北京基因组所等开发完成基于K-mer的基因组组分分析数据库

文章来源：北京基因组研究所 发布时间：2015-11-02 【字号： 小 中 大 】

在过去的几十年中，人们往往使用高度保守的基因家族进行系统进化分析，采用全基因组序列进行系统进化分析并不普遍。目前，基于是否进行序列的比对，分子系统发生树的构建分为两类。其中，不需要进行序列比对的方法是依据K-mer向量计算的距离矩阵进行系统进化分析，大量的研究证实该算法是行之有效的，尤其是对基因组中诸如蛋白编码序列等的特定区域。不仅如此，K-mer算法还在组学的其他方面，包括基因组组装、motif预测、重复序列的识别以及基因组的复杂性评估等都受到了广泛的关注。基于K-mer算法在组学中的重要表现，在这个大规模基因组数据快速积累的时代，构建一个基于K-mer算法易于存储并且将大量基因组数据可视化处理的数据库十分迫切。

为此，中国科学院北京基因组研究所基因组科学与信息重点实验室于军组和英国伦敦大学学院（UCL）肿瘤研究所王大鹏合作开发了一套基于K-mer算法的基因组组分分析数据库KGCAK。此项研究于近期发表在Biology Direct 杂志。

在这个数据库中，研究人员搜集了Ensembl、Phytozome和NCBI等几大主流基因组数据库中包括高等动植物、原生生物、真菌、细菌、病毒等在内的8000多个核基因组或者细胞器基因组，同时包括基因组不同维度的序列，主要有DNA、cDNA、CDS、氨基酸和ncRNA序列。并且还分别计算和存储了核酸序列（K从2变化到10）和氨基酸序列（K从1变化到5）的K-mer向量，以方便进行不同维度数据跨物种的系统发生树构建。此外，该数据库提供了评估不同物种基因组复杂度的交互工具，主要包括基因组基本特征参数、K-mer向量的数学参数统计、频率分布、唯一性比率，以及二维和三维空间可视化分析基因组参数和K-mer参数的交互关系等。

总的来说，该数据库通过捕获基因组序列特征并把基因组转化成更易于理解和可视化的数字K-mer向量，以期通过K-mer算法用可视化的图形和定量的数据构建一个比较基因组学的平台，将为系统发生树构建和通过基因组数据研究物种关系提供良好的参照和指引。

文章链接

热点新闻

中科院与广东省签署合作协议 ...

发展中国家科学院中国院士和学者代表座...

白春礼在第十三届健康与发展中山论坛上...

中科院江西产业技术创新与育成中心揭牌

中科院西安科学园暨西安科学城开工建设

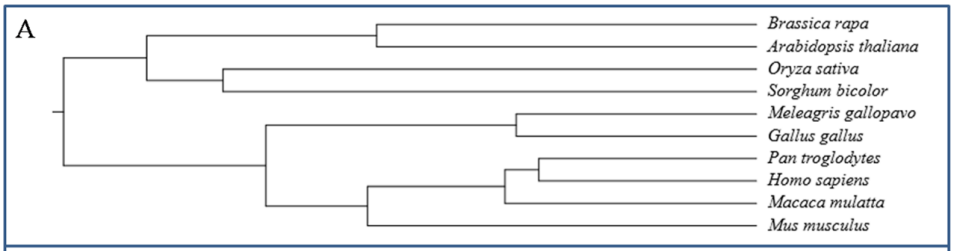
中科院与香港特区政府签署备忘录

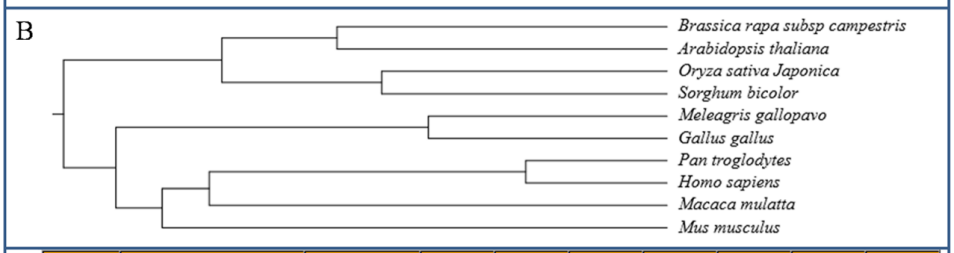
视频推荐

【新闻联播】“率先行动”计划 领跑科技体制改革

【新闻直播间】我科学家造血干细胞研究获新突破

专题推荐



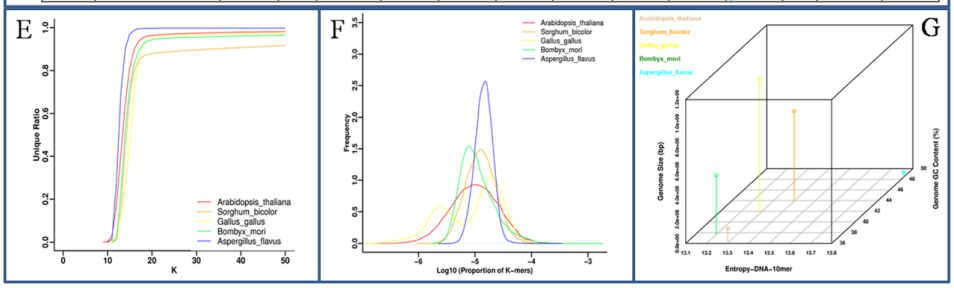


C

Location	Species	Genome Size (bp)	A Content	C Content	G Content	T Content	N Content	GC Content	Purine Content
Nuclear	Arabidopsis_thaliana	121662572	31.99%	18.04%	18.02%	31.95%	1.79%	36.06%	50.01%
Nuclear	Sorghum_bicolor	751053697	28.04%	21.96%	21.97%	28.03%	7.12%	43.93%	50.01%
Nuclear	Gallus_gallus	1118825381	29.21%	20.77%	20.77%	29.24%	6.81%	41.54%	49.99%
Nuclear	Bombyx_mori	491626853	31.15%	18.85%	18.83%	31.17%	12.18%	37.68%	49.98%
Nuclear	Aspergillus_flavus	40836880	25.80%	24.24%	24.19%	25.77%	2.27%	48.43%	50.00%

D

Location	Species	Sequence Type	K-mer	Number of K-mer	Maximum	Index of Maximum	Mean	Standarnd Deviation	Median	Lower Quantile	Upper Quantile	Ratio of Zero	Information Entropy	Distance to Even
Nuclear	Arabidopsis_thaliana	cDNA	5mer	64629149	4.28E-03	698	9.77E-04	5.71E-04	8.65E-04	5.58E-04	1.27E-03	0	6.78E+00	3.34E-04
Nuclear	Sorghum_bicolor	cDNA	5mer	48908462	3.96E-03	471	9.77E-04	4.59E-04	9.20E-04	6.38E-04	1.20E-03	0	6.83E+00	2.16E-04
Nuclear	Gallus_gallus	cDNA	5mer	39507084	3.50E-03	942	9.77E-04	5.66E-04	8.97E-04	5.30E-04	1.26E-03	0	6.77E+00	3.28E-04
Nuclear	Bombyx_mori	cDNA	5mer	17675380	3.75E-03	698	9.77E-04	3.94E-04	8.99E-04	7.37E-04	1.11E-03	0	6.86E+00	1.59E-04
Nuclear	Aspergillus_flavus	cDNA	5mer	17379768	2.64E-03	942	9.77E-04	3.30E-04	9.95E-04	7.85E-04	1.18E-03	0	6.87E+00	1.11E-04



KGCAK数据库中基本功能模块举例

（责任编辑：叶瑞优）