Gene Identification and Genome Annotation

Jutta Marzillier, PhD
Lehigh University
Department of Biological Sciences
March 15th, 2011

Outline

High Throughput Next Generation Sequencing

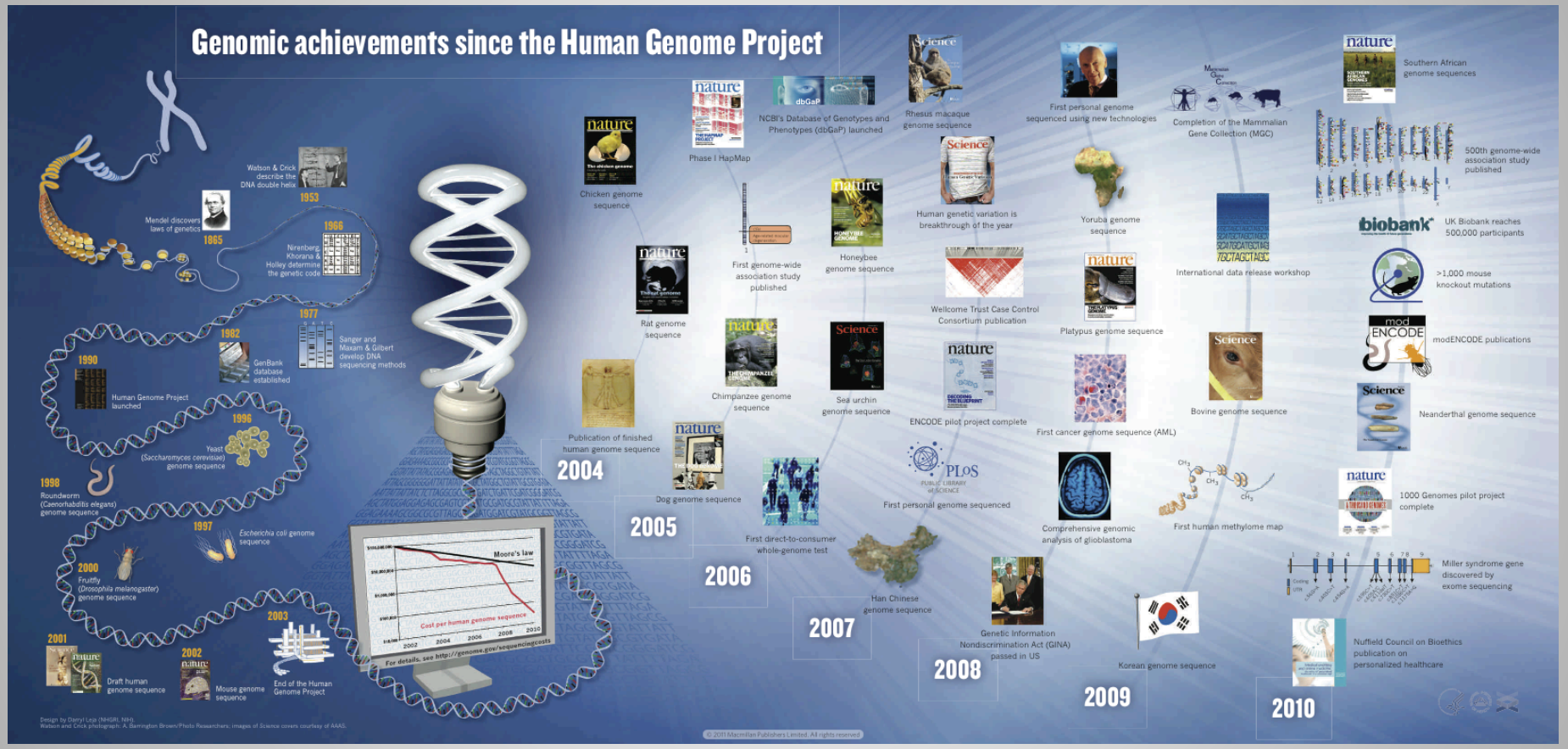
Data Collection, Organization and Interpretation

Gene Calling – Definition of a Gene

Gene Calling Software: Glimmer and Genemark

Life Sciences Thrive Trough their Co-laboration with Engineering and BioInformatics

Genomic Achievements since the Human Genome Project.

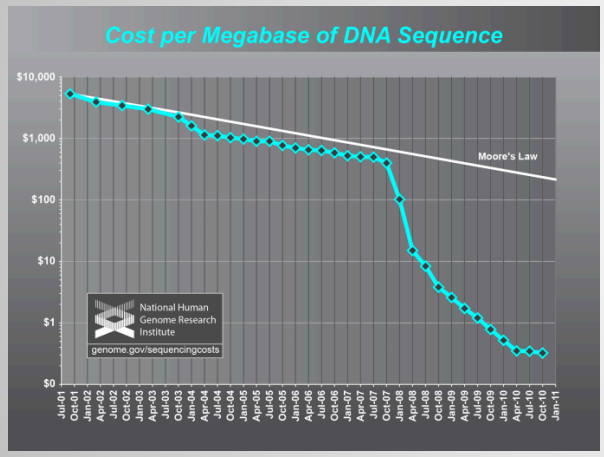


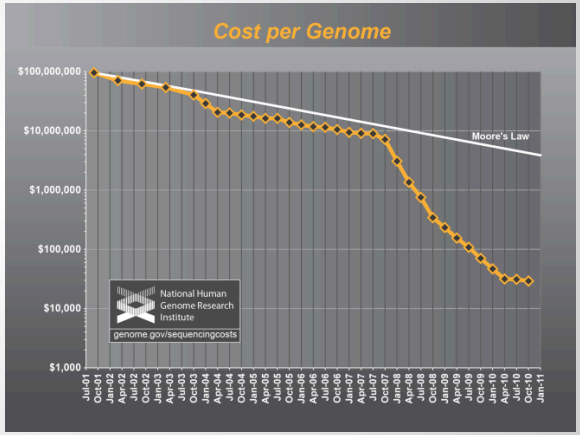
ED Green *et al. Nature* **470**, 204-213 (2011) doi:10.1038/nature09764

nature

Sudden and profound out-pacing of Moore's Law

Transition from Sanger Sequencing to Next-Generation Sequencing technologies





Moore's Law, which describes a long-term trend in the computer hardware industry that involves the doubling of 'compute power' every two years

www.genome.gov/sequencingcosts.

Dealing with Data

Science, Vol 331, 11 Feb 2011

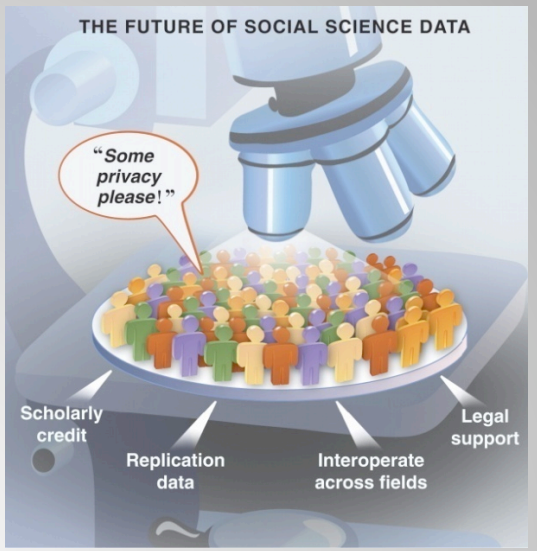
Science and technology are essential to improving public health and welfare

There is a huge influx of research data

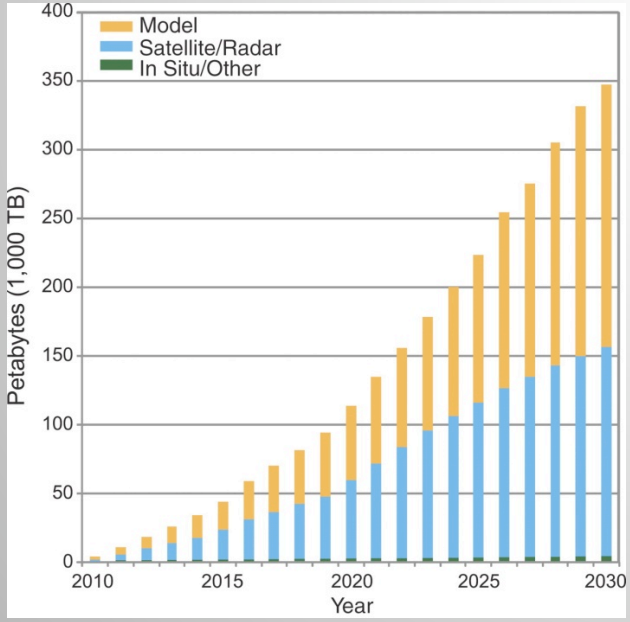
Data collection, curation and access are central to look at data in a meaningful way

Data Collection Examples

Social Sciences



Climate change



Global Health

Science, Vol 331, 11 Feb 2011

Finishing Draft Sequences using Assembly Software

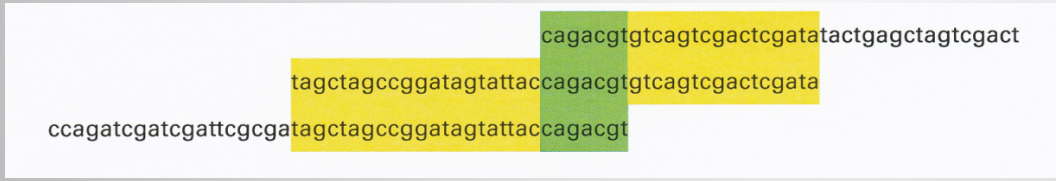
Primary sequencing reads:

cagacgtgtcagtcgactcgatatactgagctagtcgact

tagctagccggatagtattaccagacgtgtcagtcgactcgata

ccagatcgatcgattcgcgatagctagccggatagtattaccagacgt

Align sequences for homologies



Green: 3-fold coverage

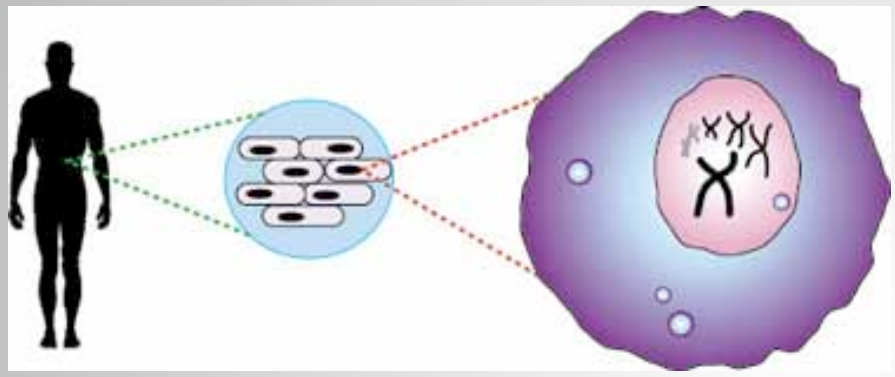
Yellow: 2-fold coverage

Fragment assembly to contig

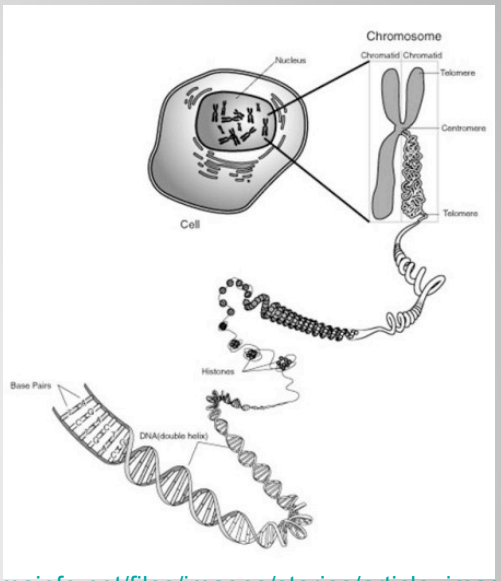
Goal for Genome Sequencing:
Sanger: 8-fold coverage
454: 30 fold coverage

TGCGGCTGCCAGATTTTGTACGGGTTTGGAAGTCGACGGAGAGAACAGCG
CGGGCCTAGAAGGCCCCGTAATGCCCCCTGAGAGCCCCGTAGACGGACG
AACGGTGCGGATCGATAGATGGCACCGGAGACAAGCGAAGACGGCCGCA
GAGCCGTCGCCGGCTGACGCCCGCGTAGGAAGATATTCGTGTGAAGTGC
GTCACATTCTACGGGTGAAACGCGAAAGTGGAAGGTTCCTTACCTATGGAG
GGGTAAGGGAGCGAGCTCCAGCGAGCGACCGCACCCCGACATAGGTTCT
TGTCGGGGTAGTCGAACGGAGAGAGACTACCCCTTTTAGCGACCTCCGGT
CGCCCAGGTAGGTACCGAACGATGAGAGAGGTACCTAGACCGTACAGGCC
GGGGTTTATCCCCCGGCCGATACAGCATGGTCATTTTGGGTAGGTACGTTA
CGTAAGCATCACTCACCAACAGAACCAGTGGTTACGTAACCGGGTACGTTA
CGTACGACTAGATACGTAACAGAACCACTAACCCGTGGCCGCCCGAAGGC
GGCCCGCAGCGGGTTACGTTTGTCAGGTATGTCACTAGGGAGGGTGAACT
ATGAATCTCAAGGAGCATCAGAAGGAGCTTCTTCGTCTACTGCTCGTGTCT
GAGTCGATGACGTATGAGCGTGCCTACTCGAAGGTGGAGACCATCCTGCA
CCTCATCCAGTCCTTCGACGCGGAGAACTCGATAAGTGAGCTGGGAGTCA
TCTGACCGGCGCGATCGTCTGCCGACCGACTGGCCTCGCATCCGCCGCG
AGGTTCTGCGGGCGGCTGGCCACCGCTGCCAGATCCGCTACGCGGACAT
CTGCACAGGGATGGCTACCGAGGTTGATCACGTCCGCTACCGCGACGAG
GCGTCACCTCTGCAGGTGTCGTGCAGACCGTGCCATGCGCGGAAGTCCG
CGATGGAAGGCGTTGCTCAGCGTGCGAAGCTGCGCGCGATGAAGAAGCG
GCCGCCGCCCCGCCACCCGGGGCGTAGAAGCAACTAGGAGGGACCAGG
CGTCCCCGAGCCCAGGAGGCGTCATGCCGGGTCCAGTGCCCAAGCGATC
GGACGAACGCGTCCGGCGCAAATCGCATACGAGTGAGCGCGAGGCTAGC

Why is the Knowledge about the Human Genome interesting?



http://www.genomenewsnetwork.org/articles/06_00/sequence_primer.shtml



http://www.pharmainfo.net/files/images/stories/article_images/

The human body has about 100 trillion cells.
Each cell harbors the same genetic information in its nucleus in form of DNA containing chromosomes.
Depending on cellular, developmental, and functional stage of a cell only a subset of genes is expressed.

How do we Identify a Gene?

DNA sequence of the insulin gene:

AGCCCTCCAGGACAGGCTGCATCAGAAGAGGCCATCAAGCAGATCACTGTCCTTCTGCCATGGCCCTG
TGGATGCGCCTCCTGCCCCTGCTGGCGCTGCTGGCCCTCTGGGGACCTGACCCAGCCGCAGCCTTTG
TGAACCAACACCTGTGCGGCTCACACCTGGTGGAAGCTCTCTACCTAGTGTGCGGGGAACGAGGCTTC
TTCTACACACCCAAGACCCGCCGGGAGGCAGAGGACCTGCAGGTGGGGCAGGTGGAGCTGGGCGGG
GGCCCTGGTGCAGGCAGCCTGCAGCCCTTGGCCCTGGAGGGGTCCCTGCAGAAGCGTGGCATTGTG
GAACAATGCTGTACCAGCATCTGCTCCCTCTACCAGCTGGAGAACTACTGCAACTAGACGCAGCCCGCA
GGCAGCCCCACACCCGCCGCCTCCTGCACCGAGAGAGATGGAATAAAGCCCTTGAACCAGCAAAA

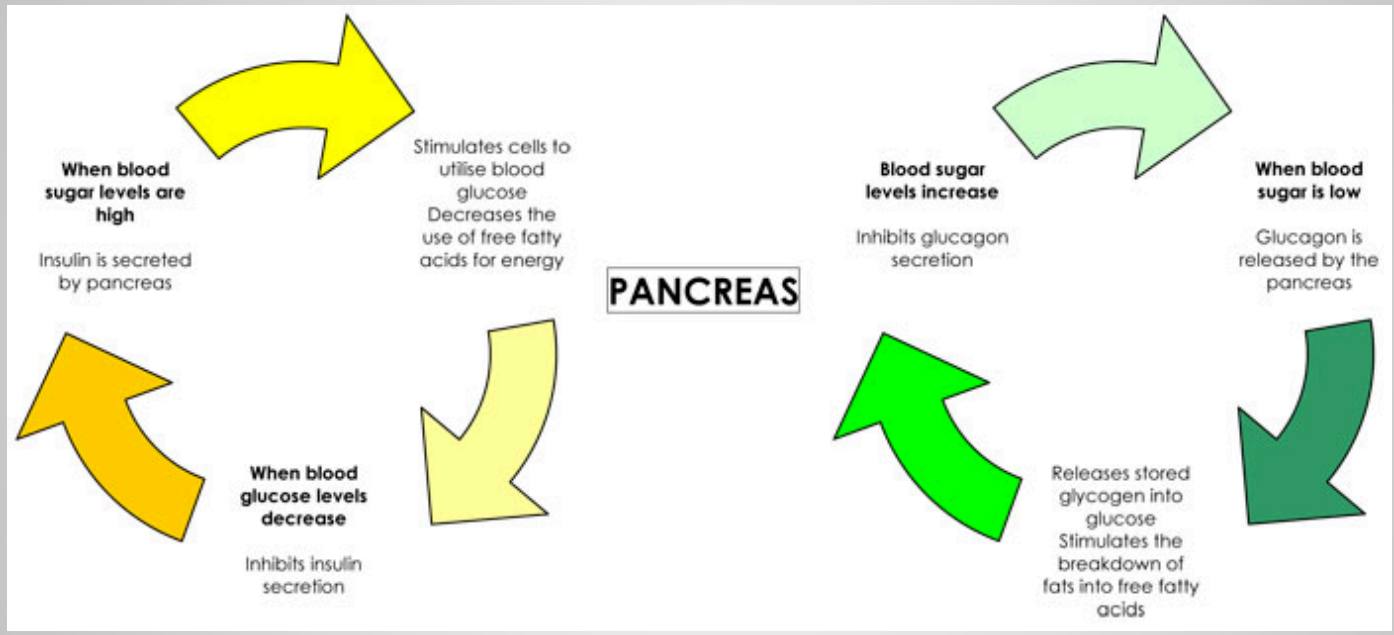
Unraveling the sequence of the bases was the ‘easy’ part with a defined endpoint

Next comes to figure out the meaning of these sequences of A’s, G’s, T’s, and C’s

Examples:

- Which stretches of DNA correspond to genes?
- When and in what tissues are these genes expressed?
- What kind of proteins do they code for?
- How do all these proteins interact with each other and function?

Example: Regulation of Insulin Secretion



http://www.google.com/imgres?imgurl=http://pilatespower.tv/Portals/0/Articles/LooseWeight-001-

The Biological 'Dogma'

DNA: storage of genetic information

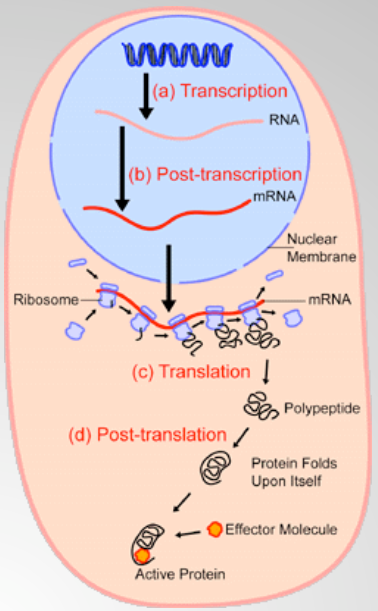
↓

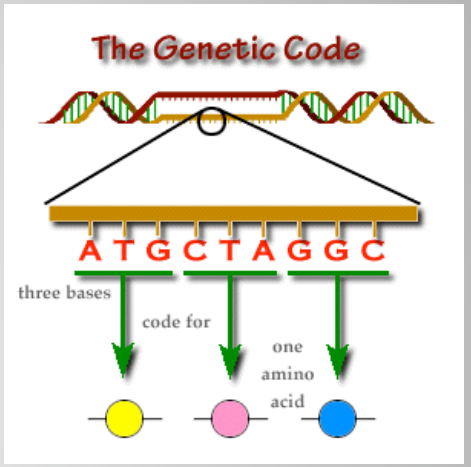
RNA: vehicle to transport genetic information into the cytoplasm where it is translated into proteins

↓

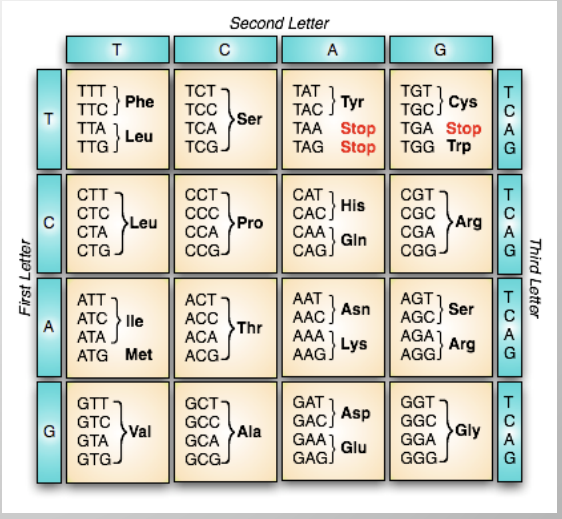
Protein: gene products made of amino acids with very specific functions, such as enzymes, hormones, motility proteins, transport, structural, etc.

Genetic code: 4^3 combinations





www.brooklyn.cuny.edu/.../GP.GeneticCode.GIF



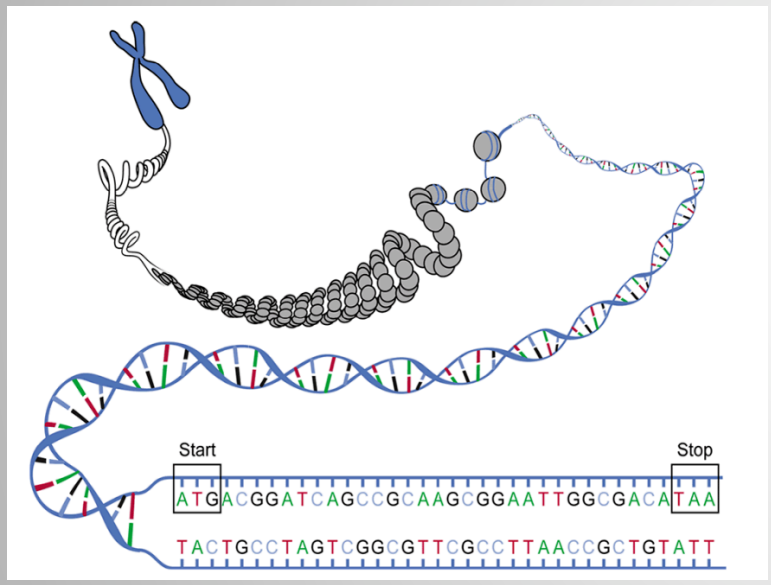
Start and Stop Codons in the Insulin Gene

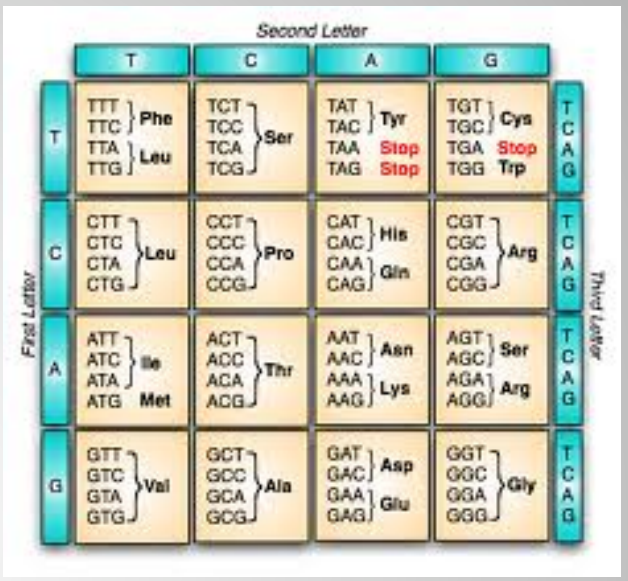
Start – ATG

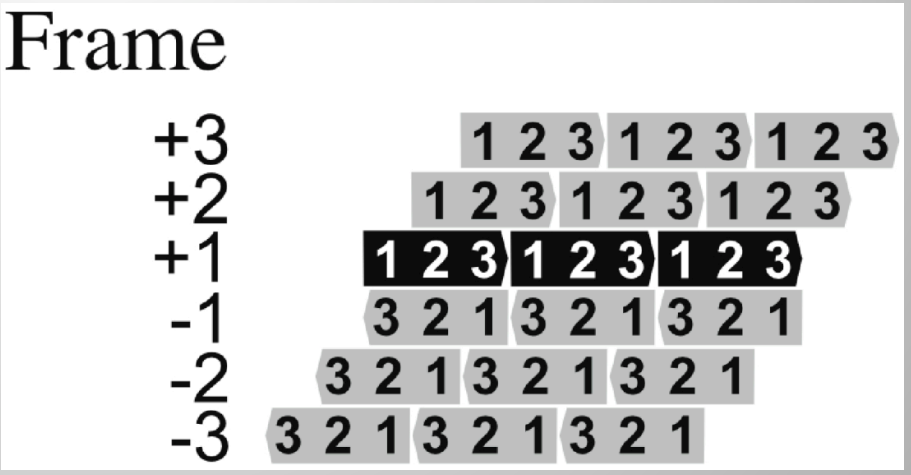
Stop – TAA, TAG, TGA

AGCCCTCCAGGACAGGCTGCATCAGAAGAGGCCATCAAGCAGA
TCACTGTCCTTCTGCCATGGCCCTGTGGATGCGCCTCCTGCCC
CTGCTGGCGCTGCTGGCCCTCTGGGGACCTGACCCAGCCGCA
GCCTTTGTGAACCAACACCTGTGCGGCTCACACCTGGTGGAAG
CTCTCTACCTAGTGTGCGGGGAACGAGGCTTCTTCTACACACCC
AAGACCCGCCGGGAGGCAGAGGACCTGCAGGTGGGGCAGGT
GGAGCTGGGCGGGGGCCCTGGTGCAGGCAGCCTGCAGCCCTT
GGCCCTGGAGGGGTCCCTGCAGAAGCGTGGCATTGTGGAACA
ATGCTGTACCAGCATCTGCTCCCTCTACCAGCTGGAGAACTACT
GCAACTAGACGCAGCCCGCAGGCAGCCCCACACCCGCCGCCT
CCTGCACCGAGAGAGATGGAATAAAGCCCTTGAACCAGCAAAA

The Genetic Code translates Nucleotide triplets into Amino Acids



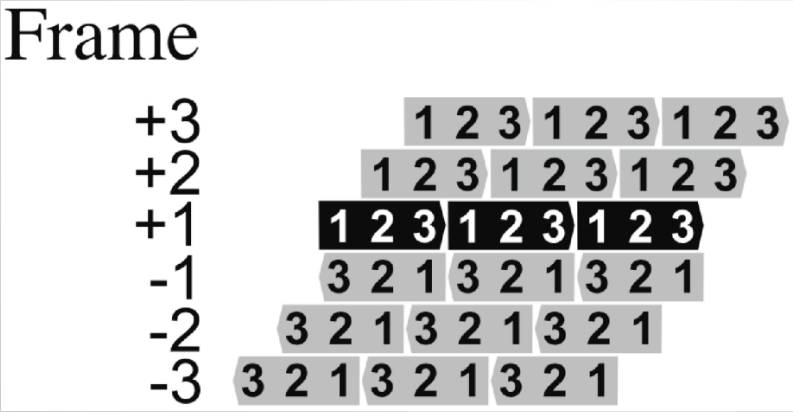




To Identify Genes in a Genome One Needs to Look for Open Reading Frames

A random sequence of nucleotides will by chance encode for a stop signal once every 20 codons (3 Stop codons out of 64).

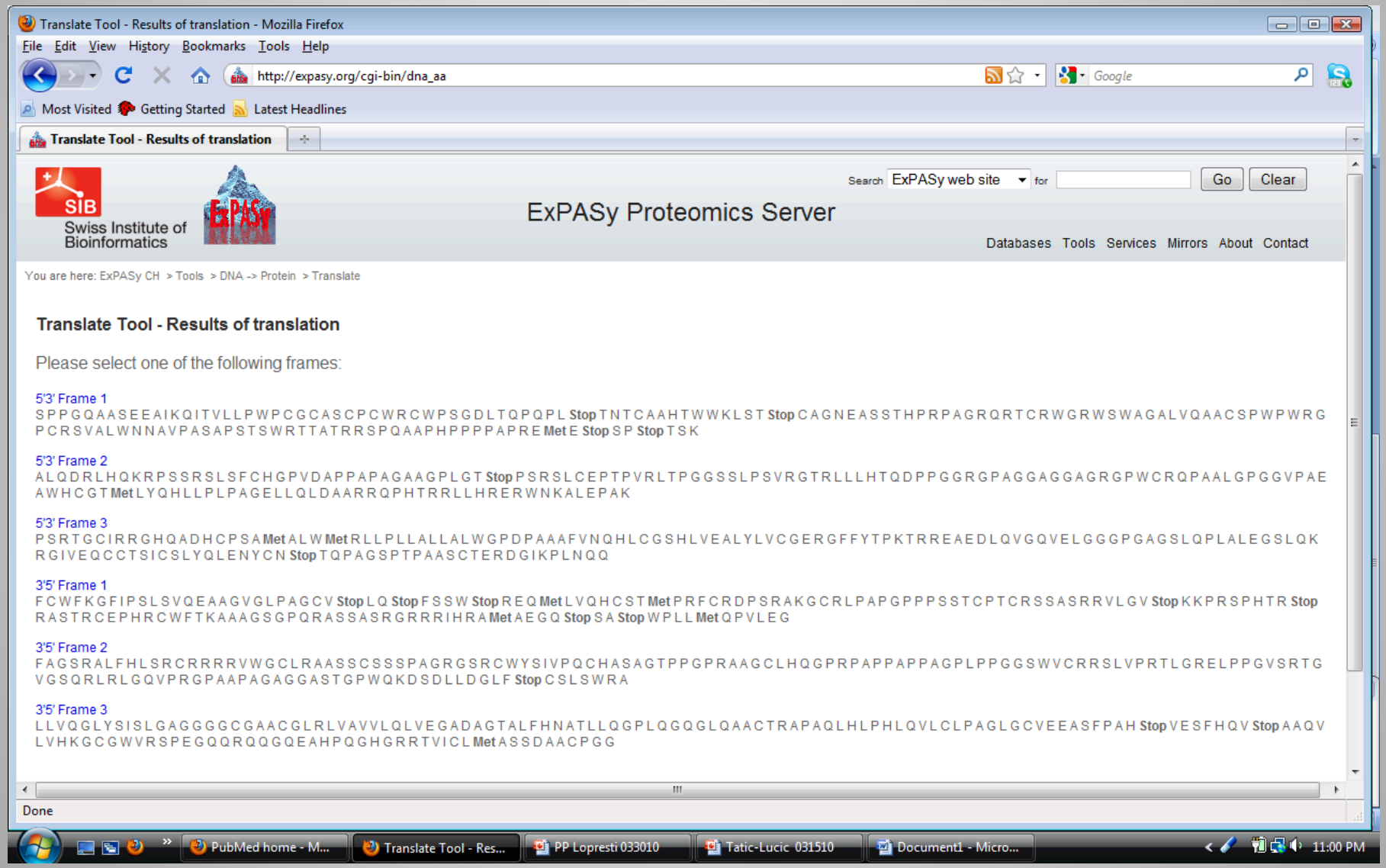
Open Reading Frames are continuous nucleotide stretches (>100 codons) that lack stop codons.



Criteria of a Gene

Open Reading Frame
Promoter Consensus Sequence
Ribosomal Binding Site
> 35 amino acids

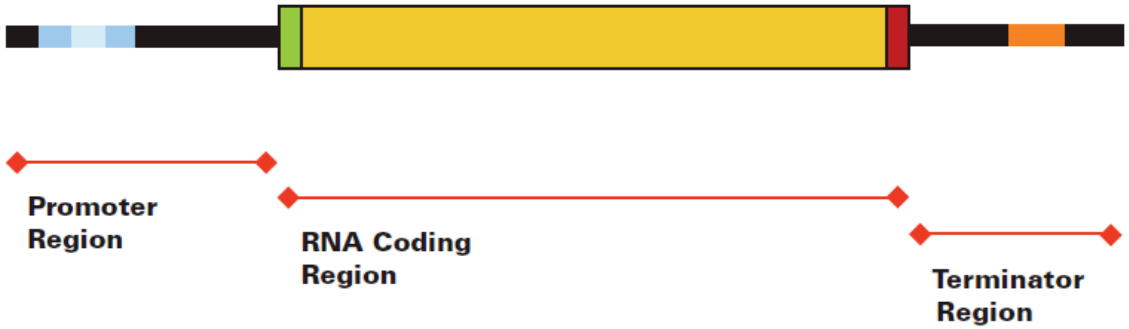
The ExPasy Proteomics Server translates proteins in all 6 Reading Frames



Basic Structure of a Gene

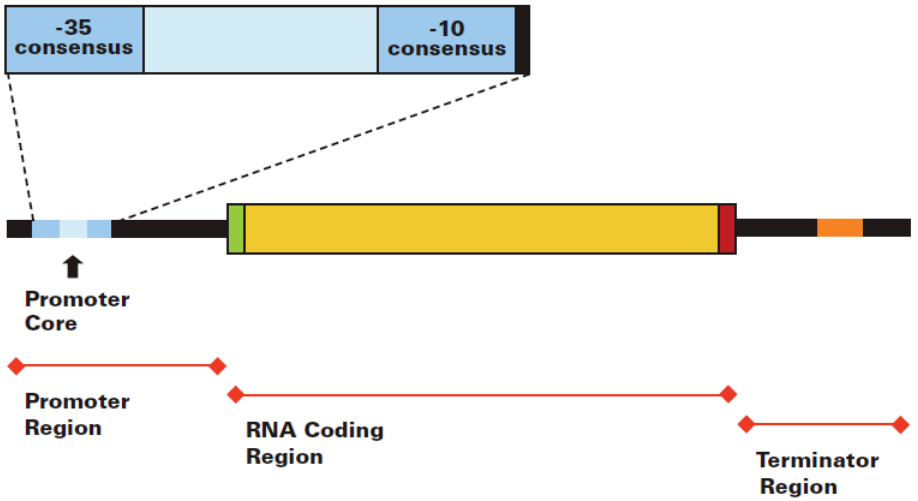
The basic structure of a gene includes the promoter region, the RNA coding region, and the terminator region, indicated in *Analyze* fig. 37.

Analyze fig. 37
Basic gene structure.



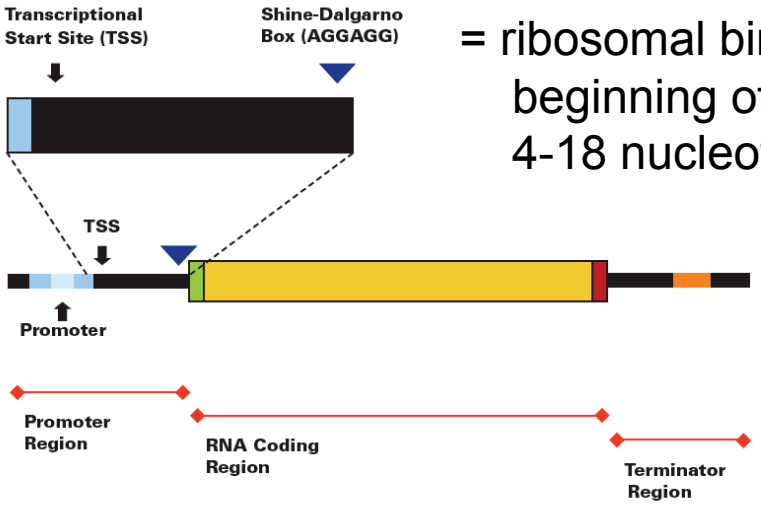
Promoter: RNA polymerase binding site, upstream of the ORF

-10 consensus sequence
- 35 consensus sequence
- the transcriptional start site
-Shine-Dalgarno sequence



Analyze fig. 38
The core of the promoter (in blue) contains the -10 and -35 consensus sequences.

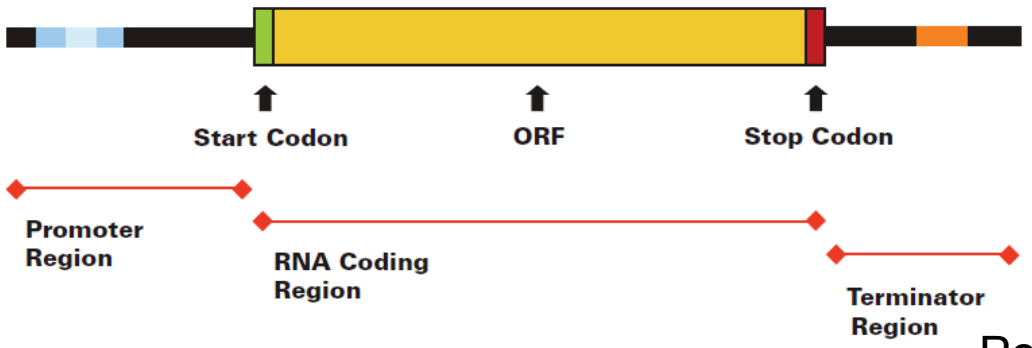
Analyze fig. 39
Functional sequence domains downstream of the core promoter include the TSS and Shine-Dalgarno sequence (or RBS).



The consensus Shine-Dalgarno sequence in mycobacterio-phage genomes is *AGGAGGA* (think of the song: "Hay, ba-by, Hay, ba-by, Hey"). In general,

= ribosomal binding site
beginning of translation
4-18 nucleotides upstream of start codon

The RNA coding region



Analyze fig. 40
The RNA coding region.
The ORF is flanked by the start codon (green) and the stop codon (red) in the RNA coding region.

Point where transcription ends

Mycobacteria have three possible start codons:
ATG – 45%
GTG – 45%
TTG – 10%

Human Insulin Gene Organization

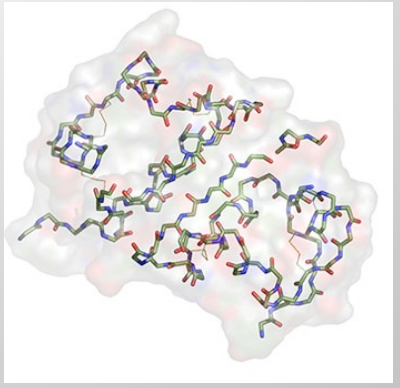
DNA/RNA sequence of insulin gene

AGCCCTCCAGGACAGGCTGCATCAGAAGAGGCCATCAAGCAGATCACTGTCCTTCTGCCATGGCCCTGTGGATGCGCCTCCTGCCCCTG
CTGGCGCTGCTGGCCCTCTGGGGACCTGACCCAGCCGCAGCCTTTGTGAACCAACACCTGTGCGGCTCACACCTGGTGGAAGCTCTCT
ACCTAGTGTGCGGGGAACGAGGCTTCTTCTACACACCCAAGACCCGCCGGGAGGCAGAGGACCTGCAGGTGGGGCAGGTGGAGCTGG
GCGGGGGCCCTGGTGCAGGCAGCCTGCAGCCCTTGGCCCTGGAGGGGTCCCTGCAGAAGCGTGGCATTGTGGAACAATGCTGTACCAG
CATCTGCTCCCTCTACCAGCTGGAGAACTACTGCAACTAGACGCAGCCCGCAGGCAGCCCCACACCCGCCGCCTCCTGCACCGAGAGAG
ATGGAATAAAGCCCTTGAACCAGCAAAA

Linear Translated Open Reading Frame

MetAlaLeuTrpMetArgLeuLeuProLeuLeuAlaLeuLeuAlaLeuTrpGlyProAspProAlaAlaAlaPheValAsnGlnHisLeuCysGlySer
HisLeuValGluAlaLeuTyrLeuValCysGlyGluArgGlyPhePheTyrThrProLysThrArgArgGluAlaGluAspLeuGlnValGlyGlnValGl
uLeuGlyGlyGlyProGlyAlaGlySerLeuGlnProLeuAlaLeuGluGlySerLeuGlnLysArgGlyIleValGluGlnCysCysThrSerIleCysSe
rLeuTyrGlnLeuGluAsnTyrCysAsnEnd

However, no information about
3D structure, function…….



Four Major Methods are used to define 'Genes'

1. Search for 'Open Reading Frames' (ORFs)

2. Presence of a consensus sequence for ribosome binding in the immediate vicinity of the start codon

3. Expressed Sequence Tags (EST) analysis

4. Homology of the putative gene to other known genes using BLAST

Genome Comparisons using 'BLAST'

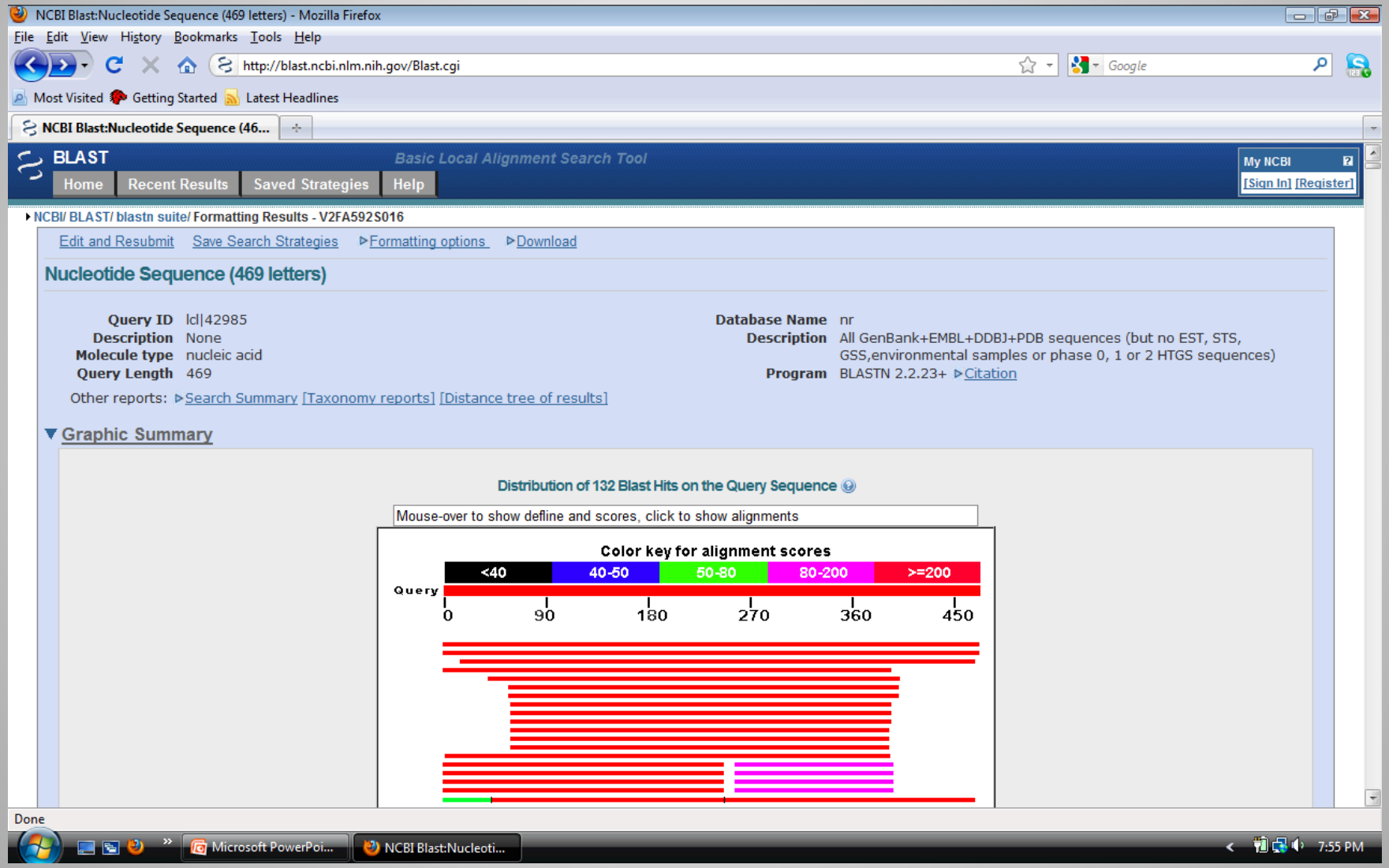
Basic Local Alignment Search Tool

An algorithm for comparing primary biological sequence information, such as the nucleotides of DNA sequences or the amino acid sequences of different proteins

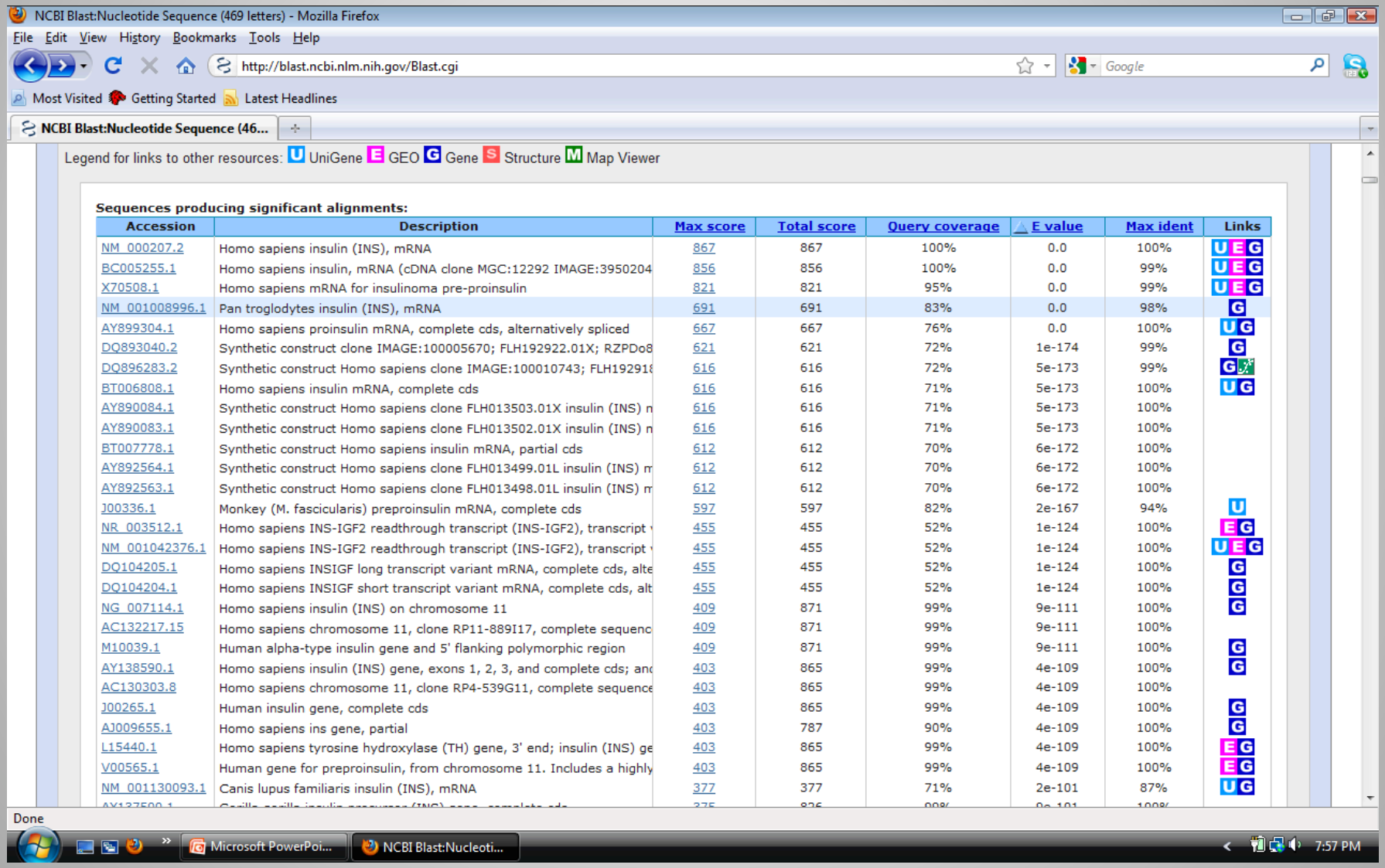
The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches.

http://blast.ncbi.nlm.nih.gov/Blast.cgi

Blast search: Human Insulin

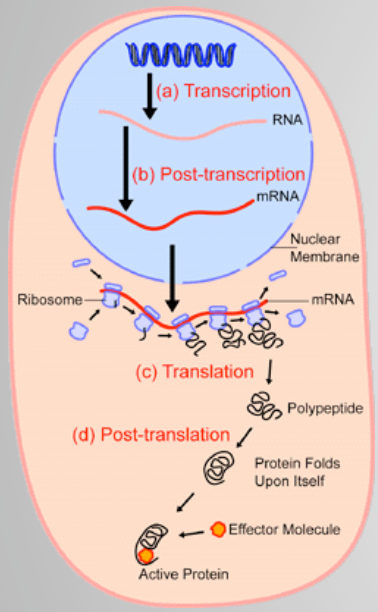


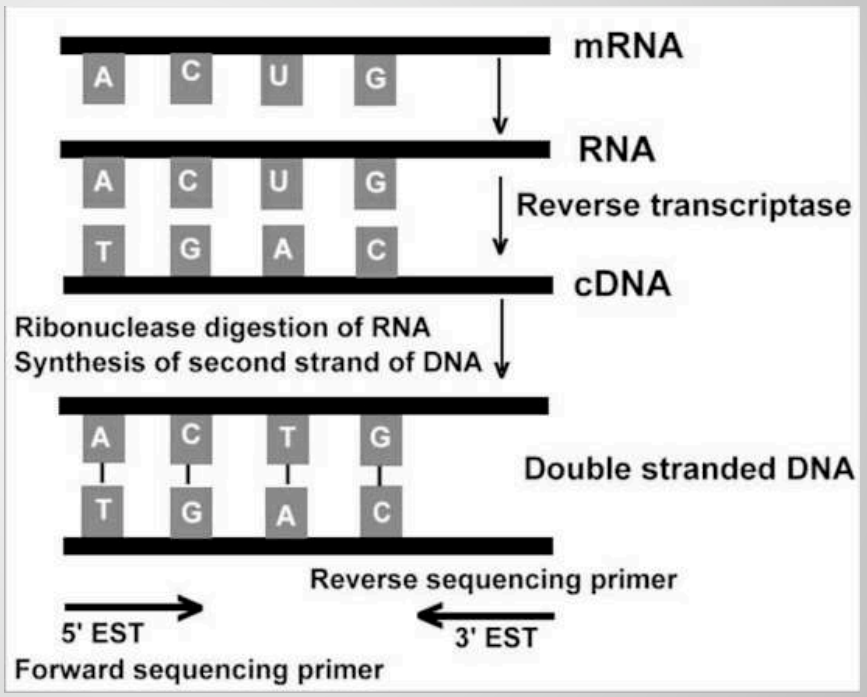
Blast Search: Human Insulin Continued



Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
NM_000207.2	Homo sapiens insulin (INS), mRNA	867	867	100%	0.0	100%	U E G
BC005255.1	Homo sapiens insulin, mRNA (cDNA clone MGC:12292 IMAGE:3950204	856	856	100%	0.0	99%	U E G
X70508.1	Homo sapiens mRNA for insulinoma pre-proinsulin	821	821	95%	0.0	99%	U E G
NM_001008996.1	Pan troglodytes insulin (INS), mRNA	691	691	83%	0.0	98%	G
AY899304.1	Homo sapiens proinsulin mRNA, complete cds, alternatively spliced	667	667	76%	0.0	100%	U G
DQ893040.2	Synthetic construct clone IMAGE:100005670; FLH192922.01X; RZPDo8	621	621	72%	1e-174	99%	G
DQ896283.2	Synthetic construct Homo sapiens clone IMAGE:100010743; FLH192918	616	616	72%	5e-173	99%	G
BT006808.1	Homo sapiens insulin mRNA, complete cds	616	616	71%	5e-173	100%	U G
AY890084.1	Synthetic construct Homo sapiens clone FLH013503.01X insulin (INS) n	616	616	71%	5e-173	100%	
AY890083.1	Synthetic construct Homo sapiens clone FLH013502.01X insulin (INS) n	616	616	71%	5e-173	100%	
BT007778.1	Synthetic construct Homo sapiens insulin mRNA, partial cds	612	612	70%	6e-172	100%	
AY892564.1	Synthetic construct Homo sapiens clone FLH013499.01L insulin (INS) m	612	612	70%	6e-172	100%	
AY892563.1	Synthetic construct Homo sapiens clone FLH013498.01L insulin (INS) m	612	612	70%	6e-172	100%	
J00336.1	Monkey (M. fascicularis) preproinsulin mRNA, complete cds	597	597	82%	2e-167	94%	U
NR_003512.1	Homo sapiens INS-IGF2 readthrough transcript (INS-IGF2), transcript	455	455	52%	1e-124	100%	E G
NM_001042376.1	Homo sapiens INS-IGF2 readthrough transcript (INS-IGF2), transcript	455	455	52%	1e-124	100%	U E G
DQ104205.1	Homo sapiens INSIGF long transcript variant mRNA, complete cds, alte	455	455	52%	1e-124	100%	G
DQ104204.1	Homo sapiens INSIGF short transcript variant mRNA, complete cds, alt	455	455	52%	1e-124	100%	G
NG_007114.1	Homo sapiens insulin (INS) on chromosome 11	409	871	99%	9e-111	100%	G
AC132217.15	Homo sapiens chromosome 11, clone RP11-889I17, complete sequenc	409	871	99%	9e-111	100%	
M10039.1	Human alpha-type insulin gene and 5' flanking polymorphic region	409	871	99%	9e-111	100%	G
AY138590.1	Homo sapiens insulin (INS) gene, exons 1, 2, 3, and complete cds; and	403	865	99%	4e-109	100%	G
AC130303.8	Homo sapiens chromosome 11, clone RP4-539G11, complete sequence	403	865	99%	4e-109	100%	
J00265.1	Human insulin gene, complete cds	403	865	99%	4e-109	100%	G
AJ009655.1	Homo sapiens ins gene, partial	403	787	90%	4e-109	100%	G
L15440.1	Homo sapiens tyrosine hydroxylase (TH) gene, 3' end; insulin (INS) ge	403	865	99%	4e-109	100%	E G
V00565.1	Human gene for preproinsulin, from chromosome 11. Includes a highly	403	865	99%	4e-109	100%	E G
NM_001130093.1	Canis lupus familiaris insulin (INS), mRNA	377	377	71%	2e-101	87%	U G

ESTs (Expressed Sequence Tags) are Used to Determine Which Protein Coding Genes are Expressed in a Particular Cell Type or Tissue





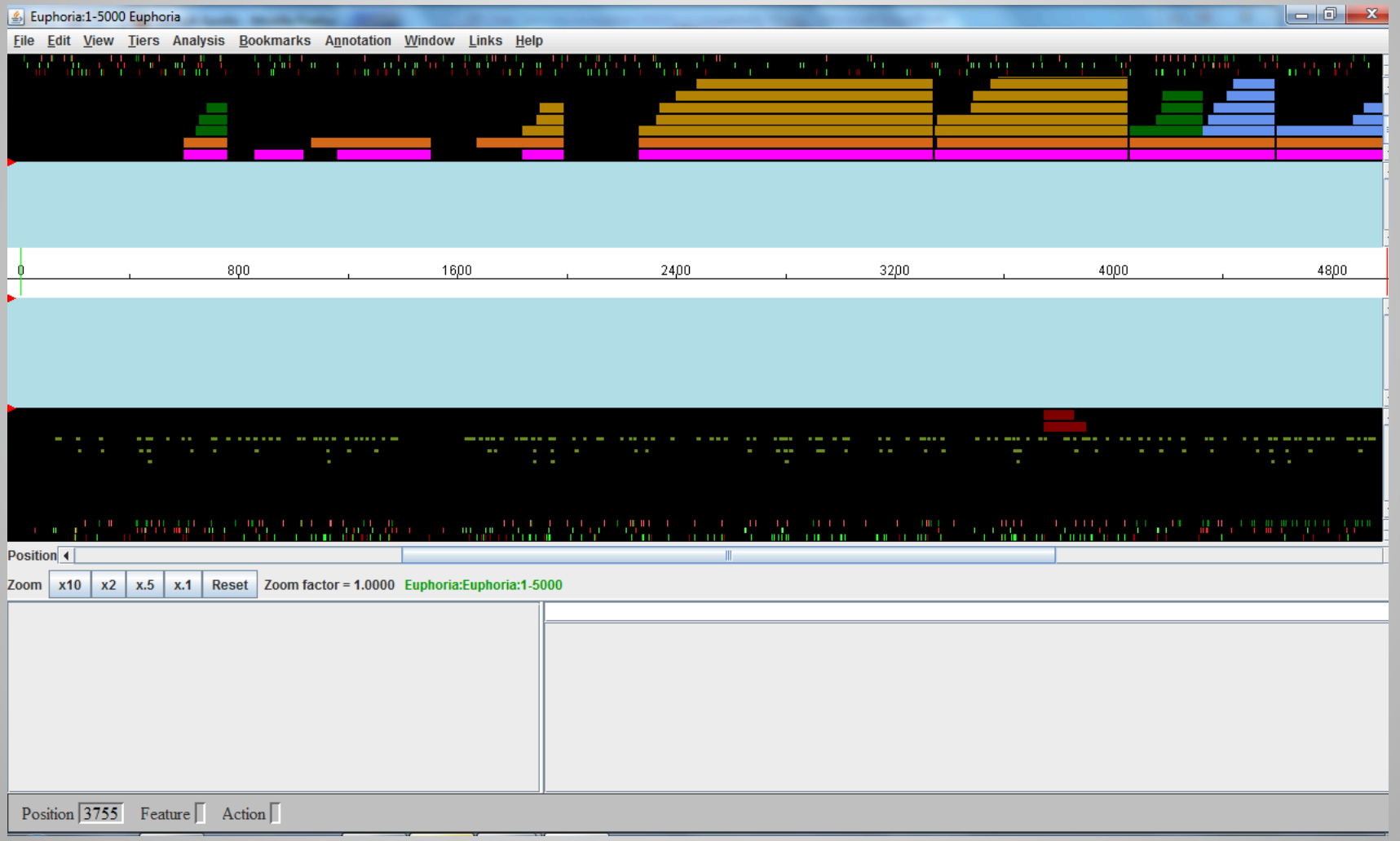
Approximately 200 different cell types
Only a subset of genes is being expressed on cellular level under certain conditions
Map expressed genes on cellular level to whole genome

Gene Calling Softwares

GLIMMER: uses Interpolated Markov models (IMMs)

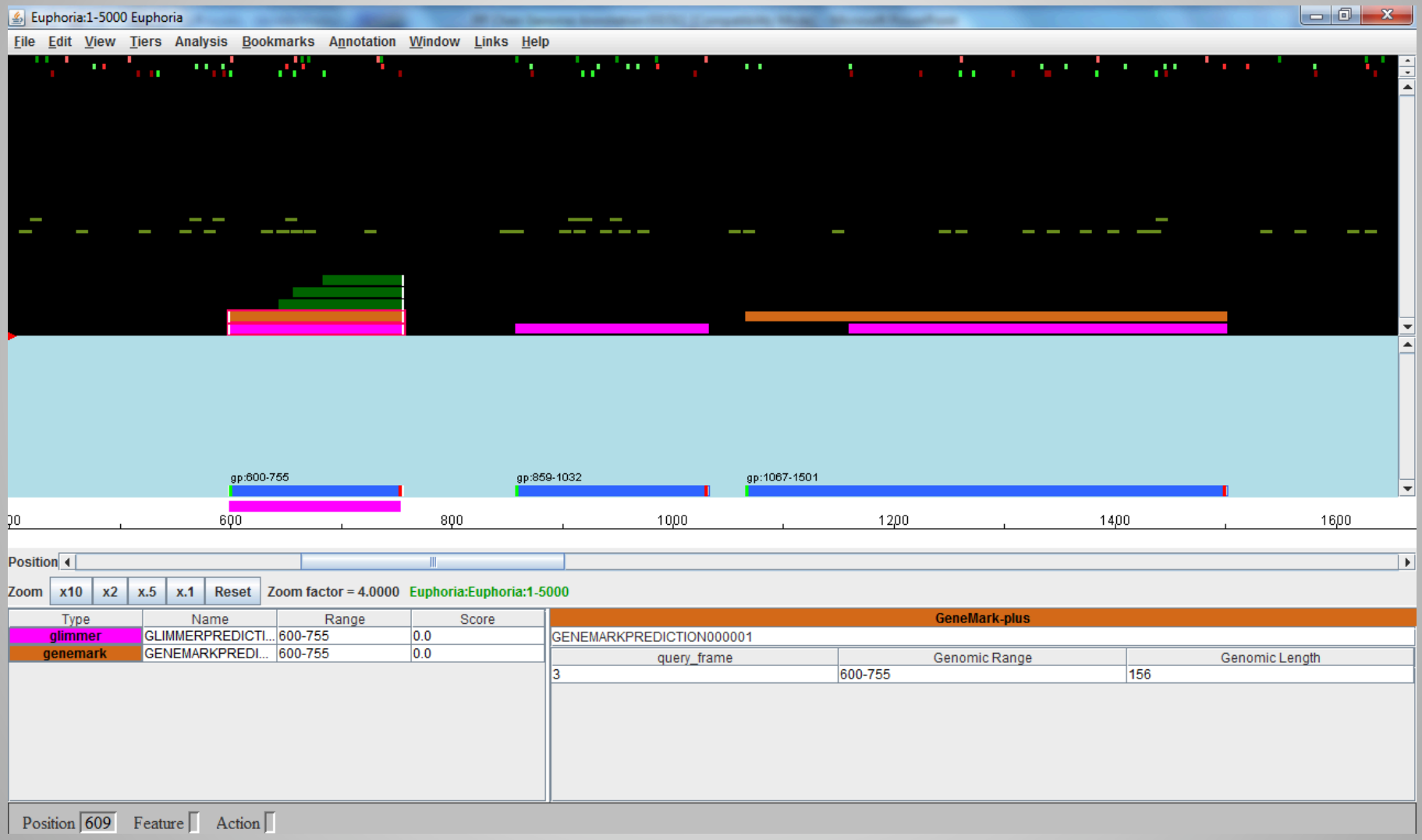
GeneMark: determine the protein-coding potential of a DNA sequence (within a sliding window) by using species specific parameters of the Markov models of coding and non-coding regions

Genome Annotation of Bacteriophage EUPHORIA using GeneMark and GLIMMER algorithms displayed in ‘Apollo’

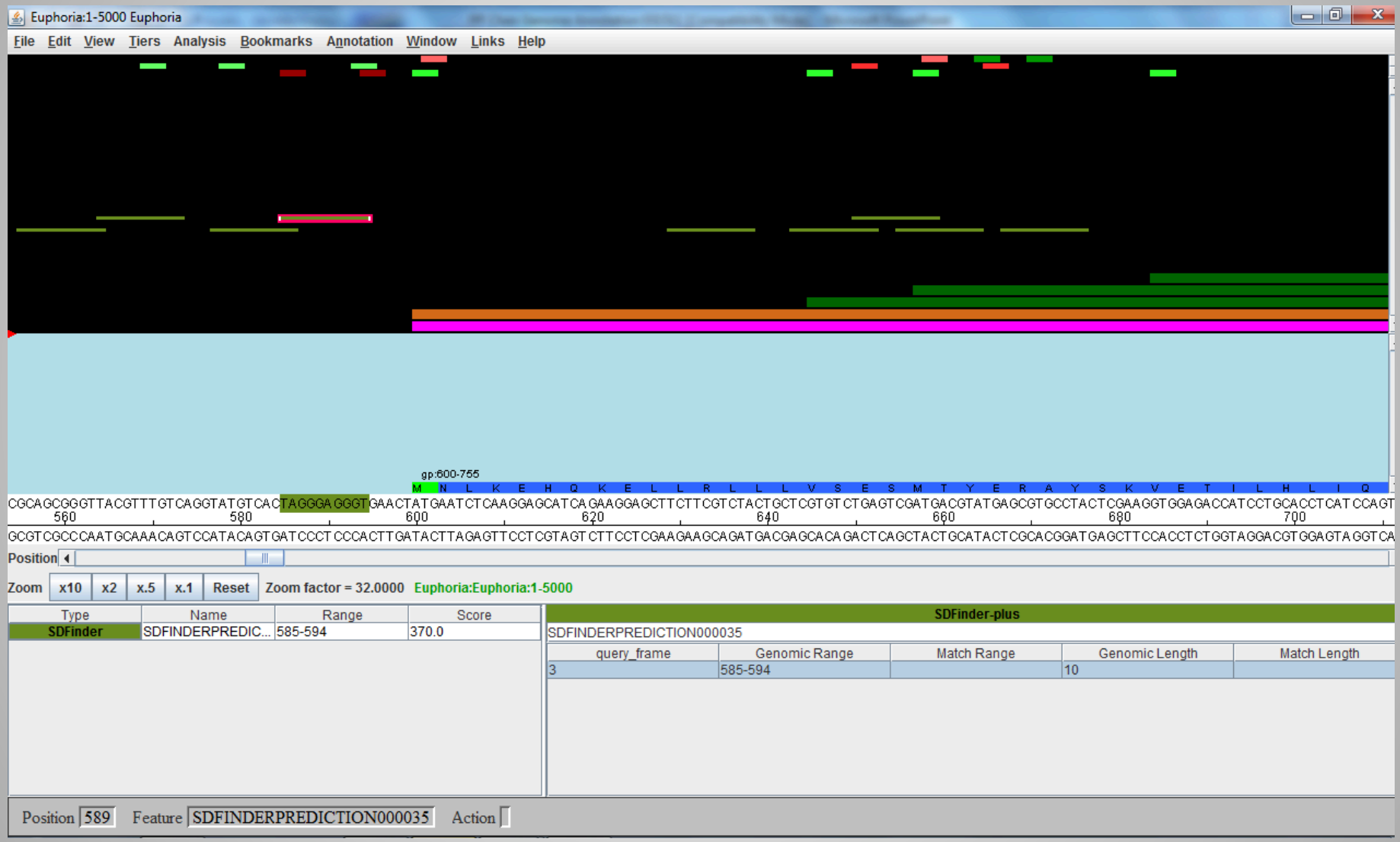


purple: GLIMMER dark brown: GeneMark light brown: GeneMark-TB

First three Gene Calls of EUPHORIA genome using GLIMMER and GeneMark Gene Prediction



Apollo: Zoom into nucleotide resolution



Bright green: Start Codon Red: Stop Codon Light green: ribosomal binding site

Summary

Genome sequencing is followed by gene identification and genome annotation

Genome annotation is the critical process by which functions are assigned to predicted proteins.

Gene annotation performed computationally should alwas be viewed As generating a hypothesis that needs to be experimentally tested.