蛋白基本注释

各数据库比对

将样本物种的蛋白序列与公共数据 gene 进行比较，通过 gene 的相似性进行功能注释。基因相似性比对主要基于 BLAST 算法。BLAST，全称 Basic Local Alignment Search Tool，即"基于局部比对算法的搜索工具"，由 Altschul 等人于 1990 年发布。Blast 能够实现比较两段核酸或者蛋白序列之间的相似性的功能，它能够快速的找到两段序列之间的相似序列并对比对区域进行打分以确定相似性的高低。将蛋白序列分别与 KOG、Swissprot、TrEMBL、GO、KEGG 库进行比对，取相似度>30%，且 e<1e-5 的注释，合并基因得到的所有注释详细信息。

各数据库说明如下：

KOG/COG: COG 是 Clusters of Orthologous Groups of proteins 的简称，KOG 为 euKaryotic Ortholog Groups。这两个注释系统都是 NCBI 的基于基因直系同源关系，其中 COG 针对原核生物，KOG 针对真核生物。COG/KOG 结合进化关系将来自不同物种的同源基因分为不同的 Ortholog 簇，目前 COG 有 4873 个分类，KOG 有 4852 个分类。来自同一 ortholog 的基因具有相同的功能，这样就可以将功能注释直接继承给同一 COG/KOG 簇的其他成员。详见 http://www.ncbi.nlm.nih.gov/COG/。

Swiss-Prot（A manually annotated and reviewed protein sequence database）搜集了经过有经验的生物学家整理及研究的蛋白序列。详见 http://www.ebi.ac.uk/uniprot/。

KEGG 是 Kyoto Encyclopedia of Genes and Genomes 的简称，是系统分析基因产物和化合物在细胞中的代谢途径以及这些基因产物的功能的数据库。它整合了基因组、化学分子和生化系统等方面的数据,包括代谢通路（KEGG PATHWAY）、药物（KEGG DRUG）、疾病（KEGG DISEASE）、功能模型（KEGG MODULE）、基因序列（KEGG GENES）及基因组（KEGG GENOME）等等。KO（KEGG ORTHOLOG）系统将各个 KEGG 注释系统联系在一起，KEGG 已建立了一套完整 KO 注释的系统，可完成新测序物种的基因组或转录组的功能注释。详见 http://www.genome.jp/kegg/。

GO(Gene Ontology)是一套国际标准化的基因功能描述的分类系统。GO 分为三大类 ontology：生物过程（Biological Process）、分子功能（Molecular Function）和细胞组分(Cellular Component)，分别用来描述基因编码的产物所参与的生物过程、所具有的分子功能及所处的细胞环境。GO 的基本单元是 term，每个 term 有一个唯一的标示符（由“GO:”加上 7 个数字组成，例如 GO:0072669）；每类 ontology 的 term 通过它们之间的联系（is_a, part_of, regulate）构成一个有向无环的拓扑结构。详见 http://www.geneontology.org/。