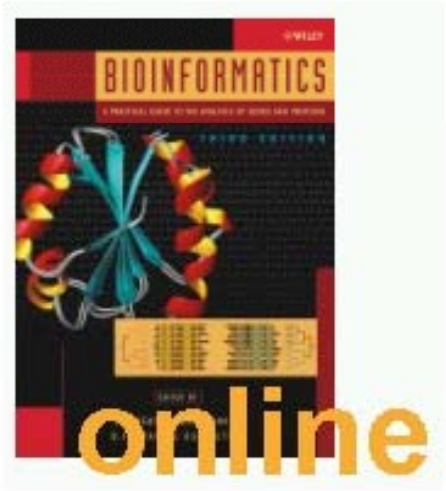
3. Predictive Methods Using DNA Sequences (2)

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



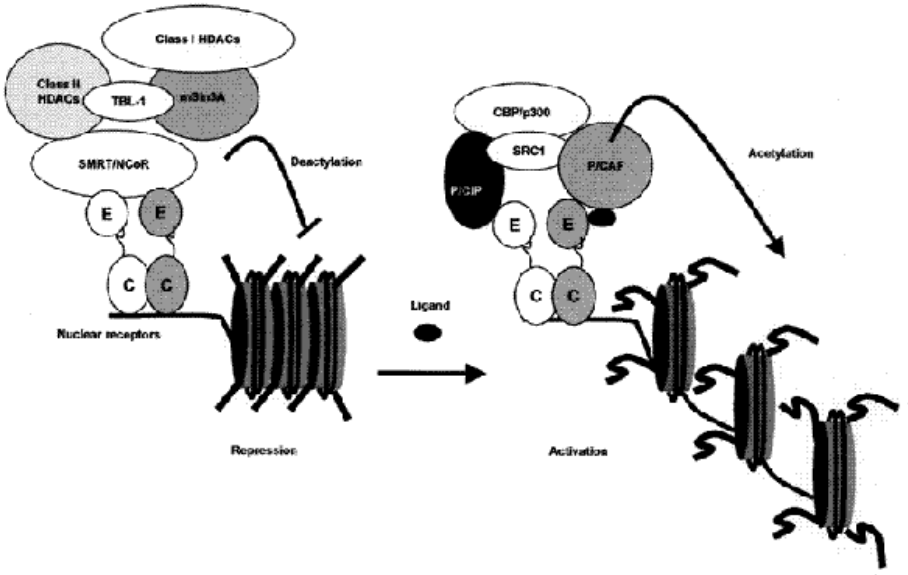
Select a Chapter: Chapter 5

Chapter 5: Predictive Methods Using DNA Sequences

- Sample Data for Problem Sets
- Internet Resources

Promoter Analysis: Characterization & Prediction

- **Regulation of gene expression**
  - Compaction of chromatin
  - **Transcriptional initiation** (*****)
    - To ensure no superfluous intermediates are synthesized
  - Polyadenylation
  - Splicing
  - mRNA stability
  - Translation initiation
  - The control of **protein activity**



Promoters (1)

- **Functional regions** immediately upstream or downstream of a **transcription start site (TSS)**
  - Immediately involved in the regulation of transcription

- The structure of **a promoter region**
  - A specific arrangement of transcription factor binding sites
    - **Regulatory** or **promoter** elements (Fickett & Hatzigeorgiou 1997)

  - **Core promoter** (next slide)
    - The region of the promoter **near the TSS where RNA polymerase II binds** is known as the
      - An **universal structure** that is responsible for basal gene transcription
        - Zhang 1998

Promoters (2)

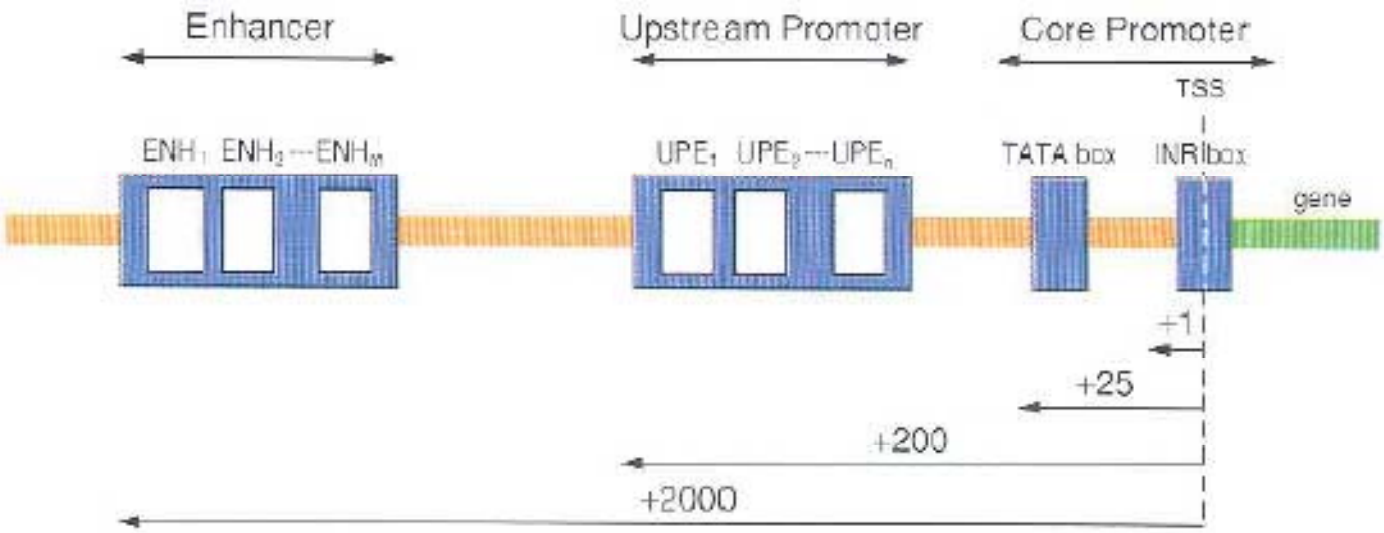


FIGURE 5.12 A schematic representation of a typical promoter. The core of the promoter consists on the TATA box and an initiator sequence. Upstream and downstream promoters (not shown here) are comprised of several binding sites or composites. Enhancers are regulatory regions distant from the regulated gene.

Promoters (3)

- **Transcriptional enhancers**
  - Nondirectional promoter regions
    - Several Kb away from the regulated gene
- **Experimental identification** of promoter regions at a genomic scale is **extremely**
  - Laborious
  - Expensive
- **Computational methods** may play an important role in the annotation of these sequences
  - Pennacchio & Rubin 2001

Promoters (4)

- **Gene finding & promoter prediction**
  - A correct **promoter** prediction ⇨ to improve gene prediction
    - A way to define better the boundaries of a gene
  - A correct **gene** prediction automatically ⇨ better predictions of upstream promoter regions
- Two problems in promoter predictions
  - **TSS** & the promoter regions (sites)
  - The detection of **transcription factor binding motifs (binding)**

Algorithms – Branzma et al. 1998

- **Pattern-driven algorithms**
  - Search for **known regulatory patterns** in genomic sequences

- **Sequence-driven algorithms**
  - To discover **unknown pattern** from sets of sequences that are functionally related

Pattern-Driven Algorithms (1)

- **The availability of collections of experimentally annotated binding sites**
  - **TRANSFAC** (Matys et al. 2003)
  - **PROMO** (Messeguer et al. 2002)

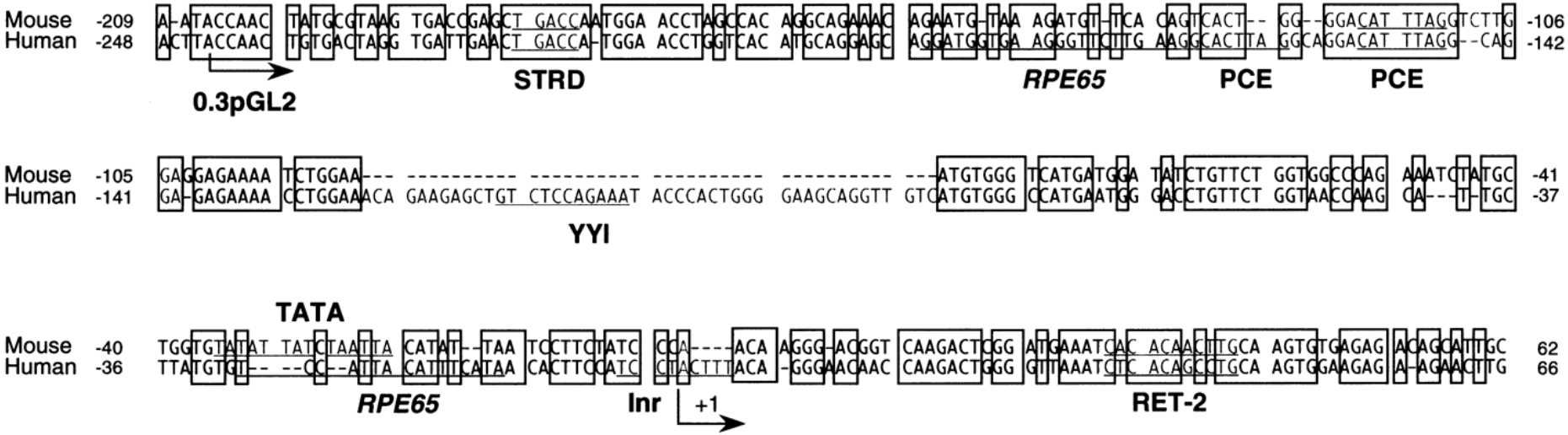
- General representations or patterns of a given binding site (Bucher 1990; Stormo 2000)
  - Simple sequence alignment of real sites ⇨ **consensus sequence** or **weight matrices** ⇨ scan genomic sequences to find new occurrences of a binding motif

TRANSFAC Matrix

```
AC   M00252
XX
ID   V$TATA_01
XX
DT   25.09.1996 (created); ewi.
DT   25.09.1996 (updated); ewi.
CO   Copyright (C), Biobase GmbH.
XX
NA   TATA
XX
DE   cellular and viral TATA box elements
XX
BF   T00794 TBP; Species: human, Homo sapiens.
BF   T00796 TBP; Species: mouse, Mus musculus.
BF   T00797 TBP; Species: fruit fly, Drosophila melanogaster.
XX
PO       A      C      G      T
01      61    145    152     31      S
02      16     46     18    309      T
03     352      0      2     35      A
04       3     10      2    374      T
05     354      0      5     30      A
06     268      0      0    121      A
07     360      3     20      6      A
08     222      2     44    121      W
09     155     44    157     33      R
10      56    135    150     48      N
11      83    147    128     31      N
12      82    127    128     52      N
13      82    118    128     61      N
14      68    107    139     75      N
15      77    101    140     71      N
XX
BA   389 TATA box elements
XX
CC   selected sequences from 502 promoters of EPD, mainly from vertebrates
```

- A **TRANSFAC matrix** entry constructed from a real collection of 389 TATA boxes (Bucher 1990)
- AC: accession; ID: Identification…
- The weight matrix corresponding to **the promoter element**, the number of sites used to build it, other additional information
- Each line in the matrix corresponds to a position in **the promoter motif**
- The last column: the consensus nucleotide at that position

Global Alignment of the Human & Mouse CRALBP Proximal Promoter Sequences



http://www.molvis.org/molvis/v4/p14/kennedy-fig6.html

Pattern-Driven Algorithms (2)

- A huge number of **false positives**, reasons
  - The **binding site sequence** for a specific transcription factor may be **highly variable**
  - The **length** of the binding sties is **very short** (typically 5-15 bp)
  - **Interaction** between transcription factors may influences **binding affinity** with the promoter sites
  - **One** binding site may be recognized by one or **more different transcription factors**

Pattern-Driven Algorithms (3)

- **Solutions** to overcome false positives
  - Concentration of predictions > isolation predictions (Werner 2000)
  - TSS nearby (Praz et al. 2002)

- **A difficult problem**
  - The correct **annotation of the first exon of a gene** (coding or non-coding) (Davuluri et al. 2001)

Pattern-Driven Algorithms (4)

- **Cooperation between transcription factors** play important role in the regulation of transcription
  - Searching for **clusters of associated sites** or **composites** can lead to a substantial improvement (Wagner 1997)
  - **Experimental** discovery & classification of new transcription factors can help to develop more complete catalogs of regulatory elements (Pennacchio & Rubin 2001)

Sequence-Driven Algorithms (1)

- **Rationale:**
  - **Common functionality** can be deduced through **underlying sequence conservation**

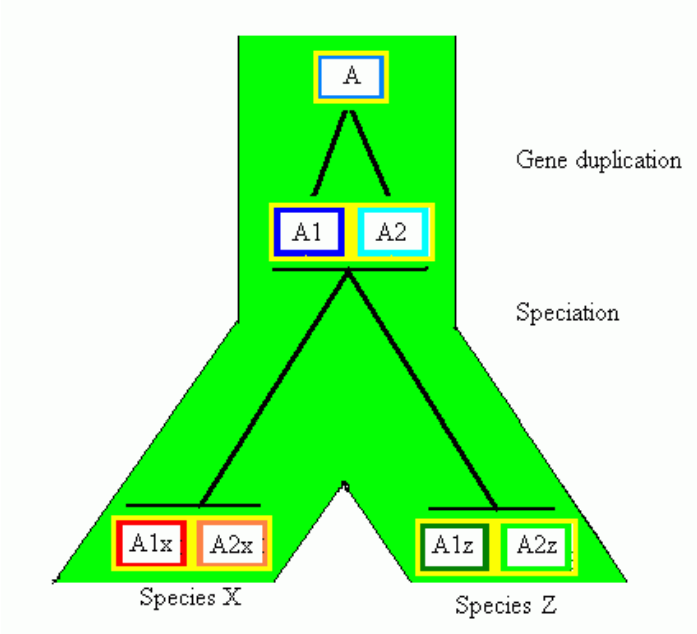
- Alignments of promoter regions of **co-regulated genes** ⇨ highlight **regulatory elements** involved in **co-regulation**
  - Heuristic techniques to identify **common motifs** in sets of **unaligned sequences**
    - **MEME** (Bailey & Elkan 1995)
    - **AlignAce** (Roth et al. 1998)

MEME (Bailey & Elkan 1995; Bailey et al. 2006)

- To discover potentially **novel motifs**
  - Searching for **conserved motifs** in sets of **unaligned**, **functionally** related sequences
  - Novel motifs might be statistic artifacts

Sequence-Driven Algorithms (2)

- Two kinds of **co-regulated sequences**
  - **Orthologous genes** from different species
  - Genes that have been experimentally determined to be **co-expressed** (Pennacchio & Rubin 2001)



Comparative Promoter Prediction (1)

- Patterns of **gene regulation** are often **conserved across species**
  - **Phylogenetic footprinting**
    - Interspecies comparisons ⇨ to identify common regulatory sequences (Wasserman et al. 2000)
      - The selection of appropriate species, **critical**

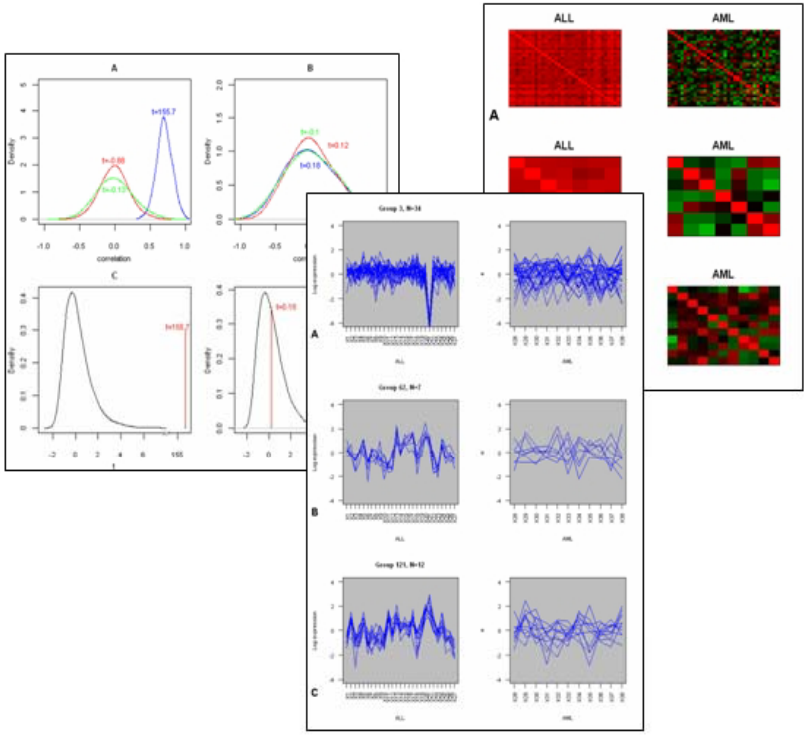
Comparative Promoter Prediction (4) – Potential Problems

- It is unclear **to what extent noncoding conserved elements** exhibit regulatory functions (Dermitzakis et al. 2002)

- There is **an important fraction** of regulatory elements that are not conserved across species

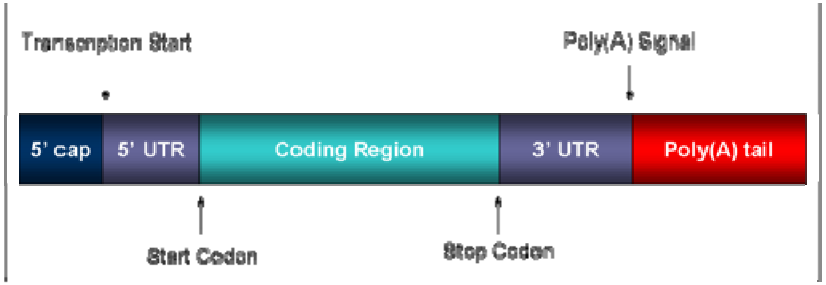
Sequence-Driven Algorithms (3)

- **Microarray data** or **expression profiling**
  - Co-expression of genes could reflect the existence of common configuration of promoter elements
  - Examples
    - **Yeast** (Chu et al. 1998; Tavazoie et al. 1999)
- In mammalian
  - Higher complexity = more difficult (Pennacchio & Rubin 2001)



Prediction of Promoter Regions (1)

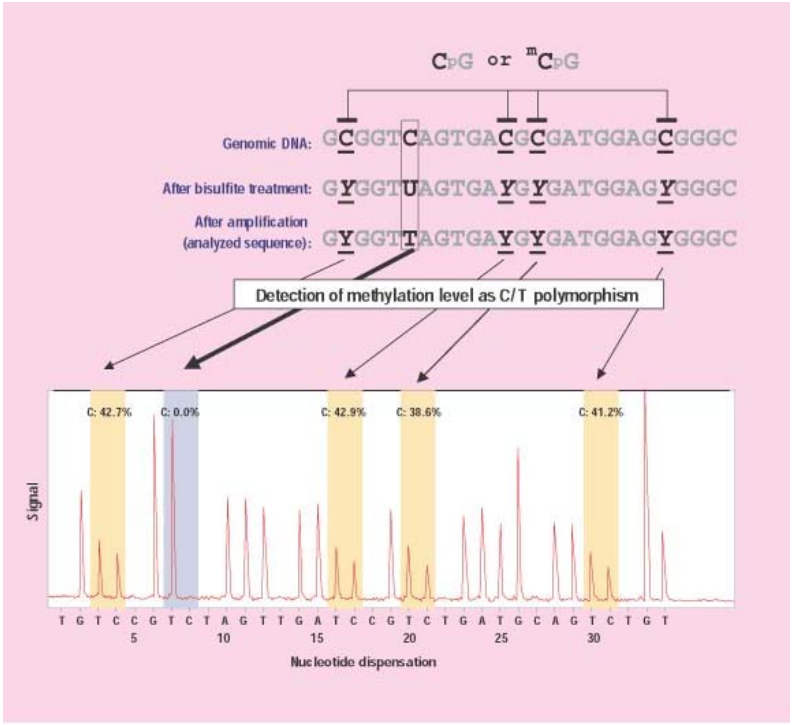
- Many methods
  - **No programs was found to be significantly superior to any other** (Fickett & Hatzigeorgiou 1997)
    - In general, quality is poor

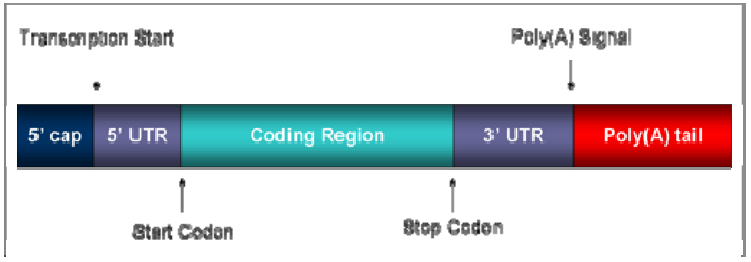


- Gene promoter regions
  - **Clusters of binding sites**
  - Methods
    - **Biased composition to locate TSS** + the region upstream containing a significant concentration of binding sites
      - Weight metrics & oligonucleotide counts (overexpresented words)

Prediction of Promoter Regions (2)

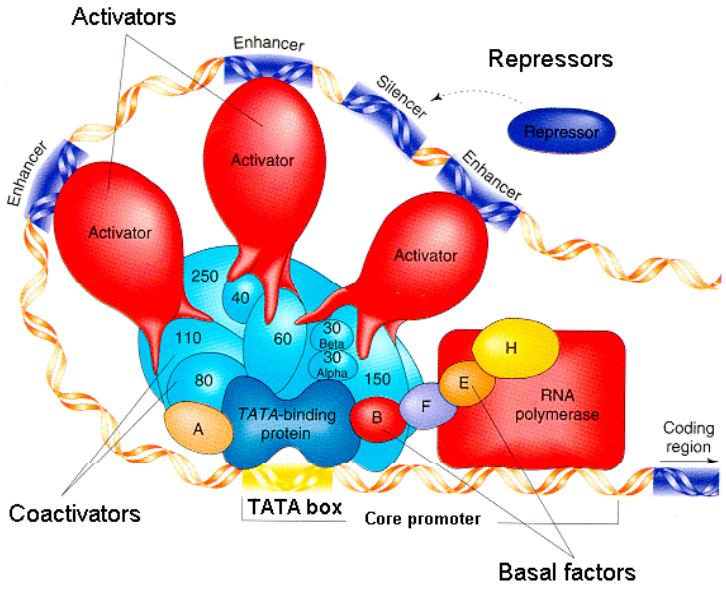
- **FirstEF** (Davuluri et al. 2001)
  - A set of discriminant functions, e.g., **CpG island** detection, donor splice site matrices…
  - To identify both **promoter regions** & **first exons (+UTRs)**





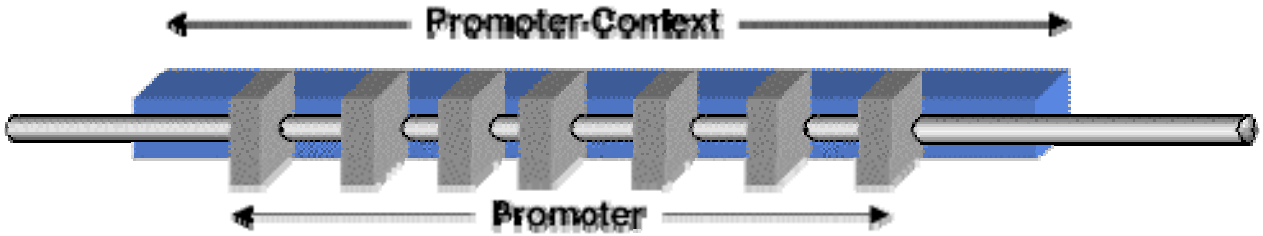
Prediction of Promoter Regions (3)

- PromoterScan (Prestridge 1995)
  - The density of **known binding sites**
  - The existence of **TATA box**



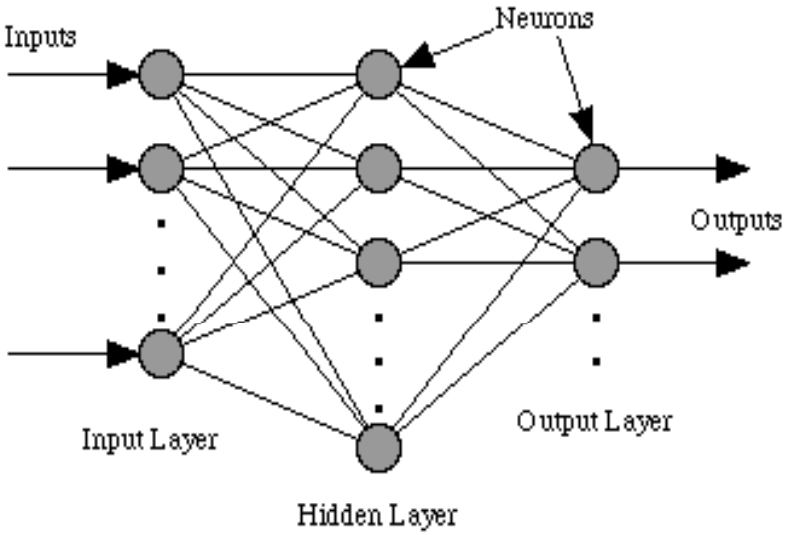
Prediction of Promoter Regions (4)

- PromoterInspector (Scherf et al. 2000)
  - The recognition of **the context of promoter** by exact pattern matching to **a consensus sequence** (rather than their exact location)
  - **Specificity**: 85%; sensitivity: 48%



Prediction of Promoter Regions (5)

- Dragon Promoter Finder (Bajic et al. 2002)
  - Artificial neural networks
    - To combine information derived from promoters, exons & introns



Strategies & Considerations (1)

- Complete genomes, usually with these information

- **Still useful**
  - The users may wish to **use different parameters**,
    - To analyze **alternative splicing** or to analyze regions apparently devoid of genes, to predict
      - **Splice signals**
      - Suboptimal exons

  - For analyzing the genomes of **organisms currently being sequenced**

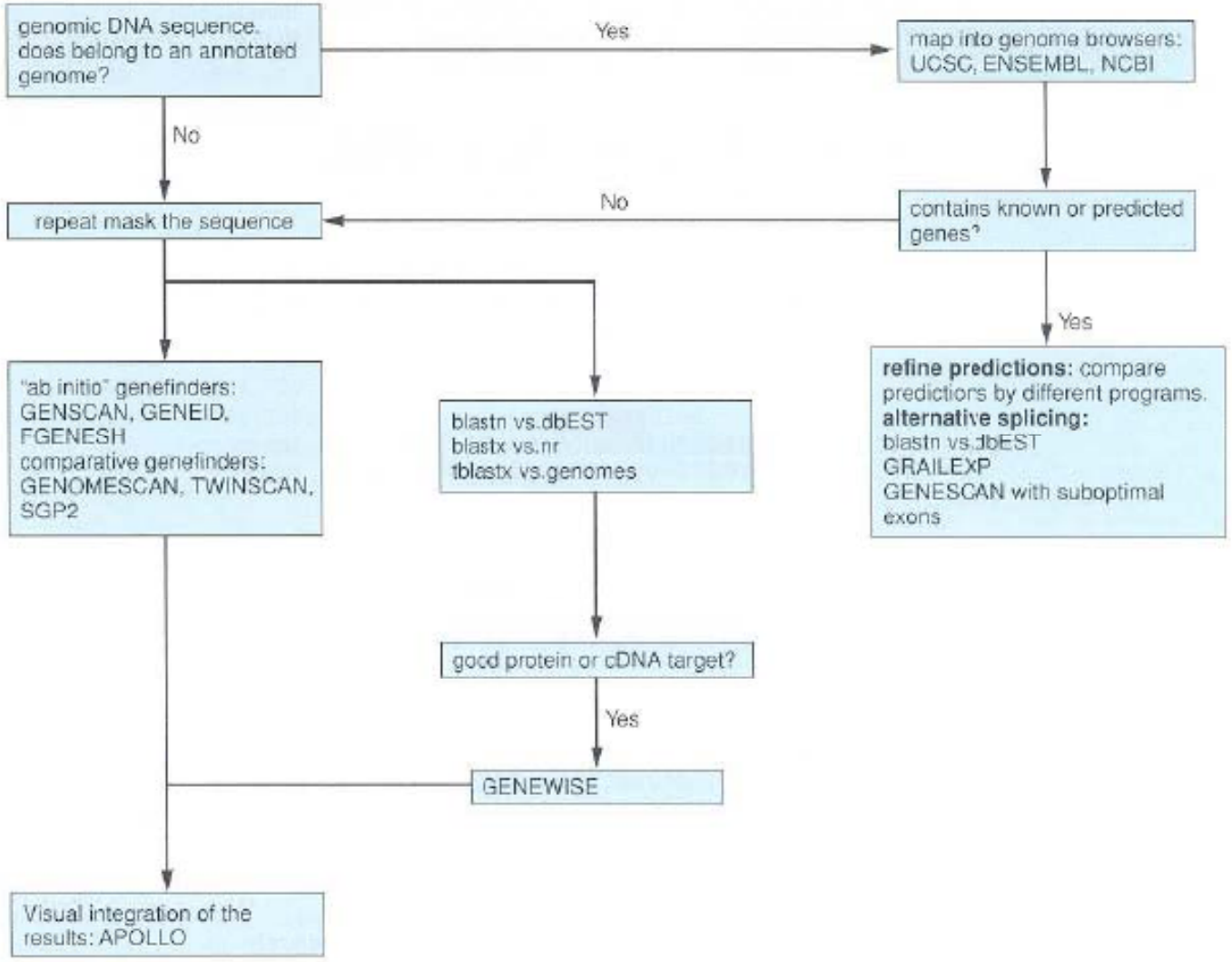
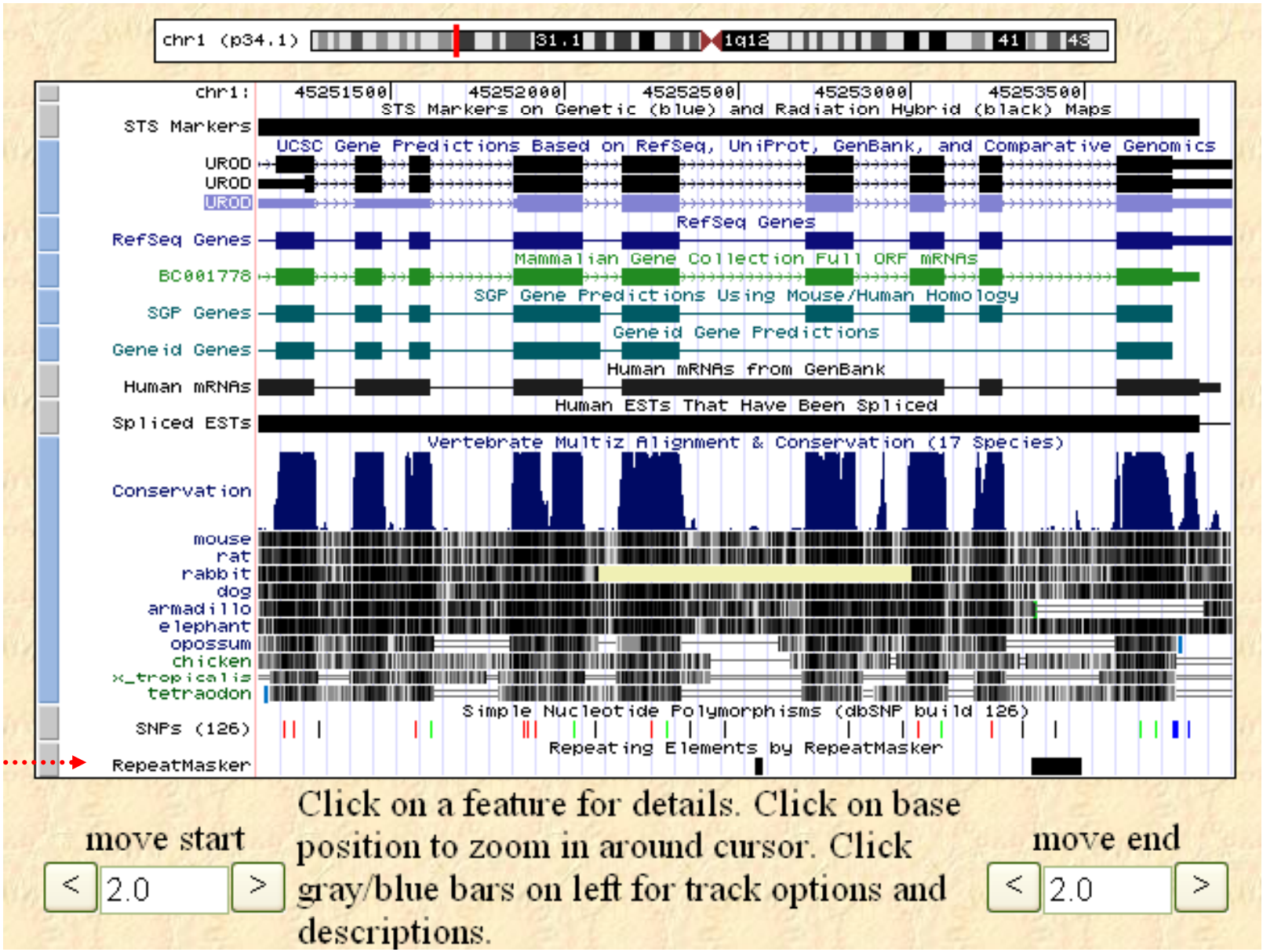


FIGURE 5.15 **Flowchart illustrating decision-making and considerations that need to be taken into account when selecting and using gene prediction programs. See main text for discussion.**

Masking the Sequences: Searching for Repeats (1)

- **RepeatMasker** (Smit & Green)
  - To find repetitive elements & **reduce false-positive prediction**
    - To mask the regions containing repeats (substitutes an N for each character in a repetitive element)
  - Gene prediction programs **ignore such stretches** in making their predictions
    - **Coding exons** tend **not to** overlap or to contain repetitive elements
    - Example (next slide), UROD: few repetitive elements
    - But most genomes ~40% repetitive elements



http://genome.ucsc.edu/cgi-bin/hgTracks?position=chr1:45251116-45253928&hgsid=92968557&knownGene=pack&hgFind.matches=uc001cnc.1,

Masking the Sequences: Searching for Repeats (2)

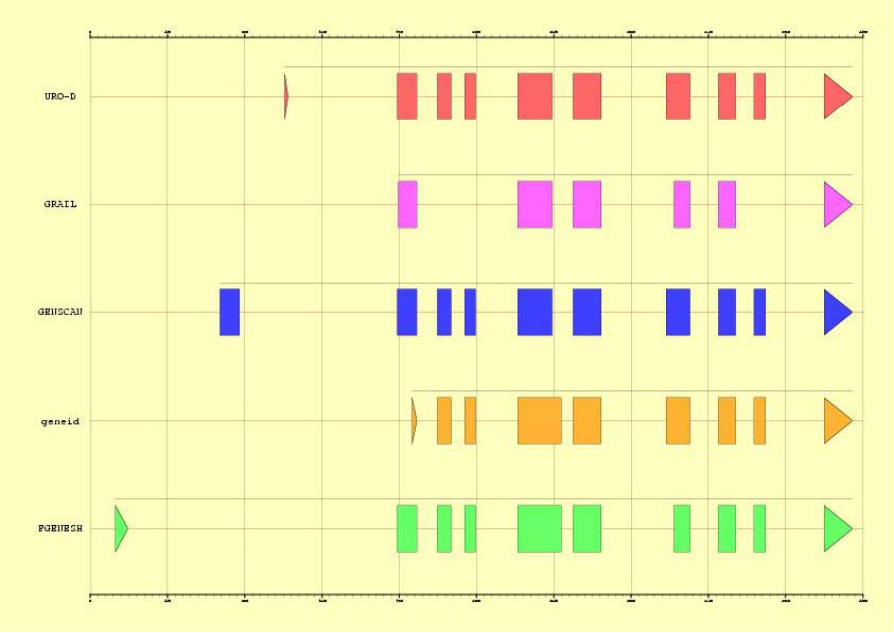
- **Dramatic effects**
  - **GeneScan** predicts 1128 genes **without masking** (human chr. 22) vs. 789 with **masking**
  - **GeneID**
    - 1179 to 730
- Although **most** of the additional exons predicted using **unmasked sequence** data are likely to be **false positives**
  - There are **times** where **coding regions** do occur in repetitive or low-complexity regions ⇨ aggressive masking of the sequence, then, could lead to missing some actual exons
- **Run both and see**

Interpreting Gene Predictions (1)

- **Internal exons** are predicted consistently **across all of the methods**
- There is **substantial disagreement** in the prediction at the 5'-end

- **GeneID:** part of the 2nd exon as the initial exon
- **GENESCAN:** no initial exon, a partial gene starting with an internal exon
- **FGENESH:** a wrong exon in a known non-coding region as being the 5'

- The **coding fraction** of the **1st exon** is **quite short,** yielding a poor conding signla ⇨ **gene boundaries** are difficult to predict



- **All the programs** show a significant expansion of the 5'-end of the gene, immediately upstream of the UROD gene

Interpreting Gene Predictions (2)

- Regions containing genes, coding exons tend to be well delineated
  - **Not always** assembled into the correct overall gene structure
- General ways to predict
  - To split **one real gene** into **gene fragments** or predict **chimeric genes**
  - In region devoid of genes, to predict coding exons, as well
- Consistent predictions by **different *ab initio* methods** ⇨ suggestive of the **actual** presence of a protein-coding gene, even **in the absence of** experimental evidence

Interpreting Gene Predictions (3)

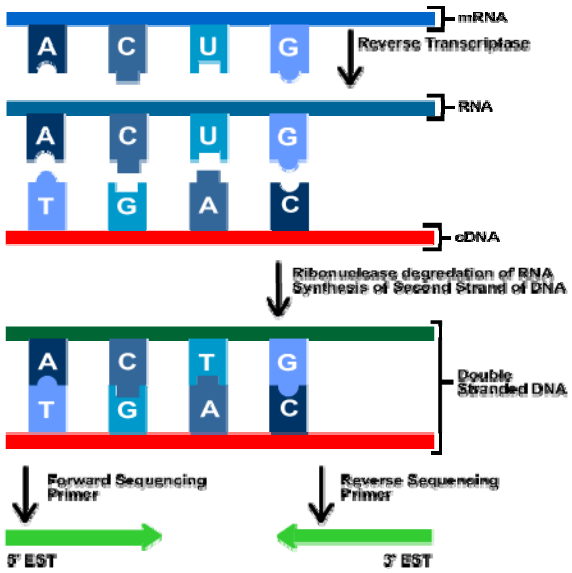
- The existence of a human mRNA confirms that it corresponds to a bona fide exon

- Conversely, inconsistent predictions ⇨ indication of a potential false positive…
  - **Highly suspicious**

EST Searches

- From the set of EST matches, the minimum number of splice forms **to which these ESTs actually correspond?** (Wheeler 2002; Eyras et al. 2004)

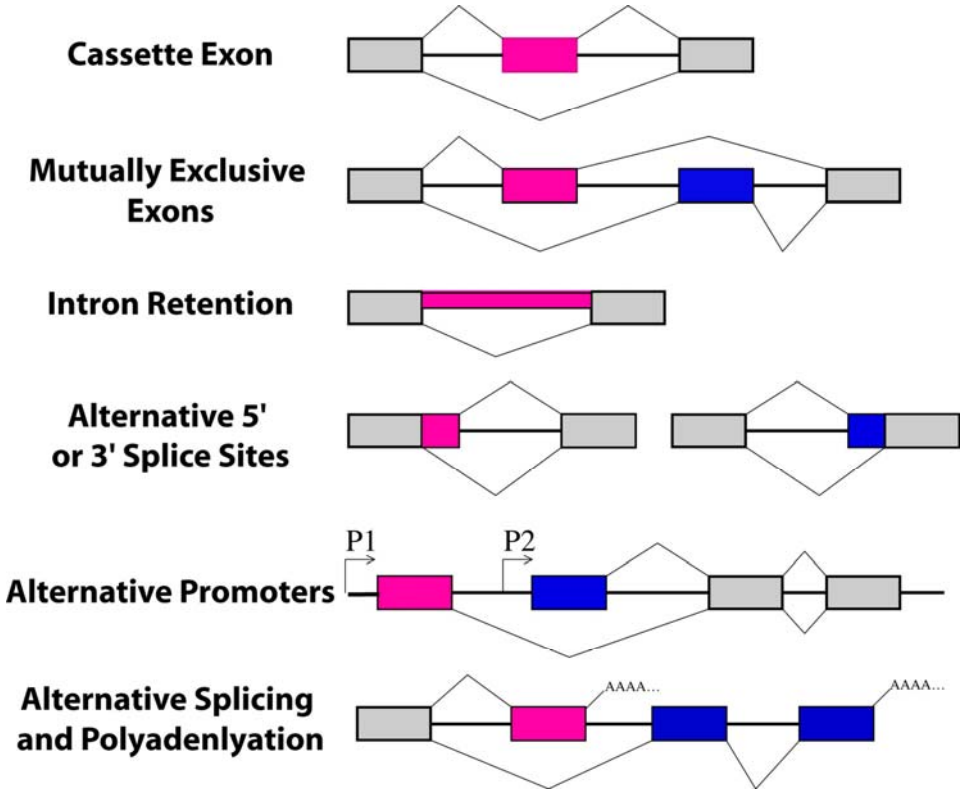
- **Facts**
  - A large incidence of **genomic contamination** & **unprocessed transcripts** in cDNA libraries
  - Similar to an EST ~ the alignment of the query sequence occurs **across a splice junction**



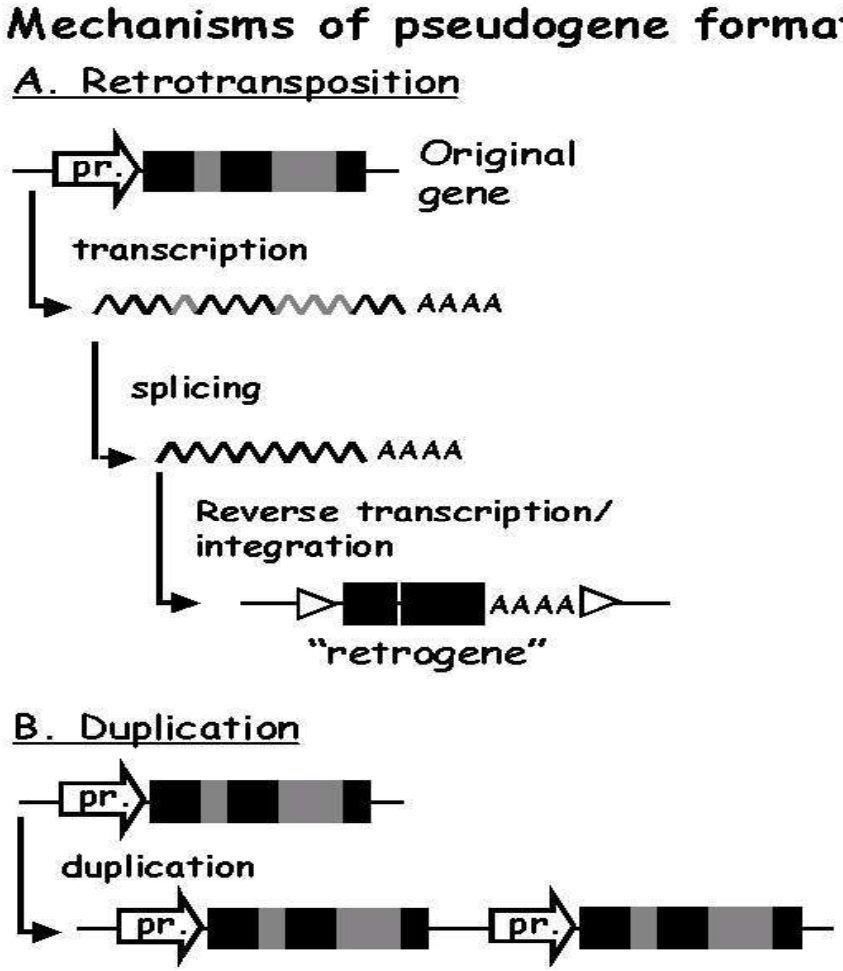
- Not all known genes are supported by EST evidence
- **3'- ESTs** tend to correspond to the 3' end of a gene (+ substantial fraction of **the UTR**)
- **5' ESTs** often are **internal** to the gene & **are mostly coding** (Guigo et al. 2000)

Predicting Genes on Top of Previous Annotations

- Major genome browsers contain predicted genes
- **Gene prediction programs**
  - In exploratory data analysis, esp. alternative splicing of known genes
    - GeneID & GENESCAN
  - Prediction on regions that are apparently devoid of genes
    - Based on experimental evidence with protein-coding region
      - GeneID (Blanco et al. 2002)
      - GAZE (Howe et al. 2002; the Sanger Center)



The Problem with Pseudogenes (Real or Not) (1)



Pseudogene features

-Absence of introns and promoter sequences

- flanked by direct repeats

- presence of poly-A tract

- randomly integrated anywhere in the genome

-Introns present

-Transcriptional regulatory elements often present

-Usually adjacent to original gene

The Problem with Pseudogenes (Real or Not) (2)

- **Many pseudogenes** are similar to **functional** paralogous genes
- ESTs **do not always** exist for actual genes
- Intronless gene in multiexon paralogous genes
  - Recent gene duplication ⇨ multiexon predictions
- **Homologues** in **another organism**
  - To compute the synonymous (Ka) vs. non-synonymous (Ks) substitution rate
  - **Ka/Ks (Fay & Wu 2003)**
    - **~1 = neutral evolution** = pseudogene
  - Conservation of overall **gene structure** in close homologs
  - Example: human vs. mouse
    - Guigo et al. 2003

Using the Right Parameters

- **Appropriate for**
  - The **species** & taxonomic group
    - Not always possible

- The use of parameters that have been optimized for one organism may produce **poor results** when used for another organisms
  - Especially when the organism is distinctly related

  - The **quality** of the prediction is seen to degrade as a function of phylogenetic distance

Promoter Prediction & Characterization (1)

- **One major problem**
  - **The transcription start site (TSS)**
    - The 5′ end of the gene is not determined accurately most of time
    - No full-length cDNAs for a large fraction of mammalian genes
    - cDNAs are almost non-existent for most of the species whose genomes are currently sequenced

- FirstEF (Davuluri et al. 2001)
  - To identify genes without full-length cDNAs
    - Or UCSC Genome Browser

Promoter Prediction & Characterization (2)

- **TRANSFAC**
  - To search against all known transcription factor binding motifs (those found in TRANSFAC)
    - Overwhelming number of prediction
    - Comparative genomic techniques to assist in processing the results
      - **To align two or more promoter sequences** from **homologous genes** will highlight the common, conserved fragments
        - Fragments that may correspond to functional elements

- Most of the conserved fragments are near the TSS + have TRANSFAC hits corresponding to actually regulatory element

Visualization & Integration Tools (1)

- A common visual interface to view different results

- **Third-party interfaces**
  - UCSC Genome Browser
  - Ensembl Genome Browser (Wolfsberg et al. 2001)

- **Local** interactive graphical tools provide a better solution
  - **The Apollo system** (Lewis et al. 2002)

Visualization & Integration Tools (2)

- Local interactive graphical tools provide a better solution
  - **GFF2PS** (Abril & Guigo 2000)
    - In high-throughput environment, especially useful

BIOINFORMATICS **APPLICATIONS NOTE** *Vol. 16 no. 8 2000*
Pages 743–744

gff2ps: visualizing genomic annotations

Josep F. Abril and Roderic Guigó*

Grup de Recerca en Informàtica Mèdica, Institut Municipal d'Investigació Mèdica (IMIM), Universitat Pompeu Fabra (UPF), C/ Dr. Aiguader, 80. 08003—Barcelona, Spain

Received on December 15, 1999; revised on February 18, 2000; accepted on February 24, 2000

Abstract

Summary: *gff2ps is a program for visualizing annotations of genomic sequences. The program takes the annotated features on a genomic sequence in GFF format as input, and produces a visual output in PostScript. While it can be used in a very simple way, it also allows for a great degree of customization through a number of options and/or customization files.*

Availability: *gff2ps is freely available at http://www1.imim.es/~jabril/GFFTOOLS/GFF2PS.html*

Contact: *jabril@imim.es*

gff2ps plots the features from different sources specified on a GFF file in a number of parallel rows (the so-called tracks here) along the length of the output page(s) (see Figure 1 for examples). Actually these are 'virtual' pages (the so-called blocks here) allowing for several blocks to be included in a single physical page, or for splitting a single block in a number of physical pages. Features can be plotted in a variety of colors and shapes and those grouped together can be visually linked in a number of ways.

gff2ps allows for a substantial amount of customiza-