4 groups of introns: group 1, II, nuclear mRNA, nuclear tRNA.

**Group I:** can be self splicing, or not self splicing (2 exchanges of phsphodiester bonds, i.e. transesterification reactions – 1st cleaves 5’ splice site, adds guanosine to the 5’ end of the intron; second cleaves 3’ site with ligation of exons). *Group I self splicing introns interrupt not only organelle primary transcripts, and some bacterial and nuclear genes.*

**Group II:** *confined to mitochondrial and chloroplast genomes, but also found in cyanobacteria (ancestors of chloroplasts) in* ***1993*** *and purple bacteria.*

*-* UAA anticodon from five diverse cyanobacteria, several major groups of chloroplasts contains a single group I intron. Conserved in secondary structure and primary sequence, and occupies the same position (**within the UAA anticodon**). *Homology of intron across chloroplasts and cyanobacteria – present in their common ancestor*.

Conserved Sequence elements: P,Q,R,S base pair to form 3 stem regions.

**Bacterial mRNAs exclusively contain group I or group II introns.** The introns are found in genes encoding thymidylate synthase, ribonucleotide reductase, and anaerobic ribonucleotide reductase. Thymidylate synthase and anaerobic ribonucleotide reductase **in trans** contain internal ORFs that encode ds-DNA endonucleases == trigger homing/site-specific movement of intron sequences to intronless alleles (**retrohoming**) or to ectopic sites (**retrotransposition**).

* Translation and antisense regulation combinations + kinetics of RNA folding can be effectively exploited to control gene expression.
* *Group II introns evolved in bacteria and invaded the nucleus of a primitive eukaryote*
* Intron encoded proteins: required for in-vivo folding of intron RNA into a catalytically active structure. Most bacterial Group II introns have an IEP.
* Associate some introns with specific genetic signals, or wit hthe host replication machinery for mobility.

Ortholog: genes in different species that evolved from a common ancestral gene by speciation. Retain function.

Homolog: gene related to a second gene by descent from a common ancestral DNA sequence (covers orthologs and paralogs)

**Combined approaches**

Maker: extrinsic + ab initio (maps protein and EST data to genome, to validate predictions)

Augustus: Incorporate hits in the form of EST alignments or protein profiles (accuracy)

\***sequence similarity** search methods to identify the locations of protein coding rgns\* ex. AGenDA

- Take in two sequences, align them, search for conserved splice sites around peaks of local sequence similarity -> identify candidate exons

*- existing limitation of* ***GeneMark.hmm*** *is that a gene that overlaps at its 3’ end with an adjacent gene in the opposite orientation can be missed. (you can recover these genes using GeneMark)*

- learn the global nucleotide frequencies

- determine nucleotide frequencies in each of the 3 codon positions using a linear relationship (global == local nucleotide frequencies)

- Initial values of frequency of occurrence of 61 codons obtained

- Adapt for GC content, modify initial value

- Markov chain model (contiguous codon frequencies) : state assignment, coding or non-coding sequence based on global frequencies.

**BASys** – queries multiple programs, database comparison; BLAST query, similarity searches, sequence analyses (PDB protein db for bacteria)

**Issues for gene prediction:**

* incomplete genes are difficult to predict
* short read length
* sequencing errors (can cause frameshifts)
* Avg length of microbial gene is 950 bp
* Illumina GA has a high mismatch rate (especially in long reads)

**Conservation genomics**

* Accrue mutations in functional parts of the genome at a slower rate
* So you can compare genomes of related species, to detect genes (SLAM, SGP, CONTRAST)
* **TWINSCAN**: mice and humans synteny (orthologous genes)

**Comparative gene finding**

* use annotations from one genome to another (projector, GeneMapper, GeneWise)

**Pseudogene** identification

Pseudogenes are a regular feature of bacterial genomes. Found in high numbers in the genomes of recently emerged bacterial pathogens, presumably because of inactivation and degradation of genes that were no longer needed in the host environment. Identified by sequence alignment

* (align homologs, search for truncated Coding Sequences).
* **\*\*IDENTIFICATION IS A PROBLEM!\*\***
* Some groups annotate shortened ORFs as a pseudogene
* Considerably shortened homologs still specifying some function
* FragGeneScan (HMM combines sequencing error models and codon usages to improve prediction of protein-coding regions in short reads)
* MetaGeneAnnotator
* MetaGene

[E. coli](https://en.wikipedia.org/wiki/E._coli) uses 83% AUG (3542/4284), 14% (612) GUG, 3% (103) UUG [6] and one or two others (e.g., an AUU and possibly a CUG)

As genomic GC% increases, the increase in frequency of valine may be explained as compensation for a deficiency of isoleucine.

**Notes:**

Bacterial proteins fold by themselves or use chaperones? Limit to the number of folds in bacterial proteins – proteins have a specific energy waterfall to follow for folding (little pathway route – induces limit on the types of 3d structures that can be taken up). ??? **paradox**

* sequence with alpha helix that’s hydrophobic (can be a signal peptide that can be intermediary to secretion)
* even though you have big pores, there are proteins that can get out and fold spontaneously.

**1D and 2D Gels**

* denaturing gel (length). SDS PAGE.
* Native (size and charge play a role). Mass spec based (peaks gives the different amino acids). Size sorting, and then pH sorting.
* Array gel x RNA gel

6 types of secretion. **Databases** are specific to specific secretion mechanisms. Popular ones III, IV, VI. Injectisomes (Type III) . SecretVI tool. Phage mediated secretion systems (toxin systems are related to phages). Why are there no archaeal pathogens, compared to bacterial pathogens (Brinkman lab has a paper on it) – more selection for, ex., cholera bacteria with virus

Dimension reduction with training in ML. Make training set representative and diverse. **Always** going to run into biases.

Motif based searching (MEME): Ungapped approximate sequence. Motif discovery algorithms used. Compare against existing. De novo searches? None for protein motif discovery found.

There are cases where you want high precision, in recall you want to capture all the people, MCC and accuracy.