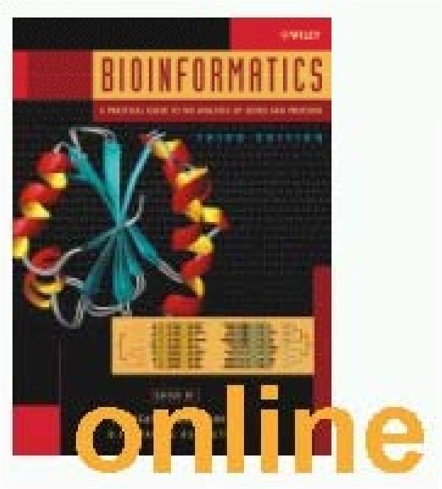
2. Predictive Methods Using DNA Sequences (1)

薛 佑 玲 Yow-Ling Shiue
國立中山大學生物醫學研究所
✉ ylshiue@mail.nsysu.edu.tw

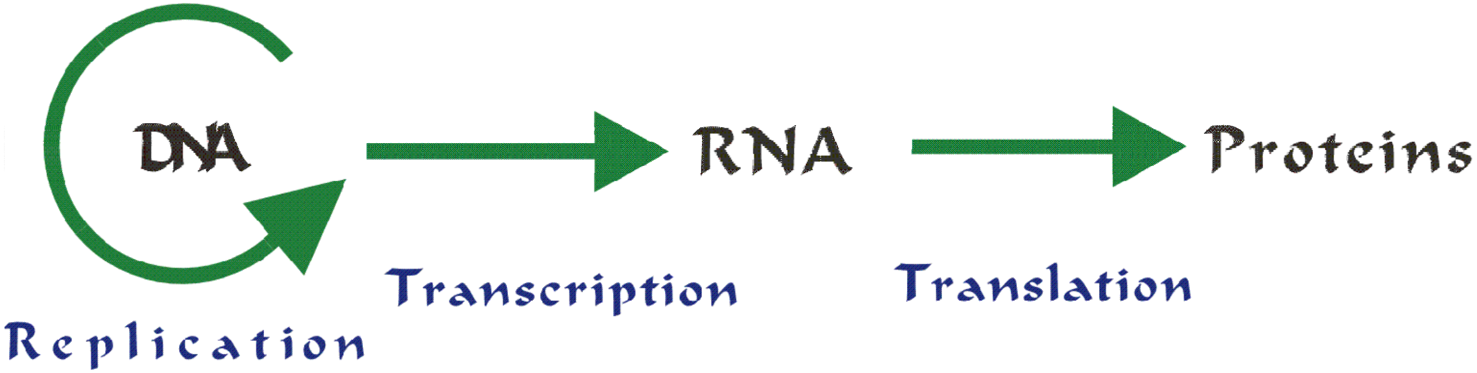


Select a Chapter: Chapter 5

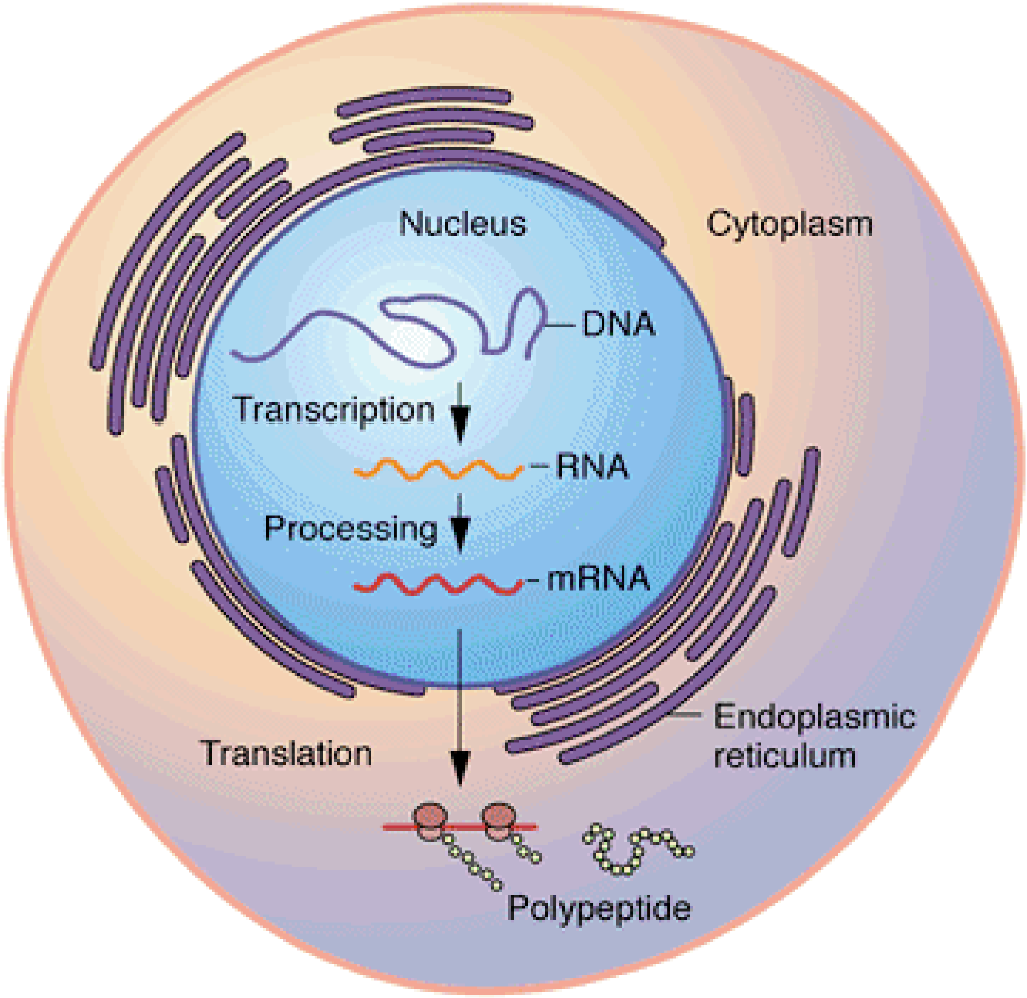
Chapter 5: Predictive Methods Using DNA Sequences

- Sample Data for Problem Sets
- Internet Resources

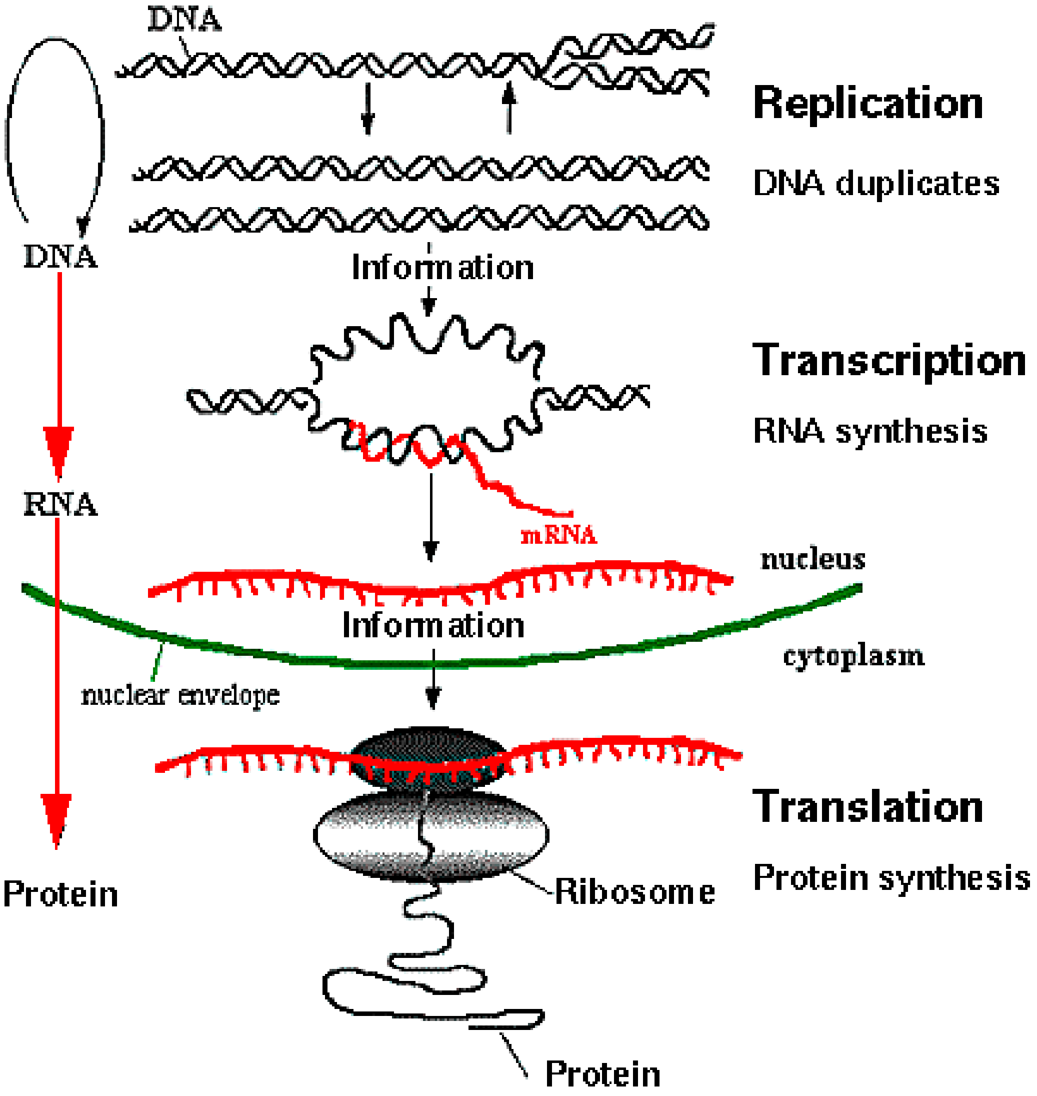
Introduction - the Central Dogma



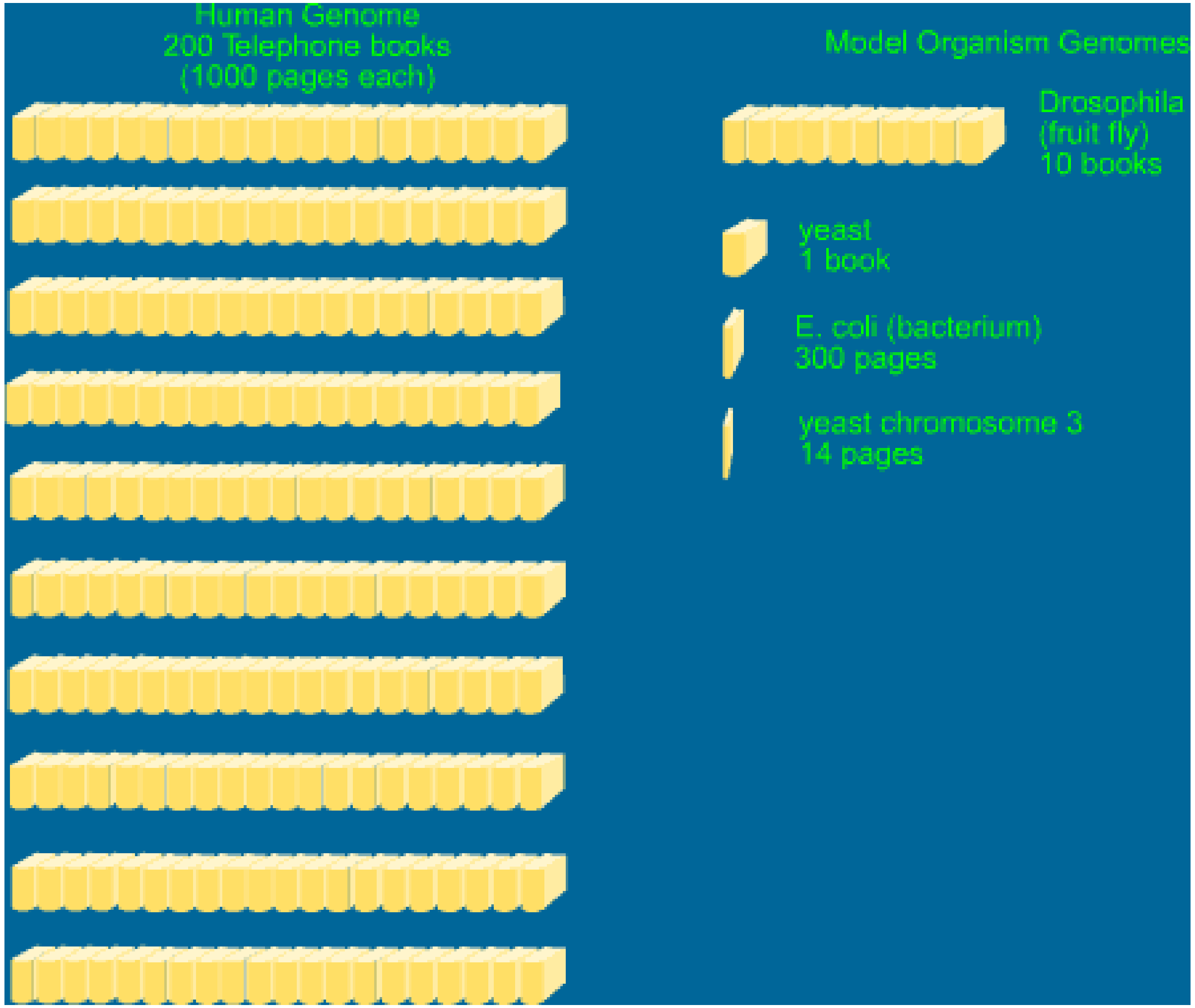
Introduction – the Central Dogma



Introduction – the Central Dogma



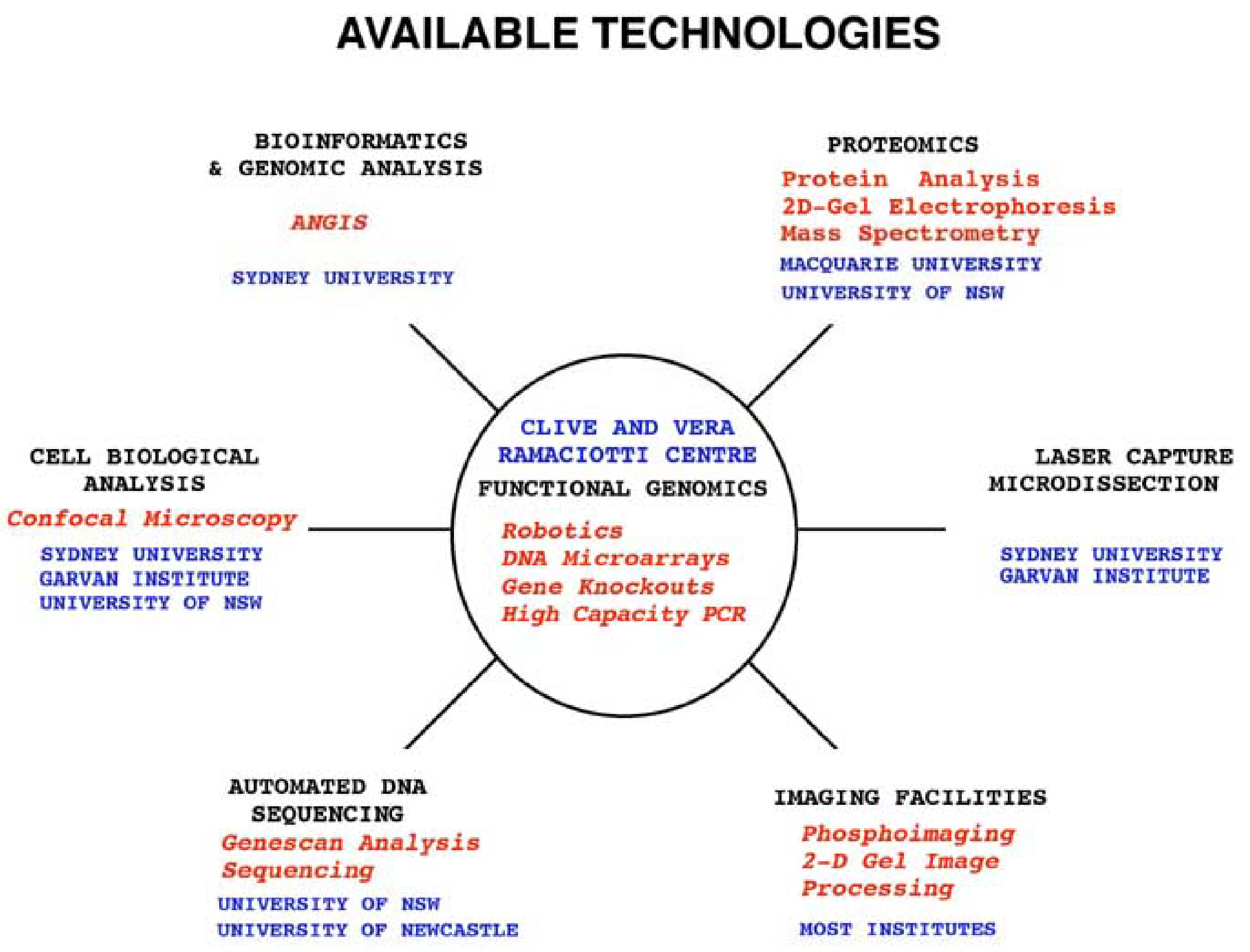
Prokaryotic vs. Eukaryotic Genomes



- <2% of vertebrate genomes code for proteins (Venter et al. 2001)

- http://www.nyu.edu/classes/ytchang/book/e001.html

Functional Genomics



AVAILABLE TECHNOLOGIES

CLIVE AND VERA RAMACIOTTI CENTRE
FUNCTIONAL GENOMICS
Robotics
DNA Microarrays
Gene Knockouts
High Capacity PCR

BIOINFORMATICS & GENOMIC ANALYSIS
ANGIS
SYDNEY UNIVERSITY

PROTEOMICS
Protein Analysis
2D-Gel Electrophoresis
Mass Spectrometry
MACQUARIE UNIVERSITY
UNIVERSITY OF NSW

CELL BIOLOGICAL ANALYSIS
Confocal Microscopy
SYDNEY UNIVERSITY
GARVAN INSTITUTE
UNIVERSITY OF NSW

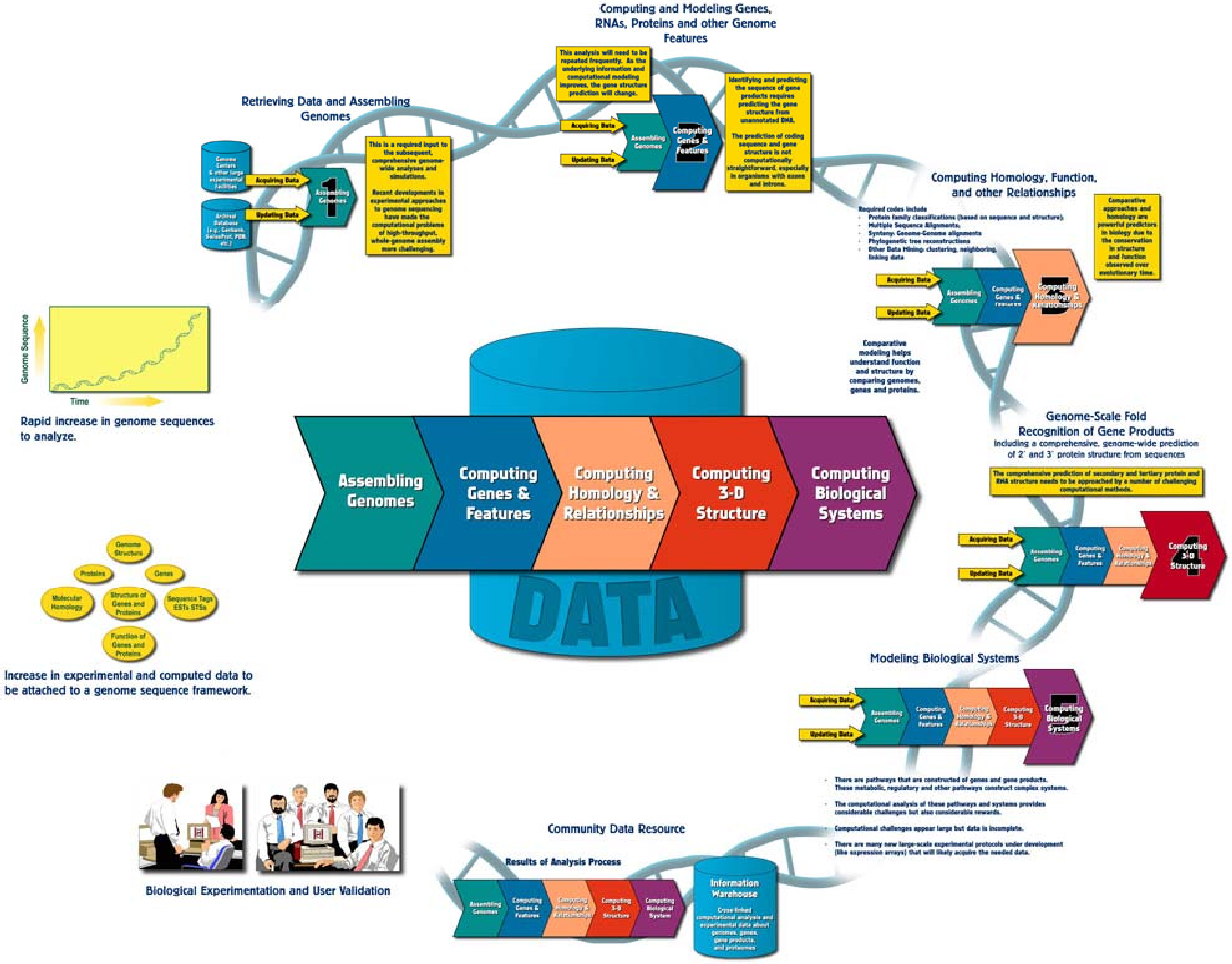
LASER CAPTURE MICRODISSECTION
SYDNEY UNIVERSITY
GARVAN INSTITUTE

AUTOMATED DNA SEQUENCING
Genescan Analysis
Sequencing
UNIVERSITY OF NSW
UNIVERSITY OF NEWCASTLE

IMAGING FACILITIES
Phosphoimaging
2-D Gel Image Processing
MOST INSTITUTES

http://www.ramaciotti.unsw.edu.au/directors_report.html

Computing the Genome Revolution: Biology for the 21st Century



Assembling Genomes
Computing Genes & Features
Computing Homology & Relationships
Computing 3-D Structure
Computing Biological Systems
DATA
Acquiring Data
Updating Data
Genome Centers & other large experimental facilities
Archival Database (e.g. Genbank, SwissProt, PDB, etc.)
1
2
3
4
5
Genome Sequence
Time
Genome Structure
Proteins
Genes
Molecular Homology
Structure of Genes and Proteins
Sequence Tags ESTs STSs
Function of Genes and Proteins
Results of Analysis Process
Computing Biological System
Information Warehouse
Cross-linked computational analysis and experimental data about genomes, genes, gene products, and proteomes

Retrieving Data and Assembling Genomes

This is a required input to the subsequent, comprehensive genome-wide analyses and simulations.

Recent developments in experimental approaches to genome sequencing have made the computational problems of high-throughput, whole-genome assembly more challenging.

Computing and Modeling Genes, RNAs, Proteins and other Genome Features

This analysis will need to be repeated frequently. As the underlying information and computational modeling improves, the gene structure prediction will change.

Identifying and predicting the sequence of gene products requires predicting the gene structure from unannotated DNA.

The prediction of coding sequence and gene structure is not computationally straightforward, especially in organisms with exons and introns.

Computing Homology, Function, and other Relationships

Required codes include

- Protein family classifications (based on sequence and structure),
- Multiple Sequence Alignments,
- Synteny: Genome-Genome alignments
- Phylogenetic tree reconstructions
- Other Data Mining: clustering, neighboring, linking data

Comparative approaches and homology are powerful predictors in biology due to the conservation in structure and function observed over evolutionary time.

Comparative modeling helps understand function and structure by comparing genomes, genes and proteins.

Genome-Scale Fold Recognition of Gene Products

Including a comprehensive, genome-wide prediction of 2˚ and 3˚ protein structure from sequences

The comprehensive prediction of secondary and tertiary protein and RNA structure needs to be approached by a number of challenging computational methods.

Modeling Biological Systems

- There are pathways that are constructed of genes and gene products. These metabolic, regulatory and other pathways construct complex systems.
- The computational analysis of these pathways and systems provides considerable challenges but also considerable rewards.
- Computational challenges appear large but data is incomplete.
- There are many new large-scale experimental protocols under development (like expression arrays) that will likely acquire the needed data.

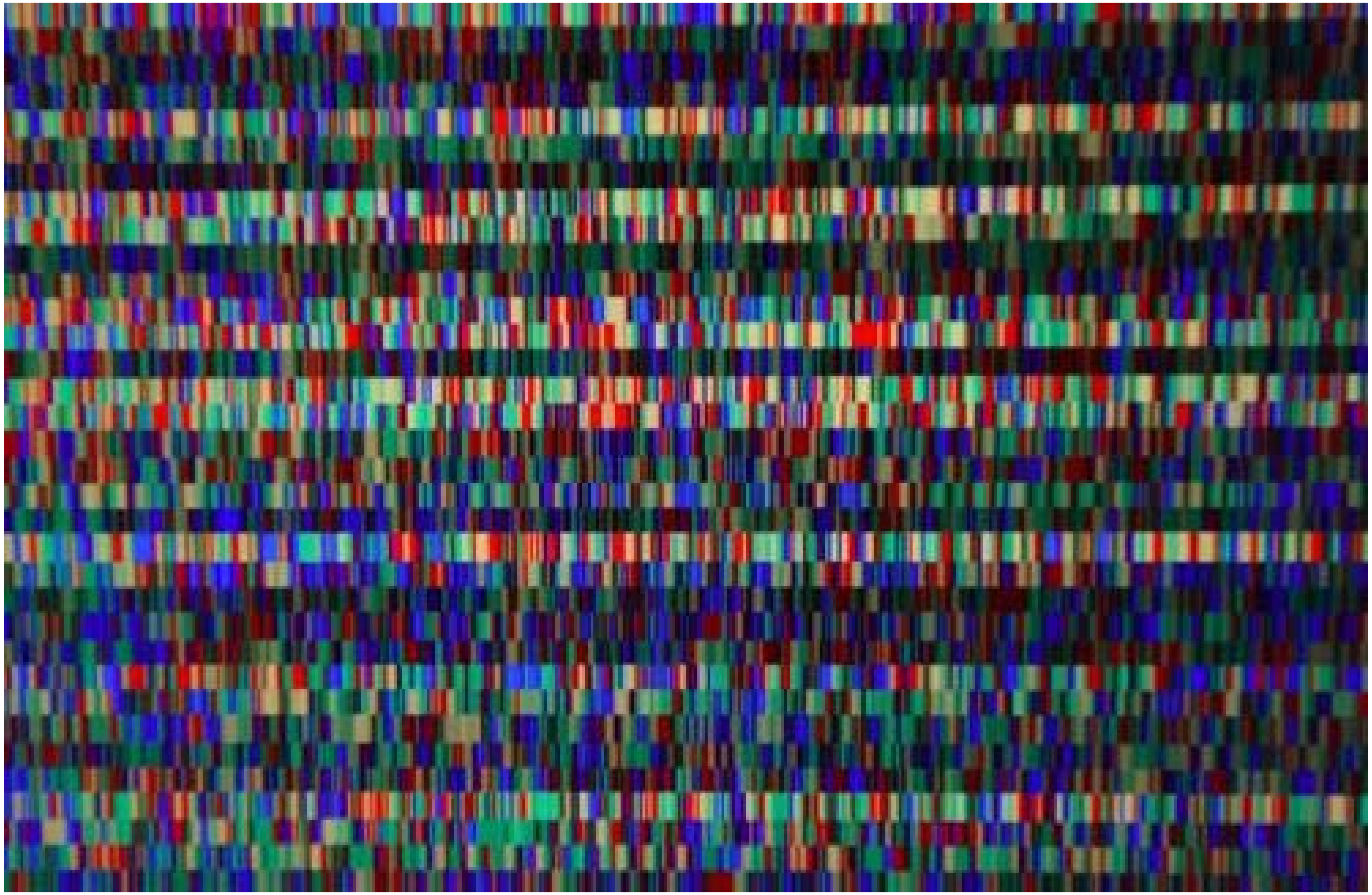
Community Data Resource

Rapid increase in genome sequences to analyze.

Increase in experimental and computed data to be attached to a genome sequence framework.

Biological Experimentation and User Validation

Gene Prediction Programs (1)



Gene Prediction Programs (2)

- Factors based
  - **Compositional bias** found in protein-coding regions
  - **Similarity** with known sequences
- But **not** accurate enough, without **cDNA sequence** data
  - Prediction = highly hypothetical

Gene Prediction Programs (3)

- Annotation of the human genome
  - Genome Browse (UCSC)
    - Kent et al. 2002
  - Ensembl (EBI)
    - Birney et al. 2004
  - Map Viewer (NCBI)

Gene Prediction Methods – Single vs. Combinatorial (1)

- **Searching by signal**
  - The analysis of sequence signals that are potentially involved in **gene specification**

- **Searching by content**
  - The analysis of regions showing **compositional bias** that has been correlated with **coding regions**

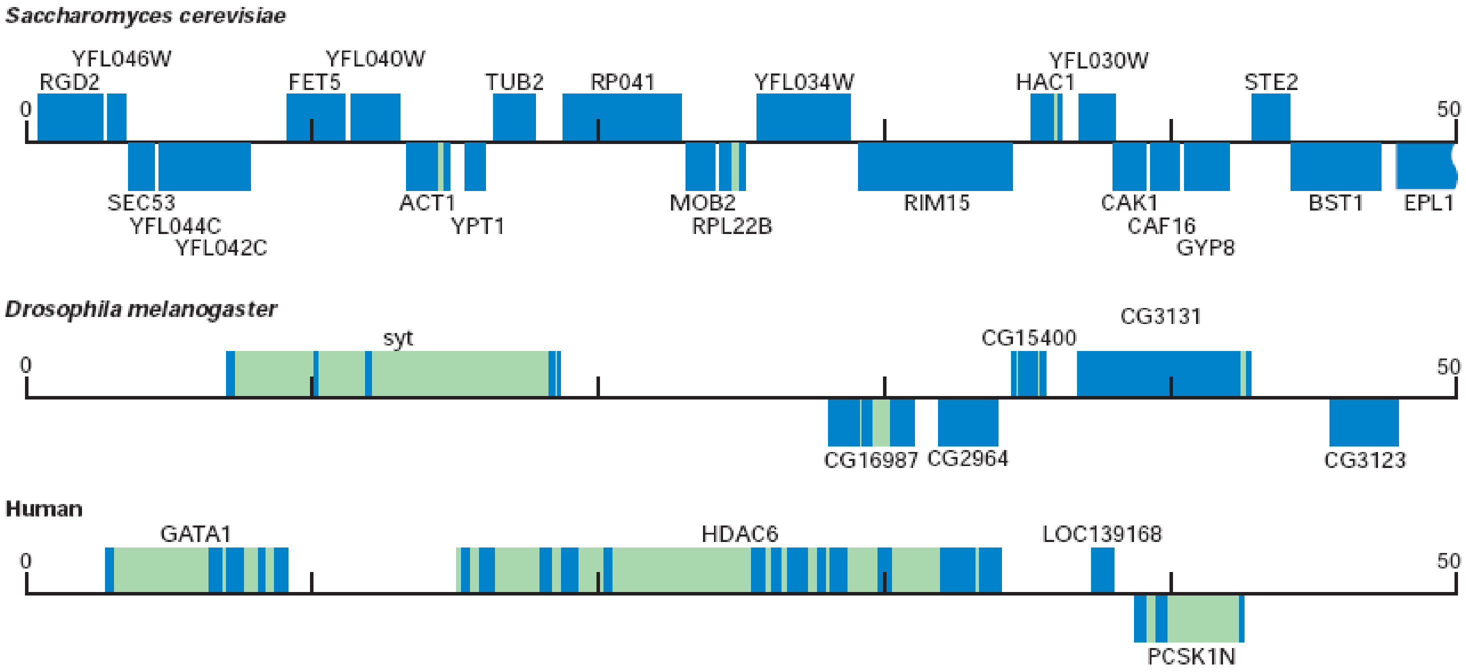
- **Example**
  - ***Ab initio*** gene prediction ~ **intrinsic** or **template** gene prediction

Gene Prediction Methods – Single vs. Combinatorial (2)

- Homolog-based gene prediction
  - Comparing sequences of interest against **known coding sequences**
- Comparative gene prediction
  - Comparing sequences of interest **anonymous genomic sequences**

- **Example**
  - **Extrinsic** or **look-up** gene prediction
    - Gene structure is predicted through **comparison with other sequences** whose **characteristics** are already known

Prokaryotic vs. Eukaryotic Genes (1)



▲ **FIGURE 9-33 Arrangement of gene sequences in representative 50-kb segments of yeast, fruit fly, and human genomes.** Genes above the line are transcribed to the right; genes below the line are transcribed to the left. Blue blocks represent exons (coding sequences); green blocks represent introns (noncoding sequences). Because yeast genes contain few if any introns, scanning genomic sequences for open reading frames (ORFs) correctly identifies most gene sequences. In contrast, the genes of higher eukaryotes typically comprise multiple exons separated by introns. ORF analysis is not effective in identifying genes in these organisms. Likely gene sequences for which no functional data are available are designated by numerical names: in yeast, these begin with Y; in *Drosophila*, with CG; and in humans, with LOC. The other genes shown here encode proteins with known functions.

Prokaryotic vs. Eukaryotic Genes (2)

- **Prokaryotic genes**
  - By **single** open reading frames **(ORFs)**
    - Usually found **adjacent** to one another

- **Eukaryotic genes**
  - Coding sequences (the **exons**) are interrupted by large, non-coding **introns**

Gene Prediction in Eukaryotes (1)

1. Identifying and scoring suitable
   - **Splice sites**, **start** & **stop signals** along the query sequence
2. Predicting candidate **exons**
   - As deduced through the detection of these **signals**

Gene Prediction in Eukaryotes (2)

3. Scoring these exons as a function of both
   - The **signals** used to detect the **exons**, as well as on
   - **Coding statistics** computed on the putative exon sequence itself

- In **homology-based & comparative methods**
  - Exon scores **factor** in the quality of **the alignment** between **the query sequence** and either **known coding sequences** or **anonymous genomic sequences**

Gene Prediction in Eukaryotes (3)

4. Assembling a subset of these candidates into a predicted gene structure
   - To maximize **a particular scoring function**
     - Dependent on the score of each of the individual exon candidates that comprise the overall predicted gene structure

Prediction of Exon-Defining Signals (1)

- Df: sequence signals
  - **Short, function DNA elements** involved in **gene specification**

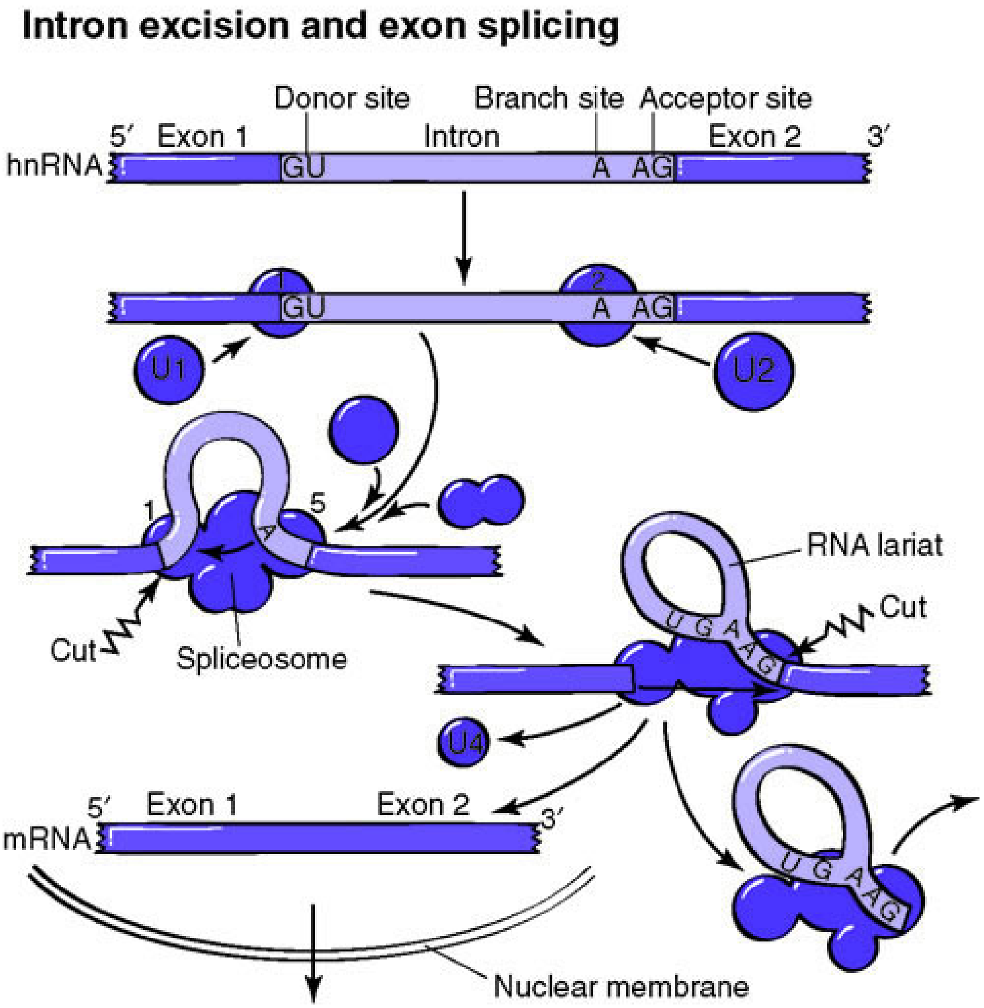
Four Basic Signals Involved in Gene Specification (1) – PWMs

1. The translational start site ($^{1}\underline{A}TG$)

2. The 5' (donor) splicing site

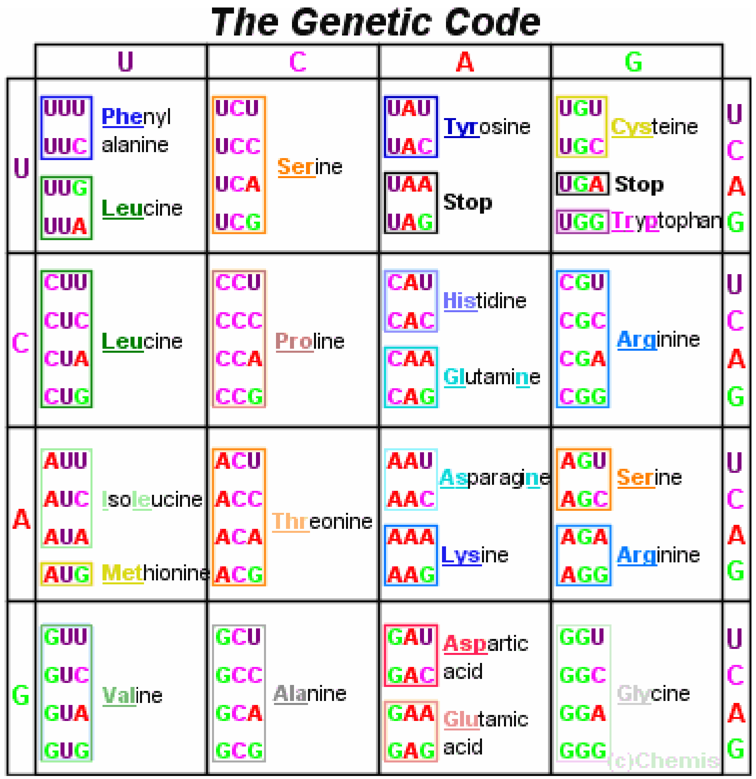
3. The 3' (acceptor) splicing site

U1, U2: ribonucleoproteins



Four Basic Signals Involved in Gene Specification (2) - PWMs

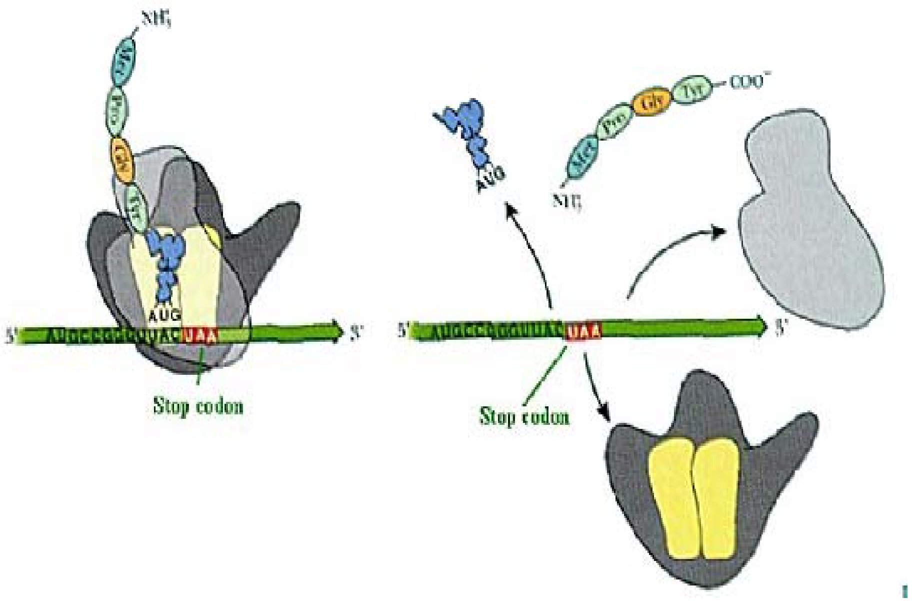
4. The stop codon



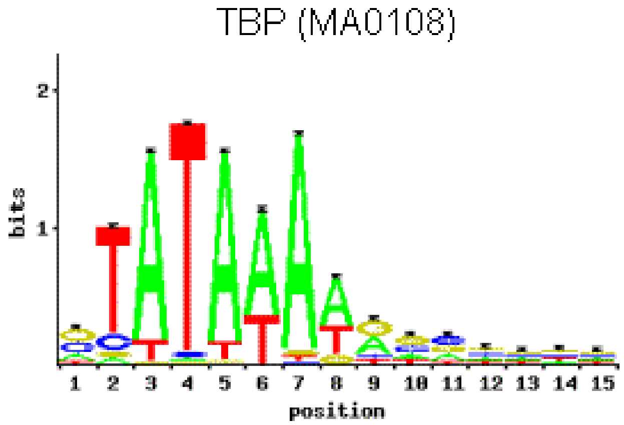
The Genetic Code

	U	C	A	G	
U	UUU UUC Phenyl alanine UUG UUA Leucine	UCU UCC UCA UCG Serine	UAU UAC Tyrosine UAA UAG Stop	UGU UGC Cysteine UGA Stop UGG Tryptophan	U C A G
C	CUU CUC CUA CUG Leucine	CCU CCC CCA CCG Proline	CAU CAC Histidine CAA CAG Glutamine	CGU CGC CGA CGG Arginine	U C A G
A	AUU AUC Isoleucine AUA AUG Methionine	ACU ACC ACA ACG Threonine	AAU AAC Asparagine AAA AAG Lysine	AGU AGC Serine AGA AGG Arginine	U C A G
G	GUU GUC GUA GUG Valine	GCU GCC GCA GCG Alanine	GAU GAC Aspartic acid GAA GAG Glutamic acid	GGU GGC GGA GGG Glycine	U C A G

When the ribosome encounters a stop codon (shown as the red triplet), there is no tRNA attracted and the ribosome separates and leaves the mRNA.



Position Weight Matrices (PWMs)



```
A [ 61  16 352   3 354 268 360 222 155  56  83  82  82  68  77 ]
C [145  46   0  10   0   0   3   2  44 135 147 127 118 107 101 ]
G [152  18   2   2   5   0  10  44 157 150 128 128 128 139 140 ]
T [ 31 309  35 374  30 121   6 121  33  48  31  52  61  75  71 ]
```

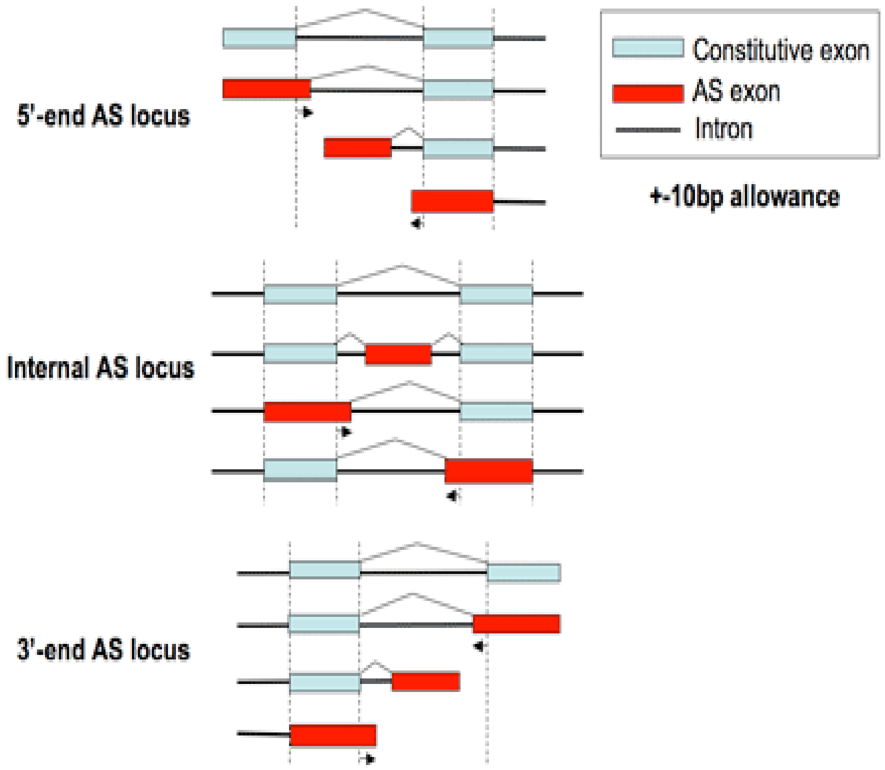
- A set of **known functional signals** and are used **to compute** the sequence signal across a sequence of interest

- **Transcriptional binding protein** (TBP) motif

Prediction & Scoring of Exons (1)

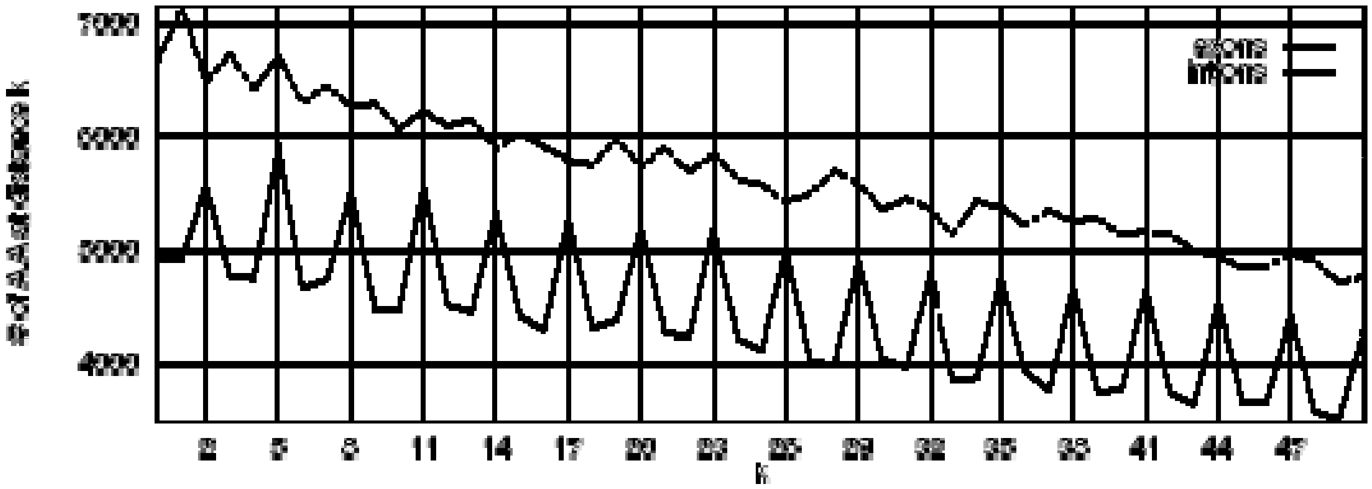
  - Sequence signals +
- **Content-based features = coding statistics**

- Three types of exons
  - **Initial exons**
    - ORFs delimited by **a start site** and a **5' (donor) site**
  - **Internal exons**
    - ORFs delimited by a **3' (acceptor) site** and **5' (donor) site**
  - **Terminal exons**
    - ORFs delimited by **a 3' (acceptor) site** and **a stop codon**



Content-based Features = Coding Statistics

Absolute frequency

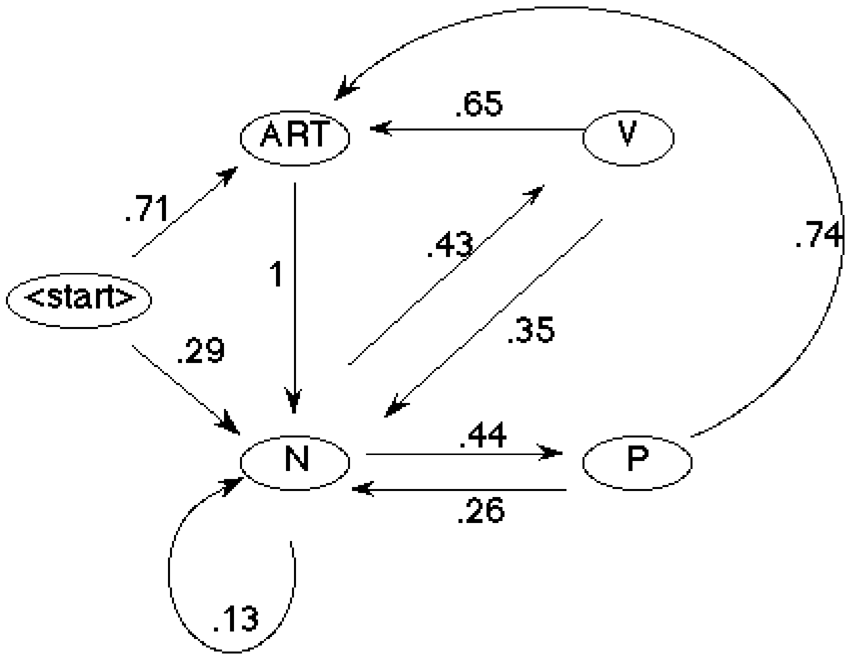


- •1761 exons
- •1753 introns
- •Clear triplet pattern in coding regions=the characteristic codon usage in coding region

- **Coding statistics**
  - The **likelihood** that a given DNA sequence codes for **a protein** or **protein fragment**
    - E.g., **Hexamer frequencies**: in the form of codon position-dependent fifth-order Markov model: most widely used

- The **uneven** distribution of amino acids in proteins, discriminate **protein-coding** regions from **non-coding regions**
  - Fickett & Tung 1992; Gelfand 1995; Guigo 1999

A Morkov Chain



- A series of observations in which the **probability** of an observation depends **on a number of previous observations**

- The number of observations defines **the "order" of the chain**
  - **[Example]** in **a first-order** Markov model, the probability of an observation depends **only on the previous observation**. **In a Markov chain of order 5**, the probability of an observation depends **on the five preceding observations**

- An edge-labeled **directed graph**; each **node**: **a "state";** edge-labels: **probabilities** of moving the state at the end of **the directed arc**.

DNA Sequences & Markov Models

- The **likelihood** of observing a particular base **at a given position** may depend on **the base preceding it**
  - In particular, in coding regions, it is well known that the probability of a given base depends on the **five preceding bases**, reflecting observed **codon biases** and **dependencies** between **adjacent codons**
  - In non-coding regions, such **dependence** is not observed
- When scanning *an* ***anonymous*** *genomic region*, one can compute how well **the local nucleotide sequence** conforms to **the fifth-order dependencies** observed **in coding regions** & assign appropriate **coding likelihood scores**

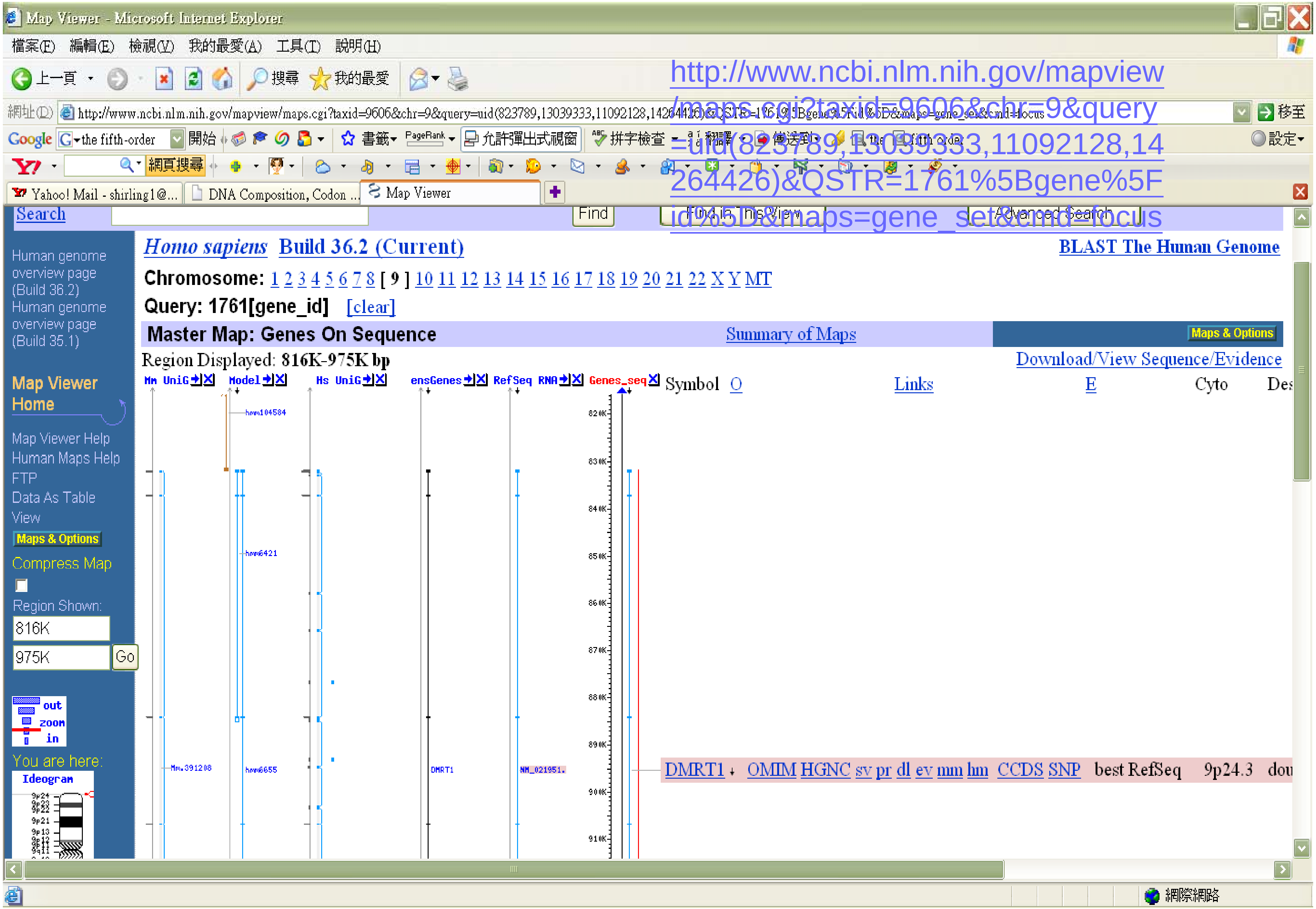
Prediction of Genes Through *Ab initio* Methods

- **Splicing genes** together into **a putative gene structure** can help to eliminate **the prediction of false exons** by simply examining whether **adjacent exons** maintain the open reading frame established by the initial exon
  - See next slide

- Main difficulty in **exon assembly**
  - Simple **combinatiorics**: the number of possible exon assemblies grows **exponentially** with the number of predicted exons for any given gene

  - **Solution**
    - Dynamic programming techniques (Bellman 1957)

http://www.ncbi.nlm.nih.gov/mapview/maps.cgi?taxid=9606&chr=9&query=uid(823789,13039333,11092128,14264426)&QSTR=1761%5Bgene%5Fid%5D&maps=gene_set&cmd=focus



Map Viewer - Microsoft Internet Explorer

檔案(F) 編輯(E) 檢視(V) 我的最愛(A) 工具(T) 說明(H)

上一頁 搜尋 我的最愛

網址(D) http://www.ncbi.nlm.nih.gov/mapview/maps.cgi?taxid=9606&chr=9&query=uid(823789,13039333,11092128,14264426)&QSTR=1761%5Bgene%5Fid%5D&maps=gene_set&cmd=focus 移至

Yahoo! Mail - shirling1@... DNA Composition, Codon ... Map Viewer

Search Find Find In This View Advanced Search

Human genome overview page (Build 36.2)
Human genome overview page (Build 35.1)

Map Viewer Home

Map Viewer Help
Human Maps Help
FTP
Data As Table View
Maps & Options
Compress Map □

Region Shown: 816K 975K Go

You are here:

Homo sapiens Build 36.2 (Current) BLAST The Human Genome

Chromosome: 1 2 3 4 5 6 7 8 [9] 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y MT

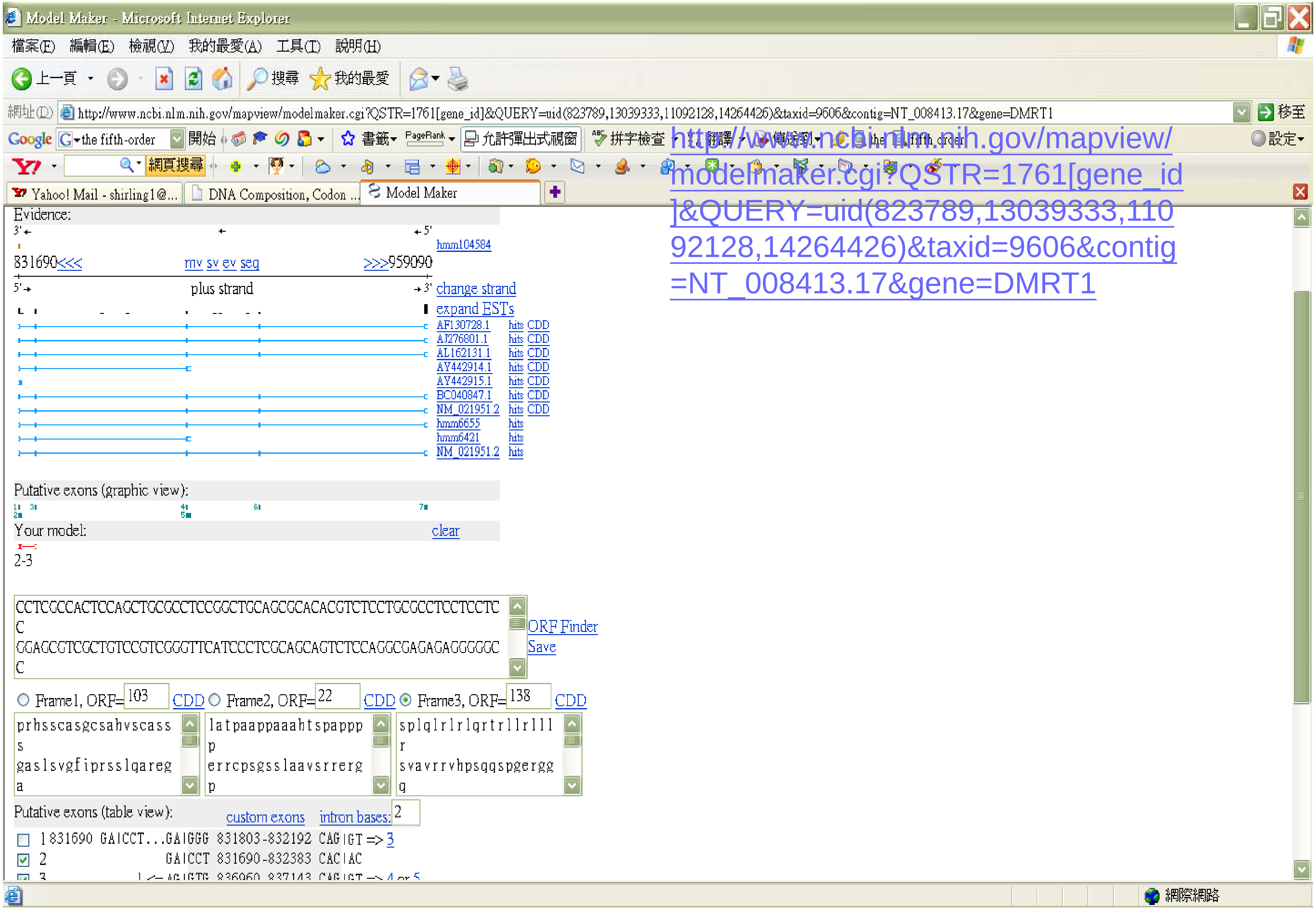
Query: 1761[gene_id] [clear]

Master Map: Genes On Sequence Summary of Maps Maps & Options

Region Displayed: 816K-975K bp Download/View Sequence/Evidence

Symbol O Links E Cyto Des

DMRT1 OMIM HGNC sv pr dl ev mm hm CCDS SNP best RefSeq 9p24.3 dou



[http://www.ncbi.nlm.nih.gov/mapview/modelmaker.cgi?QSTR=1761[gene_id]&QUERY=uid(823789,13039333,11092128,14264426)&taxid=9606&contig=NT_008413.17&gene=DMRT1](http://www.ncbi.nlm.nih.gov/mapview/modelmaker.cgi?QSTR=1761[gene_id]&QUERY=uid(823789,13039333,11092128,14264426)&taxid=9606&contig=NT_008413.17&gene=DMRT1)

Model Maker - Microsoft Internet Explorer
檔案(F) 編輯(E) 檢視(V) 我的最愛(A) 工具(T) 說明(H)
上一頁 搜尋 我的最愛
網址(D)
移至
Google the fifth-order 開始 書籤 PageRank 允許彈出式視窗 拼字檢查 翻譯 傳送到 the fifth-order 設定
網頁搜尋
Yahoo! Mail - shirling1@...
DNA Composition, Codon ...
Model Maker
網際網路

http://www.ncbi.nlm.nih.gov/mapview/modelmaker.cgi?QSTR=1761[gene_id]&QUERY=uid(823789,13039333,11092128,14264426)&taxid=9606&contig=NT_008413.17&gene=DMRT1

Evidence:

AF130728.1	hits	CDD
AJ276801.1	hits	CDD
AL162131.1	hits	CDD
AY442914.1	hits	CDD
AY442915.1	hits	CDD
BC040847.1	hits	CDD
NM_021951.2	hits	CDD
hmm6655	hits	
hmm6421	hits	
NM_021951.2	hits	

Putative exons (graphic view):

Your model: clear

2-3

```
CCTCGCCACTCCAGCTGCGCCTCCGGCTGCAGCGCACACGTCTCCTGCGCCTCCTCCTC
C
GGAGCGTCGCTGTCCGTCGGGTTCATCCCTCGCAGCAGTCTCCAGGCGAGAGAGGGGGC
C
```

ORF Finder

Save

○ Frame1, ORF= 103 CDD ○ Frame2, ORF= 22 CDD ● Frame3, ORF= 138 CDD

```
prhsscasgcsahvscass      latpaappaaahtspappp      splqlrlrlqrtrllrlll
s                        p                        r
gaslsvgfiprsslqareg      errcpsgsslaavsrrerg      svavrrvhpsqqspgergg
a                        p                        q
```

Putative exons (table view): custom exons intron bases: 2

```
□ 1 831690 GA|CCT...GA|GGG 831803-832192 CAG|GT => 3
☑ 2                 GA|CCT 831690-832383 CAC|AC
☑ 3            1 <= AG|GTG 836060-837143 CAG|GT => 4 or 5
```

Programs with Dynamic Programming for Gene Prediction

- The solution of a general problem is obtained by **the recursive solution of smaller versions of the problem** (Gelfand & Roytberg 1993)
    - Find the solution **efficiently** without having to enumerate or consider each and every possible combination of exons

- **GRAIL2**
    - Xu et al. 1994

- **FGENESH**
    - Solovyev et al. 1995

- **GENEID**
    - Guigo et al. 1992; Guigo 1998

Hidden Markov Models (HMMs) in Gene Prediction (1)

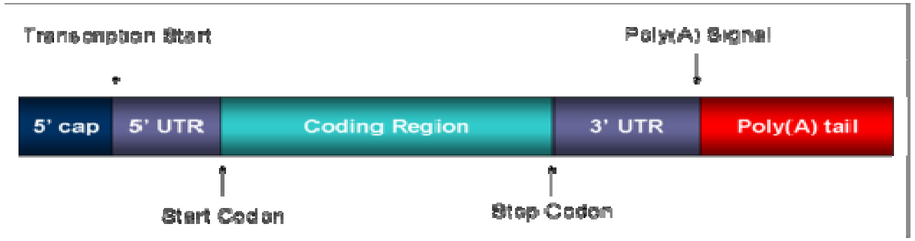
- To define highly complex patterns, e.g., **multigenic** genes
  - High **efficiency** in genome sequences

- **Applications**
  - Multiple sequence alignment **(MSA)**
  - The classification and characterization of **protein families**
  - The comparison of **protein structures**
  - The **prediction** of **gene structure**

Hidden Markov Models (HMMs) in Gene Prediction (2)

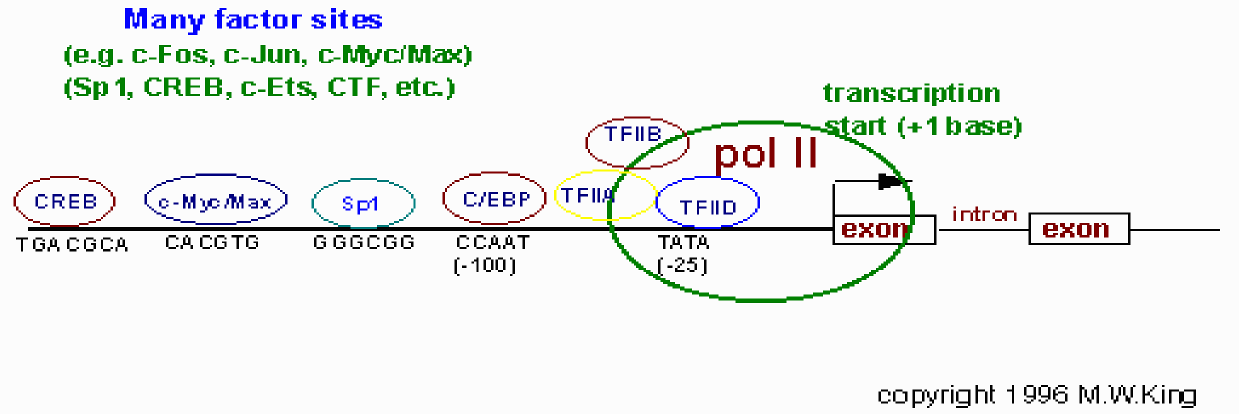
- **Input**
  - A raw **nucleotide sequence**
- **To predict**
  - Whether a given base is most likely found in
    - An intron,
    - An exon, or
    - Within an intergenic region

- **From 5' to 3' end of the gene**
  - The unique characteristics of **promoter** regions
    - Transcription start sites (TSSs), 5' UTRs, start codons, exons, splice donors, splice acceptors, stop codons, 3' UTRs, **polyA tails**



Hidden Markov Models (HMMs) in Gene Prediction (3)

- To take into account
  - The **promoter** (& its **TATA box**) must be **appear before** the **start codon**
  - An **initial exon** must follow **the start codon**
  - Introns must follow exons
  - Introns can only be followed by **internal** or **terminal** exons
  - Stop codons **cannot** interrupt the coding region
  - **PolyA signals** must appear **after the stop codon** (see previous slide)
  - An **ORF** must be **maintained throughout** actually to produce a protein

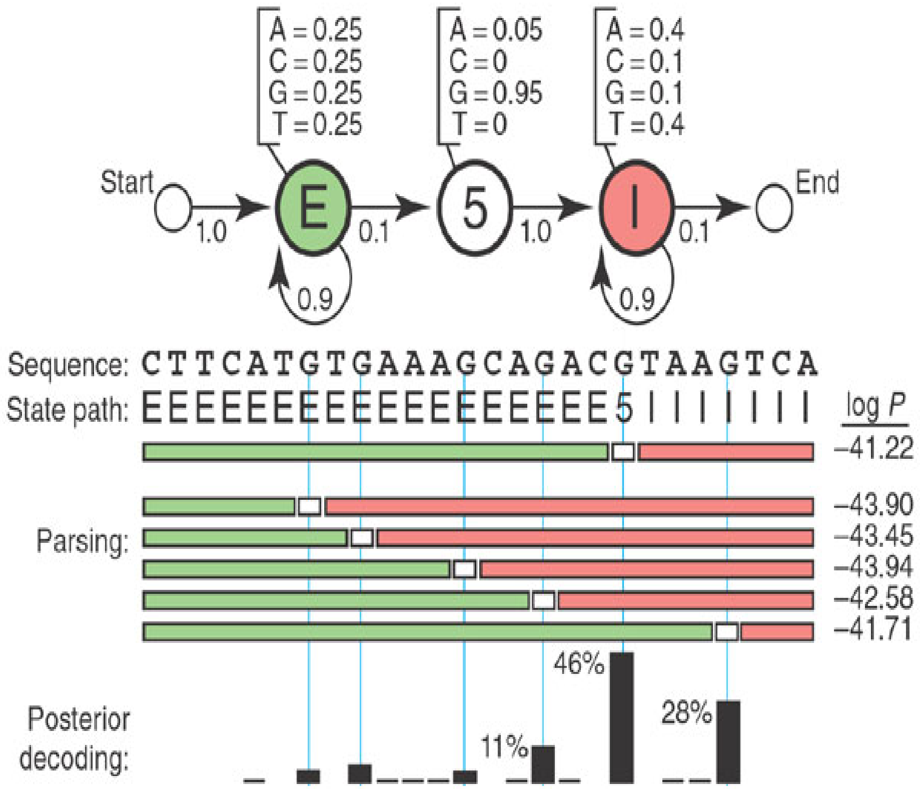


copyright 1996 M.W.King

Hidden Markov Models (HMMs) in Gene Prediction (4)

- Each of the elements
  - Exons, introns…= **states**
- The sequence characteristics & **syntactical constraints** (above two slides) allow **a transition probability** to be assigned
  - Indicating **how likely** a change of state is as one moves through the sequence, **base by base**
- **Hidden**
  - The user "sees" the nucleotide sequence **being analyzed**, but the user doesn't actually see **the states** that the individual bases are in

Hidden Markov Models (HMMs) in Gene Prediction (5)



- Each state emits a particular **kind of nucleotide sequence**, with **its own emission probability**
  - The **state emitting** the nucleotide is **hidden**
  - The *sequence itself is visible*
- The transition & **emission probabilities** are derived from **training sets**
  - Sequences for which the **correct gene structure** is already known

Hidden Markov Models (HMMs) in Gene Prediction (5)

- **Goal**
  - To develop **a set of parameters** that allows the method to be **fine tuned**
    - **Maximizing the chances that a correct prediction** is generated on **a new sequence of interest**
      - These parameters differ from organism to organism

- The success of any given HMM-based method depends on **how well** these **parameters** have been deduced from **the training set**

Programs Based on HMMs

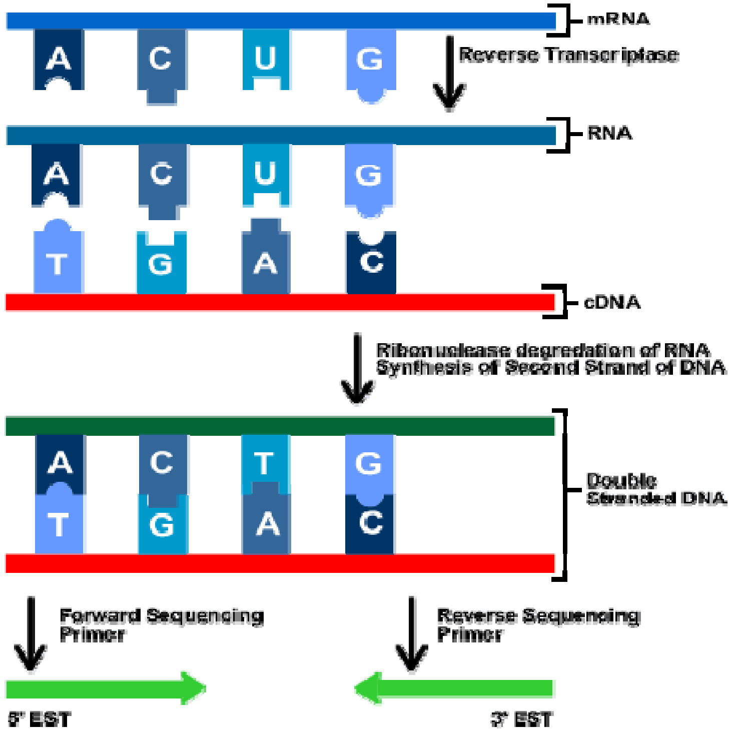
- To define **highly complex patterns**, e.g., **multigenic** genes
  - High **efficiency** in genome sequences
- GENSCAN
  - Burge & Karlin 1997
  - Annotation of **eukaryotic genomes**
- **GENIE**
  - Kulp et al. 1996
- **HMM gene**
  - Krogh 1997

Sequences Similarity-Based Prediction (1)

- Methods based on the **comparison** of **the genomic sequence** with **known coding sequences**
  - BLASTx (Gish & States 1993)
    - **ORFs** in prokaryotic genomes: useful

- The split nature of eukaryotic genes: BLASTx-like searches **do not** resolve **exon splice boundaries**
  - Solution: **combined BLASTx & *ab initio* methods**
    - GenomeScan (Yeh et al. 2001)
    - GeneID (Blanco et al. 2002)

Sequences Similarity-Based Prediction (2)


mRNA
A
C
U
G
Reverse Transcriptase
RNA
A
C
U
G
T
G
A
C
cDNA
Ribonuclease degredation of RNA
Synthesis of Second Strand of DNA
A
C
T
G
T
G
A
C
Double Stranded DNA
Forward Sequencing Primer
Reverse Sequencing Primer
5' EST
3' EST

- Expressed sequence tag (EST)
  - Valuable for identifying genes & delineating **exonic structure**
    - Alternative splicing forms
- **Example**
  - http://www.ncbi.nlm.nih.gov/mapview/modelmaker.cgi?taxid=9606&cntg=cntg&QSTR=1761[gene_id]&QUERY=uid(823789,13039333,11092128,14264426)&contig=NT_008413.17&from=831690&to=959090&strand=plus&with_est

http://www.ncbi.nlm.nih.gov/About/primer/est.html

Sequences Similarity-Based Prediction (3)

- **Mapping ESTs** to **genomic DNA sequences** with stringent parameters
  - BLAT (Kent 2002)
  - BLASTn (Altschul et al. 1990)

- **Disadvantages**
  - Exon boundaries not perfectly identified: a viable ORF is not identified

- **Specialized programs**
  - GRAIL-EXP
    - Using **splice site models**, provide a more clear solution to the problem

Sequences Similarity-Based Prediction (4)

- **Spliced alignments**
  - Aligning the genomic query against **a protein (or cDNA) target**, presumably homologous to the **protein encoded in the genomic sequence**
  - **Large gaps** corresponding to **introns** in the query sequence are only allowed at **"legal" splice junctions**
  - **Examples of programs**
    - SIM4 (Florea et al. 1998)
    - EST_GENOME (Mott 1997)
    - PROCRUSTES (Gelfand et al. 1996)
    - GENEWISE (Birney & Durbin 1997)

Comparative Gene Prediction (1)

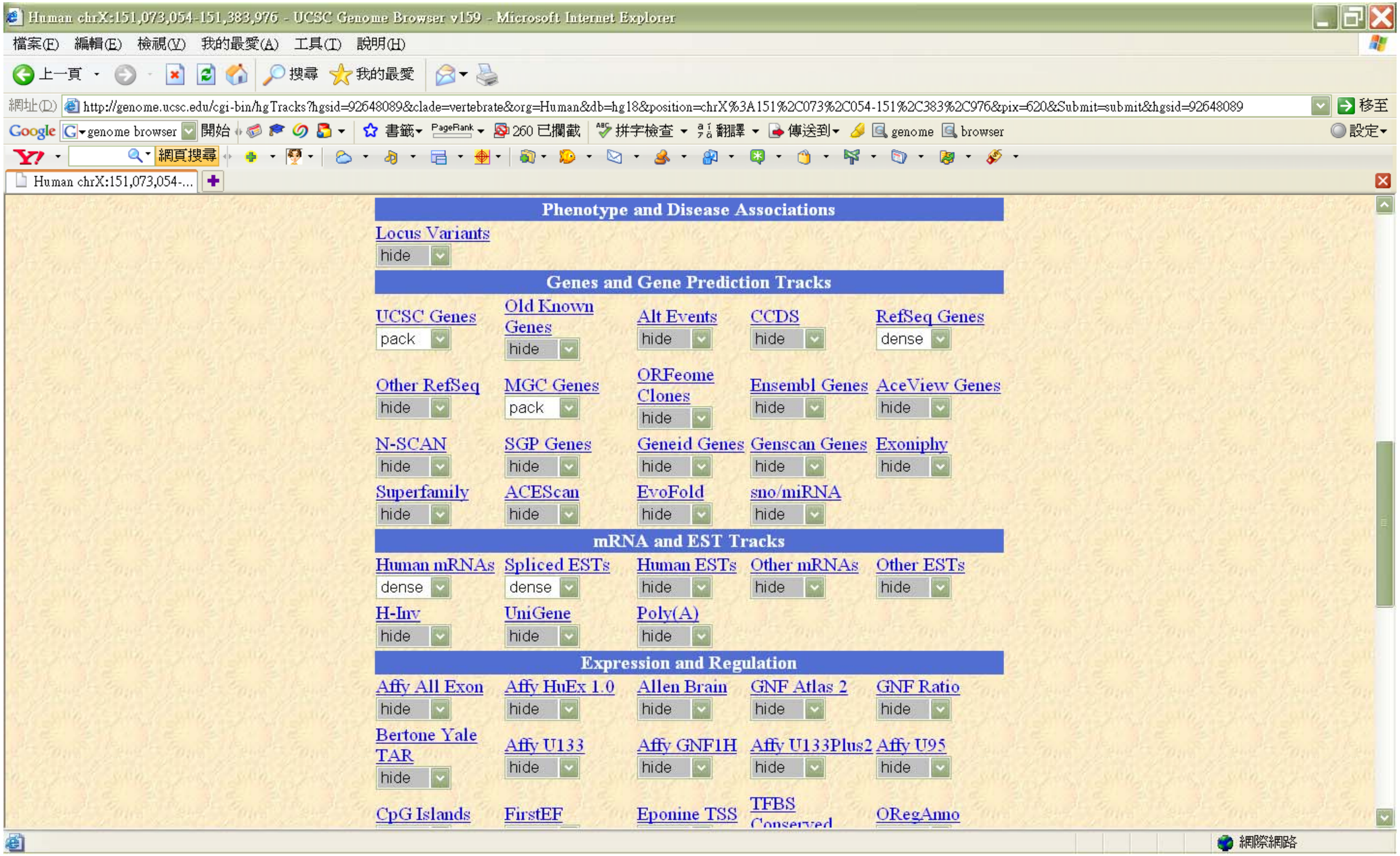
- **Rationale**
  - **Functional regions (protein-coding regions)** tend to be more **conserved** than non-protein-coding regions

- **Application**
  - To identify **protein-coding regions** in newly sequenced genomes

Comparative Gene Prediction (2)

- **Examples** for **mouse vs. human** comparative gene prediction
  - TWINSCAN (Korf et al. 2001)
    - An extension of GENSCAN (Annotation of **eukaryotic genomes)**
  - SGP-2 (Parra e tal. 2003)
    - An extension of GeneID (dynamic programming)
  - **SLAM** (Alexandersson et al. 2003)
    - **HMM-based method**: gene predictions & **sequence alignments** are performed simultaneously
- The **probability scores** calculated by each of these programs for **putative exons** are adjusted based on **comparative results**

Gene Prediction Programs – Cross-section



Human chrX:151,073,054-151,383,976 - UCSC Genome Browser v159 - Microsoft Internet Explorer
檔案(F) 編輯(E) 檢視(V) 我的最愛(A) 工具(T) 說明(H)
上一頁 搜尋 我的最愛
網址(D)
移至
Google genome browser 開始 書籤 PageRank 260 已攔截 拼字檢查 翻譯 傳送到 genome browser 設定
網頁搜尋
Human chrX:151,073,054-...
網際網路

http://genome.ucsc.edu/cgi-bin/hgTracks?hgsid=92648089&clade=vertebrate&org=Human&db=hg18&position=chrX%3A151%2C073%2C054-151%2C383%2C976&pix=620&Submit=submit&hgsid=92648089

Phenotype and Disease Associations

Locus Variants hide

Genes and Gene Prediction Tracks

UCSC Genes pack	Old Known Genes hide	Alt Events hide	CCDS hide	RefSeq Genes dense
Other RefSeq hide	MGC Genes pack	ORFeome Clones hide	Ensembl Genes hide	AceView Genes hide
N-SCAN hide	SGP Genes hide	Geneid Genes hide	Genscan Genes hide	Exoniphy hide
Superfamily hide	ACEScan hide	EvoFold hide	sno/miRNA hide	

mRNA and EST Tracks

Human mRNAs dense	Spliced ESTs dense	Human ESTs hide	Other mRNAs hide	Other ESTs hide
H-Inv hide	UniGene hide	Poly(A) hide		

Expression and Regulation

Affy All Exon hide	Affy HuEx 1.0 hide	Allen Brain hide	GNF Atlas 2 hide	GNF Ratio hide
Bertone Yale TAR hide	Affy U133 hide	Affy GNF1H hide	Affy U133Plus2 hide	Affy U95 hide
CpG Islands	FirstEF	Eponine TSS	TFBS Conserved	ORegAnno

GRAIL (1)

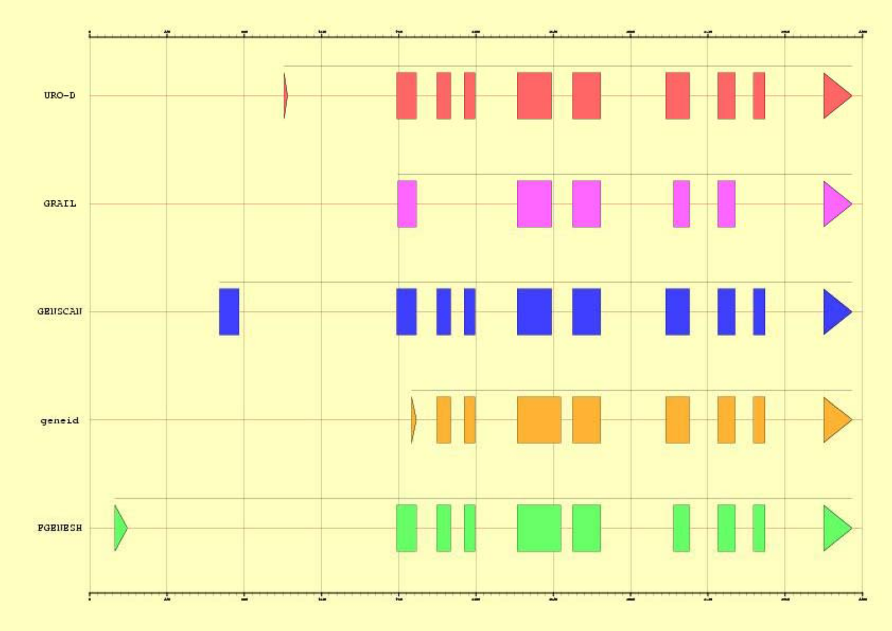
- **The Gene Recognition and Analysis Internet Link (GRAIL)**
  - Uberbacher & Mural 1991
  - To calculate the likelihood that a particular position is within a coding region by computing and integrating seven separate coding statistic measures

- **GRAIL2 (Xu et al 1994)**
  - Incorporation of information about different splice and translational signals,

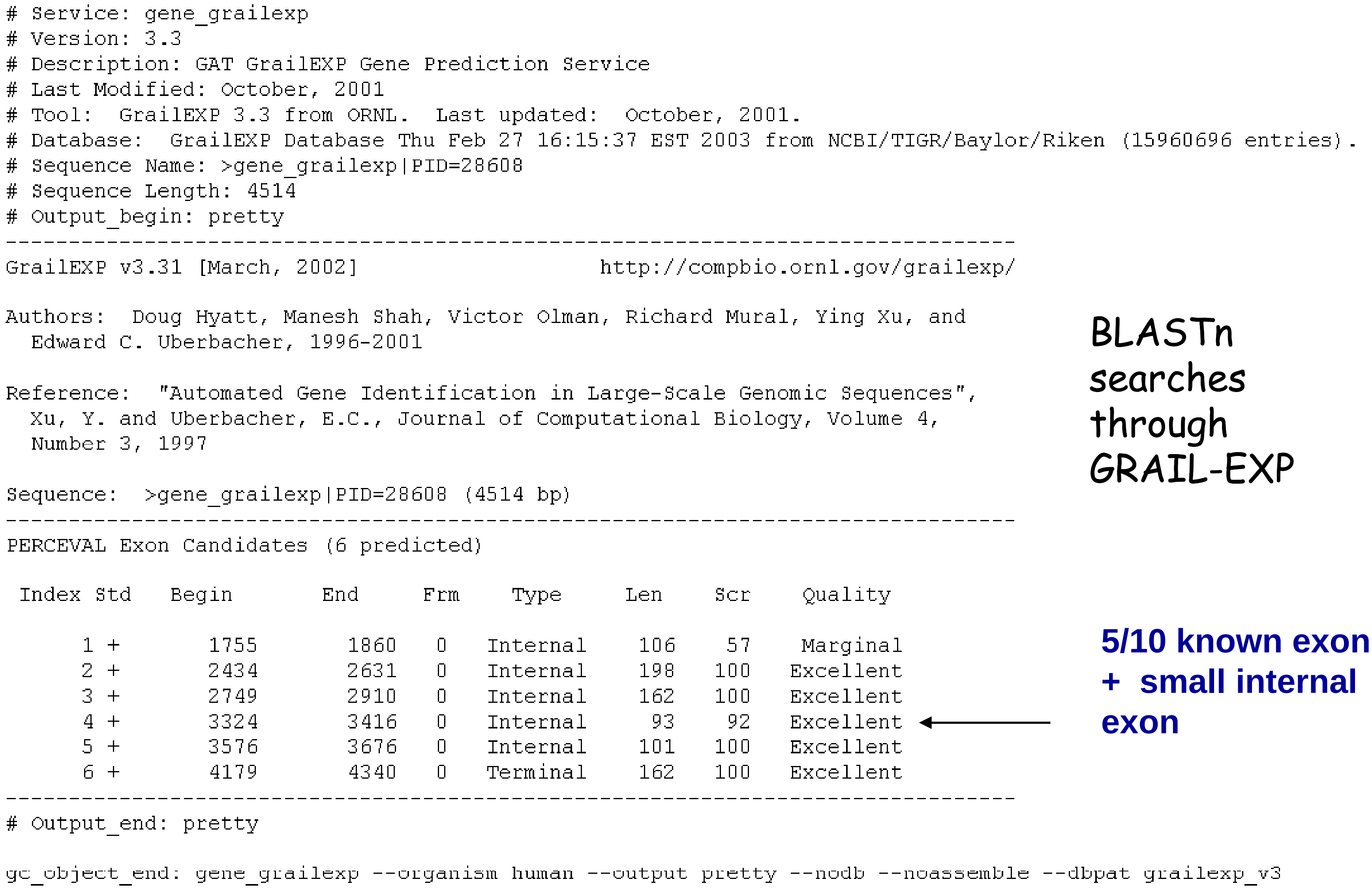
- **GRAIL-EXP (Xu & Uberbacher 1997)**
  - Incorporation of homology information
    - BLASTn searches against a database of partial & complete transcripts (ESTs)

GRAIL (2)

- **Outputs**
  - A profile along the length of the query sequence, peaks correspond to coding regions
- Example
  - The human UROD gene
  - U30787
    - FASTA format
    - An SP1 binding site, TATA box, 10 exons have been annotated to this sequence
    - Full length: 4,514 bp



EMBL annotation and genes predicted by **Grail**, **GENSCAN**, **geneid** and **FGENESH** in the sequence U30787. **First exon** is always missed in the predictions and there are some problems **to detect the donor site from exon 5**. Detection of **start codons** is a serious drawback in current gene finding programs. However, this problem can be overcome by using **homology information** to complete the gene prediction.

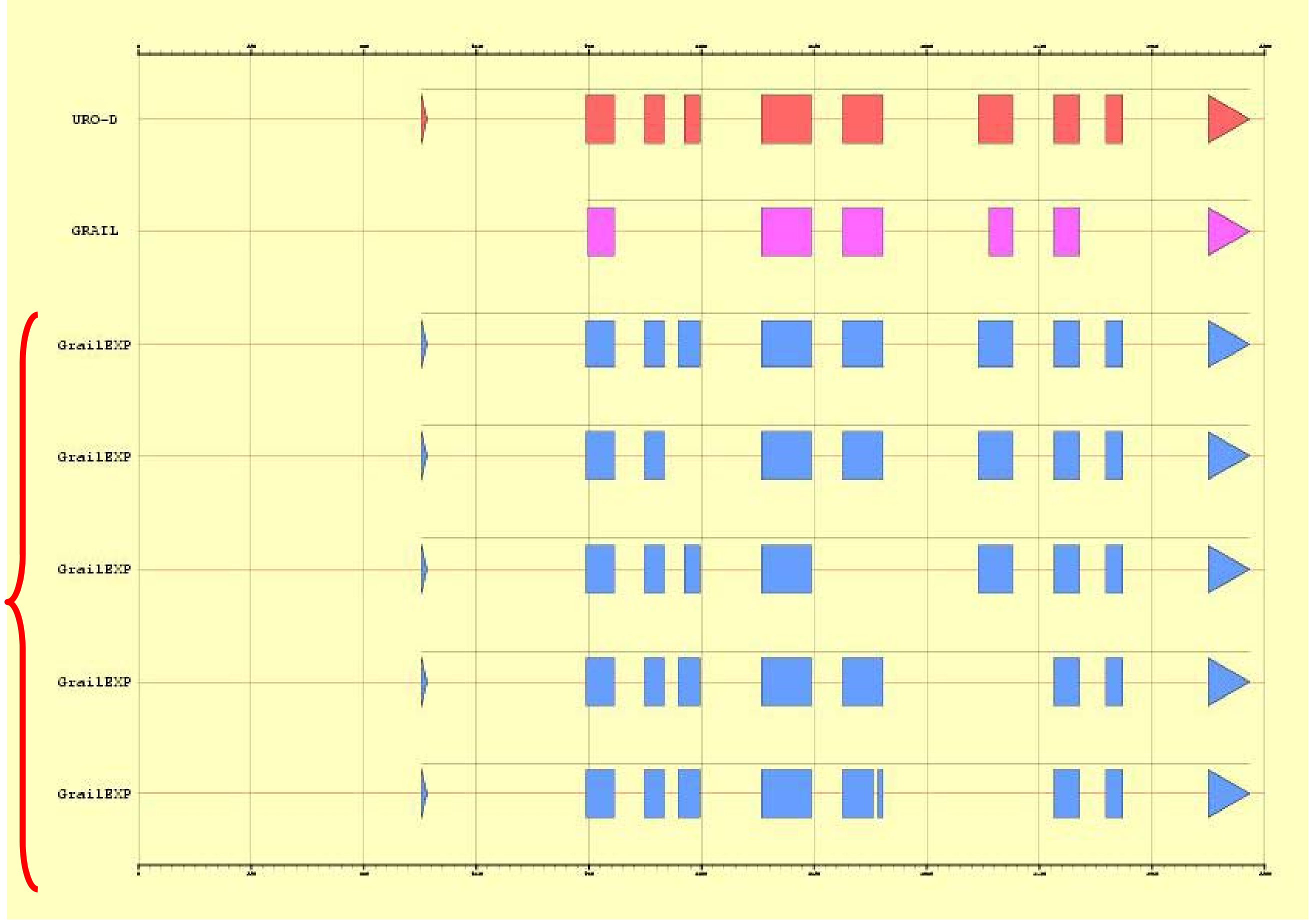


```
# Service: gene_grailexp
# Version: 3.3
# Description: GAT GrailEXP Gene Prediction Service
# Last Modified: October, 2001
# Tool:  GrailEXP 3.3 from ORNL.  Last updated:  October, 2001.
# Database:  GrailEXP Database Thu Feb 27 16:15:37 EST 2003 from NCBI/TIGR/Baylor/Riken (15960696 entries).
# Sequence Name: >gene_grailexp|PID=28608
# Sequence Length: 4514
# Output_begin: pretty
--------------------------------------------------------------------------------
GrailEXP v3.31 [March, 2002]                   http://compbio.ornl.gov/grailexp/

Authors:  Doug Hyatt, Manesh Shah, Victor Olman, Richard Mural, Ying Xu, and
  Edward C. Uberbacher, 1996-2001

Reference:  "Automated Gene Identification in Large-Scale Genomic Sequences",
  Xu, Y. and Uberbacher, E.C., Journal of Computational Biology, Volume 4,
  Number 3, 1997

Sequence:  >gene_grailexp|PID=28608 (4514 bp)
--------------------------------------------------------------------------------
PERCEVAL Exon Candidates (6 predicted)

 Index Std   Begin       End     Frm    Type     Len    Scr    Quality

      1 +      1755      1860   0   Internal   106    57    Marginal
      2 +      2434      2631   0   Internal   198   100   Excellent
      3 +      2749      2910   0   Internal   162   100   Excellent
      4 +      3324      3416   0   Internal    93    92   Excellent
      5 +      3576      3676   0   Internal   101   100   Excellent
      6 +      4179      4340   0   Terminal   162   100   Excellent
--------------------------------------------------------------------------------
# Output_end: pretty

gc_object_end: gene_grailexp --organism human --output pretty --nodb --noassemble --dbpat grailexp_v3
```

BLASTn searches through GRAIL-EXP

5/10 known exons + small internal exon

EMBL annotation vs. Gene Predicted by GRAIL & GRAIL-EXP



Five alternative predictions supported by **ESTs information**

GeneID (1)

- A program that predicts genes in **genomic sequences** using a **hierarchical approach**
  - Guigo et al. 1992; Parra et al. 2000

- Incorporation of **new information** in most recently version (Blanco et al. 2002)
  - Sequence **similarity**
  - **Experimental** data
  - Data from **other computational predictions**

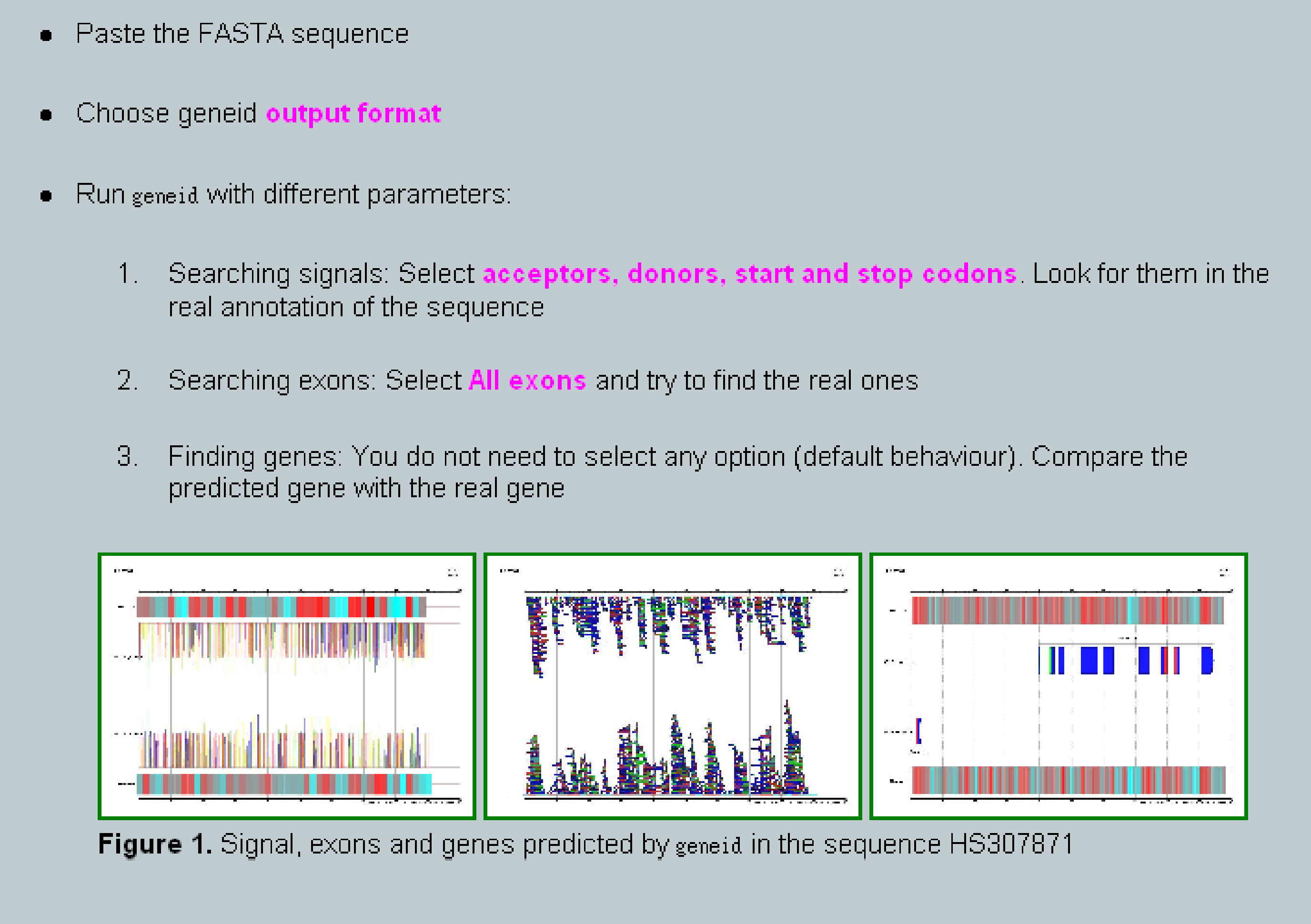
GeneID (2)

- **Step 1**
  - **Position weight matrices (PWM)**: prediction of **splice sites**, **start**, **stop** codons, **score** given

- **Step 2**
  - **Exons** are built from identified **"defining sites"** **(step 1), score** given
  - Exons are scored = sum of the scores of **the defining sites** + the score of **their coding potential**

- **Step 3**
  - Based on the set of **predicted exons**, the **gene structure is assembled**, predicting **the most likely gene structure** by **maximizing the sum of the scores** of the assembled exons

GeneID (3) - Output



- Paste the FASTA sequence
- Choose geneid **output format**
- Run geneid with different parameters:

  1. Searching signals: Select **acceptors, donors, start and stop codons**. Look for them in the real annotation of the sequence
  2. Searching exons: Select **All exons** and try to find the real ones
  3. Finding genes: You do not need to select any option (default behaviour). Compare the predicted gene with the real gene

Figure 1. Signal, exons and genes predicted by geneid in the sequence HS307871

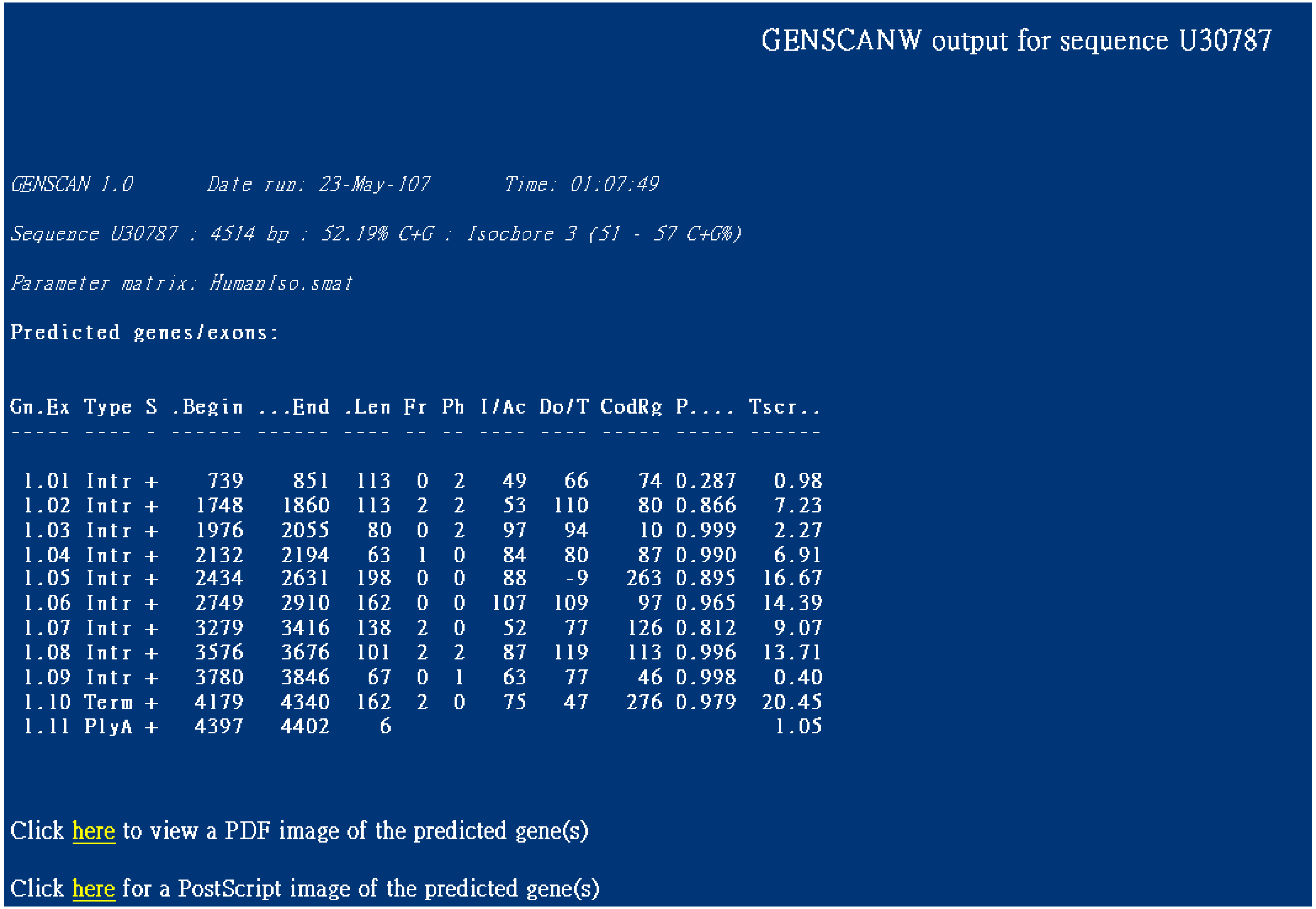
http://genome.imim.es/courses/Madrid04/exercises/genefinding1/index.html

GENESCAN (1)

- A general purpose **eukaryotic gene** prediction program
  - **Hidden Morkov Model**
    - Donor splice site modeling, ***maximal dependence decomposition***
      - **A series of weight matrices** (instead of just one) are used to capture **dependencies** between positions in these splice sites
  - **Parameters**
    - Accounting for many **higher-order properties of genomic sequences**
      - E.g., **typical gene density**, typical number of exons per gene & **the distribution of exon sizes** for different types of exons
    - **Separate sets of gene model parameters** can be used to adjust for the differences in gene density and G+C composition seen **across genomes**
  - Vertebrate, maize & Arabidopsis sequences

GENESCAN (2)

- GenomeSCAN
  - Yeh et al. 2001
  - An extension of GENESCAN
  - Incorporations of **sequence similarity to known proteins using BLASTx**
    - Higher scores for exons exhibiting **similarity to known proteins**
    - Decreased scores for **predicted exons** having little to no similarity with known proteins



GENSCANW output for sequence U30787

```
GENSCAN 1.0       Date run: 23-May-107       Time: 01:07:49

Sequence U30787 : 4514 bp : 52.19% C+G : Isochore 3 (51 - 57 C+G%)

Parameter matrix: HumanIso.smat

Predicted genes/exons:


Gn.Ex Type S .Begin ...End .Len Fr Ph I/Ac Do/T CodRg P.... Tscr..
----- ---- - ------ ------ ---- -- -- ---- ---- ----- ----- ------

 1.01 Intr +    739    851  113  0  2   49   66    74 0.287   0.98
 1.02 Intr +   1748   1860  113  2  2   53  110    80 0.866   7.23
 1.03 Intr +   1976   2055   80  0  2   97   94    10 0.999   2.27
 1.04 Intr +   2132   2194   63  1  0   84   80    87 0.990   6.91
 1.05 Intr +   2434   2631  198  0  0   88   -9   263 0.895  16.67
 1.06 Intr +   2749   2910  162  0  0  107  109    97 0.965  14.39
 1.07 Intr +   3279   3416  138  2  0   52   77   126 0.812   9.07
 1.08 Intr +   3576   3676  101  2  2   87  119   113 0.996  13.71
 1.09 Intr +   3780   3846   67  0  1   63   77    46 0.998   0.40
 1.10 Term +   4179   4340  162  2  0   75   47   276 0.979  20.45
 1.11 PlyA +   4397   4402    6                                1.05
```

Click here to view a PDF image of the predicted gene(s)

Click here for a PostScript image of the predicted gene(s)

- Gn. Ex: gene exon no.; **Type**: exon type or an identified poly A; **S**: the strand; Fr: frame; several scoring columns; **P**: probability value: P>0.99 are 97.7% accurate when the prediction matches a true, annotated exon; **0.50 to 0.99** are deemed to be correct most of the time; **9/10 correct**

FGENES (1)

- **FGENES="Find genes"**
  - 1st version: Solovyev et al. 1995
  - **Linear discriminant analysis** to identify **splice sites**, **exons**, and **promoter** elements
  - **Filtered exons** are assembled using **a dynamic programming algorithm** that searches paths of compatible exons, with the goal of maximizing the final gene score
- **FGENESH**
  - An **HMM-based** variant of FGENES

FGENES (2)

- **FGENESH+**
  - + **protein homology** (Salamov & Solovyev 2000)
- **& FGENESH-C**
  - + **cDNA homology** (Salamov & Solovyev 2000)
- Using information of **known genes & DNA sequences**
  - Better power

Discriminant Analysis in Gene Prediction (1)

- To discriminate **two or more naturally occurring groups**
  - Zhang 1997

- In the area of gene prediction, the **observables**
  - Try to discriminate whether a **particular stretch of DNA** is found in either **an intron** or **an exon** could include the presence of putative acceptor sites, donor sites, or start and stop codons

  - Two observables
    - **Splice site scores** and exon length are plotted against each other on a simple XY graph
    - Two different symbols = **two different groups**
      - **X= exon; circle= intron**

Discriminant Analysis in Gene Prediction (2)

- **Two different types of discriminant analysis** could be applied to try to separate the two states from one another
  - Linear discriminant vs. **quadratic discriminant analysis**

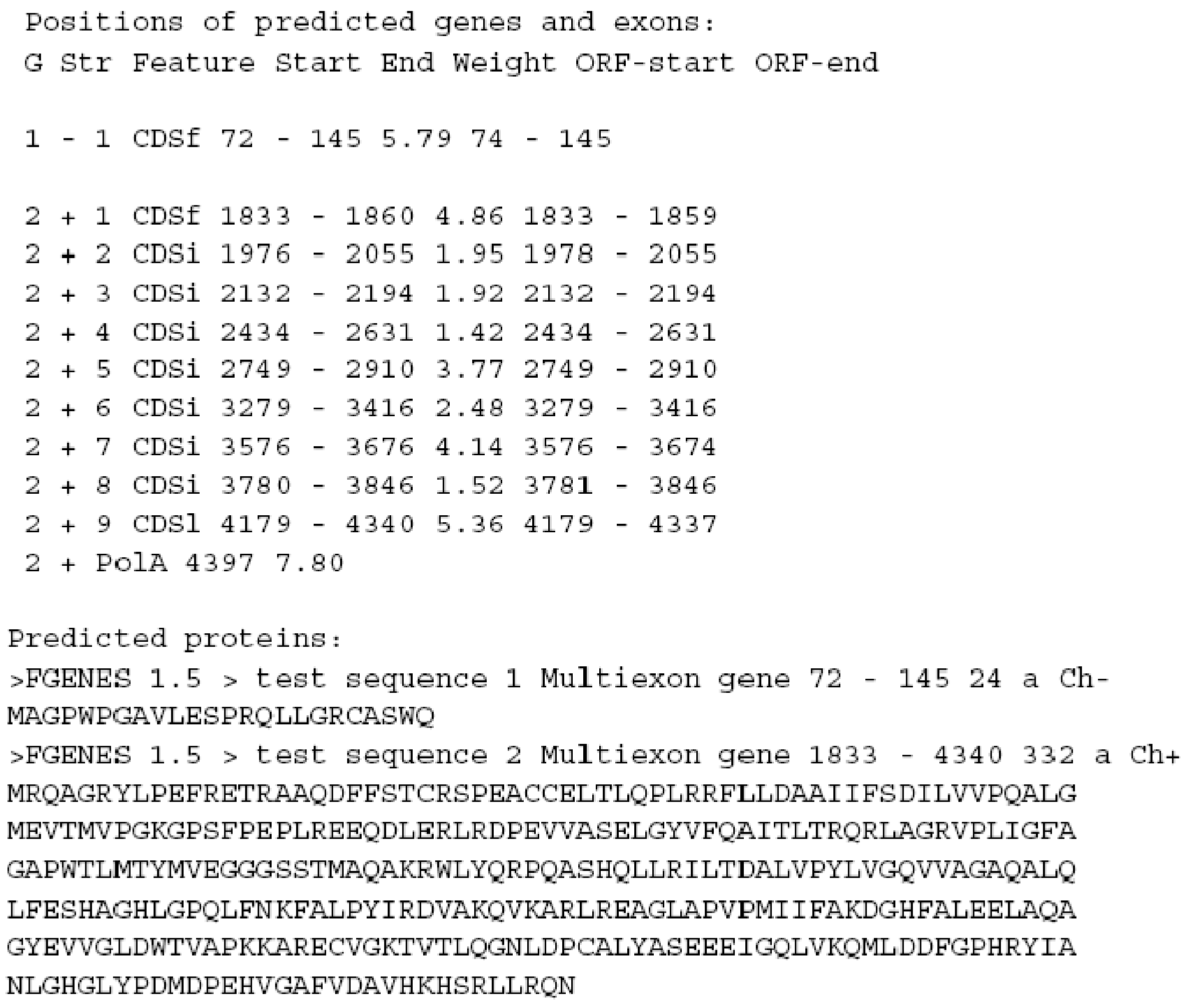
- The relationship between these two sets of observables
  - **Nonlinear or multivariate**, the resulting graph looks like a swarm of points

  - A linear function L(x) cannot adequately separate the two states
    - An appropriable number of points have been misclassified

  - The quadratic function Q(x) is capable of **completely separating the two groups in this case**

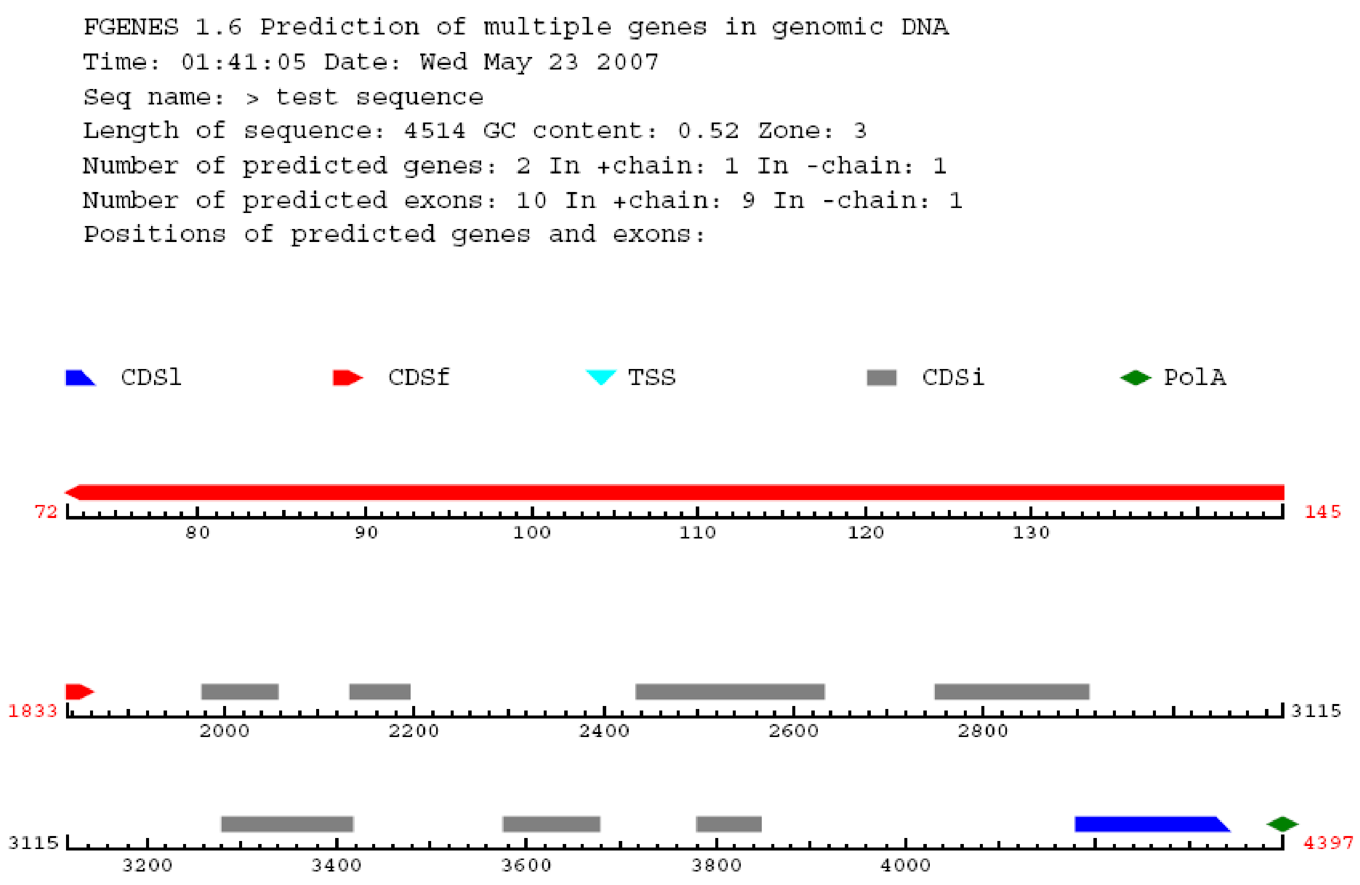
FGENESH - Output

- G=gene number;
- Strand;
- The exon number within the gene;
- The exon type; f=first; i=internal; l=last;
- The start and stop positions for the exon;
- An exon score;
- ORF start and stop positions;



```
 Positions of predicted genes and exons:
 G Str Feature Start End Weight ORF-start ORF-end

 1 - 1 CDSf 72 - 145 5.79 74 - 145

 2 + 1 CDSf 1833 - 1860 4.86 1833 - 1859
 2 + 2 CDSi 1976 - 2055 1.95 1978 - 2055
 2 + 3 CDSi 2132 - 2194 1.92 2132 - 2194
 2 + 4 CDSi 2434 - 2631 1.42 2434 - 2631
 2 + 5 CDSi 2749 - 2910 3.77 2749 - 2910
 2 + 6 CDSi 3279 - 3416 2.48 3279 - 3416
 2 + 7 CDSi 3576 - 3676 4.14 3576 - 3674
 2 + 8 CDSi 3780 - 3846 1.52 3781 - 3846
 2 + 9 CDSl 4179 - 4340 5.36 4179 - 4337
 2 + PolA 4397 7.80

Predicted proteins:
>FGENES 1.5 > test sequence 1 Multiexon gene 72 - 145 24 a Ch-
MAGPWPGAVLESPRQLLGRCASWQ
>FGENES 1.5 > test sequence 2 Multiexon gene 1833 - 4340 332 a Ch+
MRQAGRYLPEFRETRAAQDFFSTCRSPEACCELTLQPLRRFLLDAAIIFSDILVVPQALG
MEVTMVPGKGPSFPEPLREEQDLERLRDPEVVASELGYVFQAITLTRQRLAGRVPLIGFA
GAPWTLMTYMVEGGGSSTMAQAKRWLYQRPQASHQLLRILTDALVPYLVGQVVAGAQALQ
LFESHAGHLGPQLFNKFALPYIRDVAKQVKARLREAGLAPVPMIIFAKDGHFALEELAQA
GYEVVGLDWTVAPKKARECVGKTVTLQGNLDPCALYASEEEIGQLVKQMLDDFGPHRYIA
NLGHGLYPDMDPEHVGAFVDAVHKHSRLLRQN
```

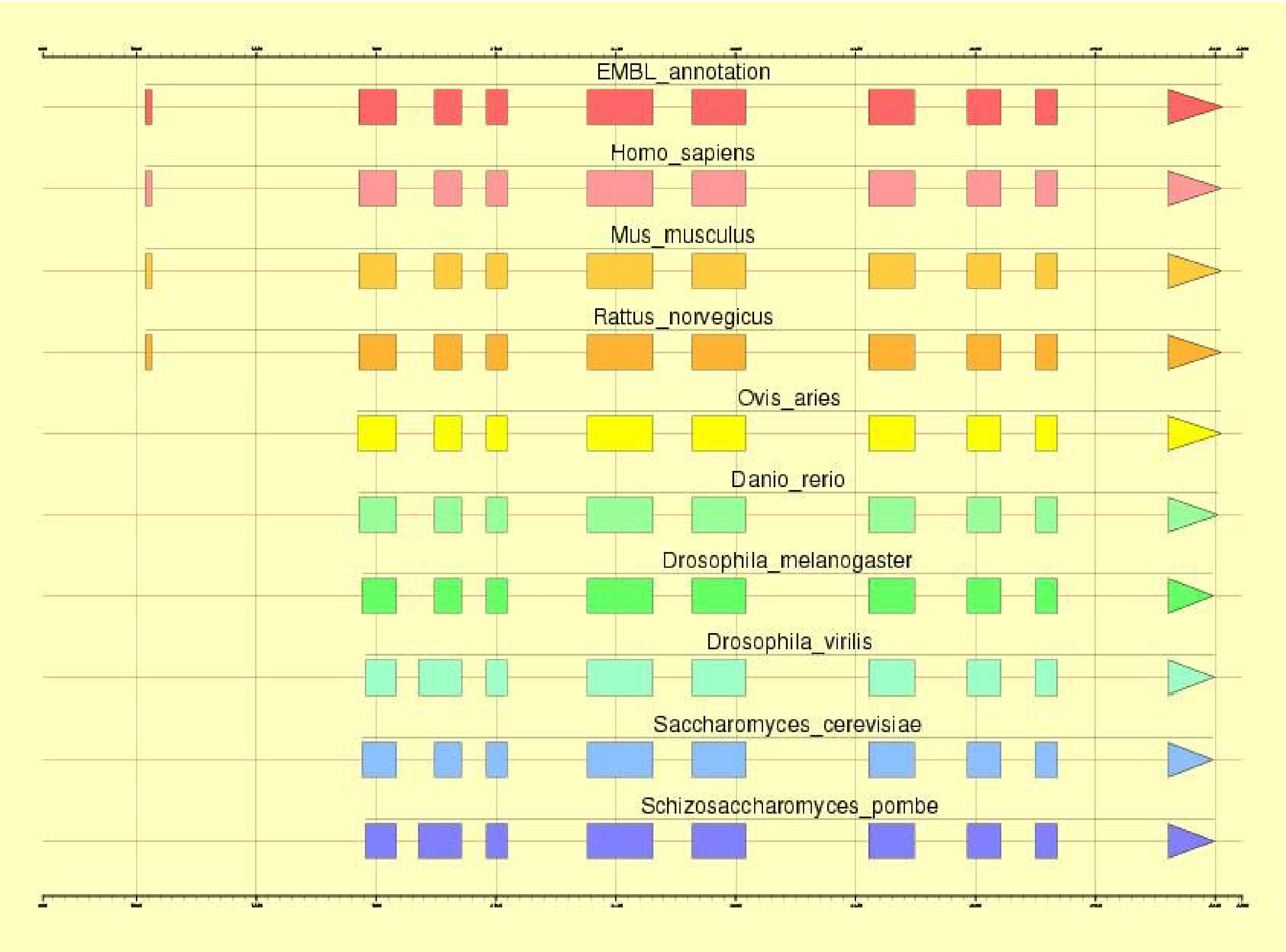
FGENESH - Output



```
FGENES 1.6 Prediction of multiple genes in genomic DNA
Time: 01:41:05 Date: Wed May 23 2007
Seq name: > test sequence
Length of sequence: 4514 GC content: 0.52 Zone: 3
Number of predicted genes: 2 In +chain: 1 In -chain: 1
Number of predicted exons: 10 In +chain: 9 In -chain: 1
Positions of predicted genes and exons:
```

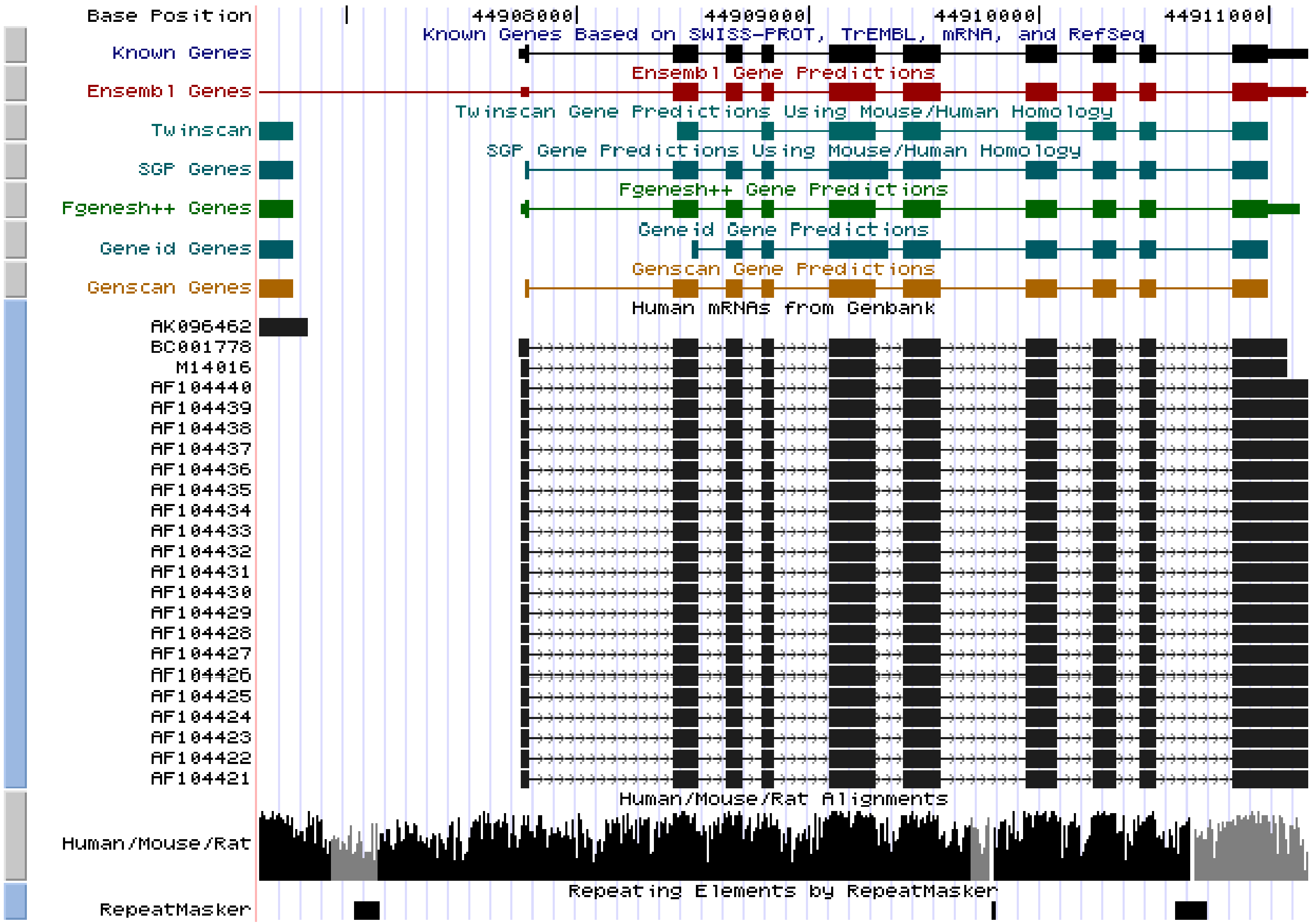
GENEWISE

- To compare **a genomic sequence** with **a protein sequence** or with **an HMM representing a protein domain**
  - At protein level while maintaining the reading frame, **regardless** of intervening introns or sequence errors that may cause frameshifts
- **Gene prediction + a homology comparison**
- Computationally **expensive** and **accurate prediction** requires the presence of **a close, homologous protein**

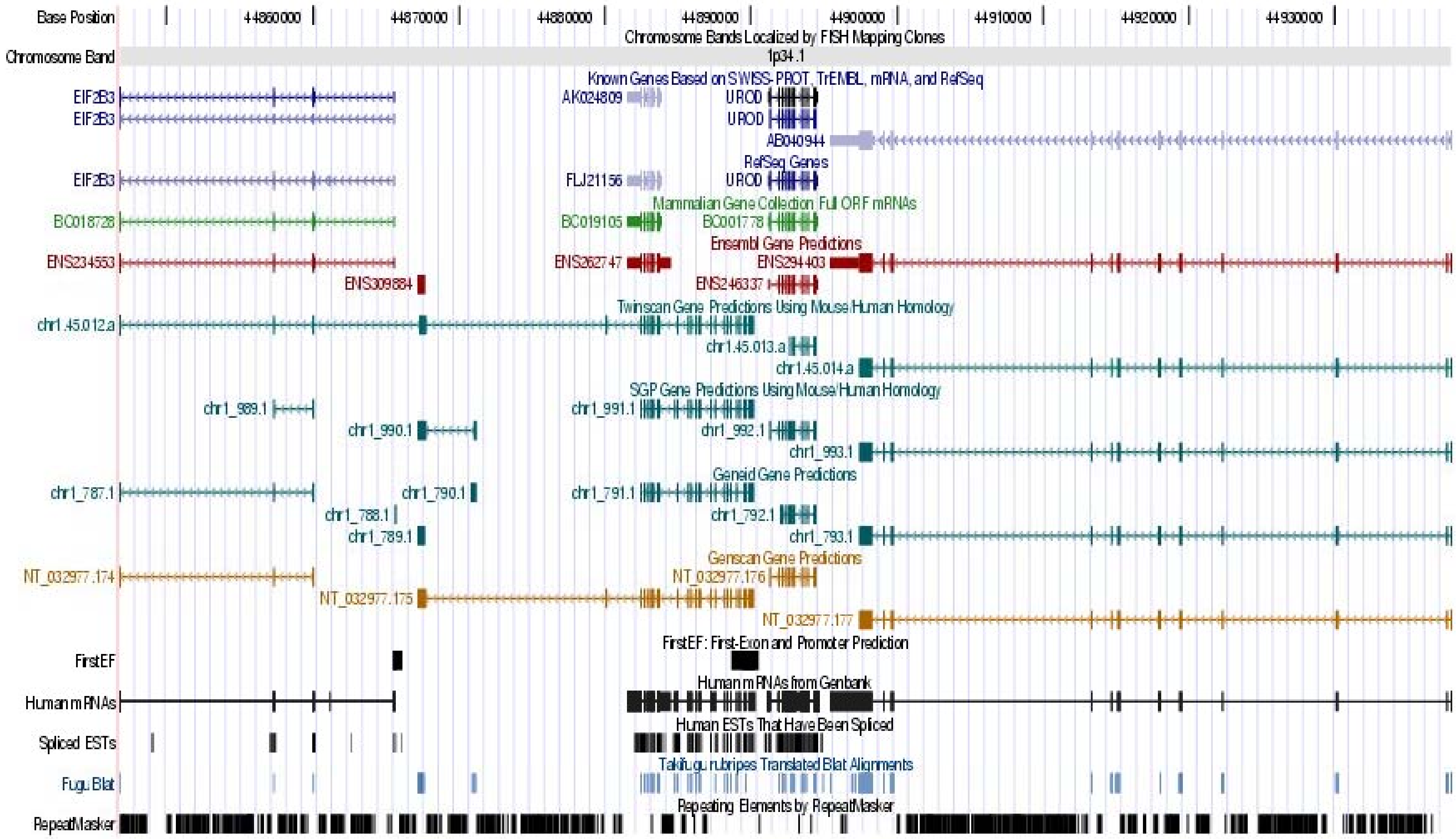


GFF2PS software

- The results of GENEWISE predictions **when progressively distant homologs** of the UROD protein are used – **POWERFUL** (in EBI)



(a) UCSC genome browser representation of the region containing the gene *uroporphyrinogen decarboxylase (URO-D)*



(b) UCSC genome browser representation of the context (100Kbps) region around the gene *uroporphyrinogen decarboxylase (URO-D)*.

How Well Do the Methods Work? (1)

- **Different methods** can produce different, & sometimes, **contradictory** results
- **Factors affecting**
  - **Species**
  - The sequence **context**
  - The existence of experimental evidence
    - **Spliced ESTs**: strong supports
- **Consistent predictions by different methods**

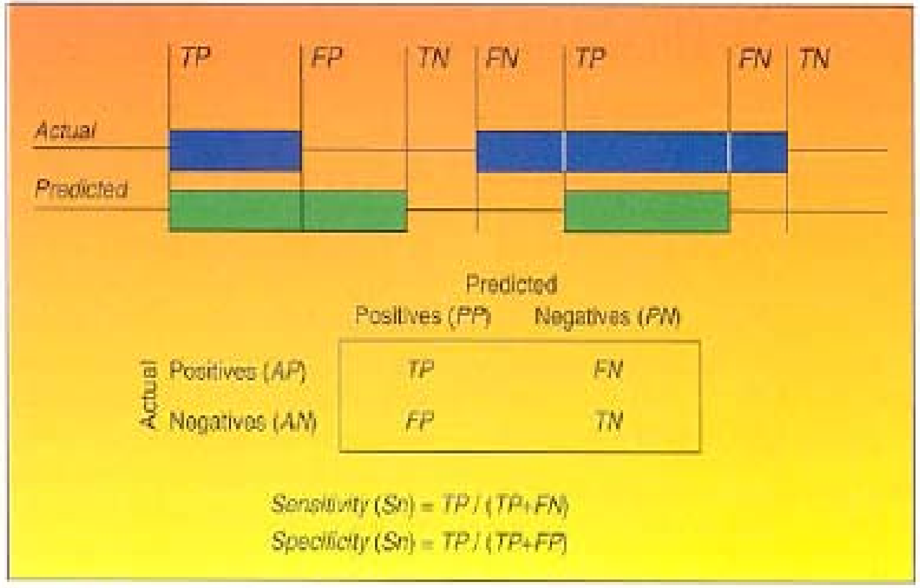
How Well Do the Methods Work? (2)

- **The reliability**
  - The accuracy of gene prediction program is usually determined using **controlled, defined data sets**
    - Comparing the prediction made by a method with the actual gene structure, determined experimentally
  - Two basic measure, a perfect prediction Sn=1; Sp=1, neither one alone provide a good measure of global accuracy
    - **Sensitivity (Sn) (0~1)**
      - The proportion of coding nucleotides, exons, or genes that have been predicted **correctly**
    - **Specificity (Sp) (0~1)**
      - The proportion of predicted coding nucleotides, exons, or genes that are **real** (the overall fraction of the prediction that is correct)

How Well Do the Methods Work? (3)

- The **reliability**
  - Correction coefficient (CC)
    - (worst) -1~1 (perfect prediction)
  - **A combined measure** of the Sn and Sp values

$$cc = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP) \times (TP+FN) \times (TN+FP) \times (TN+FN)}}$$



		Predicted	
		Positives (PP)	Negatives (PN)
Actual	Positives (AP)	TP	FN
	Negatives (AN)	FP	TN

Sensitivity (Sn) = TP / (TP+FN)

Specificity (Sn) = TP / (TP+FP)

FIGURE 5.11 Schematic representation of measures of gene prediction accuracy at the nucleotide level. In the upper portion of the figure, the four possible outcomes of a prediction are shown: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). The matrix at the bottom of the figure shows how both sensitivity and specificity are determined from these four possible outcomes, giving a tangible measure of the effectiveness of any gene prediction method. (Adapted from Burset & Guigó, 1996; Snyder & Stormo, 1997).

Dunham et al. 1999

- Chr. 22
  - Comparisons of a number of *ab initio* and comparative gene finders vs. **curated, manual annotation**

- The accuracy of *ab initio* gene finders substantially suffers when moving up in **complexity** from single gene sequence to genome-scale sequence data
  - GENSCAN CC=0.64; SGP2 CC=0.73

How Well Do the Methods Work? (3)

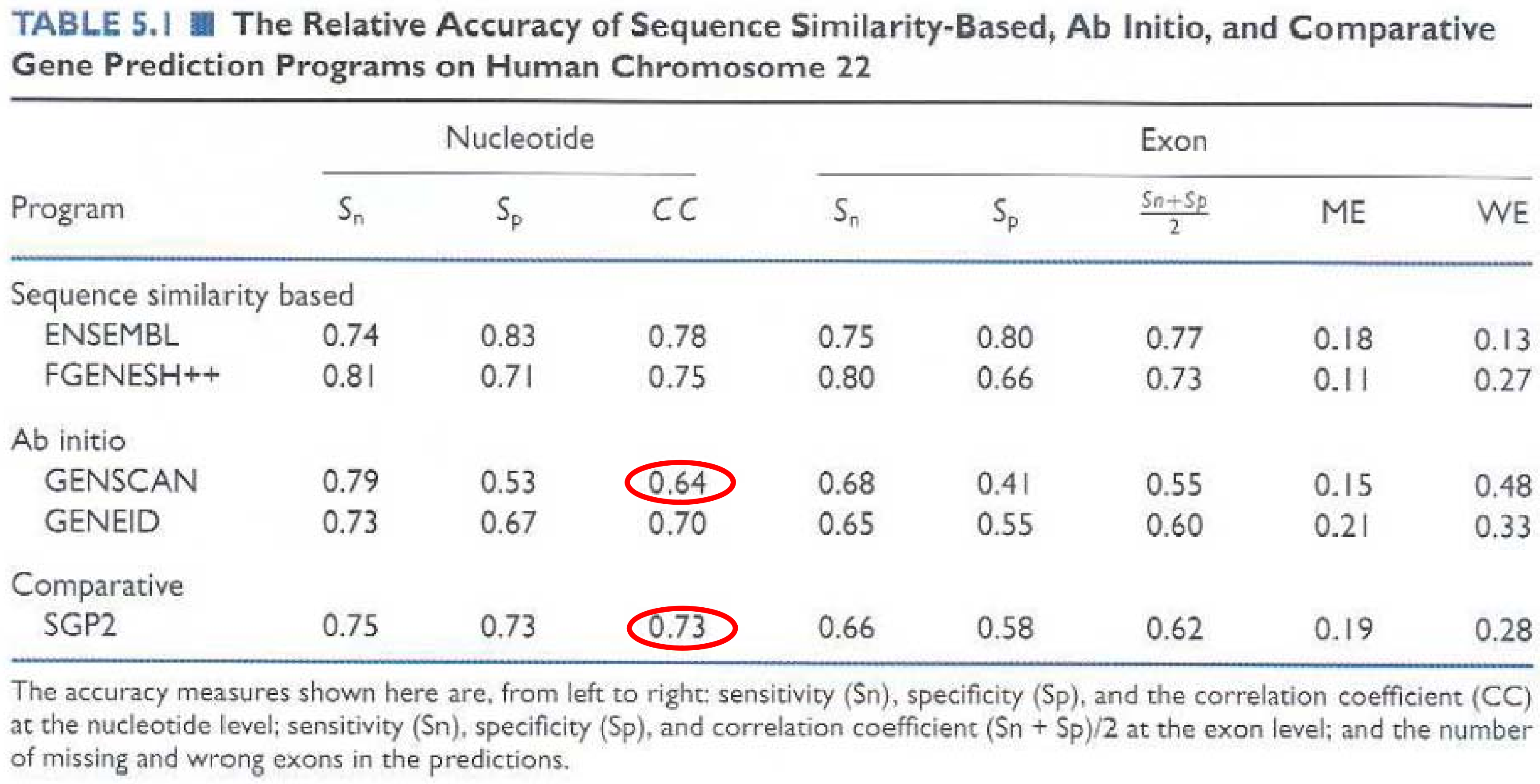


TABLE 5.1 The Relative Accuracy of Sequence Similarity-Based, Ab Initio, and Comparative Gene Prediction Programs on Human Chromosome 22

	Nucleotide			Exon				
Program	S_n	S_p	CC	S_n	S_p	$\frac{Sn+Sp}{2}$	ME	WE
Sequence similarity based								
ENSEMBL	0.74	0.83	0.78	0.75	0.80	0.77	0.18	0.13
FGENESH++	0.81	0.71	0.75	0.80	0.66	0.73	0.11	0.27
Ab initio								
GENSCAN	0.79	0.53	0.64	0.68	0.41	0.55	0.15	0.48
GENEID	0.73	0.67	0.70	0.65	0.55	0.60	0.21	0.33
Comparative								
SGP2	0.75	0.73	0.73	0.66	0.58	0.62	0.19	0.28

The accuracy measures shown here are, from left to right: sensitivity (Sn), specificity (Sp), and the correlation coefficient (CC) at the nucleotide level; sensitivity (Sn), specificity (Sp), and correlation coefficient (Sn + Sp)/2 at the exon level; and the number of missing and wrong exons in the predictions.

Dunham et al. 1999

Exercise

- http://genome.imim.es/courses/Madrid04/exercises/genefinding1/index.html