

非编码序列的重要性

The significance of
non-coding sequence

刘广斌

The University of Southern
Queensland



Guang Bin Liu presented 'The significance of non-coding sequences' in the faculty of Life Sciences Shan Dong University (15/10/2010)

从一个学说说起：

关于非编码RNA
(ncRNA)

基因组

编码蛋白质的基因

不编码蛋白质的基因

外显子 (exons)

内含子 (introns)

转录

编码RNA

+

非编码RNA (ncRNA)

蛋白质

功能?

一个基本的传统信念

基因=编码蛋白的DNA序列

遗传信息的传递： DNA→RNA→蛋白质

所以： 基因=蛋白质

基因遗传的全部或近乎全部产物都是由蛋白质来执行的。

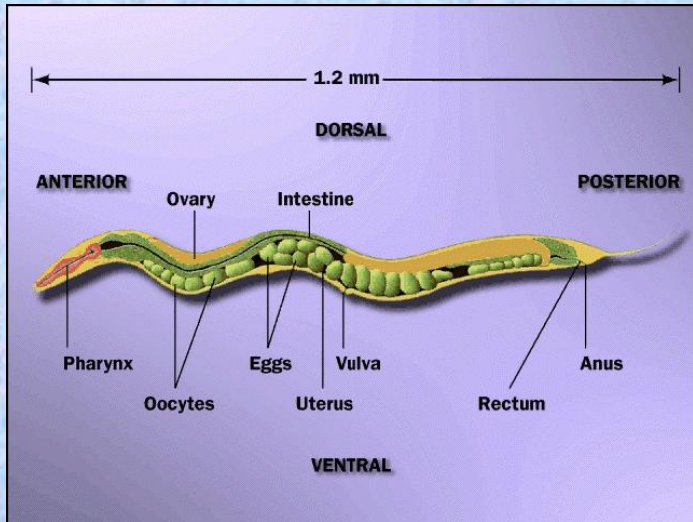
从后生动物到哺乳动物，蛋白编码基因的数量和功能范围大致相似，约两万个（最近的结果），而且这些基因编码的蛋白功能大多数与人类相似！

例如：

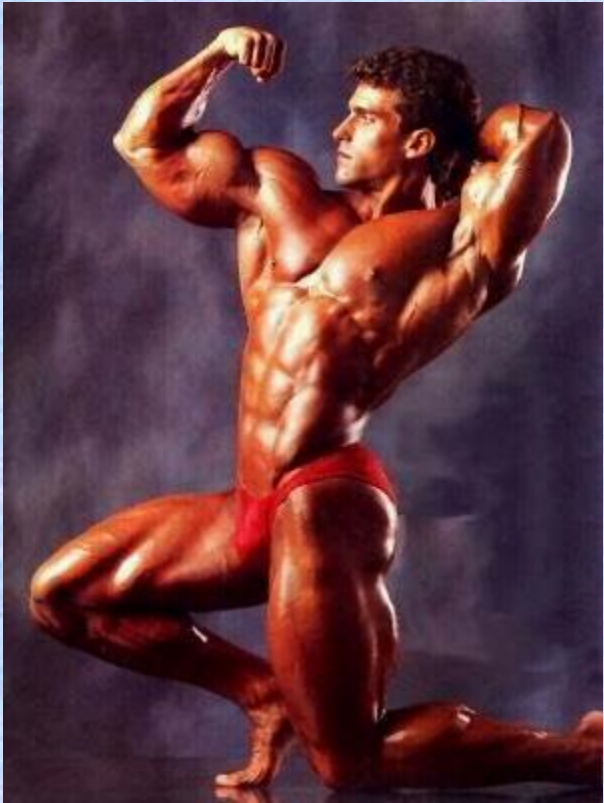
透明线虫（*Caenorhabditis elegans*）只有约1千个细胞：肌肉，神经细胞数量极少

人体约有30万亿到100万亿细胞，仅想象一下面部肌肉要如何表达人类的情感，就可推测人体的结构是无法想象的复杂。

与此可见，动物发生过程中的最基本元件是在数百万年前就已经建立起来了。



||| ? ≠



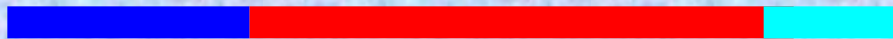
既然蛋白编码基因的数量尺度在线虫和人类之间没有大的区别，显然在高等生物中更为复杂的调控信息就存在于非编码序列中。

非编码序列的数量以及复杂性与生物的复杂性成正比，就足以说明这一点。请见下表：

粘菌	40%
线虫	75%
果蝇	83%
鱼类	89%
人类	98.8%

原核生物

5 ' 基因 3 '

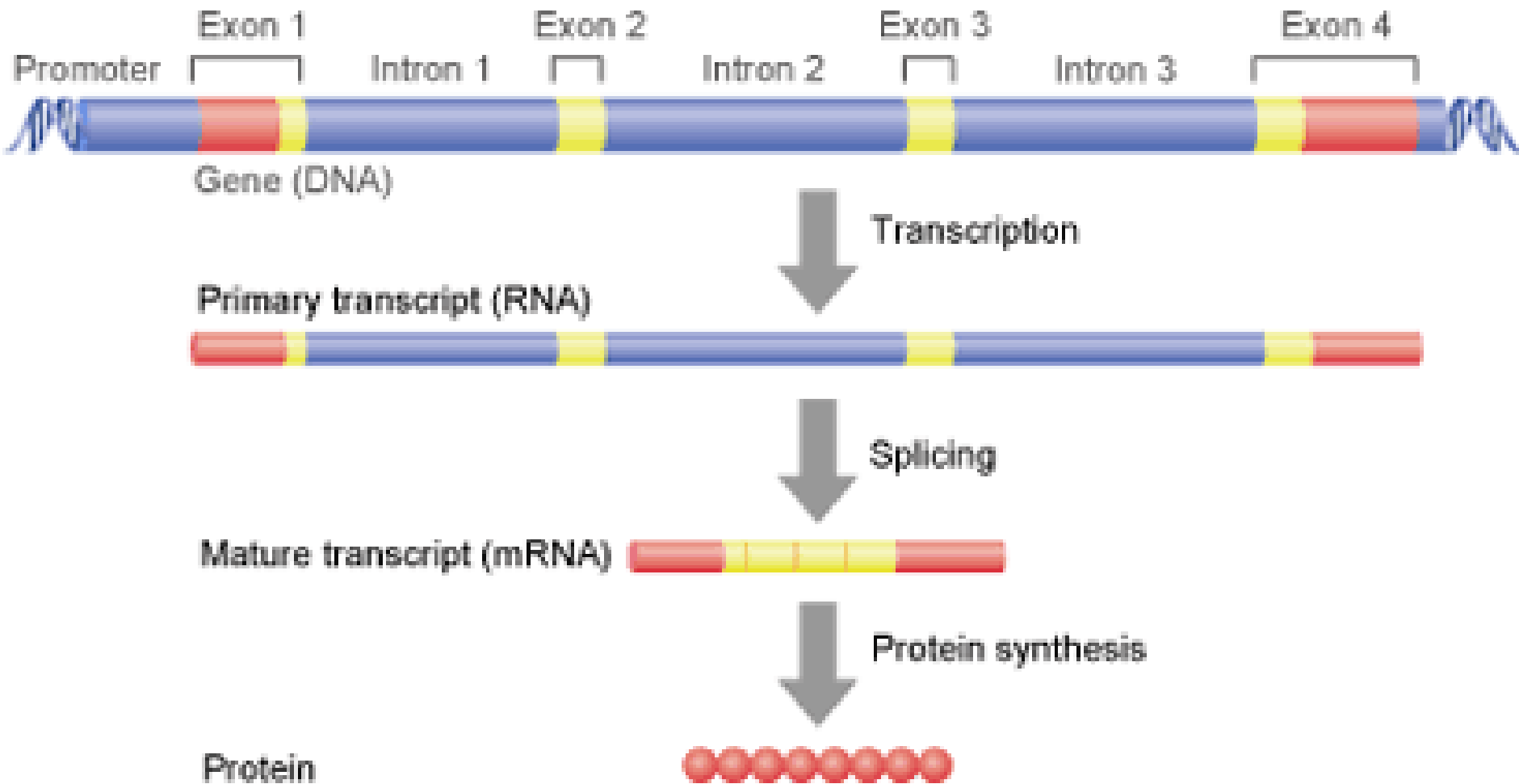


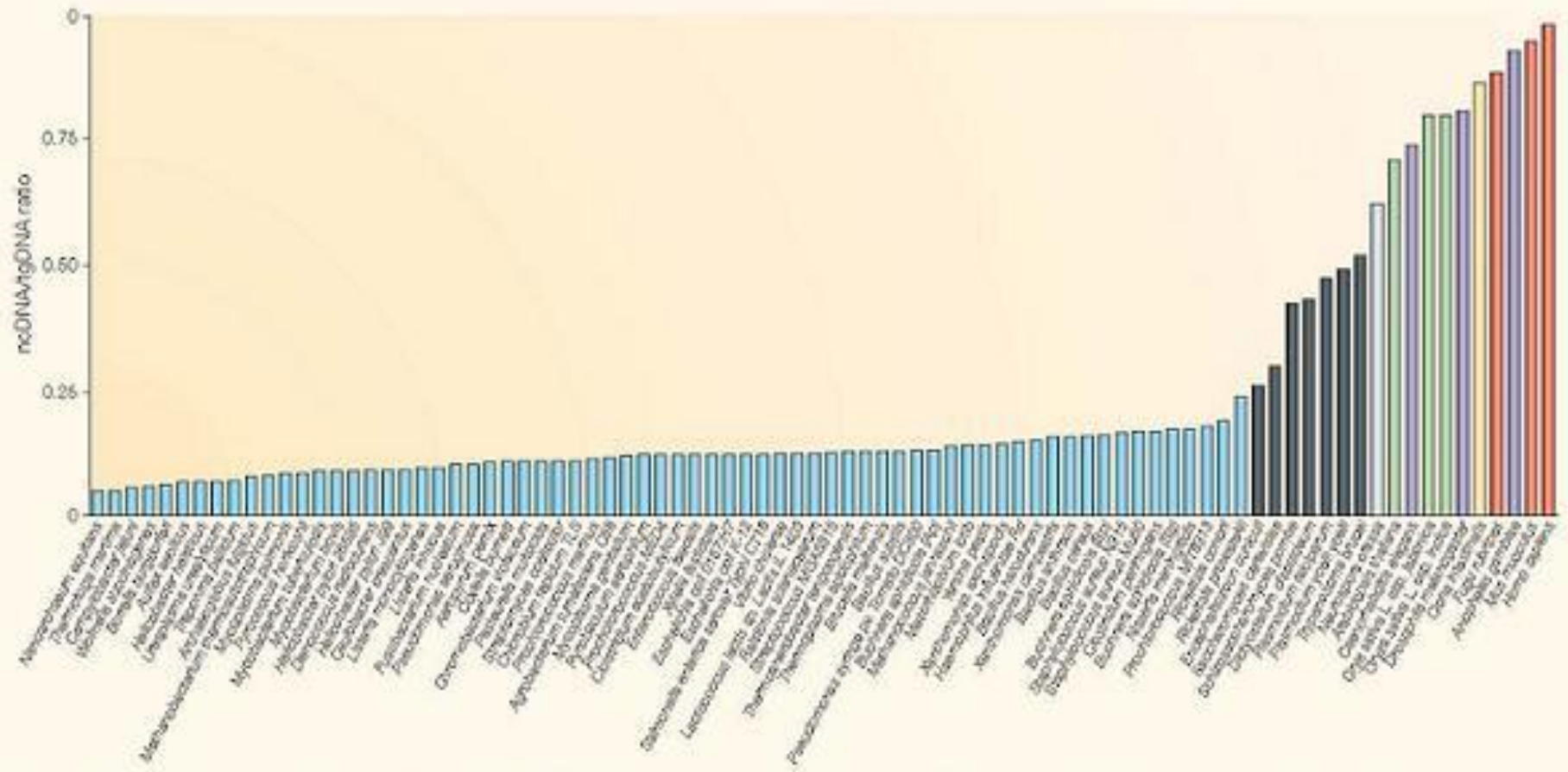
调节序列

调节序列

蛋白质： 细胞基本功能和结构
 细胞内调节因子
 对细胞外环境变化信号反应
 对细胞活动进行控制

Structure of a Gene





科学史上有过这样的例子：一个合理但未经证实的推论一旦占据了主导地位，便很能‘经得起’质疑，也很难被推翻！

在分子生物学领域也是这样，认为基因就是蛋白质的同义语的概念曾成为了本学科信念大厦的基石，而这个基石却产自于未经严格审查就舒舒服服的被接受的断言。从此人们以为有了真理，并对于任何‘挑战’真理的研究结果，观察，建议嗤之以鼻，不予理睬。不管是客观的对于正统的质疑，还是面对令人惊讶的现象的新解释，统统无效！

1977年内含子（Intron）的发现，以及人们不可思议的反应就是一个很好的例子。

在分子生物学的历史上，这大概是一个最令人惊讶的故事：人们发现在高等生物体内，基因内编码蛋白的序列被许多更长的非编码序列分成了像马赛克样的若干段。关于这些非编码的作用无非有两种可能解释：有功能，或无功能！

如今看来前者不但同样合理，而且应该有更为迷人的价值和研究意义。

两种解释必然导致截然不同的逻辑结果。

结果人们接受了后者，而抛弃了前者，仅仅因为它们‘不编码’。

内含子从而被描述为沉默的**DNA** 或 无意义序列。

这样的结论来源于未经验证但却根深蒂固的假定：基因所含的信息必需被表达为蛋白！

内含子虽然被转录，但却随后就被切掉了，离解了！

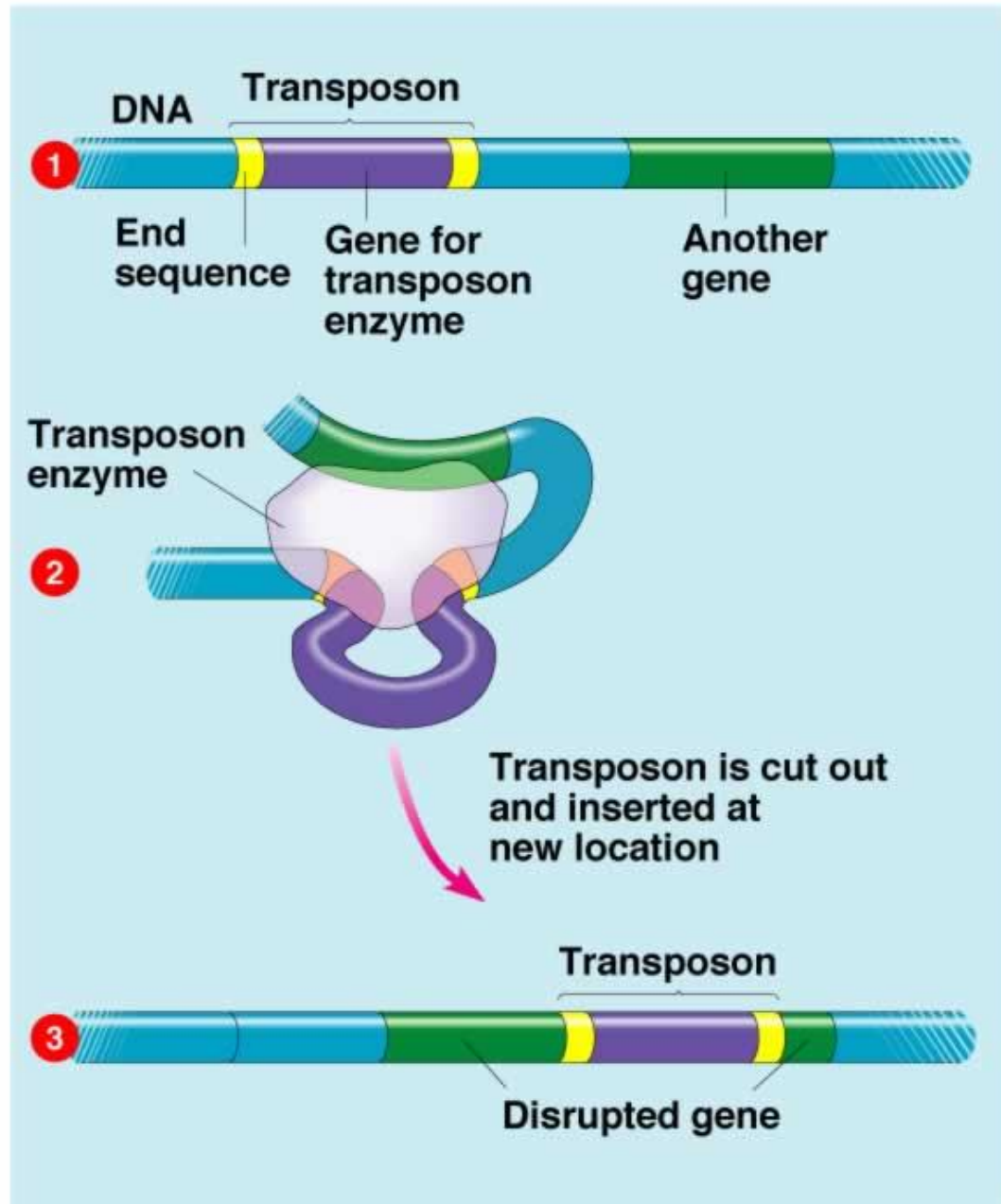
因为人们必须解释内含子存在的理由，于是就不得不推测它们为早期RNA世界蛋白编码片断的残留物，或者自私DNA（selfish DNA）转移过程的产物。

认为非编码序列的大部分是进化的废物，原因之一是：我们基因组成分的一半左右来源于DNA转座子

(Transposons) 序列的换位

(transposition)

这甚至也是Crick的观点。



大量的基因被转录时，内含子（Intron）形成**RNA**。然后内含子被切掉，丢弃- 因为它们无用！这个结论不很荒唐吗？

其实，这些被转录了的**RNA** 序列是具有功能的。非常有趣的现象！

绝大多数基因的信息是被蛋白表达的！

这个认识，在细菌是对的。但是在真核生物，就不对了。

从**1977**年以来，许多实事显示存在着大量的掩藏了的调节**RNA**，它们参与指导分化和发育的方向。揭开这些隐秘，我们目前才处于开始阶段。

非编码RNA（Non-coding RNA）是真核生物进化的关键

非编码RNA是不翻译成蛋白质的功能性的RNA，如：tRNA, rRNA, snoRNA, microRNA, siRNA 等

基因组

两个相互关联的设计

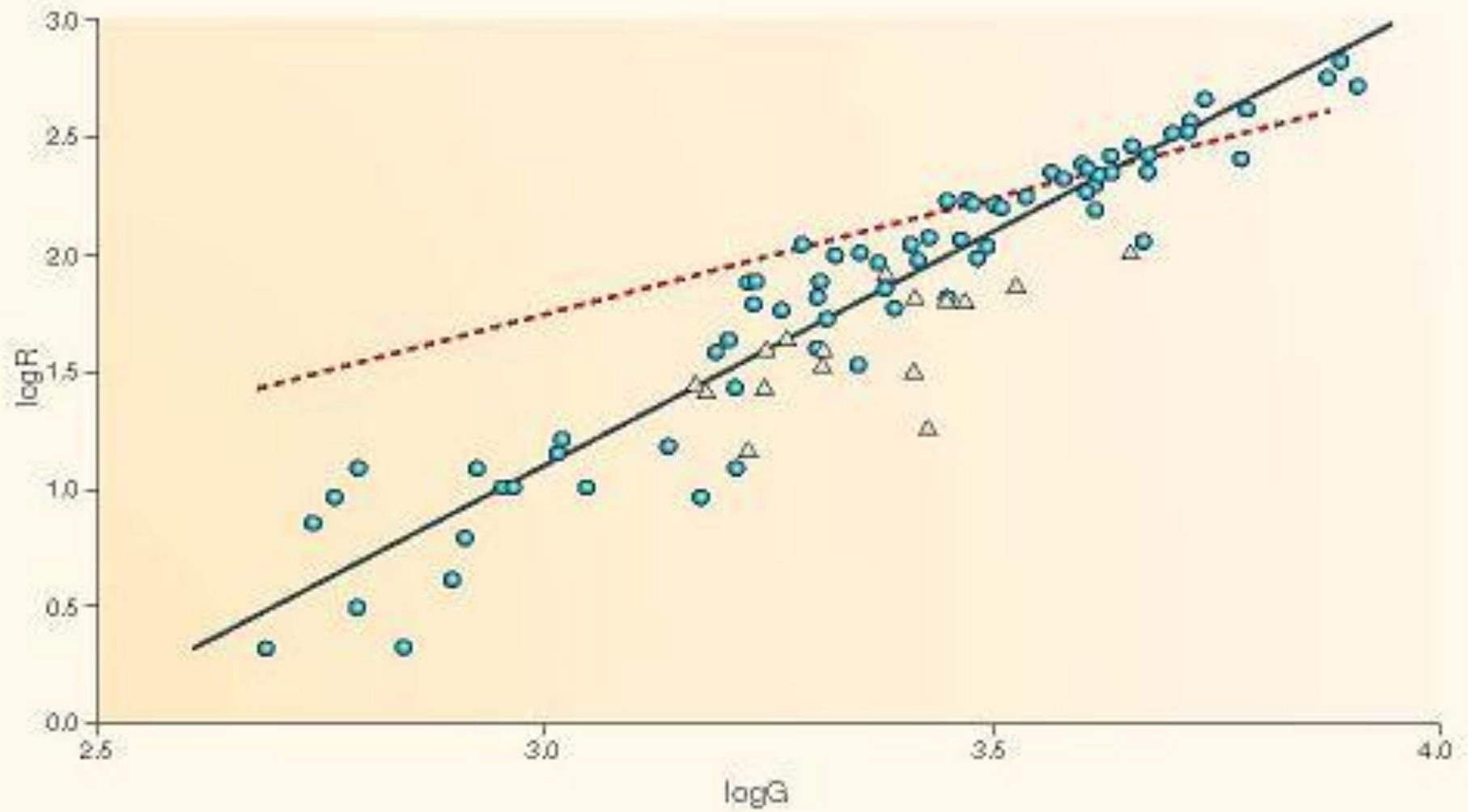
- 1 结构和功能成分(蛋白质等)
- 2 如何编排组合以构成高一级的组织+调节功能的系统。

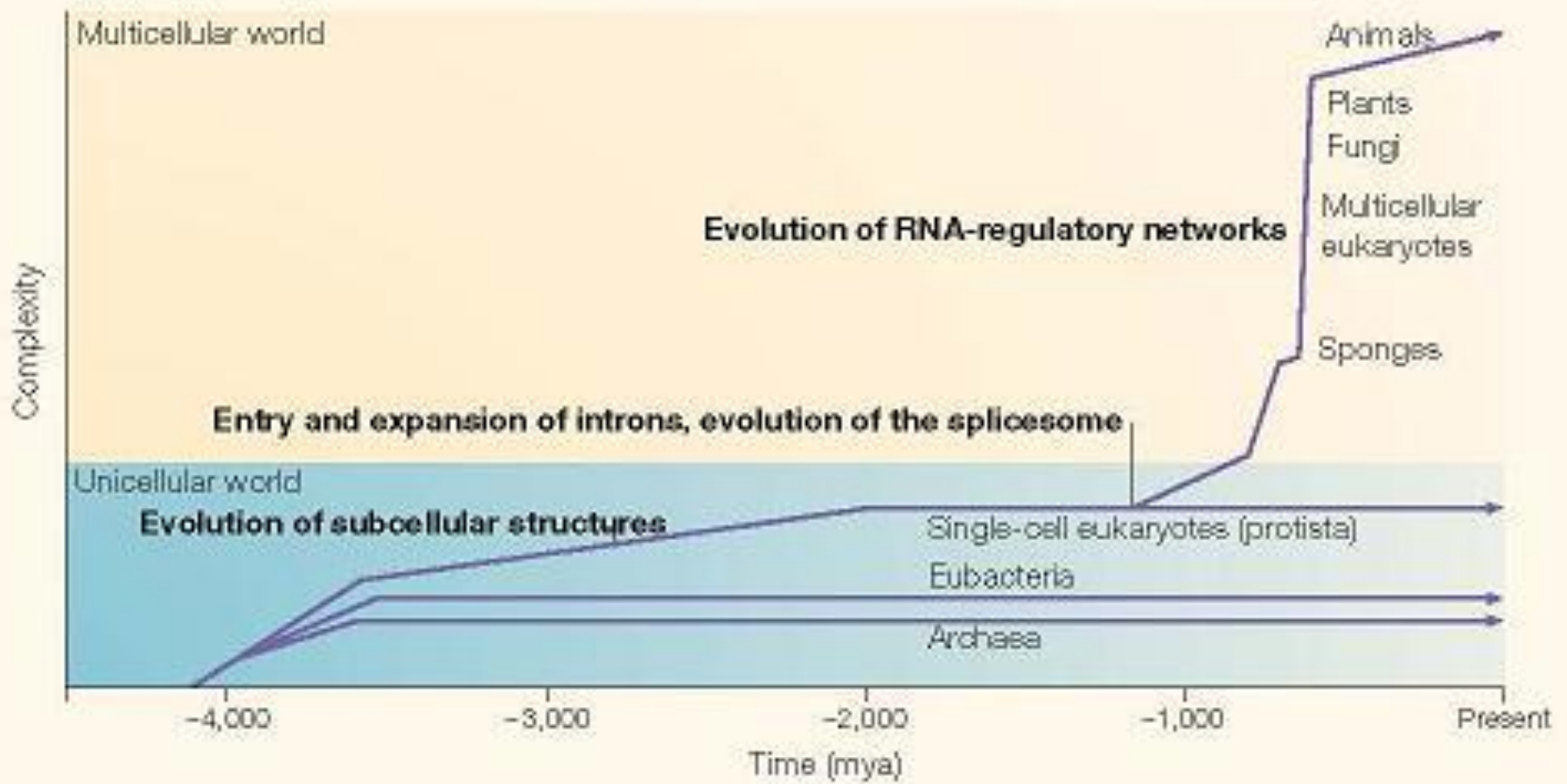
人体中有亿万个结构精当配位准确的细胞。在四维空间位置与形态功能上达到如此程度的和谐精当，是否仅靠蛋白调节因子的组合与调控序列的相互作用就可达到？很少有人试图验证这样的假设。

但有一个伴生的问题是：在个体中多少调节信息可以有效地被整合？

例如，在不同的基因拼接组合中，要产生不同的产物需要多少调节信息？随着生物系统复杂性的升高，又需要多大的控制尺度才能有效调度这些控制因子？

复杂性：控制个体或系统发生与运转所需要的最小信息量





真核生物平行表达两种形态的信息

蛋白质+调节RNA序列

ncRNA在进化选择过程中处于优势得以扩充
->一部分或许多基因仅产生RNA信号->高级
调节信号。

证据：复杂真核生物比简单真核生物明显
有更多的内含子，有更多数量的ncRNA转录，
占到了哺乳类总转录数量的一半以上。

基因转录和拼接后



大量的小信号单元



與不同的目的物结合



影响染色质的构架，转录，选择性拼接，
翻译效率及RNA的稳定性等等

这是一个数字调控系统。

如果RNA中介的调控是实，那麽为甚麽这个系统如此长时间未被认识到？

- 1、我们没有准备接受这样的概念。
- 2、这个系统在生物化学上是无法观察的—RNA是易变的。许多（即使不是全部）RNA信号转瞬即逝且很微弱，似有似无，虚虚实实的游离于实验室凝胶电泳的可见画面内外。
- 3、这个调控系统很精细：
 - 蛋白编码区突变→巨变
 - 调控区突变→可能很细微

如果不是在线虫中发现lin-4和let-7以及它们与RNAi途径的联系（这本身只是一个意外发现），都不知道我们至今能不能意识到微RNA存在于真核细胞中。

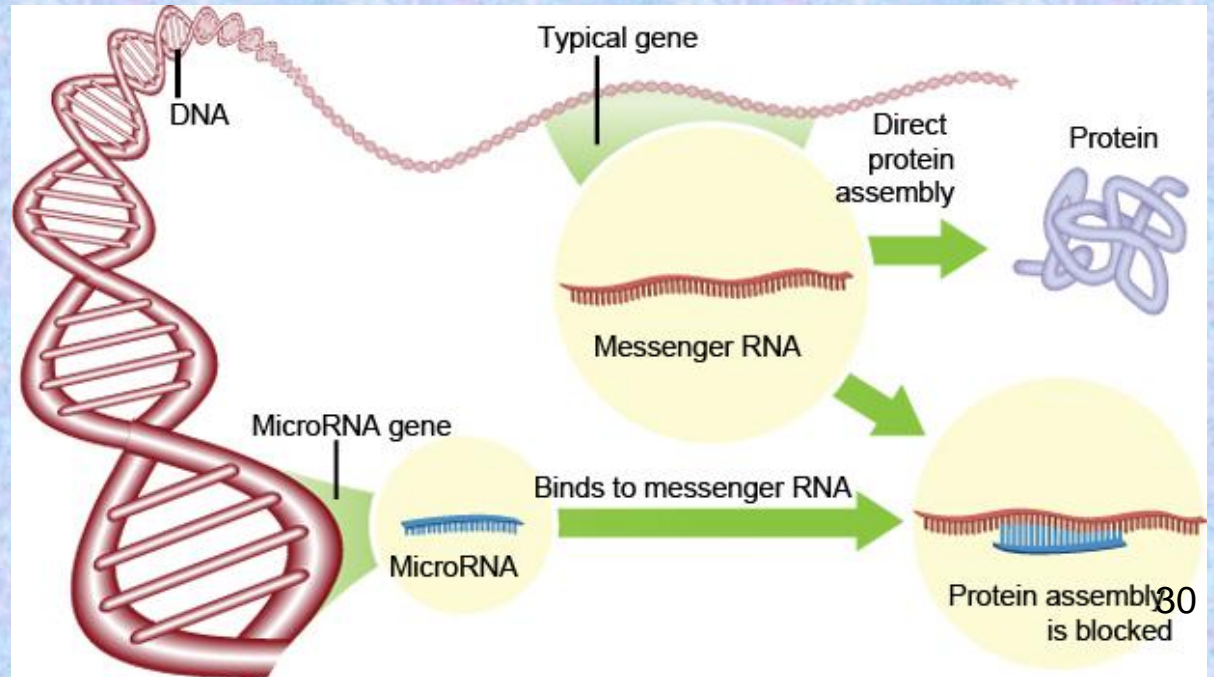
如何对ncRNA进行功能分析？

分子生物学：

查找同源序列，确定进化表现型，细胞内定位，在转基因动物中对部分表达和异常表达基因的测定。大量相关试验正在进行。

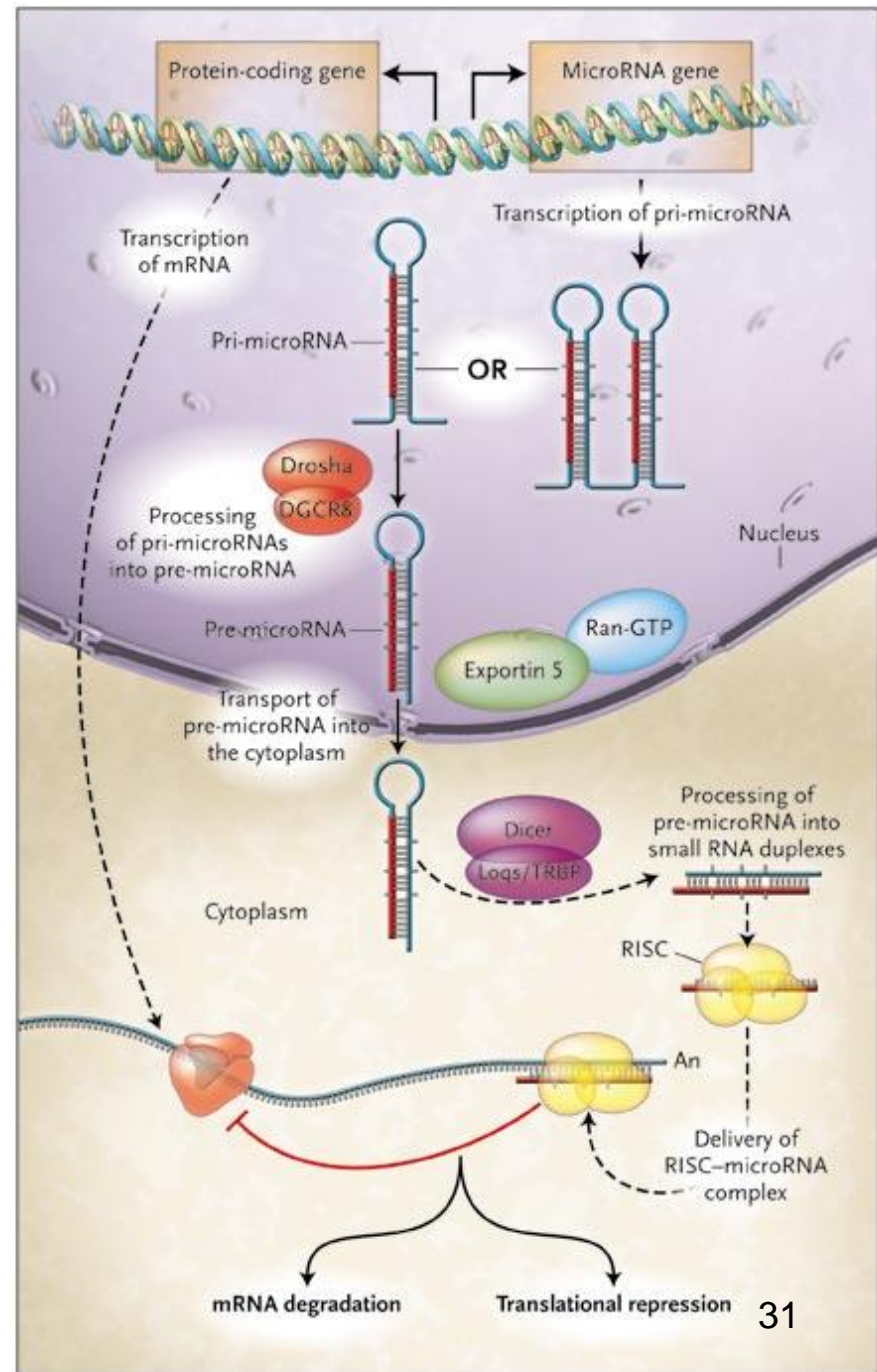
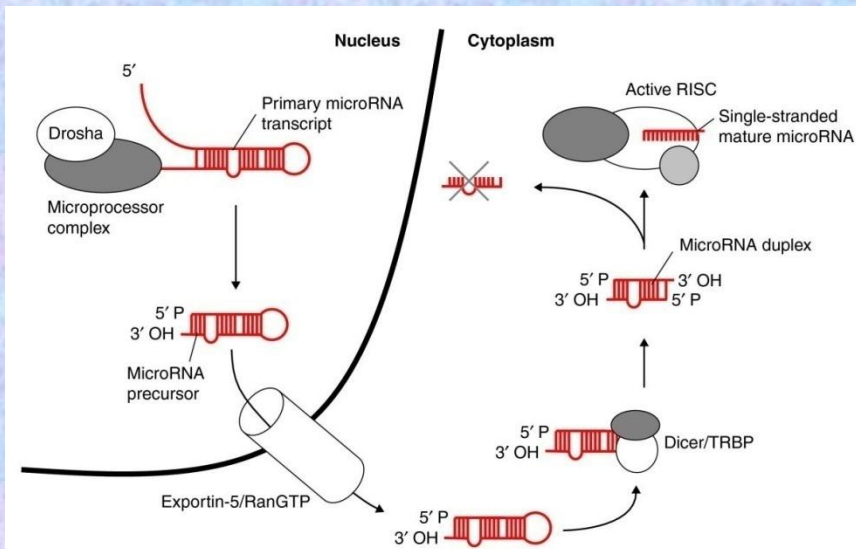
ncRNA 参与调解的两个可能机制：

1. **microRNA**调节。现已知道，**microRNA**参与调节动物和植物的所有发育过程。它们没有催化功能，因为它们只是一些大约**22**个碱基长短的序列，它们以其他**RNA**为靶的。在进化过程中，位于不同细胞的**microRNA**被以很灵活的方式激活起来，去影响不同的靶的。



这样就不需要针对每一个调节过程设计一个蛋白或一套蛋白复合体。

相反，在基因组内存在一个简洁，灵活的RNA 机制，它只需要使用一套简单通用的复合体便可以针对不同的靶的进行调节。



2. 染色质结构和遗传记忆的调节

这个调节对于复杂生物的发育至关重要。**DNA**甲基化酶以及一套相对通用的酶和酶复合体以不同的方式调控着组蛋白-它们大约有**60**个。

然而如何选择它们？何时选择， 调控哪里？

越来越多的证据指出在分化发育的过程中需要选择使用不同的酶时，**RNA**提供序列特异性的信号并可能在适当的位置充当受体， 从而正确的蛋白复合体能够在适当的位置发挥作用。

现有的证据已经指出数十个**microRNA** (当然会有更多被发现)，它们或与某些细胞器的正常功能有关，或与一些器官如眼睛的发育有关联，也有一些被证明参与疾病的过程，如心血管疾病，癌症等。

这也正好可以解释为甚麽全基因组研究结果中大多数基因的多态性 (**polymorphism**) 与疾病的联系是发生在非编码区域。

有些疾病由单一基因变异引起，比如传统的单基因疾病肇源于蛋白编码区的突变。

但有更多疾病具有更复杂的机制：它们的成因常常有许多基因成分参与。

然而毫不奇怪，大多数影响人体性状的变异发生在调节区。**90%**以上的结果支持上述看法，而且许多突变位点远离相应的蛋白编码序列。

请注意**microRNA** 的表达具有时间和空间上的高度特异性：有人调查了约**1000**个 **microRNA** 在神经系统的表达状况。发现约半数具有高度准确的空间表达：有的只在海马的一个小区，有的只在大脑皮层特定的层次表达，也有些只在小脑的浦肯野氏细胞内表达。

而且可以肯定地说，这些**microRNA** 是被转运到细胞内的特殊区域进行工作的。

于是我们可以这样认为：**microRNA**的表达是极为精确的，准确到特定的细胞，准确到细胞内的特定位置。

这个系统不仅允许相当可观的进化灵活性，从而在日益复杂的生物系统的网络内执行精准的调节，而且保证了在高度复杂系统内的有效调节处于低成本状态。

自从产生了意识，人便一直在追问我是谁，我所生存的周围环境是怎麼回事？在理解自然，预测未来的过程中，深刻地理解人类自己原本就是这题中应有之义。

现代科学，通过生物信息学所要达成的已不再仅仅是预测未来和理解大自然，它也包括重塑人类自己的过程。清楚地知道我们所想要的是甚麽，并找到合适的方法去实现它，已日益成为重要的课题。

谢 谢！