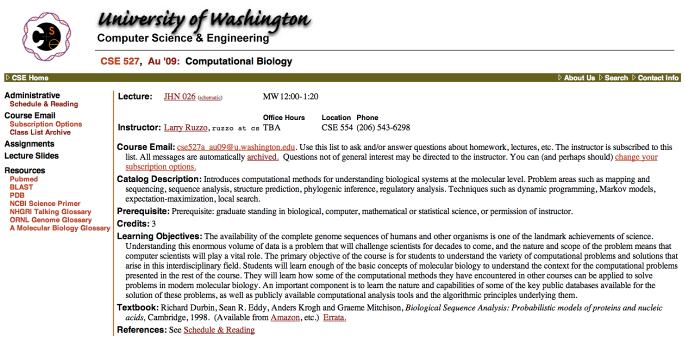
CSE527
Computational Biology

http://www.cs.washington.edu/527

Larry Ruzzo

Autumn 2009

UW CSE Computational Biology Group



Lecture: JHN 026 (schematic) MW 12:00- 1:20

	Office Hours	Location	Phone
Instructor: Larry Ruzzo, ruzzo at cs	TBA	CSE 554	(206) 543-6298

Course Email: cse527a_au09@u.washington.edu. Use this list to ask and/or answer questions about homework, lectures, etc. The instructor is subscribed to this list. All messages are automatically archived. Questions not of general interest may be directed to the instructor. You can (and perhaps should) change your subscription options.

Catalog Description: Introduces computational methods for understanding biological systems at the molecular level. Problem areas such as mapping and sequencing, sequence analysis, structure prediction, phylogenic inference, regulatory analysis. Techniques such as dynamic programming, Markov models, expectation-maximization, local search.

Prerequisite: Prerequisite: graduate standing in biological, computer, mathematical or statistical science, or permission of instructor.

Credits: 3

Learning Objectives: The availability of the complete genome sequences of humans and other organisms is one of the landmark achievements of science. Understanding this enormous volume of data is a problem that will challenge scientists for decades to come, and the nature and scope of the problem means that computer scientists will play a vital role. The primary objective of the course is for students to understand the variety of computational problems and solutions that arise in this interdisciplinary field. Students will learn enough of the basic concepts of molecular biology to understand the context for the computational problems presented in the rest of the course. They will learn how some of the computational methods they have encountered in other courses can be applied to solve problems in modern molecular biology. An important component is to learn the nature and capabilities of some of the key public databases available for the solution of these problems, as well as publicly available computational analysis tools and the algorithmic principles underlying them.

Textbook: Richard Durbin, Sean R. Eddy, Anders Krogh and Graeme Mitchison, *Biological Sequence Analysis: Probabilistic models of proteins and nucleic acids,* Cambridge, 1998. (Available from Amazon, etc.) Errata.

References: See Schedule & Reading

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

-- Chinese Proverb

Today

Admin

Why Comp Bio?

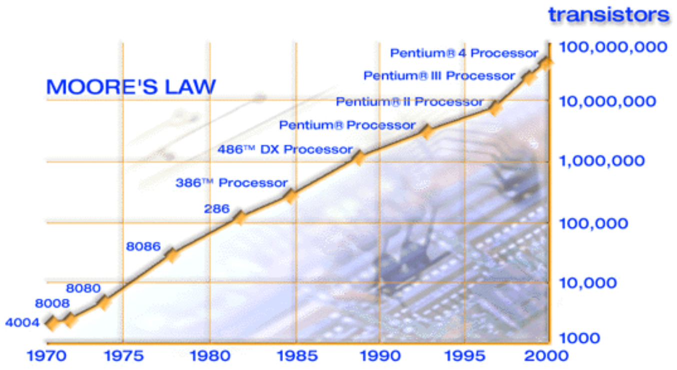
The world's shortest Intro. to Mol. Bio.

Admin Stuff

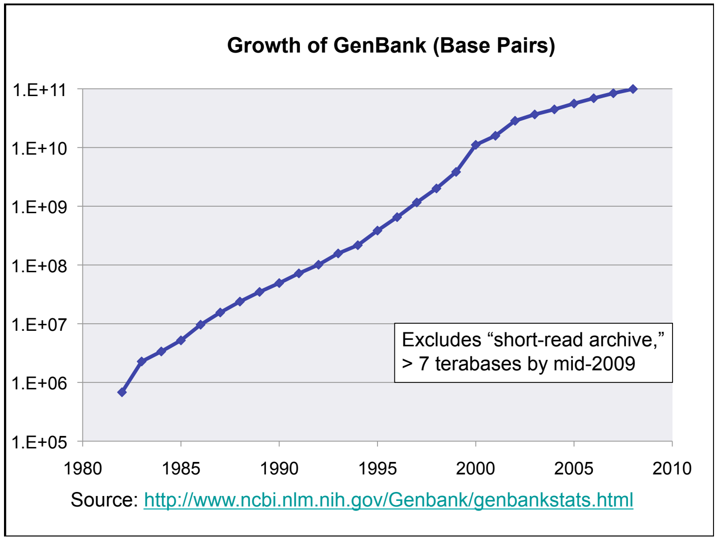
Course Mechanics & Grading

- Reading
- In class discussion
- Lecture scribes
- Homeworks ← Check web for 1st
  - reading
  - paper exercises
  - programming
- Project
- No exams

Background & Motivation



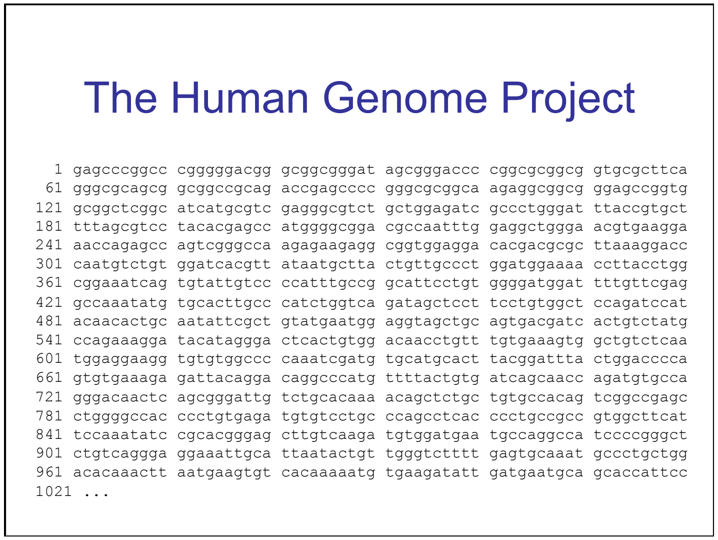
Source: http://www.intel.com/research/silicon/mooreslaw.htm



Growth of GenBank (Base Pairs)

Excludes “short-read archive,”
> 7 terabases by mid-2009

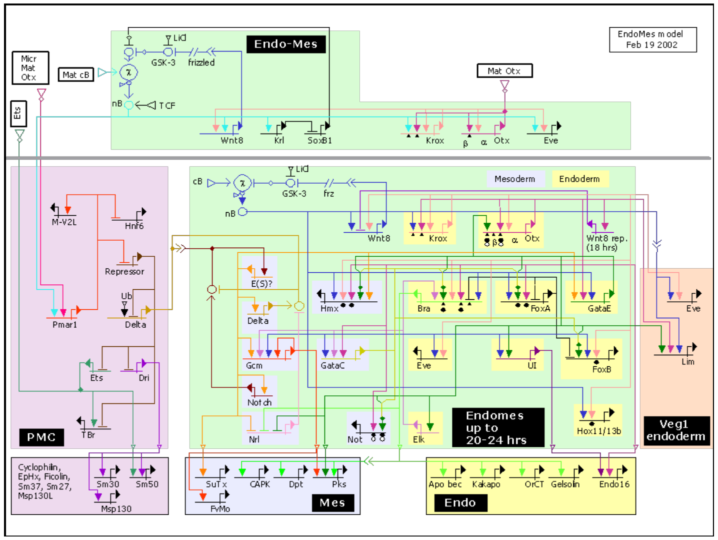
Source: http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html



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 ...
```

The sea urchin *Strongylocentrotus purpuratus*



Goals

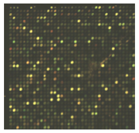
Basic biology

Disease diagnosis/prognosis/treatment

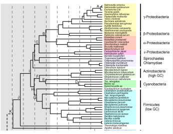
Drug discovery, validation & development

Individualized medicine

…

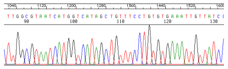


"High-Throughput BioTech"



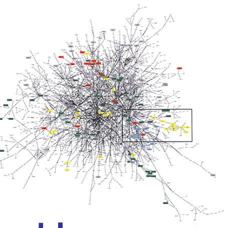
Sensors

- DNA sequencing
- Microarrays/Gene expression
- Mass Spectrometry/Proteomics
- Protein/protein & DNA/protein interaction

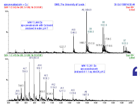


Controls

- 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, individual variation, …

AI/NLP/Text Mining

- Information extraction from text with inconsistent nomenclature, indirect interactions, incomplete/inaccurate models, …

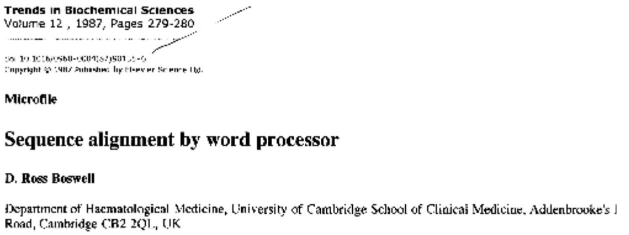
Machine learning

- System level synthesis of cell behavior from low-level heterogeneous data (DNA seq, gene expression, protein interaction, mass spec,…)

...

Algorithms

Computers in biology: Then & now



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

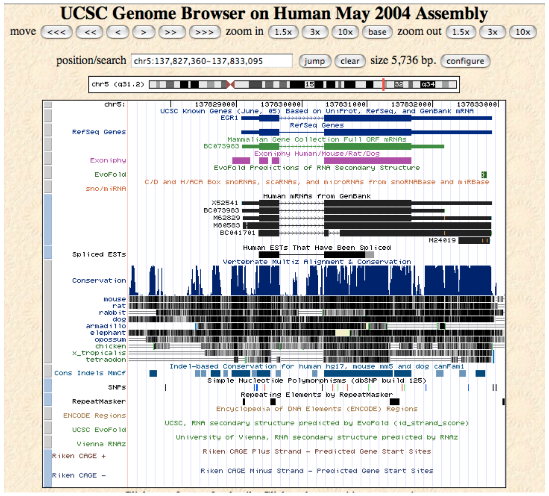
doi:10.1016/0968-0004(87)90135-6
Copyright © 1987 Published by Elsevier Science Ltd.

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)
RefSeq Genes
Exoniphy
EvoFold
sno/miRNA
Spliced ESTs
Conservation
mouse
rat
rabbit
dog
armadillo
elephant
opossum
chicken
x_tropicalis
tetraodon
Cons Indels MmCf
SNPs
RepeatMasker
ENCODE Regions
UCSC EvoFold
Vienna RNAz
Riken CAGE +
Riken CAGE -
UCSC Known Genes (June, 05) Based on UniProt, RefSeq, and GenBank mRNA
EGR1
RefSeq Genes
Mammalian Gene Collection Full ORF mRNAs
BC073983
Exoniphy Human/Mouse/Rat/Dog
EvoFold Predictions of RNA Secondary Structure
C/D and H/ACA Box snoRNAs, scaRNAs, and microRNAs from snoRNABase and miRBase
Human mRNAs from GenBank
X52541
BC073983
M62829
M80583
BC041701
M24019
Human ESTs That Have Been Spliced
Vertebrate Multiz Alignment & Conservation
Indel-based Conservation for human hg17, mouse mm5 and dog canFam1
Simple Nucleotide Polymorphisms (dbSNP build 125)
Repeating Elements by RepeatMasker
Encyclopedia of DNA Elements (ENCODE) Regions
UCSC, RNA secondary structure predicted by EvoFold (id_strand_score)
University of Vienna, RNA secondary structure predicted by RNAz
Riken CAGE Plus Strand - Predicted Gene Start Sites
Riken CAGE Minus Strand - Predicted Gene Start Sites

An Algorithm Example: ncRNAs

The “Central Dogma”:

- DNA -> messenger RNA -> Protein

Last ~5 years:

- 100s – 1000s of examples of functionally important ncRNAs

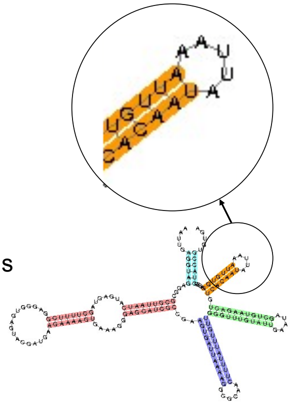
Much harder to find than protein-coding genes

Main method - Covariance Models

- ≈ 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 sequence 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,…)

Details

CENSORED

(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

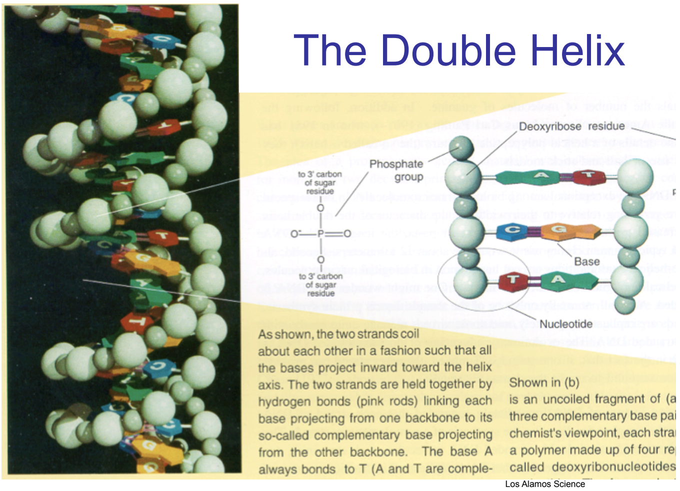
The hereditary info present in every cell

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

Human genome -- about 3 x 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 strai a polymer made up of four re called deoxyribonucleotides

Los Alamos Science

DNA

Discovered 1869

Role as carrier of genetic information - much later

4 “bases”:

adenine (A), cytosine (C), guanine (G), thymine (T)

The Double Helix - Watson & Crick 1953

Complementarity

A ←→ T C ←→ G

Visualization:

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*

I.e., genes vs their outward manifestation

Mendel

Each individual two copies of each gene

Each parent contributes one (randomly)

Independent assortment (approx, but useful)

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: 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”

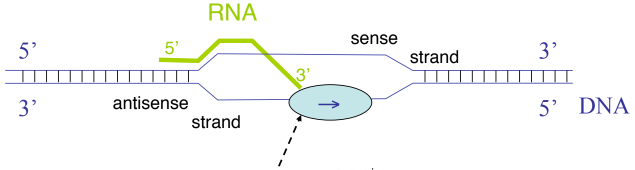
Genes encode proteins

DNA transcribed into messenger RNA

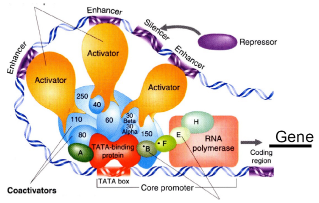
mRNA translated into proteins

Triplet code (codons)

Transcription: DNA → RNA



RNA polymerase



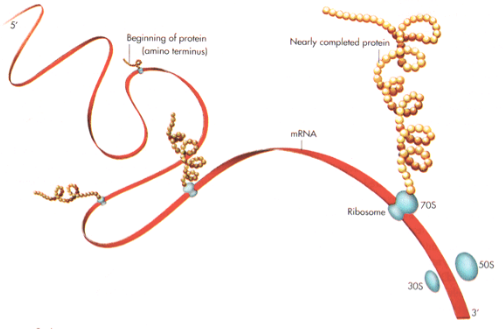
Codons & The Genetic Code

		Second Base				
		U	C	A	G	
First 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

Third Base

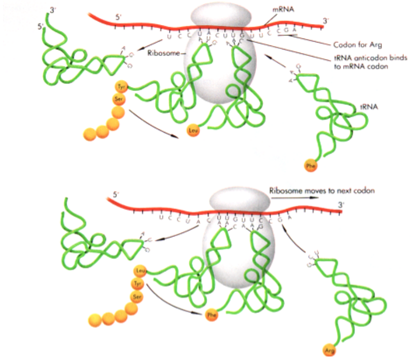
Ala	: Alanine
Arg	: Arginine
Asn	: Asparagine
Asp	: Aspartic acid
Cys	: Cysteine
Gln	: Glutamine
Glu	: Glutamic acid
Gly	: Glycine
His	: Histidine
Ile	: Isoleucine
Leu	: Leucine
Lys	: Lysine
Met	: Methionine
Phe	: Phenylalanine
Pro	: Proline
Ser	: Serine
Thr	: Threonine
Trp	: Tryptophane
Tyr	: Tyrosine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

Translation: mRNA → Protein



Watson, Gilman, Witkowski, & Zoller, 1992

Ribosomes


3'
5'
5'
mRNA
3'
U C C U A C U U G U U C C G A
Ribosome
Codon for Arg
tRNA anticodon binds to mRNA codon
tRNA
Tyr
Ser
Leu
Phe
5'
Ribosome moves to next codon
3'
Leu
Tyr
Ser
Phe
Arg

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

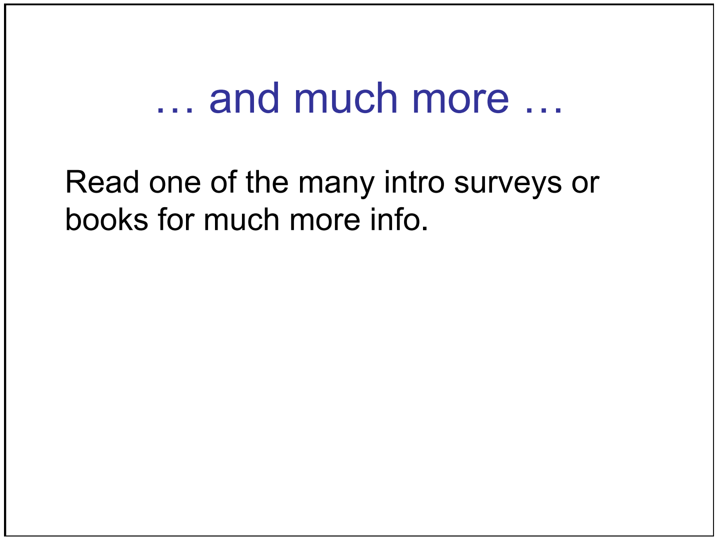
But more individual variation than expected

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.