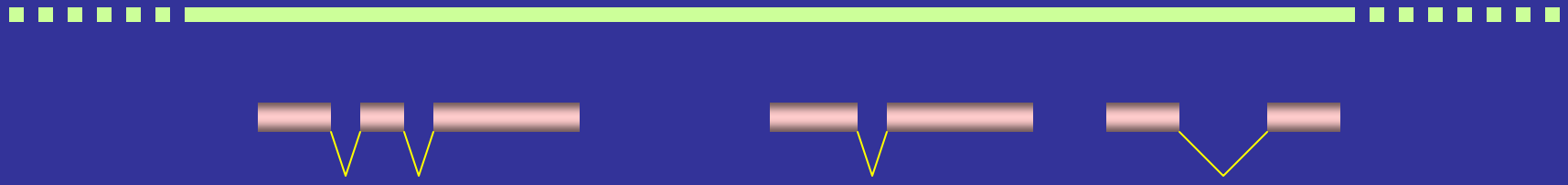
Gene Prediction & Annotation

Qi Sun

Computational Biology Service
Unit
Cornell University



Genome sequencing projects:

1. **shotgun sequencing**

2. **454 technology (~ $40 - $50k per bacterial genome)**

```
   1 aacagggtgt atctcgcaca ttctcatcca ctagtataac tgctgctgac agtaatcgaa
  61 ctagatagac tgttctggat gctatcattc gatattttga caacacggga gccatcctgt
 121 tcgttgatcc gagattcgac gagtcatgca acaagatcca gaccgttgcc tgcaaacgcc
 181 taggctgtga atgaacgact cgatcacgat cgctagtcgc acgtctgatc tcaccgattg
 241 aagccgtatt ccacagagtg cgagaaccgg tcatttactg agtggttcgg ctctgtttaa
 301 atacggaaag cccactcggg agagatatct ctccttaatg ggctatgaaa ggtatgaatg
 361 gtggcggcga accgcgtttc ccagaggctc gcgcactcca gtactccccg gaacgctggt
 421 gggcttatct tccgtgttcg ggatgggtac gggaggcaac cccaccgctg tggccgccta
 481 acgtcgagtc acggaatcga accgcgatag taccagtctc gattaactct tccaccgtgt
 541 gattacgtgc gatccagttt gcgcctggac tcgttcagcg acgagttaaa tcgatggtga
 601 atgagtcaca gtgcgtatga atgatggctt tggtctgtta gtgctcgtgg gcttaacgtc
 661 tcgttacctc gacgcgcaca ccccgagtct atcgaccgcg tcttgtacgc gggacctcgg
 721 cggtgtctct tttccaagtg ggtttcgagc ttagatgcgt tcagctctta ccccgtgtgg
 781 cgtggctacc cggcacgtgc tctctcgaac aaccggtaca ccagtggcca ccaaccgtag
 841 ttcctctcgt actatacggt cgttcttgtc agacaccatt acacacccag tagatagcag
 901 ccgacctgtc tcacgacggt ctaaacccag ctcacgacat cctttaatag gcgaacaacc
 961 tcacccttgc ccgcttctgc acgggcagga tggagggaac cgacatcgag gtagcaagcc
1021 actcggtcga tatgtgctct tgcgagtgac gactctgtta tccctagggt agcttttctg
1081 tcatcaattg cccgcatcaa gcaggctaat tggttcgcta gaccacgctt tcgcgtcagc
1141 gttcctcgtt gggaagaaca ctgtcaagct taattttgct cttgcactct tcgccgggtc
1201 tctgtcccgg ctgagatagc catagggcgc gctcgatatc ttttcgagcg cgtaccgccc
1261 cagtcaaact gcccggctat cggtgtcctc ctcccggagt gagagtcgca gtcaccgacg
1321 ggtagtattt cactgttgac tcggtggccc gctagcgcgg gtacctgtgt agtgtctcct
1381 atgtatgctg cacatcggcg accacgtctc agcgacagcc tgcagtaaag ctccataggg
1441 tcttcgcttc cccctgggtg tctccagact ccgcactgga atgtacagtt caccgggccc
1501 aacgttggga cagtgaagct ctggttaatc cattcatgca agccgctact gatgcggcaa
1561 ggtactacgc taccttaaga gggtcatagt taccccgcc gttgacaggt ccttcgtcct
1621 cttgtacgag gtgttcagat acctgcactg ggcaggattc agtgaccgta cgagtccttg
1681 cggatttgcg gtcacctatg ttgttactag acagtccgag cttccgagtc actgcgacct
1741 gctccgttcc ggagcaggca tcccttcttc cgaaggtacg ggactaactt gccgaattcc
1801 ctaacgttgg ttgctcccga caggccttgg ctttcgccgc catggacacc tgtgtcggtt
```

Strategy 1

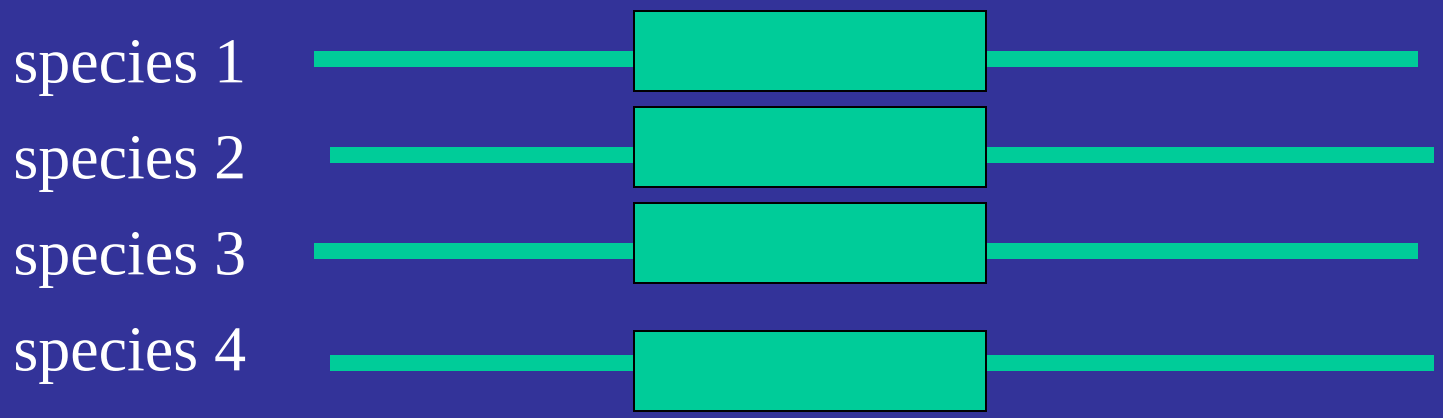
Based on similarity to known proteins

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: evolutionary conserved elements



Reference: Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES.(2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* 423:241-54

Strategy 3. gene finding programs

Glimmer

For most procaryotic genomes

http://www.cbcb.umd.edu/software/glimmer/

Fgenesh & Genscan

For some eucaryotic genomes

http://www.softberry.com

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. sequence pattern 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)

TIGR Annotation Engine Service

http://www.tigr.org/edutraining/training/annotation_engine.shtml

- **Glimmer**
- **BER HMM annotation**
- **Manatee manual curation**

Glimmer
(TIGR)

- **Glimmer can find 99% of genes in a bacterial genome.**

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

ORF vs Gene in Glimmer

Goal of Glimmer is to distinguish between ORF and Gene

ORF (Open Reading Frame)

- absence of translation “stop” codon

Gene

- start with “start” codon; end with “stop” codon
- has biological significance

Start and stop codens:

start codon: **ATG GTG TTG**

stop codon: **TAA TAG TGA**

AGGTACGATCGATGACGC**ATG**GAT
GACGCGATACGTACT**TGA**GGAC

ORF identification -- 6 frame translation

```
ATGCTTTGCTTGGAT
|||||||||||||||
TACGAAACGAACCTA

frame 1 ATG CTT TGC TTG GAT
frame 2  TGC TTT GCT TGG ATT
frame 3   GCT TTG CTT GGA TTC

frame -1 ATC CAA GCA AAG CAT
frame -1  TCC AAG CAA AGC ATG
frame -1   CCA AGC AAA GCA TGC
```

Algorithm: Interpolated Markov Model (IMM)

```
...AACTCGTAGTCGATTTACGAGAGCTAAAC
GACTCGACGACGGACGTACGGACCGACTACGA
CCCAG ...
```

Algorithm: Interpolated Markov Model (IMM)

Random

AGCTA

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

Coding

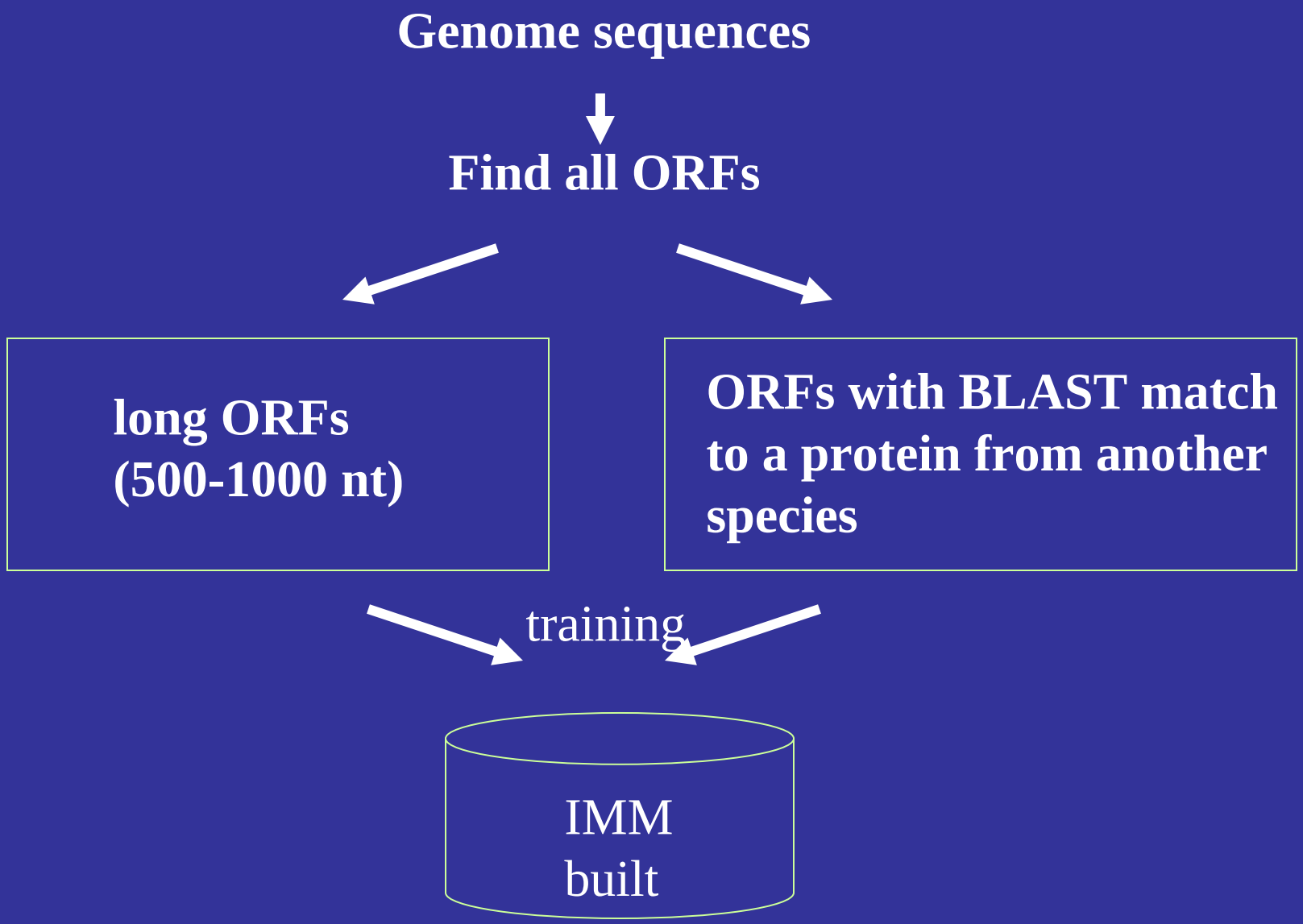
AGCTA

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

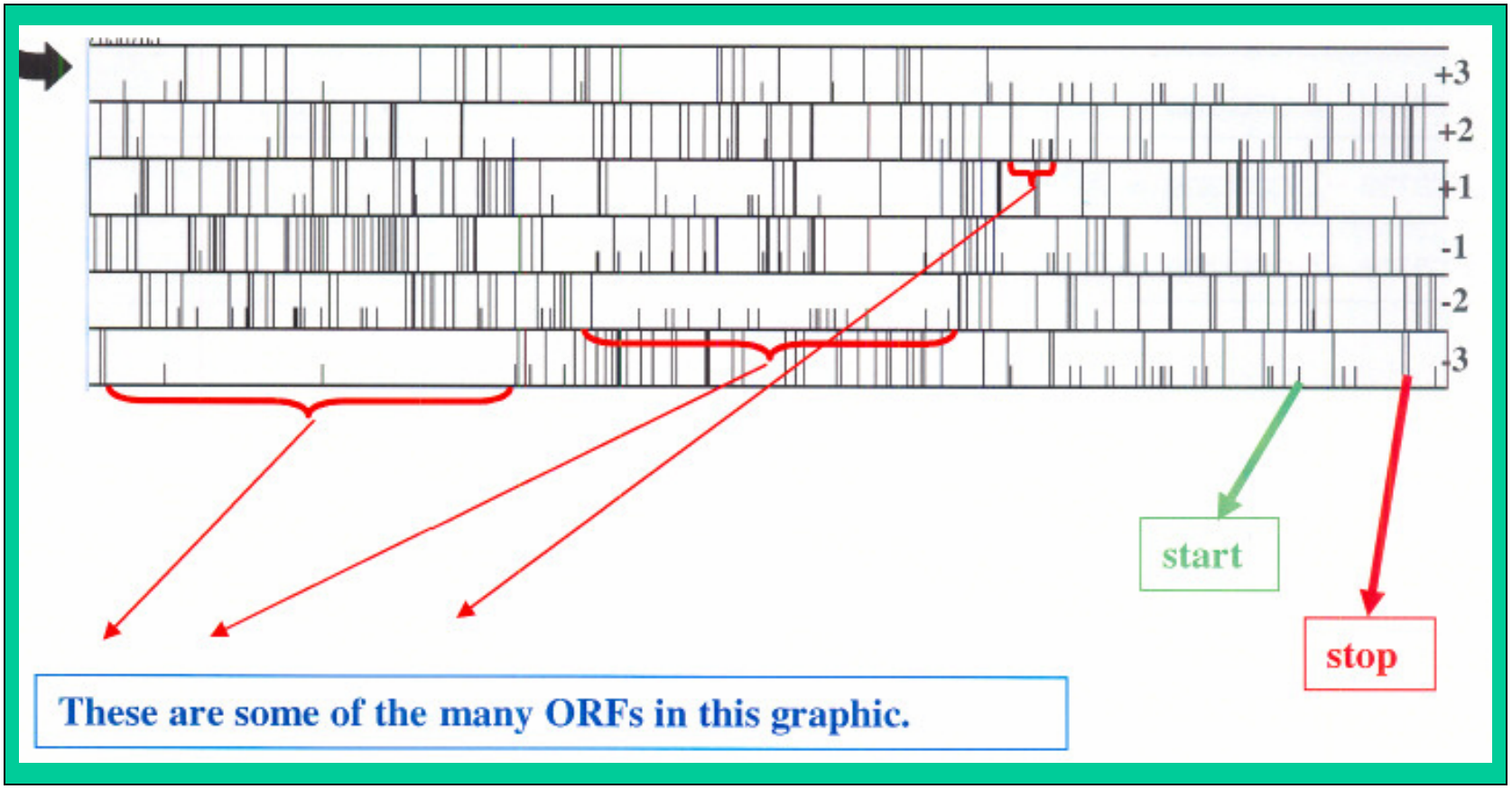
1 - 8 nt chain

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

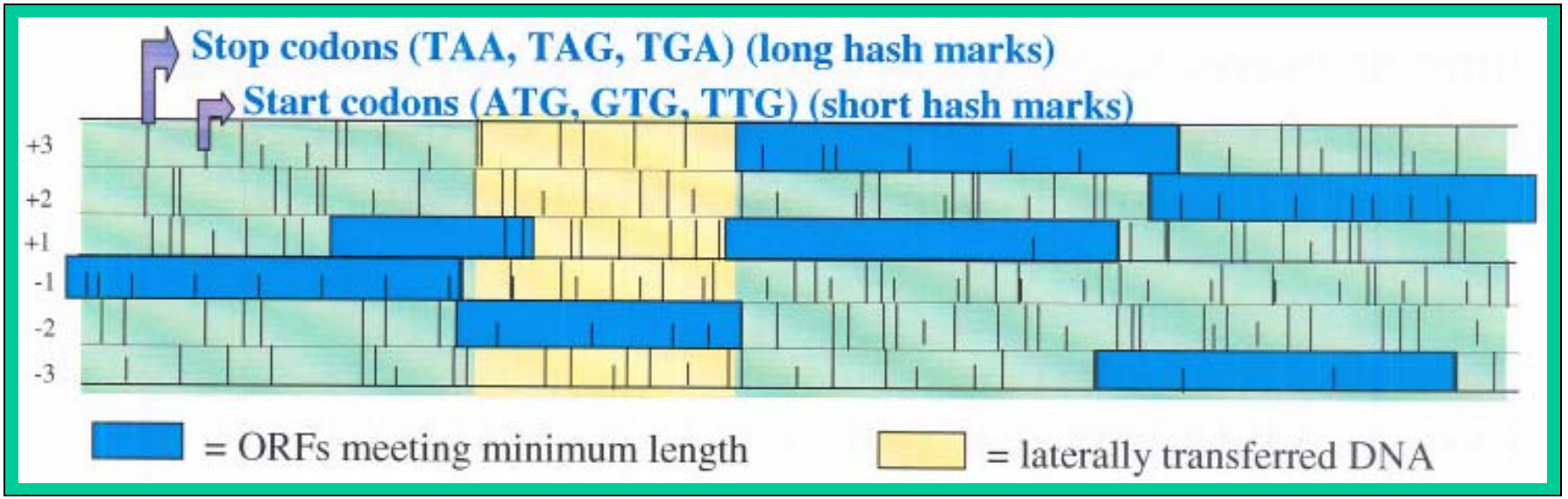
Step 1: Select training sequences

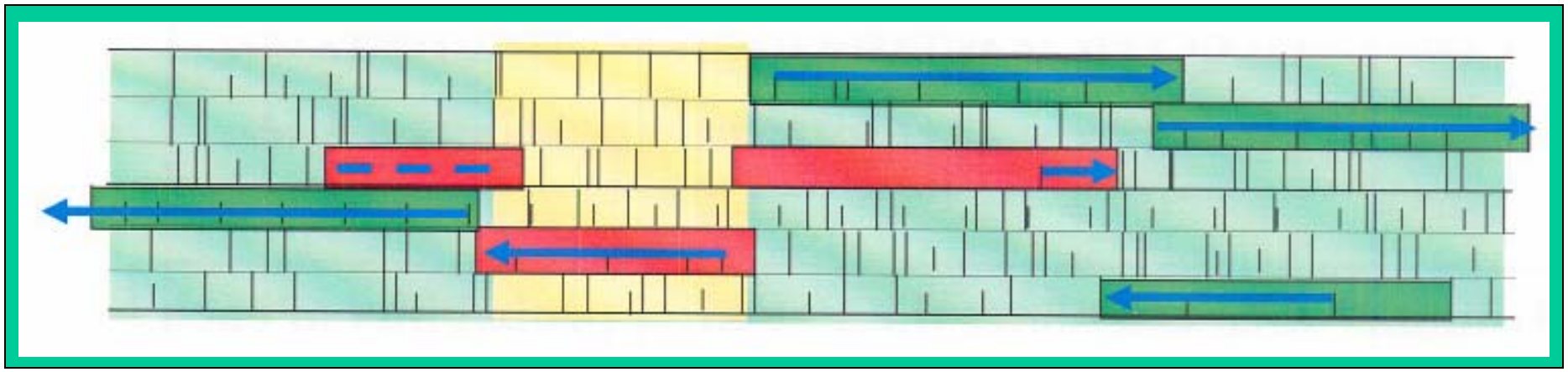


ORF

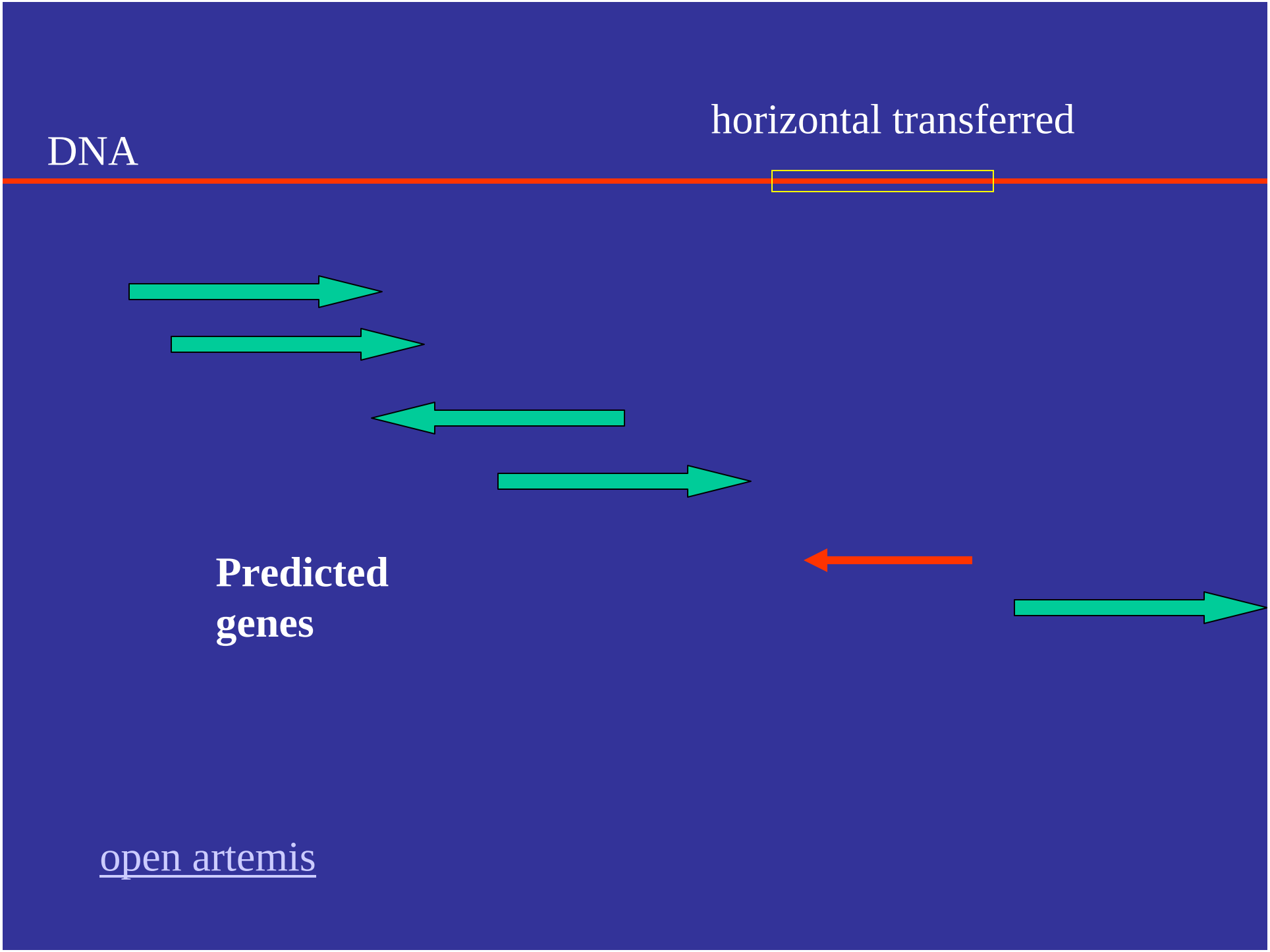

+3
+2
+1
-1
-2
-3
start
stop
These are some of the many ORFs in this graphic.

Candidate genes (from start to stop codon)





Glimmer score: red area has low score; green area has high score; dotted line has no start codon;


DNA
horizontal transferred
Predicted genes
open artemis

TIGR Annotation Engine Service

Glimmer
BLAST-extend-reprze (BER)
HMM

TIGR

MANATEE

Manual curation

Local

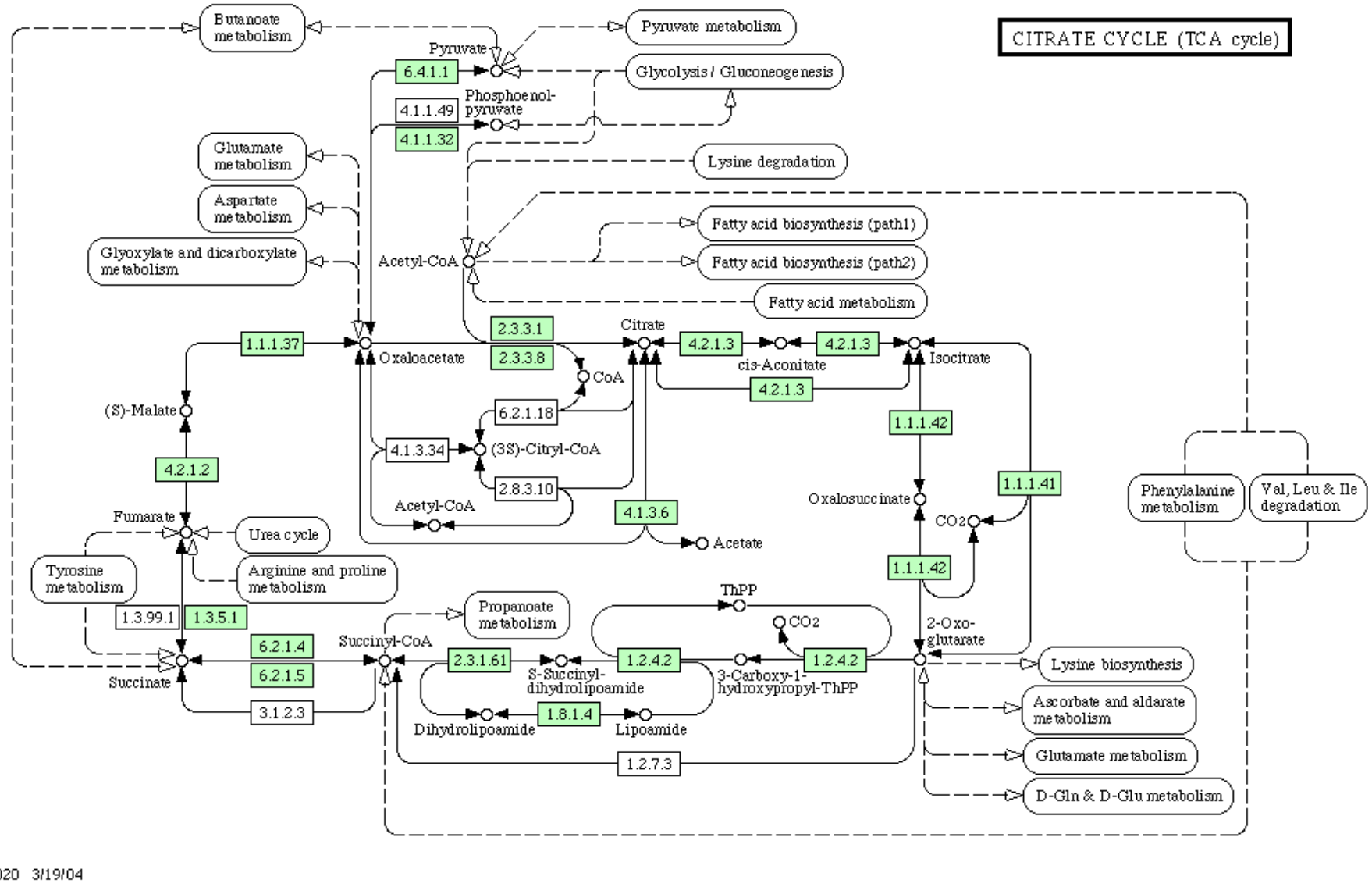
Finding the function of a new protein

1. **Experimental characterization**
   - **Microarray**
   - **Mutation analysis**
   - **Protein interaction**

2. **Homology searching**

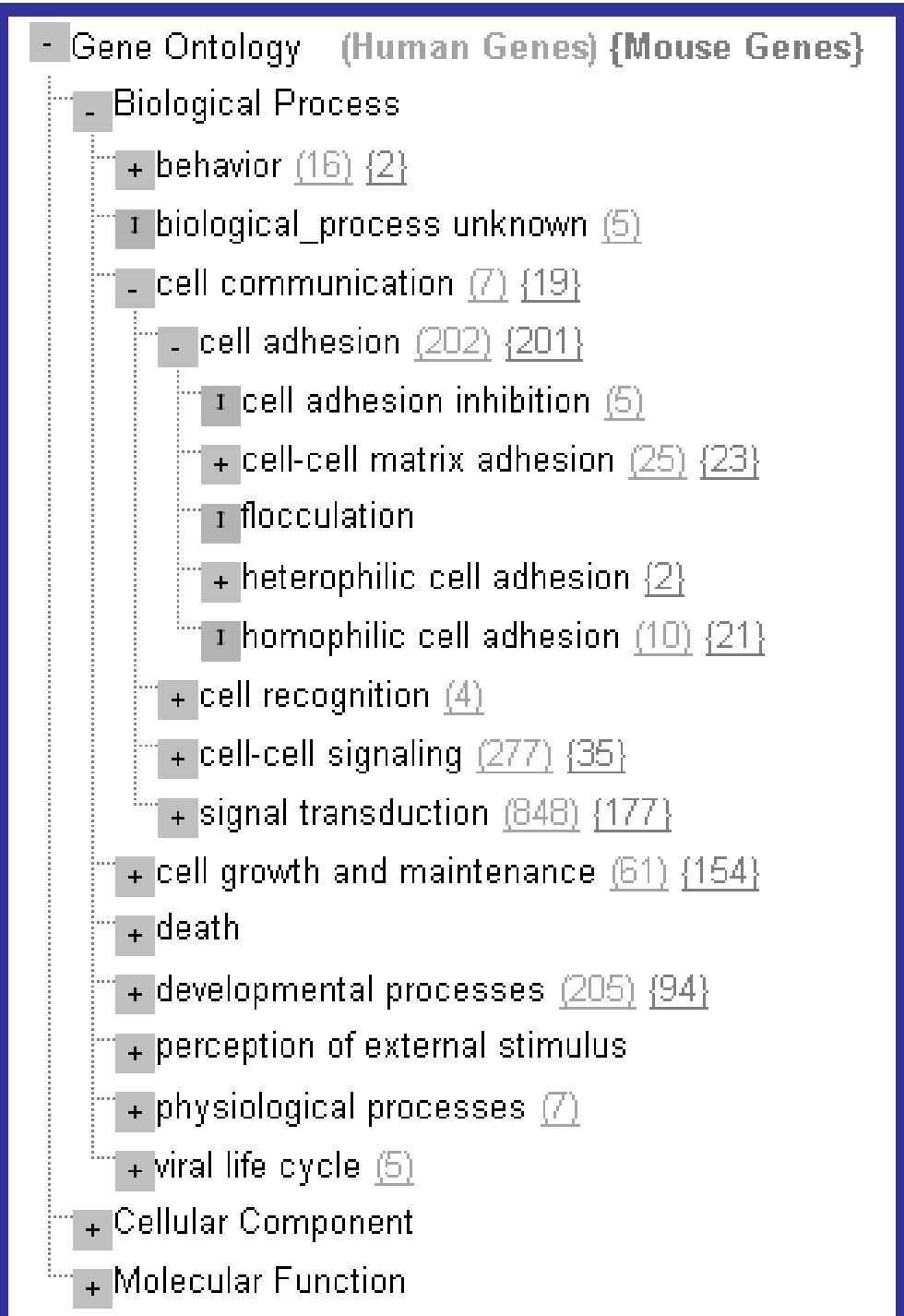
   - **BLAST**
   - **HMM**
   - **threading**

Annotation with KEGG database



Annotation with GO:

Biological Process



HMM based annotation process

DNA binding domain

GO:0003677 : DNA binding (9406)

GO:0005634 : nucleus (14440)

Interproscan can automate the annotation process

Interproscan result.