Genes, Promoters, and Position Specific Scoring Matrixes

Introduction

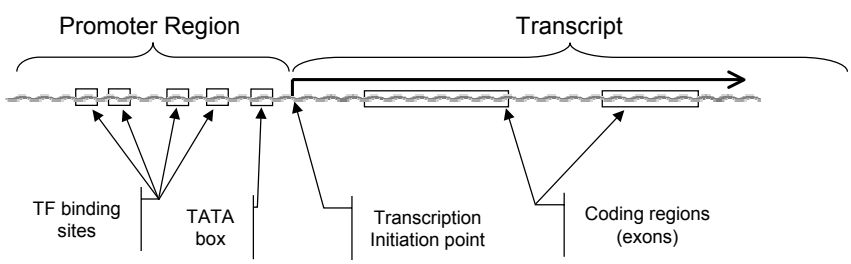
- Recall, regulation of gene expression (i.e., use) is specific
- Different genes are activated depending on cell type, external/internal signals, and past history of the cell
- How does this regulation works?

Schematics of Gene Expression

There are two processes that we need to understand

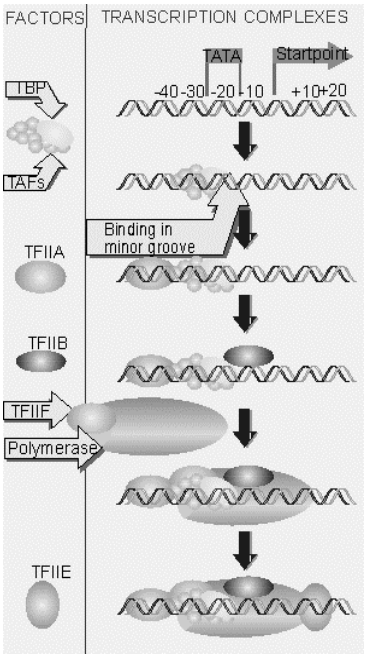
- **Basal Transcription Factors**
  - These are “common” factors that work on all genes
  - They supply the basic machinery
- **Regulatory Transcription Factors**
  - These are responsible for creating the differences in expression between genes

Layout of a Gene



Basal Transcription Factors

- A complex of many proteins
- Some parts needed for actual transcription
- Others are needed just of “jump starting” the process
- Major player - “TATA Binding Protein” (TBP) that anchors the complex

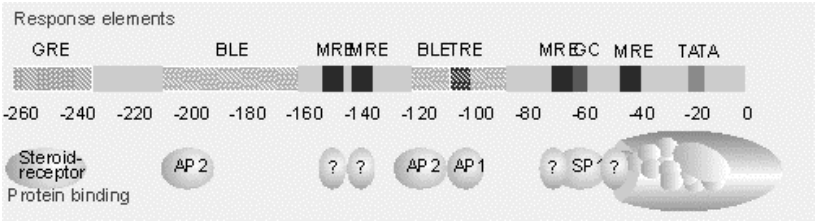


Promoter Regions

In these regions regulatory TFs bind

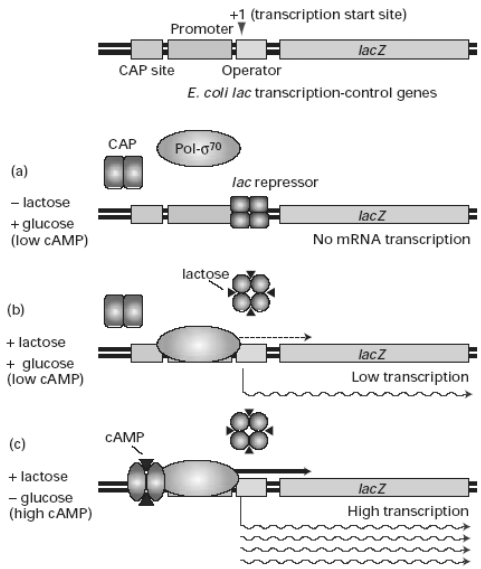
These can

- Activate/enhance transcription



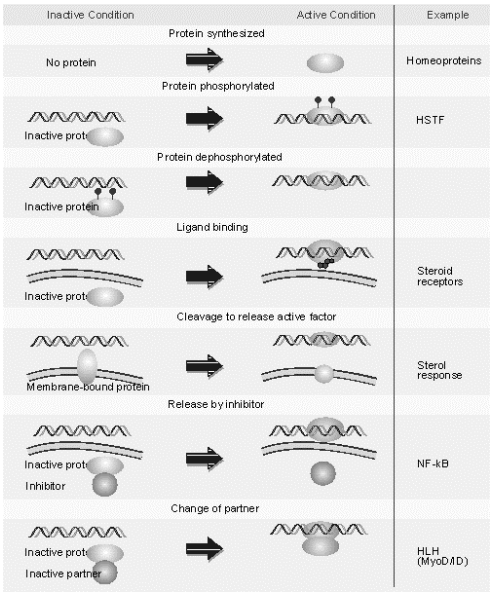
TF Binding

- TF binds to DNA in a **specific** manner
- They recognize specific sequences of DNA by interactive with the nucleotides
- There are various forms that these TFs can take



Regulation of Factors

- Many mechanisms can control the number of "active" copies of each TFs



Computational Challenges

- How to represent TF binding sites?
- Can we discover new sites of known TFs in genome sequence?
- Can we new TF binding sites from examples of promoter regions of co-expressed genes
- How do these TFs interact?

Protein Coding Regions

Sequence Analysis Tasks

- √ Calculating the probability of finding a sequence pattern
- √ Calculating the probability of finding a region with a particular base composition
- √ Representing and finding sequence features/motifs using frequency matrices

Goal

- Given a DNA or RNA sequence, find those regions that code for protein(s)
  - Direct approach: Look for stretches that can be interpreted as protein using the genetic code
  - Statistical approaches: Use other knowledge about likely coding regions

Direct Approach

Genetic codes

- The set of tRNAs that an organism possesses defines its genetic code(s)
- The universal genetic code is common to all organisms
- Prokaryotes, mitochondria and chloroplasts often use slightly different genetic codes
- More than one tRNA may be present for a given codon, allowing more than one possible translation product

Genetic codes

- Differences in genetic codes occur in start and stop codons only
- Alternate initiation codons: codons that encode amino acids but can also be used to start translation (GUG, UUG, AUA, UUA, CUG)
- Suppressor tRNA codons: codons that normally stop translation but are translated as amino acids (UAG, UGA, UAA)

Reading Frames

- Since nucleotide sequences are “read” three bases at a time, there are three possible “frames” in which a given nucleotide sequence can be “read” (in the forward direction)
- Taking the complement of the sequence and reading in the reverse direction gives three more reading frames

Reading frames

```
   TTC TCA TGT TTG ACA GCT
RF1  Phe Ser Cys Leu Thr Ala>
RF2   Ser His Val *** Gln Leu>
RF3    Leu Met Phe Asp Ser>
   AAG AGT ACA AAC TGT CGA
RF4 <Glu *** Thr Gln Cys Ser
RF5   <Glu His Lys Val Ala
RF6  <Arg Met Asn Ser Leu
```

	U	C	A	G	
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	*** ◀	*** ◀	A
	Leu	Ser	*** ◀	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met ▶	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

Reading frames

- To find which reading frame a region is in, take nucleotide number of lower bound of region, divide by 3 and take remainder (modulus 3)
- 1=RF1, 2=RF2, 0=RF3
- This is the convention used by MacVector
- Assumes first nucleotide is 1 (not 0)

Reading frames

- For *reverse* reading frames, take nucleotide number of *upper* bound of region, subtract from total number of nucleotides, divide by 3 and take remainder (modulus 3)
- 0=RF4, 1=RF5, 2=RF6
- This is because the convention MacVector uses is that RF4 starts with the last nucleotide and reads backwards

Open Reading Frames (ORF)

- Concept: Region of DNA or RNA sequence that *could* be translated into a peptide sequence (open refers to absence of stop codons)
- Prerequisite: A specific genetic code
- Definition:
  - (start codon) (amino acid coding codon)$_n$ (stop codon)
- Note: Not all ORFs are *actually* used

Open Reading Frames

- Open file **YSPTUBB** in **Sample Files** folder
- Under Analyze select Open Reading Frames
- Click box next to start/stop codons...
- Click OK

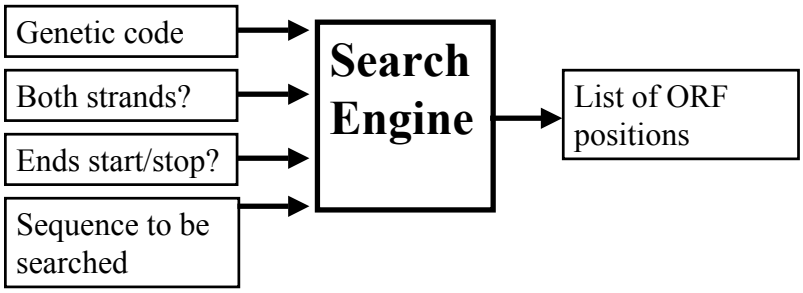
Splicing ORFs

- For eukaryotes, which have interrupted genes, ORFs in different reading frames may be spliced together to generate final product
- ORFs from forward and reverse directions cannot be combined

ORFs and Exons

- MacVector displays “annotations” to the sequence in a features table
- Open the feature table for YSPTUBB by clicking on the icon
- Note the six exons for the tubulin gene
- Does the large exon (exon 5) correspond to the large ORF in reading frame 3?

Block Diagram for Search for ORFs



Statistical Approaches

Calculation Windows

- Many sequence analyses require calculating some statistic over a long sequence looking for regions where the statistic is unusually high or low
- To do this, we define a window size to be the width of the region over which each calculation is to be done
- Example: %AT

Base Composition Bias

- For a protein with a roughly “normal” amino acid composition, the first 2 positions of all codons will be about 50% GC
- If an organism has a high GC content overall, the third position of all codons must be mostly GC
- Useful for prokaryotes
- Not useful for eukaryotes due to large amount of noncoding DNA

Fickett's statistic

- Also called TestCode analysis
- Looks for asymmetry of base composition
- Strong statistical basis for calculations
- Method:
  - For each window on the sequence, calculate the base composition of nucleotides 1, 4, 7..., then of 2, 5, 8..., and then of 3, 6, 9...
  - Calculate statistic from resulting three numbers

Codon Bias (Codon Preference)

- Principle
  - Different levels of expression of different tRNAs for a given amino acid lead to pressure on coding regions to "conform" to the preferred codon usage
  - Non-coding regions, on the other hand, feel no selective pressure and can drift

Codon Bias (Codon Preference)

- Starting point: Table of observed codon frequencies in known genes from a given organism
  - best to use highly expressed genes
- Method
  - Calculate “coding potential” within a moving window for all three reading frames
  - Look for ORFs with high scores

Codon Bias (Codon Preference)

- Works best for prokaryotes or unicellular eukaryotes because for multicellular eukaryotes, different pools of tRNA may be expressed at different stages of development in different tissues
  - may have to group genes into sets
- Codon bias can also be used to estimate protein expression level

Portion of *D. melanogaster* codon frequency table

Amino Acid	Codon	Number	Freq/1000	Fraction
Gly	GGG	11	2.60	0.03
Gly	GGA	92	21.74	0.28
Gly	GGT	86	20.33	0.26
Gly	GGC	142	33.56	0.43
Glu	GAG	212	50.11	0.75
Glu	GAA	69	16.31	0.25

Comparison of Glycine codon frequencies

Codon	E. coli	D. melanogaster
GGG	0.02	0.03
GGA	0.00	0.28
GGT	0.59	0.26
GGC	0.38	0.43