Gene finding and Genome annotation

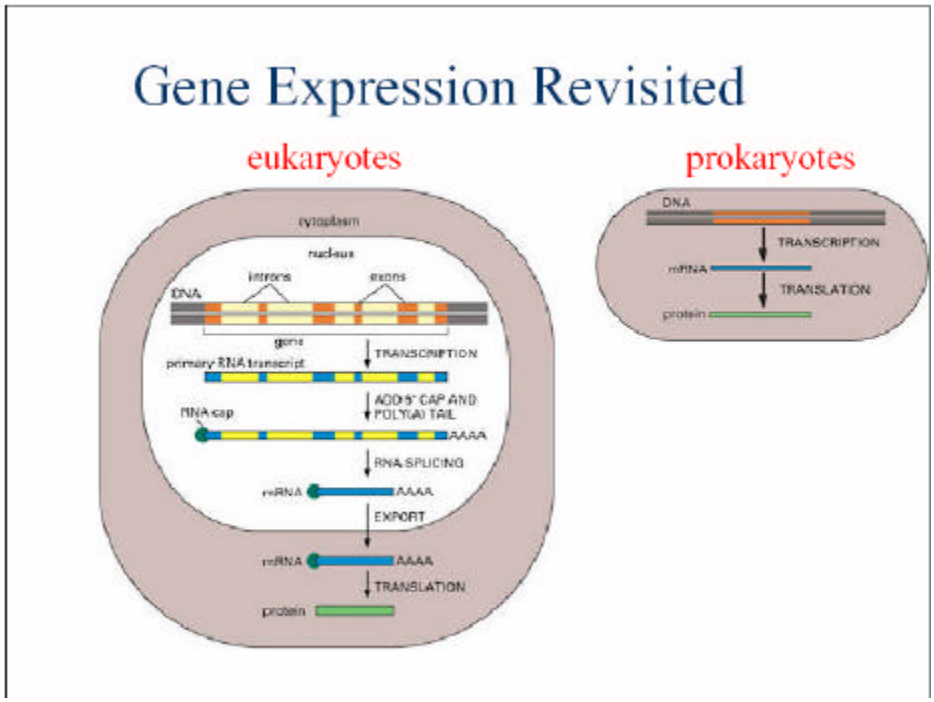
What is a Gene?

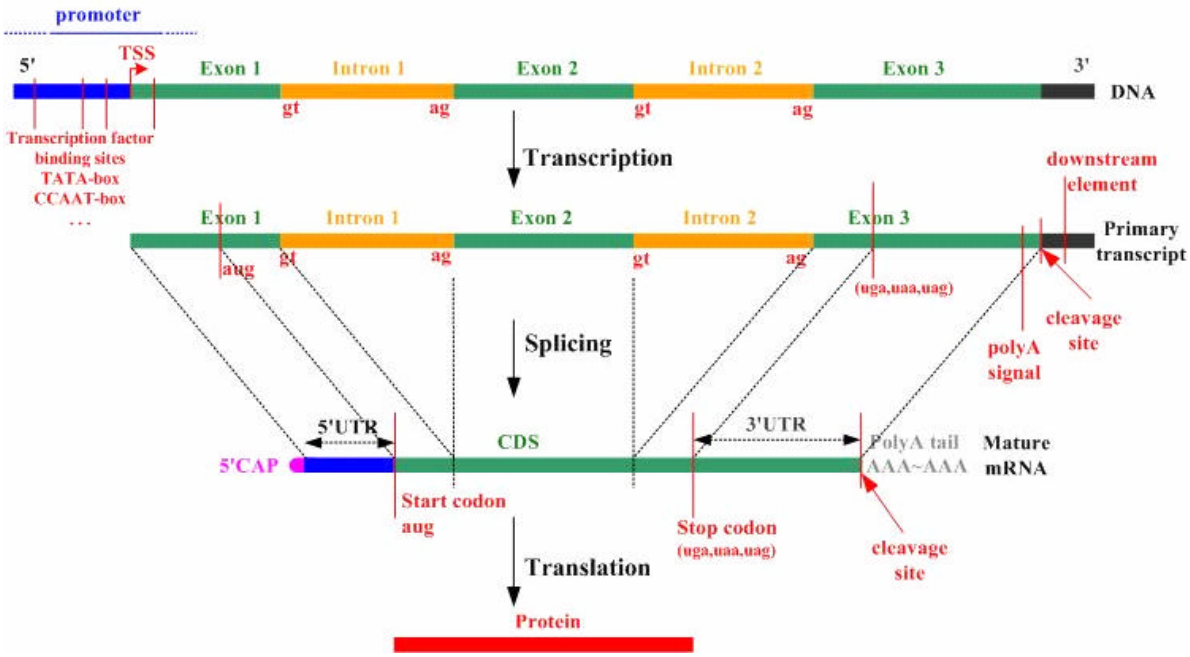
An inheritable trait associated with a region of DNA that codes for a polypeptide chain or specifies an RNA molecule which in turn have an influence on some characteristic phenotype of the organism.

Abstract concept that describes a complex phenomenon

Given uncharacterized DNA sequences how you identify:

- The protein coding regions
- Exon/intron boundaries and splice sites
- Beginning and end of translation
- Alternative splicings
- Regulatory elements (e.g. promoters)





What is annotation?

Extraction, definition, and interpretation of features on the genome sequence derived by integrating computational tools and biological knowledge.

<u>Identifiable features in the sequence</u>

How does an annotation differ from a gene?

There is no (yet known) perfect method for finding genes. All approaches rely on combining various "weak signals" together. Many annotations describe features that constitute a gene.

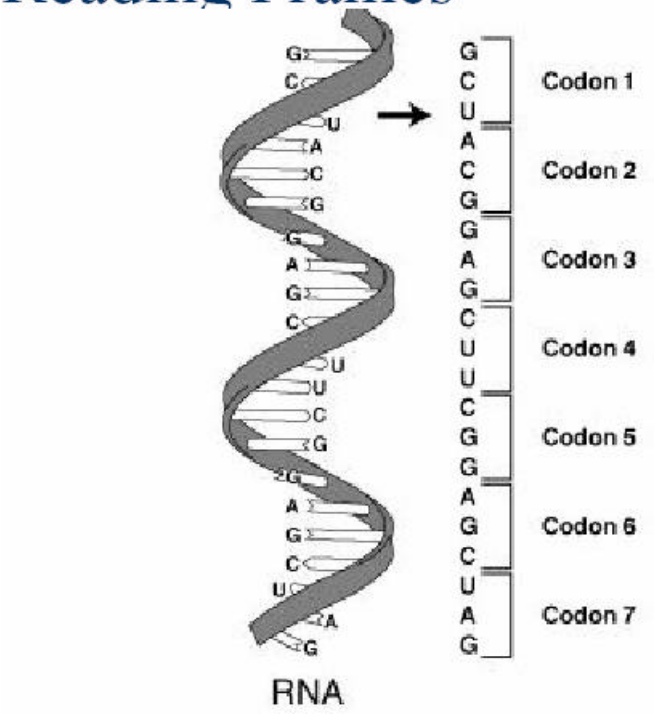
Find elements of a gene

- Coding sequences (exons)
- Promoters and start signals
- Poly-A tails and downstream signals

Assemble into a consistent gene model

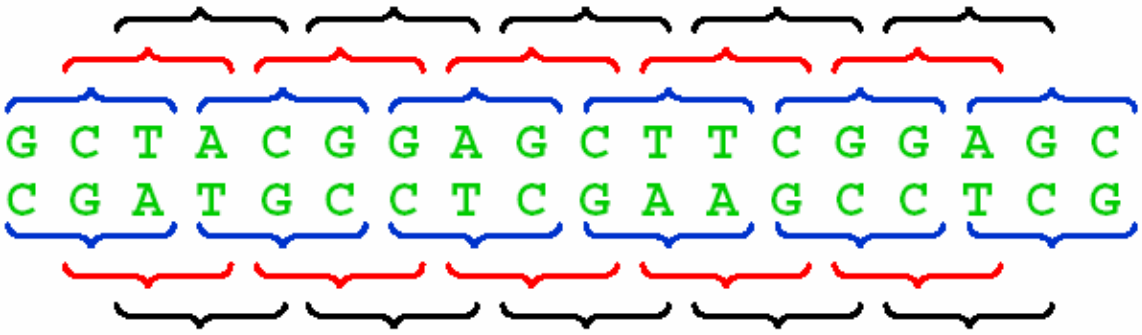
Use of homologous sequences

Reading Frames



Reading Frames

- a given sequence may encode a protein in any of the six reading frames



Approaches to finding (annotating) genes

Search by signal

Find genes by identifying the sequence (signals) involved in gene expression

Search by content

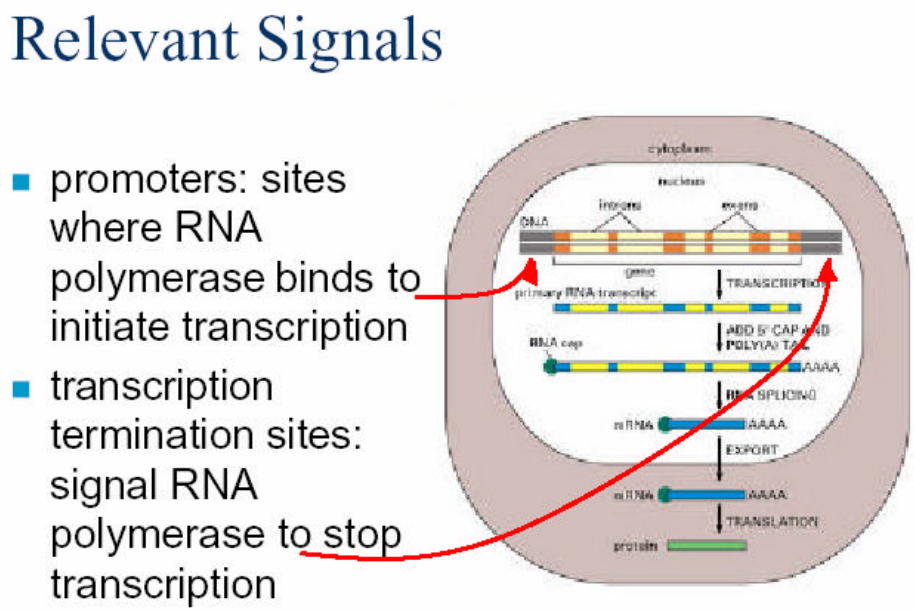
Find genes by statistical properties that distinguish protein-coding DNA from non-coding DNA

Combined

State-of-the-art systems for gene finding combine the above two strategies

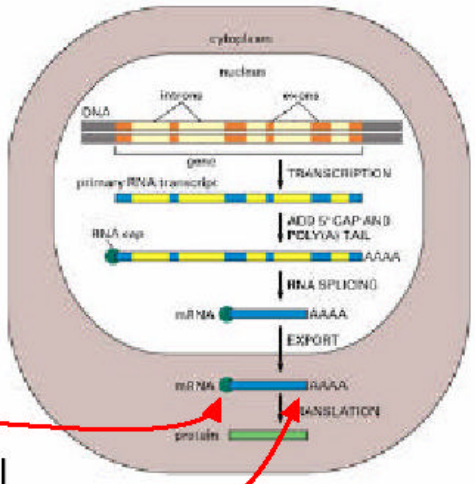
Relevant Signals for Search

Promoters (transcription initiation)
Terminators (transcription termination)
Ribosome binding sites (translation initiation)
Initiation codons (translation initiation)
Stop codons (translation termination)
Splice junctions (RNA processing)



Relevant Signals

- ribosome binding sites: where the ribosome binds to mRNA to initiate translation
- initiation codons: the first translated codons
- stop codons: the final codons in a gene



<u>Start and Stop Codons</u>

Start codons

Prokaryotes 90% time AUG is the initiation codon, but sometimes GUG or UUG is used

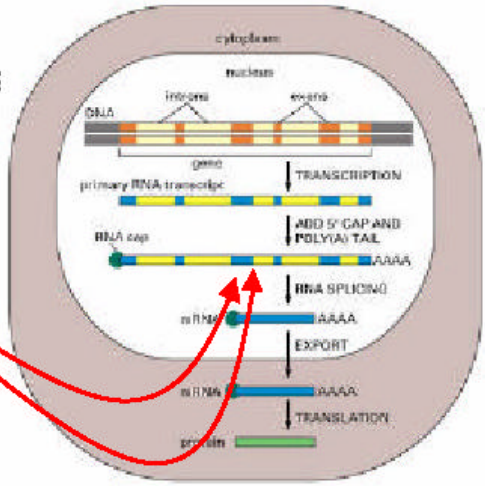
Eukaryotes AUG is almost always the initiation codon

Stop Codons

UUA, UAG, UGA always signal the end of translation

Relevant Signals

- splice junctions: sites where introns are spliced out of RNA transcripts



Splice Junctions

The splice sites on either end of an intron have different properties

The 5' splice site is called the **donor** site
The 3' splice site is call the **acceptor** site

Most Eukaryotic introns have a consensus splice signal: GU at the beginning ("donor"), AG at the end ("acceptor").

Variation does occur in the splice sites

Many AGs and GTs are not splice sites

Promoters

There are two important regions in a promoter sequence

About 10 bases before the transcription initiation site (-10 region)

The consensus sequence for the -10 region in *E. coli* is TATAAT, but few promoters actually have this sequence

About 35 bases before the transcription initiation site (-35 region)

Search by signal

To recognize promoters more general models need to be build on

Weight matrices
Probabilistic models
Neural networks
etc.

Use models to detect previously unseen instances of the signal

Search by content

Find genes by statistical properties that distinguish protein-coding DNA from non-coding DNA

Encoding a protein affects the statistical properties of a DNA sequence

Some amino acids are used more frequently than others (Leu more popular than Trp)
Different numbers of codons for different amino acids (Leu has 6, Trp has 1)
For a give amino acid, usually one codon is used more frequently than others

Codon Preference in E. Coli

AA	codon	/1000
Gly	GGG	1.89
Gly	GGA	0.44
Gly	GGU	52.99
Gly	GGC	34.55
Glu	GAG	15.68
Glu	GAA	57.20
Asp	GAU	21.63
Asp	GAC	43.26

Search by content

Inherent features
DNA exhibits certain biases that can be exploited to locate coding regions

Uneven distribution of bases
Codon bias

- CpG islands (regions of sequence that have a high proportion of CG dinucleotide pairs (p is a phoshodiester bond linking them)). They are present in promoter and exonic regions of approximately 40% of mammalian genes. Other regions of the mammalian genome contain few CpG dinucleotides and these are largely methylated.

In-phase words
Encoded amino acid sequence
Imperfect periodicity
Other global patterns

Search by content

Search by homology
Translate the DNA sequence in all reading frames
Search against protein database
High-scoring matches suggest the presence of homologous genes in DNA
The program BLASTX does just this

Intelligent methods of annotating genes

Pattern recognition methods weigh inputs and predict gene location
Content-based methods
Site-based methods
Comparative methods

Neural Networks

Hidden Markov Models

Stochastic Context-Free Grammar

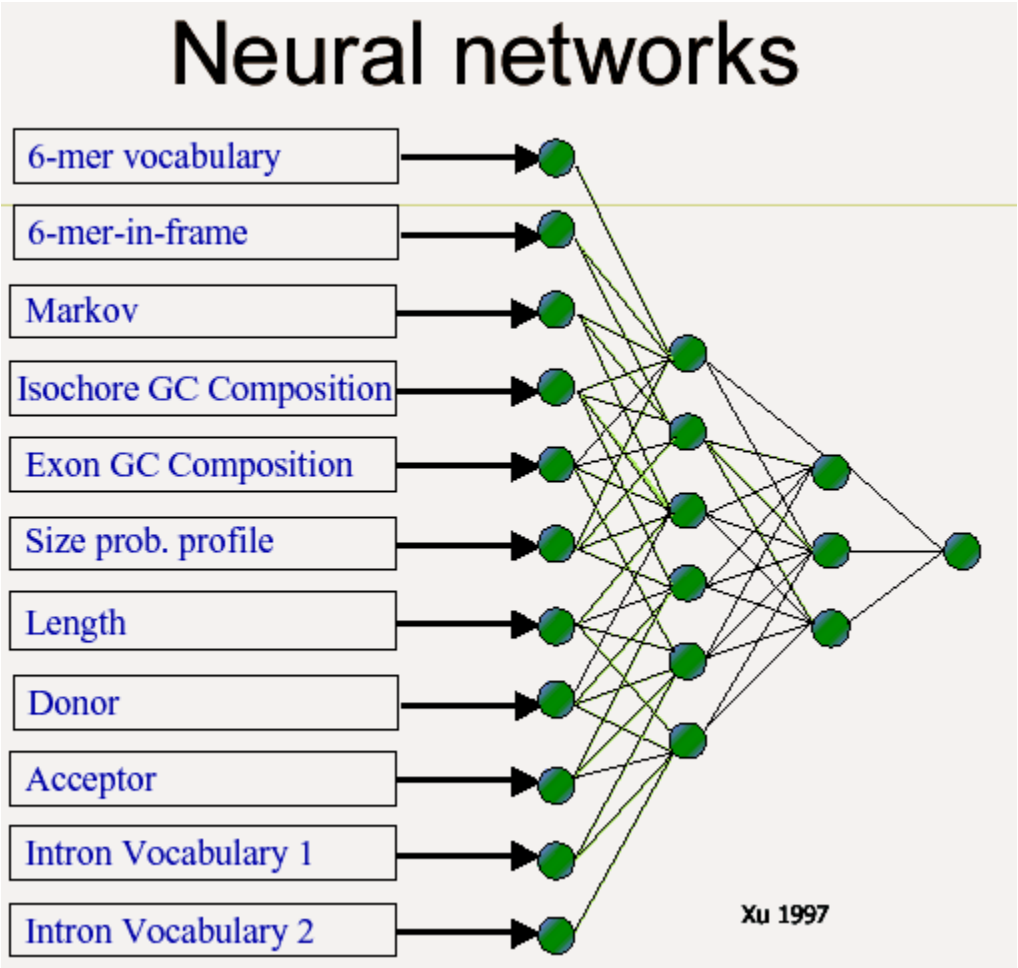
GRAIL *Uberbacher, Mural*

- GRAIL 1
  Neural network with fixed window length (100 bases)
- GRAIL 1a
  GRAIL 1 + adjacent information
- GRAIL 2
  Variable length window, contextual information
- GRAIL-EXP
  Comparison with partial and complete gene sequences

Window size trade-off

With small windows, better able to distinguish coding-region boundaries

Predictions often more accurate with larger windows since they look at more evidence



FGENEH/FGENES *Solovyev*

– Looks at several structural features
– Splice donor/acceptor sites
– Putative coding regions
– Intronic regions
– *Linear discriminant analysis* to split exon / non-exon classes
– Dynamic programming to assemble best gene structure

MZEF *Zhang*

– *Quadratic discriminant analysis*
– Exon length
– Exon-intron transitions
– Splice sites
– Branch sites
– Exon, strand, frame scores
– Detects internal exons
– No information about gene structure

GENSCAN *Burge, Karlin*

– Probabilistic model of sequence composition and gene structure
– Looks for gene structure descriptions that are consistent with the query sequence to assign probability that sequence stretch is exon, …
– Best ---> optimal
– But generates also suboptimal exons