Genome Annotation

BCB 5200 Introduction Bioinformatics I

Fall 2017

Tae-Hyuk (Ted) Ahn

Department of Computer Science
Program of Bioinformatics and Computational Biology
Saint Louis University

SAINT LOUIS UNIVERSITY™

EST. 1818

Genome Annotation

- Which sequences code for proteins and structural RNAs?
- What is the function of the predicted gene products?
- Can we link genotype to phenotype? (i.e. What genes are turned on when ? Why do two strains of the same pathogen vary in their pathogenicity?)
- Can we trace the evolutionary history of an organism from its genomic sequence and genome organization? Evolutionary history of a pathway?

After the *de-novo* genome assembly

- NCBI RefSeq project
  - The NCBI Reference Sequence (RefSeq) project provides sequence records and related information for numerous organisms, and provides a baseline for medical, functional, and comparative studies.
  - The International Nucleotide Sequence Database Collaboration (INSDC, made up of GenBank, the European Nucleotide Archive, and the DNA Data Bank of Japan) represents an archival repository of all sequences
  - The RefSeq database is a non-redundant set of reference standards derived from the INSDC databases that includes chromosomes, complete genomic molecules (organelle genomes, viruses, plasmids), intermediate assembled genomic contigs, curated genomic regions, mRNAs, RNAs, and proteins.

RefSeq – Entrez Genomes

- This process flow provides genomic, RNA, and protein RefSeq records derived from assembled and annotated whole genome sequence data submitted to the INSDC.
- This pipeline provides all of the bacterial, viral, organelle, and plasmid RefSeq records and records for some eukaryotic genomes, including plants and fungi, as data becomes publicly available.
- Protein and transcript records are instantiated from the submitted genome sequence annotation or are predicted by NCBI's bacterial or eukaryotic computational annotation process.

RefSeq – Eukaryotic Genome Annotation Pipeline

- This process flow is an automated computational method that provides a copy of the submitted genome assembly in order to provide an annotated genome.
- RefSeq records may include chromosomes, intermediate assembled scaffolds and contigs, and transcripts and proteins.
- Depending on the species, genome annotation may reflect a mixture of transcript-based RefSeq records and computationally predicted transcripts and proteins with varying levels of support from transcript or protein alignments

The NCBI Eukaryotic Genome Annotation Pipeline

RefSeq - Curation-supported RefSeq pipeline

- NCBI staff scientists provide curation support in several ways. Staff leverage the Protein Clusters database to apply consistent nomenclature to orthologous proteins, work with collaborating groups to better represent data ranging from whole genomes to paralogous genes, and react to feedback from users reporting sequence or name improvements.
- NCBI curation staff also work closely with developer staff to provide genomic region, RNA, and protein RefSeq records for a subset of species grouped under the Bilateria node. Transcript and protein records are primarily derived from cDNA records submitted to the INSDC. This process flow is supported by a combination of bioinformatics and a significant level of manual curation.

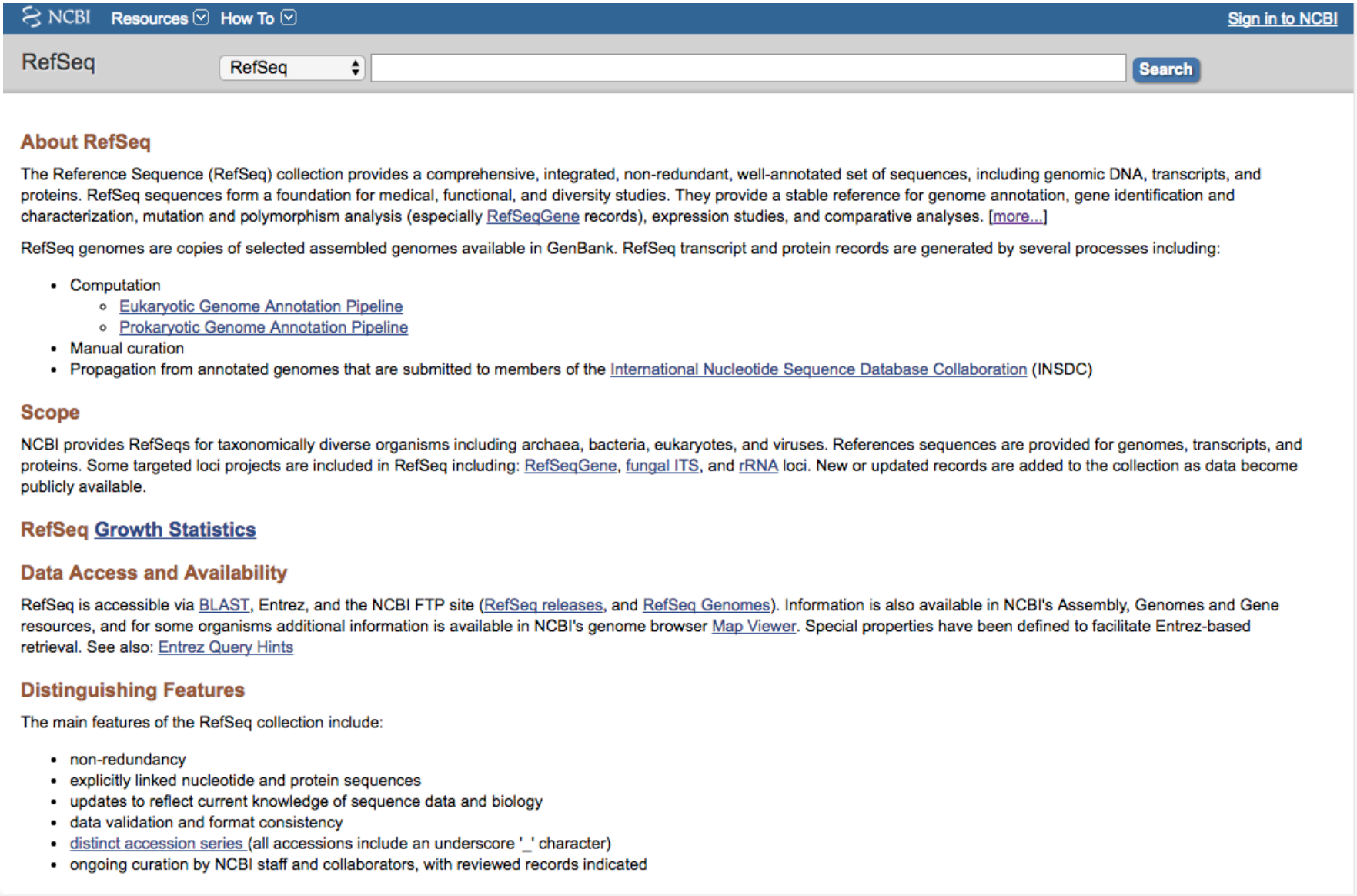
RefSeq - Collaboration

- Some RefSeq records are provided by collaborating groups. Different collaborations provide some fully annotated genomes or records for gene families or individual genes. Collaborations with official nomenclature groups, model organism databases, or other database groups also provide descriptive information, including gene symbols, names, publications, mapping data, feature annotation, database cross-references, and more.

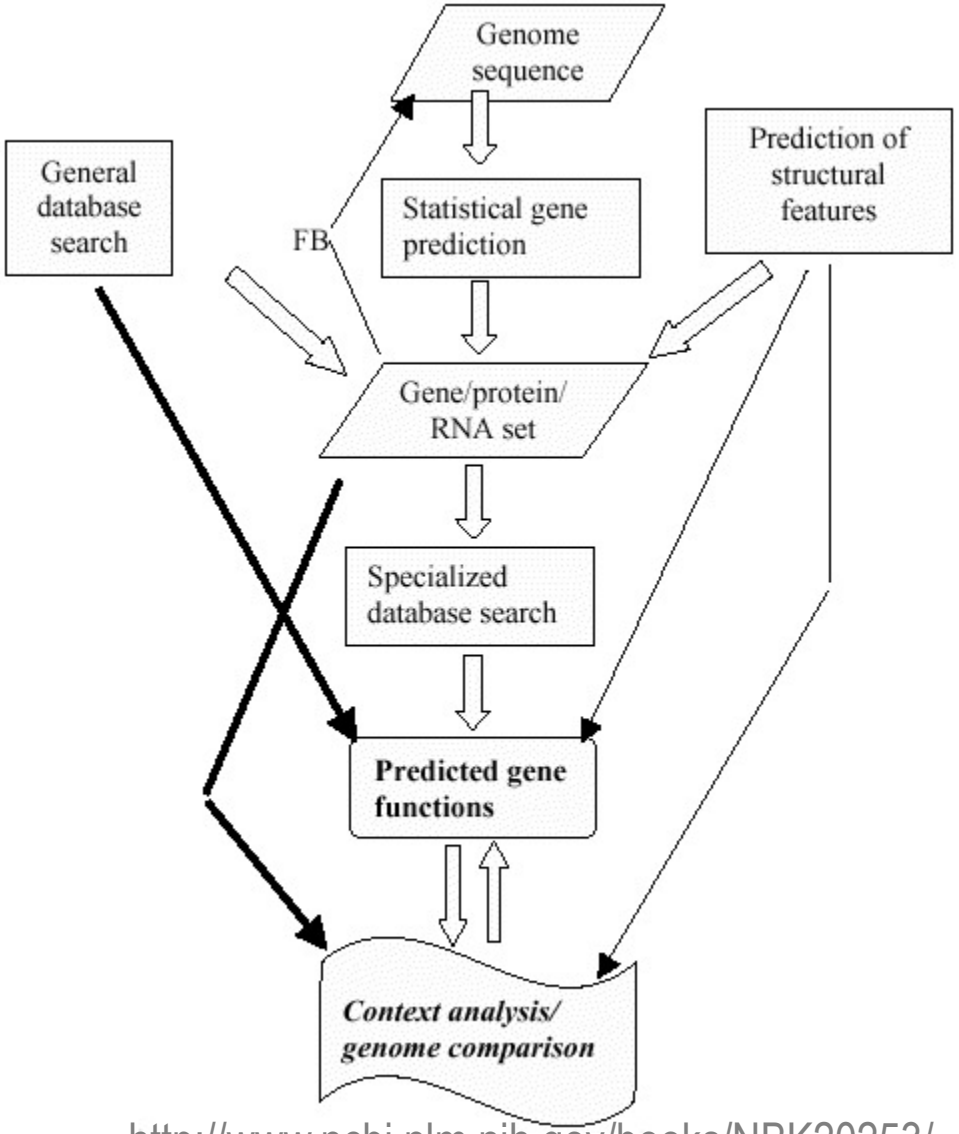
RefSeq Status Codes

Code	Description
MODEL	The RefSeq record is provided by the NCBI Genome Annotation pipeline and is not subject to individual review or revision between annotation runs.
INFERRED	The RefSeq record has been predicted by genome sequence analysis, but it is not yet supported by experimental evidence. The record may be partially supported by homology data.
PREDICTED	The RefSeq record has not yet been subject to individual review, and some aspect of the RefSeq record is predicted.
PROVISIONAL	The RefSeq record has not yet been subject to individual review. The initial sequence-to-gene association has been established by outside collaborators or NCBI staff.
REVIEWED	The RefSeq record has been reviewed by NCBI staff or by a collaborator. The NCBI review process includes assessing available sequence data and the literature. Some RefSeq records may incorporate expanded sequence and annotation information.
VALIDATED	The RefSeq record has undergone an initial review to provide the preferred sequence standard. The record has not yet been subject to final review at which time additional functional information may be provided.
WGS	The RefSeq record is provided to represent a collection of whole genome shotgun sequences. These records are not subject to individual review or revisions between genome updates.

Check the RefSeq



A generalized flow chart of genome annotation

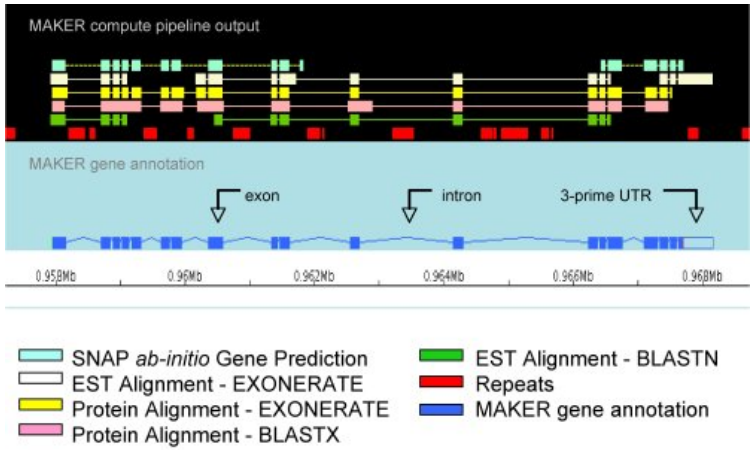


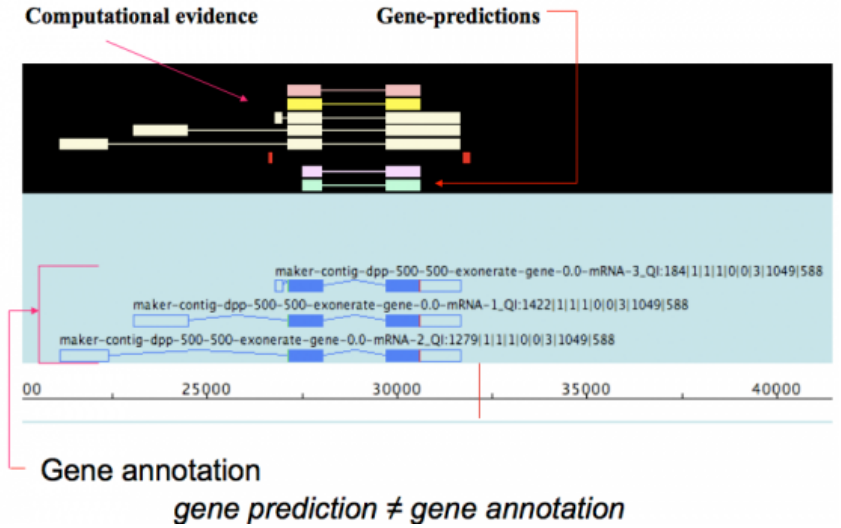
http://www.ncbi.nlm.nih.gov/books/NBK20253/

- Statistical gene prediction: use of methods like GeneMark or Glimmer to predict protein-coding genes.
- FB: feedback from gene identification for correction of sequencing errors, primarily frameshifts.
- General database search: searching sequence databases (typically, NCBI NR) for sequence similarity, usually using BLAST.
- Specialized database search: searching domain databases, such as Pfam, SMART, and CDD, for conserved domains, genome-oriented databases, such as COGs, for identification of orthologous relationship and refined functional prediction, metabolic databases, such as KEGG for metabolic pathway reconstruction, and possibly, other database searches.
- Prediction of structural features: prediction of signal peptide, transmembrane segments, coiled domain and other features in putative protein functions.

Genome Annotation Pipeline

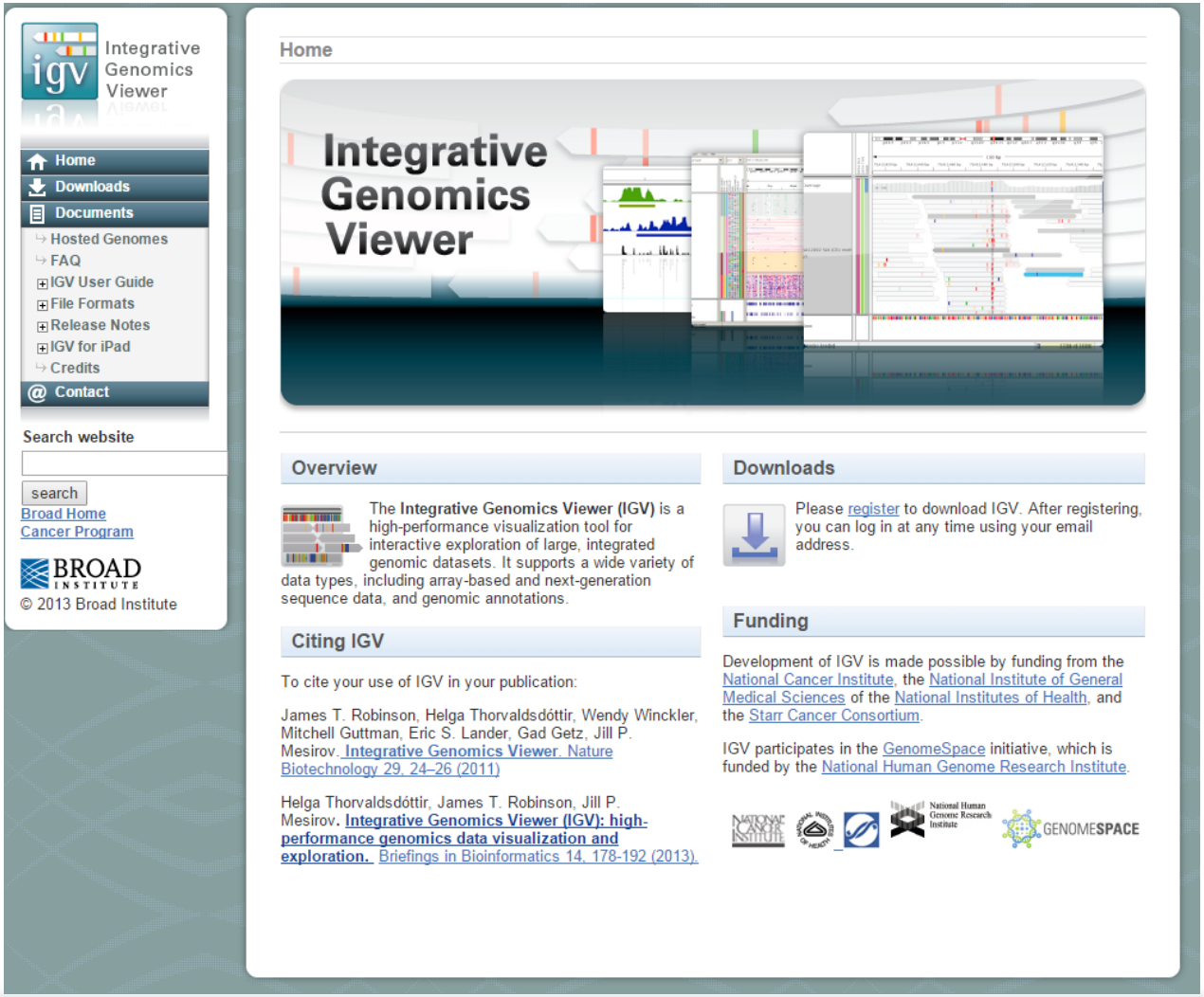
- MAKER 2 (http://www.yandell-lab.org/software/maker.html)
  - Identifies and masks out repeat elements
  - Aligns ESTs to the genome
  - Aligns proteins to the genome
  - Produces ab initio gene predictions
  - Synthesizes these data into final annotations
  - Produces evidence-based quality values for downstream annotation management





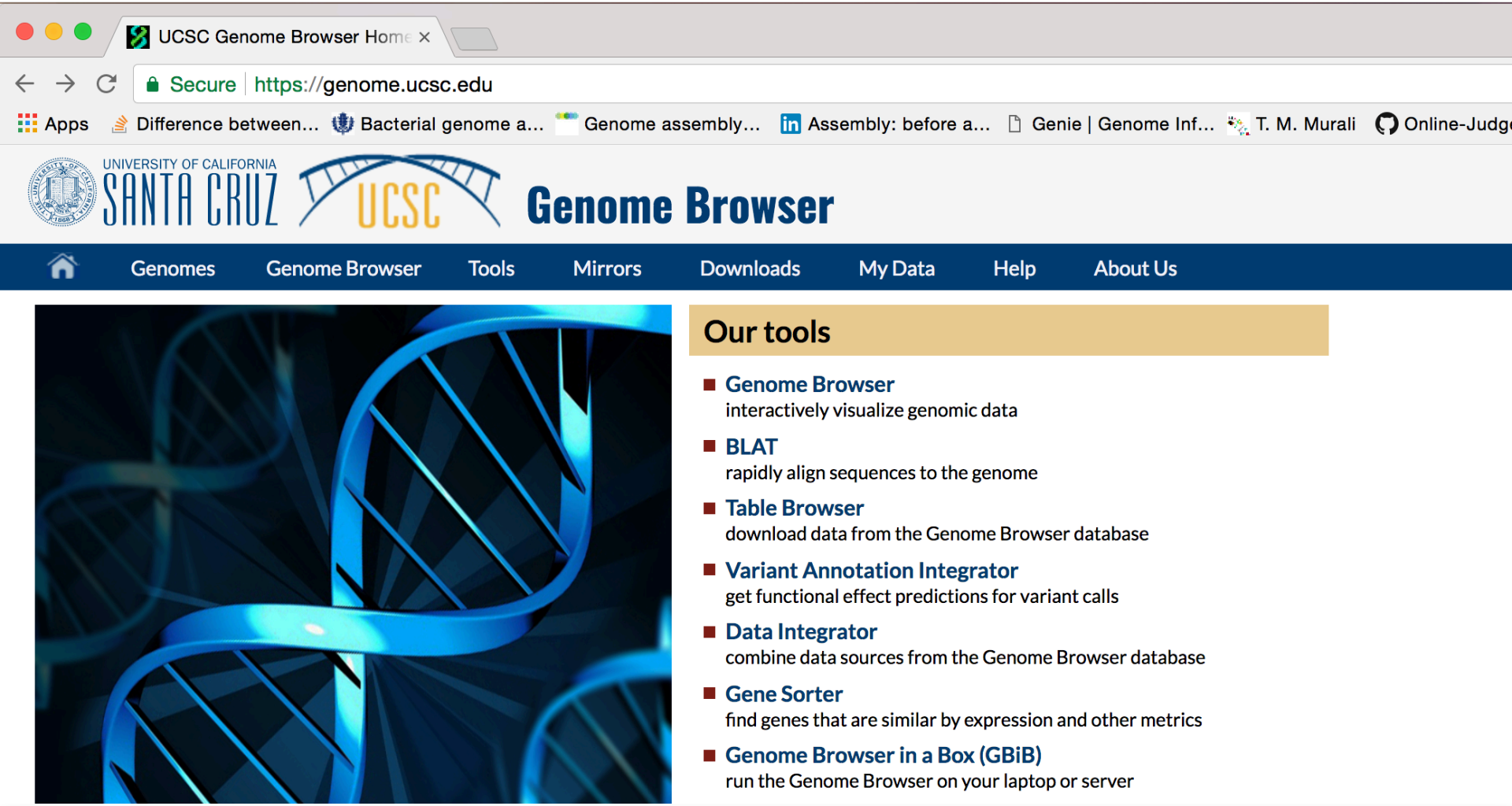
Genome Browser: IGV

- https://www.broadinstitute.org/igv/



Web: UCSC Genome Browser

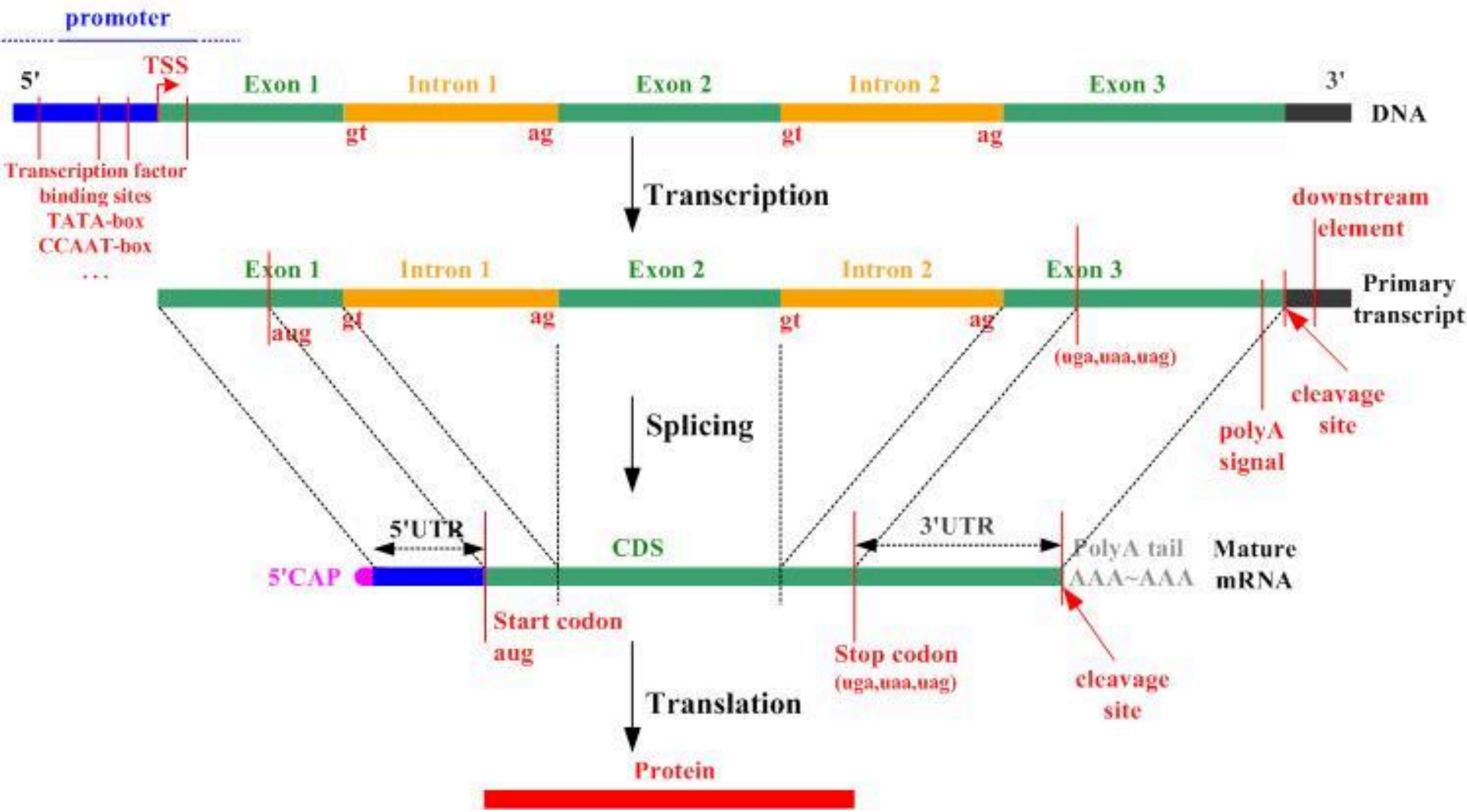
- https://genome.ucsc.edu/



OpenHelix

- http://www.openhelix.com
- http://www.openhelix.com/ucsc : homework

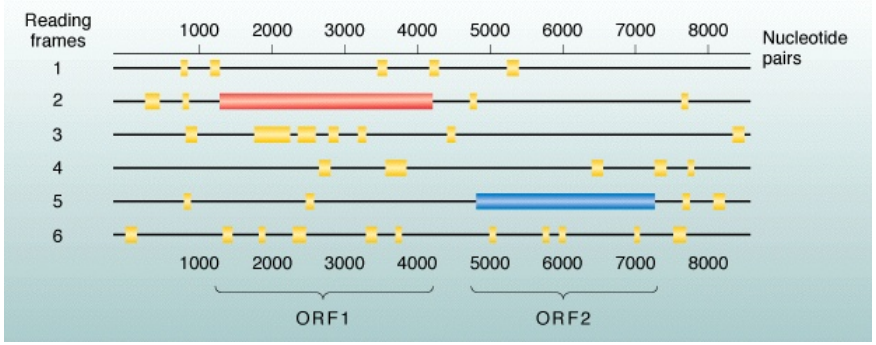
CDS And ORF


promoter
5'
TSS
Exon 1
Intron 1
Exon 2
Intron 2
Exon 3
3'
DNA
gt
ag
gt
ag
Transcription factor
binding sites
TATA-box
CCAAT-box
...
Transcription
downstream
element
Exon 1
Intron 1
Exon 2
Intron 2
Exon 3
Primary
transcript
aug
gt
ag
gt
ag
(uga,uaa,uag)
cleavage
site
polyA
signal
Splicing
5'UTR
CDS
3'UTR
PolyA tail
AAA~AAA
Mature
mRNA
5'CAP
Start codon
aug
Stop codon
(uga,uaa,uag)
cleavage
site
Translation
Protein

Gene finding

- Begins with the prediction of gene models through the

1. Identification of Open Reading Frames (ORFs)



2. Examination of base composition differences between coding vs. non-coding regions

3. Computational gene recognition (exons, introns, exo-intron boundaries) using a variety of gene-finding algorithms (GLIMMER, GRAIL, FGENEH, GENSCAN GLIMMER-HMM, etc…)

FINDING ORFs

The simplest method in prokaryotes is to scan the DNA for start and stop codons

The DNA is double stranded and each strand has three potential reading frames (codons are groups of 3 bases)

```
THE CAT ATE THE RAT     Frame 1

T HEC ATA TET HER AT    Frame 2

TH ECA TAT ETH ERA T    Frame 3
```

```
Start Codon: ATG (AUG)
Stop  Codon: TAG (UAG)
             TAA (UAA)
             TGA (UGA)
```

The scan must look at all 6 reading frames

Ex) 5'ATGATGTGATGTAAATAAATTGAT 3'

FINDING ORFs

- Any region of DNA between a start codon and a stop codon in the same reading frame could potentially code for a polypeptide and is therefore an ORF

Start AUG (methionine)

Stop UAA UAG UGA

- Problems
  - Small genes may be missed
  - The actual start codon may be internal to the ORF
  - There may be overlapping genes

Gene finding (cont')

- Another gene finding/confirmation approach is based on experimental evidence using homology

1. Alignment of Expressed Sequence Tags (EST) and full cDNA sequences with gDNA
   - Advantages: gene discovery, proof of expression, training for gene finders
   - Disadvantages: Disproportionate representations
2. Examination of protein translation profiles: Peptide sequencing, mass spectrometry, etc…

Gene finding (cont')

The gene finding task comes with various levels of difficulty in different organisms

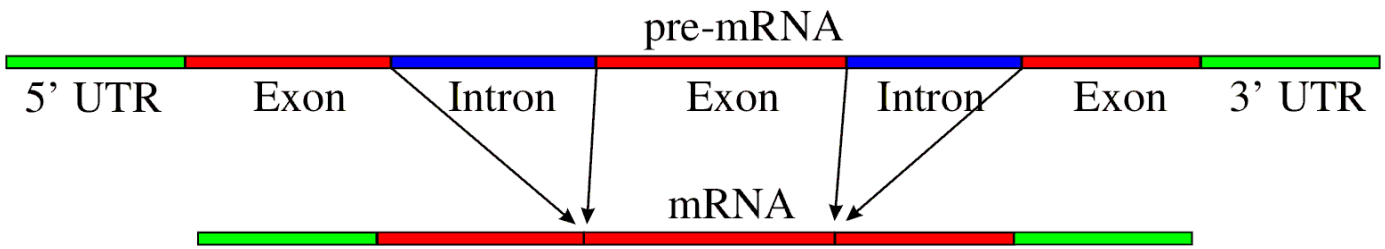
Relatively easy in bacterial and archeal genomes mostly due to:

1) High gene density (500-1000 genes/Mb)
2) Short intergenic regions
3) Lack of introns

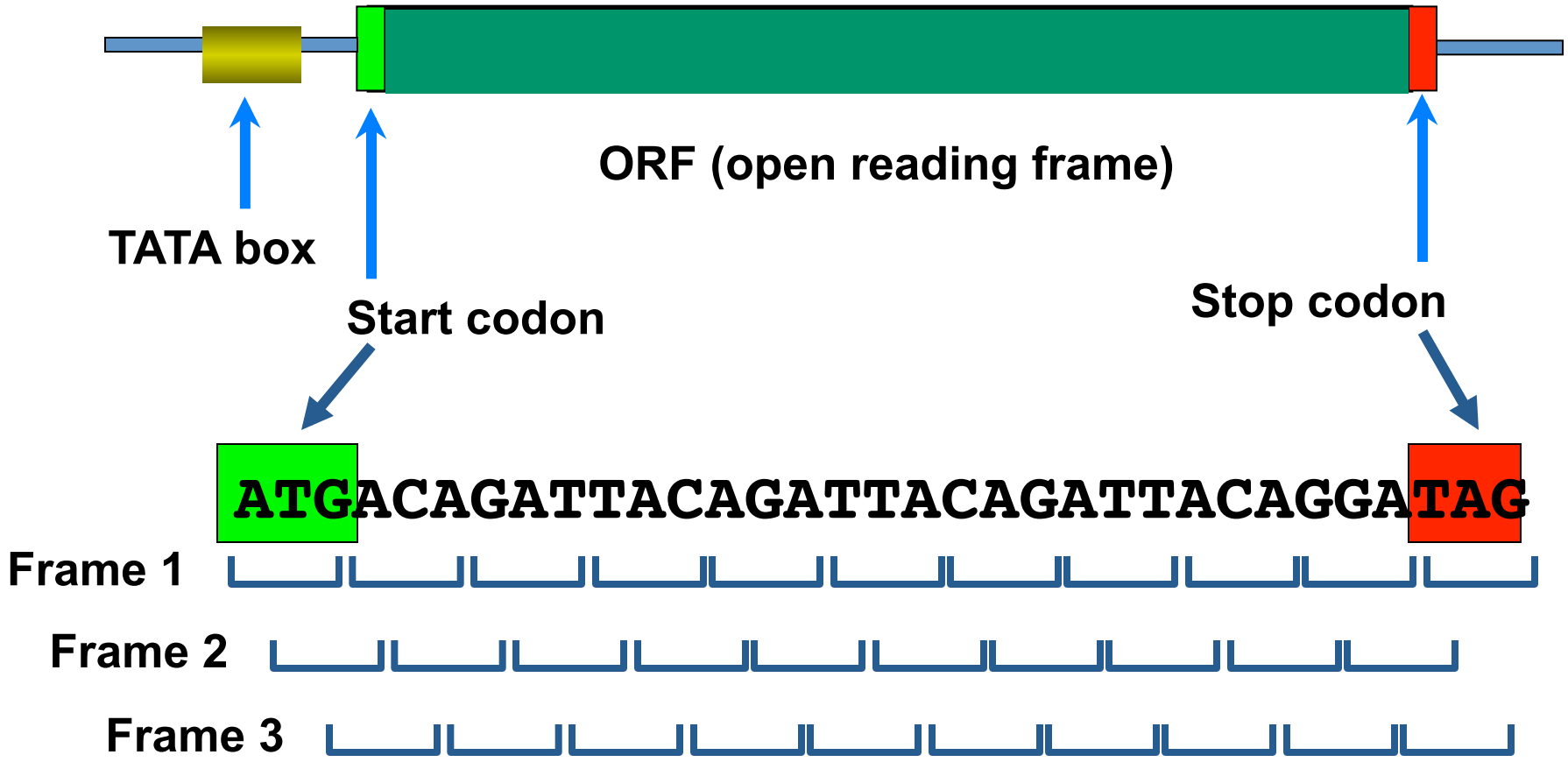
Much more difficult in eukaryotic genomes and can become major focus of activity in the annotation phase of a genome:

1) Low gene density (The human genome has a gene density of 12-15 genes/Mb)
2) Presence of repeats
3) Most eukaryotic genes have introns and exons, alternative splicing

Inaccurate predictions and false positives are common



Prokaryotic gene structure



Prokaryotes

- Advantages
  - Simple gene structure
  - Small genomes (0.5 to 10 million bp)
  - No introns
  - Genes are called Open Reading Frames (ORFs)
  - High coding density (>90%)
- Disadvantages
  - Some genes overlap (nested)
  - Some genes are quite short (<60 bp)

Gene finding approaches

1) Rule-based (e.g, start & stop codons)
2) Content-based (e.g., codon bias, promoter sites)
3) Similarity-based (e.g., orthologs)
4) Pattern-based (e.g., machine-learning)
5) *Ab-initio* methods (FFT)

Simple rule-based gene finding

- Look for putative start codon (ATG)
- Staying in same frame, scan in groups of three until a stop codon is found
- If # of codons >=50, assume it's a gene
- If # of codons <50, go back to last start codon, increment by 1 & start again
- At end of chromosome, repeat process for reverse complement

ORF Finder

- The simplest tool for finding ORFs is ORF Finder at NCBI
- It simply scans all 6 reading frames and shows the position of the ORFs which are greater than a user defined minimum size
- The genetic code used for the analysis can be altered by the user

ORF Finder

Secure https://www.ncbi.nlm.nih.gov/orffinder/

Apps | Ingenuity Pathway... | C++ version of rea... | Systems biology a... | Systems biology in... | Video Lectures | In...

NCBI Resources How To Sign in to NCBI

ORFfinder PubMed Search

Open Reading Frame Finder

ORF finder searches for open reading frames (ORFs) in the DNA sequence you enter. The program returns the range of each ORF, along with its protein translation. Use ORF finder to search newly sequenced DNA for potential protein encoding segments, verify predicted protein using newly developed SMART BLAST or regular BLASTP.

This web version of the ORF finder is limited to the subrange of the query sequence up to 50 kb long. Stand-alone version, which doesn't have query sequence length limitation, is available for Linux x64.

Examples (click to set values, then click Submit button) :

- NC_011604 Salmonella enterica plasmid pWES-1; genetic code: 11; 'ATG' and alternative initiation codons; minimal ORF length: 300 nt
- NM_000059; genetic code: 1; start codon: 'ATG only'; minimal ORF length: 150 nt

Enter Query Sequence

Enter accession number, gi, or nucleotide sequence in FASTA format:

From: **To:**

Content based gene prediction method

To overcome the limitations of ORF finder, more sophisticated programmes detect compositional biases and increase the reliability of gene detection

- These compositional biases are regular, though very diffuse, and arise for a variety of reasons
- many organisms there is a detectable preference for G or C over A and T in the third ("wobble") position in a codon
- All organisms do not utilize synonymous codons with the same frequency - consequently there is a codon bias
- There is an unequal usage of amino acids in proteins sufficient to cause a bias in all three positions of codons and increase the overall codon bias

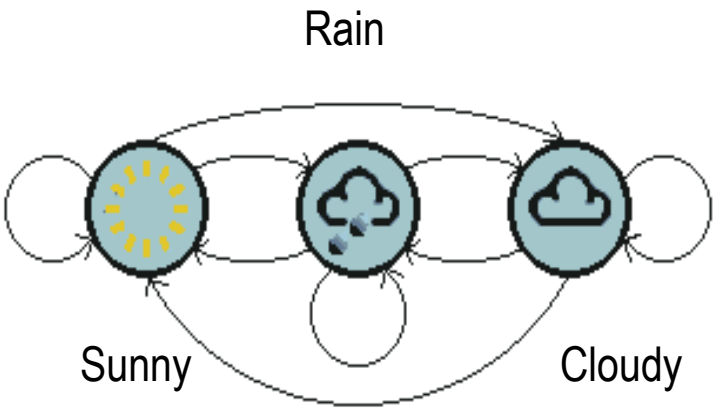
GC Contents

- The %GC content of the first two codon positions of the universal genetic code is approximately 50%, therefore, organisms which have a low or high %GC content will exhibit a marked bias at the third position of codons to achieve their overall %GC content

The most recent approaches to using compositional features to distinguish coding from non-coding regions employ 'Markov models'

- such approaches include the popular GENEMARK and GLIMMER programs

Markov Chains



States : Three states - sunny, cloudy, rainy.

weather today

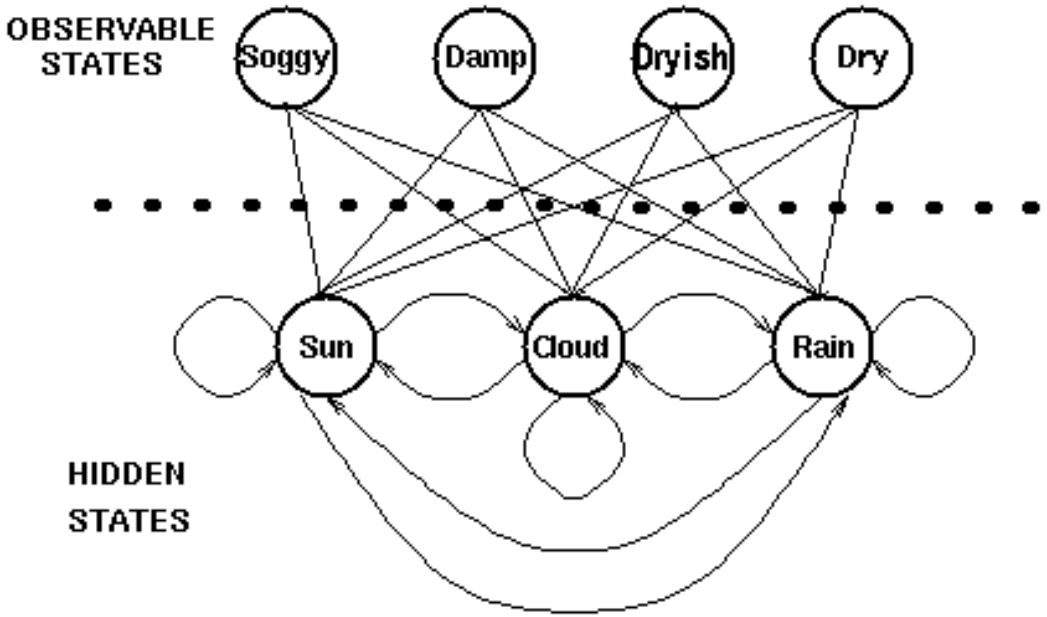
weather yesterday	Sun	Cloud	Rain
Sun	0.5	0.25	0.25
Cloud	0.375	0.125	0.375
Rain	0.125	0.625	0.375

State transition matrix : The probability of the weather given the previous day's weather.

Sun	Cloud	Rain
1.0	0.0	0.0

Initial Distribution : Defining the probability of the system being in each of the states at time 0.

Hidden Markov Models



Hidden states : the (TRUE) states of a system that may be described by a Markov process (e.g., the weather).

Observable states : the states of the process that are `visible' (e.g., seaweed dampness).

Components Of HMM

		Seaweed			
		Dry	**Dryish**	**Damp**	**Soggy**
weather	**Sun**	0.60	0.20	0.15	0.05
	Cloud	0.25	0.25	0.25	0.25
	Rain	0.05	0.10	0.35	0.50

Output matrix : containing the probability of observing a particular observable state given that the hidden model is in a particular hidden state.

Initial Distribution : contains the probability of the (hidden) model being in a particular hidden state at time t = 1.

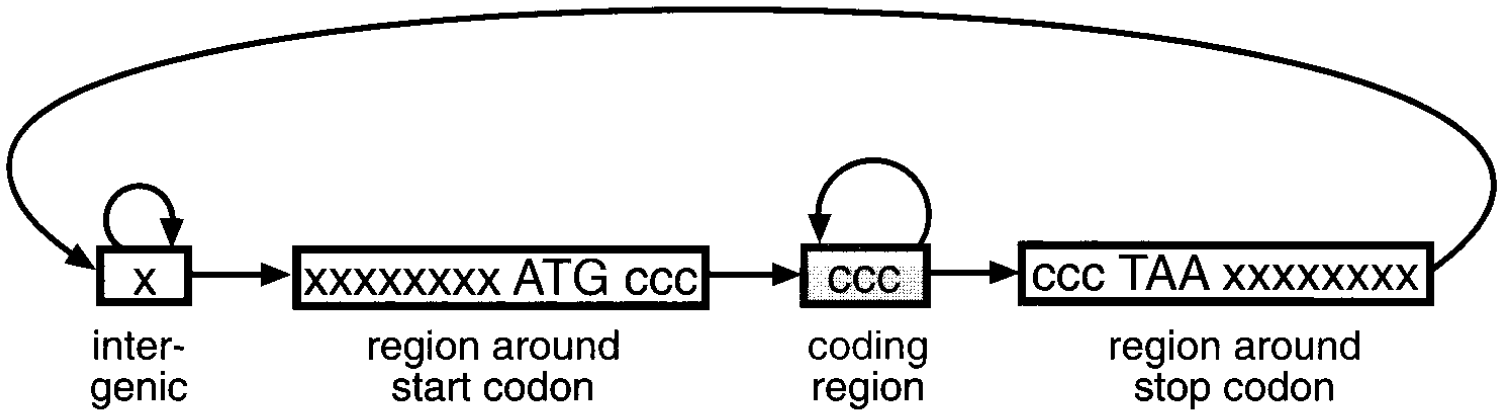
State transition matrix : holding the probability of a hidden state given the previous hidden state.

Gene Finding HMMs

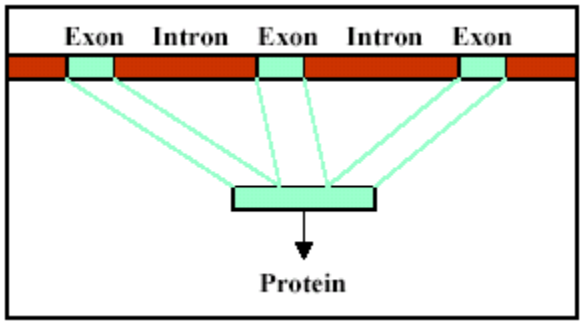
- Our Objective:
  - To find the coding and non-coding regions of an unlabeled string of DNA nucleotides
- Our Motivation:
  - Assist in the annotation of genomic data produced by genome sequencing methods
  - Gain insight into the mechanisms involved in transcription, splicing and other processes

HMMs for gene finding

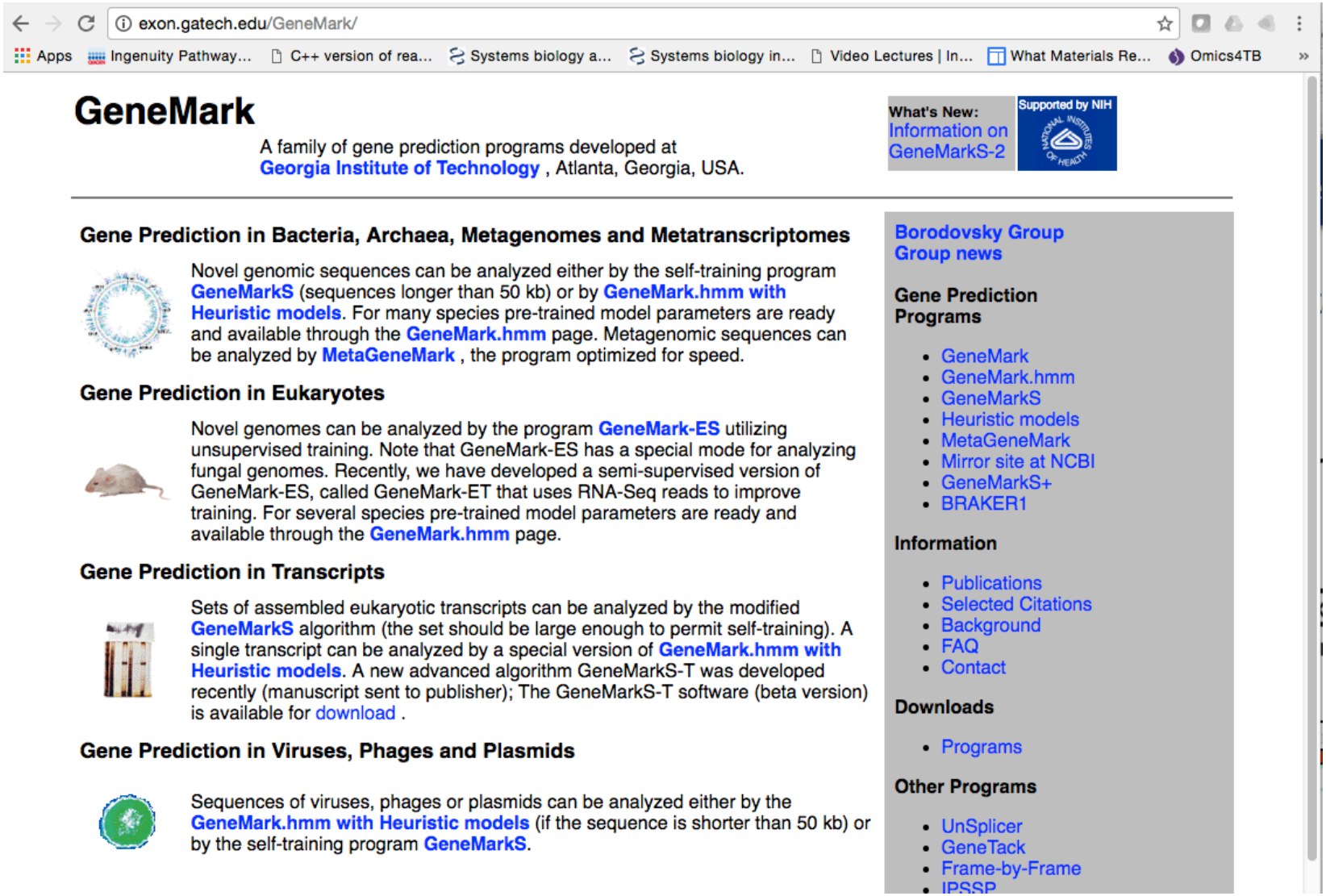
- **Training** - Expectation Maximization (EM)
- **Parsing** – Viterbi algorithm



An HMM for unspliced genes.
x : non-coding DNA
c : coding state



GeneMark


exon.gatech.edu/GeneMark/
Apps
Ingenuity Pathway...
C++ version of rea...
Systems biology a...
Systems biology in...
Video Lectures | In...
What Materials Re...
Omics4TB
GeneMark
A family of gene prediction programs developed at
Georgia Institute of Technology , Atlanta, Georgia, USA.
What's New:
Information on
GeneMarkS-2
Supported by NIH
Gene Prediction in Bacteria, Archaea, Metagenomes and Metatranscriptomes
Novel genomic sequences can be analyzed either by the self-training program GeneMarkS (sequences longer than 50 kb) or by GeneMark.hmm with Heuristic models. For many species pre-trained model parameters are ready and available through the GeneMark.hmm page. Metagenomic sequences can be analyzed by MetaGeneMark , the program optimized for speed.
Gene Prediction in Eukaryotes
Novel genomes can be analyzed by the program GeneMark-ES utilizing unsupervised training. Note that GeneMark-ES has a special mode for analyzing fungal genomes. Recently, we have developed a semi-supervised version of GeneMark-ES, called GeneMark-ET that uses RNA-Seq reads to improve training. For several species pre-trained model parameters are ready and available through the GeneMark.hmm page.
Gene Prediction in Transcripts
Sets of assembled eukaryotic transcripts can be analyzed by the modified GeneMarkS algorithm (the set should be large enough to permit self-training). A single transcript can be analyzed by a special version of GeneMark.hmm with Heuristic models. A new advanced algorithm GeneMarkS-T was developed recently (manuscript sent to publisher); The GeneMarkS-T software (beta version) is available for download .
Gene Prediction in Viruses, Phages and Plasmids
Sequences of viruses, phages or plasmids can be analyzed either by the GeneMark.hmm with Heuristic models (if the sequence is shorter than 50 kb) or by the self-training program GeneMarkS.
Borodovsky Group
Group news
Gene Prediction Programs
GeneMark
GeneMark.hmm
GeneMarkS
Heuristic models
MetaGeneMark
Mirror site at NCBI
GeneMarkS+
BRAKER1
Information
Publications
Selected Citations
Background
FAQ
Contact
Downloads
Programs
Other Programs
UnSplicer
GeneTack
Frame-by-Frame
IPSSP

Similarity-based gene finding

- Take all known genes from a related genome and compare them to the query genome via BLAST
- Disadvantages:
  - Orthologs/paralogs sometimes lose function and become pseudogenes
  - Not all genes will always be known in the comparison genome (big circularity problem)
  - The best species for comparison isn't always obvious
- Summary: Similarity comparisons are good supporting evidence for prediction validity

Ab-initio Methods

- Ab Initio gene prediction is an intrinsic method based on gene content and signal detection.
- Genomic DNA sequence alone is systematically searched for certain tell-tale signs of protein-coding genes.
- These signs can be broadly categorized as either signals, specific sequences that indicate the presence of a gene nearby, or content, statistical properties of protein-coding sequence itself.
- Ab initio gene finding might be more accurately characterized as gene prediction.
- Fast Fourier Transform based methods.
- Able to identify new genes.

Machine Learning Techniques

- Hidden Markov Model
- ANN (Artificial neural network) based method
- Bayes Networks

Eukaryotes

- Complex gene structure
- Large genomes (0.1 to 3 billion bases)
- Exons and Introns (interrupted)
- Low coding density (<30%)
  - 3% in humans, 25% in Fugu, 60% in yeast
- Alternate splicing (40-60% of all genes)
- Considerable number of pseudogenes

Finding Eukaryotic Genes Computationally

- Rule-based
  - Not as applicable – too many false positives
- Content-based Methods
  - CpG islands, GC content, hexamer repeats, composition statistics, codon frequencies
- Feature-based Methods
  - donor sites, acceptor sites, promoter sites, start/stop codons, polyA signals, feature lengths
- Similarity-based Methods
  - sequence homology, EST searches
- Pattern-based
  - HMMs, Artificial Neural Networks
- Most effective is a combination of all the above

Gene prediction programs

- Prokaryotes: Prodigal (http://prodigal.ornl.gov/)
- Rule-based programs
  - Use explicit set of rules to make decisions.
  - Example: GeneFinder
- Neural Network-based programs
  - Use data set to build rules.
  - Examples: Grail, GrailEXP
- Hidden Markov Model-based programs
  - Use probabilities of states and transitions between these states to predict features.
  - Examples: GeneMark, Genscan, GenomeScan

Combined Methods

- GRAIL (http://compbio.ornl.gov/Grail-1.3/)
- FGENEH (http://www.bioscience.org/urllists/genefind.htm)
- HMMgene (http://www.cbs.dtu.dk/services/HMMgene/)
- GENSCAN(http://genes.mit.edu/GENSCAN.html)
- GenomeScan (http://genes.mit.edu/genomescan.html)
- Twinscan (http://ardor.wustl.edu/query.html)