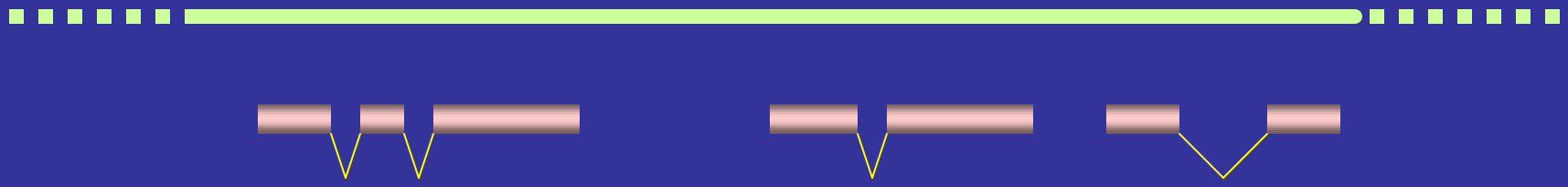
Gene Finding

Glimmer and Genscan



Strategy 1

Based on homology to known proteins

Run blastx at http://www.ncbi.nlm.nih.gov/BLAST/ or http://ser-loopp.tc.cornell.edu/cbsu/pblast.htm

BLASTX

Position		Targets	E value
230-1000	1	ref|NP_191509	1.00E-01
2035-4500	1	ref|NP_410450	4.00E-02
9500-8416	1	ref|NP_600075	2.00E-10
1774-2532	-1	ref|NP_628300	0
2600-5347	-1	ref|NP_624600	3.00E-05
5682-5864	-1	ref|NP_620000	6.00E-15
…	…	…	…

Strategy 2. Using gene finding programs

Glimmer

For most procaryotic genomes

http://www.tigr.org

Genscan

For some eucaryotic genomes

http://genes.mit.edu/GENSCAN.html

Some Other Gene Finding Systems

GeneMark: models for many individual species

– http://genemark.biology.gatech.edu/Genemark/

Genie: human, *Drosophilia*

– http://www.fruitfly.org/seq_tools/genie.html

GeneFinder: human *C. elegans*

– *http://ftp.genome.washington.edu/cgi-bin/Genefinder*

GRAIL: human, mouse

– http://grail.lsd.ornl.gov/grailexp/

Basics of Gene finding programs

1. **Search by signal**
   - a. Ribosomal binding site
   - b. Splicing site
   - c. Stop codon
   - d. Others

2. **Search by content**
   - a. Nucleotide distribution within coding region

The difficulty of gene finding

1. No clear-cut translation start, splicing signal.

2. Coding density in eucaryotes is extremely low.

	Genome Size	Density
Procaryotes	0.5 -10 Mb	90%
Eucaryotes	3300 Mb (human)	1-3% (human)

Glimmer
(TIGR)

- Glimmer can find 98% of genes in a bacterial genome.

- The program requires no other input than the genome sequence.

Algorithm: Interpolated Markov Model (IMM)

A (25%) C (25%)
G (25%) T (25%)

A

A (28%) C (23%)
G (22%) T (27%)

AGCTA

A (55%) C (15%)
G (10%) T (20%)

1 - 8 nt chain

A (?%) C (?%)
G (?%) T (?%)

Step 1: Build a Markov chain model to describe the probability of each of the 4 nucleotide after certain short prefix (contexts)

How to select training sequence?

Default solution to generate training sequences:
Using the long-orfs program to generate a run of codons that contains no stop codons (overlapping orfs are removed).

Alternative solutions for training sequences:
1. annotated sequence from a closely related species

2. doing blastx against protein database

Glimmer Step 1: Select training sequences

Program. long-orfs and extract
(identifying longest non-overlapping orf)
commands:

```
long-orfs ecoli_k12 –g 1200 > list.coord

trim the header lines of list.coord

extract ecoli_k12 list.coord >trainingseq
```

Glimmer Step 2: Building models

Program. **build-icm.exe**

command: **build-icm <trainingseq >training_model**

Glimmer Step 3: Run Glimmer

Program. glimmer2

command: glimmer2 ecoli_k12 train.model -g 150

Glimmer Step 4: Run RBSfinder.pl to redefine the RBS site

Program. rbs_finder.pl

command: rbs_finder.pl ecoli_k12 coord_file Output_file 50

GenScan

(Chris Burge, MIT)

http://genes.mit.edu/GENSCAN.html

Current available models:

human

Arabidopsis

Maize

Sensitivity (Sn) and Specificity (Sp)

	Accuracy per nucleotide			Accuracy per exon				
Method	**Sn**	**Sp**	**AC**	**Sn**	**Sp**	**(Sn+Sp)/2**	**ME**	**WE**
GENSCAN	**0.93**	**0.93**	**0.91**	**0.78**	**0.81**	**0.80**	**0.09**	**0.05**
FGENEH	**0.77**	**0.85**	**0.78**	**0.61**	**0.61**	**0.61**	**0.15**	**0.11**
GeneID	**0.63**	**0.81**	**0.67**	**0.44**	**0.45**	**0.45**	**0.28**	**0.24**
GeneParser2	**0.66**	**0.79**	**0.66**	**0.35**	**0.39**	**0.37**	**0.29**	**0.17**
GenLang	**0.72**	**0.75**	**0.69**	**0.50**	**0.49**	**0.50**	**0.21**	**0.21**
GRAILII	**0.72**	**0.84**	**0.75**	**0.36**	**0.41**	**0.38**	**0.25**	**0.10**
SORFIND	**0.71**	**0.85**	**0.73**	**0.42**	**0.47**	**0.45**	**0.24**	**0.14**
Xpound	**0.61**	**0.82**	**0.68**	**0.15**	**0.17**	**0.16**	**0.32**	0.13

Limitation of Genscan

- **Organism specific**

- **Prediction of gene boundaries are not reliable. (polyadenylation or promoter signals, the first and last exons)**