王红梅,武恩斯,朱玉凤. 固有无序蛋白质无序区和有序区氨基酸组成偏好性分析[J]. 江苏农业科学,2014,42(4):38 - 39.

固有无序蛋白质无序区和有序区氨基酸组成偏好性分析

王红梅[1], 武恩斯[2], 朱玉凤[2]

(1. 德州学院物理与电子信息学院/山东省功能大分子生物物理重点实验室,山东德州 253023;
2. 山东师范大学,山东济南 250358)

摘要:以固有无序蛋白质为研究对象,通过 CD - HIT 对数据进行去冗余处理,然后利用编程软件对数据进行统计而得到新的数据。对所有无序区及有序区的氨基酸含量进行对比,认为氨基酸 Val、Ile、Leu、Phe、Trp、Asn、Tyr、His 具有形成有序结构的偏好性;氨基酸 Pro、Ser、Gln、Asp、Lys 具有形成无序结构的偏好性。研究结论有助于进一步挖掘固有无序蛋白质的序列特征,并为固有无序蛋白质的预测提供一些借鉴。

关键词:固有无序蛋白质;功能位点;无序区;序列分析

中图分类号: Q516　　**文献标志码**: A　　**文章编号**:1002 - 1302(2014)04 - 0038 - 02

蛋白质是生物体中最重要的两类大分子之一,传统思想认为蛋白质要实现其生物功能,必须先折叠成一个稳定的三维结构,因此形成了蛋白质结构决定其功能的主流观点[1]。然而随着基因工程方法和实验技术的发展以及基因组计划的开展,在 20 世纪 90 年代初,人们发现有些蛋白质或蛋白质序列中的一部分区域在生理条件下不具有一个确定的三维结构,但是依然能够正常行使生物学功能。进一步研究发现的这类蛋白质越来越多,并逐渐形成了一种新的蛋白质类型,称为固有无序蛋白质(intrinsically disordered proteins,简称为 IDPs)[1-3]。对目前存在的大量基因库数据进行分析发现:蛋白质的无序结构与蛋白质功能之间关系密切,无序蛋白质在诸如转录、翻译、调控细胞信号转导、蛋白质磷酸化及小分子存储等过程中发挥着重要的作用;另一方面,无序蛋白质又经常与多种疾病联系在一起。与人类癌症相关的蛋白质中,无序蛋白质的含量高达 79%;在心血管疾病有关的蛋白质中,无序蛋白质的含量也高达 57%。无序区是固有无序蛋白质发挥功能的主要区域,功能位点大多分布在该区域,因此预测蛋白质的无序区成为判断蛋白质是否无序的热点问题。Romero 等在 1997 年首次对蛋白质无序区域进行预测,他们预测的准确性达到 70%,此后无序蛋白质的预测方法得到了迅速发展,目前应用于无序蛋白质序列预测的方法已经超过 50 种,并且这些预测方法的准确性普遍达到 85% 以上。

本研究基于序列分析的方法,以 DisProt 数据库中的固有无序蛋白质为研究对象,通过 CD - HIT 程序对数据进行去冗余处理,将处理后的数据利用编程软件 Matlab 7.0 进行统计而得到新的数据;对新数据进行分析,通过编程把序列的无序区和有序区分别提取出来,再分析无序区和有序区氨基酸组成的偏好性。本研究有助于进一步挖掘固有无序蛋白质的序列特征,从而为固有无序蛋白质的预测提供借鉴。

1　数据来源及去冗余处理

1.1　数据来源

本研究以固有无序蛋白质数据库 DisProt(版本 6.01)[4](http://www.disprot.org/index.php)为研究对象(发布日期为 2012 年 10 月 15 日),下载数据库中最新的固有无序蛋白质进行研究,共有无序蛋白质 684 个,无序区 1 513 个。

1.2　去冗余处理

由于蛋白质序列数据库中都含有大量的冗余序列,它们通常不能提供更多的信息,而且不利于数据的统计分析,并且由于冗余序列要占用更多的计算机存储和处理资源,因此去除这些冗余信息具有很高的实用价值,不但可以减小数据库的大

收稿日期:2013 - 08 - 23

基金项目:山东省自然科学基金(编号:ZR2010CQ041)。

作者简介:王红梅(1974—),女,山东德州人,硕士,副教授,主要从事生物信息学的研究。E - mail:whm_2327@126.com。

[2]朱立静,陈淑吟,许晓风,等. 四角蛤蜊江苏群体线粒体 *CO I* 基因片段序列研究[J]. 江苏农业科学,2010(4):33 - 35,97.

[3]姜　帆,刘佳琪,李志红,等. 基于 DNA 条形码的广西苦瓜中实蝇幼虫分子鉴定研究[J]. 植物保护,2011,37(4):150 - 153.

[4]李鹏飞,朱文斌,贺舟挺,等. 东海带鱼 DNA 条形码的建立及*COI*序列变异分析[J]. 浙江海洋学院学报:自然科学版,2013,32(1):6 - 9.

[5]Chase M W,Salamin N,Wilkinson M,et al. Land plants and DNA barcodes:short - term and long - term goals[J]. Philosophical Transactions of the Royal Society of London. Series B,Biological Sciences,2005,360(1462):1889 - 1895.

[6]任保青,陈之端. 植物 DNA 条形码技术[J]. 植物学报,2010,45(1):1 - 12.

[7]张　欣,于瑞祥,方晓明,等. 橄榄油掺假检测技术的研究进展[J]. 中国油脂,2013,38(3):67 - 71.

[8]庞晓慧,宋经元,陈士林. 应用 DNA 条形码技术鉴定中药材灯心草[J]. 中国中药杂志,2012,37(8):1097 - 1099.

[9]伏建国,杨晓军,钱　路,等. 植物 DNA 条形码技术在出入境检验检疫领域的应用[J]. 植物检疫,2012,2(02):64 - 69.

[10]高连明,刘　杰,蔡　杰,等. 关于植物 DNA 条形码研究技术规范[J]. 植物分类与资源学报,2012,34(6):592 - 606.

[11]Tamura K,Peterson D,Peterson N,et al. MEGA5:molecular evolutionary genetics analysis using maximum likelihood,evolutionary distance,and maximum parsimony methods[J]. Molecular Biology and Evolution,2011,28(10):2731 - 2739.

小、提高序列搜索的速度，而且有助于对数据的统计分析。本研究利用去冗余程序 CD－HIT[5-6]（http://weizhong－lab.ucsd.edu/cdhit_suite/cgi－bin/index.cgi）对数据进行处理，将相似度阈值设为 30%。结果显示：去冗余前，固有无序蛋白质共有 684 条序列；去冗余后，蛋白质共有 549 条序列。

2　固有无序蛋白质无序区和有序区的氨基酸组成偏好性分析

用 Matlab 编程对全部序列（去冗余后）提取无序区和有序区。无序区包括 112 个全部无序区（如 DisProtDP00001，108 个氨基酸都是无序的）以及非全部无序蛋白质（蛋白质中含有无序片段）序列中的各条无序区；无序区的氨基酸总数为 64 243，约占固有无序蛋白质氨基酸总数的 28.67%。因此可以看出：固有无序蛋白质中有序区的氨基酸数大约是无序区氨基酸数的 3.5 倍。结果表明，固有无序蛋白质的氨基酸在有序区的含量要大大高于无序区，即固有无序蛋白质的大部分组分都是有序部分。

对固有无序蛋白质中的所有无序区及有序区的氨基酸个数和含量进行对比，以分析每种氨基酸的偏好性。通过 Matlab 软件进行处理得到了固有无序蛋白质中的无序区和有序区的所有氨基酸含量及差值，详见表 1。

表 1　固有无序蛋白中有序区及无序区的氨基酸含量对比

氨基酸种类	无序区的氨基酸含量(%)	有序区的氨基酸含量(%)	有序区和无序区氨基酸含量的差值(%)
Gly	7.138 521	6.651 194	－0.487 327 00
Ala	7.961 957	7.337 648	－0.624 309 10
Val	5.399 810	6.499 442	1.099 631 99
Ile	3.220 584	5.057 353	1.836 768 93
Leu	6.568 809	8.882 839	2.314 029 60
Phe	2.395 592	3.451 015	1.055 423 66
Pro	7.107 389	5.488 507	－1.618 882 10
Met	1.797 861	2.298 594	0.500 732 82
Trp	0.619 523	1.113 144	0.493 621 64
Cys	0.837 445	1.662 575	0.825 129 99
Ser	9.123 173	7.510 823	－1.612 349 50
Thr	5.7624 96	5.752 734	－0.009 761 80
Asn	3.709 354	4.341 442	0.632 088 09
Gln	5.146 086	4.655 657	－0.490 428 70
Tyr	2.095 170	2.819 014	0.723 843 71
His	1.849 229	2.174 068	0.324 839 58
Asp	6.727 581	5.451 908	－1.275 673 10
Glu	9.901 468	7.316 670	－2.584 797 50
Lys	8.060 022	6.411 069	－1.648 953 10
Arg	4.577 931	5.124 303	0.546 371 94

由表 1 对数据进行处理的结果看出：有序区 Val、Ile、Leu、Phe、Trp、Asn、Tyr、Met、His、Cys、Arg 氨基酸含量较无序区多，表明这 11 种氨基酸比较有利于有序结构的形成；无序区 Gly、Ala、Pro、Ser、Thr、Gln、Asp、Glu、Lys 氨基酸含量较有序区多，表明这 9 种氨基酸比较有利于无序结构的形成。20 种氨基酸的平均百分含量为 5%，其中有序区的 Leu 含量为 8.88%，比无序区高 2.31 百分点；无序区的 Glu 含量为 9.90%，比有序区高 2.58 百分点。

Radivojac 对无序蛋白质中有序区与无序区各种氨基酸含量的差值也进行过研究，也认为不同的氨基酸残基具有不同的促进无序和有序结构形成的倾向：Gly、Trp、Tyr、Ile、Phe、Val、Leu、His、Thr、Asn 比较有利于有序结构的形成；Asp、Met、Lys、Arg、Ser、Glu、Pro、Gln 有利于无序的形成；其他残基的作用则比较中性[7]。

将本研究的结果与 Radivojac 的结果对比可以发现：比较有利于有序结构形成的相同氨基酸有 Val、Ile、Leu、Phe、Trp、Asn、Tyr、His；比较有利于无序结构形成的相同氨基酸有 Pro、Glu、Ser、Gln、Asp、Lys。可见统计结论大体相同，不同之处有：Radivojac 的结果认为 Gly 和 Thr 比较有利于有序结构的形成，而笔者的结果认为 Gly 和 Thr 比较有利于无序结构的形成；Radivojac 认为 Met、Arg 比较有利于无序结构的形成，而笔者的结果认为 Met、Arg 比较有利于有序结构的形成。除此之外，本研究的结果还认为，Cys 有利于有序结构的形成，Ala 有利于无序结构的形成，而 Radivojac 的工作没有得出这一结论，推测出现误差的主要原因是由于数据库更新很快，使得研究数据有所变化，其次是所用统计软件之间也有不同。但是大部分结论是相同的，由此可以认为，氨基酸 Val、Ile、Leu、Phe、Trp、Asn、Tyr、His 具有形成有序结构的偏好性；氨基酸 Pro、Ser、Gln、Asp、Lys 具有形成无序结构的偏好性。

3　结论

本研究以 DisProt 数据库中的固有无序蛋白质为研究对象，先通过程序 CD－HIT 对数据进行去冗余处理，然后利用编程软件 Matlab7.0 对数据进行统计而得到新的数据，再对数据进行分析。结果表明：氨基酸 Val、Ile、Leu、Phe、Trp、Asn、Tyr、His 具有形成有序结构的偏好性；氨基酸 Pro、Ser、Gln、Asp、Lys 具有形成无序结构的偏好性。

无序蛋白质具有独特的氨基酸组成特点，这些独特的氨基酸序列决定了其无序的结构。无序蛋白质的研究将促进人们重新认识蛋白质的结构和功能关系，也将为蛋白质的全新设计和疾病的治疗提供新的思路。相信随着研究数据的增加，对固有无序蛋白质的研究将更深入和全面，从而能够进一步加深对这类蛋白质的认识。

参考文献：

[1] Uversky V N. Natively unfolded proteins：A point where biology waits for physics[J]. Protein Science，2002，11(4)：739－756.

[2] Dunker A K，Obradovic Z，Romero P，et al. Intrinsic protein disorder in complete genomes[J]. Genome Informatics，2000，11：161－171.

[3] Dunker A K，Oldfield C J，Meng J，et al. The unfoldomics decade：an update on intrinsically disordered proteins[J]. BMC Genomics，2008，9(S2)：12－18

[4] Sickmeier M，Hamilton J A，LeGall T，et al. DisProt：the database of disordered proteins[J]. Nucleic Acids Research，2007，35(S1)：786－793.

[5] Li W，Godzik A. Cd－hit：a fast program for clustering and comparing large sets of protein or nucleotide sequences[J]. Bioinformatics，2006，22(13)：1658－1659.

[6] Li W，Jaroszewski L，Godzik A. Clustering of highly homologous sequences to reduce the size of large protein databases[J]. Bioinformatics，2001，17(3)：282－283.

[7] 黄永棋，刘志荣. 天然无序蛋白质：序列－结构－功能的新关系[J]. 物理化学学报 2010，26(8)：2061－2072.