氨基酸序列特征提取方法研究

罗林波,陈 绮
(海南大学 信息科学技术学院,海南 海口 570228)

摘 要:组成蛋白质的基本单位是氨基酸,对于蛋白质分类预测问题,氨基酸序列特征提取方法是一个非常重要的因素。对基于氨基酸组成、位置的特征提取算法如熵密度、n阶耦联组成和基于氨基酸性质的特征提取方法如自相关函数、伪氨基酸组成等方法进行了阐述,并进行了简单评价。基于氨基酸组成的方法实现简单、计算量小,且对所有的氨基酸序列都适用,但丢失了氨基酸的顺序信息以及其间的相互作用,基于氨基酸位置信息或理化特性等方法计算量非常大,科研工作者可以根据对蛋白质的不同要求选择相应的特征提取方法。

关键词:特征提取;熵密度;完全信息集
中图分类号:TP301.6 **文献标识码**:A **文章编号**:1673-629X(2010)02-0206-03

Research of Feature Extraction Methods of Amino Acid Sequence

LUO Lin-bo, CHEN Qi
(College of Information Science and Technology, Hainan University, Haikou 570228, China)

Abstract: Amino acids are the basic components of proteins. As for the prediction of protein classification, the feature extraction method of the amino acid sequence is a very important factor. It gives a clear clarification and simple evaluation on the basis of amino acid composition, location of the feature extraction algorithm, such as entropy density profile, n-order coupled composition, and feature extraction methods of amino acid nature, such as auto-correlation function, pseudo, amino acid composition and so on. Researchers may choose the corresponding feature extraction methods according to the different requirements for protein on the basis of that the method of the amino acid composition is simple, the computation load is light, and it can be applied to all amino acid sequence but the information on amino acid order and the interaction between them are lost, and also on the basis of that the computation load for information of amino acid location or physical and chemical features is heavy.

Key words: feature extraction; entropy density profile; complete information set

0 引 言

随着人类基因组计划的顺利进展,越来越多的蛋白质被测定出来,而通过实验确定其结构与功能的蛋白质则相对较少,且费时、费力、费财,实验中可能还会遇到一些目前无法解决的困难,因此探索利用理论及计算方法来研究蛋白质结构和功能具有重要意义。

如何从一条氨基酸序列提取它的有用信息,并用适当的数学方法来描述或表示这些信息,使之能正确反映序列与结构或功能之间的关系,对于蛋白质分类研究是至关重要的,也是决定分类质量的关键。目前的氨基酸序列的特征提取方法主要分为两类:一类为仅仅基于氨基酸组成和位置的方法;另一类为基于氨基酸物理化学性质的方法。

收稿日期:2009-06-25;**修回日期**:2009-09-05
作者简介:罗林波(1982-),男,湖北黄冈人,硕士研究生,研究方向为数据挖掘;陈 绮,博士,副教授,硕士生导师,研究方向为数据挖掘。

1 基于氨基酸组成和位置的特征提取算法

氨基酸是组成蛋白质的基本单位。一条蛋白质包含的基本信息是20种氨基酸的种类和排列顺序,因此基于氨基酸组成和位置的特征提取算法是最简单、最直观的方法,主要有氨基酸组成(Amino Acid Composition,AAC)、熵密度(Entropy Density Profile,EDP)、n阶耦联组成(n-Order Coupled Composition,n-OCC)和完全信息集(Complete Information Set,CIS)等。

1.1 氨基酸组成(AAC)

氨基酸组成[1]是指20种氨基酸在一条蛋白质中出现的频率。Nishikaw等人研究发现,蛋白质的折叠信息与氨基酸组成有明显的关联性,因此可以用一个20维的向量来表示一条蛋白质。假设 X 是一条蛋白

质,$f(x_i)$ 表示氨基酸 $x_i(i=1,2,\cdots,20)$ 在该序列中出现的次数,于是一条蛋白质可以表示为氨基酸组成空间的一个单位向量:

$$\boldsymbol{V}(X)=(V_1(X),V_2(X),\cdots,V_{20}(X))^{\mathrm{T}}$$

其中,$V_i(X)=f(x_i)/\sum_{i=1}^{20}f(x_i),(i=1,2,\cdots,20)$,显然 $\sum_{i=1}^{20}V_i(X)=1$。

1.2 熵密度(EDP)

Zhu 等人用熵密度表示 DNA 序列,确定基因序列中的外显子[2]。

信息熵的定义为[3]:

$$H(X)=-\sum_{i=1}^{20}f_i\log f_i$$

公式中 f_i 为第 i 种氨基酸在该蛋白质中出现的频率,则熵密度函数为:

$$S_i(x)=-\frac{1}{H(X)}V_i(X)\log V_i(X)$$
$$(i=1,2,\cdots,20)$$

于是该蛋白质 X 可表示为:

$$S(x)=(S_1(x),S_2(x),\cdots,S_{20}(x))^{\mathrm{T}}$$

1.3 n 阶耦联组成(n-OCC)

n 阶耦联组成[4]考虑了邻近的 n 个氨基酸对某个氨基酸的耦联作用。当 $n=0$ 时耦联组成退化为氨基酸组成,用一个 20 维的向量来表示;当 $n=1$ 时,耦联组成表示为一个 20×20 的条件概率矩阵:

$$\psi_1(s)=\begin{Bmatrix}P(A\mid A) & P(C\mid A) & P(D\mid A) & P(E\mid A) & \cdots & P(Y\mid A)\\ P(A\mid C) & P(C\mid C) & P(D\mid C) & P(E\mid C) & \cdots & P(Y\mid C)\\ P(A\mid D) & P(C\mid D) & P(D\mid D) & P(E\mid D) & \cdots & P(Y\mid D)\\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots\\ P(A\mid Y) & P(C\mid Y) & P(D\mid Y) & P(E\mid Y) & \cdots & P(Y\mid Y)\end{Bmatrix}_{20\times20}$$

其中 $P(a_1\mid a_2)$ 表示该蛋白质中氨基酸 a_1 出现并且氨基酸 a_2 紧邻其后的概率。此时,有 $\sum_{i=1}^{20}\sum_{j=1}^{20}p(a_i\mid a_j)=1$。当 $n>2$ 时,n 阶耦联组成用多维的条件概率矩阵表示。

1.4 完全信息集(CIS)

这种方法是由靳利霞[5]结合 FDOD 函数,用于蛋白质结构类预测的方法。完全信息集是指包含一条序列全部结构信息。设蛋白质 H 的总长度为 L,有一条长为 l 的氨基酸子序列。设每个位置上的氨基酸有 m 种选择,因此这样的子序列共有 m^l 个,令 $m(l)=m^l$,记第 i 个子序列为 $s_i^{(l)}(i=1,2,\cdots,m(l))$。计算 $s_i^{(l)}$ 在蛋白质 H 中的次数为 $n_i^{(l)}$,则 $s_i^{(l)}$ 在整个该蛋白质中出现的频率为 $f_i^{(l)}=n_i^{(l)}/(L-l+l)$。于是可以得到 H 的一个子序列分布:

$$U^l(X)=(f_1^{(l)},f_2^{(l)},\cdots,f_{m(l)}^{(l)})^{\mathrm{T}},\text{其中}\sum_{i=1}^{m(l)}f_i^{(l)}=1,l\leqslant L$$

令 Γ^l 表示所有满足 $\sum_{i=1}^{m(l)}f_i^{(l)}=1$ 的子序列分布的集合,于是:

$$\Gamma^l=\{(f_1^{(l)},f_2^{(l)},\cdots,f_{m(l)}^{(l)})^{\mathrm{T}}\mid\sum_{i=1}^{m(l)}f_i^{(l)}=1,f_i^{(l)}\geqslant0\},(l=1,2,\cdots,L)$$

因此对所有 $l(l=1,2,\cdots,L)$ 组成的集合就构成序列 H 的一个完整表示:

$$U(X)=\{U^1(X),U^2(X),\cdots,U^L(X)\mid U^l(X)\in\Gamma^l,l=1,2,\cdots,L\}$$

基于氨基酸组成和位置的特征提取算法还有残基耦联模型(Residue Couple Model, RCM)方法[6]、氨基酸组成分布方法[7]和多肽组成成分方法[8]等。

2 基于氨基酸物理化学性质特征提取方法

氨基酸的侧链决定了氨基酸的种类,20 种氨基酸侧链在形状、大小、负电性、水性以及酸碱性等方面都存在差异,正是这 20 种氨基酸的差异,使各种不同组合的氨基酸序列形成各种不同的蛋白质结构,并适应各类环境,完成其特定的生理功能。蛋白质的生物学活性和理化性质主要决定其空间结构的完整,因此仅仅知道蛋白质的氨基酸组成和它们的排列顺序并不能完全了解蛋白质的结构,需要考虑氨基酸的性质。目前,考虑氨基酸性质的主要方法有:自相关函数、伪氨基酸组成(Pseudo Amino Acid Composition, PseAA)、准序列次序作用和疏水模式组成(Hydrophob1e Pattern, HP)等。

2.1 自相关函数

氨基酸指数是定量表示 20 种氨基酸不同物理化学和生物化学性质的一组数值。采用自相关函数表示序列[9]时,先用 Kawashima 等人建立的氨基酸数据库中的氨基酸残基指数值,将蛋白质符号序列映射为数值序列:$S_h=h_1h_2h_3\cdots h_L$,式中 $h_i(i=1,2,\cdots,L)$ 为第 i 个残基的氨基酸指数。定义自相关方程:

$$r_n=\frac{1}{L-n}\sum_{i=1}^{L-n}h_ih_{i+n},n=1,2,\cdots,m\text{ 且 }m<L$$

m 是一个待定的整数,$m<L$。从而得到向量:

$$\boldsymbol{V}=(r_1,r_2,\cdots,r_n)^{\mathrm{T}}$$

2.2 伪氨基酸组成(PseAA)

伪氨基酸组成成分特征提取方法是 Chou 提出的一种基于氨基酸序列特征提取方法。伪氨基酸组成是一个(20+ λ)维的向量,其中前 20 维元素是氨基酸组成,后 λ 元素由下式得到:

$$\theta_j = \frac{1}{L-\lambda}\sum_{i=1}^{L-j}\Theta(R_i,R_j), j = 1,2,\cdots,\lambda$$

其中 R_i 为第 i 个残基相对应的理化参数值，进行归一化处理后得到一个(20 + λ)维单位向量：

$$\boldsymbol{V}(x) = (X_1, X_2, \cdots, X_{20}, X_{21}, \cdots, X_{20+\lambda})^{\mathrm{T}}$$

其中：

$$X_u = \begin{cases} \dfrac{f_u}{\sum_{i=1}^{20} f_i + \omega\sum_{j=1}^{\lambda}\theta_j} & 1 \leqslant u \leqslant 20 \\ \dfrac{\omega\theta_{u-20}}{\sum_{i=1}^{20} f_i + \omega\sum_{j=1}^{\lambda}\theta_j} & 21 \leqslant u \leqslant 20+\lambda \end{cases}$$

式中 f_u 为 20 种氨基酸在该蛋白质中出现的频率。

2.3 准序列次序作用

整合准序列次序作用[10]特征提取法是 Chou 提出的又一种氨基酸序列特征提取法，已经成功应用于蛋白质亚细胞定位预测和膜蛋白质分类研究。对于一个长度为 L 的蛋白质 $X = h_1h_2h_3\cdots h_L$，序列次序的影响可以通过下面所定义的一组序列次序相互作用因子 θ_j 来表示：

$$\theta_j = \frac{1}{N-j}\sum_{i=1}^{N-j} J_{i,i+j}, j = 1,2,\cdots,N-1$$

式中 $J_{i,k} = D(h_i,h_k)$ 为氨基酸 h_i 到 h_k 的物理化学距离，可根据氨基酸残基的疏水性、亲水性、极性和侧链体积计算出来。于是一个蛋白质可以表示成：

$$\boldsymbol{V}(x) = (X_1, X_2, \cdots, X_{20}, X_{21}, \cdots, X_{20+\lambda})^{\mathrm{T}}$$

其中：

$$X_u = \begin{cases} \dfrac{f_u}{\sum_{i=1}^{20} f_i + \omega\sum_{j=1}^{\lambda}\theta_j} & 1 \leqslant u \leqslant 20 \\ \dfrac{\omega\theta_{u-20}}{\sum_{i=1}^{20} f_i + \omega\sum_{j=1}^{\lambda}\theta_j} & 21 \leqslant u \leqslant 20+\lambda \end{cases}$$

3 其它特征提取算法

为了从氨基酸字母序列中提取更多的特征参数，人们引入了数字信号处理技术，提出了基于蛋白质数据库信息挖掘的方法，同时寻求多种方法相结合的混合方法。王勇献[11]提出了基于主成分分析的特征提取方法，Chou[12]提出了功能结构域组成的特征提取方法，Ashburner 等[13]提出了基因本体论特征提取算法，Pan 等[14]提出了利用统计信号处理技术对氨基酸序列进行特征提取的新思想，王克龙[15]提出基于离散小波变换的特征提取方法。

4 结束语

讨论了氨基酸序列特征提取的方法，基于氨基酸组成的方法如 AAC 方法和熵密度等方法实现简单、计算量小，且对所有的氨基酸序列都适用，但丢失了氨基酸的顺序信息以及其间的相互作用；基于氨基酸位置信息或理化特性等方法如 n - OCC 方法，完全信息集方法，PseAA 方法等计算量非常大，有的需要把蛋白质序列转换成一个 20 多维的向量；虽然在氨基酸序列特征提取过程中引入如序列顺序、氨基酸之间作用等信息，但是这些新引入信息的物化特性有些不是很明确，有些在物理、生物或者化学上根本无法解释；因此对于进一步预测蛋白质结构类起到的效果并不是十分明显，氨基酸序列特征提取方法有待进一步发展。

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):54 - 61.

[2] Zhu H Q, She Z S, Wang J. An EDP - based description of DNA sequences and its application in identification of exons in Human genome[C]//第二届中国生物信息学大会论文集. 北京:[出版者不详],2002:23 - 24.

[3] Shannon C E. The mathematical theory of communication [J]. Bell Sys. Tech. J, 1948, 27:623 - 656.

[4] Luo R Y, Feng Z P, Liu J K. Prediction of protein structural class by amino acid and polypeptide composition[J]. European Journal of Biochemistry, 2002(269):4219 - 4225.

[5] 靳利霞. 蛋白质结构预测方法研究[D]. 大连:大连理工大学,2002.

[6] Guo J, Lin Y L, Sun Z R. A novel method for protein subcellular localization: Combining residue - couple model and SVM [C]//Proceedings of 3rd Asia - Pacific Bioinformatics Conference. Singapore:[s. n.],2005.

[7] 施建宇，潘 泉，张绍武，等. 基于氨基酸组成分布的蛋白质同源寡聚体分类研究[J]. 生物物理学报，2006，22(1)：49 - 55.

[8] 曲 娟. 同源寡聚蛋白质的信息熵分类方法[D]. 大连:大连理工大学,2006.

[9] 张洪才. 基于支持向量机的蛋白质分类研究[D]. 西安:西北工业大学, 2003.

[10] Chou K C. Prediction of protein subcellular locations by incorporating quasi - sequence - order effect[J]. Biochemical and Biophysical Research Communications, 2000(19):477 - 483.

[11] 王勇献. 蛋白质二级结构预测的模型与方法研究[D]. 长沙:国防科技大学大学,2004.

[12] Chou K C, Cai Y D. Predicting protein structural class by functional domain Composition[J]. Biochemical and Biophysical Research Communications, 2004(321):1007 - 1009.

(下转第 212 页)

据能够基本达到 14 位的精度。

表 1 采样结果

输入信号单位:V	普通采样		过采样	
	理论值(12 位)	实际采样值(12 位)	理论值(16 位)	实际采样值(16 位)
0.1111	133	134	2141	2149~2150
0.4111	495	495	7924	7927~7929
0.8111	977	978	15634	15642~15645
1.2111	1459	1460	23344	23355~23358
1.6111	1940	1941	31054	31064~31068
2.1111	2543	2544	40692	40713~40718
2.7111	3266	3268	52257	52281~52286
3.1111	3747	3749	59967	59986~59992

3 处理器负荷分析

在该软件系统中,处理器主要的时间花在 DMA 的传输完成中断处理上,其中最占时间的是 256 个数据的循环累加。

由于 STM32 系列微控制器采用了 Cortex-M3 处理器,而 Cortex-M3 处理器的核心是基于哈佛架构的 3 级流水线内核,该内核集成了分支预测技术,所以达到了 1.25 DMIPS/MHz 的优越性能。在 56MHz 的系统时钟下,指令周期 = 1/(1.25×56) = 0.01423μs。通过反汇编可以看到,每次累加需要 6 条指令(虽然其中包含了 LDR 和 BCC 指令,但由于指令顺序经过编译器调整,内核又使用了分支预测,所以流水线能保证每个机器周期吐出一条指令[7]),256 次累加总共用到 256×6 = 1536 条指令,执行时间 = 1536×0.01423≈21.857μs。也就是说在 10ms 的采样间隔时间内,处理器只需要花费 21.857μs 去处理数据,由此可计算处理器使用率 = 0.021857/10≈0.22% 。由此可见处理器并不会因为过采样数据的处理而受到很大的影响。

4 STM32 上过采样对输入信号频率的限制

根据 Nyquist 定律,采样频率必须是输入信号的 2 倍才能将信号还原,当需要提高 p 位采样精度的时候,频率又得提高 4^p 倍。STM32 上的 ADC 能达到的最高采样率为 1MHz,如果要达到 16 位的精度,那么输入信号的频率就不能够超过 1M/2/256 = 2kHz。

5 结束语

过采样技术的应用,能够有效地利用低精度的 ADC 获得高精度的采样结果,这使得本来需要使用昂贵的外部专用 ADC 的微控制器仅使用自带的 ADC 就能够达到应用要求,在一定程度上节约了成本。

当然,过采样技术的使用必须具备一定的前提条件,并且使用过采样技术后对输入信号频率具有一定局限性。过采样技术对 CPU 的负荷也是有一定的影响的,但是由于 STM32 采用了高性能的 Cortex-M3 处理,并且采用了 DMA 在外设和 SRAM 之间进行直接数据传输,使得过采样技术不会给 STM32 处理器增加太大负荷。

参考文献:

[1] Schafer R. Discrete2time signal processing[M]. New Jersey: Prentice Hall,1999.

[2] Karema T,Ritoniemi T,Tenhunen H. A 20-bit sigma-delta D/A converter prototype for audio applications[M]//of 1991 IEEE Int. Conference on A/D and D/A Conversion. UK:[s. n.],1991:136-141.

[3] Cygnal. Application Note An18 UAS[EB/OL]. 2003. http://www. cygnal. com/suport/application. htm/<<Improving ADC resolution by Over sampling and Averaging>>

[4] 许 勇,叶 刚,卞青青,等.基于 A/D 转换最小二乘法的数据采集应用[J]. 微计算机信息,2009,4(2):280-282.

[5] 阮双喜.基于 ARM 的气象数据采集系统的研制[J].吉林大学学报,2006,24(2):222-223.

[6] 王永宏,徐 炜,郝立平.STM32 系列 ARM Cortex-M3 微控制器原理与实践[M].北京:北京航空航天大学出版社,2008.

[7] Sloss A N,Symes D. ARM 嵌入式系统开发——软件设计与优化[M].沈建华,译.北京:北京航空航天大学出版社,2005.

[8] 张 俊.匠人手记[M].北京:北京航空航天出版社,2008.

(上接第 208 页)

[13] Ashburner M,Ball C A,Blake J A,et al. Gene ontology: tool for the unification of biology[J]. Nature Genetics,2000(25): 25-29.

[14] Pan Y X,Zhang Z Z,Guo Z M,et al. Application of pseudo amino acid composition for predicting protein subcellular location: stochastic signal processing approach[J]. Journal of Protein Chemistry,2003(22):395-402.

[15] 王克龙 . 离散小波变换分析蛋白质序列相似性[D]. 成都:四川大学,2004.