DOI:10.12113/202009001

EMBOSS 和 EMBnet

罗 静 初
(北京大学 生命科学学院,北京大学生物信息中心,北京 100871)

摘 要:笔者撰写的"EMBOSS 软件包序列分析程序实例"一文,已经在《生物信息学》期刊 2021 年第 19 卷第 1 期发表。此文介绍欧洲分子生物学开放软件包(European Molecular Biology Open Software Suite, EMBOSS)。EMBOSS 是欧洲分子生物学网络组织(European Molecular Biology Network, EMBnet)于上世纪九十年代末启动的以欧洲国家为主的国际合作项目,是生物信息学领域中较早投入使用的大型开源软件包。本文基于笔者亲身经历,回顾 EMBOSS 项目的来龙去脉,讲述 EMBnet 三十多年来的发展历程,及其对生物信息开发、服务和教育培训等方面的贡献,从某个侧面为读者特别是年轻读者展示生物信息学发展早期的一段历史。

关键词:生物信息学;生物信息学软件;欧洲分子生物学开放软件包 EMBOSS;欧洲分子生物学网络组织 EMBnet
中图分类号:Q349+.53 **文献标志码**:A **文章编号**:1672-5565(2021)04-223-09

EMBOSS and EMBnet

LUO Jingchu
(*College of Life Sciences and Center for Bioinformatics*, *Peking University*, *Beijing* 100871, *China*)

Abstract: An article entitled "Application examples of EMBOSS sequence analysis program" has been published in the *Chinese Journal of Bioinformatics* (Volume19, Issue 1, 2021). It made a description of the European Molecular Biology Open Software Suite (EMBOSS). Initiated in the late 1990s by the European Molecular Biology Network (EMBnet), EMBOSS is an international collaborative project mainly among European countries. It is one of the critical open source bioinformatics software packages with popular tools used in the bioinformatics field. Based on the author's own experience, this article tries to look back the ins and outs of the EMBOSS project, to trace the route that the EMBnet was born and became an international organization, and to have an overview of the contribution that EMBnet made to the development, service, education, and training of bioinformatics. Hopefully, it may give readers, especially the younger generations, some hints about the birth and growth of bioinformatics in early days.

Keywords: Bioinformatics; Bioinformatics software; EMBOSS; EMBnet

1 生物信息学

二十世纪九十年代诞生的国际互联网(Internet),为信息时代的到来奠定了基础,也催生了生物信息学这一交叉学科。DNA 双螺旋结构模型的提出、遗传密码的破译,以及蛋白质三维空间结构的测定,开创了从分子水平上探索生命规律的新时代。DNA 测序技术的成熟和产业化,为人类基因组计划的实施提供了技术储备。

就在生命科学研究高速发展的同时,计算机科学技术也取得了突飞猛进的发展。计算机在生命科学领域中的应用,可追溯到二十世纪八十年代。《核酸研究》(Nucleic Acids Research, NAR)半月刊于 1982、1984 和 1986 年第 1 期出版专辑,刊登分子生物学数据库以及核酸和蛋白质序列分析文章。1985 年,《计算机在生物学中应用》(Computer Application to Biosciences, CABIOS)创刊,标志着计

收稿日期:2020-09-04;**修回日期**:2020-10-14.
作者简介:罗静初,男,教授,研究方向:实用生物信息技术.E-mail:luojc@ pku.edu.cn.

算机在生命科学领域中的应用日趋成熟。而生物信息学作为一个学科,则诞生于二十世纪九十年代。1988 年,美国国家生物技术信息中心(National Center for Biotechnology Information, NCBI)创建时,英文"生物信息学"(Bioinformatics)这一术语尚未广泛使用。生物医学文献摘要数据库 PubMed 检索结果表明,"Bioinformatics"在文献杂志中广泛使用,始于二十世纪九十年代初。1994 年,欧洲生物信息学研究所(European Institute of Bioinformatics, EBI)成立时,生物信息学作为一门新兴学科,已悄然而生。1998 年, CABIOS 正式改名为《生物信息学》(Bioinformatics),并成为当前生物信息学权威杂志之一。

实际上,生物信息学和计算生物学这两个学科很难严格区分,英语"Bioinformatics"和"Computational Biology"这两个术语也经常混用。PubMed 医学主题词(Medical Subject Headings, MeSH)数据库(https://www.nlm.nih.gov/mesh/)就把 Bioinformatics 与 Computational Biology 列在同一个条目中。国际计算生物学学会(International Society for Computational Biology, ISCB)官方网站(https://www.iscb.org/)则称 ISCB 为"计算生物学和生物信息学一级专业学会"(Leading Professional Society for Computational Biology and Bioinformatics)。十分有趣的是,ISCB 的官方期刊有两个,一个就是 Bioinformatics,另一个则是 Public Library of Sciences (PLOS) Computational Biology。

作为一门新兴学科,要给生物信息学下一个严格的定义,似乎并不容易。若以目前较为流行的组学数据为研究对象,则可以大体描述如下。以核酸、蛋白质等生物大分子数据为主要研究对象,以基因组、转录组、蛋白组、代谢组等组学数据和文献资料为主要研究基础,以计算机为主要研究工具,以计算机网络为主要研究环境,构建各种类型的数据库,开发新一代生物信息软件,对浩如烟海的原始数据进行存储、管理、注释、加工和提取生物信息,用于药物设计、疾病诊治、品种改良和环境治理等领域。同时,利用数理统计、模式识别、神经网络、遗传算法、支持向量机和隐马氏模型等各种理论和方法,结合分子生物学、遗传学和基因组学等生命科学各领域研究成果,对大量生物信息进行查询、搜索、比较、分析,从中获取基因和基因组复制、转录、翻译、修饰和调控等理性知识,探索生命起源、生物进化及细胞、器官、个体和群体的发生、发育、发展等生命科学中重大问题,搞清它们的基本规律和时空联系。

2 生物信息学软件

不言而喻,生物信息学软件在生物信息学领域中具有无可替代的特殊地位。首先,生物信息软件的研究开发本身就是生物信息学领域的重要组成部分;而所开发和集成的各种软件工具、应用程序和分析平台,为核酸和蛋白质序列和结构以及高通量组学数据的分析提供了必不可少的手段。

生物信息软件开发包括许多方面,核酸和蛋白质序列分析软件的开发起步较早。1977 年,英国剑桥医学研究委员会(Medical Research Council, MRC)分子生物学实验室(Laboratory of Molecular Biology, LMB) Rodger Staden 在 NAR 上发表题为"利用计算机处理序列数据"(Sequence data handling by computer)的文章,开创了生物信息软件开发和集成的先河[1]。文章介绍了他编写的核酸和蛋白质序列分析程序,包括序列存储、编辑、转换,限制性内切酶搜索、密码子使用频率统计、序列相似性比较等。程序采用人机交互运行模式,在内存仅为28 kb的 PDP11/45 小型计算机上可处理长达6 000 bp的 DNA 序列。以后 20 多年,他一直致力于生物信息软件编写,完成了分子生物学领域第一个免费软件包 Staden 的开发。序列装配是该软件包的主要模块之一,在早期基因组测序和装配中起了重要作用。2004 年 Staden 退休,该软件包由他年轻的同事 James Bonfield 继续维护(http://staden.sourceforge.net/)。

1997 年,笔者在英国伦敦癌症研究基金会(Imperial Cancer Research Fund, ICRF)从事合作研究期间,有幸参观了 MRC 实验室,拜访了 Staden 博士。2002 年,James Bonfield 博士应邀在北京大学举办的生物信息培训班上介绍了 Staden 软件包。值得一提的是,英国剑桥 MRC 实验室是国际知名分子生物学研究机构,弗雷德里克·桑格(Frederick Sanger)、弗朗西斯·克里克(Francis Crick)、詹姆斯·沃特森(James Watson)、马克斯·佩鲁茨(Max Peruts)、约翰·肯德鲁(John Kendrew)、悉尼·布伦纳(Sydney Brenner)和约翰·萨尔斯顿(John Sulston)等十多位著名诺贝尔奖获得者曾在 MRC 工作。约翰·萨尔斯顿是人类基因组计划英国团队的主要负责人,2002 年获诺贝尔生理奖,笔者有幸参加了在英国基因组园区举办的庆祝会(见图 1)。

微型计算机(简称微型机,Microcomputer)的普及,使计算机在分子生物学中的应用得到了长足的进步。基于微型机的序列分析软件于八十年代中期开始使用,后来发展成 DNAStar、PCGene、MacVector 等商业软件。1986 年,美国亚利桑那大学分子生物

学教授 David Mount 应邀为北京大学生物系(1993 年改名为生命科学学院)举办为期一个月的生物技术和基因工程讲习班,带来了他编写的程序 DNA Management (DM),作为 DNA 和蛋白质序列分析工具。DM 基于微型机编写,当时北京大学生物系仅有一台处理器为 Intel 8086 的微型机,配有512 K内存、10 M硬盘和两个软盘驱动器,外带 12 英寸单色显示器。程序 DM 成功安装在这台微型机上。Mount 教授还带来了一盒软盘,每张软盘容量为 360 kb,存放了核酸序列数据库 GenBank 和蛋白质序列数据库(Protein Information Resource, PIR)。

图 1　约翰·萨尔斯顿同事祝贺其荣获诺贝尔奖

Fig.1　Sir John Sulston at the party after the announcement of the 2002 Nobel Prize

程序 DM 采用交互式会话菜单,使用相当方便,除用文本方式输出分析结果外,还可绘制简单的图形,如环形质粒 DNA 限制性内切酶位点等。军事医学科学院基础医学研究所吴加金研究员全程参加了该讲习班。随后的几年,他领导的团队编写了基于微型机的序列分析软件“金钥匙”(Goldkey),填补了国内生物信息领域软件开发的空白[2]。Mount 教授编著的“Bioinformatics: Sequence and Genome Analysis”2001 年由美国冷泉港出版社(Cold Spring Harbor Laboratory Press)出版,2002 年由科学出版社购买版权并以影印本形式出版。复旦大学钟杨主译的中译本于 2003 年由高等教育出版社出版。2004 年,本书第 2 版出版;2006 年,同济大学曹志伟将第 1 章翻译成中文,由科学出版社出版了本书第 2 版的中文导读版。

九十年代以来,微型计算机很快普及到科研究机构、大专院校、乃至千家万户。基于微型机的生物信息软件不断涌现。除 DNAStar、MacVector 等一些商业软件外,不少由学术单位编写的软件可免费下载和使用,如序列编辑、显示和分析软件 BioEdit、多序列比对软件 ClustalW、系统树构建软件 MEGA 等。与此同时,基于 UNIX 的开源操作系统 Linux 日趋成熟,而基于 Linux 系统的软件逐步成为生物信息领域的主流软件,例如数据库搜索软件 BLAST 和 FASTA,基因组序列装配软件 PHRED/PHRAP、基因结构预测软件 GenScan 和 GeneID 等,其中最为著名的是欧洲分子生物学开放软件包(European Molecular Biology Open Software Suite, EMBOSS)。

3　EMBOSS 软件包

EMBOSS 软件包的诞生有一个鲜为人知的故事。二十世纪八十年代,美国 Wisconsin 大学遗传计算研究组(Genetics Computing Group)开发了分子生物学软件包 GCG[3]。该软件包起初基于多用户小型机系统 Vax/VMS 开发,在一台服务器上安装后,多个用户可同时使用,后来移植到 Unix 平台。GCG 整合了许多常用序列分析工具,功能相当齐全,美国和欧洲不少科研机构和高等院校均购买并安装了该商业软件,供本单位研究人员使用。经过多年开发和商业化运行,上世纪八十年代至九十年代中期,GCG 软件包成为欧美各国最为流行的基于 Unix 服务器的多用户序列分析软件。由于 GCG 软件包实际上是许多已发表算法的实现或现有程序的整合,发行初期,其源代码对外公开。欧洲生物信息学网络组织(European Molecular Biology Network, EMBnet)等许多学术机构和个人在此基础上进行了二次开发,增加了许多新程序,形成了名为 EGCG 的软件包。EGCG 最初的含义为欧洲(European)GCG,后来,因为参加开发的人员不再限于欧洲国家,该软件包的名字也就改为扩充的(Extended)GCG。

九十年代末,由于人员变更和商业模式的改变,GCG 软件包不再公开源代码,EGCG 开发不得不终止。为此,EGCG 的主要开发者 Peter Rice 和 Alan Bleasby 等决定另起炉灶,抛开 GCG 而自行开发另一套分子生物学软件包,即 EMBOSS 软件包。这一计划得到了 EMBnet 成员的大力支持和积极参与。基于前期 EGCG 软件包现有基础,EMBOSS 项目很快取得了实质性进展。1999 年 4 月,Peter Rice 在北京大学举办的讲习班上演示了 EMBOSS 的第一个程序 seqret。

之后不久,基于 Needleman-Wunsch 动态规划算法的全局比对程序 needle,基于 Smith-Waterman 算法的局部比对程序 water,以及点阵图可视化序列比对程序 dottup 和 dotmatcher 等程序也很快完成。基于 GenBank/EMBL 等核酸序列数据库、PIR 和 Swiss-Prot 等蛋白质序列数据库的格式转换和序列特征信

息提取等一系列程序为用户提供了极大方便,而字串统计、密码子分析、酶切位点鉴定、重复序列识别和 CpG 岛预测等核酸序列分析程序,以及组分统计、跨膜螺旋识别和二级结构预测等蛋白质序列分析程序,则是 EMBOSS 软件包最具特色的核酸和蛋白质序列分析程序。本世纪初,Peter Rice 领导的 EMBOSS 研发团队受聘于欧洲生物信息学研究所,完成了该软件包的主要开发和集成,编写了系统的帮助文档[4]。2009 年,Peter Rice 领导的 EMBOSS 团队得到英国生物技术和生物科学研究委员会(Biotechnology and Biological Science Research Council, BBSRC)资助,继续进行 EMBOSS 软件包的开发(见图 2)。

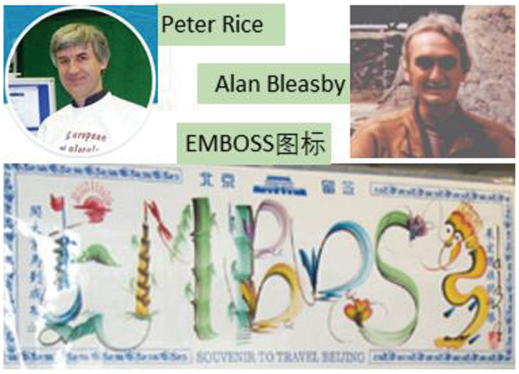

图 2 EMBOSS 软件包的主要开发者和 EMBOSS 彩绘图标

Fig.2 Major developers of EMBOSS(Peter Rice and Alan Bleasby) and the EMBOSS logo in color painting

除了 EMBOSS 开发团队自行编写的程序外,EMBOSS 还整合了不少其它常用生物信息软件包,如基于隐马尔可夫模型的蛋白质结构域序列谱构建和结构域识别软件包 HEMMER、系统发育分析软件包 Phylip 及 RNA 二级结构分析和预测软件包 VIENNA 等。2016 年发布的 EMBOSS 6.6.0 版包括 300 多个程序,十多个类别,是生物信息领域内容最为丰富、功能最为齐全的序列分析软件包,同时包括 JEMBOSS、PISE、wEMBOSS、mEMBOSS 等多个 Web 接口程序,均可免费下载安装(见图 3)。

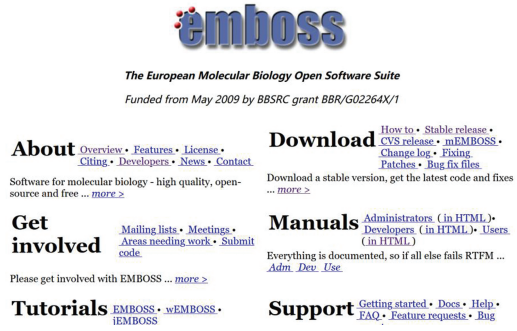

图 3 EMBOSS 软件包主页(http://emboss.open-bio.org/)

Fig.3 Main website of the EMBOSS software package

4 EMBnet

显而易见,EMBOSS 软件包的诞生,得益于欧洲分子生物学网络组织 EMBnet。EMBnet 成立于 1988 年(https://www.embnet.org/wp/about/history/),主要发起单位为德国欧洲分子生物学实验室(European Molecular Biology Laboratory, EMBL),英国 Daresbury 国家实验室,以及法国、荷兰和瑞典等几个西欧发达国家从事计算机在分子生物学中应用的学术机构和高等院校。EMBL 位于德国海德堡,是欧洲重要分子生物学实验室,由欧盟各国政府提供经费支持。1989 年五月,当时的 14 个欧盟成员国都加入了 EMBnet。

1991 年,EMBnet 获欧盟生物技术研究领域创新、开发和增长(Biotechnology Research for Innovation, Development and Growth in Europe, BRIDGE)框架计划资助,进入了快速发展时期。九十年代中期,EMBnet 成员单位达到 28 个,包括英国、德国、瑞士等西欧国家,波兰、斯洛伐克和匈牙利等东欧国家,以及以色列和土耳其等。在教育部和学校领导的大力支持下,北京大学蛋白质工程和植物基因工程(现更名为蛋白质和植物基因研究)重点实验室于 1996 年加入 EMBnet,同年加入的还有澳大利亚国家基因组信息服务中心(Australian National Genomic Information Service, ANGIS)和俄罗斯莫斯科州立大学。此后,南非、加拿大、印度、古巴等世界各大洲许多国家也纷纷加入 EMBnet。1998 年 EMBnet 成立十周年时,已经发展到 37 个成员单位(见图 4)。

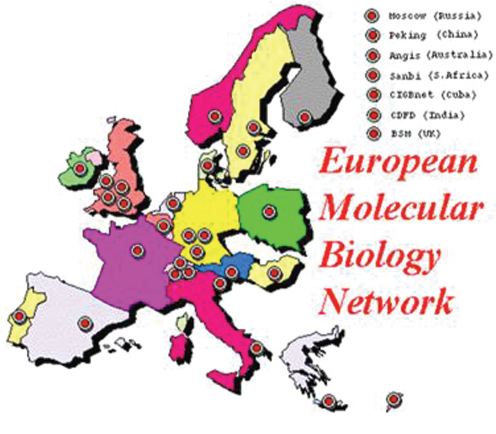

图 4 欧洲分子生物学网络组织节点分布(1997 年)

Fig.4 Distribution of EMBnet nodes (1997)

4.1 国家节点

EMBnet 成员单位称节点(Node),按成员单位的性质分为国家节点(National Node)和专业节点

(Specialist Node)两类。根据 EMBnet 章程,每个国家只能有一个国家节点,由政府部门推荐本国从事计算生物学的学术机构或高等院校,并向 EMBnet 提出申请,在 EMBnet 年会上由全体成员无记名投票,得票超过三分之二者通过,成为新的成员。每个成员单位指派一名代表,称节点负责人(Node Manager)。

由于人力资源、经费来源、硬件设施、网络环境和所在单位支持程度的差别,不同国家节点的情况各不相同。其中影响和贡献较大的有英国、荷兰、瑞士、瑞典和意大利等几个国家节点。英国的国家节点为 Daresbury 国家实验室的 SeqNet 项目组,负责人为 Alan Bleasby,也是 EGCG 和 EMBOSS 项目的主要开发者之一。荷兰的国家节点为内梅根大学(University of Nijmegen),负责人为 Jack Leunissen。由于 EMBnet 注册在荷兰,荷兰节点也承担财务管理等日常事务(见表 1)。

表 1　欧洲分子生物学网络组织国家节点(1998 年)

Table 1　National nodes of EMBnet(1998)

代码	国家	节点负责人	成员单位
AR	Argentina	Oscar Grau	Universidad Nacional de LaPlata Argentina
AU	Australia	Tim Littlejohn	University of Sydney
AT	Austria	Martin Grabner	Biocomputing Centre Vienna University
BE	Belgium	Robert Herzog	Free University of Brussels
CA	Canada	Christoph Sensen	National Research Council of Canada
CN	China	Jingchu Luo	Peking University
CU	Cuba	Ricardo Bringas	Centre for Genetic Engineering
DK	Denmark	Hans Ullitz-Moller	Danish Human Genome Centre
FI	Finland	Kimmo Mattila	Center for Scientific Computing
FR	France	Philippe Dessen	Infobiogen
DE	Germany	Martin Ebeling	German Cancer Research Centre
GR	Greece	Babis Savakis	Insitute of Molecular Biology
HU	Hungary	Endre Barta	Agricultural Biotechnology Centre
IN	India	MW Pandit	Centre for DNA Fingerprinting
IE	Ireland	Andrew Lloyd	Trinity College
IL	Israel	Leon Esterman	Weizmann Institute of Science
IT	Italy	Marcella Attimonelli	CNR-BARI
NL	Netherlands	Jack Leunissen	University of Nijmegen
NO	Norway	Karin Lagesen	Oslo University
PL	Poland	Piotr Zielenkiewicz	Institute of Biochemistry and Biophysics, PAS
PT	Portugal	Pedro Fernandes	Instituto Gulbenkian de Ciencia Rua da Quinta Grande
RU	Russia	Sergei Spirin	Moscow State University
ZA	South Africa	Wine Hide	University of the Western Cape
ES	Spain	Jose Valverde	Universidad Autonoma de Madrid
SW	Sweden	Nils-Einar Eriksson	Uppsala Biomedical Center
CH	Switzerland	Victor Jongeneel	ISREC Bioinformatics Group
TR	Turkey	Zehra Sayers	National Bioinformatics Node
UK	United Kingdom	Alan Baeasby	Daresbury Laboratory

此外,EMBnet 也不定期聘请一些兼职教员,为各成员单位举办各种类型的培训班、讲习班。如英国剑桥大学的 David Judge,英国爱丁堡大学的 Frank Wright 等。

4.2　专业节点

EMBnet 规定,每个国家除一个国家节点外,可以设立一个或多个专业节点(Specialist Node)。截止 1998 年,EMBNet 专业节点共有 9 个(见表 2)。

和国家节点一样,专业节点也指派一名节点负责人。专业节点中影响较大的有欧洲生物信息学研究所 EBI(https://www.ebi.ac.uk/)和桑格研究所(Sanger Institute, https://www.sanger.ac.uk/)。EBI 是欧洲分子生物学实验室 EMBL 设在英国的分部,建于剑桥南部小镇辛克斯顿基因组高新园区内,与桑格研究所毗邻。英国人类基因组图谱项目资源中心(Human Genome Mapping Project Resource Center)也设在该园区内。

表 2 欧洲分子生物学网络组织成员专业节点(1998 年)

Table 2 Specialist nodes of EMBnet(1998)

国家	名称	节点负责人	成员单位
欧盟	EMBL-EBI	Rodrigo Lopez	European Bioinformatics Institute
英国	Sanger Center	Peter Rice	The Sanger Centre, Wellcome Trust
英国	HGMP-RC	Martin Bishop	Human Genome Mapping Project Resource Centre
英国	UCL	Teresa Attwood	University College London
意大利	ICGEB	Sandor Pongor	International Center for Genetic Engineering and Biotechnology
荷兰	ETI	Peter Schalk	Biodiversity Center Universiteit van Amsterdam
德国	MIPS	Werner Mewes	Max Planck Institut fur Biochemie
瑞士	Hoffman-LaRoche	Daniel Doran	Pharma Preclinical Research Hoffman-LaRoche
瑞典	Pharmacia	Staffan Bergh	Pharmacia-Upjohn AB

EBI 基于 EMBL 原有几个计算生物学和生物信息学研究组扩充而成,是欧洲最大的生物信息学研究、开发和服务机构。1981 年,由 EMBL 创建的核酸序列数据库(EMBL Data Library,简称 EMBL)正式发布[5]。蛋白质结构分析、预测和设计是 EMBL 另一个重要学研究方向,知名学者 Chris Sander 任研究组负责人。上世纪八十年代至九十年代,系统分析了当时蛋白质结构数据库(Protein Data Bank, PDB)已经测定的结构,构建了一系列蛋白质结构相关数据库。

桑格研究所建于 1992 年,主要经费由英国生物医学慈善机构惠康信托基金会(Wellcome Trust, https://wellcome.ac.uk/)提供,是英国最大的基因组测序中心,承担了人类基因组计划 30% 测序任务。1999 圣诞节前夕,桑格研究所宣布完成 22 号染色体测序,这也是人类基因组计划最先完成测序装配的第一条染色体。桑格研究所和 EBI 同在辛克斯顿的基因组园区内,为数据共享和项目合作提供了极大便利,著名基因组数据库和分析系统 ENSEMBL(http://www.ensembl.org/)就是两个单位合作的结果。

4.3 运行机制

EMBnet 设有执行委员会(Executive Board),委员会由四名成员组成,除主席和秘书外,还设有一名成员负责财务。EMBnet 日常事务由执委会主席通过邮件组与节点负责人商定。执委会由全体成员在年会上无记名投票选举产生,执委会分工由四名成员商定。除执委会外,另设公共事务(Publicity and Public Relation)、教育培训(Education and Training)和技术管理(Technical Management)三个委员会,每个委员会也各由四位成员组成。无论是国家节点或专业节点负责人,均可竞选执委会和其它三个委员会委员。

EMBnet 每年举行一次年会,年会时间地点由节点负责人提出申请,经全体成员讨论通过后确定,通常是在各成员单位所在国家和城市轮流举行。不论是国家节点还是专业节点,每个成员单位每年缴纳 1 000 欧元会费,主要用于举办年会的交通食宿等费用。通常,举办年会的同时,也举行生物信息学学术讨论会或专题培训班。

4.4 软件开发

EMBnet 的宗旨,是把分散在各国的计算生物学领域人力物力组织在一起,为本国和其它国家提供数据库和软件等生物信息资源服务。值得一提的是,EGCG 及其后续的 EMBOSS 项目,就是 EMBNet 各成员单位之间合作的典范。EMBnet 的另一个合作项目,就是基于文本的数据库信息检索系统 SRS[6]。SRS 的英文原文是 Sequence Retrieval System,直译为序列提取系统,主要开发者为 EMBL 的 Thure Etzold。SRS 免费提供所有学术单位,最初用于检索 EMBL 和 GenBank 等核酸序列数据库、PIR 和 Swiss-Prot 等蛋白质序列数据库等以文本形式保存的序列和注释信息。通过对数据库条目中的关键词建立索引,以浏览器界面提供简单检索和高级检索功能。用户可通过蛋白名、基因名、物种名等基本

信息,以及序列条目中的大量注释信息,快速高效地对某个或几个数据库进行检索。SRS 后来扩充到 PubMed 文献摘要数据库、Pfam, PRINTS 和 Prosite 等蛋白质结构域和功能位点数据库,PDB, DSSP, HSSP 和 FSSP 等蛋白质结构和分类数据库。有的 SRS 服务器上安装的数据库多达几十个乃至上百个。EBI 成立后,Etzold 带领的 SRS 开发团队从德国海德堡搬到英国辛克斯顿,继续进行开发。九十年代末,许多 EMBnet 节点都安装了 SRS 系统[7]。鉴于许多生物技术和药物开发公司对 SRS 系统的需求,上世纪末,SRS 开发团队成立了软件开发公司,SRS 也成了商业软件,最新版的 SRS 8.0 不再免费提供学术单位,包括 EBI 在内的大部分 SRS 服务器不得不终止服务。

4.5　教育培训

EMBnet 的另外一个重要任务,是举办各种类型的讲习班、培训班,除了培训数据库和软件开发等生物信息领域专业人员外,也为从事分子生物学实验的最终用户进行培训。例如上面提到的开源软件 Staden 和商业软件 GCG,以及数据库检索系统 SRS 等。1997-1999 年,笔者以学生身份,先后参加了 Alan Bleasby 在英国 Daresbury 国家实验室举办的网络资源讲习班、David Judge 在剑桥大学举办的序列分析培训班、James Bonfiled 在剑桥 MRC-Lab 举办的 Staden 软件包安装使用培训班,以及 Thure Etzold 在 EBI 举办的 SRS 系统管理员培训班,为笔者日后举办多次生物信息培训班和开设“实用生物信息技术”课程积累了经验[8]。

4.6　网络刊物

自 1994 年起,EMBnet 不定期出版 EMBnet 新闻(EMBnet. news)网络刊物,并于 2010 年更名为 EMBnet 杂志(EMBnet. journal, https://journal.embnet.org/),报道各节点硬件和网络建设、数据库和软件开发、教育和培训等进展,介绍网络生物信息资源和生物信息软件使用经验。该网络刊物提供的生物信息数据库和软件使用快速指南(https://www.embnet.org/wp/quick-guides/),至今依然是生物信息初学者的简明手册(见表 3)。

表 3　EMBnet 快速指南

Table 3　EMBnet Quick Guides

内容	作者	年份
UNIX	McLysaght A & Lloyd A (IR)	2003
MySQL	Naseem A & Rahman N (PK)	2010
Perl Programing	Falquet L & Ioannidis A (CH)	2005
Perl Regular Expression	Falquet L & Ioannidis A (CH)	2006
Phylip	Valverde J (ES)	1998
BLAST	Forminghieri E (BR)	2004
EMBOSS	Mullan L (UK)	2004
wEMBOSS	Bottu G & Herzog R (BE)	2004
Phrap	Araujo M (BR)	2004
Phred	Araujo M (BR)	2004
Swiss-Prot and TrEMBL	Boeckmann B (CH)	2006
Swiss Model	Bordoli L (CH)	2004
MOLPAC	Stewart J, D'Elia D & Attwood T (IT, UK)	2014
IMGT/3Dstructure-DB and IMGT/2Dstructure-DB	Marie-Paule Lefranc (FR)	2004
IMGT-Collier-de-Perles	Marie-Paule Lefranc (FR)	2004
Amino Acid Conformation	Valverde J (ES)	2013

2012 年,EMBNet 加入了国际生物信息学教育和培训组织(Global Organization for Bioinformatics Learning, Education and Training, GOBLET)[9]。作为该组织的主要成员之一,EMBnet 为国际生物信息学的教育培训发挥了重要作用(https://www.mygoblet.org/)。本世纪初,英国伦敦大学专业节点负责人 Terresa Attwood 受聘英国曼彻斯特大学,继续进行生物信息数据库和软件开发。她为 GOBLET 编写了生物信息学详尽指南(https://www.embnet.org/wp/critical-guides/),详细介绍蛋白质序列数据库 UniProt 和 neXtProt、蛋白质结构数据库 PDB,以及生物信息领域中最常用的序列相似性数据库搜索系统 BLAST(见表 4)。

表 4 EMBnet 详尽指南

Table 4 EMBnet Critical Guides

内容	作者	年份
UNIX	Teresa Attwood	2018
BLAST	Teresa Attwood	2018
PDB	Teresa Attwood	2018
UniProt	Teresa Attwood	2018
UniProt Flat File	Teresa Attwood	2018
InterPro	Teresa Attwood	2018
neXtProt	MoniqueZahn & Teresa Attwood	2019
Genetic Disease	Marie-Blatter, Patricia Palagi & Teresa Attwood	2019
Power of Computers in Biology	Daniel Barker, Heleen Plaisier, Stevie Anne Bain & Teresa Attwood	2019

2000 年 2 月,由 EMBnet 主办的期刊生物信息学简报(Briefings in Bioinformatics, BiB)正式出版,主编为资深生物信息学家 Martin Bishop,包括笔者在内的许多 EMBnet 节点负责人应聘为编委会成员(https://academic.oup.com/bib)。2017 年,中国科学院北京基因组研究所章张应邀担任亚洲地区副主编。该杂志主要刊登生物信息软件和数据库等综述,介绍生物信息领域最新进展,是生物信息学领域极具影响力的重要杂志。

5 EMBnet 中国节点

EMBnet 成员单位的首要任务是为本国分子生物领域提供基本的生物信息资源服务。1996 年 10 月,北京大学蛋白质工程和植物基因工程重点实验室加入 EMBnet 后,该重点实验室主任兼国家 863 生物领域蛋白质工程专家组组长顾孝诚教授积极推动与北大计算中心和北大图书馆学术文献中心合作,建立了北京大学生物信息中心(Center for Bioinformatics, CBI),开始通过计算机网络,为国内用户提供数据库资源和软件工具等服务。1997 年 9 月,EMBnet 年会在意大利巴利(Bari)召开,作为 EMBnet 中国节点负责人,笔者第一次参加这次年会,结识了英国节点负责人 Alan Bleasby、荷兰节点负责人 Jack Leunissen、SRS 主要开发者 Thure Etzold 等多名欧洲生物信息学领域早期研究开发人员。

获 EMBnet 资助,1998 年 4 月在北大举办首次生物信息讲习班,来自全国各地的近百名学员参加了培训,Alan Bleasby、Jack Leunissen 和 ThureEtzold 等五位 EMBnet 节点负责人和兼职教师应邀担任培训班教师。1998 年 10 月,EMBnet 十周年纪念会在英国辛克斯顿 EBI 召开,应会议主持人 Peter Rice 邀请,笔者在会上播放了讲习班实况录像,给与会者留下了深刻印象。获国际遗传工程和生物技术中心(ICGEB)资助,1999 年 4 月在北大举办第 2 期生物信息讲习班班,10 位 EMBnet 节点负责人和兼职教员为培训班学员做报告或讲课。年过六旬的著名理论物理学家郝柏林院士以学生身份全程参加了该讲习班,并于讲习班后不久撰写了“建议尽快组建国家级的生物医学信息中心”的院士建议,和夫人张淑誉老师一起编写了国内第一本生物信息学书籍《生物信息学手册》。

2000 年 9 月,获国家自然科学基金委员会资助,以国家 863 生物领域首席专家强伯勤院士为团长的中国生物信息学代表团参加了在瑞士洛桑举行的 EMBnet 年会,访问了瑞士生物信息研究所(Swiss Institute of Bioinformatics, SIB)、Swiss-Prot 数据库、罗氏公司和苏黎世联邦理工学院。瑞士在生物信息学领域中有着特殊地位,蛋白质序列数据库 Swiss-Prot 于 1986 年诞生于瑞士日内瓦。SIB 构建的蛋白质分析专家系统(Expert of Protein Analysis System, ExPASy)则是重要生物信息资源网站,收集了几百个生物信息数据库和软件工具网址。鉴于网络带宽限制,欧洲以外其它国家访问该网站受到一定影响。为此,ExPASy 在拉美、澳大利亚等地设有镜像。这次访问的直接结果,就是与瑞士生物信息研究所商定,ExPASy 亚洲镜像就设在北大生物信息中心,为国内用户提供了极大方便。

二十多年来,作为 EMBnet 国家节点,北大生物信息中心得到国家教育和科研计算机网(CERNET)的大力支持,在生物信息资源建设、人才培训和基础教学,以及数据库和软件开发等方面做了一些工作[10]。

值得高兴的是,中国科学院北京基因组研究所大数据中心(BIGD, https://bigd.big.ac.cn/)于 2016 年成立,并于 2019 年成为国家基因组科学数据中心,而基因组所于同年加挂“国家生物信息中心”牌子。EMBnet 国家节点的任务,正在由该所年轻的生物信息学团队承担[11]。

6 结 语

作为一门交叉学科,生物信息学的诞生还不到三十年。最近十多年来,随着新一代基因组测序技术的诞生,高通量组学数据快速积累,ENSEMBL 等

各种类型的基因组数据库纷纷上网,Bowtie 和 BWA 等各种组学数据分析软件也不断涌现。作为以传统的单个基因或基因家族、单个蛋白或蛋白家族为主要分析对象的 EMBOSS 软件包,尽管它在组学数据分析中无能为力,但由组学数据分析得到的靶标基因或蛋白的深入分析依然离不开该软件包中的大量工具。目前,EMBOSS 软件包开发项目已经结束,该项目主要负责人 Peter Rice 受聘于 AXIOMEDIX 公司,担任客户部主任(https://axiomedix.com/about/team/)。作为 EMBOSS 软件包的主要开发者,Peter Rice 仍然负责维护该软件包。作为开源软件,EMBOSS 的维护开发需要生物信息领域中的同行共同努力。

本世纪初,鉴于 EBI 和 NCBI 等国际生物信息中心提供的生物信息资源越来越多,部分欧洲国家不再在人力物力上继续支持 EMBnet 国家节点,德国、瑞士、英国、比利时等国家节点先后退出 EMBnet。最近几年,根据欧洲和世界各国的实际情况,EMBnet 组织模式作了调整,有条件的成员单位可继续以国家节点方式保留会员资格,同时也吸收生物信息学研究团体和个人为为团体或个人会员。希望国内生物信息领域有志者积极参与,为国际国内生物信息学特别是生物信息资源服务和教育培训做出应有的贡献。

致　谢

感谢鲍一明、朱伟民、章张等人对本文的修改意见。2021 年 10 月,中国科学院北京基因组研究所(国家生物信息中心)成为 EMBnet 中国节点,鲍一明博士担任节点负责人。

参考文献(References)

[1]STADEN R. Sequence data handling by computer[J]. Nucleic Acids Research,1977 ,4(11):4037-4051. DOI:10.1093/nar/4.11.4037.

[2]吴加金,李伍举,雷红星,等. 核酸和蛋白质序列分析的软件系统——GOLDKEY[J]. 生物技术通讯,1994,5(4):189-193.
WU Jiajin, LI Wuju, LEI Hongxing, et al. GOLDKEY: The software package for the analysis of nucleic acid and protein sequences[J]. Letters in Biotechnology,1994,5(4):189-193.

[3]DEVEREUX J, HAEBERLI P, SMITHIES O. A comprehensive set of sequence analysis programs for the VAX[J]. Nucleic Acids Research, 1984,12(1 Pt 1):387-395. DOI:10.1093/nar/12.1part1.387.

[4]RICE P, LONGDEN I, BLEASBY A. EMBOSS: The European molecular biology open software suite[J]. Trends in Genetics, 2000, 16(6): 276-277. DOI: 10.1016/s0168-9525(00)02024-2.

[5]HAMM G H, CAMERON G N. The EMBL data library[J]. Nucleic Acids Research, 1986,14(1):5-9.

[6]ETZOLD T, ULYANOV A, ARGOS P. SRS: Information retrieval system for molecular biology data banks[J]. Methods in Enzymology,1996,266:114-128. DOI:10.1016/s0076-6879(96)66010-8.

[7]胡德华,张洁,方平. 生物信息学数据库调查分析及其利用研究[J]. 生物信息学,2005,3(1):22-25. DOI: 10.3969/j.issn.1672-5565.2005.01.006.
HU Dehua, ZHANG Jie, FANG Ping. A survey on bioinformatics databases and the investigation of their application [J]. Chinese Journal of Bioinformatics, 2005,3(1):22-25. DOI:10.3969/j.issn.1672-5565.2005.01.006.

[8]罗静初. 实用生物信息技术课程教学实例[J]. 生物技术通报,2015,31(11):102-111. DOI: 10.13560/j.cnki.biotech.bull.1985.2015.07.001.
LUO Jingchu. Teaching examples of applied bioinformatics course[J]. Biotechnology Bulletin, 2015, 31(11): 102-111. DOI: 10.13560/j.cnki.biotech.bull.1985.2015.07.001.

[9]ATTWOOD T K, BONGCAM-RUDLOFF E, BRAZAS M E, et al. GOBLET: The global organisation for bioinformatics learning, education and training[J]. PLoS Computer Biology,2015, 11(4): e1004143. DOI: 10.1371/journal.pcbi.1004143.

[10]LUO J. Bioinformatics service, education and research: The EMBnet and CBI. European Molecular Biology Network Centre of Bioinformatics[J]. Silico Biology, 2002,2(3): 173-177. DOI:10.1590/S0365-05962012000200007.

[11]National Genomics Data Center Members and Partners. Database resources of the National Genomics Data Center in 2020[J]. Nucleic Acids Research, 2020,48(D1):D24-D33. DOI:10.1093/nar/gkz913.