

文章编号 :1009-038X(2001)04-0340-04

EcoPDB :高精度大肠杆菌蛋白质结构 与对应基因序列数据集

李炜疆, 宋江宁

(无锡轻工大学 工业生物技术教育部重点实验室, 江苏 无锡 214036)

摘 要 :高质量蛋白质结构及其对应基因序列数据是研究蛋白质折叠与蛋白质编码序列关系问题的基础.通过查询 SWISS-PROT 数据库中 *E. coli* 的蛋白质,得到不同数据库中的蛋白质结构与基因序列的交叉索引表,在此基础上,删除大量冗余及不可靠数据,最后得到一个高精度数据集 EcoPDB.该数据集共有 191 个 *E. coli* 基因及其相应的精度好于 2.5 Å 的 X 射线衍射测定的 PDB 蛋白质结构数据,总残基数约 5.5 万.

关键词 :基因序列;蛋白质结构;数据集;大肠杆菌

中图分类号:Q61

文献标识码:A

EcoPDB : A High Quality Dataset of *Escherichia coli* Gene Sequences and Corresponding Protein Structures

LI Wei-jiang, SONG Jiang-ning

(Key Laboratory of Industrial Biotechnology of Ministry of Education, Wuxi University of Light Industry, Wuxi 214036, China)

Abstract :High quality data about protein structures and their gene sequences are essential to the understanding the relation between protein folding and protein coding sequences. By query about *Escherichia coli* proteins in SWISS-PROT, a cross-reference table of the protein structures and their corresponding gene sequences in different databases was obtained. Removing a large amount of redundant and uncertain data, a high quality dataset, EcoPDB, was finally constructed. This dataset includes 191 *E. coli* genes and corresponding Protein Data Bank structures determined by X-ray diffraction method with resolutions better than the 2.5 Å. The total residue number amounts to about 55 000.

Key words :coding sequence;protein structure;dataset;*E. coli*

一般认为,蛋白质高级结构是由其一级结构即氨基酸序列决定的.也就是说,蛋白质空间折叠的信息完全蕴藏在氨基酸序列中.氨基酸序列由 mRNA 序列编码,但由于遗传密码的简并性, mRNA 编

码序列与相应的氨基酸序列并非一一对应,而是多对一的关系.编码同一蛋白质的不同 mRNA 序列都使其翻译产物——氨基酸序列折叠到同样的空间结构吗?同义密码子用语是不是蛋白质折叠的“隐

收稿日期:2000-12-31; 修订日期:2001-04-12.

基金项目:国家自然科学基金项目资助课题(项目号 39760027).

作者简介:李炜疆(1964-),男,陕西榆林人,理学博士,教授,博士生导师.

变量”?

在生物体内,同义密码子翻译为氨基酸的过程差异显著^[1,2]。也就是说,编码同样氨基酸序列的 mRNA 翻译表达为新生肽链的过程是很不相同的。如果蛋白质的翻译与折叠同时进行,核酸序列的差异就会影响到蛋白质折叠的动力学过程,自然也就有可能对蛋白质空间结构产生影响。

由于蛋白质折叠的复杂性,从动力学过程进行分析,目前还不能得到确切答案。能不能越过动态过程而直接分析蛋白质结构与核酸编码序列之间的关系?随着蛋白质空间结构数据的大量积累,对这个问题进行统计研究已经成为可能。

为研究这个问题,首先需要找到蛋白质空间结构数据与核酸序列数据之间的对应关系。从现有的蛋白质结构及核酸序列基础数据库难以直接自动地找出这两者之间的联系,因为通常蛋白质的 PDB 结构数据文件并不给出相应的核酸编码序列,而准确可靠的数据是研究工作的基础。由于数据来源不同,其中包含很多不确定因素甚至是错误,具体表现形式繁杂不一,所以建立这样的数据库很难完全依赖计算机程序设计来自动完成,需要大量的人工工作。这样的基础工作对相关的研究是至关重要的,此基础数据集不仅可用于研究蛋白质折叠规律,而且可以提供其它分子生物学研究课题的有关信息。基因序列及其相应蛋白质结构数据库还可用于优化基因表达、蛋白质结构预测、分子进化等方面的研究。

1 材料与方法

1.1 材料

基本的核酸序列数据来源是 GenBank^[3~5],蛋白质空间结构数据取自 Protein Data Bank (PDB)^[6,7]。为简化数据库格式转换,基因序列取自从 GenBank 导出的大肠杆菌基因二级数据库 EcoGene^[8,9]。而核酸序列数据与蛋白质结构数据之间

的交叉索引则取自蛋白质数据库 SWISS-PROT^[10,11]。

1.2 方法

首先查询 SWISS-PROT 中有 PDB 结构数据的大肠杆菌蛋白质,得到原始数据集,共得到 307 条记录,包含蛋白质序列、基因名、以及可能相关的蛋白质结构数据文件的 PDB 码(共有 1129 个)。通过核对具体数据,剔除其中大量的冗余和不精确数据,就得到一个高精度大肠杆菌蛋白质结构与对应基因序列数据集,称为 EcoPDB。基本过程如下:

- 1) 从原始数据集中剔除所有非 X 射线衍射方法(核磁共振、理论模型等)以及结构数据分辨率大于 2.5 Å 的结构数据;
- 2) 删除膜蛋白基因及相应的 PDB 数据;
- 3) 删除含有大量不确定氨基酸残基或者缺少大量原子坐标的 PDB 数据;
- 4) 根据分辨率和 R 因子,删除对应同一基因的多余 PDB 数据;
- 5) 删除冗余蛋白质多肽链;
- 6) 进一步序列比对,删除不可靠数据。

2 结果与讨论

原始数据集经过上述筛选过程,最后保留下来 191 个基因(约占大肠杆菌基因组总基因数的 4.5%)和一一对应的 191 个蛋白质链的 PDB 结构,详细基因名及对应空间结构数据 PDB 码见表 1。表中的 PDB ID 的前四位表示蛋白质结构的 PDB 码,如果有多条蛋白质链,则用第五位字母表示链标识符。在最终的 EcoPDB 数据集中,属于 α 型蛋白质结构的 PDB 占总 PDB 数的 18.3%,属于 β 型蛋白质结构的 PDB 占总 PDB 数的 12.0% ($\alpha + \beta$)型蛋白质结构的 PDB 占总 PDB 数的 69.1%,属于膜和细胞表面蛋白质结构型的 PDB 占总 PDB 数的 0.5%,具有完整主链原子坐标并且与基因序列密码子相吻合氨基酸残基(有效残基)数约为 5.5 万。

表 1 EcoPDB 中基因及其相应蛋白质结构 PDB 码

Tab.1 The genes and the PDB codes of their corresponding protein structures in EcoPDB							
基因	PDB 码	基因	PDB 码	基因	PDB 码	基因	PDB 码
accB	1BDO	dppA	1DPE	gshB	1GSA	nagB	1DEAA
accC	1BNCA	dps	1DPSA	gst	1A0FA	nanA	1NAL1
ada	1SFE	dsbA	1FVKA	guaA	1GPMA	narL	1A04A
adk	1AKEA	dut	1EUWA	gyrB	1AJ6	nrdA	1RLR
alkA	1WJGA	era	1EGAB	hdeA	1BG8A	nrdB	1XIKA
						recA	2REB
						rffG	1BXKB
						rho	1A62
						rnhA	2RN2
						rpIL	1CTF

续表 1

基因	PDB 码	基因	PDB 码	基因	PDB 码	基因	PDB 码	基因	PDB 码
alsB	1RPJA	fabA	1MKAA	hdhA	1FMCA	nth	2ABK	rplY	1DFUP
ampC	2BLSA	fabB	1DD8A	hemB	1B4EA	ompF	2OMF	rpoA	1BDFB
ansB	4ECAA	fabD	1MLA	hemC	1PDA	ompR	1OPC	rtcA	1QMHA
araC	2ARCB	fabF	1KAS	hisJ	1HSLA	pck	1AYL	ruvA	1CUK
araF	8ABP	fabH	1EBLA	holB	1A5T	pepP	1AZ9	ruvC	1HJRA
arcB	2A0B	fabI	1QSGA	hybD	1CFZA	pfkA	1PFKA	serC	1BJNA
asd	1BRMA	fbaA	1DOSA	ilvE	1A3GA	pflB	1CM5A	sodA	1VEWA
asnA	12ASA	fcl	1BSVA	katE	1CF9A	pheA	1ECMB	sodB	1ISAA
asnB	1CT9A	fdhF	1AA6	kdtB	1B6TA	phoA	1B8JA	sodC	1ESO
aspC	1ARS	fhuD	1EFDN	leuB	1CM7A	phoB	1B00A	sucC	2SCUB
aspS	1C0AA	fis	1FIAA	livJ	2LIV	php	1BF6A	sucD	2SCUA
atpC	1AQT	fldA	1AG9A	livK	2LBP	phrB	1DNPA	talB	1ONRA
bioD	1BYI	fmt	1FMTA	malE	1ANF	potD	1POT	tar	2ASR
bioF	1BS0A	folA	1RA9	manX	1PDO	potF	1A99A	thyA	1QQQA
birA	1BIA	folE	1A8RA	map	1C24A	ppa	1OBWA	topA	1ECL
carA	1A9XB	folK	1HKA	marA	1BL0A	pstS	1IXH	treR	1BYKA
carB	1BXRA	folP	1AJ2	mdh	2CMD	pth	2PTH	trpC	1PII
cdd	1CTT	fpr	1FDR	metC	1CL1A	ptsH	1OPD	trxA	2TRXA
chbB	1IIBA	ftsY	1FTS	metF	1B5TA	ptsI	1ZYMA	trxB	1TRB
cheY	3CHY	fucA	2FUA	metG	1QQT A	ptsN	1A6JB	tufB	1EFCA
clpP	1TYFA	fucI	1FUIA	metH	1MSK	purA	1QF5A	tynA	1OACB
crp	3GAPA	fumC	1FURA	metJ	1CMCA	purD	1GSOA	ubiB	1QFJA
ctr	1F3Z	galE	1UDC	mglB	2GBP	purE	1QCZA	umuD	1UMUA
cspA	1MJC	galT	1GUQA	mgsA	1B93B	purF	1ECFB	ung	4EUGA
cybC	256BA	gapA	1GADO	mltB	1LTM	purK	1B6RA	ushA	1USH
cyoA	1CYX	glcB	1D8CA	modA	1WOD	purN	3GAR	vsr	1CW0A
cysH	1SUR	glmS	1MOQ	mtlA	1A3AA	purR	1DBQA	xerD	1A0P
cysI	2GEP	glnB	2PII	mug	1MUGA	pykF	1PKYA	xthA	1AKO
dapA	1DHPA	glpK	1BU6Y	mukB	1QH LA	pyrB	8ATCA	ycaC	1YACA
dapB	1DRW	gmd	1DB3A	murA	1UAE	pyrE	1OROA	yjgF	1QU9A
ddlB	1IOW	gor	1GESB	murB	1UXY	pyrI	8ATCB		
def	1BS4A	gpt	1A96B	murD	1UAG	qor	1QORA		
deoD	1ECPA	greA	1GRJ	mutH	1AZO	rbsB	2DRI		
dnaN	2POLA	groL	1DERA	mutL	1B63A	rbsK	1RKD		

在建库过程中具体的操作方法和一些特殊情况说明如下：

1) 一般不保留处于复合物状态的蛋白质结构，

特别是 DNA 结合状态,因为此时的蛋白质结构会产生诱导变形,可能与自然非结合状态下的结构有较大出入；

2) PDB 数据文件 1AW8 (由基因 *panD* 编码)有 A、B、D (=A)、E (=B) 四条链,如果将 B 链与 A 链连接起来,则所得序列与基因翻译序列相符,为什么一个基因编码两条链呢?进一步查询资料发现,此蛋白质自发水解成两种基团,鉴于这一特殊情况,删去该结构数据;

3) 结构数据 1TOL (对应基因 *tolA*) 的序列与基因翻译序列相差悬殊,比对发现,1TOL 序列仅 C 端 93 个氨基酸与基因翻译序列 C 端相符,而其余

部分则与翻译序列完全不符,故删去该结构数据;与此类似的还有 1ISQ (对应基因 *icd*),其基因翻译序列仅是 PDB 序列 C 端的一小部分,原因不明,故该结构数据也不保留;

4) 基因翻译序列与 PDB 序列完全相同的情况极少,仅占 EcoPDB 的 191 个基因中的 8 个.大多数序列差异主要表现为氨基酸序列两末端的微小差别,较大的差异则是由于 PDB 序列去除了信号肽.

参考文献：

[1] ORESIC M, SHALLOWAY D. Specific correlations between relative synonymous codon usage and protein secondary structure [J]. **J Mol Biol**, 1998, 281 : 31~48.

[2] 解涛,丁达夫. 同义密码子用语与蛋白质结构的关系[J]. 生物物理学报, 1998, 14 : 296~302.

[3] BENSON D A, KARSCH-MIZRACHI I, LIPMAN D J, *et al.* GenBank[J]. **Nucleic Acids Res**, 2000, 28 : 15~18.

[4] WHEELER D L, CHAPPEY C, LASH A E, *et al.* Database resources of national center for biotechnology information[J]. **Nucleic Acids Res**, 2000, 28 : 10~14.

[5] GENBANK[DB]. <http://www.ncbi.nlm.nih.gov/Web/GenBank/> 2001-04-12.

[6] BERMAN H M, WESTBROOK J, FENG Z, *et al.* The Protein Data Bank[J]. **Nucleic Acids Res**, 2000, 28 : 235~242.

[7] PDB[DB]. <http://www.rcsb.org/pdb/> 2001-04-12.

[8] ECOWEB[DB]. <http://bmb.med.miami.edu/EcoGene/EcoWeb/> 2001-04-12.

[9] RUDD K E. EcoGene : a genome sequence database for *Escherichia coli* K12[J]. **Nucleic Acids Res**, 2000, 28 : 60~64.

[10] BAIROCH A, APWEILER R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000[J]. **Nucleic Acids Res**, 2000, 28 : 45~48.

[11] SWISS-PROT[DB]. <http://www.expasy.ch/sprot/> 2001-04-12.

(责任编辑 李春丽)