第一章 什么叫生物信息学

1、从染色体讲起－DNA(四种元素组成的很长的线)

2、什么是生物信息学？

历史短（20 年）

伴随着人类遗传密码破译（人类基因组计划）而产生的一门学科

最早的定义（基本的）：主要处理跟遗传密码相关的方方面面，对基因组信息的获取、处理、存储，分配、分析和解释（－－来自美国政府文本）

（1）基因组信息的获取、处理、存储，分配：如何存好，用信息

（2）基因组信息的分析和解释：深刻挖掘基因组信息所提供的生命活动基本规律的知识

讲者的定义：破译人类遗传密码（30 亿个符号组成的一百万页的天书）

研究者建立方法去读这本书，读出哪些是功能元件，编码哪些蛋白，蛋白的功能是什么；

内涵的定义（具体化）：生物信息学是把基因组 dna 序列分析作为源头，破译隐藏在 dna 序列中的遗传语言，特别是非编码区的实质，同时在发现新基因的信息之后，进行蛋白质空间结构的模拟和预测。

自然哲学的定义（哲学上）：研究目标是揭示基因组信息结构的复杂性和遗传语言的根本规律，它是本世纪自然科学和技术科学领域中基因组、信息结构和复杂性这三个重大科学问题的有机结合。

3、测序

80 年代末，一次测 500-1000 个碱基，费用一个碱基一美金（人类总共有 30 亿个碱基对，中国承担 1 / 100：三千万）

计算机能力达到

4、4 类序列数据与生物信息学相关

碱基的数目（代表我们破译了多少遗传密码，大约 14 个月增长一倍）

各种生物的 esp 数据（表达序列的标签）：代表遗传密码中跟基因相关的一小段序列，增长更快

单核苷酸多态性（snp):代表不同人种之间，病人和正常人之间碱基的变异。

：肿瘤病人的一个碱基从 g 变为 t,就把这个信号记下来，叫做和该肿瘤相关的单核苷酸多态性

：比 esp 数据增长更快，更引起人们重视，跟人类分子水平的疾病相关

完整基因组数据（有多少生物的遗传密码被破译了）

5、典型的数据库

genebank(美国，1997 建设)：关于核酸数据库，非常全面核酸序列

embl（英国，1982 开始服务）:

ddbj（日本,1987 开始服务）：dna data bank Japan

以上 3 个数据库互相交互，更新信息。用三个都可，推荐 genebank,登录速度较快

6、方法过程

破译的遗传密码－－拼接好遗传密码－－识别基因(遗传密码中基因编码区，不超过 3%）－－翻译成蛋白质的一级序列（氨基酸序列）－－通过蛋白质的一级序列，预测蛋白质的空间结构－－预测生物学功能

7、原来的研究核心是研究编码的基因，研究蛋白，这些只占很少的部分
8、97%的遗传序列是非编码序列

tips:
1、美国有由总统签署的科研计划
一：曼哈顿计划（能源变化－使用核能）
二：阿波罗计划（离开地球，探索宇宙－登月）
三：基因组计划（上世纪 90 年代，30 亿美金，克林顿签署，破译人遗传密码，分子水平认识自身）
2、中国独立完成水稻基因组测序，水稻基因组大小是人类基因组的 1/7；该研究工作登上 science 封面
3、基因是割裂开的，由外显子和内显子组成的
对于真核生物，识别基因就是找外显子

第二章 生物信息学研究内容之一（讲解几个经典课题）

一、大规模基因组测序中的信息分析

基因组的拼接；基因的标注
DNA 测序仪：测出来是四种光密度，光密度越大说明碱基的含量越高，每个颜色代表一个碱基
光密度采样（使光密度反应碱基的表达量）－－**碱基读出**（把光密度翻译成碱基的信息，哪个光密度代表哪个碱基）
tips：
测序前，克隆遗传密码－－繁殖－－测序
克隆时前面需要加一些载体(载体数据库),实际得到的序列比真正要的序列多，所以还要把之前加的载体序列切除（序列和载体数据库比对）－－得到一段一段的序列－－**基因拼接**－－拼接后还有很多洞,还要补洞－－－得到完整的样本－－**基因标注**（找出哪些是基因，哪些是其他的功能元件）
以上每一步都要用到信息分析

1、基因拼接（没有信息生物学家参与，无法完成拼接）
举例：一本书，剪成方块，如何尽快拼出原来的书？
方法 1:买两本一样的书，一个剪成方块，一个剪成长条；在长度设置合理的情况下，反复比对长条和方块中的内容，拼出内容 30%。
方法 2：买 3 本书，用三种方式撕，统计学家计算这种情况下最大拼出 50％
最后，统计数学家得出：5 本书，可以拼出 90%（拼接的覆盖率）

0 本书，可以拼出 **99%**（10 次测序，300 亿美元，－2004）
tips：
人类基因组测序完全图（99%的覆盖率），遗留 200 多个洞

2、基因注释（真核生物：怎么发现外显子；原核生物：怎么发现基因）

方法 1:基于信号

剪切位点：是内含子和外显子之间的边界，这个边界的序列有特征，它是有信号的）

任何一个编码序列都有起始密码子和终止密码子；起始密码子是个三联体，终止密码子有好几个，每一个又是个三联体

发现密码子，就意味着基因开始和结束

只使用信号的缺点：

1、信号太短，密码子只有三联体；由于序列太短，假信号比真信号多得多，信噪比差

解决：综合利用多种信息，多种方法

2、这个信号不是一个固定的，确定的模式；这个序列可能是一种可以稍加改变的模式，因此不能简单的用序列标定

信号实际不是个序列，是个分布矩阵；根据该矩阵对序列进行打分，高于一定阈值的认为是真信号，否则，假信号

分布矩阵是科学家经过大量数据计算得到的

信号：启动子，终止子；密码子；等

方法 2:基于组成

（1）基于统计的办法

三个碱基决定一个氨基酸

如果是某段遗传密码是基因，那么这段序列按 3 个一组翻译后一定是一个蛋白质的一级序列；

：对决定蛋白质结构的氨基酸空间序列一定不是 20 种氨基酸残基的平均分布

：对蛋白质氨基酸的组成分布进行统计，真正的蛋白质中氨基酸分布是有规律的

从第一个碱基开始读，得到一个氨基酸序列；从第二个碱基开始读，得到一个；从第三个碱基开始，得到一个；

若是基因：这三个序列中一定有一个很符合正真蛋白质的氨基酸分布规律，另外两个很不符合

该方法，成功准确率很低（约 50%）

（2）基于序列比对的办法

现有的数据库中已经存在千千万万个已经发现的基因，如果你得到一个新的序列，拿到数据库里去两两比对

两个序列比较，使用动态规划算法，可以得到严格解；时间复杂度较高，实际工作中用其简化算法

简化算法:blast 算法(basic local alignment search tool)；fasta 算法

：是实验室工作的开始（得到一条序列）

该方法，无法发现新的基因

（3）神经网络的方法

目的：区别不同特征（区别外显子和内含子）

BP 方法：输入层；输出层；中间层

输入的序列包含奇数个碱基；输出为 0，1:判断序列中间的碱基是否编码基因；

通过将序列左移或右移，依次判断每个碱基的情况

实际成功率：大约 63％

练习：编程实现一个三层的神经网络

tips：

外显子和内含子交界处是有信号的：外显子 5’端有一个二联体 GU，3’端有 AG

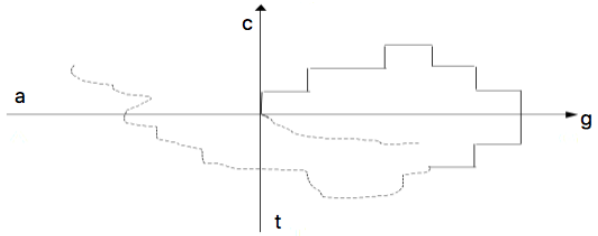
但是仅通过二联体判断，　　信息量太少，信噪比极高

（4）分维的方法（fractal　dimension 非线性、物理方法）

曲线的维度一定比一大比二小

具体方法： 把一个序列投影到二维碱基平面变成一条曲线，计算曲线的分维值；基因序列和非基因序列的分维值有明显差异，即可用来区分

具体投影过程如下：四个坐标方向分别代表四种不同碱基



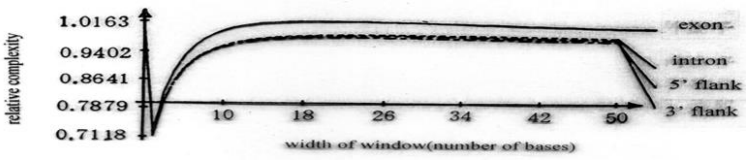
----------CGCGGCGTGTGTTATA ----------

（5）复杂性的方法（数学家）

把复杂性作为判据，来帮助理解是否编码序列和非编码序列有所不同

任取一段序列，数一联体出现的频次，二联体的频次，三联体的频次，加和，所得值为该序列的复杂度

实验结果：在序列长度大于 18 后，复杂性结果成稳态；编码序列的复杂度会更高



（6）密码学的方法

一致指数（coincident index）：指数越高，说明此处有信息

修改密码学的公式，得到计算碱基的一致指数公式

具体方法：计算序列的一致指数高低，如高，则认为是编码序列

方法的准确率约为 70%

实际使用中，是多重方法共同判断，得出可靠性更高的结果

多序列比对（都是近似办法，没有精确办法）

tips：

BLAST (Basic Local Alignment Search Tool)是一套在蛋白质数据库或 DNA 数据库中进行相似性比较的分析工具。BLAST 程序能迅速与公开数据库进行相似性序列比较。BLAST 结果中的得分是对一种对相似性的计说明。BLAST 对一条或多条序列(可以是任何形式的序列)在一个或多个核酸或蛋白序列库中进行比对。BLAST 还能发现具有缺口的能比对上的序列。 BLAST 是基于 Altschul 等人在 J.Mol.Biol 上发表的方法 (J.Mol.Biol.215:403-410(1990)),在序列数据库中对查询序列进行同源性比对工作。从最初的 BLAST 发展到现在 NCBI 提供的 BLAST2.0,已将有缺口的比对 序列也考虑在内了。BLAST 可处理任何数量的序列,包括蛋白序列和核算序列;也可选择多个数据库但数据库必须是同一类型的,即要么都是蛋白数据库要么都是核酸数据库。所查询的序列和调用的数据库则可 以是任何形式的组合,既可以是核酸序列到蛋白库中作查询,也可以是蛋白序列到蛋白库中作查询,反之亦然。 1、BLASTP 是蛋白序列到蛋白库中的一种查询。库中存在的每条已知序列将逐一地同每条所查序列作一对一的序列比对。 2、BLASTX 是核酸序列到蛋白库中的一种查询。先将核酸序列翻译成蛋白序列（一条核酸序列会被翻译成可能的六条蛋白），再对每一条作一对一的蛋白序列比对。 3、BLASTN 是核酸序列到核酸库中的一种查询。库中存在的每条已知序列都将同所查序列作一对一地核酸序列比对。 4、TBLASTN 是蛋白序列到核酸库中的一种查询。与 BLASTX 相反，它是将库中的核酸序列翻译成蛋白序列，再同所查序列作蛋白与蛋白的比对。 5、TBLASTX 是核酸序列到核酸库中的一种查询。此种查询将库中的核酸序列和所查的核酸序列都翻译成蛋白（每条核酸序列会产生 6 条可能的蛋白序列），这样每次比对会产生 36 种比对阵列。 至于如何进行,你先进入 NCBI,然后点 BLAST,选择你所需要的 BLAST 的类型,然后输入你要 BLAST 的序列,就可以看到与你所输入的序列的同源性序列了

3、遇到的问题与挑战

大问题：重复序列对于大规模基因组破译的影响

解决思路：对于人类基因组，用已经找出的独特的基因片段（唯一的 marker），分片时每片包含一个 marker；对于其他没有 marker 的其他物种基因组，首先过滤重复序列，之后对剩下的独特序列进行拼接，就能避免重复序列存在带来的拼接错误，与此同时，基因多出现在独特序列中，较少出现在重复序列。因此，以上处理过程基本可以拼接出包含基因的序列。

小问题：可变剪切所带来的蛋白质多样性的识别问题

有些大基因，可能包含很多（几十个）外显子，在成熟的过程中可以组合不同的外显子，组成蛋白；所以同一个基因，由于可变剪切的关系，会表达为不同的蛋白；这是从基因到蛋白遇到的一个复杂的问题。如何识别可变剪切的真实的蛋白产物？

由于外显子的多样性和可变剪切，使得一个基因可产生多种蛋白

（ 1、基因组拼接中的重复序列处理；2、基因标识的准确率；3、基因的可变剪接；4、蛋白质的空间结构和功能预测）

4、举例

（1） *T. tengcongensis* Complete Genome Analysis（腾冲耐热菌，微生物所在腾冲温泉中发现的）

中国破译的第一个 微生物的完整基因组（微生物所，基因组所，华大基因）

选择生物的原因：希望是耐热菌（跟生物起源有关，提升基础研究的价值）；发现耐热的酶；预估的基因量最小（300 万碱基，300 万人名币）；名族情怀（是中国发现的，中国人定名的）

：环状 DNA,
发表在基因组研究（是基因组研究的顶级期刊）
得到基因组后的研究工作：
找相似的细菌；
研究腾冲细菌能实现的生物功能（更功能 pathway 图进行比对判断即可）；
耐热菌的耐热功能可能需要基因表达特别的蛋白来实现
耐热性哟求：ran（跟造蛋白有关）的要求 CG 含量高，稳定

tips：
微生物由于功能简单，90%是基因，10%是非编码序列；而人类序列中，只有 3%是基因，97%是非编码序列；
高 AT 含量的序列很不稳定，是脆弱的（CG 三个氢键链接，AT 两个氢键链接）

二、SNP 的发现与鉴定

1、基于哪些数据可以帮助我们发现新的基因？
（1）基因组 DNA 的序列数据库：在序列中识别编码序列（拼接，组装，标示）
（2）EST(expression sequence tag：表达序列标签)数据库(dbEST:http://www.ncbi.nlm.nih.gov/genbank/dbest/)：把归属于同一个基因的 est 序列接在一块（去噪，拼接）

EST 是 完整 mRNA 中的一小段,**本身就是**要表达的**一段基因**

人类 EST 数据已经发展到超过 1000 万，人类实际有 2 万 5 千个基因，每个基因有 10 倍左右的覆盖度，所以科学家估计这些数据已经覆盖了 95%的人类基因；利用好 EST 数据就能的到基因（组装、拼接）；

作用：发现基因；发现 SNP;发现这个基因是否存在可变剪切；发现非编码 RNA

有效性：大部分新基因是靠理论方法预测出来的。比如啤酒酵母完整基因组 (约 1300 万 bp) 所包含的 6 千多个基因，大约 60% 是通过信息分析得到的。

：国际上现已出现了几个基于 EST 的基因索引如 UniGene (ftp://ncbi.nlm.nih.gov/pub/schuler/unigene) , Merck-Gene index(http://genome.wustl.edu/est/esthmpg.html) , GenExpress-index（ http://www..cshl.org) ，这些基因索引数据库(即二次数据库)构建了基因框架，极大地方便了相关研究者

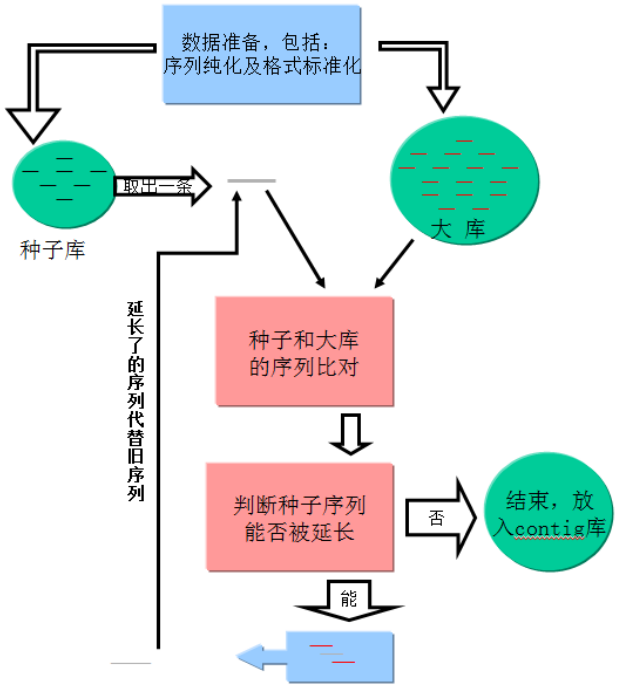
原理： 当测序获得一条 EST 序列时，它来自哪一个基因的哪个区域是未知的(随机的)，所以属于同一个基因的不同 EST 序列之间常有交叠的区域。根据这种“交叠”现象，就能找出属于同一个基因的所有 EST 序列，进而将它们拼接成和完整基因相对应的全长 cDNA 序列

实际应用：若你在实验室发现了一段核酸序列，可以送到数据库中进行比对（选择在 EST 中比对），之后不断拼接，找不到一样的之后，可能会得到基因；拼接时，寻找重合段的长度为 40（国际推荐 overlap 值为 40，低于 30 时，很容易接错），即 40 个完全一样的；两端剩余的未重合的部分长度不大于 10（否则，认为该序列本身不够可靠）；以上过程也叫基因电脑克隆（ 以一个序列片段为线索，通过它和整个数据库的比较，还原出全序列原貌）

应用流程：

（1） EST 数据库质量相对较低 因此在进行 Contig 电脑组装之前，需要探测并去除 EST 数据库中的污染序列

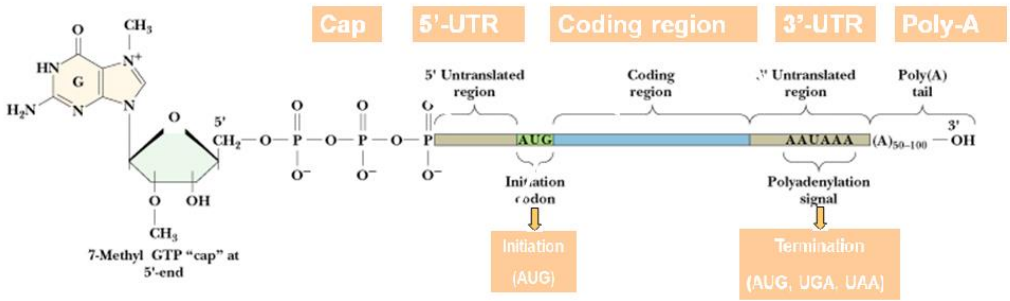
Silone:基因电脑克隆流程图



Tips：

1、**一个成熟 RNA 的结构**（准备放在核糖体上制造蛋白的）；

一定有的部分：cap(最前端的帽子结构：焦磷酸，三个）+ 5‘的非翻译区（UTR）+ **起始密码子（甲硫氨酸:AUG）+ 编码区+终止密码子（三选一：AUG,UGA,UAA）+** 3‘的非翻译区（UTR） + 尾巴（poly-A：若干个 A 连起来的）



2、SNP(**Single Nucleotide Polymorphisms:单核苷酸的多态性**)

：是拼接后，发现两个基本一样的拼接产物中，个别碱基的变化；可以拿来进一步鉴定它是否标志着正常人/病人导致的、不同人种导致的；去查两个拼接的 EST 来源，可以基本判定；

2、新基因发现的主要原理？

（1）使用 DNA 数据，发现基因的过程是个识别问题；

（2）使用 EST 数据，发现基因的过程拼接问题；

3、新基因发现的关键步骤是什么？

（1）使用 DNA 数据，步骤：组装--数学、物理的方法识别出编码区

（2）使用 EST 数据，步骤：数据的纯化（数据的处理、标准化）--归属于同一个基因的序列拼接在一起

三、完整基因组的比较研究（一个新的研究方向）

1、该研究解决什么问题？

研究生命是从哪里起源的？生命是如何进化的？遗传密码是如何起源的？估计最小独立生活的生物至少需要多少基因，这些基因是如何使它们活起来的？-- 完整基因组序列的比较研究是解决这些问题的重要途径

举例：讲者通过对鼠和人基因组的研究，发现两者拥有的基因总数及种类都基本一致；但是基因在染色体上的分布式完全不同的；由此，推测从遗传密码到表型之间，不仅仅跟遗传密码的总量有关，不仅仅跟包含多少基因有关，而跟这些遗传密码如何组织在一起有很大关系；（其实高等脊椎动物的遗传密码量都差不多 30 亿）--因为不同基因，它的左邻右舍不同，它的调控模式就不同；调控模式不同，它的表达时空特征就不同；时空特征不同就决定它的表型不同；所以，这是一个只有在完整基因组水平上比较才能给我们的启示。

四、基于序列数据的生物进化研究（当前面临的问题）

1、进化论研究

进化论研究的核心：是描述生物进化的历史(系统进化树)和探索进化过程的机制。自本世纪中叶以来，随着分子生物学的不断发展，进化论的研究也进入了分子水平。

当前分子进化的研究，已是进化论研究的重要手段，并建立了一套依赖于核酸、蛋白质序列信息的理论方法。

2、

序列相似性比较（序列的两两比对（常用程序包 BLAST、FASTA），确定这个序列的生物属性：判断该序列与那个序列相似）—>序列同源性比较（多序列比较(常用程序包 CLUSTAL)：计算量大，只有近似算法）—>选择合适的构建进化树的方法（有多种构建方法，常用软件包 PYLIP,MEGA）—>稳定性检验(构建过程要随机重复成百上千次，来判断分支是否稳定；通用方法 Boostrap)

相关软件及下载地址：

序列分析和多序列比较
BLAST Web site http://www.ncbi.nlm.nih.gov/BLAST/
FASTA at EBI http://www2.ebi.ac.uk/fasta3/
CLUSTALW software ftp://ftp-igbmc.u-strasbg.fr/pub/ClustalW
HMMER software http://hmmer.wustl.edu/
SAM profile software http://www.cse.ucsc.edu/research/compbio/sam.html
BCM Search Launcher
http://kiwi.imgen.bcm.tmc.edu:8088/searchlauncher/launcher.html

系统进化树构建和稳定性分析
PHYLIP http://evolution.genetics.washington.edu/phylip.html
Hennig86 http://www.vims.edu/~mes/hennig/software.html
MEGA/METREE http://www.bio.psu.edu/faculty/nei/imeg
GAMBIT http://www.lifesci.ucla.edu/mcdbio/Faculty/Lake/Research/Programs/
MacClade http://phylogeny.arizona.edu/macclade/macclade.html
PAUP http://onyx.si.edu/PAUP/
GCG software package http://www.gcg.com/

3、遇到的问题：

基因在演化的过程中，会发生横向迁移（Lateral Gene Transfer,客观事实）；所以在研究基因演化时，确保基因时没有受到横向迁移的影响；

4、在基因迁移客观存在的前提下，如何构建进化树

方法一:在选材时,去掉跟横向迁移有关的序列;已经存在数据库(COG),中只保留纵向演化的序列，去掉了横向迁移

方法二： 用完整的基因组进行比对（碱基水平的比对），不考虑基因是由纵向演化还是横向迁移得来；

遇到问题：基因组 size 不同，不能做符号和符号之间的比较，发展新的基于完整基因组的信息比对

提取符号特征，对 size 归一化；基本提取序列之间联体符号出现的频率

另一种思路：基于基因组中的部分，即只考虑基因组中的基因，基因也可以部分表达所以基因组信息

Tips:

1、相似跟同源是两个概念

第三章 生物信息学研究内容之二（进入 21 世纪后的研究热点：基因组研究从测序走向功能）

从遗传密码出发，而了解生物生活的功能；就是要完成静态的遗传密码，像 4 维的时空的信息的动态表达谱进行转换；20 世纪得到遗传密码，而 21 世纪要了解这个码在哪个时间，哪些组织里表达哪些功能

一、干细胞作为基因组研究的重要选材

- Totipotent stem cell: 全能干细胞, 如受精卵
- pluripotent stem cell: 多能干细胞，如囊胚中的内囊细胞
- multipotent stem cell: 专能干细胞，如造血干细胞

二、snp 研究的国籍大协作

1、snp 研究

SNP 研究是基因组领域理论成果走向应用的关键步骤。是联系基因型和表现型之间关系的桥梁。

2、重要发现：对 y 染色体的 snp 分析

实验：取了 100 中不同人群，其中包括黑猩猩、大猩猩，非洲原住民，等；想找出一个变化的碱基位点（snp），可以把所有样本分成两类；最后，找到一个位点可以把样本分成两类，一类包含三者（黑猩猩、大猩猩，非洲原住民），另一类是其他人类样本；

结论：

1、非洲的古老人群，比其他人更为古老；

2、人是从非洲开始迁徙的；

3、这个碱基的突变速率（大约 20 万年才会发生突变），即最近的一波人类迁移是从非洲分布到世界各地的；

3、snp 数据库

1、SNP Consortium's database（http://snp.cshl.org/index.html）：最新版本中（May 2001）包含有 1,034,034 个 SNP, 它们全部被锚定到人类基因组的工作草图上

2、NCBI SNP database （http://www.ncbi.nlm.nih.gov/SNP/overview.html）：这一数据库是由美国国立卫生研究院（NIH）的国家人类基因组研究所和国家生物技术信息中心共同建立和维护的。到 2001 年 9 月 12 日该数据库共收录了 3,053,511 个 SNP，其中人的 SNP 有 3,052,574。将这些数据进行整理，**去掉冗余**，使每个 **SNP 都是唯一**的。此时的 SNP 被称为 reference SNP 或 refSNP, 对每个这样的 SNP 都设定一个标号，称为 rs ID。整理后，人类的 refSNP 总数为 1,805,951。

3、The Human Genic Bi-Allelic Sequences Database（http://hgbase.cgr.ki.se/）：它是由一个欧洲协作组构建的。这一数据库收录了人基因组中所有已知的序列变化，包括：SNPs、**序列的插入和缺失(Indels)**、简单重复序列等。

4、The Allele Frequency Database(ALFRED)（http://alfred.med.yale.edu/alfred/index.asp）：它是人类群体**等位基因频率数据库**。当前它存有人类等位基因频率表 3561 份，SNP 233 个

5、The Protein Mutant Database(PMD)（ http://pmd.ddbj.nig.ac.jp/）：它不是核酸突变数据库，而是**蛋白突变数据库**。库中收录了蛋白质特定位点的氨基酸突变信息，以及这些突变对蛋白质结构功能的影响。现存有突变记录 23,167 条，代表了 115,766 个氨基酸突变。

三、DNA 芯片的广泛应用

1、基因芯片：**核酸水平**了解特定时间特定特定组织基因表达的重要工具；

芯片分析的流程：基因芯片的扫描（获得芯片上基因位点的光密度值）：对每个位点进行光密度的扫描时，要先进行网格化（确定芯片每一个杂交点的位置。具体说来，它要做三件事情，即：发现每一个杂交点；按照信号的强弱等级分割信号和背景的边界；分别读取信号和背景的光强度），以保证位点光密度的准确度量－> **一定要进行**

正则化（矫正其他因素的影响），使得光密度值能准确反映基因的表达量一>进行信息提取：可以通过聚类方法（把表达特征一致的那些个点放在一起（可能说明这些基因时协同起作用的））；也可以通过分类的方法（基于一些现有知识进行分类）

芯片分析的软件包：

（1）TIGR(The Institute for Genomic Research)芯片数据分析软件包： MultipleExperimentViewer (TMEV)是用 **Java** 语言设计的。用于对芯片数据标准化及进行聚类和距离代数的分析；ArrayViewer 是一个简化的芯片数据分析软件；Spotfinder 是用于芯片信号收集和图象处理的。

（2）Stanford 大学 Brown 实验室芯片数据分析软件：该实验室最早建立了实用的 Microarray 设备及其运行软件： ScanAlyze 用于荧光图象的处理； SMD Package 用于原始数据的储存和标准化；Cluster、XCluster 用于多层次聚类分析和构建自组织图； SAM 用于有监管的学习过程；TreeView 用于从分析结果建图； KNNimpute 用于评估缺失的芯片信息。

做正则化处理

（3）EBI(欧洲分子生物学研究所)的表达数据分析工具：是该实验室正在建立与完善的一套大型软件系统。其功能包括：聚类分析，特征模式发现，图象可视化，浏览基因分类信息，萃取调控序列，研究蛋白相互作用以及将分析结果连接到其他工具等

标准表达谱数据：从 Brown 实验室和 NCBI **的** GEO(Gene Expression Omnibus)网点可取得典型的基因表达谱数据集

可改进的问题：基因芯片光密度数据标准化上进行改进和提高；如何处理大量表达不明显的点（出现两倍或两倍以上的变化，认为有表达差异），在稀疏矩阵中，这些变化小于两倍的点如何赋值；大规模基因功能表达谱数据挖掘和知识发现：这是表达谱研究成功与否的关键。只有找到成百上千个表达水平发生变化基因之间在实现生物功能上的关联，才能充分揭示基因功能表达谱数据蕴含的丰富信息

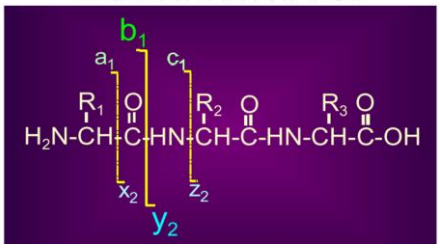
2、功能基因组信息分析的进一步工作 ：必然是获取基因调节网络的知识。探讨生物系统中各部分、各层次的相互作用，从而进入系统生物学的领域。

3、**蛋白质组学技术**：它与基因芯片一样是功能基因组的研究手段，与基因芯片不同的是，它是在蛋白质水平获取基因功能表达谱（使用二维凝胶电泳技术和测序质谱技术）是在蛋白质水平上检测基因表达的变化

二维凝胶电泳技术：把含有几千个蛋白的样品经过二维凝胶电泳进行粗分，分成大概 10 个左右一组的蛋白；因为数据库中有标准谱的存在，所以分离出的单点可以直接确定是何种蛋白

测序质谱技术：输入是几个或几十个蛋白组成的，第一步：经过特别的酶（用哪个酶切，酶切位点是固定的），把蛋白水解成片段 —> 经过印象色谱？，让小片段,根据质量重新排序 -> 小肽进行第一次测序，测整个小肽的质量（在两个电极之间加电压，每个小片给一个电荷，通过记录时间，判断质量(通过荷质比)）-> 第二次测序，把小片段用物理的方法（惰性气体冲击；激光把平面弄成高能状态，使其具有电荷）再打断，同样每个都加上电荷，再加电极，跑起来记时->得到两个量（小肽的质量和数量）->最终，测出来的序是水解后多肽的序，不是蛋白的序，把小肽的序拼起来就得到了蛋白质

二级质谱中的离子种类



问题：打断位点不都是我们想要的；电荷不是每次都只给一个；找哪个是 b 离子

缺点：测序的质谱得到的两个量（数量和质量），而比对时的标准谱，没有标准的数量值，是认为设定的；大量的计算是浪费的；无法测出新的蛋白；

4、蛋白质芯片：把蛋白质固定在芯片上，是研究蛋白质功能的，主要研究蛋白与蛋白以及蛋白与配体（药物）的相互作用

困难：得到纯样品；并且保持蛋白的天然构象

5、大规模基因功能表达谱的分析

随着人类基因组测序逐渐接近完成，人们自然会提出如下的问题：即使我们已经获得了人的完整基因图谱，那我们对人的生命活动能说明到什么程度呢？人们进一步提出了一系列由上述数据所不能说明的问题，例如：基因表达的产物是否出现与何时出现；基因表达产物的量是多少；是否存在翻译后的修饰过程，若存在是如何修饰的；基因敲除（knock-out）或基因过度表达的影响是什么；多基因差异表达与表现型关系如何等等。概括这些问题，其实质应该是：知道了核酸序列和基因，我们依然不知道它们是如何发挥功能的，或者说它们是如何按照特定的时间、空间进行基因表达的，表达量有多少

第四章 生物信息学研究内容之三（最近两三年，该领域最热和最值得关注的问题；最重要的研究方向和国际上在做什么）

一、系统生物学的出现与发展

重要科学家：Dr.leroy hood

基因组研究的三个阶段：

1、测序基因组（得到基因组的静态序列）

2、功能基因组（得到遗传密码在特定组织特定时间的表达情况）

3、系统生物学（通过特定基因的表达模式构建出生命活动的基本单位）

系统生物学：整合所有来自各个层次的资料数据（搜集数据）—> 用海量数据建立生物活动的模型（建立模型）—> 预测正常生命活动未来的发展以及对外来刺激的反应（预测行为）

后基因组研究对象的多层次：

基因组－转录组、蛋白质组、相互作用组、定位组（比如蛋白质分布在不同位置，如核糖体，等）、折叠子组、代谢组、表型组

系统生物学本质：

分子生物学（1 维）：序列（有基因）—> 结构（蛋白折叠）->功能

系统生物学（2 维）：多种生物大分子共同作用—>形成网络—>功能

系统生物学:两类参量(生物分子和分子间的相互作用)-- 形成网络 -- 构建生物活动的基本单元

系统生物学的研究示例：

1、病毒的调控过程

2、半乳糖代谢通路研究

3、心脏建模

功能基因组发展趋势：更好整合生物过程不同阶段的分散数据（基因组+转录组+蛋白质组+代谢组）；对复杂生物过程的更好的模拟；生物过程动态研究

系统生物学的研究思路：（信息整合）多信息融合构建功能图谱；

系统生物学研究方法的创新点以及研究方法:生物复杂系统的突显性规律，如钙波（不是简单的子系统的加和）；功能基因组多层次系统的贯穿特性；系统和系统，层次和层次的互相作用

方法：支持向量机；主成分分析；功能子系统的建模；多信息融合

应用：肿瘤研究

二、生物网络

1、生物大分子相互作用

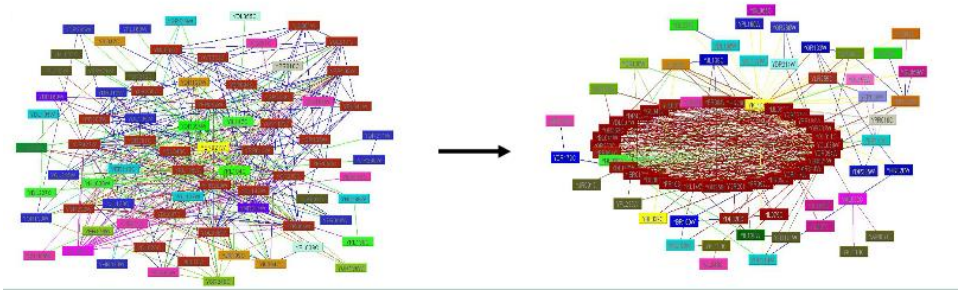
举例：蛋白相互作用；DNA 与蛋白；

目前有二十多种手段（27 种）可以得到蛋白相互作用数据，如 MS 测序质谱技术、酵母双杂交技术；方法的准确率不高，所以确定的数据得到不易，至少要两种方法同时确认

陈院士实验室工作：

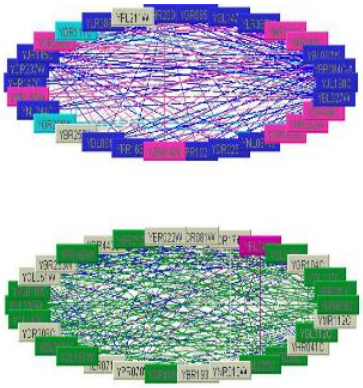
（1）用图论的方法，把蛋白网络转化为图，从而避免了极端稀疏矩阵的问题；使用**谱分析方法**（有相互作用的点被不断拉近，否则相互排斥，距离变远），需要证明方法收敛

Spectrum Analysis algorithm

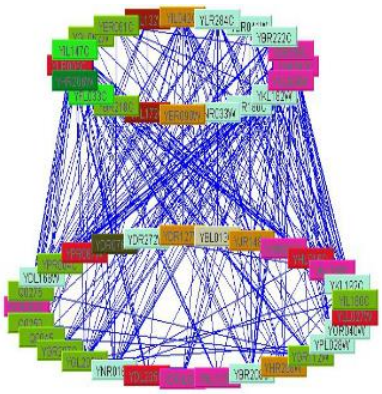


产生团结构，团内元素的相互作用远大于团之外的相互作用

产生二面图，两个集合间的所有元素之间相互作用较强，集合间元素相互作用稀疏



A Clique



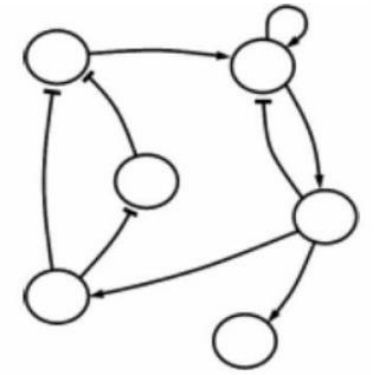
b Bipartite

实验结果解释：在完全没考虑生物背景的情况下：团 1 都是核糖体蛋白、团 4 是参与蛋白的可变剪切

（2）clustering tree：受上一个方法结果的影响（团）—〉考虑聚类的方法

2、基因表达调控网络（有向图）

定义：一组调控因子如何调控一套基因表达的过程；其中元素包括 cDNA,mRNA,NcRNA,蛋白，小分子



尖头表示增强；末端断线表示抑制；基因调控存在自调控

陈院士实验室工作：转录调控

转录模式：一个转录因子调控多个基因的转录；两个转录因子同时调控两个基因的转录；多个因子调控一个基因的转录

3、代谢网络

元素：代谢物；边：代谢物之间的反应

陈院士实验室工作：通过构造进化树的方法解决问题

4、信息传导网络

二、非编码序列，非编码 RNA，非编码基因

基因 1911：是传递遗传特征的元件

基因组 1920：一个物种基因的组合；现在看来，基因组远大于一组基因的总和，基因组包含 4 大类元件（基因，调节元件（启动子，增强子，抑制子，等），与整个基因组复制相关的元素（端粒，复制起始位点），非功能序列）

基因组构成：基因和可调节源（基因前的启示序列）占 2%；非编码序列占 98%（20%左右的 intron（基因的内含子），10%卫星 DNA（satellite），2%小微型 DNA，40%的重复序列（可迁移源），1%假基因，25%什么都不属于的非编码非重复的序列）

SNP（表现生物之间的差异）：98%的 SNP 分布在非编码序列

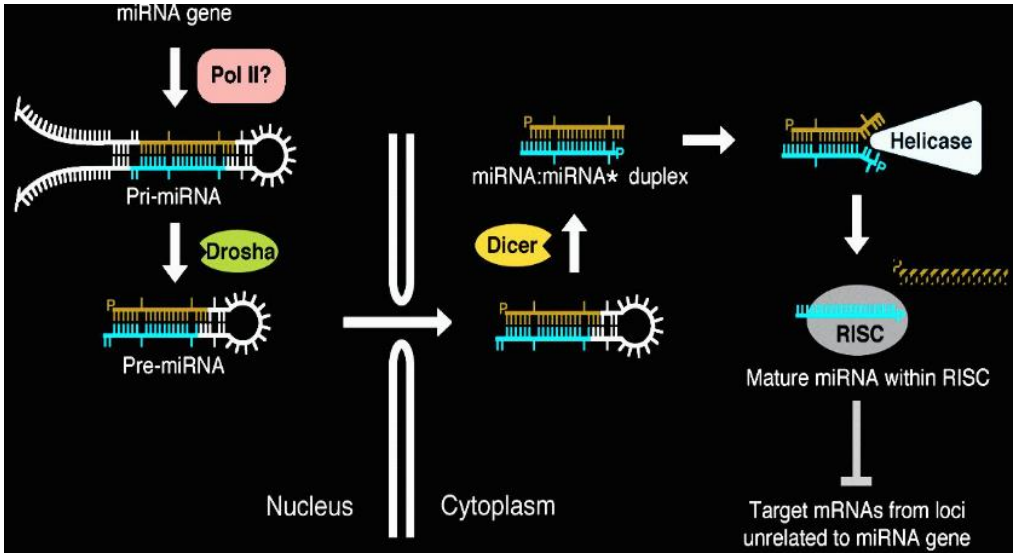
1、非编码序列（基因组水平上的）：为编码蛋白质的序列（占 97%）；基因组上不编码基因的序列

重复序列有两种：串级重复（重复单元链接在一起，如，CACACACACA）：微型 DNA，小微型 DNA；散在重复（重复是不连接在一起的）：ALU 家族，SINE

2、非编码 RNA（RNA 层次上）：是非编码序列的转录产物

非编码序列中 2%是 RNA

microRNA：RNA 干涉现象引入的，外源的叫做小的双股 RNA，内源的叫做 microRNA（21-24，在基因组上的长度 400-1000 个核苷酸） ！！microRNA 家族的发现（1000 多种）；从基因组直接转录出来的叫做 **microRNA 的前前体（pro microRNA**：400-1000 个核苷酸），经过一次酶（Drosha）切，得到 **microRNA 前体（pre micro RNA**：40-90 个核苷酸），前体从细胞核释放到细胞质里，被 Dicer 酶剪切，成为**双股 microRNA 成体**；再经过裂解酶，变成两股链，其中一股链招募了几个蛋白，形成了成熟的 **microRNA 的转录抑制复合物**（RISC），这个复合物变成工厂去识别 target 进行剪切



RNA 干涉（RNAi:RNA interference）：由 RNA 导致的基因沉默现象；是一项快速、高效、便于操作的是使靶基因失活的技术，它可以像基因敲除一样非常有效的鉴定特定基因的功能

NON-coding RNA: RNomics; Long nc RNA

3、非编码基因：某些非编码 RNA 具有明确的生物学功能，有明确功能的非编码 RNA 对应在基因组上的位置，叫做非编码基因