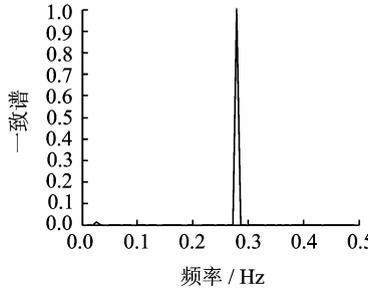
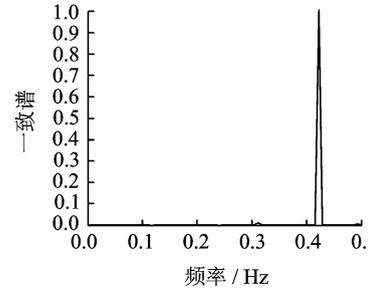
P02152|MYG_CALJA;P02154|MYG_LAGLA;P02155|MYG_SAISC;P02190|MYG_SHEEP;

P02199|MYG_APTFO;P11343|MYG_LUTLU;P02173|MYG_ORCOR;P68278|MYG_PHOPH;

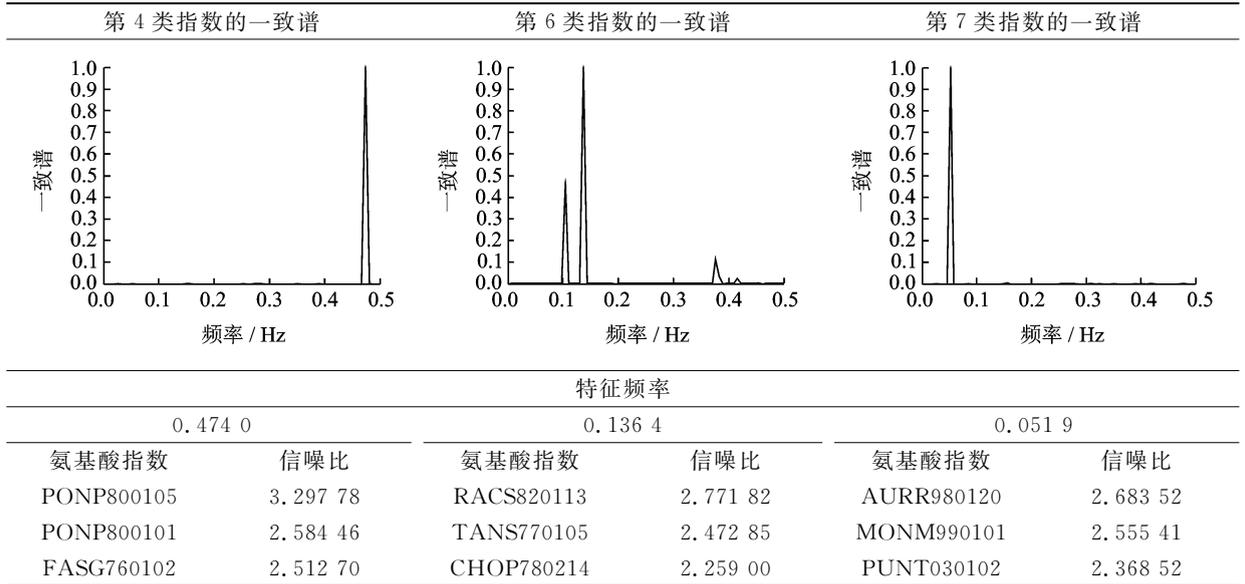
P63114|MYG_OTOME;P02145|MYG_PANTR;P68084|MYG_PAPAN;

首先取一个蛋白质做氨基酸指数筛选,不妨取P02192|MYG_BOVIN-Myoglobin,得到6个主要特征频率,能够在特征频率处出现最高峰的氨基酸指数共计73个,限于篇幅,每个特征频率对应的氨基酸指数列出3个在表2中。表2中的图形为由式(2)计算的“一个蛋白质关于多种氨基酸指数的一致谱”图形。图形经过归一化处理。

表 2 P02192|MYG_BOVIN-Myoglobin 的特征频率及筛选出的与之对应的氨基酸指数

第 1 类指数的一致谱		第 2 类指数的一致谱		第 3,5 类属性的一致谱	
					
特征频率		特征频率		特征频率	
0.376 6		0.279 2		0.422 1	
氨基酸指数	信噪比	氨基酸指数	信噪比	氨基酸指数	信噪比
NAGK730103	2.501 04	PALJ810112	2.562 69	CHAM830108	2.886 06
PALJ810115	2.458 98	GEIM800105	2.490 08	DAYM780101	2.845 56
TANS770101	2.371 91	LIFS790101	2.375 593	JUKT750101	2.715 59

续 表



若使用信号方法研究蛋白质,可以选用表 3 中的 6 个氨基酸指数研究蛋白质,因为这 6 个氨基酸指数在 6 个特征频率处的信噪比最大。

表 3 P02192|MYG_BOVIN-Myoglobin 的特征频率及与之对应的信噪比最大的氨基酸指数

特征频率	所选属性	信噪比
0.051 9	AURR980120	2.368 52
0.136 4	RACS820113	2.77182
0.279 2	PALJ810112	2.562 69
0.376 6	NAGK730103	2.501 04
0.422 1	CHAM830108	2.886 06
0.474 0	PONP800105	3.297 78

对于肌红蛋白组,重复以上步骤,针对所有筛选出的氨基酸指数,统计不同的氨基酸指数和各自出现的频率,按出现频数从高到低排,列出出现频数最高的 5 个氨基酸指数: PALJ810112, RACS820113, JUNJ780101, FASG760101, PONP800105,它们对应的物理化学、生物化学性质分别为: α/β 类中 β 折叠的归一化频率; $\theta(i)$ 的值;序列频率;分子量; β 折叠附件的疏水性。

第二组蛋白质:从 SCOP 数据库(<http://www.ncbi.nlm.nih.gov/COG/grace/shokog.cgi?KOG0117>)中选取,具体如下:

HsM5453806, Hs20357572, Hs21450675, 7300657, Hs20357578, CE250177300657, Hs20357575, Hs5031755, Hs14771929, Hs20357578。

此处按出现频数从高到底,列出出现频数最高的 5 个氨基酸指数:BUNA790101,OOBM850102,

OOBM850101, NAKH920106, LEVM760103。它们对应的物理化学、生物化学性质分别为: α -NH 化学转换;形成反转的优化倾向;优化 β 结构成圈状平衡常数。

3 结束语

本文提出了“蛋白质关于多种氨基酸指数一致谱”的概念,该方法可以用于发现蛋白质的特征频率,并建立了针对一个或多个蛋白质,用于筛选 AAindex 中具有较好信号特征的氨基酸指数的方法。从实验结果可知,不同的蛋白质所筛选出来的氨基酸指数往往不同,而且多数指数反映了蛋白质的空间结构的性质。使用能够产生比较明显的信号特征的氨基酸指数研究蛋白质,对于使用诸如傅里叶变换,或小波变换等波谱学方法研究蛋白质是有益的。

本文的工作还解决了“高维共鸣识别”中氨基酸指数的选择问题^[28],使用波谱学方法研究蛋白质时,蛋白质的结合位点、空间结构、功能等在波谱学中都会有所反映^[2,26,29],蛋白质的傅里叶变换谱有时还会呈现出一定的周期性^[30,31],序列中某些片段的周期性的改变往往预示着片段上生物属性的改变。这些研究很多情况下都与氨基酸的数字表达有关,都面临 AAindex 中氨基酸指数选择的问题。

目前,机器学习方法在生物信息学中被广为使用,其中,如何生成好的特征是一个非常重要的环节。利用本文的方法,可以通过筛选氨基酸指数,

对蛋白质在频率域上提取具有生物学意义的特征频谱,建立特征向量,或者作为特征向量的部分分量,增加特征向量的信息含量,提高性能。氨基酸指数的筛选,也是蛋白质数字代替的一种优化手段。

致谢 衷心感谢南京大学数学系王嘉松教授在撰写本文过程中给予的帮助。

参考文献:

- [1] Kawashima S, Pokarowski P, Pokarowska M, et al. AAindex: amino acid index database, progress report 2008[J]. *Nucleic Acids Research*, 2008, 36(S1): D202-D205.
- [2] Norbert N, Huseyin S. Digital signal processing techniques: Calculating biological functionalities[J]. *Journal of Proteomics and Bioinformatics*, 2011(4): 260-268.
- [3] Nanni L, Lumini A. A new encoding technique for peptide classification [J]. *Expert Systems with Applications*, 2011, 38(4): 3185-3191.
- [4] Venkatarajan M S, Braun W. New quantitative descriptors of amino acids based on multidimensional scaling of a large number of physical-chemical properties[J]. *Molecular Modeling Annual*, 2001, 7(12): 445-453.
- [5] Atchley W R, Zhao J, Fernandes A D. Solving the protein sequence metric problem[J]. *Proceedings of the National Academy of Science of the United States of America*, 2005, 102(18): 6395-6400.
- [6] Rockovsky S, Scheraga H A. On the information content of protein sequences [J]. *Journal of Biomolecular Structure and Dynamics*, 2011, 28(4): 593-594.
- [7] Mizuguchi K, Deane C M, Blundell T L, et al. JOY protein sequence structure representation and analysis [J]. *Bioinformatics*, 1998, 14(7): 617-623.
- [8] Rockovsky S. Global characteristics of protein sequences and their implications[J]. *Proceedings of the National Academy of Sciences*, 2010, 107(19): 8623-8626.
- [9] Leary R H, Rosen J B, Jambeck P. An optimal structure-discriminative amino acid index for protein fold recognition [J]. *Biophysical Journal*, 2004, 86(1): 411-419.
- [10] Chou K C. Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology[J]. *Current Proteomics*, 2009, 6(4): 262-274.
- [11] Ma X, Guo J, Wu J, et al. Prediction of RNA-binding residues in proteins from primary sequence using an enriched random forest model with a novel hybrid feature [J]. *Proteins: Structure, Function, and Bioinformatics*, 2011, 79(4): 1230-1239.
- [12] Wu J, Liu H, Duan X, et al. Prediction of DNA-binding residues in proteins from amino acid sequences using a random forest model with a hybrid feature [J]. *Bioinformatics*, 2009, 25(1): 30-35.
- [13] Wang Jiasong, Yan Ming. Numerical methods in bioinformatics: An introduction [M]. Beijing: Science Press, 2013.
- [14] Cosic I. Macromolecular bioactivity: Is it resonant interaction between macromolecules theory and applications [J]. *Biomedical Engineering, IEEE Transactions on*, 1994, 41(12): 1101-1114.
- [15] Katoh K, Misawa K, Kuma K, et al. MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform [J]. *Nucleic Acids Research*, 2002, 30(14): 3059-3066.
- [16] Ding Y S, Zhang T L, Chou K C. Prediction of protein structure classes with pseudo amino acid composition and fuzzy support vector machine network [J]. *Protein and Peptide Letters*, 2007, 14(8): 811-815.
- [17] Liu H, Yang J, Wang M, et al. Using Fourier spectrum analysis and pseudo amino acid composition for prediction of membrane protein types [J]. *The Protein Journal*, 2005, 24(6): 385-389.
- [18] Wang M, Yang J, Liu G P, et al. Weighted-support vector machines for predicting membrane protein types based on pseudo-amino acid composition [J]. *Protein Engineering Design and Selection*, 2004, 17(6): 509-516.
- [19] Wang B, Chen P, Zhang J. Protein interface residues prediction based on amino acid properties only [M]// *Bio-inspired Computing and Applications*. Berlin Heidelberg: Springer, 2012: 448-452.
- [20] Hu J, Zhang F. Improving protein localization prediction using amino acid group based physicochemical encoding [M]// *Bioinformatics and Computational Biology*. Berlin Heidelberg: Springer, 2009: 248-258.
- [21] Han P, Zhang X, Feng Z P. Predicting disordered regions in proteins using the profiles of amino acid indices [J]. *BMC Bioinformatics*, 2009, 10(S1): S42.
- [22] Zhang T, Ding Y, Chou K C. Prediction of protein sub-cellular location using hydrophobic patterns of amino acid sequence [J]. *Computational Biology and Chemistry*, 2006, 30(5): 367-371.
- [23] Pirogova E, Cosic I. Examination of amino acid indexes within the resonant recognition model [C]// *Con-*

- ference of the Victorian Chapter of the IEEE EMBS. [S. l.]: IEEE, 2001:1-4.
- [24] Lazović J. Selection of amino acid parameters for Fourier transform-based analysis of proteins[J]. *Computer Applications in the Biosciences: CABIOS*, 1996, 12(6):553-562.
- [25] Bloomfield P. *Fourier analysis of time series*[M]. New Jersey: John Wiley and Sons, Inc., 2000.
- [26] 翁诗甫. 傅里叶变换红外光谱分析[M]. 2版. 北京:化学工业出版社, 2010.
Weng Shifu. *Analysis of Fourier transform infrared spectroscopy*[M]. Second Edition. Beijing: Chemical Industry Press, 2010.
- [27] Saha I, Maulik U, Bandyopadhyay S, et al. Fuzzy clustering of physicochemical and biochemical properties of amino acids[J]. *Amino Acids*, 2012, 43(2): 583-594.
- [28] 赵剑, 王嘉松, 刘国庆. 高维共鸣识别——蛋白质比较的新方法[J]. *计算机工程与应用*, 2013, 49(9): 224-228.
- Zhao Jian, Wang Jiasong, Liu Guoqing. Higher dimensional resonance recognition model-novel method for protein comparison[J]. *Computer Engineering and Applications*, 2013, 49(9):224-228.
- [29] Vaidyanathan P P. Genomics and proteomics a signal processor's tour[J]. *Circuits and Systems Magazine, IEEE*, 2004, 4(4):6-29.
- [30] Rockovsky S. 'Hidden' sequence periodicities and protein architecture[J]. *Proceedings of the National Academy of Sciences*, 1998, 95(15):8580-8584.
- [31] Eisenberg D, Weiss R M, Terwilliger T C. The hydrophobic moment detects periodicity in protein hydrophobicity[J]. *Proceedings of the National Academy of Sciences*, 1984, 81(1):140-144.
- 作者简介:**赵剑(1975-),男,讲师,研究方向:信号处理、生物信息学,E-mail:zhaojian@njut.edu.cn;许金涛(1991-),本科生,研究方向:应用数学。