Gene Structure & Gene Finding: Part II

David Wishart

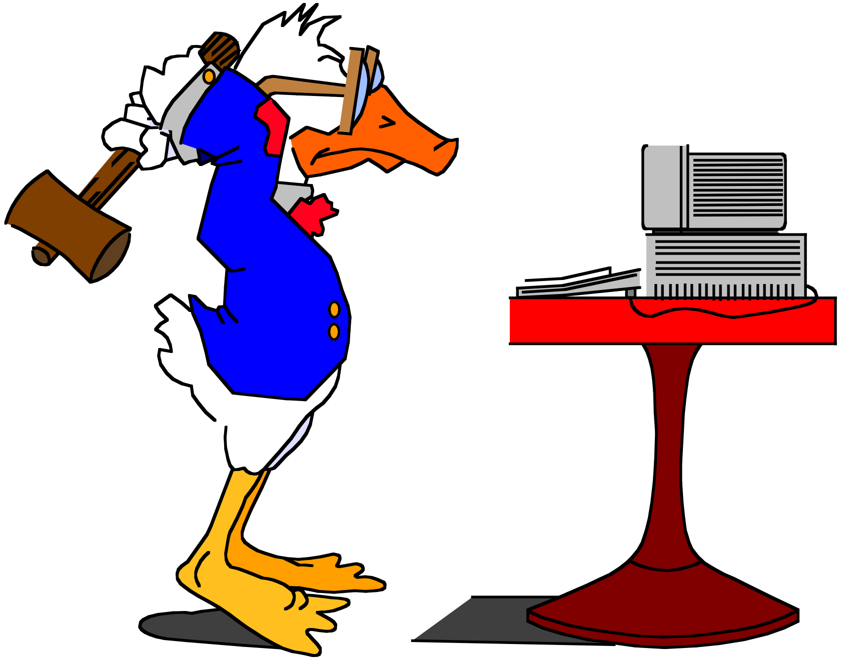
david.wishart@ualberta.ca

Contacting Me…

- 200 emails a day – not the best way to get an instant response
- *Subject line: Bioinf 301 or Bioinf 501*
- Preferred method…
  - Talk to me after class
  - Talk to me before class
  - Ask questions in class
  - Visit my office after 4 pm (Mon. – Fri.)
  - Contact my bioinformatics assistant – Dr. An Chi Guo (anchiguo@gmail.com)

Lecture Notes Available At:

- **http://www.wishartlab.com/**
- ***Go to the menu at the top of the page, look under Courses***



Outline for Next 3 Weeks

- Genes and Gene Finding (Prokaryotes)
- Genes and Gene Finding (Eukaryotes)
- Genome and Proteome Annotation
- Fundamentals of Transcript Measurement
- Introduction to Microarrays
- More details on Microarrays

Assignment Schedule

- Gene finding - genome annotation
  - (Assigned Oct. 31, due Nov. 7)
- Microarray analysis
  - (Assigned Nov. 7, due Nov. 19)
- Protein structure analysis
  - (Assigned Nov. 21, due Nov. 28)

Each assignment is worth 5% of total grade, 10% off for each day late

Objectives*

- Learn key features of eukaryotic gene structure and transcript processing
- Learn/memorize a few key eukaryotic gene signature sequences
- Learn about RNA→ cDNA preparation
- Review algorithms and web tools for eukaryotic gene identification
- Measuring/assessing gene prediction (limitations, methods)

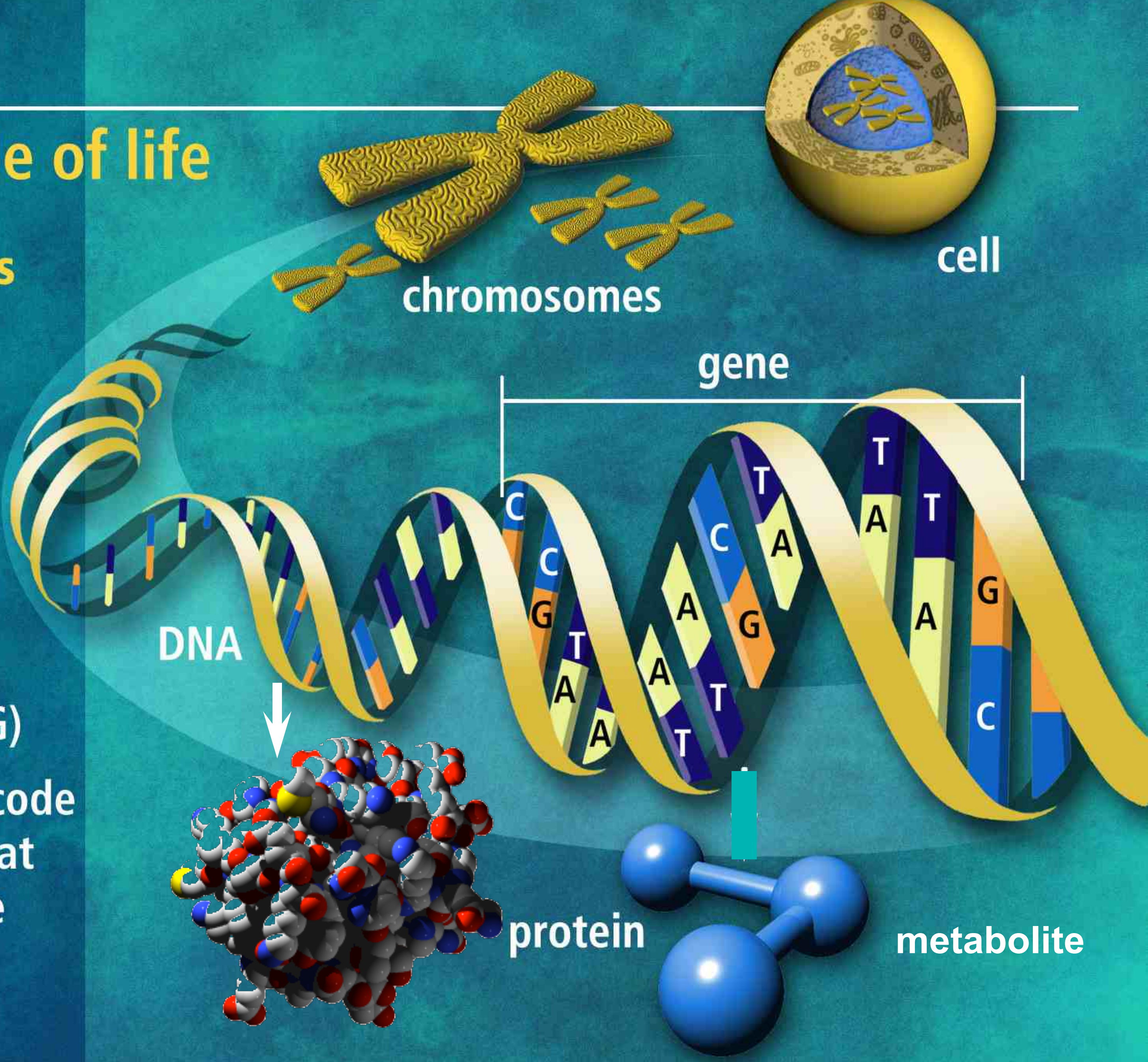
DNA

the molecule of life

Trillions of cells

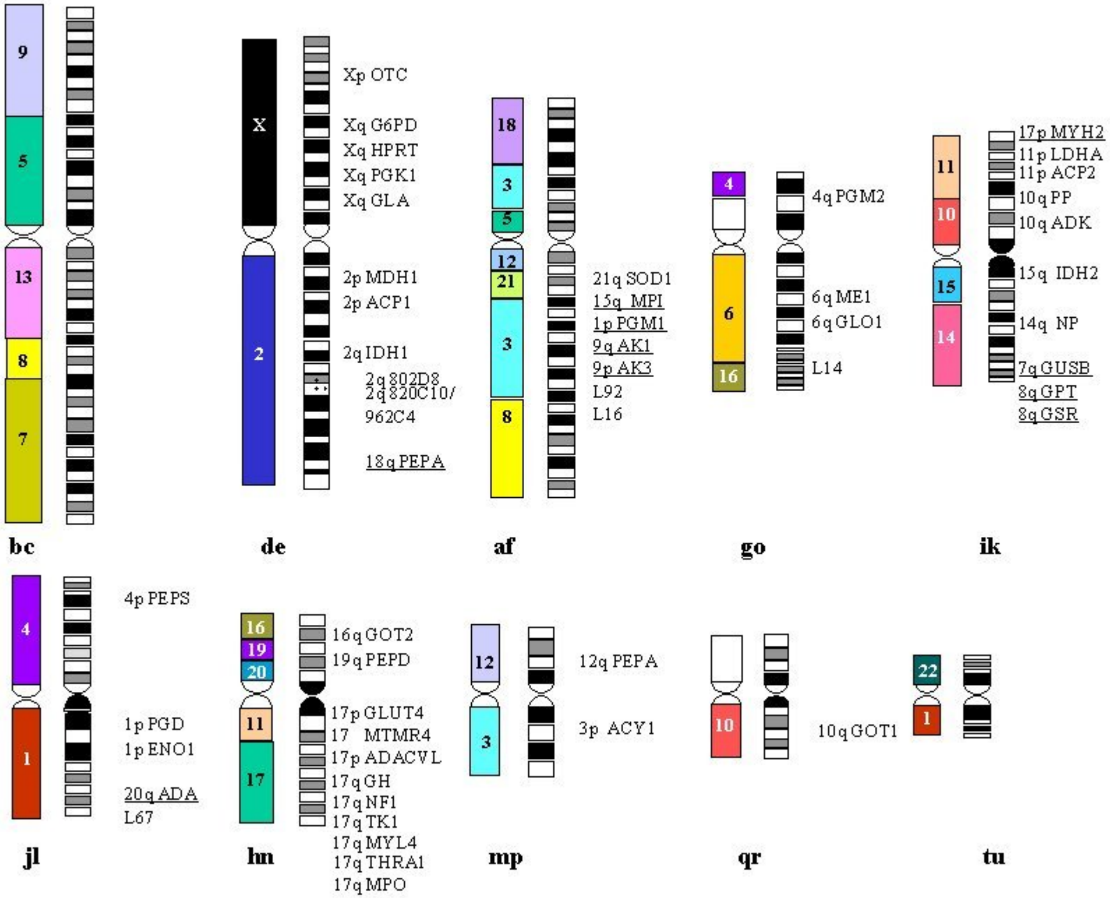
Each cell:

- 46 human chromosomes
- 2 m of DNA
- 3 billion DNA subunits (the bases: A, T, C, G)
- 23,000 genes code for proteins that perform all life functions



Y-GA 98-090R

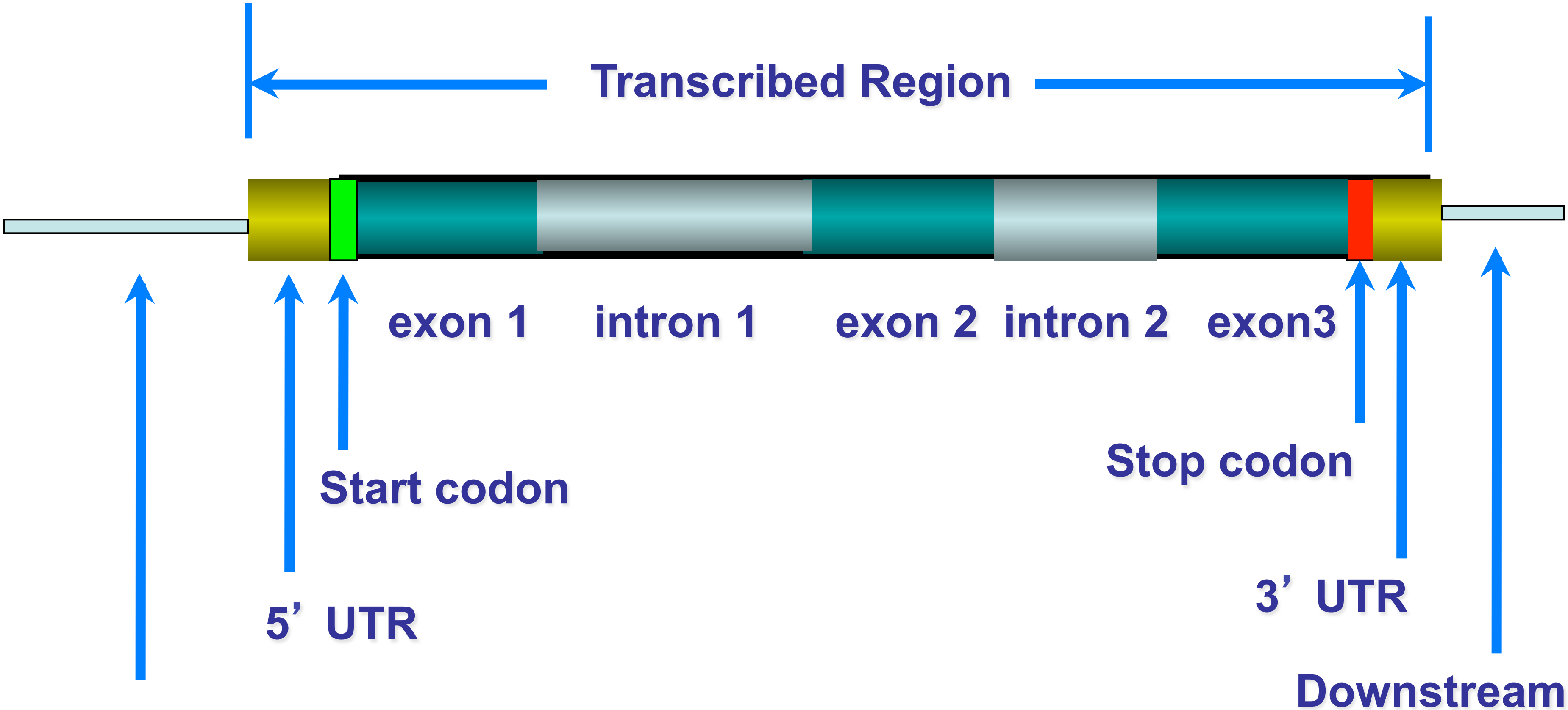
Gene Finding in Eukaryotes



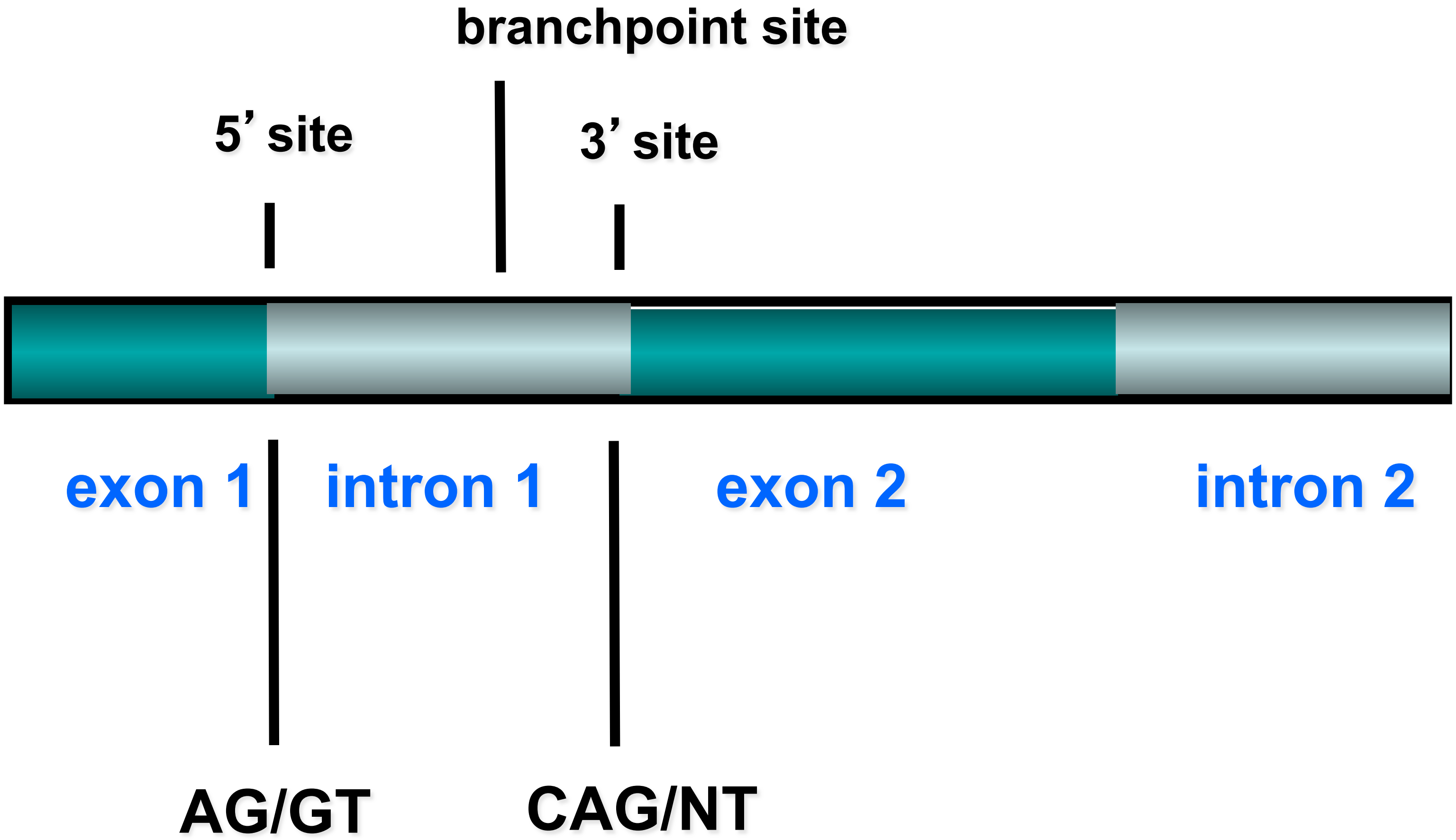
Eukaryotes*

- Complex gene structure
- Large genomes (0.1 to 10 billion bp)
- Exons and Introns (interrupted)
- Low coding density (<30%)
  - 3% in humans, 25% in Fugu, 60% in yeast
- Alternate splicing (40-60% of all genes)
- High abundance of repeat sequence (50% in humans) and pseudo genes
- Nested genes: overlapping on same or opposite strand or inside an intron

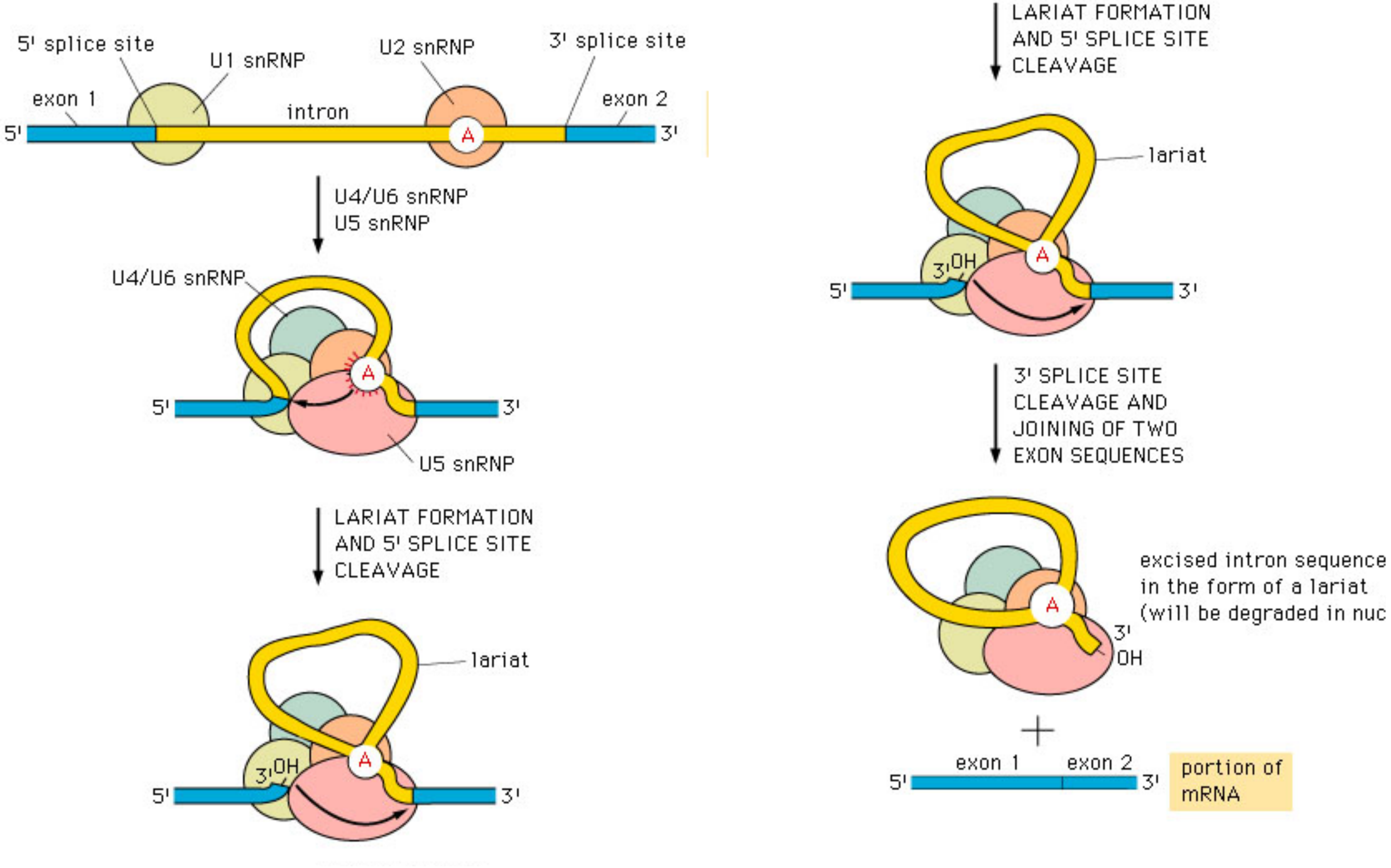
Eukaryotic Gene Structure*



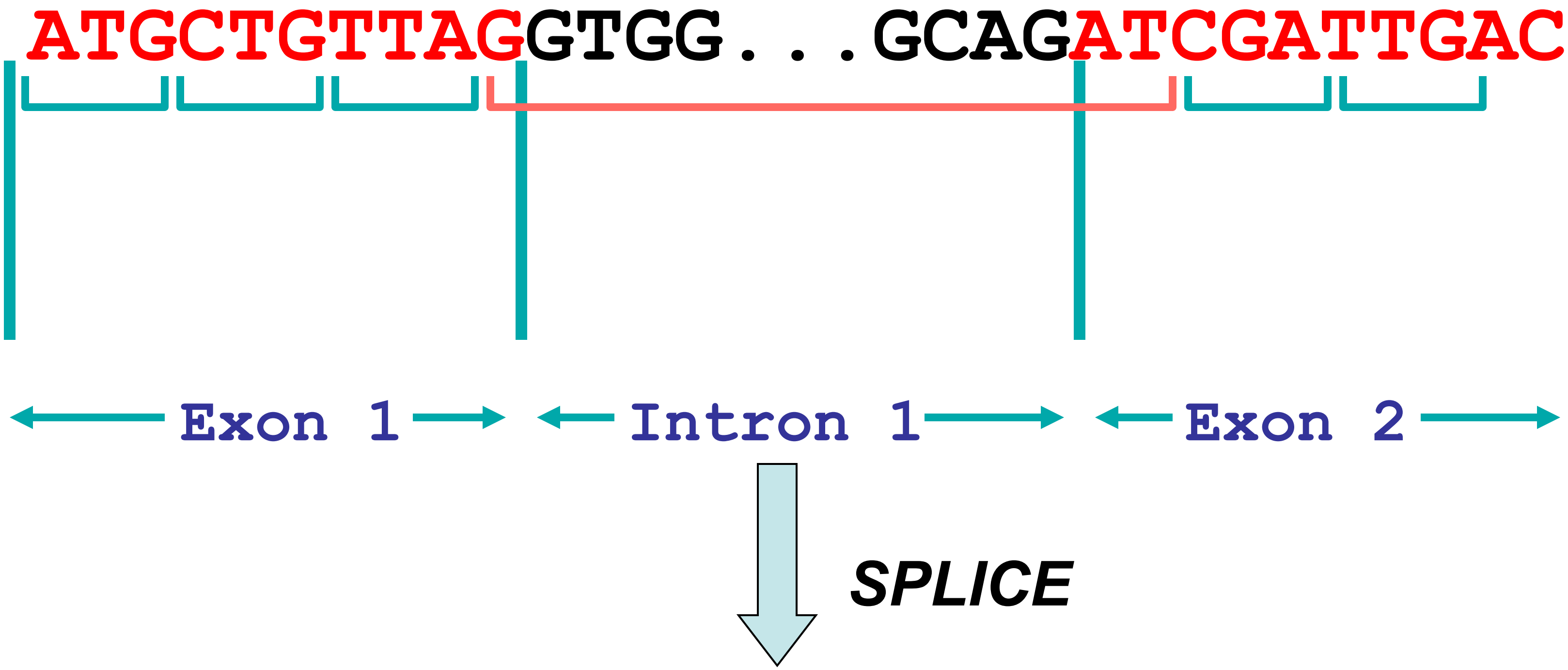
Eukaryotic Gene Structure*



RNA Splicing*



Exon/Intron Structure (Detail)



ATGCTGTTAGGTGG...GCAGATCGATTGAC

Exon 1

Intron 1

Exon 2

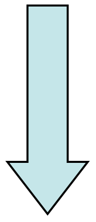
SPLICE

ATGCTGTTAGATCGATTGAC

Intron Phase*

- A codon can be interrupted by an intron in one of three places

Phase 0:	ATGATTGTCAG...CAGTAC
Phase 1:	ATGATGTCAG...CAGTTAC
Phase 2:	ATGAGTCAG...CAGTTTAC



SPLICE

AGTATTTAC

Repetitive DNA*

- Moderately Repetitive DNA
  - Tandem gene families (250 copies of rRNA, 500-1000 tRNA gene copies)
  - Pseudogenes (dead genes)
  - Short interspersed elements (SINEs)
    - 200-300 bp long, 100,000+ copies, scattered
    - Alu repeats are good examples
  - Long interspersed elements (LINEs)
    - 1000-5000 bp long
    - 10 - 10,000 copies per genome

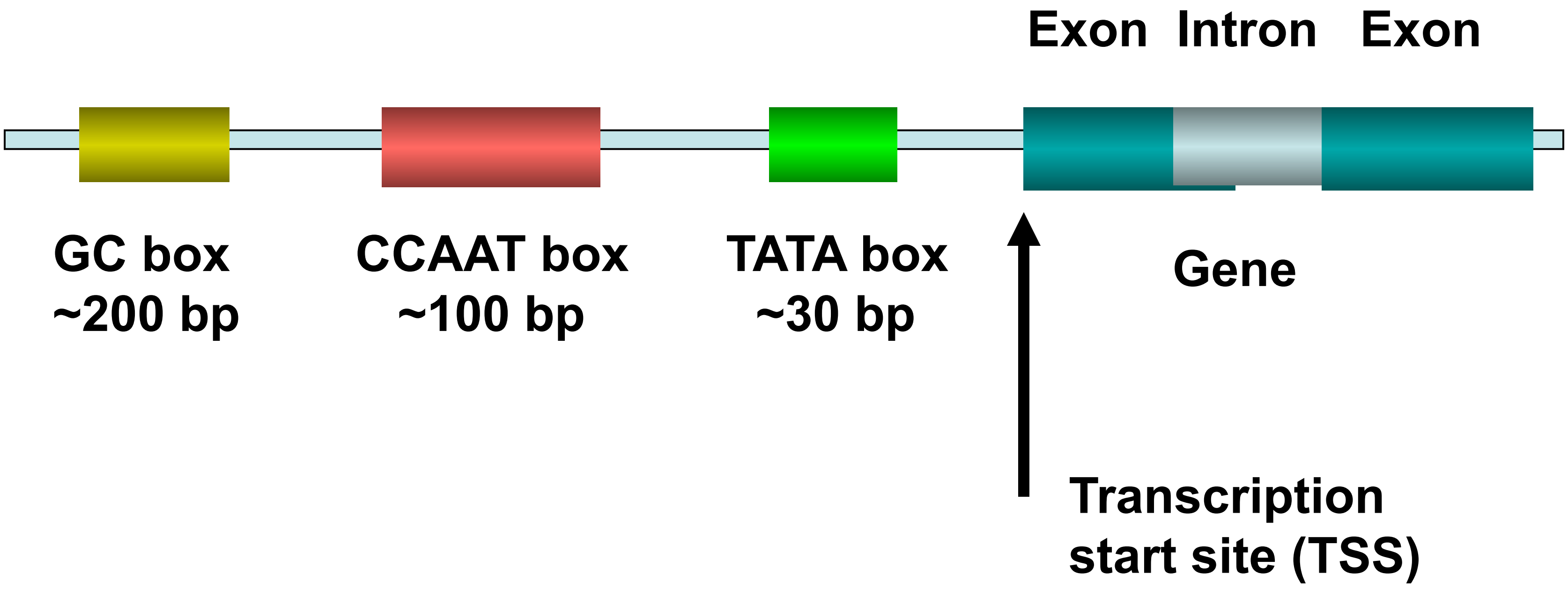
Repetitive DNA*

- Highly Repetitive DNA
  - Minisatellite DNA
    - repeats of 14-500 bp stretching for ~2 kb
    - many different types scattered thru genome
  - Microsatellite DNA
    - repeats of 5-13 bp stretching for 100’s of kb
    - mostly found around centromere
  - Telomeres
    - highly conserved 6 bp repeat (TTAGGG)
    - 250-1000 repeats at end of each chromosome

Key Eukaryotic Gene Signals*

- Pol II RNA promoter elements
  - Cap and CCAAT region
  - GC and TATA region
- Kozak sequence (Ribosome binding site-RBS)
- Splice donor, acceptor and lariat signals
- Termination signal
- Polyadenylation signal

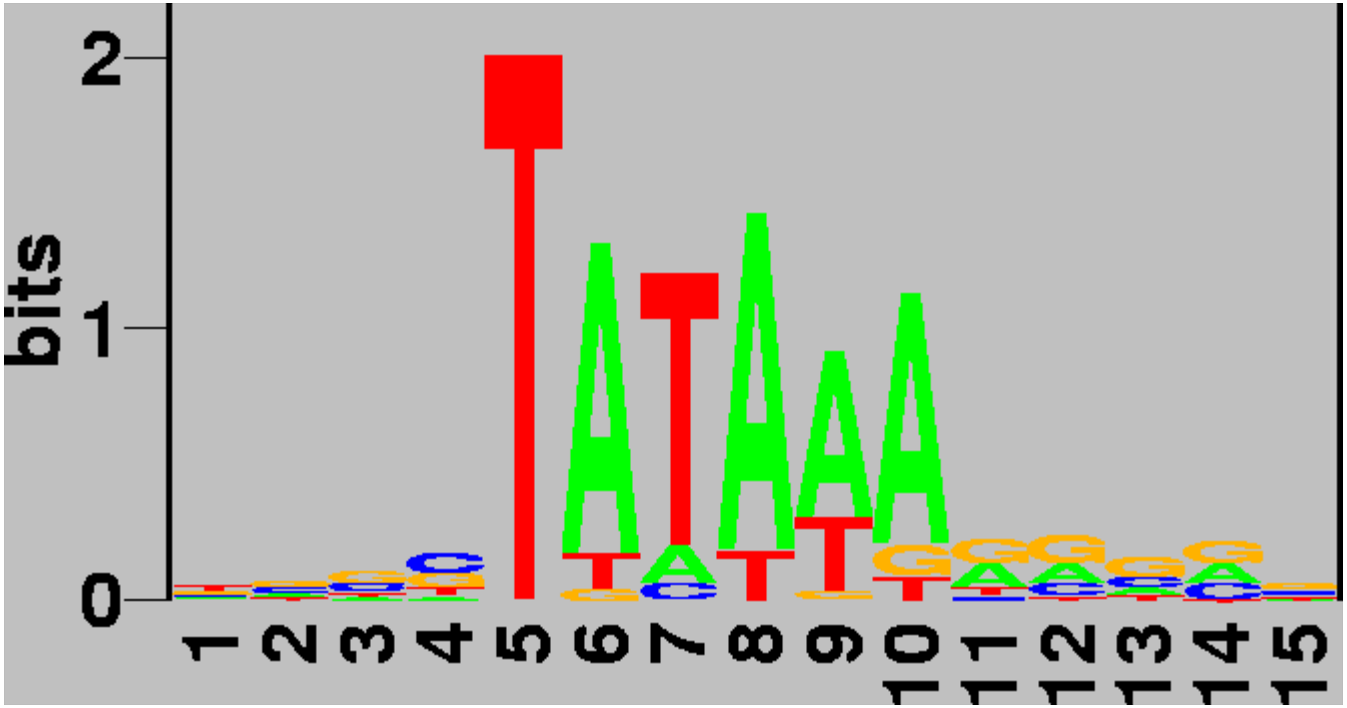
Pol II Promoter Elements*



Pol II Promoter Elements*

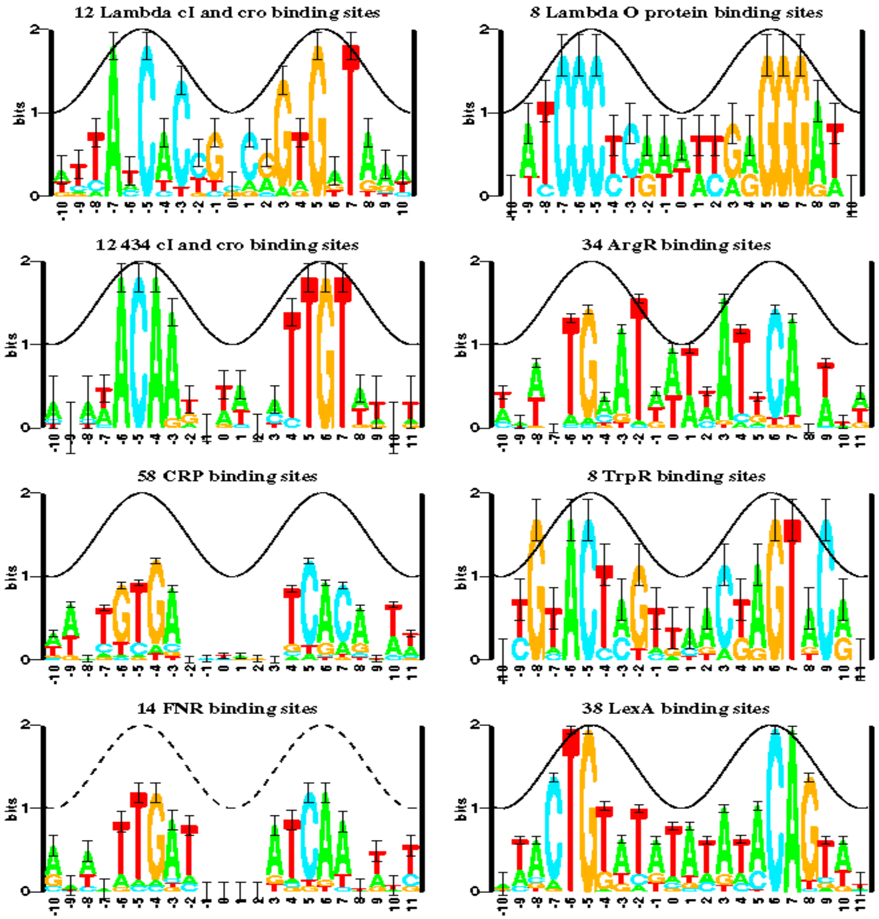
- Cap Region/Signal
  - n C A G T n G
- TATA box (~ 25 bp upstream)
  - T A T A A A n G C C C
- CCAAT box (~100 bp upstream)
  - T A G C C A A T G
- GC box (~200 bp upstream)
  - A T A G G C G n G A

Pol II Promoter Elements



TATA box is found in ~70% of promoters

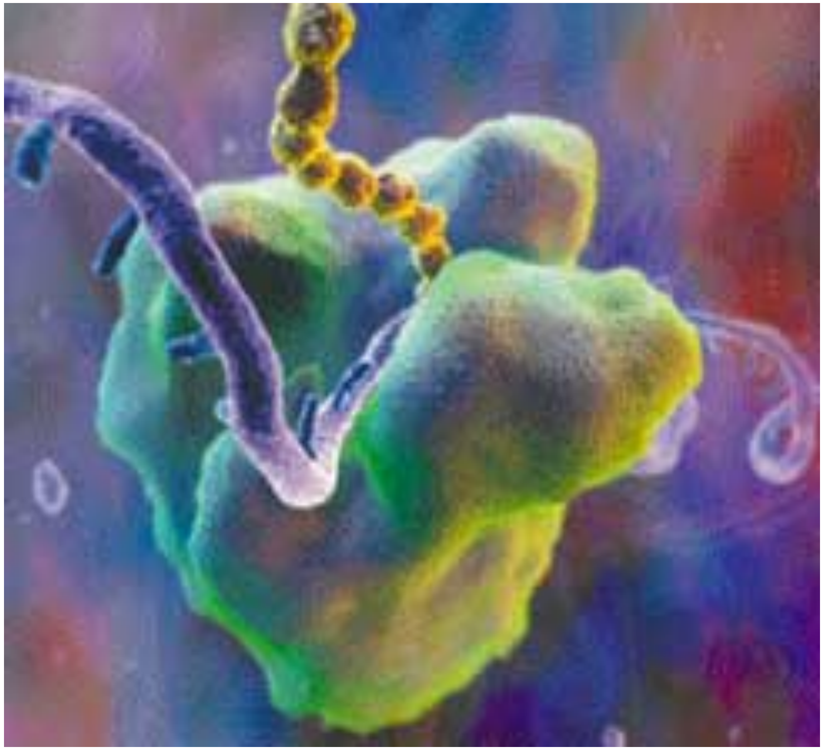
WebLogos



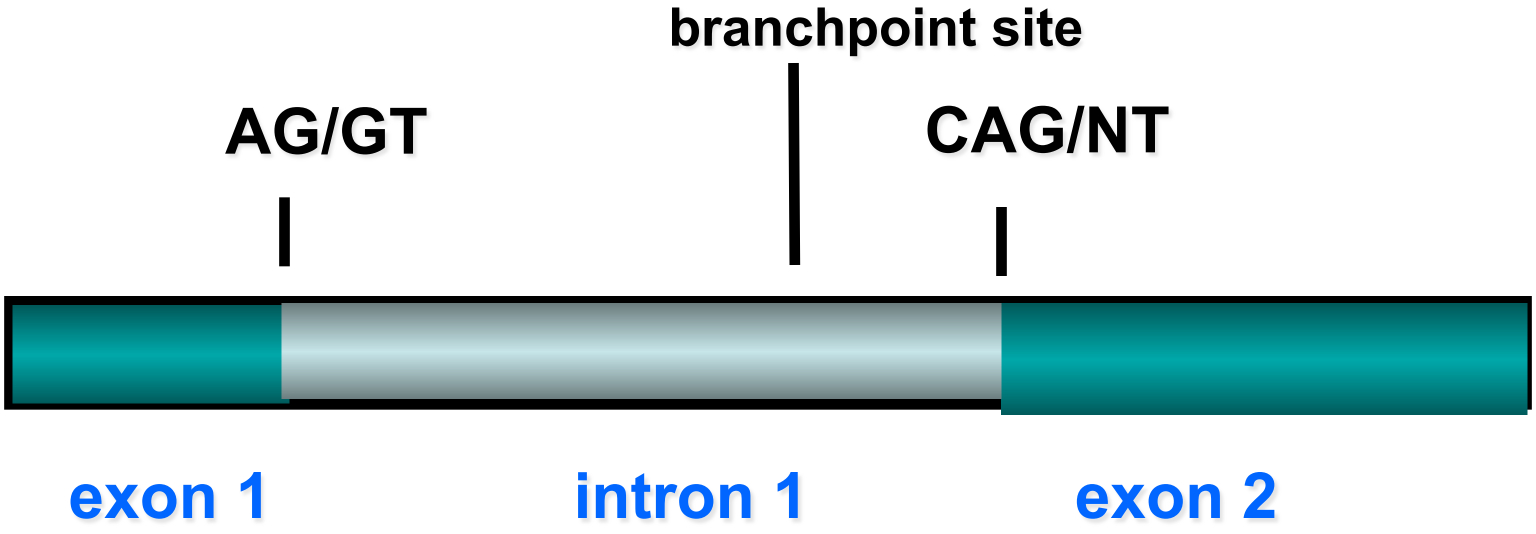
http://weblogo.berkeley.edu/

Kozak (RBS) Sequence*

-7	-6	-5	-4	-3	-2	-1	0	1	2	3
A	G	C	C	A	C	C	A	T	G	G



Splice Signals*



branchpoint site

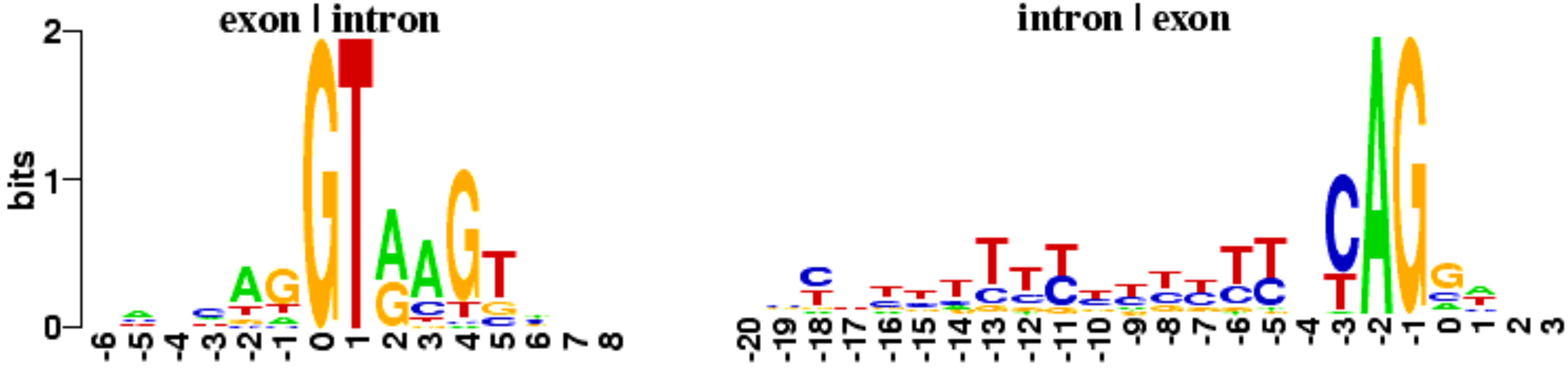
AG/GT

CAG/NT

exon 1

intron 1

exon 2



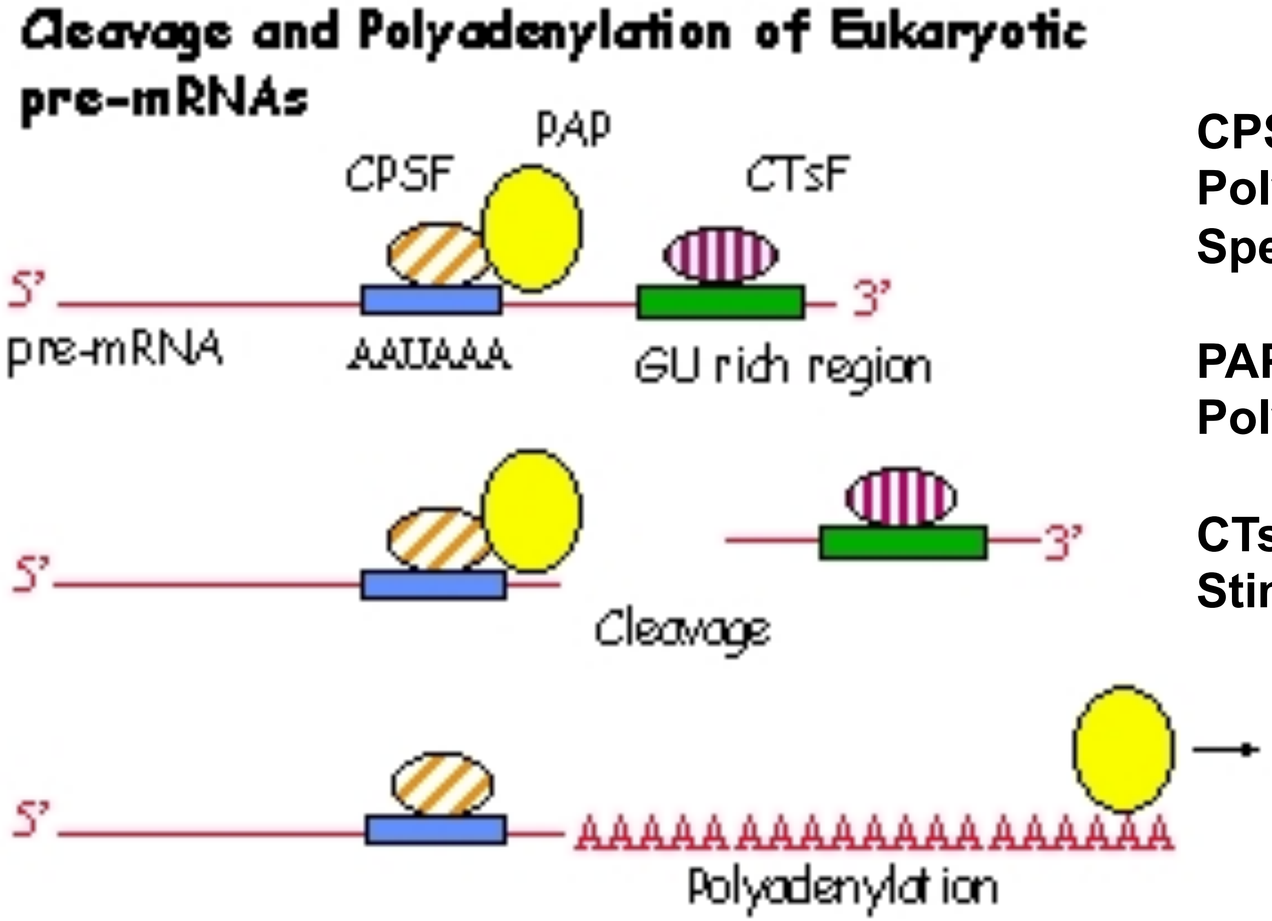
Splice Sites*

- Not all splice sites are real
- ~0.5% of splice sites are non-canonical (i.e. the intron is not GT...AG)
- It is estimated that 5% of human genes may have non-canonical splice sites
- ~50% of higher eukaryotes are alternately spliced (different exons are brought together)

Miscellaneous Signals*

- Polyadenylation signal
  - AATAAA or ATTAAA
  - Located 20 bp upstream of poly-A cleavage site
- Termination Signal
  - AGTGTTCA
  - Located ~30 bp downstream of poly-A cleavage site

Polyadenylation*



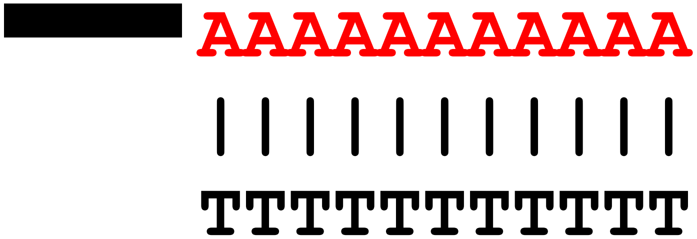
CPSF – Cleavage & Polyadenylation Specificity Factor

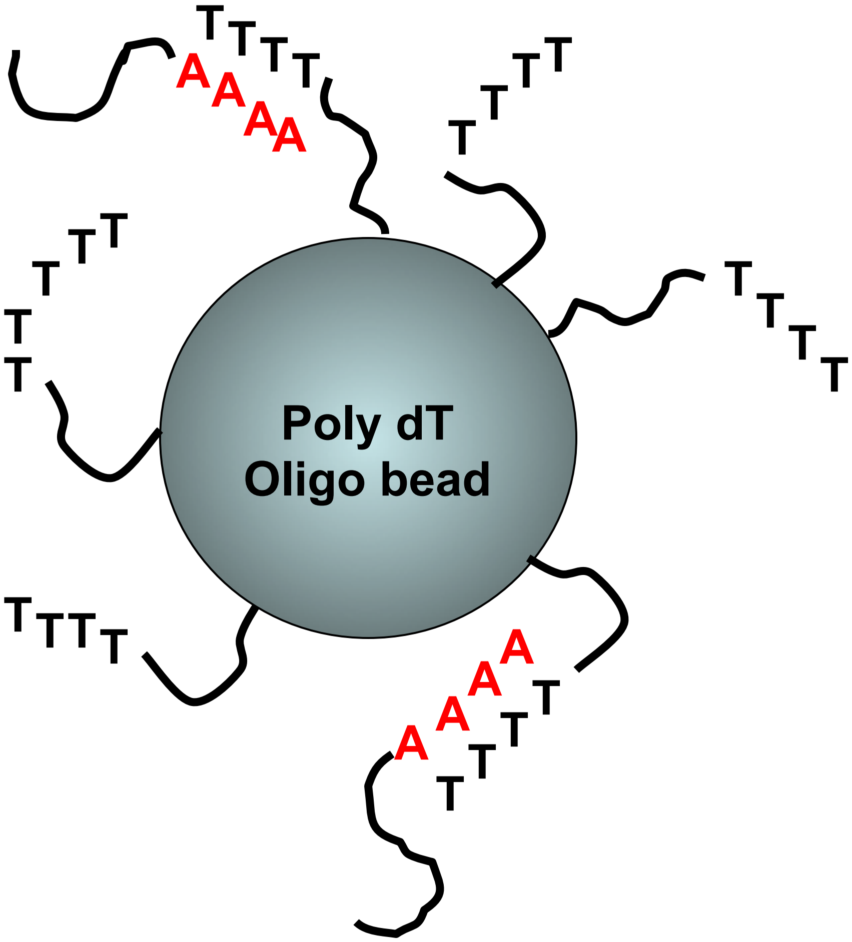
PAP – Poly-A Polymerase

CTsF – Cleavage Stimulation Factor

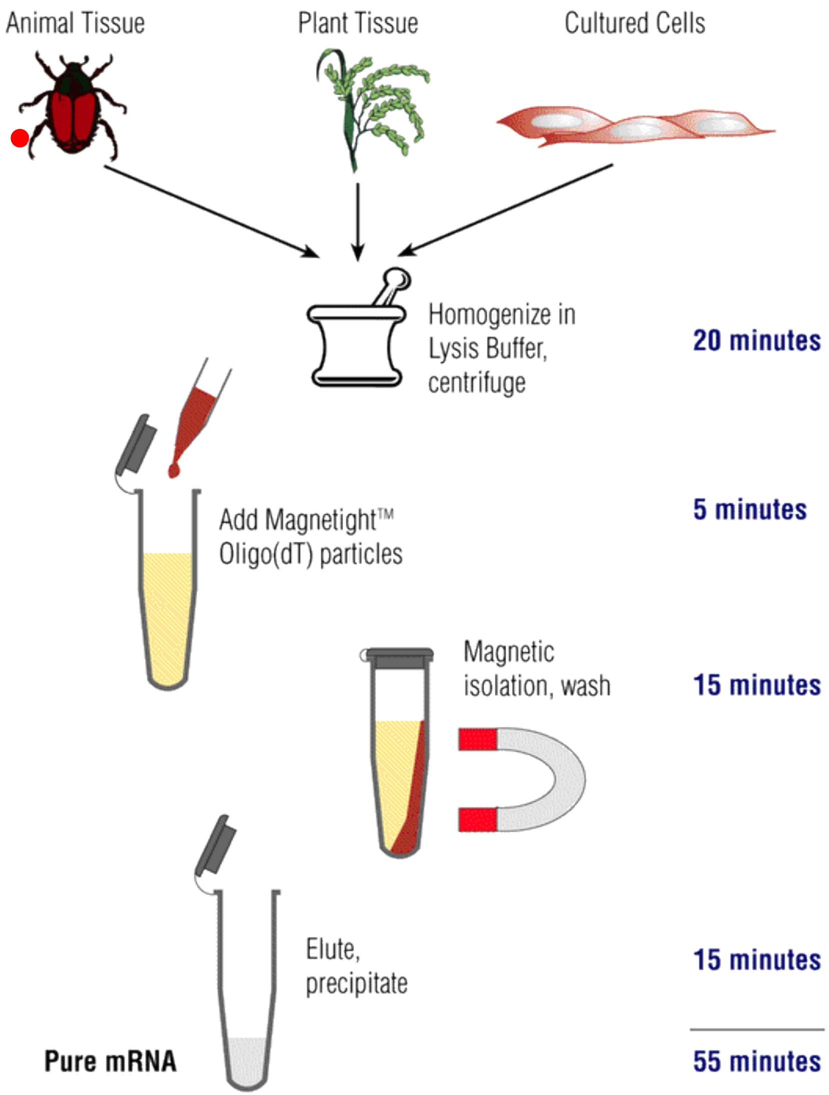
Why Polyadenylation is Really Useful

Complementary Base Pairing



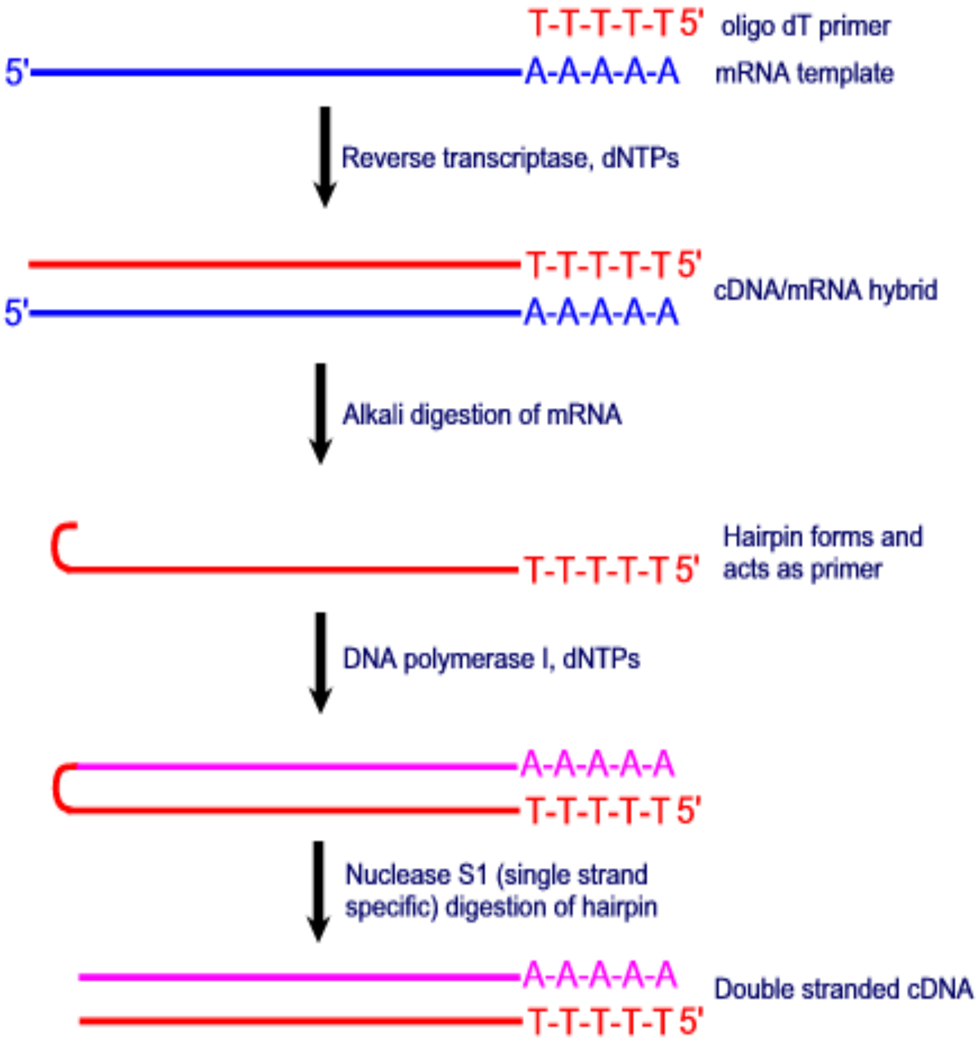


mRNA isolation*



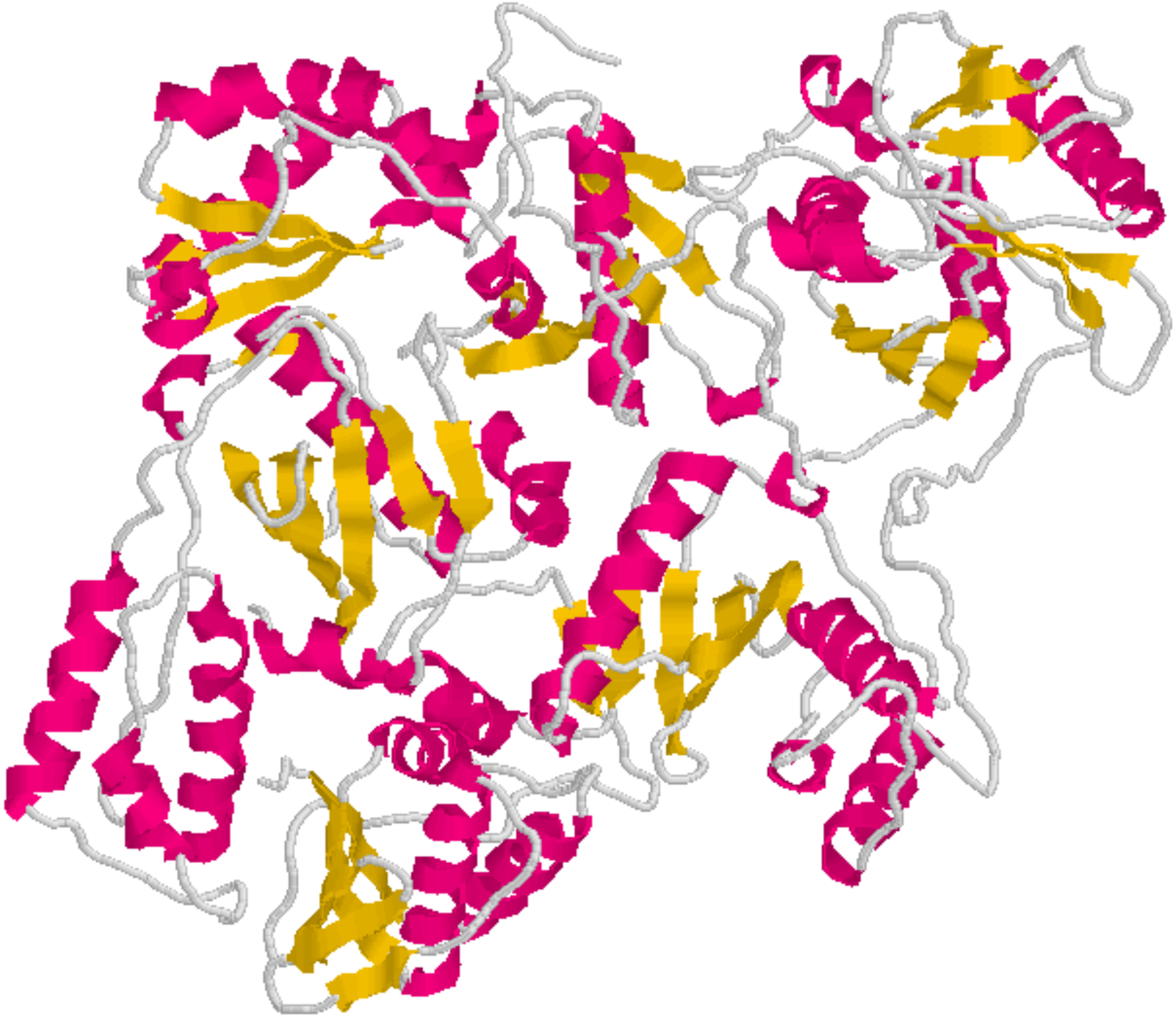
- Cell or tissue sample is ground up and lysed with chemicals to release mRNA
- Oligo(dT) beads are added and incubated with mixture to allow A-T annealing
- Pull down beads with magnet and pull off mRNA

Making cDNA from mRNA*



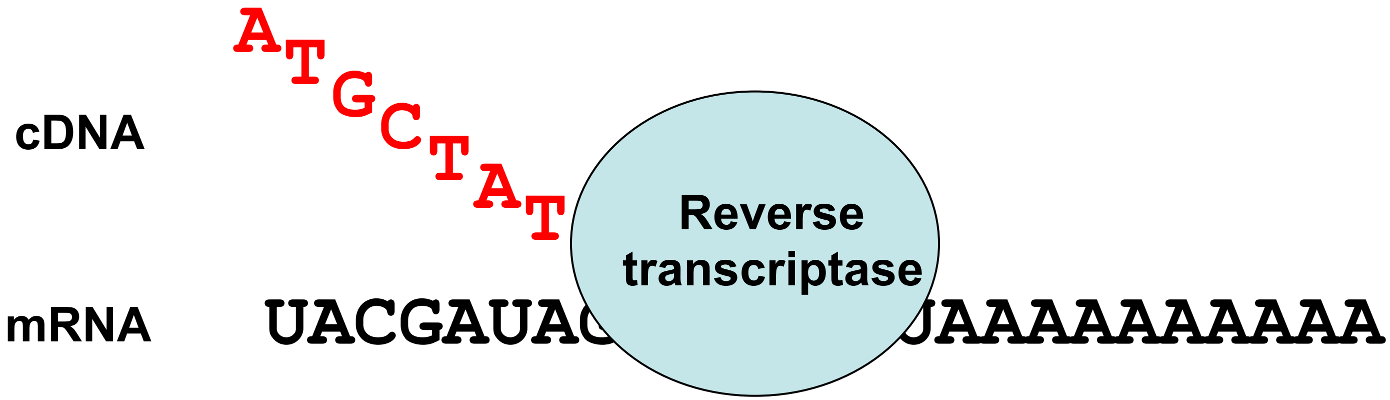
- cDNA (i.e. complementary DNA) is a single-stranded DNA segment whose sequence is complementary to that of messenger RNA (mRNA)
- Synthesized by reverse transcriptase

Reverse Transcriptase



Finding Eukaryotic Genes Experimentally

- Convert the spliced mRNA into cDNA



- Only expressed genes or expressed sequence tags (EST’s) are seen
- Saves on sequencing effort (97%)

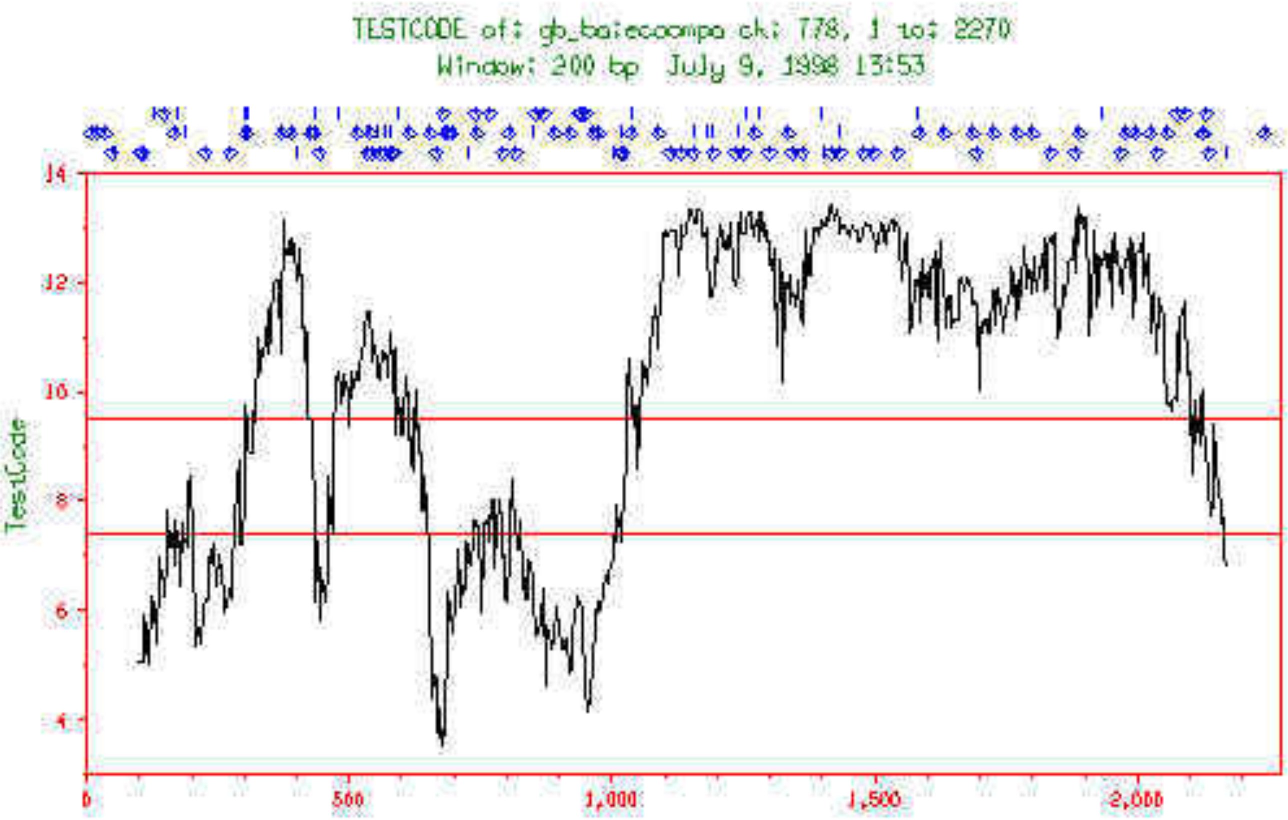
Finding Eukaryotic Genes Computationally*

- Content-based Methods
  - GC content, hexamer repeats, composition statistics, codon frequencies
- Site-based Methods
  - donor sites, acceptor sites, promoter sites, start/stop codons, polyA signals, lengths
- Comparative Methods
  - sequence homology, EST searches
- Combined Methods

Content-Based Methods*

- CpG islands
  - High GC content in 5’ ends of genes
- Codon Bias
  - Some codons are strongly preferred in coding regions, others are not
- Positional Bias
  - 3rd base tends to be G/C rich in coding regions
- Ficketts Method
  - looks for unequal base composition in different clusters of i, i+3, i+6 bases - TestCode graph

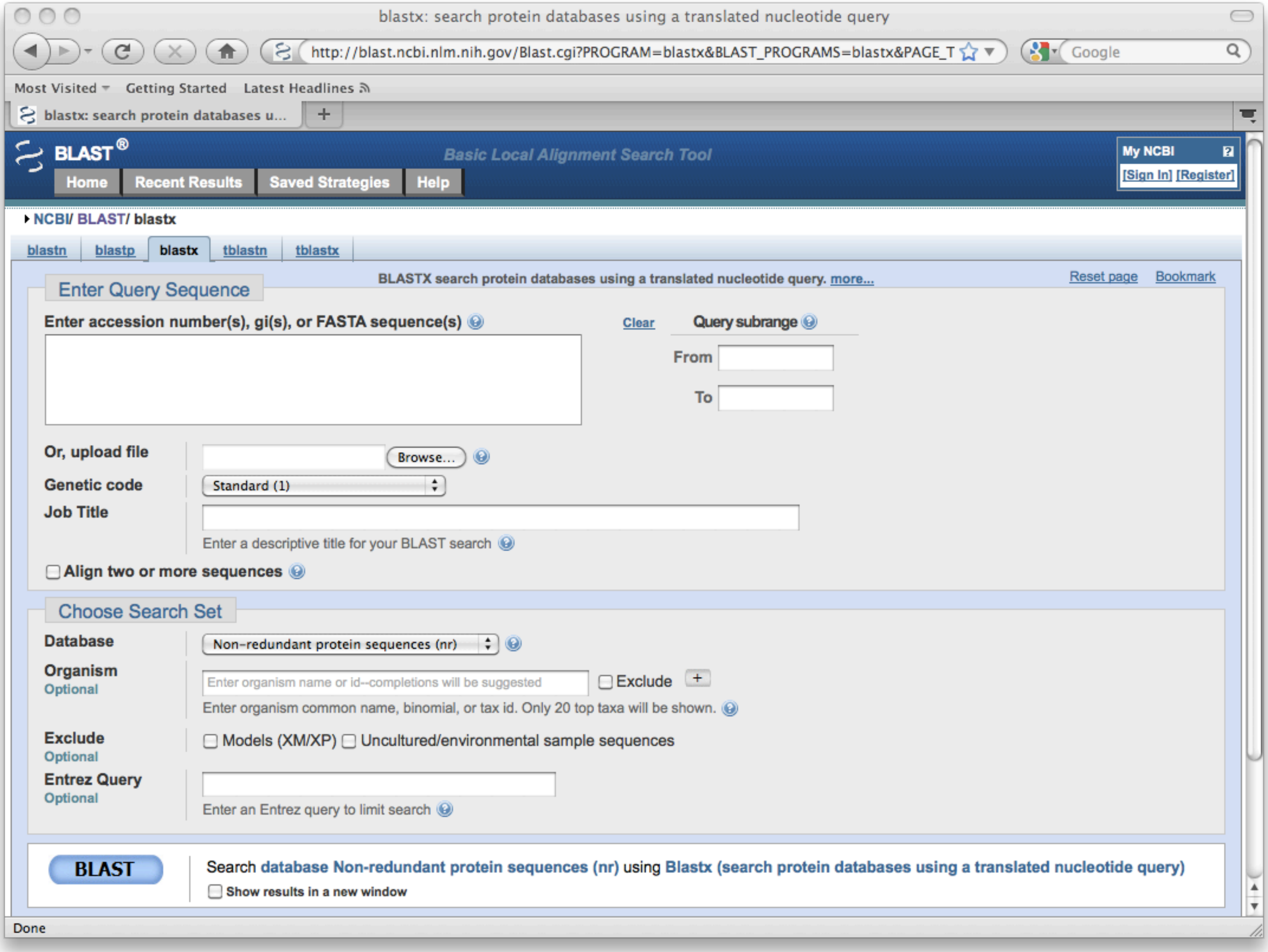
TestCode Plot



Comparative Methods*

- Do a BLASTX search of all 6 reading frames against known proteins in GenBank
- Assumes that the organism under study has genes that are homologous to known genes (used to be a problem, in 2001 analysis of chr. 22 only 50% of genes were similar to known proteins)
- BLAST against EST database (finds possible or probable 3’ end of cDNAs)

BLASTX



blastx: search protein databases using a translated nucleotide query

http://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastx&BLAST_PROGRAMS=blastx&PAGE_T

NCBI/ BLAST/ blastx

BLASTX search protein databases using a translated nucleotide query. more...

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s)

Query subrange

From

To

Or, upload file

Genetic code: Standard (1)

Job Title

Enter a descriptive title for your BLAST search

□ Align two or more sequences

Choose Search Set

Database: Non-redundant protein sequences (nr)

Organism
Optional

Enter organism name or id--completions will be suggested □ Exclude +

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Exclude
Optional

□ Models (XM/XP) □ Uncultured/environmental sample sequences

Entrez Query
Optional

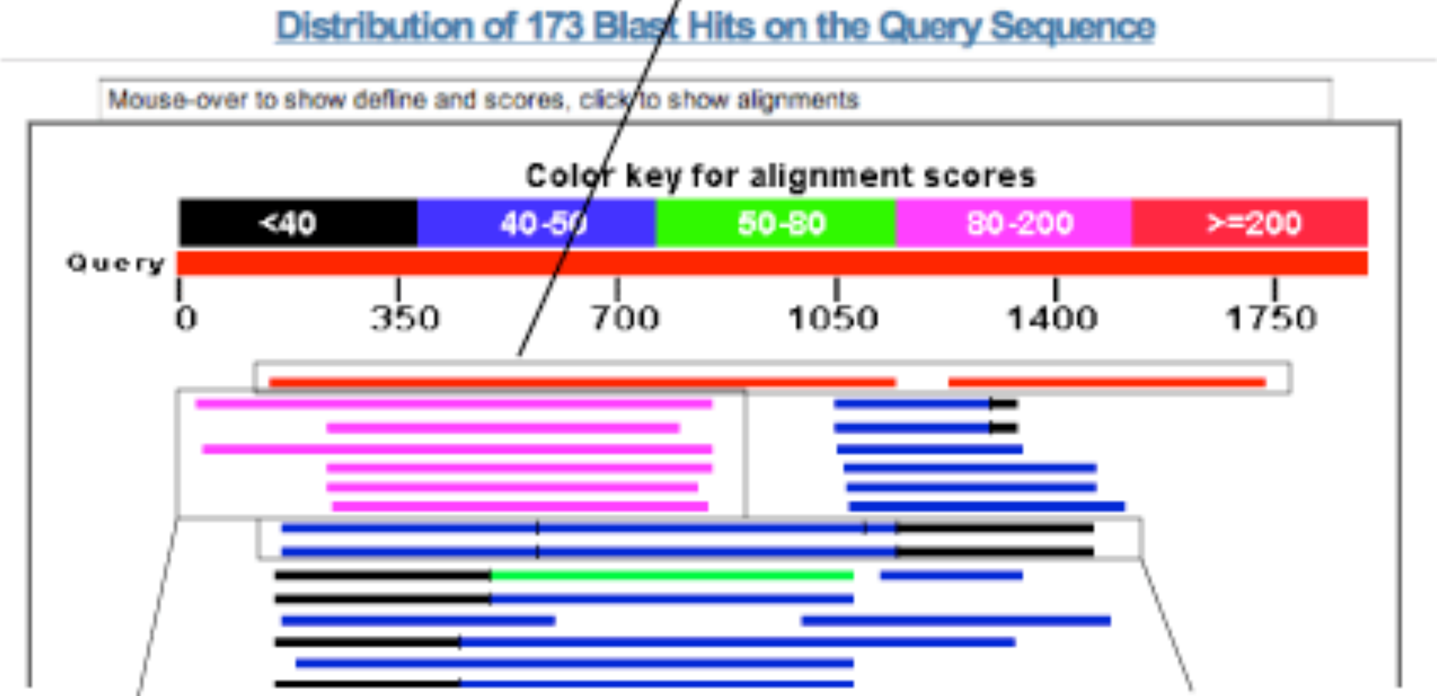
Enter an Entrez query to limit search

Search database Non-redundant protein sequences (nr) using Blastx (search protein databases using a translated nucleotide query)

□ Show results in a new window

BLASTX Output

The protein sequences, predicted by computer translation and deposited in GenBank. These were expected to match.



Distribution of 173 Blast Hits on the Query Sequence

Mouse-over to show defline and scores, click to show alignments

Hypothetical or putative proteins from other fungi

Hypothetical proteins from rice. One is from a transposon.

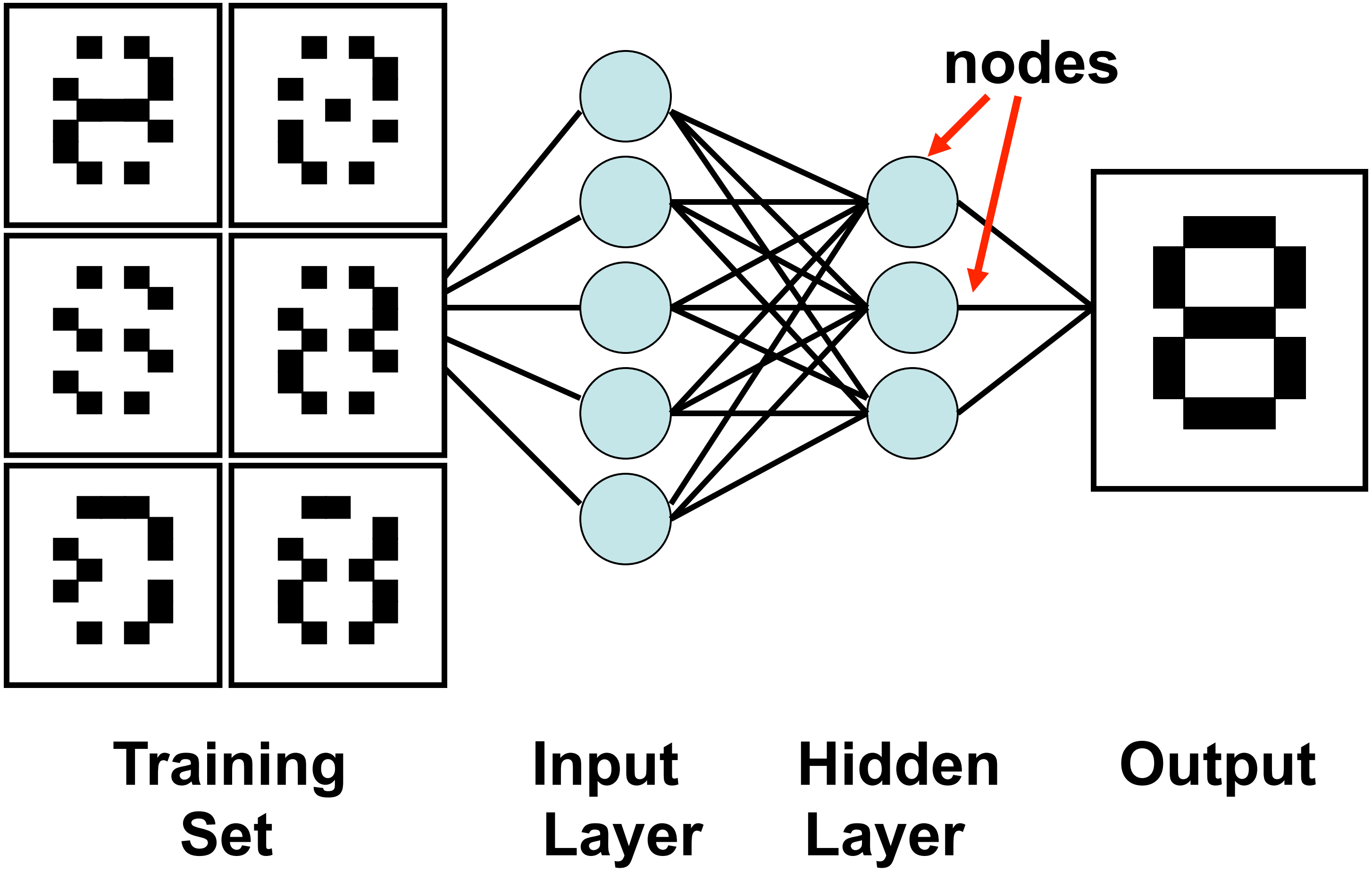
Site-Based Methods*

- Based on identifying gene signals (promoter elements, splice sites, start/ stop codons, polyA sites, etc.)
- Wide range of methods
  - consensus sequences
  - weight matrices
  - neural networks
  - decision trees
  - hidden markov models (HMMs)

Neural Networks

- Automated method for classification or pattern recognition
- First described in detail in 1986
- Mimic the way the brain works
- Use Matrix Algebra in calculations
- Require “training” on validated data
- *Garbage in = Garbage out*

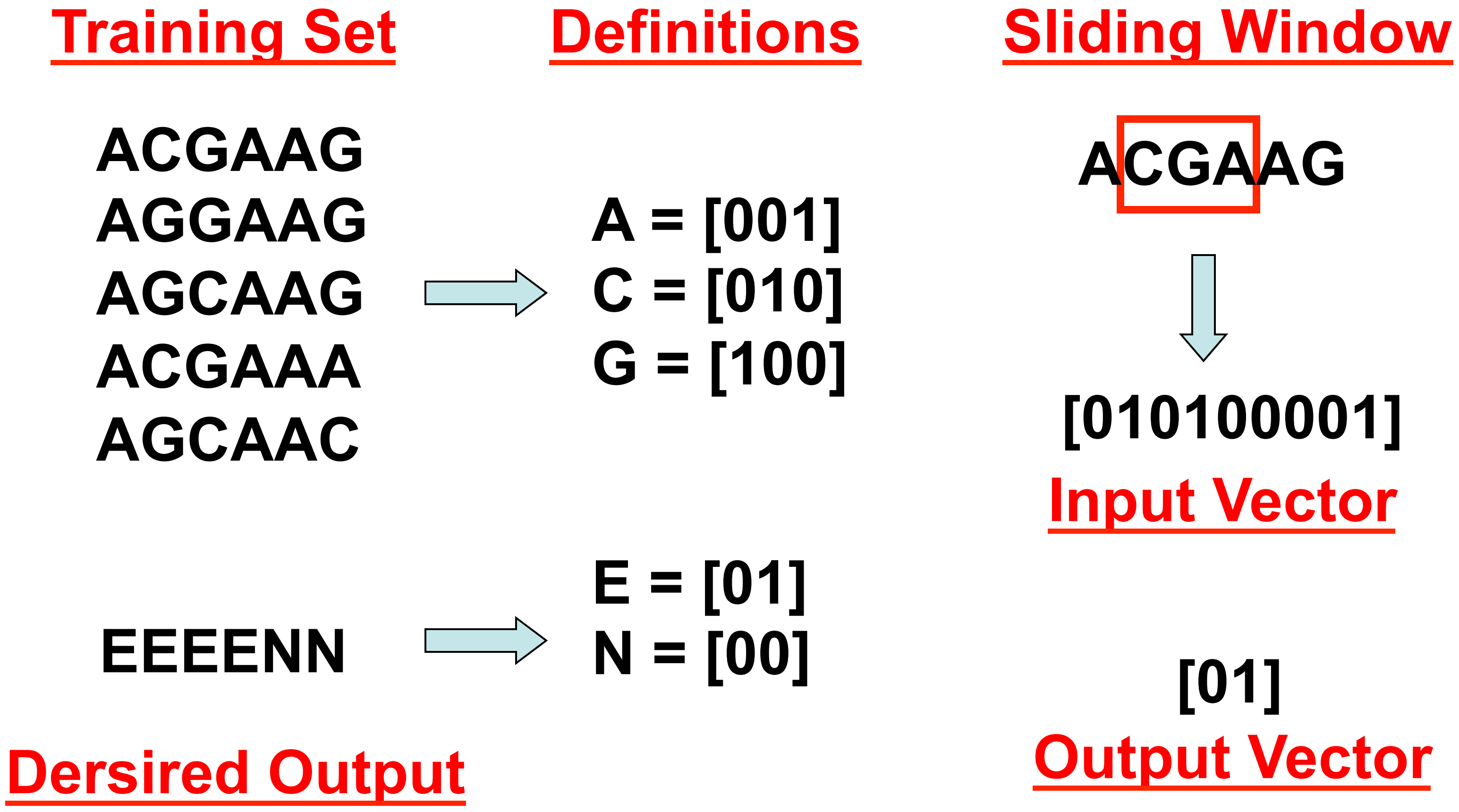
Neural Networks



Neural Network Applications

- Used in Intron/Exon Finding
- Used in Secondary Structure Prediction
- Used in Membrane Helix Prediction
- Used in Phosphorylation Site Prediction
- Used in Glycosylation Site Prediction
- Used in Splice Site Prediction
- Used in Signal Peptide Recognition

Neural Network*



Training Set

ACGAAG
AGGAAG
AGCAAG
ACGAAA
AGCAAC

Definitions

A = [001]
C = [010]
G = [100]

Sliding Window

ACGAAG

[010100001]

Input Vector

EEEENN

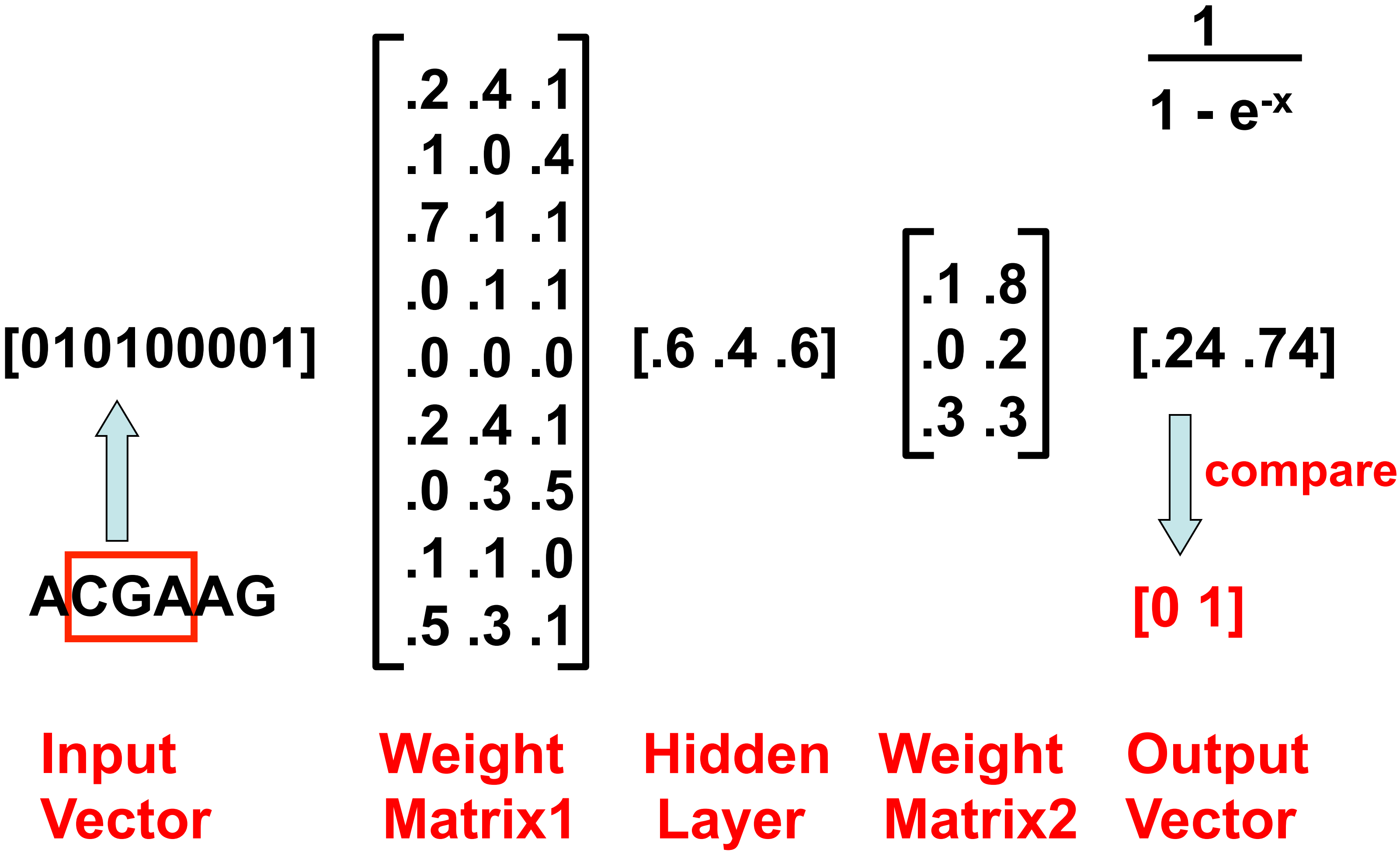
E = [01]
N = [00]

[01]

Output Vector

Dersired Output

Neural Network Training*



$$\frac{1}{1 - e^{-x}}$$

ACGAAG

[010100001]

$$\begin{bmatrix} .2 & .4 & .1 \\ .1 & .0 & .4 \\ .7 & .1 & .1 \\ .0 & .1 & .1 \\ .0 & .0 & .0 \\ .2 & .4 & .1 \\ .0 & .3 & .5 \\ .1 & .1 & .0 \\ .5 & .3 & .1 \end{bmatrix}$$

[.6 .4 .6]

$$\begin{bmatrix} .1 & .8 \\ .0 & .2 \\ .3 & .3 \end{bmatrix}$$

[.24 .74]

compare

[0 1]

Input Vector

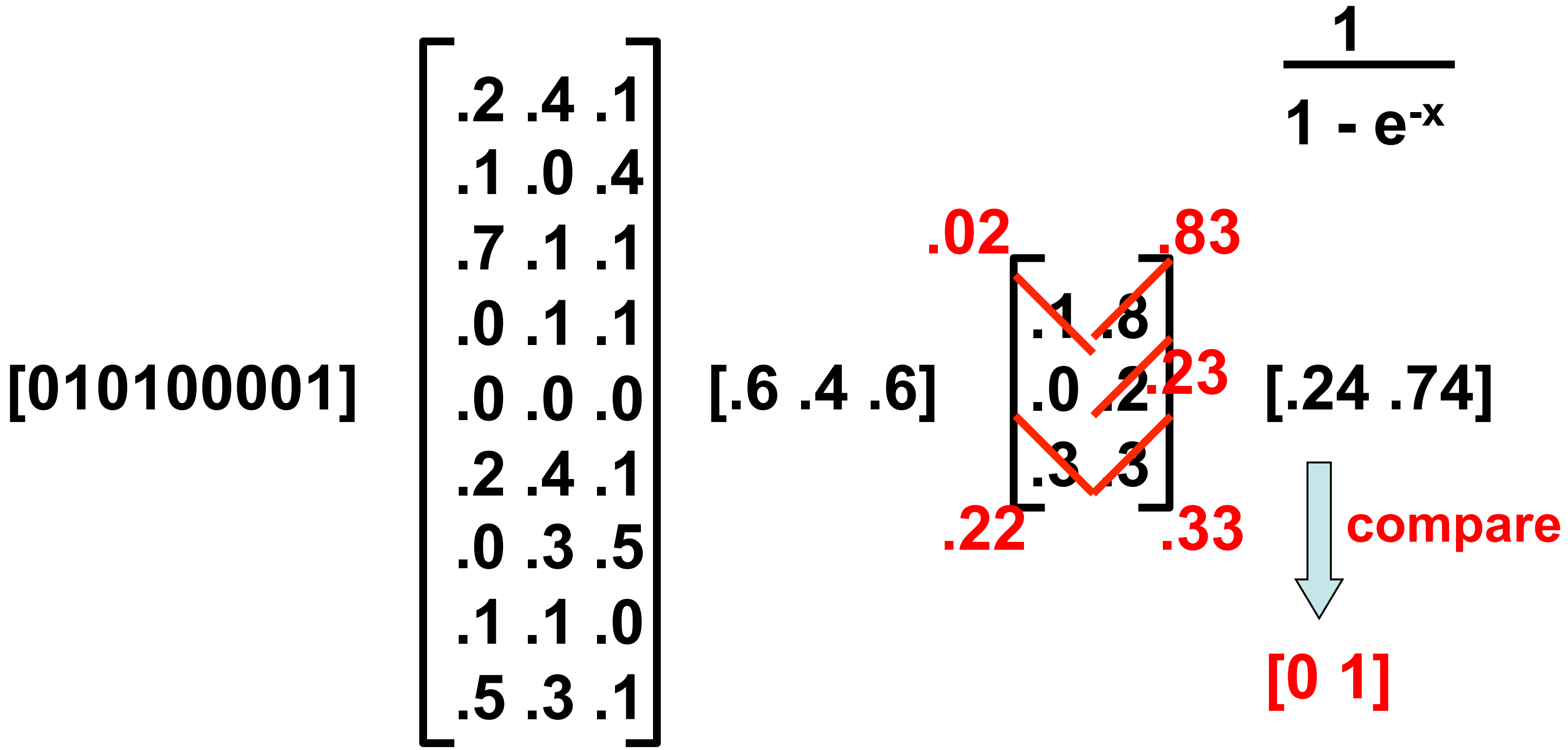
Weight Matrix1

Hidden Layer

Weight Matrix2

Output Vector

Back Propagation*



$$[010100001] \quad \begin{bmatrix} .2 & .4 & .1 \\ .1 & .0 & .4 \\ .7 & .1 & .1 \\ .0 & .1 & .1 \\ .0 & .0 & .0 \\ .2 & .4 & .1 \\ .0 & .3 & .5 \\ .1 & .1 & .0 \\ .5 & .3 & .1 \end{bmatrix} \quad [.6\ .4\ .6] \quad \begin{bmatrix} .1 & .8 \\ .0 & .2 \\ .3 & .3 \end{bmatrix} \quad [.24\ .74]$$

$$\frac{1}{1 - e^{-x}}$$

Input Vector

Weight Matrix1

Hidden Layer

Weight Matrix2

Output Vector

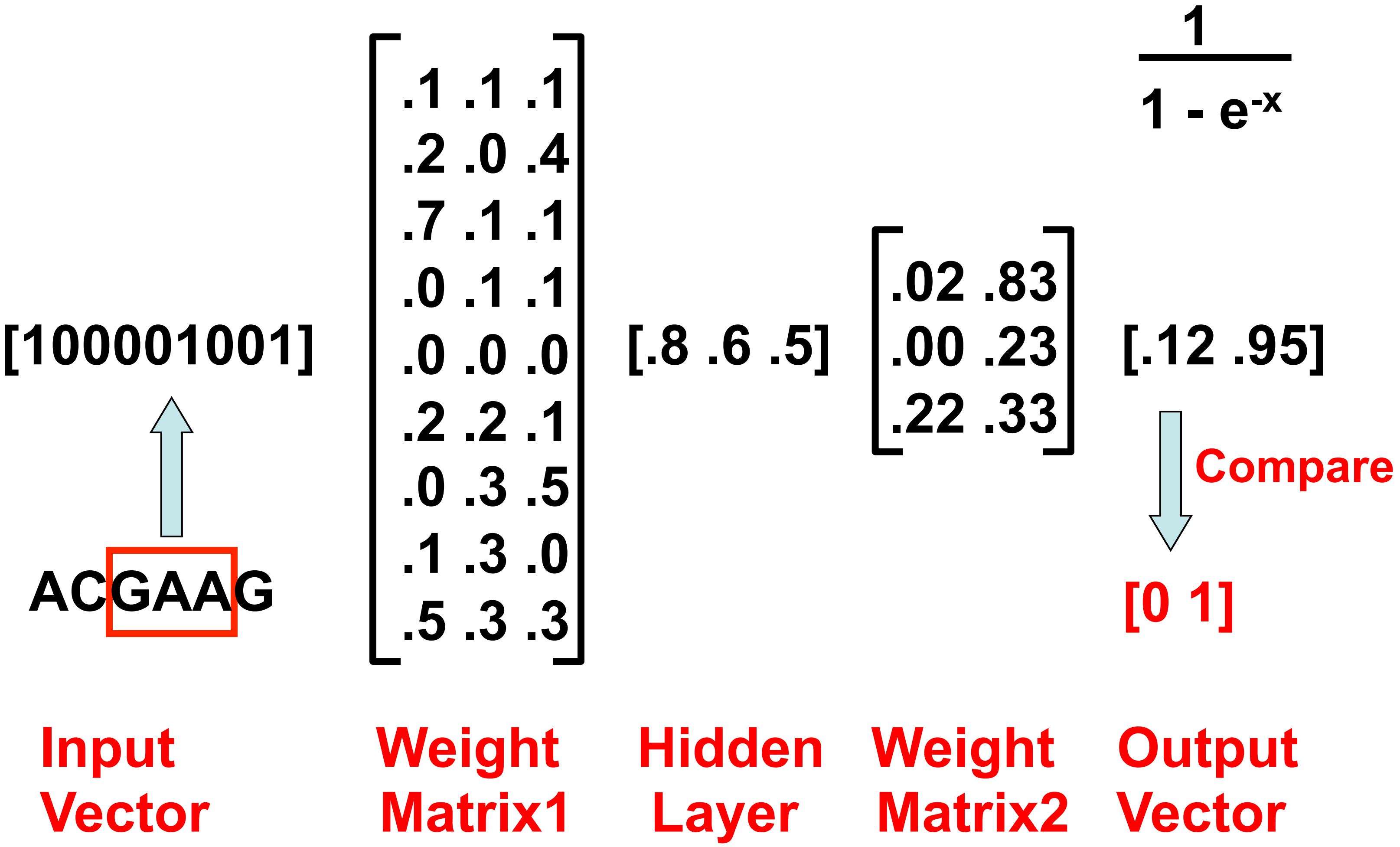
Calculate New Output*

$$\frac{1}{1 - e^{-x}}$$

$$\underset{\text{Input Vector}}{[010100001]} \quad \underset{\text{Weight Matrix1}}{\begin{bmatrix} .1 & .1 & .1 \\ .2 & .0 & .4 \\ .7 & .1 & .1 \\ .0 & .1 & .1 \\ .0 & .0 & .0 \\ .2 & .2 & .1 \\ .0 & .3 & .5 \\ .1 & .3 & .0 \\ .5 & .3 & .3 \end{bmatrix}} \quad \underset{\text{Hidden Layer}}{[.7\ .4\ .7]} \quad \underset{\text{Weight Matrix2}}{\begin{bmatrix} .02 & .83 \\ .00 & .23 \\ .22 & .33 \end{bmatrix}} \quad \underset{\text{Output Vector}}{[.16\ .91]}$$

$$[.16\ .91] \xrightarrow{\text{Converged!}} [0\ 1]$$

Train on Second Input Vector*



$$\frac{1}{1 - e^{-x}}$$

[100001001]

ACGAAG

$$\begin{bmatrix} .1 & .1 & .1 \\ .2 & .0 & .4 \\ .7 & .1 & .1 \\ .0 & .1 & .1 \\ .0 & .0 & .0 \\ .2 & .2 & .1 \\ .0 & .3 & .5 \\ .1 & .3 & .0 \\ .5 & .3 & .3 \end{bmatrix}$$

[.8 .6 .5]

$$\begin{bmatrix} .02 & .83 \\ .00 & .23 \\ .22 & .33 \end{bmatrix}$$

[.12 .95]

Compare

[0 1]

Input Vector

Weight Matrix1

Hidden Layer

Weight Matrix2

Output Vector

Back Propagation*

$$\frac{1}{1 - e^{-x}}$$

$$[010100001] \quad \begin{bmatrix} .1 & .1 & .1 \\ .2 & .0 & .4 \\ .7 & .1 & .1 \\ .0 & .1 & .1 \\ .0 & .0 & .0 \\ .2 & .2 & .1 \\ .0 & .3 & .5 \\ .1 & .3 & .0 \\ .5 & .3 & .3 \end{bmatrix} \quad [.8\ .6\ .5] \quad \begin{bmatrix} \overset{.01}{\cancel{.02}} & \overset{.84}{\cancel{.83}} \\ .00 & \cancel{.23}\,.24 \\ \underset{.21}{\cancel{.22}} & \underset{.34}{\cancel{.33}} \end{bmatrix} \quad [.12\ .95]$$

$$\Big\downarrow \text{ compare}$$

$$[0\ 1]$$

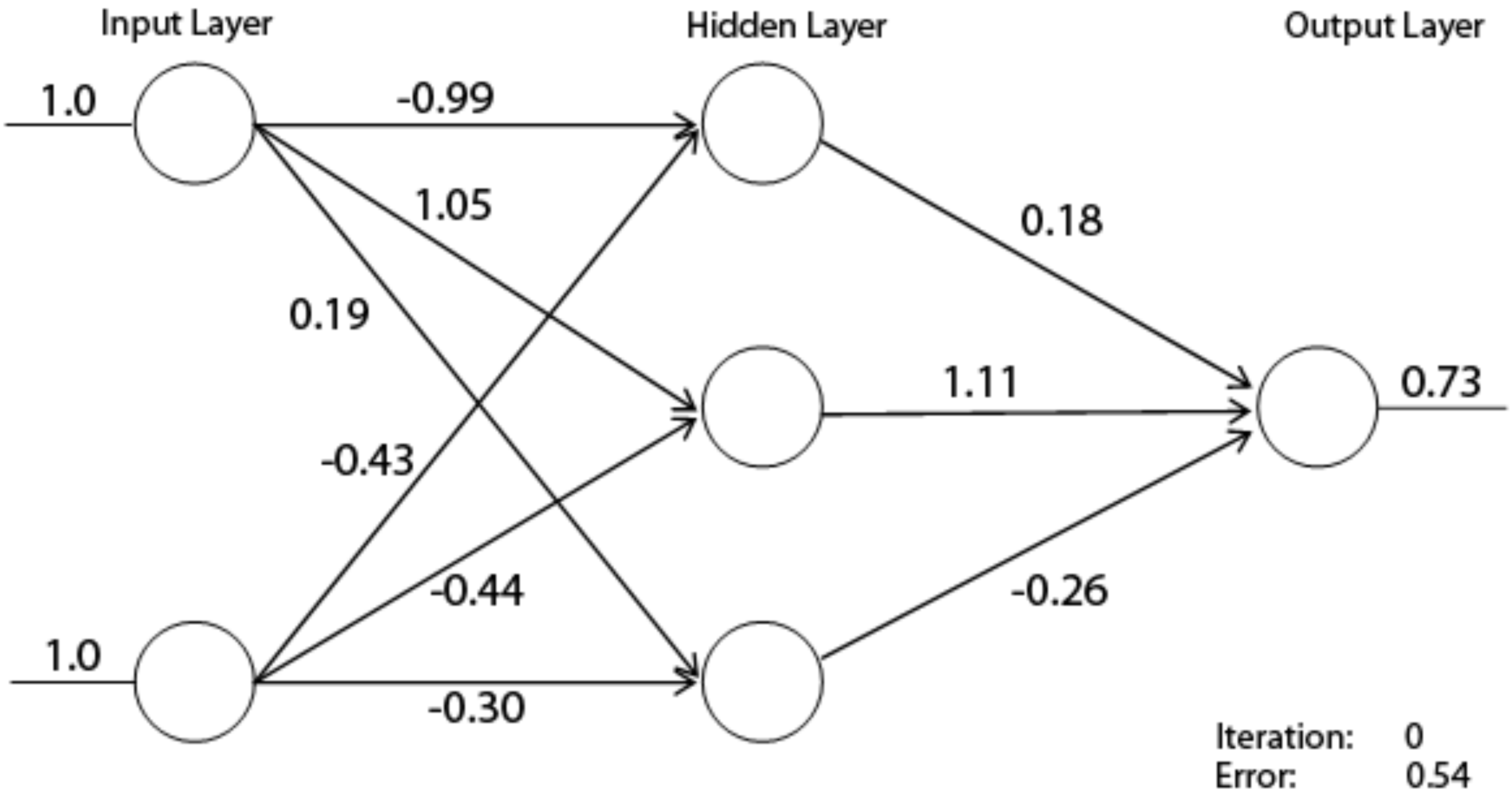
Input Vector	Weight Matrix1	Hidden Layer	Weight Matrix2	Output Vector

After Many Iterations….

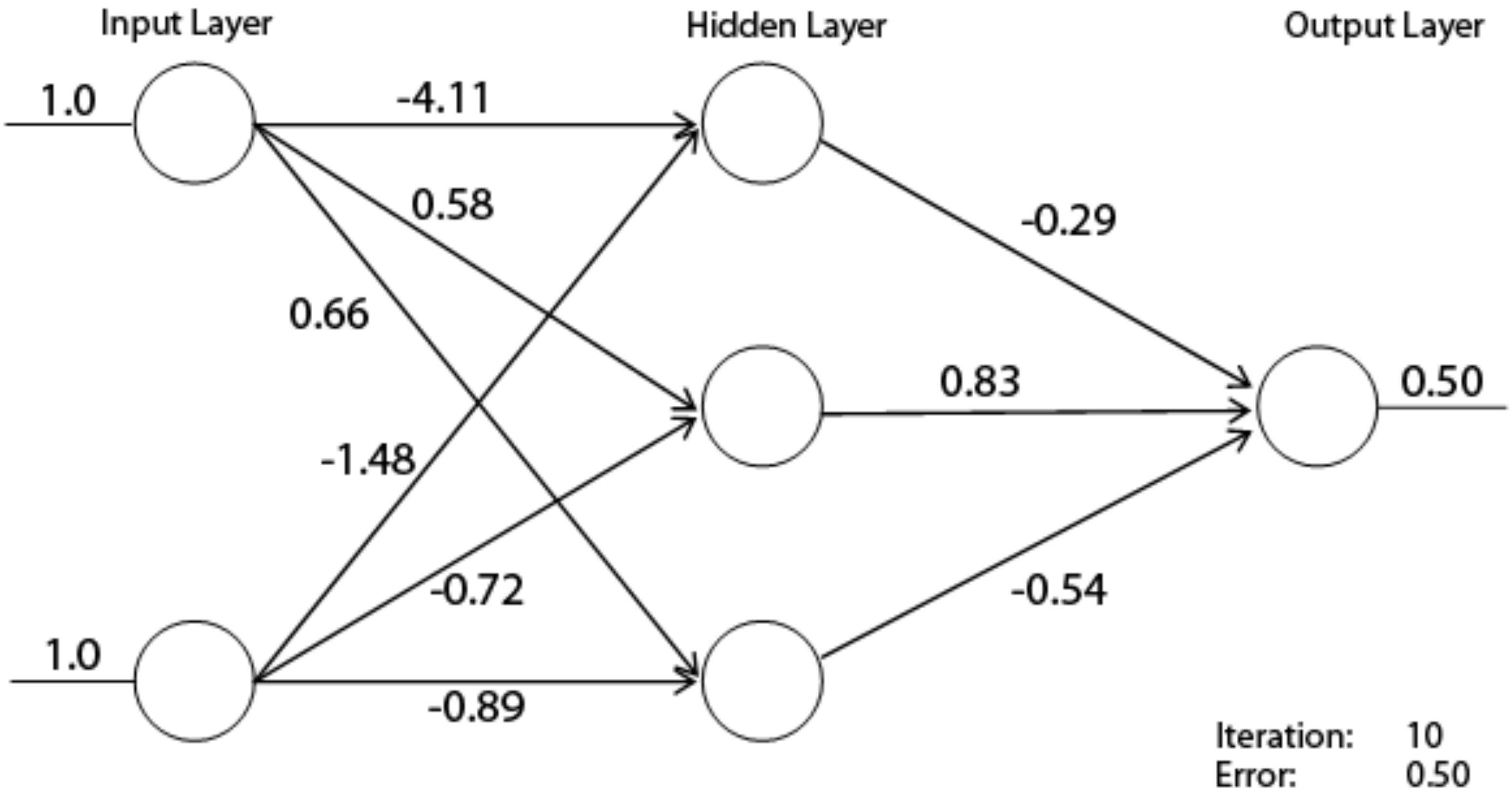
$$\begin{bmatrix} .13 & .08 & .12 \\ .24 & .01 & .45 \\ .76 & .01 & .31 \\ .06 & .32 & .14 \\ .03 & .11 & .23 \\ .21 & .21 & .51 \\ .10 & .33 & .85 \\ .12 & .34 & .09 \\ .51 & .31 & .33 \end{bmatrix} \qquad \begin{bmatrix} .03 & .93 \\ .01 & .24 \\ .12 & .23 \end{bmatrix}$$

Two "Generalized" Weight Matrices

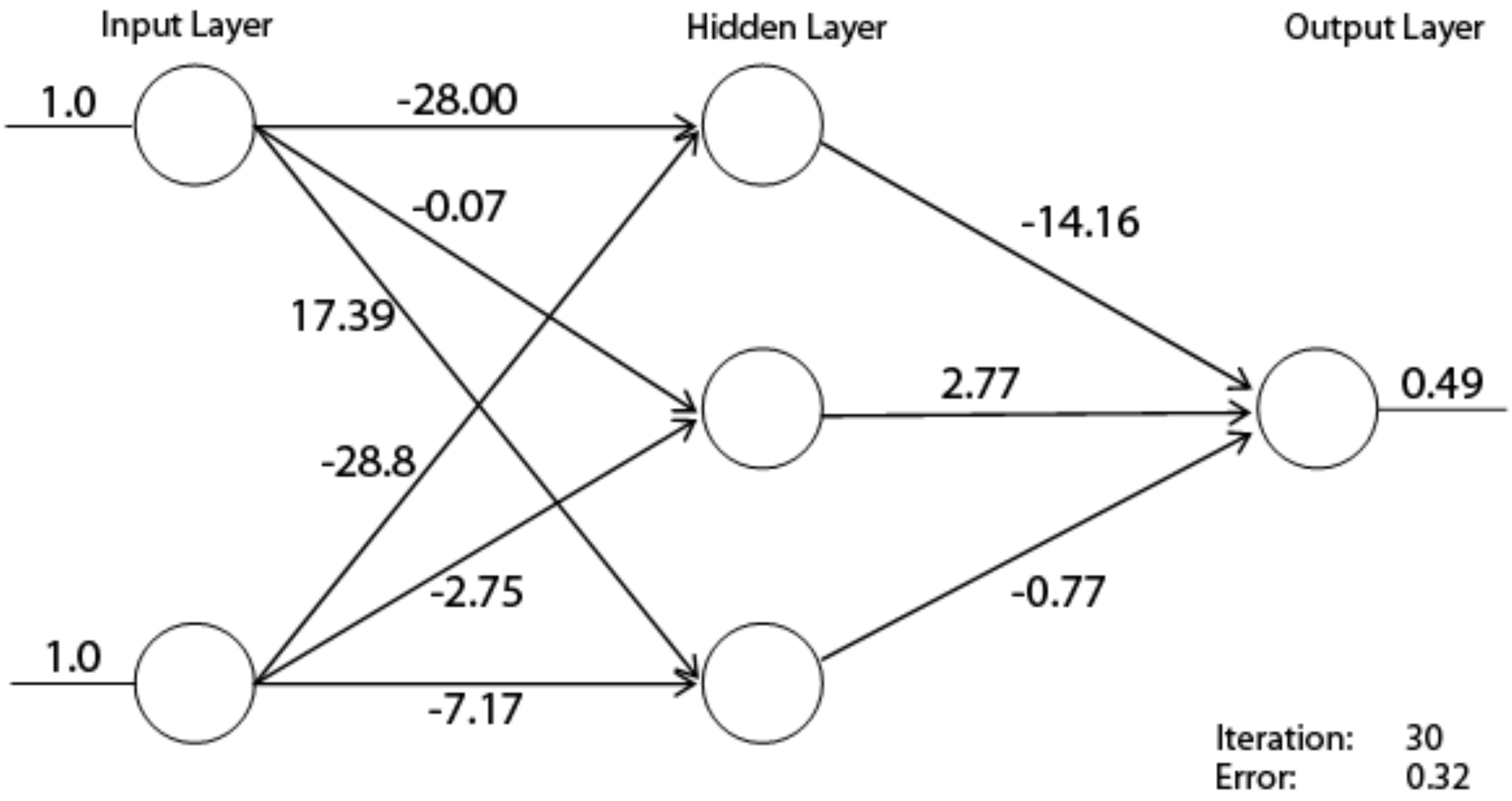
Neural Network in Action



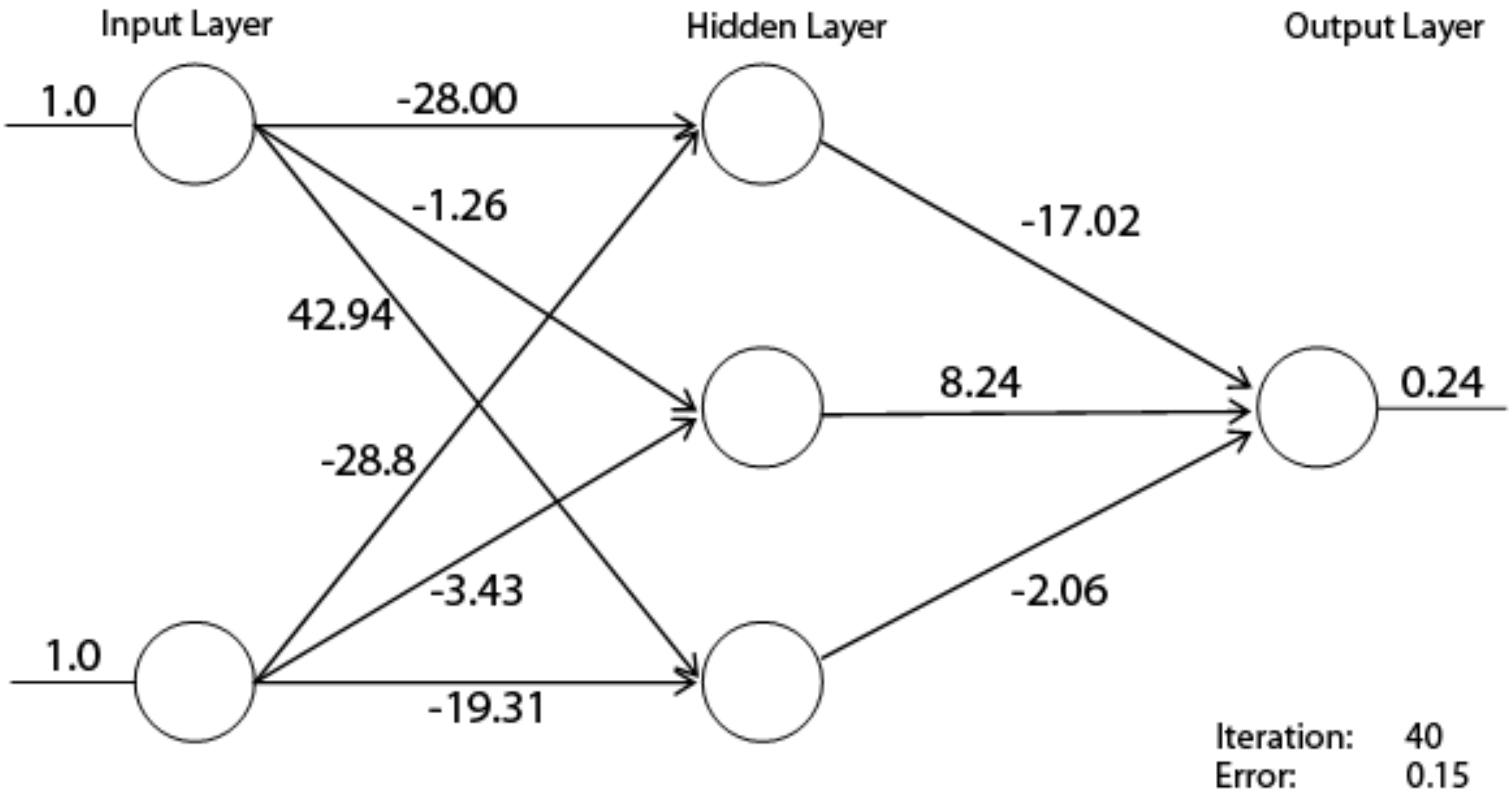
Neural Network in Action



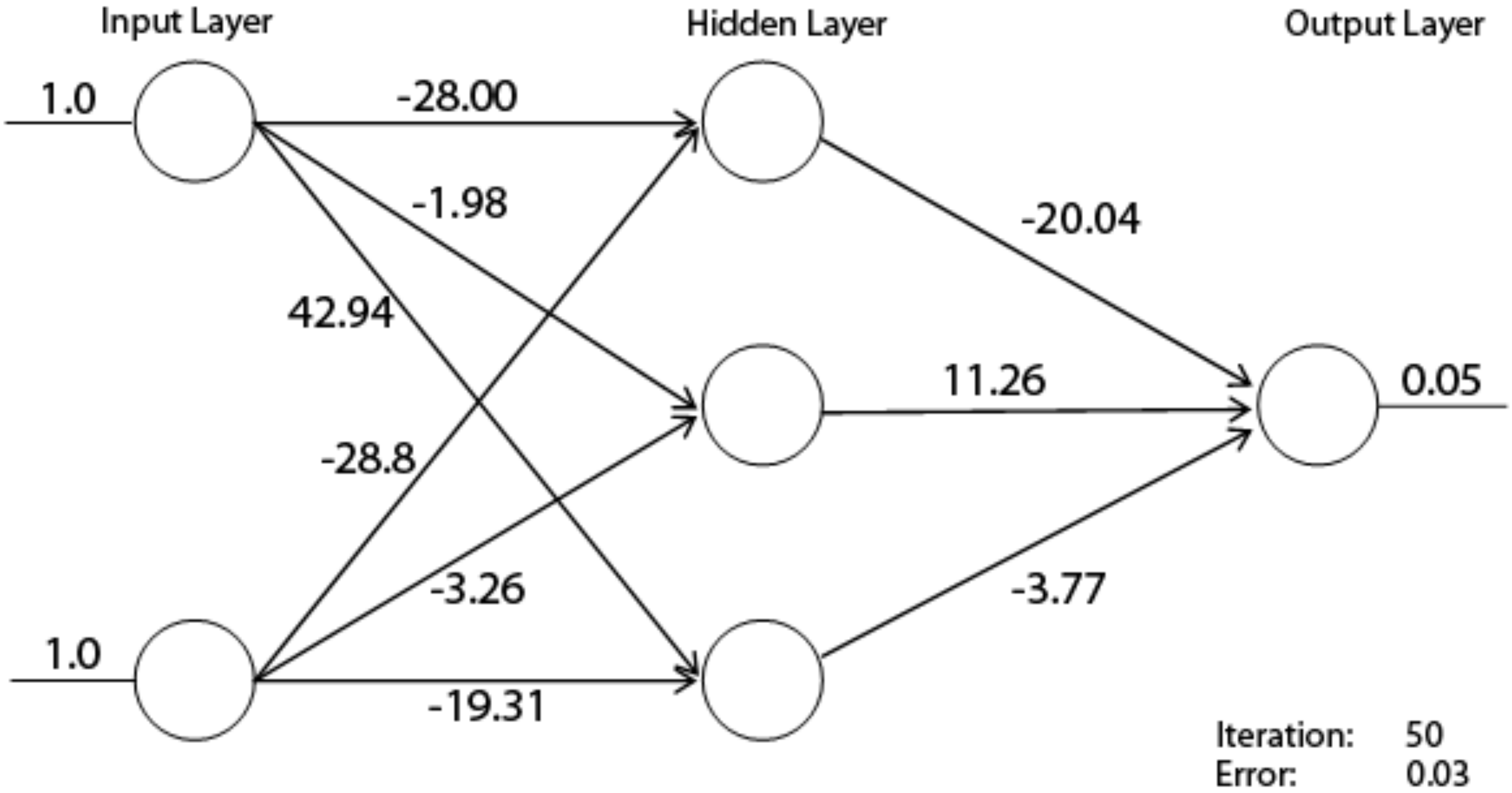
Neural Network in Action



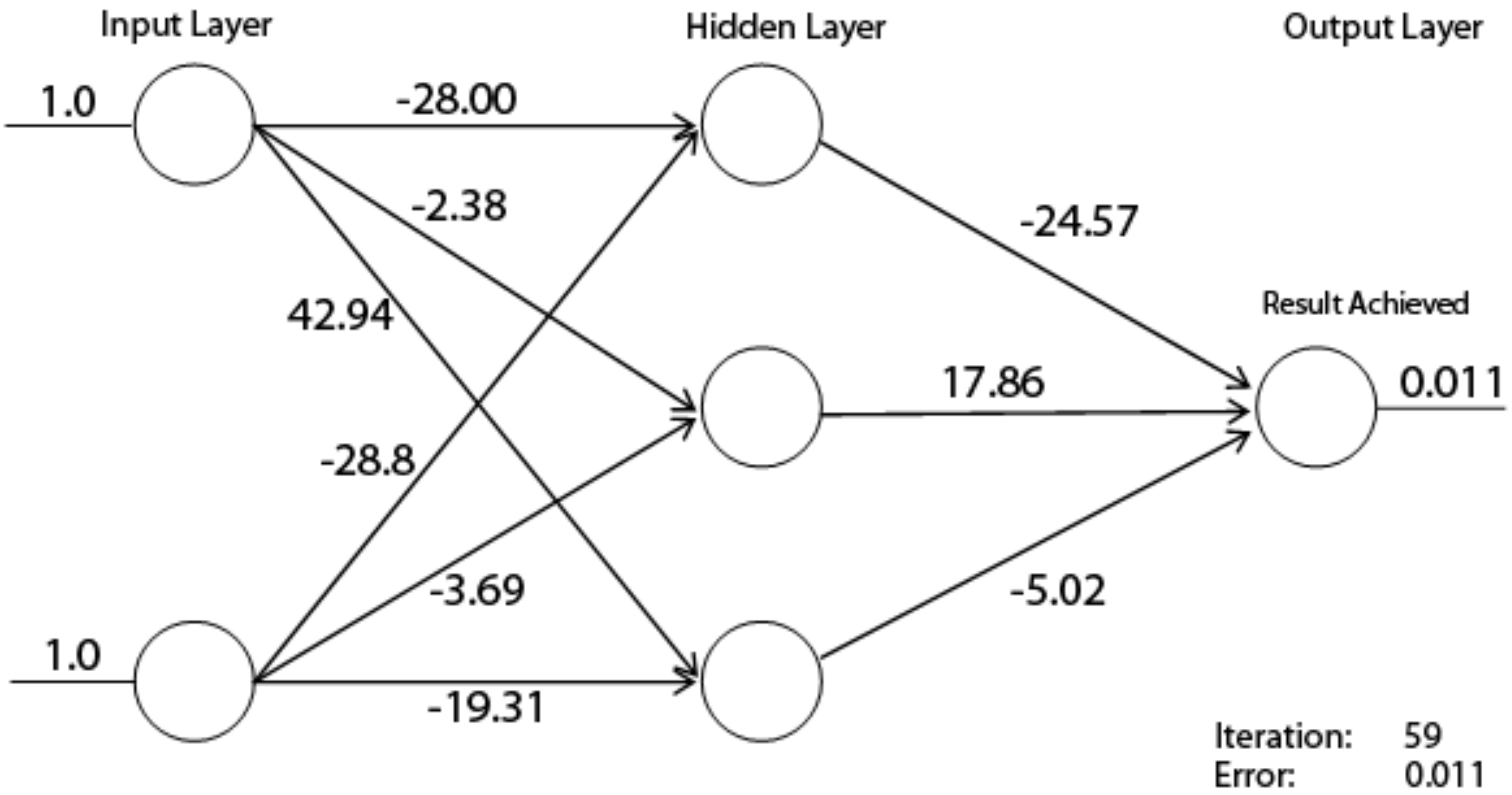
Neural Network in Action



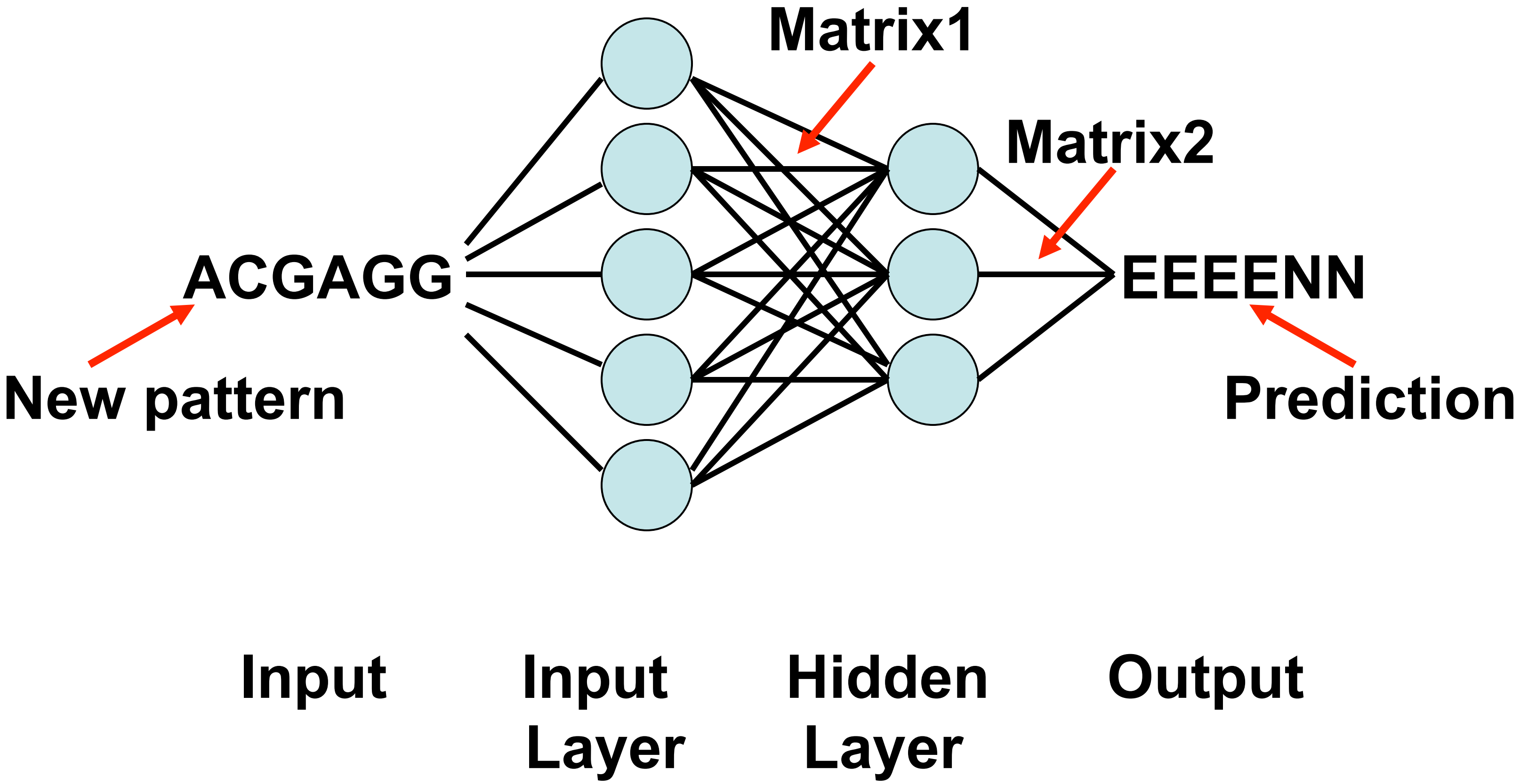
Neural Network in Action



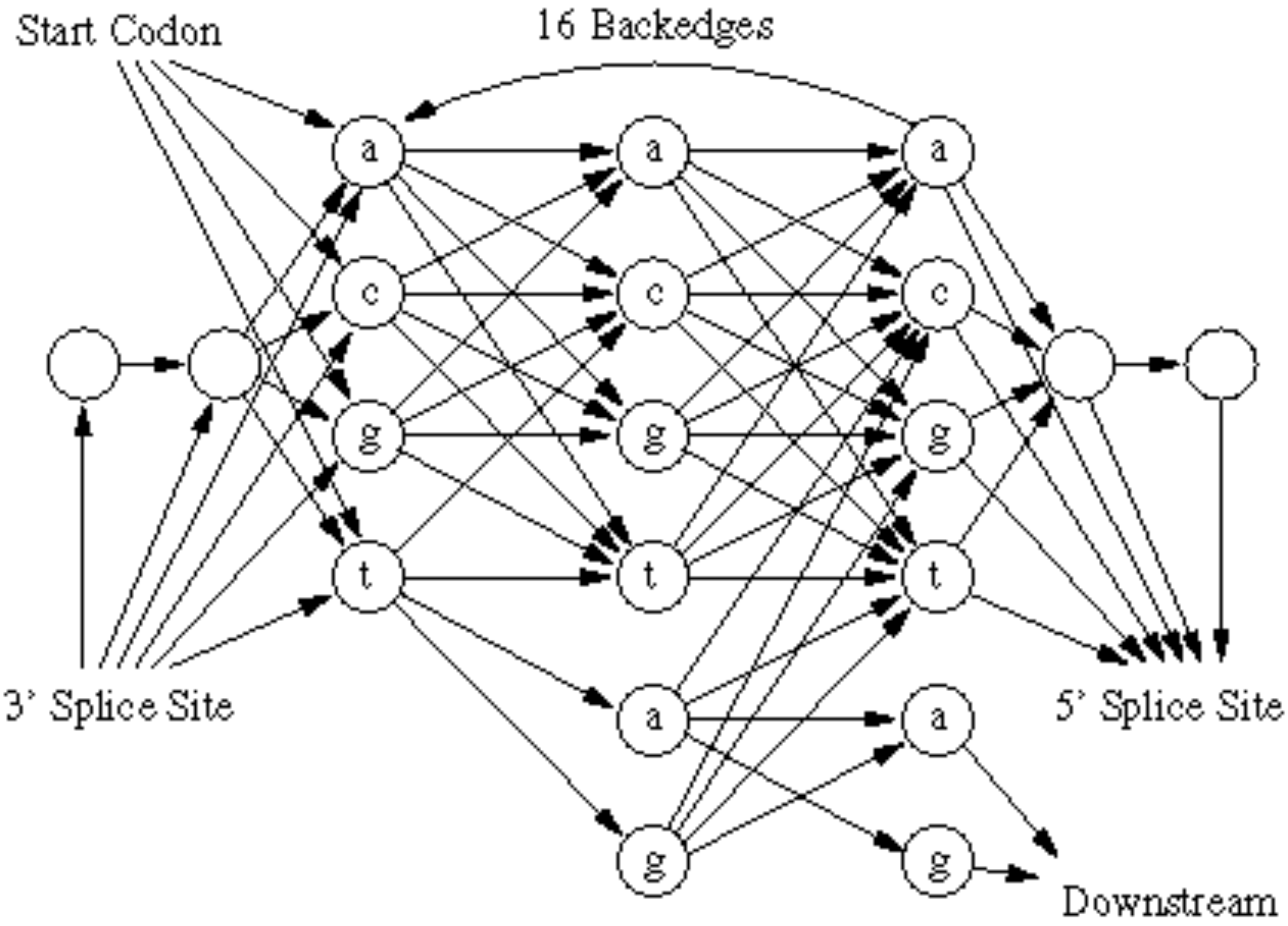
Neural Network in Action



Neural Networks



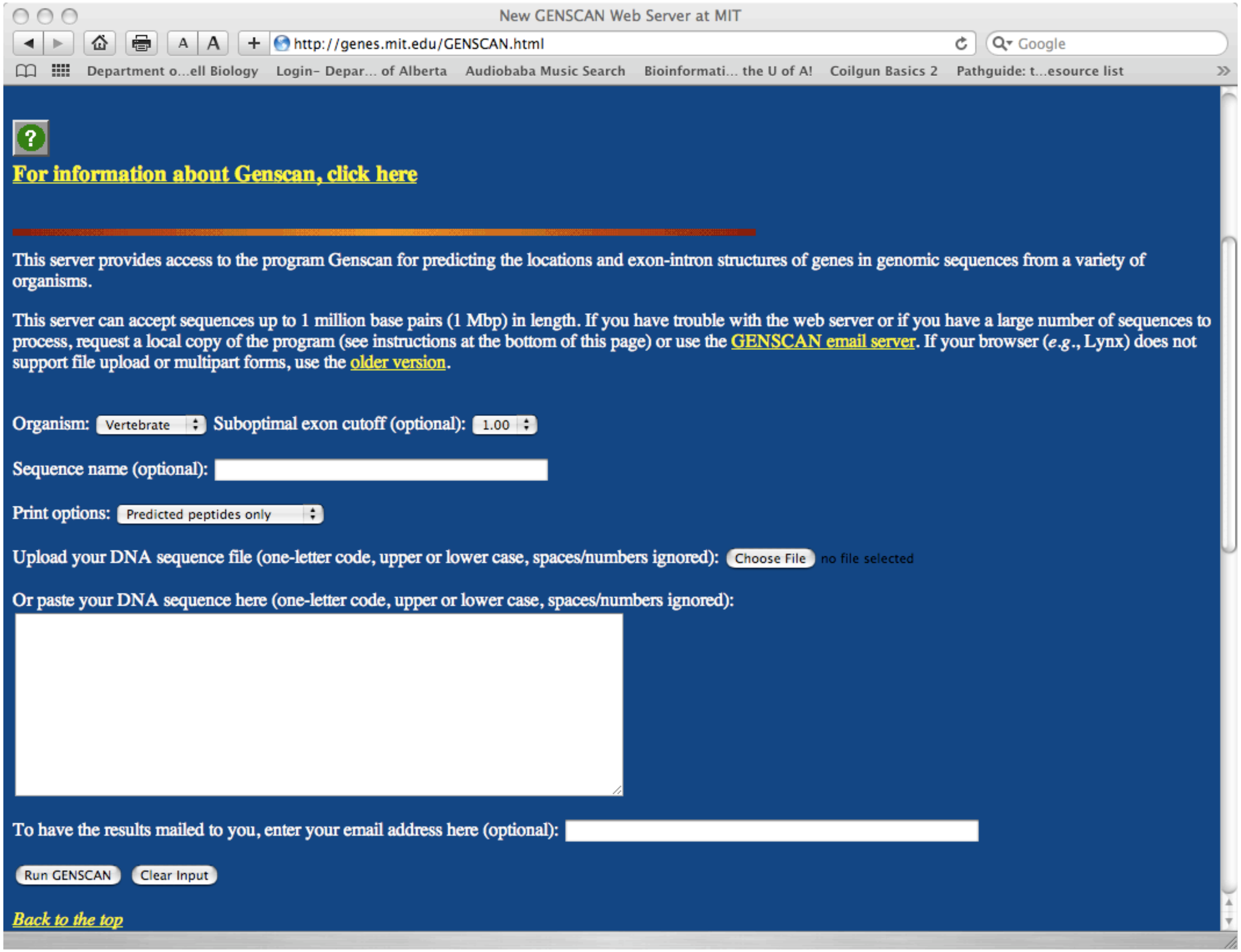
HMM for Gene Finding



Combined Methods

- Bring 2 or more methods together (usually site detection + composition)
- GrailEXP (http://compbio.ornl.gov/Grail-1.3/)
- GeneMark-E (http://exon.biology.gatech.edu/)
- HMMgene (http://www.cbs.dtu.dk/services/HMMgene/)
- GENSCAN(http://genes.mit.edu/GENSCAN.html)
- GRPL (GeneTool/BioTools)

Genscan*



New GENSCAN Web Server at MIT
http://genes.mit.edu/GENSCAN.html
Google
Department o...ell Biology
Login– Depar... of Alberta
Audiobaba Music Search
Bioinformati... the U of A!
Coilgun Basics 2
Pathguide: t...esource list

For information about Genscan, click here

This server provides access to the program Genscan for predicting the locations and exon-intron structures of genes in genomic sequences from a variety of organisms.

This server can accept sequences up to 1 million base pairs (1 Mbp) in length. If you have trouble with the web server or if you have a large number of sequences to process, request a local copy of the program (see instructions at the bottom of this page) or use the GENSCAN email server. If your browser (*e.g.*, Lynx) does not support file upload or multipart forms, use the older version.

Organism: Vertebrate Suboptimal exon cutoff (optional): 1.00

Sequence name (optional):

Print options: Predicted peptides only

Upload your DNA sequence file (one-letter code, upper or lower case, spaces/numbers ignored): Choose File no file selected

Or paste your DNA sequence here (one-letter code, upper or lower case, spaces/numbers ignored):

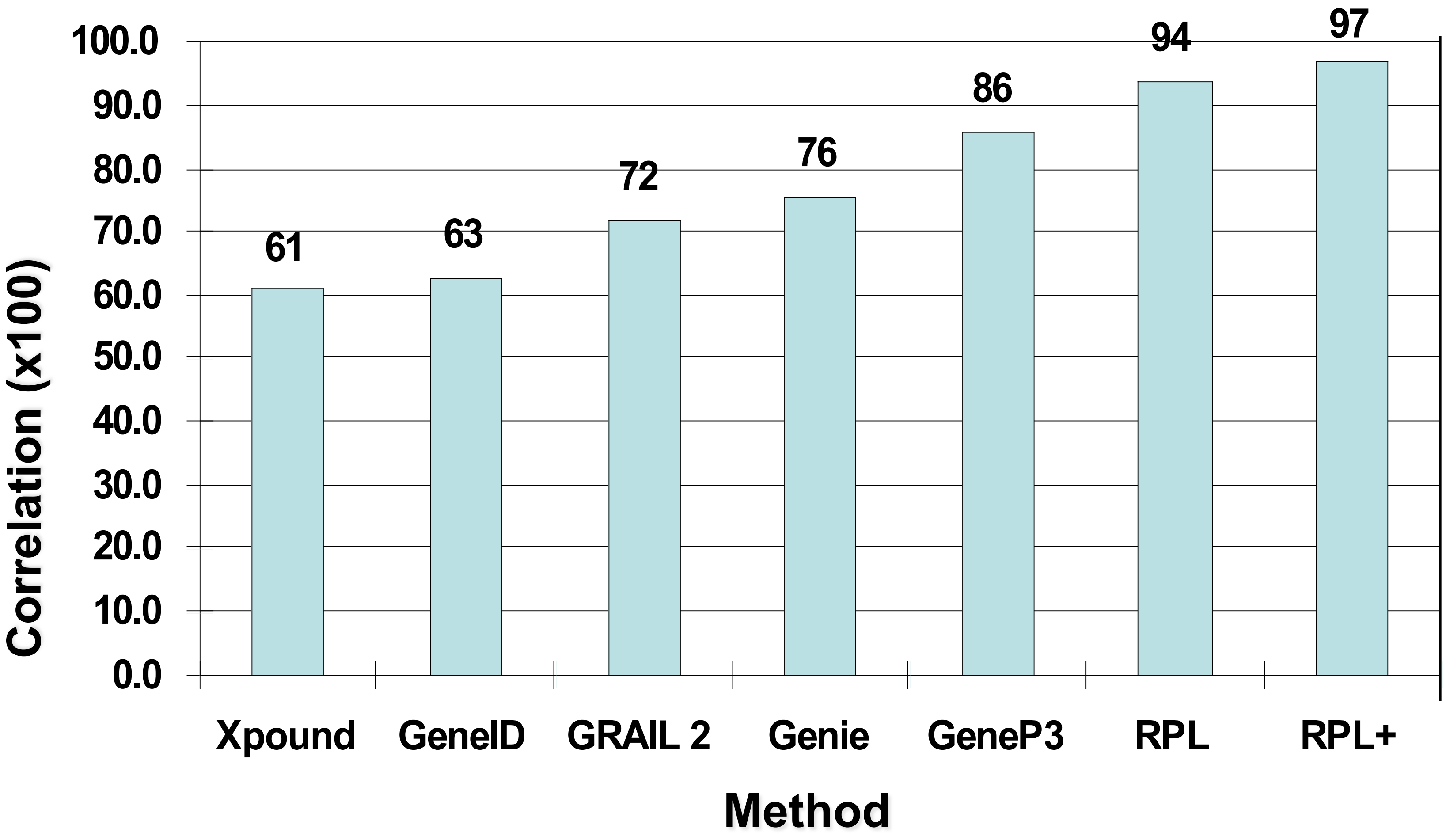
To have the results mailed to you, enter your email address here (optional):

Back to the top

How Do They Work?*

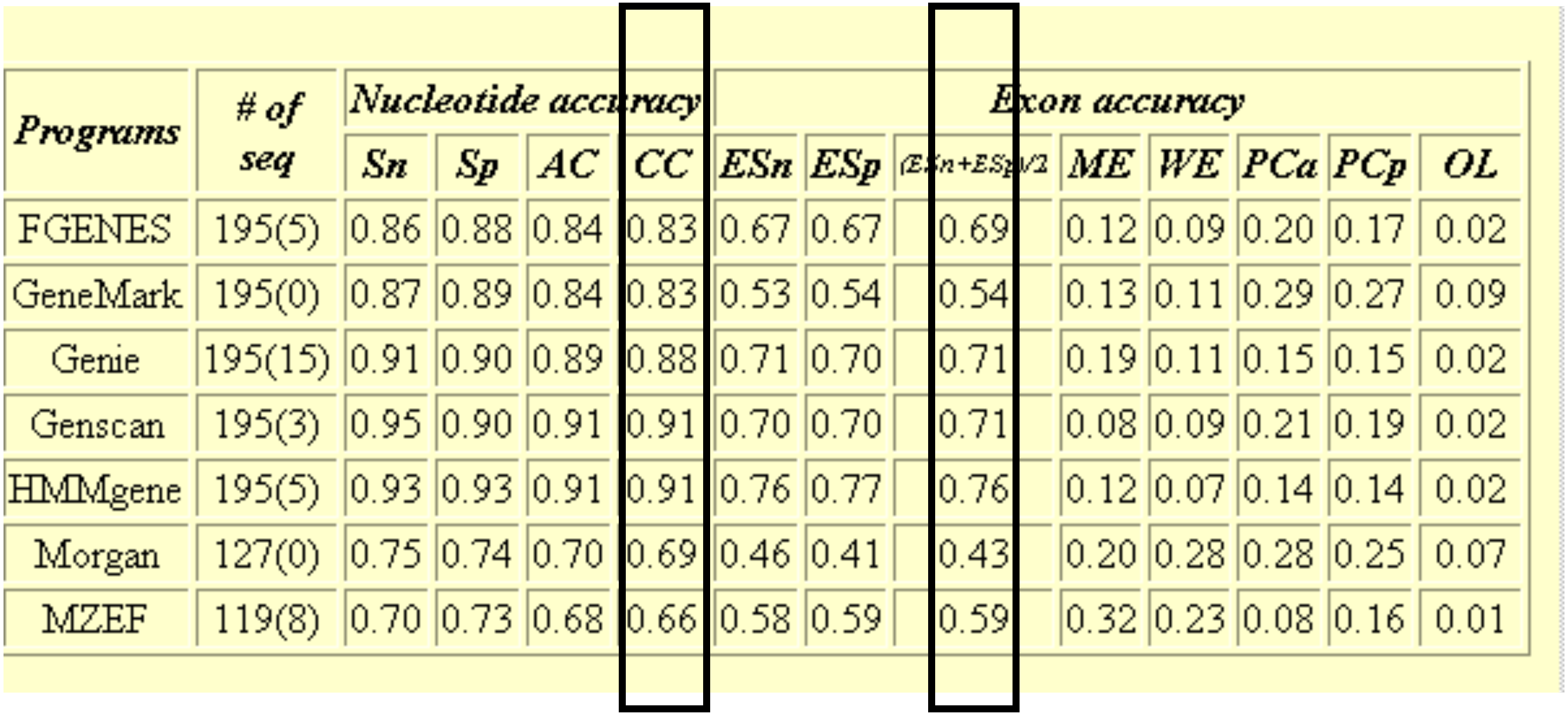
- GENSCAN
  - 5th order Hidden Markov Model
  - Hexamer composition statistics of exons vs. introns
  - Exon/intron length distributions
  - Scan of promoter and polyA signals
  - Weight matrices of 5' splice signals and start codon region (12 bp)
  - Uses dynamic programming to optimize gene model using above data

How Well Do They Do?



Burset & Guigio test set (1996)

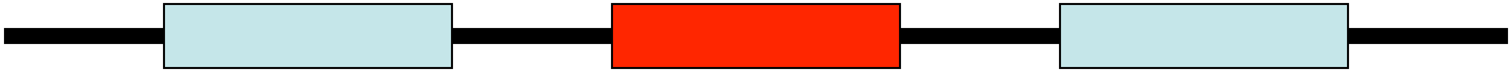
How Well Do They Do?*



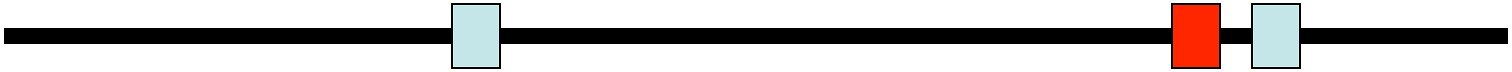
Programs	# of seq	Nucleotide accuracy				Exon accuracy								
		Sn	Sp	AC	CC	ESn	ESp	(ESn+ESp)/2	ME	WE	PCa	PCp	OL
FGENES	195(5)	0.86	0.88	0.84	0.83	0.67	0.67	0.69	0.12	0.09	0.20	0.17	0.02
GeneMark	195(0)	0.87	0.89	0.84	0.83	0.53	0.54	0.54	0.13	0.11	0.29	0.27	0.09
Genie	195(15)	0.91	0.90	0.89	0.88	0.71	0.70	0.71	0.19	0.11	0.15	0.15	0.02
Genscan	195(3)	0.95	0.90	0.91	0.91	0.70	0.70	0.71	0.08	0.09	0.21	0.19	0.02
HMMgene	195(5)	0.93	0.93	0.91	0.91	0.76	0.77	0.76	0.12	0.07	0.14	0.14	0.02
Morgan	127(0)	0.75	0.74	0.70	0.69	0.46	0.41	0.43	0.20	0.28	0.28	0.25	0.07
MZEF	119(8)	0.70	0.73	0.68	0.66	0.58	0.59	0.59	0.32	0.23	0.08	0.16	0.01

"Evaluation of gene finding programs" S. Rogic, A. K. Mackworth and B. F. F. Ouellette. Genome Research, 11: 817-832 (2001).

Easy vs. Hard Predictions

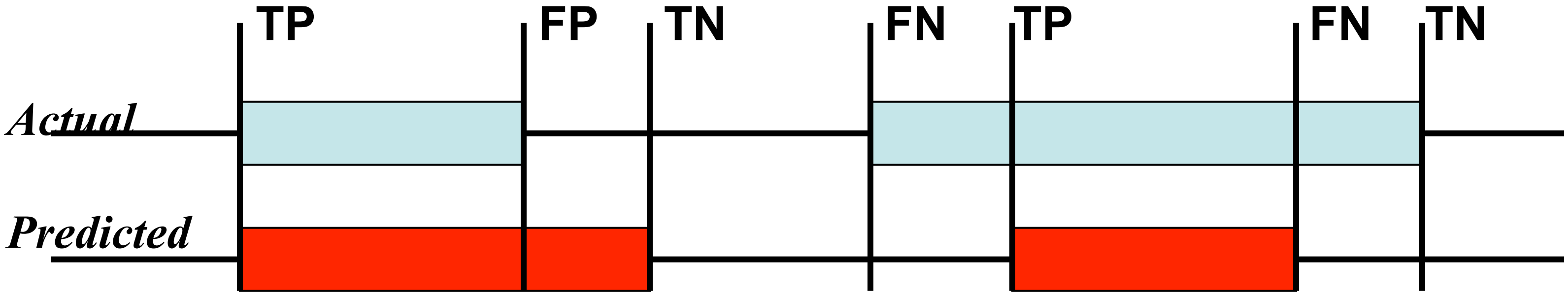


3 equally abundant states (easy)
BUT random prediction = 33% correct



Rare events, unequal distribution (hard)
BUT "biased" random prediction = 90% correct

Gene Prediction (Evaluation)*



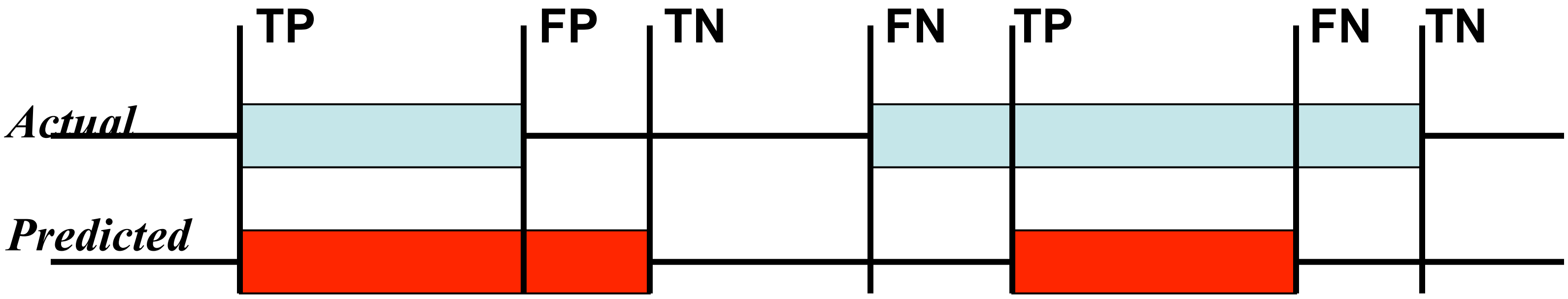
Sensitivity ***Measure of the % of false negative results (sn = 0.996 means 0.4% false negatives)***

Specificity ***Measure of the % of false positive results***

Precision ***Measure of the % positive results***

Correlation ***Combined measure of sensitivity and specificity***

Gene Prediction (Evaluation)



Sensitivity or Recall $Sn = TP/(TP + FN)$

Specificity $Sp = TN/(TN + FP)$

Precision $Pr = TP/(TP + FP)$

Correlation

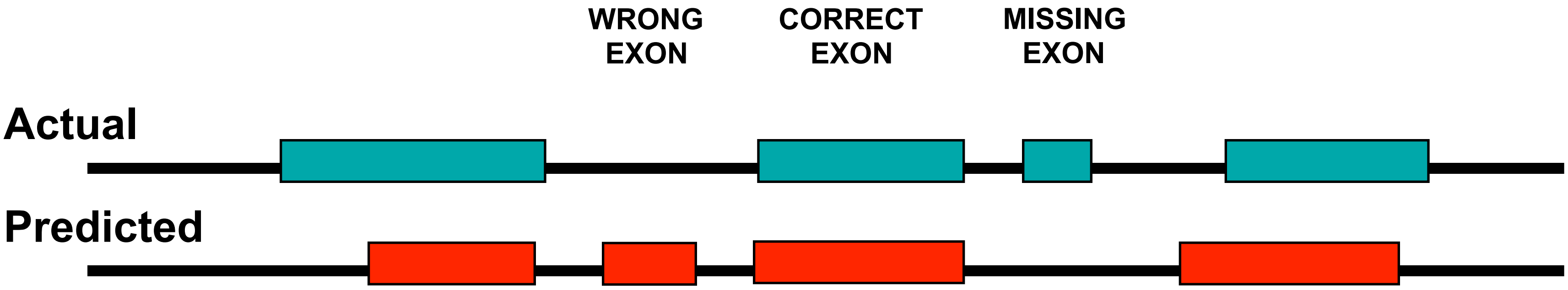
$$CC = (TP*TN-FP*FN)/[(TP+FP)(TN+FN)(TP+FN)(TN+FP)]^{0.5}$$

This is a better way of evaluating

Different Strokes for Different Folks

- Precision and specificity statistics favor conservative predictors that make no prediction when there is doubt about the correctness of a prediction, while the sensitivity (recall) statistic favors liberal predictors that make a prediction if there is a chance of success.
- Information retrieval papers report precision and recall, while bioinformatics papers tend to report specificity and sensitivity.

Gene Prediction Accuracy at the Exon Level *



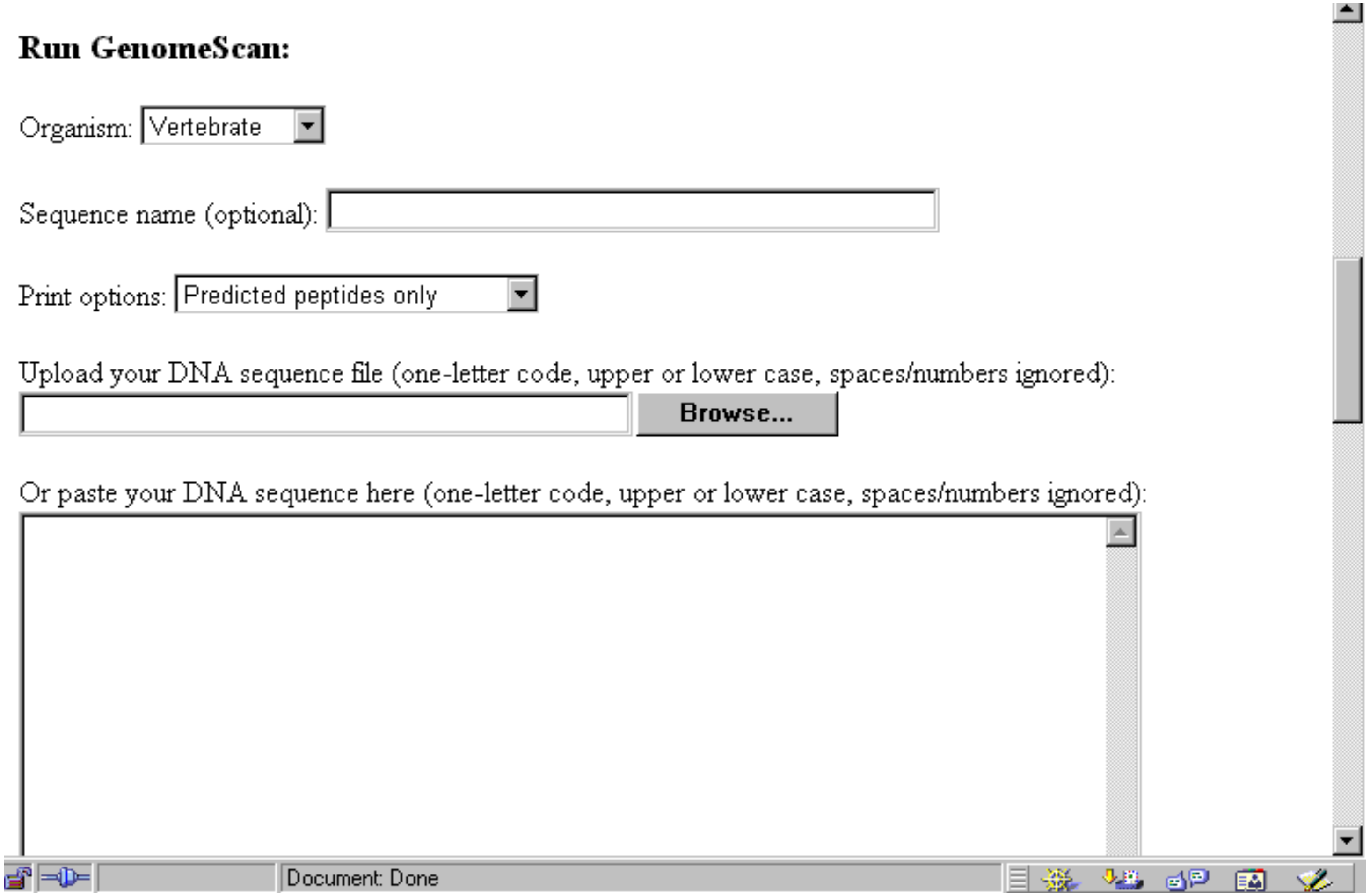
Sensitivity $$S_n = \frac{\text{number of correct exons}}{\text{number of actual exons}}$$

Specificity $$S_p = \frac{\text{number of correct exons}}{\text{number of predicted exons}}$$

Better Approaches Are Emerging...

- Programs that combine site, comparative and composition (3 in 1)
  - GenomeScan, FGENESH++, Twinscan
- Programs that use synteny between organisms
  - ROSETTA, SLAM, SGP
- Programs that combine predictions from multiple predictors
  - GeneComber, Augustus

GenomeScan - http://genes.mit.edu/genomescan.html



Run GenomeScan:

Organism: Vertebrate

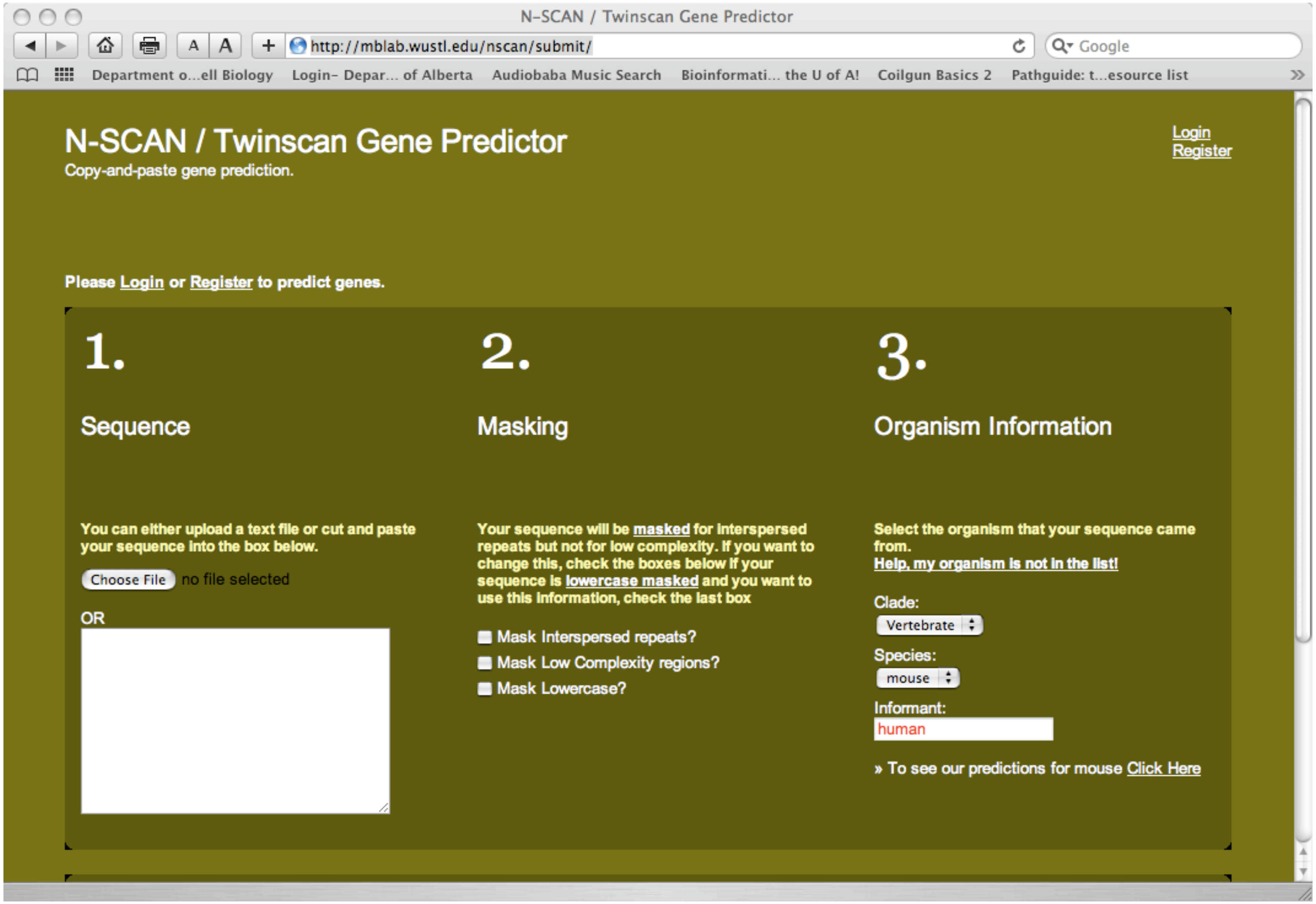
Sequence name (optional):

Print options: Predicted peptides only

Upload your DNA sequence file (one-letter code, upper or lower case, spaces/numbers ignored):

Or paste your DNA sequence here (one-letter code, upper or lower case, spaces/numbers ignored):

TwinScan - http://mblab.wustl.edu/nscan/submit/ (requires Login)


N-SCAN / Twinscan Gene Predictor
http://mblab.wustl.edu/nscan/submit/
Google
Department o...ell Biology
Login- Depar... of Alberta
Audiobaba Music Search
Bioinformati... the U of A!
Coilgun Basics 2
Pathguide: t...esource list

N-SCAN / Twinscan Gene Predictor

Copy-and-paste gene prediction.

Login
Register

Please Login or Register to predict genes.

1.

Sequence

You can either upload a text file or cut and paste your sequence into the box below.

Choose File no file selected

OR

2.

Masking

Your sequence will be masked for interspersed repeats but not for low complexity. If you want to change this, check the boxes below if your sequence is lowercase masked and you want to use this information, check the last box

□ Mask Interspersed repeats?
□ Mask Low Complexity regions?
□ Mask Lowercase?

3.

Organism Information

Select the organism that your sequence came from.
Help, my organism is not in the list!

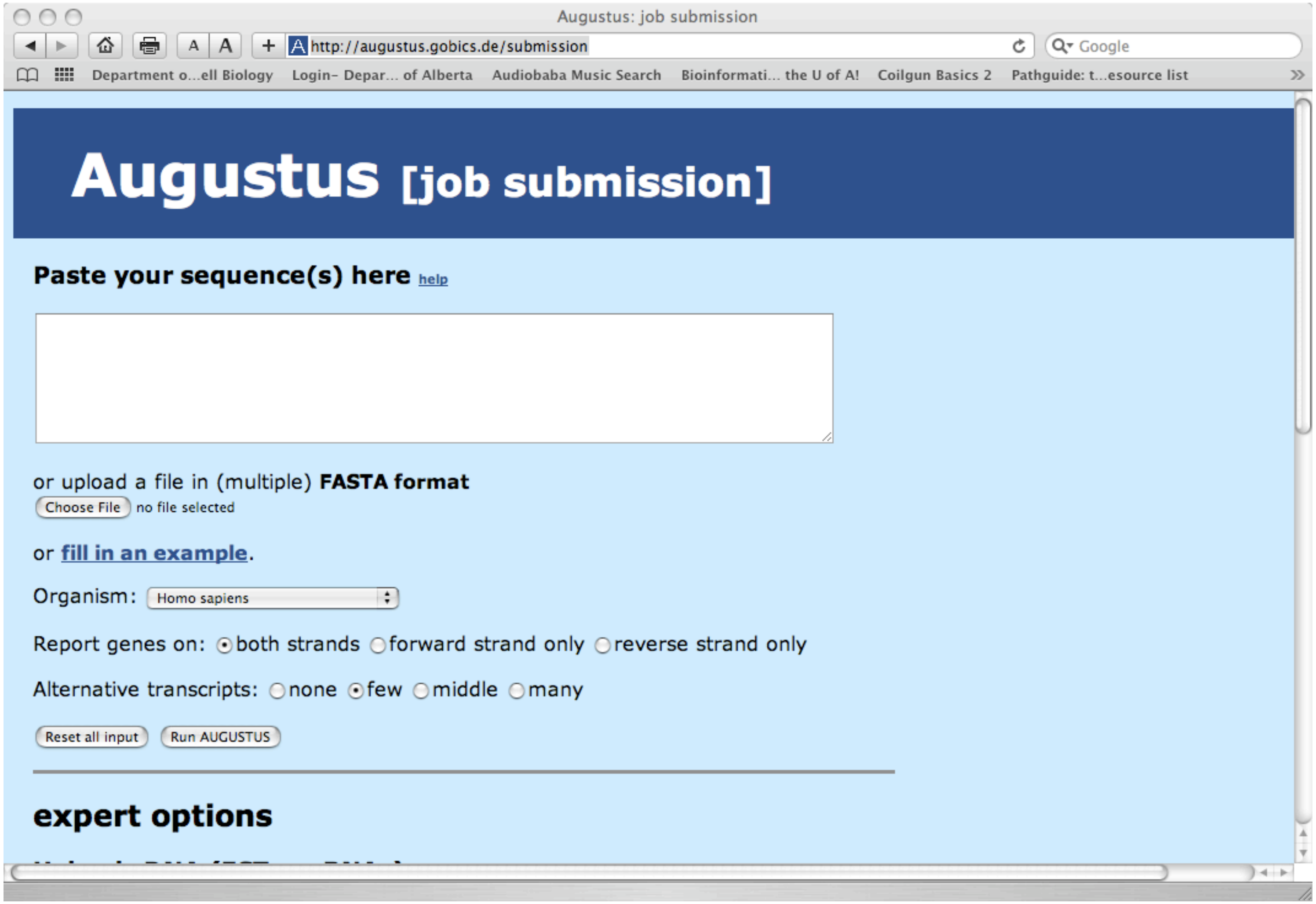
Clade: Vertebrate

Species: mouse

Informant: human

» To see our predictions for mouse Click Here

Augustus – http://bioinf.uni-greifswald.de/augustus/submission/


Augustus: job submission
http://augustus.gobics.de/submission
Google
Department o...ell Biology
Login– Depar... of Alberta
Audiobaba Music Search
Bioinformati... the U of A!
Coilgun Basics 2
Pathguide: t...esource list

Augustus [job submission]

Paste your sequence(s) here help

or upload a file in (multiple) **FASTA format**

no file selected

or fill in an example.

Organism: Homo sapiens

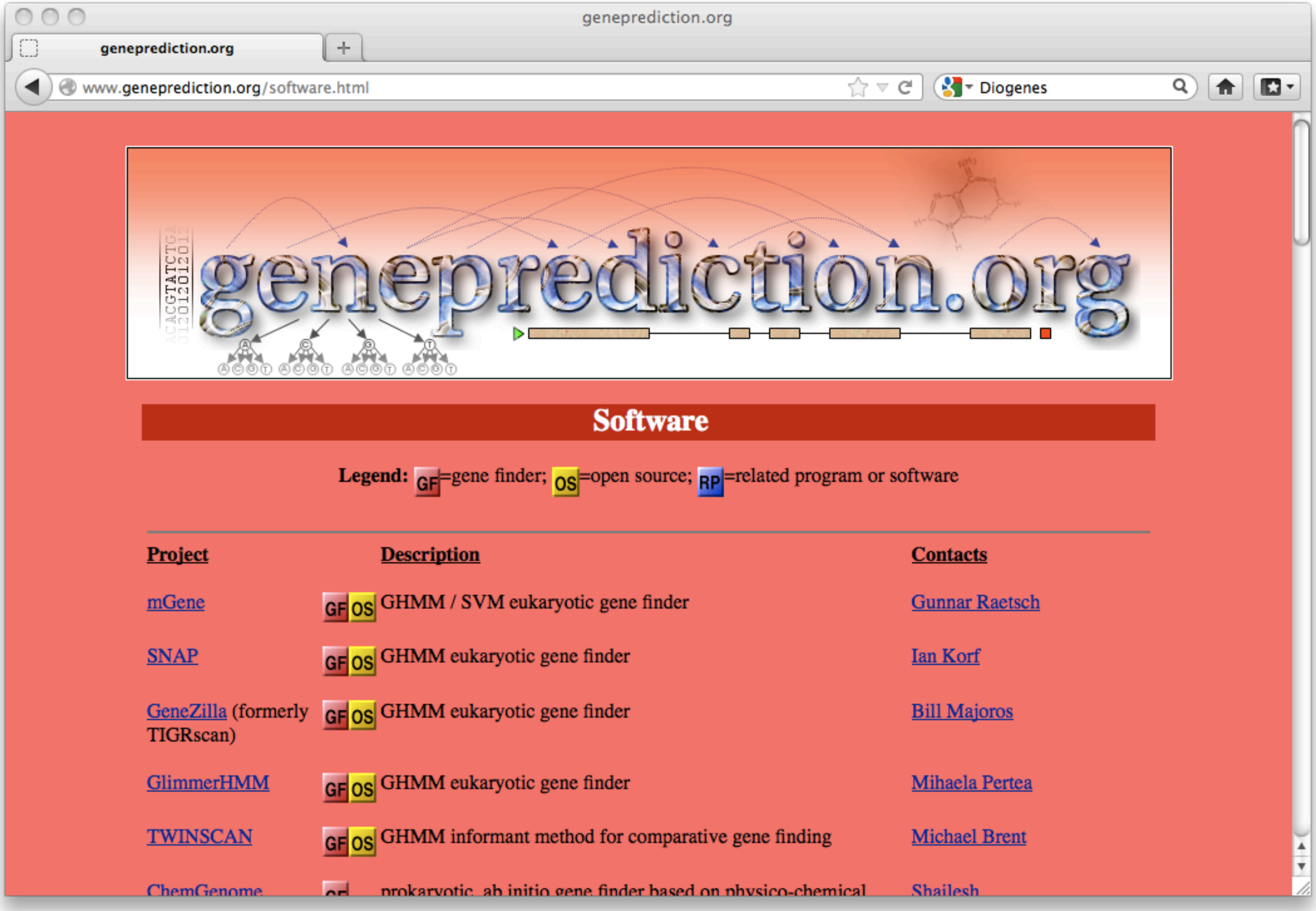
Report genes on: ● both strands ○ forward strand only ○ reverse strand only

Alternative transcripts: ○ none ● few ○ middle ○ many

Reset all input
Run AUGUSTUS

expert options

Even More Tools…



Software

Legend: GF=gene finder; OS=open source; RP=related program or software

Project		**Description**	**Contacts**
mGene	GF OS	GHMM / SVM eukaryotic gene finder	Gunnar Raetsch
SNAP	GF OS	GHMM eukaryotic gene finder	Ian Korf
GeneZilla (formerly TIGRscan)	GF OS	GHMM eukaryotic gene finder	Bill Majoros
GlimmerHMM	GF OS	GHMM eukaryotic gene finder	Mihaela Pertea
TWINSCAN	GF OS	GHMM informant method for comparative gene finding	Michael Brent
ChemGenome	GF	prokaryotic, ab initio gene finder based on physico-chemical	Shailesh

An active list of gene prediction programs (prok and euk)

Gene Finding with GenScan & Company

- Go to your preferred website
- Paste in the DNA sequence of your favorite EUKARYOTIC genome (this won’t work for prokaryotic genomes and it won’t necessarily work for viral or phage genomes)
- Press the submit button
- Output will typically be presented in a new screen or emailed to you

Outstanding Issues*

- Most Gene finders don’t handle UTRs (untranslated regions)
- ~40% of human genes have non-coding 1st exons (UTRs)
- Most gene finders don’t’ handle alternative splicing
- Most gene finders don’t handle overlapping or nested genes
- Most can’t find non-protein genes (tRNAs)

Bottom Line...

- Gene finding in eukaryotes is not yet a "solved" problem
- Accuracy of the best methods approaches 80% at the exon level (90% at the nucleotide level) in coding-rich regions (much lower for whole genomes)
- Gene predictions should always be verified by other means (cDNA sequencing, BLAST search, Mass spec.)
- *Homework: Try testing some of the web servers I have mentioned today*

How Many Genes in the Human Genome?

- **1969 – 2,000,000**
- **1999 – 100,000**
- **2000 - ~50 researchers placed bets and guessed between 27,462 to 153,478 genes**
- **2001 – 30-40,000**
- **2003 – 23,299 (ENSEMBL)**
- **2004 – 20-25,000**
- **2008 – 21,787 (Genome Consortium)**
- **2012 - 20,687 protein-coding genes determined by *in vitro* gene expression in multiple cell lines (not by computers)**