KEGG简介

　　KEGG(Kyoto Encyclopedia of Genes and Genomes)是系统分析基因功能、基因组信息数据库，它有助于研究者把基因及表达信息作为一个整体网络进行研究。基因组信息存储在GENES数据库 里，包括完整和部分测序的基因组序列；更高级的功能信息存储在PATHWAY数据库里，包括图解的细胞生化过程如代谢、膜转运、信号传递、细胞周期，还包 括同系保守的子通路等信息；KEGG的另一个数据库是LIGAND，包含关于化学物质、酶分子、酶反应等信息。可以免费获取。KEGG提供的整合代谢途径 (pathway)查询十分出色，包括碳水化合物、核苷、氨基酸等的代谢及有机物的生物降解，不仅提供了所有可能的代谢途径，而且对催化各步反应的酶进行 了全面的注解，包含有氨基酸序列、PDB库的链接等等。KEGG是进行生物体内代谢分析、代谢网络研究的强有力工具。

　　KEGG简介
　　KEGG简介一、KECC概况（全基因组及代谢途径数据库）

　　虽然决定生物体基因分类的基因组测序工程有了飞速的发展，但对单个基因功能的研究仍然相差甚远。同时活细胞的生物学功能是许多分子相互作用的结果，不能仅仅归功于单个基因或单个分子。日本教育、科学、体育、文化部人类基因组计划于1995年5月建立了KEGG工程。KEGG将基因组信息和高一级的功能信息有机地结合起来，通过对细胞内已知生物学过程的计算机化处理和将现有的基因功能解释标准化，对基因的功能进行系统化的分析。KEGG的另一个任务是一个将基因组中的一系列基因用一个细胞内的分子相互作用的网络连接起来的过程，如一个通路或是一个复合物，通过它们来展现更高一级的生物学功能。

　　其目的是由细胞或生物体的基因组信息去了解其较高层次的功能与作用之生物信息资源，也就是整理出现存的调控网络，并建立其中每个组件与基因间的关系，一但研究者找到 基因即可透过KEGG，让研究者由基因组至细胞层次做一整合性连结，并对生命现象做in silico 分析。虽然 KEGG 的工作受到肯定，但距离理想目标还有一大段距离，因为反应路径图上的每一个关系都应有文献资料做后盾，我们不能只相信一张可能带有个人偏见的路径图。目前 数据库中虽已建立各基因与其它数据库的关系，但它缺少的是讨论到路径正确性或调控机制的数据。

　　KEGG是基因组破译方面的数据 库。在后基因时代一个重大挑战是如何使细胞和有机体在计算机上完整的表达和演绎，让计算机利用基因信息对更高层次和更复杂细胞活动和生物体行为做出计算推 测。为达到此目的，人们建立了一个在相关知识基础上的网络推测计算工具。在给出染色体中一套完整的基因的情况下，它可以对蛋白质交互（互动）网络在各种细 胞活动起的作用做出预测。

　　二、KEGG的组成

　　KEGG现在由6个各自独立的数据库组成，分别是基因数据 库(GENES database)、通路数据库(PATHWAY database)、配体化学反应数据库(NGAND database)、序列相似性数据库(SSDB)、基因表达数据库(EXPRESSION)、蛋白分子相互关系数据库(BRITE)等。
　　1.基因数据库(GENES database) 含有所有已知的完整的基因组和不完整的基因组。有细菌、蓝藻、真核生物等生物体的基因序列，如人、小鼠、果蝇、Arabidopsis等等。总共240943个条目，其总的条目数远远大于SWISS—PROT数据库。基因数据库(GENES database)含有关于每个基因的最低限度的信息，并且在不断地更新和改进，同时还可作为通往其他相关信息的路径。

　 　2.通路数据库(PATHWAY database) 储存了基因功能的相关信息，通过图形来表示细胞内的生物学过程，例如代谢，膜运输，信号传导和细胞的生长周期。在通路数据库(PATHWAY database)中，有一部分由ortholog group图表组成的保守的亚通路(通路基序)信息。亚通路是由染色体位置偶联的基因编码的，它对预测基因的功能有很大的作用。

　　(1)代谢通路：目前在通路数据库(PATHWAY database) 中代谢通路是建立得最好的，有大约90个参考代谢途径的图形。每个参考代谢途径是一个由酶或EC号组成的网络。利用如下方法可通过计算机构建出生物体特有 的代谢通路：先根据基因的序列相似性和位置相关性确定基因组中酶的基因，然后合理地安排EC号，最后将基因组中的基因和参照通路中用EC号编号的基因产物 结合起来。图4—5—19为通路数据库的页面尔例。

　　(2)Ortholog group图表：在KEGG中目前有83个ortholog图表。Ortholog对单个基因的序列的相似性进行识别，检查功能组(如保守的亚通路或分子 复合物)中所有组分。KEGG ortholog grope图表表达了三个特征：①一个生物体是否具备组成一个功能组的完整基因；②这些基因是否偶联在染色体上；③在不同生物体中的 orthologous基因是什么。例如在基因组中的一个基因簇编码代谢通路中的功能相关的酶簇。在KEGG中，这样的相关簇首先被一个启发式的图表比较 算法检测，然后手工编辑为ortholog group图表。现有两种类型的图表比较法，基因组—通路和基因组—基因组比较法。一个ortholog grope图表是这样一组比较的组成部分，它代表了一个通路的保守的部分，也就是通常说的通路基序。

　　(3)蛋白—蛋白反应：KKGG通路表达的重点在于由基因产物的构成的网络上，其中包含大多数蛋白和功能性RNAs。代谢通路是蛋白—蛋白的间接相互作用(实际上是酶—酶相互作用)形成的网络。而调节通路是蛋白—蛋白的直接相互作用(如结合，磷酸化)和另一种蛋白—蛋白的间接相互作用(通过基因表达与转录因子及基因的翻译产物相关联)构成的网络。一般的蛋白—蛋白的相互作用包括了上面所述的这三种形式的相互作用，它是一个抽象的网络，但是它在与基因组信息的连接中起到关键作用，这样网络中的节点(基因产物)与基因织中的节点(基因)就可以直接相连。有了这样一个蛋白—蛋白相互作用的网络，就可以增加手工绘制的参考通路图了。

　　3.配体数据库(LIGAND database) 包括了细胞内的化学复合物，酶分子和酶反应的信息。

　　三、KEGG的使用

　　KEGG提供了java的图形工具用于浏览基因组图谱，比较两个基因组图谱，操作表达图谱，还可作为比较序列、图表、通路的计算工具。

　 　KEGG需要各种各样的计算工具用来维护基因数据库(GENES database)，尤其是从GenBank中提取信息和对基因功能的系统化解释。网络注释工具和其他计算机工具一起用来分配EC号，ortholog识 别符，合并文献中的新的实验证据，并且对以通路结构为基础的推断做出解释。Ortholog识别号可以作为查找工具，自动比较通路基因组和基因产物的基 因。

　　GENES的主要检索系统是DBGET／LinkDB系统，另外也有其他进入数据库的办法。包括Java虚拟的基因组图谱浏 览器和文件分层浏览器(用于将基因目录进行功能性分层)。表达浏览器是Java图形浏览器中的一种，它可以分析从cDNA微序列或寡核苦酸序列实验中得到 的基因表达文件。从这样的功能性基因组实验中得到的大量数据将对基因组序列进行补充，这样有助于理解更高一级的细胞的生物学功能。利用与KEGG的通路数 据和基因组图谱数据相连接的一个表达图谱浏览器的预备版本，用户可以检查一组共同调节的基因是否在通路上也有相互联系或是否由染色体上的一群基因编码。

　　四、KEGG的Object Identifier
　　本部分内容设定了隐藏,需要回复后才能看到
　　五、例子

　　KEGG中每一个Object（除基因外）都含有一个唯一的KEGG标识符，它包括一个5位数的号码并有一个大写字母作为前缀，如K05032和D00336 ，或由一个有2-4的字母代码开头的标识符，如map00010和br08301 。

　　每个KEGG的数据库资料都配有一个独特的标识符，如下图所示：

　　Release Database Object Identifier
　　1995 KEGG PATHWAY map number
　　 KEGG GENES locus\_tag / GeneID
　　 KEGG ENZYME EC number
　　 KEGG COMPOUND C number
　　2000 KEGG GENOME organism code / T number
　　2001 KEGG REACTION R number
　　2002 KEGG ORTHOLOGY K number
　　2003 KEGG GLYCAN G number
　　2004 KEGG RPAIR A number
　　2005 KEGG BRITE br number
　　 KEGG DRUG D number
　　2007 KEGG MODULE M number
　　 KEGG DISEASE H number

　 　下面我以一个例子来介绍怎样使用pathway在KEGG上找有关氨基酸代谢的代谢图。首先打开KEGG PATHWAY,找到 Amino Acid Metabolism，然后找相应的氨基酸代谢途径就可以了。如点击Glutamate metabolism就进入了其代谢图，我们可以点击Pathway entry获得更多的信息。

　　六、展望

　 　现在，KEGG的通路数据库(PATHWAY database)中关于调节通路的部分和代谢通路相比还相差较多。这是因为代谢通路，特别是中间代谢，在大部分生物(从哺乳动物到细菌)中都是保守的。 这样就可以徒手画出一条参考通路，然后用计算机构建出许多生物体特异的通路来。但是调节通路非常多样化，而且很难归为一个统一的参考通路图。因此需要给每 一个生物分别画一个通路图。同时还需要鉴别有共同通路或集群的生物群体以及通路图可以合并起来的生物群体。例如人和小鼠共同的凋亡的通路图表，细菌、 archaea和真核生物的三种核糖体的集群图表。

　　另一个在调节通路中存在的问题是缺乏合适的调节通路中的功能识别符。在代谢通 路中，EC号是节点(酶)的识别符，同时它也是与基因组信息连接的关键。KEGG目前致力于使ortholog识别符具有EC号的功能。这样 ortholog识别符将可以识别调节通路中的节点(蛋白)同时与基因组信息相连接。随着不断的改进，ortholog识别符将取代代谢通路中的EC号， 这样就可以区分对应一个EC号的多个基因。例如：一个酶复合物的不同亚单位或是在不同条件下基因的不同表达。