生物信息计算

章 乐 主编

科学出版社

生物信息计算

章　乐　主编

科学出版社
北　京

内 容 简 介

本书重点关注生物信息学中的计算问题，立足于生物学及工程方向前沿技术和问题的研究，向读者展示生物学新方向并提供相关的应用实例。本书详细介绍高性能计算和生物医药大数据研究，Hadoop 和 Spark 在基因大数据挖掘方向的应用，天津、长沙、广州超算中心的生物医药健康大数据平台。从最基本的云计算理论体系入手，介绍云计算在健康大数据方向的部署和应用，包括云计算体系和架构及健康云计算体系结构，并对其中的关键技术进行讨论。针对生物影像技术，不仅对传统的生物影像处理技术进行详述，同时结合人工智能、机器学习技术在此方面的应用进行解读和挖掘。在基因序列比对问题上，针对双序列比对和多序列比对进行展开。最后重点介绍蛋白质结构生物学问题，从蛋白质结构预测、比对等方面展开问题，并对相关理论和技术进行详细的解读。

本书适用于生物医学、生物信息学、计算机科学、计算机工程和相关专业的高年级本科生及低年级研究生，也适用于使用 GO 语言、具有坚实的计算和编程技能并希望能够开展生物信息计算的研究者和实践者。

图书在版编目(CIP)数据

生物信息计算 / 章乐主编. — 北京：科学出版社, 2018. 5
ISBN 978-7-03-057077-2

Ⅰ. ①生… Ⅱ. ①章… Ⅲ. ①生物信息论 Ⅳ. ①Q811.4

中国版本图书馆 CIP 数据核字 (2018) 第 064832 号

责任编辑：张 展 黄明冀 / 责任校对：梁晶晶
责任印制：罗 科 / 封面设计：墨创文化

科 学 出 版 社 出版
北京东黄城根北街16 号
邮政编码：100717
http://www.sciencep.com

成都锦瑞印刷有限责任公司印刷
科学出版社发行 各地新华书店经销
*
2018 年 5 月第 一 版 开本：B5（720×1000）
2018 年 5 月第一次印刷 印张：10
字数：200 千字
定价：39.00 元
（如有印装质量问题，我社负责调换）

《生物信息计算》编委会名单

主　编：章　乐

副主编：强　彦　彭绍亮

编　委：蒲　讯　陶文静　刘治平　吴红艳　王晓伟

序

交叉学科的发展不是多学科简单的加减，而是建立一种新的学科体系，包括对基础学科的再认识和再创造。

生物信息学是在生命科学的研究中，以计算机为工具对生物信息进行储存、检索和分析的科学。它是当今生命科学和自然科学的重大前沿领域之一。

章乐教授是一位在生物信息领域工作多年的优秀教师,由章乐教授主编的这本《生物信息计算》分别从计算机的三个传统二级学科的角度来阐述该领域的前沿技术。第一章探讨生物信息计算在软件工程上的应用范例。第二和第三章探讨生物信息计算在体系结构上的应用与研究。第四，五和六章探讨生物信息计算在计算机科学应用上的成果。

通过本书，读者可进一步体会和分析生物和计算机学科发展的方向和趋势，使读者能够在这个瞬息万变的时代赶上技术发展的步伐。最重要的是，该书的出版可以很好地吸引更多的优秀学生和研究人员加入到生物信息学这个令人激动的研究领域中来，推动生物信息学的持续发展。

希望本书的出版能够对生物信息学科的发展起到良好的促进作用！

中国科学院院士　陈润生

2017 年 11 月

试读结束：需要全本请在线购买：www.ertongbook.com

前　言

随着生物工程、医药和信息学的不断交叉融合，特别是近几年生物工程、医药信息标准化的发展为生物医学信息学打下了良好的发展基础，加上近年来软硬件发展迅猛，很多发达国家对医学信息学和生物信息学的发展前所未有地重视。生物信息学通常是在分子水平上采用应用数学、信息学、统计学、计算机科学、人工智能、化学和生物化学来解决生物学中所遇到的信息问题。采用这些技术的主要原因是生物学问题往往规模庞大，用人力无法进行鉴别和演算。计算生物学的研究往往和系统生物学重叠，主要涉及序列比对、基因发现、基因组聚焦、蛋白质结构路线、蛋白质结构预测等诸多领域。医学信息学则是信息学、计算机科学、医学卫生保健等相互交叉的学科，它主要研究医学卫生领域的数据（包括资源、设备、方法等），从而优化健康和医学数据的采集、存储、检索和信息利用，它是处理医学健康数据的工具。当前，生物学数据量和复杂性不断增长，每 14 个月，基因研究产生的数据就会翻一番，单纯地依靠观察和实验已难以应付。因此，必须依靠大规模计算机模拟技术，从海量信息中获取最为有用的数据。各种计算方法已开始广泛应用于药物研究，以及研发创新的、具有自主知识产权的疾病靶标、信息学分析系统等。同时，运用计算生物学，科学家有望直接破译核酸序列中的遗传语言规律，模拟生命体内的信息流过程，从而认识代谢、发育、进化等一系列规律，最终为人类造福。

本书以计算机科学技术的三个二级学科为纲要，以数据库设计为例讨论生物信息计算在软件工程学科上的应用，以云计算和高性能计算为例讨论生物信息计算在系统结构学科上的应用，以基因、蛋白质和影像分析为例讨论生物信息计算在计算机应用学科上的应用。系统而全面地将生物医学信息学相关理论与高性能计算、生物医药大数据、云计算、影像处理、序列比对、蛋白质组学等多方向的技术实践编写成本书，书中列举的应用更加具有创新性和针对性。为说明和总结书中描述的主要技术和概念，使用了大量的图片和具体的源码实例，同时在各章结束处例出了详尽的参考文献。

本书对生物信息计算领域的理论体系和相关应用实践提供了实用性的介绍。“实用”一词应该被理解为让读者/学生借助从书中获得的知识来制定切实可行的项目，书中囊括了相当有实战操作性的技术和典型应用。因此，本书不仅适合其原有目标受众(如生物医学、生物信息学、计算机科学、计算机工程和相关专业的

高年级本科生和低年级研究生)作为指导教材使用，也适合具有坚实的计算和编程技能并希望能够开展生物信息计算的研究者和实践者参考阅读。

感谢重庆市中迪医疗信息科技股份有限公司，浙江微松冷链科技有限公司，国家超级计算长沙中心、化学生物传感与计量学国家重点实验室，北京以利天诚科技有限公司在本书写作过程中的支持，尤其感谢他们提供的超算资源和大数据实验研发平台——冰山大数据实验与科研系统(www.ylitech.com)，感谢国家科研重大研究专项 2018ZX10201002，重点研发计划 2017YFB0202602、2017YFC1311003、2016YFC1302500 和国家自然科学基金 61372138、61772543 以及重庆杰出青年基金 cstc2014jcyjjq40003 等项目的支持。

目　录

第 1 章　生物医学信息的知识库工程技术

1.1　现有生物医学信息知识库系统简介

生物医学是医学的分支，主要是将生物技术和其他自然科学理论应用于临床实践。生物医学信息学致力于在生物学、生物医学科学、医学和医疗保健领域的工作实践中，对计算机科学、信息科学、信息学、认知科学和人机交互进行研究与应用。

当前，生物医学和生物医学信息学的研究日新月异。发表的生物医学文献越来越多[1]。生物学领域内的信息爆炸，使研究人员难以掌握最新的生物医学知识和网络信息资源[2]，因此构建生物医学领域的知识库具有重要意义。

生物医学知识库系统针对生物医学信息数据量大、数据类型复杂、资源分布不平衡、利用程度低下、标准不统一等问题。在标准的语义网络中，整合生物医学本体和多类型文本资源，并融合多层次生物信息数据，构建统一、规范化、机构化、开放共享、高效准确和可自动更新的生物医学信息知识库系统。

调研文献可知[3]，发达国家和地区高度重视生物医学信息知识库的建设。1986 年，美国国家医学图书馆(National Library of Medicine，NLM)主持了一项长期研究和开发计划，即一体化医学语言系统(unified medical language system，UMLS)。该研究计划旨在建立一个可计算化和可持续发展的生物医学检索语言集成系统与机读情报资源指南系统。其目的在于提高计算机程序“理解”用户提问中生物医学词汇含义的能力，并利用这种理解能力帮助用户检索和获取相关的文献资源[4]。由美国国家医学图书馆建立的美国医学文献在线分析和检索系统 Medline 是当前国际上最权威的生物医学文献数据库[5]。Medline 2004 数据库包含 12 500 000 条记录，并且还在以每年 50 万条记录的速度增加。21 世纪，包括美国国家生物技术信息中心(National Center for Biotechnology Information，NCBI)和欧洲生物信息研究所(European Bioinformatics Institute，EBI)在内的国际大型生物信息中心都在建设生物医学知识库。一些大型的信息技术公司也开展了生物医学知识库的开发，如 Watson 肿瘤治疗和临床应用系统，以及 GenGo、IPA、Pathway Studio 等专业平台。

国内方面，由于生物医学信息知识库建设起步不久，目前建成的专病医学知识库存在信息源单一和结构各异等问题。由中国医学科学院医学信息研究所/图书

试读结束：需要全本请在线购买：　www.ertongbook.com

馆开发的中国生物医学文献服务系统(SinoMed)，是集检索、免费获取、个性化服务和全文传递服务于一体的中西文整合文献服务系统。华中科技大学博士生游俊等设计构建的 BMELD 系统是多功能的学术期刊全文文献平台。该平台维护神经科学、生物信息学、生物医学光子学、蛋白质组学、医学影像专题等多个研究领域，可以为神经影像诊断和功能识别提供新的、准确的预测工具[6]。

下面按照数据库的功能对主要的生物数据库做表 1-1 所示的分类。

表 1-1 主要生物数据库

生物数据库	细 分
核酸序列数据库	Genbank：核酸序列数据库
	EMBL：核酸序列数据库
	dbEST：表达序列标签数据库
	Refseq：参考序列数据库
	UniGene：基因座数据库
蛋白质序列及相关数据库	SwissProt：蛋白序列数据库
	trEMBL：计算机注释的蛋白质数据库
	Interpro：蛋白质结构域和功能位点的整合数据库
	Pfam：结构域数据库
	PRINTS：蛋白质家族的指纹和模体数据库
	PROSITE：蛋白质功能位点数据库
	UniProt：全球蛋白资源数据库
蛋白质结构相关数据库	PDB：蛋白质三维结构数据库
	HSSP：同源派生的蛋白质二级结构数据库
	DSSP：蛋白质二级结构构象参数数据库
特定基因或蛋白质的数据库	ENZYME：酶数据库
	GPCRDB：内源性 G 蛋白偶联受体列表数据库
	REBASE：限制性内切酶和甲基化酶
人类基因突变及疾病相关数据库	OMIM：在线人类孟德尔遗传学数据库
基因功能数据库	GO：基因本体论数据库
	GOA：将 GO 应用到 SWISS-PROT，TREMBL 中的非冗余蛋白数据集中

1.2　本　　体

1.2.1　本体简介

生命科学领域，为了理解和治疗一种疾病，通常需要访问多个生命科学数据库。例如，为了了解抗痉挛药物，可能需要查询抗痉挛药物所含有的主要化合物、化合物的副作用[副作用数据库(side effect resource，SIDER)]、通路(KEGG pathway 数据库)、服用药物时会受影响的组织(DrugBank 数据库)等多个数据库。然而，这些数据库可能会使用不同的术语来表示同一个概念。例如，在不同的数据库里分别用“name”或者“DrugName”来表示药物名称。这让信息查找更加麻烦，数据库使用者需要熟悉各个数据库模式，类似于人类的方言，同时这也使计算机查找无章可循。

本体用一套控制字汇(controlled vocabulary)来描述实体与实体之间的联系和逻辑关系。本体类似于计算机的普通话，只要所有的用户都使用和遵守这套控制词汇，计算机就可以保证对事物的理解的一致性，并在此基础上对事物进行查找，以及对事物之间的逻辑关系进行推理。

1.2.2　GO 基因本体

GO 是基因本体联合会(Gene Onotology Consortium)为了让各种数据库中基因产物功能描述一致，在不同生物数据库中的查询具有一致性，并且在不同水平查询基因产物的特性而诞生的。它旨在建立一个适用于多个物种，可以对基因和蛋白质功能进行限定和描述的，并能随着研究不断深入而更新的语义词汇标准。GO 是多种生物本体语言中的一种，提供了三层结构的系统定义方式，用于描述基因产物的功能。目前，GO 包括三个部分：生物学过程(biological process)、分子功能(molecular function)及细胞组件(cellular component)。细胞组件用于描述亚细胞结构、位置和大分子复合物，如核仁、端粒和识别起始的复合物等；分子功能用于描述基因、基因产物个体的功能，如与碳水化合物结合、ATP 水解酶活性等；生物学过程指分子功能的有序组合，达成更广的生物功能，如有丝分裂或嘌呤代谢等。

GO 的结构是一个有向无环图，有点类似于分类树，不同点在于 GO 的结构中一个术语可以有不止一个父节点(图 1-1)。例如，生物过程中的己糖合成(hexose biosynthesis)有两个父节点，它们分别是己糖代谢(hexose metabolism)和单糖合成

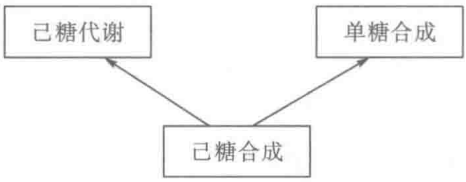

图 1-1 节点可以有多个父节点

(monosaccharide biosynthesis)，这是因为生物合成是代谢的一种，而己糖又是单糖的一种。

基因产物可能分别具有分子生物学上的功能、生物学途径和在细胞中的组件作用。当然，它们也可能在某一个方面有多种性质。例如，细胞色素 C 在分子功能上体现为电子传递活性，在生物学途径中与氧化磷酸化和细胞凋亡有关，在细胞中存在于线粒体中和线粒体内膜上。GO 中最基本的概念是术语(term)。GO 里面的每一个 entry 都有一个唯一的数字标记，形如 GO: nnnnnnn，还有一个术语名，例如“cell”“fibroblast growth factor receptor binding”或“signal transduction”。以下是一个术语示例。

gene_ontology.obo 示例：

```
[Term]
id: GO: 0000003
name: reproduction
namespace: biological_process
alt_id: GO: 0019952
alt_id: GO: 0050876
def: “The production by an organism of new individuals that contain some portion of their genetic material inherited from that organism.” [GOC: go_curators, GOC: isa_complete, ISBN: 0198506732]
subset: goslim_generic
subset: goslim_pir
```

```
subset: goslim_plant

subset: gosubset_prok

exact_synonym: "reproductive physiological process" []

xref_analog: Wikipedia: Reproduction

is_a: GO: 0008150!biological_process
```

1. 语义关系

基因本体论组织类似于图，语义作为图的节点，语义之间的关系为图中的边。因此，一旦产生新的语义，它与其他语义之间的关系也会同时被定义。语义之间的关系有三种：is a、part of 和 regulates。关系表示几点约定如图 1-2 所示：①“语义”用图论的术语“节点”表示；②习惯用父子节点来表示语义之间的关系，其中父节点离根节点较近，表示相对宽泛的语义，而子节点离叶子节点较近，相对父节点，其语义所代表的内容更为具体；③图 1-2 中的实线表示节点之间的关系；④虚线表示推理关系，该关系未被显式定义。其中，I 表示关系 is a；P 表示关系 part of；R 表示关系 regulates。

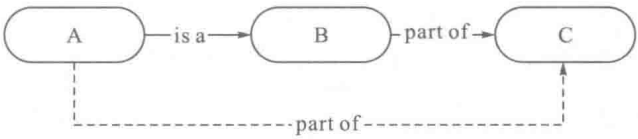

图 1-2　关系表示

该图表示：A 是一个 B，B 是 C 的一部分，因此可以推理出 A 是 C 的一部分。

part of 关系。GO 图具有树的性质，但与其不同的是，GO 图中节点不但可能具有多个子节点，而且可能具有多个父节点，且与不同的父节点具有不同的关系，如图 1-3 所示：线粒体(mitochondrion)便有两个父节点，因为线粒体既是一种细

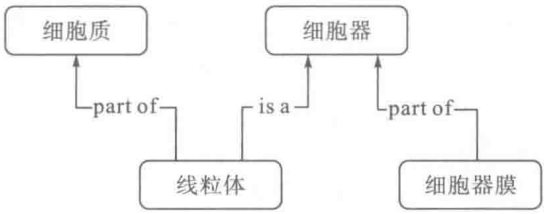

图 1-3　part of 关系

胞器(organelle)，又是细胞质(cytoplasm)的一部分。同样，细胞器(organelle)也有两个子节点，因为线粒体是一种细胞器(organelle)，细胞器膜(organelle membrane)是细胞器的一部分。

图 1-4 包含了上述的三种关系。

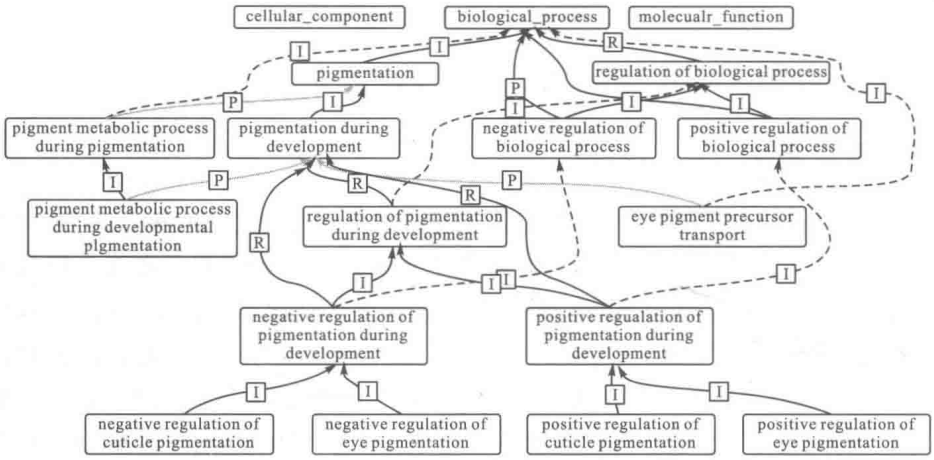

图 1-4 Go 本体的部分关系

数据来源：http：//geneontology.org/page/ontology-structure。

图中，I 表示 is a 关系；P 表示 part of 关系；R 表示 regulate 关系。

2. 关系的推理

(1) is a 关系的传递性。即 is a · is a → is a，如图 1-5 所示。

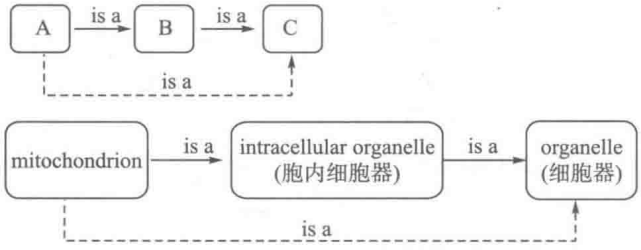

图 1-5 is a · is a 关系的推理

线粒体是一种胞内细胞器(intracellular organelle)，而胞内细胞器是一种细胞器官，由此可以推出：线粒体是一种细胞器官。

(2) part of 关系的传递性。即 part of· part of → part of。例如，线粒体是细胞质的一部分，细胞质又是细胞(cell)的一部分，从而可得出：线粒体是细胞的

一部分。

(3) part of· is a → part of, is a· part of → part of。即关系 is a 与 part of 组合后，其关系均为 part of(图 1-6)。例如，线粒体是细胞内器官，细胞内器官是细胞的一部分，因此线粒体是细胞的一部分。

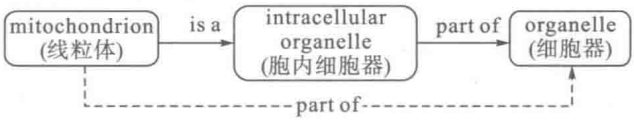

图 1-6　关系的推理

(4) 倒函数关系如图 1-7 所示。

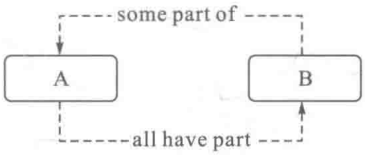

图 1-7　倒函数关系

(5) 其他的关系推导请参考 http://geneontology.org/page/ontology-relations。为了解决基因相关描述在各个数据库的表述不一致性，GO 项目最初是由 1988 年对三个模式生物数据库的整合开始：the FlyBase(果蝇数据库 Drosophila)，the Saccharomyces Genome Database(酵母基因组数据库，SGD)和 the Mouse Genome Informatics(小鼠基因组数据库，MGI)。从此，GO 不断发展扩大，现在已是包含数十个动物、植物、微生物的数据库(详见 GO Consortium Page)。一个基因或蛋白质可从三个层面进行注解(annotation)，首先是构成在细胞内的特定组件的基因或蛋白质，其次是此组件在分子功能上所扮演的角色，最后是基因或蛋白质参与的生物的文献资料及序列比较资讯为基础，将所有的真核生物的基因或蛋白质都在此系统(Gene Ontology)下作注解与分类(classification)。

1.2.3　本体建模

本体不仅适用于基因及其产物的描述，更因其具有知识表述的一致性、计算机可读性、知识的可推理性等优点，而被广泛应用于生物医学信息的知识库工程领域。下面讲述如何用 Protégé 工具对本体进行建模。

二元关系是通过 subject(S)、predicate(P)、Object(O)的三元组的形式来表示

试读结束：需要全本请在线购买：　www.ertongbook.com