研究论文

DOI:10.3969/j.issn. 1673–1689.2018.11.007

基于功率谱的蛋白质序列特征提取新方法

梁启浩， 李 阳， 唐旭清*

（江南大学 理学院，江苏 无锡 214122）

摘要：采用分层聚类和熵评价方法进行基于功率谱的蛋白质序列特征提取新方法研究。具体包含以下 3 个内容：首先，基于经典的 HP 模型给出了氨基酸序列的数值序列表达；其次，采用离散傅里叶变换方法获取蛋白质序列的特征频谱，构造 12 维特征向量；最后，利用分层聚类法获取蛋白质序列的分层结构。这种新方法将基于功率谱的 DNA 序列特征提取方法推广到蛋白质序列上。通过基于 19 条动物线粒体脱氢酶亚基 1 和亚基 4，以及 11 条 β 珠蛋白等 3 组数据的分层结构比较实验，结果表明，新方法在数据系统的分层结构的信息提取上优于基于功率谱的 DNA 序列分析方法。因此，新方法对确定未知基因的结构与功能有重要的生物意义。

关键词：DNA 序列；功率谱；分层聚类；蛋白质序列；熵

中图分类号：TP 391；O 29　　**文献标志码**：A　　**文章编号**：1673—1689(2018)11—1160—06

Feature Extraction Method of Protein Sequences Based on Power Spectrum

LIANG Qihao， *Li Yang*， *TANG Xuqing**

（School of Science，Jiangnan University，Wuxi 214122，China）

Abstract：Based on the power spectrum，a new way for extracting the protein sequences feature was proposed by applying the hierarchical clustering and entropy evaluation. It contained the following three main parts. Firstly，the numerical expression of amino acid sequences was given by the classical HP model. Then，the characteristic spectrum of protein sequence was obtained by using the discrete Fourier transform，and a 12-dimensional feature vector was constructed to represent the protein sequence spectral. Finally，the hierarchical clustering method was used to obtain the structure of protein sequences. The way is a new extension from DNA sequence to the protein sequence. By testing and comparing on three sets of data，their hierarchical structures shown that the new method is better than the DNA sequence analysis method based on power spectrum for extracting the structural information of the data system. This method has important biological significance in determining the structure and function of the unknown genes.

Keywords：DNA sequence，power spectrum，hierarchical clustering，protein sequence，entropy

收稿日期：2016–01–22

基金项目：国家自然科学基金项目(11371174)；江苏省普通高校研究生科研创新计划项目(KYLX15_1188)。

* **通信作者**：唐旭清(1963—)，男，安徽望江人，工学博士，教授，硕士研究生导师，主要从事智能计算，生物信息学，生态系统建模与仿真方面的研究。E–mail：txq5139@jiangnan.edu.cn

引用本文：梁启浩，李阳，唐旭清. 基于功率谱的蛋白质序列特征提取新方法[J]. 食品与生物技术学报，2018，37(11)：1160–1165.

蛋白质序列特征提取是指依据研究的目的提取序列信息,并使用数学方法描述,建立可以反映序列结构和空间信息的特征向量,进而表达其功能[1]。如何从复杂的序列中挖掘有用的信息是生物信息学的研究方向之一,信号频谱分析技术基于自动信息处理,广泛应用于特征提取的各个领域,比如周期性分析、蛋白质编码区预测和基因识别等方面[2-3]。Yin 等[2]将信号处理与分析方法引入 DNA 序列相似性分析中。Hota 等[4]基于快速离散傅里叶变换(Fast discrete Fourier transform,DFT)和小波变换(Wavelet transform,WT),从功率谱等信号处理方法的角度对基因识别进行了研究。王其强等[5]基于功率谱将信号处理与分析方法应用于 P53 家族基因的三周期性特征分析。这些研究对于大数据中 DNA 序列处理过程中的特征提取有重要的意义。

蛋白质存在于所有的生物细胞中,是生命的物质基础之一,蛋白质序列的研究具有极其重要的意义。蛋白质空间结构的所有信息均隐藏在氨基酸序列中,因此研究蛋白质的氨基酸序列组成已经成为生物信息学研究领域的关键问题之一[6]。聚类分析技术已广泛应用于蛋白质序列信息处理的各个方面,如分析蛋白质间的亲缘关系,提取蛋白质结构信息、功能信息等[7-8],其目的是简约数据信息系统、降低系统复杂度。文献[9]通过 Voss 映射将 DNA 序列转换为数字序列,采用功率谱方法提取 DNA 序列的特征信息从而进行 DNA 序列聚类分析,其中特征信息提取的核心是由离散傅里叶变换的序列特征频谱的 $j(j=1,2,3)$ 阶矩构造的一个 12 维的特征向量,并采用传统的非加权组平均法(UPGMA)得到不同物种基于这种相似关系的系统发生树。在此基础上,本文结合基于信号频谱分析技术与层次聚类方法,将 DNA 序列数据推广到蛋白质序列数据,进行蛋白质序列的特征提取与物种的系统发生树(或分层结构)研究。

1 材料与方法

1.1 数据来源

本文从 NCBI 网站中下载了文献[10]中 19 种动物的 ND1、ND4 的蛋白质序列(NADH dehydrogenase subunit1 是线粒体 NADH 脱氢酶亚基 1 的简写、NADH dehydrogenase subnits4 是线粒体 NADH 脱氢酶亚基 4 的简写,分别表示为数据 1 与数据 2)进行研究,具体的数据有 Gibbon(NC_002082.1),Gorilla(NC_011120.1),Human(NC_012920.1),Chimp(NC_001643.1),Pygmy Chimp(NC_001644.1),Sumatran Orang(NC_002083.1),Bornean Orang(NC_001646.1),Hedgehog(NC_002080.2),Rat(AC_000022.2),Mouse(NC_005089.1),Donkey(NC_001788.1),Horse(NC_001640.1),Cow(NC_006853.1),Baleen whale(NC_001601.1),Fin whale(NC_001321.1),Cat(NC_001700.1),Gray seal(NC_001602.1),Harbor seal(NC_001325.1),Rhino(NC_001779.1)。同时下载了文献[11]中 11 种 β 珠蛋白蛋白质序列(表示为数据 3),分别为 Human(AAA16334),Gorilla(CAA43421),Chimpanzee(CAA26204),Lemur(AAA36822),Rabbit(CAA24251),Goat(AAA30913),Bovine(CAA25111),Mouse(CAA24101),Rat(CAA29887),Opossum(AAA30976),Gallus(CAA23700)进行研究。并同时找到 3 种数据所对应的 DNA 序列与文献[9]做比较。

1.2 符号序列的数字表达 HP 模型

随着生物信息学的发展,对于 DNA 序列中碱基进行数值化的映射有很多,如 Voss 映射、实数映射、Z-curve 映射及朱平等[12]建立的映射 $\varphi:GF(7^3)\rightarrow C_{343}$ 等。

本文考虑氨基酸的物理和化学性质,在详细的 HP 模型中将 20 种氨基酸分成 4 大类,分别为极性亲水性(PQ),极性疏水性(PR),非极性亲水性(SQ)和非极性疏水性(SR),且 $PQ=\{G\}$,$PR=\{A,V,L,I,P,F\}$,$SQ=\{S,T,C,N,Q,K,R,H,D,E\}$,$SR=\{W,M,Y\}$,由此就将一条给定长度的蛋白质序列转化为一条由 4 类氨基酸构成的 4 元序列。

对含有 n 个氨基酸的蛋白质序列 $s=s_1s_2\cdots s_n$,其中为组成此蛋白质序列的氨基酸,进行数据化定义

$$\alpha_i=\begin{cases}0 & s_i\in PQ\\ 1 & s_i\in PR\\ 2 & s_i\in SQ\\ 3 & s_i\in SR\end{cases}$$

由此即可将任意一条蛋白质序列转化为一条由 0、1、2、3 构成的 4 元序列,记作:$X(s)=\alpha_1\alpha_2\cdots\alpha_n$。这样产生的序列为基因序列的指示序列。

1.3 基于功率谱的蛋白质特征向量提取

里叶变换能将满足一定条件的某个函数表示

成三角函数(正弦和/或余弦函数)或者它们的积分的线性组合。使用离散傅里叶变换,其显著的优点是使隐藏或潜伏在原始数据中的信息经周期性变换之后变得清晰。下文的研究以极性亲水性(*PQ*)为例说明,极性疏水性(*PR*),非极性亲水性(*SQ*)和非极性疏水性(*SR*)相似。

利用离散的傅里叶变换及上述的指示序列,可以将基因序列数据进行离散化

$$U_{PQ}(n)=\sum_{n=0}^{N-1}u_{PQ}(n)\mathrm{e}^{-\mathrm{i}\frac{2\pi nk}{N}}=$$

$$\sum_{n=0}^{N-1}u_{PQ}(n)\left(\cos\frac{2\pi nk}{N}-\mathrm{i}\sin\frac{2\pi nk}{N}\right),k=0,1,\cdots,N-1$$

这样就能得到复数序列$\{U_{PQ}(k)\},k=0,1,\cdots,N-1$。对这个复数序列取模的平方,以定义序列的功率谱

$$P_{PQ}(k)=|U_{PQ}(k)|^2,k=0,1,\cdots,N-1$$

类似地,可得 $P_{PR}(k)$、$P_{SR}(k)$和 $P_{SQ}(k)$,且原氨基酸序列的功率谱为这 4 个子序列的功率谱之和。即

$$P(k)=P_{PQ}(k)+P_{PR}(k)+P_{SQ}(k)+P_{SR}(k),k=0,1,\cdots,N-1$$

生物序列数学表示的目的之一就是分析生物序列的相似性[13],但是不同的蛋白质序列长度不同,因此仅通过功率谱不能进行相似性分析,进而去聚类。解决这种问题需要通过功率谱构造向量,而相似性则可以通过计算两向量之间的欧式距离得到。一般认为距离越小,两序列就越相似。对于极性且亲水性(*PQ*)类氨基酸,在文献[9]中定义 j 阶矩为

$$M_j^{PQ}=\frac{1}{N_{PQ}^{j-1}(N-N_{PQ})^{j-1}}\sum_{k=1}^{\left[\frac{N}{2}\right]}(P_{PQ}(k))^j$$

类似地,可获得 M_j^{PR}、M_j^{SQ} 和 M_j^{SR}。因此通过 M_j^{PQ}、M_j^{PR}、M_j^{SQ}、M_j^{SR} 可以构建 12 维向量,也称为基于矩的蛋白质序列特征向量,即(M_1^{PQ},M_1^{PR},M_1^{SQ},M_1^{SR},M_2^{PQ},M_2^{PR},M_2^{SQ},M_2^{SR},M_3^{PQ},M_3^{PR},M_3^{SQ},M_3^{SR})。

这样每一条蛋白质序列都会得到一个 12 维的特征向量。特征向量间的距离按欧式距离进行计算,即给定蛋白质序列 i 和 j 的特征向量分别为 M^i 和 M^j,则两条序列之间的欧氏距离记为

$$\|M^i-M^j\|^2=\sqrt{\sum_{k=1}^{3}(M_k^{PQ_i}-M_k^{PQ_j})^2+\sum_{k=1}^{3}(M_k^{PR_i}-M_k^{PR_j})^2+\sum_{k=1}^{3}(M_k^{SQ_i}-M_k^{SQ_j})^2+\sum_{k=1}^{3}(M_k^{SR_i}-M_k^{SR_j})^2}$$

以下将在上面特征向量的基础上开展研究。

1.4 聚类

聚类分析是进行数据分析的一个基本方法,在数据挖掘、模式识别、生物信息学和统计学等领域都有广泛的研究与应用[14]。它是探索或提取隐含在数据中的新规律和新知识的重要手段。本文将采用基于文献[15]得到的层次聚类方法,在有限集 X=$\{x_1,x_2,\cdots,x_n\}$上定义一个标准化的度量矩阵 $\boldsymbol{d}$,令

$$D=\{\boldsymbol{d}(x,y)|x,y\in X\}=\{d_0,d_1,\cdots,d_m\}$$

其中 $d_0=0<d_1<\cdots<d_m$。其算法如下:

算法 A

S1 输入 n 个样本,$i\Leftarrow 0$;

S2 构造 n 个类,每个类中只含有一个样本,记为 $X(d_i)=C=\{c_1,c_2,\cdots,c_n\}$;

S3 $A\Leftarrow C$,$i\Leftarrow i+1$,$C\Leftarrow\varnothing$;

S4 $B\Leftarrow\varnothing$;

S5 对于任意的 $c_j\in A$,令 $B\Leftarrow B\cup\{c_j\}$,$A\Leftarrow A/c_j$;

S6 $\forall c_k\in A$,如果存在 $x_j\in c_j$,$y_k\in c_k$,使得 $d(x_j,x_k)\leqslant d_i$,则 $B\Leftarrow B\cup\{c_j\}$,$A\Leftarrow A/c_k$;

S7 $C\Leftarrow\{B\}\cup C$;

S8 若 $A\neq\phi$,则转 S4;

S9 $X(d_i)=C$,输出 $X(d_i)$;

S10 直到 $C\neq\{X\}$,否则转 S3;

S11 结束。

通过算法 A,可获得数据系统的分层(或层次)结构。

1.5 聚类结果评价

本文采用熵作为聚类效果的度量标准,熵值表示同一类对象在聚类簇集中的分散程度。处于同一类中的对象熵值越高越分散,相应的聚类效果就越差。用标准公式计算簇 i 的熵,其计算公式为

$$e_i=-\sum_{j=1}^{k}p_{ij}\log_2 p_{ij}$$

其中,$p_{ij}=\frac{m_{ij}}{m_i}$,$m_i$ 为簇 i 中对象的个数,m_{ij} 为簇 i 中类 j 的个数。簇集合的总熵用每个簇熵的加权和表示

$$\mathrm{e}=\sum_{i=1}^{k}\frac{m_i}{m}e_i$$

其中,k 为簇的个数,m 为数据集中对象的总数。熵值的含义是:熵值越小相应的聚类效果就越好。在

理想情况下,每一类的所有对象应分布于不同的簇中,此时 e=0。

2 实验结果比较与分析

将本文的方法用于 ND1、ND4 与 β 珠蛋白序列的相似性分析。选取这 3 种蛋白质序列来检验本文方法的好坏,一是因为这 3 种数据在生物上具有重要意义,研究的结果有实际应用价值,ND1 与 ND4 参与线粒体氧化磷酸化的电子传递 [10],β 珠蛋白是位于红细胞内的一种更大的蛋白质(血红蛋白)的一个组件(亚基)[11],对所有的生物体都是至关重要的。二是由于这 3 种数据已被广泛研究。将本文结果与文献[9]构建的结果进行比较,使本文的方法更具有说服力。

本文将数据分为 3 类时两种方法的错分率和熵值进行比较研究。图 1~3 为采用 3 种数据构造 12 维特征向量进行聚类分 3 类时的分层结构。表 1,2 是采用本文方法与文献[9]中的方法对实验数据进行聚类所得到的结果的对错统计情况,表 3 为采用两种方法将 3 种数据分 3 类时的熵值对比。

将数据 1 分为 3 类时,文献[10]中标准的分为 3 类时的层次结构为(Hedgehog,Donkey,Horse,Cow,Baleen whale,Fin whale,Cat,Gray seal,Harbor seal,Rhino),(Rat,Mouse),(Gorilla,Gibbon,Human,Chimp,Pygmy Chimp,Sumatran Orang,Bornean Orang),从图 1 和 2 可以看出,将本文的方法应用到数据 1 和数据 2 所得到的结果与文献[10]中的结果是基本一致。人(Human),苏门答腊猩猩(Sumatran Orang),婆罗洲猩猩(Bornean Orang),长臂猿(Gibbon),大猩猩(Gorilla),侏儒黑猩猩(Pygmy Chimp),黑猩猩(Chimp)彼此之间的线粒体 NADH 脱氢酶很相似不是一种偶然,都属于灵长目。与其他哺乳类的动物进化关系最远的物种 Mouse 和 Rat 的线粒体 NADH 脱氢酶与其他物种最不相似,为啮齿目。19 种数据中除(Rat,Mouse)属于脊索动物门以外,Hedgehog 与另外 16 种都属于脊椎动物门,所以本文将数据中的 Hedgehog 先与灵长目聚为一类是可行的。而文献[9]的方法将数据 1 中本该属于鲸目、偶蹄目的 Cow 与啮齿目 Mouse 聚为 1 类,数据 2 中文献[9]的方法将啮齿目 Rat 和其他两类混在一起,而 Mouse 单独分为 1 类,这与进化事实相悖。因此,采用本文的方法得到的结果更好一些。

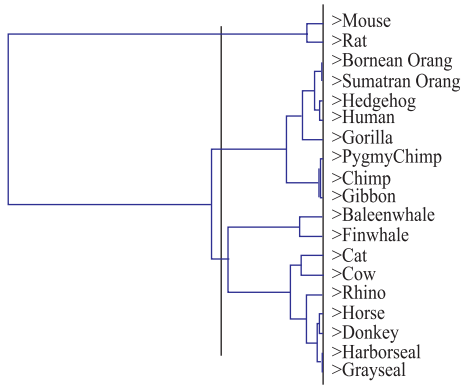

图 1 数据 1 分 3 类时的分层结构

Fig. 1 Hierarchical structure of Data 1

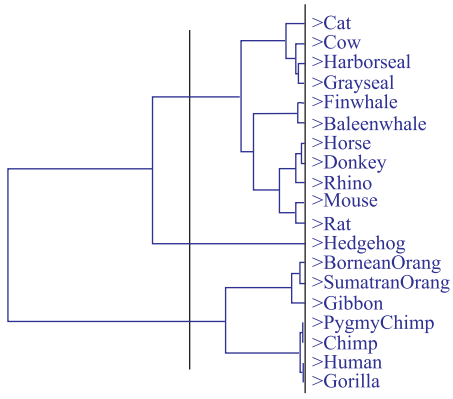

图 2 数据 2 分 3 类时的分层结构

Fig. 2 Hierarchical structure of Data 2

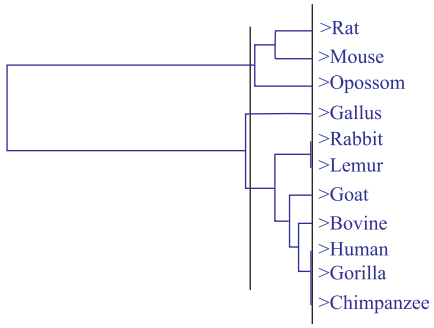

图 3 数据 3 分 3 类时的分层结构

Fig. 3 Hierarchical structure of Data 3

将数据 3 分为 3 类时,文献[11]中标准的分为 3 类时的结果为(Human,Gorilla,Chimpanzee,Lemur,Rabbit,Goat,Bovine,Mouse,Rat),(Opossum),(Gallus)。从图 3 可以看出,将本文所得结果与文献[11]中的结果相比:哺乳动物中的(Mouse,Rat),(Human,Gorilla,Chimpanzee)和(Lemur,Rabbit)分别是聚类过程中距离最近的物种,它们处于同一分支上,即

亲缘关系最近;Gallus 作为 11 个物种中唯一的非哺乳动物和其他的哺乳动物之间的进化距离很远;本文将(Mouse,Rat)先与负鼠目 Opossum 聚为一类,文献[11]后将(Mouse,Rat)与 Opossum 聚为一类,有一定的差别。而采用文献[9]方法的聚类过程中,将 Chimpanzee 与 Gorilla 这两个物种单独分为一类,此与文献[11]差别较大,同时与进化事实不一致。因此,采用本文的方法研究 β 珠蛋白的结果好于文献[9]的方法。

表 1　数据 1 和数据 2 分 3 类时聚类结果对错统计

Table 1　Error statistics of Clustering result for Data 1 and Data 2

数据	方法	动物品种						样本总数
		鲸目、偶蹄目		啮齿目		灵长目		
		对	错	对	错	对	错	
数据 1	本文	9	1	2	0	7	0	19
	文献[9]	8	2	1	1	0	7	19
数据 2	本文	9	1	0	2	7	0	19
	文献[9]	9	1	1	1	0	7	19

表 2　数据 3 分 3 类时聚类结果对错统计

Table 2　Error statistics of Clustering result for Data 3

数据	方法	动物品种						样本总数
		灵长、啮齿目		鸡形总目		负鼠目		
		对	错	对	错	对	错	
数据 1	本文	7	2	1	0	1	0	11
数据 2	文献[9]	7	2	1	0	0	1	11

表 3　本文方法与文献[9]的 3 类数据熵值比较

Table 3　Compare the entropy values of our method with literature[9] on 3 Data sets

方法	数据	簇 1	簇 2	簇 3	簇熵加权和
本文	数据 1	0	0	0.543 6	0.228 9
	数据 2	0.684 0	0	0	0.396 0
	数据 3	0.918 3	0	0	0.250 4
文献[9]	数据 1	1.271 8	1	0	1.176 2
	数据 2	1.253 3	0	0	1.121 4
	数据 3	0.543 6	0	0	0.395 3

由表 1,2 可以看出将 3 种数据分别分为 3 类时,采用本文方法样本被分错类的数目分别为1 个、3 个、2 个,错分率分别为 1/19=5.3%、3/19=15.8%、2/19=10.5%。与采用文献[9]的方法相比(样本被分错类的数目分别为 10 个、9 个、3 个,错分率分别为 10/19=52.6%、9/19=47.4%、3/19=15.8%),准确率分别提高了 47.3%、31.6%、5.3%。可见将本文的方法应用于这 3 种数据的特征提取时能够提高聚类质量。

由表 3 可看出将 3 种数据分别分为 3 类时,采用本文方法簇熵加权和分别为 0.228 9、0.396 0、0.250 4。与采用文献[9]的方法相比(簇熵加权和分别为 1.176 2、1.121 4、0.395 3),用本文方法产生的结果簇集具有最小的熵值,说明本文方法聚类结果较优。

因此,对于经常用于蛋白质氨基酸序列分析的小批量数据,由图 1~3 与表 1~3 可以看出用本文方法计算出的错分率与熵值都远低于文献[9]中的方法,即频率域上蛋白质序列的特征提取得到的相似度远远大于基于 DNA 序列的特征提取的相似度,本文的方法是有效的,之所以在蛋白质序列水平上得到的层次结构要比 DNA 序列更好,是因为对蛋白质序列的相似性进行分析,本质上是对组成蛋白质氨基酸序列的差异性比较,同一种蛋白质 DNA 序列比氨基酸序列长 3 倍,随着序列长度的增加,计算越来越复杂,并且序列到数字的映射以及对特征信息处理过程的信息缺失都会影响方法的准确率[2]。

通过本文方法来提取蛋白质序列的特征信息有 3 个优点:一是不用直接比较蛋白质序列而是去考虑经过离散傅里叶变换后这些蛋白质序列所对应的 12 维特征向量,特征向量是从组成蛋白质序列的氨基酸中提取出来的信息,这样蛋白质序列的

比较就转化成了向量之间的比较;二是因为氨基酸序列的亲疏水特性、极性非极性跟蛋白质的结构有一定的关系,所以基于氨基酸的这一特性对其分类进而简化蛋白质序列,再提取特征向量可以包含了更多的生物信息;三是基于层次聚类构建的分层结构能非常直观地反映蛋白质之间的进化关系。尽管如此,在提取序列的特征向量时仍然会不可避免地伴随着某些蛋白质序列结构方面的信息丢失。这也正是目前在蛋白质的研究中面临的一大挑战。

3 结 语

本文将基于 DNA 序列在频率域上的表示和离散傅里叶变换构造特征向量方法相结合以表征物种的特征方法进行推广,提出基于功率谱的蛋白质序列特征提取新方法。在研究的过程中采用经典 HP 模型对蛋白质的氨基酸序列数值化;由离散傅里叶变换将序列离散化,根据定义计算序列的功率谱构造蛋白质序列的特征向量距离;采用基于距离的层次聚类算法获取分层结构以考察蛋白质序列的相似性。选取 3 种不同的物种数据进行了新方法实验,以及与文献[9]中方法的比较研究。实验结果表明新方法是可行、有效的,且基于蛋白质的氨基酸序列在频率域上的特征提取方法优于基于 DNA 序列。这些研究结果对确定未知基因的结构与功能有重要的生物意义,对大规模的生物数据的自动信息处理具有应用价值。

参考文献:

[1] NAKASHIMA H,NISHIKAWA K. Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies[J]. **Journal of Molecular Biology**,1994,238(1):54-61.

[2] YIN C,CHEN Y,YAU S S T. A measure of DNA sequence similarity by Fourier transform with applications on hierarchical clustering[J]. **Journal of Theoretical Biology**,2014,359:18-28.

[3] YIN C,YAU S S T. An improved model for whole genome phylogenetic analysis by Fourier transform [J]. **Journal of Theoretical Biology**,2015,382:99-110.

[4] HOTA M K,SRIVASTAVA V K. Identification of protein-coding regions using modified Gabor-wavelet transform with signal boosting technique[J]. **International Journal of Computational Biology and Drug Design**,2010,3(4):259-270.

[5] WANG Qiqiang,TAN Chengjie,ZHU Ping,et al. A study on P53 genes' characteristics based on the 3-base periodicity [J]. **Acta Biophysica Sinica**,2013,29(4):296-309.(in Chinese)

[6] MU Z,WU J,ZHANG Y. A novel method for similarity/dissimilarity analysis of protein sequences [J]. **Physica A:Statistical Mechanics and Its Applications**,2013,392(24):6361-6366.

[7] MEI Juan,HE Sheng,LI Weijiang. Detection of communities in the yeast protein-protein interaction network based on graph clustering[J]. **Journal of Food Science and Biotechnology**,2011,30(1):95-100.(in Chinese)

[8] ZHANG Kun,TANG Xuqing. Research on the connection bias of amino acids based on probability transition matrix [J]. **Journal of Food Science and Biotechnology**,2012,32(1):106-111.(in Chinese)

[9] HOANG T,YIN C,ZHENG H,et al. A new method to cluster DNA sequences using Fourier power spectrum [J]. **Journal of Theoretical Biology**,2015,372(1):135-145.

[10] YU C,HE R L,YAU S S T. Protein sequence comparison based on K-string dictionary[J]. **Gene**,2013,529(2):250-256.

[11] ZOU S,WANG L,WANG J. A 2D graphical representation of the sequences of DNA based on triplets and its application[J]. **Bioinformatics and Systems Biology**,2014,DIO:10.118611687-4153-2014-1.

[12] YAN Y Y,ZHU P. Extended triplet set C343 of DNA sequences and its application to the p53 gene [J]. **Chinese Physics B**,2011,20(1):018701.

[13] ZHAO Jingjing,QI Bin,DING Lijuan,et al. Based on RSCU and QRSCU research codon bias of F/10 and G/11 xylanase[J]. **Journal of Food Science and Biotechnology**,2010,29(5):755-764.(in Chinese)

[14] TANG X Q,ZHU P. Hierarchical clustering problems and analysis of fuzzy proximity relation on granular space [J]. **IEEE Transactions on Fuzzy Systems**,2013,21(5):814-824.

[15] TANG X Q,ZHU P,CHENG J X. The structural clustering and analysis of metric based on granular space [J]. **Pattern Recognition**,2010,43(11):3768-3786.