CSE427
Computational Biology

http://www.cs.washington.edu/427

Larry Ruzzo

Winter 2008

He who asks is a fool for five minutes, but he who does not ask remains a fool forever.

-- Chinese Proverb

This week

Admin

Why Comp Bio?

The world's shortest Intro. to Mol. Bio.

Admin Stuff

Course Mechanics & Grading

- Reading
- In class discussion
- Homeworks
  - reading
  - paper exercises
  - programming
- Small Project?
- No exams

Digression: Evolution & scientific literacy

"human beings, as we know them, developed from earlier species of animals"
(avoiding the now politically charged word "evolution")

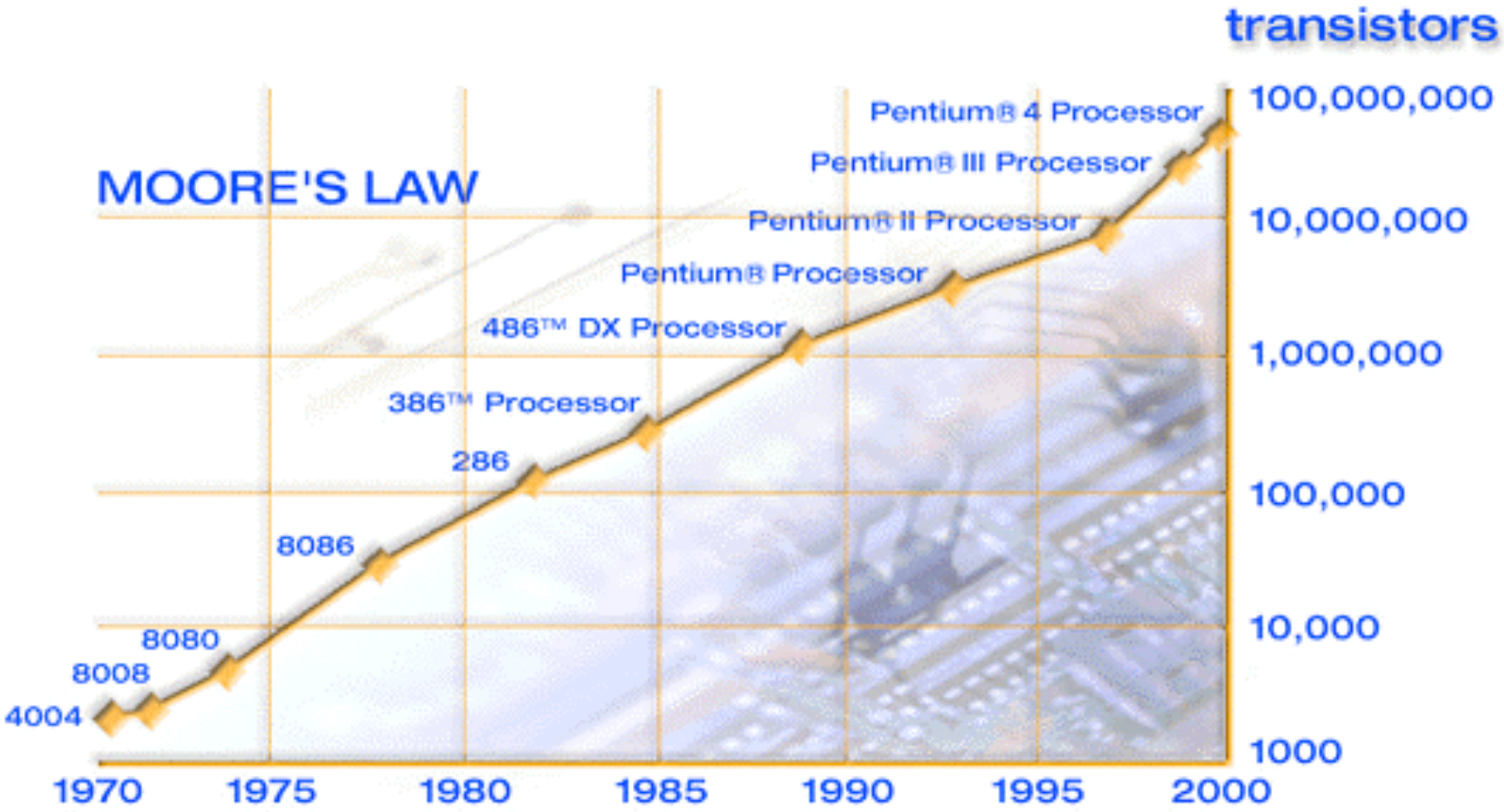
from 1985 to 2005, the % of Americans

- rejecting: declined from 48% to 39%
- accepting: also declined 45% to 40
- uncertain: increased 7% to 21%

In a 2005 survey,the proportion of adults who accept evolution in 34 European countries and Japan, the United States ranked 33rd, just above Turkey.

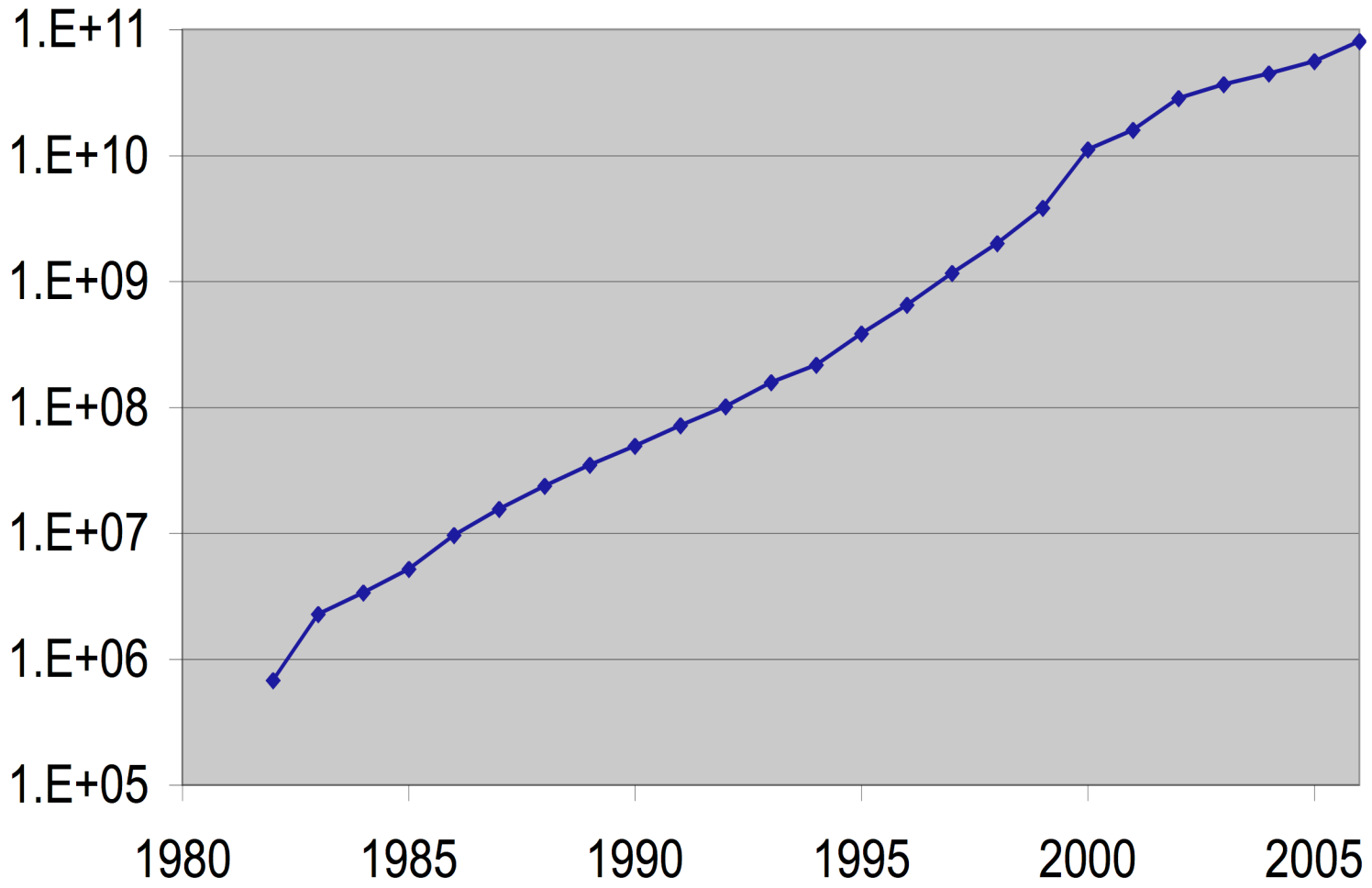
http://biology.plosjournals.org/perlserv/?request=get-document&doi=10.1371/journal.pbio.0040167

Background & Motivation



Source: http://www.intel.com/research/silicon/mooreslaw.htm

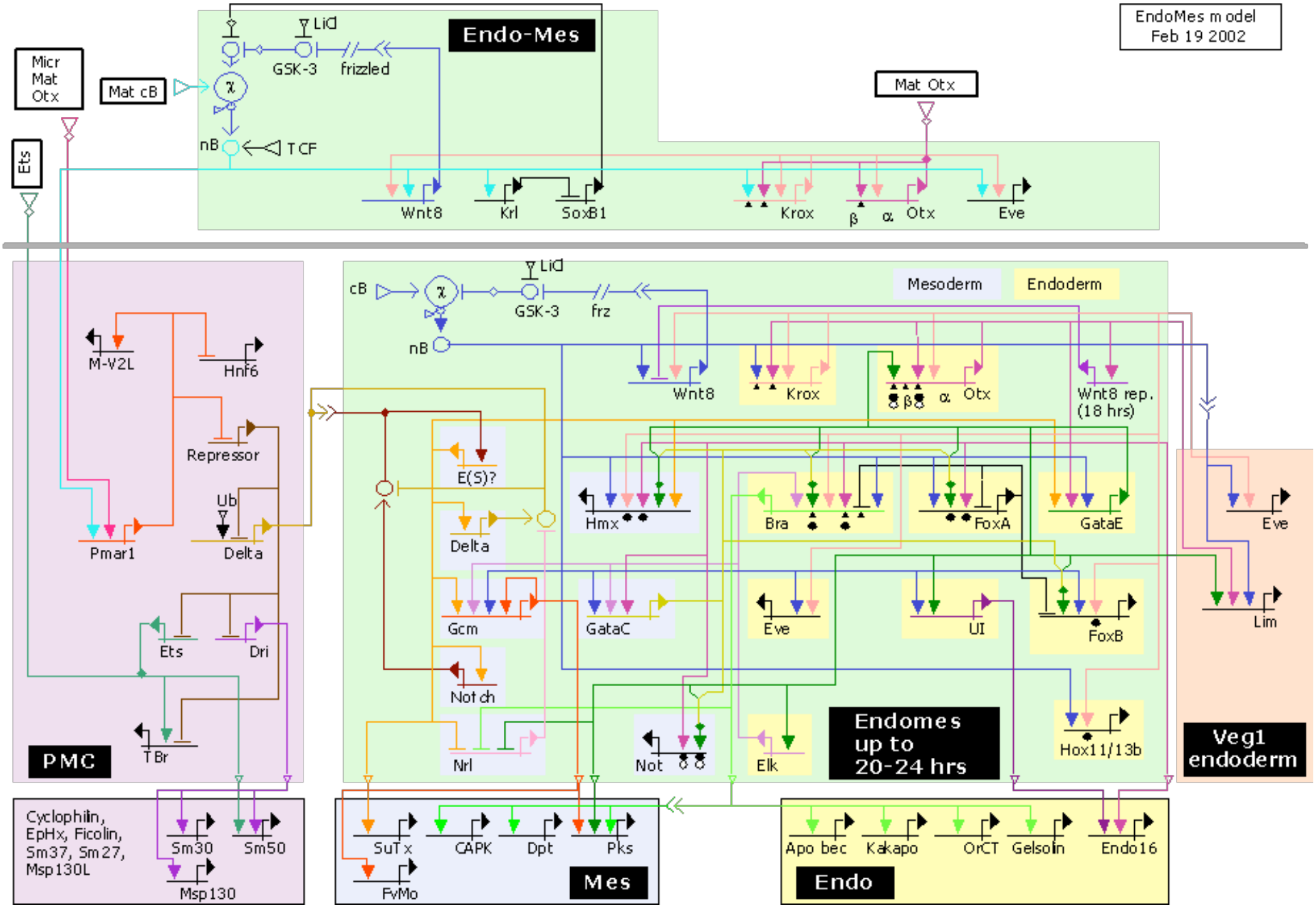
Growth of GenBank (Base Pairs)



Source: http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html

The Human Genome Project

```
   1 gagcccggcc cgggggacgg gcggcgggat agcgggaccc cggcgcggcg gtgcgcttca
  61 gggcgcagcg gcggccgcag accgagcccc gggcgcggca agaggcggcg ggagccggtg
 121 gcggctcggc atcatgcgtc gagggcgtct gctggagatc gccctgggat ttaccgtgct
 181 tttagcgtcc tacacgagcc atggggcgga cgccaatttg gaggctggga acgtgaagga
 241 aaccagagcc agtcgggcca agagaagagg cggtggagga cacgacgcgc ttaaaggacc
 301 caatgtctgt ggatcacgtt ataatgctta ctgttgccct ggatggaaaa ccttacctgg
 361 cggaaatcag tgtattgtcc ccatttgccg gcattcctgt ggggatggat tttgttcgag
 421 gccaaatatg tgcacttgcc catctggtca gatagctcct tcctgtggct ccagatccat
 481 acaacactgc aatattcgct gtatgaatgg aggtagctgc agtgacgatc actgtctatg
 541 ccagaaagga tacataggga ctcactgtgg acaacctgtt tgtgaaagtg gctgtctcaa
 601 tggaggaagg tgtgtggccc caaatcgatg tgcatgcact tacggattta ctggacccca
 661 gtgtgaaaga gattacagga caggcccatg ttttactgtg atcagcaacc agatgtgcca
 721 gggacaactc agcgggattg tctgcacaaa acagctctgc tgtgccacag tcggccgagc
 781 ctggggccac ccctgtgaga tgtgtcctgc ccagcctcac ccctgccgcc gtggcttcat
 841 tccaaatatc cgcacgggag cttgtcaaga tgtggatgaa tgccaggcca tccccgggct
 901 ctgtcaggga ggaaattgca ttaatactgt tgggtctttt gagtgcaaat gccctgctgg
 961 acacaaactt aatgaagtgt cacaaaaatg tgaagatatt gatgaatgca gcaccattcc
1021 ...
```


EndoMes m odel
Feb 19 2002
Endo-Mes
LiCl
GSK-3
frizzled
Micr
Mat
Otx
Mat cB
nB
TCF
Ets
Mat Otx
Wnt8
Krl
SoxB1
Krox
β
α
Otx
Eve
cB
GSK-3
frz
nB
LiCl
Mesoderm
Endoderm
Wnt8
Krox
β
α
Otx
Wnt8 rep.
(18 hrs)
M-V2L
Hnf6
Repressor
Ub
Pmar1
Delta
E(S)?
Delta
Hmx
Bra
FoxA
GataE
Eve
Gcm
GataC
Eve
UI
FoxB
Lim
Ets
Dri
Notch
TBr
Nrl
Not
Elk
Hox11/13b
PMC
Endomes
up to
20-24 hrs
Veg1
endoderm
Cyclophilin,
EpHx, Ficolin,
Sm37, Sm27,
Msp130L
Sm30
Sm50
Msp130
SuTx
CAPK
Dpt
Pks
FvMo
Mes
Apo bec
Kakapo
OrCT
Gelsolin
Endo16
Endo


The sea urchin *Strongylocentrotus purpuratus*

Goals

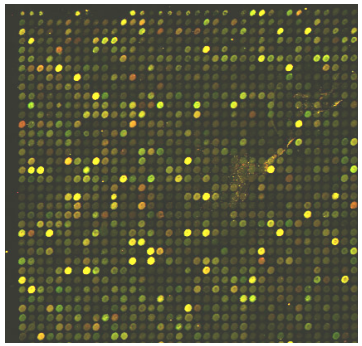
Basic biology

Disease diagnosis/prognosis/treatment

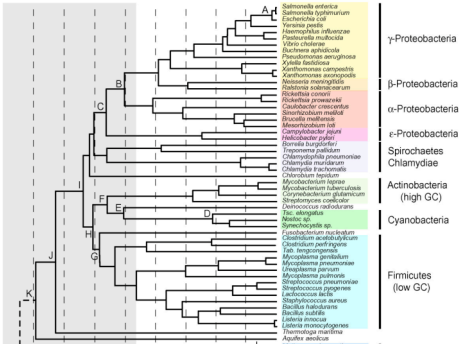
Drug discovery, validation & development

Individualized medicine

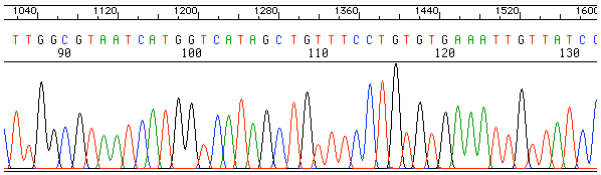
…

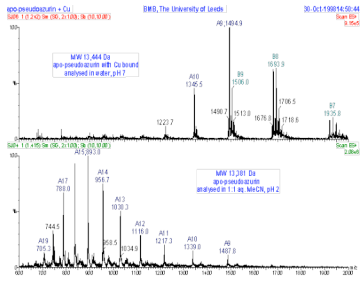


“High-Throughput BioTech”



- Sensors
  - DNA sequencing
  - Microarrays/Gene expression
  - Mass Spectrometry/Proteomics
  - Protein/protein & DNA/protein interaction
- Controls
  - Cloning
  - Gene knock out/knock in
  - RNAi





***Floods* of data**

“Grand Challenge” problems

What's all the fuss?

The human genome is "finished"…

Even if it were, that's only the beginning

Explosive growth in biological data is revolutionizing biology & medicine

"All pre-genomic lab techniques are obsolete"

(and computation and mathematics are crucial to post-genomic analysis)

CS Points of Contact & Opportunities

- Scientific visualization
  - Gene expression patterns
- Databases
  - Integration of disparate, overlapping data sources
  - Distributed genome annotation in face of shifting underlying genomic coordinates
- AI/NLP/Text Mining
  - Information extraction from journal texts with inconsistent nomenclature, indirect interactions, incomplete/inaccurate models,…
- Machine learning
  - System level synthesis of cell behavior from low-level heterogeneous data (DNA sequence, gene expression, protein interaction, mass spec,…)
- ...
- Algorithms

Computers in biology: Then & now

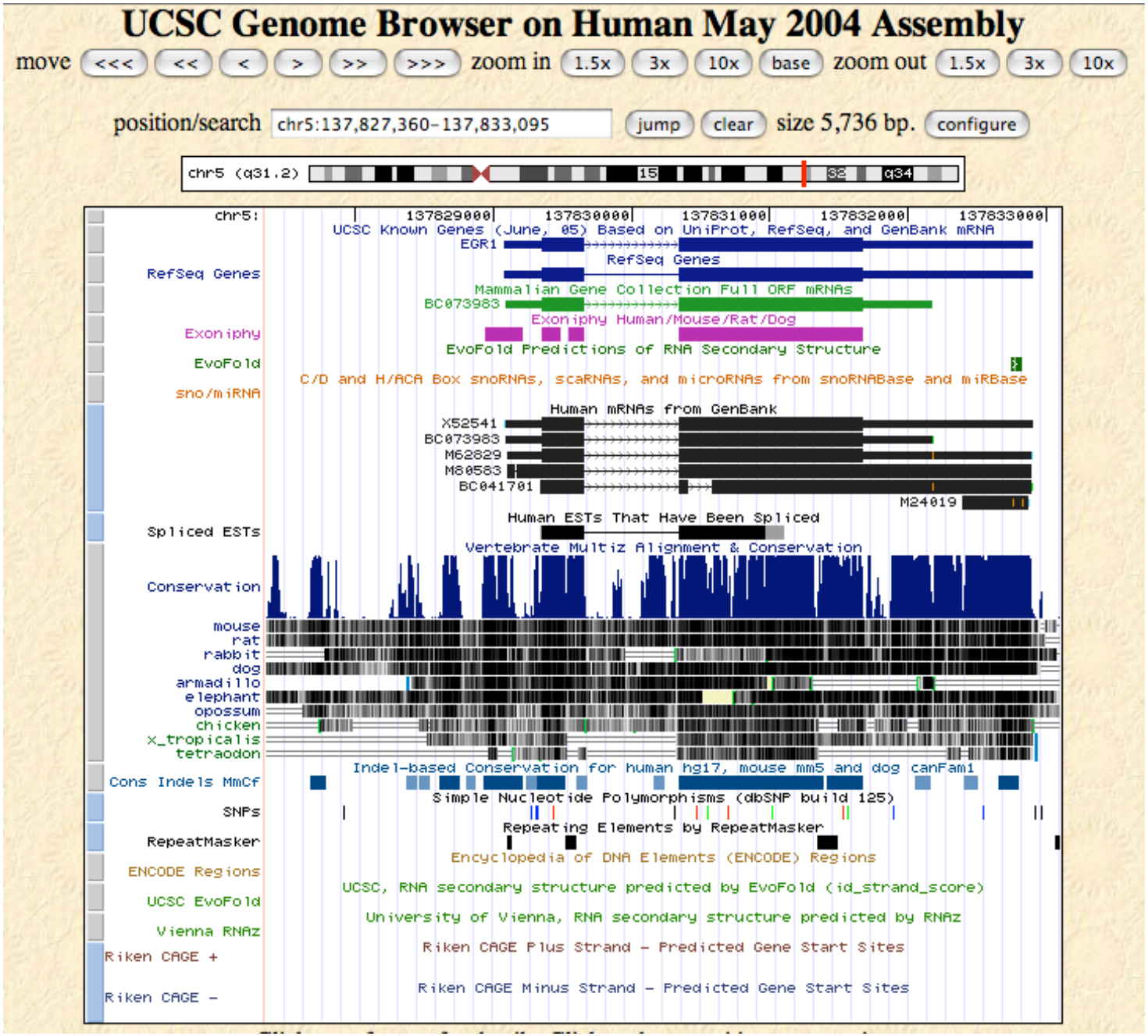
Trends in Biochemical Sciences
Volume 12 , 1987, Pages 279-280

Microfile

Sequence alignment by word processor

D. Ross Boswell

Department of Haematological Medicine, University of Cambridge School of Clinical Medicine, Addenbrooke's . Road, Cambridge CB2 2QL, UK


UCSC Genome Browser on Human May 2004 Assembly
move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x
position/search chr5:137,827,360-137,833,095 jump clear size 5,736 bp. configure
chr5 (q31.2)
chr5:
137829000 137830000 137831000 137832000 137833000
UCSC Known Genes (June, 05) Based on UniProt, RefSeq, and GenBank mRNA
EGR1
RefSeq Genes
RefSeq Genes
Mammalian Gene Collection Full ORF mRNAs
BC073983
Exoniphy Human/Mouse/Rat/Dog
Exoniphy
EvoFold Predictions of RNA Secondary Structure
EvoFold
C/D and H/ACA Box snoRNAs, scaRNAs, and microRNAs from snoRNABase and miRBase
sno/miRNA
Human mRNAs from GenBank
X52541
BC073983
M62829
M80583
BC041701
M24019
Human ESTs That Have Been Spliced
Spliced ESTs
Vertebrate Multiz Alignment & Conservation
Conservation
mouse
rat
rabbit
dog
armadillo
elephant
opossum
chicken
x_tropicalis
tetraodon
Indel-based Conservation for human hg17, mouse mm5 and dog canFam1
Cons Indels MmCf
Simple Nucleotide Polymorphisms (dbSNP build 125)
SNPs
Repeating Elements by RepeatMasker
RepeatMasker
Encyclopedia of DNA Elements (ENCODE) Regions
ENCODE Regions
UCSC, RNA secondary structure predicted by EvoFold (id_strand_score)
UCSC EvoFold
University of Vienna, RNA secondary structure predicted by RNAz
Vienna RNAz
Riken CAGE Plus Strand - Predicted Gene Start Sites
Riken CAGE +
Riken CAGE Minus Strand - Predicted Gene Start Sites
Riken CAGE -

Published online 9 July 2007

Nucleic Acids Research, 2007, Vol. 35, No. 14 ***4809–4819***
doi:10.1093/nar/gkm487

Identification of 22 candidate structured RNAs in bacteria using the CMfinder comparative genomics pipeline

Zasha Weinberg[1,*], Jeffrey E. Barrick[2,3], Zizhen Yao[4], Adam Roth[2], Jane N. Kim[1], Jeremy Gore[1], Joy Xin Wang[1,2], Elaine R. Lee[1], Kirsten F. Block[1], Narasimhan Sudarsan[1], Shane Neph[5], Martin Tompa[4,5], Walter L. Ruzzo[4,5] and Ronald R. Breaker[1,2,3]

[1]Department of Molecular, Cellular and Developmental Biology, [2]Howard Hughes Medical Institute, [3]Department of Molecular Biophysics and Biochemistry, Yale University, Box 208103, New Haven, CT 06520-8103, USA
[4]Department of Computer Science and Engineering and [5]Department of Genome Sciences, University of Washington, Box 352350, Seattle, WA 98195-2350, USA

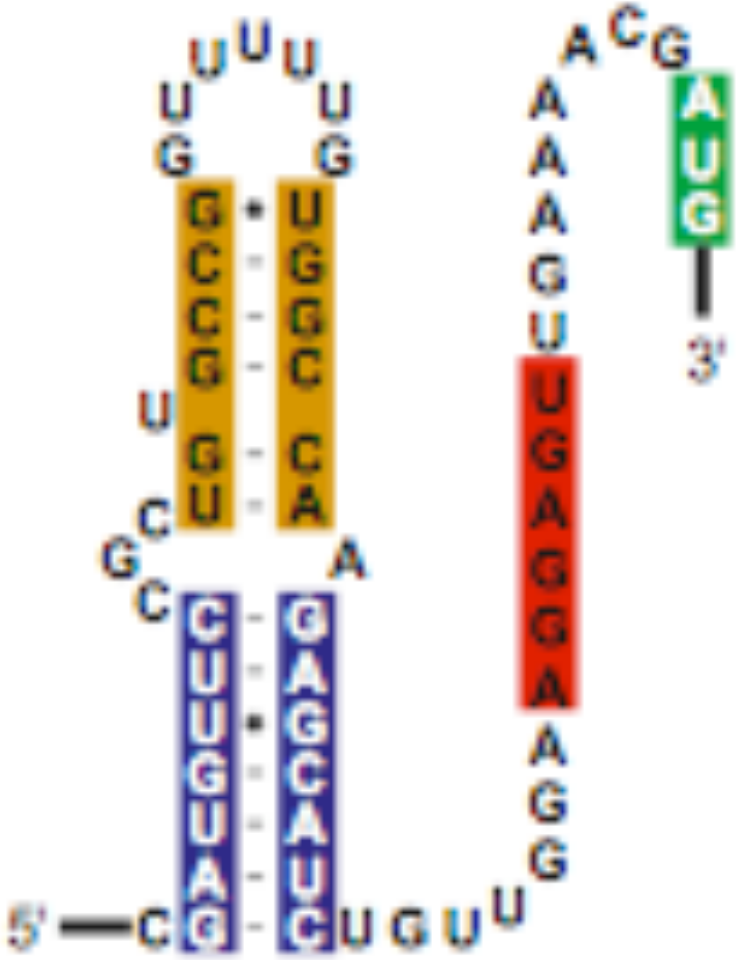
Letter

Comparative genomics beyond sequence-based alignments: RNA structures in the ENCODE regions

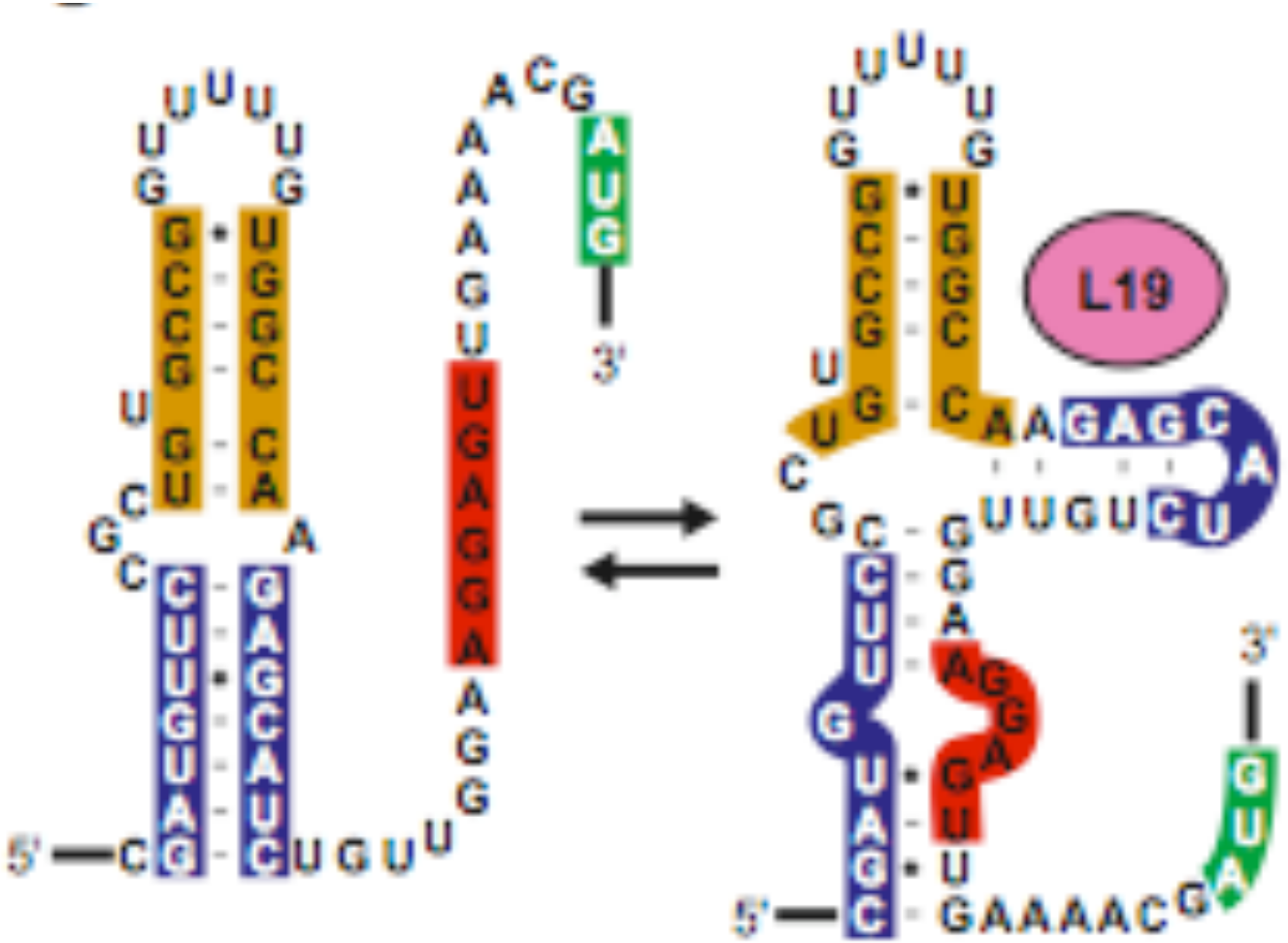
Elfar Torarinsson,[1,2] Zizhen Yao,[3] Eric D. Wiklund,[4] Jesper B. Bramsen,[4] Claus Hansen,[5] Jørgen Kjems,[4] Niels Tommerup,[5] Walter L. Ruzzo,[3,6] and Jan Gorodkin[1,7]

[1] Section for Genetics and Bioinformatics, IBVH, Faculty of Life Sciences, University of Copenhagen, 1870 Frederiksberg C, Denmark; [2]Department of Natural Sciences, Faculty of Life Sciences, University of Copenhagen, 1871 Frederiksberg C, Denmark; [3]Department of Computer Science and Engineering, University of Washington, Seattle, Washington 98195-2350, USA; [4]Department of Molecular Biology, University of Aarhus, 8000 Aarhus, Denmark; [5]Department of Cellular and Molecular Medicine, Wilhelm Johannsen Centre for Functional Genome Research, University of Copenhagen, 2200 Copenhagen N, Denmark; [6]Department of Genome Sciences, University of Washington Seattle, Washington 98195-5065, USA

An RNA Structure

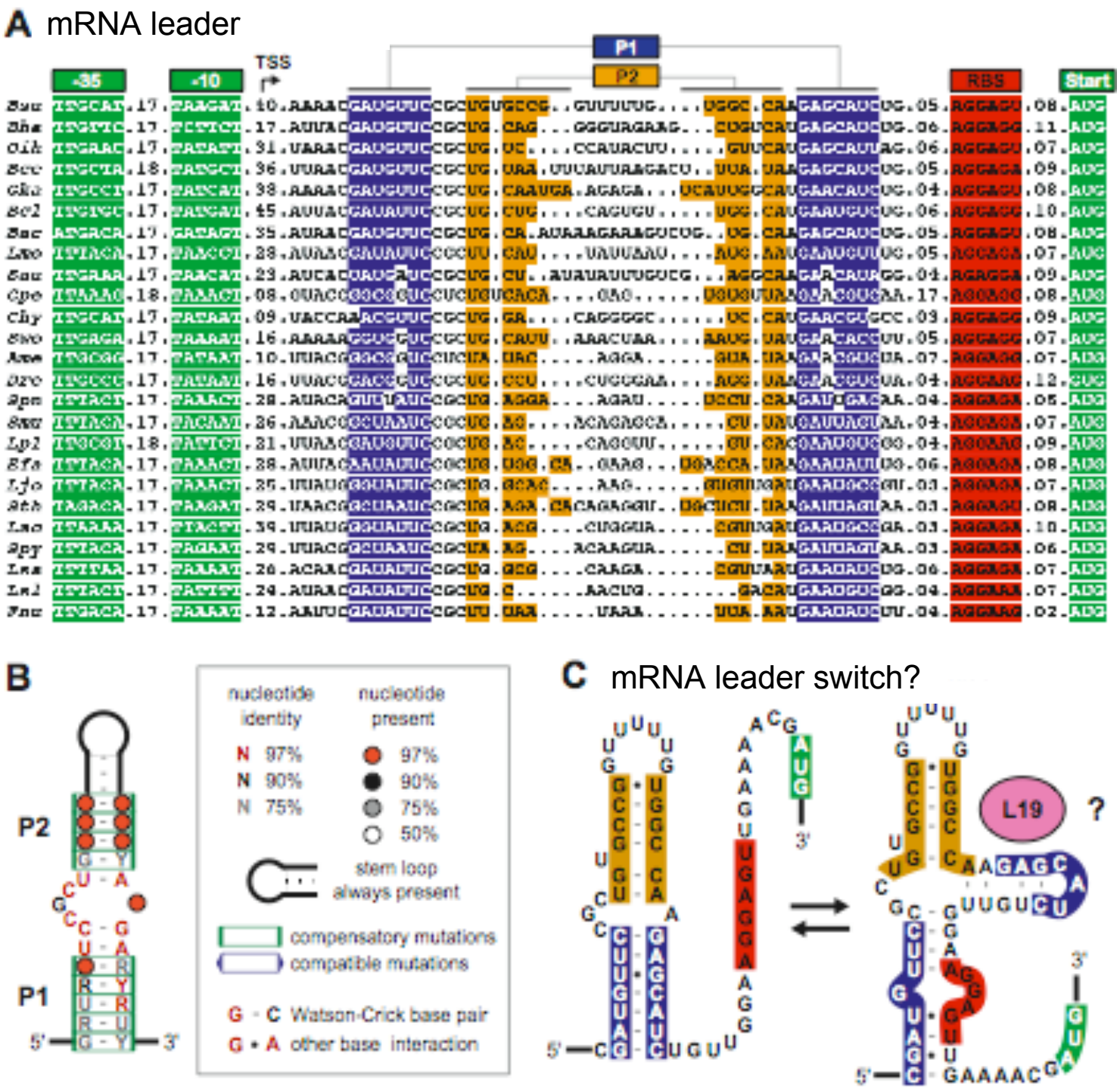


An RNA Sensor & On/Off Switch



L19 absent: Gene On

L19 present: Gene Off


A mRNA leader
P1
P2
TSS
-35
-10
RBS
Start
B
P2
P1
5'
3'
nucleotide identity
nucleotide present
N 97%
N 90%
N 75%
97%
90%
75%
50%
stem loop always present
compensatory mutations
compatible mutations
G - C Watson-Crick base pair
G • A other base interaction
C mRNA leader switch?
AUG
3'
5'
L19
?
GAAAAC

An RNA Grammar

$S \rightarrow LS \mid L$

$L \rightarrow s \mid$ "dFd"

$F \rightarrow LS \mid$ "dFd"

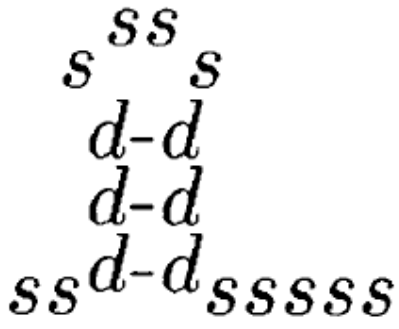
"dFd" means Watson-Crick base pair:

aFu | *uFa* | *gFc* | *cFg*

paren-like nesting

a) $S \rightarrow LS \rightarrow LLLLLLLS \rightarrow LLLLLLLL$
$\rightarrow ssLssssss \rightarrow ssdFdssssss$
$\rightarrow ssdddFddddssssss$
$\rightarrow ssdddLSddddssssss$
$\rightarrow ssdddLLLLddddssssss$
$\rightarrow ssddddssssddddssssss$

b)



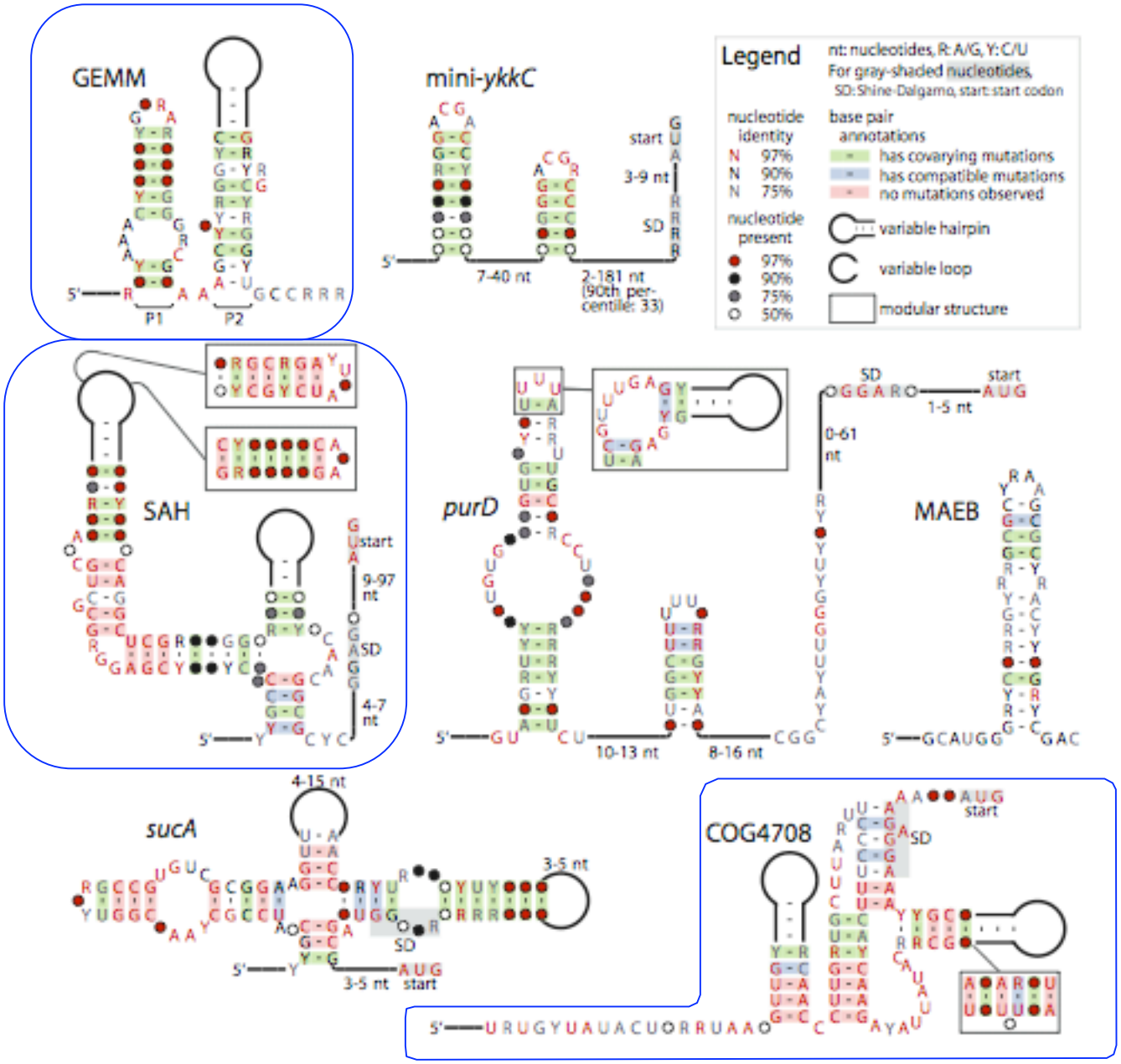
c) $F \rightarrow dFd \rightarrow ddFdd \rightarrow ddLSdd$
$\rightarrow ddLLdd \rightarrow ddLsdd \rightarrow dddFdsdd$

Actually, a *Stochastic* CFG

Associate probabilities with rules:

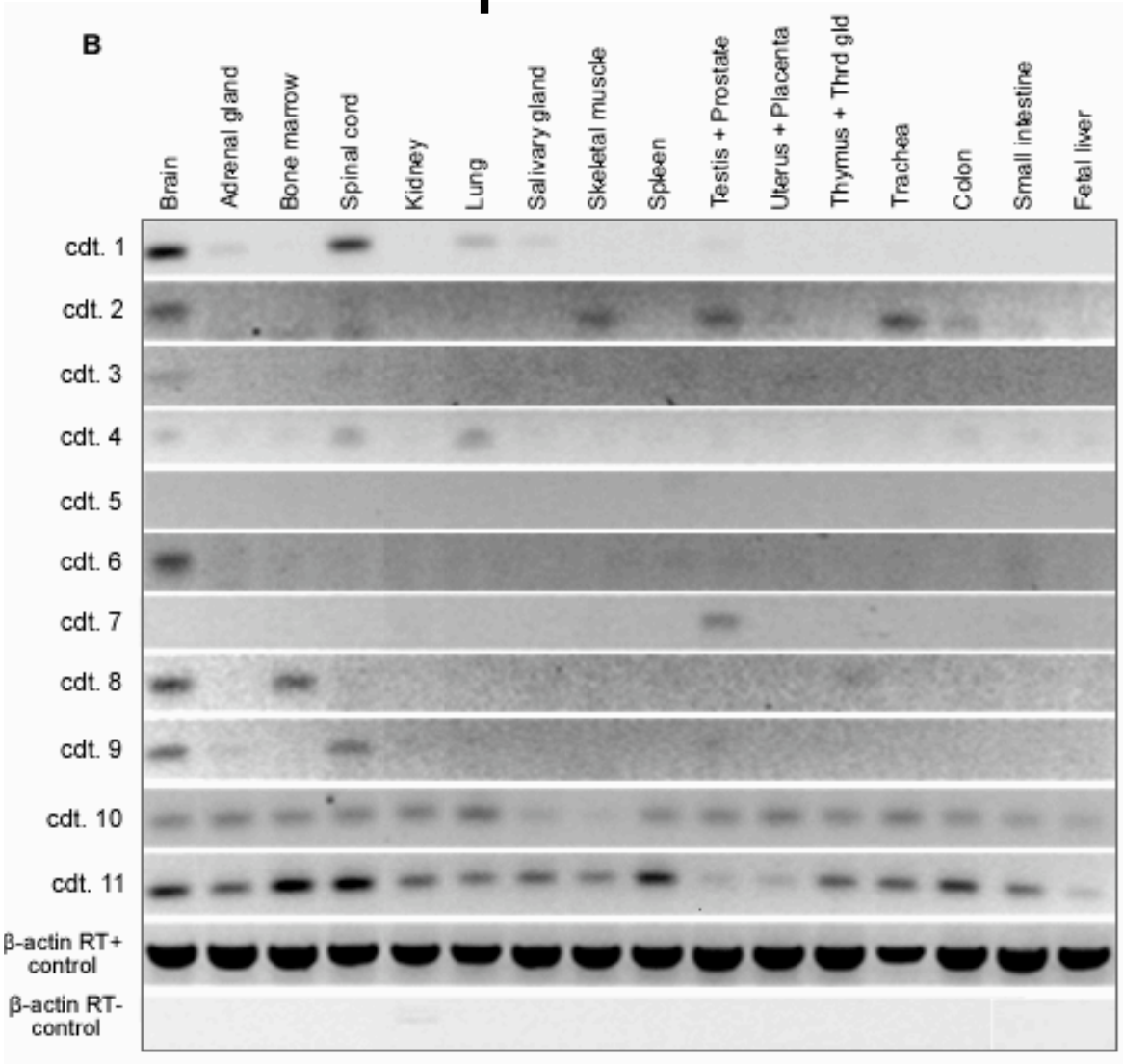
$S \rightarrow LS$	(0.87)	$\mid L$	(0.13)
$L \rightarrow S$	(0.89*p(s))	$\mid dFd$	(0.11*p(dd))
$F \rightarrow LS$	(0.21)	$\mid dFd$	(0.79*p(dd))

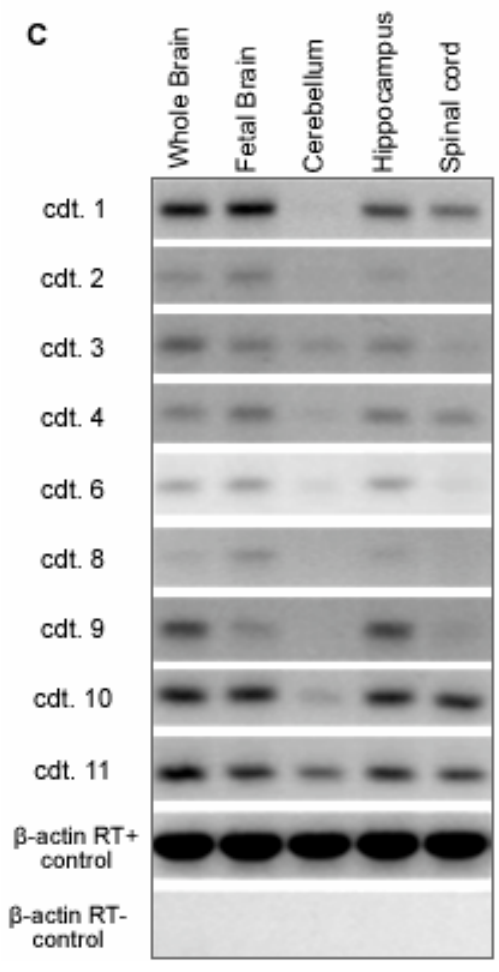
Where p(s) & p(dd) are the probabilities of the specific single/paired nucleotides, perhaps from empirical data or a model of sequence evolution



boxed = confirmed riboswitch (+2 more)

Experimental Validation





Bottom Line

CFG technology is a key tool for RNA description, discovery and search

A very active research area. (Some call RNA the “dark matter” of the genome.)

Huge compute hog: results above represent hundreds of CPU-years, and smart algorithms can have a big impact

An Algorithm Example: ncRNAs

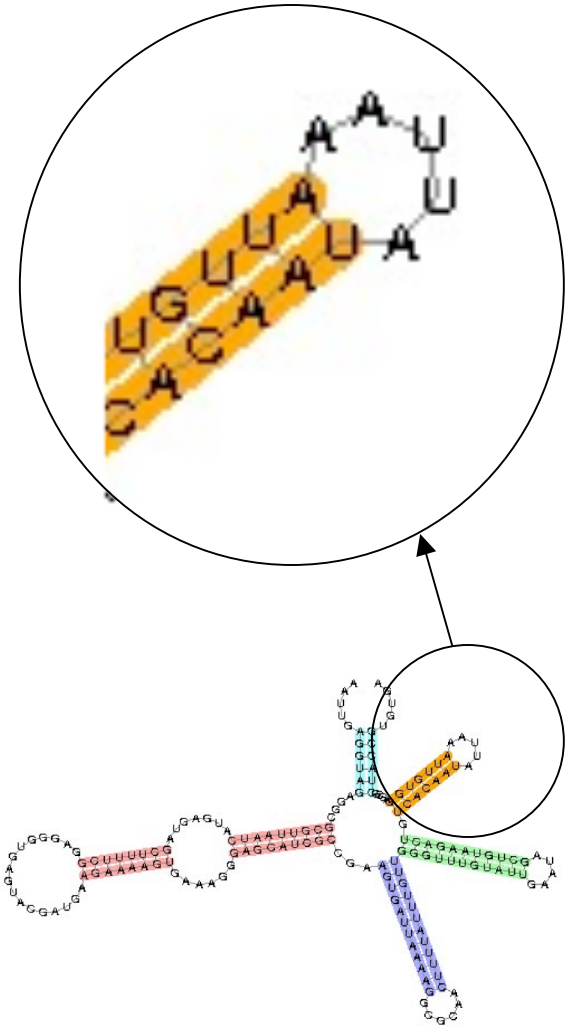
The “Central Dogma”:
DNA -> messenger RNA -> Protein

Last ~5 years: 100s – 1000s of examples of functionally important ncRNAs

Much harder to find than protein-coding genes

Main method - Covariance Models (based on stochastic context free grammars)

Main problem - Sloooow … $O(nm^4)$



“Rigorous Filtering” - Z. Weinberg

Convert CM to HMM
(AKA: stochastic CFG to stochastic regular grammar)

Do it so HMM score *always* ≥ CM score

Optimize for most aggressive filtering subject to constraint that score bound maintained

A large convex optimization problem

Filter genome sequences with (fast) HMM, run (slow) CM only on sequences above desired CM threshold; guaranteed not to miss anything

Newer, more elaborate techniques pulling in key secondary structure features for better searching
(uses automata theory, dynamic programming, Dijkstra, more optimization stuff,…)

CENSORED

Details
(but stay tuned…)
Plenty of CS here

Results

Typically 200-fold speedup or more

Finding dozens to hundreds of new ncRNA genes in many families

Has enabled discovery of many new families

Newer, more elaborate techniques pulling in key secondary structure features for better searching
(uses automata theory, dynamic programming, Dijkstra, more optimization stuff,…)

More Admin

Course Focus & Goals

Mainly sequence analysis

Algorithms for alignment, search, & discovery

Specific sequences, general types (“genes”, etc.)

Single sequence and comparative analysis

Techniques: HMMs, EM, MLE, Gibbs, Viterbi…

Enough bio to motivate these problems, including very light intro to modern biotech supporting them

Math/stats/cs underpinnings thereof

Applied to real data

A *VERY* Quick Intro To Molecular Biology

The Genome

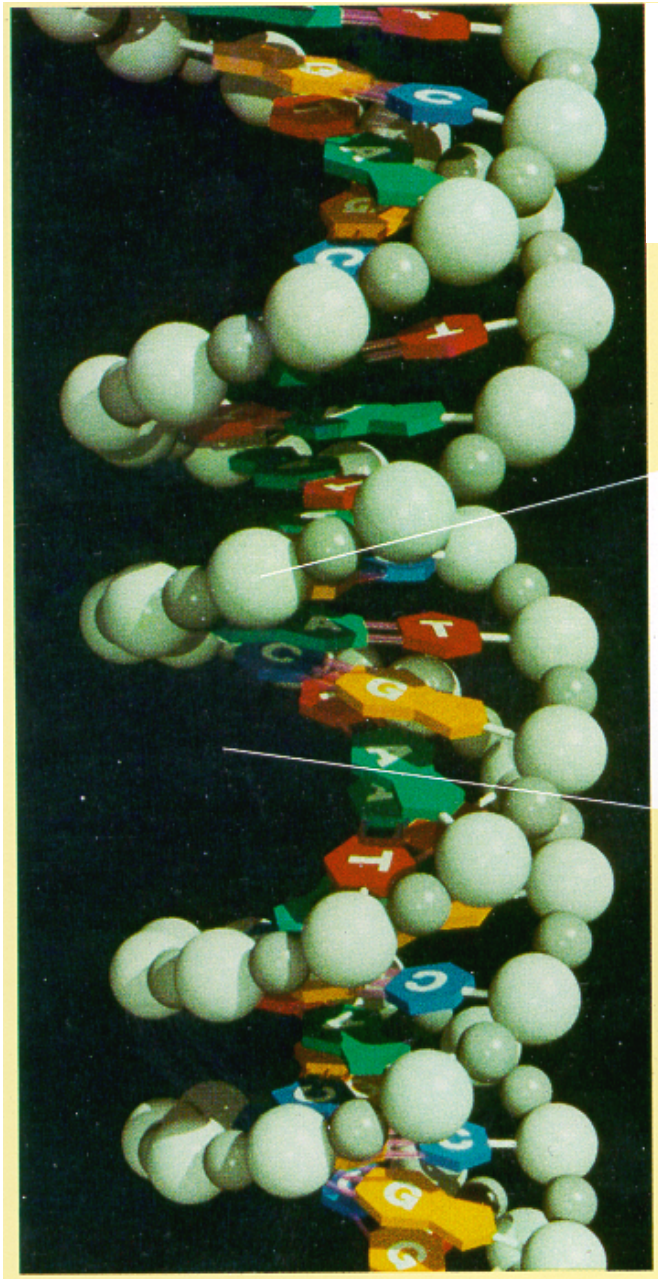
The hereditary info present in every cell

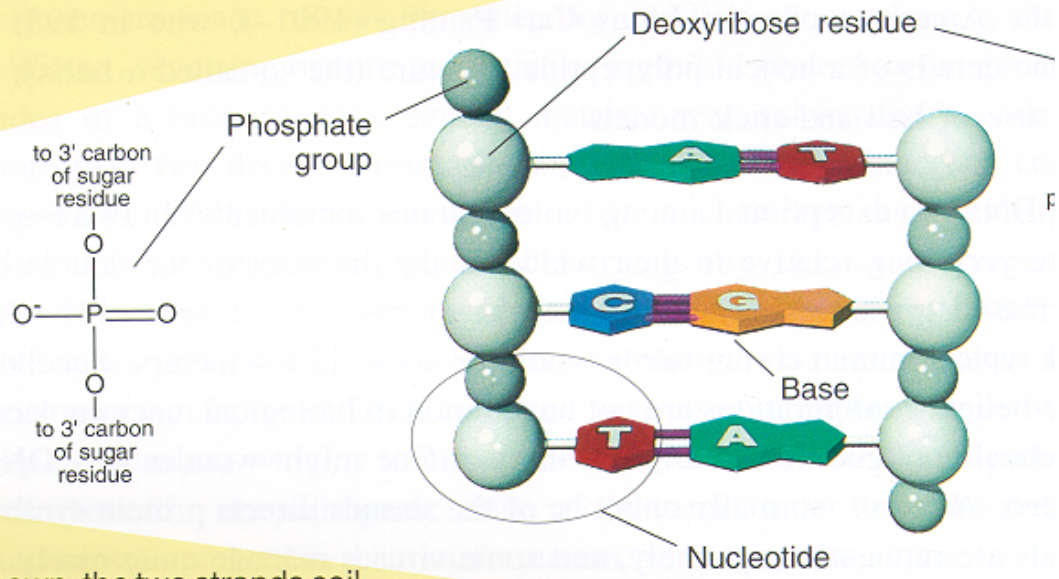
DNA molecule -- a long sequence of *nucleotides* (A, C, T, G)

Human genome -- about 3×10^9 nucleotides

The genome project -- extract & interpret genomic information, apply to genetics of disease, better understand evolution, …

The Double Helix





As shown, the two strands coil about each other in a fashion such that all the bases project inward toward the helix axis. The two strands are held together by hydrogen bonds (pink rods) linking each base projecting from one backbone to its so-called complementary base projecting from the other backbone. The base A always bonds to T (A and T are comple-

Shown in (b) is an uncoiled fragment of (a three complementary base pai chemist's viewpoint, each stra a polymer made up of four re called deoxyribonucleotides

DNA

Discovered 1869

Role as carrier of genetic information - much later

The Double Helix - Watson & Crick 1953

Complementarity

A ⟷ T C ⟷ G

Visualizations:

http://www.rcsb.org/pdb/explore.do?structureId=123D

Genetics - the study of heredity

A *gene* -- classically, an abstract heritable attribute existing in variant forms (*alleles*)

Genotype vs *phenotype*

Mendel

- Each individual two copies of each gene
- Each parent contributes one (randomly)
- Independent assortment

Cells

Chemicals inside a sac - a fatty layer called the *plasma membrane*

Prokaryotes (bacteria, archaea) - little recognizable substructure

Eukaryotes (all multicellular organisms, and many single celled ones, like yeast) - genetic material in nucleus, other organelles for other specialized functions

Chromosomes

1 pair of (complementary) DNA molecules (+ protein wrapper)

Most prokaryotes have just 1 chromosome

Eukaryotes - ~~all~~ most cells have same number of chromosomes, e.g. fruit flies 8, humans & bats 46, rhinoceros 84, …

Mitosis/Meiosis

Most “higher” eukaryotes are *diploid* - have homologous pairs of chromosomes, one maternal, other paternal (exception: sex chromosomes)

Mitosis - cell division, duplicate each chromosome, 1 copy to each daughter cell

Meiosis - 2 divisions form 4 *haploid* gametes (egg/sperm)

- *Recombination/crossover* -- exchange maternal/paternal segments

Proteins

Chain of amino acids, of 20 kinds

Proteins:the major functional elements in cells

- Structural/mechanical
- Enzymes (catalyze chemical reactions)
- Receptors (for hormones, other signaling molecules, odorants,…)
- Transcription factors
- …

3-D Structure is crucial: the protein folding problem

The “Central Dogma”

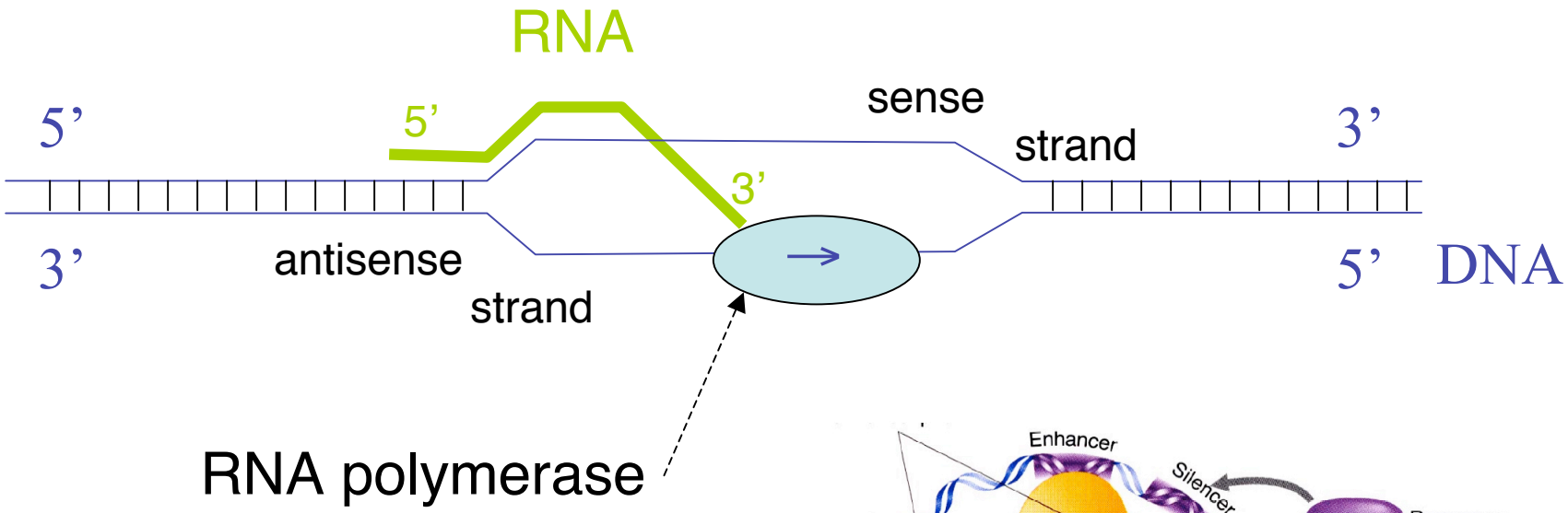
Genes encode proteins

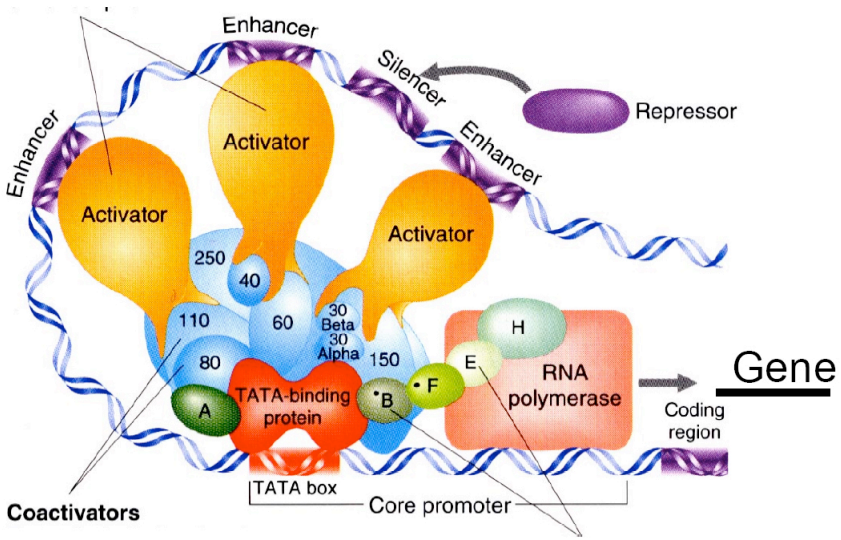
DNA transcribed into messenger RNA

mRNA translated into proteins

Triplet code (codons)

Transcription: DNA → RNA



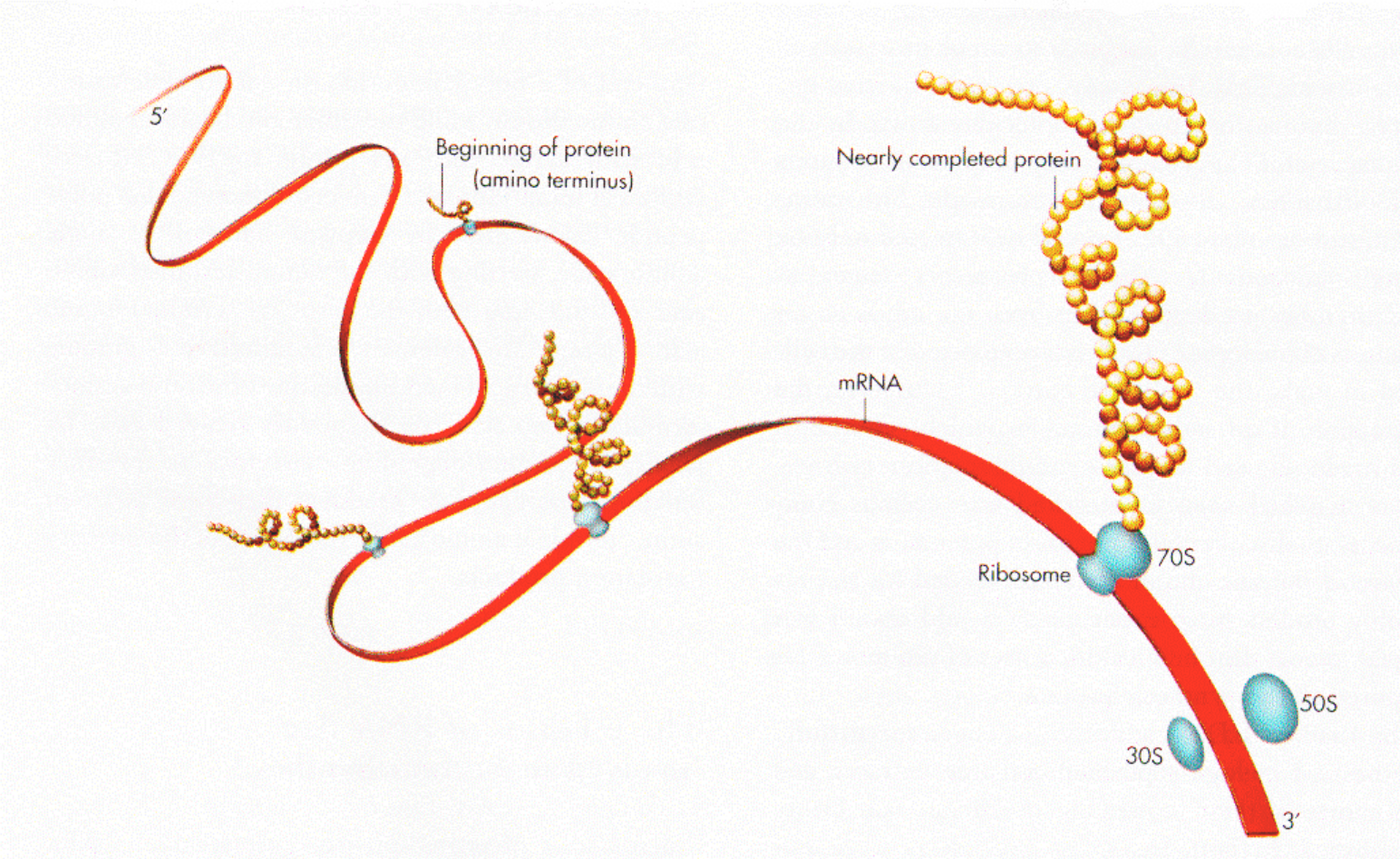


Codons & The Genetic Code

First Base	Second Base: U	Second Base: C	Second Base: A	Second Base: G	Third Base
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	Stop	Stop	A
	Leu	Ser	Stop	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met/Start	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

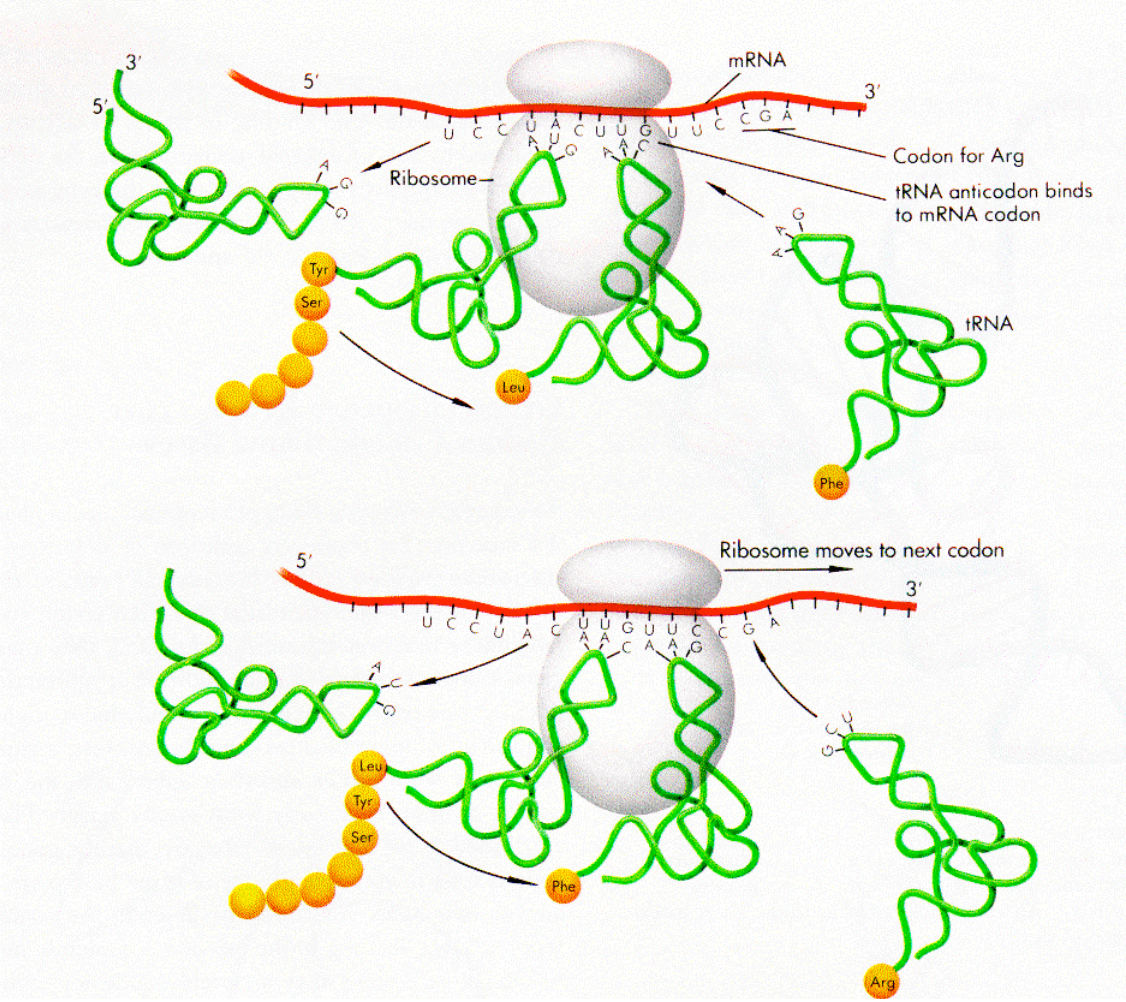
Ala : Alanine
Arg : Arginine
Asn : Asparagine
Asp : Aspartic acid
Cys : Cysteine
Gln : Glutamine
Glu : Glutamic acid
Gly : Glycine
His : Histidine
Ile : Isoleucine
Leu : Leucine
Lys : Lysine
Met : Methionine
Phe : Phenylalanine
Pro : Proline
Ser : Serine
Thr : Threonine
Trp : Tryptophane
Tyr : Tyrosine
Val : Valine

Translation: mRNA → Protein



Watson, Gilman, Witkowski, & Zoller, 1992

Ribosomes



Watson, Gilman, Witkowski, & Zoller, 1992

Gene Structure

Transcribed 5’ to 3’

Promoter region and transcription factor binding sites (usually) precede 5’ end

Transcribed region includes 5’ and 3’ untranslated regions

In eukaryotes, most genes also include *introns*, spliced out before export from nucleus, hence before translation

Genome Sizes

	Base Pairs	Genes
Mycoplasma genitalium	580,073	483
MimiVirus	1,200,000	1,260
E. coli	4,639,221	4,290
Saccharomyces cerevisiae	12,495,682	5,726
Caenorhabditis elegans	95,500,000	19,820
Arabidopsis thaliana	115,409,949	25,498
Drosophila melanogaster	122,653,977	13,472
Humans	3.3×10^9	~25,000

Genome Surprises

Humans have < 1/3 as many genes as expected

But perhaps more proteins than expected, due to *alternative splicing, alt start, alt end*

Protein-wise, all mammals are just about the same

But more individual variation than expected

And many more *non-coding RNAs* -- more than protein-coding genes, by some estimates

Many other non-coding regions are highly conserved, e.g., across all vertebrates

90% of DNA is transcribed (< 2% coding)

Complex, subtle “epigenetic” information

… and much more …

Read one of the many intro surveys or books for much more info.