基因组信息资源及应用

基因组学和生物信息学前沿讲座

中科院研究生院玉泉校区综合楼404

2014年6月16日

罗静初

北京大学生命科学学院

luojc@pku.edu.cn

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

国际生物信息中心

- NCBI – http://www.ncbi.nlm.nih.gov/
  National Center for Biotechnology Information
  美国国家生物技术信息中心
- EBI – http://www.ebi.ac.uk/
  European Bioinformatics Institute
  欧洲生物信息学研究所
- ExPASy – http://www.expasy.org/
  Expert of Protein Analysis System (SIB)
  瑞士蛋白质分析专家系统（瑞士生物信息研究所）

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10多名工作人员，目前有工作人员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1月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。2015年4月，将在北京举行的Biocuration大会上(http://biocuration2015.tilsi.org/)作报告。

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。

生物信息网络教程

- EBI – http://www.ebi.ac.uk/training/online/course-list
  EBI 生物信息网络教程
- NCBI – http://www.ncbi.nlm.nih.gov/About/primer/
  NCBI 生物信息培训教程
- SIB – http://edu.isb-sib.ch/
  瑞士生物信息学研究所培训课
- BTN – http://www.biotnet.org/
  欧洲生物信息培训网
- EMBER – www.ember.man.ac.uk/
  英国曼切斯特大学生物信息学自学网站

生物信息文献文档

- PuBMed – http://www.ncbi.nlm.nih.gov/pubmed/
  NCBI生物医学文摘数据库
- PMC – http://www.ncbi.nlm.nih.gov/pmc/
  NCBI生物医学全文数据库
- Bookshelf – http://www.ncbi.nlm.nih.gov/books/
  NCBI生物医学书刊数据库
- Protein Spotlight - http://www.expasy.org/spotlight/
  蛋白质故事精选
- MoM - http://www.rcsb.org/pdb/101/motm_archive.do
  Molecule of the Month 生物大分子月刊

生物信息数据库

- GenBank – http://www.ncbi.nlm.nih.gov/genbank/
  NCBI核酸序列数据库
- RefSeq – http://www.ncbi.nlm.nih.gov/refseq/
  NCBI参考序列数据库
- UniProt – http://www.uniprot.org/
  EBI/SIB蛋白质序列数据库
- PDB – http://www.rcsb.org/
  蛋白质结构数据库
- Ensembl – http://www.ensembl.org/
  EBI/Sanger基因组数据库

生物信息数据库专刊

- 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；2003年起，每年7月1日专辑刊登生物信息分析网络平台，Gary Benson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。

生物信息分析平台和软件工具

- WebLab - http://weblab.cbi.pku.edu.cn/
  生物信息分析平台
- EMBOSS - http://emboss.sourceforge.net/
  欧洲分子生物学开源软件包
- BLAST - http://www.ncbi.nlm.nih.gov/blast/
  数据库搜索系统
- MEGA - http://www.megasoftware.net/
  系统发生分析软件
- SPDBV - http://spdbv.vital-it.ch/
  蛋白质分子模拟软件
- NAR - http://nar.oxfordjournals.org/content/41/W1.toc
  NAR杂志生物信息网络分析平台专辑

基因组综合数据资源和注释系统

简称	内容	网址
NCBI	NCBI基因组数据资源	http://www.ncbi.nlm.nih.gov/sites/genome
JGIDB	美国联合基因组研究	http://genome.jgi-psf.org/
UCSC	基因组生物信息学综合网站	http://genome.ucsc.edu/
ENSEMBL	EBI/Sanger基因组数据库	http://www.ensembl.org/
Gramene	植物基因组数据资源	http://www.gramene.org/
EuPathDB	真核生物病源体	http://eupathdb.org/
NMPDR	美国微生物数据资源	http://www.nmpdr.org/
GeneDB	英国Sanger病源体基因数据库	http://www.genedb.org/
GenoList	法国巴斯德研究所微生物资源	http://genolist.pasteur.fr/

模式生物基因组

- 人类基因组计划指定的六大模式生物
- 已完成基因组测序的代表性模式动物
- 已完成基因组测序的代表性模式植物
- 已完成基因组测序的代表性模式微生物
- 已完成基因组测序的主要宏基因组

模式生物基因组基本信息

物种名	英文	学名	基因组	染色体	基因
人	Human	*Homo sapians*	3300	23	20800
小鼠	Mouse	*Mus musculus*	3500	20	23000
果蝇	Fruitfly	*Drosophila Melanogaster*	169	6	14000
拟南芥	Thale Cress	*Arabidopsis thaliana*	136	5	27500
线虫	Worm	*Caenorhabtitis elegans*	103	6	20500
酵母	Yeast	*Saccharomyces celevisiae*	12.16	10	6700
大肠杆菌	E. coli	*Escherichia coli*	4.64	0	4169

基因组大小和蛋白质编码基因数均为近似值。
大肠杆菌数据来自http://genolist.pasteur.fr/，其余来自http://www.ensembl.org/

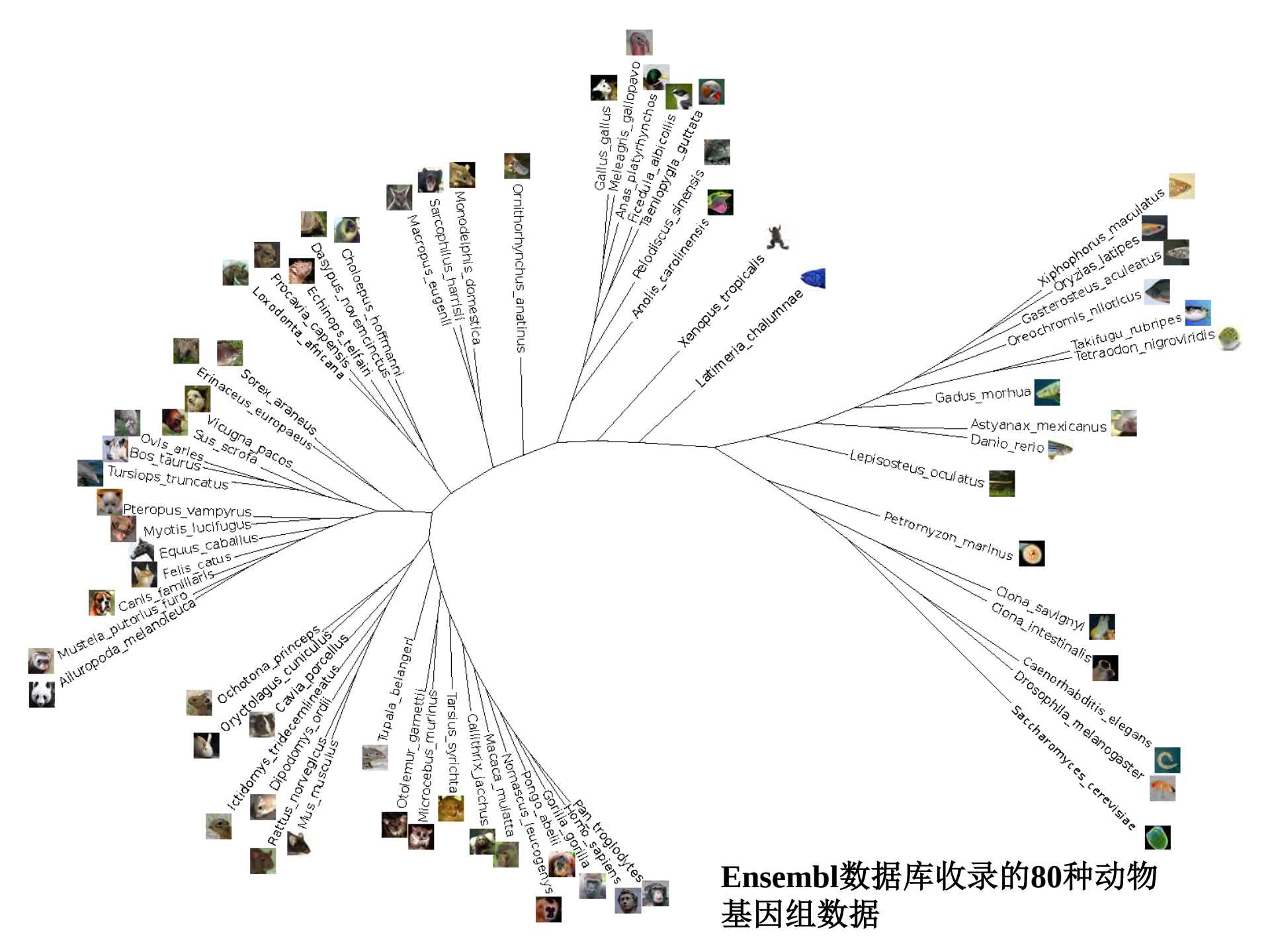
生物医学常用模式生物

- 哺乳类：小鼠(Mouse)、大鼠(Rat)
- 鸟类：鸡(Chicken)
- 两栖类：爪蟾(Frog)
- 鱼类：斑马鱼(Zebrafish)
- 昆虫类：果蝇(Fruit fly)、水蚤(Water flea)
- 线虫类：线虫(Round worm)
- 变形虫类：阿米巴(Amoeba)
- 真菌类：链孢霉(Neurospora crassa)
- 酵母类：裂殖酵母(Fission yeast)、芽殖酵母(Budding yeast)

http://www.nih.gov/science/models/

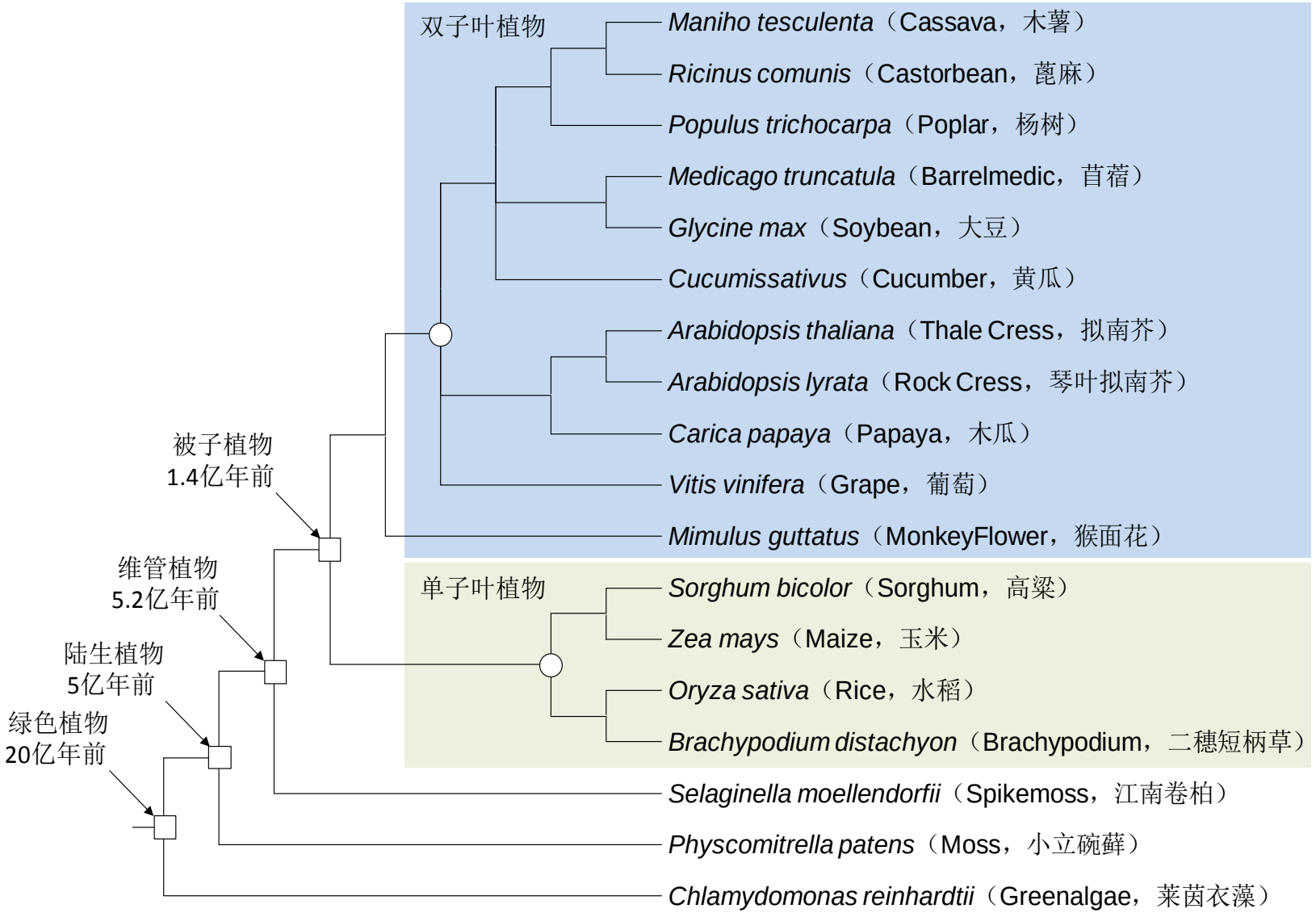
常用模式生物资源网站

物种	资源	网址
小鼠	Mouse Genome Informatics	http://www.informatics.jax.org/
大鼠	Rat Genome Database	http://rgd.mcw.edu/
爪蟾	Xenopus Resorce	http://www.xenbase.org/
斑马鱼	Zebrafish Database	http://zfin.org/
果蝇	Drosophila genes and genomes	http://flybase.org/
水蚤	Water Flea Genome Database	http://wfleabase.org/
线虫	Worm Base	http://www.wormbase.org/
变形虫	Dictyostelid Database	http://dictybase.org/
链孢霉	*Neurospora crassa* Database	http://www.broadinstitute.org/annotation/genome/neurospora/MultiHome.html
裂殖酵母	Pombe Net	http://www.pombe.net
芽殖酵母	Yeast Genome Database	http://www.yeastgenome.org/



Ensembl数据库收录的80种动物基因组数据

Phytozome数据库收录的部分植物基因组



人类基因组相关信息

染色体	RefSeq登录号	序列长度	蛋白质	RNA	基因	完成日期	PubMed ID
1	NC_000001.10	249,250,621	3129	451	3251	2006 May	16710414
2	NC_000002.11	243,199,373	1957	214	2175	2005 Apr	15815621
3	NC_000003.11	198,022,430	1651	167	1728	2006 Apr	16641997
4	NC_000004.11	191,154,276	1184	96	1340	2005 Apr	15815621
5	NC_000005.9	180,915,260	1400	127	1493	2004 Sep	15372022
6	NC_000006.11	171,115,067	1561	154	1791	2003 Oct	14574404
7	NC_000007.13	159,138,663	1476	207	1711	2003 Jul	12853948
8	NC_000008.10	146,364,022	1122	94	1216	2006 Jan	16421571
9	NC_000009.11	141,213,431	1214	176	1407	2004 May	15164053
10	NC_000010.10	135,534,747	1282	140	1283	2004 May	15164054
11	NC_000011.9	135,006,516	1927	103	2015	2006 Mar	16554811
12	NC_000012.11	133,851,895	1584	119	1596	2006 Mar	16541075
13	NC_000013.10	115,169,878	505	71	632	2004 Apr	15057823
14	NC_000014.8	107,349,540	938	129	1373	2003 Feb	12508121
15	NC_000015.9	102,531,392	989	227	1170	2006 Mar	16572171
16	NC_000016.9	90,354,753	1270	143	1279	2004 Dec	15616553
17	NC_000017.10	81,195,210	1755	183	1648	2006 Apr	16625196
18	NC_000018.9	78,077,248	473	52	507	2005 Sep	16177791
19	NC_000019.9	59,128,983	2029	171	1910	2004 Apr	15057824
20	NC_000020.10	63,025,520	861	85	834	2001 Dec	11780052
21	NC_000021.8	48,129,895	381	73	419	2000 May	10830953
22	NC_000022.10	51,304,566	708	99	820	1999 Dec	10591208
X	NC_000023.10	155,270,560	1324	127	1533	2005 Mar	15772651
Y	NC_000024.9	59,373,566	137	62	397	2003 Jun	12815422

巴斯德研究所微生物基因组数据库

细菌名	学名	网址	**RefSeq**	长度
大肠杆菌	Escherichia coli K-12	http://genolist.pasteur.fr/Colibri/	NC_000913	4639675
枯草杆菌	Bacillus subtilis 168	http://genolist.pasteur.fr/SubtiList/	NC_000964	4215606
兰细菌	Synechocystis PCC6803	http://genolist.pasteur.fr/CyanoList/	NC_000911	3573470
兰细菌	Anabaena PCC7120	http://genolist.pasteur.fr/CyanoList/	NC_003272	6413771
发光杆菌	Photorhabdus luminescens	http://genolist.pasteur.fr/PhotoList/	NC_005126	5688987
无乳链球菌	Streptococcus agalactiae	http://genolist.pasteur.fr/SagaList/	NC_004368	2211485
肺炎链球菌	Streptococcus pneumonia R6	http://genolist.pasteur.fr/StreptoPneumoList/	NC_003098	2038615
肺炎链球菌	Streptococcus pneumonia TIGR4	http://genolist.pasteur.fr/StreptoPneumoList/	NC_003028	2160842
金黄葡萄球菌	Staphylococcus aureus N315	http://genolist.pasteur.fr/AureoList/	NC_002745	2814816
金黄葡萄球菌	Staphylococcus aureus	http://genolist.pasteur.fr/AureoList/	NC_002758	2878529
无害李氏杆菌	Listeria innocua	http://genolist.pasteur.fr/ListiList/	NC_003212	3011208
李氏杆菌	Listeria monocytogenes	http://genolist.pasteur.fr/ListiList/	NC_003210	2944528
麻风分枝杆菌	Mycobacterium leprae	http://genolist.pasteur.fr/Leproma/	NC_002677	3268203
结核分枝杆菌	Mycobacterium tuberculosis	http://genolist.pasteur.fr/TubercuList/	NC_000962	4411532
牛结核杆菌	Mycobacterium bovis	http://genolist.pasteur.fr/BoviList/	NC_002945	4345492
溃疡分枝杆菌	Mycobacterium ulcerans	http://genolist.pasteur.fr/BuruList/	NC_008611	5631606
幽门螺杆菌	Helicobacter pylori 26695	http://genolist.pasteur.fr/PyloriGene/	NC_000915	1667867
幽门螺杆菌	Helicobacter pylori J99	http://genolist.pasteur.fr/PyloriGene/	NC_000921	1643831
嗜肺军团菌	Legionella pneumophila	http://genolist.pasteur.fr/LegioList/	NC_006368	3503610
肺支原体	Mycoplasma pulmonis	http://genolist.pasteur.fr/MypuList/	NC_002771	963879

生物信息应用实例

- Blast数据库搜索策略
- 植物转录因子数据库
- 植物特异转录因子SBP基因家族起源和演化
- 水稻基因组多倍化分析

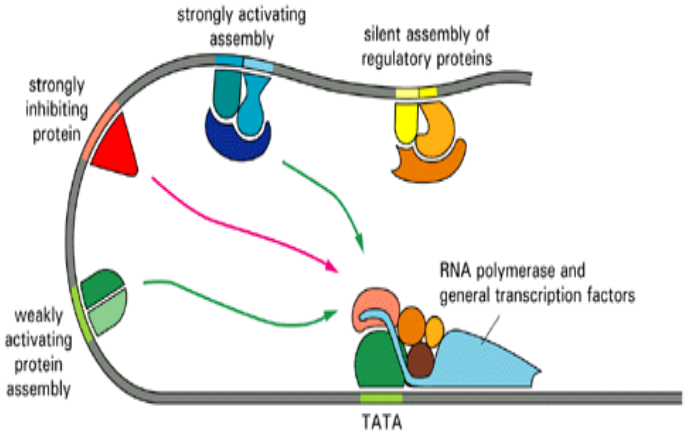
植物转录因子数据库

- 数据来源：基因组数据库、RefSeq、PlantGDB
- 方法：文献阅读、家族分类、预测、注释
- 结果：83个物种、59个家族、129288个转录因子
- 用途：基因转录调控研究、农作物育种

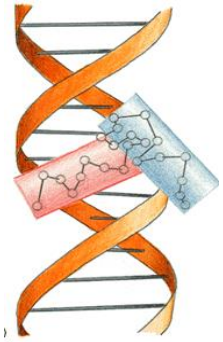
http://planttfdb.cbi.pku.edu.cn/

转录调控是基因调控的主要机制

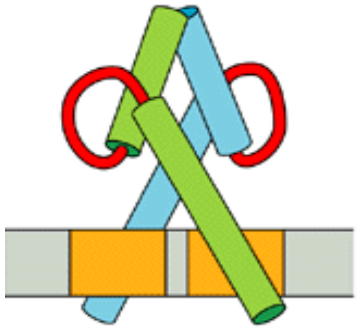
- 转录调控通过顺式作用元件和反式作用因子相互作用实现
- 反式作用因子通称转录因子
- 转录因子包括通用转录因子和特异转录因子
- 特异转录因子分为不同家族



螺旋-回折-螺旋
Helix-Turn-Helix

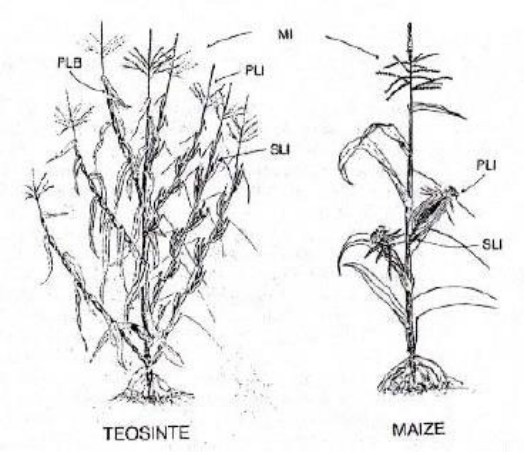


螺旋-回环-螺旋
Helix-Loop-Helix

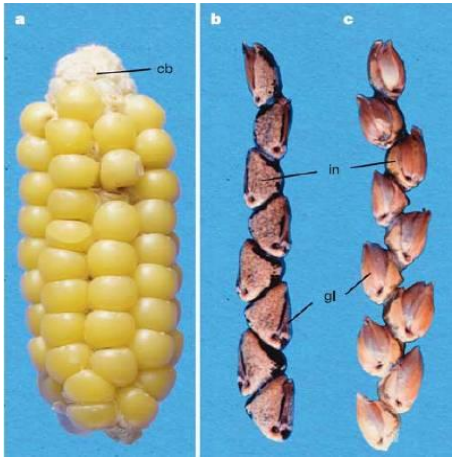


锌指结构
Zinc Finger

转录因子与多种农作物性状相关



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



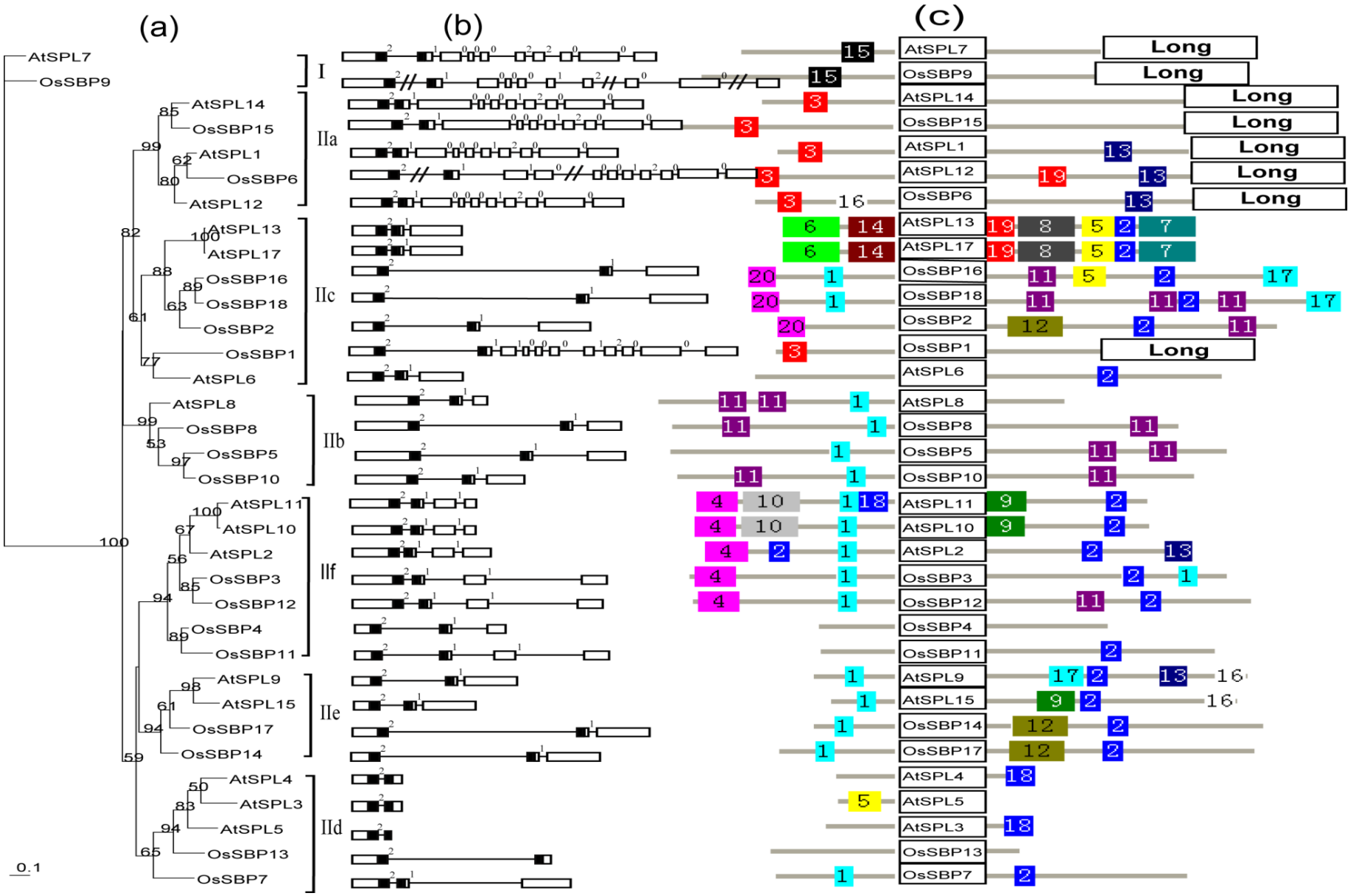
Wang *et al*. (2005)
Nature, 436:714

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

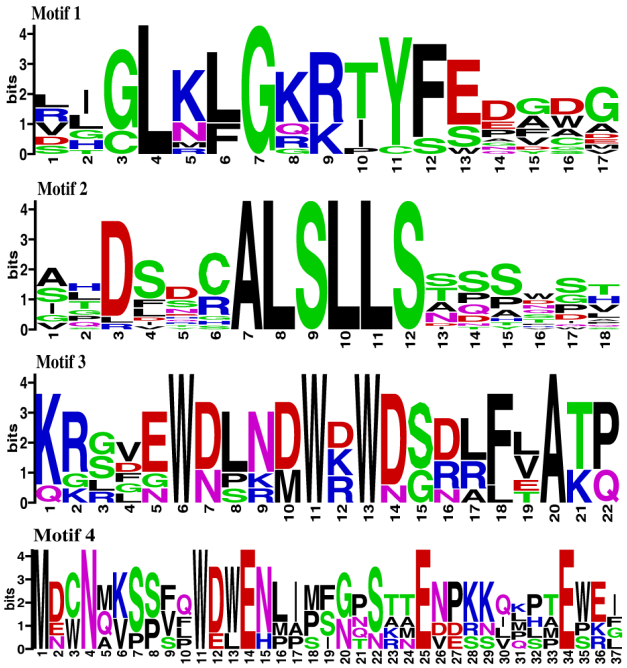
Kinishi *et al* (2006) Science, 312:1392
Li *et al* (2006) *Science*, 312:1936

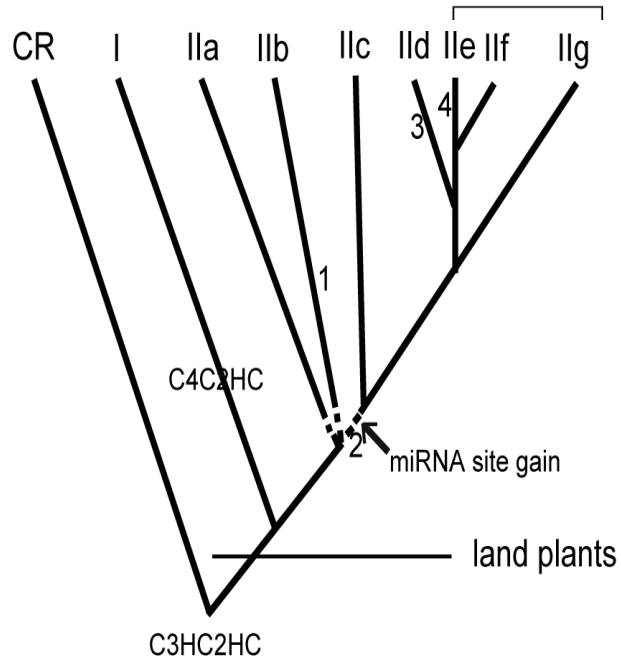
Jiao *et al* (2010) Nature Genetics, 42:541
Miura et al (2010) Nature Genetics, 42:545

拟南芥和水稻SBP转录因子分析



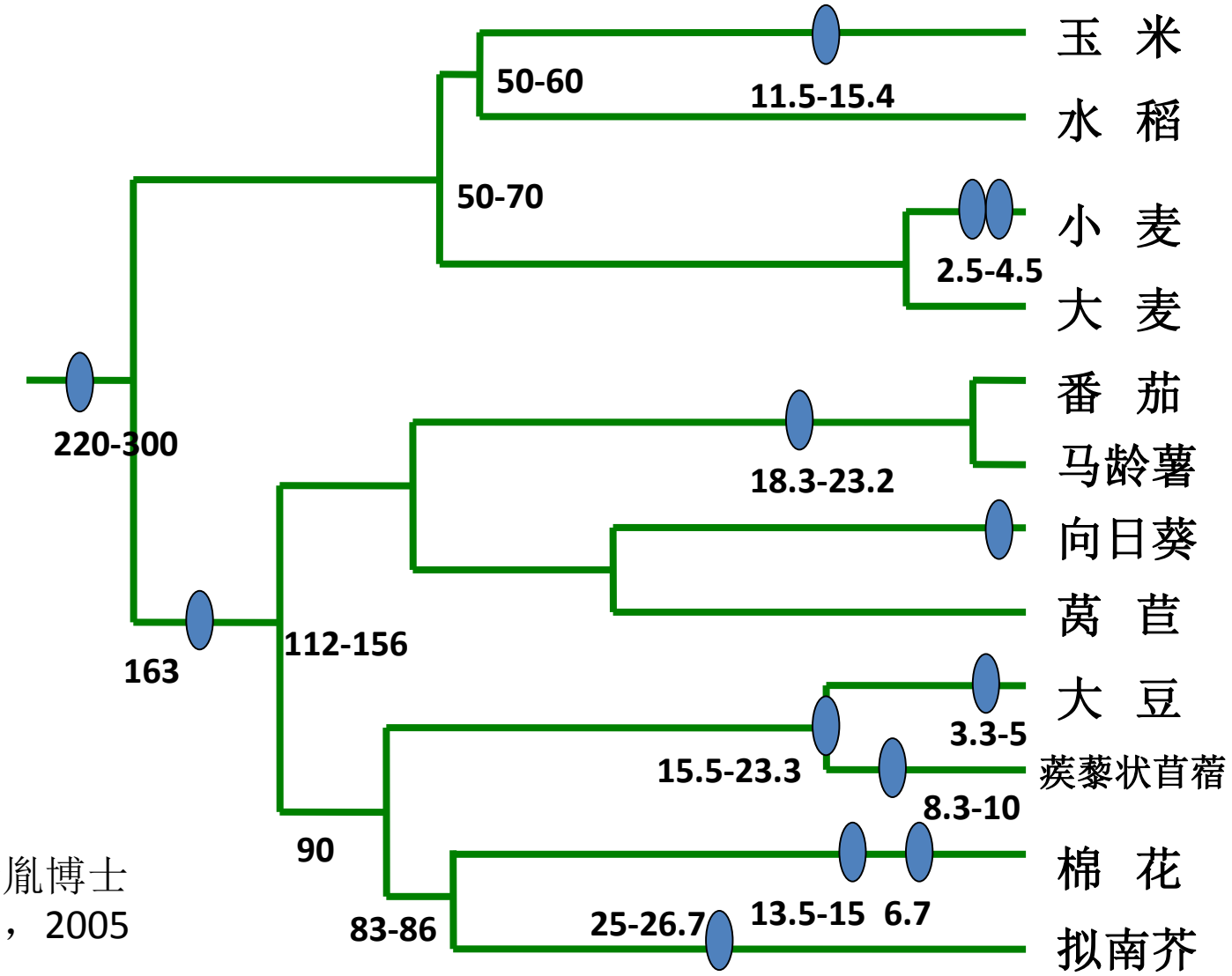
拟南芥和水稻SBP转录因子分析





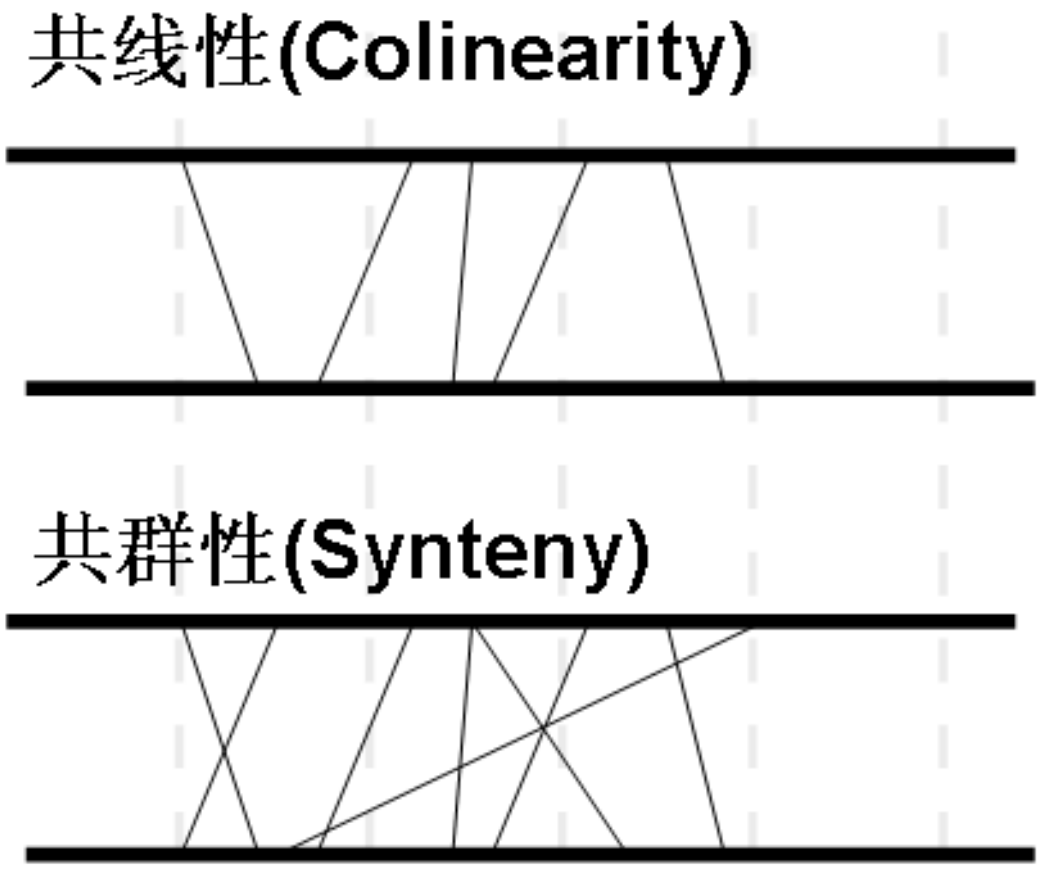
Guo et al. 2005, Bioinformatics. 21:2568-9.
Gao et al. 2006, Bioinformatics. 22:1286-7.
Zhu et al, 2007, Bioinformatics, 23:1307-8.
Guo et al, 2008, NAR, 36:D966-9.
Guo et al, 2008, Gene, 418:1-8.
Zhang et al., 2011, NAR, 39:D1114-7.

被子植物祖先基因组复制事件

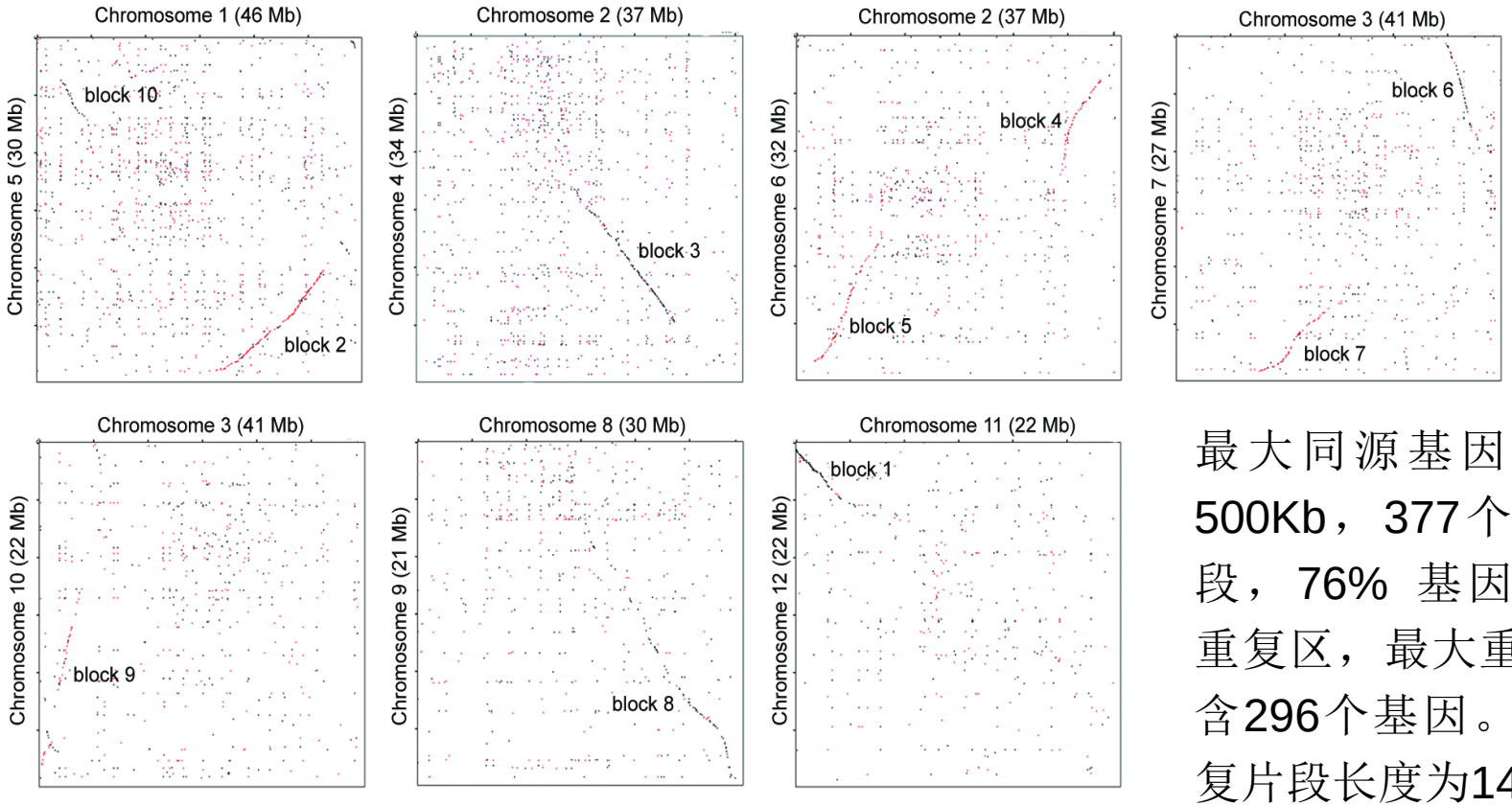


王希胤博士
论文，2005

基因组水平基因复制



水稻祖先基因组复制事件分析



最大同源基因对间隔500Kb，377个重复片段，76% 基因分布在重复区，最大重复片段含296个基因。最大重复片段长度为14Mb

Wang et al. 2005, New Phytol. 165:937-46.
Wang et al. 2006, BMC Bioinformatics. 7:447.
Shi et al. 2006, Gene. 376:199-206.
Li et al. 2009, BMC Bioinformatics. 10(Suppl 6):S8.

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
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
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
顾孝诚	项目指导、论文修改	北京大学
高歌	水稻转录因子数据库、植物转录因子数据库	北京大学
陈新	植物转录因子数据库	北京大学
张禾	植物转录因子数据库维护、和WebLab接口	北京大学
靳进朴	植物转录因子数据库更新、注释	北京大学
赵义	植物转录因子数据库注释	北京大学

植物特异转录因子SBP家族分析

郭安源 – 华中科技大学生命科学学院教授

朱其慧 – 哈佛大学博士后

顾孝诚 – 北京大学生命科学学院教授

葛颂 – 中科院植物所研究员

杨继 – 复旦大学生命科学学院教授

水稻祖先基因组复制分析

王希胤 – 河北理工大学生命科学学院教授

史晓黎 – 美国德州

葛颂 – 中科院植物所研究员

郝柏林 – 复旦大学生命科学学院教授

Applied Bioinformatics Course

Welcome

Welcome to ABC - the web site of Applied Bioinformatics Course. We'll learn, step by step, the ABCs of:

- How to access various bioinformatics resources on the Internet.
- How to query and search biological databases.
- How to use bioinformatics tools to analyze your own DNA or protein sequences.
- How to construct a phylogenetic tree for the bunch of sequences at your hand.
- How to predict the three dimensional structure of your favorite protein.

And lots more!

How we learn

We will run the course in a training room. Each student will have a PC connected into the Internet. We start with introducing the international bioinformatics resources around the world, for example, NCBI and EBI. We then use the WebLab bioinformatics platform developed by CBI, to get familiar with dozens of bioinformatics tools through hands-on practice. We will do a lot of exercises for sequence alignment, database similarity search, motif finding, gene prediction, as well as phylogenetic tree construction and molecular modeling. Finally, we will focus on several projects to solve real biological problems. You are encouraged to bring your own problems to discuss and, hopefully, to solve during the course! Please read the article Teaching the ABC of Bioinformatics [PDF] to know more about this course.

What you need before the course

- A desktop PC or laptop hooked to the Internet.
- Good background of biochemistry and molecular biology - you may try the pretest [English, Chinese] to see how good at it you are.
- Ability to read in English such as the contents of this page.
- At least three hours every week to have group discussions and to stick on the Internet to do exercises and homework assignments.

What you gain from the course