Genome Sequence Annotation

Pablo A. Moreno C.
Ingeniero Civil en Biotecnología, UCH.
Ingeniero de Proyectos

Andrés O. Aravena D.
Ingeniero Civil Matemático, UCH.
Ingeniero Jefe LBMG

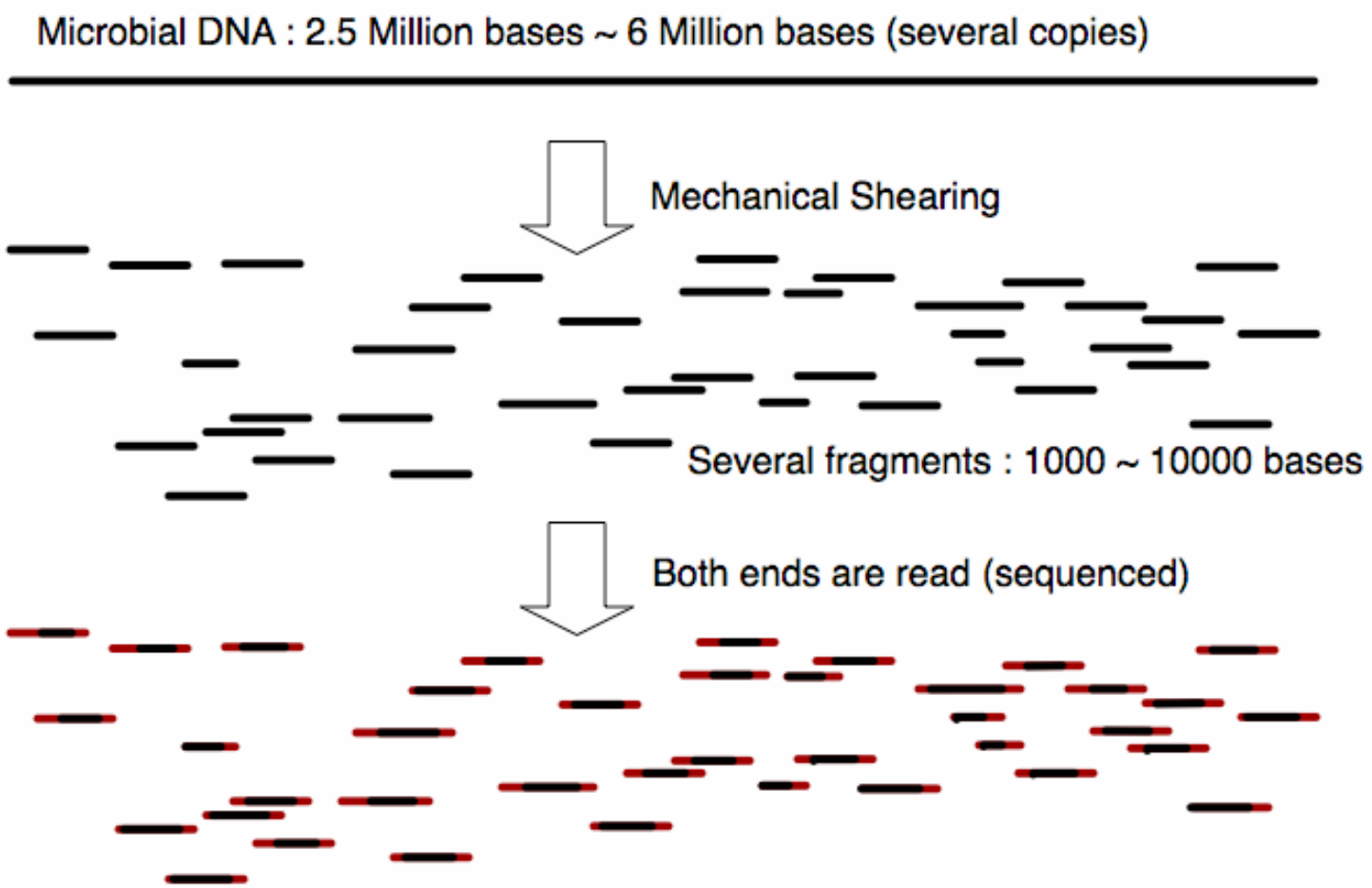
Juan A. Ugalde C.
Ingeniero en Biotecnología Molecular, UCH.
Ingeniero de Proyectos

Senior Advisors:
Mauricio Gonzales, Ph.D
Biology, INTA-UCH
Alejandro Maass, Ph.D
Mathematics, DIM-UCH

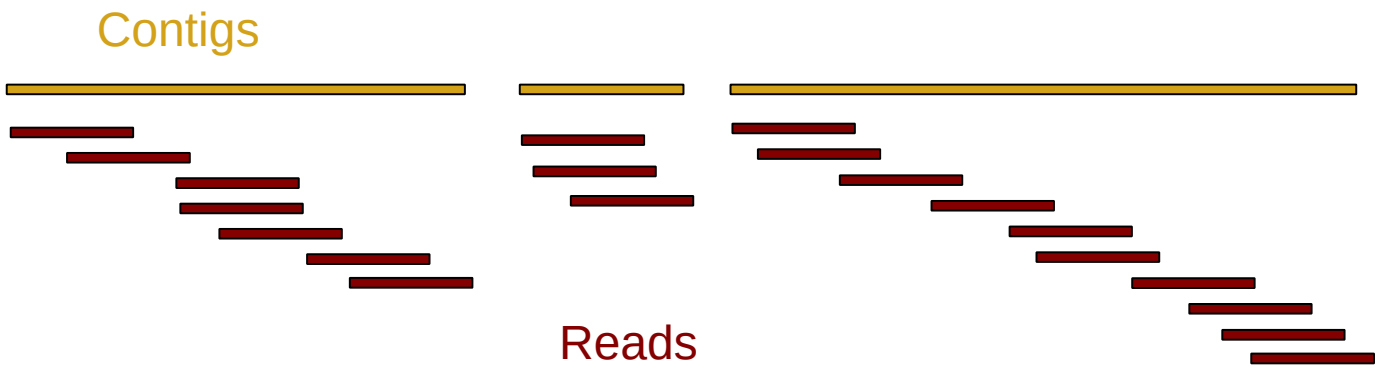
The input: genome assembly

- Sanger Method.
- Technology limitation on lengths.
- Walking primers.
- Shot-gun sequencing.

Shot-gun sequencing

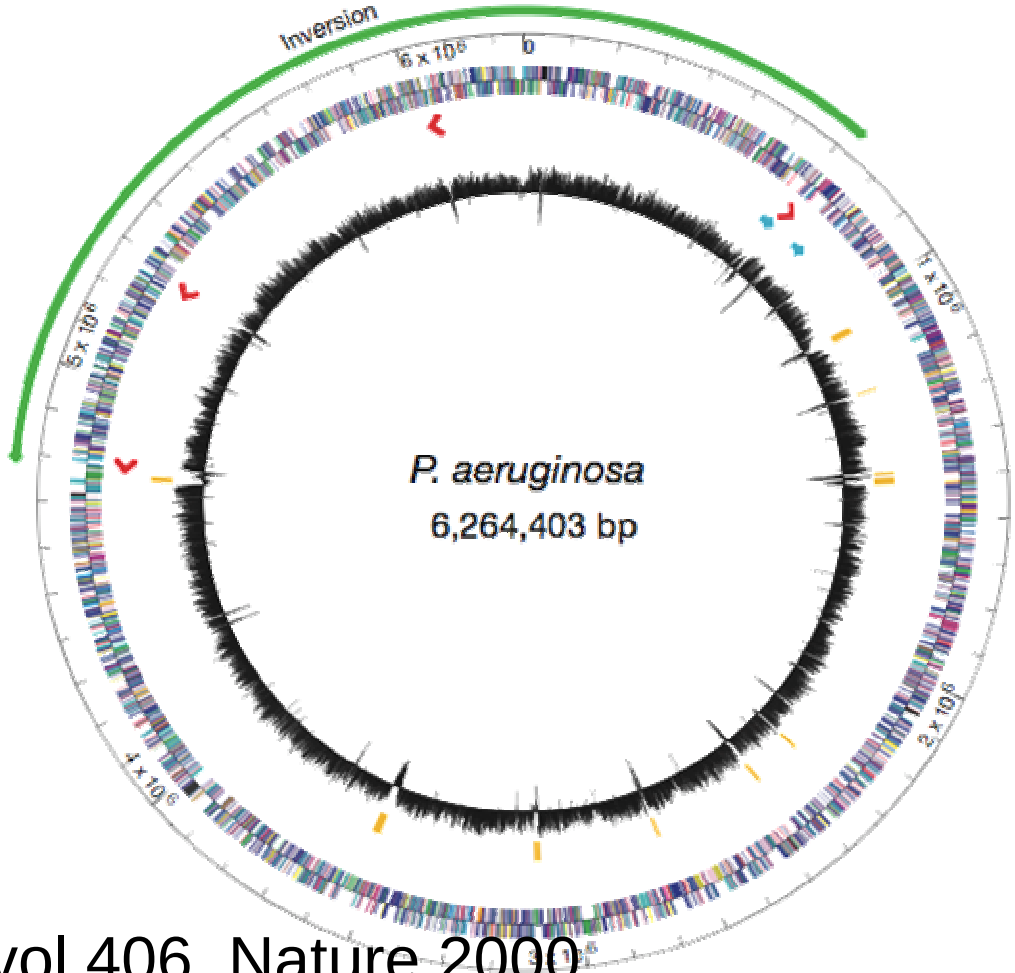


Assembly

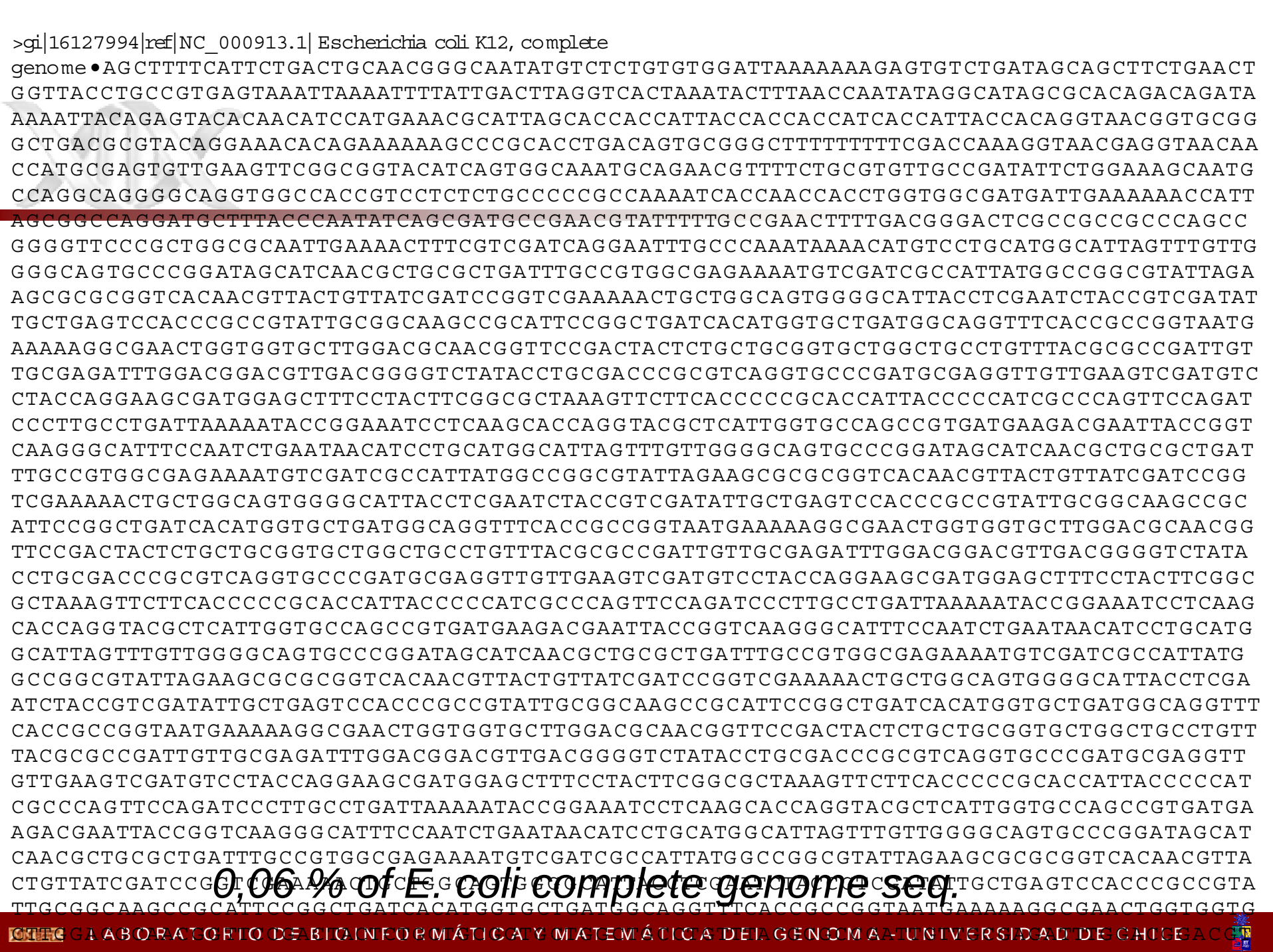


ACTTAGCGC...CTAGCTATCTATCTACTA...GTCA

Circular Genome



Stover et al, vol 406, Nature 2000

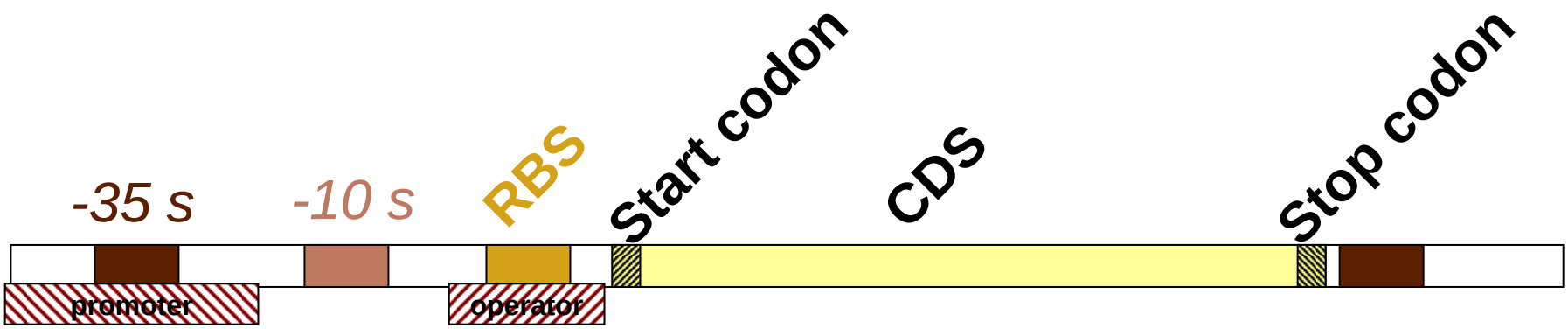


0,06 % of E. coli complete genome seq.

Sequence Features

- CDSs and accompanying signals.
- RNA Products
  - rRNA (16S, 23S).
  - tRNAs
  - ssRNAs
- Regulatory Binding Sites
- Insertion Sites
- Other Signals (especially in Eukaryonts)

Coding Sequences & Co.

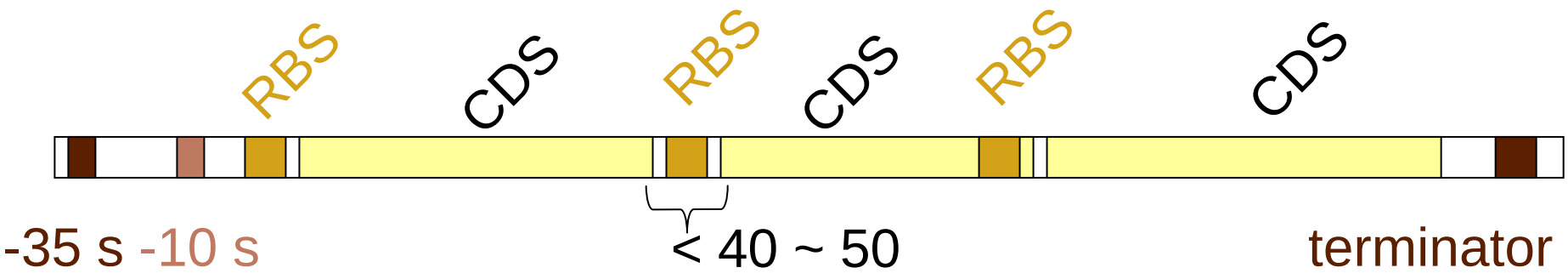


Transcription

Translation

"A CDS is a Coding ORF,
Not all ORFs are CDSs."

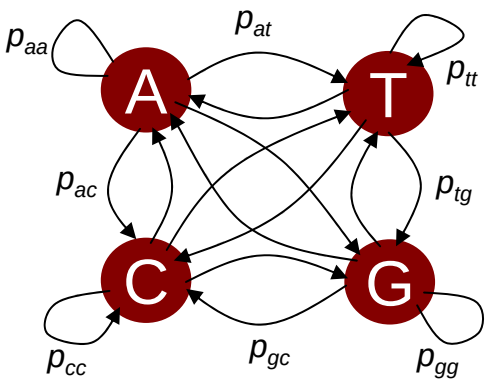
Related CDSs: Operons



Polycistronic transcripts; Operons.

Main approaches

- Automatic Marking:
  - Intrinsic: sequence statistics.
  - Extrinsic: alignment-based.
- Manual revision.



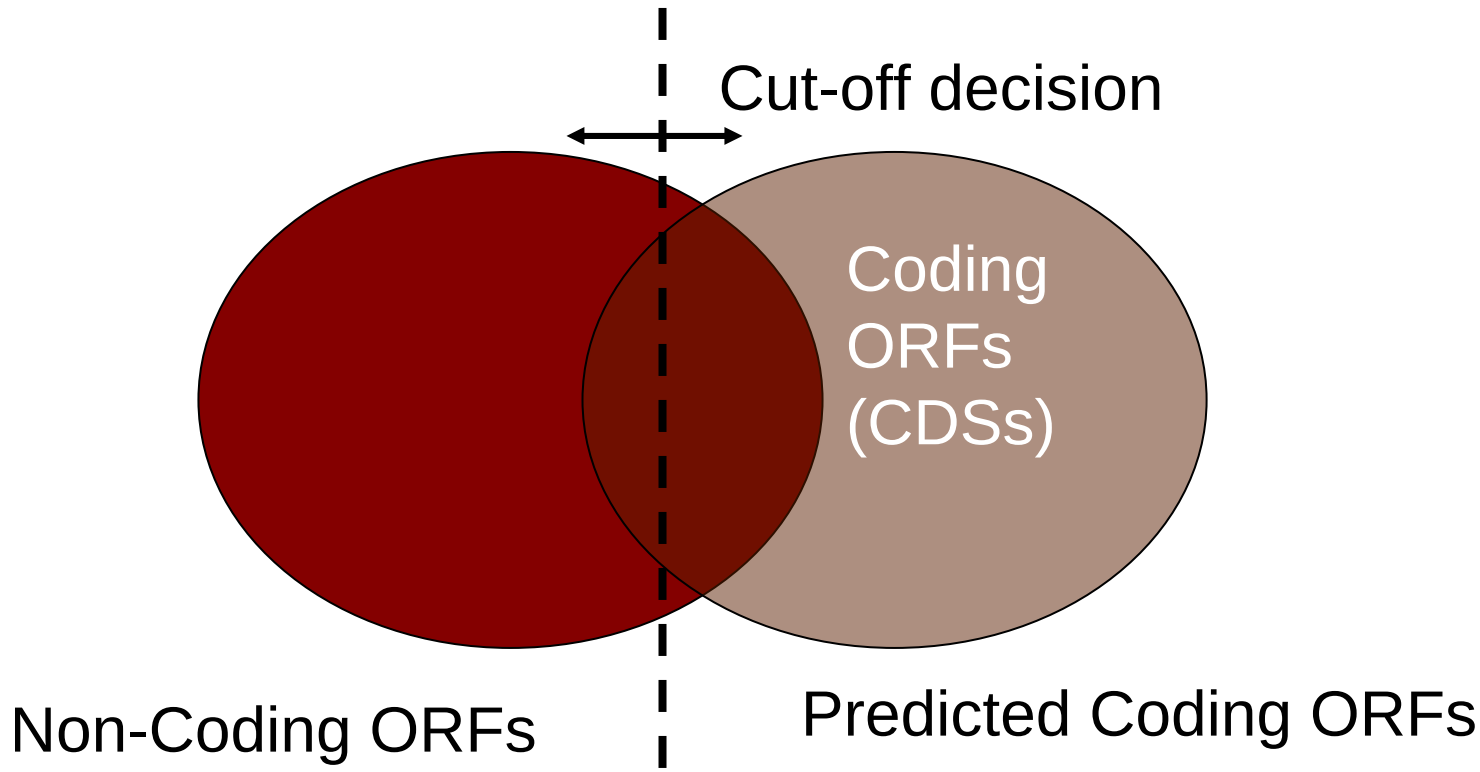
$P(\text{ATCGA}) = \pi_a p_{at} p_{tc} p_{cg} p_{ga}$

```
 Score =  907 bits (2344), Expect = 0.0
 Identities = 462/478 (96%), Positives = 462/478 (96%)

Query: 99  VPGLGPKRVKALWHELDVETVDQLTRAAREGRIRSIPGFGEKTEARILEALQAQIAAVPR 158
       VPGLGPKRVKALWHELDVETVDQLTRAAREGRIRSIPGFGEKTEARILEALQAQIAAVPR
Sbjct: 1   VPGLGPKRVKALWHELDVETVDQLTRAAREGRIRSIPGFGEKTEARILEALQAQIAAVPR 60

Query: 159 FPIXXXXXXXXXLVRYLQNVPGVRRVVVAGSFRRGRDTVGDLD.....................
       FPI     LVRYLQNVPGVRRVVVAGSFRRGRDTVGDLD.................
Sbjct: 61  FPIAVAAPYAAALVRYLQNVPGVRRVVVAGSFRRGRDTVGDLD.....................


.....................LIGSREPMDVDMPRIIRHARERGCFLEIDAQPERLDLVDIHARTA 518
       .............LIGSREPMDVDMPRIIRHARERGCFLEIDAQPERLDLVDIHARTA
.....................LIGSREPMDVDMPRIIRHARERGCFLEIDAQPERLDLVDIHARTA 420

Query: 519 KEEGVLLAVNSDAHSIHDFDNLRFGLQQAQRGWLEMKDVLNTRTLEELRPLLAATMSR 576
       KEEGVLLAVNSDAHSIHDFDNLRFGLQQAQRGWLEMKDVLNTRTLEELRPLLAATMSR
Sbjct: 421 KEEGVLLAVNSDAHSIHDFDNLRFGLQQAQRGWLEMKDVLNTRTLEELRPLLAATMSR 478
```

Extrinsic Methods (CDSs)

- Search against gene or protein DBs.
  - Sequence Alignments (Blasts, etc)
  - Position weighted matrices (Pfam, etc)

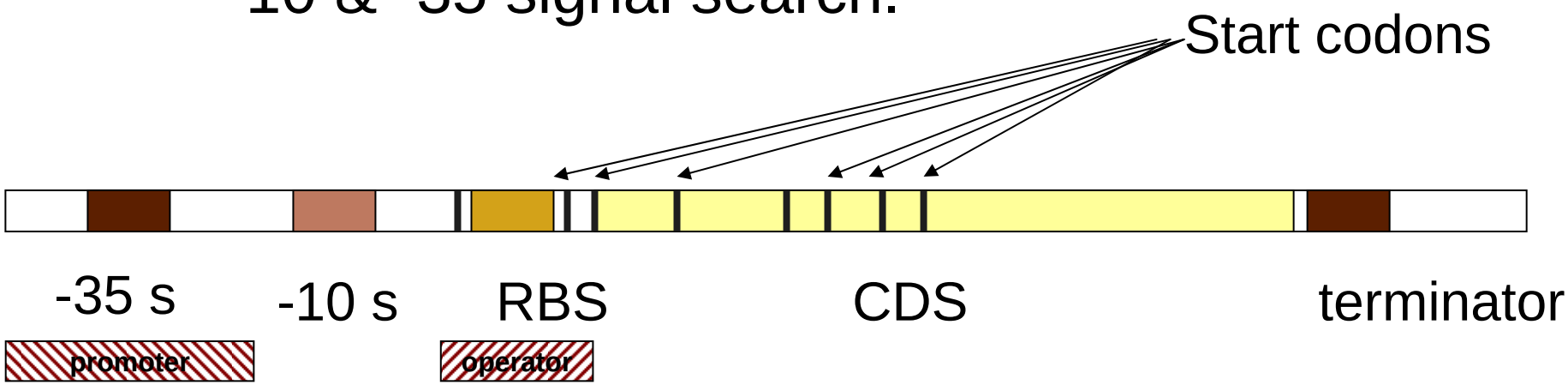
Class overlap



Glimmer: Up to 18% false pos.

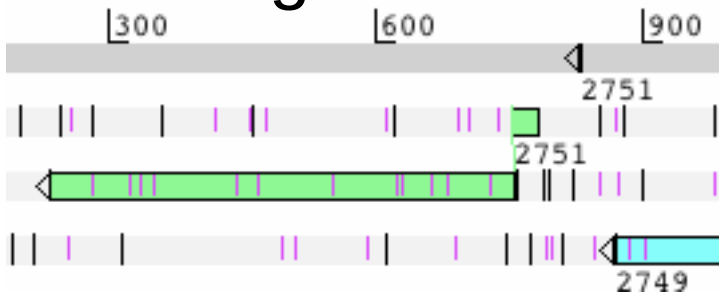
Start codon issue

- Difficult problem.
- Integrate up-stream signal search.
  - RBS.
  - -10 & -35 signal search.



Frame shift issue

- One gene, more than one reading frame.
- Manual Joining.
- Sequence correction needed.
  - Compare against sequence quality.
  - Re-sequence some short tags.



CDS Marking Software

- Glimmer (1997 - 2004)
  - Variable Order Markov models.
- Critica (1999 - 2000)
  - Codon statistics of known coding sequences.
- ProKoV (2002)
  - Markov models/Bayes

CDS Marking Software

- ZCurve (2003)
  - Sequence transformation.
- AMIGene (2003)
  - Prokov, heuristic to select CDSs from ORFs.
- Priam (2003)
  - Position-weighted matrix from known E.C. classes.
- FrameD (2003)
  - DAG, designed for frame shifts and overlaps.

Other Seq. Features Markers

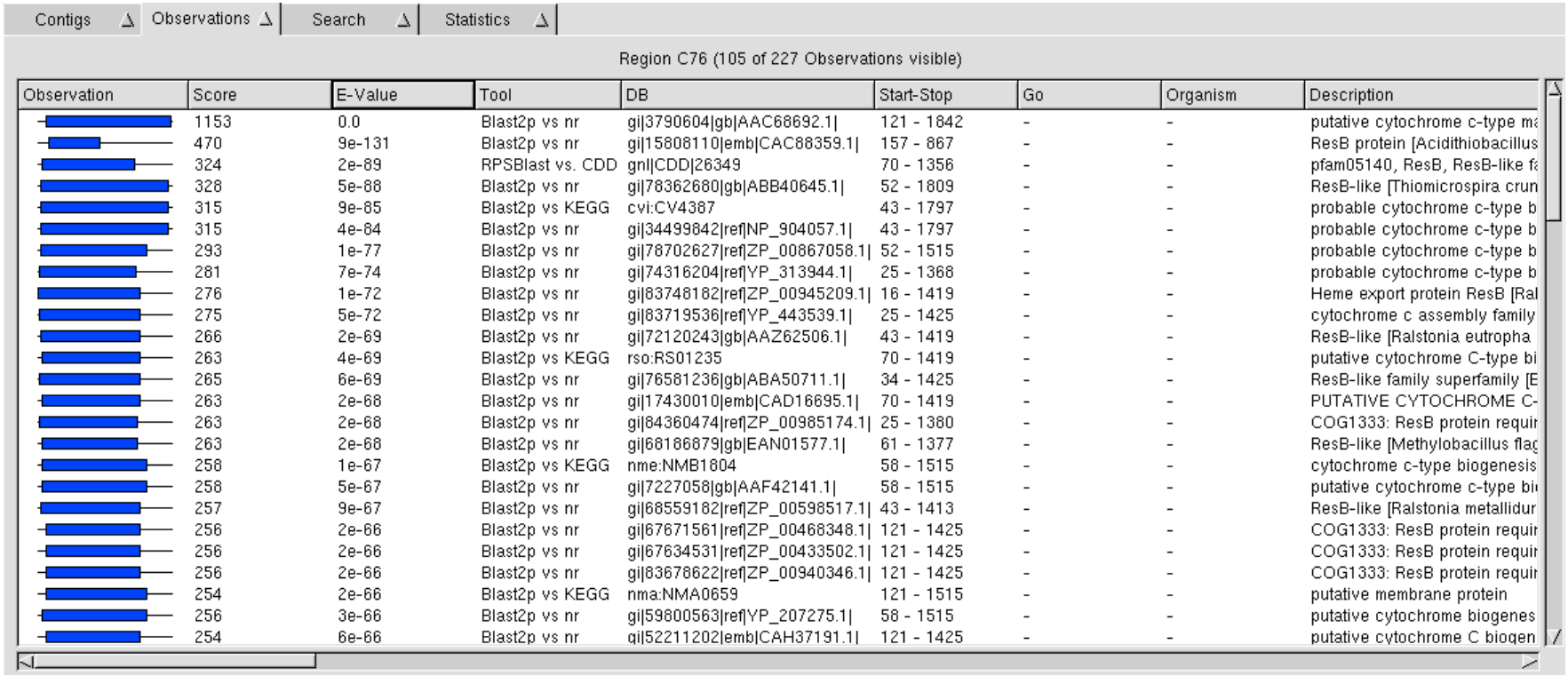
- TRNAscan SE (2000)
- RBSfinder (2001)
  - Markov Model for consensus RBSs.
- Search for RNAs(1999)
- GSFinder (2004)
  - Z-Curve method.
- TermIt (1990)
  - Single cistron or polycistron statistics.
- TransTerm (2000)
  - Whole genome statistics.


>gi|16127994|ref|NC_000913.1| Escherichia coli K12, complete genome
AGCTTTTCATTCTGACTGCAACGGGCAATATGTCTCTGTGTGGATTAAAAAAAGAGTGTCTGATAGCAGCTTCTGAACT
GGTTACCTGCCGTGAGTAAATTAAAATTTTATTGACTTAGGTCACTAAATACTTTAACCAATATAGGCATAGCGCACAGACAGAT
AAAAATTACAGAGTACACAACATCCATGAAACGCATTAGCACCACCATTACCACCACCATCACCATTACCACAGGTAACGGT
GCGGGCTGACGCGTACAGGAAACACAGAAAAAAGCCCGCACCTGACAGTGCGGGCTTTTTTTTTCGACCAAAGGTAACGAG
GTAACAACCATGCGAGTGTTGAAGTTCGGCGGTACATCAGTGGCAAATGCAGAACGTTTTCTGCGTGTTGCCGATATTCTGGA
AAGCAATGCCAGGCAGGGGCAGGTGGCCACCGTCCTCTCTGCCCCCGCCAAAATCACCAACCACCTGGTGGCGATGATTGA
AAAAACCATTAGCGGCCAGGATGCTTTACCCAATATCAGCGATGCCGAACGTATTTTTGCCGAACTTTGACGGGACTCGCCG
CCGCCCAGCCGGGGTTCCCGCTGGCGCAATTGAAAACTTTCGTCGATCAGGAATTTGCCCAAATAAAACATGTCCTGCATGG
CATTAGTTTGTTGGGGCAGTGCCCGGATAGCATCAACGCTGCGCTGATTTGCCGTGGCGAGAAAATGTCGATCGCCATTATG
GCCGGCGTATTAGAAGCGCGCGGTCACAACGTTACTGTTATCGATCCGGTCGAAAAACTGCTGGCAGTGGGGCATTACCTCG
AATCTACCGTCGATATTGCTGAGTCCACCCGCCGTATTGCGGCAAGCCGCATTCCGGCTGATCACATGGTGCTGATGGCAGGT
TTCACCGCCGGTAATGAAAAAGGCGAACTGGTGGTGCTTGGACGCAACGGTTCCGACTACTCTGCTGCGGTGCTGGCTGCCTG
TTTACGCGCCGATTGTTGCGAGATTTGGACGGACGTTGACGGGGTCTATACCTGCGACCCGCGTCAGGTGCCCGATGCGAGG
TTGTTGAAGTCGATGTCCTACCAGGAAGCGATGGAGCTTTCCTACTTCGGCGCTAAAGTTCTTCACCCCCGCACCATTACCCC
CATCGCCCAGTTCCAGATCCCTTGCCTGATTAAAAATACCGGAAATCCTCAAGCACCAGGTACGCTCATTGGTGCCAGCCGT
GATGAAGACGAATTACCGGTCAAGGGCATTTCCAATCTGAATAACATCCTGCATGGCATTAGTTTGTTGGGGCAGTGCCCGG
ATAGCATCAACGCTGCGCTGATTTGCCGTGGCGAGAAAATGTCGATCGCCATTATGGCCGGCGTATTAGAAGCGCGCGGTCAC
AACGTTACTGTTATCGATCCGGTCGAAAAACTGCTGGCAGTGGGGCATTACCTCGAATCTACCGTCGATATTGCTGAGTCCACC
CGCCGTATTGCGGCAAGCCGCATTCCGGCTGATCACATGGTGCTGATGGCAGGTTTCACCGCCGGTAATGAAAAAGGCGAAC
TGGTGGTGCTTGGACGCAACGGTTCCGACTACTCTGCTGCGGTGCTGGCTGCCTGTTTACGCGCCGATTGTTGCGAGATTTGG
ACGGACGTTGACGGGGTCTATACCTGCGACCCGCGTCAGGTGCCCGATGCGAGGTTGTTGAAGTCGATGTCCTACCAGGAAG
CGATGGAGCTTTCCTACTTCGGCGCTAAAGTTCTTCACCCCCGCACCATTACCCCCATCGCCCAGTTCCAGATCCCTTGCCTG
ATTAAAAATACCGGAAATCCTCAAGCACCAGGTACGCTCATTGGTGCCAGCCGTGATGAAGACGAATTACCGGTCAAGGGCAT
TTCCAATCTGAATAACATCCTGCATGGCATTAGTTTGTTGGGGCAGTGCCCGGATAGCATCAACGCTGCGCTGATTTGCCGTGG
CGAGAAAATGTCGATCGCCATTATGGCCGGCGTATTAGAAGCGCGCGGTCACAACGTTACTGTTATCGATCCGGTCGAAAAA
CTGCTGGCAGTGGGGCATTACCTCGAATCTACCGTCGATATTGCTGAGTCCACCCGCCGTATTGCGGCAAGCCGCATTCCGG
CTGATCACATGGTGCTGATGGCAGGTTTCACCGCCGGTAATGAAAAAGGCGAACTGGTGGTGCTTGGACGCAACGGTTCCGA
CTACTCTGCTGCGGTGCTGGCTGCCTGTTTACGCGCCGATTGTTGCGAGATTTGGACGGACGTTGACGGGGTCTATACCTGCG
ACCCGCGTCAGGTGCCCGATGCGAGGTTGTTGAAGTCGATGTCCTACCAGGAAGCGATGGAGCTTTCCTACTTCGGCGCTAA
AGTTCTTCACCCCCGCACCATTACCCCCATCGCCCAGTTCCAGATCCCTTGCCTGATTAAAAATACCGGAAATCCTCAAGCAC
CAGGTACGCTCATTGGTGCCAGCCGTGATGAAGACGAATTACCGGTCAAGGGCATTTCCAATCTGAATAACATCCTGCATGG
CATTAGTTTGTTGGGGCAGTGCCCGGATAGCATCAACGCTGCGCTGATTTGCCGTGGCGAGAAAATGTCGATCGCCATTATGG
CCGGCGTATTAGAAGCGCGCGGTCACAACGTTACTGTTATCGATCCGGTCGAAAAACTGCTGGCAGTGGGGCATTACCTCGAA
TCTACCGTCGATATTGCTGAGTCCACCCGCCGTATTGCGGCAAGCCGCATTCCGGCTGATCACATGGTGCTGATGGCAGGTT
TCACCGCCGGTAATGAAAAAGGCGAACTGGTGGTGCTTGGACGCAACGGTTCCGACTACTCTGCTGCGGTGCTGGCTGCCTG
TTTACGCGCCGATTGTTGCGAGATTTGGACGGACGTTGACGGGGTCTATACCTGCGACCCGCGTCAGGTGCCCGATGCGAGG
TTGTTGAAGTCGATGTCCTACCAGGAAGCGATGGAGCTTTCCTACTTCGGCGCTAAAGTTCTTCACCCCCGCACCATTACCCC
CATCGCCCAGTTCCAGATCCCTTGCCTGATTAAAAATACCGGAAATCCTCAAGCACCAGGTACGCTCATTGGTGCCAGCCGT
GATGAAGACGAATTACCGGTCAAGGGCATTTCCAATCTGAATAACATCCTGCATGGCATTAGTTTGTTGGGGCAGTGCCCGG
ATAGCATCAACGCTGCGCTGATTTGCCGTGGCGAGAAAATGTCGATCGCCATTATGGCCGGCGTATTAGAAGCGCGCGGTCAC
AACGTTACTGTTATCGATCCGGTCGAAAAACTGCTGGCAGTGGGGCATTACCTCGAATCTACCGTCGATATTGCTGAGTCCACC
CGCCGTATTGCGGCAAGCCGCATTCCGGCTGATCACATGGTGCTGATGGCAGGTTTCACCGCCGGTAATGAAAAAGGCGAAC
TGGTGGTGCTTGGACGCAACGGTTCCGACTACTCTGCTGCGGTGCTGGCTGCCTGTTTACGCGCCGATTGTTGCGAGATTTGG
ACGGACGTTGACGGGGTCTATACCTGCGACCCGCGTCAGGTGCCCGATGCGAGGTTGTTGAAGTCGATGTCCTACCAGGAAG
CGATGGAGCTTTCCTACTTCGGCGCTAAAGTTCTTCACCCCCGCACCATTACCCCCATCGCCCAGTTCCAGATCCCTTGCCTG
ATTAAAAATACCGGAAATCCTCAAGCACCAGGTACGCTCATTGGTGCCAGCCGTGATGAAGACGAATTACCGGTCAAGGGCAT
TTCCAATCTGAATAACATCCTGCATGGCATTAGTTTGTTGGGGCAGTGCCCGGATAGCATCAACGCTGCGCTGATTTGCCGTGG
CGAGAAAATGTCGATCGCCATTATGGCCGGCGTATTAGAAGCGCGCGGTCACAACGTTACTGTTATCGATCCGGTCGAAAAA
CTGCTGGCAGTGGGGCATTACCTCGAATCTACCGTCGATATTGCTGAGTCCACCCGCCGTATTGCGGCAAGCCGCATTCCGG
CTGATCACATGGTGCTGATGGCAGGTTTCACCGCCGGTAATGAAAAAGGCGAACTGGTGGTGCTTGGACGCAACGGTTCCGA
CTACTCTGCTGCGGTGCTGGCTGCCTGTTTACGCGCCGATTGTTGCGAGATTTGGACGGACGTTGACGGGGTCTATACCTGCG
ACCCGCGTCAGGTGCCCGATGCGAGGTTGTTGAAGTCGATGTCCTACCAGGAAGCGATGGAGCTTTCCTACTTCGGCGCTAA
AGTTCTTCACCCCCGCACCATTACCCCCATCGCCCAGTTCCAGATCCCTTGCCTGATTAAAAATACCGGAAATCCTCAAGCAC
CAGGTACGCTCATTGGTGCCAGCCGTGATGAAGACGAATTACCGGTCAAGGGCATTTCCAATCTGAATAACATCCTGCATGG
CATTAGTTTGTTGGGGCAGTGCCCGGATAGCATCAACGCTGCGCTGATTTGCCGTGGCGAGAAAATGTCGATCGCCATTATGG
CCGGCGTATTAGAAGCGCGCGGTCACAACGTTACTGTTATCGATCCGGTCGAAAAACTGCTGGCAGTGGGGCATTACCTCGAA
TCTACCGTCGATATTGCTGAGTCCACCCGCCGTATTGCGGCAAGCCGCATTCCGGCTGATCACATGGTGCTGATGGCAGGTT
TCAC
0,06 % of E. coli complete genome seq.
LABORATORIO DE BIOINFORMÁTICA Y MATEMÁTICA DE GENOMAS UNIVERSIDAD DE CHILE

CDSs to Functions

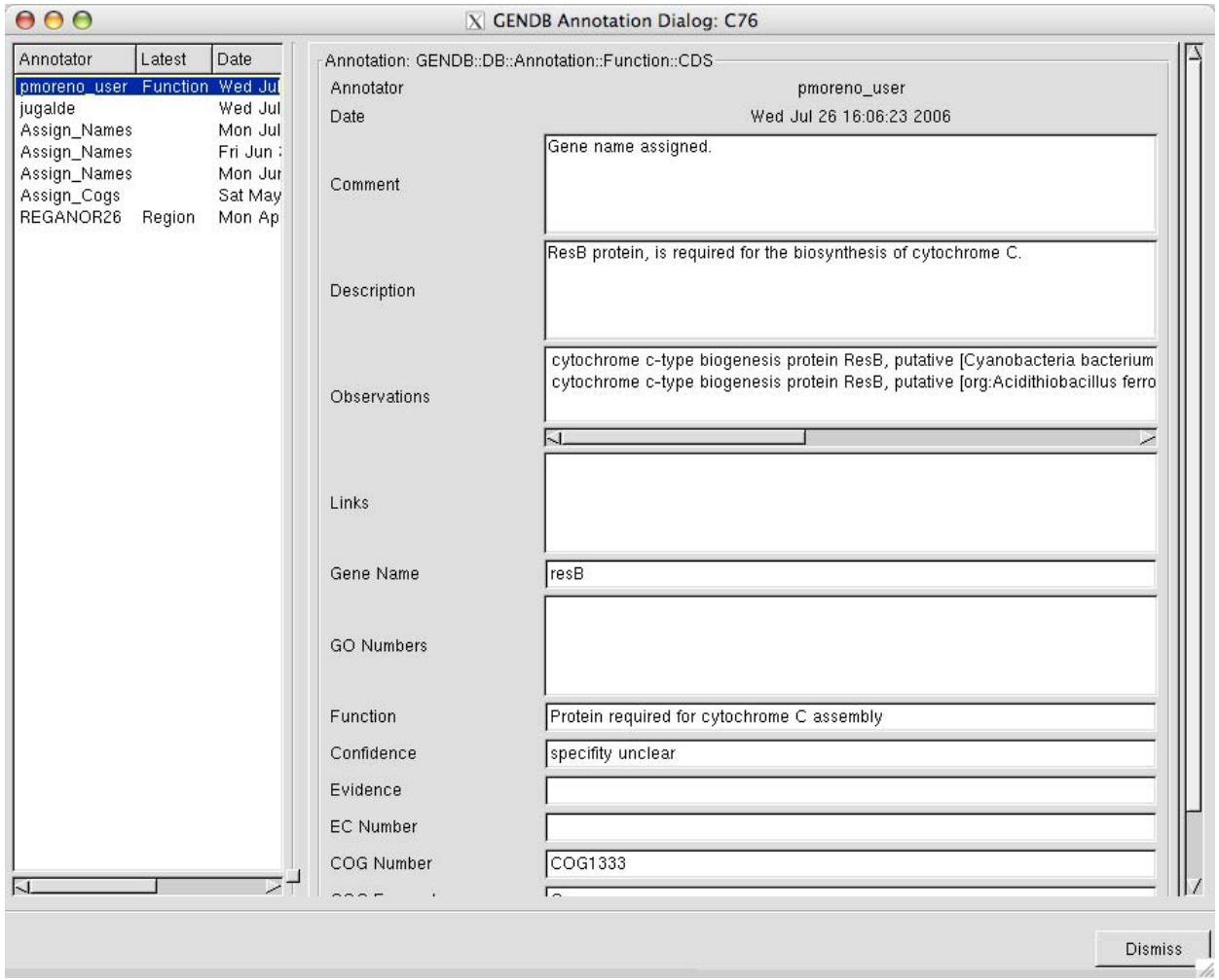
- “Black” CDSs say nothing
- Search Tools
  - Blast (t-n, -x, -p)
  - Emboss
  - InterProScan (pfam, hmmer)
- Sequence Databases & Controlled Vocabularies
  - COG, KOG.
  - NCBI, Swiss-Prot, etc.
  - GO, EC.
- Assignment Method!!

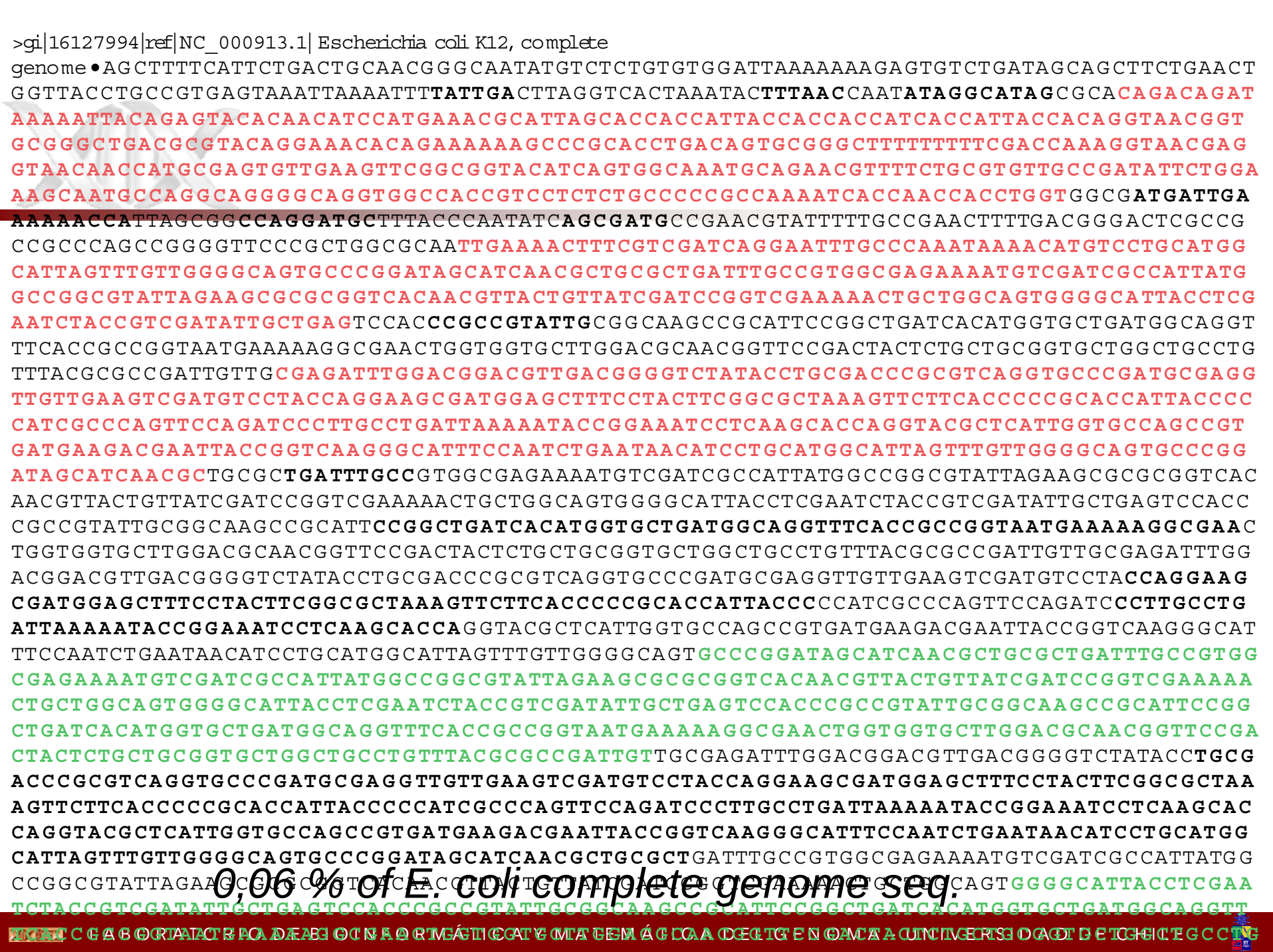
Observations



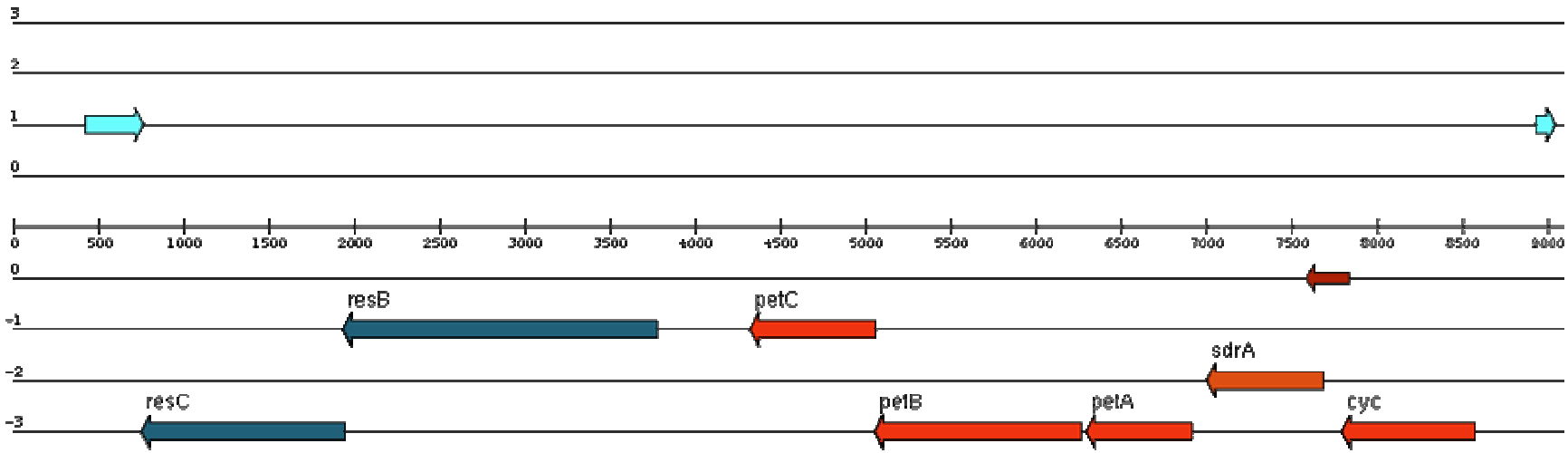
Observation	Score	E-Value	Tool	DB	Start-Stop	Go	Organism	Description
	1153	0.0	Blast2p vs nr	gi|3790604|gb|AAC68692.1|	121 - 1842	-	-	putative cytochrome c-type ma
	470	9e-131	Blast2p vs nr	gi|15808110|emb|CAC88359.1|	157 - 867	-	-	ResB protein [Acidithiobacillus
	324	2e-89	RPSBlast vs. CDD	gnl|CDD|26349	70 - 1356	-	-	pfam05140, ResB, ResB-like fa
	328	5e-88	Blast2p vs nr	gi|78362680|gb|ABB40645.1|	52 - 1809	-	-	ResB-like [Thiomicrospira crun
	315	9e-85	Blast2p vs KEGG	cvi:CV4387	43 - 1797	-	-	probable cytochrome c-type b
	315	4e-84	Blast2p vs nr	gi|34499842|ref|NP_904057.1|	43 - 1797	-	-	probable cytochrome c-type b
	293	1e-77	Blast2p vs nr	gi|78702627|ref|ZP_00867058.1|	52 - 1515	-	-	probable cytochrome c-type b
	281	7e-74	Blast2p vs nr	gi|74316204|ref|YP_313944.1|	25 - 1368	-	-	probable cytochrome c-type b
	276	1e-72	Blast2p vs nr	gi|83748182|ref|ZP_00945209.1|	16 - 1419	-	-	Heme export protein ResB [Ral
	275	5e-72	Blast2p vs nr	gi|83719536|ref|YP_443539.1|	25 - 1425	-	-	cytochrome c assembly family
	266	2e-69	Blast2p vs nr	gi|72120243|gb|AAZ62506.1|	43 - 1419	-	-	ResB-like [Ralstonia eutropha
	263	4e-69	Blast2p vs KEGG	rso:RS01235	70 - 1419	-	-	putative cytochrome C-type bi
	265	6e-69	Blast2p vs nr	gi|76581236|gb|ABA50711.1|	34 - 1425	-	-	ResB-like family superfamily [E
	263	2e-68	Blast2p vs nr	gi|17430010|emb|CAD16695.1|	70 - 1419	-	-	PUTATIVE CYTOCHROME C-
	263	2e-68	Blast2p vs nr	gi|84360474|ref|ZP_00985174.1|	25 - 1380	-	-	COG1333: ResB protein requir
	263	2e-68	Blast2p vs nr	gi|68186879|gb|EAN01577.1|	61 - 1377	-	-	ResB-like [Methylobacillus flag
	258	1e-67	Blast2p vs KEGG	nme:NMB1804	58 - 1515	-	-	cytochrome c-type biogenesis
	258	5e-67	Blast2p vs nr	gi|7227058|gb|AAF42141.1|	58 - 1515	-	-	putative cytochrome c-type bi
	257	9e-67	Blast2p vs nr	gi|68559182|ref|ZP_00598517.1|	43 - 1413	-	-	ResB-like [Ralstonia metallidur
	256	2e-66	Blast2p vs nr	gi|67671561|ref|ZP_00468348.1|	121 - 1425	-	-	COG1333: ResB protein requir
	256	2e-66	Blast2p vs nr	gi|67634531|ref|ZP_00433502.1|	121 - 1425	-	-	COG1333: ResB protein requir
	256	2e-66	Blast2p vs nr	gi|83678622|ref|ZP_00940346.1|	121 - 1425	-	-	COG1333: ResB protein requir
	254	2e-66	Blast2p vs KEGG	nma:NMA0659	121 - 1515	-	-	putative membrane protein
	256	3e-66	Blast2p vs nr	gi|59800563|ref|YP_207275.1|	58 - 1515	-	-	putative cytochrome biogenes
	254	6e-66	Blast2p vs nr	gi|52211202|emb|CAH37191.1|	121 - 1425	-	-	putative cytochrome C biogen

Annotations



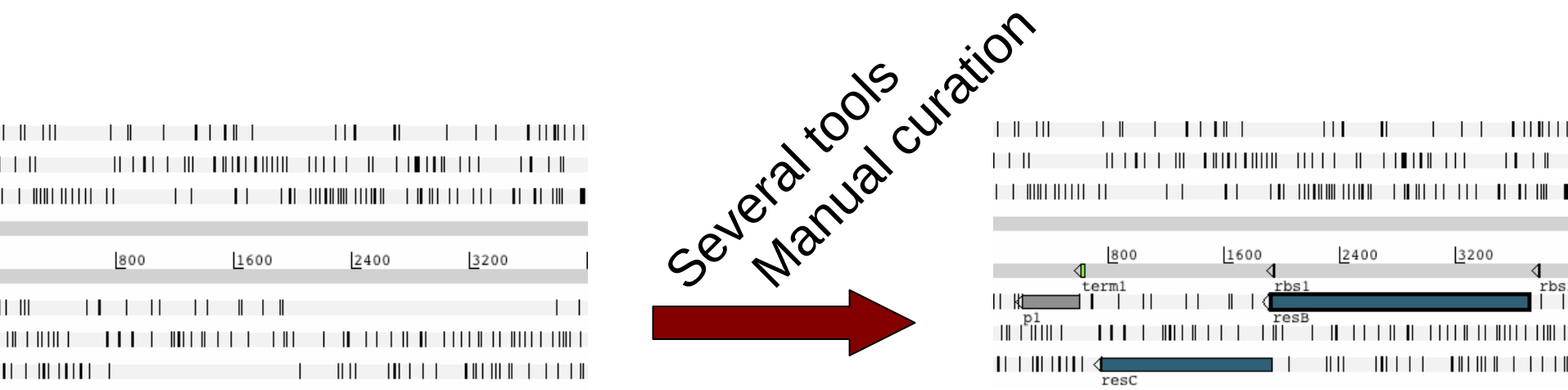

>gi|16127994|ref|NC_000913.1| Escherichia coli K12, complete
genome • AGCTTTTCATTCTGACTGCAACGGGCAATATGTCTCTGTGTGGATTAAAAAAAGAGTGTCTGATAGCAGCTTCTGAACT
GGTTACCTGCCGTGAGTAAATTAAAATTTTATTGACTTAGGTCACTAAATACTTTAACCAATATAGGCATAGCGCACAGACAGAT
AAAAATTACAGAGTACACAACATCCATGAAACGCATTAGCACCACCATTACCACCACCATCACCATTACCACAGGTAACGGT
GCGGGCTGACGCGTACAGGAAACACAGAAAAAAGCCCGCACCTGACAGTGCGGGCTTTTTTTTTCGACCAAAGGTAACGAG
GTAACAACCATGCGAGTGTTGAAGTTCGGCGGTACATCAGTGGCAAATGCAGAACGTTTTCTGCGTGTTGCCGATATTCTGGA
AAGCAATGCCAGGCAGGGGCAGGTGGCCACCGTCCTCTCTGCCCCCGCCAAAATCACCAACCACCTGGTGGCGATGATTGA
AAAAACCATTAGCGGCCAGGATGCTTTACCCAATATCAGCGATGCCGAACGTATTTTTGCCGAACTTTTGACGGGACTCGCCG
CCGCCCAGCCGGGGTTCCCGCTGGCGCAATTGAAAACTTTCGTCGATCAGGAATTTGCCCAAATAAAACATGTCCTGCATGG
CATTAGTTTGTTGGGGCAGTGCCCGGATAGCATCAACGCTGCGCTGATTTGCCGTGGCGAGAAAATGTCGATCGCCATTATG
GCCGGCGTATTAGAAGCGCGCGGTCACAACGTTACTGTTATCGATCCGGTCGAAAAACTGCTGGCAGTGGGGCATTACCTCG
AATCTACCGTCGATATTGCTGAGTCCACCCGCCGTATTGCGGCAAGCCGCATTCCGGCTGATCACATGGTGCTGATGGCAGGT
TTCACCGCCGGTAATGAAAAAGGCGAACTGGTGGTGCTTGGACGCAACGGTTCCGACTACTCTGCTGCGGTGCTGGCTGCCTG
TTTACGCGCCGATTGTTGCGAGATTTGGACGGACGTTGACGGGGTCTATACCTGCGACCCGCGTCAGGTGCCCGATGCGAGG
TTGTTGAAGTCGATGTCCTACCAGGAAGCGATGGAGCTTTCCTACTTCGGCGCTAAAGTTCTTCACCCCCGCACCATTACCCC
CATCGCCCAGTTCCAGATCCCTTGCCTGATTAAAAATACCGGAAATCCTCAAGCACCAGGTACGCTCATTGGTGCCAGCCGT
GATGAAGACGAATTACCGGTCAAGGGCATTTCCAATCTGAATAACATCCTGCATGGCATTAGTTTGTTGGGGCAGTGCCCGG
ATAGCATCAACGCTGCGCTGATTTGCCGTGGCGAGAAAATGTCGATCGCCATTATGGCCGGCGTATTAGAAGCGCGCGGTCAC
AACGTTACTGTTATCGATCCGGTCGAAAAACTGCTGGCAGTGGGGCATTACCTCGAATCTACCGTCGATATTGCTGAGTCCACC
CGCCGTATTGCGGCAAGCCGCATTCCGGCTGATCACATGGTGCTGATGGCAGGTTTCACCGCCGGTAATGAAAAAGGCGAAC
TGGTGGTGCTTGGACGCAACGGTTCCGACTACTCTGCTGCGGTGCTGGCTGCCTGTTTACGCGCCGATTGTTGCGAGATTTGG
ACGGACGTTGACGGGGTCTATACCTGCGACCCGCGTCAGGTGCCCGATGCGAGGTTGTTGAAGTCGATGTCCTACCAGGAAG
CGATGGAGCTTTCCTACTTCGGCGCTAAAGTTCTTCACCCCCGCACCATTACCCCCATCGCCCAGTTCCAGATCCCTTGCCTG
ATTAAAAATACCGGAAATCCTCAAGCACCAGGTACGCTCATTGGTGCCAGCCGTGATGAAGACGAATTACCGGTCAAGGGCAT
TTCCAATCTGAATAACATCCTGCATGGCATTAGTTTGTTGGGGCAGTGCCCGGATAGCATCAACGCTGCGCTGATTTGCCGTGG
CGAGAAAATGTCGATCGCCATTATGGCCGGCGTATTAGAAGCGCGCGGTCACAACGTTACTGTTATCGATCCGGTCGAAAAA
CTGCTGGCAGTGGGGCATTACCTCGAATCTACCGTCGATATTGCTGAGTCCACCCGCCGTATTGCGGCAAGCCGCATTCCGG
CTGATCACATGGTGCTGATGGCAGGTTTCACCGCCGGTAATGAAAAAGGCGAACTGGTGGTGCTTGGACGCAACGGTTCCGA
CTACTCTGCTGCGGTGCTGGCTGCCTGTTTACGCGCCGATTGTTGCGAGATTTGGACGGACGTTGACGGGGTCTATACCTGCG
ACCCGCGTCAGGTGCCCGATGCGAGGTTGTTGAAGTCGATGTCCTACCAGGAAGCGATGGAGCTTTCCTACTTCGGCGCTAA
AGTTCTTCACCCCCGCACCATTACCCCCATCGCCCAGTTCCAGATCCCTTGCCTGATTAAAAATACCGGAAATCCTCAAGCAC
CAGGTACGCTCATTGGTGCCAGCCGTGATGAAGACGAATTACCGGTCAAGGGCATTTCCAATCTGAATAACATCCTGCATGG
CATTAGTTTGTTGGGGCAGTGCCCGGATAGCATCAACGCTGCGCTGATTTGCCGTGGCGAGAAAATGTCGATCGCCATTATGG
CCGGCGTATTAGAAGCGCGCGGTCACAACGTTACTGTTATCGATCCGGTCGAAAAACTGCTGGCAGTGGGGCATTACCTCGAA
TCTACCGTCGATATTGCTGAGTCCACCCGCCGTATTGCGGCAAGCCGCATTCCGGCTGATCACATGGTGCTGATGGCAGGTT
0,06 % of E. coli complete genome seq.

Real genome view



Environments & Pipelines

- Need to integrate several tools.
- Need to automate this integration.
- Avoid file storage, migrate data to a relational database.



Annotation Pipeline

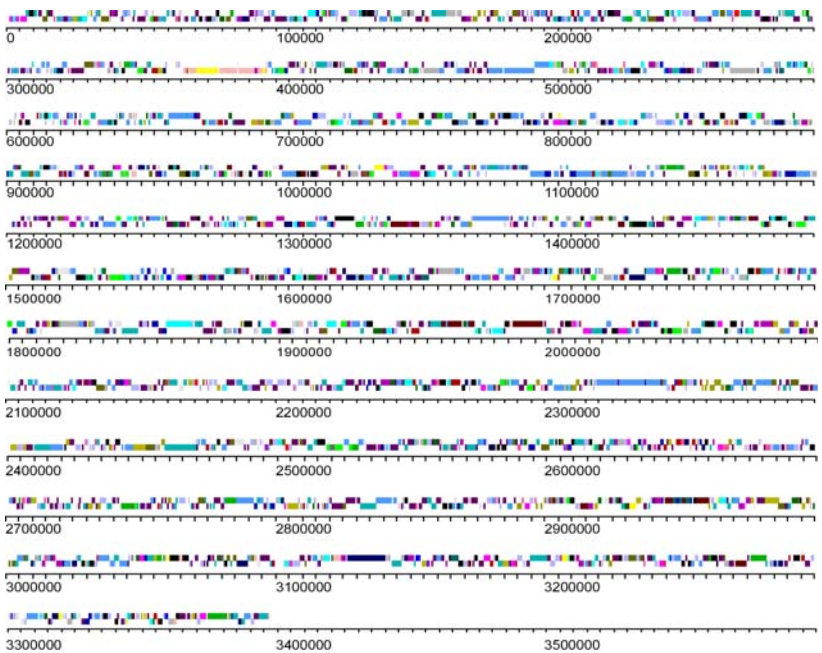
- Integrates several annotation tools.
- Perl, C++, Python plug-ins
- SQL Database backend
- Client-Server
- Grid Job Submission

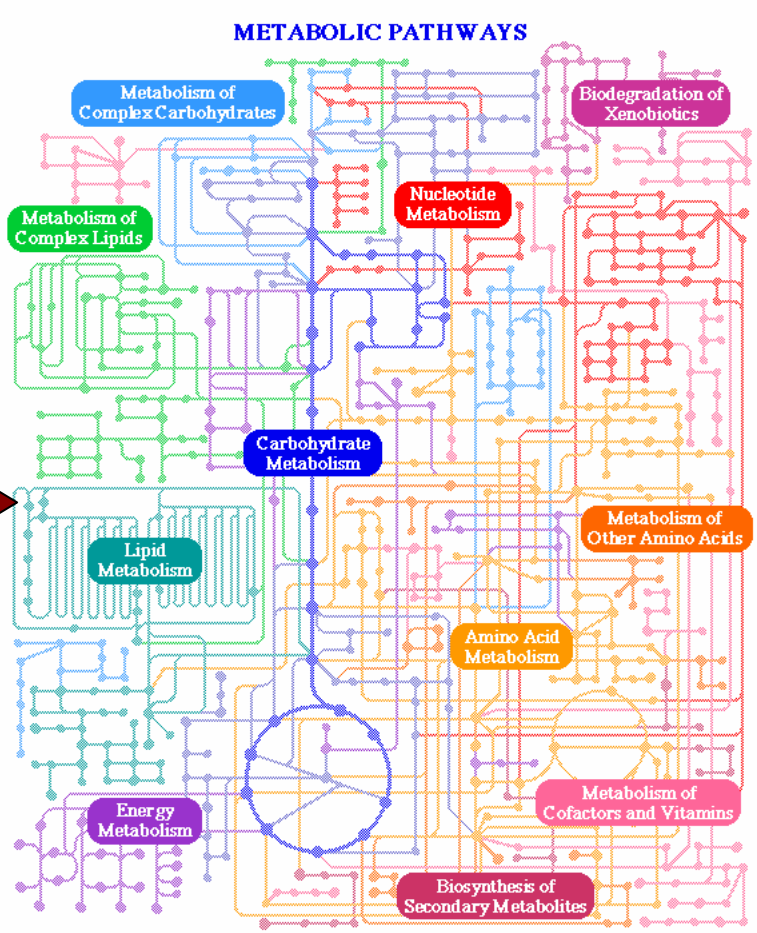
Systems Biology

Systems Biology

- *"Systems biology is an academic field that seeks to integrate high-throughput biological studies to understand how biological systems function. By studying the relationships and interactions between various parts of a biological system (e.g. metabolic pathways, organelles, cells, physiological systems, organisms etc.) it is hoped that eventually an understandable model of the whole system can be developed."* -Wikipedia

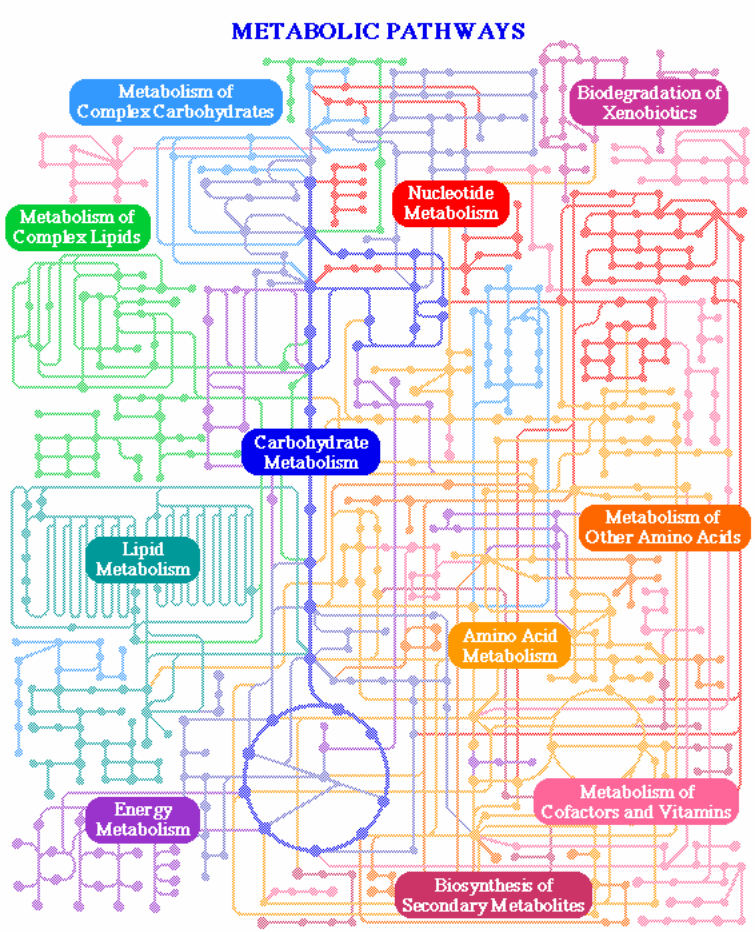
Gene Reg. & Metabolic Pathways





Populating the network

- Kinetics Data.
- Stationary Fluxes.
- Gene Expression.
- Protein-DNA Interactions.
- Physical constraints.
- *New pathways inferences.*



Simulating “the Cell”

- Dynamic.
- Stationary.
- As an Optimization LP problem.
- Parameter Estimation.