12): 751~ 756
12 Paulsen H, Robertson J M, Wintermeyer W. Topological arrangement of two transfer RNAs on the ribosome: fluorescence energy measurements between A and P site bound $tRNA^{Phe}$. J Mol Biol, 1983, **167** (2): 411~ 426
13 Ryaboval L A, Selivanova O M, Baranov V I, *et al*. Does the channel for nascent peptide exist inside the ribosome? Immune electron microscopy study. FEBS Lett, 1988, **226** (2): 255~ 260
14 Yonath A, Leonard K R, Wittmann H G, *et al*. A tunnel in the large ribosomal subunit revealed by three-dimensional image reconstruction. Science, 1987, **236** (4803): 813~ 816

Recent Advance in the Study of Ribosomal Model.

LIU Shu-Qun, LIU Ci-Quan (*Laboratory of Cellular and Molecular Evolution, Kunming Institute of Zoology, The Chinese Academy of Sciences, Kunming* 650223, China).

Abstract With the technological developments of cryoelectron microscope, X-ray diffraction and the growing data available on various components of ribosome, some marvelously intricate structural models of the *Escherichia coli* 70S ribosome have been reconstructed. The picture of the ribosomal model are detailed, including the placement of the mRNA, the arrangement of the A-site and P-site tRNAs and the peptidytransferase within the interface gap as well as the path of nascent polypeptide chain, which results in a better understanding of the structure and function of ribosome as well as the translational process.

Key words ribosomal model, high resolution, placement

后基因组时代的基因组功能注释

解 涛 梁卫平 丁达夫[1)]
(中国科学院上海生物化学研究所, 上海 200031)

摘要 基因组功能注释是后基因组时代功能基因组学研究的热点领域. 从基因组功能注释的研究内容与研究手段出发, 重点综述了生物信息学在该领域方法学上的研究进展, 并展望了今后的发展前景.

关键词 后基因组, 基因组, 基因组功能注释
学科分类号 Q71

早在 1920 年, Winkles 从 GENes 和 chromosOMEs铸成 GENOME (基因组) 一词. 随着人类基因组计划的实施, 开创了以图谱制作与序列测定为目的的序列基因组学时代[1]. 宏伟的人类基因组计划自 20 世纪 90 年代初正式启动以来, 已经提前完成了高密度的遗传图谱和物理图谱的制作, 测序工作正在紧张的进行当中. 模式生物的基因组计划也开展得如火如荼. 自 1995 年第一个细菌基因组——流感嗜血杆菌 (*H. influenzae*) 全基因组序列发表以来[2], 已经完成了包括大肠杆菌、酿酒酵母在内的十余种微生物的全基因组序列. 1998 年 12 月, 第一个多细胞真核生物线虫的基因组在美国《Science》杂志上发表[3]. 与此同时, DNA 序列数据库一直处在指数增长之中, GenBank的容量每隔 14 个月就翻一番. 1998 年 12 月 15 日释放的 110.0 版包含碱基数约 22 亿, 序列数约 304 万 (http://www.ncbi.nlm.nih.gov/web/Genbank/index.html). 在 1995 年左右, 一些眼光长远的科学家预见到序列爆炸的大趋势, 提出了"后基因组" 的概念[4], 即在基因组静态的碱基序列弄清楚之后, 转入对基因组动态的生物学功能的研究即"功能基因组学"[5]. 可见, 由于人类基因组计划的顺利进展, 提供了以往不可想象的巨量的生物学信息资源, 推动了世纪之交的生物学走向以功能基因组学为标志的后基因组时代, 从根本上改变了传统生物学的思维方式. 人类基因组计划 1998 年~ 2003 年的新五年规划[6]在重申主要目标是测定人基因组全序的同时, 多处强调了基因组研究从传统的序列基因组学向功能基因组学的转

1) 通讯联系人.
Tel: (021) 64374430-255, E-mail: dingdafu@server.shcnc.ac.cn
收稿日期: 1999-03-29, 修回日期: 1999-08-16

移. 基因组功能注释 (genome annotation) 是功能基因组学的主要研究目标, 其包括应用生物信息学方法, 高通量地注释基因组所有编码产物的生物学功能. 该领域已经成为后基因组时代的研究热点[7,8].

1 基因组功能注释的研究内容与方法

顾名思义, 基因组功能注释的研究对象是基因组序列, 其研究内容可分为以下三个层次.

1.1 基因组组成元素的识别

首先要预测基因组的全部编码区或称“开放阅读框架 (open reading frame, ORF)”. ORF 的识别手段可以分为两大类: 一类是评估未知 DNA 片段的编码可能性, 称为概率型方法, 如应用隐马尔可夫模型的 GENSCAN[9]; 另一类是通过同源性比较搜寻蛋白质库或 dbEST 库找寻编码区[10]. 需要指出的是, EST 测序的飞速发展, 使得 dbEST 中的记录已经超过一百多万条. 对于人基因组来说, 理论上接近所有的基因都在 dbEST 库中有对应的 EST. 这种方法越来越受到重视, 不仅因为它可以判断一段 DNA 中是否包含 ORF, 而且能精确地给出该基因的内含子和外显子的剪切模式. 在线虫基因组的 ORF 识别中, 综合运用了上述两种手段[3]. 总的来看, 原核基因组的基因识别正确率较高; 真核生物比较低, 方法学上仍需要改进. 非编码区包括各类重复序列、基因表达调控序列等, 对它们的注释同样具有重要意义. 相对编码区而言, 这方面的工作较少.

1.2 注释所有 ORF 产物的功能

这是目前基因组功能注释的主要层次. 对于已有实验证据的基因产物只需将功能描述与相应基因关联即可. 对于无实验证据的基因, 从生物信息学[11]研究的角度出发, 目前主要有三大类方法可用于大通量的基因组功能注释工作: a. 用最大相似的同源基因的功能注释咨询序列; b. 用模体 (MOTIF) 搜索, 因为模体往往是功能相关的保守序列; c. 用 Tatusov 等的 COG (cluster of orthologous group) ——直系同源簇方法[12], 即用不同种族的基因成对相似聚类法把它们划分成各种直系同源簇, 从而可以用同一簇中的已知基因注释未知基因的功能. 在序列分析之外, 还有两个新兴领域对基因组功能注释意义重大: 结构基因组的研究与蛋白质组的研究. 它们正在使得基因组功能注释发生深刻的变化. 有关具体问题下面还将细述.

1.3 基因之间相互作用及比较基因组学研究

基因组的各组成基因在序列水平, 有位置排列的顺序关系; 在转录、表达水平, 又有基因、基因产物之间的相互作用. 因此完整地了解基因的功能必然要研究其在生物体代谢途径中的地位, 并尽可能揭示它们之间相互调控的机制, 绘制出调控网络的图式. 比较基因组研究不仅可以揭示生命的起源、进化等重大生物学问题, 还具有不可低估的实用价值. 比如通过细菌、真核生物的比较基因组研究, 有望筛选出只在细菌中保守的基因, 作为广谱抗菌素的药靶. 目前该层次的研究正处于起步阶段.

2 当前基因组功能注释的主要进展

后基因组时代的到来必然要求基因组功能注释工作成为功能基因组学研究中的重要组成部分. 我们在这里主要讨论当前最受关注的第二个层次, 即应用生物信息学方法进行 ORF 功能预测问题.

2.1 最大序列相似性搜索

基于序列比较的最大相似法为序列基因组学解决了许多问题, 在各种基因及蛋白质的进化、结构、催化等特性的研究中取得了很多成果. 但是现在经大规模基因组比较资料发现这肯定会导致错误. 比如 1998 年的网络杂志《In Silico Biology》第一期[13]中, 列举了大量此类错误. 错误的根源在于“同源= 功能相似”的假定. 相似比较没有解析各种族基因间的进化关系, 如趋同和趋异、重复 (duplication)、基因缺失 (gene lose)、基因水平转移 (gene horizontal transfer) 等. 由于其具有大通量与自动化的优势, 与线虫基因组测序同期完成的线虫与酵母之间的直系同源体的搜索[14]仍然采用这种方案. 为减少错误, 实际运用中作了改进: 设立了几个同源性指标等级, 如 P 值从 10^{-100} 到 10^{-10}之间有 4 档, 另外还有同源区域的长度比例条件. 这样包含了一对多、多对多的直系同源关系, 部分改善了最高相似法的结果, 但没有从根本上解决问题.

2.2 序列模体搜索

序列模体搜索的是查找序列上的局部特征. 在序列整体同源性不明显的情况下, 模体搜索可以提高功能预测的灵敏度, 模体分析一般由两部分组成: 首先收集现有的蛋白质家族, 通过蛋白质家族各成员的多重联配来构造模体数据库, 而后通过搜索该数据库预测未知蛋白质的功能. 典型的模体数

据库有 Prosite[15] 等. 越来越多的事实表明, 模体本身具有层次性, 在一个蛋白质家族具有相同的模体的情况下, 亚家族可能具有各自特异的模体, 它们与功能的联系更为特异[16]. 而目前现有的模体库在制作时没有深入考虑进化关系, 其形成的模体往往不是功能特异的. 这成为用模体搜索法作基因组功能注释的最大障碍.

2.3 COG 方法

Tatusov 等[12] 的 COG 方法是在基因组水平上找寻直系同源体, 从而预测未知 ORF 的生物学功能, 所谓直系同源 (ortholog) 是指不同物种中由同一个祖先基因特化而来的对应基因, 相应旁系同源 (paralog) 是指基因组内基因复制形成的多个基因[17]. 一般而言, 直系同源之间保持了同样的功能, 旁系同源则进化出不同的功能. 因此确定直系同源对功能注释的可靠性很重要. COG 的构建者提出了三项考核标准, 即 A 基因组的某个基因 a 是 B 基因组中基因 b 的直系同源需满足: 第一, a 是 b 在 A 基因组中同源性最高的基因; 第二, 若 C 基因组与 B 基因组在系统发育树上的距离大于 A 到 B 的距离, c 是 b 在 C 基因组中同源性最高的基因, 则要求 ab 之间的同源性好于 cb 之间的同源性; 第三, ab 的同源区域大于各序列长度的 60%. 若三条件在 ab 互换时也成立, 则 ab 两基因互为直系同源. 他们以 7 种全基因组序列已知的生物为对象, 用 BLAST 交错搜寻, 构造出 720 个 COG, 由于 "直系同源= 功能相似" 比 "同源= 功能相似" 更接近于生物学的客观实际, 从而可以将功能信息从 COG 的一个成员传递到 COG 中其他功能未知的成员. 该方法充分利用了全基因组已知的优势, 大大提高了功能注释的准确度. 目前其考察的基因组已扩大到 8 个.

2.4 进化分析方法

最近, COG 方法也面临挑战. COG 的核心即直系同源的判断方法仍在序列相似性比较的框架内. Eisen[18] 主张用较为严格的进化分析的方法划分直系同源. 具体方案是: 先找寻同源性为基础的蛋白质家族, 再用进化分析方法将其分为亚家族, 并用亚家族中已知蛋白质的功能描述注释该亚家族中功能未知的成员. 由于基因树与物种进化关系形成的种族树之间常有矛盾, Page 等[19] 发展了和谐树 —— "RECONCILED TREE" 方法来识别基因重复与基因丢失等进化事件. 和谐树反映了基因在基因组载体上的进化历程 (图 1). 进化分析的范围可以从单个基因组扩大到具有不同系统发育位置的多个基因组. 进化分析方法涉及的分析工具较多, 无法自动化操作, 难以实现高通量的功能注释.

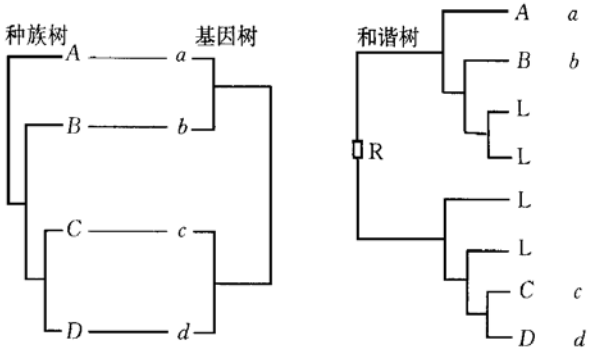

图 1 基因树、种族树与和谐树

图中基因树与种族树是不和谐的. 若引入一次基因重复 (R) 和四次基因丢失 (L) 则构成和谐树. *a*~ *d*: 表示基因; *A* ~ *D*: 表示各种族.

2.5 进化印记搜索

在上述方案的基础上, 我们研究小组发展出一种利用生物分子进化印记 —— 直系同源体特异的模体注释基因组功能的简便与有效的方案[20]. 该方案综合了进化分析的准确与模体搜索的快速的特点, 对 5 个家族检验获得初步成功, 显示出该方案具有潜在的优势.

2.6 亚细胞定位

蛋白质的功能与其亚细胞定位密切相关. 蛋白质序列分析有助于推测亚细胞定位. 而亚细胞定位所提供的信息往往可以在同源性分析得出的结果模棱两可时起到 "一锤定音" 的效果. 目前预测未知蛋白质的亚细胞定位的方法主要是从蛋白质的氨基酸组成出发. Reinhardt 等[21] 将蛋白质按来源分为真核、原核两大类, 用神经网络法根据蛋白质的氨基酸组成来判断该蛋白的亚细胞定位, 结果比较好, 其中原核生物蛋白质的定位准确率达到 81%. Andrade 等[22] 指出用整个蛋白质的氨基酸组成显得比较粗糙, 决定蛋白质亚细胞定位的主要因素是蛋白质表面氨基酸的性质. 他们用主成分分析法研究蛋白质在核内、胞质、胞外的分布, 总准确率高于前一种方法.

2.7 结构基因组学

结构基因组学的兴起使得三维结构模建和结构类的识别成为基因组功能注释的一个重要方面. 越来越多的例子表明, 同样的三维结构可以由很不相

似的序列折叠而成，而三维结构尤其是关键部分的三维结构是决定蛋白质生物学功能的基础. 如果能够模拟出未知基因的蛋白质产物的三维结构，就可以根据结构与功能的关系作出功能注释. 由于目前从头预测三维结构尚难达到实际应用的程度，而同源模建要求有一定程度的序列同源性的模板蛋白，所以很多未知 ORF 的蛋白质产物无法模建出可信度高的结构. 在这种情况下结构类的识别较有实际意义. 由于结构类与蛋白质超家族有对应关系，故可根据蛋白质所属的超家族对其功能作出初步的推测. 目前的结构类识别方法研究的热点领域是“穿线”法——Threading[23]. 有若干研究小组正通过实验与模拟方法系统地分析基因组上所有基因产物的空间结构，因此赋予结构基因组学以新的含义. 类似于序列模体的概念，由蛋白质特定区域形成的空间上的三维模体得到越来越多的重视，三维模体搜索方法发展得很快[24]，有望成为一种新的功能注释的信息来源. 进一步，结构基因组学的研究可以深入探求蛋白质为何具有特定的生物学功能. Bryant 等给出了一个实例[25]. PTEN 基因编码一个 403 个氨基酸残基的蛋白质，已有文献报道其 123、124、129 位的突变可能导致 Cowden 病. 由于其结构尚未解出，不能理解致病机制. 用 BLAST 搜索，找不到具有已知结构的同源序列. 他们用“同源”的可传递性找到 PTEN 的一个有结构信息的同源蛋白 Cdc14b2，其 PDB 编号为 1VHR，编码一个磷酸酶. 通过“穿线”法作出序列——结构联配. 比较之后，发现 PTEN124 位的半胱氨酸与 1VHR 磷酸酶活性位点的半胱氨酸对应. 而 1VHR 该位点突变为丝氨酸会破坏其磷酸酶活性. 由此推断 PTEN 该位点突变为精氨酸导致 Cowden 病的机制也是磷酸酶的活性的破坏.

2.8 蛋白质组学

蛋白质组是生命状态的直接体现，随发育阶段、特定组织甚至所处的环境的变迁而变化，反映了蛋白质后加工等作用，蕴藏着巨量的动态的生命活动信息[26]. 序列分析难以处理的没有任何同源序列的“孤儿”基因，有望从蛋白质组的表达变化规律中找到其生物学功能的线索，进而揭示出它在整个功能网络中的地位[27]. 目前，蛋白质组的核心技术 2D-Gel 和质谱分析发展很快，可以一次分离几千甚至上万蛋白质点和鉴定出翻译后加工的机制. 随着蛋白质组技术的日益成熟，其不仅可以作为现有功能注释的鉴定和补充，甚至可以独立地完成基因组的功能注释.

3 基因组功能注释的展望

随着基因组序列数据的积累和生物信息学的飞速发展，将会有更灵敏、更有效的算法出现，功能注释的可靠性会不断提高，范围会不断扩大. 随着完整基因组数量的增加，比较基因组学也将提供更多的生物进化历程的信息.

我们认为，基因组功能注释有几个值得重视的方向. a. 非编码区的功能注释. 由于目前测定的基因组多是单细胞原核生物，非编码区比例很小，所以研究工作相对较少. 而高等真核生物基因组的 90% 以上是非编码区. 其中有很多是具有生物学功能意义的片段，它们对于全面理解基因组功能，尤其是了解各相关基因之间的调控关系是不可缺少的. 随着多细胞真核生物的基因组序列的出现，可以预计，非编码区的功能注释将成为新的热点. b. 人类物理图谱、基因图谱的利用. 1998 年 10 月 23 日出版的《Science》的基因组专辑发表了国际上几个知名研究机构共同努力完成的包含人 3 万个基因的物理图[28]，准确度比以往提高了 2~ 3 倍. 通过这张图，可以将功能注释与基因定位数据联系起来，再通过 OMIM (人遗传疾病表型) 等数据库查询临床表型. 这具有非常重大的理论和实际意义. 在以上各分析方法的基础上，今后的基因组功能注释将向更高层次发展，即确定所有基因组成分在生物体功能网络上的地位，并进而从根本上沟通基因型与表型，即整体生物学. 完整的生物学功能是在生化途径 (biochemical pathway) 例如代谢途径、调控途径等中体现出来的. Karp 等[29]制作的大肠杆菌的完整代谢图谱——ECOCYC 就是一个典型代表. 日本京都大学的 Bono 等[30]重建了 *E. coli*, *H. influenzae* 等 5 种全基因组已测定的微生物的 20 种氨基酸的代谢途径. 在这个过程中，一些原来没有功能信息或标定错误的 ORF 得到了正确的功能描述. 他们将陆续积累的研究结果以超文本形式发放到 INTERNET 上，取名 KEGG (Kyoto Encyclopedia of Genes and Genomes). 目前已发展到 9.0 版. 今后此类研究将引起更大的关注. 表 1 中列出了目前互联网上有关基因组功能注释的一些 WWW 站点地址.

表 1 有关基因组功能注释的 WWW 站点

名称	说明	WWW 地址
AAT	基因组分析和注释工具	Http://genome.cs.mtu.edu/aat
COG	直系同源体簇分析数据库	Http://www.ncbi.nlm.nih.gov/COG/
EcoCyc	大肠杆菌的基因与代谢	Http://ecocyc.pangeasystems.com/ecocyc/ecocyc.html
KEGG	京都基因与基因组百科全书	Http://www.genome.ad.jp/kegg/
OMIM	在线人类孟德尔遗传资料	Http://www.ncbi.nlm.nih.gov/Omim/
PEDENT	完整基因组的生物信息学分析	Http://pedant.mips.biochem.mpg.de/frishman/pedant.html
SAS	结构为基础的基因组序列分析	Http://www.biochem.ucl.ac.uk/cgi-bin/sas/query.cgi
WIT	基因组代谢途径重构	Http://www.cme.msu.edu/WIT/

参 考 文 献

1 Rowen L, Mahairas G, Hood L. Sequencing the human genomes. Science, 1997, **278** (5338): 605~607

2 Fleischmann R, Adams M D, White O, *et al*. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd, Science, 1995, **269** (5223): 496~512

3 *C. elegans* Sequencing Consortium. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. Science, 1998, **282** (5396): 2012~2017

4 Nowak R. Genetics: entering the postgenome era. Science, 1995, **270** (5235): 368~369

5 Hieter P, Boguski M. Functional genomics: it's all how you read it. Science, 1997, **278** (5338): 601~602

6 Collins F, Patrinos A, Jordan E, *et al*, New goals for the U. S. human genome project: 1998-2003. Science, 1998, **282** (5389): 682~689

7 Overton G C, Bailey C, Crabtree J, *et al*. The GAIA software framework for genome annotation. Pac Symp Biocomput, 1998, (3): 291~302

8 Bork P, Dandekar T, Diaz-Lazcoz Y, *et al*. Predicting function: from genes to genomes and back. J Mol Biol, 1998, **283** (4): 707~725

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

10 Bailey L C Jr, Searls D B, Overton G C. Analysis of EST-driven gene annotation in human genomic sequence. Genome Res, 1998, **8** (4): 362~376

11 Boguski M. Bioinformatics. Curr Opin Genet Dev, 1994, **4** (3): 383~388

12 Tatusov R L, Koonin E V, Lipman D J. A genomic perspective on protein families. Science, 1997, **278** (5338): 631~637

13 Galperin M, Koonin E. Sources of systematic error in functional annotation of genomes: domain rearrangement, non-orthologous gene displacement, and operon disruption. http://www.bioinfo.de/isb/1998/01/0007/

14 Chervitz S, Aravind L, Sherlock G, *et al*. Comparison of the complete protein sets of worm and yeast: orthology and divergence. Science, 1998, **282** (5396): 2018~2028

15 Hofmann K, Bucher P, Falquet L, *et al*. The PROSITE database, its status in 1999. Nucleic Acids Res, 1999, **27** (1): 215~219

16 Nevill-Manning M, Wu T D, Brutlag D L. Highly specific protein sequence motif for genome analysis. Proc Natl Acad Sci USA, 1998, **95** (11): 5865~5874

17 Fitch W. Distinguishing homologous from analogous proteins. Syst Zool, 1970, **19** (2): 99~113

18 Eisen J. A phylogenomic study of the MutS family of proteins. Nucleic Acids Res, 1998, **26** (18): 4291~4300

19 Page R, Charleston M A. From gene to organismal phylogeny: reconciled trees and the gene tree/species tree problem. Mol Phylogenet Evol, 1997, **7** (2): 231~240

20 解 涛, 陈 洁, 丁达夫. 基因组功能预测的进化印记方法. 生物化学与生物物理学报, 1999, **31** (4): 433~439
Xie T, Chen J, Ding D F. Acta Biochimica et Biophysica Sinica, 1999, **31** (4): 433~439

21 Reinhardt A, Hubbard T. Using neural networks for prediction of the subcellular location of proteins. Nucleic Acids Res, 1998, **26** (9): 2230~2236

22 Andrade M, Donoghue S I, Rost B. Adaptation of protein surfaces to subcellular location. J Mol Biol, 1998, **276** (2): 517~525

23 Mirny L A, Shakhnovich E I. Protein structure prediction by threading: Why it works and why it does not. J Mol Biol, 1998, **283** (2): 507~526

24 陈 洁, 汤海旭, 丁达夫. 用于蛋白质分子设计的三维模体搜索, 生物物理学报, 1998, **13** (4): 639~646
Chen J, Tang H X, Ding D F. Acta Biophysica Sinica, 1998, **13** (4): 639~646

25 Marchler-Bauer A, Addess K J, Chappey C, *et al*. MMDB: Entrez's 3D structure database. Nucleic Acids Res, 1999, **27** (1): 240~243

26 Humphery-Smith I, Cordwell S J, Blackstock W P. Proteome research: Complementarily and limitations with respect to the DNA and RNA worlds. Electrophoresis, 1997, **18** (8): 1217~1242

27 Patel V, Corbett J, Dunn J, *et al*. Protein profiling in cardiac tissue in response to the chronic effects of alcohol. Electrophoresis, 1997, **18** (15): 2788~2794

28 Deloukas P, Schuler G D, Gyapay G, *et al*. A physical map of 30000 human genes. Science, 1998, **282** (5389): 744~746

29 Karp P, Riley M, Paley S M, *et al*. Ecocyc: encyclopedia of *Escherichia coli* genes and metabolism. Nucleic Acids Res, 1999, **27** (1): 55~58

30 Bono H, Ogata H, Goto S, *et al*. Reconstruction of amino acid biosynthesis pathways from the complete genome sequence. Genome Res, 1998, **8** (3): 203~210

Genome Annotation in the Postgenome Era. XIE Tao, LIANG Wei-Ping, DING Da-Fu (*Shanghai Institute of Biochemistry, The Chinese Academy of Sciences, Shanghai* 200031, China).

Abstract Genome annotation using bioinformatics tools becomes one of the most active research fields in the postgenome era. The new development in this area was reviewed. Various levels of genome annotation are discussed, while predicting protein function in genome scale is well discussed.

Key words postgenome, genome, genome annotation