CS273A

Lecture 2: Protein Coding Genes

MW 1:30-2:50pm in Clark **S361*** (behind Peet's)

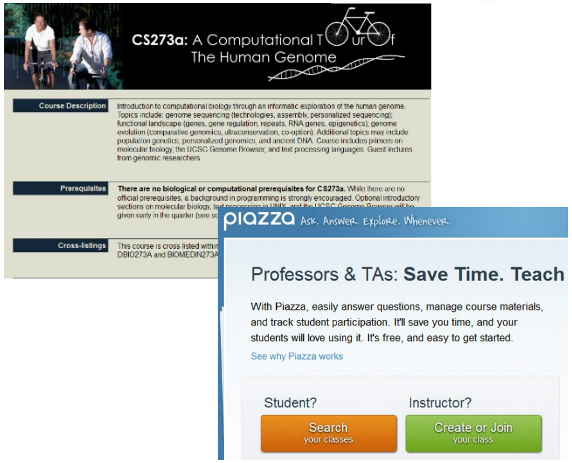
Profs: Serafim Batzoglou & Gill Bejerano

CAs: Karthik Jagadeesh & Johannes Birgmeier

* Handful of lectures/primers elsewhere: track on website/piazza

Announcements

- http://cs273a.stanford.edu/
  - Course guidelines, office hours, etc.
  - Lecture 1 is posted
  - Problem set 1 rolls out next week



CS273a: A Computational Tour of The Human Genome

Course Description

Introduction to computational biology through an informatic exploration of the human genome. Topics include: genome sequencing (technologies, assembly, personalized sequencing); functional landscape (genes, gene regulation, repeats, RNA genes, epigenetics); genome evolution (comparative genomics, ultraconservation, co-option). Additional topics may include population genetics, personalized genomics, and ancient DNA. Course includes primers on molecular biology, the UCSC Genome Browser, and text processing languages. Guest lectures from genomic researchers.

Prerequisites

There are no biological or computational prerequisites for CS273a. While there are no official prerequisites, a background in programming is strongly encouraged. Optional introductory sections on molecular biology, text processing in UNIX, and the UCSC Genome Browser will be given early in the quarter (see sc

Cross-listings

This course is cross listed within
DBIO273A and BIOMEDIN273A

piazza Ask. Answer. Explore. Whenever.

Professors & TAs: **Save Time. Teach**

With Piazza, easily answer questions, manage course materials, and track student participation. It'll save you time, and your students will love using it. It's free, and easy to get started.

See why Piazza works

- Course communications via Piazza
  - Auditors please sign up too

- The first tutorial this Friday in Beckman B-302 from 2:00pm-3:30pm. It's the only one some students should consider skipping. While they may be familiar with the first half of the Molecular Biology 101 lecture, we also cover gene regulation and genome rearrangements.

- CAs will be sending out a Doodle poll via Piazza to identify ideal times for office hours. Students can contact them via Piazza for questions.

Class Goals

READING THE ENTIRE GENOME CODE...
Listen to this part: 'AGGCTAATC CGCATAACTG'
Wow!
S. Harris

- Meet your genome (learn to surf, learn the surf)
- Understand genomic tools (theory, applications)
- DIY (pose questions, write & run tools, understand answers)

Class Topics

(0) Genome context:
cells, DNA, central dogma

(1) Genome content / genome function:
genes, gene regulation, repeats, epigenetics

(2) Genome sequencing:
technologies, assembly/analysis, technology dependence

(3) Genome evolution:
evolution = mutation + selection, modes of evolution, comparative genomics, ultraconservation, exaptation

(4) Population genomics:
Tracking human migration patterns via neutral evolution

(5) Genomics of human disease:
disease susceptibility, cancer genomics, personal genomics

(6) Genome “output” (organism) evolution:
Evolutionary developmental biology (“evo-devo”)

Genome Context

```
ATATTGAATTTTCAAAAATTCTTACTTTTTTTTTGGATGGACGCAAAGAAGTTTAATAATCATATTACATGGCATTACCACCATATA
ATCCATATCTAATCTTACTTATATGTTGTGGAAATGTAAAGAGCCCCATTATCTTAGCCTAAAAAAACCTTCTCTTTGGAACTTTC
AATACGCTTAACTGCTCATTGCTATATTGAAGTACGGATTAGAAGCCGCCGAGCGGGCGACAGCCCTCCGACGGAAGACTCTCCTC
GCGTCCTCGTCTTCACCGGTCGCGTTCCTGAAACGCAGATGTGCCTCGCGCCGCACTGCTCCGAACAATAAAGATTCTACAATACT
TTTTATGGTTATGAAGAGGAAAAATTGGCAGTAACCTGGCCCCACAAACCTTCAAATTAACGAATCAAATTAACAACCATAGGATG
ATGCGATTAGTTTTTTAGCCTTATTTCTGGGGTAATTAATCAGCGAAGCGATGATTTTTGATCTATTAACAGATATATAAATGGAA
CTGCATAACCACTTTAACTAATACTTTCAACATTTTCAGTTTGTATTACTTCTTATTCAAATGTCATAAAAGTATCAACAAAAAT
TAATATACCTCTATACTTTAACGTCAAGGAGAAAAACTATAATGACTAAATCTCATTCAGAAGAAGTGATTGTACCTGAGTTCAA
TAGCGCAAAGGAATTACCAAGACCATTGGCCGAAAAGTGCCCGAGCATAATTAAGAAATTTATAAGCGCTTATGATGCTAAACCGG
TTGTTGCTAGATCGCCTGGTAGAGTCAATCTAATTGGTGAACATATTGATTATTGTGACTTCTCGGTTTTACCTTTAGCTATTGAT
GATATGCTTTGCGCCGTCAAAGTTTTGAACGATGAGATTTCAAGTCTTAAAGCTATATCAGAGGGCTAAGCATGTGTATTCTGAAT
TAAGAGTCTTGAAGGCTGTGAAATTAATGACTACAGCGAGCTTTACTGCCGACGAAGACTTTTTCAAGCAATTTGGTGCCTTGATG
GAGTCTCAAGCTTCTTGCGATAAACTTTACGAATGTTCTTGTCCAGAGATTGACAAAATTTGTTCCATTGCTTTGTCAAATGGATC
TGGTTCCCGTTTGACCGGAGCTGGCTGGGGTGGTTGTACTGTTCACTTGGTTCCAGGGGGCCCAAATGGCAACATAGAAAAGGTAA
AAGCCCTTGCCAATGAGTTCTACAAGGTCAAGTACCCTAAGATCACTGATGCTGAGCTAGAAAATGCTATCATCGTCTCTAAACCA
TTGGGCAGCTGTCTATATGAATTAGTCAAGTATACTTCTTTTTTTTACTTTGTTCAGAACAACTTCTCATTTTTTTTCTACTCATAA
TAGCATCACAAAATACCCAATAATAACGAGTAGTAACACTTTTATAGTTCATACATGCTTCAACTACTTAATAAATGATTGTATGA
TGTTTTCAATGTAAGAGATTTCGATTATCCACAAACTTTAAAACACAGGGACAAAATTCTTGATATGCTTTCAACCGCTGCGTTTT
TACCTATTCTTGACATGATATGACTACCATTTTGTTATTGTACGTGGGGCAGTTGACGTCTTATCATATGTCAAAG.TTGCGAA
CTTGGCAAGTTGCCAACTGACGAGATGCAGTAACACTTTTATAGTTCATACATGCTTCAACTACTTAATAAATGATTGTATGATAA
TTTCAATGTAAGAGATTTCGATTATCCACAAACTTTAAAACACAGGGACAAAATTCTTGATATGCTTTCAACCGCTGCGTTTTGGA
CTATTCTTGACATGATATGACTACCATTTTGTTATTGTACGTGGGGCAGTTGACGTCTTATCATATGTCAAAGTCATTTGCGAAGT
TGGCAAGTTGCCAACTGACGAGATGCAGTTTCCTACGCATAATAAGAATAGGAGGGAATATCAAGCCAGACAATCTATCATTACAT
AGCGGCTCTTCAAAAAGATTGAACTCTCGCCAACTTATGGAATCTTCCAATGAGACCTTTGCGCCAAATAATGTGGATTTGGAAAA
GTATAAGTCATCTCAGAGTAATATAACTACCGAAGTTTATGAGGCATCGAGCTTTGAAGAAAAAGTAAGCTCAGAAAAACCTCAAT
GCTCATTCTGGAAGAAAATCTATTATGAATATGTGGTCGTTGACAAATCAATCTTGGGTGTTTCTATTCTGGATTCATTTATGTAC
CAGGACTTGAAGCCCGTCGAAAAAGAAAGGCGGGTTTGGTCCTGGTACAATTATTGTTACTTCTGGCTTGCTGAATGTTTCAATAT
CACTTGGCAAATTGCAGCTACAGGTCTACAACTGGGTCTAAATTGGTGGCAGTGTTGGATAACAATTTGGATTGGGTACGGTTTCG
GTGCTTTTGTTGTTTTGGCCTCTAGAGTTGGATCTGCTTATCATTTGTCATTCCCTATATCATCTAGAGCATCATTCGGTATTTTC
TCTTTATGGCCCGTTATTAACAGAGTCGTCATGGCCATCGTTTGGTATAGTGTCCAAGCTTATATTGCGGCAACTCCCGTATCATT
GCTGAAATCTATCTTTGGAAAAGATTTACAATGATTGTACGTGGGGCAGTTGACGTCTTATCATATGTCAAAGTCATTTGCGAAGT
TGGCAAGTTGCCAACTGACGAGATGCAGTAACACTTTTATAGTTCATACATGCTTCAACTACTTAATAAATGATTGTATGATAATG
TCAATGTAAGAGATTTCGATTATCCACAAACTTTAAAACACAGGGACAAAATTCTTGATATGCTTTCAACCGCTGCGTTTTGGATA
ATTCTTGACATGATATGACTACCATTTTGTTATTGTTTATAGTTCATACATGCTTCAACTACTTAATAAATGATTGTATGATAATG
TCAATGTAAGAGATTTCGATTATCCTTATAGTTCATACATGCTTCAACTACTTAATAAATGATTGTATGATAATGTTTTCAATGTA
GATTTCGATTATCCTTATAGTTCATACATGCTTCAACTACTTAATAAATGATTGTATGATAATGTTTTCAATGTAAGAGATTTCGA
TCCTTATAGTTCATACATGCTTCAACTACTTAATAAATGATTGTATGATAATGTTTTCAATGTAAGAGATTTCGATTATCCTTATA
CATACATGCTTCAACTACTTAATAAATGATTGTATGATAATGTTTTCAATGTAAGAGATTTCGATTATCCTTATAGTTCATACATG
CAACTACTTAATAAATGATTGTATGATAATGTTTTCAATGTAAGAGATTTCGATTATCCTTATAGTTCATACATGCTTCAACTAC
TAAATGATTGTATGATAATGTTTTCAATGTAAGAGATTTCGATTATCCTTATAGTTCATACATGCTTCAACTACTTAATAAATGAT
```

Organism – Cell - Genome

10^{13} different cells in an adult human. The cell is the basic unit of life.

DNA = linear molecule inside the cell that carries instructions needed throughout the cell’s life ~ long string(s) over a small alphabet

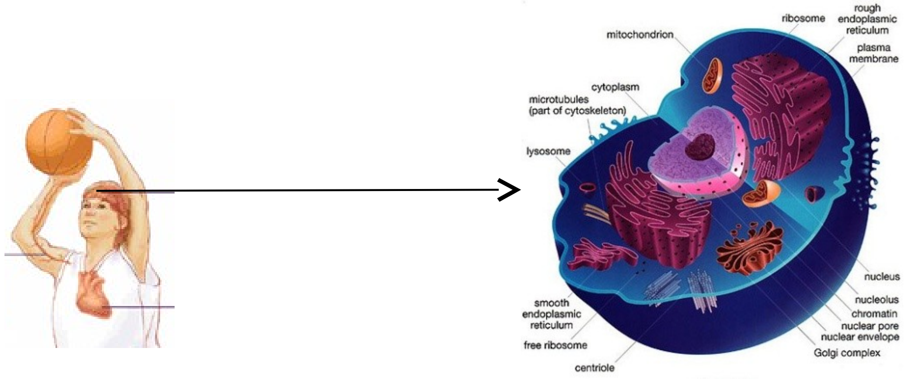
Alphabet (nucleotides/bases) {A,C,G,T}

Strings (chromosomes) of length 10^4-10^{11}

□ “instruction”

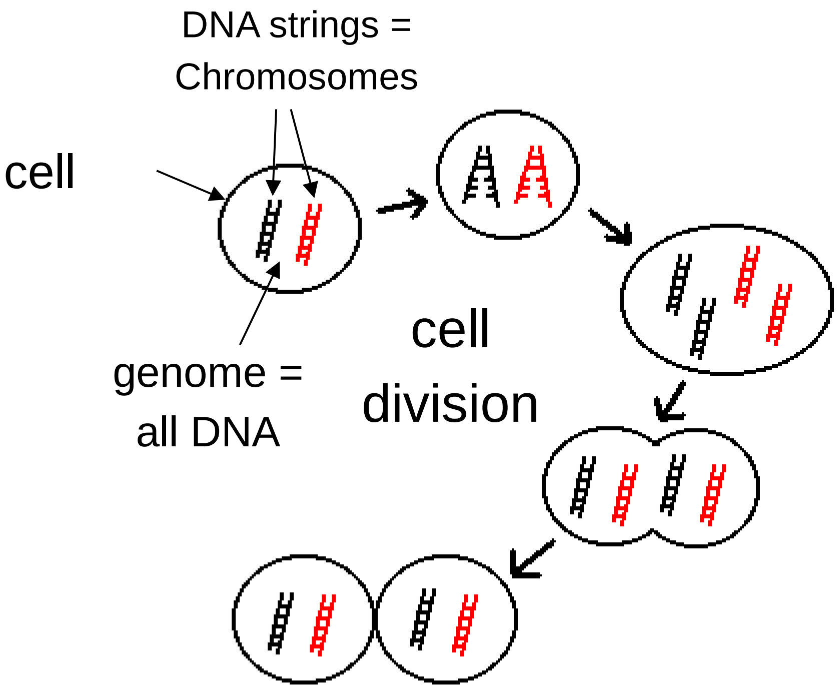
Genome:

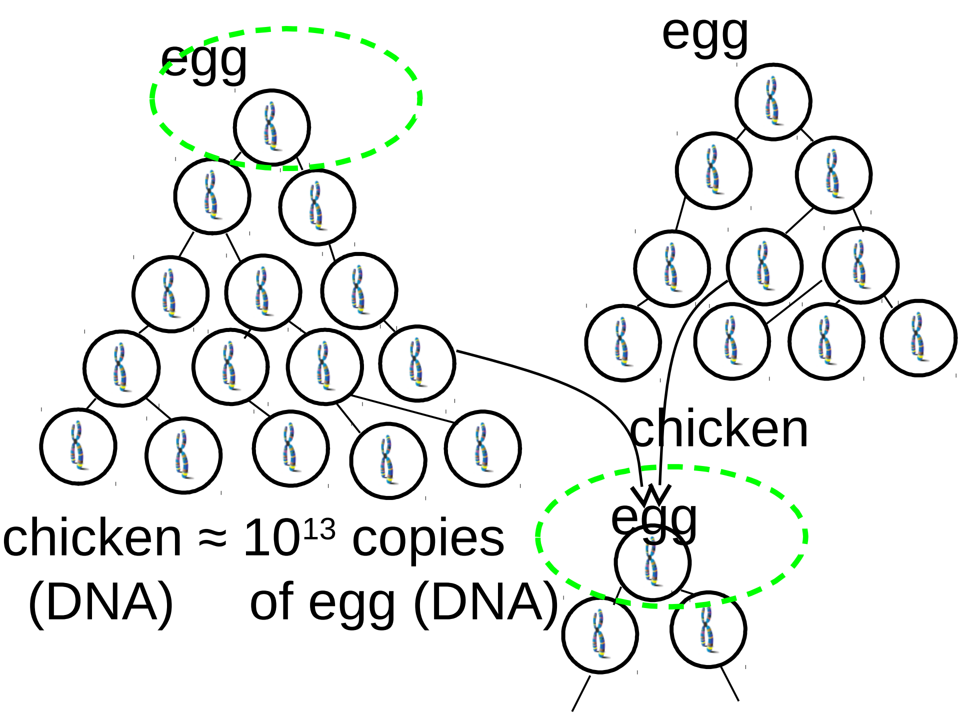
...ACGTACGACTGACTAGCATCGACTACGACTAGCAC...



One Cell, One Genome, One Replication

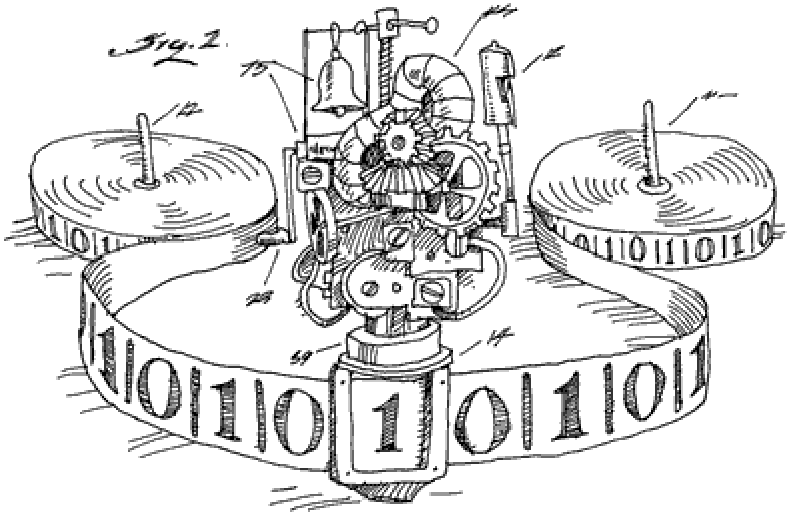
- Every cell holds a copy of all its DNA = its genome.
- The human body is made of ~10^{13} cells.
- All originate from a *single* cell through *repeated* cell divisions.

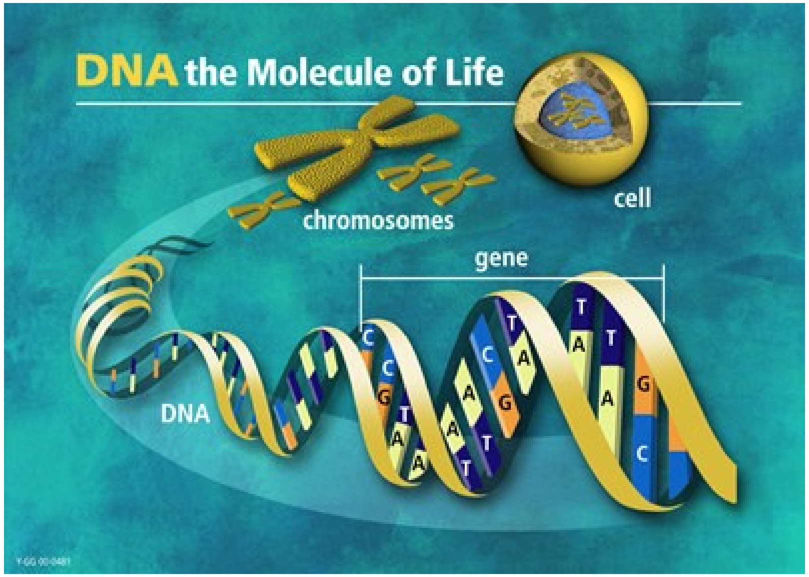




What will we study?

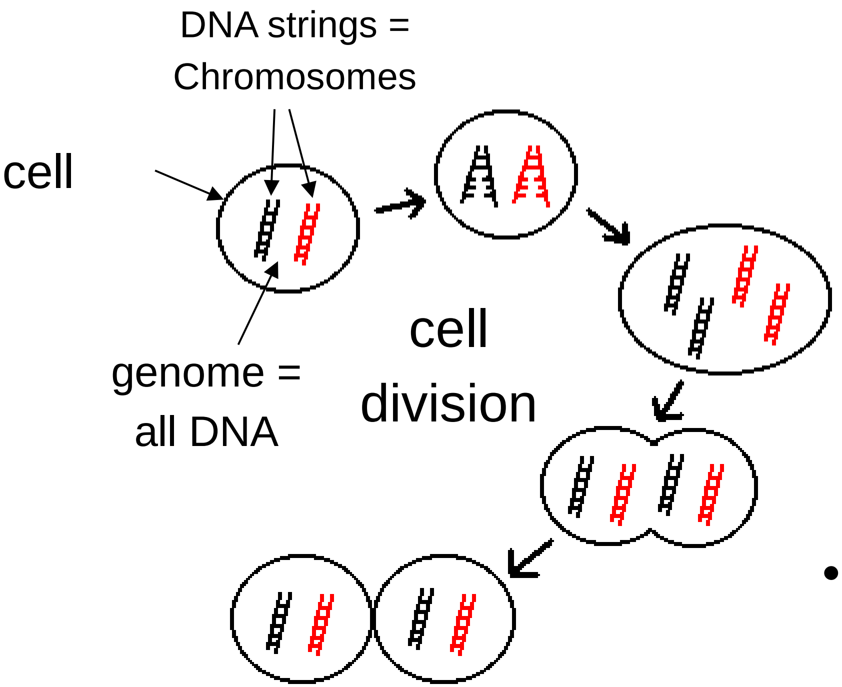
The most amazing “Turing tape” in existence, your genome.





How to Read The Genome

- Genome = DNA.
- Genome is broken up into several strings = chromosomes.
- Humans: Females= (2*chr.1-22)+XX Males= (2*chr.1-22)+XY

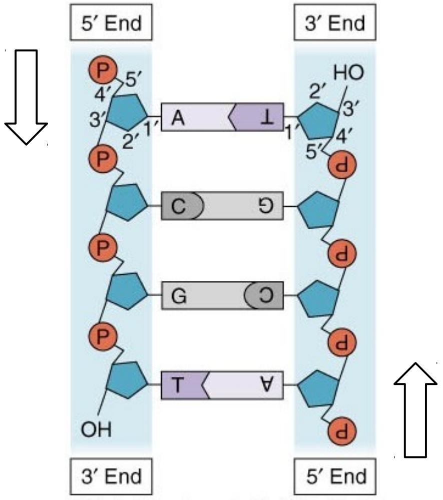


DNA strings = Chromosomes

cell

genome = all DNA

cell division



- DNA is double stranded.
- Complementation is rigid.
- Information can be read off of either strand.

- Every cell contains 2 copies of your genome, one from mom, one from dad.

The Biggest Challenge in Genomics…

… is computational:

How does this encode *this*

```
GCTAGATCGCCTGGTA
GCTTTGCGCCGTCAAA
GTCTTGAAGGCTGTGA
TCAAGCTTCTTGCGAT
CCCGTTTGACCGGAGC
CTTGCCAATGAGTTCT
CAGCTGTCTATATGAA
TCACAAAATACGCAAT
```

Program

Output

This "coding" question has profound implications for our lives

Class Topics

(0) Genome context:
cells, DNA, central dogma

(1) Genome content / genome function:
genes, gene regulation, repeats, epigenetics

(2) Genome sequencing:
technologies, assembly/analysis, technology dependence

(3) Genome evolution:
evolution = mutation + selection, modes of evolution,
comparative genomics, ultraconservation, exaptation

(4) Population genomics:
Tracking human migration patterns via neutral evolution

(5) Genomics of human disease:
disease susceptibility, cancer genomics, personal genomics

(6) Genome “output” (organism) evolution:
Evolutionary developmental biology (“evo-devo”)

Genome Content

```
TATTGAATTTTCAAAAATTCTTACTTTTTTTTGGATGGACGCAAAGAAGTTTAATAATCATATTACATGGCATTACCACCATATA
ATCCATATCTAATCTTACTTATATGTTGTGGAAATGTAAAGAGCCCCATTATCTTAGCCTAAAAAAACCTTCTCTTTGGAACTTTC
AATACGCTTAACTGCTCATTGCTATATTGAAGTACGGATTAGAAGCCGCCGAGCGGGCGACAGCCCTCCGACGGAAGACTCTCCTC
GCGTCCTCGTCTTCACCGGTCGCGTTCCTGAAACGCAGATGTGCCTCGCGCCGCACTGCTCCGAACAATAAAGATTCTACAATACT
TTTTATGGTTATGAAGAGGAAAAATTGGCAGTAACCTGGCCCCACAAACCTTCAAATTAACGAATCAAATTAACAACCATAGGATG
ATGCGATTAGTTTTTTAGCCTTATTTCTGGGGTAATTAATCAGCGAAGCGATGATTTTTGATCTATTAACAGATATATAAATGGAA
CTGCATAACCACTTTAACTAATACTTTCAACATTTTCAGTTTGTATTACTTCTTATTCAAATGTCATAAAAGTATCAACAAAAAT
TAATATACCTCTATACTTTAACGTCAAGGAGAAAAAACTATAATGACTAAATCTCATTCAGAAGAAGTGATTGTACCTGAGTTCAA
TAGCGCAAAGGAATTACCAAGACCATTGGCCGAAAAGTGCCCGAGCATAATTAAGAAATTTATAAGCGCTTATGATGCTAAACCGG
TTGTTGCTAGATCGCCTGGTAGAGTCAATCTAATTGGTGAACATATTGATTATTGTGACTTCTCGGTTTTACCTTTAGCTATTGAT
GATATGCTTTGCGCCGTCAAAGTTTTGAACGATGAGATTTCAAGTCTTAAAGCTATATCAGAGGGCTAAGCATGTGTATTCTGAAT
TAAGAGTCTTGAAGGCTGTGAAATTAATGACTACAGCGAGCTTTACTGCCGACGAAGACTTTTTCAAGCAATTTGGTGCCTTGATG
GAGTCTCAAGCTTCTTGCGATAAACTTTACGAATGTTCTTGTCCAGAGATTGACAAAATTTGTTCCATTGCTTTGTCAAATGGATG
TGGTTCCCGTTTGACCGGAGCTGGCTGGGGTGGTTGTACTGTTCACTTGGTTCCAGGGGGCCCAAATGGCAACATAGAAAAGGTAA
AAGCCCTTGCCAATGAGTTCTACAAGGTCAAGTACCCTAAGATCACTGATGCTGAGCTAGAAAATGCTATCATCGTCTCTAAACCA
TTGGGCAGCTGTCTATATGAATTAGTCAAGTATACTTCTTTTTTTTACTTTGTTCAGAACAACTTCTCATTTTTTTCTACTCATAA
TAGCATCACAAA[illegible]CAATAATAACGAGTAGTAACACTTTTATAGTTCA[illegible]TGCTTCAACTAC[illegible]TAATAAATGATTG[illegible]ATGA
TGTTTTCAATG[illegible]AAGAGAT[illegible]
TACCTATTCTT[illegible]
GGCAAGTTGCCAA[illegible]
CAATGTAAGAGATTTCGATTATCCACAAACTTTAAAACACAGGGACAAAATTCTTGATATGCTTTCAACCGCTGCGTTTTGGATAC
TTCTTGACATGATATGACTACCATTTTGTTATTGTACGTGGGGCAGTTGACGTCTTATCATATGTCAAAGTCATTTGCGAAGTTCT
CAAGTTGCCAACTGACGAGATGCAGTTTCCTACGCATAATAAGAATAGGAGGGAATATCAAGCCAGACAATCTATCATTACATTTA
GGCTCTTCAAAAAGATTGAACTCTCGCCAACTTATGGAATCTTCCAATGAGACCTTTGCGCCAAATAATGTGGATTTGGAAAAAGA
TAAGTCATCTCAGAGTAATATAACTACCGAAGTTTATGAGGCATCGAGCTTTGAAGAAAAAGTAAGCTCAGAAAAACCTCAATACA
CATTCTGGAAGAAAATCTATTATGAATATGTGGTCGTTGACAAATCAATCTTGGGTGTTTCTATTCTGGATTCATTTATGTACAAC
GACTTGAAGCCCGTCGAAAAAGAAAGGCGGGTTTGGTCCTGGTACAATTATTGTTACTTCTGGCTTGCTGAATGTTTCAATATCAA
TTGGCAAATTGCAGCTACAGGTCTACAACTGGGTCTAAATTGGTGGCAGTGTTGGATAACAATTTGGATTGGGTACGGTTTCGTTG
CTTTTGTTGTTTTGGCCTCTAGAGTTGGATCTGCTTATCATTTGTCATTCCCTATATCATCTAGAGCATCATTCGGTATTTTCTTC
TTATGGCCCGTTATTAACAGAGTCGTCATGGCCATCGTTTGGTATAGTGTCCAAGCTTATATTGCGGCAACTCCCGTATCATTAAT
GAAATCTATCTTTGGAAAAGATTTACAATGATTGTACGTGGGGCAGTTGACGTCTTATCATATGTCAAAGTCATTTGCGAAGTTCT
CAAGTTGCCAACTGACGAGATGCAGTAACACTTTTATAGTTCATACATGCTTCAACTACTTAATAAATGATTGTATGATAATGTTT
ATGTAAGAGATTTCGATTATCCACAAACTTTAAAACACAGGGACAAAATTCTTGATATGCTTTCAACCGCTGCGTTTTGGATACCT
CTTGACATGATATGACTACCATTTTGTTATTGTTTATAGTTCATACATGCTTCAACTACTTAATAAATGATTGTATGATAATGTTT
ATGTAAGAGATTTCGATTATCCTTATAGTTCATACATGCTTCAACTACTTAATAAATGATTGTATGATAATGTTTTCAATGTAAGA
TTCGATTATCCTTATAGTTCATACATGCTTCAACTACTTAATAAATGATTGTATGATAATGTTTTCAATGTAAGAGATTTCGATTA
TTATAGTTCATACATGCTTCAACTACTTAATAAATGATTGTATGATAATGTTTTCAATGTAAGAGATTTCGATTATCCTTATAGTT
ACATGCTTCAACTACTTAATAAATGATTGTATGATAATGTTTTCAATGTAAGAGATTTCGATTATCCTTATAGTTCATACATGCTT
CTACTTAATAAATGATTGTATGATAATGTTTTCAATGTAAGAGATTTCGATTATCCTTATAGTTCATACATGCTTCAACTACTTAA
ATGATTGTATGATAATGTTTTCAATGTAAGAGATTTCGATTATCCTTATAGTTCATACATGCTTCAACTACTTAATAAATGATTG
```

Genomes, Genes & Proteins

The most visible instructions in our genome are Genes.

Genes explain exactly HOW to synthesize any protein.

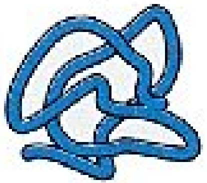
Proteins are the work horses of every living cell.

🟩 gene

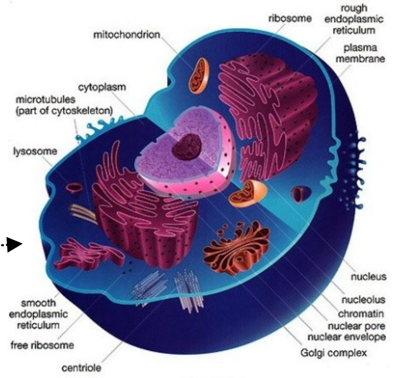
Genome:

...ACGTACGACTGACTAGCATCGACTACGACTAGCAC...

linear (folded) molecule

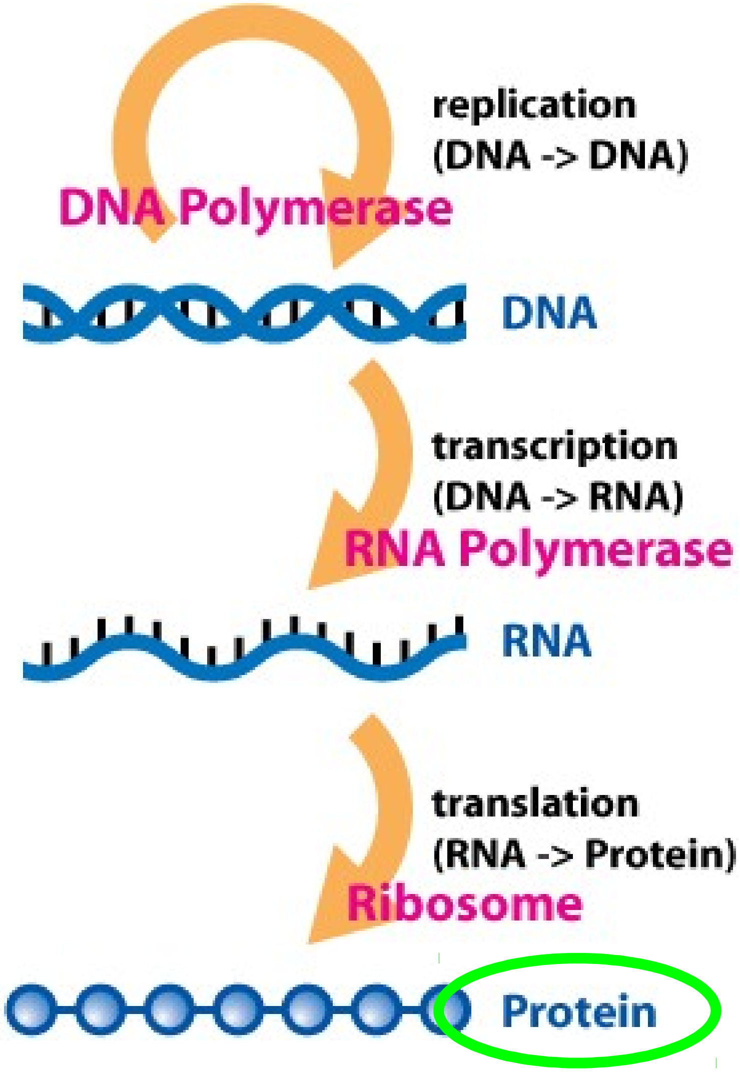


protein



cell

Central Dogma of Biology



genome

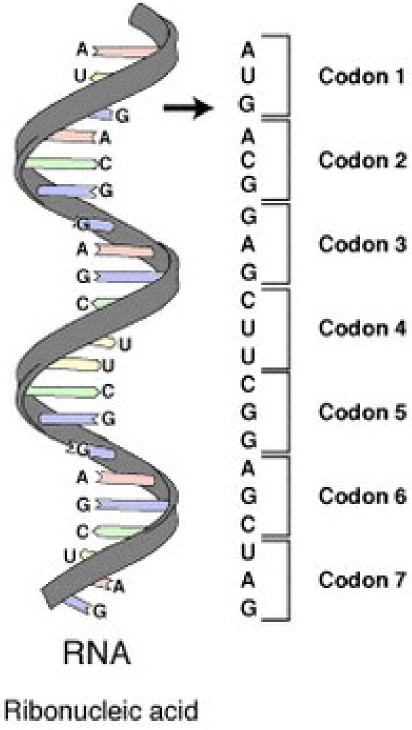
{A,C,G,T}

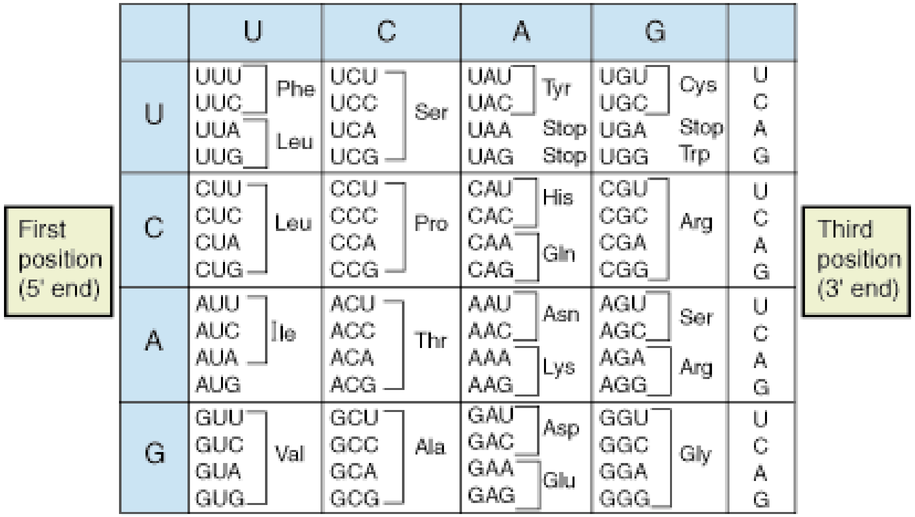
{A,C,G,U}

next page

{20 letters}

Translation: The Genetic Code





First position (5' end)	U		C		A		G		Third position (3' end)
U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys	U
	UUC		UCC		UAC		UGC		C
	UUA	Leu	UCA		UAA	Stop	UGA	Stop	A
	UUG		UCG		UAG	Stop	UGG	Trp	G
C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	Gln	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	Lys	AGA	Arg	A
	AUG		ACG		AAG		AGG		G
G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	Glu	GGA		A
	GUG		GCG		GAG		GGG		G

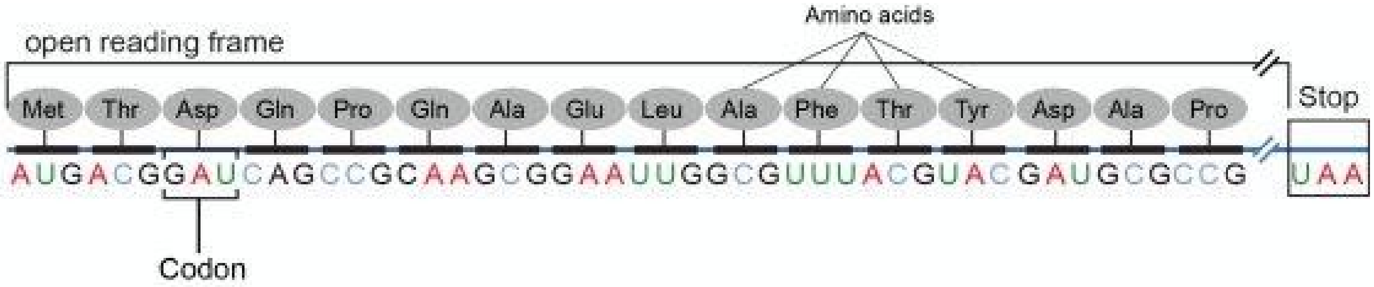
Amino acid names:

Ala = alanine
Arg = arginine
Asn = asparagine
Asp = aspartate
Cys = cysteine

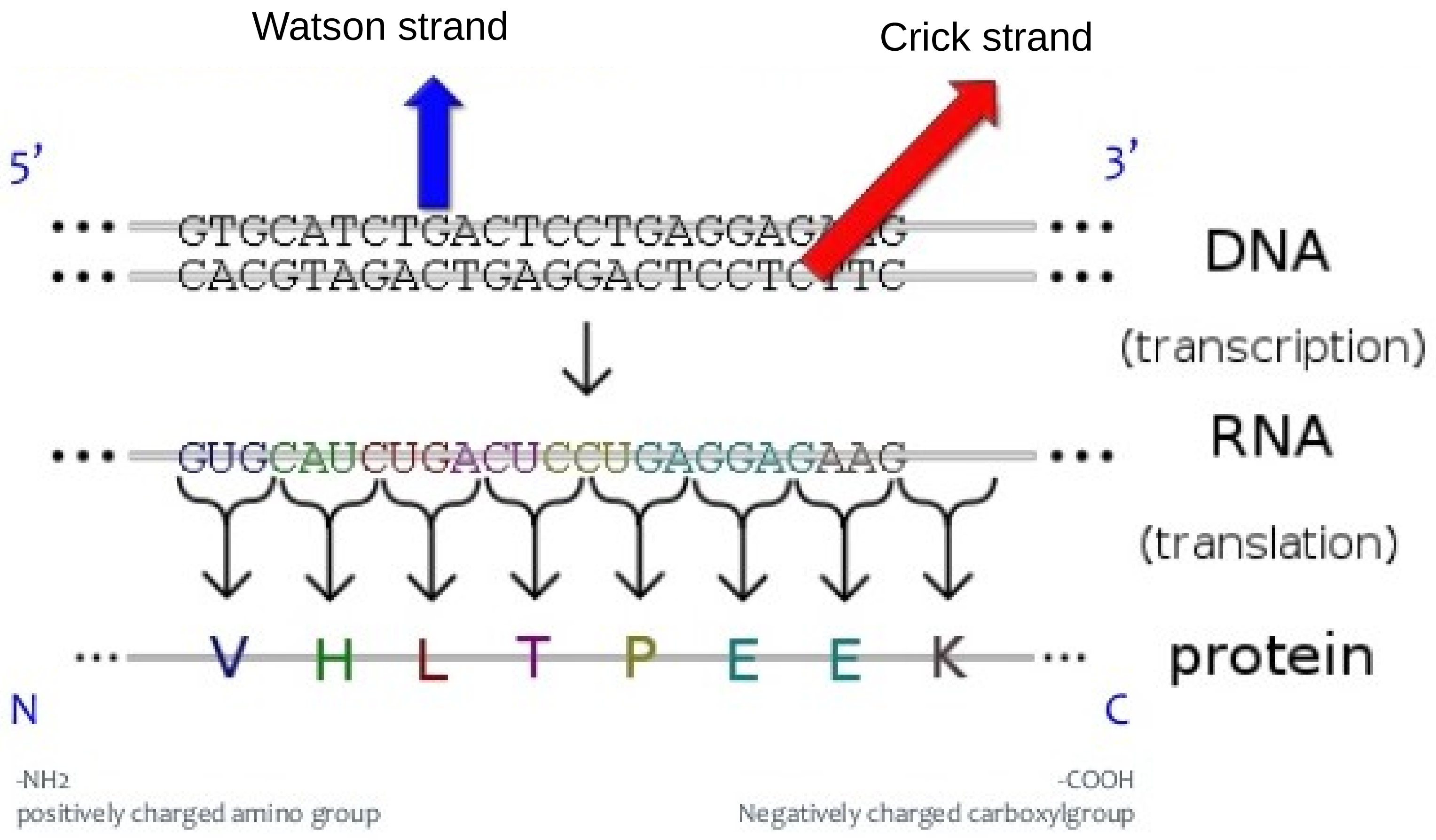
Gln = glutamine
Glu = glutamate
Gly = glycine
His = histidine
Ile = Isoleucine

Leu = leucine
Lys = lysine
Met = methionine
Phe = phenylalanine
Pro = proline

Ser = serine
Thr = threonine
Trp = tryptophan
Tyr = Tyrosine
Val = valine



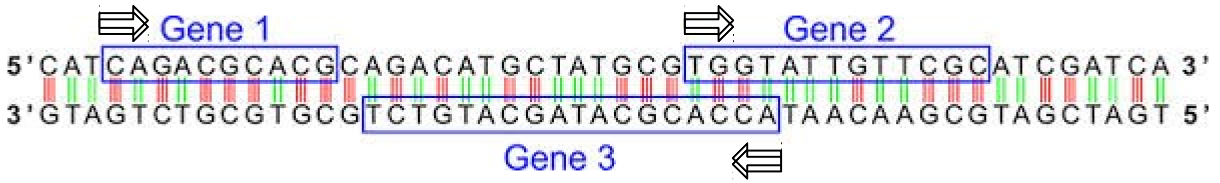
Genes Can Be Encoded on Either Strand



Watson strand

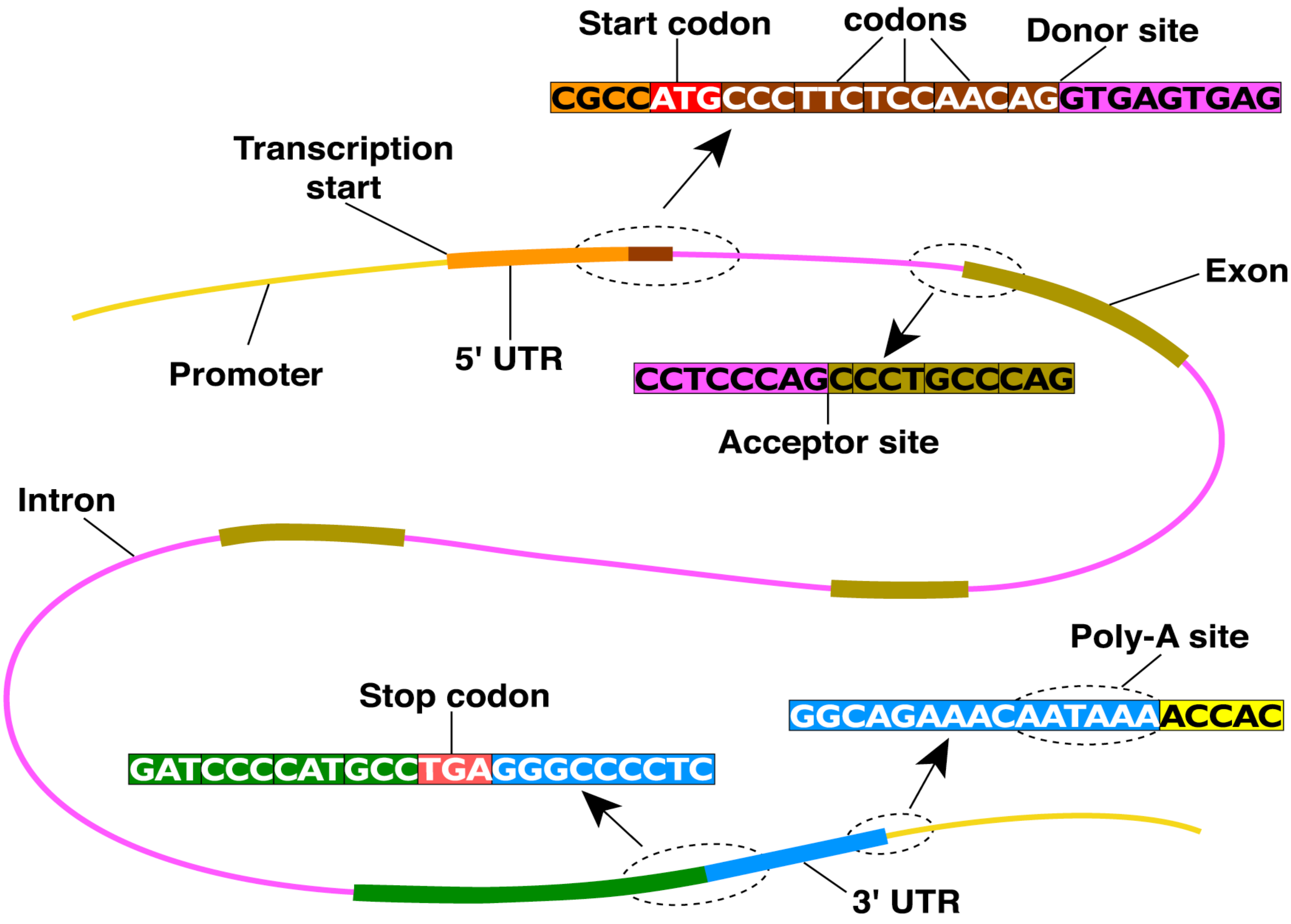
Crick strand

5'
3'
... GTGCATCTGACTCCTGAGGAGAAG ...
... CACGTAGACTGAGGACTCCTCTTC ...
DNA
(transcription)
... GUGCAUCUGACUCCUGAGGAGAAG ...
RNA
(translation)
... V H L T P E E K ...
protein
N
C
-NH2
positively charged amino group
-COOH
Negatively charged carboxylgroup

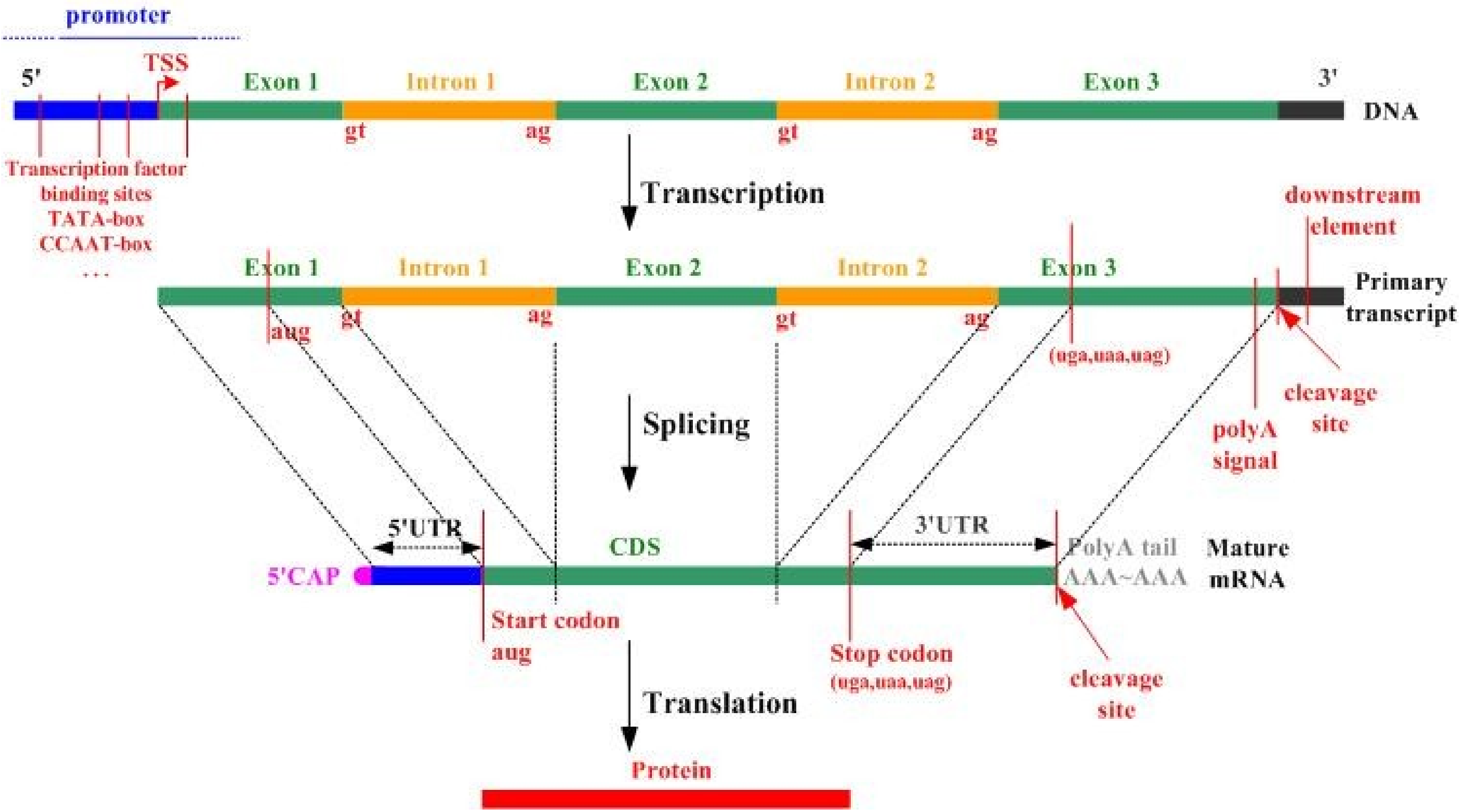


Gene 1
Gene 2
5'CATCAGACGCACGCAGACATGCTATGCGTGGTATTGTTCGCATCGATCA 3'
3'GTAGTCTGCGTGCGTCTGTACGATACGCACCATAACAAGCGTAGCTAGT 5'
Gene 3

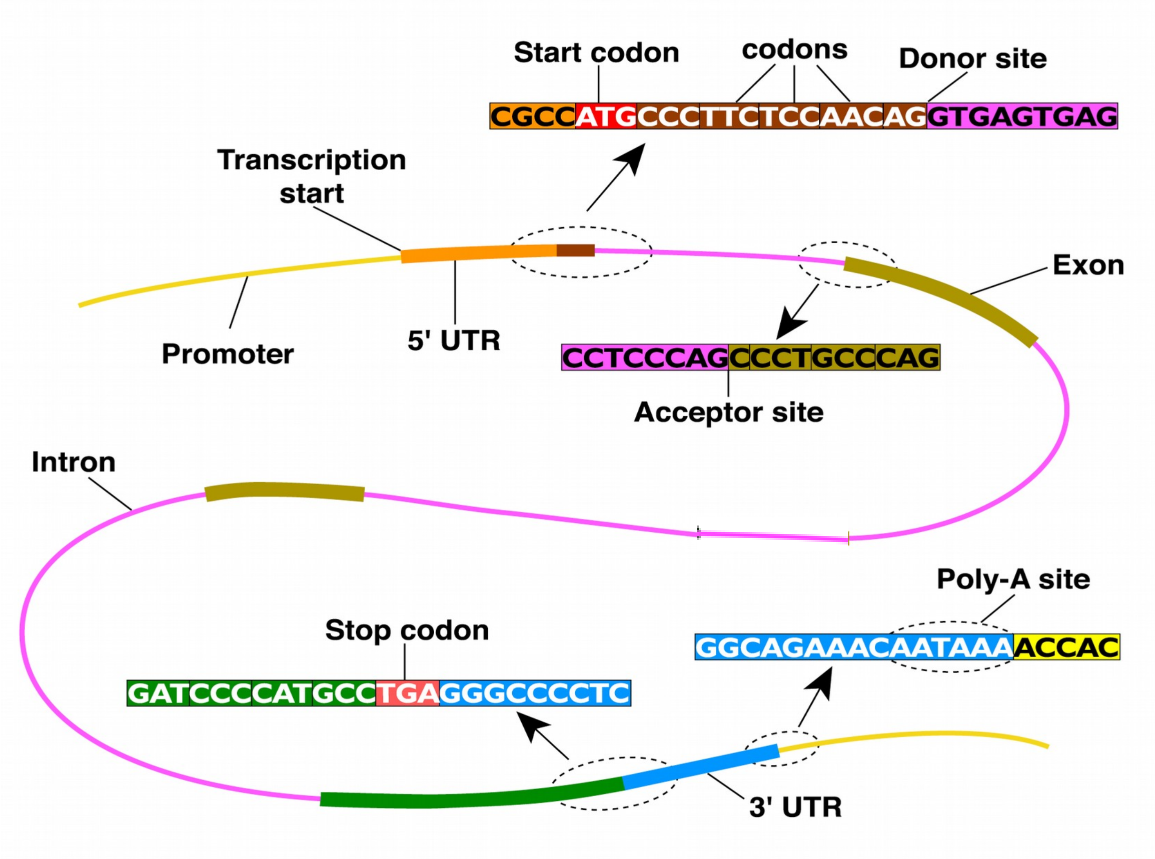
Gene Structure

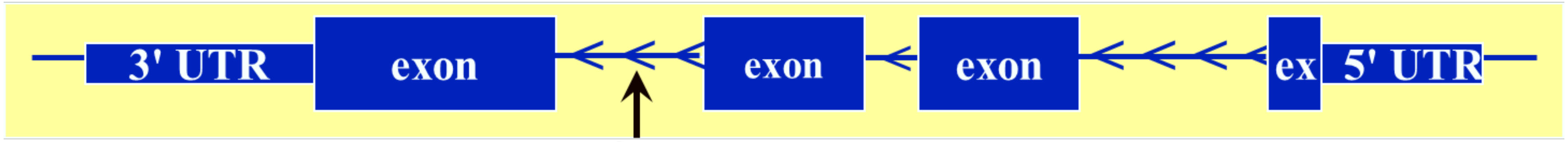


Gene Splicing



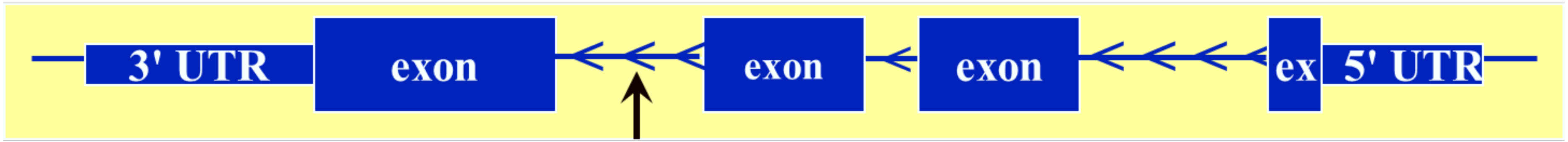
Visualizing Gene Structure



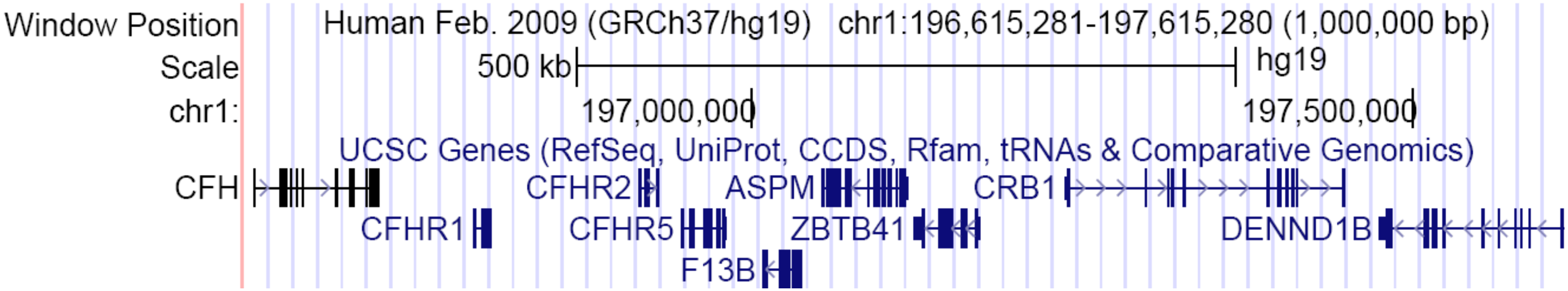


Intron, and direction of transcription <<< or >>>

Genes in the Human Genome



<u>Intron,</u> and <u>direction</u> of transcription <<< or >>>


Window Position
Human Feb. 2009 (GRCh37/hg19) chr1:196,615,281-197,615,280 (1,000,000 bp)
Scale
500 kb
hg19
chr1:
197,000,000
197,500,000
UCSC Genes (RefSeq, UniProt, CCDS, Rfam, tRNAs & Comparative Genomics)
CFH
CFHR2
ASPM
CRB1
CFHR1
CFHR5
ZBTB41
DENND1B
F13B

UCSC primer

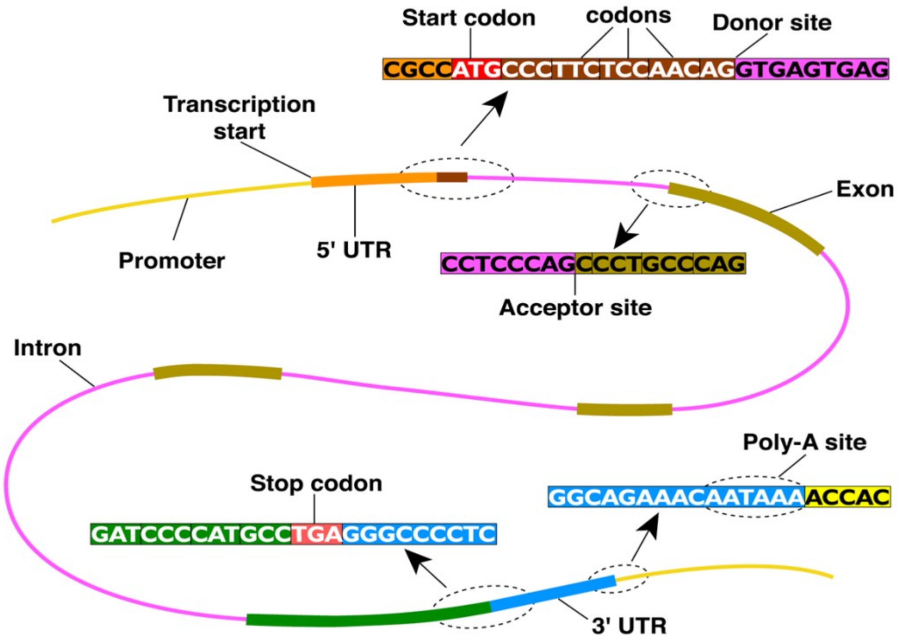
There are ~20,000 protein coding genes in the human genome.

(Even half way through sequencing the human genome,
Researchers thought there will be well over 100,000 genes).

Gene Finding

Computational Challenge:

“Find the genes, the whole genes, and nothing but the genes”

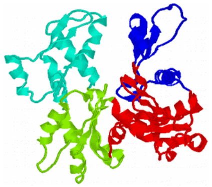


Understand Biology Write discovery tools

(<u>Our</u>) answer depends on our understanding, data & tools

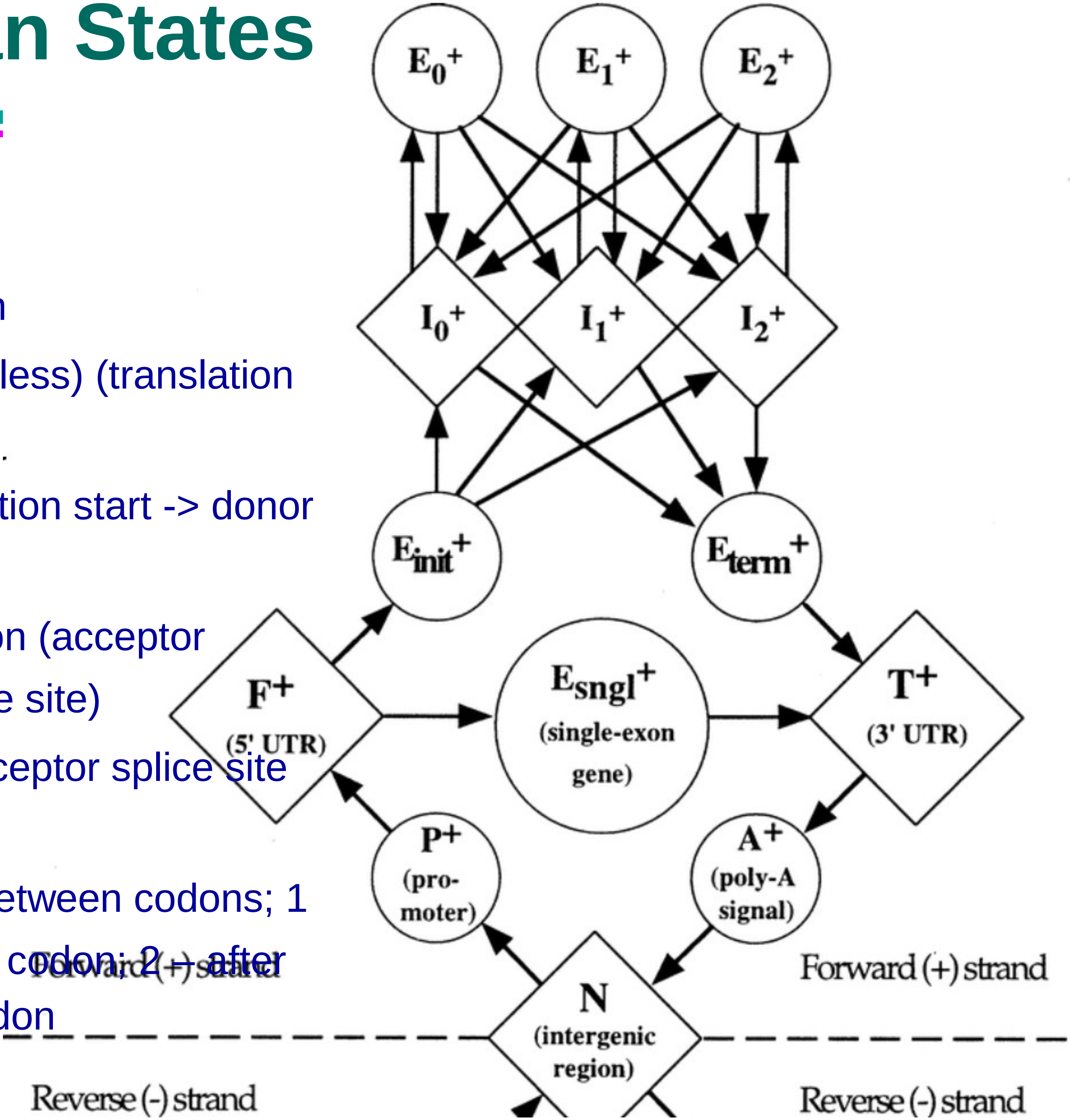
Gene prediction approaches

- Rule-based programs
  - Use explicit set of rules to make decisions.
  - Example: GeneFinder
- Neural Network-based programs
  - Use data set to build rules.
  - Examples: Grail, GrailEXP
- Hidden Markov Model-based programs
  - Use probabilities of states and transitions between these states to predict features.
  - Examples: Genscan, GenomeScan

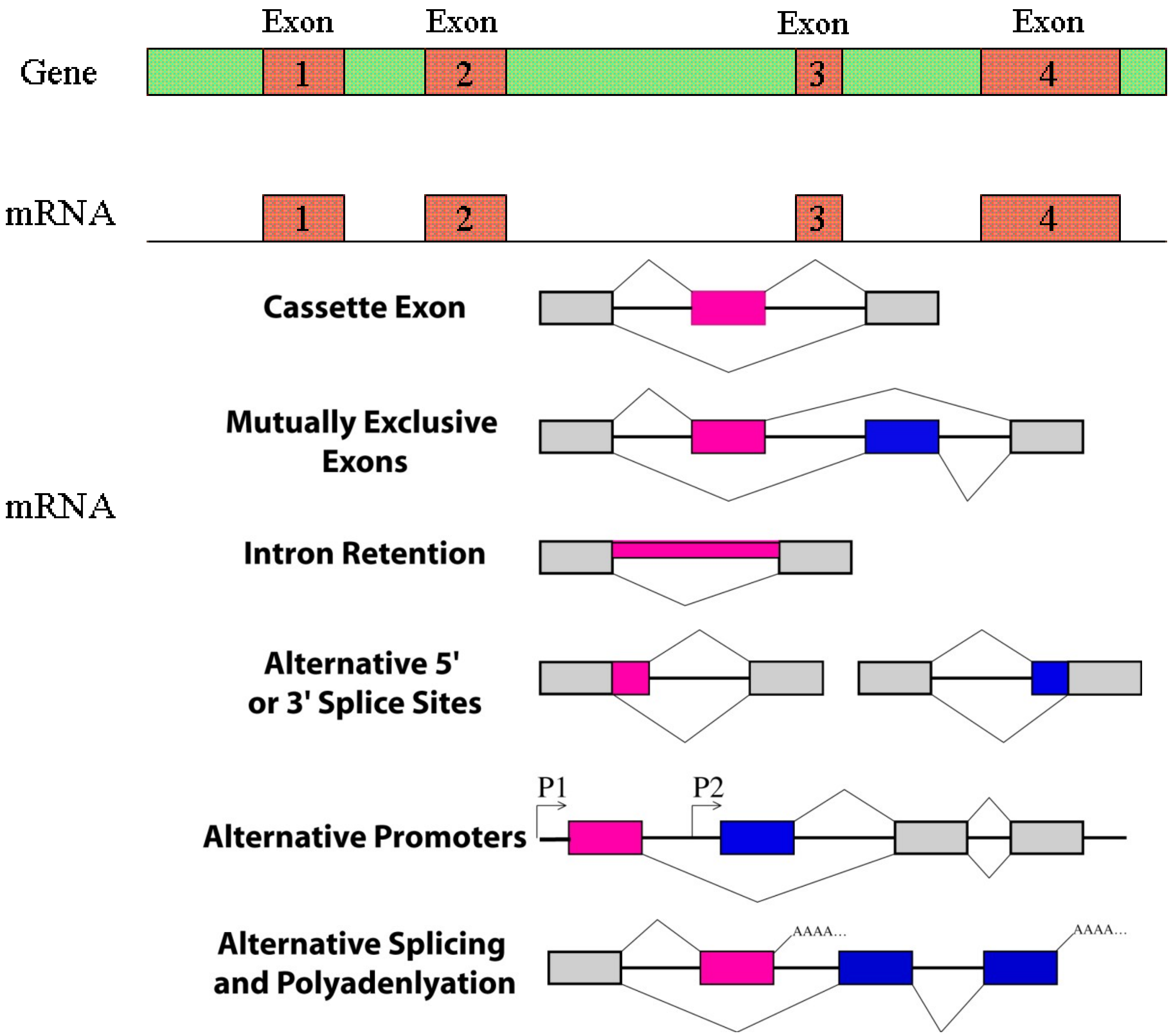


GenScan States

- N - intergenic region
- P - promoter
- F - 5’ untranslated region
- E_{sngl} – single exon (intronless) (translation start -> stop codon)
- E_{init} – initial exon (translation start -> donor splice site)
- E_k – phase k internal exon (acceptor splice site -> donor splice site)
- E_{term} – terminal exon (acceptor splice site -> stop codon)
- I_k – phase k intron: 0 – between codons; 1 – after the first base of a codon; 2 – after the second base of a codon



Alternative Splicing



Gene

Exon 1 · Exon 2 · Exon 3 · Exon 4

mRNA

1 · 2 · 3 · 4

mRNA

Cassette Exon

Mutually Exclusive Exons

Intron Retention

Alternative 5' or 3' Splice Sites

Alternative Promoters

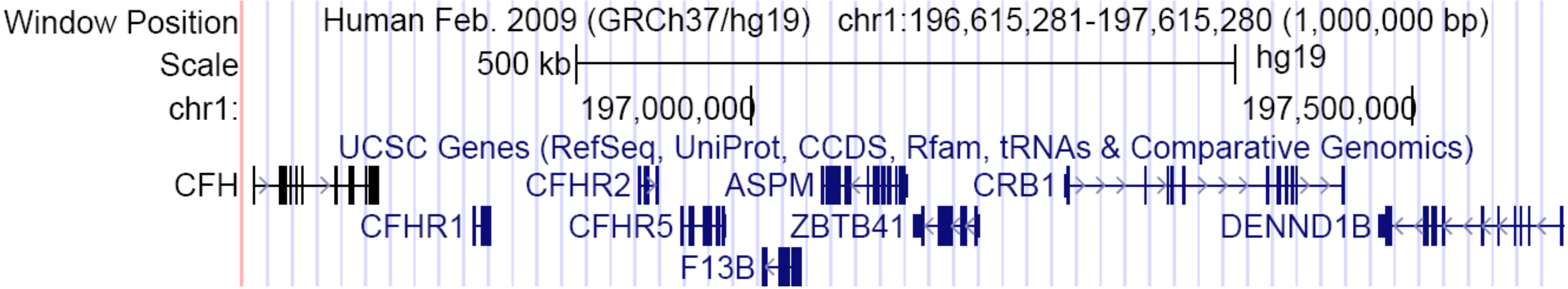
P1 · P2

Alternative Splicing and Polyadenlyation

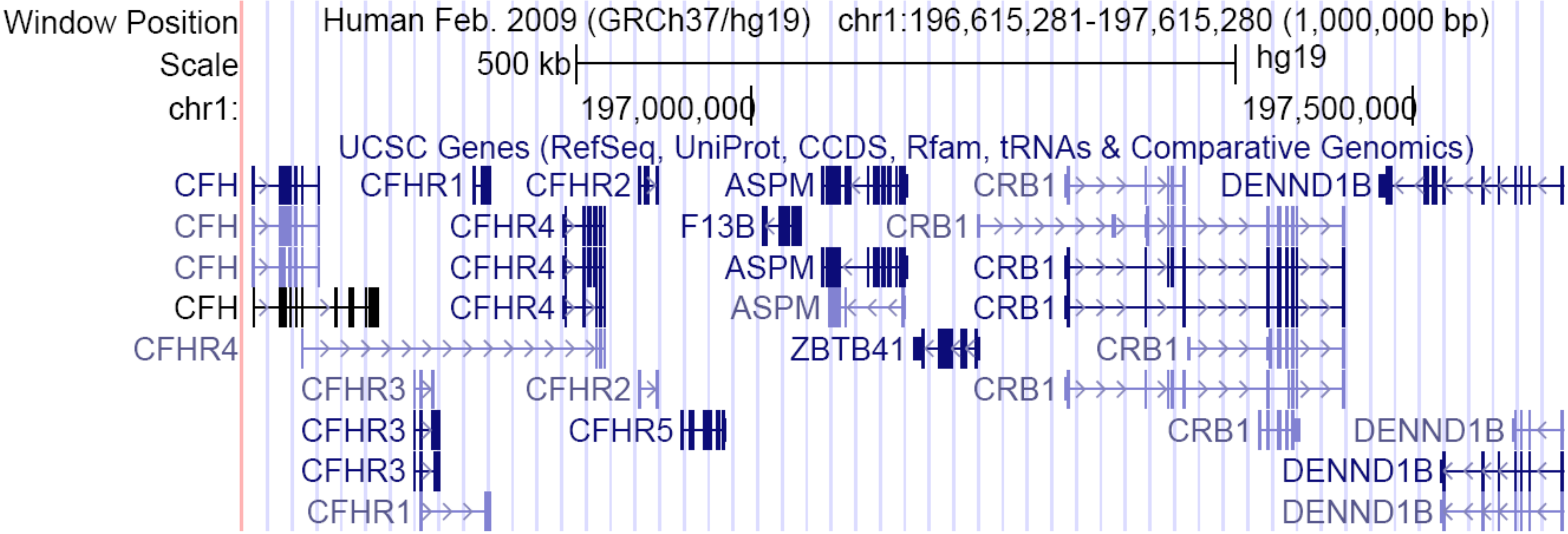
AAAA... · AAAA...

Genes in the Human Genome

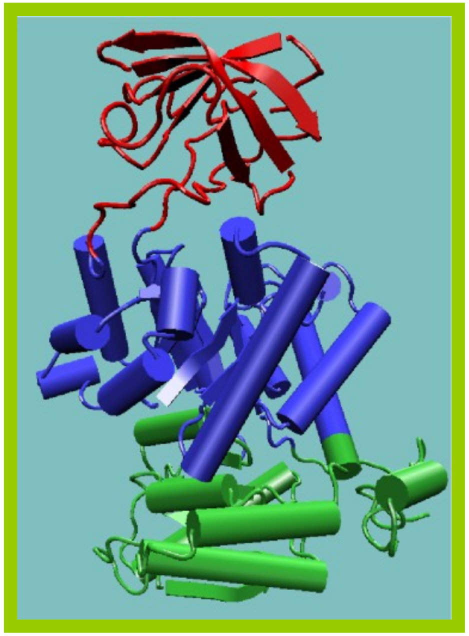
When you only show one transcript per gene locus:


Window Position
Scale
chr1:
Human Feb. 2009 (GRCh37/hg19) chr1:196,615,281-197,615,280 (1,000,000 bp)
500 kb
hg19
197,000,000
197,500,000
UCSC Genes (RefSeq, UniProt, CCDS, Rfam, tRNAs & Comparative Genomics)
CFH
CFHR2
ASPM
CRB1
CFHR1
CFHR5
ZBTB41
DENND1B
F13B

If you ask the GUI to show you all well established gene variants:



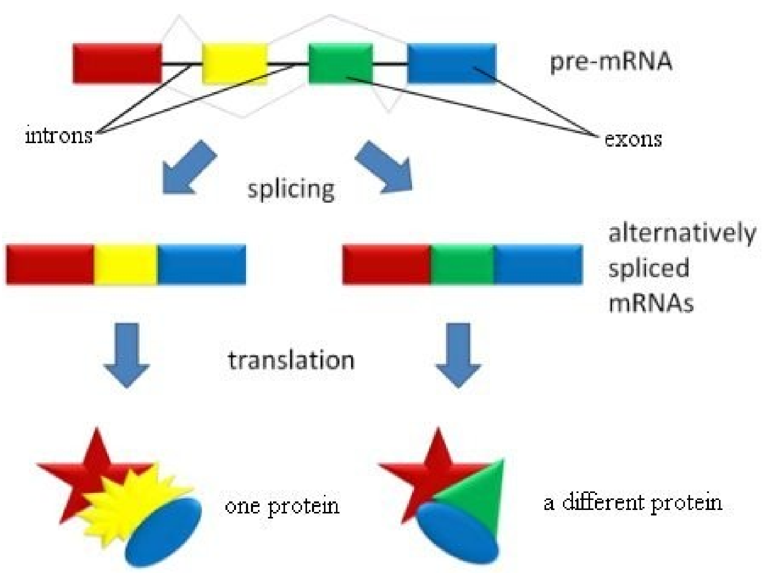
Protein Domains

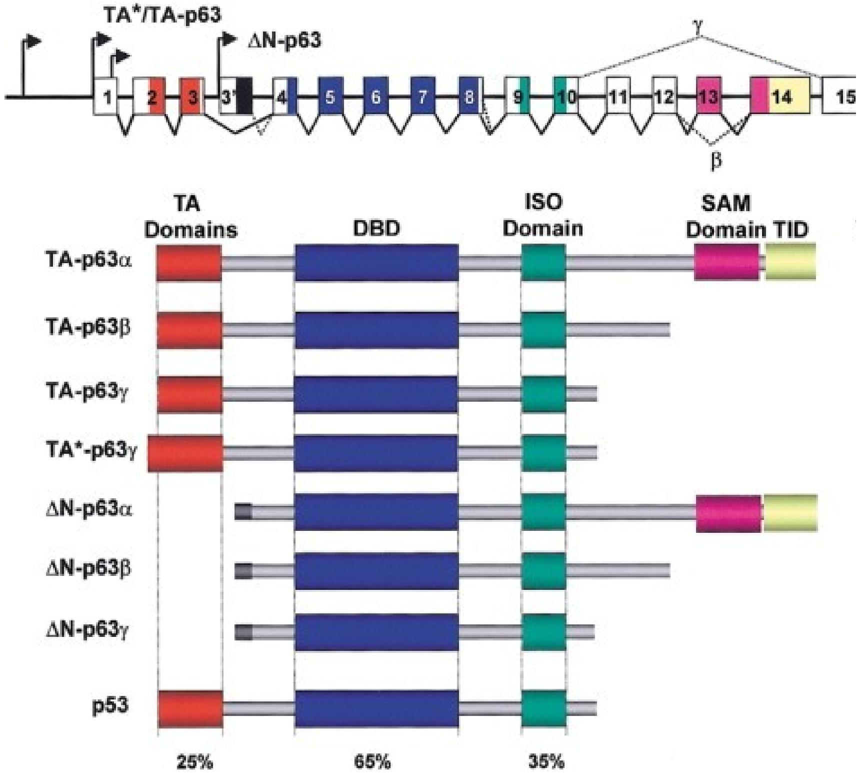


```
SKSHSEAGSAFIQTQQLHAAMADTFLEHMCRLDIDSAPITARNT
GIICTIGPASRSVETLKEMIKSGMNVARMNFSHGTHEYHAETIK
NVRTATESFASDPILYRPVAVALDTKGPEIRTGLIKGSGTAEVE
LKKGATLKITLDNAYMAACDENILWLDYKNICKVVEVGSKVYVD
DGLISLQVKQKGPDFLVTEVENGGFLGSKKGVNLPGAAVDLPAV
SEKDIQDLKFGVDEDVDMVFASFIRKAADVHEVRKILGEKGKNI
KIISKIENHEGVRRFDEILEASDGIMVARGDLGIEIPAEKVFLA
QKMIIGRCNRAGKPVICATQMLESMIKKPRPTRAEGSDVANAVL
DGADCIMLSGETAKGDYPLEAVRMQHLIAREAEAAMFHRKLFEE
LARSSSHSTDLMEAMAMGSVEASYKCLAAALIVLTESGRSAHQV
ARYRPRAPIIAVTRNHQTARQAHLYRGIFPVVCKDPVQEAWAED
VDLRVNLAMNVGKAAGFFKKGDVVIVLTGWRPGSGFTNTMRVVP
VP
```

A protein domain is a subsequence of the protein that folds independently of the other portions of the sequence, and often confers to the protein one or more specific functions.

Alt. Splicing and Protein Repertoire





Alternative splicing often produces protein variants that have a different domain composition, and thus perform different functions.

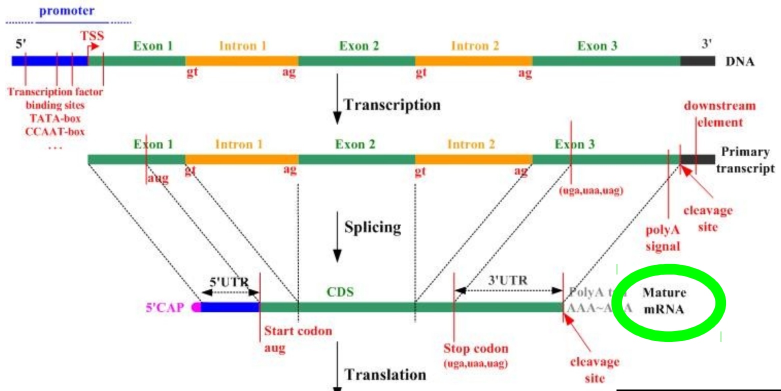
What if we want to predict all splice variants that are ever made? Can we even do it from sequence alone?

Common Problems

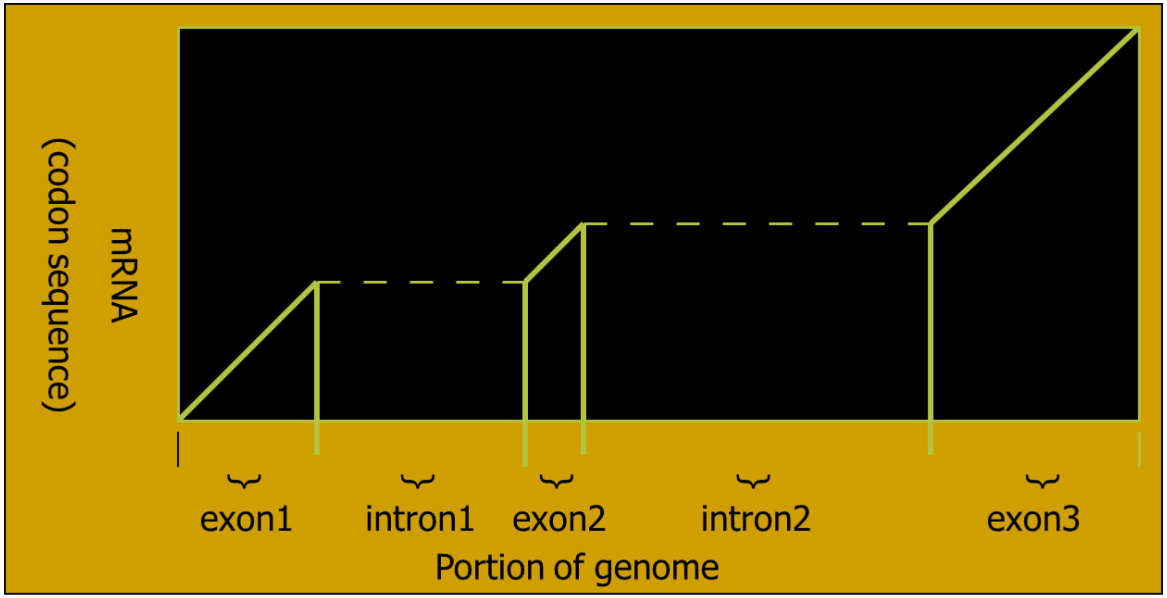
- Common problems with gene finders
  - Fusing neighboring genes
  - Spliting a single gene
  - Miss exons or entire genes
  - Overpredict exons or genes

- Other challenges
  - Nested genes
  - Noncanonical splice spites
  - Pseudogenes
  - Different isoforms of same gene

We can sequence all mRNA of a given cell



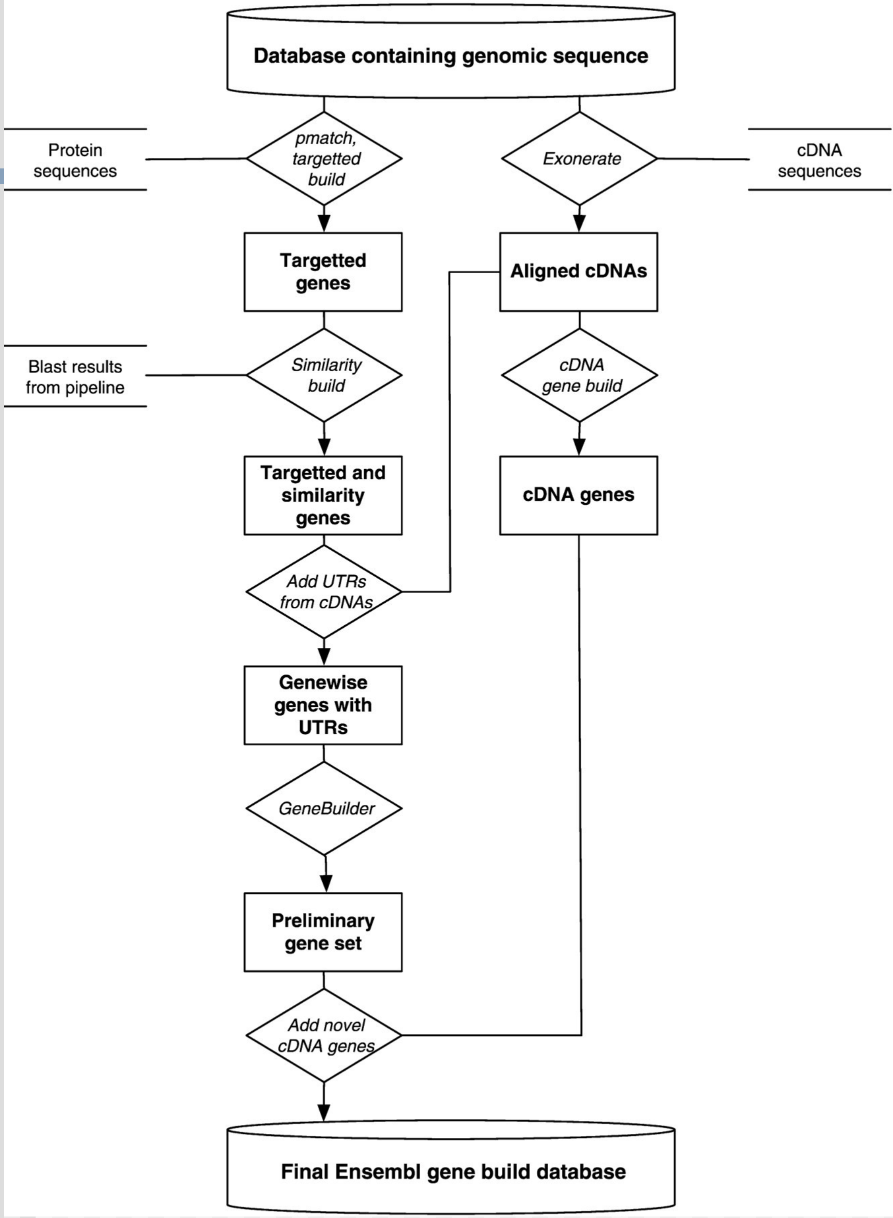
(Great, but not all genes/isoforms are expressed in all cells. Some are very exotic).



AUGGUG - - - -**GGCCCUUUGGGA** - - - - - **CACUAA**

GTGAGG**ATGGTA**AATA**GGGCAT - - - GGA**TTGAG**CACUAA**TAA

Gene Annotation System



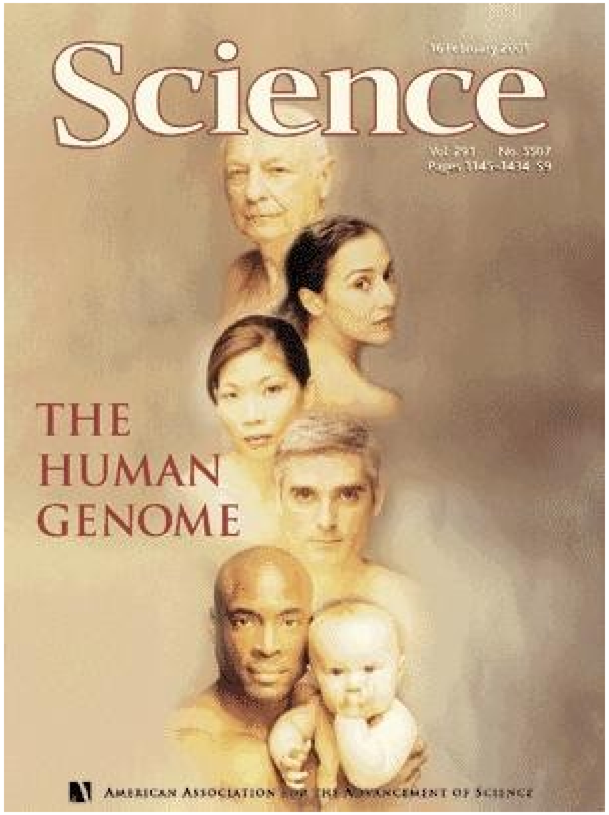
- All Ensembl gene predictions are based on experimental evidence

- Predictions based on manually curated Uniprot/Swissprot/Refseq databases

- UTRs are annotated only if they are supported by EMBL mRNA records

Val Curwen, et al. The Ensembl Automatic Gene Annotation System Genome Res., (2004) ***14*** *942 - 950.*

First full draft of the Human Genome



Human Genome Consortium
(HGC)

Celera

2001

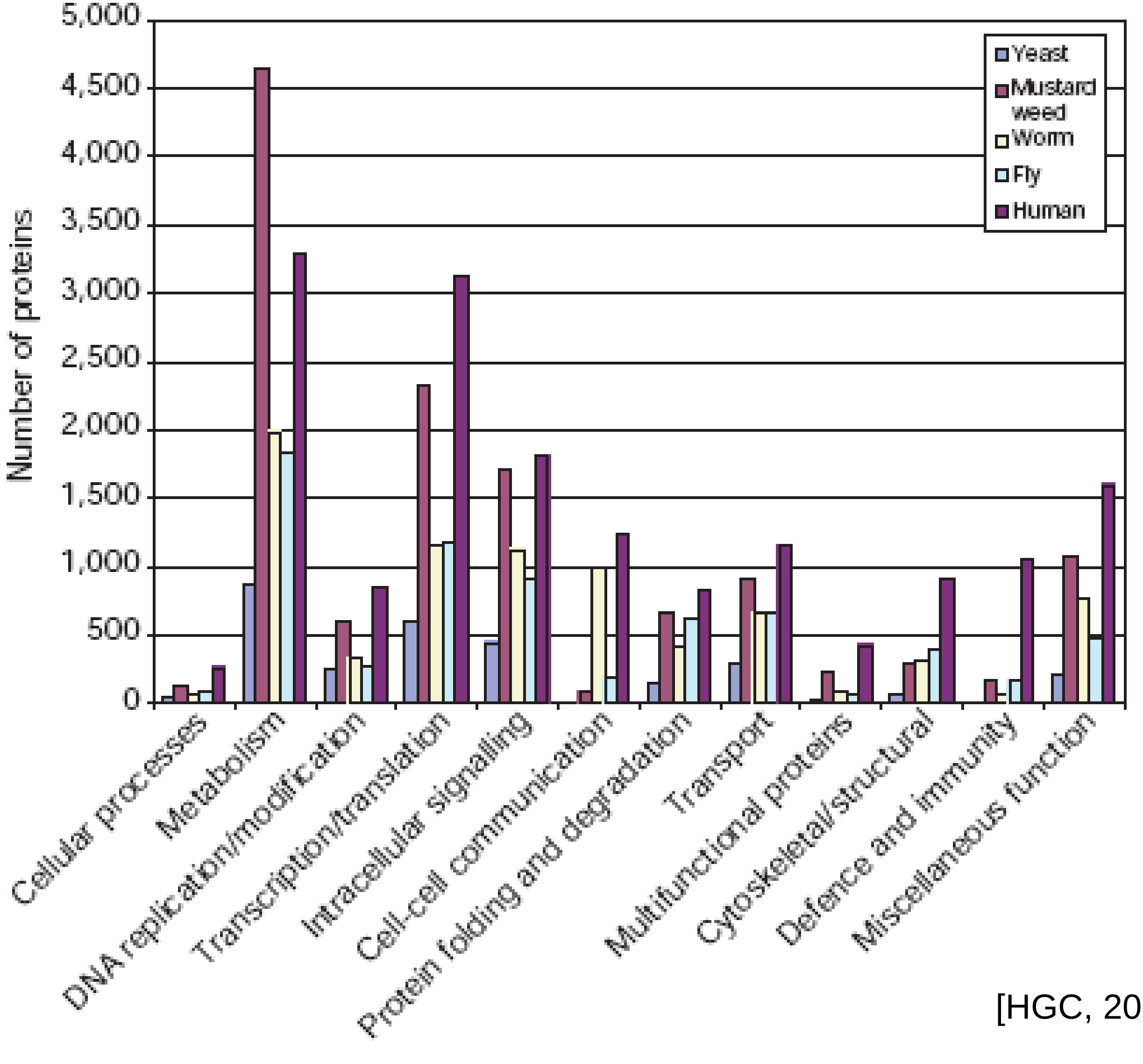
Everything in Genomics is a *Moving* Target

- The genomes (ie, assemblies)
- Their annotations
- Our understanding of Biology
- The portals

Conclusion:
write code
that can be
run...

and rerun
and rerun
and rerun
and rerun

Biological Functions of the Human Gene Set



Focus on
the X axis:

[HGC, 2001]

Molecular Functions of the Human Gene Set

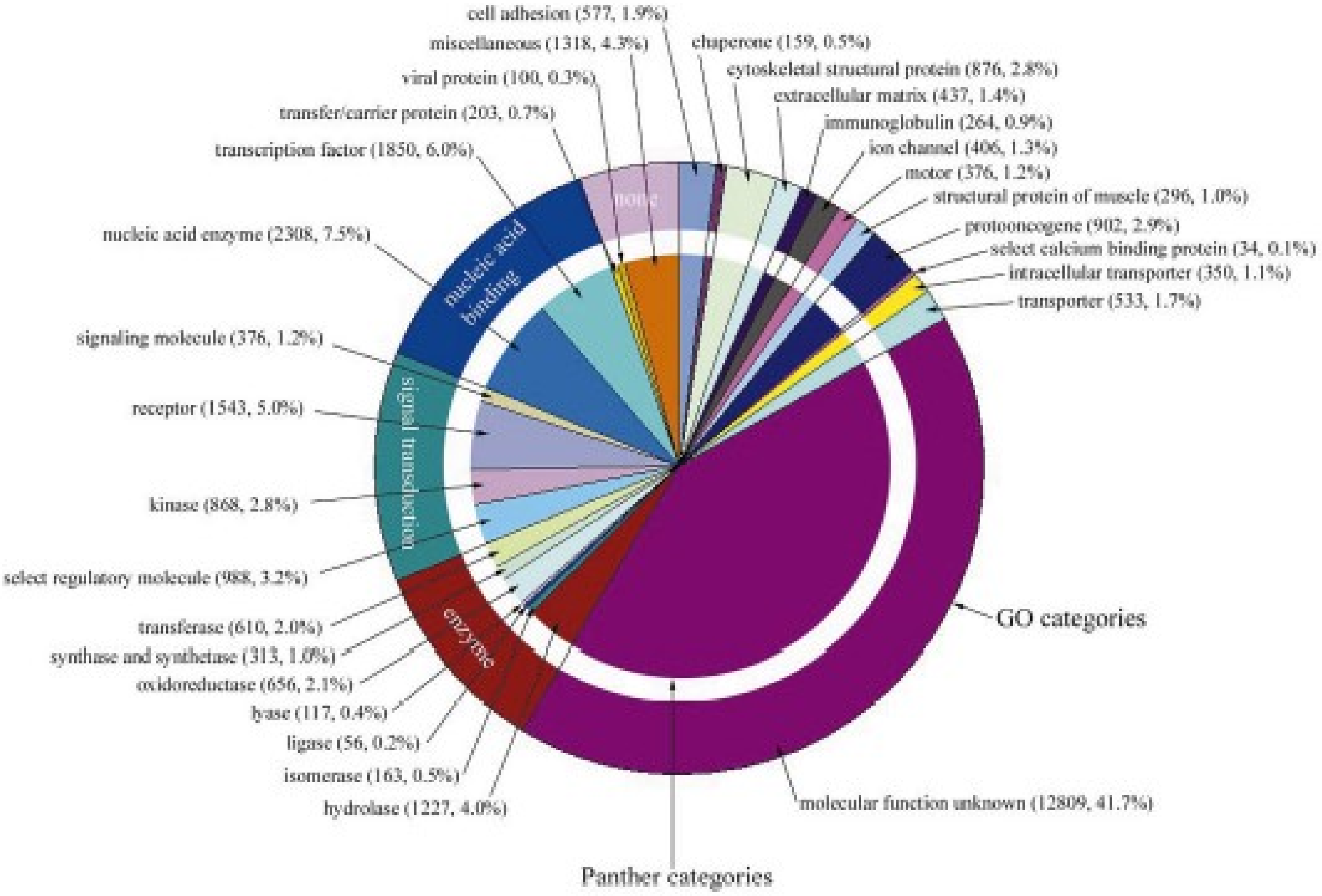
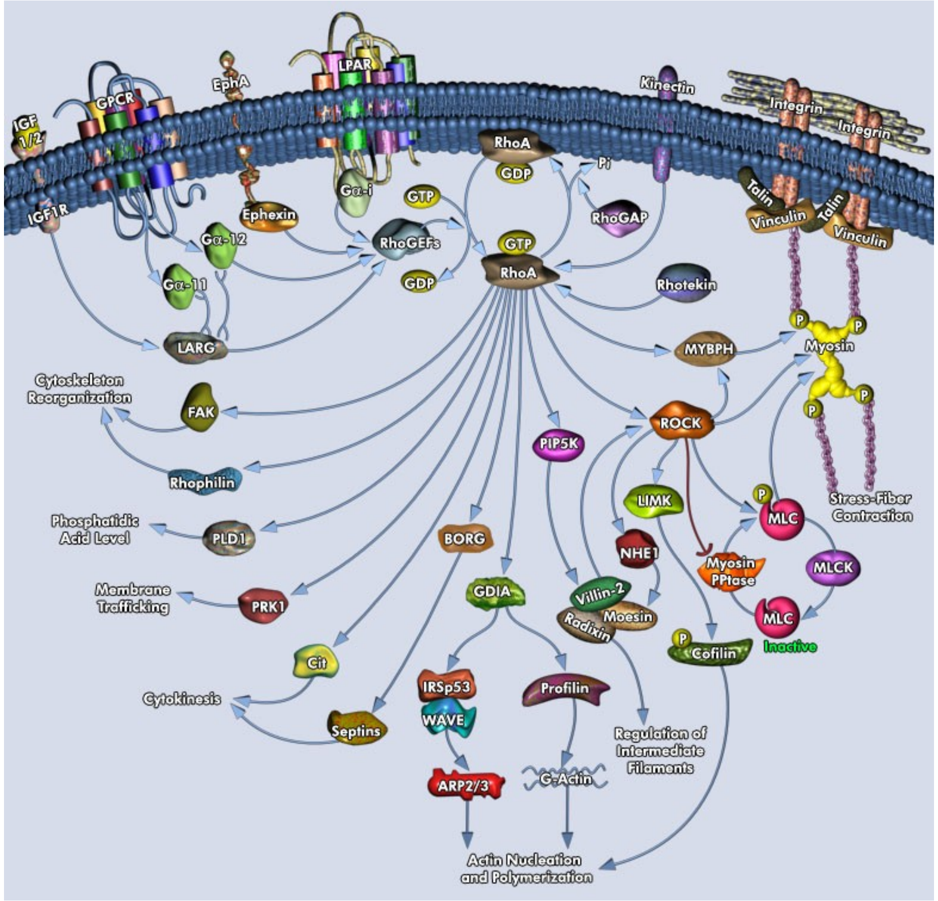


Fig. 15. Distribution of the molecular functions of 26,383 human genes. Each slice lists the numbers and percentages (in parentheses) of human gene functions assigned to a given category of molecular function. The outer circle shows the assignment to molecular function categories in the Gene Ontology (GO) (*179*), and the inner circle shows the assignment to Celera's Panther molecular function categories (*116*).

[Celera, 2001]

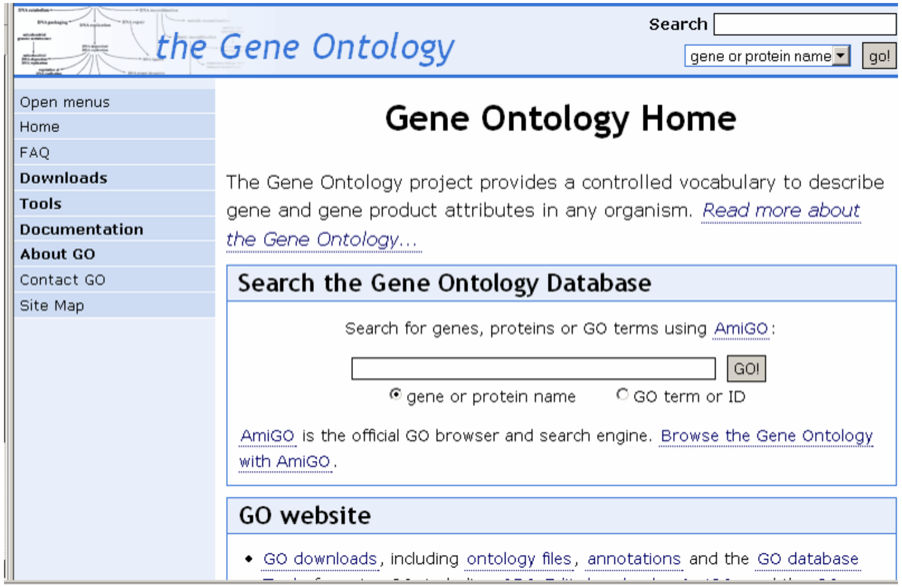
Biological vs. Molecular Function: Pathways



Proteins with very different molecular functions participate to manifest a single biological function, for example: a pathway.

Gene Sets

- Gene Ontology (“GO”)
  - Biological Process
  - Molecular Function
  - Cellular Location
- Pathway Databases
  - KEGG
  - BioCarta
  - Broad Institute
- Multiple others



Gene Ontology Home

The Gene Ontology project provides a controlled vocabulary to describe gene and gene product attributes in any organism. Read more about the Gene Ontology...

Search the Gene Ontology Database

Search for genes, proteins or GO terms using AmiGO:

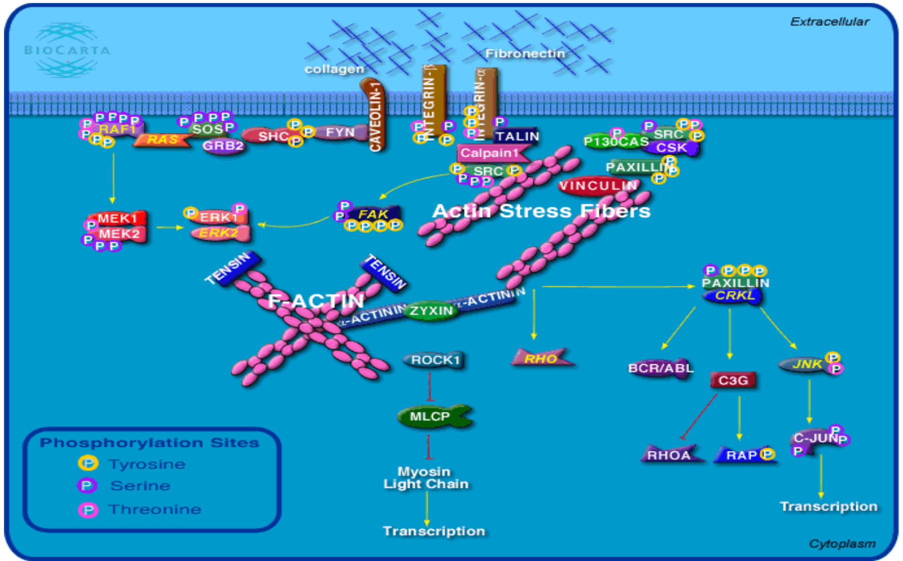
GO!

● gene or protein name ○ GO term or ID

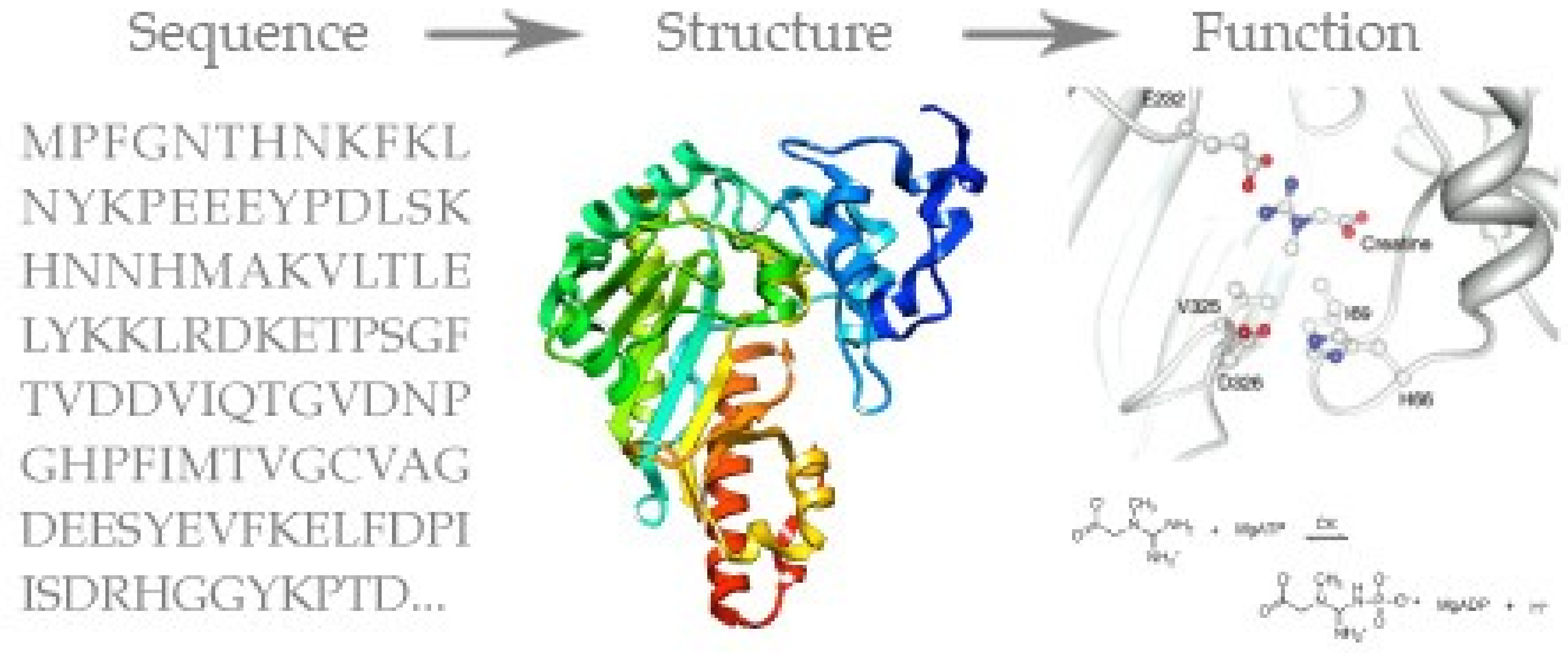
AmiGO is the official GO browser and search engine. Browse the Gene Ontology with AmiGO.

GO website

- GO downloads, including ontology files, annotations and the GO database

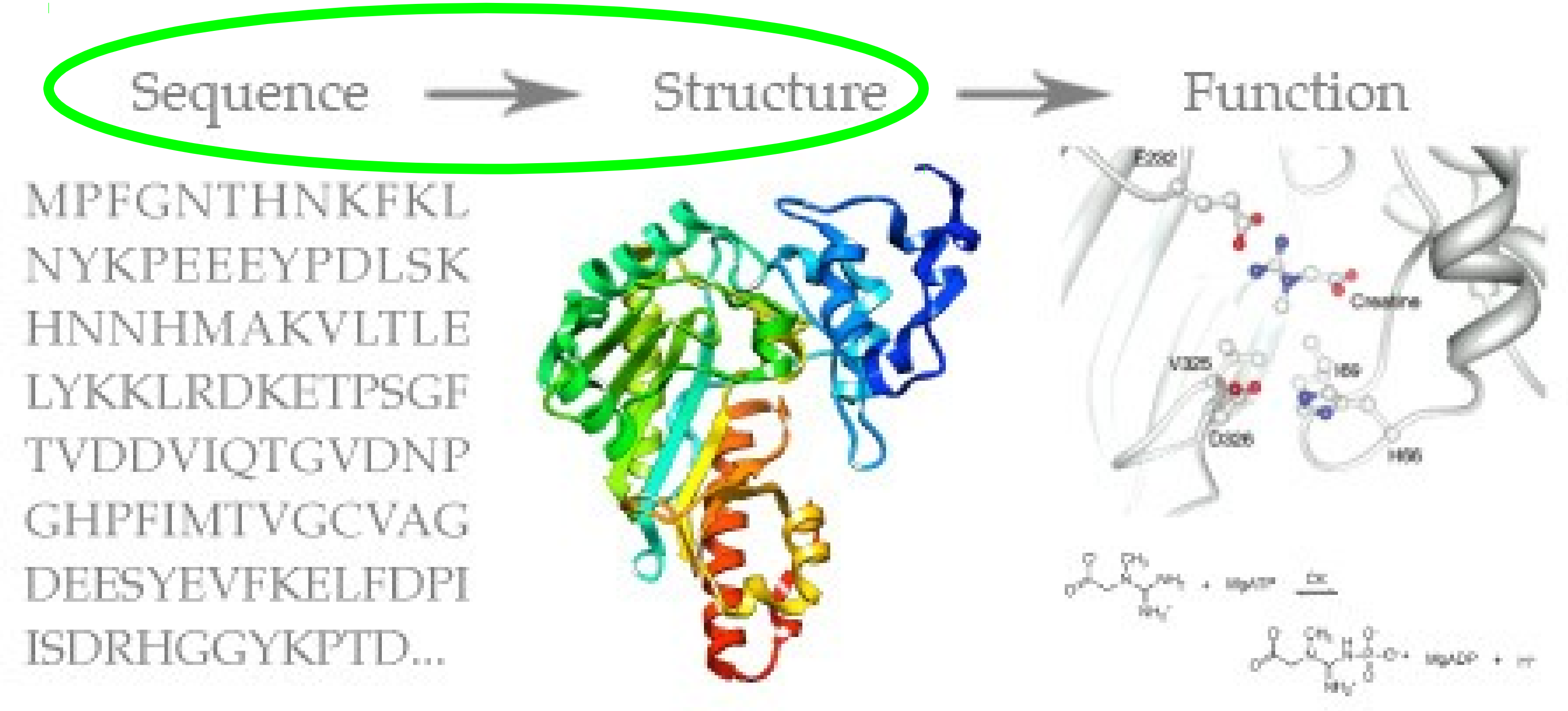


Genes & Their Functions



Gene (DNA) sequence determines protein (AA) sequence,
which determines protein (3D) structure,
which determines protein's function.

Protein Folding



Protein folding is the challenge of deducing protein structure from protein sequence.

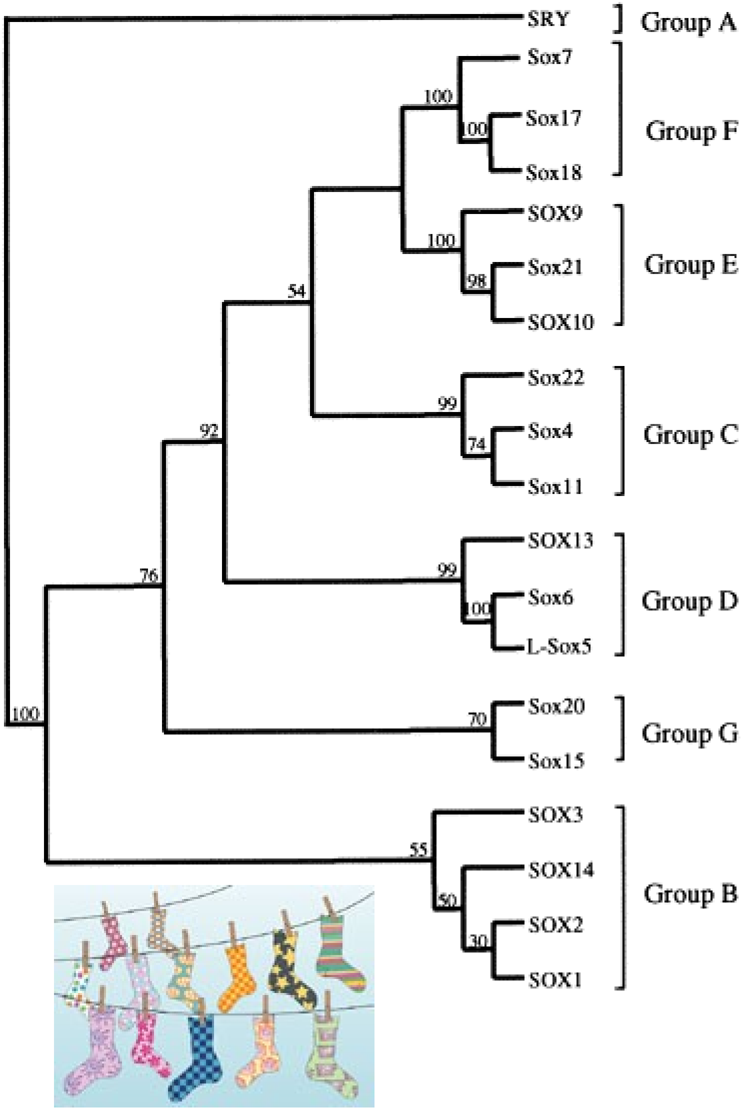
Gene Families, Gene Names

Genes (proteins) come in families.

Genes of the same family have similar sequences.

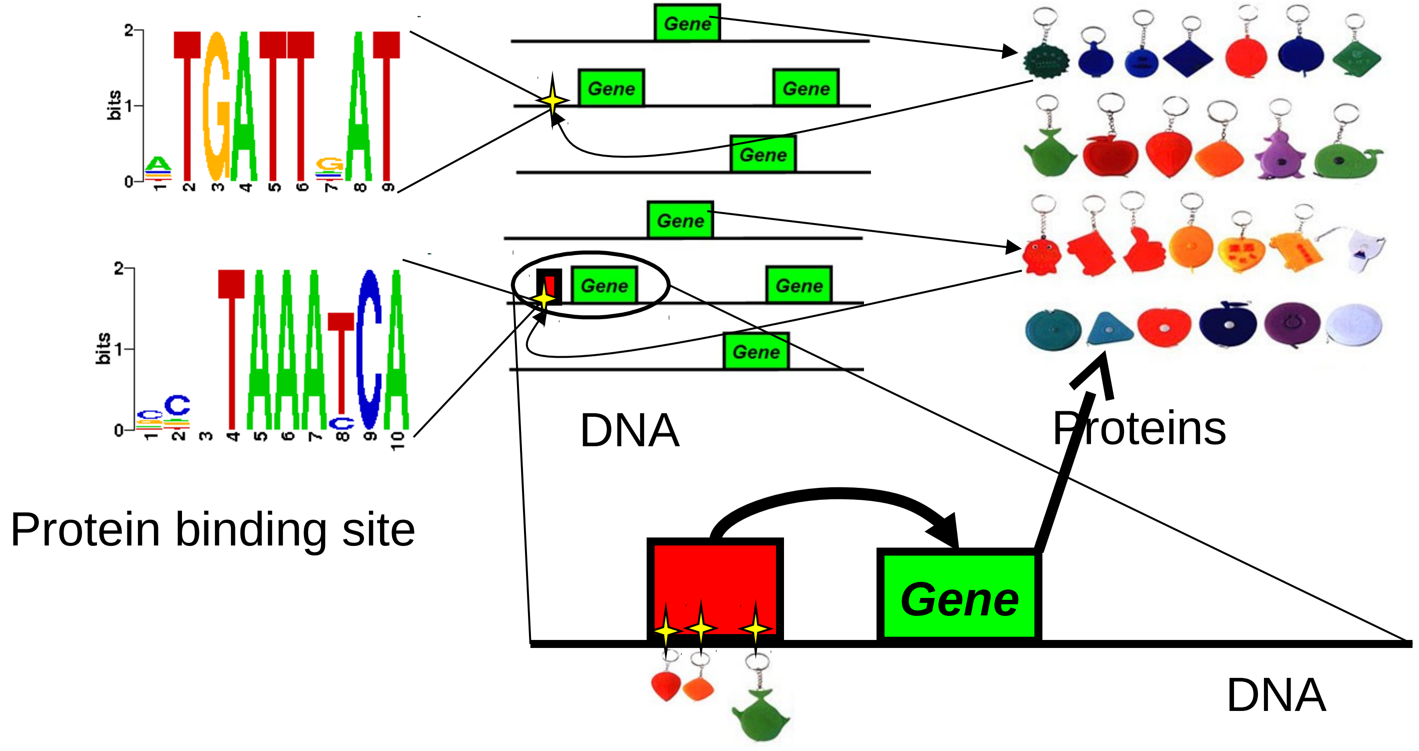
Which is why the fold into similar structure and perform similar functions.

Genes of the same family will typically have a “family name” followed by a (sequential) number or “first name”.



Some “Special” Functions: Gene Regulation

2,000 different proteins can bind specific DNA sequences.



Protein binding site

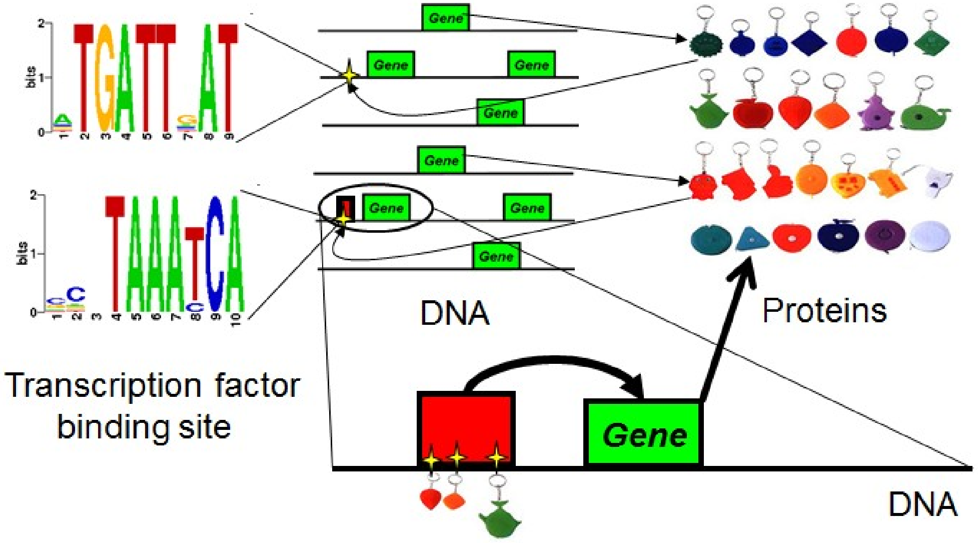
DNA

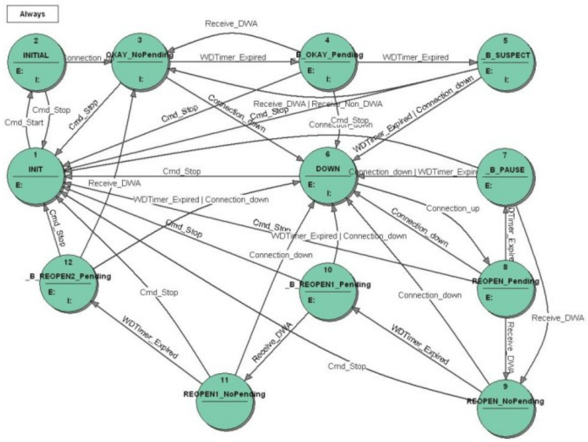
Proteins

DNA

Proteins that regulate the transcription of other proteins
are called transcription factors.

The Importance of Gene Regulation



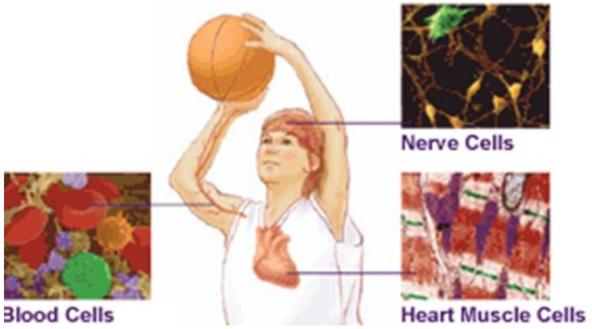


The looks & capabilities of different cells are determined by the subset of genes they express.

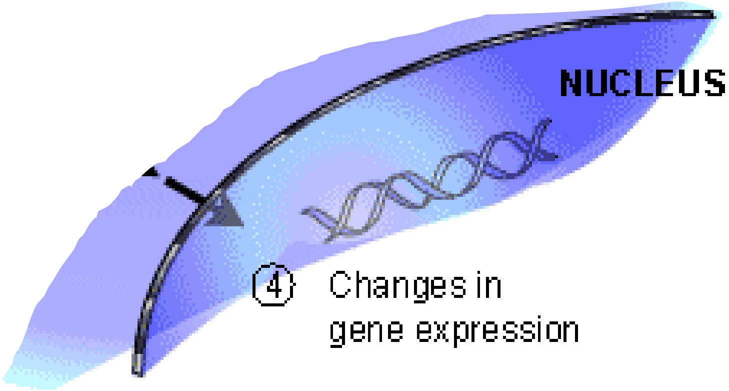
Different cell types express very different gene repertoires (from the same genome).

To change its behavior a cell can change its transcriptional program.

Think of it as a giant state machine…

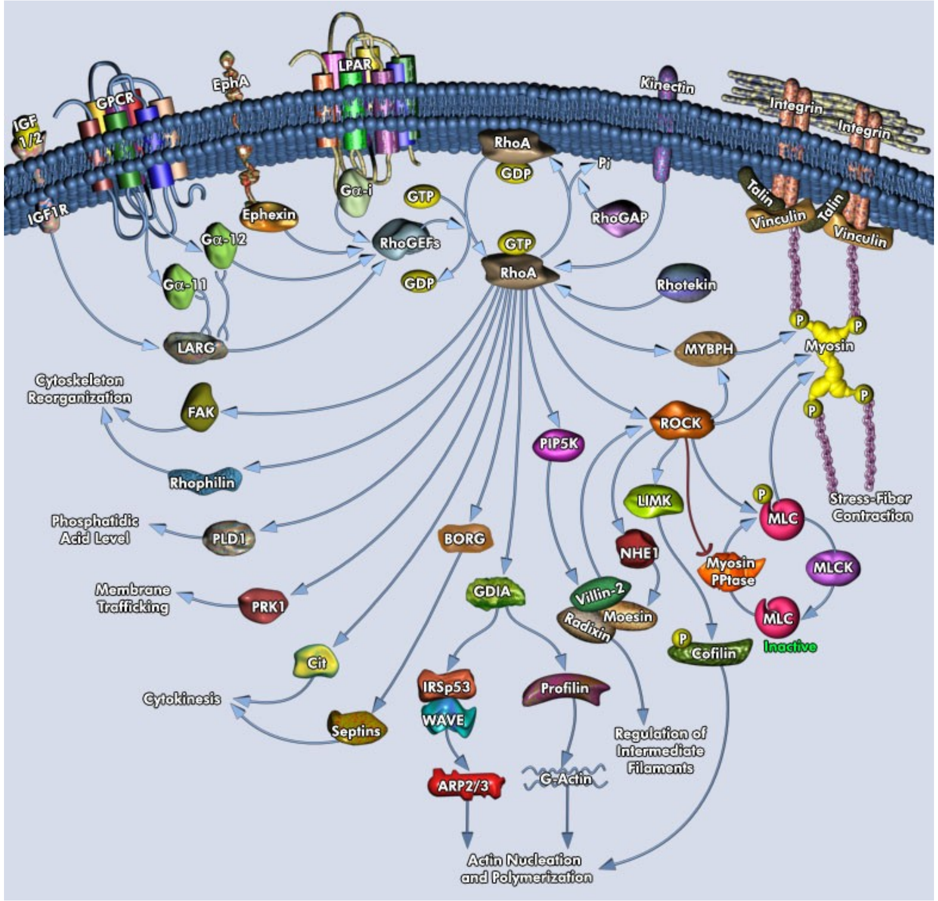


“Special” Function: Cell Signaling



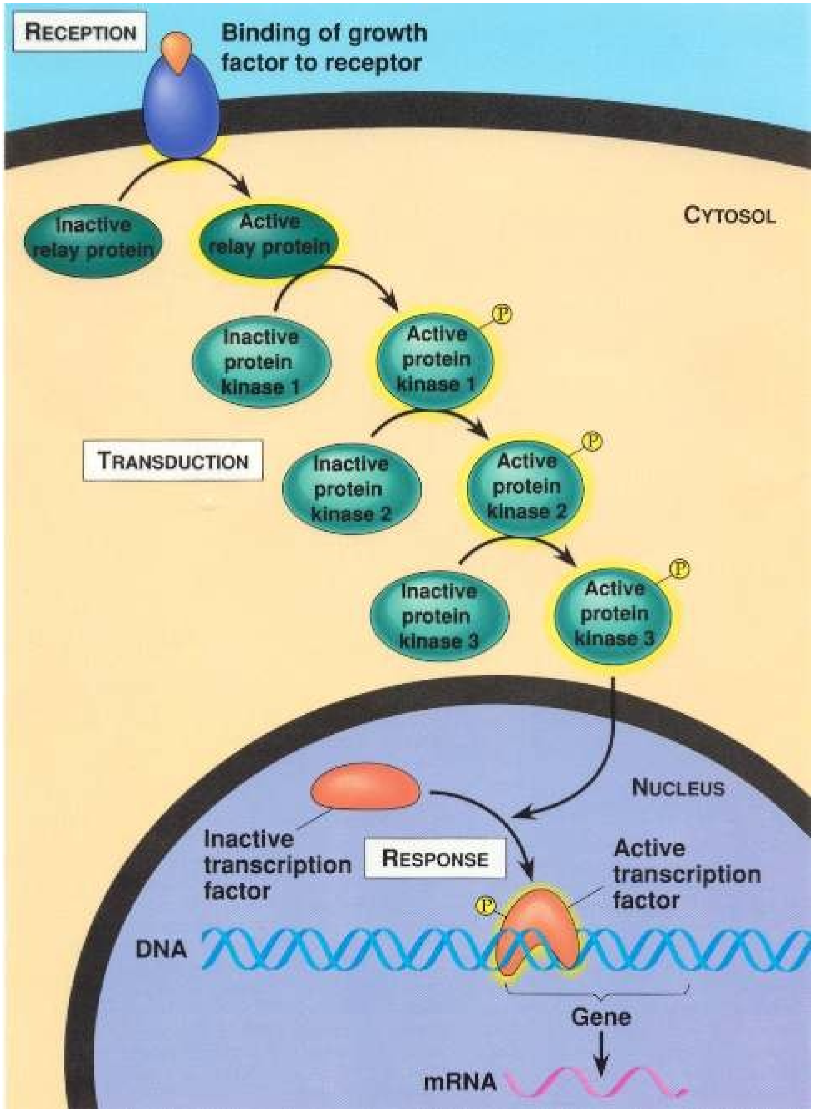
Cells also talk with each other. They send and receive messages, and change their behavior according to messages they receive.

Biological vs. Molecular Function: Pathways



Proteins with very different molecular functions participate to manifest a single biological function, for example: a pathway.

Signal Transduction



Now its an even bigger state machine of individual state machines (=cells) talking with each other, orchestrating their individual activities.