

## 基于混合并行遗传算法的大豆蛋白质二级结构预测

孟翔燕<sup>1</sup>, 孟军<sup>2</sup>, 葛家麒<sup>1</sup>

(<sup>1</sup>东北农业大学理学院, 黑龙江 哈尔滨 150030; <sup>2</sup>国家大豆工程技术中心, 黑龙江 哈尔滨 150030)

**摘要:**大豆蛋白质是人类生活不可或缺的物质,对大豆蛋白质二级结构预测是能够准确预测蛋白质分子三维空间结构功能的关键步骤。将聚类分析、并行处理技术和遗传算法相结合,提出基于混合并行遗传算法(HPGA)的蛋白质二级结构预测方法,充分考虑蛋白质序列两端氨基酸对中间氨基酸结构的影响,蛋白质疏水性对二级结构的影响。在整合、改进前人算法的基础上使得计算复杂度降低1个数量级,使得预测准确率达到74%左右。

**关键词:**大豆蛋白质;并行遗传算法;蛋白质二级结构预测;疏水序列

中图分类号:Q518.1

文献标识码:A

文章编号:1000-9841(2009)02-0200-04

## Soybean Protein Secondary Structure Prediction Based on Hybrid Parallel Genetic Algorithm

MENG Xiang-yan<sup>1</sup>, MENG Jun<sup>2</sup>, GE Jia-qi<sup>1</sup>

(<sup>1</sup>College of Science, Northeast Agricultural University; Harbin 150030, Heilongjiang; <sup>2</sup>National Research Center of Soybean Engineering and Techniques of China; Harbin 150086, Heilongjiang, China)

**Abstract:** Soybean protein is indispensable material in human life. Soybean protein secondary structure prediction is the key step of protein 3D structure and function prediction. This paper combined cluster analysis and parallel technique with genetic algorithm, then proposed protein secondary structure prediction based on hybrid parallel genetic algorithm, in which effects on the behavior of any amino acid in a protein sequence caused by the adjacent amino acid and hydrophobe were considered. On the basis of the former algorithms integrated and improved, the computational complexity is decreased with one magnitude order, and the average prediction accuracy is about 74%.

**Key words:** Soybean protein; Parallel genetic algorithm; Protein secondary structure prediction; Hydrophobic sequence

大豆作为战备粮食与平衡营养的优质蛋白质食品,一直是世界各国不容忽视的作物之一。联合国粮农组织早在20世纪60年代就把大豆作为解决人类营养不足的最优秀作物,号召世界各国扩大种植和利用<sup>[1]</sup>。同时,大豆含有丰富的大豆蛋白质、大豆油、碳水化合物与纤维素等,已广泛应用于营养食品、医疗保健、药物载体和日常用品等诸多领域,因此,大豆作为一种来源充足、营养丰富、用途广泛、产品及多功能多样化的可再生资源,对它进行深入和系统开发,具有重要的现实意义和应用前景<sup>[1]</sup>。

随着大豆生物学的不断发展和信息技术的进步,应用生物信息学的手段来研究大豆的思想孕育而生。因为二级结构是一级结构预测蛋白质三维空间结构功能的关键步骤并且通过分析蛋白质二级

结构,能确认蛋白质功能单位或者结构域,可以为遗传操作提供目标,为设计新的蛋白质或改造已有蛋白质提供可靠的依据,同时为新的药物分子设计提供合理的靶分子结构。所以,大豆蛋白质二级结构的预测不仅有助于了解蛋白质的功能及其作用机制,对于正确预测蛋白质的空间结构更具有非常重要的意义。大豆蛋白质二级结构预测的精度是蛋白质二级结构预测中至关重要的部分,如果精度达不到70%以上,对于经过二级结构预测后实现以上的功能的研究更难。因此,如何提高大豆蛋白质二级结构预测的精度是当务之急。

针对已有研究工作中的不足<sup>[2-4]</sup>,对评测数据和遗传算法进行改进和优化,将聚类分析、并行处理技术和遗传算法相结合,提出基于混合并行遗传算

收稿日期:2008-11-24

基金项目:东北农业大学科技创新资助项目(CXZ010-3)。

作者简介:孟翔燕(1975-),女,硕士,讲师,研究方向为生物信息学。E-mail: mxy20040407@126.com。

通讯作者:孟军,教授,博士生导师。E-mail: merd@mail.neau.edu.cn。

法的大豆蛋白质二级结构预测方法 (Protein Secondary Structure Prediction based on Hybrid Parallel Genetic Algorithm, PSSP-HPGA), 它是由基于 HPGA 的蛋白质二级结构模式提取算法和蛋白质二级结构预测算法共同构成, 模式提取算法主要功能在于利用已知二级结构的蛋白质的训练样本, 通过遗传操作提取高适应度的模式集合; 预测算法主要功能在于利用上述提取的模式集合, 对未知结构蛋白质加以预测。结果表明预测准确率至少提高了 5%。

1 大豆蛋白质评测数据预处理

表 1 训练样本和测试样本

Table 1 Training samples and test samples

训练样本 PDB-ID	1avu	1avw	1avx	1ba7	1bbi	1bte	1bya	1byb	1byc	1byd
Training samples	1f8n	1fgm	1fgo	1fgq	1fgr	1fgt	1fhf	1fs1	1fxz	1g9f
PDB-ID	1hu9	1hyp	1ik3	1ipj	1ipk	1jnk	1sbf	1lnh	1no3	1oaf
	1oag	1q6c	1q6d	1q6e	1q6f	1q6g	1rrh	1rrl	1s6i	1s6j
测试样本 PDB-ID										
Test samples	1sbd 1sbe 1k9b 1uij 1uik									

Sequence and Secondary Structure

1 WDSLPEDELLL GIFSCLCLPE LLKVSGVCKR WYRLASDESL WPSIKLQSSD  
XXCCXHHHHHH HHHCCXHHH HHHHHCCXHH HHHHHHCXHHH XXEEEEEXCC  
51 GEIFEVDVEI AKQSVTIKTM LEDLG  
CXEEEEHHH HHCXHHHHHH HHHCX

图 1 蛋白质 1FS1 的氨基酸及其二级结构数据

Fig. 1 Examples of PDB-DSSP-COUNT data of protein 1FS1

考虑到各个氨基酸位点处二级结构主要是由于序列上相邻的残基的局部作用而产生的, 因此每个氨基酸位点上的二级结构类型应当与它两侧相邻近的氨基酸的类型及性质相关, 如图 2 所示, 中间的氨基酸 L 受到其两侧的氨基酸 A 的影响, 使得氨基酸 L 在二级结构中形成 H 结构, 换句话说, 氨基酸 L 只受到其两侧氨基酸 A 的影响, 与其它 \* 位置上的氨基酸无关, 这就是试验所提取的模式形式<sup>[6]</sup>。对每条蛋白质结构序列采用滑动窗口技术进行切片处理<sup>[7]</sup>, 采用切片长度为 9。在每一蛋白质序列中如果含有未知结构的氨基酸位点, 则在处理时舍去空位记录, 不计入蛋白质结构序列切片数据库中。对蛋白质结构序列 1FS1, 以长度为 9 进行切片处理, 所得结果如表 2 所示。

考虑蛋白质亲疏水性对蛋白质结构的影响, 疏水值序列在一定程度上反映了氨基酸序列所折叠成二级结构所需的作用力规则, 按照如下原则: (i)

评测数据的选择是评价方法准确率的重要环节。所用的评测数据来源于 PDB 数据库, 由于数据量较大, 系统评测时, 选取其中的 40 个序列的 19067 个氨基酸作为训练样本, 选取 5 个序列的 1398 个氨基酸作为测试样本, 如表 1 所示。通过 DSSP 在线程序对其计算, 以提取氨基酸序列及其对应二级结构, 二级结构为 Q8 形式, 将 Q8 简化为 Q3 形式: (i) Helices 包括 G, H, I 三类记为 H; (ii) Sheets 包括 B 和 E 二类记为 E; (iii) Coils 包括 S, T 和 C 三类记为 C<sup>[5]</sup>。以训练样本中大豆蛋白质 1FS1 为例, 仿照上述转化后结果如图 1 所示。

A\*\*L\*\*\*A→H

图 2 影响氨基酸位点二级结构类型图

Fig. 2 Illustration of effect on the secondary structure type of the amino acid locus

将氨基酸 R、N、D、Q、E、H 和 K 归为 0 级; (ii) 将氨基酸 C、I、L、F 和 V 归为 1 级; (iii) 将氨基酸 G、P、S、T、W 和 Y 归为 2 级; (iv) 将氨基酸 A 和 M 归为 3 级<sup>[8]</sup>, 将氨基酸序列转化为疏水值序列, 结果如表 2 所示。通过这种转化以后, 问题已经转变为在疏水值序列中发现某几个位点, 这些位置的氨基酸对于其二级结构的形成起主导作用, 即这些位置的氨基酸在某个化学性质上有一定的规律, 而其他位置的氨基酸对于其二级结构的形式不起作用或者作用较小。将蛋白质氨基酸序列转化为疏水值序列, 不仅可以降低遗传算法搜索空间, 提高遗传算法搜索效率, 更主要的是考虑蛋白质亲疏水性对蛋白质结构的影响, 有效提高蛋白质二级结构预测准确率。

表2 氨基酸序列转化为疏水值序列结果

Table 2 Results of converting amino acid sequence into hydrophobic sequence

氨基酸序列	疏水值序列	二级结构	氨基酸序列	疏水值序列	二级结构
Amino acid sequence	Hydrophobic sequence	Secondary structure	Amino acid sequence	Hydrophobic sequence	Secondary structure
DELLLGIFS	001112112	H	ELLLGIFSC	011121121	H
LLLGIFSCL	111211211	H	LPELLKVSG	120110122	H
PELLKVSGV	201101221	H	IFEVDVEIA	110101013	E
FEVDVEIAK	101010130	H	EVDVEIAKQ	010101300	H
VTIKTMLED	121023100	H	TIKTMLEDL	210231001	H

## 2 基于混合并行遗传算法的蛋白质二级结构模式提取算法

### 2.1 混合并行遗传算法

2.1.1 并行处理技术 并行遗传算法主要考虑按照某种原则将种群分群,由于主要工作是利用每个长度为L的蛋白质结构序列切片,来预测该切片的中心位点的二级结构,因此考虑按照中心位点疏水值将种群分群。如果蛋白质结构序列切片长度为9,那么每个子群搜索空间大小从 $4^9$ 降为 $4^8$ ,进一步降低了搜索空间,提高运算率,达到降低计算复杂度的目的。以蛋白质结构序列1FS1为例按照蛋白质结构序列切片中心位点的疏水值将种群分群,分群结果如表3所示。

表3 疏水值序列数据集分群——分群结果

Table 3 Group-group result of hydrophobic value dataset

疏水级别	二级结构	样本序列
Hydrophobic level	Secondary structure	Samples sequence
0	H	201101221 010101300
	E	110101013
	C	NULL
1	H	001112112 111211211 120110122
	E	101010130
	C	NULL
	NULL	NULL
2	H	011121121 121023100
	E	NULL
	C	NULL
	NULL	NULL
3	H	210231001
	E	NULL
	C	NULL

2.1.2 群体初始化产生 提取模式是基于一定样本数据的提取过程,因此有一定先验知识,群体初始化产生方法主要是一部分从样本数据集中选取一定代表性样本,另一部分随机产生,这样做可以提高算法初始群体的多样性。从样本数据集中选取一定代表性样本作为种群的一部分,主要通过聚类分析 K-

CLARANS方法对每个子群的二级结构所对应的样本集进行分类,从每类中选取一个中心点,作为初始种群。

2.1.3 编码 采用样本序列是蛋白质疏水值序列,因此要设计疏水值序列的编码方式,考虑氨基酸的疏水级别为4级,可以用两位二进制编码表示各个疏水级别,同样二级结构为3模式二级结构,可以用两位二进制编码表示各个二级结构,在这里,由于二进制编码对应四个不同数值,在进行产生初始种群时直接将反映二级结构位置的编码为00的个体删除。蛋白质疏水级别和二级结构具体编码如表4。将蛋白质疏水值序列进行二进制编码,例如“300210230 H”编码为110000100100101101。

表4 蛋白质疏水值和二级结构具体编码表

Table 4 Concrete coding of protein hydrophobic value and its secondary structure

编码方式	蛋白质疏水值	Q3 二级结构
Coding method	Protein hydrophobic value	Q3 secondary structure
00	0	-
01	1	H
10	2	E
11	3	C

2.1.4 适应度函数 利用遗传算法对蛋白质二级结构进行预测,适应度函数的设计主要体现个体代表性强和区分度好等特点,利用已知结构的蛋白质序列的信息,确定蛋白质序列的适应度函数。为了能够使得个体代表性强,区分度好,设计适应度函数时考虑两个函数自信度函数:

$$confidence = \frac{n_{\max}}{n_H + n_E + n_C}, n_{\max} = \max(n_H, n_E, n_C)$$

其中 $n_H$ 、 $n_E$ 和 $n_C$ 表示该染色体的邻近样本在其对应的H、E和C组中相似性较高的数量。

$$鉴别率函数: DR = \frac{n_{Highest} - n_{Second}}{n_H + n_E + n_C}$$

其中 $n_{Highest}$ 的定义和上式的 $n_{\max}$ 相同,而 $n_{Second}$ 则表示是 $n_H$ 、 $n_E$ 和 $n_C$ 之中次高的数量。

综合考虑自信度和鉴别率两个函数,确定混合并行遗传算法的适应度函数:

$$fitness = confidence * DR$$

2.1.5 其他遗传算子的设计 选择算子选取基于精英保留策略的轮盘赌选择;交叉算子选取多点交叉;变异算子选取分层次变异;终止条件选取最大迭代代数。

2.2 模式提取算法

模式提取算法步骤如下:

Step1:将已转换为蛋白质疏水值序列的样本数据集,按照中间位置疏水级别分为 4 个子群,对每个子群进行独立遗传操作提取模式规则;

Step2:根据分群后样本数据集,产生初始群体;

Step3:对种群轮盘赌选择,产生新种群;

Step4:进行多点交叉,分层次变异,产生新种群;

Step5:从新种群选取适应度最高的个体,存储在模式规则集合中;

Step6:随机产生一个个体共同构成子代(新种群);

Step7:若不满足终止条件,转到 Step3,若满足终止条件则输出模式规则集合。

3 蛋白质二级结构预测算法

利用模式提取算法提取模式后建立模式集,以此为参照对未知结构蛋白质序列进行二级结构预测,具体算法步骤如下:

Step1:将未知二级结构氨基酸序列转换为疏水值序列,对其进行蛋白质序列切片处理形成长度为 L 的蛋白质疏水值序列切片;

Step2:根据每个切片序列中间疏水值,去寻找对应的模式规则集合;

Step3:计算未知结构序列切片与对应的模式规则集合的各个模式规则的相似性,取其最大值,若该最大值大于所给定阈值,则认为此未知结构的序列切片中间氨基酸为其所比对模式的结构,否则认为此氨基酸结构未知。

4 测试与结果

以测试样本 1K9B 为例,采用 PSSP-HPGA 对其预测,预测结果如图 3 所示(下划线位点错误),对 1K9B 预测准确率达 81.03%,已所提供的测试样本,采用 PSSP-HPGA 方法平均准确率在 74.23%左右,如表 5 所示,验证了 PSSP-HPGA 方法的有效性。

Sequence and Secondary Structure  
1 KPABDQCADT KSNPPQDRBS DMRLNSEHSA FKSEIGALSY PAQGFFVDIT  
XXXXXXCCEEE XXXCCCEEEE EEEEECCCC CCCEEECCCC CCEEEEXXX  
51 DFCYEPAK  
XXXXXXXX

图 3 蛋白质 1K9B 的预测结果  
Fig.3 Prediction result of protein 1K9B

表 5 测试样本预测准确率  
Table 5 Prediction accuracy of test samples/%

PDB- ID	1sbd	1sbe	1k9b	1uij	1uik	AVG
准确率 Prediction accuracy	71.2	72.42	81.03	75.2	71.3	74.23

5 结论

针对大豆蛋白质同源性特征,提出基于混合并行遗传算法的蛋白质二级结构预测方法,该方法应用混合并行遗传算法提取有用模式,能够利用模式集合来预测蛋白质二级结构,充分考虑蛋白质序列两端氨基酸对中间氨基酸结构的影响;蛋白质疏水性对二级结构的影响,在整合、改进前人算法的基础上使得计算复杂度降低 1 个数量级,使得预测准确

率达到 74% 左右。可见,该方法是目前蛋白质二级结构预测方法的改善与补充。

参考文献

[1] 杨文钰,雍太文,任万军,等. 发展套作大豆,振兴大豆产业[J]. 大豆科学,2008,27(1):1-7. ( Yang W Y,Yong T W,Ren W J,et al. Develop relay- planting soybean revitalize soybean industry[J]. Soybean Science,2008,27(1):1-7. )  
[2] Huang H C. An evolutionary approach to finding schemas for 3- class protein secondary structure prediction[J]. Proceedings of the Computational Systems Bioinformatics,2003,488-491. ( 下转第 209 页 )

- [14] Guo W, Cai C, Wang, C, et al. A microsatellite based gene – rich linkage map reveals genome structure, function and evolution in *Gossypium* [J]. *Genetics*, 2007, 176: 527 – 541.
- [15] 李永强, 李宏伟, 高丽锋, 等. 基于表达序列标签的微卫星标记 (EST – SSRs) 研究进展 [J]. *植物遗传资源学报*, 2004, 5 (1): 91 – 95. ( Li Y Q, Li H W, Gao L F, et al. Progress of simple sequence repeats derived from expressed sequence tags [J]. *Journal of Plant Genetic Resources*, 2004, 5 (1): 91 – 95. )
- [16] Varshney R K, Graner A, Sorrells M E. Genic microsatellite markers in plants: features and applications [J]. *Trends in Biotechnology*, 2005, 23 (1): 48 – 55.
- [17] Rota L R, Kantety R V, Yu J K. Nonrandom distribution and frequencies of genomic and EST – derived microsatellite markers in rice, wheat and barley [J]. *BMC Genomics*, 2005, 6: 23.
- [18] Song Q J, Marek L F, Shoemaker R C, et al. A new integrated genetic linkage map of the soybean [J]. *Theoretical and Applied Genetics*, 2004, 109: 122 – 128.
- [19] Temnykh S, De Clerck G, Lukashova A, et al. Computational and experimental analysis of microsatellites in rice (*Oryza sativa* L.) [J]. *Genome Research*, 2001, 11: 1441 – 1452.
- [20] Dreisigacker S, Zhang P, Warburton M L, et al. SSR and pedigree analyses of genetic diversity among CIM – MYT wheat lines targeted to different megaenvironments [J]. *Crop Science*, 2004, 44: 381 – 388.
- 
- (上接第 203 页)
- [3] Chu Y W, Sun C T. Regularity of secondary protein structures: A genetic algorithm approach [C]. *Proceedings of the 5<sup>th</sup> World Congress on Intelligent Control and Automation*, 2004: 2104–2108.
- [4] 邹先霞, 陈孝卫, 许龙飞. 基于关联规则与遗传算法的蛋白质二级结构预测 [J]. *计算机工程与应用*, 2006, 42 (15): 152–156. (Zou X X, Chen X W, Xu L F. Protein secondary structure prediction based on association rules and genetic algorithm [J]. *Computer Engineering and Applications*, 2006, 42 (15): 152–156. )
- [5] 张宁, 张涛. 蛋白质二级结构预测样本集数据库的设计与实现 [J]. *生物信息学*, 2006, 4 (4): 163–166. (Zhang N, Zhang T. Design and implementation of the database of data sets for prediction of protein secondary structures [J]. *China Journal of Bioinformatics*, 2006, 4 (4): 163–166. )
- [6] Chu Y W, Sun C T. A hybrid genetic algorithm approach for protein secondary structure [C]. *Proceedings of the 6<sup>th</sup> World Congress on Intelligent Control and Automation*, 2006: 3320–3324.
- [7] 杜耀华, 王正志, 倪青山. 基于滑动窗口的原核转录起始点计算定位方法 [J]. *生物物理学报*, 2006, 22 (5): 360–366. (Du Y H, Wang Z Z, Ni Q S. Computational location of transcription start sites in prokaryotic genome based on sliding window [J]. *Acta Biophysica Sinica*, 2006, 22 (5): 360–366. )
- [8] 黄敏, 沈辉, 肖奕. 不同类蛋白质氨基酸疏水序列周期性 [J]. *生物物理学报*, 2000, 16 (4): 755–760. (Huang M, Shen H, Xiao Y. Correlation properties of protein sequences of different types [J]. *Acta Biophysica Sinica*, 2000, 16 (4): 755–760. )
- 

## 立足黑龙江 辐射全中国 聚焦大农业 促进快发展

### 欢迎订阅 2009 年《黑龙江农业科学》

《黑龙江农业科学》是黑龙江省农业科学院主办的综合性科技期刊。内容丰富、栏目新颖、信息量大、可读性强,读者群大、发行面广,是全国优秀期刊、黑龙江省优秀期刊。现已被《中国科学引文数据库》《中国核心期刊(遴选)数据库》、CNKI 系列数据库、万方数据库、重庆维普中文科技期刊数据库和华艺电子出版事业群等多家权威数据库收录。

《黑龙江农业科学》为双月刊,单月 10 日出版,国内外公开发行。国内邮发代号 14 – 61,每期定价 8.00 元,全年 48.00 元;国外由中国国际图书贸易总公司发行,发行代号 BM8321,每期定价 8.00 美元,全年 48.00 美元。

热忱欢迎广大农业科研工作者、农业院校师生、国营农场及农业技术推广人员、管理干部和广大农民群众踊跃订阅。全国各地邮局均可订阅。漏订者可汇款至本刊编辑部补订。汇款写明订购份数,收件人姓名、详细邮寄地址及邮编。

另外,编辑部现有少量 2007 年合订本珍藏版。每册 80.00 元,邮费 10.00 元,共计 90.00 元,售完为止。

地址:哈尔滨市南岗区学府路 368 号《黑龙江农业科学》编辑部

邮编:150086 电话:0451 – 86668373 电子函件:nykx13579@sina.com