BIOC4004 - Industrial Biochemistry
Lecture 15 - Mon Mar 01, 04

Topics for the Day:

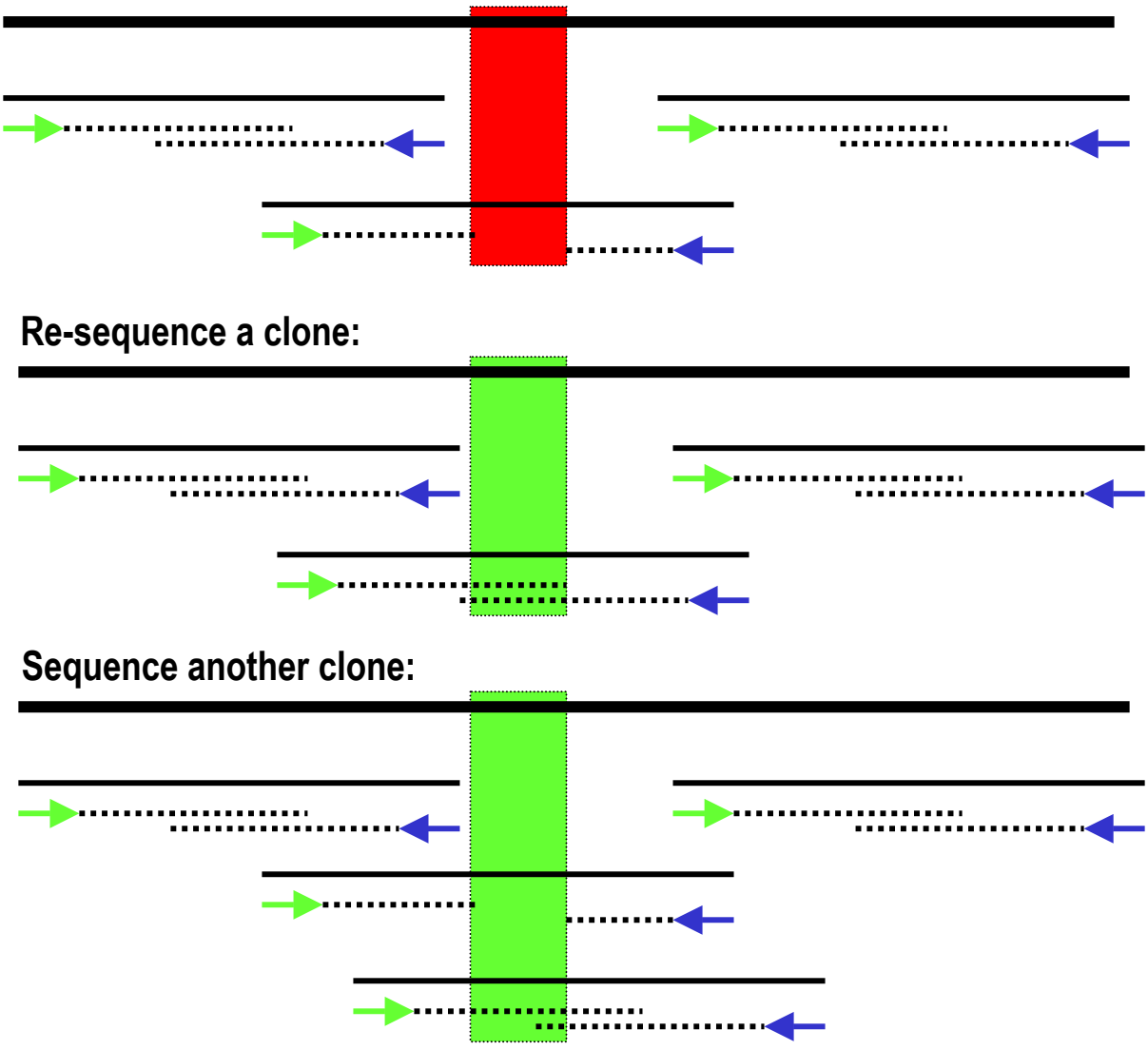
- Genome projects
  - gaps and scaffolds
  - analysis and genome annotation
  - Bioinformatics assignment

Physical Gaps vs. Sequencing Gaps (Pt.I)

Sequencing Gaps (ie. "Fake Gaps")

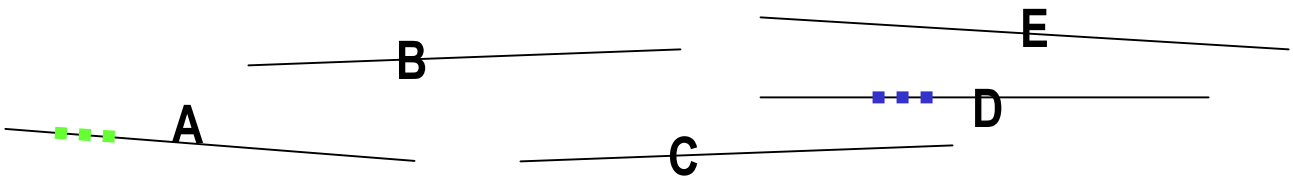
• Sequencing gaps result from non-overlap of sequencing runs which **should** overlap

• If there are no physical gaps in your library and you still have gaps after sequence assembly, there **must** be sequencing gaps



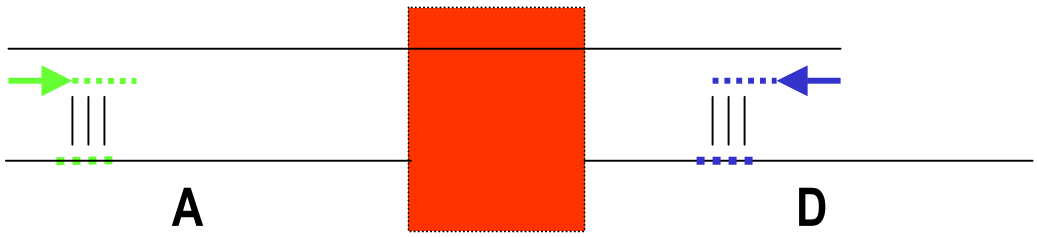
• Sequencing gaps are a nuisance, but can be dealt with because at least the gaps are "contained" in the library (**you know all the pieces in the puzzle are there**)

Scaffolding using cosmid/fosmids (or "dealing with Physical Gaps")

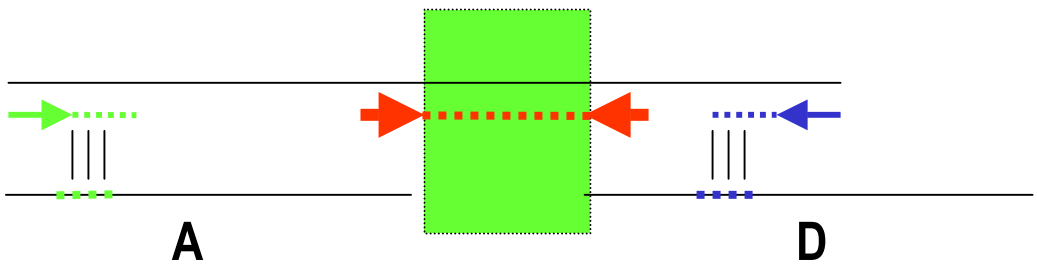


- contigs are not physical entities, they are "constructs" made from sequence data
- clone inserts are actual physical entities

Cosmid Insert



- if cosmid end-sequences match regions in two different contigs, the contigs **have to be adjacent to each other**

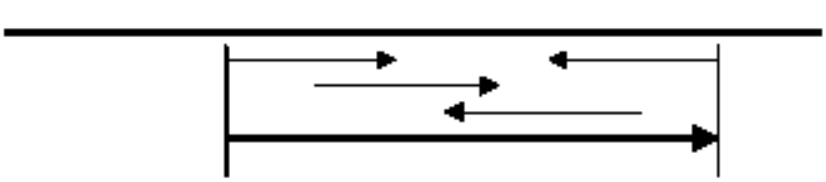


new primer flanking gap

- To close the gap, make new primers, primer walking off the cosmid

- Gap closure (physical or sequencing gaps) is always conceptually the same:
  - **Sequencing Gaps:**
    - contained within clones in existing library
  - **Physical Gaps:**
    - contained within a clone from a different library
    - people make extra libraries to cover all the bases

Contig Assembly




```
******************** Contig 1 *********************
                .    :    .    :    .    :    .    :    .    :    .    :
F1+          ATGGATGCAATACTGAATTACAGGTCAGAAGATACTGAAGATTACTACACATTACTGGGA
F2+                           ACAGGTCAGAAGATACTGAAGATTACTACACATTACTGG-A
F3+                                                GATTACTACACATTACTGGGA
F4-                                                     CATTACTGGGA
F5-                                                                  A
                ____________________________________________________________
consensus    ATGGATGCAATACTGAATTACAGGTCAGAAGATACTGAAGATTACTACACATTACTGGGA

                .    :    .    :    .    :    .    :    .    :    .    :
F1+          TGTGATGAACTATCTTCGGTTGAACAAAACCTGGGCAGA
F2+          TGTGATGAACTATCTTCGGTTGAACAAATCCTGG-CAGAAT
F3+          TGTGATGAACTATCTTCGCTTGAACAAATCCTGG-CAGAATTTAAAGTCAGAGCTCTGGA
F4-          TGTGATAAACTATCTTCGGTTGAACAAATCCTGG-CAGAATTTAAAGTCAGAGCTCTGGA
F5-          TGTGATGAACTATCTTCGGTTGAACAAATCCTGG-CAGAATTTAAAGT-AGAGCTCTGGA
                ____________________________________________________________
consensus    TGTGATGAACTATCTTCGGTTGAACAAATCCTGG-CAGAATTTAAAGTCAGAGCTCTGGA

                .    :    .    :    .    :    .    :    .    :    .    :
F3+          A
F4-          ATGTCACCCAGACAAGCATCC
F5-          ATGTCACCCAGACAAGCATCCTGAAAACCCC
                ____________________________________________________________
consensus    ATGTCACCCAGACAAGCATCCTGAAAACCCC

>Contig1
ATGGATGCAATACTGAATTACAGGTCAGAAGATACTGAAGATTACTACACATTACTGGGA
TGTGATGAACTATCTTCGGTTGAACAAATCCTGGCAGAATTTAAAGTCAGAGCTCTGGA
ATGTCACCCAGACAAGCATCCTGAAAACCCC
```

- Used to piece together sequence from different subclones
- Specialized programs for assembly of shotgun clones
  - CAP, PHRAP

Aligning Coding Sequences

```
gaccagctgcag
||||  |
gacctccaa

D   Q   L   Q
gac,cag,ctg,cag
|||  |||  |||
gac,---,ctc,caa
D   -   L   Q
```

```
gaccagctgcag
||||  |  ||
gacct-c--caa
```

Final assembly must take coding sequences into consideration

PHRED and PHRAP and contig assembly

Phred Score	Error probability	Likelihood that the base is correct
10	1:10	90%
20	1:100	99%
30	1:1,000	99.9%
40	1:10,000	99.99%
50	1:100,000	99.999%
60	1:1,000,000	99.999%

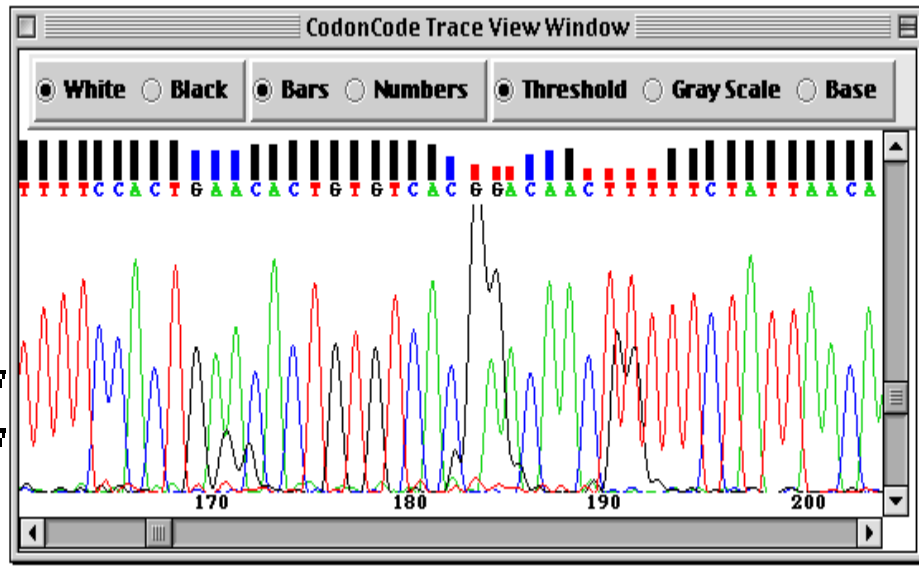
Black: q≥30 (99.9%)

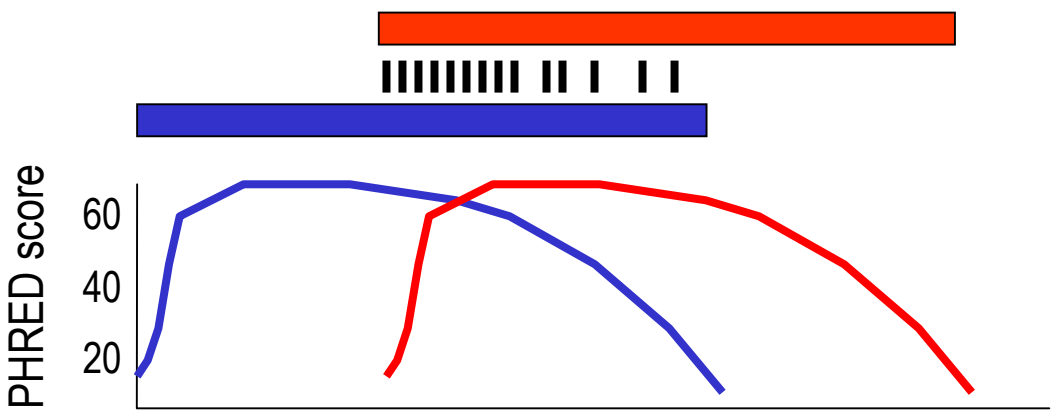
Blue: q≥20 (99%)

Red: q<20 (<99%)

Tall = high quality

Short = low quality





Annotation of genome sequence (Pt. I)

The Sequence is only the beginning:

…A house may be built out of stones…
…A pile of stones is not a house…(H. Poincare)

- At 90-95% genome coverage you will already be able to identify > 99% of the genes
- "Genome Light" concept:
    - many sequencing projects are not aiming to get the sequence down to the last nucleotide (or one single contig)
    - 2-5 X coverage will allow you to discover most of the genes anyway
    - If 10% of the effort will get you 90% of the sequence, do only 10% of the work !!!!
        - (many companies do this)
    - If you only have partial sequence for a gene of interest, you can obtain the remaining bits through other means (e.g. PCR)
- Finished Sequence:
    - no gaps
    - less than 1/10,000 error rate (0.01%)
    - sequence on **BOTH** strands
    - assembly tested by restriction digests

Annotation of genome sequence (Pt. II)

• Annotation does not have to wait until you have completed the genome sequence

Aims of Annotation:

• **identification of all putative genes**

• assigning putative function to genes
  • based on the literature
  • based on Dbase homology
  • based on matches to known structural motifs

• identification of potential regulatory sequences in genes

• identification of interesting sequence features
  • gene families
  • mobile genetic elements
  • sequence repeats

• determination of genome organization
• determination of metabolic or regulatory networks

Annotation Stage 1. Gene Prediction (Part I)

- Identification of open reading frames using standard methods
  - – ORF scanning looking for > 50 AA translation products on all reading frames
    - based on ATG and Stop codons
      - sequence composition problem:
        - more ATGs in AT-rich genomes
    - longer ORFs likely to be real genes
    - eukaryotic problem: introns break coding sequences up
  - – Use codon-bias to help prediction
    - coding regions will exhibit codon bias, non-coding will not
    - codon bias affects base-composition
  - – Look for sequence motifs:
    - promoter sequences & high densities of TF binding sites
    - CpG islands (housekeeping genes)
    - intron/exon splice sites
    - poly-A signals
    - Translation signals (Shine-Delgarno, Kozak)

Annotation Stage 1. Gene Prediction (Part II)

• Gene prediction in prokaryotes is “relatively” straightforward because of lack of introns.

– however AT-rich genomes can be tricky

– “lots of short ORFs”

• Eukaryotic gene prediction is still a big challenge.

– Exon prediction still a little iffy

– exons don’t necessarily start or end in complete codons

– Splicing sequences are not fully conserved, there are always far more "potential splice sites" than real ones

- can only figure real splice sites by comparing genomic sequence to cDNA sequence

– Promoter predictions are equally dicey

– Which ATG is used ? Not always the first one…

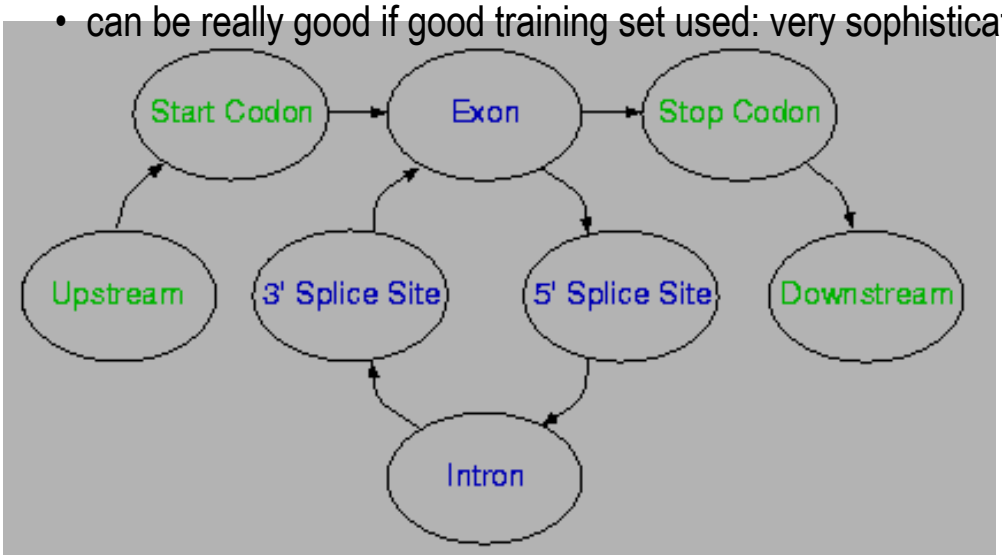
Annotation Stage 1. Gene Prediction (Part III)

Hidden Markov Modeling (HMM)

• More sophisticated algorithms for finding genes

• Use a "training set" of **KNOWN** genes from organism

• HMM program trained on known sequences:
  • establish statistical rules to evaluate unknown parameters of genes (ie. model genes provide statistical distributions on codon bias, nucleotide frequencies, etc.)
  • looks for "hidden" rules that may not be immediately apparent
  • Once HMM program learns what a gene looks like, looks for sequences that fulfil these statistical rules
    • somewhat black-boxish
    • need good training set
    • can be really good if good training set used: very sophisticated modeling



A simple HMM for eukaryotic gene prediction

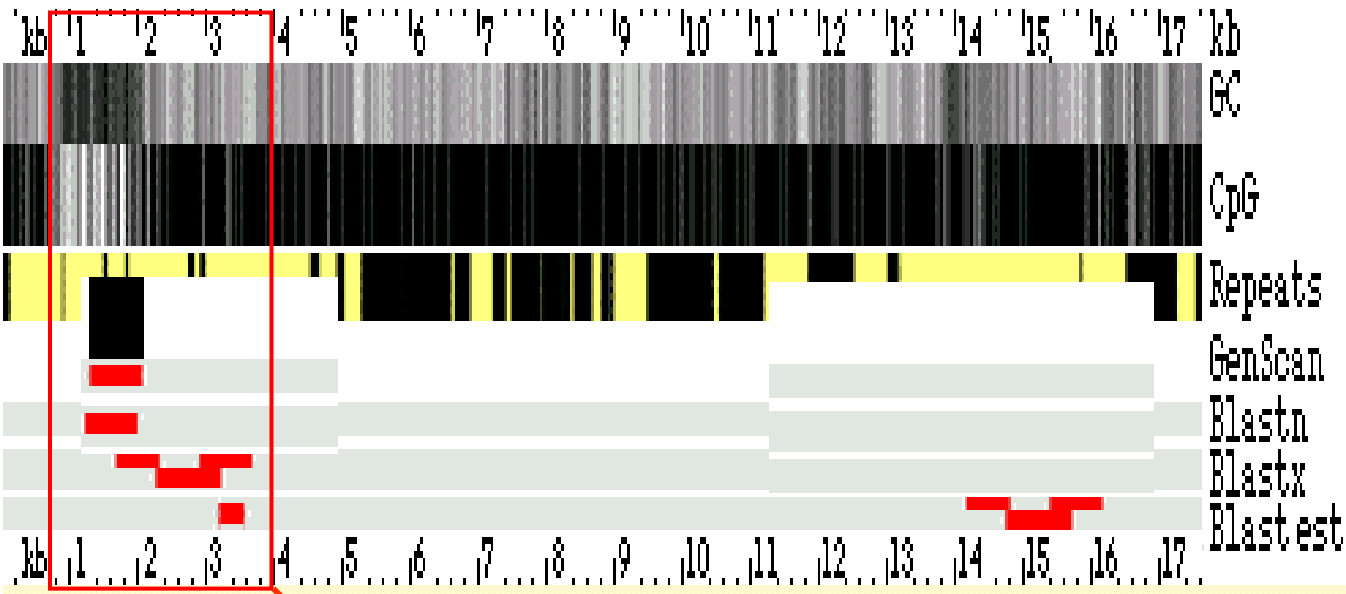
The best way to perform gene predictions is to use a combinatorial approach:
• HMM
• Similarity searches
• Motif finding

Annotation Stage 1. Gene Prediction (Part IV)

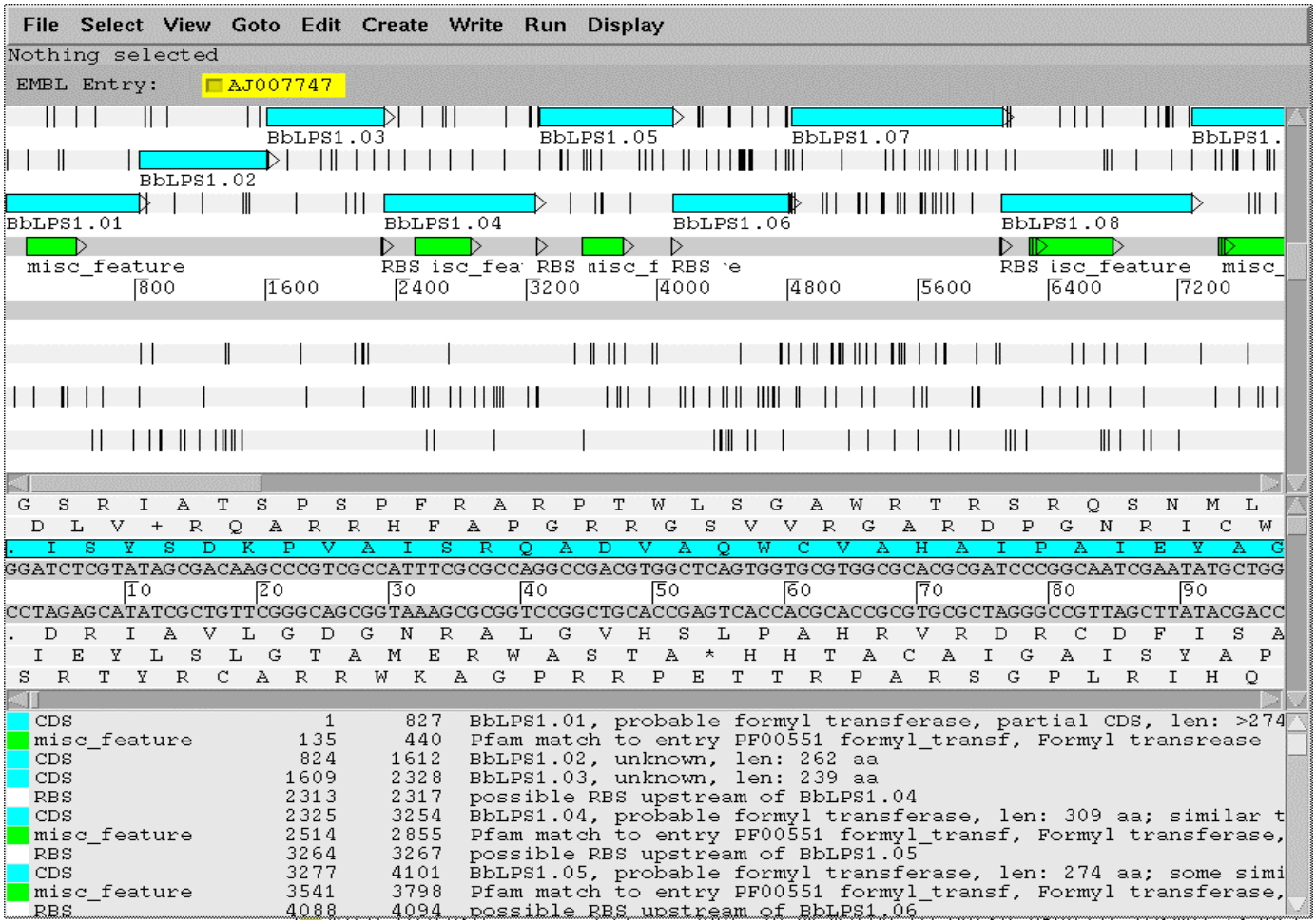
We are visual animals !!!!!

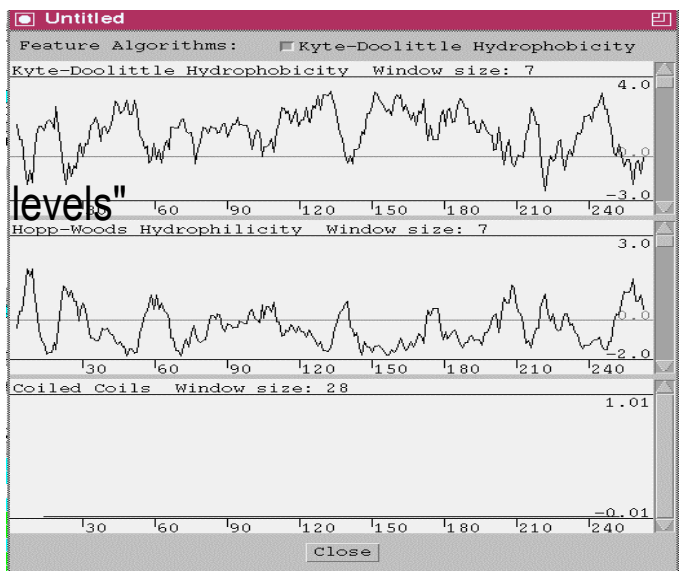
- The main idea of graphical annotation tools is assembling information from a large number of independent “lines of evidence” for gene presence as bands stacked in register

- Examples of "Bands"
  - **GC isochores**: areas with higher G+C content
    - can be a useful gene predictor (human genes are G+C rich)
  - CpG islands: found in many “housekeeping” promoters
  - RepeatMasker: look for sequence repeats
  - GenScan: ORF prediction based on HMM model
  - Blastn vs. NR Nucleic Acid database
  - Blastx vs NR Protein database
  - Blastn vs EST database



Sequence annotation using Artemis (The Sanger Centre)





- can view sequence at different "zoom levels"
- show ORFs
- searchable motifs
- BLAST search results
- Sequence calculations
  - Base composition
  - Hydrophobicity

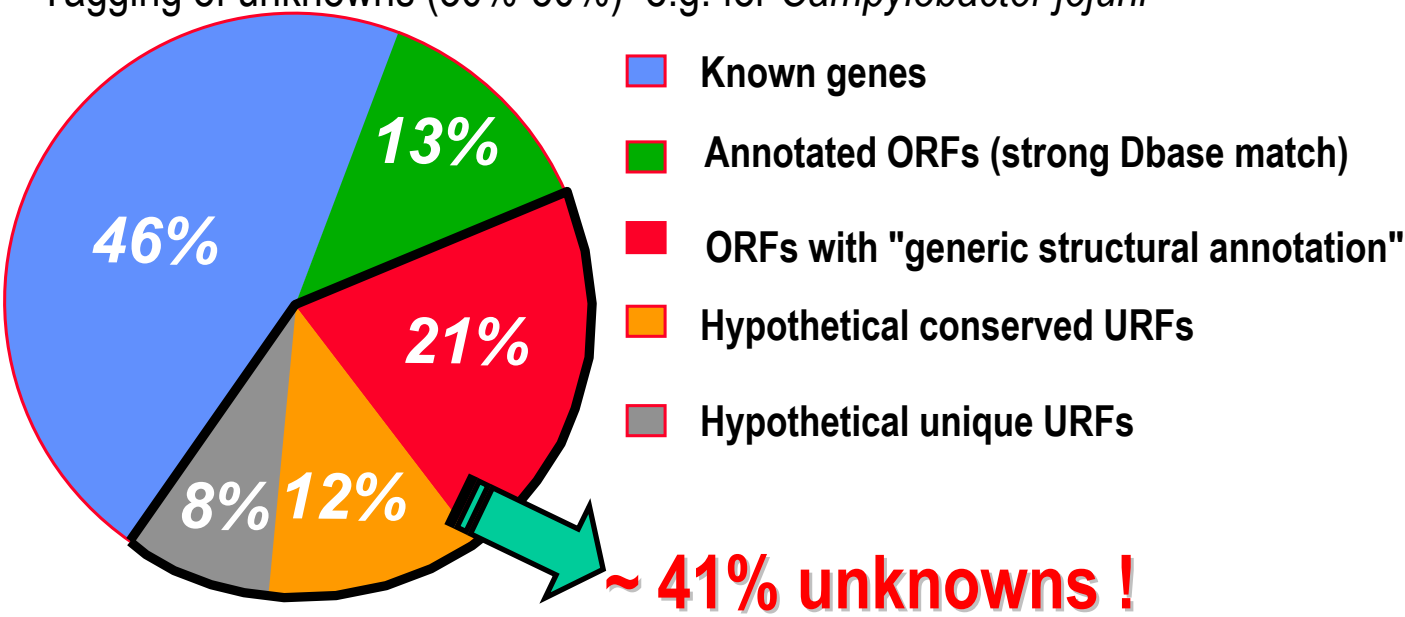
Annotation Stage 2. Gene Annotation

A human has to go through it, gene by gene (ie. It is hand curated)

1. **Identify candidate genes**
   - Exons, Introns, Promoters, etc.…
2. **Confirm and characterize candidate genes in order of declining quality of support: known genes, unknown but well-supported genes, possible genes, pseudogenes**
   - Homology searches (BLAST)
3. **Annotate using info from various sources**
   - Literature, Dbase matches
   - Similarity to known gene families (COGs, TIGR fams, Pfams)
   - "Structural Annotation"

- Avoid "**annotation drift**" from bad GenBank data

Tagging of unknowns (30%-50%) e.g. for *Campylobacter jejuni*



Public Human Genome project:

• Used Ensemble and Genie programs

• Merged that data with known sequences from various Dbases

Ensemble

• uses prediction of Genscan (an HMM program)

• checks these predictions against ESTs, mRNAs and protein motifs in known databases

• Ensemble predicted 35,500 genes

Genie

• tries to match 5' end ESTs with 3' end ESTs to make full-length predictions

In The End:

Came up with total of 31,778 predicted proteins

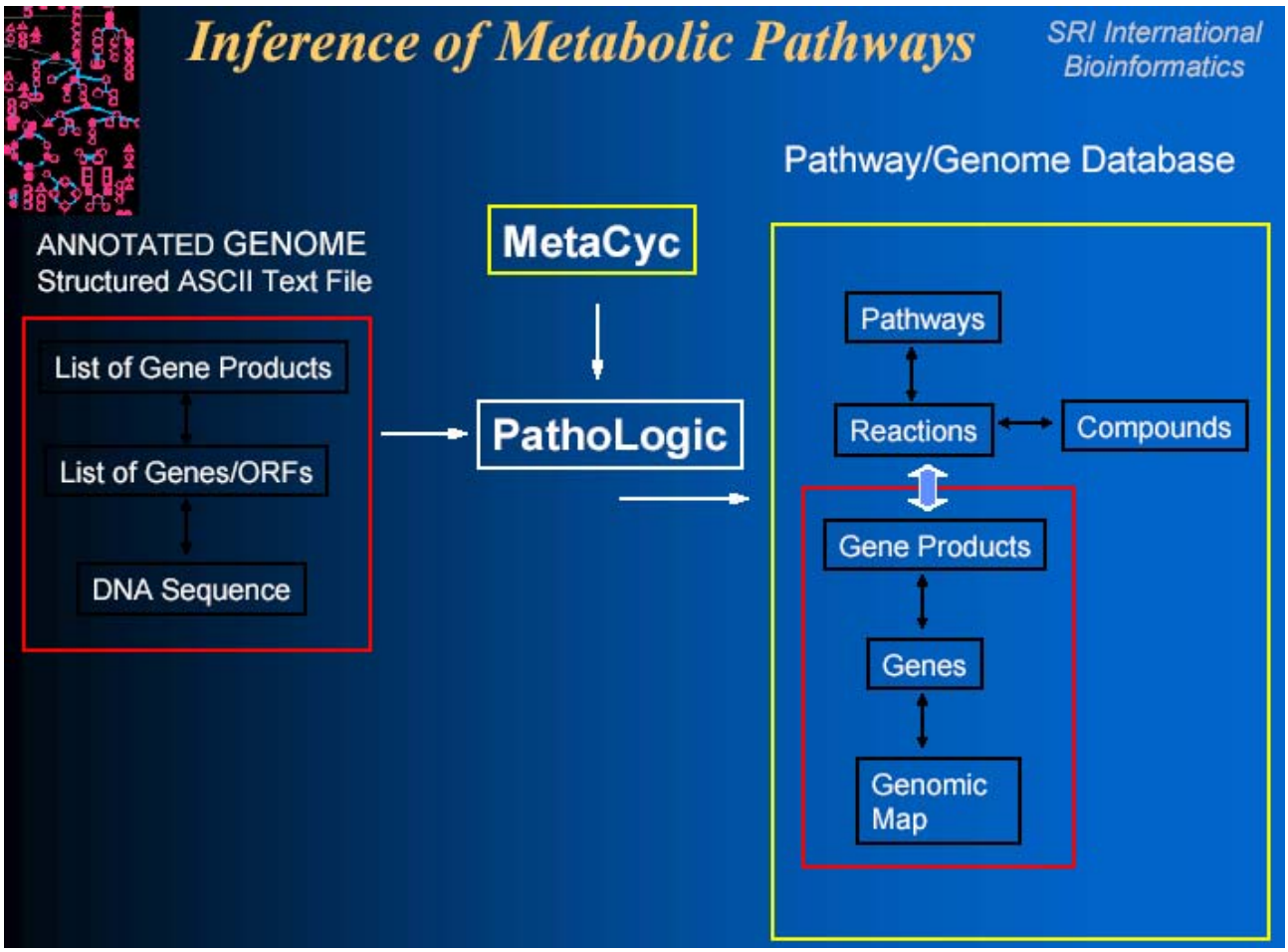
• 14,882 from known genes

• 12,839 from Ensemble

• 4,057 from Ensemble-Genie

Annotation Stage 3. Pathways and Networks

http://MetaCyc.org



- MetaCyc is a comprehensive metabolic pathway DB
- Literature based
- Detailed information on each pathway
- Goal is to contain an example of every different metabolic pathway
- Freely available

Bioinformatics Assignment:

1) Pick a gene. Obtain both its protein and nucleotide sequence in FASTA format. Give me the database source, provide its accession number. Provide a restriction map for enzymes in the MCS of pUC19. What enzymes could we use to clone **all** of the gene into pUC19.

2) From the sequence annotation, give any information available about the protein's function. Search the literature and tell me a recent (within the last 2 years) finding about the protein.

3) Use one of the secondary structure prediction tools to predict the secondary structure across the whole sequence. Does it have any putative membrane-spanning segments? Tell me something about this protein's structure.

4) Find its domain architecture using the Pfam site. Give me the names of any domains found in your protein. What are the known functions of these domains.

5) Run BLAST on your sequence against the GenPept Dbase. Are there any **distant** relatives of your sequence (ie. A related gene, **NOT** the same gene in a different organism) ? Name one and tell me why you think it is a "distant relative".

6) Obtain two homologs to your protein from different species and perform a full-length sequence alignment. Show me the alignment. Anything interesting ?

Human	Protein Kinase C
Drosophila	Estrogen Receptor
Arabidopsis	DNA polymerase
Mouse	Beta-tubulin
Saccharomyces cerevisiae	Acetyl CoA carboxylase
	HMG-CoA reductase
	Phosphoglycerate kinase
	Wilm's Tumour factor