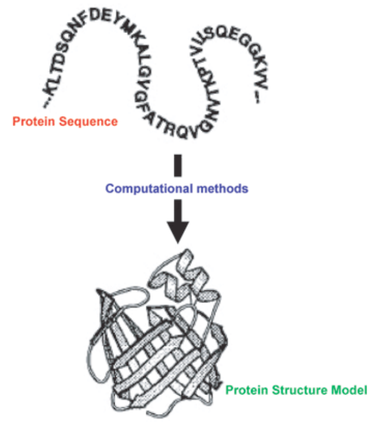
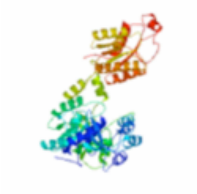
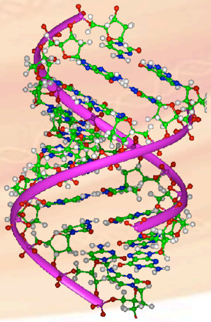




# Bioinformatics Computational Methods 1 - BIOL 6308



Million years
3.4
5.1
H aactggtacg tggtcg tacgaactgg H
C aaccggttcg cggtcg ttcgaaccgg C
G aacctgttcg ctgtcg ttcgaacctg G
O atctggtacg tggtcg tacgatctgg O
Million years
7.0
2.2

October 8th 2013

http://155.33.203.128/cleslin/home/teaching6308F2013.php

# Last Time

- Filters in UNIX
- Pattern matching in UNIX
- Regular Expressions in UNIX
- Regular Expressions in Perl
- Special Characters
- Interpolation
- Anchors
- Metacharacters
- Alternative Characters
- Retrieving what was Matched
- Greediness
- Parsing a single entry GenBank File

# Genomes

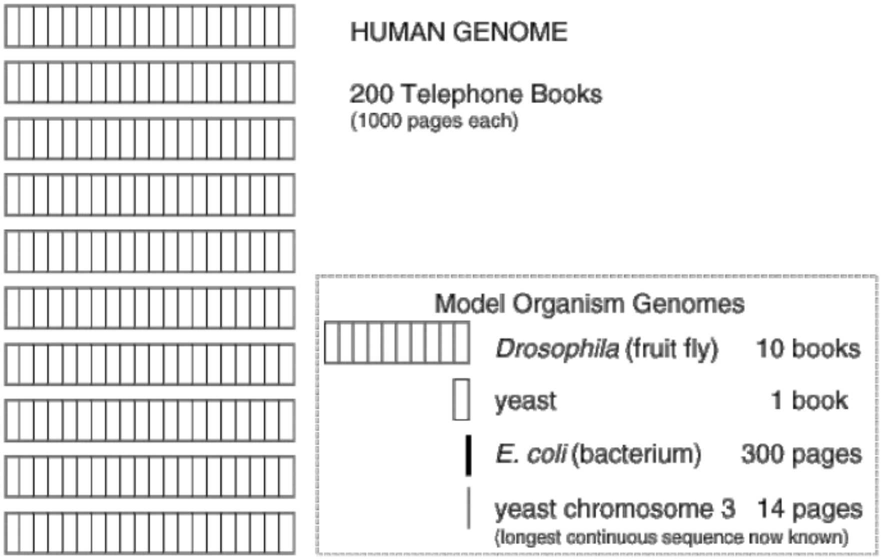
# Why Study Genomes?

- Understand biological processes
- Understand pathological processes
- Diagnose, prevent, and cure diseases

- Outline
  - **Genome Organization**
  - **Locating Genes in Genomes**
  - What About Other Genomic Features
  - And now the Genomes…

# Challenges of DNA

- DNA has only four letters
- Strung together with NO obvious "punctuation"
- No signals to say “a gene starts (or ends) here”
- How do we make sense of so much information?


HUMAN GENOME
200 Telephone Books
(1000 pages each)
Model Organism Genomes
Drosophila (fruit fly) 10 books
yeast 1 book
E. coli (bacterium) 300 pages
yeast chromosome 3 14 pages
(longest continuous sequence now known)

We'll discuss each, so remember this scale

# What Else is in the DNA? (Besides Genes)

- ~The Entire genome does note coded for proteins
- There are areas that: (non-coding)
  - **Do not** code for genes at all
  - **Regulate** the genes (transcriptional and translational regulation)
    - When should a protein be expressed?
      - night/day, fetus/adult
    - Where should a protein be expressed?
      - eye, lung, muscle, brain
  - transcribed into functional noncoding RNA

- More than 98% of the human genome does not encode protein sequences

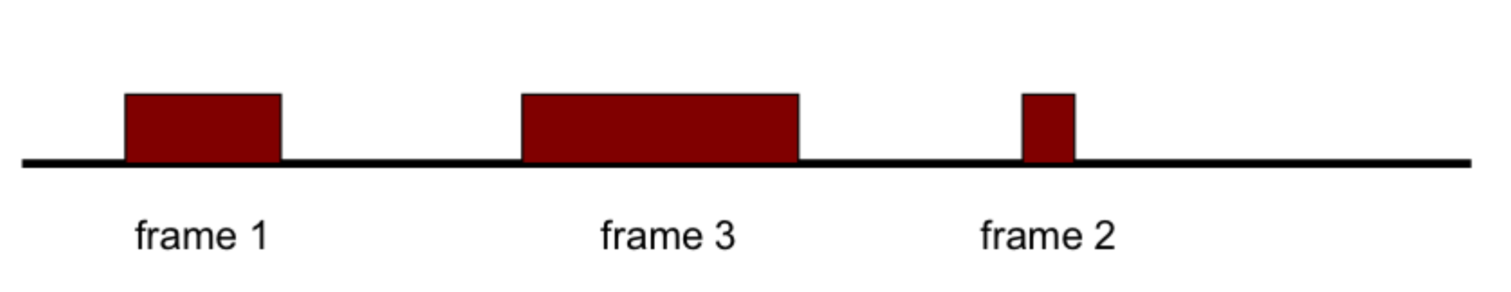
# Genome Organization

# What Characterizes Cells

- Different patterns of proteins also characterize different cells
- The relationship between DNA content and protein content is **not direct**
  - Really, any ideas?
    - Not all DNA codes for proteins
    - Conversely, some genes exist in multiple copies
  - Even more complicated in eukaryotes
    - Many genes produce several different proteins by **alternative splicing**
- Think of it this way:
  - **The amount of protein sequence information in a cell, much less the number and pattern of different proteins expressed, cannot easily be estimated from the genome size**
  - **Intuitive?**
    - **More like calculated**

# Simplified Gene Structure Rules

- Here's a simple framework to describe genes
  - Later, we'll see how this is oversimplified
  - But this is a start
- Each coding region (**exon**) has a **fixed translation frame** (no gaps allowed)
- **All exons** of a gene are on the **same strand**
- Neighboring exons of a gene can have **different reading frames**



# Genes in Bacteria

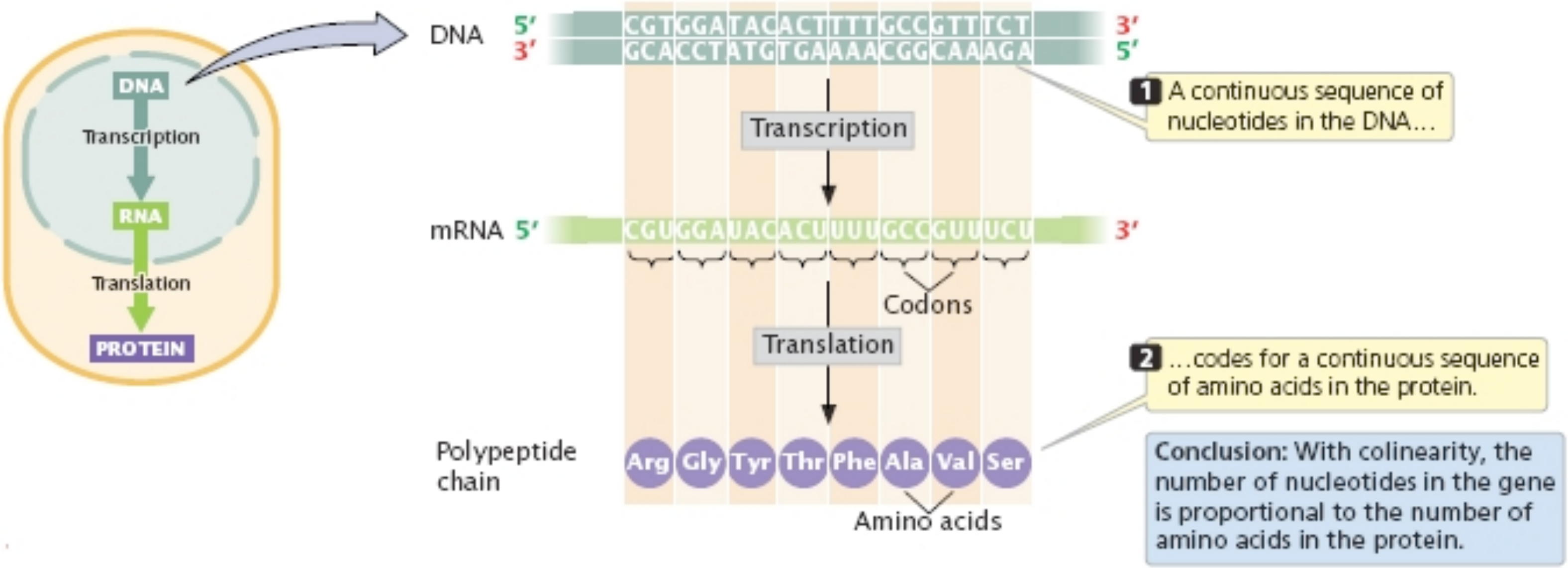
- A single gene coding for a particular protein:
  - Corresponds to a sequence of nucleotides along one or more regions of a molecule of DNA
  - The DNA sequence is **collinear** with the protein sequence
- **Double-stranded DNA -** genes may appear on **either** strand
- Bacterial genes are continuous regions of DNA**
  - Functional protein-coding unit is a string of 3N nucleotides encoding a string:
    - of N amino acids, or
    - of N nucleotides encoding a structural RNA molecule
- Such a string - with annotations – form typical entry in an archive of genetic sequences
  - **Think of the GenBank Flat Files**

****Bacterial Genomics: Unexpected introns in the 16S**
Nature Reviews Microbiology 10, 310 (May 2012) | doi:10.1038/nrmicro2786

# The Collinearity of Gene and Protein Structures

- Watson and Crick's **structure** for **DNA**
  - Together with Sanger's **demonstration** that **protein sequences were unique and specific**
  - Made it seem **likely** that **DNA** sequence specified **protein** sequence
- Yanofsky provided better evidence in 1964:
  - He showed that the relative distances between mutations in DNA were proportional to the distances between amino acid substitutions in E. coli tryptophan synthase

# Collinearity of Genes


DNA
Transcription
RNA
Translation
PROTEIN
DNA
5′
3′
CGT GGA TAC ACT TTT GCC GTT TCT
GCA CCT ATG TGA AAA CGG CAA AGA
3′
5′
Transcription
mRNA 5′
CGU GGA UAC ACU UUU GCC GUU UCU
3′
Codons
Translation
Polypeptide chain
Arg Gly Tyr Thr Phe Ala Val Ser
Amino acids

**1** A continuous sequence of nucleotides in the DNA…

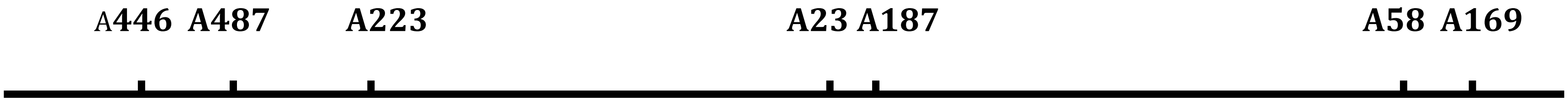
**2** …codes for a continuous sequence of amino acids in the protein.

**Conclusion:** With colinearity, the number of nucleotides in the gene is proportional to the number of amino acids in the protein.

# Collinearity

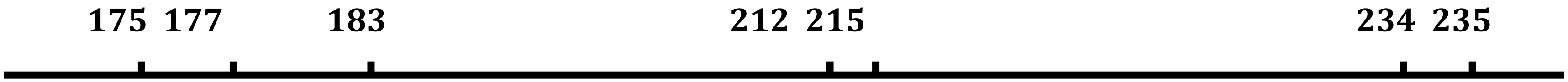
The tryptophan synthase (*trpA*) cistron

## Positions of mutant loci on genetic recombination map



A446 A487 A223 A23 A187 A58 A169

## Positions of altered amino acids in the protein chain



175 177 183 212 215 234 235

The genetic map and the amino acid sequence are collinear. The mutations in the gene and the changed amino acids in the protein appear in the same relative positions.

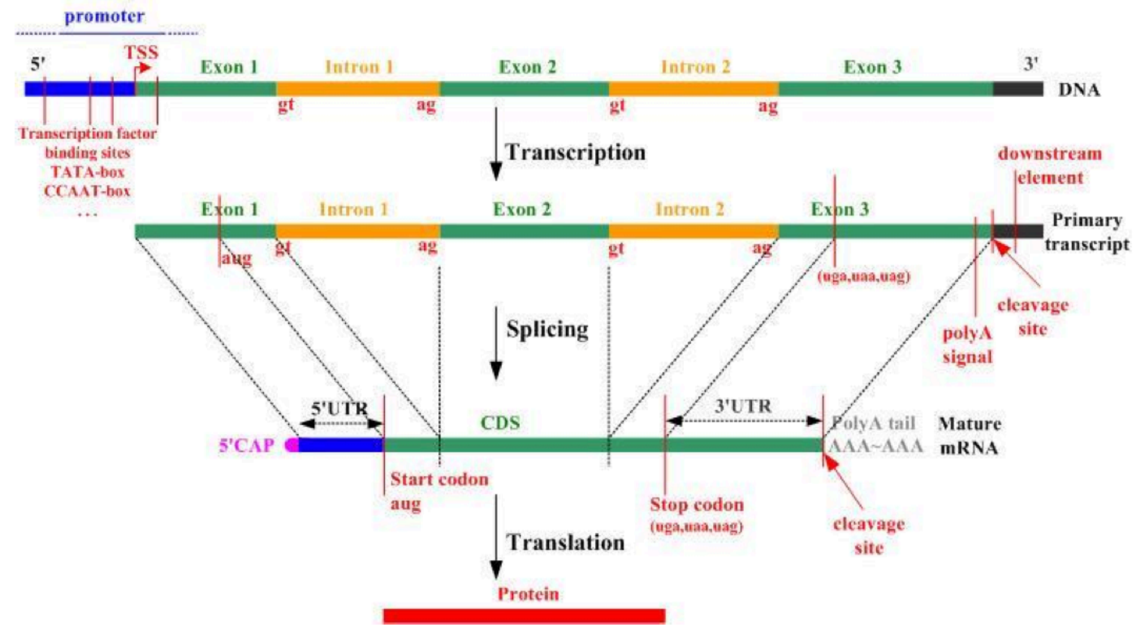
http://www.jbc.org/content/280/46/e43.full

# Genes - Eukaryotes

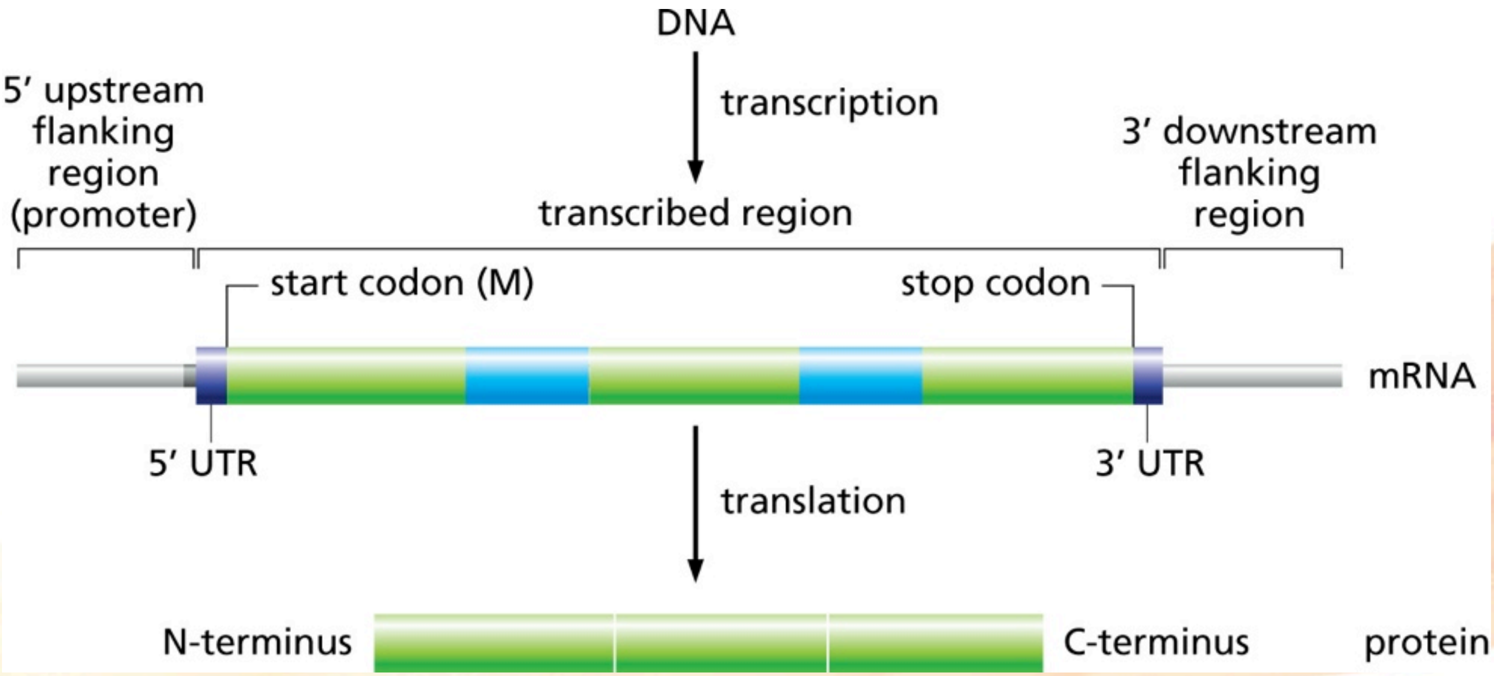
- nt sequences that encode a.a. sequences of individual proteins organized in a more **complex manner**
- Frequently genes appear split into separated segments in the genomic DNA
  - Exon:
    - stretch of DNA **retained** in the mature **mRNA** that the ribosome translates into protein
  - Intron
    - intervening region **between** two **exons**
- **Cellular machinery** splices together the segments of initial RNA transcripts, based on signal sequences flanking the exons in the sequences themselves
- Many introns are **very long**
  - Some cases **substantially** longer than the exon

# Genes - Eukaryotes

- Much more Complex!



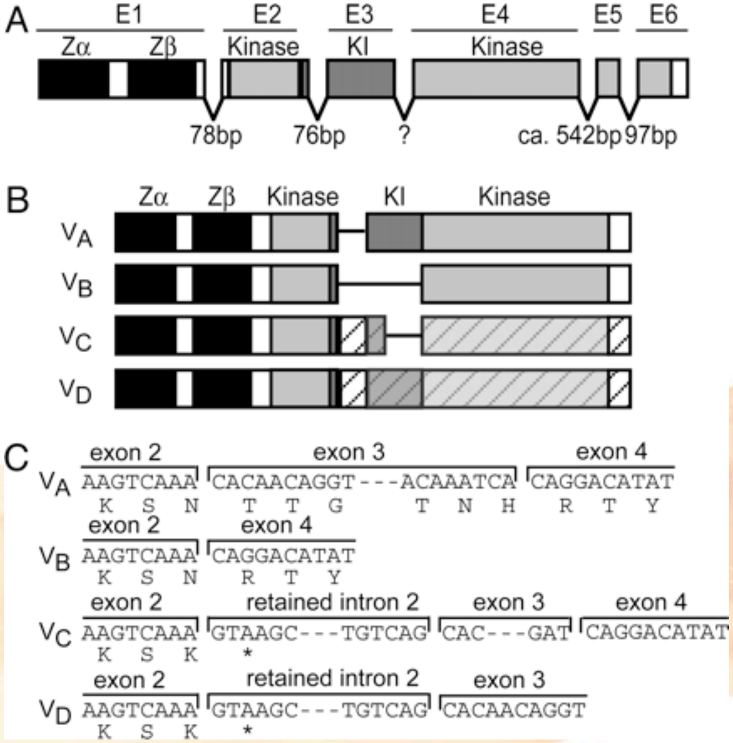
# Genes – mRNA Eukaryotes



Understanding Bioinformatics

# Splice Variants - Eukaryotes

The double-stranded RNA (dsRNA)-dependent protein kinase (PKR)



http://www.pnas.org/content/102/5/1602.full

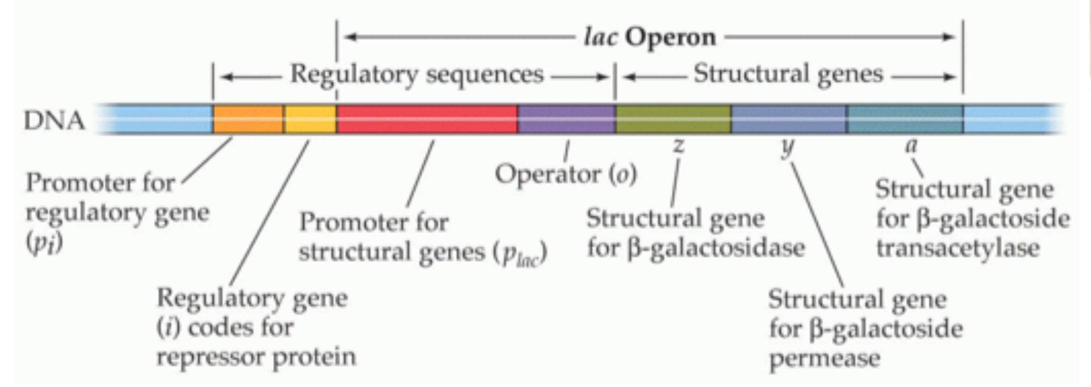
http://www.premierbiosoft.com/tech_notes/gene-splicing.html

# Control Regions - Prokaryotes

- Many control regions of DNA lie near the segments coding for proteins
- Contain signal sequences:
  - Serve as **binding sites** for the molecules that transcribe the DNA sequence, or sequences that bind regulatory molecules that **can block transcription**
    - **aka transcription factors**
- Bacterial genomes contain examples of **contiguous genes**:
  - Coding for several proteins that catalyze successive steps in an integrated sequence of reactions
    - All under the control of the same regulatory sequence
    - Known as **Operons**
  - One can readily understand the utility of a parallel control mechanism

# Operon

- Region of bacterial DNA that regulates gene expression in bacteria
- Consists 4 major parts: structural genes, regulatory gene, the promoter gene, and operator
- Structural genes code for enzymes needed:
  - In a chemical reaction
  - Transcribed at the same time to produce specific enzymes
- Regulatory codes for specific regulatory protein called repressor
  - Capable of attaching to the operator and blocking transcription
- Promoter where RNA polymerase binds to begin transcription
- Operator controls whether or not transcription will occur


*lac* Operon
Regulatory sequences
Structural genes
DNA
Promoter for regulatory gene ($p_i$)
Regulatory gene ($i$) codes for repressor protein
Promoter for structural genes ($p_{lac}$)
Operator ($o$)
$z$
Structural gene for β-galactosidase
$y$
Structural gene for β-galactoside permease
$a$
Structural gene for β-galactoside transacetylase

# Control Regions - Eukaryotes

- Eukaryotic chromosomes contain complexes of DNA with histones
- Chromatin remodelling is an important mechanism of transcriptional control
- Reversible chemical modification of histones occurs by variety of reactions including:
    - Deacetylation
    - Methylation
    - Decarboxylation
    - Phosphorylation
    - Ubiquitinylation
- Leads to alterations of the DNA-histone interactions that render transcription initiation sites more or less accessible

http://www.nature.com/scitable/topicpage/chromatin-remodeling-in-eukaryotes-1082

# Hereditary Information

- Just how hereditary information is **stored, passed on and implemented** is perhaps the **fundamental problem of biology**
- Three types of maps have been essential
  - 1. Linkage maps of genes
  - 2. Banding patterns of chromosomes
  - 3. **DNA sequences**
- Represent three **very different types of data**
- Genes, as discovered by Mendel, were entirely abstract entities
- Chromosomes are physical objects - banding patterns provide visible landmarks
- DNA sequences - we dealing directly with stored hereditary information in its **physical form**
  - So how do we analyze DNA?

# FYI – Gene Map

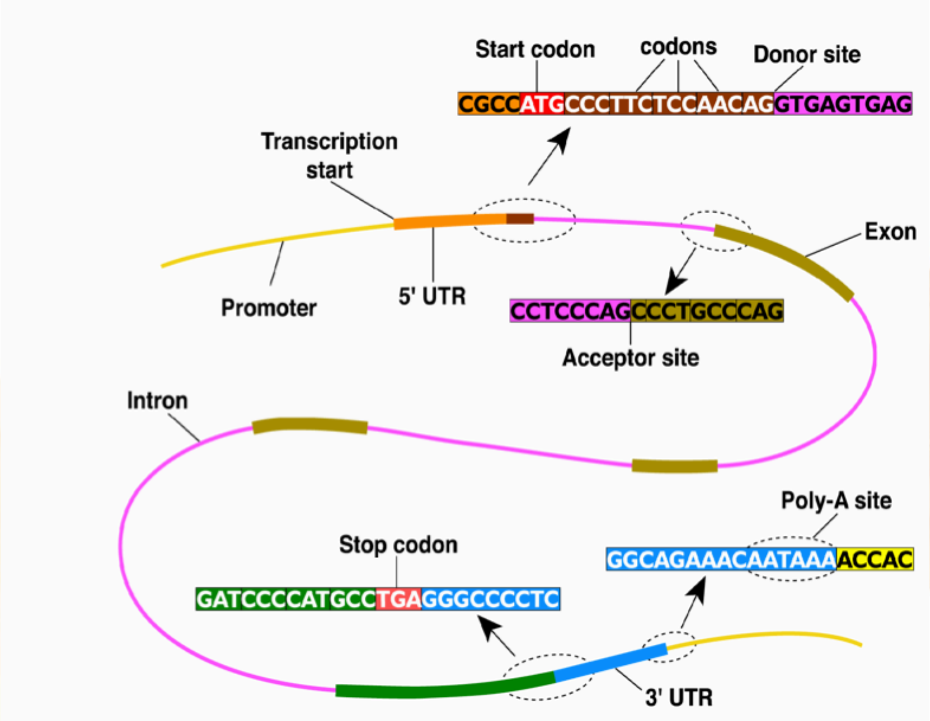
- A gene map is classically determined by observed patterns of heredity. Linkage groups and recombination frequencies can detect whether genes are on the same or different chromosomes, and, for genes on the same chromosome, how far apart they are
- The principle is that the farther apart two linked genes are, the more likely they are to recombine, by crossing over during meiosis. Indeed, two genes on the same chromosome but very far apart will appear to be unlinked
- The unit of length in a gene map is the Morgan, defined by the relationship that 1 cM corresponds to a 1% recombination frequency. (We now know that 1 cM ∼1 × 106 bp in humans, but it varies with the location in the genome and with the distance between genes.)
- Deletions of substantial segments of DNA are observable in changes in banding patterns. (Smaller deletions are observable by fluorescent in-situ hybridization, or FISH)
- The observation of banding patterns was crucial to the identification of chromosomes as the vessels of heredity

# FYI - Chromosome Banding Pattern Maps

- Chromosomes are physical objects. Banding patterns are visible features on them. The nomenclature is as follows: in many organisms, chromosomes are numbered in order of size, 1 being the largest. The two arms of chromosomes, separated by the centromere, are called the p (petite = short) arm and q (queue) arm. Regions within the chromosome are numbered p1, p2, . . . and q1, q2 . . . outward from the centromere
- Subsequent digits indicate subdivisions of bands. For example, certain bands on the q arm of human chromosome 15 are labelled:
  - 15q11.1, 15q11.2, 15q12
- Originally bands 15q11 and 15q12 were defined; subsequently 15q11 was divided into 15q11.1 and 15q11.2

# A Protein-Coding Gene

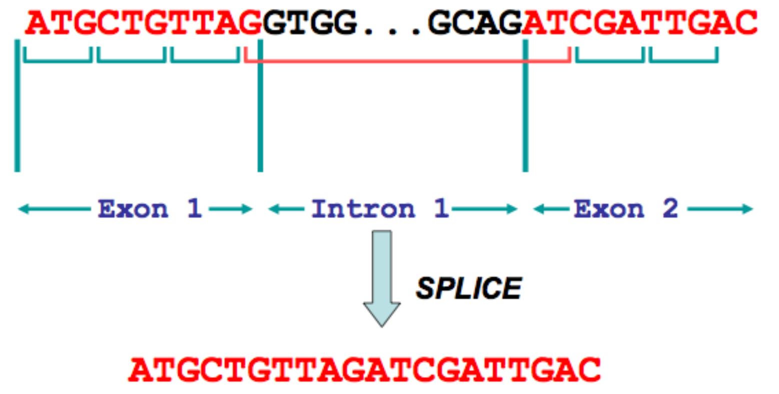
- We begin by understanding what codes for a protein



- Then look at the Exon and Introns

# Exon/Intron

- Exon/Intron structure detail



ATGCTGTTAGGTGG. . .GCAGATCGATTGAC

Exon 1

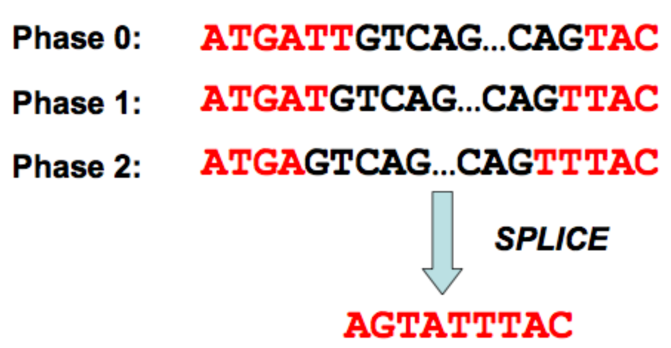
Intron 1

Exon 2

SPLICE

ATGCTGTTAGATCGATTGAC

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



Phase 0: ATGATTGTCAG...CAGTAC

Phase 1: ATGATGTCAG...CAGTTAC

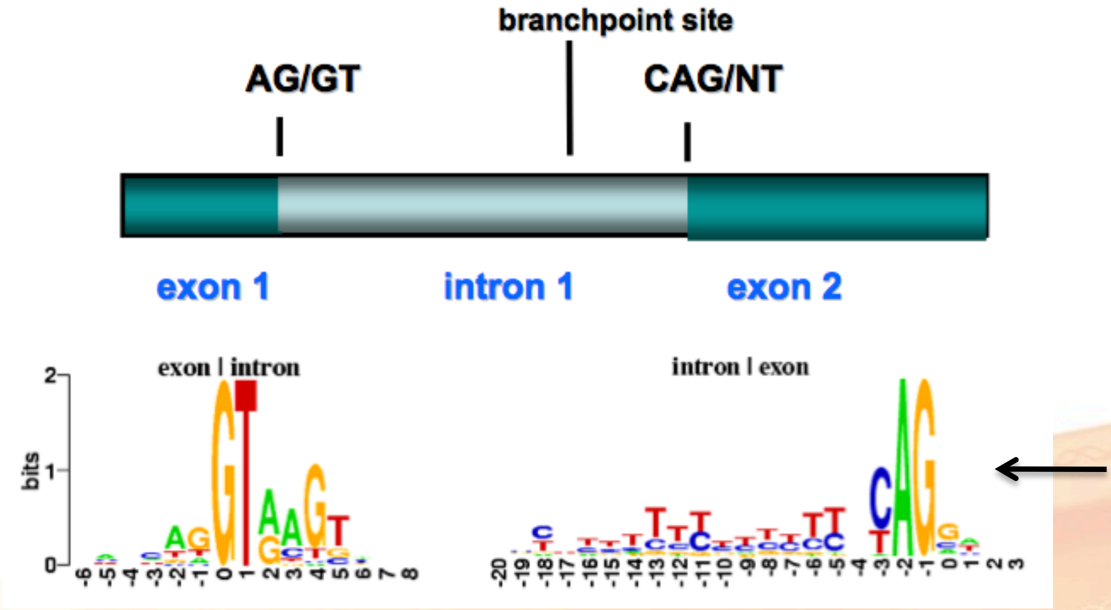
Phase 2: ATGAGTCAG...CAGTTTAC

SPLICE

AGTATTTAC

How do we know where these splices occur?

# Splice Signals in Detail



← What is this?

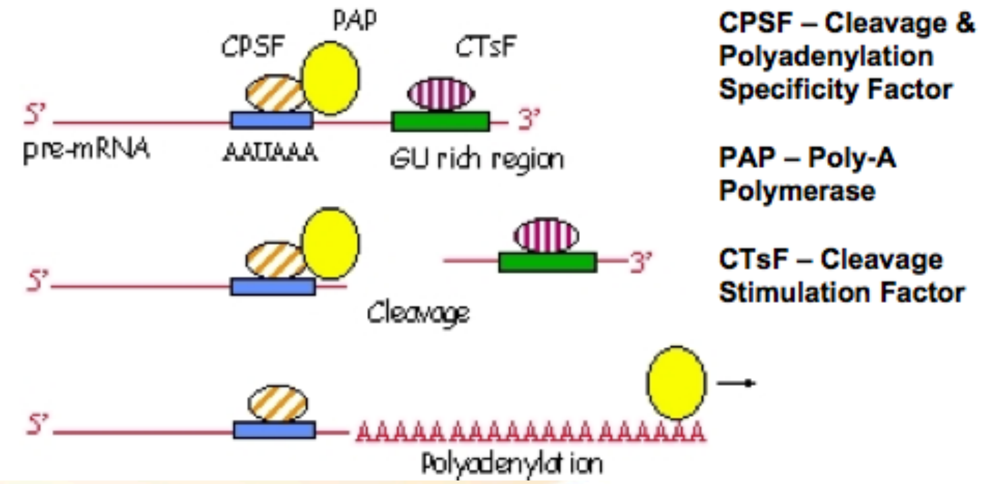
Information Content

We'll come back to this mathematical model..

- 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

# Miscellaneous Signals

- Polyadenylation signal
  - A A T A A A or A T T A A A–
  - Located 20 bp upstream of poly-A cleavage site
  - Cleavage and Polyadenylation of Eukaryotic proceeds as follows



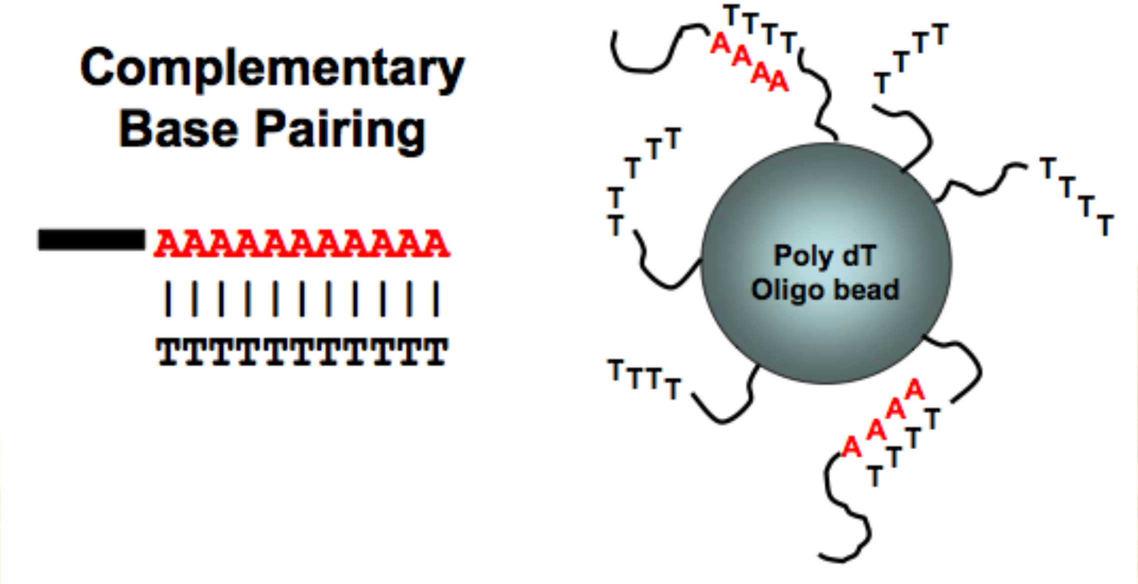
**CPSF – Cleavage & Polyadenylation Specificity Factor**

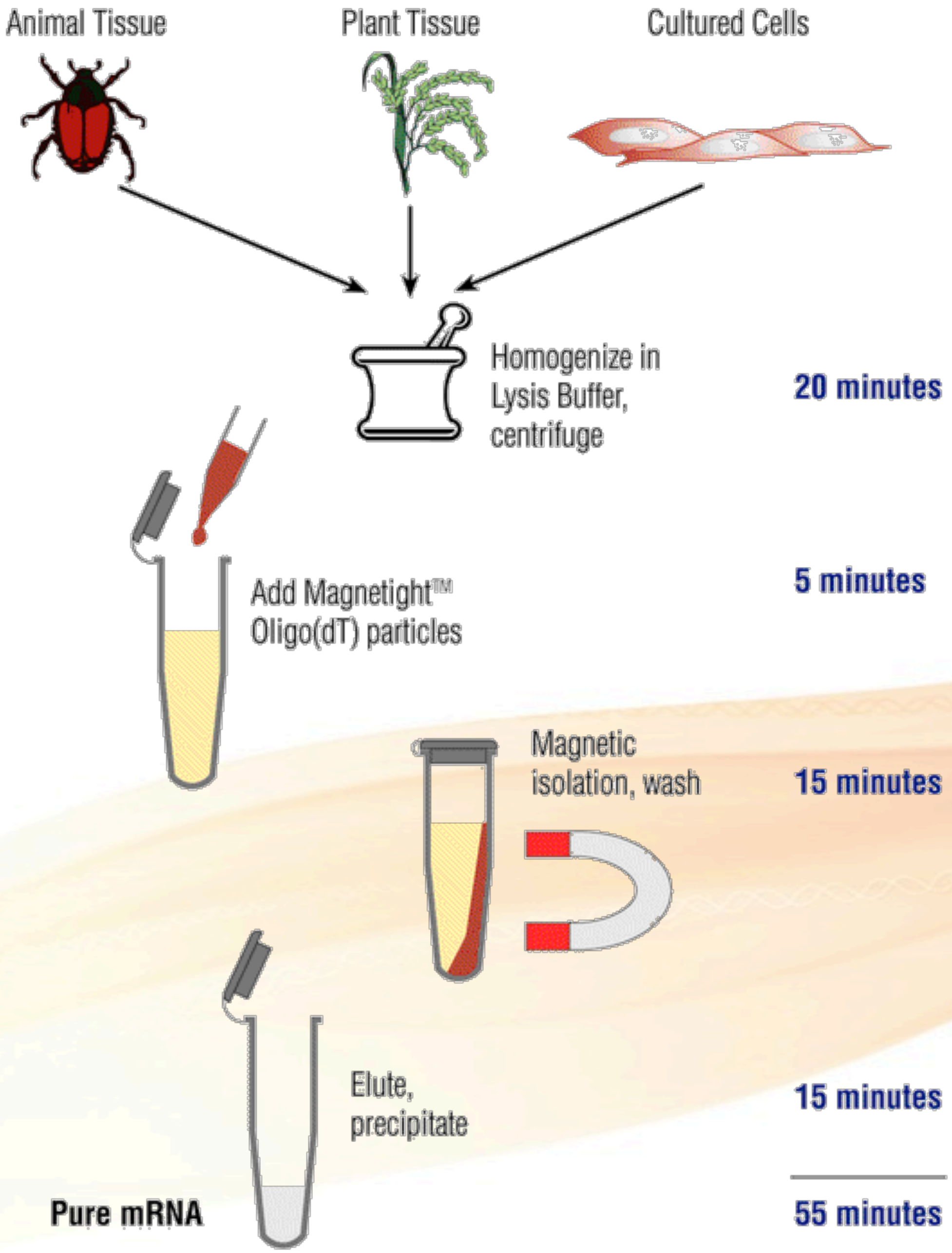
**PAP – Poly-A Polymerase**

**CTsF – Cleavage Stimulation Factor**

How Polyadenylation Useful?

# Polyadenylation

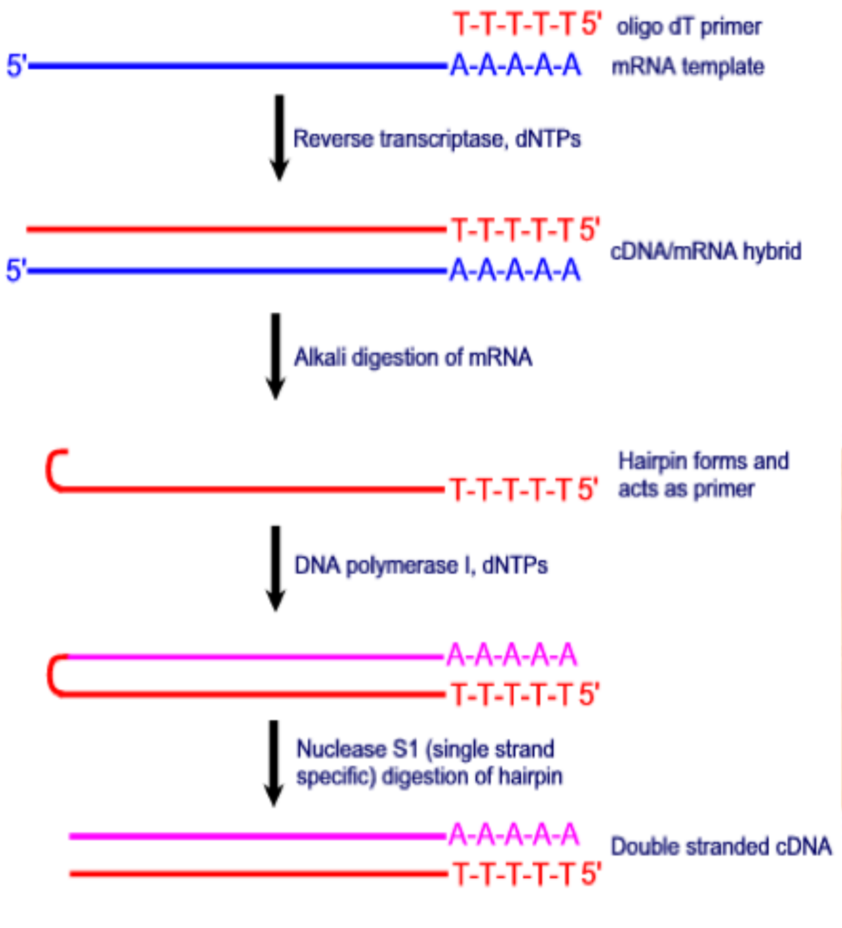




# mRNA Isolation

- 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
- Now have mRNA
- This has been a very useful molecular tool

# Making cDNA from mRNA

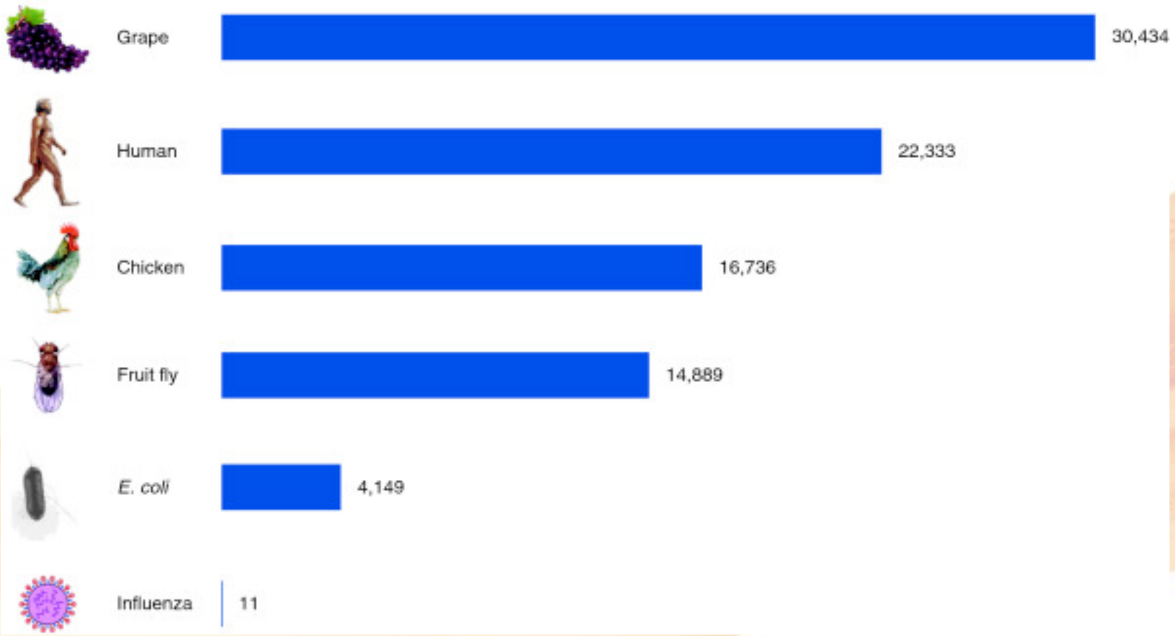


- Once you have mRNA, then what?
- cDNA (i.e. complementary DNA) is a single-stranded DNA segment whose sequence is complementary to that of messenger RNA (mRNA)
- Synthesized by reverse transcriptase

As a molecular tool,
why must we convert to cDNA?

# Locating Genes in Genomes

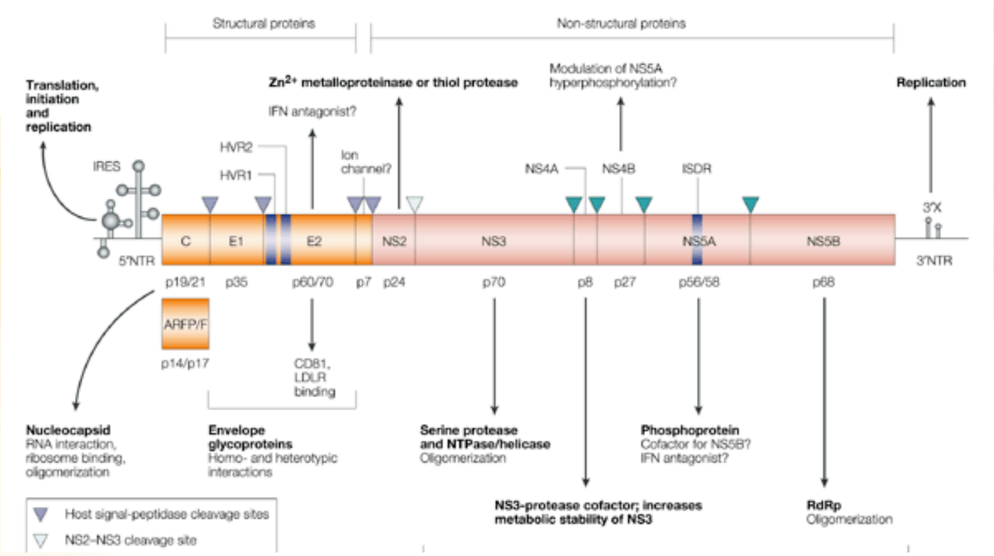
# Varying Gene Counts in Organisms



http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2898077/

# Locating Genes in Genomes

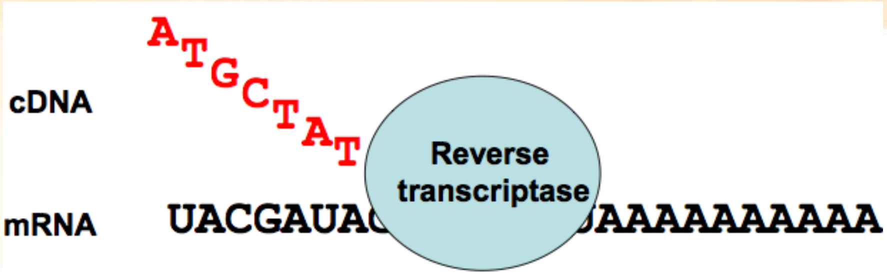
- Computer programs can pick out open reading frames (ORFs)
  - ORF is a potential protein-coding region: initiation codon (ATG) and ends with a stop codon
- Approaches to identifying genes combine three possible approaches:
  - Use coding regions from other organisms
  - Identify genes from the properties of the DNA sequences themselves
  - Comparative genome approach



The boxed area corresponds to the single open reading frame of the hepatitis C virus (HCV) genome

# Detection Genes 1$^{st}$ Method

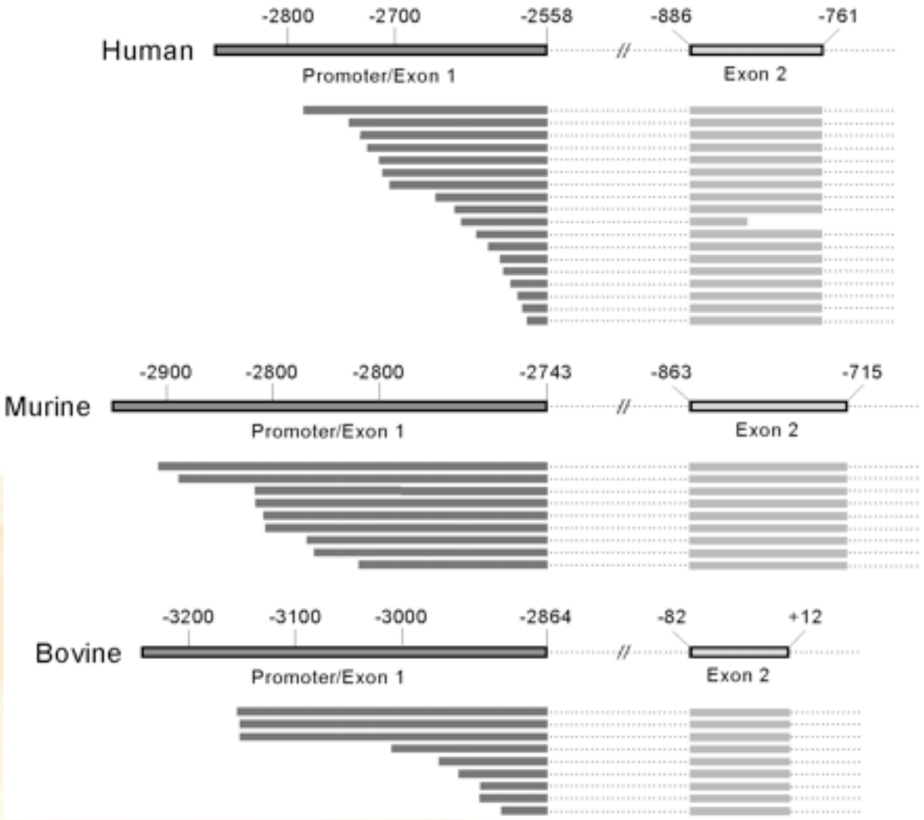
- Detection of regions similar to known coding regions from other organisms found in a database
- These regions may encode amino acid sequences similar to known proteins, or may be similar to ESTs
- Because ESTs are derived from mRNA, they correspond to genes known to be transcribed
  - ESTs
    - **Convert the spliced mRNA into cDNA**
    - Only expressed genes or expressed sequence tags (EST's) are seen
    - **Used to** save on sequencing effort (97%)



# Detection Genes 1$^{st}$ Method

- These approaches are able to find biologically relevant genes
  - Sounds great, but what's the issue?
  - Cannot identify genes that code for proteins not found in DB
- Use BLAST (Basic Local Alignment Search Tool) in this category
  - Do a BLASTx search of all 6 reading frames against known proteins in GenBank
    - What other issue do we have here?
    - Assumes organism under study has genes homologous to known genes
  - BLAST against EST database (finds possible or probable 3’ end of cDNAs)

# BLAST Against a EST Database - Genes



http://www.iovs.org/content/48/8/3872.full

# ESTs in 2010 on… RNA-Seq

- For many years:
  - ESTs - The standard method for determining the sequence of transcribed genes
  - Or full-length complementary DNA (cDNA) sequences using conventional Sanger sequencing technology
- Recently a new experimental method, RNA-Seq, has emerged:
  - Number of advantages over conventional EST sequencing
    - Uses next-generation sequencing (NGS) technologies that can sample the mRNA with fewer biases
    - It generates far more data per experiment that can be used as a direct measure of the level of gene expression

# Major Drawback RNA-Seq

- Major drawback of RNA-Seq over conventional EST sequencing:
  - The sequences themselves are much shorter
  - Typically 25–50 nt versus several hundred nucleotides with older technologies
- Critical step in an RNA-Seq experiment
  - Mapping the NGS ‘reads’ to the reference transcriptome
  - What's the problem?
  - Transcriptomes are incomplete:
    - even for well-studied species such as human and mouse
    - RNA-Seq analyses are forced to map to the reference genome as a proxy for the transcriptome
- We'll Come back to NGS

# Detection Genes 2nd Method

- *Ab initio* methods
  - Seek to identify genes **from the properties of the DNA sequences** themselves
- Computer-assisted annotation of genomes is more complete and accurate for bacteria than for eukaryotes
- Bacterial genes are relatively easy to identify because?
  - Are contiguous
  - Lack introns characteristic of eukaryotic genomes
  - Intergene spaces are small
- In higher organisms, identifying genes is harder
  - Identification of exons is one problem, assembling them is another
  - Alternative splicing patterns present a particular difficulty

# *ab initio* Gene Discovery

- Protein-coding genes have recognizable features
- Software can be designed to scan genome and identify these features
- Some of these programs work well:
  - Especially in bacteria
  - Simpler eukaryotes with smaller and more compact genomes
- Much harder for higher eukaryotes, where:
  - There are long introns
  - Genes can be found within introns of other genes, etc.
- We tend to do **OK** finding protein coding regions, but miss a lot of non-coding 5’ exons

# *ab initio* Gene Discovery

- Validating predictions and refining gene models
- Standard types of evidence for validation of predictions include:
  - Match to previously annotated cDNA
  - Match to EST from same organism
  - Similarity of nucleotide or conceptually translated protein sequence to sequences in GenBank
    - Translation works better—why?
- Protein structure prediction match to a PFAM domain
- Associated with recognized promoter sequences, ie TATA box, CpG island
- Known phenotype from mutation of the locus
  - Expensive endeavor

# *ab initio* Gene Discovery—Approaches

- Most gene-discovery programs use of some form machine learning algorithm
- Machine learning requires a training set of input data that computer uses to "learn" how to find a pattern
- Two common machine learning approaches used in gene discovery (and many other bioinformatics applications) are:
  - Artificial neural networks (ANNs)
  - Hidden Markov models (HMMs)

- A gene finding problem can be decomposed into two problems
  - Identification of **coding potential of a region** in a particular frame
  - Identification of **boundaries between coding** and **non-coding regions**

# Framework for *ab initio* Gene Identification in Eukaryotic Genomes

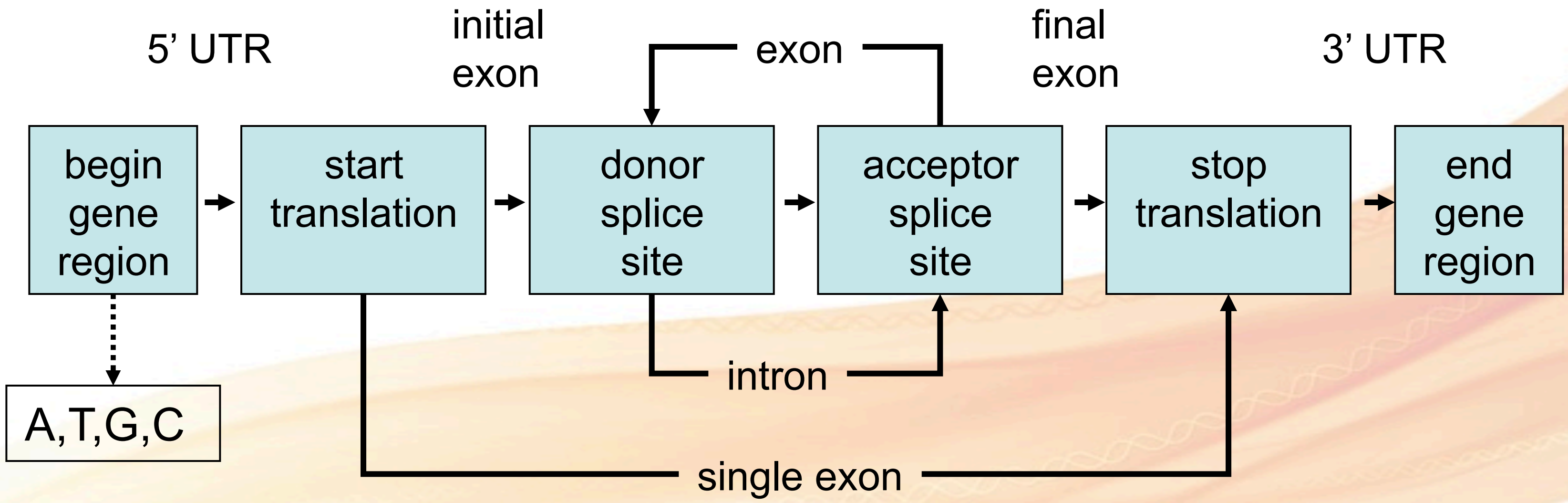
- Initial ( 5') exon starts with transcription start point preceded by a core **promotor site** such as the TATA box typically ~30 bp upstream
  - It is free of in-frame stop codons
  - Ends immediately before a dinucleotide GT splice signal
  - Occasionally a non-coding exon precedes the exon that contains the initiator codon
- Internal exons are free of in-frame stop codons
  - They begin immediately after an AG splice signal
  - End immediately before a GT splice signal
- Final (3') exon starts immediately after an AG splice signal
  - Ends with a stop codon
  - Followed by a polyadenylation signal sequence
  - Occasionally a non-coding exon follows the exon that contains the stop codon

# GENESCAN

- Designed to predict:
  - Complete gene structures in genomic sequences
  - Including exons, introns, promoter and poly-adenylation signals
- Differs from the majority of existing gene finding algorithms:
  - Allows for **partial** genes as well as complete genes
  - Occurrence of **multiple genes** in a single sequence
    - On **either** or **both** DNA strands
- Based on a **probabilistic model** of gene structure compositional properties
- **Does not** make use of protein sequence homology information
  - Uses GHMM
- Output of program is list of one or more (or possibly zero) predicted genes together with the corresponding peptide sequences

# *ab initio* Gene Discovery— GHMMs

An example state diagram for an GHMM for gene discovery is this simplified version of one used by *Genescan*:



Each box and arrow has associated *transition probabilities*, and *emission probabilities* for emission of nucleotides (dotted arrow). These are *learned* from examples of known gene models and provide the probability that a stretch of sequence is a gene

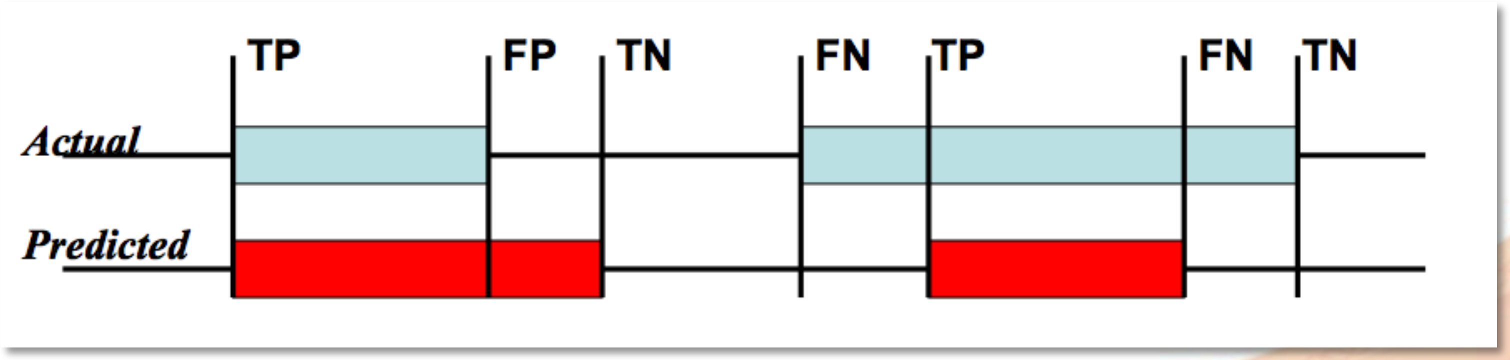
adapted from Gibson and Muse, *A Primer of Genome Science*

There is a great deal to learn before we thoroughly jump into HMM, so sit tight

# Evaluation of Methods

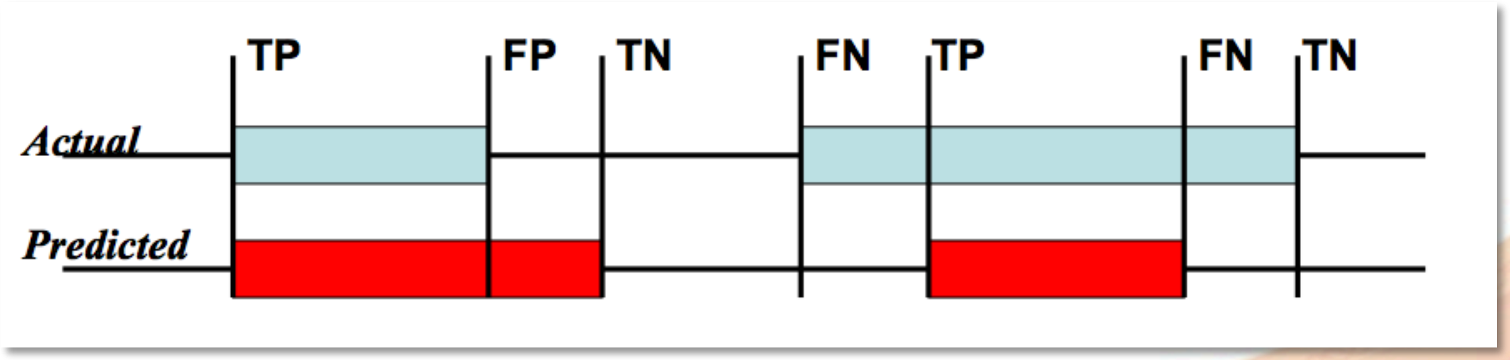
- Sensitivity (Sn) and specificity (Sp) are two common measures used for bioinformatics tools
  - Used:
    - for gene finding tools
    - any many other programs and diagnostics
  - Important to understand Sn and Sp since used throughout bioinformatics papers and industry
- Performance can be evaluated at nucleotide level and at exon level
  - Accuracy:
    - nucleotide level reflects how close predicted sequence and real coding sequence are in an alignment
    - exon level reflects how well signals (e.g., start codons, stop codons, and spice sites) are identified

# Accuracy at Nucleotide Level



**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

# Accuracy at Nucleotide Level

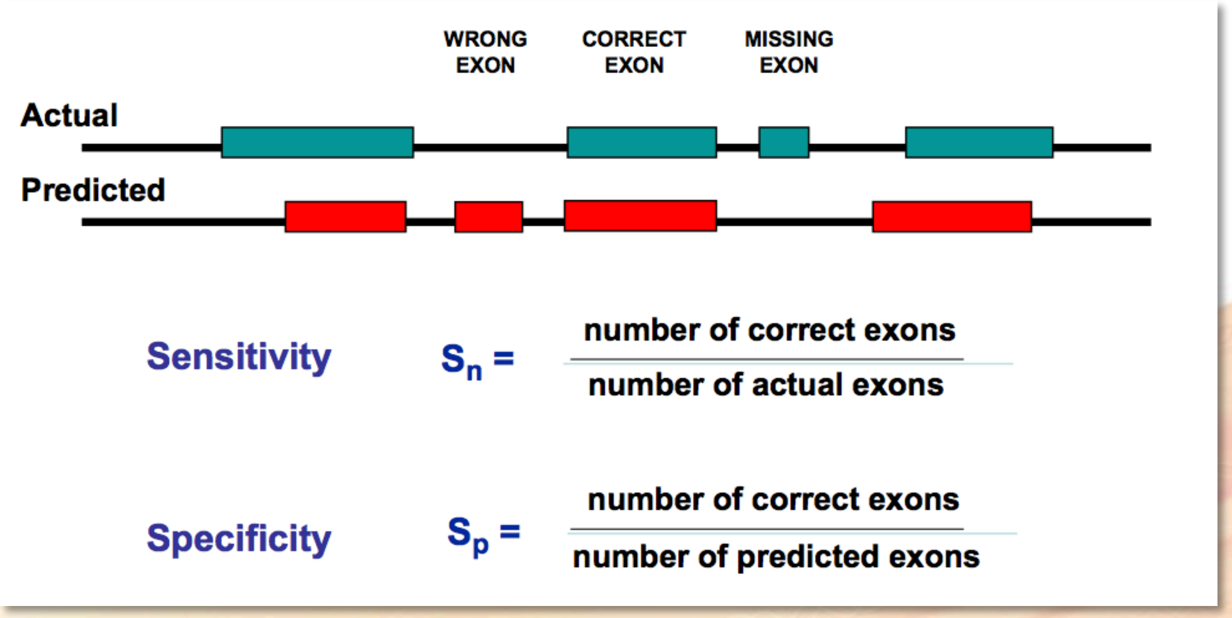


**Sensitivity or Recall** *Sn=TP/(TP + FN)*
**Specificity** *Sp=TN/(TN + FP)*
**Precision** *Pr=TP/(TP + FP)*
***Correlation*** *CC=(TP*TN-FP*FN)/[(TP+FP)(TN+FN)(TP+FN)(TN+FP)]*$^{0.5}$

# Different Strokes for Different Folks

- **Precision** and **specificity** statistics **favor conservative predic**tors 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
- As a side note:
  - Information retrieval papers report **precision** and **recall**
  - Bioinformatics papers tend to report **specificity** and **sensitivity**

# Accuracy at Exon Level



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

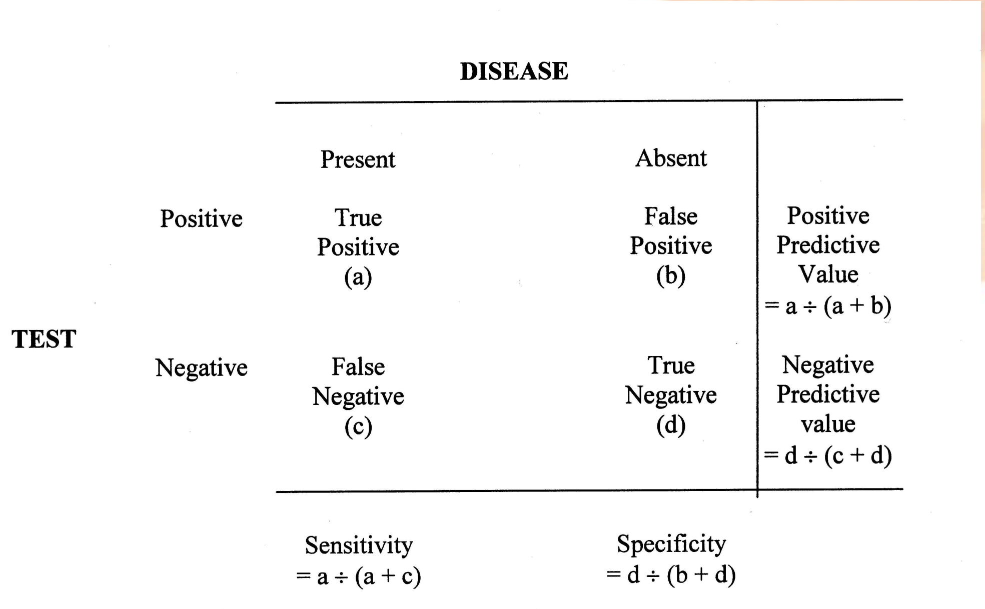
Specificity $$S_p = \frac{\text{number of correct exons}}{\text{number of predicted exons}}$$

Sn and Sp .. segue time!

# More on Sensitivity and Specificity and How They are Utilized

# Diagnostic Tests

- Sensitivity (Sn)
- Specificity (Sp)
- **Positive** and **Negative** Predictive Values of tests
- Used in Industry and Academics everywhere



| | | DISEASE | | |
|---|---|---|---|---|
| | | Present | Absent | |
| TEST | Positive | True Positive (a) | False Positive (b) | Positive Predictive Value = a ÷ (a + b) |
| | Negative | False Negative (c) | True Negative (d) | Negative Predictive value = d ÷ (c + d) |
| | | Sensitivity = a ÷ (a + c) | Specificity = d ÷ (b + d) | |

Adopted from J. Ebrahim

# Classifiers

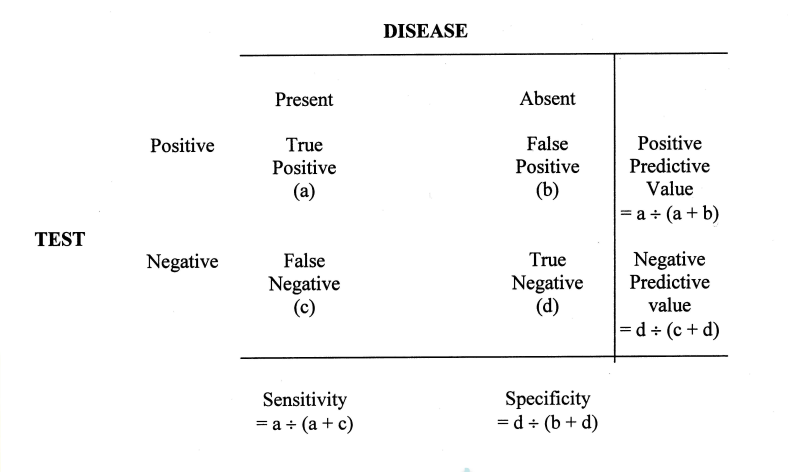
- A classifier assigns an object to one of a predefined set of categories or classes
- Examples:
  - Disease present of absent
  - Section of DNA either exon or not
  - Variant is either present or absent
  - Gene Either:
    - Lost or Not
    - Gained or Not
    - Amplified or Not
  - Two proteins are structurally similar or not
  - Etc.
- This talk: only two classes, “positive” and “negative”

# Two Types of Errors

- False Postive ("false alarm"), FP
  - Exon is called but is not truly coding
- False Negative ("miss"), FN
  - Exon is not called, but is there in the sequence

# Confusion Matrix

- **confusion matrix** is a specific table layout that allows visualization of the performance of an algorithm
  - *false positives*, *false negatives*, *true positives*, and *true negatives*



| | | DISEASE | | |
|---|---|---|---|---|
| | | Present | Absent | |
| TEST | Positive | True Positive (a) | False Positive (b) | Positive Predictive Value = a ÷ (a + b) |
| | Negative | False Negative (c) | True Negative (d) | Negative Predictive value = d ÷ (c + d) |
| | | Sensitivity = a ÷ (a + c) | Specificity = d ÷ (b + d) | |

| True class | Predicted class | |
|---|---|---|
| | positive | negative |
| positive (#P) | TP | FN (#P - #TP) |
| negative (#N) | FP | TN (#N - #FP) |

# Example: 3 classifiers

|  | Predicted |  |
|---|---|---|
| True | pos | neg |
| pos | 40 | 60 |
| neg | 30 | 70 |

|  | Predicted |  |
|---|---|---|
| True | pos | neg |
| pos | 70 | 30 |
| neg | 50 | 50 |

|  | Predicted |  |
|---|---|---|
| True | pos | neg |
| pos | 60 | 40 |
| neg | 20 | 80 |

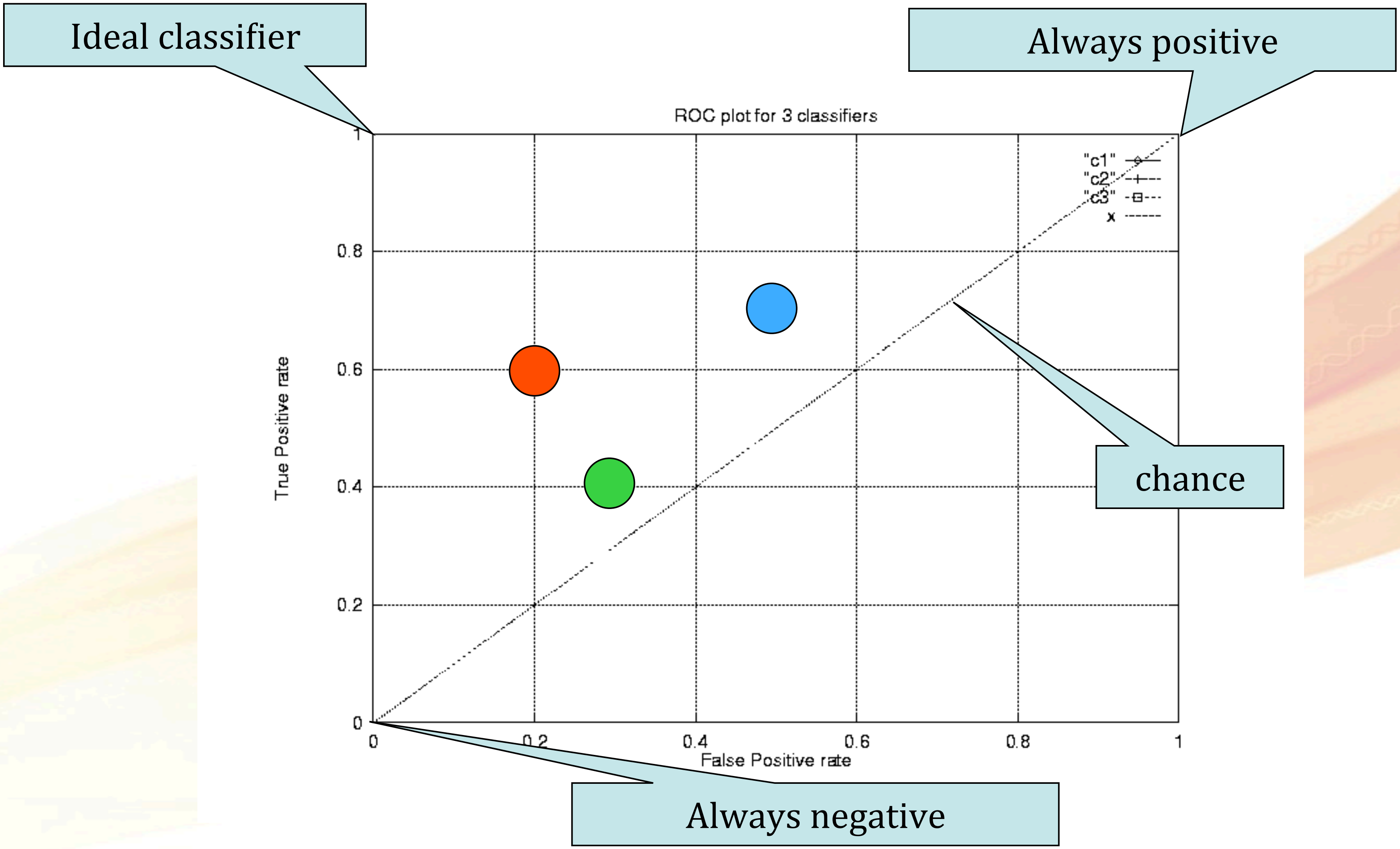
Classifier 1
TPR = 0.4
FPR = 0.3

Classifier 2
TPR = 0.7
FPR = 0.5

Classifier 3
TPR = 0.6
FPR = 0.2

Adopted from Robert Holte

# ROC Plot for the 3 Classifiers



Ideal classifier

Always positive

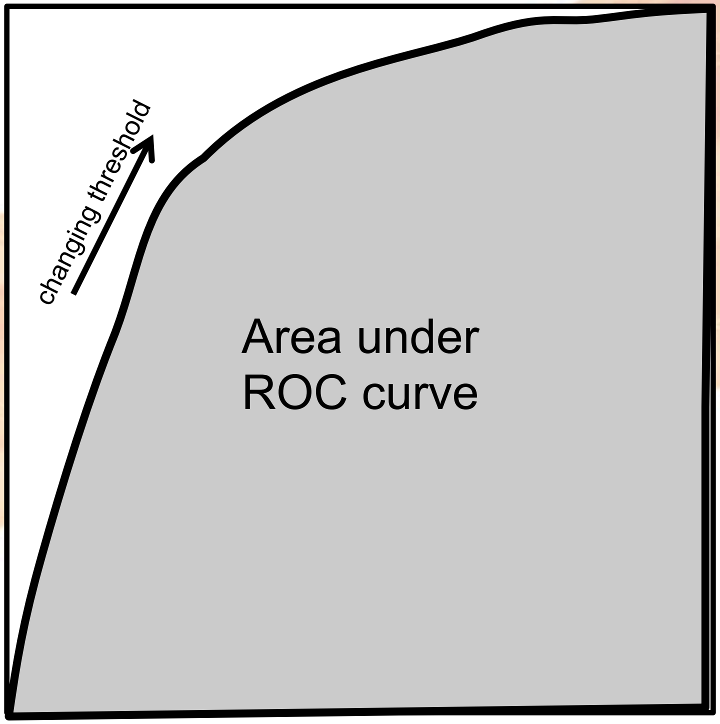
chance

Always negative

Robert Holte

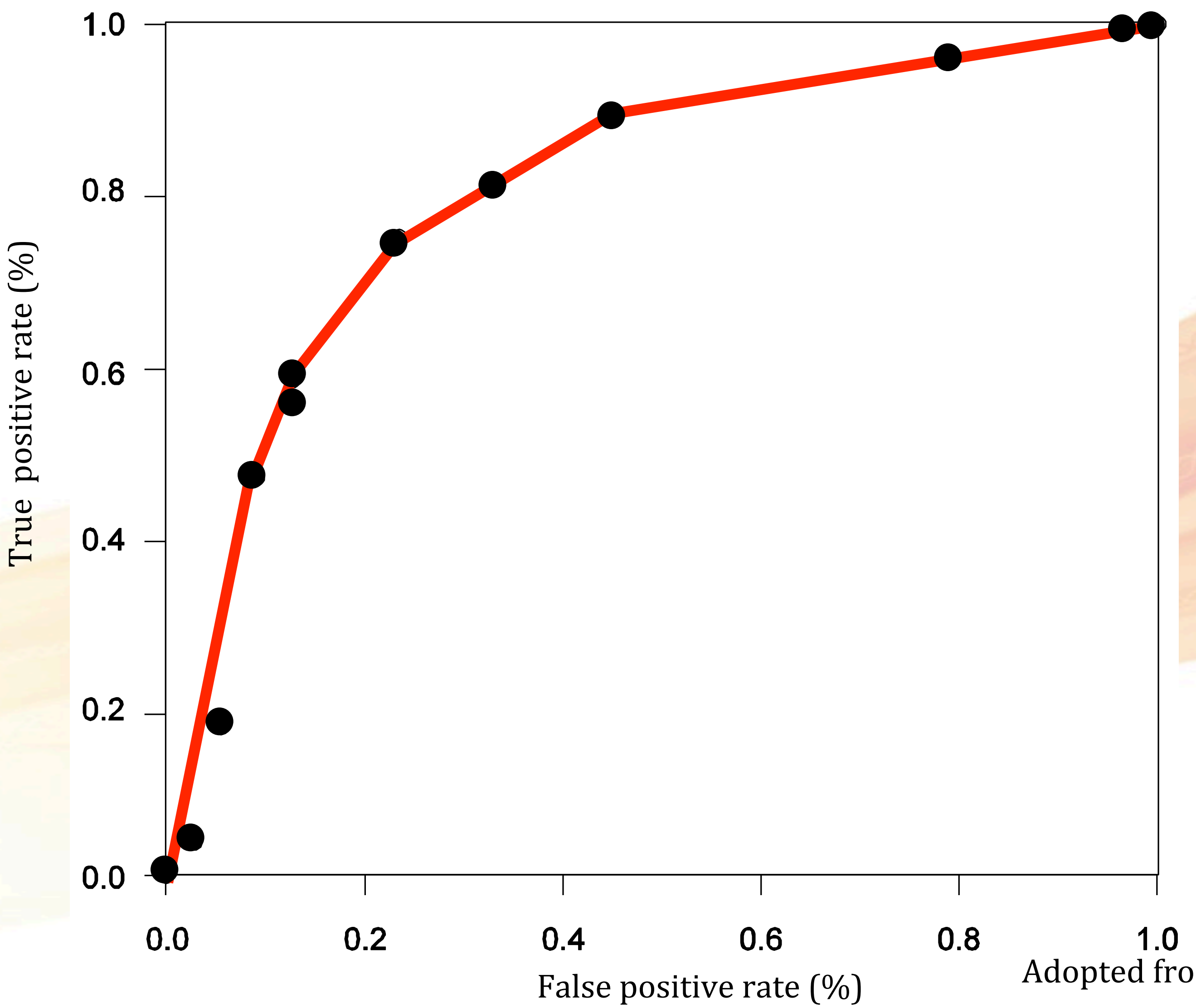
# Creating an ROC Curve

- A **classifier produces a single ROC point**
- If the classifier has a "**sensitivity**" parameter, varying it produces a series of ROC points (confusion matrices)
  - In Bioinformatics, this is usually a threshold
  - Changing the threshold, changes the plot point
  - Example Thresholds
    - Blast score
    - Log2 ratio
    - LOD
    - etc.



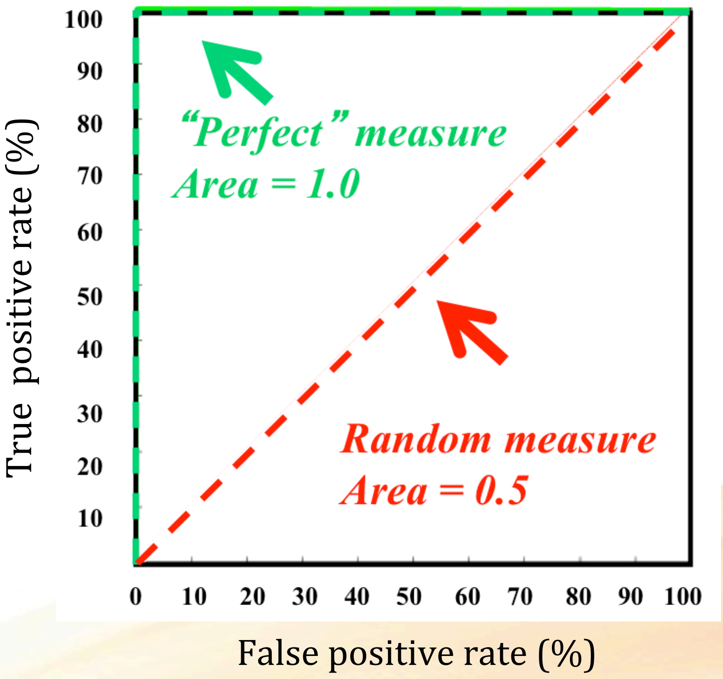
Adopted from Robert Holte

# More Points Plotted - Forms a Curve

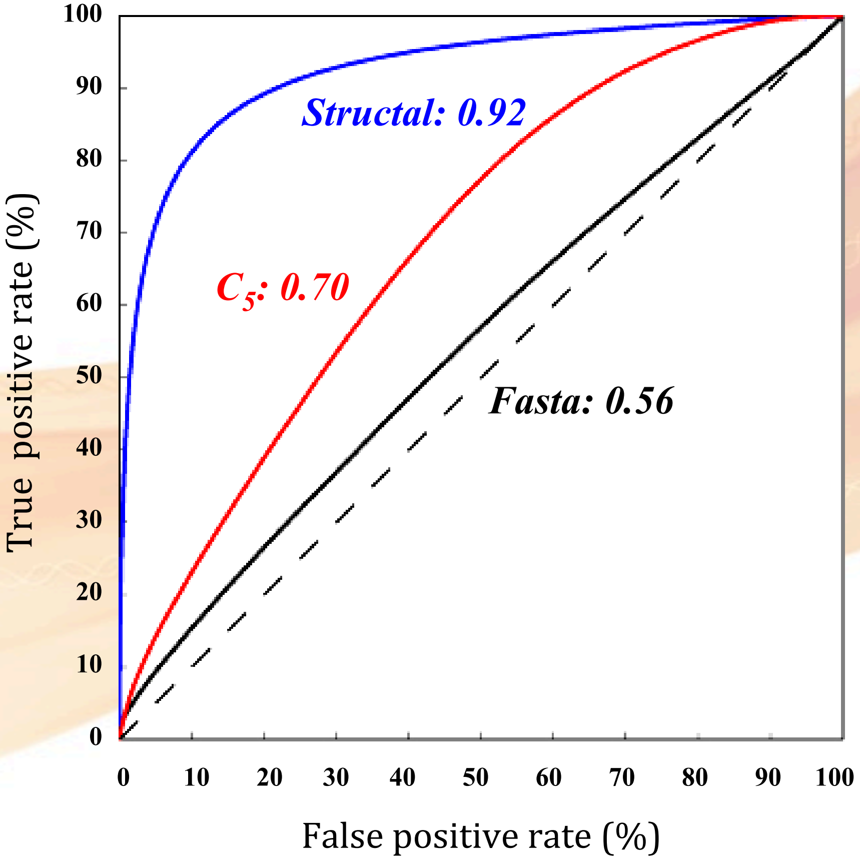


Adopted from Robert Holte

# The Meaning of a ROC Curve and an Example



Just and Example used to determine Structure conservation



**Y axis** is Seen = Sensitivity & True Positives

**X axis** is Seen = False Positive Rate **or**
**1 - Specificity**

Adopted from Dennis R. Livesay

# Example Ssed to Determine Structure Conservation

- For each pair P of proteins in dataset, perform alignment and record score: S(P)
- Rank all pairs according to their scores, from highest to lowest
- Scan ranked pairs, and record rate of true positives and true negatives for various thresholds

Classifier threshold

# Thresholds ROC curve

| Prediction | True Call |
|---|---|
| 1.00 | Yes |
| 0.99 | Yes |
| 0.98 | Yes |
| 0.97 | Yes |
| 0.96 | No |
| 0.95 | No |
| 0.93 | Yes |
| 0.91 | Yes |
| 0.89 | No |
| 0.87 | No |
| 0.85 | No |
| 0.83 | No |
| 0.83 | Yes |
| 0.81 | No |
| 0.77 | No |
| 0.74 | No |
| 0.73 | No |
| 0.70 | No |
| 0.69 | No |
| 0.67 | Yes |
| 0.62 | No |
| 0.56 | No |
| 0.54 | No |
| 0.53 | No |

$$\text{Sensitivity} = TPR = TP/\left(TP+FN\right) \qquad \text{1-Specificity} = FPR = FP/\left(FP+TN\right)$$

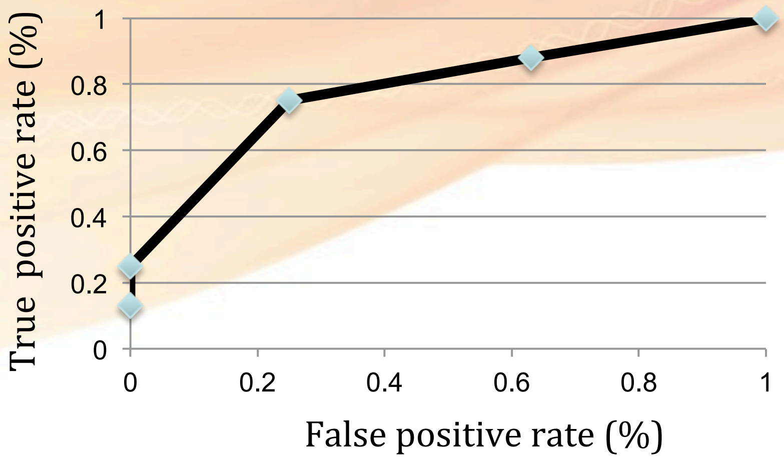
$$=1/\left(1+7\right)=0.13 \qquad =0/\left(0+16\right)=0.00$$

$$=2/\left(2+6\right)=0.25 \qquad =0/\left(0+16\right)=0.00$$

$$=6/\left(6+2\right)=0.75 \qquad =4/\left(4+12\right)=0.25$$

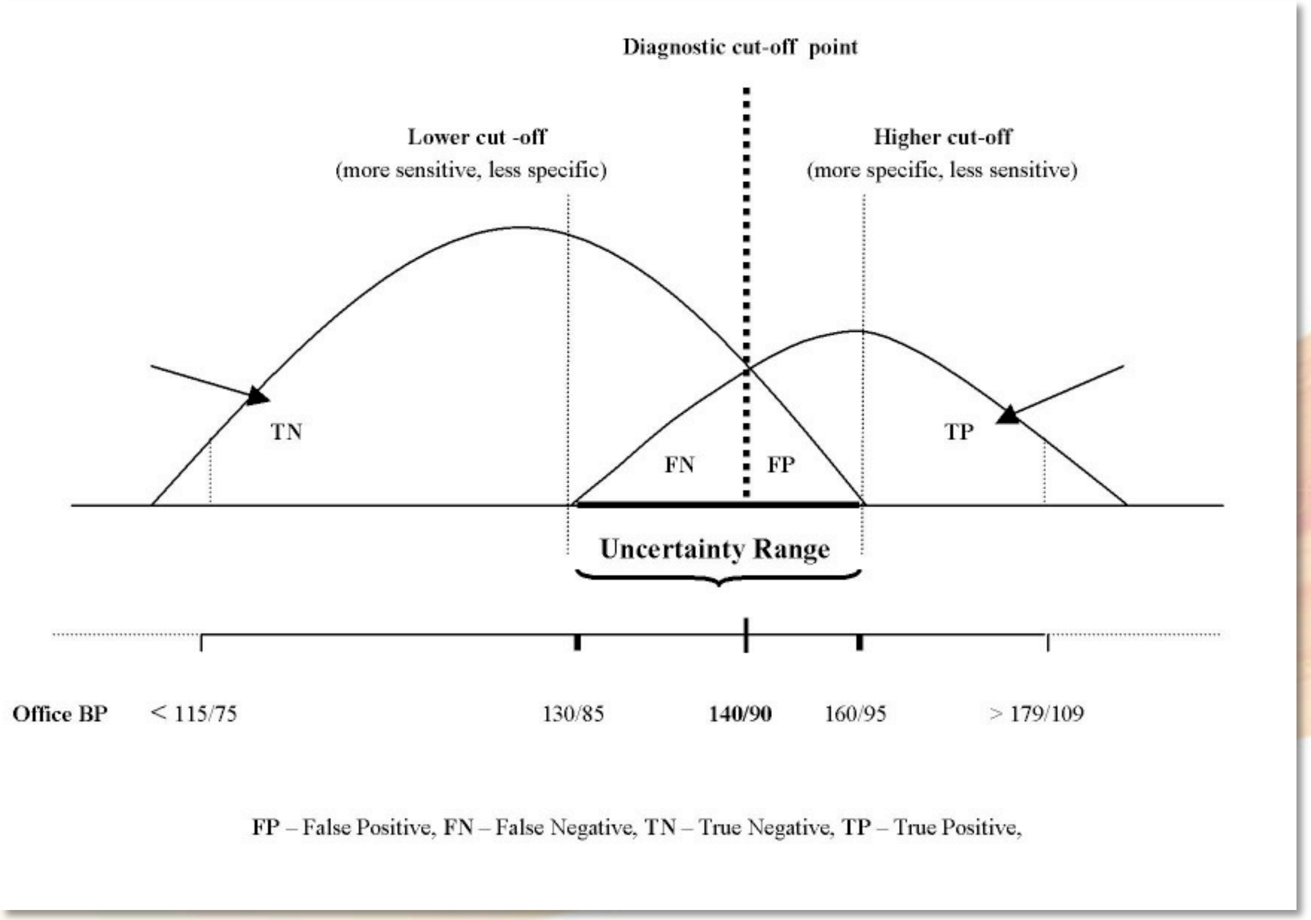
$$=7/\left(7+1\right)=0.88 \qquad =10/\left(10+6\right)=0.63$$

$$=8/\left(8+0\right)=1.00 \qquad =16/\left(16+0\right)=1.00$$



Adopted from Dennis R. Livesay

# Determining a Threshold is a Balancing Act


Diagnostic cut-off point
Lower cut -off
(more sensitive, less specific)
Higher cut-off
(more specific, less sensitive)
TN
FN
FP
TP
Uncertainty Range
Office BP
< 115/75
130/85
140/90
160/95
> 179/109
FP – False Positive, FN – False Negative, TN – True Negative, TP – True Positive,

When trying to get more TP, you get more FP
When trying to get more TN, you get more FN

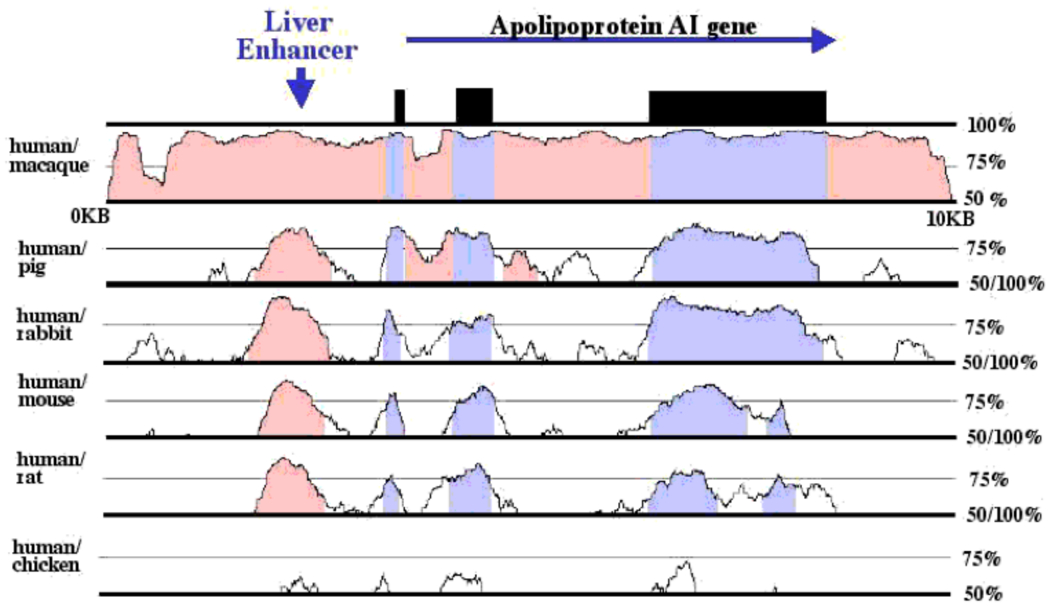
Back to Gene Finding!

# Despite its Performance on Coding Exons

- Genscan's gene-level accuracy (the proportion of genes for which it correctly predicts every exon) on the human genome was only about 10%!
- One reason for the low gene-level accuracy is that typical human genes contain 5 to 10 exons
- Even at 80% accuracy per exon, the likelihood of getting all the exons correct for any particular gene is low

# Detection Genes $3^{rd}$ Method

- Comparative genomics approach
- Locate genes based on observation:
  - Force of natural selection makes genes and other functional elements mutate at a **slower rate** than other parts of a genome
  - With more genomes sequenced in related species, genes can be identified by comparing these genomes to detect this conservation



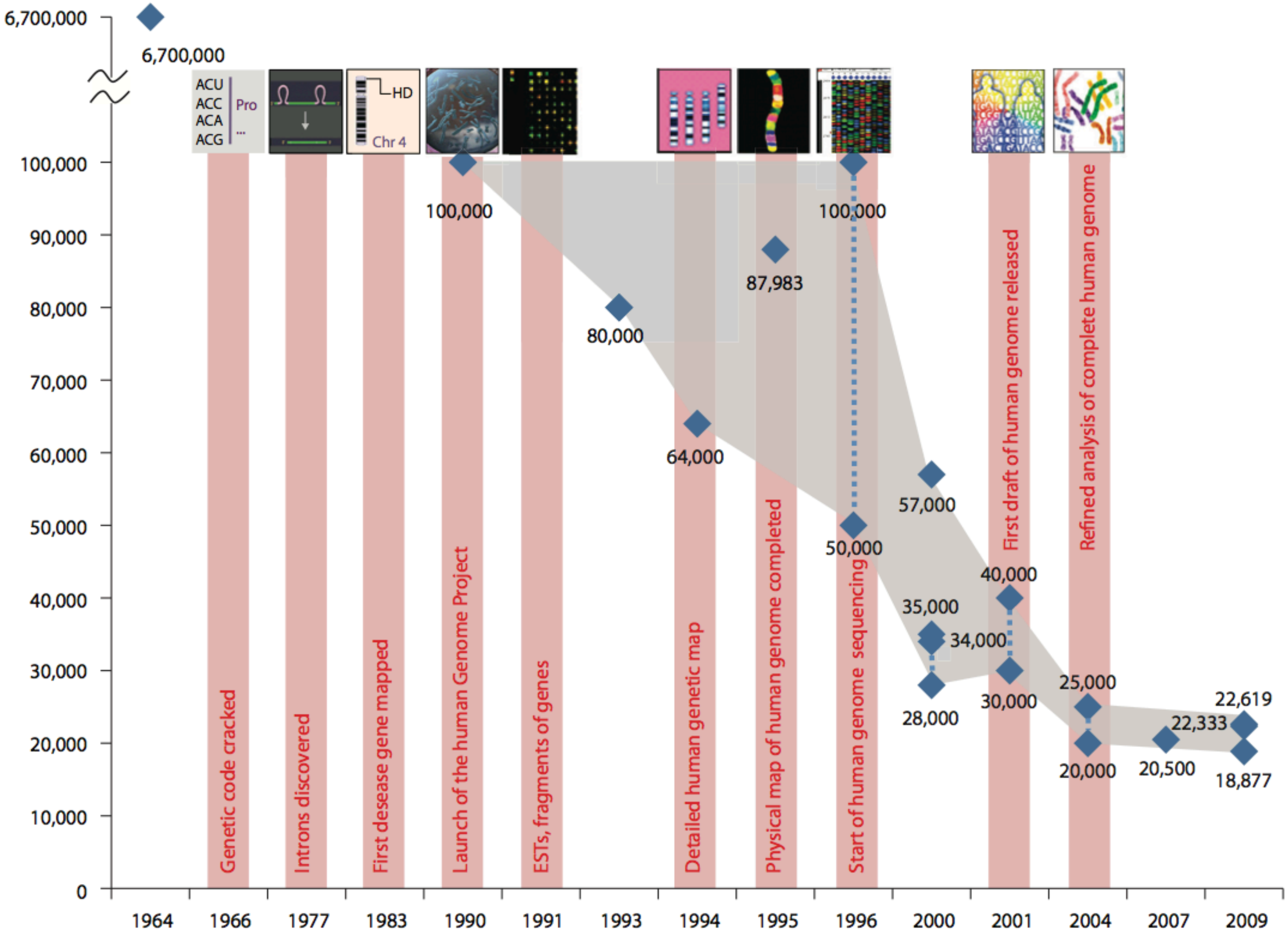
This is becoming easier, why?

# Comparative Gene Finders

- Use patterns of conservation between two related species
  - Such as human and mouse
  - To predict the location and structure of protein-coding genes
- Biggest effect of using two genomes at once was to reduce the number of **false-positive predictions**
  - Using human-mouse alignments, Twinscan [27], a dual-genome gene finder, predicted 25,600 human genes versus 45,000 predicted by Genscan

http://www.ncbi.nlm.nih.gov/pubmed/11473003

# The Trend of Human Gene # Counts



http://genomebiology.com/2010/11/5/206

# One Day We'll Know…

- Although the near-finished human genome sequence now covers 99% of the euchromatic (or gene-containing) genome at 99.999% accuracy (GRCh37.p12)
- Exact number of human genes is **still unknown**
- Leading repositories of genome annotation
  - Ensembl
    - Coding genes:
      - 20,769 Primary assembly
      - 2,516 Alternative assembly
    - Genscan gene predictions: 48,461

http://useast.ensembl.org/Homo_sapiens/Info/Annotation#genebuild

# Closing Statements on Gene Prediction

- All coding regions have non-random sequence characteristics, based partly on **codon usage preferences (see following slides)**
- Empirically:
  - Statistics of **hexanucleotides** perform best in distinguishing coding from non-coding regions
  - Starting from a set of known genes from an organism as a **training set**, pattern-recognition programs can be tuned **to particular genomes**
- Accurate gene detection is a crucial component of genome sequence analysis
  - **~40% of human genes have non-coding 1st exons (UTRs)**
  - **Most gene finders do not handle alternative splicing**
  - **Most can't find non-protein genes (tRNAs)**
- **Best methods approach 80% at the exon level (90% at the nucleotide level) in coding-rich regions (sensitivity)**
- **Gene predictions should always be verified by other means (cDNA sequencing, BLAST search, Mass spec, RNA-Seq)**

5 minute break…

# Sequence Coding Statistics

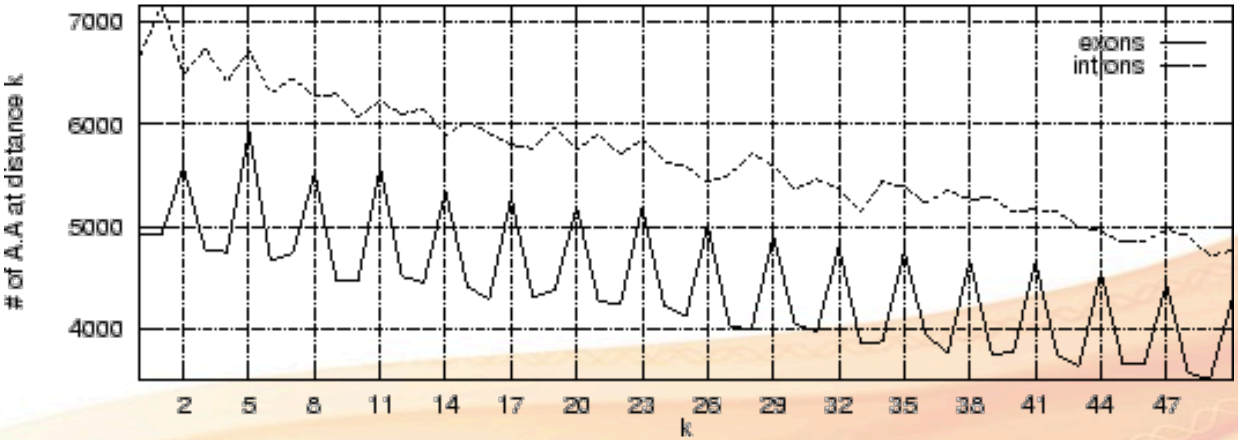
# Sequence Composition Bias in Coding Sequences

- Let's perform a very simple exercise:
  - Given a nucleotide sequence
  - Compute the number of times that the nucleotide **A** (Adenine) appears at a distance **k** from another nucleotide **A**
  - And let's do that for every possible **k**, from 0 to the length of the sequence For instance if the sequence is: TAAGAGACTCATAAGT

| K | # |
|---|---|
| 0 | 2 |
| 1 | 3 |
| 2 | 2 |
| 3 | 2 |
| 4 | 1 |
| 5 | 2 |
| 6 | 1 |
| 7 | 2 |
| 8 | 2 |
| 9 | 1 |
| 10 | 2 |
| 11 | 1 |
| 12 | 0 |
| 13 | 0 |
| 14 | 0 |
| 15 | 0 |

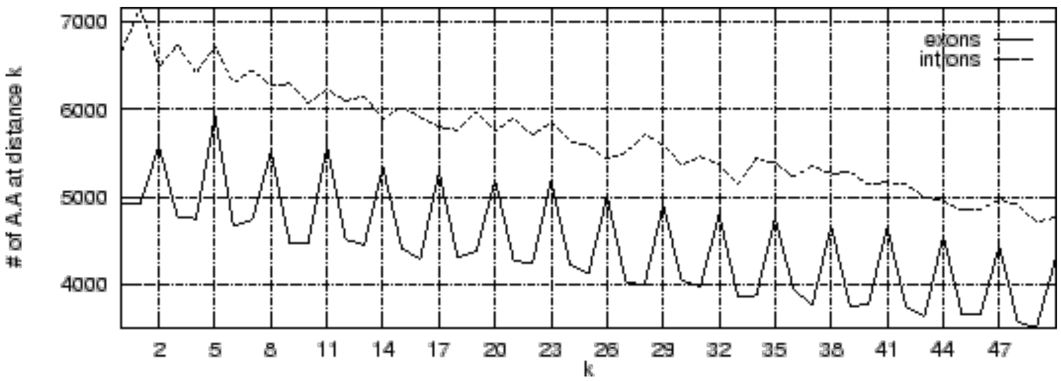
# Large Scale

- Repeat previous analysis - 500 exon  - 500 intron human sequences
- Plot **frequency** of occurrence of pairs **A … A** at each possible distance **k**



- See a clear periodic pattern arises from the set of exon sequences
- The nucleotide **A** is more likely to be found at distance *k=2,5,8, …* from another **A** than at other distances
- Periodic pattern is absent in the intronic sequences
- Anything interesting about the peaks?

# Large Scale



- Nucleotide pairs at a distance of *k=2,5,8, ...* nucleotides
  - At the **same codon position**, whereas nucleotide pairs at other distances, are not
- Periodic pattern reflects:
  - Proteins use different a.a. with different frequencies
  - Synonymous codons are used with different frequencies to code for a given amino acid
- Causes coding sequences to exhibit a **strong codon bias**, which is (mostly) absent in non-coding sequences
- **The codon bias** causes the periodic pattern observed in coding sequences
- Is this a pattern of A .. A?
  - Pattern is characteristic of the **16** pairs of nucleotides, and not only of the **pair A ... A**

# Extending This Idea

- How are we going to **measure** the strength of the periodic pattern in a sequence problem
- We can measure the **likelihood** of the sequence being coding
- A measure of DNA sequence **periodicity** is known as a **sequence coding statistic**
- Measure either:
  - **codon usage bias**
  - base compositional bias between codon position
  - or periodicity in **base occurrence**
- A coding statistic or coding measure is defined:
  - As a **function** that computes given a DNA sequence a **real number** related to the **likelihood** that the **sequence** is **coding for a protein**

# An Example of Coding Statistics: Codon Usage

- Unequal usage of codons in the coding regions appears to be a universal feature of the genomes across the phylogenetic spectra
- This bias caused:
  - Uneven usage of:
    - the amino acids in the existing proteins
    - **synonymous** codons
- Any ideas why there is uneven usage?
- The bias in the usage of synonymous codons correlates with the abundance of the corresponding tRNAs
  - The correlation is particularly strong for highly expressed genes
  - Key to codon optimization
- Codon usage is specific of the taxonomic group

http://www.genscript.com/codon_opt.html

# Comparing the Frequency of Codons

- By comparing the frequency of codons in a region of a species genome, read in a given frame with the typical frequency of codons in the species genes
  - Possible to estimate a likelihood of the region coding for a protein in such a frame
  - Regions in which codons are used with frequencies similar to the typical species codon frequencies are likely to code for genes
- This idea was first introduced by Staden and McLahlan (1982)
- In practice, the likelihood can be computed in a number of different ways
  - How do we do this?
  - We need a model

# Our First Model!

# Log-Likelihood Ratio - *LLR*

- Let ***F(c) =*** frequency (probability) of codon ***c*** in the genes of the species under consideration (use published data)
- Then, given a sequence of codons , $C = C_1 C_2 \cdots C_m$ and assuming independence between adjacent codons

$$P(C) = F(C_1)F(C_2)\cdots F(C_m)$$

is the probability of finding the sequence of codons $C$ knowing that $C$ codes for a protein

- For instance, if ***S*** is the sequence ***S*** = AGGACG, when read in frame 1, it results in the sequence $C_1^1 = \mathtt{AGG}$ $C_2^1 = \mathtt{ACG}$
- Then $P^1(S) = P(C^1) = F(\mathtt{AGG})F(\mathtt{ACG})$
- Substituting the appropriate values from published data, we obtain

$$P^1(S) = P(C^1) = 0.022 \times 0.038 = 0.000836$$

# Log-Likelihood Ratio - *LLR*

- On the other hand, let ***F₀(c)*** be the frequency of codon ***c*** in a non-coding sequence

$$P_0(S) = P_0(C) = F_0(C_1)F_0(C_2)\cdots F_0(C_m)$$

is the probability of finding the sequence ***S*** if ***C*** is non-coding

- Assuming the random model of coding DNA is $F_0(c) = 1/64 = 0.0156$ for all codons
- $\boldsymbol{P_0}$ for the above sequence of codons ***C*** would be:

$$P_0(C) = 0.0156 \times 0.0156 = 0.000244$$

- The LLR ratios for S coding in frame 1, $\boldsymbol{LP^1}$ is

$$LP^1(S) = \log(0.000836/0.000244) = \log(3.43) = 0.53$$

We will see LLR everywhere in bioinformatics

# The Other Frames

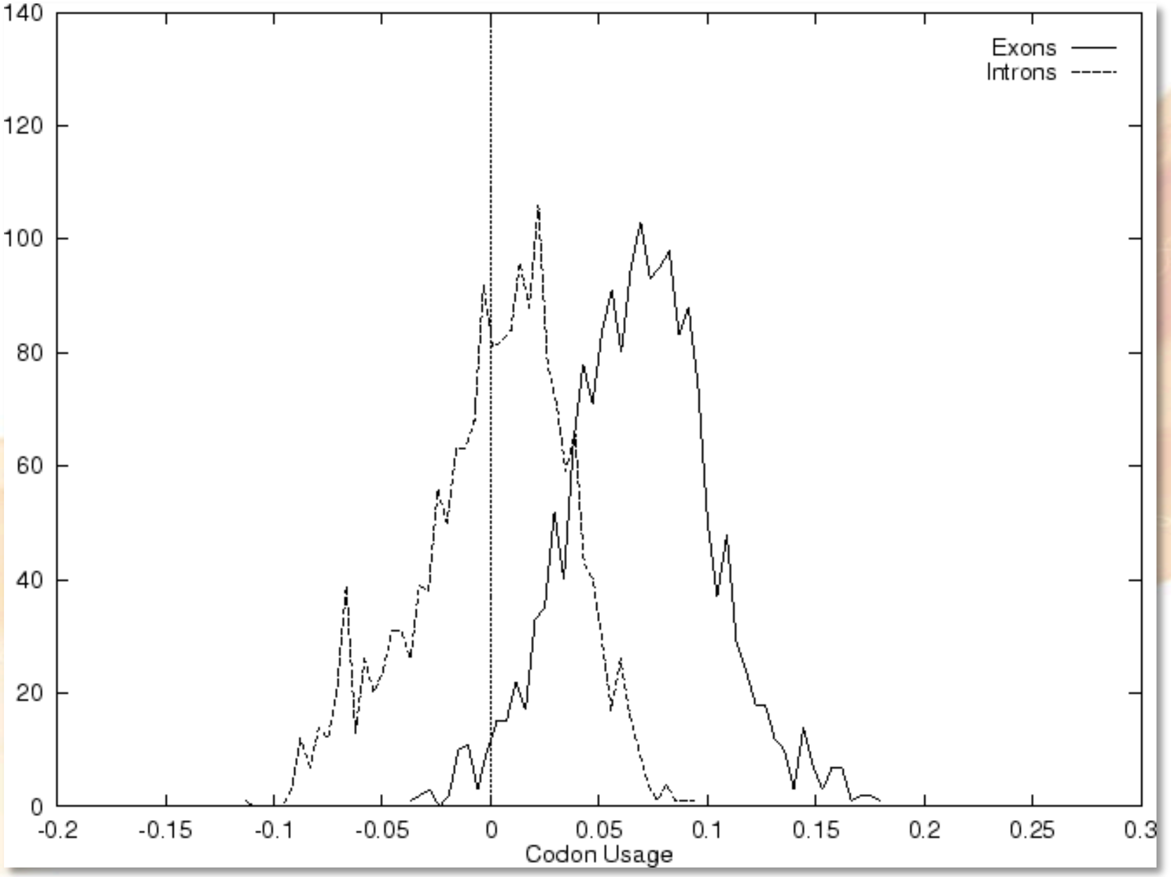
- The log-likelihood ratios for ***S*** coding in frames 2 , and 3 ($\boldsymbol{LP^2}$ and $\boldsymbol{LP^3}$) are computed in a similar way
- Here log-likelihood ratios in the three frames computed on a real exon, and on a real intron sequence

| **exon sequences** | | | **intron sequences** | | |
|---|---|---|---|---|---|
| coding frame | non coding frames | | frame 1 | frame2 | frame 3 |
| 24.06 | -16.3 | -3.16 | -14.36 | -23.74 | -19.67 |

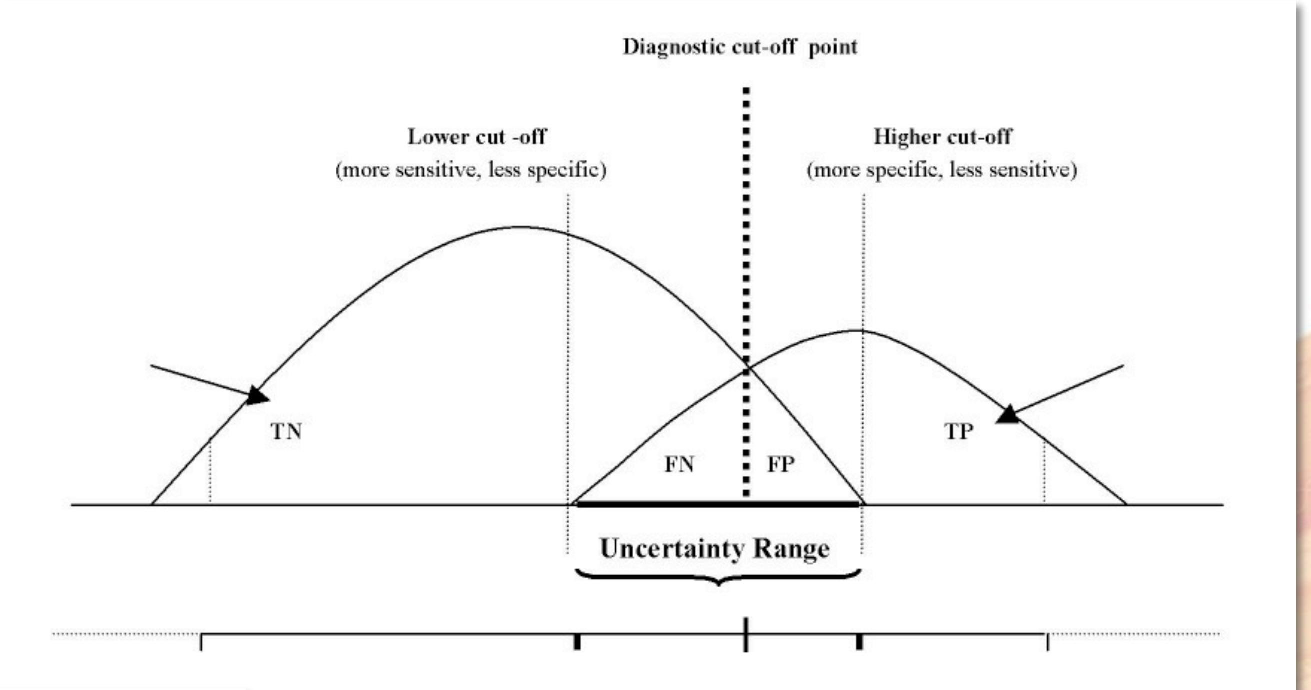
- Here LLR is indeed greater than zero in the coding frame of the exon sequence
- While is smaller than zero in the non-coding frames of the exon sequence and in all frames of the intron sequence

# The Distribution of the Scores of the Codon Usage

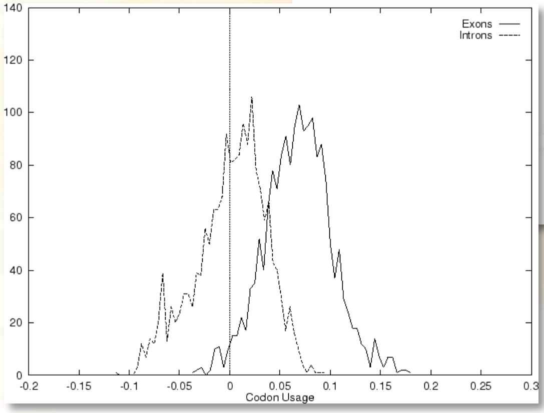
- Although the distributions are clearly distinct, there is substantial overlap between the Codon Usage scores in the sets of intron and exon sequence
- This is a general situation for all coding statistics, and many **models in general**
- **We've seen this before, right?**



# Remember…



FP – False Positive, FN – False Negative, TN – True Negative, TP – True Positive,



# Take Home

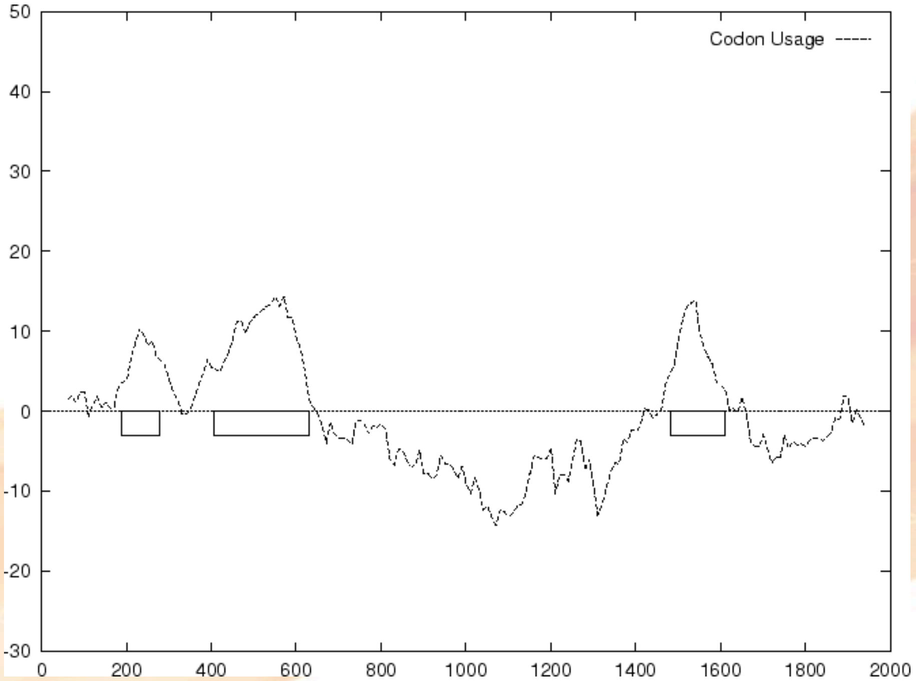
- ***As it can be seen in this case,*** log-likelihood ratio:
  - Greater than zero in the coding frame of the exon sequence
  - Smaller than zero in the non-coding frames of the exon sequence and in all frames of the intron sequence
- In practice:
  - The problem is **not** usually to determine the likelihood a given sequence is coding or not
  - **But to locate the (usually small) coding regions within large genomic sequences**
- How could we do this?

# Typical Way

- Compute the value of a coding statistic in successive windows (an sliding window)
- Record the value of the statistic for each of the windows
- This generates a profile along the sequence in which peaks may point to the coding regions and valleys to the non-coding ones

# 2000 bp Genomic Region Coding for the Human β-globin Gene

- Plot the result of:
  - Sliding a window of length 120 bp
  - Distance between consecutive windows being 10 bp
  - Computing in the three different frames
    - Plotting the highest value obtained
- What are the boxes?
- Codon usage log-likelihood profile reproduces fairly well the exonic structure of this gene



# For Thursday

- Complete Perl lab from last week
- Go over the solutions, I will post later tonight
- Be ready for a quiz
- Go through slides, and come with questions about any uncertainties
- Read: under "Required Readings"
  - "Genomes for All" by George Church
    *http://155.33.203.128/teaching/BIOL6308-Fall2013/local/Literature/GenomesForAll_church.pdf*
  - A General Discription of the Broad's Automated Genome Annotation for Eukaryotic Genomes
    *http://www.broadinstitute.org/science/projects/fungal-genome-initiative/gene-finding-methods*
  - ***Optional:*** "Between a chicken and a number of human genes" by Steven L Salzberg