第 22 卷第 3 期
2003 年 5 月

无锡轻工大学学报
Journal of Wuxi University of Light Industry

Vol. 22 No. 3
May 2003

文章编号:1009-038X(2003)03-0007-05

用氨基酸组分预测含顺式肽键蛋白质

宋江宁, 李炜疆
(江南大学 工业生物技术教育部重点实验室,江苏 无锡 214036)

摘 要:以含有 820 个蛋白质结构数据的 SLCTBASE 数据库作为样本库,根据蛋白质序列氨基酸百分含量与顺式肽键的关联信息,构建了基于序列整体氨基酸含量的顺式构象势函数. 利用此函数拟合得出的 20 种氨基酸的顺式构象因子值,进行两类蛋白质序列的分类和预测. 训练集和测试集相同时和不相同时的预测准确率分别可达 66% 和 64%.

关键词:顺式肽键;氨基酸百分含量;蛋白质二级结构;统计关联
中图分类号:Q 61 **文献标识码**:A

Predicting Proteins Containing Cis Peptide Bonds Using Amino Acid Compositions

SONG Jiang-ning, LI Wei-jiang
(The Key Laboratory of Industrial Biotechnology, Ministry of Education, Southern Yangtze University, Wuxi 214036, China)

Abstract: Based on the sampling database SLCTBASE with 820 protein spatial structure data, the cis peptide potential based on the correlation information between protein sequence amino acid content and cis peptide was successfully constructed. Factor values of twenty amino acids for cis peptide conformation were used to classify and predict the two protein datasets having cis peptide bonds and without cis peptide bonds. The predicting accuracy was as high as 66% when train datasets and prediction datasets are the same, while the accuracy was 64% in the case of using different train and prediction datasets.

Key words: cis peptide bond; amino acid composition; protein secondary structure; statistical correlation

对于蛋白质多肽链而言,由于连接相邻两个氨基酸残基的 C—N 键具有部分双键的性质,一般情况下,肽键不能够自由旋转. 它的肽键二面角 ω 基本上只能取 2 个值:ω 为 180°(反式构象)和 ω 为 0°. 肽链中的肽键绝大多数都采取反式构象,只有极少数的肽键采用顺式构象,并且多为含有脯氨酸(Proline)的肽键,主要原因是脯氨酸中的四氢吡咯环引起的空间位阻部分消去了反式构象的优势,因而由脯氨酸形成的肽键相对较易跨越两种构象间的势垒而形成顺式构象.

顺式肽键或者反式肽键基本上属于一种空间结构特征,仅从蛋白质数据库中的氨基酸序列无法确定肽键是顺势构象还是反式构象,并且原始蛋白质结构数据库中也没有给出顺式肽键的记录. 根据

收稿日期:2002-10-11; 修回日期:2003-03-11.
基金项目:工业生物技术教育部重点实验室访问学者基金项目资助课题.
作者简介:宋江宁(1978-),男,山东莱阳人,发酵工程博士研究生.
万方数据

蛋白质结构数据库中的原子坐标来计算肽键扭角 ω 的值来判断某一肽键是否为顺式肽键. 序列决定结构,结构决定功能,这是一种基本认可的假设. 蛋白质结构由蛋白质序列决定,从原理上讲,蛋白质序列隐含了蛋白质折叠后的空间结构,理论上可以从氨基酸序列计算出自然折叠的蛋白质结构. 以往的统计分析结果也表明,蛋白质一维氨基酸序列中或多或少地隐藏了蛋白质二级结构信息. 目前对于顺式肽键和反式肽键统计预测方面的研究很少,结果也并不满意.

Frömmel 和 Preissner 发现,反式和顺式脯氨酸肽键周围序列存在着较为明显的特征差异[1],研究了半胱氨酸周围序列(±6 个氨基酸)物理化学特征,总结得到了 6 种不同的模式,使用相同的训练集和测试集时预测脯氨酸顺式肽键可达 75% 的准确率. 不过他们的方法描述不是很清晰,不易准确验证,在更大的数据集上的表现尚有待核实.

Jabs 和 Weiss 等人详细地计算了 43 个非脯氨酸顺式肽键的一些几何参数[2],发现高精度结构数据中顺式肽键的一些几何参数与相应的反式肽键有很大差异,例如顺式肽键的构造显著改变了氨基上 N 原子相关的键角.

Pal 和 Chakrabarti 研究了 PDB 数据库中蛋白质顺式肽键的分布特征——残基偏好性、局部构象、结合氢和其它包含顺式肽键的稳定相互作用[3],通过比较含有脯氨酸和不含有脯氨酸的顺式肽键,发现这两类键在构象、二级结构中的位置以及与分子中心的关系和残基的相对可及性上都存在差异.

含顺式肽键蛋白质序列有着自己不同的序列特征,利用这些序列特征能够在某种程度上预测含顺式肽键的氨基酸序列. 已有的研究表明,脯氨酸顺式肽键与其周围氨基酸序列间存在着一定的关联,在此基础上构建的简单的顺式构象函数,可以将顺式脯氨酸信噪比由 5.4% 提高到 30% 以上[4]. 顺式肽键对蛋白质整体结构都有很大影响,而已有的研究只利用了局部序列的信息,是否可以利用序列整体信息对顺式肽键进行某种预测呢? 作者尝试利用蛋白质中氨基酸组分预测蛋白质结构中是否含有顺式肽键.

1 材料与方法

1.1 数据库

选择一个高精度非同源的蛋白质空间结构样板库 SLCTBASE[5](此库的蛋白质结构测定方法均为 X 射线衍射,测量精度均好于 2.5 Å)作为研究的样本数据库. 库中的结构数据来自于 PDB 数据库[6],序列同源性都低于 25%. 总共含有 820 条蛋白质多肽链,188 551 个氨基酸残基,其中共有 523 条顺式肽键,见表 1.

含有顺式肽键的数据集有 337 条蛋白质多肽链(95 154 个氨基酸残基),脯氨酸形成的顺势肽键有 419 条,占到总顺式肽键数目的 80.1%;不含有顺式肽键的数据集有 483 条蛋白质多肽链(93 376 个氨基酸残基).

表 1 SLCTBASE 数据库中含有顺式肽键的 PDB 文件

Tab.1 The PDB Files containing cis peptide bonds in SLCTBASE

PDB								
1A2ZA	1AUOA	1ECD	1HPM	1LTSA	1PVC2	1U9AA	2GSAA	5HPGA
1A3K	1AUVA	1ECEA	1HSBA	1LTSD	1PYTB	1UAE	2GSTA	5NLL
1A6O	1AV4	1ECFA	1HSBB	1LUCA	1QNF	1UCH	2I1B	6LDH
1A8E	1AVMA	1EDE	1HSSA	1LUCB	1QORA	1UDC	2ILK	7AHLA
1A8P	1AX4A	1EDG	1HTRB	1MHYG	1RA9	1UKZ	2KAUC	7RSA
1A8UA	1AYL	1EFVB	1HUCB	1MOLA	1RCY	1UMAH	2MCM	8FABA
1A9S	1BAM	1ESC	1HXN	1MAT	1REQA	1VCAA	2MHR	8FABB
1A17	1BFD	1EUR	1HYXH	1MHLC	1REQB	1VHH	2MPRA	8RUCA
1AAZA	1BDB	1EZM	1HYXL	1MKAA	1RFS	1VID	2MSBA	8TLNE
1AC5	1BGP	1FINA	1IAKA	1MLA	1RGEA	1VNC	2MTAH	
1ADEA	1BLSA	1FINB	1IAKB	1MPGA	1RHS	1VPSA	2NADA	
1ADOA	1BMDA	1FJMA	1IDK	1MPP	1RKD	1VRTA	2OVO	

万方数据

续表 1

PDB							
1AERA	1BNDB	1FMTA	1IDO	1MSK	1RLAA	1VSD	2PGD
1AERB	1BPYA	1FNC	1ILR1	1NAR	1RMG	1WHTA	2PKAB
1AF7	1BTL	1FROA	1IMBA	1NBAA	1ROM	1WBA	2RN2
1AGJA	1CEC	1FUA	1INP	1NDH	1RSY	1WDNA	2SGA
1AH7	1CELA	1FUPA	1IOV	1NIF	1RVAA	1WGJA	2SICI
1AI3	1CEM	1FUS	1IVYA	1NLS	1SFP	1WHO	2SICE
1AIJH	1CGT	1FVKA	1JDBB	1NNC	1SFTA	1XAA	2SIL
1AIJM	1CHD	1G3P	1JDBC	1NSCA	1SMD	1XER	2TGI
1AIJS	1CHKA	1GAI	1JDW	1NSJ	1SNC	1XGSA	2TRCP
1AIR	1CLC	1GAL	1JER	1NULA	1SOXA	1XJO	2TRXA
1AISA	1CNSA	1GCB	1JETA	1OBR	1SRA	1XNB	2VPFA
1AJSA	1CNV	1GDHA	1JFRA	1OBWA	1SVB	1XYZA	2WEA
1AJSB	1CPO	1GDOA	1JPC	1ONEA	1TABI	1YAIA	3BCT
1AK1	1CPT	1GEN	1KNB	1OPS	1TCA	1YTBA	3BTOA
1AKO	1CSEE	1GIFA	1KOBA	1OPY	1TDE	1ZIN	3CHY
1ALY	1CTF	1GLQA	1KPF	1OYC	1TDTA	1ZXQ	3DFR
1AOFA	1CTN	1GNWA	1KSIA	1PAZ	1TF4A	2ACK	3FIB
1AOA	1CUK	1GOF	1KTE	1PBE	1TGXA	2AAA	3FRUA
1AOCA	1CVL	1GP1A	1LAM	1PBGA	1THG	2AYH	3FRUB
1AOEA	1DANH	1GPMA	1LAUE	1PGS	1THV	2BBKH	3GRS
1AORA	1DANT	1GSA	1LBEA	1PHG	1TIF	2CAE	3MINA
1AOZA	1DAPA	1GYM	1LCL	1PHP	1TIIA	2CBA	3MINB
1APA	1DBQA	1HA1	1LDG	1PLC	1TIID	2CMD	3NUL
1APYA	1DIN	1HCL	1LED	1PNKB	1TMKA	2CTC	3PTE
1AQ0A	1DKZA	1HCZ	1LGHA	1PPN	1TMY	2EBN	3TGL
1AQ6A	1DORA	1HFC	1LIT	1PRCC	1TN3	2ERL	4AAHA
1AQZA	1DPE	1HGXA	1LKI	1PRCH	1TTQA	2GEP	4HTCH
1AUIA	1DRW	1HIAB	1LST	1PRCM	1TTQB	2GMFA	4SGBE
1AUK	1EAYA	1HPLA	1LT5D	1PUD	1TVDA	2GSQ	5CROO

1.2　统计方法

对数据库中的某一蛋白质 k,定义其特征参数 Q_k 如下:

$$Q_k = \begin{cases} +1, \text{如果蛋白质 k 含有顺式肽键,} \\ -1, \text{如果蛋白质 k 不含有顺式肽键.} \end{cases} \tag{1}$$

通过一个简单的与蛋白质序列氨基酸组分百分含量相关的线性方程来求得相应的 Q_k 值,即:

$$Q_k \approx \sum p_a^{(k)} V_a \tag{2}$$

V_a 表示 a 这个氨基酸的百分含量对于序列中顺式肽键构象的影响,v_a 对于所有的蛋白质而言为一常数;$p_a^{(k)}$ 表示 a 这种氨基酸在某一蛋白质序列中的百分含量,a = 1,2,3,…,20.

为求得常数 V_a 的值,最小化下式:

$$Z = \sum_k (Q_k - \sum V_a p_a^{(k)})^2 \tag{3}$$

k 表示对于所有的序列都求和,要求其极小值,对上式取偏导,令 $\partial Z / \partial v_b$,化简,得到:

$$\sum_a (\sum_k p_a^{(k)} p_b^{(k)}) V_a = \sum_k Q_k p_b^{(k)} \tag{4}$$

k 遍历数据集中所有的蛋白质序列,解方程(4),可以得到常数 V_a 的值.

对于每个蛋白质序列,有一 V_a 值,代入方程(2),就可以求出相应的 Q_k 值. 用 Q_k 值去拟合这两类蛋白质序列,在此,设定阈值等于 0,如果 $Q_k > 0$,就预测此序列含有顺式肽键;如果 $Q_k < 0$,就预

万方数据

测此序列没有顺式肽键.

2 结果与讨论

2.1 样本数据同时作为训练集和测试集

将样本库中的这两类数据集既作为训练集同时又作为测试集,看分类预测效果如何.

计算得到20种氨基酸的 V_a 值见表2.

表2 20种氨基酸拟合的 V_a 值(训练集和测试集相同)

Tab.2 V_a values of 20 amino acids with the same train set and test set

氨基酸	拟合的 V_a 值	氨基酸	拟合的 V_a 值	氨基酸	拟合的 V_a 值
F	0.766	L	−1.029	I	−0.348
M	2.990	V	−0.832	S	−1.132
P	11.659	T	1.335	A	−0.381
Y	4.957	H	0.511	Q	−5.191
N	−0.240	K	−2.412	D	−1.513
E	−2.539	C	−5.129	W	−0.250
R	−1.997	G	0.268		

拟合得到的 V_a 值的绝对值越大,说明此种氨基酸对于形成相应的肽键的影响(贡献)越大.从表2可以看出,对于顺式肽键构象作出贡献的有Phe,Met,Pro,Thr,Tyr,His,Gly等7种氨基酸,其中Met,Pro,Tyr是3种对于顺式构象影响最大的氨基酸.其余对于反式肽键构象作出贡献的13种氨基酸中,Ser,Asn,Lys,Asp,Glu,Cys,Arg 7种影响更大些.

求出氨基酸的 V_a 值后,做进一步的预测,研究单纯用蛋白质序列的氨基酸百分含量区分含有顺式和反式肽键的两类蛋白质序列的可能性.

预测结果见表3.

表3 不同阈值、相关系数下的预测准确率(训练集和测试集相同)

Tab.3 The prediction accuracy of different thresholds and correlation coefficients with the same train set and test set

阈值 Z_0	相关系数	准确率/%	阈值 Z_0	相关系数	准确率/%
−0.36	0.327	60.2	−0.32	0.308	60.7
−0.28	0.322	62.4	−0.20	0.308	63.8
−0.18	0.310	64.4	−0.12	0.292	65.1
0	0.258	65.2	0.02	0.270	65.9
0.10	0.258	65.7	0.20	0.179	62.4

万方数据

预测的最佳准确率是65.9%,但是相关系数不是最大的;在阈值为0时,准确率为65.2%,相关系数为0.258;相关系数最大是0.327,但是阈值为−0.36,离开0较远.

2.2 样本数据分成2份——训练集和测试集不同

为考察预测分类结果对于所选数据集的依赖性,将样本数据分成两部分——训练集和测试集不同.训练集中含顺式肽键的数据集包括168条蛋白质多肽链,有50 320个氨基酸残基;不含顺式肽键的数据集包括241条蛋白质多肽链,有47 034个氨基酸残基.测试集中含顺式肽键的数据集包括169条蛋白质多肽链,有44 834个氨基酸残基;不含顺势肽键的数据集有242条蛋白质多肽链,46 342个氨基酸残基.

结果见表4,5.

表4 20种氨基酸拟合的 V_a 值(训练集和测试集不同)

Tab.4 V_a values of 20 amino acids with the different train set and test set

氨基酸	拟合的 V_a 值	氨基酸	拟合的 V_a 值	氨基酸	拟合的 V_a 值
F	3.687	L	0.780	I	−1.276
M	−3.049	V	−2.445	S	−1.056
P	11.516	T	0.238	A	0.623
Y	5.819	H	−0.324	Q	−6.185
N	0.905	K	−3.511	D	−0.046
E	−2.258	C	−6.423	W	0.829
R	−2.660	G	−0.300		

表5 不同阈值、相关系数下的预测准确率(训练集和测试集不同)

Tab.5 The prediction accuracy of different thresholds and correlation coefficients with the different train set and test set

阈值 Z_0	相关系数	准确率/%	阈值 Z_0	相关系数	准确率/%
−0.30	0.190	56.4	−0.20	0.184	58.2
−0.10	0.179	59.8	0.00	0.182	61.3
0.08	0.218	63.7	0.10	0.211	63.5
0.20	0.118	60.3	0.30	0.117	60.6

对比表2和表4可以发现,有4种氨基酸对于形成顺式肽键构象作出了较大的贡献:Phe,Pro,Thr和Tyr;不利于形成顺式肽键构象的有Ile,Val,Ser,Asn,Lys,Asp,Glu和Arg,这些氨基酸百分含量越多,则越不容易形成顺式肽键构象.

由表 5 可以看出,训练集和测试集在不同情况下,预测准确率在阈值为 0.08 时,最高可以达到 63.7%,相关系数为 0.218,也是最大;而当阈值等于 0 时,准确率为 61.3%,此时的相关系数为 0.182.

3 结　论

1) 氨基酸序列决定了蛋白质的结构和功能,而序列则首先受到氨基酸组成的约束和限制.通过简单的统计分析预测方法,用序列中 20 种氨基酸的百分含量较为成功地区分了两类含有顺式肽键和反式肽键的蛋白质序列,训练集和测试集相同时和不相同时的预测准确率分别可达 66%和 64%,差别不大,这说明预测结果不依赖于数据集.

2) 使用 20 个参数,通过拟合 20 种氨基酸的 V_a 值来预测和分类两类蛋白质序列,能够达到 66%的预测准确率,说明单纯用氨基酸百分含量来区分两类含顺式肽键和不含顺式肽键实际可行.

3) 20 种氨基酸中,对于顺式肽键构象作出较大贡献的有 Phe,Pro,Thr 和 Tyr 等 4 种氨基酸;不利于形成顺式肽键构象的有 8 种氨基酸 Ile,Val,Ser,Asn,Lys,Asp,Glu 和 Arg.

4) 这一分类的可行性对于进一步预测蛋白质序列中究竟是哪 2 个氨基酸残基之间形成了顺式肽键也有着很大的帮助.

5) 这种预测只是利用了全局的序列信息(global sequence information),而没有考虑周围局部序列提供的信息(local sequence information or flanking sequence information).可以肯定地说,顺式肽键序列周围也蕴含着顺式肽键形成的信息.如果在预测时,还加入形成顺式肽键的局部序列特征信息(单残基与顺式构象的关联、二肽频率与顺式构象的关联)应该可以进一步提高顺式肽键的预测准确率,这种简单的统计预测分析方法还有待进一步的提高和改善.

参考文献:

[1] Frömmel C, Preissner R. Prediction of prolyl residues in cis-conformation in protein structures on the basis of amino acid sequence[J]. **FEBS Letters**,1990,277(1～2):159－163.

[2] Jabs A, Weiss M S, Hilgenfeld R. Non-proline cis peptide bonds in proteins[J]. **J Mol Biol**,1999,286:291－304.

[3] Pal D, Chakrabarti P. Cis peptide bonds in proteins: residues involved, their conformations, interactions and locations[J]. **J Mol Biol**,1999,294:271－288.

[4] 李炜疆,张颖. 脯氨酸顺式肽键与氨基酸序列关联的统计分析[J]. 内蒙古大学学报,2001,32(1):12－16.

[5] 李炜疆,张颖. 蛋白质结构样板库的构建及其总体特征[J]. 内蒙古大学学报,2000,31(5):471－478.

[6] Berman, H M, Westbrook J, Feng Z, *et al*. The Protein Data Bank[J]. **Nucleic Acids Res**,2000, 28:235－242.

(责任编辑 杨 萌)

万方数据