BioSun 3.0：一个综合性辅助分子生物学实验设计的软件系统

查磊 应晓敏 曹源 李伍举*
军事医学科学院基础医学研究所计算生物学中心，北京 100850

摘 要 本文介绍了一个综合性辅助分子生物学实验设计的软件系统BioSun。BioSun为一套辅助分子生物学实验设计软件，功能全面，并为生物学研究人员提供了易用的环境。使用BioSun分析数据和设计分子生物学实验可以加快实验进程，提高研究效率。

关键词： BioSun；生物信息学；计算机辅助设计

BioSun 3.0 : An Integrated Software System for Aiding Molecular Biology Experiment Design

Cha Lei, Ying Xiaomin, Cao Yuan, Li Wuju
Center of Computational Biology, Beijing Institute of Basic Medical Sciences, Academy of Military Medical Sciences, Beijing 100850, China

Abstract: In this paper, a comprehensive software system for aiding design of molecular biology experiments, named BioSun, is introduced. BioSun, as a software system for molecular biology experiments developed by our Lab, offers an easy-to-use environment and Integrated functions for biologists. By using BioSun to analyze data and design molecular biology experiments, we can not only accelerate the process of the related experiments, but also improve their success rate.

Keywords: BioSun ; Bioinformatics; Computer-aided design

BioSun软件是军事医学科学院基础医学研究所计算生物学中心自行开发的一套功能较为全面、使用方便和界面友好的生物信息学软件系统，并已获得软件著作权证书。先后推出了1.0和2.0两个版本[1][2]，目前最新版本为3.0。该软件系统分为序列编辑与管理、序列比对、蛋白质相关分析、DNA/RNA相关分析和微阵列分析五个部分，不仅包括序列数据库管理、蛋白质基本性质分析、双重及多重序列比对、构建系统进化树、酶切位点分析、质粒绘图、蛋白质功能位点分析和开放阅读框搜索等常见功能，还包括了自行开发并已得到实验验证的一些特色功能，例如RNA二级结构预测[3]、辅助寡核苷酸微阵列探针设计[4]、pBV220及pPIC9载体中外源基因高效表达设计[5][6]等功能，并在实验中得到了广泛应用。

1. 序列编辑与管理

1）序列编辑

为了方便用户处理序列，BioSun软件提供了两个可视化序列工作区，支持复制、粘贴、撤销等常用功能。另外，还提供序列字符大小写转换、DNA-RNA相互转换、片段查找、片段插入、片段提取、序列反向、序列互补和序列字符过滤等功能，有助于用户可视化操作序列。

2）序列格式支持

在序列格式方面，BioSun软件支持单个及批量导入目前国际上流行的FASTA、GeneBank和EBI等格式数据，同时兼容早期推出的Goldkey软件格式数据。此外，通过BioSun序列数据库可以实现各种序列格式之间转换，从而为其它软件提供数据支持。

3）数据库管理

数据是生物信息学的基础，良好的数据库管理将为数据分析提供方便，为此，BioSun软件提供了完善的序列数据库管理功能，可以识别目前国际上流行的FASTA、GeneBank和EBI等格式数据，同时支持数据库的导入、合并、查询以及数据变换等操作。不仅如此，BioSun还提供了一些经过手工收集整理的数据库以方便使用。例如：用于蛋白质功能位点分析的Prosite数据库，

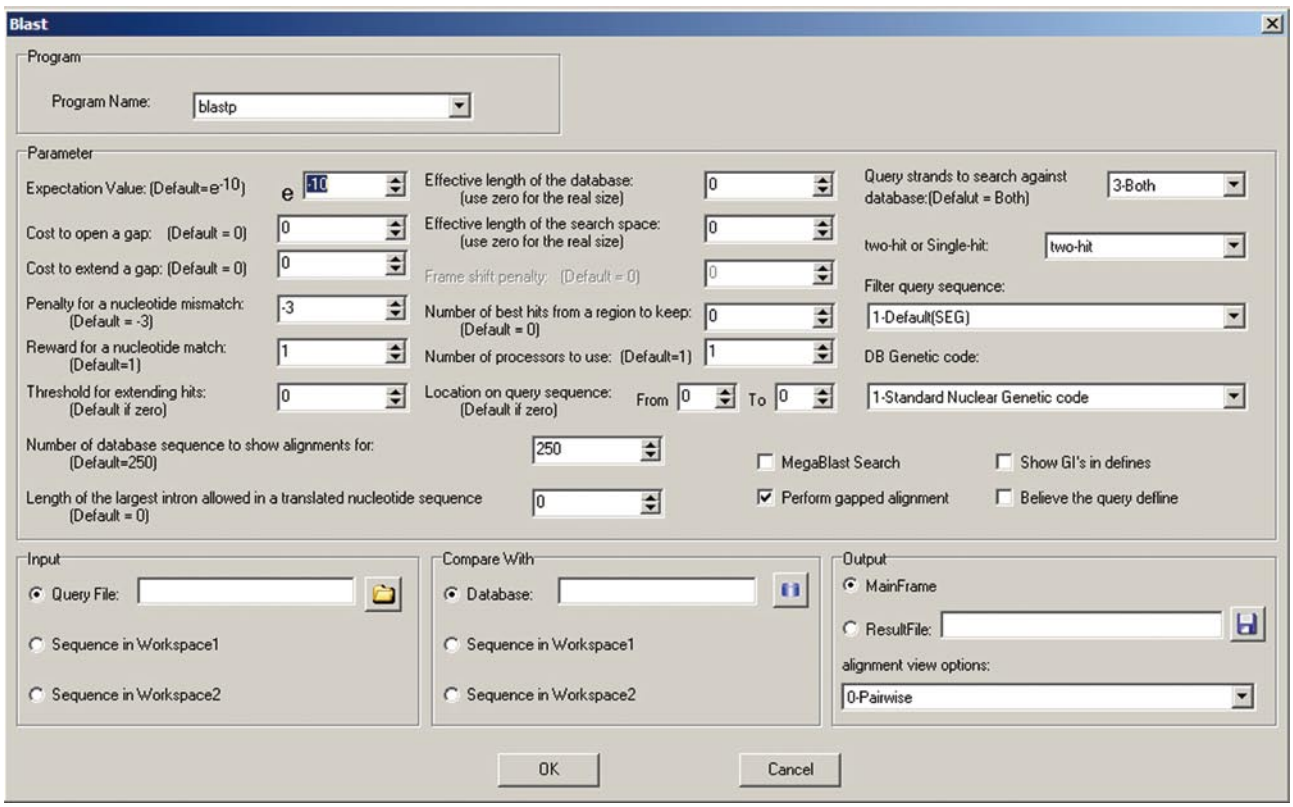

图1 Blast模块界面

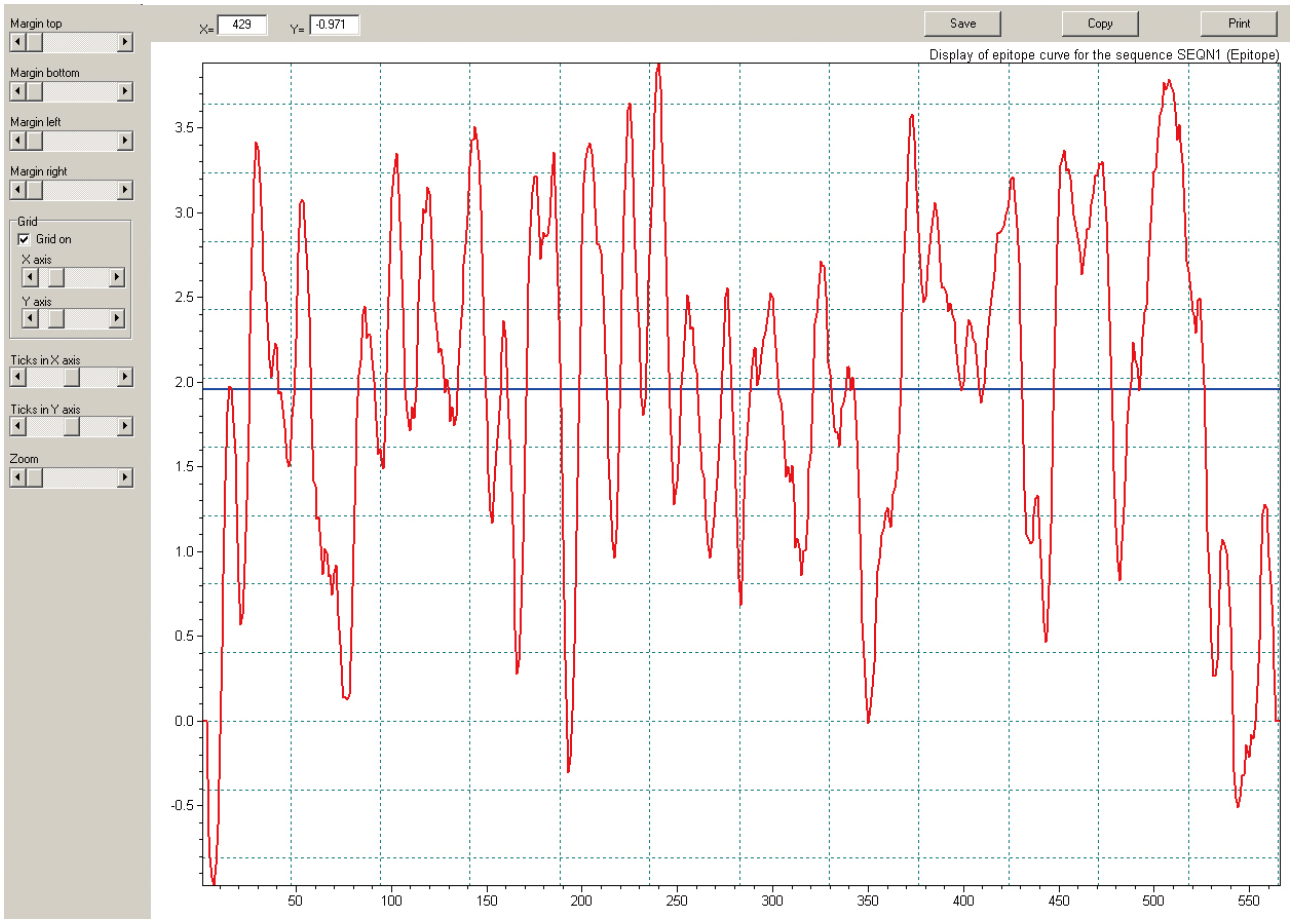

图2 B细胞抗原表位预测

用于酶切位点分析的rebase和subenzyme数据库等。

2. 序列比对模块

1）常用序列比对

序列比对是一项基本的生物信息学技术，在RNA二级结构预测、蛋白质三维结构预测、基于序列的进化树构建、PCR引物特异性分析和序列功能研究等方面均得到广泛应用。为此，BioSun软件提供了多种形式的的序列比对，包括识别工作区序列的重复片段、计算两个工作区序列相似性、第一工作区对数据库比对和数据库对数据库比对等功能。

2）Blast比对

Blast是生物信息学中的一个常用软件，其基本运行方式是命令行形式，涉及参数很多，使用复杂。为此，BioSun软件对Blast进行了封装，提供了一个简洁易用的界面和参数选择方式，并且可以选择单序列对单序列、单序列对多序列和多序列对多序列等多种比对方式，界面如图1所示。

3）ClustalW/X和进化树构建

多序列比对是两个序列比对的延展，也是进化树构建的基础，常用软件为ClustalW/X。另外，TreeView则是用于进化树展示的常用软件。为了便于用户构建进化树，BioSun软件实现了ClustalW/X和TreeView的无缝连接，在一个统一的界面里可以完成从多序列比对到进化树构建的整个过程。

3. 蛋白质相关分析

1）基本性质分析

蛋白质基本性质分析主要是对工作区序列或数据库内的蛋白质序列进行分析，主要包括长度、分子量、等电点、酸性氨基酸比例[DEY]、碱性氨基酸比例[KHR]、中性氨基酸比例[NQCTS]、非极性氨基酸比例[WGAFVPLMI]、半胱氨酸比例[C]、芳香簇氨基酸比

例[FYW]和脂肪簇氨基酸比例[ACDEGHIKLMNPQRSTV]。

2）蛋白质相关预测

在蛋白质分析模块中，提供了信号肽预测、跨膜区预测、B细胞抗原表位预测和T细胞抗原表位预测等多种功能，并且在B细胞抗原表位预测中，提供了Epitope、Hopp-Woods、Accesbility、Antigenicity和HPLC等多种曲线类型，图2所示的是根据我们提出的基于多参数的抗原表位预测方法对甲型H1N1流感病毒的HA段做出的预测曲线。上述分析对理解蛋白质功能具有重要意义，例如，基于跨膜区分析可以预测一个蛋白是否为跨膜蛋白，从而考虑是否有可能作为药物靶标。

3）蛋白质二级结构预测

蛋白质二级结构预测是指根据蛋白质序列的一级结构预测序列中各个氨基酸残基所处的二级结构状态，主要有α-螺旋、β-折叠、β-转角和无规卷曲4种状态。在蛋白质结构与功能关系分析方面具有参考价值。目前已有多种预测方法，BioSun软件采用Garnier方法对蛋白质二级结构进行预测。

4）蛋白质功能位点分析

BioSun提供的功能位点分析主要包括两个方面，首先是以Prosite数据库为基础，检索工作区内蛋白质序列含有的可能功能位点。Prosite数据库含有1139个模式，涉及翻译后的各种修饰位点、蛋白质功能域和多种类型的蛋白质家族信号等20大类，并且大部分是从文献上手工收集获得。其次是检索指定数据库中序列是否含有指定的蛋白质功能位点，这一步主要是分析用户所生成的模式的特异性。

5）蛋白质三维结构的展示

PDB数据库是存储蛋白质三维结构的数据库，BioSun可以对PDB文件进行三维展示，并提供了多种显示方式，可以进行旋转、缩放、移动和保存等操作，便于用户定制图形大小和从不同角度观察蛋白质的结构。

4. DNA与RNA分析

1）基本分析

包括碱基数目统计或对应的比例、计算可能的K个碱基组成（K-mer）、DNA模式分析、DNA序列翻译为蛋白质序列、蛋白质序列反向翻译为RNA序列等常用功能，其中K-mer组成分析在生物信息学研究中具有广泛应用。

2）转录因子结合位点定位

转录因子结合位点分析在基因表达调控研究中具有重要应用，BioSun软件提供了序列匹配、位置矩阵、位置矩阵加核心区三种定位方式来查找转录因子结合位点，为寻找转录因子的靶基因提供了便利。

3）酶切位点分析及显示

酶切位点分析在基因工程研究中是一项基本分析，可以用于基因重组、基因鉴定和核酸多态性研究等方面。BioSun软件提供了比较全面的酶切位点分析功能，包括构建限制性酶切位点数据库和多种方式显示酶切图谱等操作。

4）PCR引物设计及报告

PCR是一种在体外模拟DNA复制过程的扩增技术，可用于病毒鉴定与分类等诸多方面，其关键在于一对人工合成的寡核苷酸引物，BioSun提供了寡核苷酸引物的设计功能，可以通过设置PCR引物长度、GC含量、PCR扩增模板长度、解链温度、检索范围等多种参数，设计出合适的引物，并可以对引物进行质量评估。

5）RNA二级结构预测

RNA许多功能的实现需要借助一定的二级结构，在很多实验中，RNA二级结构有着重要意义，但是鉴于RNA分子具有降解快和晶体难于获得等特点，用实验方法测定其结构比较难，而通过计算机算法来预测是一个行之有效的方法。BioSun不仅提供了常用的基于最小自由能的RNA二级结构预测算法，同时也提供了我们提出的基于螺旋区堆积或螺旋区分布的二级结构预测算法。

6）密码子相关分析

密码子偏性是指生物体中编码同一种氨基酸的同义密码子的使用存在偏好现象，由于这一现象与遗传信息的载体DNA和生物功能分子蛋白质相关联，所以具有重要的生物学意义。BioSun软件提供了E.coli和Yeast系统中的密码子偏性指标计算，并可以对稀有密码子进行显示，这些分析在外源基因高效表达实验中具有重

要应用。

7）外源基因高效表达设计

外源基因的表达水平受多种因素影响，如启动子强度、载体性质、密码子偏性和mRNA稳定性等。基于我们提出的E.Coli系统中pBV220载体和Yeast系统中pPIC9载体的外源基因高效表达数学模型，BioSun软件提供了外源基因高效表达评价和自动化设计模块，目前该算法已得到多例实验验证[7-9]。同时，应用该模块也可以对其它类似载体的设计提供参考，该模块界面如图3所示。

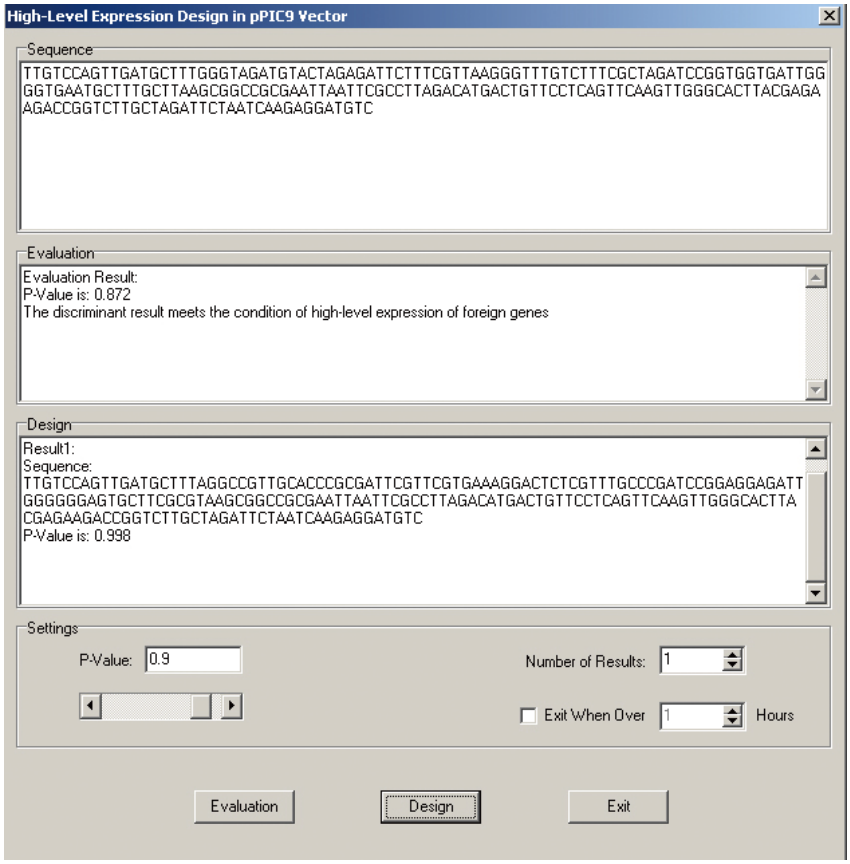

图3 pPIC9外源基因高效表达设计界面

5. 微阵列分析

目前，对于微阵列，BioSun提供了2种分析，分别是辅助寡核苷酸微阵列探针设计功能和基于基因表达谱的不同样本状态下的差异基因识别。并且在探针设计功能中，BioSun还整合了序列对数据库比较功能，使设计出的探针可直接用于合成。此外，探针设计功能还可用于设计某一种属的通用或特异探针。

6. 和其它同类软件的比较及应用情况

目前，不少公司与研究机构都推出了自己的生物信息学分析软件，通过与DNA SIS MAX 2.05、DNAStar 5.0、VectorNTI9.1、BioEdit7.0等常用生物信息学软件进行比较[2]，BioSun3.0在序列编辑与管理、核酸相关分析和蛋白质相关分析等3个生物信息学主要部分中，不仅提供了其它软件绝大部分的功能，还根据军事医学科学院基础医学研究所计算生物学中心长期的工作积累，提供了包含诸如外源基因高效表达设计、寡核苷酸芯片探针设计和蛋白质抗原表位预测等功能，这些功能都得到了多例实验验证和支持。

BioSun3.0作为自主开发的一套生物信息学软件系统，涵盖了生物信息学所需的各项常用功能，基本满足了院校和科研人员的常规需求。自BioSun软件推出以来，目前已推广到北京大学生命科学学院、军事医学科学院、中国中医研究院、北京农林科学院、汕头医学院、宁夏医学院、江南大学和湖南景达生物制药有限公司等生物医学研究单位，具体用户列表见我们的网站。BioSun软件的应用主要涉及抗原表位预测[10]、寡核苷酸芯片探针设计[11]、外源基因高效表达设计[12]和ncRNA[13]等相关研究。与国外的生物信息学软件相比，在实现同类软件常见功能的基础上，增加了贴近实验而又有自己特色的功能。今后，我们将根据研究工作的进展和实验中的实际需求，本着贴近实验和方便易用的原则，不断丰富BioSun软件的功能。有关BioSun软件的更新消息将及时发布在我们的网站上，网址为：http://www.biosun.org.cn/biosun/intro.htm或http://ccb.bmi.ac.cn/biosun/intro.htm。

参考文献:

[1]李伍举,应晓敏.BioSun:计算机辅助分子生物学实验设计的软件系统.军事医学科学院院刊,2004,10(5):401-404.

[2]查磊,应晓敏,曹源,等.BioSun2.0：一个综合性的辅助分子生物学实验设计软件.军事医学科学院院刊,2006,30(5):461-464.

[3]李伍举,吴加金.基于螺旋区随机堆积的RNA二级结构预测.生物物理学报,1996,12:213-218.

[4]Li WJ,Huang J,Fan M,et al.MProbe:Computer-aided probe design for oligonucleotide microarray experiment. Appl.Bioinform,2002,1(3): 163 -166.

[5]李伍举,吴加金.pBV220载体中外源基因表达水平定量分析.病毒学报,1997,2:126-133.

[6]Bingli Wu, Lei Cha, Zepeng Du,et al.Construction of mathematical model for high-level expression of foreign genes in pPIC9 vector and its verification.Biochemical and Biophysical Research Communications,2007,354:498-504.

[7]裴武红,沈倍奋,李伍举,等.计算机辅助设计使重组Ricin-A链在E.coil中的高效表达.细胞与分子免疫学杂志,1998,14(1):33-36.

[8]裴武红,胡美茹,李伍举,等人.FKBP12基因克隆及其表达产物的生物活性.中国生物化学与分子生物学报,2000,16(3):322-325.

[9]刘淑红,孙长凯,张玉梅,等人.NR1靶片段的原核表达.军事医学科学院院刊,2002,26 (3):188-190.

[10]刘燕华,贾杨,王景林.E型肉毒神经毒素（BoNT）基因序列分析及其B细胞表位预测.军事医学科学院院刊,2006,30(5):419-423.

[11]相磊，陈小玲，李伍举.口蹄疫病毒分型诊断探针设计初报.动物医学进展，2008，29(1)：1-5.

[12]柳伟强,牛建章,李晓霞,等.计算机辅助设计改善重组人神经珠蛋白(hNGB)在E.coli中的表达.河北大学学报（自然科学版）,2005,25(6)：639-643.

[13]李华,应晓敏,查磊,等.mRNA的比较研究.生物物理学报,2006,22(2):110-116.

收稿时间：2009年5月7日

作者信息

查磊

军事医学科学院基础医学研究所，研究实习员，主要研究方向为计算生物学。

李伍举

军事医学科学院基础医学研究所，研究员，主要研究方向为计算生物学。