Sequence Annotation

As new DNA sequence data becomes available, we seek to identify interesting features in this raw text.

The most interesting features are *genes*, the portions of the chromosome which describe how to make proteins.

Since genes and the promoter sites associated with them make promising drug candidates, there is a considerable pressure to quickly identify them computationally.

Indeed, automatic annotation is becoming a big game. Every genome sequencing project is expected to do annotation prior to publication.

The Wellcome Trust is investing at least 8 million pounds in the *Ensembl* project (http://www.ensembl.org/), providing automatic annotation of the human genome.

Transcription

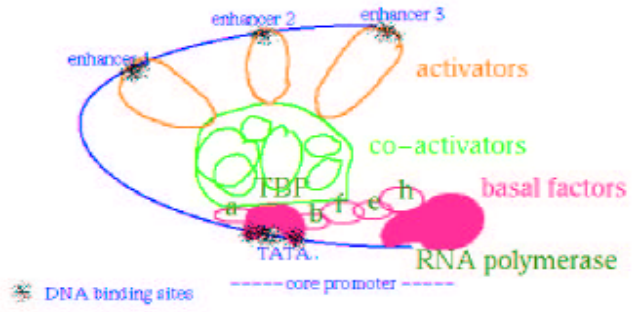
Transcription is the process of copying the portion of the DNA containing a gene into RNA.

Something has to happen to instigate transcription at genes and not at non-coding regions. Thus there must be signals in the DNA sequence which tell where to start the transcription.

An enzyme called *RNA polymerase* binds to specific patterns at approximately 10 and 35 bases before the gene to start transcription.

Other binding sites upstream from before the gene called *promoters* help signal when to express or inhibit the gene from expressing as RNA.

Transcription stops when it encounters a DNA palindrome flanking repeated As, forming a 'knot'.

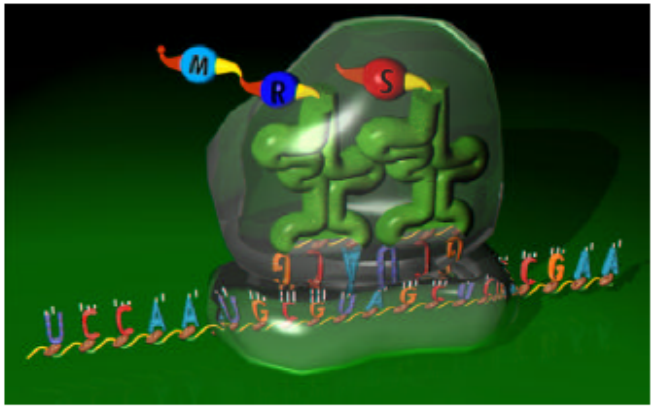


Translation

Translation is the process of building proteins according to the template RNA.

A complex molecule called a *ribosome* works its way along an RNA molecule, grabbing the appropriate amino acid for the next codon and adding to the end of the given protein.

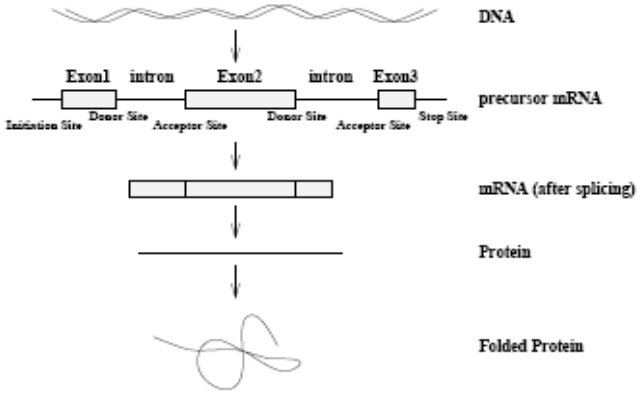
The appropriate bases get to the right places by essentially random motion, guided by electrostatic forces. Binding sites ensure that the right things stick together when they bang into each other.



Introns and Exons

Gene recognition in higher organisms (eukaryotes) is complicated by the presence of *introns*, or non-coding regions.

The coding regions of genes are called *exons*.

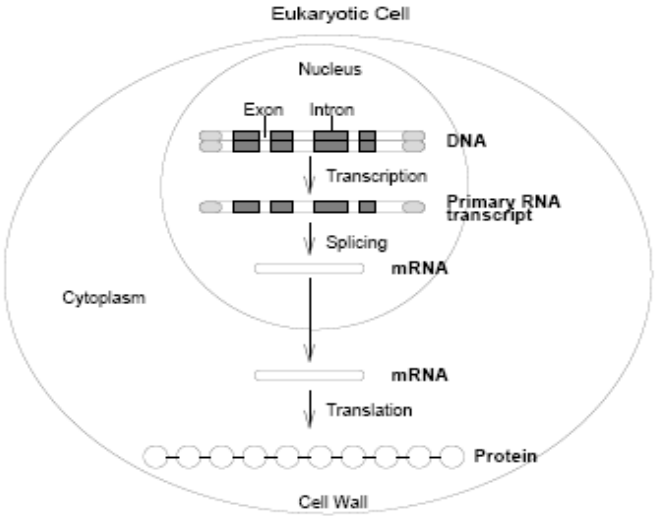


There is controversy about why introns exist. Presumably it is easier to evolve new genes by shuffling small parts, i.e. exons.

Some theorize that prokaryotes originally also had introns, but lost them.

Synthesis on a cellular level

DNA in Eukaryotes resides in the cell's *nucleus*, but proteins are translated outside the nucleus.



Issues of how proteins/RNA cross membrane boundaries are critical in understanding their function, and designing drugs.

Features which ease gene prediction

In general, introns are flanked by *donor* and *acceptor* sites GT and AG – however, such pairs should each happen by chance every $4^2 = 16$ bases.

Genes start with ATG and end with a stop codon (TAA, TAG, or TGA) – however, such codons should happen every $64/3 \approx 20$ codons.

The length of all coding regions must be a multiple of three – however coding regions can be split over multiple exons.

The distribution of base triples and heximers differs between coding and non-coding regions – but you need a sufficiently long enough region to trust statistical variations.

Problems which complicate gene prediction

Gene transfer mechanisms often introduce extra copies of genes into genomes, which then diverge through evolution. Distinguishing broken *pseudo-genes* from working genes is a difficult problem.

Sequencing errors can step on donor/acceptor sites and cause apparent frame shifts.

Exons can be separated by several thousand bases.

Genes can overlap each other, appear in different reading frames and on different strands.

Exons can be assembled in multiple ways through *alternative splicing*.

Laboratory based approaches to gene prediction

The traditional way to find genes was to do it in the laboratory.

One method is to extract and sequence RNA, since most RNA is expressed to code for proteins.

A problem with such laboratory methods is that relatively few genes tend to dominate the population of expressed sequences, and hence one discovered duplicates instead of new genes.

Directly sequencing proteins is a difficult procedure, but is becoming easier through mass spectrometry.

Feature based approaches to gene prediction

Gene recognition systems such as *Grail*, *GeneID*, and *GeneParser* work by searching for various ad hoc features of genes, and then identifying regions which score high enough.

Typical features include codon bias, donor / acceptor sites, and coding frame length.

Since stop codons should occur every 20 codons or so, long *open reading frames* or ORFs without stop codons are strongly suggestive of genes.

Dynamic programming can be used to identify the highest scoring regions.

The best gene recognition systems tend to be species-specific, trained on examples of known genes in the given organism.

Homology based approaches to gene prediction

Biology is an inherently finite discipline. There are only a given number of genes in each of a given number of species.

Further, because of evolution, we would assume that there are strong homologies between genes in related species.

Homology-based gene prediction systems such as *Procrustes* scan databases find similarities to previously identified coding regions.

Such homology-based approaches can only identify previously known genes, of course, but the fraction of known genes is growing rapidly.

A different homology-based approach to identify totally unknown genes is to compare two whole genomes and look for conserved regions, on the theory that sequence is only conserved if it is important.