文章编号:1674-2869(2010)05-0045-04

蛋白质空间结构数字特性统计分析及应用

章社生,何 康,范 宁,晏 臻,王 星
(武汉理工大学理学院统计学系,湖北 武汉 430070)

摘 要:研究了蛋白质分子的结构,从RCSB公共数据库中收集蛋白质PDB数据文件,利用统计分析和数据挖掘知识,建立了以蛋白质形心为原点的蛋白质原子空间坐标系,从蛋白质的数字特征入手,讨论了五类蛋白质(肌蛋白、血蛋白、激素、抗体、生物膜)的结构特性及数字特征的分布,其中激素分子相对其他几类蛋白质较小,原子的分布也相对集中;并讨论了20种残基的结构特性,构造出蛋白质数字特征能量函数,其结论有助于蛋白质生物功能开发和蛋白质设计研究.

关键词:蛋白质;氨基酸;数字特征;生物统计
中图分类号:O213 文献标识码:A doi:10.3969/j.issn.1674-2869.2010.05.013

0 引 言

蛋白质是构成生命的物质基础,它是与各种形式的生命活动紧密联系在一起的物质.在催化生命体内各种反应进行、调节代谢、抵御外来物质入侵及控制遗传信息等方面都起着至关重要的作用,是生命科学中极为重要的研究对象.蛋白质是由一条或多条多肽链组成的生物大分子,每一条多肽链有数十到数百个氨基酸残基不等;各种氨基酸残基按一定的空间顺序排列.不同的蛋白质空间结构有不同的生命功能.揭示蛋白质的生命活动规律,研究蛋白质的折叠,设计具有特定功能的蛋白质,都需要了解蛋白质空间结构.文献[1]介绍了X射线晶体学、二维核磁共振(2D-NMR)和低温冷冻电镜等蛋白质空间结构的实验测定方法.应用这些方法,实验室已测定大量蛋白质空间结构,并以PDB文件形式贮存在公共数据库中,免费供世界各地研究者使用.文献[2]应用统计分析方法,利用数据挖掘中的数据分布拟合理论对生物科学领域中的蛋白质侧链空间结构进行统计分析.以世界上广泛使用的生物分子三维结构数据库PDB为基础,利用多氨酸残基侧链碳原子间距离的统计分析方法,通过正交试验设计和信息论中的熵函数等相关知识,给出了不同位置、不同氨基酸残基种类对侧链结构的影响.文献[3-4]用统计和几何方法给出了氨基酸在蛋白质空间结构中的深度计算,并利用PDB数据库得到了不同氨基酸在蛋白质中的深度倾向性因子,并得到了这些倾向性因子与氨基酸的物理、化学综合特性的相关性质.根据蛋白质空间结构和蛋白质生物性质,国内外学者建立了多种蛋白质折叠模型和蛋白质设计模型[5-9];这些模型一般应用能量函数进行计算,利用蛋白质空间结构的数值特征是构造能量函数的一种途径.

本文根据PDB数据文件计算蛋白质空间结构的数值特征,构造数值特征的能量函数.PDB收集的蛋白质数据来源于X光晶体衍射和核磁共振的数据,经过整理和确认后存档而成.蛋白质种类众多,分类方式各异.按分子形状分类,可分为球状蛋白质和纤维状蛋白质.鉴于大多数蛋白质属于球状蛋白质,如血红蛋白、肌红蛋白、酶、抗体等[10-11],本文主要选取了球蛋白作为研究对象,并将其分为五大类,即血红蛋白、肌蛋白、激素、抗体、生物膜的成分,分别抽样进行计算与分析.文中叙述数字特征的计算原理及血红蛋白等五类蛋白质的数字特征,讨论氨基酸的数字特征,给出了数值特征能量函数的构造原理.

1 蛋白质数字特征

本文只讨论每一个蛋白质PDB文件中关于原子(ATOM)部分的数据.

1.1 数据处理方法

从数据库中查询选取出属于肌蛋白、血蛋白、激素、抗体、生物膜共五类的部分蛋白质,并按类

收稿日期:2010-03-04
基金项目:国家自然科学基金(69773023)资助项,武汉理工大学自主创新研究基金(批准号09140716101)资助项目
作者简介:章社生(1955-),男,湖北鄂州人,教授,博士.研究方向:生物数学.

别存放(每类选取 60 ~ 100 个蛋白质),然后按下面步骤计算数字特征.

a. 对于第 i 个蛋白质分子,提取出 PDB 文件中所有 ATOM 的立体坐标数据,其中(x_{ij},y_{ij},z_{ij})为第 j 个原子的立体坐标.

b. 计算该分子的立体中心(形心)坐标$(x_i,y_i,\bar{z_i})$:将形心平移到坐标原点,相应平移后原子坐标为$(x_{ij}',y_{ij}',z_{ij}')$.

c. 计算该蛋白质分子第 j 个原子到中心点距离 r_{ij}的期望与标准差. 该蛋白质分子内原子到形心距离的数学期望与标准差分别为:

$$E(i) = \frac{1}{n_i}\sum_{j-1}^{n_i} r_{ij}, \sigma(i) = \sqrt{\frac{1}{n_i}\sum_{j-1}^{n_i}(r_{ij} - E(i))^2}$$

d. 计算该类蛋白质分子的数学期望与标准差的均值:

$$E = \frac{1}{m}\sum_{i-1}^{m} E(i), \sigma = \frac{1}{m}\sum_{i-1}^{m}\sigma(i)$$

其中 m 表示所考察的该类蛋白质的蛋白质分子个数.

e. 统计每类蛋白质各分子的数学期望和方差(标准差),分析每类蛋白质数字特征的概率分布情况.

1.2 数据结果分析

1.2.1 血红蛋白 血红蛋白原子到形心距离的数学期望约为 22.69,平均标准差约为 7.30. 对属于血红蛋白,所考察的蛋白质分子的距离数学期望在 9.785 到 48.115 之间,标准差在 2.748 到 16.324 之间,约 40% 的蛋白质分子数学期望在 10 到 20 范围之间,43% 的分子落在 20 到 30 区域内. 然而 80% 标准差在 3 到 10 内,且大多集中在 5 左右. 因此推测,属于血红蛋白的蛋白质分子,其原子到分子形心距离的数学期望集中分布在 15 ~ 30 之间,分子的结构较为密集、聚中.

1.2.2 肌蛋白 肌蛋白原子到形心距离的数学期望约为 22.85,平均标准差为 8.25. 所考察的属于该类蛋白的蛋白质分子的距离数学期望在 10.989 到 104.242 之间,标准差在 3.918 至 58.768 之间,两者的极差均较大. 观察距离数学期望和标准差的,大约 80% 的肌蛋白分子的原子到中心距离在 10 到 30 之间,标准差在 4 到 10 之间. 总体上看,分布仍然比较集中,波动不大,但有几种肌蛋白分子偏离均值较远,分子内部原子到形心的平均距离可达到 100 左右.

1.2.3 抗体 抗体原子到形心距离的数学期望约为 25.98,平均标准差为 8.89. 所考察的抗体蛋白质分子的距离数学期望分布于 3.036 到 51.928,标准差在 1.500 到 19.842,约 80% 的抗体的距离数学期望在 20 到 40 之间,整体上没有很大的波动.

1.2.4 激素 激素原子到形心距离的数学期望约为 18.13,平均标准差 6.57,对属于该类的蛋白质分子来说,距离数学期望仍集中在 10 到 30,标准差较均匀地分布在 2 到 10 之间. 激素分子相对其他几类蛋白质较小,原子的分布也相对集中.

1.2.5 生物膜的成分 生物膜原子到形心距离的数学期望约为 20.45,平均标准差为 6.57,该类分子的原子到形心距离的数学期望在 10 到 20 附近较多,也有分子在 60 附近,分布体现的规律性不强,这可能是由于样本数量不足所导致,也可能是生物界中自身的差异多所致. 通过上述不同类蛋白质的数据分析,如表 1 所示,激素蛋白的整体数学期望最小,且方差也是较小;其后依次是生物膜成分、血红蛋白、肌蛋白、抗体. 在这五类蛋白质中,原子到其形心的平均距离较大者,这种距离的平均偏差一般也较大.

表 1 五类蛋白质的数字特征对比表

Table 1 Statistical features of five protein

类别	数学期望	标准差
血红蛋白	22.688 1	7.302 1
肌蛋白	22.851 3	8.253 4
抗体	25.977 2	8.889 8
生物膜成分	20.452 2	6.572 0
激素	18.132 9	6.572 1

2 氨基酸的数字特征

进一步研究蛋白质分子的数字特征,考虑蛋白质的组成成分氨基酸. 基于氨基酸的种类众多,只考虑 20 种天然的氨基酸. PDB 文件中氨基酸以残基序列进行记录,为此笔者研究各类蛋白质中属于同一种残基的原子的数字特征,进而进行定性与定量分析.

2.1 数据处理方法

计算各类蛋白质分子中原子到相应分子中心的距离 r_{ij},将所有考察的原子的距离依据各原子的残基名分类,统计各类(残基)中原子到形心距离的数学期望(平均值)与标准差.

2.2 数据结果分析

对于血红蛋白,不同残基下的数学期望差异较小,大致都在 28 左右波动;并且离散程度也无明显的差异. 因此笔者认为残基的不同对原子到形心距离的影响相对弱. 另外,在组成血红蛋白的原子中,残基 MET 出现次数最少,而 LEU、LYS 较多. 对于肌蛋白,数学期望差异仍是不显著,大体

在35到40之间;相对的标准差差异较小,这跟血红蛋白的情况类似. 在蛋白质的组成中,残基GLU、LEU、LYS出现较多,TRP、CYS较少. 类似地,对激素、抗体、生物膜成分三类蛋白,同类蛋白数学期望和标准差的分布都较为集中,没有大的波动. 可以推断,对于同种蛋白质,残基对其原子到相应蛋白质分子中心距离的影响不大. 对于激素,残基LEU贡献显著,CYS、TRP出现频率较小;抗体中具有残基SER、LEU的原子较多,具有残基MET、CYS的较少;对于生物膜的成分,LEU、ΛRG最多,CYS最小.

此外,笔者研究对于同一种残基,不同类的蛋白中数字特征及原子个数的差异问题. 残基ALA和CYS对应的数字特征如表2所示,由表可知,对于残基ALA,在血蛋白、肌蛋白、抗体和生物膜成分中出现的频率高于激素. 对于残基CYS,肌蛋白出现的频率远高于激素. 另外,在不同类型的蛋白质中,其氨基酸的数字特征各不相同.

表2　五类蛋白不同残基的对比表

Table 2　Residue statistical features of five protein

类别	期望值(ΛLΛ)	标准差(ΛLΛ)	个数(ΛLΛ)	期望值(CYS)	标准差(CYS)	个数(CYS)
血红蛋白	28.218	14.157	14 498	28.896	16.598	3 564
肌蛋白	35.753	26.022	10 020	40.033	24.158	2 089
激素	29.740	16.345	3 962	22.693	14.170	1 542
抗体	30.485	15.063	15 607	29.657	14.325	7 373
生物膜	30.979	12.476	11 290	25.286	13.432	1 568

一般的,各类蛋白中残基LEU出现最为频繁,CYS较小. 对于同种蛋白质,残基对其原子到相应蛋白质分子中心距离的影响不大. 另外,笔者研究各残基中原子到形心距离的分布情况,发现频数随着距离的增大而递减. 基于以上的数据分析,从侧面证实了不同类型的蛋白质的特征差异性与一致性. 同时也说明了不同的蛋白质的组成不同,对应的数字特征也不同. 这也许可以从另一个角度提供组合蛋白质的思路.

3　数字特征能量函数

设Eij为第i种蛋白质的第j种残基($i=1\sim5$, $1\sim20$)的期望,P_{ij}为第i种蛋白质的第j种残基期望的发生概率,定义为:

$$P_{ij}=\frac{E_{ij}}{\sum_{i,j}E_{ij}}=P_iP_{ij}\quad P_i=\frac{E_{i.}}{\sum_{i,j}E_{ij}}$$

$$p_{ij}=\frac{E_{ij}}{E_{i.}}\quad E_{i.}=\sum_j E_{ij}$$

式中P_i为第i种蛋白质发生的概率,p_{ij}为第i种蛋白质第j种残基在第i种蛋白质发生的条件下的条件概率. $E_{i.}$为第i种蛋白质的期望,它为E_{ij}对所有的j求和. 定义期望能量函数如下:

$$Ve(i,j)=-\log\frac{P_{ij}}{P_0}\quad P_0=E(P_{ij})$$

这里P_0为P_{ij}的概率平均值. 上式建立了能量函数与蛋白质种类和残基种类之间的关系,它可以用于蛋白质设计. 根据上面五类蛋白不同残基的对比表给出的残基ALA和CYS的预期,笔者容易求出概率p_{ij},其结果列于表3. 由表可知,肌蛋白中残基CYS的期望概率最大,激素中残基CYS的期望概率最小. 相对残基ALA的期望概率,CYS的期望概率比较分散. 文献[7]认为比较分散概率有助于蛋白质设计.

表3　五类蛋白不同残基的期望概率

Table 2　Rcsiduc cxpcctation of fivc protcins

类别	期望概率(ΛLΛ)	个数(ΛLΛ)	期望概率(CYS)	个数(CYS)
血红蛋白	0.181 8	14 498	0.197 1	3 564
肌蛋白	0.230 4	10 020	0.273 1	2 089
激素	0.191 6	3 962	0.154 8	1 542
抗体	0.196 4	15 607	0.202 3	7 373
生物膜	0.199 6	11 290	0.172 5	1 568

4　结　语

本文对蛋白质分子的结构特性进行了量化处理,利用统计分析,数据挖掘知识,从蛋白质的数字特征入手,讨论五类蛋白质的特点,进而根据20种残基分组深入研究,从不同角度分析得出了一系列的结论,为蛋白质的结构数学化提供了思路,也为组合氨基酸生成蛋白质提供了数据支持.

在数学上,数字特征的计算原理是非常成熟的. 但在生物中,有许多生物数字特征计算工作没有完成. 蛋白质是研究得较多的生物对象,但笔者查阅了国内外文献资料,没有发现完整研究蛋白质数字特征计算的文章. 至今为止,人们已测量的蛋白质数据是海量的,通过数字特征计算是揭示蛋白质空间结构生物性质的途径之一. 用数字特征构造能量函数是生物数据二次挖掘,该能量函数能用于蛋白质设计. 另外,本文工作还有极大的拓展空间,例如,有更多种类蛋白质的数字特征需要计算,DNA、RNA等生物基团的数字特征也需要计算和分析.

参考文献:

[1]　江凡. 蛋白质空间结构的实验技术和理论方法[J]. 物理,2007,36(4):272-279.

[2]　王昕,毛炳蔚,王福伟,等. 蛋白质空间结构的统计分析[J]. 山西大同大学学报:自然科学版,2008,24(5):3-8.

[3] 沈世镒,胡刚,张华.氨基酸在蛋白质空间结构中的深度倾向性因子[J].生物数学学报,2007(7):305-310.
[4] 沈世镒,胡刚,张华.蛋白质空间形态特征分析与计算方法[J].工程数学学报,2006,22(2):225-234.
[5] 胡敏,彭群生.一种基于空间密度特征的蛋白质结构相似性判定方法[J].工程图学学报,2005,26(1):90-95.
[6] 王仲君,王能超,毛黎明.基于自回避搜索遗传算法的蛋白质折叠研究[J].武汉理工大学学报,2005,27(8):91-95.
[7] Faraggi E, Yang Y, Zhang S, et al. Predicting continuous local structure and the effect of its substitution for secondary structure in fragment-free protein structure prediction[J]. Structure, 2009, 17: 1515-1527.
[8] Liang S, Wang G, Zhou Y. Refining near-native protein-protein docking decoys by local re-sampling and energy minimization[J]. Proteins, 2009, 76: 309-316.
[9] Xue B, Faraggi E, Zhou Y. Predicting residue-residue contact maps by a two-layer, integrated neural-network method[J]. Proteins, 2009, 76: 176-183.
[10] 张佑红,陈龙,靖志强,等.不同周期Sf9细胞琥珀酸脱氢酶酶活的研究[J].武汉工程大学学报,2009,30(5):4-6.
[11] 奚强,李俊,林丫丫,等.L-核糖的合成[J].武汉工程大学学报,2009,30(5):18-20.

Statistical analysis and application of the protein structure

ZHANG She-sheng, HE Kang, FAN Ning, YAN Zhen, WANG Xing
(Department of Statistics, School of Science, Wuhan University of Technology, Wuhan 430070, China)

Abstract: The article mainly deals with the structure of protein molecules. With the PDB files collected from the RCSB public database and the knowledge of statistical analysis and data mining. We build a spatial coordinate system with the geometrical center of a specific protein molecule being the origin. After discussing the features of five kinds of protein (muscle protein, blood protein, hormones, antibodies, biomembrane), we study the structural characteristics as well as the distribution of the features respectively. Consequently, in terms of our analytical system, hormone molecules are relatively smaller and the distributions of their atoms are more concentrated compared with others. Meanwhile, the detailed discussion on the structural characteristics of 20 kinds of the amino acid residues is conducted. Furthermore, we develop energy function based on the features of these residues, which will contribute to the development of protein biological function as well as the design research.

Key words: protein; the amino acid; features; biostatistics

本文编辑:张　瑞