2 Perlgut L E, Byers D L, Jope R S *et al*. Formation of triple-stranded bovine DNA *in vitro*. *Nature*, 1975; **254**: 86
3 Raae A J, Kleppe K. T4 Polynucleotide ligase catalyzed joining on triple-stranded nucleic acids. *Biochemistry*, 1978; 17(14): 2939
4 Lee J S, Johnson D A, Morgan A R. Complexes formed by (pyrimidine)$_n$·(purine)$_n$ DNAs on lowering the pH are three-stranded. *Nucl Acids Res*, 1979: 6(9): 3073
5 Letai A G, Palladino M A, Fromm E *et al*. Specificity in Formation of Triple-stranded nucleic acid helical complexes: Studies with agarose-linked polyribonucleotide affinity columns. *Biochemistry*, 1988; 27(26): 9108
6 Mirkin S M, Lyamichev V I, Drushlyak K N *et al*. DNA H-form requires a homopurine-homopyrimidine mirror repeat. *Nature*, 1987; **330**: 495
7 白春礼,叶坚,龚立三等. DNA 变异结构的扫描隧道显微镜研究,科学通报, 1990; (24): 1841
8 Pilch D S, Levenson C, Shafer R. Structural analysis of the $(dA)_{10}$·$2(dA)_{10}$ triplex helix. *Proc Natl Acad Sci USA*, 1990; 8: 1942
9 Rajagopal P, Feigon J. NMR studies of triple-stranded formation from the homopurine homopyrimidine deoxyoligonucleotides d(G-A)$_4$ and d(T-C)$_4$. *Biochemistry*, 1989; 28: 7859
10 Arnott S, Seising E. Structures for the polynucleotide complexes poly (dA)·poly(dT) and poly(dT)·poly(dA)·poly(dT). *J Mol Biol*, 1974; 88: 509
11 Griffith J, Borter C, Christiansen G *et al*. The Structure of three stranded joints catalyzed by the recA protein. In: Richardson C C *et al*. eds, *UCLA symposia on molecular and cellular biology*, new series, vol. **127**: *Molecular mechanisms in DNA replication and recombination*, New York: Wiley-Liss, 1990: 105—114
12 Htun H, Dahlberg J. E. Topology and formation of triple-stranded H-DNA. *Science*, 1989; 243: 1571
13 Gago F, Richards W G. One left-handed strand in DNA-oligonucleotide complexes. *FEBS Letters*, 1989; 242(2): 270
14 Arnott S, Bond P J, Seising E *et al*. Models of triple-stranded polynucleotides with optimised stereochemistry. *Nucl Acids Res*, 1976; 3(10): 2459
15 Strobel S A, Dervan P B. Site-specific cleavage of a yeast chromosome by oligonucleotide-directed triple-helix formation. *Science*, 1990; **249**: 73
16 Cooney M, czernuszewicz G, Postel E H *et al*. Site-specific oligonucleotide binding represses transcription of the human c-myc gene *in vitro*, *Science*, 1988; 241: 4566
17 Hayes S, Szybalski W Control of short leftward transcripts for the immunity and ori regions in induced coliphage lambda. *Mol Gen Genet*, 1973; **129**: 275

从一级结构预测蛋白质稳定性*

——Guruprasad, Reedy 和 Pandit 方法

郭宇立 倪逸声
(北京大学生物系,北京 100871)

提 要

介绍了一种从一级结构预测蛋白质稳定性的方法。Guruprasad, Reedy 和 Pandit 对 32 种稳定蛋白质和 12 种不稳定蛋白质进行了统计分析,发现存在这样一些二肽,它们在稳定的和不稳定蛋白质中的出现频率是明显不同的。通过一系列的统计学计算处理,计算出所有 400 种二肽各自对蛋白质稳定性(或不稳定性)的影响大小,给它们设计了一个二肽不稳定性权值 (DIWV)。对一个给定的蛋白质,与它的序列长度相一致的这些 DIWV 的加和能帮助区分不稳定蛋白质和稳定蛋白质。这种方法对如何提高蛋白质的稳定性具有一定的指导意义。

我们根据 Guruprasad 等人的方法计算了几个已知序列的蛋白质的稳定性指数,并由此推出它们的稳定性。

收稿日期: 1992-04-04 修回日期: 1992-07-13
* 863 计划资助项目。

关键词　蛋白质稳定性，稳定性预测，一级结构

1　预测蛋白质的稳定性是理论和实践上的重要课题

自然界中的蛋白质，在生物体内的稳定性很不相同.蛋白质制品用于医疗和工业生产时，它的稳定性是一个重要指标．随着蛋白质工程技术的产生和应用，人们希望能将一些具有高效生物活性但不稳定的蛋白质通过改造提高稳定性．这对理论和实践都有重要意义．长效药物能大大降低用药量，稳定的酶制剂可应用于工业生产，降低生产成本．蛋白质的稳定性与其结构密切相关，蛋白质的三级结构是由一级结构决定的．所以，人们希望能通过对蛋白质一级结构与其稳定性之间关系的研究找到它们之间的有机联系．通过对已知在生物体内半衰期的各类蛋白质序列的统计分析，提出了一系列阐述蛋白质一级结构与其稳定性关系的假说．研究的情况表明，构象完整程度[1]、蛋白质的序列特性、整体特点和蛋白质在细胞中的位置[2-4]对决定细胞内蛋白质的稳定性很重要．此外，N末端[5]，蛋白质与水分子的关系[6]，蛋白质分子中的构象[7,8]，对蛋白质的稳定性也有重要影响．

2　蛋白质的一级结构贮藏着稳定性的信息

Rogers 等人通过对不同蛋白质序列中各种氨基酸出现频率的统计分析，提出了 PEST 假说[9]．这个假说指出，在不稳定的蛋白质中存在着由 P(Pro)，E(Glu)，S(Ser) 和 T(Thr) 组成的区域．这促使 Guruprasad，Reedy 和 Pandit 对各类蛋白质中氨基酸的出现频率进行了进一步的统计研究．他们按照 Rogers 等人的分类，选择了 32 种生物体内半衰期大于 16h 的蛋白质作为稳定的蛋白质，12 种生物体内半衰期小于 5h 的蛋白质作为不稳定的蛋白质(表 1)，对它们的序列进行了分析．他们指出，PEST 假说并不全面，只是某些氨基酸出

表 1　所研究蛋白质的半衰期和它们的稳定性指数 (II)

编号	蛋白质	半衰期 (h)	II
	(A) 稳定蛋白质		
1	ADK	133	29.4
2	ADH	139	17.4
3	AAT	66	36.3
4	CAI	72	25.1
5	CPA	137	21.5
6	CAT	80	27.5
7	CHY	50	15.3
8	CIS	128	20.8
9	CCC	26	11.4
10	AAD	118	34.4
11	DPA	80	14.2
12	DHF	97	27.4
13	ELA	46	23.0
14	FER	61	23.8
15	GPD	84	16.1
16	HEM	209	7.0
17	HEM	209	6.1
18	ABP	65	23.1
19	LDH	171	25.6
20	LYS	16	19.9
21	MYO	127	9.1
22	PAR	41	22.8
23	PGK	207	22.2
24	PLA	78	15.1
25	PYK	181	21.2
26	RNA	61	52.2
27	SOD	41-200	7.0
28	STI	40	27.4
29	SUB	84	12.0
30	THI	155	5.6
31	PPI	122	19.7
32	TRY	44	20.6
	(B) 不稳定蛋白质		
33	EIA	0.5	100.3
34	α-酪蛋白	2-5	57.7
35	β-酪蛋白	2-5	96.5
36	c-fos	0.5	78.8
37	c-myc	0.5	92.2
38	v-myb	0.5	74.5
39	HSP70	1-2	44.1
40	HMG-CoA	1.5-3	54.5
41	ODC	0.5	52.3
42	P730	1.9	50.4
43	PAT	0.5	53.9
44	p53	0.5	70.0

现频率不同的部分反映，而且包含着错误的成分．但 Guruprasad 等人同时注意到 PEST 假说中提到的与带正电荷的氨基酸侧面相接的倾向于聚集的负电荷区域的存在．这一点提示 Guruprasad 等人意识到蛋白质的稳定性很可能是由特定序列的某些氨基酸的排列决定的．他们进一步意识到形成序列的最小单位——二肽在蛋白质中的出现频率可能是一个判断蛋白质稳定性的有效因子．他们对以上 12 种不稳定和 32 种稳定的蛋白质进行了统计分析，搜寻了所有 400 种可能的二肽在两种蛋白质中的出现频率．统计分析的结果表明，存在这样一些二肽，它们在不稳定的蛋白质和稳定的蛋白质中的出现频率是明显不同的．Guruprasad, Reedy 和 Pandit 通过一系列的统计学计算处理，根据所有 400 种二肽在这两组蛋白质中的不同出现情况，计算出它们各自对蛋白质稳定性(或不稳定性)的影响大小，给它们设计了一个不稳定性权值．对一个给定的蛋白质，与它的序列长度相一致的这些二肽权值的加和能帮助区分不稳定蛋白质和稳定蛋白质[10]．Guruprasad, Reedy 和 Pandit 的方法（以下简称 GRP 法）与以往的假说相比有许多优点：简便，能定量，不要进行复杂的计算，预测精度较高，有一定的指导意义．

3 GRP 方法

Guruprasad 等人通过对蛋白质序列的统计分析处理提出了 GRP 方法，GRP 法的统计学处理是这样进行的：先对以上两组蛋白质序列进行统计分析，统计出所有 20 种氨基酸和 400 种二肽各自在两组蛋白质中的出现频率．由下列方程式计算出两组蛋白质中分别的二肽预期出现频率

$$[N^e(xy)]:N^e(xy) = [N^0(x)/T]*[N^0(y)/T]\sum_{x=1}^{20}\sum_{y=1}^{20} \cdot N^0(xy) \cdots\cdots\cdots\cdots (1)$$

其中 $N^0(xy)$ 为观察到的二肽 xy 的出现频率，$N^0(x)$ 和 $N^0(y)$ 分别是氨基酸 x 和 y 在该组蛋白质中的出现频率．T 为该组蛋白质的总氨基酸数目．

再根据下式计算出每一组蛋白质中二肽 (xy) 的实际出现频率与预期出现频率的 χ 平方值 $[\chi^2(xy)]$:

$$\chi^2(xy) = [N^0(xy) - N^e(xy)]/N^e(xy) \cdots (2)$$

每一组蛋白质的 χ^2 平均值由下式计算:

$$\chi^2 = (1/400)^{400}\sum_{xy=1}^{400}\chi^2(xy) \cdots (3)$$

χ^2 就作为对每一组蛋白质筛选有意义的二肽的置信界限．

再由下式计算出 $\chi^2_{us-s}(xy)$:

$$\chi^2_{us-s}(xy) = [N^0_{us}(xy) - N^0_s(xy)]^2/N^0_s(xy) \cdots\cdots\cdots\cdots (4)$$

通过上面的计算处理，根据下面的三个条件挑选出三组二肽:

满足 $\chi^2_{us}(xy) \geqslant \chi^2_{us}$ 的二肽，它们对蛋白质的不稳定性有意义，记为 us 组;

满足 $\chi^2_s(xy) \geqslant \chi^2_s$ 的二肽，它们对蛋白质的稳定性有意义，记为 s 组;

满足 $\chi^2_{us-s}(xy) \geqslant \chi^2_{us-s}$ 的二肽，它们也是导致蛋白质不稳定性的因素，记为 us-s 组．

所有这三套二肽中，用下式计算出每一种二肽可能的出现频率 $P(xy)$:

$$P(xy) = N^0_{(xy)}/N^e_{(xy)} \cdots\cdots (5)$$

以上三套二肽中的每一套进一步根据 $P(xy)$ 与平均数的显著差别分为两个子集．从所有三套二肽中得到的相应的可能出现频率被用来订出列于表 2 的条件．再将满足表 2 中每一个条件的二肽分别根据每个二肽的 χ^2 值和

表 2 分类依据权值

条件		不稳定性权值
	如果二肽满足以下条件:	
1	$P_{us} \geqslant 1.50$ and $P_s < 1.50$	+13.34
2	$P_{us} < 1.50$ and $P_s \geqslant 1.50$	−1.88
3	$P_{us} \leqslant 0.64$ and $P_s > 0.64$	−6.54
4	$P_{us} > 0.64$ and $P_s \leqslant 0.64$	+24.68
5	$P_{us-s} \geqslant 1.50$	+20.26
6	$P_{us-s} \leqslant 0.64$	−7.49
7	不满足以上条件	+1.0

它的 $P(xy)$ 值进行了分类. 对应于每一个条件的不稳定性权值通过将满足该条件的所有二肽 (xy) 的 $N^0_{(xy)}$ 加和处理后得到. 对 ith 条件的影响因子 IF_i 从下式得到:

$$IF_i = (V_{usi} - V_{si})/V_{si} \cdots\cdots (6)$$

其中 V_{usi} 和 V_{si} 分别为在不稳定蛋白质和稳定蛋白质中满足 ith 条件的二肽的出现频率的校正值. 对 ith 条件的影响因子进行了处理,并通过下式得到不稳定性权值 IWV_i:

$$IWV_i = 2 + (IF_i/|LIF|) \cdots (7)$$

其中 LIF 是观察到的最低影响因子, 对应每一个条件的不稳定性权值列于表 2 中.

每一个二肽对不稳定性的影响对应于这个二肽所满足条件(可能是多个条件)的不稳定权值加和而成,称作二肽的不稳定性权值 (DIWV).

如二肽 CM, 既满足表 2 中条件 1, 又满足条件 5,所以 $DIWV_{CM} = 13.34 + 20.26 = 33.6$

所有 400 种组合的二肽的 DIWV 列在 GRP 真值表[10]中. 根据 GRP 真值表中的 DIWV, 通过下式可计算出蛋白质的不稳定性指数 (II):

$$II = (10/L) \sum_{i=1}^{L-1} DIWV(x_i y_{i+1}) \cdots (8)$$

其中 x_iy_{i+1} 是二肽, L 是序列长度, 10 是换算系数. 在统计分析中使用的两组 44 种蛋白质的不稳定性指数列于表 1 中的第 4(II) 栏.

4 GRP 方法是预测蛋白质稳定性的有效方法

稳定的与不稳定的两套蛋白质中二肽的实际和预期的出现频率有助于确定分别在这两套蛋白质中占据优势的二肽. GRP 真值表指出 400 种可能的二肽中有 81 种(约 20%) 的 DIWV > 1, 它们可能与蛋白质不稳定性有关,另有 70 种(约 18%) 的 DIWV < 1, 它们可能对稳定性有贡献, 其余 249 种二肽 DIWV = 1. 目前尚不知道这些二肽的出现(或缺失)是如何使蛋白质稳定或易于降解的. 但无论如何, 它们组合出现的净效应通过不稳定性指数显示它能作为决定蛋白质稳定性的有用指标.

从表 1 可以看出, 所有不稳定的蛋白质 $II > 40$, 相反,所有稳定的蛋白质(除了 RNaseA 这个例外) $II < 40$. RNaseA 有相对较高的 II 值 (52.2) 的原因可能是 RNaseA 中有四对二硫键, Cys 基团之间的高交联度有效地补偿了二肽的影响, 使蛋白质趋于稳定. 统计分析中所用的蛋白质均无如此之高的二硫键,因此,在已知序列的情况下,可以根据 $II <$ 40 还是 >40 直接判断一个蛋白质稳定与否. Guruprasad 等人用 II 值成功地验证了另外 22 种蛋白质的稳定性.

Guruprasad 等人将 GRP 方法与早先提出的其它假说进行了比较,特别是同 PEST 假说进行了比较. 对几个实例的分析,显示 GRP 方法具有比 PEST 假说高得多的可靠性. Na^+, K^+-ATPase 中有三个 PEST 区域. 根据 PEST 假说,它是不稳定的,但 II 值显示蛋白质是稳定的,这与实验结果一致(它的 II 值为 31.67, 半衰期为 9—40h). 另外, Rogers 等人报道 β-酪蛋白在其 N 末端(残基 1—25) 只有一个具有 RF 值的 PEST 区域, 删除这一区域将导致残留的蛋白质变得稳定, 但 II 值则显示相反的结论.

Guruprasad 等人用 GRP 方法对点突变改变蛋白质的稳定性进行了下列可能正确的推测. GRP 真值表中包括了高正值和负值的二肽(最高的 58.24, 最低的 −15.91) 在不稳定的蛋白质中进行点突变有可能导致一个高正值的二肽变成一个负值的二肽,从而使蛋白质稳定。为产生足够大的 II 值改变, 一个以上氨基酸的替换可能是必要的. 但只有那些能提高蛋白

表 3　应用 GRP 方法计算的几个蛋白质的 II 值

蛋白质	II 值	稳定性
表皮生长因子	79.0	不稳定
胰高血糖素	56.0	不稳定
溶菌酶	18.7	稳定
猪胰蛋白酶	32.4	稳定
大鼠胰蛋白酶	25.8	稳定

质稳定性又不影响功能的替换才是有用的.

正如 Guruprasad 等人指出的，除了以上与序列有关的性质外，一些因素如二硫桥的出现、配基的结合、蛋白酶的识别机制等，与蛋白质在生物体内的稳定性有关，因此，在生物体内表现出来的蛋白质的稳定性是由数个这样的因素综合作用的结果. Guruprasad 等人的工作指出序列特异性因素是一个在决定蛋白质稳定性上起重要作用的有意义的因素之一. 他们的研研成果有助于开辟一个新的前景，关于二肽的特殊性质的知识有望应用于对现存蛋白质的改造以及设计新的具有理想稳定性的蛋白质.

我们应用 GRP 方法计算了几个已知序列的蛋白质的 *II* 值，并由此推出它们的稳定性(见表 3):

由 *II* 值显示，大鼠的胰蛋白酶比猪的胰蛋白酶更稳定，这与事实是相符的. Guruprasad 等人指出有可能通过替换几个氨基酸降低 *II* 值提高蛋白质的稳定性，这是一个例证.

5 GRP 方法有待进一步完善

GRP 方法也还存在一些尚待解决的问题: a. 能进行 *II* 值计算的蛋白质是由一条肽链组成的. 而如胰岛素等由两条以上肽链组成的蛋白质如何进行 *II* 值计算还有待于进一步探讨. b. 二硫桥丰度太高的蛋白质不适于 *II* 值计算. c. 我们从表 1 中可以看到，*II* 值与半衰期(即稳定性)之间并无线性关系. 但从总体上看，仍存在以下趋势: *II* 值越低，蛋白质越趋向稳定; 蛋白质的结构越相似，它们之间根据 *II* 值推断的稳定性的可比性就越高. 尽管存在以上缺点，由统计学方法推导出的 GRP 方法仍然包含着客观规律，很有启发性.

本校生化专业班的宋洪军同学参加了部分计算工作，戴勇、陈一友同学参加了讨论，特此表示感谢.

参 考 文 献

1 Goldberg A L, St John A C. Intracellular protein degradation in mammalian and bacterial cells (Part 2). *Annu Rev Biochem*, 1976; 45: 747

2 Rechsteiner M, Rogers S, Rote K. Protein structure and intracellular stability. *Trends Biochem Sci*, 1987; 12: 390

3 Pontremli S, Melloni E. Extralysosomal protein degradation. *Annu Rev Biochem*, 1986; 55: 455

4 Finley D, Varshavsky A. The Ubiquitin system: functions and mechanisms. *Trends Biochem Sci*, 1985; 10: 343

5 Bachmair A, Finley D, Varshavsky A. In vivo half-life of a protein is a function of its amino-terminal residue. *Science*, 1986; 234: 179

6 Vriend G. Berendsen H J C, van der Zee J Rob *et al.* Stabilization of the neutral protease of Bacillus stearothermophilus by removal of a buried water molecule. *Protein Engineering*, 1991; 4: 941

7 Ferraz C, Heitz F, Widada J, S *et al.* Conformational stability of human skeletal tropomyosins modified by site-directed mutagenesis. *Protein Engineering*, 1991; 4: 561

8 Wilson K S, Vorgias C E, Tanaka I *et al.* The thermostabilty of DNA-binding protein HU from baciili. *Protein Engineering*, 1990; 4: 11

9 Rogers S, Wells R, Rechesteiner M. Amino acid sequences common to rapidly degraded proteins: The PEST hypothesis. *Science* 1986; 234: 364

10 Guruprasad K, Reddy B V B, Pandit M W. Correlation between stability of a protein and its dipeptide composition: a novel approch for predicting in vivo stability of a protein from its primary sequence. *Protein Engineering*, 1990; 4: 155