Bioinformatics Databases — Resource and Application

生物信息数据库—资源和应用

百迈客第二届全国功能基因组峰会

2015年11月11-12日 北京

罗静初

北京大学生命科学学院

北京大学生物信息中心

luojc@pku.edu.cn

http://abc.cbi.pku.edu.cn/

汇报提纲

- 国际生物信息数据库
- 国内生物信息数据库
- 北大生物信息中心数据库
- 植物转录因子数据库
- 数据库应用实例

Bioinformatics Databases

- Can fill the gap between bioinformaticians and wet-lab biologists
- Can turn the omics data into knowledge
- Can make a bridge from big data to big discoveries

Historical Overview

- The 1st biological database Protein Data Bank (PDB) was started in 1977.
- The two DNA sequence databases EMBL and GenBank were founded in early 1980’.
- The two protein sequence databases PIR and Swiss-Prot were built in middle 1980’.

核酸序列数据库

- 1979年，美国科学基金会NSF召集会议，对建立核酸序列数据库达成共识。
- 1980年，欧洲分子生物学实验室核酸序列数据库EMBL正式宣告诞生，1982年6月，EMBL第1版正式对外发布。
- 1982年，Walter Goad等竭力推动，美国国家健康研究院NIH、科学研究基金会NSF、能源部DOE和国防部DOD等共同资助，创建核酸序列数据库GenBank，洛斯阿拉莫斯国家实验室负责运行。1987-1992年，由InteliGentics公司负责分发。
- 1992年起，美国国家生物技术信息中心NCBI接管GenBank，负责核酸序列收集、储存、管理、注释、分发，并开发基于网络浏览器的数据库检索系统。
- 1986年，日本国立遗传研究所NIG建立日本DNA数据库DDBJ，并和GenBank、EMBL共同成立国际核酸序列数据库协会INSDC。

蛋白质序列数据库

- 1984年，美国国家生物医学基金会NBRF Winona Barker和Robert Ledley获NIH资助，建立了蛋白质序列数据库PIR；1988年，NBRF和德国、日本联合成立了国际蛋白质序列数据库。
- 1986年，瑞士日内瓦大学Amos Bairoch创建了蛋白质序列数据库Swiss-Prot，该数据库具有大量注释信息和交叉链接。
- 1995年，欧洲生物信息学研究所Rolf Apweiler创建了蛋白质序列数据库TrEMBL，收集从EMBL翻译得到的蛋白质序列。
- 2003年，Swiss-Prot、TrEMBL和PIR合并，建立了国际蛋白质知识库UniProt，统一收集、管理、注释、发布蛋白质序列数据。UniProt包括UniProtKB/UniRef/UniParc三部分和Swiss-Prot/TrEMBL两个子库。

美国国家生物技术信息中心NCBI

- 1988年11月，由已故参议员Claude Peper提议成立。位于华盛顿北郊马里兰州，隶属NIH下的NLM。成立初期仅8名工作人员，现已增加到500多名。
- 运用最新的计算机和信息技术，创建方便实用的生物信息存储和分析系统，开发先进的生物信息处理方法，整合国际公共数据库资源，为生物医学领域提供内容丰富、更新及时的生物信息资源。
- David Lipman任NCBI主任，2003年当选为美国科学院院士，2004年获ISCB颁发的Senior Accomplishment Award。2009年应邀参加在北京举行的亚太地区生物信息学大会，作关于流感病毒起源和演化的报告。2013年获白宫Open Science奖。

NCBI Databases

- PubMed – Biomedical literatures
- PMC – PubMed Central
- Bookshelf – Free online books
- GenBank – Nucleic acid sequences
- RefSeq – Reference sequences (DNA, RNA, Protein)
- CDD – Conserved Domain Database
- SRA – NGS sequence read archive
- Genome – Genomic sequences and annotations
- UniGene – Unique RNA transcripts and ESTs
- SNP – Single nucleotide polymorphism
- Taxonomy – Classification of biological species
- PubChem – Small molecules and drug compounds
- Flu – Influenza virus resources

欧洲生物信息学研究所EBI

- 成立于1994年，坐落在英国剑桥南部12英里Wellcome基金会基因组园区内。欧洲分子生物学实验室EMBL下属单位，研究人员主要来自英国、德国、法国等西欧各国。
- 仅次于NCBI的国际生物信息中心，为欧洲各国和世界各地用户提供生物信息资源服务，并从事生物信息研究开发。核酸序列数据库EMBL、蛋白质序列数据库UniProt和基因组数据库Ensembl由EBI负责管理发布。
- 第一任主任为剑桥大学果蝇遗传学家Michael Ashburner，Graham Cameran任副主任。2003年，著名英国生物信息学家Janet Thornton接任EBI主任，2011年，Rolf Apweiler（蛋白组学）和Ewan Birney（基因组学）任副主任。2015年，Rolf Apweile和Ewan Birney任共同主任。

EBI Databases

- ENA – European Nucleotide archive
- Ensembl – Genomic sequences and annotations
- Expression Atlas – Differential and Baseline Expression
- Array Express – NGS functional genomics experiments
- DGVa – Database of Genomic Variants archive
- TreeFam – Database of animal gene trees
- Rfam – Database of non-coding RNA families
- UniProt – Database of protein sequences
- InterPro – Classification of proteins families and domains
- Pfam – Collection of protein families and domains
- Pride – Proteome identification database
- IntAct – Database of molecular interaction
- PDBe – Macromolecular 3D structures
- PDBeChem – Chemical Components in the PDB

Database Journals

- Nuclei Acids Research (NAR) Database issue - Every January since 1996
- The Journal of Biological Databases and Curation (Database) - since 2009
- Bioinformatics - Application Notes
- BMC Bioinformatics

生物信息数据库专辑和专刊

- NAR – http://www.oxfordjournals.org/nar/database/c/
  Nucleic Acids Research杂志1982、1984、1986年第1期专集刊登分子生物学数据库以及序列分析等文章。1996年起，每年1月1日专辑刊登生物信息数据库论文，Michael Galperin任主编。
- Database – http://database.oxfordjournals.org/
  The Journal of Biological Databases and Curation
  《生物数据库及审编》，网络杂志，2009年开始出版，每年1卷，不分期，主编David Landsman，中国编委朱伟民。

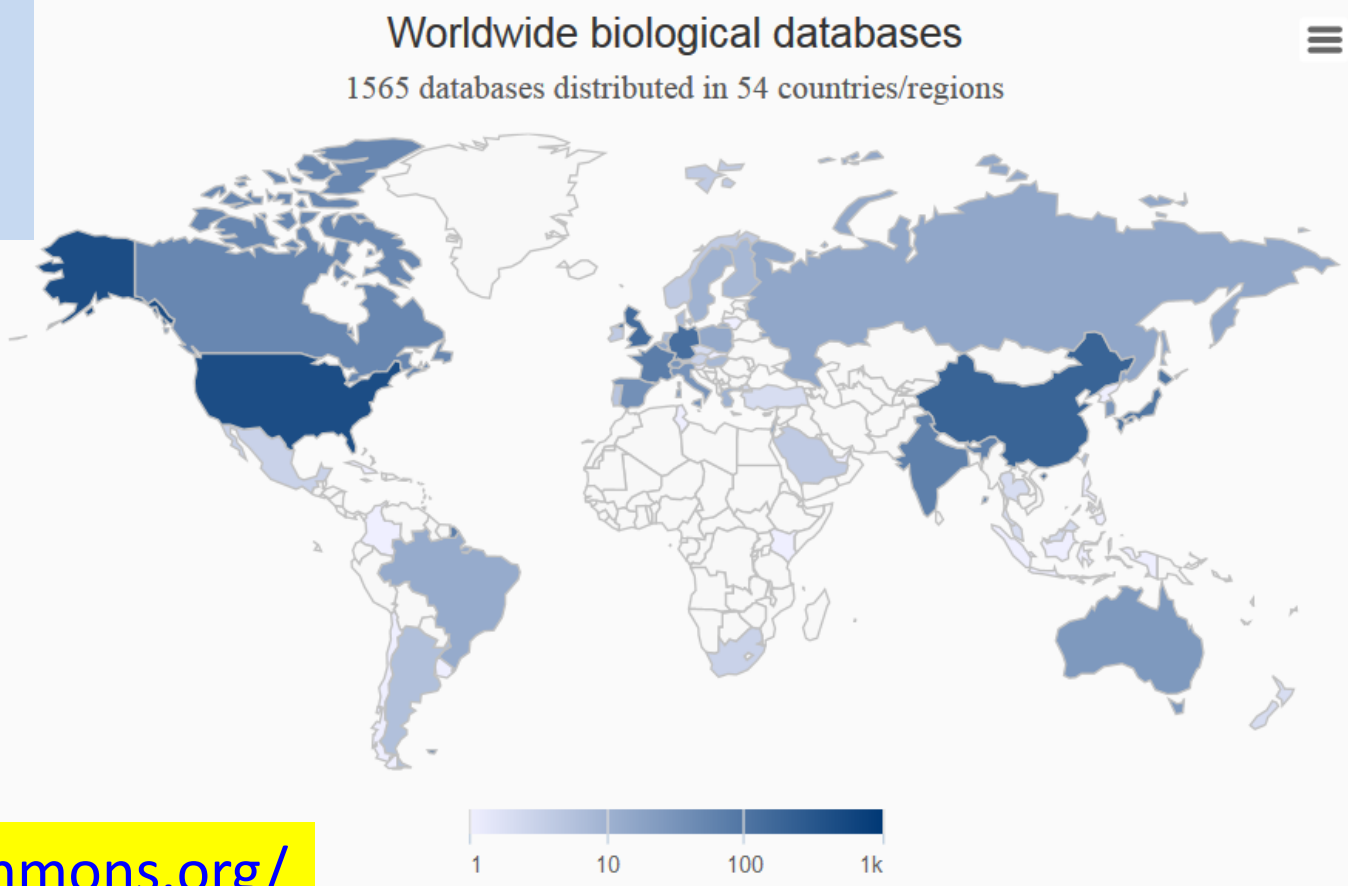
Database Commons a catalogue of biological databases

Home | Browse | Search | Statistics | Submit | Help | Database name

1565 databases
413 organisms
54 countries/regions
14 categories

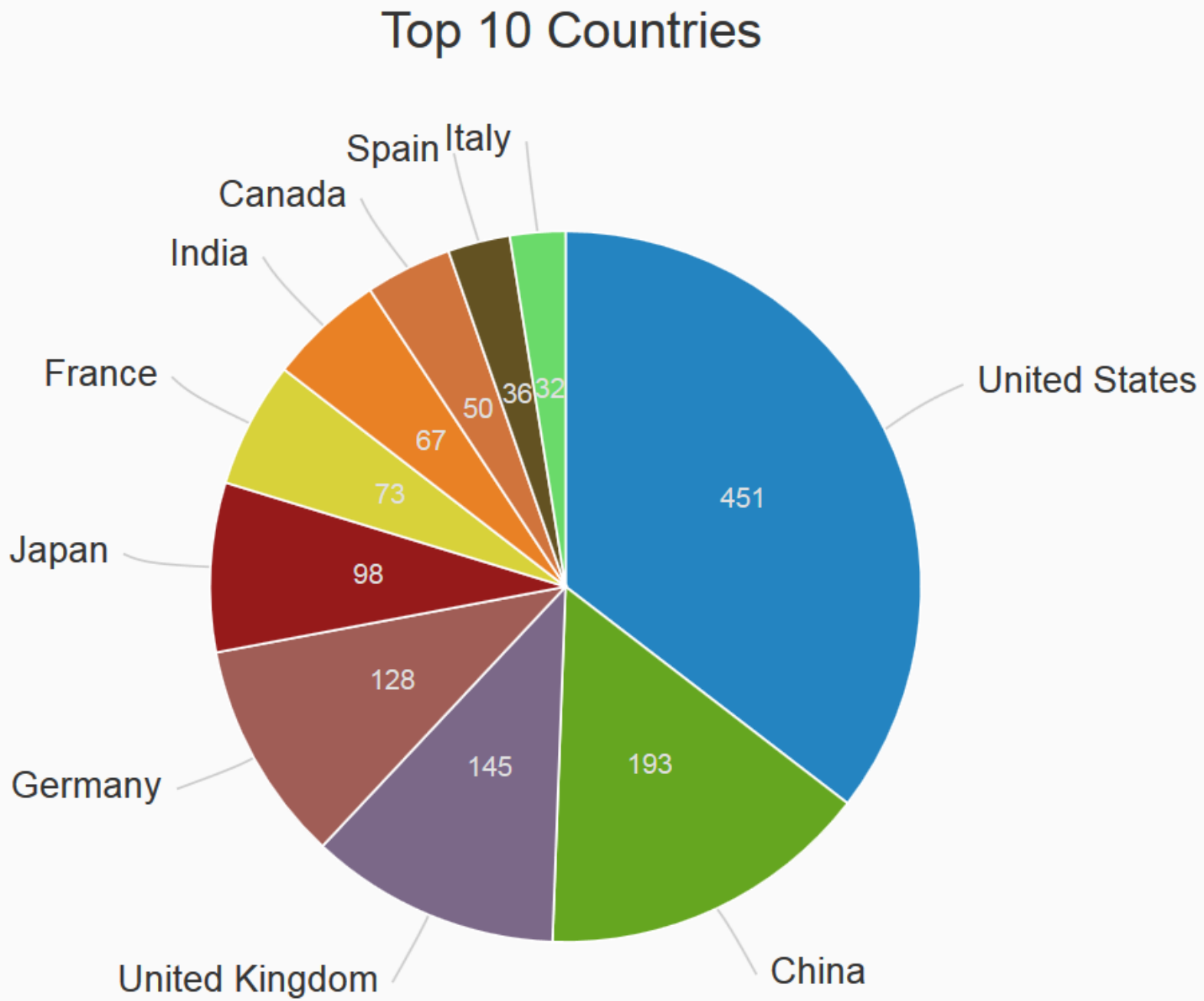
News & updates

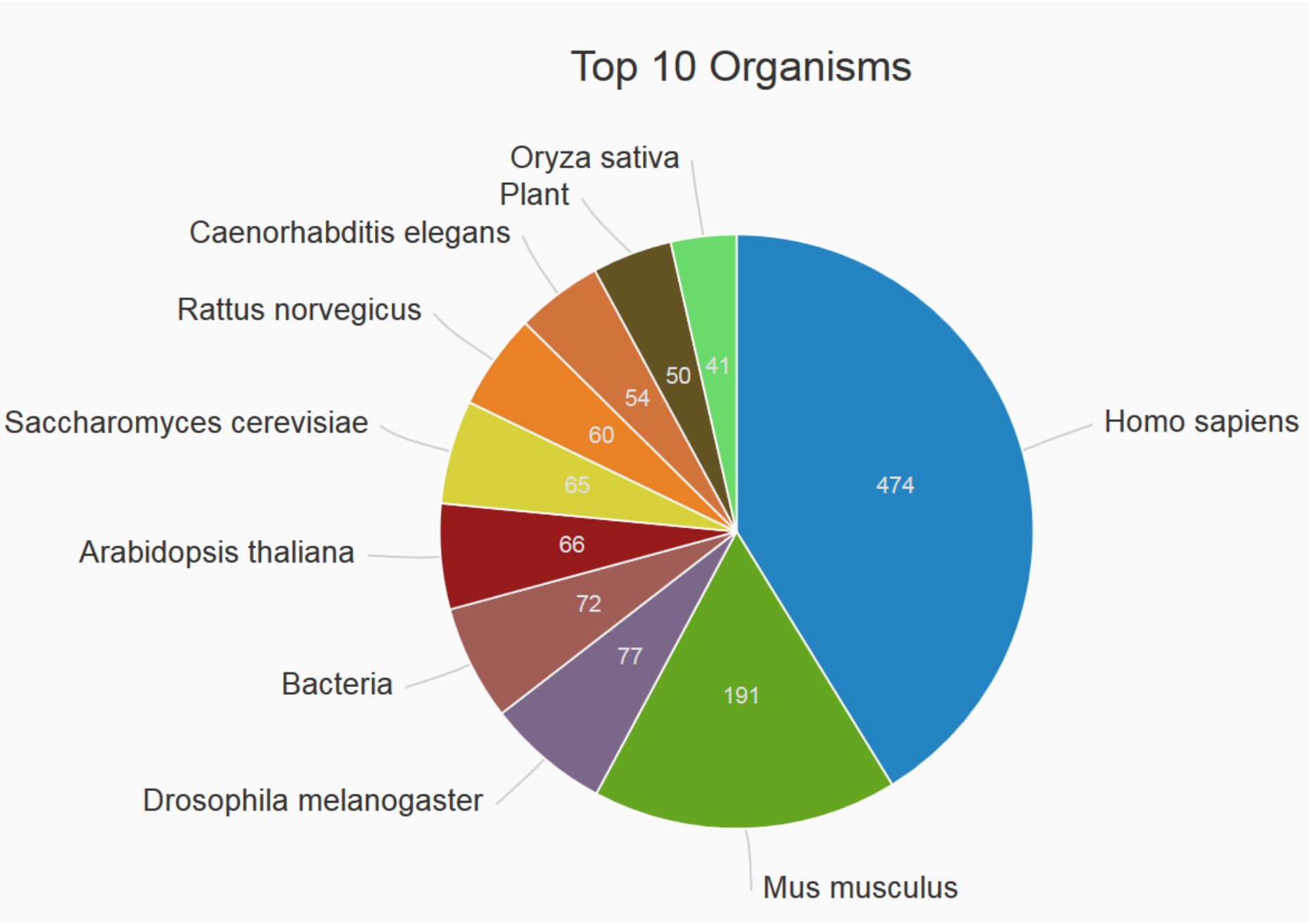
- Light-weight visualizations of worldwide databases
- More than 1000 databases incorporated
- Latest additions:
  - Epitome
  - TED
  - SoyGD
  - RMD
  - Plant MPSS databases

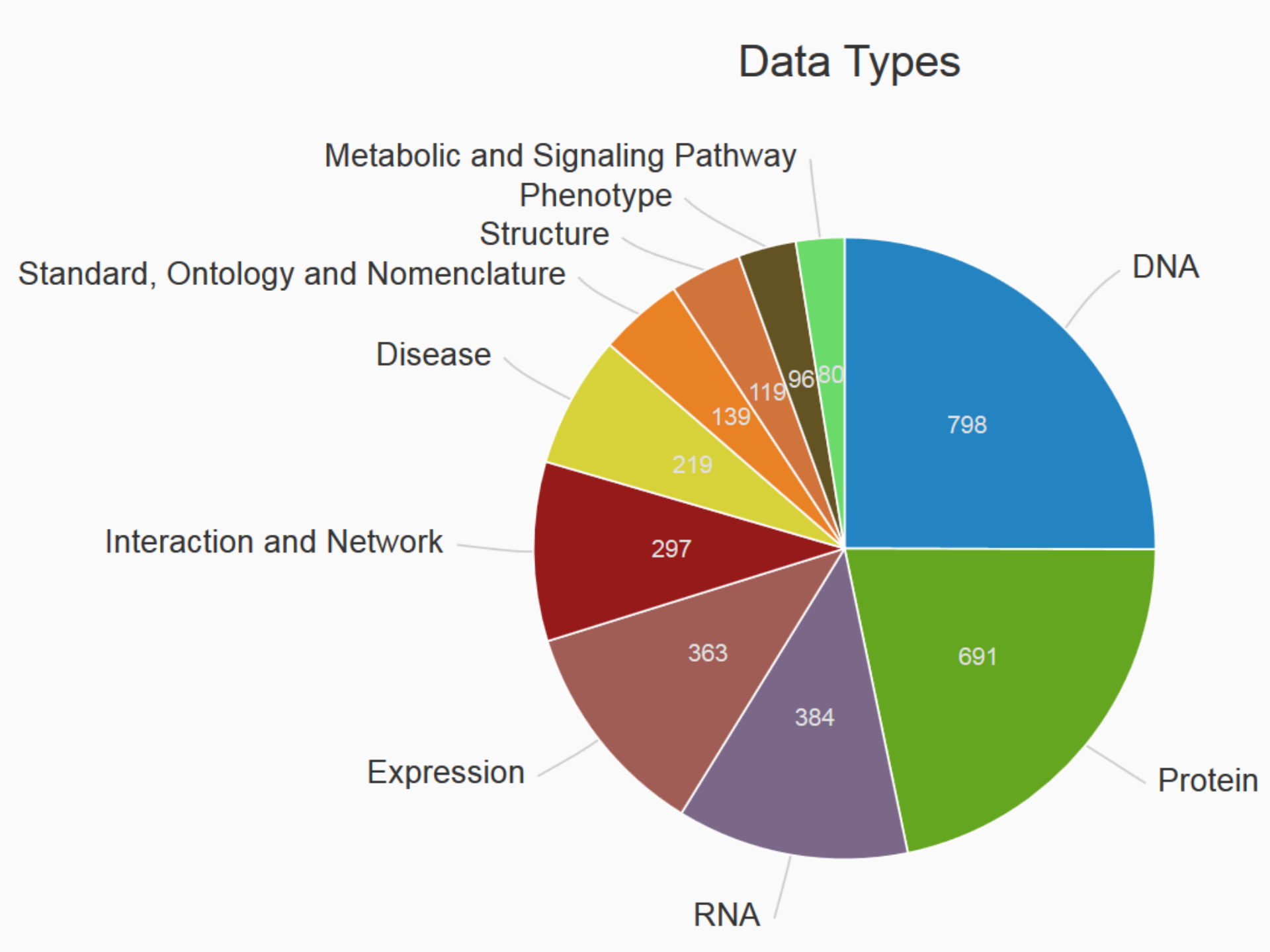


http://databasecommons.org/

Zhang Lab, Beijing Institute of Genomics, Chinese Academy of Sciences
中科院基因组所章张课题组


Top 10 Countries
United States
451
China
193
United Kingdom
145
Germany
128
Japan
98
France
73
India
67
Canada
50
Spain
36
Italy
32


Top 10 Organisms
Oryza sativa
Plant
Caenorhabditis elegans
Rattus norvegicus
Saccharomyces cerevisiae
Arabidopsis thaliana
Bacteria
Drosophila melanogaster
Mus musculus
Homo sapiens
41
50
54
60
65
66
72
77
191
474


Data Types
Metabolic and Signaling Pathway
Phenotype
Structure
Standard, Ontology and Nomenclature
Disease
Interaction and Network
Expression
RNA
Protein
DNA
798
691
384
363
297
219
139
119
96
80

我国生物信息数据库论文

- 总计416篇，其中发表在国内期刊上的27篇
- 2014和2015两年150篇，占总数36%
- 2011-2015五年内257篇，占总数62%
- 133篇发表在NAR数据库专辑上，54篇发表在Database专刊上，229篇发表在其它期刊上

Databases from China

- NAR Database papers from China
  - The PubMed list of database papers published on NAR by authors from China.
- Database papers from China
  - The PubMed list of papers published on the journal Database by authors from China.
- Other database papers from China
  - The PubMed list of database papers published on other journals by authors from China.

http://abc.cbi.pku.edu.cn/databases.php

国内部分生物信息数据库

简称	内容	网址
Z-Curve	Z曲线数据库	天津大学张春霆、高峰
CVTree	组分矢量细菌分类	复旦大学郝柏林、左光宏
NonCode	非编码RNA	生物物理所陈润生、计算所赵屹
BRAD	芸薹植物基因组	蔬菜所王晓武
DoGSD	犬类基因组SNP	昆明动物所张亚平、基因组所赵文明
SorGSD	高粱基因组SNP	植物所景海春、基因组所赵文明
rVarBase	人类基因组变异调控元件	心理所王晶
AnimalTFDB	动物转录因子数据库	华中科技大学郭安源

The Z-curve Database of Genomes

Home Browse Plotting Z-curve Gene recognition Identifying oriC Genome segmentation Special issue Search About

» Home » Information of Eukaryota

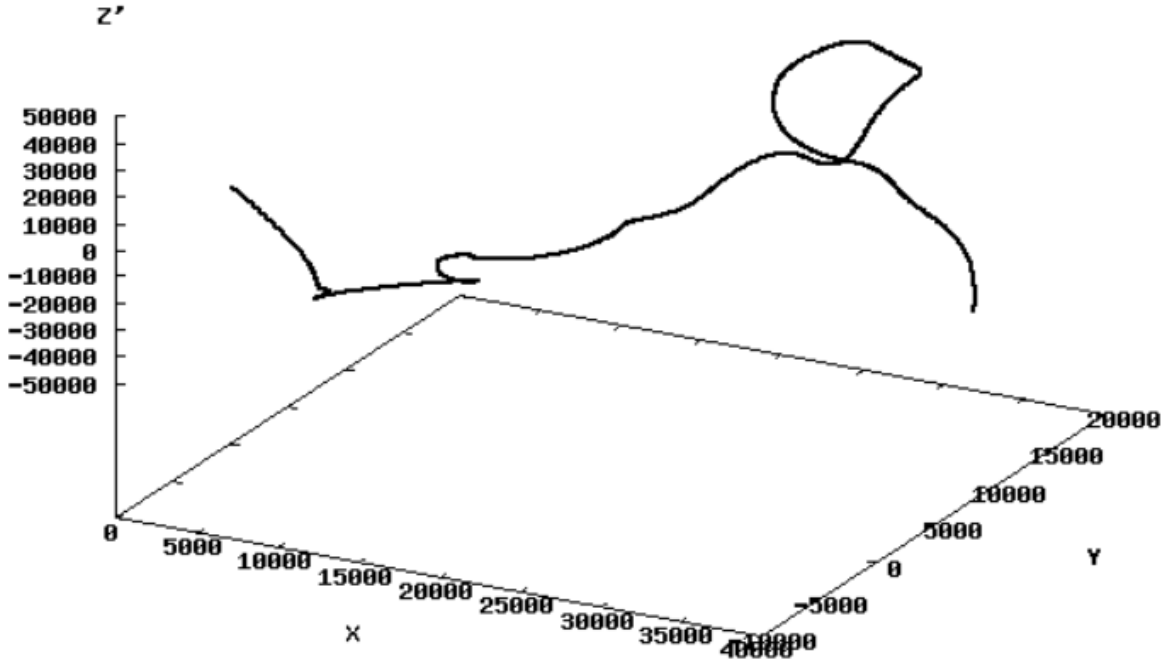
the Z curve

3D X Y Z Z' XY SRS segments

AC	AE005173
DE	Arabidopsis thaliana chromosome 1 bottom arm
SQ	Length: 14668883 bp; 4744006 A; 4718470 T; 2608943 G; 2597461 C; 3 N.

Arabidopsis thaliana chromosome 1 bottom arm, complete sequence.

http://tubic.tju.edu.cn/zcurve/



天津大学张春霆院士Z曲线数据库

CVTree3: Composition Vector Tree Version 3

>>> Previous Version

Description:

CVTree constructs whole-genome based phylogenetic trees without sequence alignment by using a Composition Vector (CV) approach. It was first developed to infer evolutionary relatedness of microbial organisms and then successfully applied to viruses, chloroplasts, and fungi. CVTree3 makes comparison with taxonomy and reports tree-branch monophyleticity from domain to species. Please read the *Online User's Manual* for details.

Reference:

Ji Qi, Bin Wang, Bailin Hao (2004) Whole proteome prokaryote phylogeny without sequence alignment: a K-string composition approach, J Mol Evol, 58: 1 –11

Guanghong Zuo, Bailin Hao (2015) CVTree3 web server for whole-genome-based and alignment-free prokaryotic phylogeny and taxonomy, Genomics Proteomics & Bioinformatics, being submitted.

Load/Create Project:

Enter Project Number to reload a previously completed project for checking or changing parameters and re-run. A blank input creates a new project. A project will be kept for 7 days after the last run.

Load/Create Project | Example

Announcement:

CVTree3 web server is under final testing. *Feedback and criticism* welcome!

复旦大学郝柏林院士CVtree数据库

http://tlife.fudan.edu.cn/archaea/cvtree/cvtree3/cvtree.html

生物数据库审编Biocuration

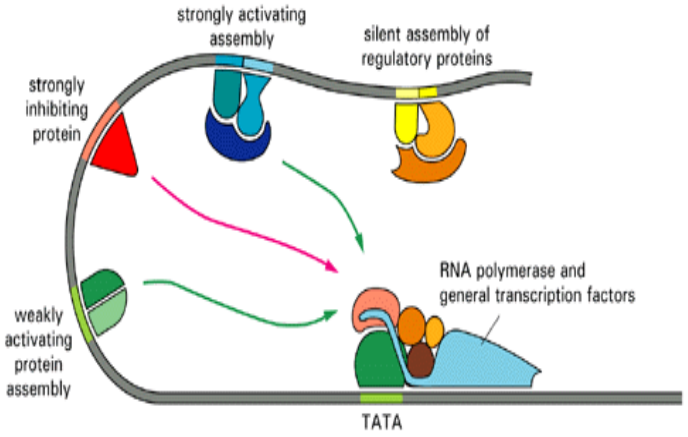
- International Society for Biocuration (ISB) http://www.biocurator.org/
- Alex Bateman为执委会主席
- 2015年11月，中科院基因组所章张当选执委会委员
- 2015年4月，第8届ISB国际会议在北京召开，中国生物审编学会成立，于军当选主席
- 2016年4月，第9届ISB国际会议将在瑞士日内瓦召开

北京大学生物信息中心部分数据库

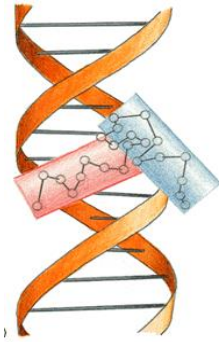
简称	内容	网址
PlantTFDB	植物转录因子数据库	http://planttfdb.cbi.pku.edu.cn/
LSD	植物叶片衰老数据库	http://psd.cbi.pku.edu.cn/
AHD	拟南芥激素及相关基因数据库	http://ahd.cbi.pku.edu.cn/
SeedGeneDB	种子发育相关基因数据库	http://sgdb.cbi.pku.edu.cn/
SPD	哺乳动物分泌蛋白数据库	http://spd.cbi.pku.edu.cn/
AutismKB	孤独症相关基因数据库	http://autismkb.cbi.pku.edu.cn/
HomeoDB	同源异形框相关基因数据库	http://homeodb.cbi.pku.edu.cn/

转录调控是基因调控的主要机制

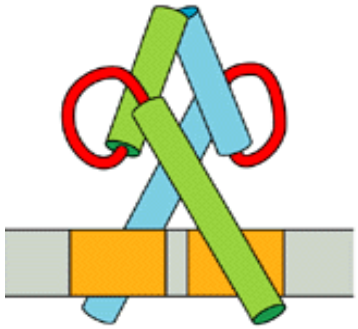
- 转录调控通过顺式作用元件和反式作用因子相互作用实现
- 反式作用因子通称转录因子
- 转录因子包括通用转录因子和特异转录因子
- 特异转录因子分为不同家族



螺旋-回折-螺旋
Helix-Turn-Helix

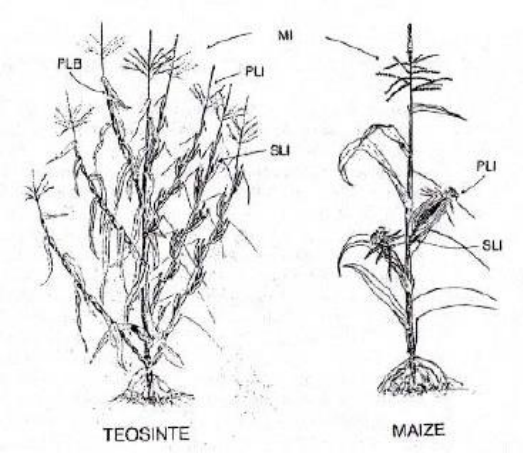


螺旋-回环-螺旋
Helix-Loop-Helix

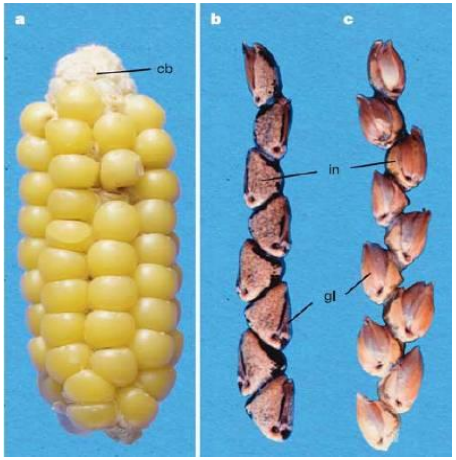


锌指结构
Zinc Finger

转录因子与多种农作物性状相关



Doebley *et al*.(1997)
Nature 386:485

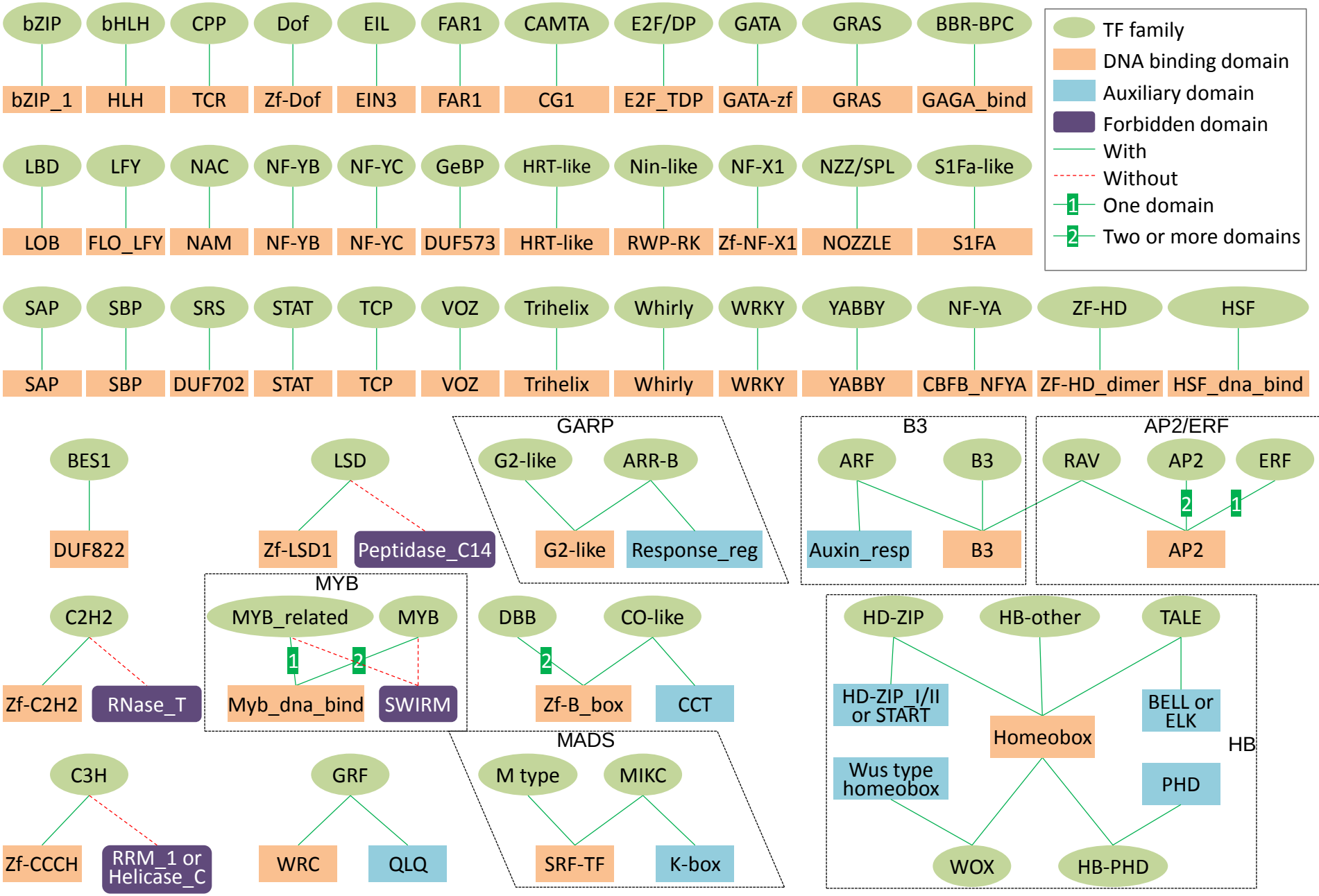


Wang *et al*. (2005)
Nature, 436:714

Manning *et al* (2006) Nature Genetics, 38:948

Kinishi *et al* (2006) Science, 312:1392
Li *et al* (2006) *Science*, 312:1936

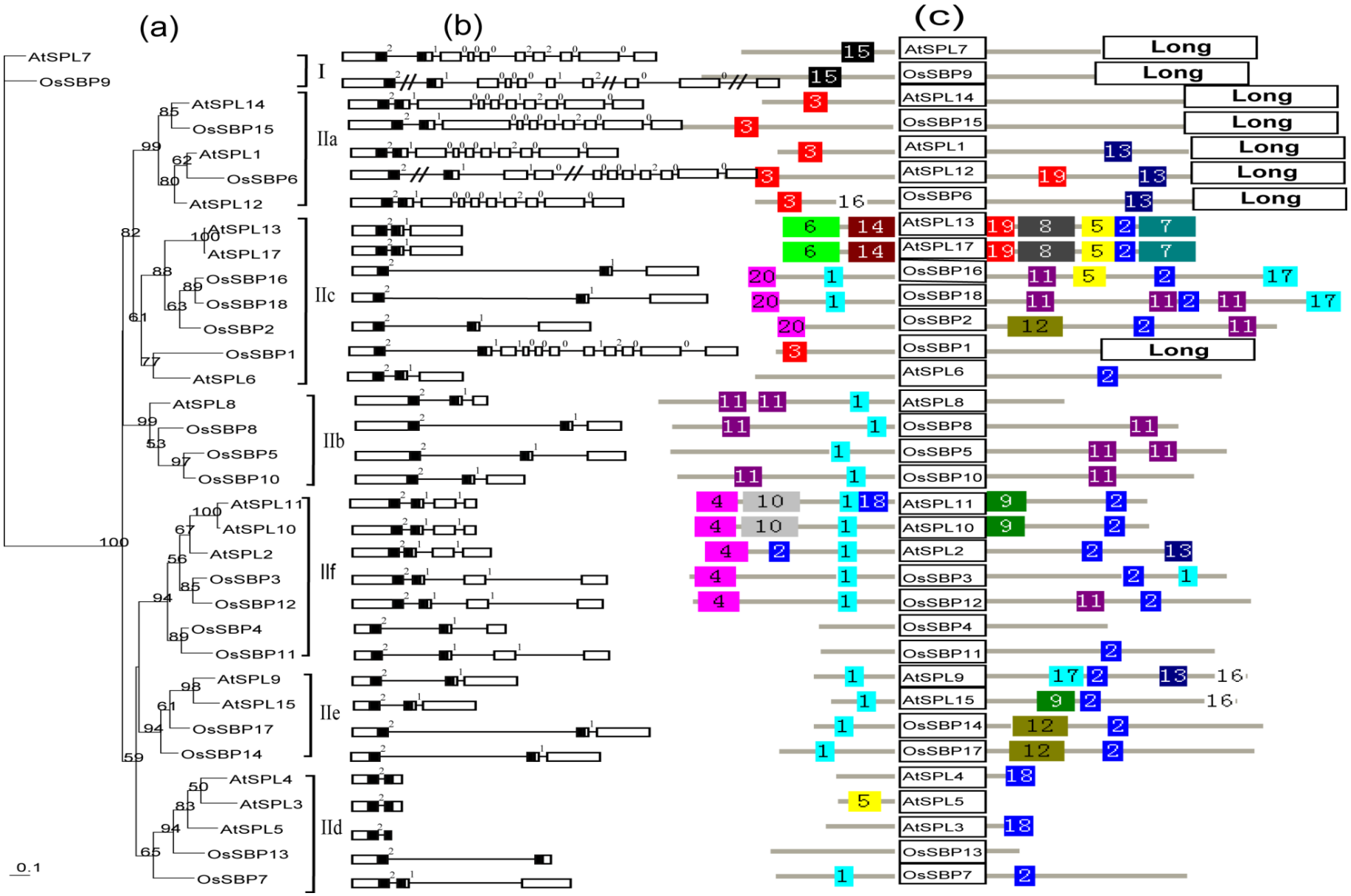
Jiao *et al* (2010) Nature Genetics, 42:541
Miura et al (2010) Nature Genetics, 42:545


bZIP
bHLH
CPP
Dof
EIL
FAR1
CAMTA
E2F/DP
GATA
GRAS
BBR-BPC
bZIP_1
HLH
TCR
Zf-Dof
EIN3
FAR1
CG1
E2F_TDP
GATA-zf
GRAS
GAGA_bind
LBD
LFY
NAC
NF-YB
NF-YC
GeBP
HRT-like
Nin-like
NF-X1
NZZ/SPL
S1Fa-like
LOB
FLO_LFY
NAM
NF-YB
NF-YC
DUF573
HRT-like
RWP-RK
Zf-NF-X1
NOZZLE
S1FA
SAP
SBP
SRS
STAT
TCP
VOZ
Trihelix
Whirly
WRKY
YABBY
NF-YA
ZF-HD
HSF
SAP
SBP
DUF702
STAT
TCP
VOZ
Trihelix
Whirly
WRKY
YABBY
CBFB_NFYA
ZF-HD_dimer
HSF_dna_bind
TF family
DNA binding domain
Auxiliary domain
Forbidden domain
With
Without
1
One domain
2
Two or more domains
BES1
DUF822
LSD
Zf-LSD1
Peptidase_C14
GARP
G2-like
ARR-B
G2-like
Response_reg
B3
ARF
B3
Auxin_resp
B3
AP2/ERF
RAV
AP2
ERF
2
1
AP2
C2H2
Zf-C2H2
RNase_T
MYB
MYB_related
MYB
1
2
Myb_dna_bind
SWIRM
DBB
CO-like
2
Zf-B_box
CCT
HD-ZIP
HB-other
TALE
HD-ZIP_I/II or START
Homeobox
BELL or ELK
HB
Wus type homeobox
PHD
WOX
HB-PHD
C3H
Zf-CCCH
RRM_1 or Helicase_C
GRF
WRC
QLQ
MADS
M type
MIKC
SRF-TF
K-box

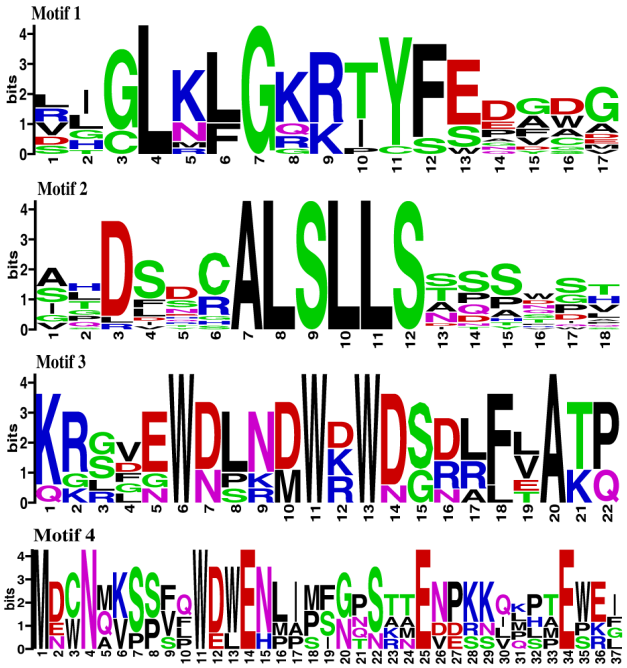
植物转录因子家族分类规则

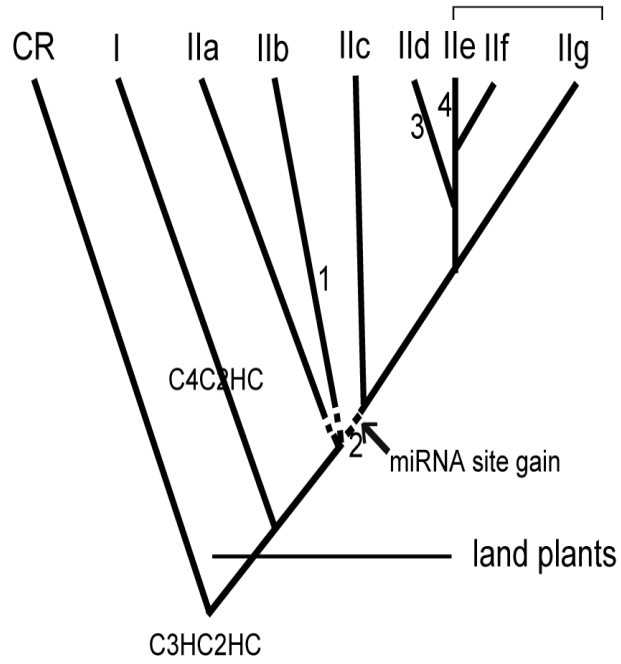
规　则	家族数	实　例
仅根据DNA结合结构域DBD	36	bHLH-HLH, EIL-EIN, SBP, WRKY
根据DBD和辅助结构域	9	GRF-(WRC/QLQ) MIKC-(SRF-TF/K-box)
根据DBD和禁止结构域	4	C2H2-(Zf-C2H2/Rnase_T)
根据DNA结合结构域数目	5	AP2-(2 AP2), ERF-(1 AP2)
超家族	6	MYB, GRAP, MADS B3, AP2/ERF, Homeobox

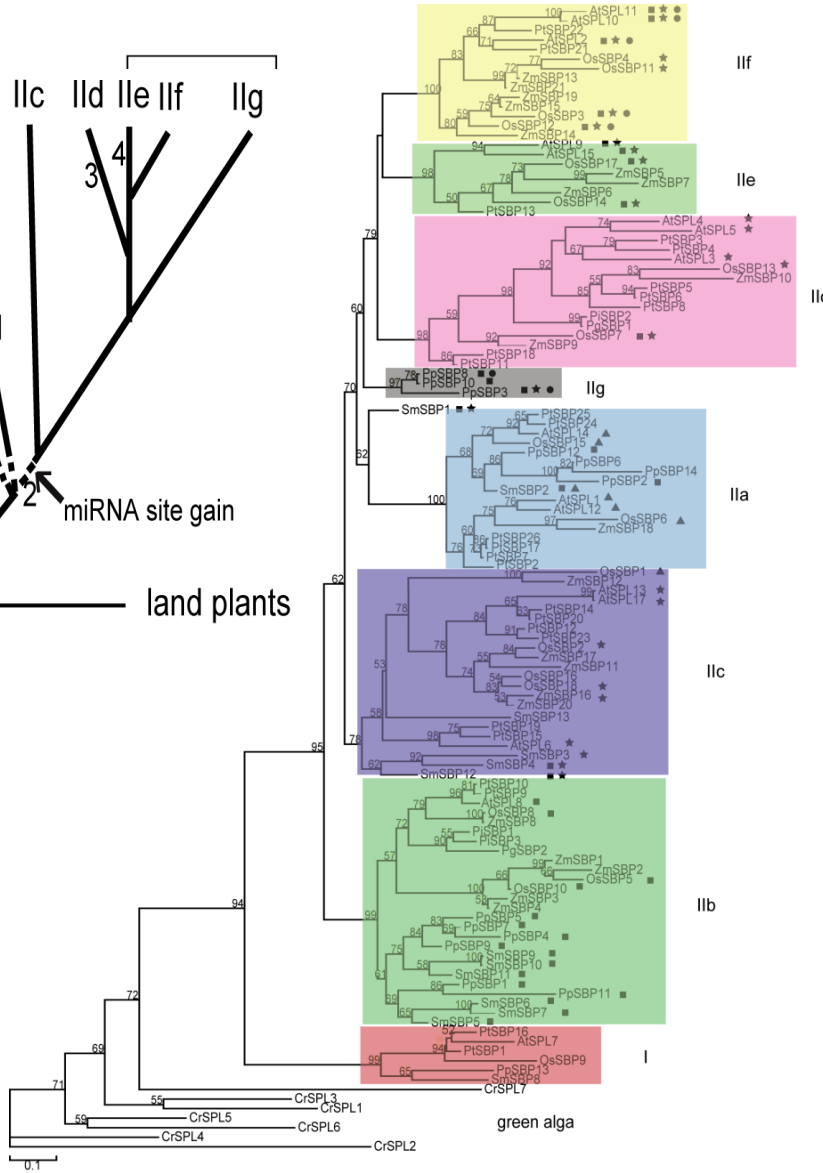
拟南芥和水稻SBP转录因子分析



拟南芥和水稻SBP转录因子分析







Guo et al. 2005, Bioinformatics. 21:2568-9.
Gao et al. 2006, Bioinformatics. 22:1286-7.
Zhu et al, 2007, Bioinformatics, 23:1307-8.
Guo et al, 2008, NAR, 36:D966-9.
Guo et al, 2008, Gene, 418:1-8.
Zhang et al., 2011, NAR, 39:D1114-7.

植物转录因子数据库课题组

姓名	主要工作	现在工作单位
何坤	拟南芥转录因子数据库、植物转录因子数据库	孟山都公司
郭安源	拟南芥转录因子数据库、植物转录因子数据库	华中科技大学教授
朱其慧	杨树转录因子数据库	哈佛大学
钟应福	水稻转录因子数据库	LifeTech公司
刘翟	拟南芥转录因子数据库	微生物所研究员
刘小川	植物转录因子数据库直系同源基因预测	美国博士后
张禾	植物转录因子数据库维护、和WebLab接口	密歇根大学
赵义	植物转录因子数据库注释	诺禾致源公司
顾孝诚	项目指导、论文修改	北京大学教授
高歌	水稻转录因子数据库、植物转录因子数据库	北京大学课题组长
陈新	植物转录因子数据库	北京大学
靳进朴	植物转录因子数据库更新、注释	北京大学博士后