计算分子生物学(网络版), 2014 年, 第 3 卷, No.10, 1-9
Jisuan Fenzi Shengwuxue (Online), 2014, Vol.3, No.10, 1-9
http://cmb.biopublisher.cn

研究论文

Research Article

利用动态生物数据融合策略对 *E.coli* K-12 进行 In Silico 蛋白组学功能重新注释

Gopal Ramesh Kumar ✉ Thankaswamy Kosalai Subazini Chinnasamy Perumal Rajadurai Kandavel Palani Kannan

生物信息学实验室, AU-KBC 研究中心, 安娜大学, 金奈, 600044, 印度

✉ 通讯作者: gramesh@au-kbc.org

计算分子生物学, 2014 年, 第 3 卷, 第 10 篇 doi: 10.5376/cmb.cn.2014.03.0010

建议最佳引用格式:

Kumar et al., 2014, In silico Proteomic Functional Re-annotation of Escherichia coli K-12 Using Dynamic Biological Data Fusion Strategy, Computational Molecular Biology, Vol.4, No.4 34-43 (doi: 10.5376/cmb.2014.04.0004)

摘 要 大肠杆菌，是广大生物学研究着最喜欢的模型生物之一，最初是在 1997 年注释，并在 2007 年完成重新注释。虽然在大肠杆菌基因组上，已经进行了多年的深入研究，但在完整和准确的生物功能上的研究信息并不可用。在大肠杆菌中，因为缺乏功能信息，约有 40%的蛋白质序列被注释为假定蛋白。因此，这些蛋白序列需要利用更先进的计算方法去获取它的生物学功能。在这里，我们采用“动态生物数据融合策略”，对大肠杆菌 K-12 完整蛋白质组进行了重新注释。它是一种计算策略，我们通常应用于与异构生物数据源相结合，最大限度地提高知识共享和生成数据集的交集。本研究对大肠杆菌 K-12 的功能重新注释结果有助于我们获取高质量、完整的蛋白质组数据。我们已经更新了以前注释的所有的蛋白质编码基因，并试图在可能的情况下分析新的或更精确的蛋白功能。约 29%的大肠杆菌的蛋白质序列，先前被注释为你不清楚或未知功能(即无功能)，现在已被注释为清楚或已知的功能。此外，重新分析也导致了对已发现是假阳性或错误注释的蛋白序列的修订。这个研究的注释结果信息可以作为数据库，“REC-DB”，这仍然是一个有用的、数据得到更新、信息更准确的数据库。REC-DB 是公开在 http://recdb.bioinfo.au-kbc.org.in/recdb/。

关键词 大肠杆菌; 重新注释; 假设蛋白; 置信水平; 系统发生; 基序

In silico Proteomic Functional Re-annotation of Escherichia coli K-12 Using Dynamic Biological Data Fusion Strategy

Gopal Ramesh Kumar ✉ Thankaswamy Kosalai Subazini Chinnasamy Perumal Rajadurai Kandavel Palani Kannan

Bioinformatics Lab, AU-KBC Research Centre, M.I.T Campus of Anna University, Chromepet, Chennai 600044, India

✉ Corresponding author, gramesh@au-kbc.org

Abstract Escherichia coli, one of the favorite model organisms, was initially annotated in 1997 and re-annotated in 2007. Although years of intensive research is being carried out on E. coli genome, still complete and accurate functional information of this organism is not available. In E. coli, about 40% of the protein sequences have been annotated as hypothetical proteins, because of lack of information. Hence, such sequences require advanced computational strategies and derive clues on their biological role. Herein, we have carried out re-annotation of the complete proteome of E. coli K-12 using “Dynamic biological data fusion method”. It is a computational strategy we typically applied for combining the heterogeneous biological data sources to maximize knowledge sharing and generating the intersection of data sets. Functional re-annotation results reported in this paper help us to present high quality data on complete proteome of E. coli K-12. We have updated all the protein coding genes from previous annotation work and tried to assign new or more precise functions, wherever possible. About 29% of the protein sequences of E. coli which have been previously annotated

收稿日期：2014 年 8 月 7 日
接受日期：2014 年 9 月 26 日
发表日期：2014 年 10 月 8 日

as unclear/unknown (hypothetical; without functions) have now been assigned with clear/known functions. Further, the analysis also resulted in the revision of the protein sequences that have been found to be false positive or poorly annotated. Information from this work is made available as a database, “REC-DB", which will remain a useful repository with accurate and updated functional information. Availability: REC-DB is publicly available at http://recdb.bioinfo.au-kbc.org.in/recdb/.

Keywords *E.coli*; Re-annotation; Hypothetical proteins; Confidence level; Phylogenetics; Motif

自从注释的大肠杆菌 K−12 基因组发表在 1997 发表后，在基因组学领域的研究以快速的步伐在不断扩大(Blattner et al., 1997)，这导致了序列信息和相关生物学数据库呈指数增长(Serres et al, 2001)。尽管通过生化实验对大肠杆菌基因组的属性已经研究几十年了，但对于该模式生物的完整和准确的功能信息仍然是不可用的。全球的几个基因组计划已经完成，其中许多研究都是持续了很久。但是由于其分析是基于过时的数据或不适当的序列模型，因此有些功能注释结果并不是完整的。此外，这些信息多年没有更新，这样一个质量较差的注释结果将导致在我们的基因组的认知上存在偏差(Salzberg., 2007)。这种不完整的注释，导致在它们的基因组中广泛存在未知功能的蛋白。传统的功能富集是通过对整个基因组进行 BLAST 分析，但这种方式得出的结果在大多数情况下并不是完整的，因为数据库的功能蛋白并非时常更新。因此，有必要通过重新注释，经常对基因组功能进行更新，否则数据库中提供的信息将是过时的。在基因组中，例如假定蛋白或者保守的假定蛋白，以及功能不确定的蛋白都会导致分析结果的不确定性，在一般的微生物基因组，这种不确定性约占 30~40%。假定蛋白和保守的假定蛋白质的功能必须要得到预测，因为它们很可能在微生物的细胞生理学中发挥了至关重要的作用。假定蛋白是未知功能的，也没有同源性和实验验证功能的蛋白，保守假定蛋白是未知功能，但是具有同源性的蛋白(Wood et al., 2001; Riley et al., 2006)。这些未知蛋白可能参与基因的表达调控、细胞信号转导、宿主与寄生虫相互作用和复杂的次生代谢产物(包括抗生素和生物活性化合物的合成)，因此保守假设蛋白的生化研究可以发现非常有意义的新的药物生物分子(Galperin and Koonin, 2010; Roberts et al., 2001)。在预测蛋白质的功能的过程中，使用不同的生物信息学工具有助于更简单和有效的对蛋白进行注释。

目前，对于大肠杆菌的生物信息学研究已经在快速的增加(Serres et al., 2001)，这也有助于更好地理解这种微生物以及其他生物编码的功能(Karp et al., 2007)。因此，目前最重要的是需要提供更新和更准确的大肠杆菌的功能注释信息，以保证能够提供给科学界使用。功能的重新注释，是一种对之前已注释的基因组信息进行重新注释的过程，其能够为基因组的研究提供更进一步的支持(Rajadurai et al., 2011)。这个过程通常涉及到各种功能预测的计算技术，这样的功能预测也可以实现使用更先进的高通量技术，然而，这样的技术是非常费力和昂贵的(Valencia., 2005)。因此，In Silico 功能再分析将有助于实现快速和可靠的功能预测，功能重新注释可以潜在地提供更高层次的细胞过程，如代谢、运输、致病性和调节的过程，从而促进个别蛋白质在蛋白质组水平上的功能说明(Zheng et al., 2002)。此外，重新注释的结果也将有助于了解蛋白质的动态相互作用和代谢过程的基本机制，所有的生物过程都是通过大的有序配合物或级联蛋白来实现的。它也有助于确定新的蛋白质的功能，为传染病和遗传疾病的治疗提供真正有效的新疗法。此前，包括 *Mycoplasma pneumoniae* (Dandekar et al. 2000), *Mycobacterium tuberculosis H37Rv* (Camus et al., 2002), *Campylobacter jejuni* (Gundogdu et al., 2007), *Geobacter sulfurreducens* (Ashok et al., 2014) and *Saccharomyces cerevisiae* (Wood et al., 2001)等在内的几种生物的基因组，使用不同的计算策略被成功地重新注释，现在它们已经作为生物研究中有用的注释信息。

同样，基因组分析工作已经成功地分析了大肠杆菌所有可用的注释，并对这些功能信息做了简要的介绍。然而，在他们的分析结束时，他们报告中依然存在约 14%的未知序列，他们报告了他们的分析结果，如文本文件和表格。此外，这些分析工作是在 1997 年和 2005 年进行的，这时候，一种新的注释方法出现了(Blattner et al. 1997; Riley et al., 2006)。因此，这项研究工作旨在于采用多元、动态的生物数据融合策略对大肠杆菌 K-12 的全蛋白组序列进行重新注释。在对基因组进行重新注释之前，

对现有的几个蛋白数据库进行仔细的比对和分析以确定可用的公共数据，进而获得对大肠杆菌研究有用的数据信息。在这个重新注释的工作中，一个动态的生物数据融合策略已被实施，以用来执行对基因组序列的功能预测。这个数据处理策略主要是通过从数据库中结合异构数据形成动态数据集，从而在最大限度的提高信息的共享(Elmore et al., 2003)。

近年来，出现了很多全基因组序列和相关的蛋白质数据库(Pearson and Lipman, 1988)，能够提供有用的与近亲属相关的注释信息。In Silico 重新注释允许进行的质量控制、系统数据的更新和简单的数据分析以及更全面的比较分析，为整个研究界提供了一个宝贵的资源(Rajadurai et al., 2011)。虽然，为了促进生命科学的研究，一些基因组和蛋白质组数据库是公开的，但是还是有一部分公开数据是有缺陷的。在这些生物序列数据库中最常见的问题是注释信息缺乏可靠性、序列冗余和错误的标注等。最近，在四种公共蛋白质序列数据库的微量注释水平，UniProtKB/Swiss-Prot 、GenBank NR 、UniProtKB/TrEMBL 和 KEGG 已被调查和鉴定出存在错误注释的酶家族，而一个更大的问题是在注释过程(Schnoes et al., 2009)。

为了缓解错误的注释和冗余的问题，使用一组共同的、可控制的数据环境来重新注释基因和蛋白质，进而描述一个基因或蛋白质的功能是必要的。因此，破译基因组中所有编码产物的确切功能是在基因组时代的巨大挑战(Bock and Gough, 2004; Altschul et al., 1990)。因此，为了克服这些挑战，使用“动态的生物数据融合”策略对大肠杆菌 K-12 的基因组重新注释，从各种可用的数据库的生物数据集成为一个独特的信息源。此外，在置信区间的水平上对未知蛋白进行功能注释，从而促进了在研究上获得更准确的信息。

1 结果

从 EcoCyc 数据库下载的大肠杆菌 K-12 菌株的原始序列的注释信息，经鉴定有 4290 的蛋白质序列。在这些序列中，2560 个序列有明确的注释信息，剩余的 1730 个序列是未知的、未知的假设、预测、保守和假定的功能(表 1, 图 1A)。随着 In Silico 的重新注释，先前的几类注释信息发生了改变，一是对未知蛋白赋予了功能，而是对已经注释的蛋白进行功能注释的修改，三是更新蛋白功能信息。

表 1 *E. coli* K-12 的基因组信息

Table 1 Genome features of *E. coli* K-12

Sequence Category	Number of Sequences	Percentage (%)
Total No. of protein Sequences	4290	100
i. Sequences with unknown functions:		
Predicted	1033	24.08
Conserved	401	9.35
Putative	212	4.94
Conserved + Hypothetical	18	0.42
Hypothetical	59	1.38
Conserved + Predicted	2	0.05
Conserved + Putative	1	0.02
Hypothetical + Predicted	2	0.05
Putative + Predicted	2	0.05
Total Sequences with unknown functions	1730	40
ii. Sequences with clear functions	2560	60

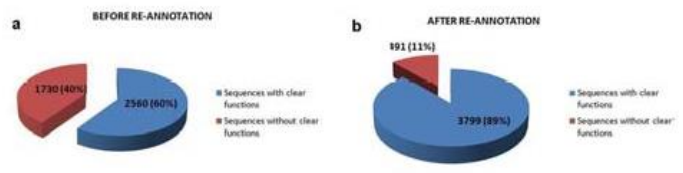

图 1 EcoCyc 基因组数据

注：a, EcoCyc 数据中已知序列和未知序列数据的百分比；b, 重新注释后的基因组数据，已知序列和未知序列数据的百分比

(REC-DB 数据)

Figure 1 EcoCyc Genome Data

Note: a. A pie chart describing the percentage of known and unknown sequences in the original data downloaded from EcoCyc. b. Genome data after re-annotation. A pie chart representing the percentage of known and unknown sequences after re-annotation (Rec-DB data)

1.1 未知蛋白的功能注释

在 1730 个未知的蛋白质序列中欧，有 1239 个序列现在已被分配具有明确的功能。从重新注释的结果分析，约 156 个蛋白质的功能没有明确的定义。例如 ec2389 蛋白序列的功能在重新注释过程中，利用不容的工具注释的功能是不同的(表 2)，从 Pfam 中注释为“metallo-beta lactomase 超级家族”、COG 注释为“锌依赖的水解酶”、ProDom 注释为“可能的水解酶”，但是在 BLAST 和 ScanProsite 中并未有注释信息。在些 Pfam 和 ProDom 的注释中产生类似的功能，但是 Pfam 给了一个不同的功能。需要注意的问题是，ProDom 给出的注释是“可能”的一种功能，并且突出显示在表 2 中，因此该分析结果不予考虑。同样的，蛋白质序列 4267 在 Pfam 和 ScanProsite 分析中注释为“硫”的功能，但其他分析工具注释为“乙酰转移”的功能。对于某些蛋白，虽然注释着不同的功能，但是其功能描述似乎是相似的，但是其他的并没有明确的定义。在这种情况下，几乎很难对其功能做出准确的定义。而进一步的功能注释，必须在设计生化实验验证的情况下进行。两个样本中，相同序列通过不同的工具分析具有不同的功能，均由手工标注，如表 2 所示。

例如保守假定蛋白 ec1270 现在已经被预测为核糖核酸内切酶，而且在 REC-DB 数据库的其他数据现在是可用的。根据以前的注释，有 1730 个未知序列(40%的序列)，由于重新注释，未知蛋白序列减少到 491 个(图 2b)。这表明，只有 11%的蛋白质被认为的未知或没有功能。因此，大肠杆菌的整体重新注释的结果显示，对于未知序列的分析有效率达到了 29%。

表 2 重新注释问题样品集合

Table 2 Sample problem sets after re-annotation

Results	Sequence ID 2389[a]	Sequence ID 4267[a]
PFAM	Metallo-beta lactomase[b]	Thiolase[b]
COG	Zinc dependent Hydrolases[b]	Acetyl CoA Acetyl transferases[b]
SCANPROSITE	No Hits	Thiolase enzymes[b]
BLAST	No Hits	Acetyl CoA Acetyl transferases[b]
PRODOM	Probable Hydrolase[c]	Probable Acetyl Transferase[c]

注: a 表示 Rec-DB 数据控的序列 ID; b 表示不同工具注释的不同功能; c 表示功能注释为“可能”

Note: a Sequence ID in Rec-DB database; b Different functions from different tools; c Functions predicted with a negative term “Probable”

1.2 蛋白功能的改变(修改蛋白功能)

对于不完全注释的序列进行重注释，结果导致现有蛋白注释的功能发生新的改变。例如(补充表 S1)，蛋白质序列 ec1034 最初被标注为 MEND-MONOMER 补体(2377281-2375611)，但重注释后它被作为 2-琥珀酰-6-羟基-2,4-环己二烯-1-羧酸酯合成酶。因此，这种蛋白功能的修改是非常有用的，其可以促进科学界对蛋白进行更精确和可靠的功能定位。这些数据在 REC-DB 数据库中是可信的(http://recdb.bioinfo.au-kbc.org.in/recdb/).。

1.3 蛋白功能的更新

基于数据库中可用的信息对大肠杆菌的重新注释，也导致对先前已注释功能的蛋白的相关功能进行了更新。在最初的注释中，一些蛋白质被指定具有一定的功能，但是实际上并不是足够的了解它们的实际的生物作用。我们的重新注释研究，有助于对这部分蛋白序列增加更详细的功能。例如(补充表 S1)，蛋白质序列 ec1915 曾被注释过一个功能–“还原酶”，但这并不是个详细的功能注释。但我们的研究结果帮助我们重新注释为“氧化还原酶钼

蝶呤结合域”。对这种类型的蛋白质数据的完整列表可在 REC-DB 数据库查看。这些蛋白功能的更新信息，反过来也将有助于研究人员深入的研究这些蛋白相关的分子系统。这些完整的重新注释的功能信息在补充表 2 中可查阅到。从 REC-DB 数据库中，我们能够检索到清晰注释功能信息的基因(补充表 S2, A)，这些基因都是经过重新注释过的结果。其次，假定基因的靶点也可以获得，这些基因都是没有未知序列，且没有进行功能的预测(补充表 S2, B)，那么它们都需要进一步的进行功能的注释(补充表 S2, C)。

1.4 不确定的结果

虽然我们的重新注释研究发现在更新大肠杆菌基因组功能信息上是有效的，但是仍然存在一些不确定的结果，这些难以确定的结果如下。

1.4.1 传递突变

传递突变是一种分析结果现象，即基于序列的相似性，一个蛋白的功能被转移到另外一个搜索到的具有相似序列上，但是这个蛋白最原始的名称是错误的(Salzberg, 2007)。随着更多基因组注释工作的完成和 BLAST 的执行，某些蛋白序列的功能发生了转移改变。这些分析误差也是众所周知的，通过数据库，在基因组数据资源中有成千上万这样的功能传递错误。因此，在这样的错误注释的信息被传播的情况下，通过使用重新注释的序列数据库，然后传递的这些错误的注释信息，进而导致了假阳性的功能预测的出现。这些错误的注释信息仍然难以处理，更无法对这些蛋白的注释信息作出关键性的决定，因此，这样的问题，还有待科学界的专家给出处理建议。

1.5 REC-DB

本研究结果已经作为公共数据库在线发表，命名为“REC-DB-重新注释的大肠杆菌数据库”。一些新增的功能已经添加到这个数据库，以便于搜索蛋白功能。在这个数据库，用户可以检索重新注释的大肠杆菌基因组数据，可以查询 REC-DB 登录号(如 ec001)，或者通过选择 GenBank ID (GI.No.90111633)，或者查询 Gene ID (GENE-ID. 948195)。在查询过程中，用户可能会在搜索选项中发现“Null”、“No GI”和“No Gene id”，如果是“Null”，那则意味着没有 REC-DB 注释的功能；如果是“No GI”，那则意味着没有 GenBank ID；如果是“No Gene id”，那则意味着在 REC-DB 数据库中没有 gene_id。

2 讨论

虽然基因组项目具有让我们更好地了解生物的潜能，但是缺乏更新以及准确的基因组功能注释限制了它对生物体进行进一步研究的能力。因此，在 In Silico 功能蛋白质组学的重注释中，结合自 1997 年开始的大量原注释数据，我们试图大量更新大肠杆菌 K-12 整个序列的功能。这一研究将会得到很多大肠杆菌 K-12 编码的分子蛋白的功能。

用常规 BLAST 程序来分析一个单一序列(http://www.ebi.ac.uk/Tools/BLAST/)，将会根据比对度而产生大量的结果，并伴随着 E-值、百分比、相似性、BLAST 分数和序列长度这些参数的产生。从 BLAST 获得的结果会包含大量地比对分数，和最佳的 E 值 1×10^{-6} 到 1×10^{-52} (Gabriel et al., 2008)。这需要大量的人为干预来获得所需的比对选项，从而得到所需的结果。因此，利用常规的 BLAST 程序来分析大肠杆菌的整个蛋白质组将会变得繁琐麻烦(Aravindhan et al., 2009; Hulo et al., 2004)。具有良好结构和简明搜索方式的 AIM-BLAST 使得我们能够更好地对大肠杆菌的完整基因进行高效的序列比对(Aravindhan et al. 2009)。先前预测和注释的大肠杆菌 K-12 的蛋白序列已经通过利用各种不同的方法进行手工重分析，如基于相似性的的搜索方法(BLASTP)，基于模式的搜索方法(ScanProsite)，基于系统发育分类的搜素(COG)，基于域的搜索(ProDom)，基于蛋白家族的搜索(Pfam)。每种方法都有不同的侧重点和收集到不同的信息集。此外，由于这些数据库被设计用来解决特定的问题，因此也包括了固有的优势和弱点(Rust et al., 2002)。

我们的重注释策略帮助预测近 29%的蛋白序列功能。例如(附表 S1)，ec0903 序列编码的假定蛋白已经被重新注释为甲酸脱氢酶。关键的是，ec0903 的 Refseq ID yp_001165334.1 之前被报道为假定蛋白且被用于人类传染病的疫苗预测(Xiang and He, 2009)。重注释也有助于修改以前注释的蛋白质。在一些情况下，原始的注释是比较广泛和不精确的。在分析这些序列中，我们发现它们被注释成一个假阳性的功能。

我们手动进行重注释后，发现几乎 29%的基因在之前没有被确定，现在已经赋予其一个已知的功能。REC-DB 数据应该对于分析大肠杆菌基因组编码的产物是很有用的。因此，我们相信，我们的重注释应该有助于科学界对大肠杆菌的研究。

3 材料与方法

3.1 蛋白序列

大肠杆菌 K-12 的完整蛋白组序列可以在 EcoCyc 数据库中下载到(Keseler et al., 2005)，并对其进行分析。先前对大肠杆菌功能基因组的分析表明共有 4290 个蛋白序列，但只有 60%被发现具有已知的功能。剩下 40%的功能是未知的。未知基因的蛋白功能，例如假定蛋白和保守假定蛋白必须被再次分析，因为它们在微生物的细胞生理学上具有很重要的作用。假定蛋白就是指那些与其它序列不具相似性或者缺乏实验证据支持的蛋白。保守假定蛋白是指那些没有系统发育分布的蛋白(Tao et al., 1999)。

3.2 序列的功能注释

完整的基因组由 4290 个蛋白序列组成，其中 2560 个蛋白序列的功能已知，1730 个序列功能是未知的(表 1)。计算机在分析序列和重测序扮演中很重要的角色，在于它可以缩短分析大量数据的时间和结合多种方法(Nascimento and Bazzan, 2005)。在这里，完整的大肠杆菌蛋白质组的功能重注释是通过利用先进的重注释策略，它整合了许多清晰高效的序列分析方法(图 2)。

3.2.1 AIM 进行相似性搜索

进化上相关的蛋白通常被称为同系物，而相近的同系物通常具有相似的功能(Ofran et al., 2005)。基于这个概念，凡是基于同源性或相似性的功能注释转移仍然是一个原始的预测位置蛋白功能的方法。反过来，由于在蛋白质数据库中新的生物信息的越来越快的积累，这种相似性的搜索方法也将有助于修改或更新以前注释的功能。

BLAST 是一种基本的局部比对搜索工具，是人们最喜欢和使用最普遍的生物信息学程序之一，用于识别生物序列中间的相似性(Altschul et al., 1990)。这个工具仍然存在计算密集和费时的问题，因为他们使用大量的数据进行传输，这需要人为干预来分析 BLAST 的结果，而且 BlAST 比对分析的 E 值阈值是 1×10^{-6} 到 1×10^{-52} (Gabriel et al., 2008)。为了克服这些困难，我们开发了一个程序 AIM-BLAST，其已经整合了 EBI (的欧洲生物信息学研究所)的 AJAX 和 SOAP 服务，能够支持在重注释分析中实现多序列比对。此外，AIM-BLAST 具有增强功能，即对 Blast 结果进行自动解析，呈现为“一个序列一个功能”的方式与人工自动管理方式。

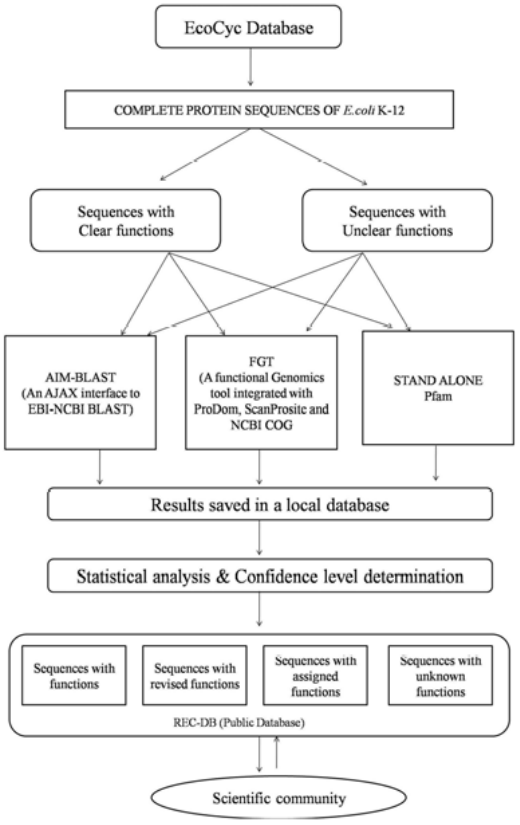

图 2 重测序的“动态生物数据融合”策略示意图

注: 这个流程图描述程序是一步一步进行大肠杆菌 K-12 菌株重新诠释的步骤

Figure 2 Schematic representation of the Re-annotation strategy “Dynamic biological data fusion”

This flow chart describes the step by step procedures for carrying out the re-annotation of E.coli K12 strain.

3.2.2 序列和注释模式搜索的控制

对基因组序列功能的注释描述，对于研究人员在实验室的分析和计算推论来说是非常重要的。单纯的基于 BLAST 分析的重注释并不会有助于进行准确的功能重新预测和注释。在某些情况下，这些注释可能并不是一致的、不完整的，或者是错误的(Karp et al., 2007)。因此，基于若干不同策略的序列分析，例如模式的搜索、系统发育分析搜索、基于结构域的搜索和基于同源家族的搜索，能有效地注释大肠杆菌 K-12 蛋白质组的信息。当比较蛋白质序列时，他们可能并不具有相似的具有明确功能的短的序列区域。因此，确定这样特别的蛋白序列信息，有助于对一些相似的蛋白进行重注释功能的初步预测。一些数据库致力于找出这种模式是可用的，ScanProsite (http://expasy.org/tools/scanprosite/)被选为在蛋白数据库中通过这种特殊的模式里寻找比对位点。这个工具利用 ProRules 环境依赖性来对模

板进行注释，通过扫描蛋白序列发现功能和内部结构域残基，实现可能的模式以及对功能进行预测(Hulo et al., 2004)。

3.2.3 基于系统发育分类的方法

同源蛋白簇数据库包含的分类信息，是基于蛋白序列的同源性而建立的(Tatusov et al., 2000)。COGs 数据库是最好的基于基因组的进化而对蛋白序列进行功能注释的门户数据库。为了促进蛋白功能注释的研究，COGs 已经被分类为 17 个广泛的功能类别，包括普通的功能预测，通常只是作为生活活性的预测，是一类可行的非特异性的 COGs 数据。此外，已知功能的同源蛋白簇已经用来代表特定的细胞系统和生化途径。因此，采用 COG 数据库的 COGNITOR 程序对序列进行分析，能够在基因组进化的基础上更深入的了解蛋白的功能。

3.2.4 基于结构域的搜索

基于家族的分类也仍然是为生物序列提供功能注释的一个重要手段。Pfam 是一个集多序列比对和 HMM 分析来预测蛋白家族的的程序。蛋白质的功能信息可以通过 HMMs 比较 Pfam 数据库中的序列来实现的(Wu et al., 2003)。基于这些蛋白家族的功能，利用 Pfam 数据库对大肠杆菌的全基因组进行分析和功能预测。虽然 Pfam 呗 FGT 程序进行了集成，到那会 Pfam 的分析依然分别进行，且阈值设定为默认的 0.001。这是因为，Pfam 分析一般会消耗更多的时间，而且在 FGT 上运行 Pfam 去分析一个大样本的蛋白数据，通常会影响到 FGT 的性能，进而降低整体的运算过程。同时，在 Pfam 服务商直接的运行单独的序列，也是一个无聊枯燥的过程。因此，Pfam FTP 文件被下载并安装到本地系统，一个单独的 Pfam 被设计来重新支持大规模序列的重测序分析。

3.3 注释的过程

我们选择 ScanProsite、COG 和 ProDom 对完整的大肠杆菌蛋白质组进行重新分析，因为它们是利用不同的策略来分析蛋白质的生物功能的，但是也有些分析并不是能够普遍使用这些分析工具的。这些工具并不允许在一个实例中进行多重序列比对。此外，对于每一个单一的序列的搜索，这些工具会产生许多匹配，用户必须仔细理解它们，并选择最佳的比对结果。选择最佳比对结果后，用户必须将适当的功能复制到本地数据库进行最终注释。此外，当在一个给定的时间使用这些工具分析序列时，用户必须同时打开和处理多个浏览器，这又会消耗过多的人力和时间。因此，使用所有这些工具来分析整个大肠杆菌的蛋白质组数据时，这又是一个巨无聊的过程。

3.4 内部的功能基因组学分析工具

了解多个复杂的处理分析工具的同时，一个简单而且新颖的分析系统 Bioinfotracker (一个功能基因组分析工具 FGT)诞生了，其是通过同时使用不容的在线的功能预测工具，用于在本地执行的蛋白质序列的注释(Kumar et al., 2009)。FGT 是一个结构良好的、灵活的、高度系统化功能分析程序，可供我们进行大规模的蛋白质注释。不同的在线工具，期分析的策略往往也是不一样的，例如 ScanProsite、ProDom、COG 和 Pfam 的功能都集成到 FGT 这个工具中了。一旦某个序列被提交到这个工具上，序列会被转发，并提交到不同的服务器，但是工具集成及其过程是在单个服务器上串联进行的。在 ScanProsite 工具上，首先会对我们使用的模式进行 PRSOTE 收集性扫描。COG 分析会设置的 E 值，Pram 分析设置的 E 值为 0.001。完成之后，当结果判断为可用的时候，这些分析工具会自动解析这些结果，并选择最佳的功能注释，并从相应的服务器上获取对应的结果。此外，所提交的序列的结果最好提供了一个简单的表格形式，这将更容易进行解释。因此，FGT 分析工具是非常支持开展对大肠杆菌 K-12 进行生物体完整的蛋白组序列的重新注释。

在这里，大肠杆菌 K-12 的 In Silico 功能蛋白组学的重新注释采用了严密的和熟练的注释的过程，涉及来自不同数据库的生物学数据进行动态融合(图 2)。大肠杆菌 K-12 的全基因序列从 EcoCyc 数据库下载获得，对所有具有明确功能或者无明确功能的蛋白序列都进行了初步的分析。然后，所有序列被提交到 AIM BLAST 进行 BLAST 分析，在 FGT 分析工具上进行 Prom、ScanProsite 和 COG 分析，以及以独立的蛋白家族为基础的 Pfam 分析。所有 5 个分析工具分析得到的结果被储存在本地数据库，以便未来进行进一步的分析。当所有序列的分析结果是可信的时候，就进行统计学分析，以确定预测的功能的置信水平。

作者贡献

GRK 完成整体数据的管理、软件的开发和论文的写作；TKS 完成软件的开发和论文的写作；CPR 完成数据分析和解释；LPL 参与软件的开发和数据

的分析，全体作者都阅读并同意最终的文本。

致谢

作者非常感谢 Aravindan Ganesan、Kalyanamoorthy Subha 和 R Sathish Kumar 对本研究的建议，以及在论文撰写过程中给予的帮助。

参考文献

Altschul S.F., Gish W., Miller W., Myers E.W., and Lipman D.J., 1990, Basic local alignment search tool. J. Mol. Biol., 215: 403-410

Ashok S., Venil S., Nupoor C., and Kumar G.R., 2014, Prediction and classification of ABC transporters in Geobacter sulfurreducens PCA using computational approaches. Current Bioinformatics, 9(2): 166-172

Aravindhan G., Kumar G.R., Kumar R.S., and Subha K., 2009, AJAX Interface: A Breakthrough in Bioinformatics Web Applications. Proteomics Insights, 2: 1-7

Aravindhan G., Kumar R.S., Subha K., Subazini T.K., Dey A., Kant K., and Kumar G.R., 2009, AIM-BLAST-AJAX Interfaced Multisequence Blast, Proteomics Insights, 2: 9-13

Blattner F.R., Plunkett G., Bloch C.A., Perna N.T., Burland V., and Riley M. et al., 1997, The complete genome sequence of Escherichia coli K-12. Science, 277: 1453-1474

Bock J.R., and Gough D.A., 2004, In silico biological function attribution: a different perspective. Drug Discov Today Biosilico, 2: 30-37

Camus J.C., Pryor M.J., Médigue C., and Cole S.T., 2002, Re-annotation of the genome sequence of Mycobacterium tuberculosis H37Rv, Microbiology, 148: 2967-2973

Dandekar T., Huynen M., Regula J.T., Ueberle B., Zimmermann C.U., and Andrade M.A., 2000, Re-annotating the Mycoplasma pneumoniae genome sequence: adding value, function and reading frames. Nucl. Acids Res., 28: 3278-3288

Elmore M.T., Potok T.E., and Sheldon F.T., 2003, Dynamic Data Fusion Using An Ontology-Based Software Agent System. Proceedings of the 7th World Multiconference on Systemics, Cybernetics and Informatics, 1-6

Finn R.D., Tate J., Mistry J., Coggill P.C., Sammut S.J., Hotz H.R, Ceric G., Forslund K., Eddy S.R., Sonnhammer E.L, and Bateman A., 2008,The Pfam protein families database. Nucl. Acids Res., 36: 281-288

Gabriel M.H., and Kristen L., 2008, Choosing BLAST options for better detection of orthologs as reciprocal best hits. Bioinformatics, 24: 319-324

Galperin M.Y., and Koonin E.V., 2010, From complete genome sequence to 'complete' understanding? Trends Biotechnol., 28: 398-406

Gundogdu O., Bentley S.D., Holden M.T., Parkhill J., Dorrell N., and Wren B.W., 2007, Re-annotation and re-analysis of the Campylobacter jejuni NCTC11168 genome sequence. BMC Genomics, 8: 162-170

Hulo N., Sigrist C.J., Le Saux V., Langendijk-Genevaux P.S., Bordoli L., Gattiker A., De Castro E., Bucher P., and Bairoch A., 2004, Recent improvements to the PROSITE database. Nucl. Acids Res., 32: 134-137

Karp P.D., Keseler I.M., Shearer A., Latendresse M., Krummenacker M., Paley S.M., and Paulsen I., 2007, Multidimensional annotation of the Escherichia coli K-12 genome. Nucl. Acids Res., 35: 7577-90

Keseler I.M., Collado-Vides J., Gama-Castro S., Ingraham J., Paley S., Paulsen I.T., Peralta-Gil M., and Karp P.D., 2005, EcoCyc: a comprehensive database resource for Escherichia coli. Nucl. Acids Res., 33: 334-337

Kumar G.R., Aravindhan G., Subazini T.K., and Kumar R.S., 2009, Bioinfotracker: A novel system for advanced genome functional insight. Journal of Bioinformatics and Sequence Analysis, 1(3): 046-049

Nascimento L.V., and Bazzan A.L., 2005, An agent-based system for re-annotation of genomes. Genet. Mol. Res., 4: 571-580

Ofran Y., Punta M., Schneider R, and Rost B., 205, Beyond annotation transfer by homology: novel protein-function prediction methods to assist drug discovery, Drug Discov. Today Biosilico., 10: 1475-1482

Pearson W.R, and Lipman D.J., 1988, Improved tools for biological sequence comparison, Proc. Natl. Acad. Sci. USA, 85: 2444-2448

Rajadurai C.P., Subazini T.K., and Kumar G.R., 2011, An integrated re-annotation approach for functional predictions of hypothetical proteins in microbial genomes, Current Bioinformatics, 6(4): 450-461

Riley M., Abe T., Arnaud M.B., Berlyn M.K., Blattner F.R., Chaudhuri R.R., and Glasner J.D., 2006, Escherichia coli K-12: a cooperatively developed annotation snapshot, Nucl. Acids Res., 34: 1-9

Roberts R.J., Chang Y.C., and Hu Z. Rachlin J.N., Anton B.P.,

Pokrzywa R.M., Choi H.P., Faller L.L., Guleria J., Housman G., Klitgord N., Mazumdar V., McGettrick M.G., Osmani L., Swaminathan R., Tao K.R., Letovsky S., Vitkup D., Segrè D., Salzberg S.L., Delisi C., Steffen M., Kasif S., 2011, COMBREX: a project to accelerate the functional annotation of prokaryotic genomes. Nucl. Acids Res., 39: D11-D14

Rust A.G., Mongin E., and Birney E., 2002, Genome annotation techniques: new approaches and challenges. Drug Discov. Today Biosilico., 7: 70-76

Salzberg S.L., 2007, Genome re-annotation: a wiki solution? Genome Biol., 8: 1-5

Schnoes A.M., Brown S.D., Dodevski I., and Babbitt P.C., 2009, Annotation Error in Public Databases: Misannotation of Molecular Function in Enzyme Superfamilies, PLoS Comput. Biol., 5: e1000605

Serres M.H., Gopal S., Nahum L.A., Liang P., Gaasterland T., and Riley M.A., 2001, functional update of the Escherichia coli K-12 genome. Genome Biol., 2: 1-7

Servant F., Bru C., Carrère S., Courcelle E., Gouzy J., Peyruc D., and Kahn D., 2002, ProDom: Automated clustering of homologous domains. Briefings in Bioinformatics, 3: 246-251

Tao H., Bausch C., Richmond C., Blattner F.R., and Conway T., 1999, Functional genomics: expression analysis of Escherichia coli growing on minimal and rich media. J. Bacteriol., 181: 6425-6440

Tatusov R.L., Natale D.A., Garkavtsev I.V., Tatusova T.A., Shankavaram U.T., Rao B.S., Kiryutin B., Galperin M.Y., Fedorova N.D., and Koonin E.V., 2001, The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res., 129: 22-28

Valencia A., 2005, Automatic annotation of protein function, Curr. Op. Struct. Biol., 15: 267-274

Wood V., Rutherford K.M., Ivens A., Rajandream M.A., and Barrell B., 2001, A re-annotation of the Saccharomyces cerevisiae genome. Comp. Funct. Genomics, 2: 143-154

Wu C.H., Huang C.H., Yeh L.S., and Barker W.C., 2003, Protein family classification and functional annotation. Comput. Biol. Chem., 27: 37-47

Xiang Z., and He Y., 2009, Vaxign: a web-based vaccine target design program for reverse vaccinology. Procedia in Vaccinology, 1: 23-29

Zheng Y., Roberts., and Kasif S., 2002, Genomic functional annotation using co-evolution profiles of gene clusters, Genome Biol., 3: 1-9

