# 前 言

目前，在谷歌上用“systems biology（系统生物学）”作为检索词，可以检索到1400万条信息，用“pathway analysis（途径分析）”作为检索词，可以检索到200万条信息。这些数据深刻地说明了系统数据分析（systems data analysis）在现代生物科学中无处不在。近年来，任何关于微矩阵的表达和SNP的文章都会有一个与途径、本体富集（ontology enrichment）或者生物网络有关的章节。尽管这是一个新兴领域，但是系统生物学已经被广泛地应用于基础研究、临床前药物研究到转化医学以及个体化的医疗保健等领域。

系统生物学在维基百科中的概念是：在生物系统中采用整合（integration）而不是还原（reduction）的新视角，系统地研究复杂的相互作用。从实际的角度来看，这意味着我们要将积累的生物学知识整合成可用计算机阅读的格式，还要创造出一些工具，利用这些工具可以将这些格式化了的信息应用到生物实验和化学实验的数据分析中。从20世纪70年代开始，生物化学成为第一个将信息录入数据库的学科，例如BRENDA、EMP/MPW和KEGG数据库。多年来，科学家把调节和信号组成的部分以蛋白质相互作用数据库（例如HPRD和BIND数据库）的形式增加到生物化学中。同时，关于细胞生物过程和蛋白质功能的理解也有一定发展，这其中最出名的就是GO数据库。

函数分析（functional analysis）与高通量的，或者说“组学”的实验生物学是不可分割的，从20世纪90年代后期得到了快速的发展。在基因组规模内，对数以千计的嘈杂组学数据进行分析对于当时的湿实验室①的研究者来说是很难的，而且统计学工具仅仅能帮助我们减少数据的复杂性，却不能帮助我们理解潜在的生物学原理。渐渐地，生物信息学家和湿实验室生物学家开始用已有的途径和过程信息的数据库实现实验数据的绘制（map）和优先化（prioritization）（富集分析）。后来从多年的图论和物理学研究中借鉴了一些经验，生物学网络被加入到分析工具箱里。

本书汇编了函数分析方法及其应用，由学术领域、政府研究机构、医药产业和生物信息学公司的一线从业人员编写。该书的内容分为三个部分。第一部分阐明蛋白质、化合物和基因间的相互作用。第二部分主要介绍了与函数分析相关的工具。第三部分讲述了函数分析的应用。

二元的相互作用是细胞内蛋白质、基因和RNA行使功能的基本单位。在函数分析中，这些相互作用是计算机模拟的细胞生命即网络和途径中的重要组成成分。这些函数分析结果的质量取决于网络和途径组成的相互作用的质量。因此，揭示相互作用的方法具有重要的意义。在这里，我们列举了一些著名的湿实验室和干实验室②获得相互作用信息时用到的工具。我们可以通过生物学实验或者已经发表的实验数据等多种方式获得蛋白质的相互作用信息。湿实验室工具包括可以全面分析蛋白质与蛋白质之间物理上的相互作用的蛋白

① 湿实验室：主要指实验室内的实验研究

② 干实验室：主要指计算机模拟和理论分析

质芯片（Invitrogen 公司），和一个可以更准确检测转录因子和基因启动子之间的相互作用的芯片（A-Star 公司，新加坡）。在基于计算机的方法中，不同形式的文本挖掘和自然语言处理起到了越来越重要的作用。Linguamatics 公司是一个专注于从 PubMed 和其他数据库中进行文本挖掘相互作用信息并且发展迅速的公司，这里的研究人员用一个章节介绍了与文本挖掘相关的内容。当然，我们可以在实验数据里挖掘出相互作用信息，然而手工注释的过程是缓慢的、耗费精力的但却是高质量的。在本书中专门介绍注释技术的专家来自两个方面，一部分专家来自生物学领域，以蛋白质-蛋白质以及蛋白质-DNA 或者与 RNA 的相互作用和代谢为研究重点（Metalogic 公司，俄罗斯），另一部分专家来自化学领域，以蛋白质-化合物的相互作用为研究重点（GVK 公司，印度）。

多数函数分析的方法（并不是全部）可分为两类：基因名称表富集分析（EA）和生物学网络。第一类包含许多步骤，我们利用这些步骤计算出我们感兴趣的数据集在特定的途径、细胞学过程、蛋白质功能、疾病靶向和生物标记物中的富集程度。基因名称表富集分析的结果是在数据内部对于实体进行基于特定打分方法的定量排序。在这里，BMS 公司研究人员介绍了关于基因名称表富集分析中经典的 GSEA 方法；斯坦福国际研究院（SRI International）的研究人员用一个章节详细地描述了生物本体是基因名称表富集分析实体的原始资料。生物网络用动态的方式表示数据，它用从实验数据获得的基因和蛋白质作为结点，用它们之间的联系作为边。与静态的途径和流程不同的是，网络是动态的并且在数据库中是唯一的，提供了较高的分辨率。正因为如此，网络是一个提出新假设时可以选择的工具。关于网络分析方法的六个章节构成了本书中分析部分的绝大多数内容。Merck 公司科技工作者撰写的章节全面综述了网络生物学，包括其概念和用网络对基因名称表富集分析时所用的工具。在讲述 bioPixies 软件的章节中，研究人员总结了概率网络图在分析大量固有复杂组学数据库中的应用，并且介绍了一个由普林斯顿大学（Princeton University）开发的著名的贝叶斯网络包：bioPixies。GeneGo 公司提供了一个基于知识的网络生成工具包，作为 MetaCore 数据挖掘套件的一部分。网络分析中的下一步，即动态网络模型，由俄罗斯系统生物学研究所（the Institute of Systems Biology）和英国的爱丁堡大学（University of Edinburg）一个合作小组提供。来自德国慕尼黑 Biomax 公司的小组描述了一个在 BioXM 软件的知识管理环境中实现网络分析的语义法，这个语义法与众不同并且有着强大的功能。网络分析集还包括一个章节，由西雅图系统生物研究所（Institute of Systems Biology）的科技人员详细地介绍了非常流行且开源的网络可视化工具 Cytoscape 和与其快速增长的一套应用程序插件。函数分析部分还用一个章节介绍了一个由英国 Infosense 公司开发的非常实用的集成软件包：KDE，其目的是从众多函数分析应用中建立强大且灵活的分析工作流程。

在函数分析的应用部分，我们从用户的角度（无论是基础研究还是药物开发/管理领域）选择了一些函数分析的方法。来自于转化基因组研究院（the Translational Genomics Institute）的小组人员提出了一个先进的技术，这个技术把 siRNA/化合物的筛选和途径数据分析结合起来应用到癌症研究领域。费城的天普大学（Temple University）和意大利几个研究型大学组成的小组描述了大规模的 miRNA 研究中网络分析的应用。斯克利普斯基因组研究所（Scripps Genomics Research）和加州大学圣地亚哥分校（UCSD）的人类

遗传学家小组提供了在SNP遗传相关性研究中途径工具的用法。还有两个章节专门介绍了疾病研究中函数工具和数据库的大规模应用。囊性纤维化治疗基金会（Cystic Fibrosis Foundation Therapeutics）领导的小组开发了一个以疾病为中心的数据分析环境MetaMiner，这个数据分析环境结合了详细疾病知识库和包含基因名称表富集分析、网络和相互作用组工具的工具箱。在Strategic Medicine公司小组撰写的转化医学相关章节中加入了临床上对途径分析的理解。最后，美国国家毒理学研究中心（食品及药物管理研究所FDA）研究小组提供了一款软件套件：ArrayTrack，它被美国食品及药物管理研究所研究人员和新药物应用评估人员广泛使用。这个套件最初被用于芯片表达分析，后来它又结合了许多公共统计工具和商业途径工具而被应用于其他的组学分析。

函数数据分析（Functional data analysis）作为一门快速发展的学科，几乎每周都会出现新的网络算法和软件工具，其应用范围也随着每一个新的DNA、RNA和蛋白质实验进入市场而扩大。因此，我们不能，也没有打算在本书中收录所有的工具。相反，我们试着把焦点集中在已建立的方法和一些在市场上常见的软件包，让读者对这个新兴领域中的议题和应用有更广泛的了解。

（高友鹤　尹剑锐　译）

# Preface

Recent Google queries for "systems biology" and "pathway analysis" fetched over 14 and 2 million entries, respectively. These numbers speak volumes about the ubiquity and popularity of systems data analysis in modern bioscience. These days, any microarray expression or SNP-based article would feature a chapter on pathways, ontology enrichment, and/or biological networks. Although a very young field, the applications of systems biology now spread widely from basic research and pre-clinical drug discovery to translational research and personalized health care.

Systems biology "focuses on the systematic study of complex interactions in biological systems, thus using a new perspective (integration instead of reduction) to study them" (Wikipedia). From a practical view point, it means integration of accumulated biological knowledge in a computer-readable format and the creation of tools for applying this structured information to the analysis of biological and chemical experimental data. Starting in the 1970s, biochemistry was the first field from which information was put into databases such as BRENDA, EMP/MPW, and, later, KEGG. Over the years, regulation and signaling components were added to biochemistry in the form of protein interaction databases such as HPRD and BIND, and comprehensive ontologies of cellular processes and protein functions were developed, the best known of which is Gene Ontology (GO).

Functional analysis is inseparable from high-throughput, or "OMICs," experimental biology, which has been rapidly evolving since the late 1990s. The "genome-wide", noisy OMICs assays with thousands of data points were nearly illegible for wet lab researchers at the time, and statistical tools only helped to reduce data complexity but not to understand the underlying biology. Gradually, bioinformaticians and wet lab biologists started to use databases of pathways and processes available to them for mapping and prioritization of experimental data (enrichment analysis). Later, biological networks were added to analysis toolboxes, borrowing from years of research in graph theory and physics.

This book represents a compilation of methods of functional analysis and its applications, written by practitioners from academy, government research organizations, the pharmaceutical industry, and bioinformatics companies. The content is divided into three sections. The first section deals with elucidation of protein, compound, and gene interactions. The second section is devoted to analytical tools of functional analysis. Applications of functional analysis is covered in the third section.

Binary interactions are the elementary units of functional organization of proteins, genes, and RNA in the cell. In functional analysis, these interactions play the role of "building blocks" in our computer-readable approximation of cellular life – pathways and networks. The result of functional analysis is only as good as the quality of interactions that pathways and networks are assembled from. Therefore, the methods of elucidation of interactions have a foremost importance. Here we present several prominent wet lab and dry lab tools for extraction of interaction information. There are many ways to obtain protein interactions, both from experiments and from already published experimental literature. Wet lab tools include profiling of physical protein–protein interactions with protein arrays (Invitrogen) and one of the most accurate versions of detection of transcription factor – gene promoter interactions (A-Star, Singapore). Among the computer-based methods, different types of text mining and language processing play an increasingly important role. We present a chapter on it by Linguamatics, a fast growing company specializing in text mining interaction information from PubMed and other sources. And, of course, the interactions can be dug out of experimental articles by slow, often painful, but high quality manual annotation, which is a technology all its own. We have chapters on annotation technology from specialists in biology-centered protein–protein, protein–DNA/RNA interactions and metabolism (MetaLogic, Russia), and chemistry-centered protein–compound interactions (GVK, India).

Many (but by no means all) methods of functional analysis can be divided into two groups: gene list enrichment analysis (EA) and biological networks. The first group comprises several procedures for calculation of relative enrichment of the data set of interest into components of certain pathways, cellular processes, protein function, disease targets, biomarkers, etc. EA results in quantitative ranking of ontology entities within the data set(s) based on a certain scoring. We present here the "classical" GSEA method of EA (by BMS) and a detailed chapter on biological ontologies as a source of entities for EA by SRI International. Biological networks represent the data in a dynamic way, using genes and proteins from experimental data points as "nodes" and connections between them as "edges". Unlike static pathways and processes, networks are dynamic and are unique for the data set, providing a higher resolution analysis. Because of this, networks are a tool of choice for applications which require new hypothesis generation. Six chapters on methods of network analysis comprise the bulk of the analysis section for this book. The chapter from Rosetta/Merck Co. provides a comprehensive overview of network biology, concepts, and tools for gene lists analysis using networks. A chapter from bioPixies reviews probabilistic networks applied for analysis of large, inherently noisy OMICs data sets, and describes bioPixie, a well-known Bayesian network package developed at Princeton University. A knowledge-based network generation toolkit is presented by GeneGo, Inc. as part of its MetaCore data mining suite. The next step in network analysis, dynamic network modeling, is presented by a collaborative team from the Institute of Systems Biology (Russia) and University of Edinburg (UK). A team from Biomax, Gmbh (Munich, Germany) describe a very different yet powerful semantic approach to network analysis realized in the BioXM Knowledge Management Environment. The network analysis set is concluded with a detailed chapter on the very popular open source "community" network visualization tool Cytoscape and its fast growing set of application plug-ins (Institute of Systems Biology, Seattle). The analysis section also includes a chapter on a very useful integration package, KDE, by Infosense, Ltd. (UK), which is designed for building robust and flexible analytical workflows from multiple functional analysis applications.

In the Applications section, we selected methods dealing with functional analysis from the user perspective, both in basic research and drug discovery/regulation. A group from the Translational Genomics Institute (Tgen, Phoenix) presents an advanced technology on a combined siRNA/compound screening, coupled with pathway data analysis which was used in cancer research. Application of network analysis in large-scale miRNA studies is described by a group from Temple University (Philadelphia) and several research universities in Italy. A team of human geneticists from Scripps Genomics Research and UCSD present usage of pathway tools in SNP genetic associations studies. Two chapters are devoted to larger scale applications of functional tools and databases for disease research. The group led by Cystic Fibrosis Foundation Therapeutics presents MetaMiner (Cystic Fibrosis), a disease-centered data analysis environment which combines a detailed disease knowledge base with a toolbox of EA, networks, and interactome tools. Clinical perspective on pathway analysis is added in the chapter on translational research submitted by Strategic Medicine, Inc. Finally, a group from the National Center for Toxicological Research, an FDA research institute, presents its software suite ArrayTrack, which is widely applied by FDA researchers and new drugs applications reviewers. ArrayTrack, primarily designed for microarray expression analysis and expanded into other OMICs assays, combines a number of statistical tools with public and commercial pathway tools.

Functional data analysis is evolving quickly as a discipline. Novel network algorithms and software tools are published almost weekly, and the scope of applications expands with every new DNA, RNA, or protein assay hitting the market. Therefore, we could not and had no intention to pack as many tools as possible into this volume. Instead, we tried to focus on the established methods and software packages we see in the marketplace every day and provide readers with a broad understanding of issues and applications of this fascinating new field.

*Yuri Nikolsky*
*Julie Bryant*