科技前沿

生物信息学在昆虫学研究中的应用*

张 赞[1] 刘金定[1,2] 黄水清[2] 李 飞[1**]

(1. 南京农业大学植物保护学院 南京 210095; 2. 南京农业大学信息科学技术学院 南京 210095)

摘 要 随着深度测序和基因芯片技术的不断发展,基因组、转录组、表达谱数据大量积累。目前,至少有 10 多个昆虫的基因组已被测序,30 多个昆虫的转录组数据被报道。显然,传统的生物统计学方法无法处理如此海量的生物数据。量变引发质变,生物数据的大量积累催生了一门新兴学科,生物信息学。生物信息学融合了统计学、信息科学和生物学等各学科的理论和研究内容,在医学、基础生物学、农业科学以及昆虫学等方面获得了广泛的应用。生物信息学的目标是存储数据、管理数据和数据挖掘。因此,建立维护生物学数据库、设计开发基于模式识别、机器学习、数据挖掘等方法的生物软件,以及运用上述工具进行深度的数据挖掘,是生物信息学的重要研究内容。本文首先简要介绍了生物信息学的历史、研究现状及其在昆虫学科中的应用,然后综述了昆虫基因组学和转录组学的研究进展,最后对生物信息学在昆虫学研究中的应用前景进行了展望。

关键词 生物信息学,昆虫学,基因组学,转录组学

Application of bioinformatics in entomology

ZHANG Zan[1] LIU Jin-Ding[1,2] HUANG Shui-Qing[2] LI Fei[1**]

(1. Department of Entomology, Nanjing Agricultural University, Nanjing 210095, China;
2. College of Information Science and Technology, Nanjing Agricultural University, Nanjing 210095, China)

Abstract The rapid development of deep sequencing and microarray technologies over the past few decades has produced a huge amount of new biological data on insects. Genomes of hundreds of species have been partially sequenced and transcriptome information has been obtained from thousands of species in recent years. The genomes of at least tens of insect species have been fully sequenced and complete transcriptome data for more than 30 species has been obtained. The deep analysis of these data requires the assistance of bioinformatics, a new cross discipline that involves statistics, computer science, information science and molecular biology. The major goals of bioinformatics are storing, managing and mining biological data. The construction and maintenance of biological databases, development of biological software and algorithms using pattern recognition, data mining, machine learning and visualization, and the implementation of tools for data mining, are the main tasks of bioinformatics. In this review, we introduce the history and progress of bioinformatics and its application to entomology. Progress in research on insect genomics, transcriptome and small RNA research are also briefly introduced.

Key words bioinformatics, entomology, genomics, transcriptome

1 生物信息学学科的诞生

生物信息学是在分子生物学和信息科学共同发展基础之上产生的一门交叉学科。在分子生物学领域, Watson 和 Crick 于 1953 年发现了 DNA 双螺旋结构,开辟了现代分子生物学的新纪元,而遗传中心法则的提出推动了分子生物学的大发展。随后,限制性内切酶发现、重组 DNA 克隆技术的实现是新一代深度测序技术的基础,是海量生物数据产生的重要推动力。然而,对海量生物

* 资助项目:中央高校科研基本科研业务费(KYJ200908,1806J0063)。

**通讯作者,E-mail:lifei@njau.edu.cn

收稿日期:2011-11-19,接受日期:2011-12-20

数据的使用却面临着存储、管理、分析上的困难，因此在迫切需求的驱动下催生了生物信息学。在信息科学领域，随机字符串、理论计算和语法定义等计算理论的提出，以及计算机技术的迅猛发展也是生物信息学这一交叉学科产生的重要基础。

1961 年，计算科学首次被应用于基因和蛋白质的进化分析中（Ingram，1961），后逐渐被广泛应用于分子生物学领域。1977 年世界上第 1 个物种——噬菌体 Φ－X174 的基因组序列被测定（Gibson *et al.*，2010），生命科学开启了基因组时代。1980 年开发了新的测序仪，从技术上基本实现了开展大规模测序工作的可能性。同时，由于癌症研究工作的停滞不前，科学家提出了人类基因组测序计划的设想。1990 年人类基因组计划正式启动，随后发现拼接、管理和分析如此海量的基因组数据是一个非常复杂的问题。为了处理人类基因组计划产生的大量序列数据，计算机科学和数学开始被引入基因组序列的拼接和后续处理中。因此，可以认为人类基因组计划直接催生了生物信息学。

此后，更多的物种基因组测序计划开始进行，不断积累了大量的生物数据。据 NCBI 收录的基因组数据统计显示，目前已经完成基因组序列拼接的真核物种已经达到了 400 多个，正在进行基因组测定的真核物种达到 700 多个。截止到 2010 年，GenBank 数据库中储存的数据量已经达到 286 多亿碱基对，相对于 2009 年增长了 12.6%，其中转录组测序数据相对于 2009 年增长了 9 倍（Benson *et al.*，2011）。成指数级增长的生物数据量和缓慢的信息挖掘速度之间存在巨大差距，这种差距进一步推动了生物信息学的飞速发展。

2 生物信息学的研究概况

生物信息学包含了生物数据的获取、处理、存储、分发、分析和挖掘等方面研究内容，通过运用数学、计算机科学的各种工具，来阐明和理解大量数据所包含的生物学意义。生物信息的发展可以分为两个阶段：基因组时代和后基因组时代。在基因组时代，生物信息学的主要研究内容包括序列拼接和对比、序列的分子进化分析、蛋白质空间结构的预测、基因的预测和非编码 DNA 功能研究等。在后基因组时代，表达谱分析、转录组分析、代谢网络分析以及药物靶点筛选等成为生物信息学的重要研究方向。

2.1 序列对比和进化研究

序列对比是生物信息学的基础。根据算法应用对象的差异性，其可分为两个序列之间的相似性分析和多个（3 条或 3 条以上）序列相似性分析。2 个序列之间的对比软件主要采用动态规划算法，例如：Blast（Altschul *et al.*，1990）、Fasta（Pearson and Lipman 1988）、BLAT（Kent，2002）等，目前这类算法已经比较成熟。多个序列的相似性分析的算法比较多，各有利弊（Demkin，2009）。比较常用的有 CLUSTAL（Chenna *et al.*，2003）、SAGA（Notredame and Higgins，1996）、Needlman-Wunsch（Needleman and Wunsch，1970）等，此外还有 Smith-Waterman（Smith and Waterman，1981）、Malign（Schneider and Mastronarde，1996）、T-coffee（Notredame *et al.*，2000）和 IterAlign（Brocchieri and Karlin，1998）等。由这些算法开发出的软件比较常用的有 GeneDoc、ClustalW、ClustalX 等。

2.2 蛋白质空间结构的预测

蛋白质结构预测是在生物信息学诞生早期的一个热门领域。经过 40 多年的应用和发展，逐渐形成了蛋白质一级序列的理化性质分析、二级结构预测、三级空间结构预测等完整的体系。蛋白质二级结构分析包含 α 螺旋、β 折叠、β 转角、无规卷曲、motif 等蛋白质局部结构组件的预测。常用的软件有 InterProScan（Quevillon *et al.*，2005）、PredictProtein（Rost *et al.*，2004）、nnPredict（Kneller *et al.*，1990）和 SOPMA（Geourjon and Deleage，1995）等。蛋白质三级空间结构的预测主要分成 2 种策略：同源建模和从头预测。同源建模算法一般首先进行模板匹配，然后根据已知模板的空间架构搭建蛋白质的骨架结构，最后利用其它信息将蛋白质的侧链及环区补充完整。从头预测的主要原理是将蛋白质所有空间构象都预测出来，根据自然状态下蛋白质的构想维持本身自由能最低的条件来选择最终的预测结果。目前开发的软件主要有 SWISS-MODEl（Arnold *et al.*，2006）、CPHmodels（Nielsen *et al.*，2010）、EsyPred3D（Lambert *et al.*，2002）、MODELLER（Fiser and Sali，2003）、THREADER（Jones *et al.*，1992）和 3D-PSSM（Kelley *et al.*，2000）等。虽然

该领域研究较早,但是由于蛋白质空间结构复杂,又和外界环境紧密相关,研究难度较大。从头预测的错误率比较高,计算量比较大,现有的算法还远不能满足实际研究的需要。

2.3　基因组分析

基因组分析主要包括基因组序列的拼接、基因预测、基因功能的注释等。序列的拼接是生物信息学的基础工作之一。由于测序技术的限制,测序的原始结果都是大量的短序列(reads),只有经过序列拼接后才能对基因组的信息做进一步的分析。序列拼接应用较多的算法有 Hamilton path 和 Eulerian path 算法。以相应算法开发的软件有:CAP3(Huang and Madan, 1999)、PCAP(Huang *et al.*, 2003)、RePS(Wang *et al.*, 2002)、Phusion(Mullikin and Ning, 2003)、Atlas(Havlak *et al.*, 2004)、Celera Assembler(Myers *et al.*, 2000)和 EULER(Pevzner *et al.*, 2001)、Fragment Gluer(Pevzner *et al.*, 2004)等。尽管目前序列的拼接算法很多,但是由于 reads 长度的限制和基因组中的重复序列的干扰,基因组序列的拼接技术仍然不尽人意。

基因预测是在拼接出基因组序列后,预测与基因有关的信息,如基因在基因组中的位置、基因的结构信息、选择性剪接、单核苷酸多态性等。目前开发出的基因预测算法可分为三类:从头预测算法、同源分析算法和二者相结合的算法。从头算法主要是基于序列本身的特征信息的提取和学习。同源分析算法是基于未知序列和已知序列的相似性的算法。相关的软件有 GenScan(Burge and Karlin, 1997)、Fgenesh(Salamov and Solovyev, 2000)、GenomeScan(Yeh *et al.*, 2001)、EST2Genome(Mott, 1997)、TwinScan(Korf *et al.*, 2001)等。

基因功能注释主要是分析基因的结构和其所编码蛋白质的功能。一般是通过与已知蛋白编码基因的序列进行同源比对分析,对未知基因序列进行功能预测。这一算法的原理是基于比较基因组学,认为序列相似的基因在功能上往往具有相似性。经常使用的数据库主要两类:一是核酸数据库(GenBank、EMBL、DDBJ);二是蛋白质数据库(SWISS-PROT、PIR、MIPS、PDB、SCOP 和 CATH 等)(商英璠等,2005)。

2.4　转录组分析

转录组学通常指某一物种或者是某组织中所有 RNA 转录本的总和。不同组织细胞的转录组表达情况和同一组织不同时间转录组的表达情况均有很大差异,对转录组的分析可以很好地研究系统发育基因、特异性状相关基因(如抗药性)等。在医学上,转录组表达谱还可以用于疾病基因的诊断。目前,通常采用和基因组测序相似的技术获得转录组序列,并进行拼接用于进一步的分析。转录组序列拼接算法主要分为两类:有参考基因组序列的拼接方法和无参考基因组序列的拼接方法。参考基因组序列的拼接算法开发的软件主要有 Cufflinks(Trapnell *et al.*, 2010)和 Scripture(Guttman *et al.*, 2010),无参考基因组序列的拼接算法开发的软件主要有 ABySS(Birol *et al.*, 2009)、SOAP denovo(Li *et al.*, 2009)、Trinity(Grabherr *et al.*, 2011)等。此外,在转录组中还含有非编码 RNA,近年来发现这些非编码 RNA 包含很多种类,它们在基因转录后调控着蛋白基因的表达。

3　昆虫生物信息学的研究现状

随着昆虫学研究的不断深入、昆虫生物数据不断增长,生物信息学在昆虫学研究中的应用价值越来越明显。特别是在模式昆虫果蝇的基因组测序成功以后,大量医学昆虫、经济昆虫和农业昆虫的基因组和转录组均相继被测序,昆虫的基因数据正进入了飞速积累阶段。昆虫的种类繁多、进化关系复杂、个体发育系统多样,生物信息学在昆虫学研究中必将大有用武之地。

3.1　昆虫基因组学研究进展

根据 NCBI 数据库统计,迄今为止有超过 50 个昆虫的基因组序列已经完成拼接或正在绘制基因组草图。下面简略介绍已在 NCBI 发布基因组序列的物种。

3.1.1　果蝇科　目前果蝇属已经有 13 个物种的基因组序列获得基因组序列,2000 年黑腹果蝇 *Drosophila melanogaster* 的基因组序列通过全基因组鸟枪法(whole genome shotgun sequencing,WGS)和 Clone-based 的方法获得(Adams *et al.*, 2000),是最早获得基因组序列的昆虫。2005 年,果蝇科的第 2 个物种拟暗果蝇 *D. pseudoobscura* 的基因组

被测序(Richards *et al.*, 2005)。2007 年由 16 个国家的研究机构联合发起对 10 种果蝇的基因组进行测序。通过和已知的 2 种果蝇基因组进行比对,发现在果蝇中进化最快的是和嗅觉、味觉、解毒和代谢有关的基因,表明果蝇基因的进化与环境变化有很大的关系。同时发现 *D. willistoni* 是 12 种果蝇中唯一没有硒蛋白的昆虫(Bai *et al.*, 2007)。

3.1.2 蚊科 冈比亚按蚊 *Anopheles gambiae* 基因组的草图及分析结果于 2002 年 10 月发表(Holt *et al.*, 2002),采用全基因组鸟枪法进行测序,发现了大量的单核苷酸多态性位点。2007 年,埃及伊蚊 *Aedes aegypti* 的基因组测序完成,获得的基因组序列比冈比亚安蚊的基因组大 5 倍左右。在埃及伊蚊中发现了大量的转座子。与此同时,应用 MPSS(massively parallel signature sequencing)以及基因芯片技术预测编码蛋白质编码基因的结构,并测定了埃及伊蚊的表达谱(Nene *et al.*, 2007)。2010 年按蚊科的另一物种 *Anopheles darlingid* 的基因组被巴西科学家测序完成。目前拼接成 18 629 条 contigs 序列,约 173 Mbp(base-pair),同时在该基因组中发现 44 条 miRNA 前体(Mendes *et al.*, 2010)。

3.1.3 蜂 2003 年由美国国家人类基因组研究中心(NHGRI) 和农业部(USDA) 资助的蜜蜂 *Apis mellifera* 基因组测序计划完成。采用全基因组鸟枪法和 Clone-based 测序方法,供获得 135M 的基因组数据,此次测序的文库都来自雄峰的 DNA。2007 年,贝勒医学院人类基因组测序中心(Human Genome Sequencing Center at Baylor College of Medicine, HGSC)使用全基因组鸟枪法获得了 *Nasonia giraulti* 和 *Nasonia longicornis* 2 个和黄蜂近源的昆虫基因组序列。2010 年另一个物种 *Nasonia vitripennis* 的基因组序列也被测序成功。经过联合分析,在这 3 个密切相关的寄生蜂的基因组中发现了昆虫 DNA 甲基化基团、蜂类的特殊基因(多种毒液的基因)以及 Pox viruses 之间的横向基因转移(Werren *et al.*, 2010)。2011 年美国马里兰大学成功测序获得 *Megachile rotundata* 的基因组序列。目前 *Bombus terrestris*、*Bombus impatiens* 以及 *Apis florea* 的基因组序列也在测序和拼接之中。

3.1.4 蚕科 2003 年,西南农业大学和中科院北京基因组研究所公布了家蚕 *Bombyx mori* 的基因组的框架图,这是我国获得的第 1 个昆虫的全基因组序列。该框架图覆盖了家蚕全基因组的 95.54%,精度达到了 99.95%,采用全基因组鸟枪测序法,共获得 16 948 条完整家蚕的基因和 7 285 条基因片段,并在第 2 年完成了家蚕的全基因图谱的"精细图"(Xia *et al.*, 2004)。2004 年,日本也发布了家蚕的基因组序列(Mita *et al.*, 2004)。2009 年,西南农业大学和重庆大学等联合测得 39 种家蚕和 11 种野蚕的遗传变异图谱,发现家蚕和野蚕之间有明显的遗传分离,表明家蚕经过单一且短暂的驯养过程,且在此后很少与野蚕之间有基因交流(Xia *et al.*, 2009)。

3.1.5 蚁科 2010 年 8 月,印第安蚂蚁和弗罗里达弓背蚁的基因组由深圳华大基因和纽约大学医学院联合测序完成。通过基因组对比,发现虽然这 2 种社会性昆虫处于不同社会等级,但具有相似的基因图谱,且均通过表观遗传学调控基因的表达,从而实现不同个体的发育。此外,2 种蚂蚁基因组的测序为研究衰老和行为表观遗传学建立了新的模型(Bonasio *et al.*, 2010)。2011 年 1 月,阿根廷蚂蚁 *Linepithema humile* (Smith *et al.*, 2011)、红色收获蚁 *Pogonomyrmex barbatus* (Smith *et al.*, 2011)的基因组序列分别由蚂蚁基因组学联盟通过 WGS 测序方法获得。2011 年 2 月 10 日切叶蚁 *Atta cephalotes* 的基因组被华盛顿大学医学院测序中心测序完成(Suen *et al.*, 2011)。2011 年 4 月红火蚁 *Solenopsis invicta* 的基因组由洛桑大学测序完成(Wurm *et al.*, 2011)。

除了以上昆虫外,赤拟谷盗 *Tribolium castaneum* (Richards *et al.*, 2008)、人体虱 *Pediculus humanus corporis* (Kirkness *et al.*, 2010)和黑森瘿蚊等昆虫的基因组也已被测序完成或正在拼接全基因组草图。

3.2 昆虫转录组学研究进展

由于序列表达标签(EST)概念的提出,cDNA 序列于 20 世纪 90 年代开始大规模测序(Adams *et al.*, 1991)。目前 NCBI 的 UniGene 数据库已经收集了 140 个物种共 2 691 966 条序列,其中包含 14 个昆虫的转录组数据,这主要得益于新一代测序技术的进步,其中绝大部分数据都是近 2 年获得的。此外还有大量的转录组数据被收录到 SRA

(sequence read archive)数据库中。由于拼接质量有所不同,转录组数据可分为基于基因组拼接的转录组和从头拼接的转录组。

3.2.1　基于基因组拼接的转录组　利用 Solexa 1G sequencer 测序技术,黑腹果蝇(野生型的卵巢和睾丸、*bam* 突变型的卵巢和睾丸)4 种生殖腺的转录组被测得。获得的 reads 通过质量过滤后,利用软件 ELAND 将其比对到黑腹果蝇的基因组序列上,然后进行拼接。在野生型果蝇的转录组中发现了一系列调控性别特征形成的基因,这证明了由 RPKM 的方法识别转录组表达量的可行性,同时,发现果蝇中的基因选择性剪切多发生在未分化富含 *bam* 突变的睾丸细胞中(Gan *et al.*, 2010)。对埃及伊蚊胚胎发育 4 个时间段的转录组分析,发现了 2 个新的合子激动蛋白轻链基因,这 2 个基因在埃及伊蚊合子早期被表达。其中 AaKLC1 在库蚊和果蝇中存在同源基因,而 AaKLC2 仅在库蚊中有同源基因(Biedler and Tu, 2010)。利用 RNA-Seq 和 Tiling microarray 获得了 30 个不同发育阶段的果蝇转录组,最终发现了 111 195 个新的功能元件,包括蛋白编码基因、非编码基因的转录本、选择性剪接和 RNA 编辑等(Graveley *et al.*, 2011),这些发现大大推动了果蝇发育的研究。

3.2.2　从头拼接的转录组　通过 454 测序技术,O'Neil 等(2010)获得了 2 种蝴蝶 *Erynnis propertius* 和 *Papilio zelicaon* 的转录组,利用 CAP3 和 Celera Assembler 将 reads 拼接成平均长度大于 714 bp 的 contigs。为了验证转录组的覆盖度,拼接好的 contigs 被用来和家蚕的蛋白序列匹配,并提出"orthology hit ratio"来具体评估转录组序列在整个基因组的覆盖度。采用相同的测序技术,5 龄的胡椒蛾 *Biston betularia* 的转录组序列被公布,该转录组利用 Roche Newbler 软件拼接,利用家蚕的基因数据库进行注释,发现 5 条工业黑变病相关基因(Van't Hof and Saccheri, 2010)。2010 年 6 月蝴蝶 *Euphydryas editha* 的幼虫,蛹和成虫的混合转录组通过测序,共获得 864 056 条 reads,利用 Roche Newber 拼接成 14 244 个 contigs (Mikheyev *et al.*, 2010),发现了 10 个微卫星重复的多态性位点。浙江大学昆虫研究所利用 Illumina 测序技术获得了烟粉虱 *Bemisia tabaci* 卵、幼虫、蛹和成虫的混合转录组(Wang *et al.*, 2010)。利用 SOAP denovo 拼接获得 unique 序列,通过各种已知的数据库进行注释,最终获得 27 290 条注释序列。采用同样的技术路线,褐飞虱的转录组数据也在同一时期测序完成(Xue *et al.*, 2010)。共获得了 85 526条 unigenes 和 6 个 DGE 数据,为研究褐飞虱的发育、翅二态性等提供了数据基础。利用 Illumina 技术获得小菜蛾 *Plutella xylostella* 的序列,获得了 172 660 条 contigs,发现与 *Diadegma semiclausum* 寄生有关的基因 928 个,此外还发现寄主的抗菌肽在被寄生以后表达量上升(Etebari *et al.*, 2011)。

3.3　昆虫数据库

自 2000 年起,*Nucleic Acids Research* 杂志在每年的第一期都集中发布各种数据库的更新及最新创建的数据库。据不完全统计,2009 年到 2010 年间创建的生物学数据库总数达 1 000 多个。在这里简要介绍几个常用的昆虫学数据库。

3.3.1　Flybase　Flybase(http://flybase.org)是 1992 年美国美国国立卫生研究院国家人类基因组研究中心建立的,用于收集和发布果蝇整合的基因和基因组。该数据库主要收集了黑腹果蝇的数据。Flybase 主要数据类型包括:序列水平的基因模型,基因产物功能的分子分类,突变表型,突变损伤和染色体畸变,基因表达模式,转基因插入及解剖图形等数据。可以使用基因的名称、DNA 或蛋白质的序列、基因的功能、表型等进行搜索(Drysdale, 2008)。目前该网站也包含其他 12 种果蝇的基因组、基因、CDS 区、内含子等数据。

3.3.2　VectorBase　VectorBase (http://www.vectorbase.org/)是一个关于人类病原无脊椎载体的数据库。它是一系列人类病原载体基因组的数据库门户网站,由美国国家过敏和传染病研究所的生物信息资源中心建立。目前包含按蚊、库蚊等 8 种疾病载体的序列、图片、抗药性等数据库。该数据库主要提供 8 种人类传染病携带物种的基因组信息以及目前正在测序的 18 个相关物种的测序进展。提供 Blast、Clustalw、HMMER 和词汇搜索等分析工具,并接受相关数据的提交。

3.3.3　家蚕数据库　Silkdb (http://silkworm.genomics.org.cn/)是家蚕基因组的资源网站。该网站由西南大学蚕学与系统生物学研究所维护,提供包括家蚕基因的功能注释、基因产物、ESTs 以

及表达芯片数据等信息。同时该网站还提供 BLAST、SilkMap、Wego、BmArray、Clustalw、SMS 等数据分析工具。SIlkSatDb(http://www.cdfd.org.in/SILKSAT/index.php)是一个收集家蚕微卫星重复序列的数据库,由印度新德里政府生物技术部门维护,提供家蚕的 WGS 和 EST 上的微卫星序列的快速搜索以及引物的开发和验证。此外与蚕相关的数据库还有 KaikoBase、SilkBase 和 WildSilkBase 等。

3.3.4 BeeBase BeeBase(http://hymenopteragenome.org/beebase/)是以蜜蜂 *Apis mellifera* 及其 2 个病原体基因组为主的数据库。该数据库由 USDA ARS 维护,目前的数据有 *A. mellifera*、*Bombus terrestris*、*B. impatiens* 的基因组及其相关的注释信息。该数据库也提供了 BLAST、PSI-BLAST、Apollo Annotation Tools 以及 Genome Browsers 等查询和分析工具。

3.3.5 ButterflyBase ButterflyBase(http://butterflybase.ice.mpg.de/)是关于鳞翅目昆虫的数据库。该数据库由德国耶拿马克斯-普朗克化学生态学研究所维护。目前该数据库引入了 NCBI 的 dbEST 数据库中所有的关于鳞翅目的 cDNA 序列数据,并且在不断的增加关于家蚕基因组的信息。该数据库提供注释搜索、引物设计、微卫星搜索和序列比对等分析工具和服务,并接受相关数据的提交。

此外还有 SPODObase、LocustDB、LepTree、FlyNets 和 AnoBase 等昆虫数据库,在此不一一介绍。

4 生物信息学的前景和展望

随着测序技术的飞速发展,预计未来 10 年左右将形成昆虫基因组测序的高峰期。但昆虫种类繁多、进化关系复杂、基因组杂合度大,给全基因组测序工作和分析工作造成很大的困难,生物信息学在此领域必将有很大的用武之地。然而,尽管昆虫学的基因数据获得了大量的积累,并且正成指数级增加,但相对于自然界丰富的物种来说,目前已获得的遗传信息仍然是非常渺小的一部分,还有很大发展空间。从昆虫的生物数据产生、储存、管理和分析挖掘角度看,生物信息学在昆虫领域的应用仍然相对不足,因此,生物信息学在昆虫学领域的应用仍处于起步阶段,有很大的发展空间。张春霆院士在 2000 年撰文指出,科学数据的大量积累将导致重大的科学规律的发现。昆虫基因数据的巨大积累也将对害虫控制和资源昆虫利用等具有重要的推动作用,有可能催生出全新的害虫控制技术,开辟资源昆虫利用的新领域。

参考文献(References)

Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF, George RA, Lewis SE, Richards S, Ashburner M, Henderson SN, Sutton GG, Wortman JR, Yell MD, Zhang Q, Chen LX, Bron RC, Rogers YH, Blazej RG, Champe M, Pfeiffer BD, Wan KH, Doyle C, Baxter EG, Helt G, Nelson CR, Gabor GL, Abril JF, Agbayani A, An HJ, Andrews-Pfannkoch C, Baldwin D, Ballew RM, Basu A, Baxendale J, Bayraktaroglu L, Beasley EM, Beeson KY, Benos PV, Berman BP, Bhari D, Bolshakov S, Borkova D, Botchan MR, Bouck J, Brokstein P, Brottier P, Burtis KC, Busam DA, Butler H, Cadieu E, Center A, Chra I, Cherry JM, Cawley S, Dahlke C, Davenport LB, Davies P, de Pablos B, Delcher A, Deng Z, Mays AD, Dew I, Dietz SM, Dodson K, Doup LE, Downes M, Dugan-Rocha S, Dunkov BC, Dunn P, Durbin KJ, Evangelista CC, Ferraz C, Ferriera S, Fleischmann W, Fosler C, Gabrielian AE, Garg NS, Gelbart WM, Glasser K, Glodek A, Gong F, Gorrell JH, Gu Z, Guan P, Harris M, Harris NL, Harvey D, Heiman TJ, Hernez JR, Houck J, Hostin D, Houston KA, Howl TJ, Wei MH, Ibegwam C, Jalali M, Kalush F, Karpen GH, Ke Z, Kennison JA, Ketchum KA, Kimmel BE, Kodira CD, Kraft C, Kravitz S, Kulp D, Lai Z, Lasko P, Lei Y, Levitsky AA, Li J, Li Z, Liang Y, Lin X, Liu X, Mattei B, McIntosh TC, McLeod MP, McPherson D, Merkulov G, Milshina NV, Mobarry C, Morris J, Moshrefi A, Mount SM, Moy M, Murphy B, Murphy L, Muzny DM, Nelson DL, Nelson DR, Nelson KA, Nixon K, Nusskern DR, Pacleb JM, Palazzolo M, Pittman GS, Pan S, Pollard J, Puri V, Reese MG, Reinert K, Remington K, Saunders RD, Scheeler F, Shen H, Shue BC, Siden-Kiamos I, Simpson M, Skupski MP, Smith T, Spier E, Spradling AC, Stapleton M, Strong R, Sun E, Svirskas R, Tector C, Turner R, Venter E, Wang AH, Wang X, Wang ZY, Wassarman DA, Weinstock GM, Weissenbach J, Williams SM, WoodageT, Worley KC, Wu D, Yang S, Yao QA, Ye J, Yeh RF, Zaveri JS, Zhan M, Zhang G, Zhao Q, Zheng L, Zheng XH, Zhong FN, Zhong W, Zhou X, Zhu S, Zhu X, Smith HO, Gibbs RA, Myers EW, Rubin GM, Venter

JC, 2000. The genome sequence of *Drosophila melanogaster*. *Science*, 287(5461):2185—2195.

Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merril CR, Wu A, Olde B, Moreno RF, 1991. Complementary DNA sequencing: expressed sequence tags and human genome project. *Science*, 252(5013):1651—1656.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ, 1990. Basic local alignment search tool. *J. Mol. Biol.*, 215(3):403—410.

Arnold K, Bordoli L, Kopp J, T Schwede, 2006. The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. *Bioinformatics*, 22(2): 195—201.

Bai Y, Casola C, Feschotte C, Betran E, 2007. Comparative genomics reveals a constant rate of origination and convergent acquisition of functional retrogenes in Drosophila. *Genome Biol.*, 8(1):R11.

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW, 2011. GenBank. *Nucleic. Acids Res.*, 39(Database issue):D32—37.

Biedler JK, Tu Z, 2010. Evolutionary analysis of the kinesin light chain genes in the yellow fever mosquito Aedes aegypti:gene duplication as a source for novel early zygotic genes. *BMC Evol. Biol.*, 10:206.

Birol I, Jackman SD, Nielsen CB, Qian JQ, Varhol R, Stazyk G, Morin RD, Zhao Y, Hirst M, Schein JE, Horsman DE, Connors JM, Gascoyne RD, Marra MA, Jones SJ, 2009. De novo transcriptome assembly with ABySS. *Bioinformatics*, 25(21):2872—2877.

Bonasio R, Zhang G, Ye C, Mutti NS, Fang X, Qin N, Donahue G, Yang P, Li Q, Li C, Zhang P, Huang Z, Berger SL, Reinberg D, Wang J, Liebig J, 2010. Genomic comparison of the ants *Camponotus floridanus* and *Harpegnathos saltator*. *Science*, 329(5995):1068—1071.

Brocchieri L, Karlin S, 1998. A symmetric-iterated multiple alignment of protein sequences. *J. Mol. Biol.*, 276(1): 249—264.

Burge C, Karlin S, 1997. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.*, 268 (1):78—94.

Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, Higgins DG, Thompson JD, 2003. Multiple sequence alignment with the clustal series of programs. *Nucleic. Acids Res.*, 31(13):3497—3500.

Demkin VV, 2009. Bioinformatic analysis of nucleotide sequences records retrieved from GenBank. *Mol. Gen. Mikrobiol. Virusol.*, 2:36—39.

Drysdale R, 2008. FlyBase: a database for the *Drosophila* research community. *Methods Mol. Biol.*, 420:45—59.

Etebari K, Palfreyman RW, Schlipalius D, Nielsen LK, Glatz RV, Asgari S, 2011. Deep sequencing-based transcriptome analysis of *Plutella xylostella* larvae parasitized by *Diadegma semiclausum*. *BMC Genomics*, 12:446.

Fiser A, Sali A, 2003. Modeller:generation and refinement of homology-based protein structure models. *Methods Enzymol.*, 374:461—491.

Gan Q, Chepelev I, Wei G, Tarayrah L, Cui K, Zhao K, Chen X, 2010. Dynamic regulation of alternative splicing and chromatin structure in *Drosophila* gonads revealed by RNA-seq. *Cell Res.*, 20(7):763—783.

Geourjon C, Deleage G, 1995. SOPMA: significant improvements in protein secondary structure prediction by consensus prediction from multiple alignments. *Comput. Appl. Biosci.*, 11(6):681—684.

Gibson DG, Glass JI, Lartigue C, Noskov VN, Chuang RY, Algire MA, Benders GA, Montague MG, Ma L, Moodie MM, Merryman C, Vashee S, Krishnakumar R, Assad-Garcia N, Andrews-Pfannkoch C, Denisova EA, Young L, Qi ZQ, Segall-Shapiro TH, Calvey CH, Parmar PP, Hutchison CA, Smith HO, Venter JC, 2010. Creation of a bacterial cell controlled by a chemically synthesized genome. *Science*, 329(5987):52—56.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A, 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.*, 29(7):644—652.

Graveley BR, Brooks AN, Carlson JW, Duff MO, Lolin JM, Yang L, Artieri CG, van Baren MJ, Boley N, Booth BW, Brown JB, Cherbas L, Davis CA, Dobin A, Li R, Lin W, Malone JH, Mattiuzzo NR, Miller D, Sturgill D, Tuch BB, Zaleski C, Zhang D, Blanchette M, Dudoit S, Eads B, Green R E, Hammonds A, Jiang L, Kapranov P, Langton L, Perrimon N, Sler JE, Wan KH, Willingham A, Zhang Y, Zou Y, Andrews J, Bickel PJ, Brenner SE, Brent MR, Cherbas P, Gingeras TR, Hoskins RA, Kaufman TC, Oliver B, Celniker SE, 2011. The developmental transcriptome of *Drosophila melanogaster*. *Nature*, 471 (7339):473—479.

Guttman M, Garber M, Levin JZ, Onaghey JD, Obinson JR, Adiconis X, Fan L, Koziol MJ, Gnirke A, Nusbaum C,

Rinn JL, Ler ES, Regev A, 2010. Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat. Biotechnol.*, 28(5):503—510.

Havlak P, Chen R, Durbin KJ, Egan A, Ren Y, Song XZ, Weinstock GM, Gibbs RA, 2004. The Atlas genome assembly system. *Genome Res.*, 14(4):721—732.

Holt RA, Subramanian GM, Halpern A, Sutton GG, Charlab R, Nusskern DR, Wincker P, Clark AG, Ribeiro JM, Wides R, Salzberg SL, Loftus B, Yell M, Majoros WH, Rusch DB, Lai Z, Kraft CL, Abril JF, Anthouard V, Arensburger P, Atkinson PW, Baden H, de Berardinis V, Baldwin D, Benes V, Biedler J, Blass C, Bolanos R, Boscus D, Barnstead M, Cai S, Center A, Chaturverdi K, Christophides GK, Chrystal MA, Clamp M, Cravchik A, Curwen V, Dana A, Delcher A, Dew I, Evans CA, Flanigan M, Grundschober-Freimoser A, Friedli L, Gu Z, Guan P, Guigo R, Hillenmeyer ME, Hladun SL, Hogan JR, Hong YS, Hoover J, Jaillon O, Ke Z, Kodira C, Kokoza E, Koutsos A, Letunic I, Levitsky A, Liang Y, Lin JJ, Lobo NF, Lopez, JR Malek JA, McIntosh TC, Meister S, Miller J, Mobarry C, Mongin E, SD Murphy, O'Brochta DA, fannkochCP, Qi R, Regier MA, Remington K, Shao H, Sharakhova MV, Sitter CD, Shetty J, Smith TJ, Strong R, Sun J, Thomasova D, Ton LQ, Topalis P, Tu Z, Unger MF, Walenz B, Wang A, Wang J, Wang M, Wang X, Woodford KJ, Wortman JR, Wu M, Yao A, Zdobnov EM, Zhang H, Zhao Q, Zhao S, Zhu SC, Zhimulev I, Coluzzi M, della Torre A, Roth CW, Louis C, Kalush F, Mural RJ, Myers EW, Adams MD, Smith HO, Broder S, Gardner MJ, Fraser CM, Birney E, Bork P, Brey PT, Venter JC, Weissenbach J, Kafatos FC, Collins FH, Hoffman SL, 2002. The genome sequence of the malaria mosquito *Anopheles gambiae*. *Science*, 298(5591):129—149.

Huang X, Madan A, 1999. CAP3:A DNA sequence assembly program. *Genome Res.*, 9(9):868—877.

Huang X, Wang J, Aluru S, Yang SP, Hillier L, 2003. PCAP:a whole-genome assembly program. *Genome Res.*, 13(9):2164—2170.

Ingram VM, 1961. Gene evolution and the haemoglobins. *Nature*, 189:704—708.

Jones DT, Taylor WR, Thornton JM, 1992. A new approach to protein fold recognition. *Nature*, 358(6381):86—89.

Kelley LA, MacCallum RM, Sternberg MJ, 2000. Enhanced genome annotation using structural profiles in the program 3D-PSSM. *J. Mol. Biol.*, 299(2):499—520.

Kent WJ, 2002. BLAT-the BLAST-like alignment tool. *Genome Res.*, 12(4):656—664.

Kirkness EF, Haas BJ, Sun W, Braig HR, Perotti MA, Clark JM, Lee SH, Robertson HM, Kennedy RC, Elhaik E, Gerlach D, Kriventseva EV, Elsik CG, Graur D, Hill CA, Veenstra JA, Walenz B, Tubio JM, Ribeiro JM, Rozas J, Johnston JS, Reese JT, PopadicA, Tojo M, Raoult D, Reed DL, Tomoyasu Y, Kraus E, Mittapalli O, Margam VM, Li HM, Meyer JM, Johnson RM, omero-Severson JR, Vanzee JP, Alvarez-Ponce D, Vieira FG, Aguade M, Guirao-Rico S, Anzola JM, Yoon KS, Strycharz JP, Unger MF, Christley S, Lobo NF, Seufferheld MJ, Wang N, Dasch GA, Struchiner CJ, Madey G, Hannick LI, Bidwell S, Joardar V, Caler E, Shao R, Barker SC, Cameron S, Bruggner RV, Regier A, Johnson J, Viswanathan L, Utterback TR, Sutton GG, Lawson D, Waterhouse RM, Venter JC, Strausberg RL, Berenbaum MR, Collins FH, Zdobnov EM, Pittendrigh BR, 2010. Genome sequences of the human body louse and its primary endosymbiont provide insights into the permanent parasitic lifestyle. *PNAS*, 107(27):12168—12173.

Kneller DG, Cohen FE, Langridge R, 1990. Improvements in protein secondary structure prediction by an enhanced neural network. *J. Mol. Biol.*, 214(1):171—182.

Korf I, Flicek P, Duan D, Brent MR, 2001. Integrating genomic homology into gene structure prediction. *Bioinformatics*, 17 (Suppl. 1):S140—148.

Lambert C, Leonard N, De Bolle X, Depiereux E, 2002. ESyPred3D: Prediction of proteins 3D structures. *Bioinformatics*. 18(9):1250—1256.

Li R, Yu C, Li Y, Lam TW, Yiu SM, Kristiansen K, Wang J, 2009. SOAP2:an improved ultrafast tool for short read alignment. *Bioinformatics*, 25(15):1966—1967.

Mendes ND, Freitas AT, Vasconcelos AT, Sagot MF, 2010. Combination of measures distinguishes pre-miRNAs from other stem-loops in the genome of the newly sequenced *Anopheles darlingi*. *BMC Genomics*, 11:529.

Mikheyev AS, Vo T, Wee B, Singer MC, Parmesan C, 2010. Rapid microsatellite isolation from a butterfly by de novo transcriptome sequencing: performance and a comparison with AFLP-derived distances. *PLoS ONE*, 5(6):e11212.

Mita K, Kasahara M, Sasaki S, Nagayasu Y, Yamada T, Kanamori H, Namiki N, Kitagawa M, Yamashita H, Yasukochi Y, Kadono-Okuda K, Yamamoto K, Ajimura M, Ravikumar G, Shimomura M, Nagamura Y, Shin IT, Abe H, Shimada T, Morishita S, Sasaki T, 2004. The genome sequence of silkworm, *Bombyx mori*. *DNA Res.*,

11(1):27—35.

Mott R, 1997. EST_GENOME: a program to align spliced DNA sequences to unspliced genomic DNA. *Comput. Appl. Biosci.*, 13(4):477—478.

Mullikin JC, Ning Z, 2003. The phusion assembler. *Genome Res.* 13(1):81—90.

Myers EW, Sutton GG, Delcher AL, Dew IM, Fasulo DP, Flanigan MJ, Kravitz SA, Mobarry CM, Reinert KH, Remington KA, Anson EL, Bolanos RA, Chou HH, Jordan CM, Halpern AL, onardiSL, Beasley EM, Bron RC, Chen L, Dunn PJ, Lai Z, Liang Y, Nusskern DR, Zhan M, Zhang Q, Zheng X, Rubin GM, Adams MD, Venter JC, 2000. A whole-genome assembly of *Drosophila*. *Science*, 287(5461):2196—2204.

Needleman SB, Wunsch, CD, 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, 48(3):443—453.

Nene V, Wortman JR, Lawson D, Haas B, Kodira C, Tu ZJ, Loftus B, Xi Z, Megy K, Grabherr M, Ren Q, Zdobnov EM, Lobo NF, Campbell KS, Brown SE, Bonaldo MF, Zhu J, Sinkins SP, Hogenkamp DG, Amedeo P, Arensburger P, Atkinson PW, Bidwell S, Biedler J, Birney E, Bruggner RV, Costas J, Coy MR, Crabtree J, Crawford M, Debruyn B, Decaprio D, Eiglmeier K, Eisenstadt E, El-Dorry H, Gelbart WM, Gomes SL, Hammond M, Hannick LI, Hogan JR, Holmes MH, Jaffe D, Johnston JS, Kennedy RC, KooH, Kravitz S, Kriventseva EV, Kulp D, Labutti K, Lee E, Li S, Lovin DD, Mao C, Mauceli E, Menck CF, Miller JR, Montgomery P, Mori A, Nascimento AL, Naveira HF, Nusbaum C, O'Leary S, Orvis J, Pertea M, Quesneville H, Reidenbach KR, Rogers YH, Roth CW, Schneider JR, Schatz M, Shumway M, Stanke M, Stinson EO, Tubio JM, Vanzee JP, Verjovski-Almeida S, Werner D, White O, Wyder S, Zeng Q, Zhao Q, Zhao Y, Hill CA, Raikhel AS, Soares MB, Knudson DL, Lee NH, Galagan J, Salzberg SL, Paulsen IT, Dimopoulos G, Collins FH, BirrenB, Fraser-Liggett CM, Severson DW, 2007. Genome sequence of *Aedes aegypti*, a major arbovirus vector. *Science*, 316(5832):1718—1723.

Nielsen M, Lundegaard C, Lund O, Petersen TN, 2010. CPHmodels-3.0 – remote homology modeling using structure-guided sequence profiles. *Nucleic. Acids. Res.*, 38(Web server issue):W576—581.

Notredame C, Higgins DG, 1996. SAGA:sequence alignment by genetic algorithm. *Nucleic. Acids. Res.*, 24(8):1515—1524.

Notredame C, Higgins DG, Heringa J, 2000. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.*, 302(1):205—217.

O'Neil ST, Dzurisin JD, Carmichael RD, Lobo NF, Emrich SJ, Hellmann JJ, 2010. Population-level transcriptome sequencing of nonmodel organisms *Erynnis propertius* and *Papilio zelicaon*. *BMC Genomics*, 11:310.

Pearson WR, Lipman DJ, 1988. Improved tools for biological sequence comparison. *PNAS*, 85(8):2444—2448.

Pevzner PA, Tang H, Tesler G, 2004. De novo repeat classification and fragment assembly. *Genome Res.*, 14(9):1786—1796.

Pevzner PA, Tang H, Waterman MS, 2001. An Eulerian path approach to DNA fragment assembly. *PNAS*, 98(17):9748—9753.

Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, Lopez R, 2005. InterProScan:protein domains identifier. *Nucleic. Acids Res.*, 33(Web server issue):W116—120.

Richards S, Gibbs RA, Weinstock GM, Brown SJ, Denell R, Beeman RW, Gibbs R, Bucher G, Friedrich M, Grimmelikhuijzen CJ, Klingler M, Orenzen ML, Roth S, Schroder R, Tautz D, Zdobnov EM, Muzny D, Attaway T, Bell S, Buhay CJ, Chrabose MN, Chavez D, Clerk-Blankenburg KP, Cree A, Dao M, Davis C, Chacko J, Dinh H, Dugan-Rocha S, Fowler G, Garner TT, Garnes J, Gnirke A, Hawes A, Hernez J, Hines S, Holder M, Hume J, Jhangiani SN, Joshi V, Khan ZM, Jackson L, Ovar CK, Kowis A, Lee S, Lewis LR, Margolis J, Morgan M, Nazareth LV, NguyenN, Okwuonu G, Parker D, Ruiz SJ, Santibanez J, Savard J, Scherer SE, Schneider B, Sodergren E, Vattahil S, Villasana D, White CS, Wright R, Park Y, Lord J, Oppert B, Brown S, Wang L, Weinstock G, Liu Y, Worley K, Elsik CG, Reese JT, Elhaik E, Lan G, Graur D, Arensburger P, Atkinson P, Beidler J, Demuth JP, Drury DW, Du YZ, Fujiwara H, Maselli V, Osanai M, Robertson HM, Tu Z, Wang JJ, Wang S, Song H, Zhang L, Werner D, Stanke M, Morgenstern B, Solovyev V, Kosarev P, Brown G, Chen HC, Ermolaeva O, Hlavina W, Kapustin Y, Kiryutin B, Kitts P, Maglott D, Pruitt K, Sapojnikov V, Souvorov A, Mackey AJ, Waterhouse RM, Wyder S, Kriventseva EV, Kadowaki T, Bork P, Ara M, Bao R, Beermann A, Berns N, Bolognesi R, Bonneton F, Bopp D, Butts T, Chaumot A, Denell RE, Ferrier DE, Gordon CM, Jindra M, Lan Q, Lattorff HM, Laudet V, von Levetsow C, Liu Z, Lutz R, Lynch JA, da Fonseca RN, Posnien N, Reuter R, Schinko

JB, Schmitt C, Schoppmeier M, Shippy TD, Simonnet F, Marques-Souza H, Tomoyasu Y, Trauner J, Van der Zee M, Vervoort M, Wittkopp N, Wimmer EA, Yang X, Jones AK, Sattelle DB, Ebert PR, Nelson D, G Scott J, Muthukrishnan S, Kramer KJ, Arakane Y, Zhu Q, Hogenkamp D, Dixit R, Jiang H, Zou Z, Marshall J, Elpidina E, Vinokurov K, Oppert C, Evans J, Lu Z, Zhao P, Sumathipala N, Altincicek B, Vilcinskas A, Williams M, Hultmark D, Hetru C, Hauser F, Cazzamali G, Williamson M, Li B, Tanaka Y, Predel R, Neupert S, Schachtner J, Verleyen P, Raible F, Walden KK, Angeli S, Foret S, Schuetz S, Maleszka R, Miller SC, Grossmann D, 2008. The genome of the model beetle and pest *Tribolium castaneum*. *Nature*, 452(7190):949—955.

Richards S, Liu Y, Bettencourt BR, Hradecky P, Letovsky S, Nielsen R, Thornton K, Hubisz MJ, Chen R, Meisel RP, Couronne O, Hua S, Smith MA, Zhang P, Liu J, Bussemaker HJ, van Batenburg MF, Howells SL, Scherer SE, Sodergren E, Matthews BB, Crosby MA, Schroeder AJ, Ortiz-Barrientos D, Rives CM, Metzker ML, Muzny DM, Scott G, Steffen D, Wheeler DA, Worley KC, Havlak P, Durbin KJ, Egan A, Gill R, Hume J, Morgan MB, Miner G, Hamilton C, Huang Y, Waldron L, Verduzco D, Clerc-Blankenburg KP, Dubchak I, Noor MA, Anderson W, White KP, Clark AG, Schaeffer SW, Gelbart W, Weinstock GM, Gibbs RA, 2005. Comparative genome sequencing of Drosophila pseudoobscura: chromosomal, gene, and cis-element evolution. *Genome Res.*, 15(1):1—18.

Rost B, Yachdav G, Liu J, 2004. The PredictProtein server. *Nucleic Acids Res.*, 32(Web server issue):W321—326.

Salamov AA, Solovyev VV, 2000. Ab initio gene finding in *Drosophila* genomic DNA. *Genome Res.*, 10(4):516—522.

Schneider TD, Mastronarde DN, 1996. Fast multiple alignment of ungapped DNA sequences using information theory and a relaxation method. *Discrete Appl. Math.*, 71(1/3):259—268.

Smith CD, Zimin A, Holt C, Abouheif E, Benton R, Cash E, Croset V, Currie CR, Elhaik E, Elsik CG, Fave MJ, Fernes V, Gadau J, Gibson JD, Graur D, Grubbs KJ, Hagen DE, Helmkampf M, Holley JA, Hu H, Viniegra AS, Johnson BR, Johnson RM, Khila A, Kim JW, Laird J, Mathis KA, Moeller JA, Munoz-Torres MC, Murphy MC, Nakamura R, Nigam S, Overson RP, Placek JE, Rajakumar R, Reese JT, Robertson HM, Smith CR, Suarez AV, Suen G, Suhr EL, Tao S, Torres CW, van Wilgenburg E, Viljakainen L, Walden KK, Wild AL, Yell M, Yorke JA, Tsutsui ND, 2011. Draft genome of the globally widespread and invasive Argentine ant (*Linepithema humile*). *PNAS*, 108(14):5673—5678.

Smith CR, Smith CD, Robertson HM, Helmkampf M, Zimin A, Yell M, Holt C, Hu H, Abouheif E, Benton R, Cash E, Croset V, Currie CR, Elhaik E, Elsik CG, Fave MJ, Fernes V, Gibson JD, Graur D, Gronenberg W, Grubbs KJ, Hagen DE, Viniegra AS, Johnson BR, Johnson RM, Khila A, Kim JW, Mathis KA, Munoz-Torres MC, Murphy MC, Mustard JA, Nakamura R, Niehuis O, Nigam S, Overson RP, Placek JE, Rajakumar R, Reese JT, Suen G, Tao S, Torres CW, Tsutsui ND, Viljakainen L, Wolschin F, Gadau J, 2011. Draft genome of the red harvester ant *Pogonomyrmex barbatus*. *PNAS*, 108(14):5667—5672.

Smith TF, Waterman MS, 1981. Identification of common molecular subsequences. *J. Mol. Biol.*, 147(1):195—197.

Suen G, Teiling C, Li L, Holt C, Abouheif E, Bornberg-Bauer E, Bouffard P, Caldera EJ, Cash E, Cavanaugh A, Denas O, Elhaik E, Fave MJ, Gadau J, Gibson JD, Graur D, Grubbs KJ, Hagen DE, Harkins TT, Helmkampf M, Hu H, Johnson BR, Kim J, Marsh SE, Moeller JA, Munoz-Torres MC, Murphy MC, Naughton MC, Nigam S, Overson R, Rajakumar R, Reese JT, Scott JJ, Smith CR, Tao S, Tsutsui ND, Viljakainen L, Wissler L, Yell MD, Zimmer F, Taylor J, Slater SC, Clifton SW, Warren WC, Elsik CG, Smith CD, Weinstock GM, Gerardo NM, Currie CR, 2011. The genome sequence of the leaf-cutter ant *Atta cephalotes* reveals insights into its obligate symbiotic lifestyle. *PLoS Genet.*, 7(2):e1002007.

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L, 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.*, 28(5):511—515.

Van't Hof AE, Saccheri IJ, 2010. Industrial melanism in the peppered moth is not associated with genetic variation in canonical melanisation gene candidates. *PLoS ONE*, 5(5):e10889.

Wang J, Wong GK, Ni P, Han Y, Huang X, Zhang J, Ye C, Zhang Y, Hu J, Zhang K, Xu X, Cong L, Lu H, Ren X, He J, Tao L, Passey DA, Yang H, Yu J, Li S, 2002. RePS: a sequence assembler that masks exact repeats identified from the shotgun data. *Genome Res.*, 12(5):824—831.

Wang XW, Luan JB, Li JM, Bao YY, Zhang CX, Liu SS,

2010. De novo characterization of a whitefly transcriptome and analysis of its gene expression during development. *BMC Genomics*, 11:400.

Werren JH, Richards S, Desjardins CA, Niehuis O, Gadau J, Colbourne JK, Beukeboom LW, Desplan C, Elsik CG, Grimmelikhuijzen CJ, Kitts P, Lynch JA, Murphy T, Oliveira DC, Smith CD, van de Ze L, Worley KC, Zdobnov EM, Aerts M, Albert S, Anaya VH, Anzola JM, Barchuk AR, Behura SK, Bera AN, Berenbaum MR, Bertossa RC, Bitondi MM, Bordenstein SR, Bork P, Bornberg-Bauer E, Brunain M, Cazzamali G, Chaboub L, Chacko J, Chavez D, Childers CP, Choi JH, Clark ME, Claudianos C, Clinton RA, Cree AG, Cristino AS, Dang PM, Darby AC, de Graaf DC, Devreese B, Dinh HH, Edwards R, Elango N, Elhaik E, Ermolaeva O, Evans JD, Foret S, Fowler GR, Gerlach D, Gibson JD, Gilbert DG, Graur D, Grunder S, Hagen DE, Han Y, Hauser F, Hultmark D, Hunter HC, Hurst GD, Jhangian SN, Jiang H, Johnson RM, Jones AK, Junier T, Kadowaki T, Kamping A, Kapustin Y, Kechavarzi B, Kim J, Kiryutin B, Koevoets T, Kovar CL, Kriventseva EV, Kucharski R, Lee H, Lee SL, Lees K, Lewis LR, Loehlin DW, Logsdon JM, Lopez JA, Lozado RJ, Maglott D, Maleszka R, Mayampurath A, Mazur DJ, McClure MA, Moore AD, Morgan MB, Muller J, Munoz-Torres MC, Muzny DM, Nazareth LV, Neupert S, Nguyen NB, Nunes FM, Oakeshott JG, Okwuonu GO, Pannebakker BA, Pejaver VR, Peng Z, Pratt SC, Predel R, Pu LL, Ranson H, Raychoudhury R, Rechtsteiner A, Reese JT, Reid JG, Riddle M, Robertson HM, Romero-Severson J, Rosenberg M, Sackton TB, Sattelle DB, Schluns H, Schmitt T, Schneider M, Schuler A, Schurko AM, Shuker DM, Simoes ZL, Sinha S, Smith Z, Solovyev V, Souvorov A, Springauf A, Stafflinger E, Stage DE, Stanke M, Tanaka Y, Telschow A, Trent C, Vattathil S, Verhulst EC, Viljakainen L, Wanner KW, Waterhouse RM, Whitfield JB, Wilkes TE, Williamson M, Willis JH, Wolschin F, Wyder S, Yamada T, Yi SV, Zecher CN, Zhang L, Gibbs RA, 2010. Functional and evolutionary insights from the genomes of three parasitoid *Nasonia* species. *Science*, 327(5963):343—348.

Wurm Y, Wang J, Riba-Grognuz O, Corona M, Nygaard S, Hunt BG, Ingram KK, Falquet L, Nipitwattanaphon M, Gotzek D, Dijkstra MB, Oettler J, Comtesse F, Shih CJ, Wu WJ, Yang CC, Thomas J, Beaudoing E, Praderv S, Flegel V, Cook ED, Fabbretti R, Stockinger H, Long L, Farmerie WG, Oakey J, Boomsma JJ, Pamilo P, Yi SV, Heinze J, Goodisman MA, Farinelli L, Harshman K, Hulo N, Cerutti L, Xenarios I, Shoemaker D, Keller L ,2011. The genome of the fire ant *Solenopsis invicta*. *PNAS*, 108(14):5679—5684.

Xia Q, Guo Y, Zhang Z, Li D, Xuan Z, Li Z, Dai F, Li Y, Cheng D, Li R, Cheng T, Jiang T, Becquet C, Xu X, Liu C, Zha X, Fan W, Lin Y, Shen Y, Jiang L, Jensen J, Hellmann I, Tang S, Zhao P, Xu H, Yu C, Zhang G, Li J, Cao J, Liu S, He N, Zhou Y, Liu H, Zhao J, Ye C, Du Z, Pan G, Zhao A, Shao H, Zeng W, Wu P, Li C, Pan M, Yin X, Wang J, Zheng H, Wang W, Zhang X, Li S, Yang H, Lu C, Nielsen R, Zhou Z, Xiang Z, 2009. Complete resequencing of 40 genomes reveals domestication events and genes in silkworm (*Bombyx*). *Science*, 326(5951):433—436.

Xia Q, Zhou Z, Lu C, Cheng D, Dai F, Li B, Zhao P, Zha X, Cheng T, Chai C, Pan G, Xu J, Liu C, Lin Y, Qian J, Hou Y, Wu Z, Li G, Pan M, Li C, Shen Y, Lan X, Yuan L, Li T, Xu H, Yang G, Wan Y, Zhu Y, Yu M, Shen W, Wu D, Xian g Z, Yu J, Wang J, Li R, Shi J, Li H, Su J, Wang X, Zhang Z, Wu Q, L i J, Zhang Q, Wei N, Sun H, Dong L, Liu D, Zhao S, Zhao X, Meng Q, Lan F, Huang X, Li Y, Fang L, Li D, Sun Y, Yang Z, Huang Y, Xi Y, Qi Q, He D, Huang H, Zhang X, Wang Z, Li W, Cao Y, Yu Y, Yu H, Ye J, Chen H, Zhou Y, Liu B, Ji H, Li S, Ni P, Zhang J, Zhang Y, Zheng H, Mao B, Wang W, Ye C, Wong GK, Yang H, 2004. A draft sequence for the genome of the domesticated silkworm (*Bombyx mori*). *Science*, 306(5703):1937—1940.

Xue J, Bao YY, Li BL, Cheng YB, Peng ZY, Liu H, Xu HJ, Zhu ZR, Lou YG, Cheng, Zhang CX, 2010. Transcriptome analysis of the brown planthopper *Nilaparvata lugens*. *PLoS ONE*, 5(12):e14233.

Yeh RF, Lim LP, Burge CB, 2001. Computational inference of homologous gene structures in the human genome. *Genome Res.*, 11(5):803—816.

商英璠,肖晓旦,刘雁书,2005. 生物信息数据库与查询检索的简介. 医学信息学,18(4):328—331.

张春霆,2000. 生物信息学的现状与展望. 院士论坛,22(6):17—20.