

A large, faint, light gray DNA double helix is positioned on the left side of the slide, behind the title.

Genome Sequence Annotation

Pablo A. Moreno C.

Ingeniero Civil en Biotecnología, UCH.
Ingeniero Jefe de Proyectos

Juan A. Ugalde C.

Ingeniero en Biotecnología Molecular, UCH.
Ingeniero de Proyectos

Andrés Aravena D.

Ingeniero Civil Matemático, UCH.
Ingeniero Jefe LBMG

Senior Advisors:

Mauricio Gonzales, Ph.D

Biology, INTA-UCH

Alejandro Maass, Ph.D

Mathematics, DIM-UCH



Today's Menu

- Genome Sequencing.
- Genome Assembly.
- *Genome Annotation.*
 - Region Marking.
 - Functional Annotation.
- A few others.





The input: genome assembly

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





Sequencing Process

mRNA

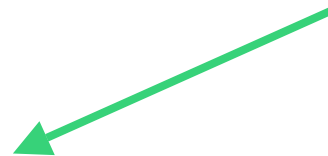


cDNA



ABI 3700
DNA Sequencer

DNA



ATGCCTGA

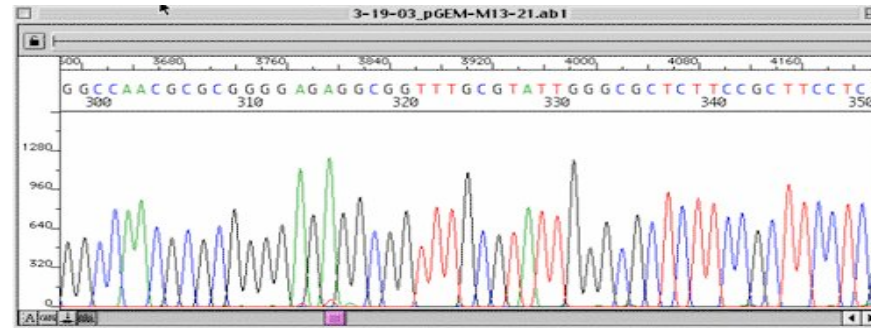
Chromatograms
Sequences
Qualities





From Chromatograms to Sequences

Chromatograms



Base Calling

Sequences

G A T C A G G C T A C G A T C G A T C T A G C

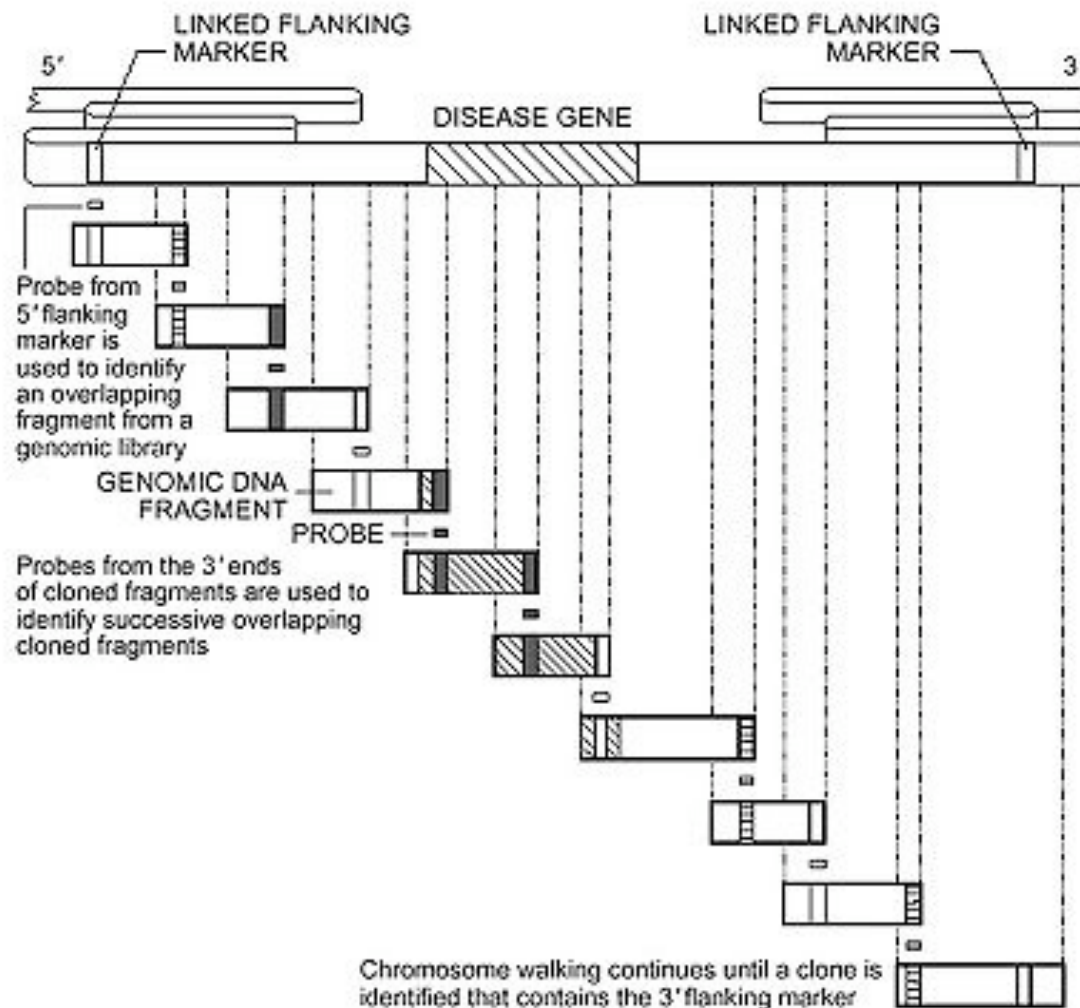
Qualities

7 15 22 25 38 46 56 66 64 75 71 68 52 46 68 53 66 65 65 45 32





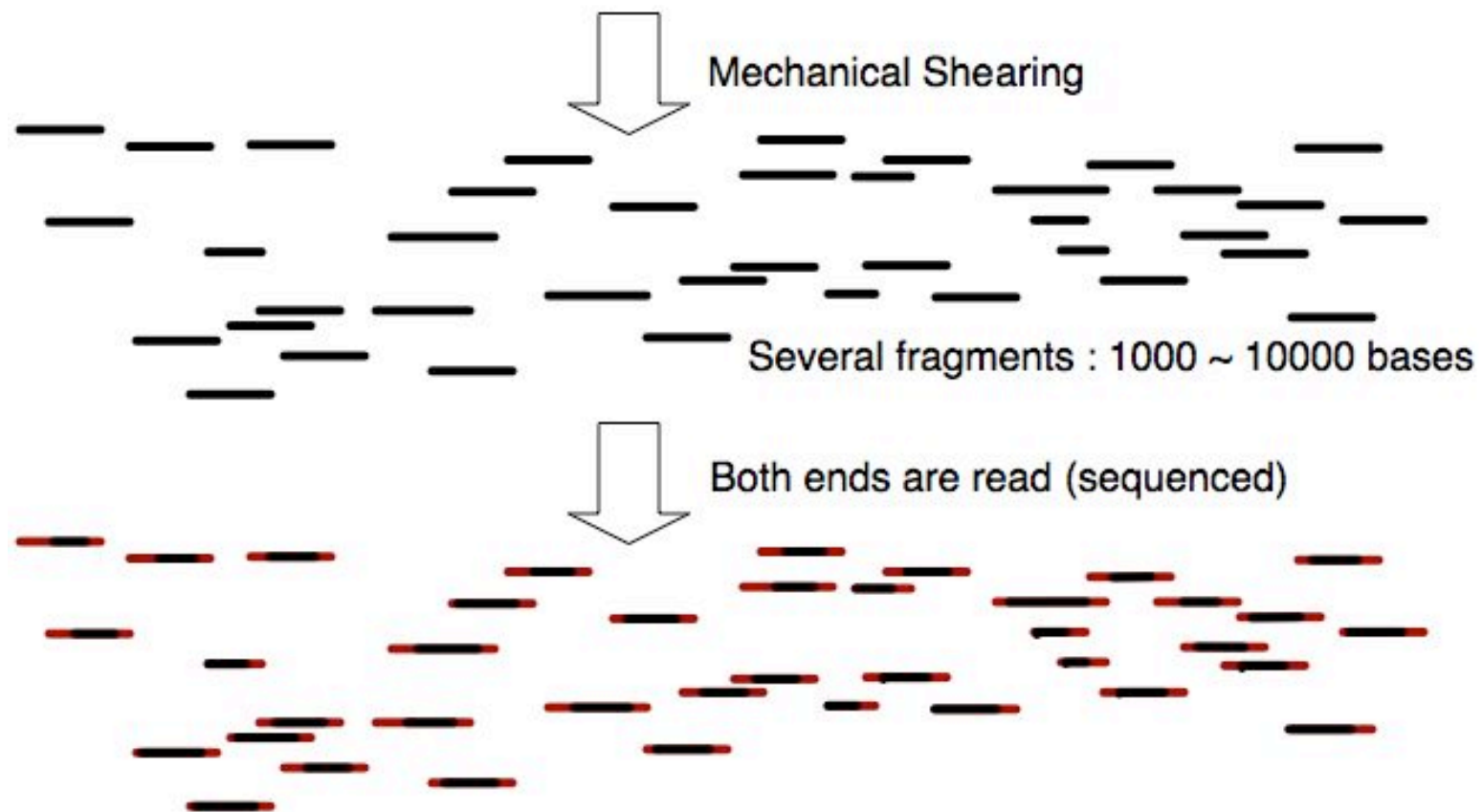
Walking Primers





Shot-gun sequencing

Microbial DNA : 2.5 Million bases ~ 6 Million bases (several copies)



DNA assembly



Sequencing

CAGTCTGCTA TATGCTATCGCA
GCTATCGACTAG ATTCGAGTCAG
AGAGCATACTG CAGTCTGCTA
GGTATGCATCG

Assembly

ACTTAGCGCATCTAGCTATCTATCTACTATCGACATCATCA



EST assembly

Sequencing



CAGTCTGCTA TATGCTATCGCA
GCTATCGACTAG ATTCGAGTCAG
AGAGCATACTG CAGTCTGCTA
GGTATGCATCG

Assembly

Gen1: ACTTAGCGCATCTAGCTAT

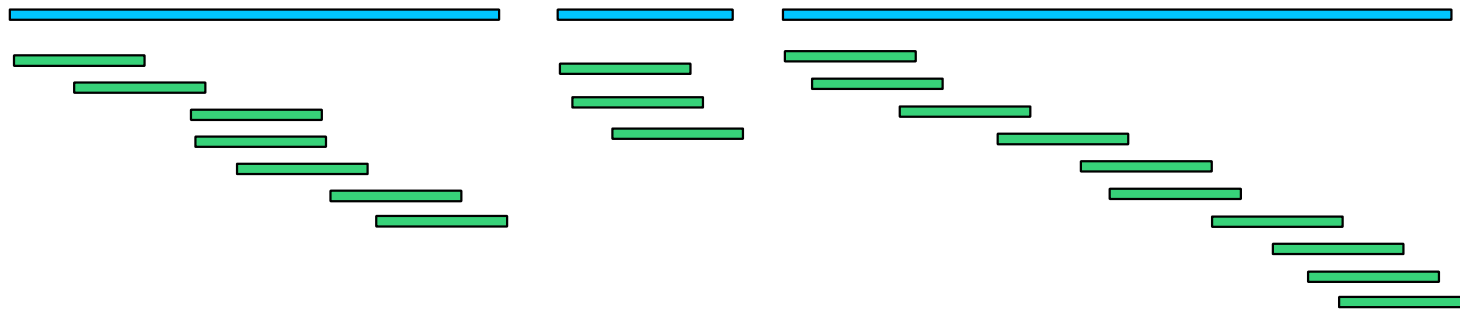
Gen2: GATTCAGCTTACATCTGAA

Gen3: TTGACTAGCTTTCATTCATCCTA





DNA assembly



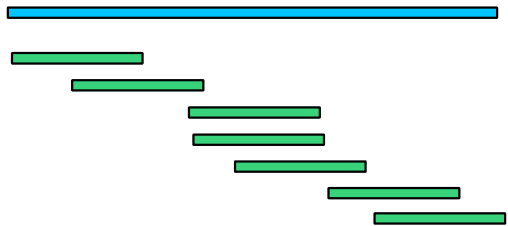
ACTTAGCGC...CTAGCTATCTATCTACTA...GTCA



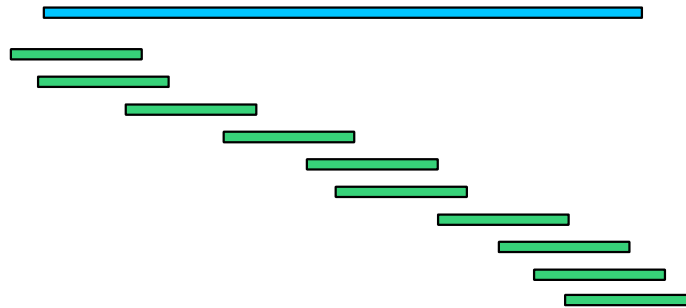


EST assembly

Gen1



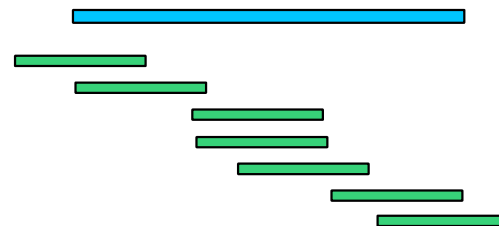
Gen2



Gen3

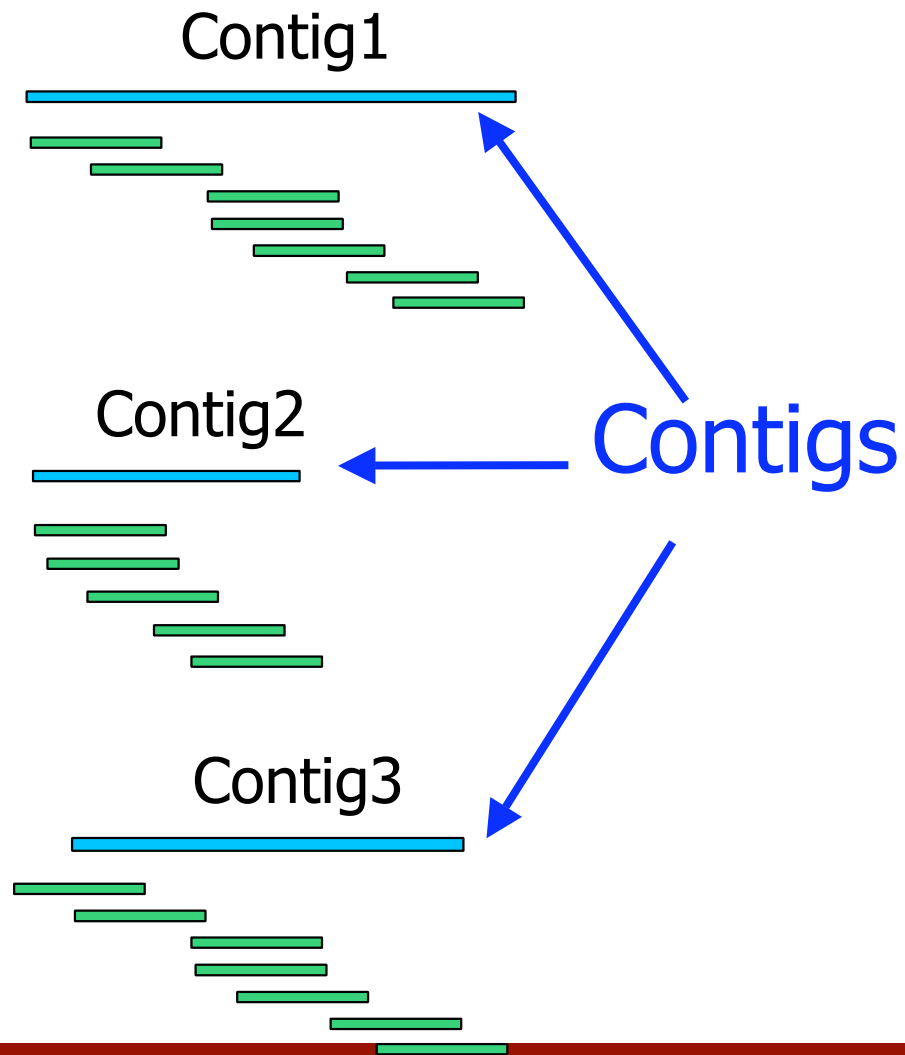


Gen4

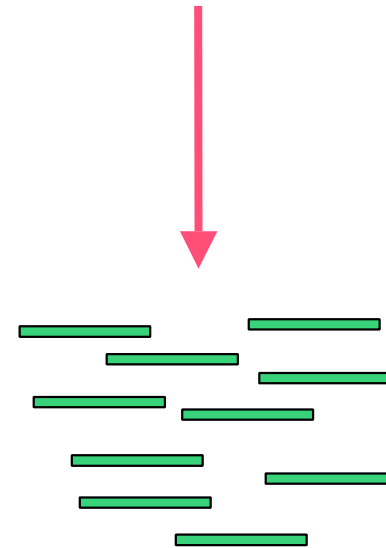




Sequence assembly



Singletons





Programs

- In real life, Heuristics.
- ClustalW, ClustalX
 - Few sequences, proteins
- Phrap, PCAP
 - Genomic assembly
- Phrap, CAP3
 - EST assembly





Example (~ClustalW)

	110	120	130	140	150
G3433	- A S V A A V A A G V G G	K L P S S R Y K G V V P Q P N G R W G A Q I Y	-	ER H L R V W L G T F A G	
G3390	- - - - G E R R G G G G G	K L P S S K Y K G V V P Q P N G R W G A Q I Y	-	ER H Q R V W L G T F T G	
G3391	C G G G G G G A G G A G G	K L P S S K F K G V V P Q P N G R W G A Q I Y	-	ER H Q R V W L G T F A G	
G3432	- - - S G A A A V S V G G	K L P S S R Y K G V V P Q P N G R W G A Q I			
G3389	- - - - - - - - - - - -	- - - S S R Y K G V V P Q P N G R W G A Q I Y	-	ER H A R V W L G T F P D	
G3388	- - - - - - - - - - - -	- Q - S S R Y K G V V P Q P N G R W G A Q I Y	-	ER H A R V W L G T F P D	
G867	- - - - - - - - - - - -	K L P S S K Y K G V V P Q P N G R W G A Q I Y	-	E K H Q R V W L G T F N E	
G1930	- - - - - - - - - - - -	K L P S S R F K G V V P Q P N G R W G A Q I Y	-	E K H Q R V W L G T F N E	
G3451	- - - - - - - - - - - -	K L P S S K Y K G V V P Q P N G R W G A Q I Y	-	E K H Q R V W L G T F N E	
G3452	- - - - - - - - - - - -	K L P S S K Y K G V V P Q P N G R W G A Q I Y	-	E K H Q R V W L G T F N E	
G3453	- - - - - - - - - - - -	K L P S S K Y K G V V P Q P N G R W G A Q I Y	-	E K H Q R V W L G T F N E	
G993	- - - - - - - - - - - -	K L P S S K Y K G V V P Q P N G R W G A Q I Y	-	E K H Q R V W L G T F N E	
G9	- - - - - - - - - - - -	K L P S S K Y K G V V P Q P N G R W G A Q I Y	-	E K H Q R V W L G T F N E	
G2690	- - - - - - - - - - - -	- - - T I K F K G V V Q Q Q N G H W G A Q I Y	-	A D H R R I W L G T F K S	
G2687	- - - - - - - - - - - -	- - - T V V A S G N V K Y K G V V Q Q Q N G H W G A Q I Y	-	A D H K R I W L G T F K S	
AP004178	- - - - - - - - - - - -	E K E H M F D K V V T P S D V G K L N R L V I P K Q	Y A E K Y F P L D A A		
AL662987	- - - - - - - - - - - -	E K E H M F D K V V T P S D V G K L N R L V I P K Q	H A E K Y F P L D S A		
G1957	- - - - - - - - - - - -	E K E H M F D K V V T P S D V G K L N R L V I P K Q	H A E R Y F P L D S S		
G1010	- - - - - - - - - - - -	D R E H M F D K V V T P S D V G K L N R L V I P K Q	H A E R F F P L D S S		
Consensus		K L P S S K Y K G V V P Q P N G R W G A Q I Y		E . H . R V W L G T F	

AP2 domain



Example (Phrap-Consed)

350 360 370

TTGGGTTTTTCCGACTCAAACAACATCGT

CONSensus

PU3_plate21_019.ab1 ttggggtttttccgactcaaacaacatcgt

PU4_plate50_L19.ab1 ttggggtttttccgactcaaacaacatcgt

PU4_plate41_F02.ab1 ttggggtttttccgactcaaacaacatcgt

PU2_plate16_D23.ab1 ttggggtttttccgactcaaacaacatcgt

PU4_plate42_D20.ab1 ttggggtttttccgactcaaacaacatcgt

PU2_plate45_M17.ab1 ttggggtttttccgactcaaacaacatcgt

PU4_plate11_E06.ab1 ttggggtttttccgactcaaacaacatcgt

PU4_plate44_B16.ab1 ttggggtttttccgactcaaacaacatcgt

PU3_plate6_B22.ab1 ttggggtttttccgactcaaacaacatcgt

PU2_plate28_K11.ab1 ttggggtttttccgactcaaacaacatcgt

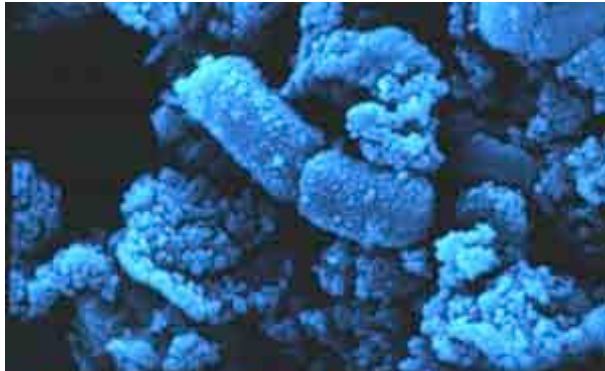
PU3_plate9_J14.ab1 ttggggtttttccgactcaaacaacatcgt

PU4_plate36_L23.ab1 ttggggtttttccgactcaaacaacatcgt





LBMG



Shotgun



Sequencing



CAGTCTGCTA TATGCTATCGCA
GCTATCGACTAG ATTCGAGTCAG
AGAGCATACTG CAGTCTGCTA
GGTATGCATCG

30.000 seqs





Estimated coverage

CAGTCTGCTA TATGCTATCGCA
GCTATCGACTAG ATTCGAGTCAG
AGAGCATACTG CAGTCTGCTA
GGTATGCATCG

30.000 sequences

...ACTTAGCGCATCTAGCTATCTATCTACTATC...

500 good quality bases
per seq



Expected 3 Mb bacterial DNA



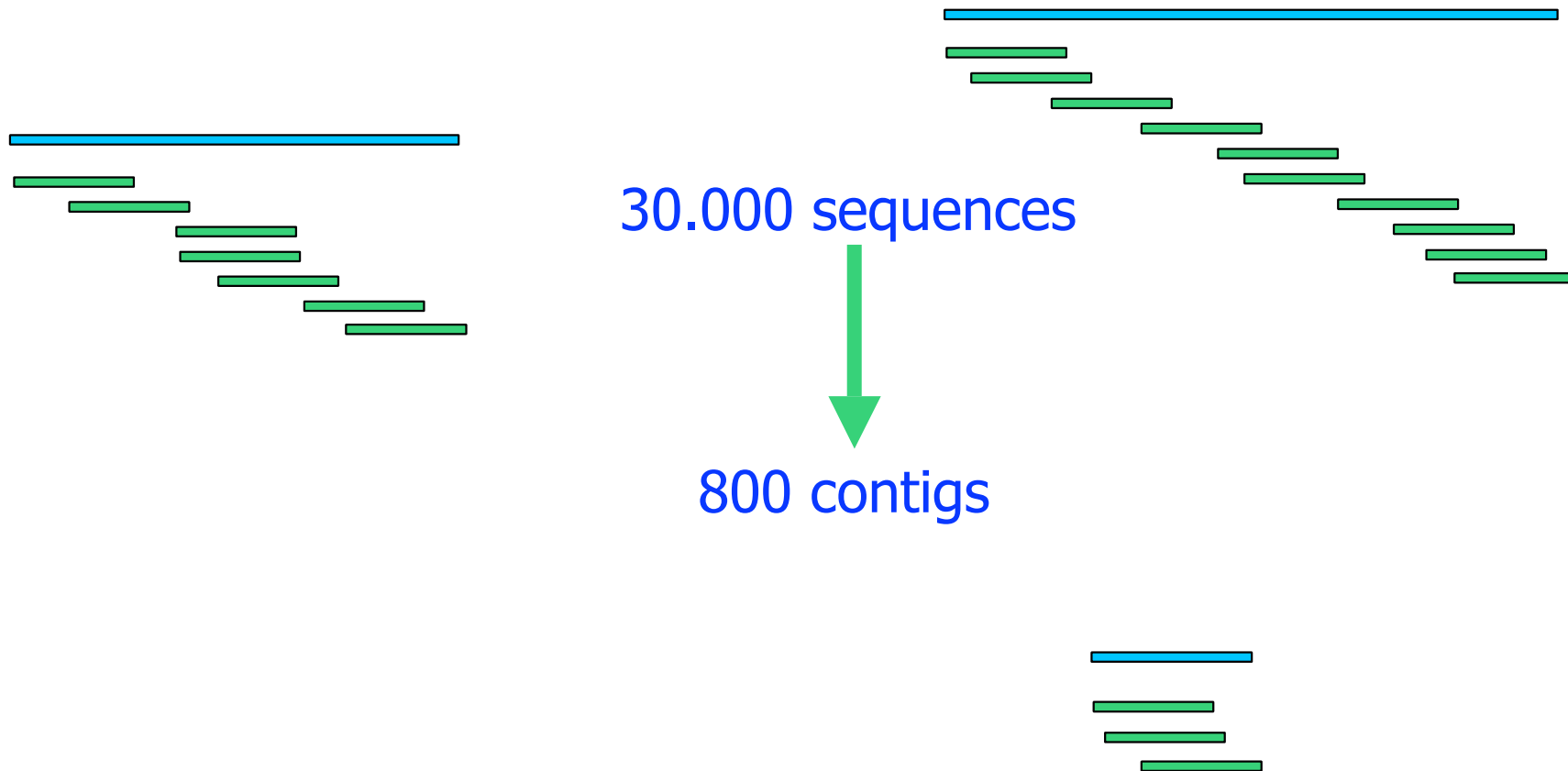
Expected 5x coverage

$$(30.000 * 500) / 3.000.000$$



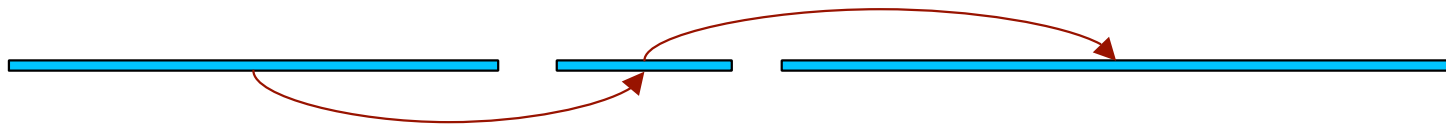


Contig assembly





Scaffolds

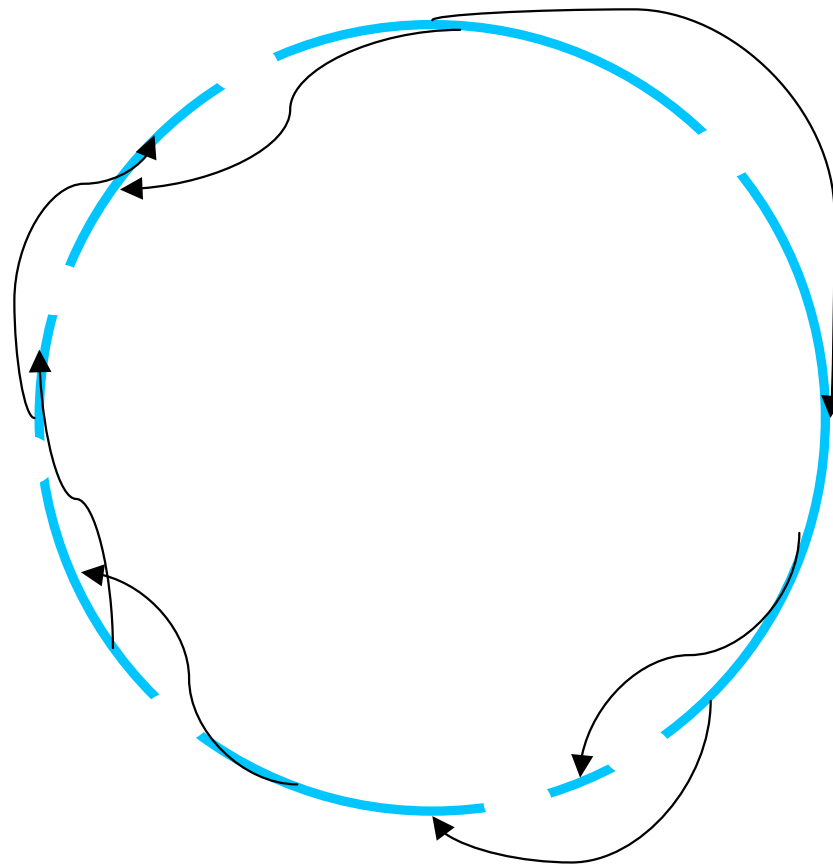


- Use extra information to order contigs
 - Clone name
 - Template alignment
 - Operons



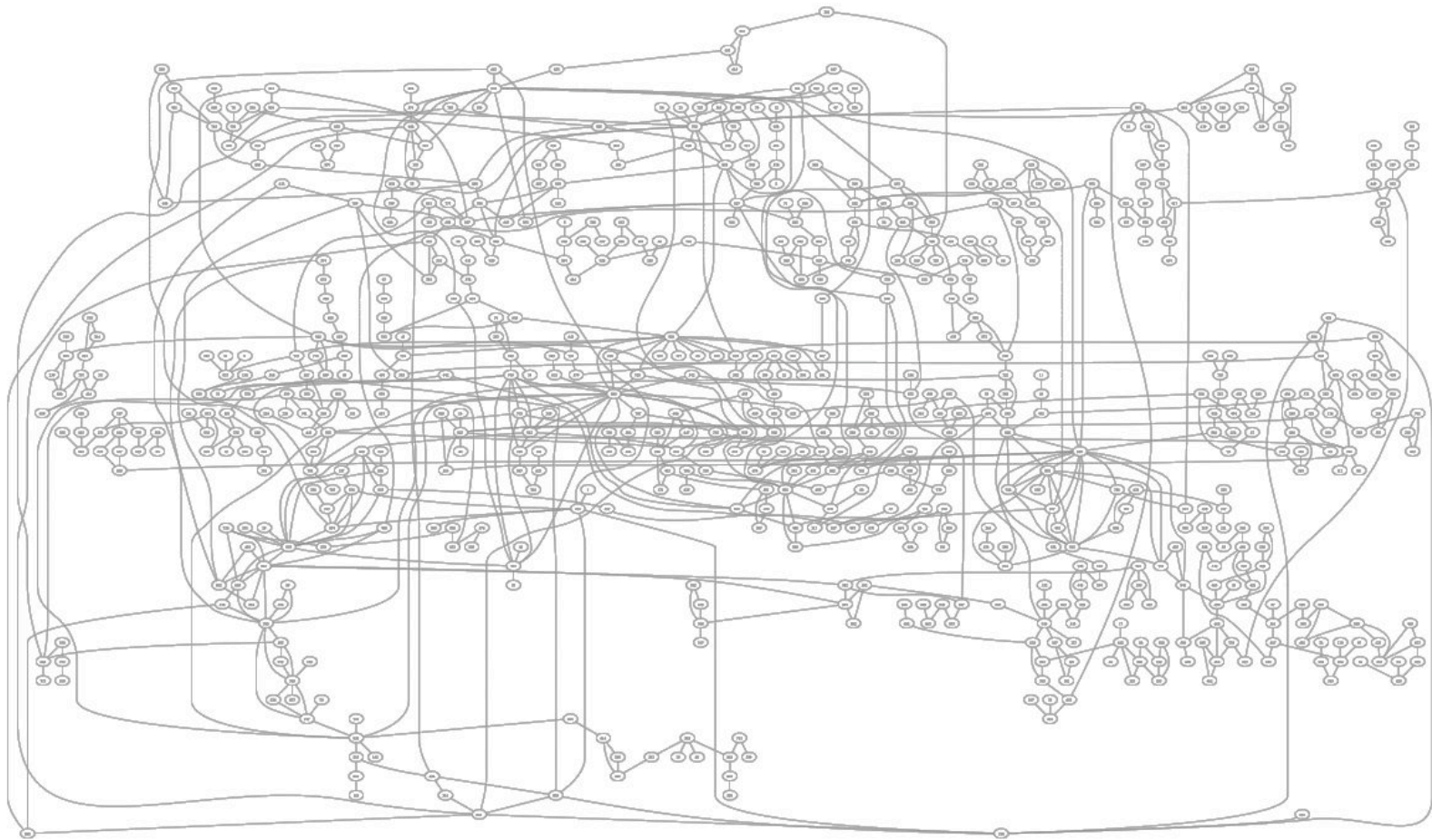


Ideal scaffold



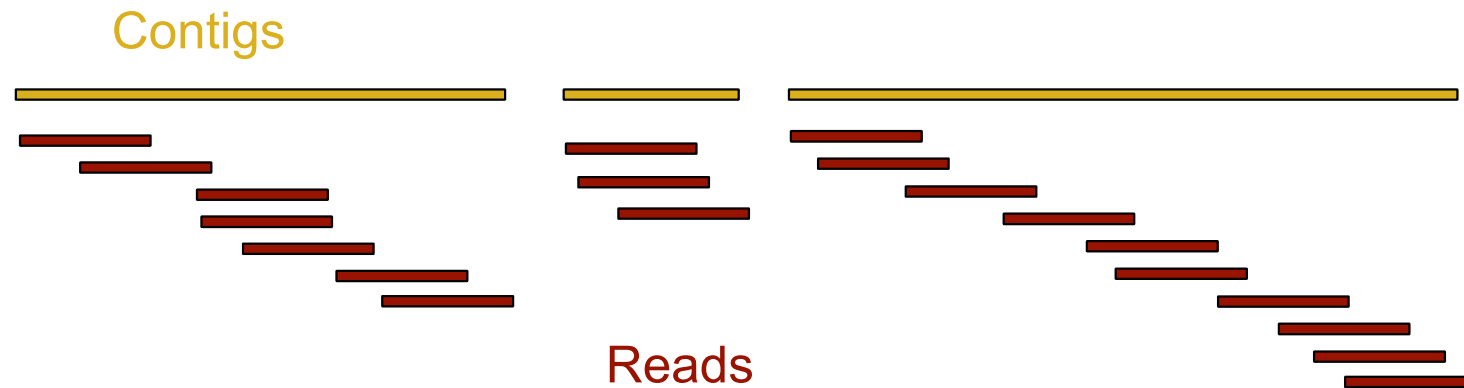


Real scaffold





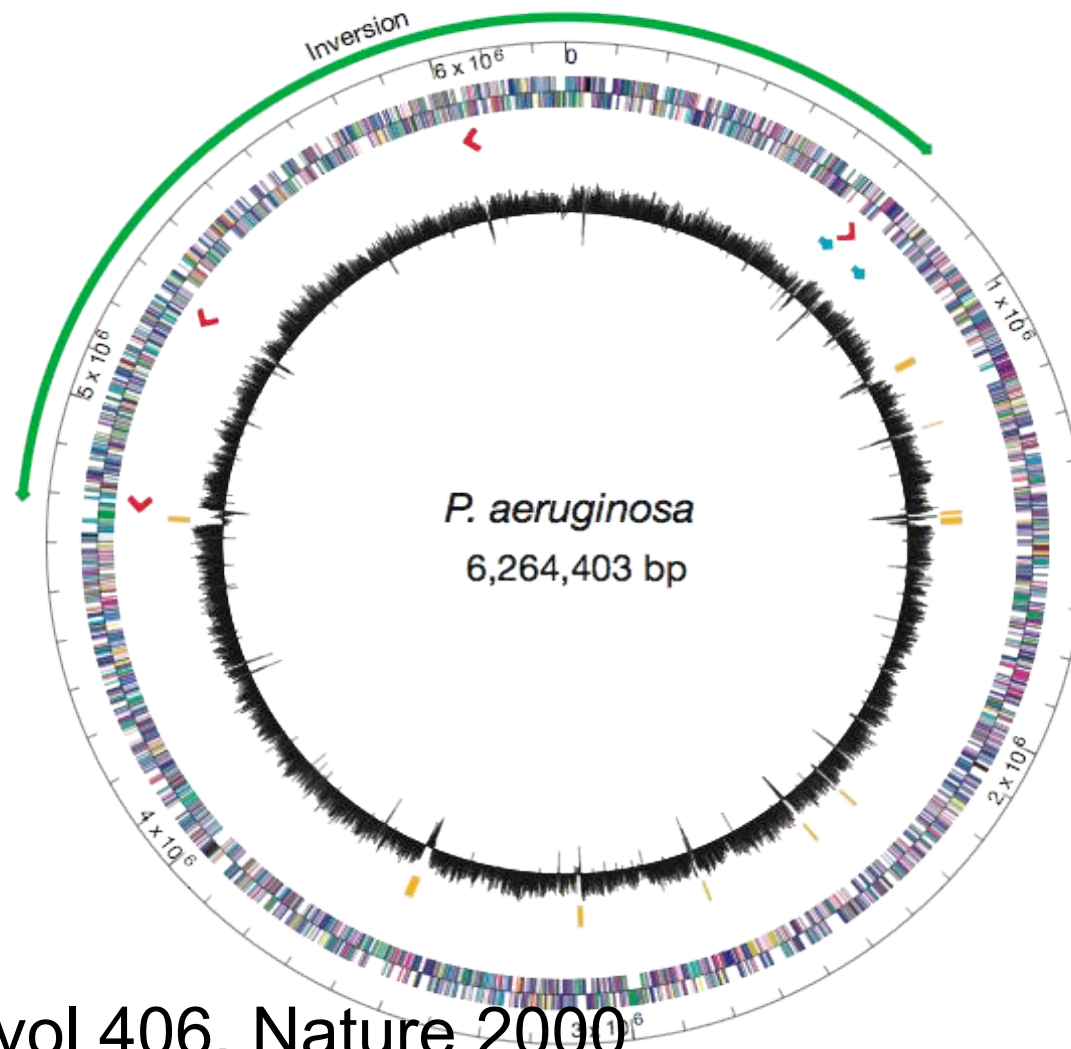
Assembly



ACTTAGCGC...CTAGCTATCTATCTACTA...GTCA



Circular Genome



Stover et al, vol 406, Nature 2000



>gi|16127994|ref|NC_000913.1| Escherichia coli K12, complete genome

AGCTTTTTCATTCTGACTGCAACGGGCAATATGTCTCTGTGTGGATTAAAAAAGAGTGTCTGATAGCAGCTTCTGAACTGGTTACCTGCCGTGAGTAAAT
TAAAAATTTTATTGACTTAGGTCACATAAATACTTTAACCAATATAGGCATAGCGCACAGACAGATAAAAAATTACAGAGTACACAACATCCATGAAACGCAT
TAGCACCACCATTACCACCACCATCACCATTACCACAGGTAACGGTGCAGGGCTGACGCGTACAGGAAACACAGAAAAAAGCCCGCACCTGACAGTGCGGG
CTTTTTTTTTTCGACCAAAGGTAACGAGGTAACAACCATGCGAGTGTGGAAGTTCGGCGGTACATCAGTGGCAAATGCAGAACGTTTTCTGCGTGTGCGG
ATATTCTGGAAAGCAATGCCAGGCAGGGGAGGTGGCCACCGTCCTCTCTGCCCCGCCAAAATCACCAACCACCTGGTGGCGATGATTGAAAAAACCAT
TAGCGGCCAGGATGCTTTACCCAATATCAGCGATGCCGAACGTATTTTTTGCCGAACTTTTGACGGGACTCGCCGCCGCCAGCCGGGGTTCCCGCTGGCG
CAATTGAAAACTTTTCGTCGATCAGGAATTTGCCCAAATAAAACATGTCCTGCATGGCATTAGTTTTGTTGGGGCAGTGCCCGGATAGCATCAACGCTGCGC
TGATTTGCCGTGGCGAGAAAATGTCGATCGCCATTATGGCCGGCGTATTAGAAGCGCGCGGTACACAACGTTACTGTTATCGATCCGGTCGAAAACTGCT
GGCAGTGGGGCATTACCTCGAATCTACCGTCGATATTGCTGAGTCCACCCGCCGTATTGCGGCAAGCCGCATTCCGGCTGATCACATGGTGCTGATGGCA
GGTTTCACCGCCGGTAATGAAAAAGGCGAACTGGTGGTGCTTGGACGCAACGGTTCGACTACTCTGCTGCGGTGCTGGCTGCCTGTTTACGCGCCGATT
GTTGCGAGATTTGGACGGACGTTGACGGGGTCTATACCTGCGACCCGCGTCAGGTGCCCGATGCGAGGTTGTTGAAGTCGATGTCCTACCAGGAAGCGAT
GGAGCTTTCCTACTTCGGCGCTAAAGTTCTTCACCCCCGCACCATTACCCCCATCGCCCAGTTCAGATCCCTTGCTGATTAAAAATACCGGAAATCCT
CAAGCACCAGGTACGCTCATTGGTGCCAGCCGTGATGAAGACGAATTACCGGTCAAGGGCATTTCGAATCTGAATAACATCCTGCATGGCATTAGTTTGT
TGGGGCAGTGCCCGGATAGCATCAACGCTGCGCTGATTGCGGTGGCGAGAAAATGTCGATCGCCATTATGGCCGGCGTATTAGAAGCGCGCGGTACAAA
CGTTACTGTTATCGATCCGGTCGAAAACTGCTGGCAGTGGGGCATTACCTCGAATCTACCGTCGATATTGCTGAGTCCACCCGCCGTATTGCGGCAAGC
CGCATTCGGCTGATCACATGGTGCTGATGGCAGGTTTCACCGCCGGTAATGAAAAAGGCGAACTGGTGGTGCTTGGACGCAACGGTTCGACTACTCTG
CTGCGGTGCTGGCTGCCTGTTTACGCGCCGATTGTTGCGAGATTTGGACGGACGTTGACGGGGTCTATACCTGCGACCCGCGTCAGGTGCCCGATGCGAG
GTTGTTGAAGTCGATGTCCTACCAGGAAGCGATGGAGCTTTCCTACTTCGGCGCTAAAGTTCTTCACCCCCGCACCATTACCCCCATCGCCCAGTTCAG
ATCCCTTGCTGATTAAAAATACCGGAAATCCTCAAGCACCAGGTACGCTCATTGGTGCCAGCCGTGATGAAGACGAATTACCGGTCAAGGGCATTTCGA
ATCTGAATAACATCCTGCATGGCATTAGTTTGTGGGGCAGTGCCCGGATAGCATCAACGCTGCGCTGATTTGCCGTGGCGAGAAAATGTCGATCGCCAT
TATGGCCGGCGTATTAGAAGCGCGCGGTACACAACGTTACTGTTATCGATCCGGTCGAAAACTGCTGGCAGTGGGGCATTACCTCGAATCTACCGTCGAT
ATTGCTGAGTCCACCCGCCGTATTGCGGCAAGCCGCATTCCGGCTGATCACATGGTGCTGATGGCAGGTTTCACCGCCGGTAATGAAAAAGGCGAACTGG
TGGTGCTTGGACGCAACGGTTCGACTACTCTGCTGCGGTGCTGGCTGCCTGTTTACGCGCCGATTGTTGCGAGATTTGGACGGACGTTGACGGGGTCTA
TACCTGCGACCCGCGTCAGGTGCCCGATGCGAGGTTGTTGAAGTCGATGTCCTACCAGGAAGCGATGGAGCTTTCCTACTTCGGCGCTAAAGTTCTTCAC
CCCCGCACCATTACCCCCATCGCCCAGTTCAGATCCCTTGCTGATTAAAAATACCGGAAATCCTCAAGCACCAGGTACGCTCATTGGTGCCAGCCGTG
ATGAAGACGAATTACCGGTCAAGGGCATTTCGAATCTGAATAACATCCTGCATGGCATTAGTTTGTGGGGCAGTGCCCGGATAGCATCAACGCTGCGCT
GATTTGCCGTGGCGAGAAAATGTCGATCGCCATTATGGCCGGCGTATTAGAAGCGCGCGGTACACAACGTTACTGTTATCGATCCGGTCGAAAACTGCTG
GCAGTGGGGCATTACCTCGAATCTACCGTCGATATTGCTGAGTCCACCCGCCGTATTGCGGCAAGCCGCATTCCGGCTGATCACATGGTGCTGATGGCAG
GTTTCACCGCCGGTAATGAAAAAGGCGAACTGGTGGTGCTTGGACGCAACGGTTCGACTACTCTGCTGCGGTGCTGGCTGCCTGTTTACGCGCCGATTG
TTGCGAGATTTGGACGGACGTTGACGGGGTCTATACCTGCGACCCGCGTCAGGTGCCCGATGCGAGGTTGTTGAAGTCGATGTCCTACCAGGAAGCGATG
GAGCTTTCCTACTTCGGCGCTAAAGTTCTTCACCCCCGCACCATTACCCCCATCGCCCAGTTCAGATCCCTTGCTGATTAAAAATACCGGAAATCCTC
AAGCACCAGGTACGCTCATTGGTGCCAGCCGTGATGAAGACGAATTACCGGTCAAGGGCATTTCGAATCTGAATAACAGATCACATGGTGCTGATGGCAG

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

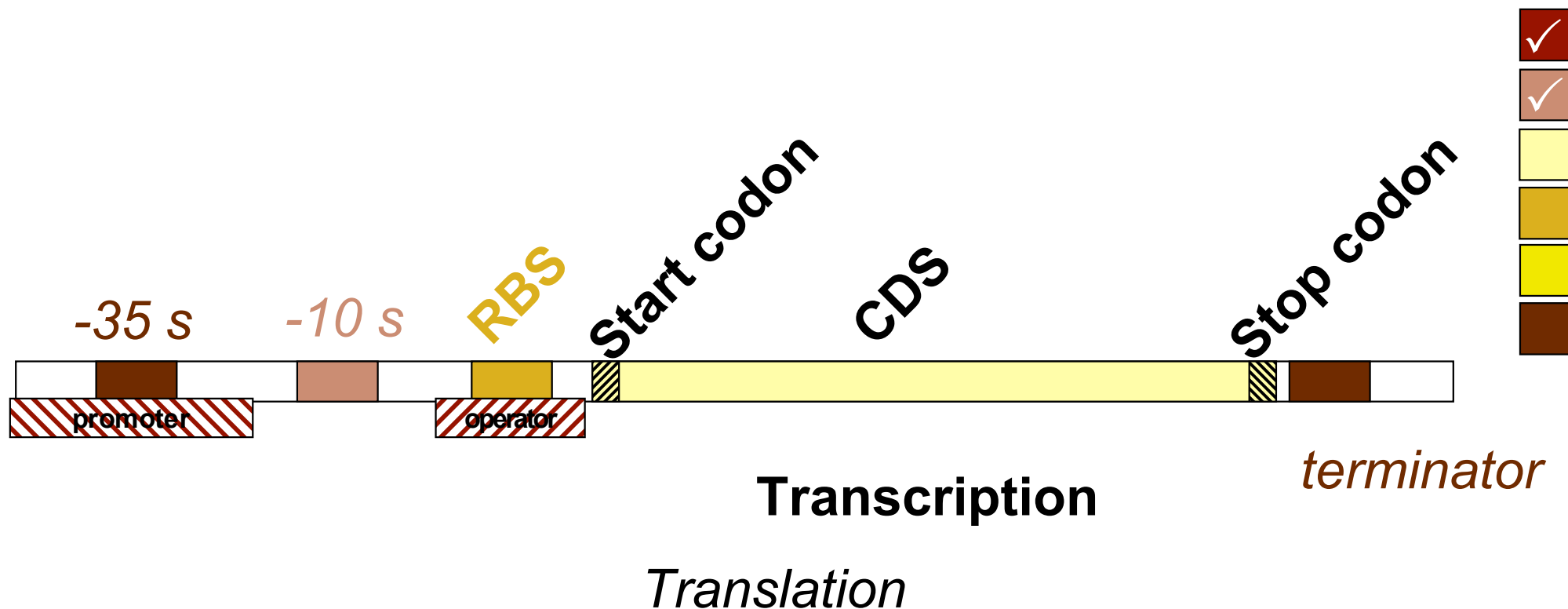


Sequence Features

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

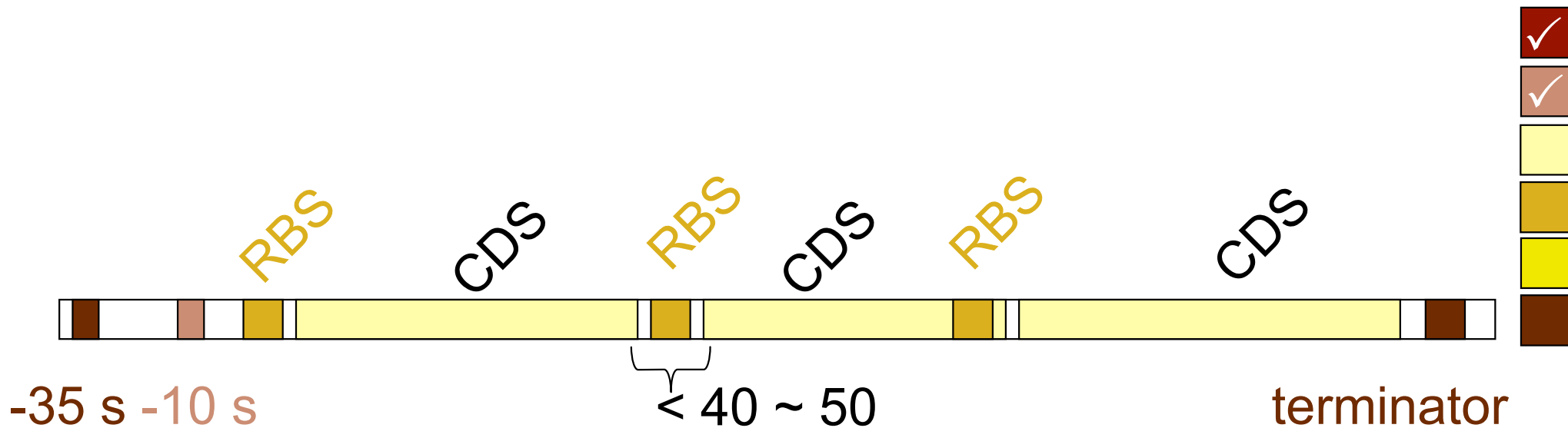


Coding Sequences & Co.



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

Related CDSs: Operons

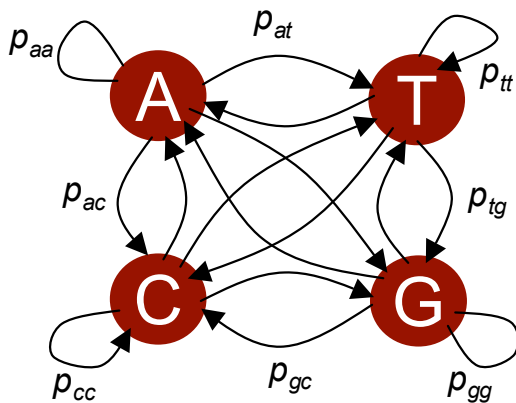


Polycistronic transcripts; Operons.



Main approaches

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



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

```
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.....

.....LIGSREPMVDMPRIIRHARERGCFLAIDAQPERLDLVDIHARTA 518
.....LIGSREPMVDMPRIIRHARERGCFLAIDAQPERLDLVDIHARTA
.....LIGSREPMVDMPRIIRHARERGCFLAIDAQPERLDLVDIHARTA 420

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





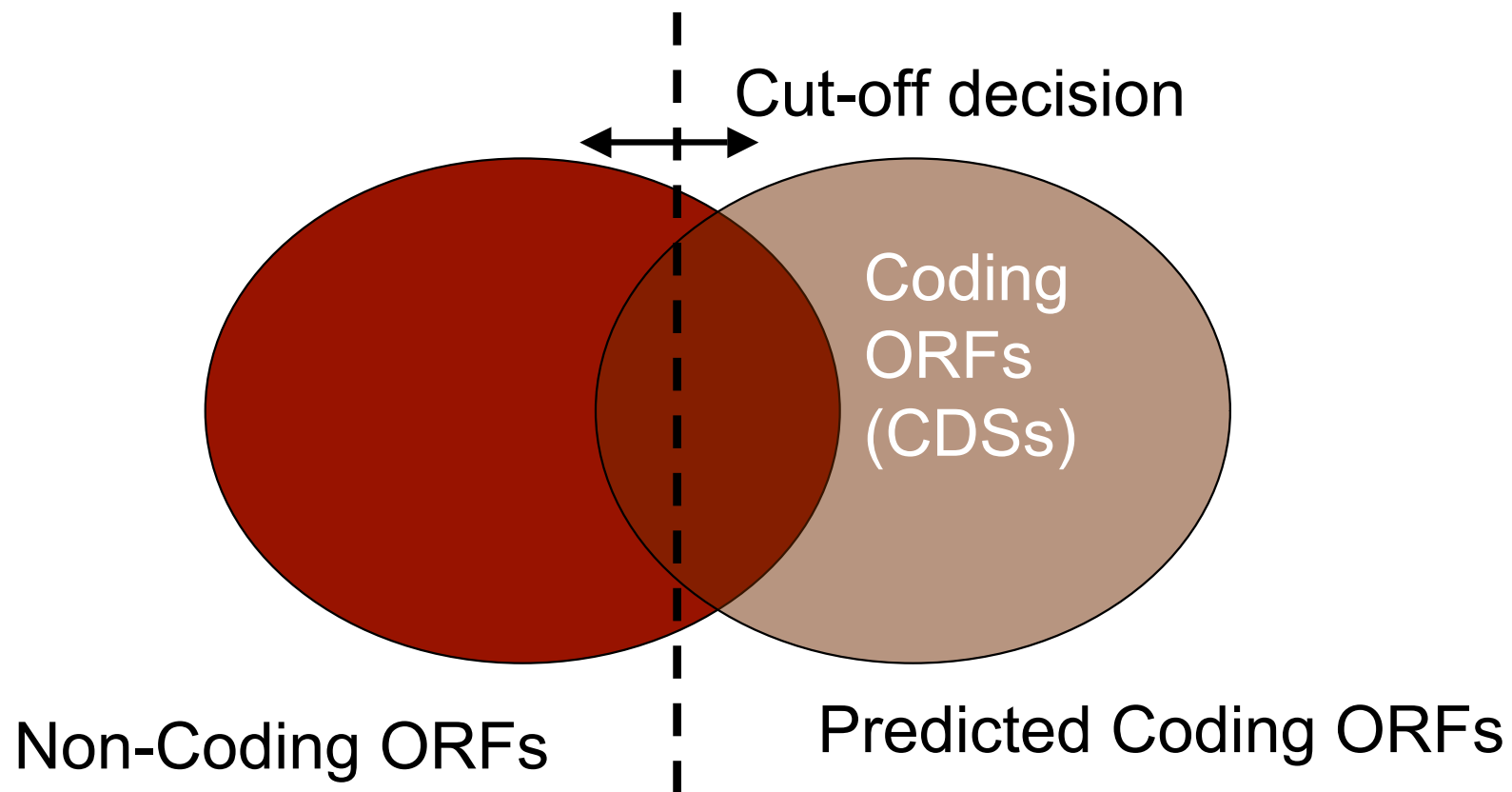
Extrinsic Methods (CDSs)

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





Class overlap



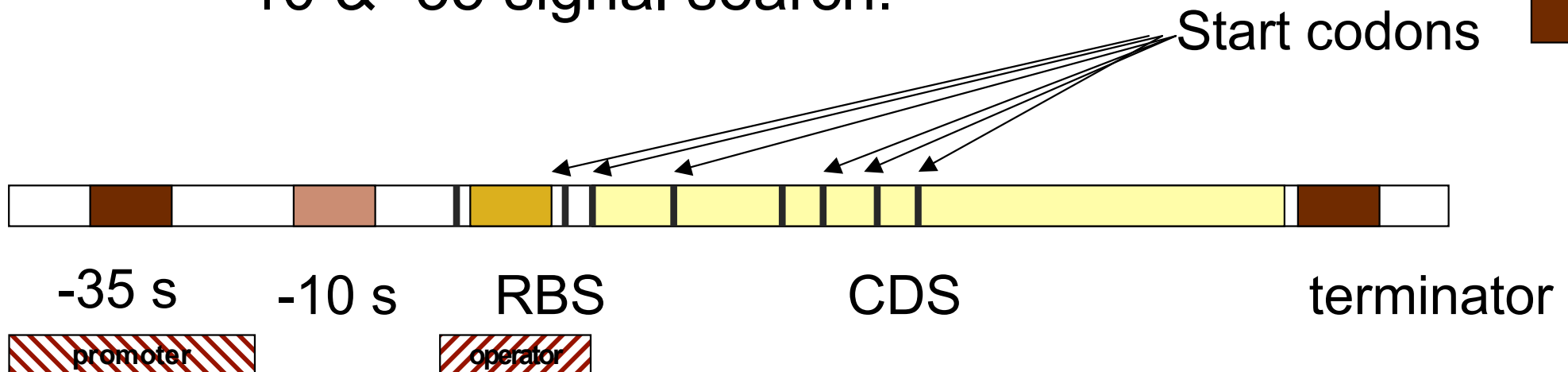
Glimmer: Up to 18% false pos.





Start codon issue

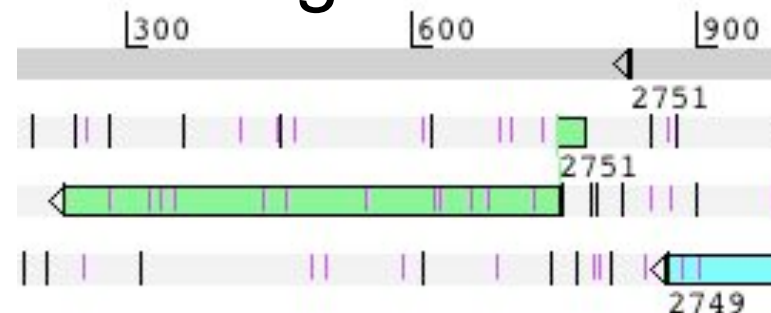
- Difficult problem.
- Integrate up-stream signal search.
 - RBS.
 - -10 & -35 signal search.





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

AGCTTTTCATTCTGACTGCAACGGGCAATATGTCTCTGTGTGGATTAAAAAAGAGTGTCTGATAGCAGCTTCTGAACTGGTTACCTGCCGTGAGTAAAT
TAAAAATTTTATTGACTTAGGTCACTAAATACTTTTAACCAATATAGGCATAGCGCACAGACAGATAAAAAATTACAGAGTACACAACATCCATGAAACGCAT
TAGCACCACCATTACCACCACCATCACCATTACCACAGGTAACGGTGCAGGCTGACGCGTACAGGAAACACAGAAAAAGCCCGCACCTGACAGTGCGGG
CTTTTTTTTTTCGACCAAAGGTAACGAGGTAACAACCATGCGAGTGTGGAAGTTCGGCGGTACATCAGTGGCAAATGCAGAACGTTTTCTGCGTGTGCGG
ATATTCTGGAAAGCAATGCCAGGCAGGGGAGGTGGCCACCGTCCTCTCTGCCCCGCCAAAATCACCAACCACCTGGTGGCGATGATTGAAAAAACCAT
TAGCGGCCAGGATGCTTTACCCAATATCAGCGATGCCGAACGTATTTTTTGCCGAACCTTTTGACGGGACTCGCCGCCGCCAGCCGGGGTTCCCGCTGGCG
CAATTGAAAACCTTCGTCGATCAGGAATTTGCCCAAATAAAACATGTCCTGCATGGCATTAGTTTGTGTTGGGGCAGTGCCCGGATAGCATCAACGCTGCGC
TGATTTGCCGTGGCGAGAAAATGTCGATCGCCATTATGGCCGGCGTATTAGAAGCGCGCGGTACACAACGTTACTGTTATCGATCCGGTCGAAAACTGCT
GGCAGTGGGGCATTACCTCGAATCTACCGTCGATATTGCTGAGTCCACCCGCCGTATTGCGGCAAGCCGCATTCCGGCTGATCACATGGTGCTGATGGCA
GGTTTCACCGCCGGTAATGAAAAAGGCGAACTGGTGGTGCTTGGACGCAACGGTTCGACTACTCTGCTGCGGTGCTGGCTGCCTGTTTACGCGCCGATT
GTTGCGAGATTTGGACGGACGTTGACGGGGTCTATACCTGCGACCCGCGTCAGGTGCCCCGATGCGAGGTTGTTGAAGTCGATGTCCTACCAGGAAGCGAT
GGAGCTTTCCTACTTCGGCGCTAAAGTTCTTCACCCCCGCACCATTACCCCCATCGCCCAGTTCAGATCCCTTGCTGATTAAAAATACCGGAAATCCT
CAAGCACCAGGTACGCTCATTGGTGCCAGCCGTGATGAAGACGAATTACCGGTCAAGGGCATTTCGAATCTGAATAACATCCTGCATGGCATTAGTTTGT
TGGGGCAGTGCCCGGATAGCATCAACGCTGCGCTGATTGCGGTGGCGAGAAAATGTCGATCGCCATTATGGCCGGCGTATTAGAAGCGCGCGGTACAAA
CGTTACTGTTATCGATCCGGTCGAAAAACTGCTGGCAGTGGGGCATTACCTCGAATCTACCGTCGATATTGCTGAGTCCACCCGCCGTATTGCGGCAAGC
CGCATTCGGGTGATCACATGGTGCTGATGGCAGGTTTCACCGCCGGTAATGAAAAAGGCGAACTGGTGGTGCTTGGACGCAACGGTTCCGACTACTCTG
CTGCGGTGCTGGCTGCCTGTTTACGCGCCGATTGTTGCGAGATTTGGACGGACGTTGACGGGGTCTATACCTGCGACCCGCGTCAGGTGCCCGATGCGAG
GTTGTTGAAGTCGATGTCCTACCAGGAAGCGATGGAGCTTTCCTACTTCGGCGCTAAAGTTCTTCACCCCCGCACCATTACCCCCATCGCCCAGTTCCAG
ATCCCTTGCTGATTAAAAATACCGGAAATCCTCAAGCACCAGGTACGCTCATTGGTGCCAGCCGTGATGAAGACGAATTACCGGTCAAGGGCATTTCGA
ATCTGAATAACATCCTGCATGGCATTAGTTTGTGGGGCAGTGCCCGGATAGCATCAACGCTGCGCTGATTTGCCGTGGCGAGAAAATGTCGATCGCCAT
TATGGCCGGCGTATTAGAAGCGCGCGGTACACAACGTTACTGTTATCGATCCGGTCGAAAAACTGCTGGCAGTGGGGCATTACCTCGAATCTACCGTCGAT
ATTGCTGAGTCCACCCGCCGTATTGCGGCAAGCCGCATTCCGGCTGATCACATGGTGCTGATGGCAGGTTTCACCGCCGGTAATGAAAAAGGCGAACTGG
TGGTGCTTGGACGCAACGGTTCCGACTACTCTGCTGCGGTGCTGGCTGCCTGTTTACGCGCCGATTGTTGCGAGATTTGGACGGACGTTGACGGGGTCTA
TACCTGCGACCCGCGTCAGGTGCCCGATGCGAGGTTGTTGAAGTCGATGTCCTACCAGGAAGCGATGGAGCTTTCCTACTTCGGCGCTAAAGTTCTTCAC
CCCCGCACCATTACCCCCATCGCCCAGTTCAGATCCCTTGCTGATTAAAAATACCGGAAATCCTCAAGCACCAGGTACGCTCATTGGTGCCAGCCGTG
ATGAAGACGAATTACCGGTCAAGGGCATTTCGAATCTGAATAACATCCTGCATGGCATTAGTTTGTGGGGCAGTGCCCGGATAGCATCAACGCTGCGCT
GATTTGCCGTGGCGAGAAAATGTCGATCGCCATTATGGCCGGCGTATTAGAAGCGCGCGGTACACAACGTTACTGTTATCGATCCGGTCGAAAAACTGCTG
GCAGTGGGGCATTACCTCGAATCTACCGTCGATATTGCTGAGTCCACCCGCCGTATTGCGGCAAGCCGCATTCCGGCTGATCACATGGTGCTGATGGCAG
GTTTCACCGCCGGTAATGAAAAAGGCGAACTGGTGGTGCTTGGACGCAACGGTTCCGACTACTCTGCTGCGGTGCTGGCTGCCTGTTTACGCGCCGATTG
TTGCGAGATTTGGACGGACGTTGACGGGGTCTATACCTGCGACCCGCGTCAGGTGCCCGATGCGAGGTTGTTGAAGTCGATGTCCTACCAGGAAGCGATG
GAGCTTTCCTACTTCGGCGCTAAAGTTCTTCACCCCCGCACCATTACCCCCATCGCCCAGTTCAGATCCCTTGCTGATTAAAAATACCGGAAATCCTC
AAGCACCAGGTACGCTCATTGGTGCCAGCCGTGATGAAGACGAATTACCGGTCAAGGGCATTTCGAATCTGAATAACAGATCACATGGTGCTGATGGCAG

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



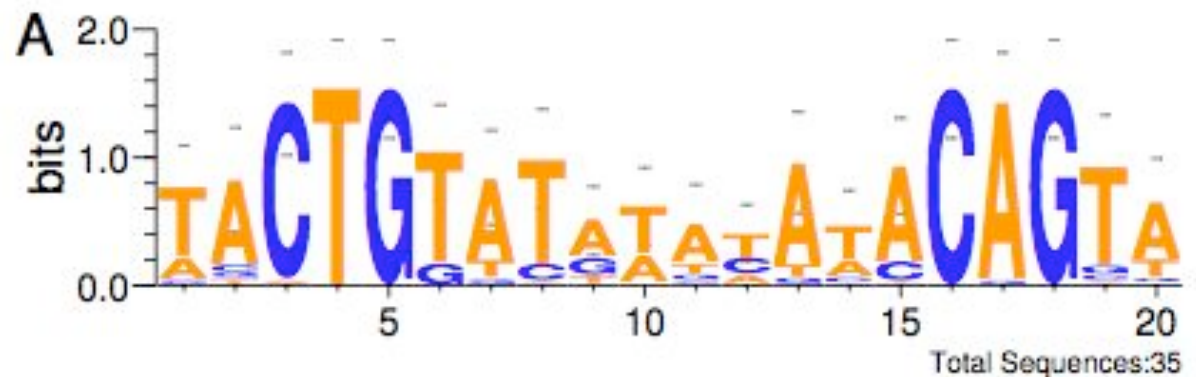
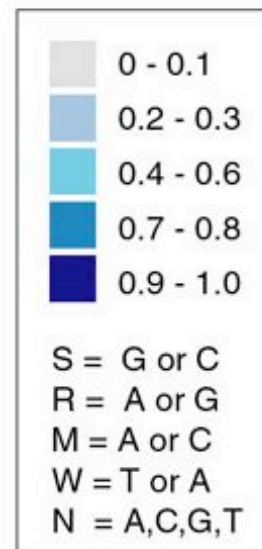
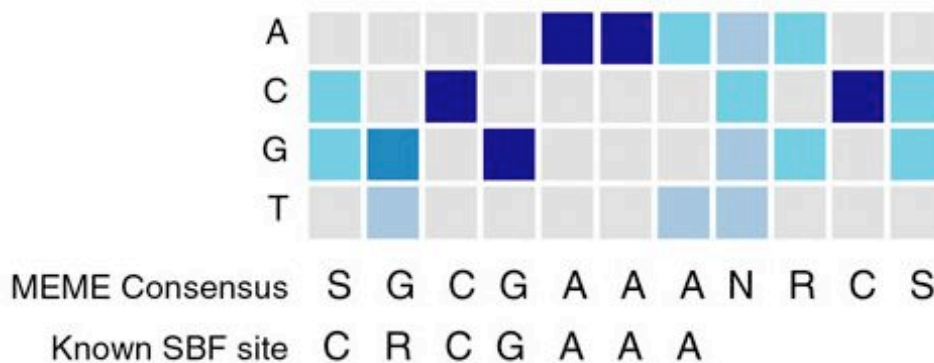
CDSs to Functions

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



Position Weighted Matrix

A.





GO Example

MAP kinase activity

Accession:GO:0004707

Synonyms:

MAPK

mitogen activated kinase

Definition:

Catalysis of the phosphorylation of proteins. Mitogen-activated protein kinase; a family of protein kinases that perform a crucial step in relaying signals from the plasma membrane to the nucleus. They are activated by a wide range of proliferation- or differentiation-inducing signals; activation is strong with agonists such as polypeptide growth factors and tumor-promoting phorbol esters, but weak (in most cell backgrounds) by stress stimuli.

Term Lineage

[Graph view.](#)

GO:0003673 : Gene Ontology (146200)

④ GO:0003674 : molecular function (97507)

① GO:0003824 : catalytic activity (32256)

① GO:0016301 : kinase activity (5076)

① GO:0004672 : protein kinase activity (3030)

① GO:0004674 : protein serine/threonine kinase activity (1854)

① GO:0004702 : receptor signaling protein serine/threonine kinase activity (257)

① **GO:0004707 : MAP kinase activity (90)**

① GO:0016740 : transferase activity (9890)

① GO:0016772 : transferase activity, transferring phosphorus-containing groups (5162)

① GO:0016773 : phosphotransferase activity, alcohol group as acceptor (3682)

① GO:0004672 : protein kinase activity (3030)

① GO:0004674 : protein serine/threonine kinase activity (1854)

① GO:0004702 : receptor signaling protein serine/threonine kinase activity (257)

① **GO:0004707 : MAP kinase activity (90)**

① GO:0004871 : signal transducer activity (6386)

① GO:0005057 : receptor signaling protein activity (598)

① GO:0004702 : receptor signaling protein serine/threonine kinase activity (257)

① **GO:0004707 : MAP kinase activity (90)**





Observations

Region C76 (105 of 227 Observations visible)								
Contigs	Observations	Search	Statistics					
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 m
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 f
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 fla
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	qil52211202 emb CAH37191.1	121 - 1425	-	-	putative cytochrome C biogen





Annotations

GENDB Annotation Dialog: C76

Annotator	Latest	Date
pmoreno_user	Function	Wed Jul
jugalde		Wed Jul
Assign_Names		Mon Jul
Assign_Names		Fri Jun :
Assign_Names		Mon Jun
Assign_Cogs		Sat May
REGANOR26	Region	Mon Ap

Annotation: GENDB::DB::Annotation::Function::CDS

Annotator: pmoreno_user

Date: Wed Jul 26 16:06:23 2006

Comment: Gene name assigned.

Description: ResB protein, is required for the biosynthesis of cytochrome C.

Observations: cytochrome c-type biogenesis protein ResB, putative [Cyanobacteria bacterium
cytochrome c-type biogenesis protein ResB, putative [org:Acidithiobacillus ferro

Links:

Gene Name: resB

GO Numbers:

Function: Protein required for cytochrome C assembly

Confidence: specificity unclear

Evidence:

EC Number:

COG Number: COG1333

Dismiss



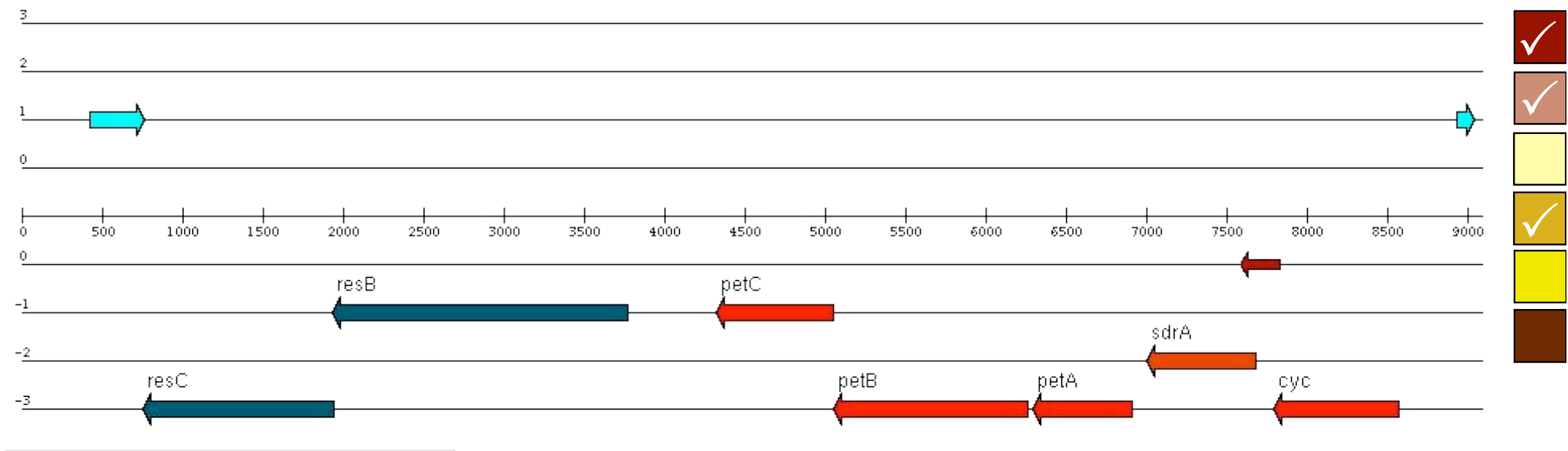
>gi|16127994|ref|NC_000913.1| Escherichia coli K12, complete genome

AGCTTTTCATTCTGACTGCAACGGGCAATATGTCTCTGTGTGGATTAAAAAAGAGTGTCTGATAGCAGCTTCTGAACTGGTTACCTGCCGTGAGTAAAT
TAAAATTTT**TATTG**ACTTAGGTCACATAAATACT**TTTAACCAATATAGGCATAGCGCA****CAGACAGATAAAAAATTACAGAGTACACAACATCCATGAAACGCAT**
TAGCACCACCATTACCACCACCATCACCATTACCACAGGTAACGGTGCGGGCTGACGCGTACAGGAAACACAGAAAAAGCCCGCACCTGACAGTGCGGG
CTTTTTTTTTTCGACCAAAGGTAACGAGGTAACAACCATGCGAGTGTTGAAGTTCGGCGGTACATCAGTGGCAAATGCAGAACGTTTTCTGCGTGTTGCCG
ATATTCTGGAAAGCAATGCCAGGCAGGGGAGGTGGCCACCGTCCTCTCTGCCCCGCCAAAATCACCAACCACCTGGTGGCGATGATTGAAAAAACCAT
TAGCGGGCCAGGATGCTTTACCCAATATC**AGCGAT**GCCGAACGTATTTTTTGCCGAACTTTTGACGGGACTCGCCGCCGCCAGCCGGGGTTCCCGCTGGCG
CAATTGAAAACCTTCCTCGATCAGGAATTTGCCCAAATAAAACATGTCCTGTCATGGCATTAGTTTGTGGGGCAGTGCCCGGATAGCATCAACGCTGCGC
TGATTTGCCGTGGCGAGAAAATGTCGATCGCCATTATGGCCGGCGTATTAGAAGCGCGCGGTACACAACGTTACTGTTATCGATCCGGTCGAAAACTGCT
GGCAGTGGGGCATTACCTCGAATCTACCGTCGATATTGCTGAGTCCAC**CCGCCGTATTG**CGGCAAGCCGCATTCCGGCTGATCACATGGTGCTGATGGCA
GGTTTCACCGCCGGTAATGAAAAAGGCGAACTGGTGGTGCTTGGACGCAACGGTTCGACTACTCTGCTGCGGTGCTGGCTGCCTGTTTACGCGCCGATT
GTTG**CGAGATTTGGACGGACGTTGACGGGGTCTATACCTGCGACCCGCGTCAGGTGCCCGATGCGAGGTTGTTGAAGTCGATGTCCTACCAGGAAGCGAT**
GGAGCTTTCCTACTTCGGCGCTAAAGTTCTTCACCCCCGCACCATTACCCCCATCGCCAGTTCAGATCCCTTGCTGATTAAAAATACCGGAAATCCT
CAAGCACCAGGTACGCTCATTGGTGCCAGCCGTGATGAAGACGAATTACCGGTCAAGGGCATTTCGAATCTGAATAACATCCTGCATGGCATTAGTTTGT
TGGGGCAGTGCCCGGATAGCATCAACGCTGCGCT**GATT****TGCC**GTGGCGAGAAAATGTCGATCGCCATTATGGCCGGCGTATTAGAAGCGCGCGGTACAAA
CGTTACTGTTATCGATCCGGTCGAAAACTGCTGGCAGTGGGGCATTACCTCGAATCTACCGTCGATATTGCTGAGTCCACCCGCCGTATTGCGGCAAGC
CGCAT**TCCGGCTGATCACATGGTGCTGATGGCAGGTTTCACCGCCGGTAATGAAAAAGGCGAACTGGTGGTGCTTGGACGCAACGGTTCGACTACTCTG**
CTGCGGTGCTGGCTGCCTGTTTACGCGCCGATTGTTGCGAGATTTGGACGGACGTTGACGGGGTCTATACCTGCGACCCGCGTCAGGTGCCCGATGCGAG
GTTGTTGAAGTCGATGTCCTACCAGGAAGCGATGGAGCTTTCCTACTT**CGGCGCTAAAGTTCTTCACCCCCGCACCATTACCCCCATCGCCAGTTCCAG**
AT**CCCTTGCTGATTAAAAATACCGGAAATCCTCAAGCACCAGGTACGCTCATTGGTGCCAGCCGTGATGAAGACGAATTACCGGTCAAGGGCATTTC**CA
ATCTGAATAACATCCTGCATGGCATTAGTTTGTGGGGCAGT**GCCCGGATAGCATCAACGCTGCGCTGATTTGCCGTGGCGAGAAAATGTCGATCGCCAT**
TATGGCCGGCGTATTAGAAGCGCGCGGTACACAACGTTACTGTTATCGATCCGGTCGAAAACTGCTGGCAGTGGGGCATTACCTCGAATCTACCGTCGAT
ATTGCTGAGTCCACCCGCCGTATTGCGGCAAGCCGCATTCCGGCTGATCACATGGTGCTGATGGCAGGTTTCACCGCCGGTAATGAAAAAGGCGAACTGG
TGGTGCTTGGACGCAACGGTTCGACTACTCTGCTGCGGTGCTGGCTGCCTGTTTACGCGCCGATTGTTGCGAGATTTGGACGGACGTTGACGGGGTCTA
TACCTGCGACCCGCGTCAGGTGCCCGATGCGAGGTTGTTGAAGTCGATGTCCTACCAGGAAGCGATGGAGCTTTCCTACTT**CGGCGCTAAAGTTCTTCAC**
CCCCGCACCATTACCCCCATCGCCAGTTCAGATCCCTTGCTGATTAAAAATACCGGAAATCCTCAAGCACCAGGTACGCTCATTGGTGCCAGCCGTG
ATGAAGACGAATTACCGGTCAAGGGCATTTCGAATCTGAATAACATCCTGCATGGCATTAGTTTGTGGGGCAGTGCCCGGATAGCATCAACGCTGCGCT
GATTTGCCGTGGCGAGAAAATGTCGATCGCCATTATGGCCGGCGTATTAGAAGCGCGCGGTACACAACGTTACTGTTATCGATCCGGTCGAAAACTGCTG
GCAGT**GGGGCATTACCTCGAATCTACCGTCGATATTGCTGAGTCCACCCGCCGTATTGCGGCAAGCCGCATTCCGGCTGATCACATGGTGCTGATGGCAG**
GTTTCACCGCCGGTAATGAAAAAGGCGAACTGGTGGTGCTTGGACGCAACGGTTCGACTACTCTGCTGCGGTGCTGGCTGCCTGTTTACGCGCCGATTG
TTGCGAGATTTGGACGGACGTTGACGGGGTCTATACCTGCGACCCGCGTCAGGTGCCCGATGCGAGGTTGTTGAAGTCGATGTCCTACCAGGAAGCGATG
GAGCTTTCCTACTT**CGGCGCTAAAGTTCTTCACCCCCGCACCATTACCCCCATCGCCAGTTCAGATCCCTTGCTGATTAAAAATACCGGAAATCCTC**
AAGCACCAGGTACGCTCATTGGTGCCAGCCGTGATGAAGACGAATTACCGGTCAAGGGCATTTCGAATCTGAATAACAGATCACATGGTGCTGATGGCAG

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

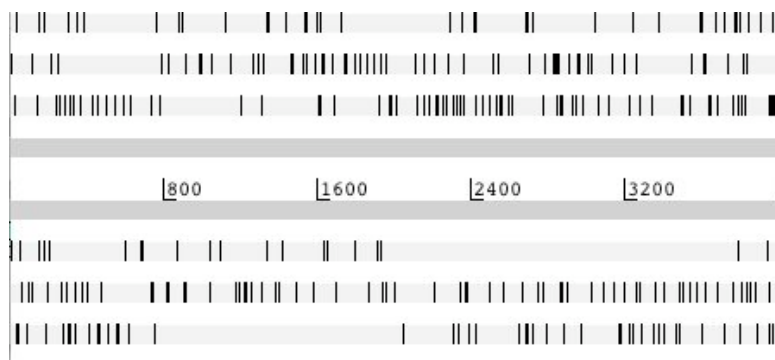


Real genome view

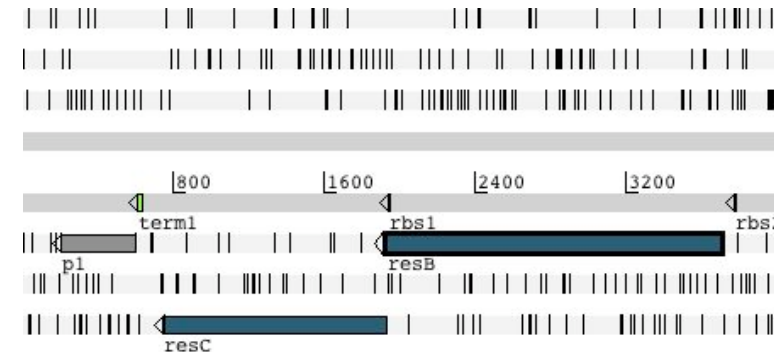


Environments & Pipelines

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



Several tools
Manual curation





Annotation Pipeline

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





Systems Biology

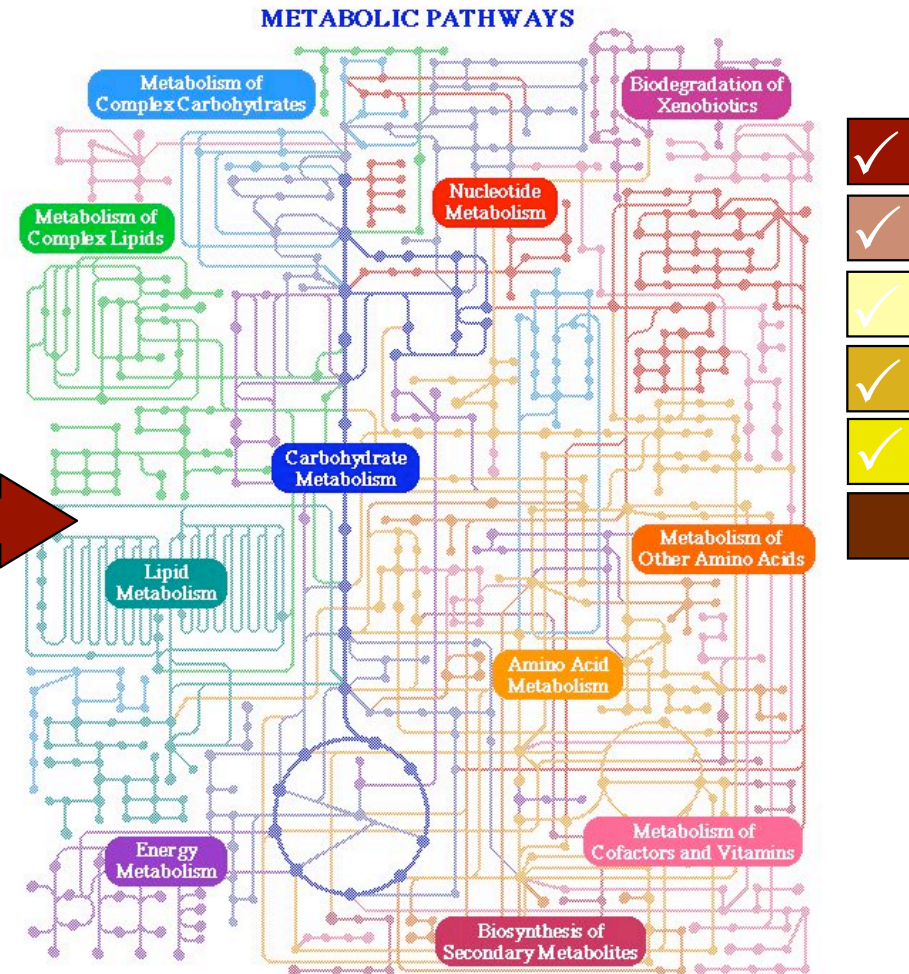
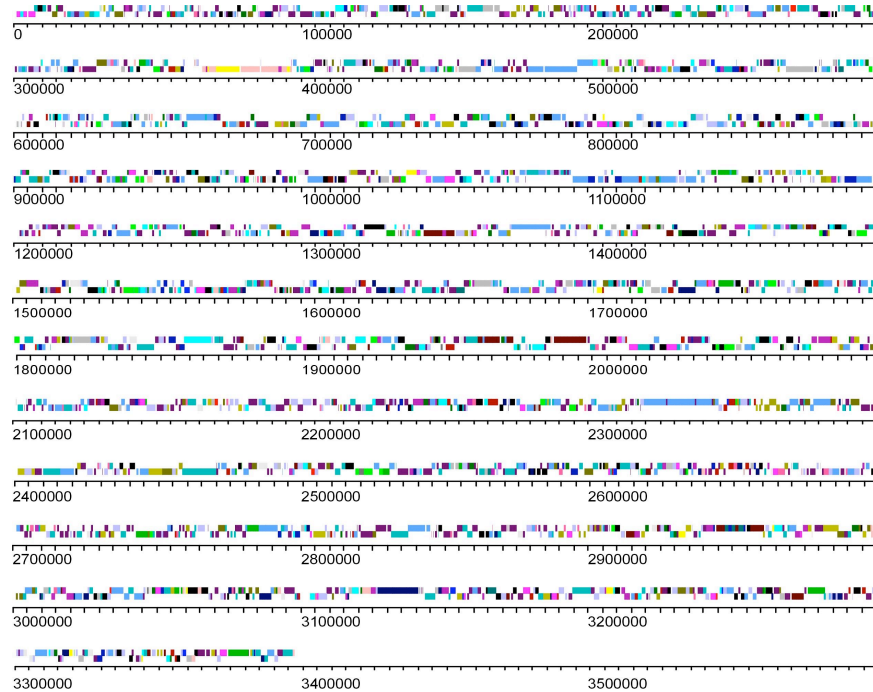


Systems Biology

- *“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*

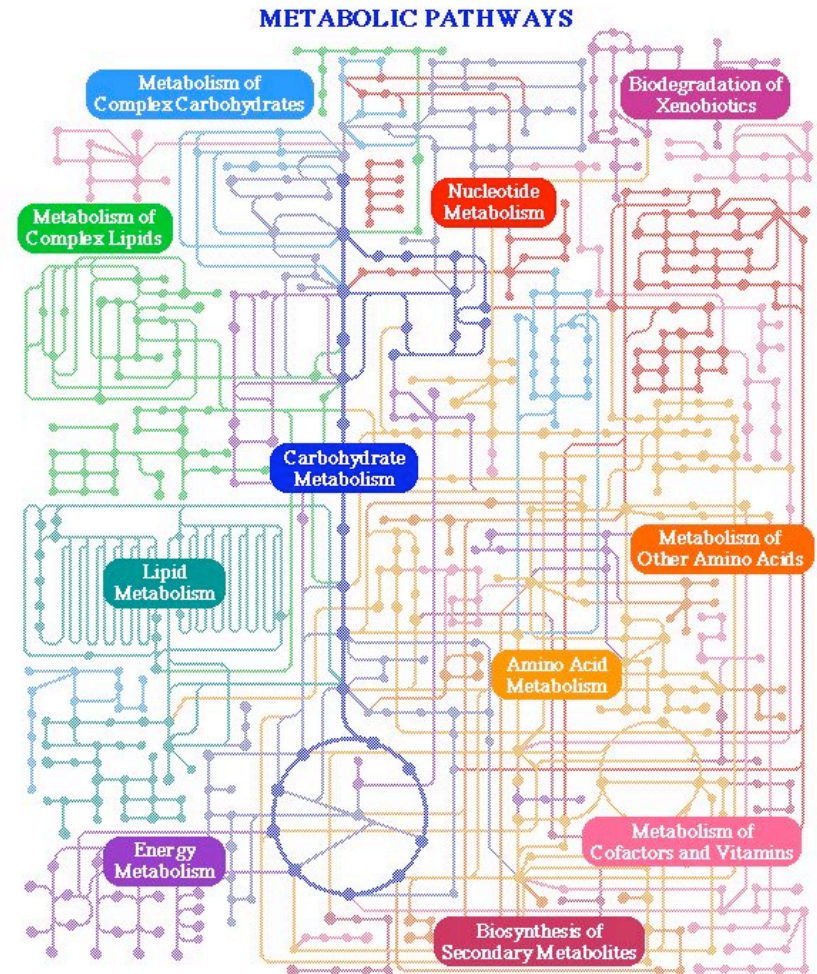


Gene Reg. & Metabolic Pathways



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.

