一、生物信息学的多学科本质

在过去的几十年间，生物医学经历了一场重要的变革。一些技术革新，如高通量测序技术已经容许人们在整个基因组水平上研究细胞、组织和完整生物个体的各种分子，也使得生物医学的资料倍增。通过开发特殊的计算机程序和开发旨在组织和分析这些生物学知识的技术方法，使其将获得的全部信息转化成有用知识时所遇到的挑战正在被逐步克服。

生物信息学和计算机生物学包括多种学科的理论和应用知识，例如计算机科学、数学、统计学、物理学和工程学，并利用这些学科知识来解决生物医学问题，同时改进对生物医学现象的理解。虽然很多情况下生物信息学和计算机生物学被认为是同义词，但是根据美国国家卫生研究院（NationalInstituteofHealth，NIH）生物医学咨询科技团体（Biomedical Information Science Technology Initiative，BISTI）（http://www.bisti. nih.gov/bistic2.cfm）的解释，其实它们是不完全相同的（表1）。

表1 生物信息学与计算生物学的异同

	生物信息学	计算生物学
定义	是研究、开发和应用计算机工具和方法来扩展生物学、医学、行为学和健康知识应用的科学，包括获得、储存、组织、存档、分析或使这些材料形象化；	是指开发和应用数据分析及理论的方法、数学建模和计算机仿真技术等；
相异点	是典型交叉学科，涉及的学科包括数学、统计学、化学、物理学、生物学和计算机科学等； 以生命科学中的现象和规律作为研究对象，以解决生物学问题为最终目标，数学和计算机仅仅是解决问题的工具和手段； 研究领域相当广泛，几乎渗透到现代生物学研究的每一个领域。	作为组织和分析生物学资料的第一步，它通常先建立数据库，例如，大量有序和一致的资料库，这通常要和计算机程序联系起来，并用来存储所有类型的生物记录，如DNA或蛋白质序列的记录。 为了处理这些资料，需要不断开发特定的软件，以更新、查询、检索和储存这个系统中的资料组分。 现在很多这样的数据库是相互联系的，通过简单的查询就可以得到完整且详细的信息。
相异点	最基本的目的都是培养新的生物学洞察力和建立整体的观点，以期从中获得一致的生物学机制（http://www.bisti.nih.gov/bistic2.cfm）。	

国际核苷酸序列数据库合作组织（The International Nucleotide Sequence Database Collaboration，INSDC）（www.insdc.org）的开发经历已超过18个年头了。目前它覆盖了日本的DNA数据库（DDBJ），欧洲分子生物学实验室（EMBL）和在美国国家生物技术信息中心（NCBI）的GenBank。这个组织有项政策，就是人们可以免费和不受限制地利用他们的信息。GenBank是美国国家卫生研究院位于NCBI中的遗传序列数据库（http://www.ncbi.nlm.nih.gov/Genbank/index.html）。这一综合性的数据库包含超过26万种已鉴定的生物和已经公布的最终DNA序列数据。欧洲分子生物学实验室的核苷酸序列的数据库，也被称为EMBL-银行（http://www.ebi.ac.uk/embl），它包含欧洲的主要核苷酸序列资源。在日本，DDBJ数据库（http://www.ddbj.nig.ac.jp）只是DNA数据库银行。大部分DNA和RNA序列资源主要来源于研究者的个人提交，或者来自不同类型的测序计划，包括 cDNA测序、基因组测序和专利申请。向GenBank提交序列主要有两种方法：Bankit和Sequin。

Bankit是基于网页提交的工具，一般少数简单的序列推荐用Bankit来提交。每天这三个数据库中的有关信息都要进行交换和更新，旨在保证它们可以纳入最新可用的序列数据并能够达到全球覆盖。这种技术的创新促进了很多建设性方案的形成，并随着资料的迅速积累而在生物学团队中扩散开来。

目前，GenBank已成为世界权威的核酸序列登记数据库。科研人员研究测定的核酸序列在正式发表之前，一般都先到GenBank注册，各类学术刊物一般也要求提供序列的GenBank登录号。现在收录在GenBank中的已测基因组全序列的病毒种类达几千种，而且每年高速递增。

随着国际核苷序列联合数据库（INSDC http://www.insdc.org）的完成，为了收集各种生物学数据，出现并形成了一股巨大的社会力量，以发展和完善各种数据库及其工具。这些数据库包括序列数据库、代谢途径数据库、蛋白质组学数据库、细胞器数据库、人类疾病数据库、植物数据库和免疫生物学数据库等。人们付出了巨大的努力，目的就是通过一个可靠和适当的方式，为科学界提供可以利用的分子数据。一个很好的例子是美国国家人类基因组研究所的分子生物学数据库汇集（The Molecular Biology Database Collection）。它是一个每年都更新的公共数据库，并在期刊《核酸研究》（*Nucleic Acids Research*）（http://nar.oxfordjournals.org）发表论文，每年介绍上百个数据库。

由于这些资料库越看越复杂，科学家开始使用基于知识发现和资料发掘的伽玛技术来从这些数据库中提取信息。数据库的知识发现（knowledge discovery in databases，KDD）是一个计算方法，它存在于基础的数据库结构中，如资料选择、预处理、转化和降维插值等。这种知识可以用来搜索模式规律、联系规则、短暂的结果和数据之间合理的联系等，并可以搜索平常不被专家认识到的一些数据。从KDD获得的成果是重要的信息系统，并能够被决策系统所采用。

数据库开采发掘的方法作为知识发现方法的一部分，也是非常有用的，它可以探索大量数据，大体上包括：

(a) 数据的探索；

(b) 方式或模型的解释；

(c) 利用其它数据集模型验证上面得到的方法。

人们已经从这些模型和在对它们的应用过程中获得了许多宝贵信息。这些信息涉及生物分子的发现模式、生物医学文献中的文本挖掘、数据整合和基因组序列的概率模型等。人们利用这些资源帮助了很多的医学研究项目，包括提出研究方案以及设计大型实验。

计算机与以实验为基础的方式、方法整合是对整个生物医学领域的一项重大挑战。虽然某些领域，如基因表达领域和系统发生学领域的实验人员每天都使用一些基于计算机的技术方法，但大多数生物学家仍远远没有有效使用电脑。正如图1所示，华盛顿大学发起的生理人计划（The physiome project）是全世界几个松散联系的研究组织共同努力的结果。它的成立是为了促进数据库建设、综合定量描述和建模（http://nsr.bioeng.washington.edu）的发展，这也为解释生命组学（Physiome）提供了条件。生物学家不能忽视生命组学这个关键领域的进展。生物医学整体的发展依赖于其更广泛地联系生物信息学和计算生物学。

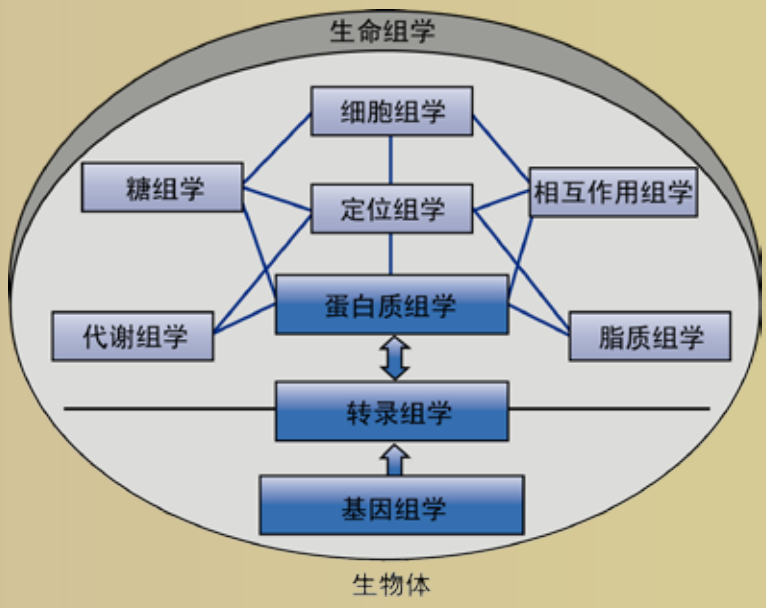

图1 生命组学中生物组织不同领域之间的关系。

二、癌症生物信息学

癌症是由于一些遗传改变和表观遗传改变而导致的疾病，在它最简单的形式中，癌症是一种由于一个细胞基因组变化而导致的遗传性疾病。这种遗传改变包括点突变、插入突变、缺失突变和染色体易位等。这些基因的变化可以导致细胞和组织生长异常，这就是肿瘤的表型特征。虽然控制癌症发生的分子机制研究一直是过去50年来研究的重点，其中包括使用了很多分子生物学手段的研究，但对人类癌症的发生机制还了解不多。尽管在细胞转化基础上，人们已经了解了许多分子遗传学和表观遗传学改变，但导致肿瘤表型的复杂过程才刚刚开始被人们理解。目前，遗传学上，癌症的基础研究正经历着一场变革。