

α -螺旋中 2,3 位氨基酸残基对螺旋结构稳定性的影响

付明哲, 王联结*, 刘选治

(陕西科技大学生命科学与工程学院, 西安 710021)

摘要: 用自编的计算机软件对 967 条非冗余蛋白质结构序列中进行统计分析, 并扩展了 Chou-Fasman 法则, 建立了一种 α -螺旋模体(Motif)的方法。结果发现在四残基 α -螺旋结构中, 第 2 位和第 3 位氨基酸残基共同影响其稳定性。

关键词: 蛋白质; 二级结构; α -螺旋

中图分类号: Q811.4

文献标识码: A

文章编号: 1672-352X (2011)06-0838-04

Impact of 2, 3 α -helical amino acid residues on the stability of the helical-structure

FU Ming-zhe, WANG Lian-jie, LIU Xuan-zhi

(College of Life Science and Engineering, Shaanxi University of Science & Technology, Xi'an 710021)

Abstract: We established a α -helix motif method by using computer software exploited by ourselves to do a statistical analysis of the structures of 967 non-redundant protein sequences. We also extended the Chou-Fasman rules. As a result, we found that α -helix in the second and third amino acid residues commonly affect its structural stability.

Key words: protein; secondary structure; α -helix

伴随着世界第一个蛋白质的三维立体结构的测定和解析, 蛋白质的结构预测也在同一时期开始发展^[1], 并为蛋白质从序列到空间结构的研究架起一座桥梁, 指导蛋白质的从头设计、药物开发和核酸功能研究^[2]。现今的蛋白质结构预测已经成为分子生物学和生物信息学的一个重要研究任务。

从 Chou-Fasman 以各氨基酸残基对某二级结构的倾向性为基础进行蛋白质的二级结构预测开始到现在被广泛接受的机器学习算法和神经网络算法^[3-12], 蛋白质结构预测的准确性不断提高, 但算法与软件设计越来越复杂, 而解释却越来越模糊。用较复杂的算法提高相对较小的几个百分点只会使所得的数据更加难以用生命科学的方法进行解释, 蛋白质结构数据的生物学意义, 即数据所解释的生命现象和所揭示内在规律不能得到体现^[13-14]。当模体(Motif)的概念提出后, 模体识别已经成为一种从生物序列中提取有用信息的有效方法。作者从构建一种 α -螺旋 Motif 出发, 研究 Motif 中各氨基酸残

基对螺旋的稳定性的影响, 利用数据的生物学意义, 来研究蛋白质序列中的所含的规律。

1 材料与方法

1.1 数据库

蛋白质结构信息数据库是进行蛋白质结构、功能与序列信息研究的基础。DSSP 数据库是一个存放蛋白质二级结构分类数据的数据库^[15]。DSSP(的 definition of secondary structure of proteins) 是由 Wolfgang Kabsch 和 Chris Sander 所设计的, 用于对蛋白质结构中氨基酸残基进行二级结构构象分类的标准化算法, 依据 PDB(protein data bank 蛋白质数据库)格式的原子坐标定义蛋白质的三维结构, 依靠静电学定义进行的氢键识别, 并根据氢键模式分为 8 种不同的二级结构构象^[16-17]。

作者所使用的数据库是一套经实验测定的, 其序列之间的同源性小于 25%的非冗余 DSSP 数据库的子集, 共 967 种蛋白质结构序列, 并按照 DSSP

收稿日期: 2011-06-27

作者简介: 付明哲, 男, 硕士研究生。E-mail: fumingzhegk@163.com

* 通讯作者: 王联结, 男, 教授。E-mail: wanglianjie2@yahoo.com.cn

数据库中 8 种不同的二级结构构象分类进行分析。

表 1 DSSP 数据库中的蛋白质二级结构及其标识
Table 1 Secondary structure of protein and its identification in DSSP database

DSSP 标识 DSSP identification	蛋白质二级结构 Secondary structure of protein
G	3 转角螺旋 (亦即 3_{10} 螺旋)。最短长度为 3 个残基。
H	4 转角螺旋 (α 螺旋)。最短长度为 4 个残基。
I	5 转角螺旋 (π 螺旋)。最短长度为 5 个残基。
E	平行的 β 折叠, 或 / 及反平行的折叠形态(延伸链)。最短长度为 2 个残基。
B	独立 β 桥内残基 (一对 β 折叠氢键)。
T	氢键转角 (3、4 或 5 个转角)。
S	弯曲 (唯一非氢键的指定)
空格 Blank	所有不是以上形态的残基, 在 DSSP 都是以空格来指定的

1.2 理论基础

1.2.1 α -螺旋 本文的研究对象是螺旋中的 α -螺旋结构, α -螺旋是蛋白质中最典型、含量最丰富的二级结构类型, 在 DSSP 数据库中标识为“H”, 它最少由四个氨基酸残基组成, 残基侧链外展, 每圈螺旋由 3.6 个氨基酸残基构成。

螺旋结构中除了 α -螺旋之外还有 3_{10} 螺旋与 π 螺旋, 由于 3_{10} 螺旋与 π 螺旋中的氢键与 α -螺旋中的氢键并不是在相同位置的氨基酸残基之间形成, 所以当我们以 α -螺旋结构为主要研究对象时就要排除 3_{10} 螺旋与 π 螺旋的干扰, 即在数据库中的匹配结构仅为“H”。选取 4 个氨基酸残基长度的结构序列, 当它们在数据库中为连续的螺旋结构(即“HHHH”结构)时即为选取的匹配结构。

1.2.2 Chou-Fasman 理论^[18-20] Chou 和 Fasman 于 1974 年对 29 个蛋白质用 X 光衍射得到的数据进行数学统计, 得到了各种氨基酸残基在蛋白质中形成各种二级结构的倾向性因子。

Chou-Fasman 原始参数。P 表示各种二级结构的倾向性因子。

$$P_{i,t} = f_{i,t} / f_t$$

$$\text{其中: } f_{i,t} = n_{i,t} / n_i, f_t = n_t / n_{\text{total}}$$

i : 表示 20 种氨基酸残基, t : 表示 8 种二级结构, $f_{i,t}$: 表示氨基酸残基 i 出现二级结构 t 的概率, f_t : 表示二级结构 t 出现的概率, $n_{i,t}$: 表示氨基酸残基 i 在二级结构 t 中出现的次数, n_i : 表示氨基酸残基 i 出现的总次数, n_t : 表示二级结构 t 中的氨基酸残基个数, n_{total} : 表示被统计的蛋白质中氨基酸残基

的总数。

Chou-Fasman 参数扩展。Chou-Fasman 原始参数的意义是氨基酸残基 i 对二级结构 t 的倾向性因子参数, 所统计的结果是一种氨基酸残基对一种二级结构的正向影响, 当我们把所统计的结构改变时就可以得到这种氨基酸残基对此二级结构的负向影响。公式不变, 但统计参数改变, 螺旋倾向性统计结果详见表 2。

表 2 Chou-Fasman 参数值及优化结果 (H)
Table 2 Parameter values of Chou-Fasman and its optimization

氨基酸残基 Amino acid residue	正向影响因子 Positive influence factor	负向影响因子 Negative influence factor
E	1.40	0.82
M	1.29	0.87
A	1.45	0.79
L	1.33	0.85
K	1.15	0.93
F	1.00	1.00
Q	1.30	0.86
W	1.08	0.96
I	1.06	0.97
V	0.92	1.04
D	0.82	1.08
H	0.86	1.06
R	1.21	0.90
T	0.77	1.11
S	0.76	1.11
C	1.00	1.00
Y	0.96	1.02
N	0.72	1.13
P	0.44	1.26
G	0.46	1.25

通过上表我们可以明显的看出, 在 20 种氨基酸中, E, M, A, L, Q, R 具有明显的螺旋倾向性; D, H, T, S, N, P, G 具有明显的阻碍螺旋形成的倾向性; 而 K, F, W, I, V, C, Y 则在不具有明显的螺旋倾向性时, 同样不会对螺旋的形成产生阻碍作用。

1.3 Motif 统计方法

基于以上理论, 作者设计 1 种 α -螺旋 Motif。在 Motif 中确定第 1 位和第 4 位氨基酸残基, 形成“确定氨基酸残基”+“不确定氨基酸残基”+“不确定氨基酸残基”+“确定氨基酸残基”这样一种 Motif, 并在非冗余数据库中搜索匹配结构, 得到“全 H”和“非全 H”二个子集, 然后对这两个子集进行 Chou-Fasman 统计。计算第 2 位与第 3 位的氨基酸残基 Chou-Fasman 平均值, 当第 1 位氨基酸残基与第 4 位氨基酸残基之间形成螺旋时, 即连续 H 结

构, 统计它们正向影响因子大于负向影响因子的概率, 并用 Chou-Fasman 原始法则进行检验; 当第 1 位氨基酸残基与第 4 位氨基酸残基之间的螺旋结构被破坏时, 即不连续 H 结构, 统计它们正向影响因子小于负向影响因子的概率, 并用 Chou-Fasman 原

始法则进行检验。

2 结果与分析

表 3 是分别以 Chou-Fasman 扩展法则为依据的统计结果。

表 3 Motif 统计结果

Table 3 Statistical result of Motifs

Motif	全 H	非全 H	Motif	全 H	非全 H	Motif	全 H	非全 H	Motif	全 H	非全 H
A**A	71.59%	54.03%	D**M	81.18%	59.31%	G**C	76.47%	59.06%	I**P	71.43%	54.17%
A**C	68.33%	47.69%	D**N	72.46%	51.76%	G**D	70.27%	54.64%	I**Q	76.03%	55.84%
A**D	72.73%	55.67%	D**P	71.43%	53.92%	G**E	69.64%	53.65%	I**R	65.87%	56.74%
A**E	80.61%	56.43%	D**Q	68.33%	53.70%	G**F	77.78%	55.68%	I**S	68.89%	54.83%
A**F	73.19%	57.67%	D**R	78.68%	58.99%	G**G	69.41%	55.85%	I**T	70.14%	54.76%
A**G	75.81%	50.42%	D**S	77.68%	52.54%	G**H	75.00%	56.64%	I**V	63.43%	57.70%
A**H	78.46%	51.66%	D**T	77.04%	53.93%	G**I	65.58%	52.32%	I**W	56.67%	54.87%
A**I	65.71%	59.49%	D**V	77.11%	56.31%	G**K	75.00%	52.17%	I**Y	66.67%	56.83%
A**K	76.86%	57.62%	D**W	90.48%	59.52%	G**L	69.14%	54.43%	K**A	74.94%	58.25%
A**L	75.49%	57.78%	D**Y	71.91%	57.91%	G**M	65.57%	49.52%	K**C	68.75%	48.44%
A**M	68.06%	52.49%	E**A	77.26%	56.28%	G**N	74.29%	54.31%	K**D	73.33%	54.62%
A**N	74.34%	54.26%	E**C	67.35%	48.00%	G**P	60.00%	51.17%	K**E	76.49%	54.13%
A**P	81.48%	50.18%	E**D	77.56%	51.40%	G**Q	82.76%	52.62%	K**F	70.86%	56.97%
A**Q	75.16%	57.04%	E**E	77.45%	52.51%	G**R	69.37%	51.89%	K**G	74.16%	45.10%
A**R	73.74%	54.71%	E**F	79.31%	63.23%	G**S	72.16%	54.78%	K**H	71.74%	54.73%
A**S	70.61%	55.82%	E**G	75.51%	46.44%	G**T	60.67%	54.02%	K**I	70.08%	54.26%
A**T	69.88%	57.56%	E**H	78.43%	51.72%	G**V	67.07%	50.14%	K**K	77.31%	56.62%
A**V	70.87%	57.31%	E**I	67.03%	57.60%	G**W	70.37%	59.68%	K**L	70.81%	55.72%
A**W	76.92%	57.50%	E**K	76.76%	48.46%	G**Y	65.06%	55.13%	K**M	68.18%	54.60%
A**Y	69.70%	61.08%	E**L	72.39%	56.09%	H**A	80.31%	53.70%	K**N	86.67%	50.97%
C**A	72.73%	53.64%	E**M	74.36%	56.41%	H**C	62.50%	54.84%	K**P	69.23%	49.77%
C**C	42.11%	55.70%	E**N	76.63%	47.41%	H**D	77.08%	55.34%	K**Q	70.37%	63.08%
C**D	68.75%	46.00%	E**P	68.42%	46.07%	H**E	87.01%	54.29%	K**R	74.58%	58.60%
C**E	82.76%	52.59%	E**Q	79.41%	52.69%	H**F	60.98%	61.90%	K**S	82.31%	52.43%
C**F	76.47%	58.82%	E**R	75.40%	49.30%	H**G	77.42%	53.43%	K**T	73.85%	57.17%
C**G	66.67%	56.30%	E**S	78.21%	55.04%	H**H	60.87%	59.52%	K**V	72.56%	53.95%
C**H	90.00%	58.49%	E**T	80.35%	54.22%	H**I	74.00%	58.93%	K**W	82.50%	65.22%
C**I	75.00%	58.33%	E**V	73.21%	56.53%	H**K	64.79%	62.02%	K**Y	75.00%	54.47%
C**K	69.23%	51.22%	E**W	62.71%	63.08%	H**L	68.52%	57.42%	L**A	73.97%	51.55%
C**L	66.67%	53.45%	E**Y	79.71%	53.76%	H**M	73.33%	48.68%	L**C	73.13%	51.13%
C**M	84.21%	38.71%	F**A	70.10%	59.57%	H**N	85.71%	49.76%	L**D	67.30%	54.25%
C**N	62.50%	55.56%	F**C	42.11%	68.33%	H**P	75.00%	56.27%	L**E	74.63%	58.60%
C**P	75.00%	51.75%	F**D	71.11%	53.23%	H**Q	60.53%	54.35%	L**F	64.89%	54.26%
C**Q	77.78%	54.35%	F**E	78.11%	57.47%	H**R	76.92%	48.91%	L**G	74.03%	52.97%
C**R	83.33%	44.00%	F**F	61.82%	58.70%	H**S	76.79%	53.14%	L**H	70.00%	51.01%
C**S	82.35%	50.44%	F**G	71.23%	54.39%	H**T	72.41%	59.92%	L**I	67.02%	55.31%
C**T	61.11%	70.21%	F**H	64.44%	57.23%	H**V	74.63%	55.10%	L**K	76.45%	54.00%
C**V	66.67%	63.55%	F**I	64.54%	58.23%	H**W	54.55%	67.80%	L**L	62.05%	55.15%
C**W	33.33%	53.13%	F**K	72.93%	56.63%	H**Y	60.47%	62.34%	L**M	68.57%	58.37%
C**Y	59.09%	46.43%	F**L	67.00%	58.18%	I**A	65.87%	55.57%	L**N	71.71%	51.39%
D**A	77.57%	54.50%	F**M	63.38%	54.92%	I**C	64.44%	52.63%	L**P	90.91%	54.46%
D**C	62.50%	56.99%	F**N	68.97%	56.87%	I**D	71.31%	55.46%	L**Q	71.79%	58.11%
D**D	73.48%	54.71%	F**P	44.44%	52.95%	I**E	73.23%	57.69%	L**R	72.97%	55.57%
D**E	73.96%	53.48%	F**Q	81.00%	56.49%	I**F	63.51%	58.40%	L**S	70.31%	56.57%
D**F	71.88%	55.35%	F**R	73.33%	55.15%	I**G	75.00%	53.90%	L**T	72.94%	59.57%
D**G	73.61%	51.71%	F**S	68.42%	61.04%	I**H	59.46%	52.47%	L**V	68.52%	57.29%
D**H	69.77%	55.23%	F**T	75.27%	54.95%	I**I	66.27%	53.81%	L**W	68.81%	58.96%
D**I	76.61%	55.40%	F**V	67.81%	57.65%	I**K	70.52%	57.17%	L**Y	65.64%	57.60%
D**K	75.08%	49.03%	F**W	54.17%	60.58%	I**L	62.68%	56.53%	M**A	69.27%	56.23%
D**L	68.20%	55.20%	F**Y	55.42%	51.60%	I**M	61.06%	50.00%	M**C	100.00%	61.29%

平均值 Average

全 H 结构时平均值为 71.87%
71.87% in all H structure

非全 H 结构时平均值为 54.96%
54.96% in not all H structure

表 3 第 1 列的数据所显示的是 Motif 的类型; 第 2 列依据扩展后的 Chou-Fasman 法则, 当螺旋结构维持时, α -螺旋第 2 位和第 3 位的氨基酸残基正向影响因子之和大于负向因子之和的概率, 平均概率为 71.87%; 第 3 列是依据扩展后的 Chou-Fasman 法则, 当螺旋结构被破坏时, Motif 中第 2 位和第 3 位的氨基酸残基正向影响因子之和小于负向因子之和的概率, 平均概率为 54.96%, 在统计时不连续的 H 结构被统计为螺旋结构被破坏, 从而造成了统计值偏小。而 C**C, C**W, F**C, F**P, F**W, M**C 这几组统计结果明显带有特殊性是因为其 Motif 统计过程中出现的这种结构次数偏少, 造成统计结果不具有一般性效果。综上可以得出: 当一个序列的第 1 位与第 4 位固定时, 第 2 位与第 3 位共同对螺旋作用。当它们的 Chou-Fasman 的正向因子大于其负向因子时, 有利于螺旋结构的稳定。

3 结论

本文基于 Chou-Fasman 法则并对其进行扩展后, 通过利用 Motif 对 967 条同源性小于 25% 的非冗余蛋白质进行分析, 得出结论: 在四肽 α -螺旋中, 其二、三位的氨基酸残基的二级结构倾向性会对螺旋结构的稳定性产生影响, 当第 2 位和第 3 位的氨基酸残基的 Chou-Fasman 正向影响因子起主导作用时, 即二、三位的 Chou-Fasman 正向因子之和大于负向因子, 有利于螺旋结构的稳定; 第 2 位和第 3 位的氨基酸残基的 Chou-Fasman 负向影响因子起主导作用时, 不利于螺旋结构的稳定。因此我们设想, 在短肽的折叠中, 也会存在类似于 Chou-Fasman 所提出的螺旋核与折叠核这样的一种核心, 凭借核心的倾向性而影响短肽的折叠。

参考文献:

- [1] 阎隆飞, 孙之荣. 蛋白质分子结构[M]. 北京: 清华大学出版社, 1999: 221.
- [2] 来鲁化. 蛋白质的结构预测与分子高计[M]. 北京: 北京大学出版社, 1993: 100-105.
- [3] Chou P Y, Fasman G D. Prediction of protein conformation[J]. Biochemistry, 1974, 13(2): 211-222.
- [4] Chou P Y & Fasman G D. Prediction of the secondary structure of proteins from their amino acid sequence [J]. Adv Enzymol, 1978, 47: 45-148.
- [5] Mayer B, Klein C Th. Influence of solvation on the helix-forming tendency of nonpolar amino acids [J]. Mol Struct: Theochem, 2000, 532, 213-226.
- [6] Mi nor D L, Kim P S. Measurement of the beta-sheet-forming propensities of amino acids [J]. Nature, 1994, 367: 660-663.
- [7] Kim C A, Berg J M. Thermodynamic-sheet propensities measured using a zinc-finger host peptide [J]. Nature 1993, 362: 267-270.
- [8] Tomoko S N, Akio O J. Peptide design of a competitive inhibitor for HMG-CoA reductase based on statin structure [J]. Mol Struct: Theochem, 1997, 419, 155.
- [9] Padmanabham S, York E J, Stewart J M, et al. [α]-Helix formation by peptides of defined sequence [J]. Mol Biol, 1996, 251, 726.
- [10] Smith C K, Withka G M, Regan L A. Thermodynamic scale for the beta-sheet forming tendencies of the amino acids [J]. Biochemistry, 1994, 33: 5510-5517.
- [11] Jiang B, Guo T, Peng LW, et al. Synthesis and folding preferences of γ -amino acid oligopeptides: stereo chemical control in the formation of a reverse turn and a helix[J]. Biopolymer, 1998, 45: 35.
- [12] Padmanabhan R, Marqusee S, Ridgeway T, et al. Relative helix-forming tendencies of nonpolar amino acids [J]. Nature, 1990, 344: 268-270.
- [13] 陈双平, 郑浩然, 刘海燕, 等. 蛋白质中的关联规则发现及应用[J]. 生物物理学报, 2006, 22(3): 171-176.
- [14] Ivankov D N, Finkelstein A V. Prediction of protein folding rates from the amino acid sequence predicted secondary structure [J]. PNAS, 2004, 101(24): 8942-8944.
- [15] Hobohm U, Sander C. Enlarged representative set of protein structures [J]. Protein Sc, 1994, 3(3): 522-524.
- [16] Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features[J]. Biopolymers, 1983, 22(12): 2577-2637.
- [17] Kabsch W. and Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen - bonded and geometrical features[J]. Biopolymers, 1983, 22: 2577-2637.
- [18] Qian N and Sejnowski T J. Predicting the secondary structure of globular proteins using neural network models [J]. J Mol Biol, 1988, 202: 865-884.
- [19] Kinjo A R, Nishikawa K. Predicting secondary structures, contact numbers, and residue-wise contact orders of native protein structures from amino acid sequences using critical random networks [J]. Biophysics, 2005, 1: 67-74.
- [20] Sun Z R, Rao X Q, Peng L W, et al. Prediction of protein super secondary structures based on the artificial neural network method[J]. Protein Engineering, 1997, 10(7): 763-769.