ORFs and genes

Please sit in row K or forward
Question: why do some strains of Vibrio cause cholera and others don’t?

Methods

- Genome sequencing: dideoxy method
- Genome sequencing: shotgun method and assembly
- Gene finding
- Comparing genomes
  - Alignment (dynamic programming version)
  - Best reciprocal hit method

Mechanisms

- Genes and the cellular processes for making protein
- Genome evolution: how do genomic differences arise?
Topics for today

• Short summary of protein production in bacteria
• Finding open reading frames (ORFs)
• Determining which ORFs are genes, take 1
What happens when you ingest pathogenic Vibrio cholerae?

Infectious dose: $10^3-10^8$ cells
Vibrio in the gut

- Pathogenesis due to excretion of cholera enterotoxin
- Uses flagellum to move near intestinal wall
A bacterial operon is a set of adjacent genes transcribed together on the same mRNA.
Vibrio cholerae Flagellar Synthesis and Virulence. Anastasia R. Rugel and Karl E. Klose
How do motA, motB (and flagellin) get made?

1. transcription

DNA → RNA

2. translation

RNA → Peptide sequence

3. protein folding

Peptide sequence → Folded protein
Transcription basics: initiation

RNA polymerase

Promoter → Termination sequence → Promoter

Promoter

Termination sequence
Transcription basics: elongation

Newly synthesized mRNA

Template strand

Coding strand

Promoter

Termination sequence
Transcription basics: termination

Termination sequence
Translation basics: initiation

mRNA

Ribosome binding site  Coding sequence  Coding sequence

Ribosome
Translation basics: elongation

<table>
<thead>
<tr>
<th>First letter</th>
<th>T</th>
<th>C</th>
<th>A</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>TTT</td>
<td>Phe F</td>
<td>TCT Ser S</td>
<td>TAT Tyr Y</td>
<td>TGT Cys C</td>
</tr>
<tr>
<td>TTC</td>
<td>Phe F</td>
<td>TCC Ser S</td>
<td>TAC Tyr Y</td>
<td>TGC Cys C</td>
</tr>
<tr>
<td>TTA</td>
<td>Leu L</td>
<td>TCA Ser S</td>
<td>TAA Stop</td>
<td></td>
</tr>
<tr>
<td>TGG</td>
<td>Leu L</td>
<td>TCG Ser S</td>
<td>TAG Stop</td>
<td>TGG Trp W</td>
</tr>
<tr>
<td>CTT</td>
<td>Leu L</td>
<td>CCT Pro P</td>
<td>CAT His H</td>
<td>CGT Arg R</td>
</tr>
<tr>
<td>CTC</td>
<td>Leu L</td>
<td>CCC Pro P</td>
<td>CAC His H</td>
<td>CGC Arg R</td>
</tr>
<tr>
<td>CTA</td>
<td>Leu L</td>
<td>CCA Pro P</td>
<td>CAA Gln Q</td>
<td>CGA Arg R</td>
</tr>
<tr>
<td>CTG</td>
<td>Leu L</td>
<td>CCG Pro P</td>
<td>CAG Gln Q</td>
<td>CGG Arg R</td>
</tr>
<tr>
<td>ATT</td>
<td>Ile I</td>
<td>ACT Thr T</td>
<td>AAT Asn N</td>
<td>AGT Ser S</td>
</tr>
<tr>
<td>ATC</td>
<td>Ile I</td>
<td>ACC Thr T</td>
<td>AAC Asn N</td>
<td>AGC Ser S</td>
</tr>
<tr>
<td>ATA</td>
<td>Ile I</td>
<td>ACA Thr T</td>
<td>AAA Lys K</td>
<td>AGA Arg R</td>
</tr>
<tr>
<td>ATG</td>
<td>Met M</td>
<td>ACG Thr T</td>
<td>AAG Lys K</td>
<td>AGG Arg R</td>
</tr>
<tr>
<td>GTT</td>
<td>Val V</td>
<td>GCT Ala A</td>
<td>GAT Asp D</td>
<td>GGT Gly G</td>
</tr>
<tr>
<td>GTC</td>
<td>Val V</td>
<td>GCC Ala A</td>
<td>GAC Asp D</td>
<td>GGC Gly G</td>
</tr>
<tr>
<td>GTA</td>
<td>Val V</td>
<td>GCA Ala A</td>
<td>GAA Glu E</td>
<td>GGA Gly G</td>
</tr>
<tr>
<td>GTG</td>
<td>Val V</td>
<td>GCG Ala A</td>
<td>GAG Glu E</td>
<td>GGG Gly G</td>
</tr>
</tbody>
</table>
### Translation basics: termination

<table>
<thead>
<tr>
<th>First letter</th>
<th>T</th>
<th>C</th>
<th>A</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>TTT Phe F</td>
<td>TCT Ser S</td>
<td>TAT Tyr Y</td>
<td>TGT Cys C</td>
</tr>
<tr>
<td></td>
<td>TTC Phe F</td>
<td>TCC Ser S</td>
<td>TAC Tyr Y</td>
<td>TGC Cys C</td>
</tr>
<tr>
<td></td>
<td>TTA Leu L</td>
<td>TCA Ser S</td>
<td>TAA Stop</td>
<td>TGA Stop</td>
</tr>
<tr>
<td></td>
<td>TTG Leu L</td>
<td>TCG Ser S</td>
<td>TAG Stop</td>
<td>TGG Trp W</td>
</tr>
<tr>
<td>C</td>
<td>CTT Leu L</td>
<td>CCT Pro P</td>
<td>CAT His H</td>
<td>CGT Arg R</td>
</tr>
<tr>
<td></td>
<td>CTC Leu L</td>
<td>CCC Pro P</td>
<td>CAC His H</td>
<td>CGC Arg R</td>
</tr>
<tr>
<td></td>
<td>CTA Leu L</td>
<td>CCA Pro P</td>
<td>CAA Gln Q</td>
<td>CGA Arg R</td>
</tr>
<tr>
<td></td>
<td>CTG Leu L</td>
<td>CCG Pro P</td>
<td>CAG Gln Q</td>
<td>CGG Arg R</td>
</tr>
<tr>
<td></td>
<td>ATT Ile I</td>
<td>ACT Thr T</td>
<td>AAT Asn N</td>
<td>AGT Ser S</td>
</tr>
<tr>
<td></td>
<td>ATC Ile I</td>
<td>ACC Thr T</td>
<td>AAC Asn N</td>
<td>AGC Ser S</td>
</tr>
<tr>
<td></td>
<td>ATA Ile I</td>
<td>ACA Thr T</td>
<td>AAA Lys K</td>
<td>AGA Arg R</td>
</tr>
<tr>
<td></td>
<td>ATG Met M</td>
<td>ACG Thr T</td>
<td>AAG Lys K</td>
<td>AGG Arg R</td>
</tr>
<tr>
<td>G</td>
<td>GTT Val V</td>
<td>GCT Ala A</td>
<td>GAT Asp D</td>
<td>GGT Gly G</td>
</tr>
<tr>
<td></td>
<td>GTC Val V</td>
<td>GCC Ala A</td>
<td>GAC Asp D</td>
<td>GGC Gly G</td>
</tr>
<tr>
<td></td>
<td>GTA Val V</td>
<td>GCA Ala A</td>
<td>GAA Glu E</td>
<td>GGA Gly G</td>
</tr>
<tr>
<td></td>
<td>GTG Val V</td>
<td>GCG Ala A</td>
<td>GAG Glu E</td>
<td>GGG Gly G</td>
</tr>
</tbody>
</table>
Transcription and translation in bacteria: a view of the whole process
Topics for today

• Short summary of protein production in bacteria
• Finding open reading frames (ORFs)
• Determining which ORFs are genes, take 1
Transcription and translation from the perspective of gene finding: things we could look for

- Ribosome binding site
- Sequence of codons
- Sequence of codons
- Transcription termination sequence
- Promoter
- DNA
- mRNA
Transcription and translation from the perspective of gene finding: things we could look for
The codon structure of coding sequence provides a strong signal of a gene

Open reading frame (ORF): a region between a start codon and next in-frame stop
Protein coding genes consist of long ORFs

A Vibrio flagellin gene (one of several)
Protein coding genes can occur on either strand

Gene 1: coding strand is on top
5' – AATGCCGTGCTTGTAGACGTAGGCTTAGATCGTCATGGG – 3'
3' – TTACGGCACGAACATCTGCATCCGAATCTAGCAGTACCC – 5'

Gene 2: coding strand is on bottom

ATG = start codon
TGA, TAG, TAA = Stop codons
How many orfs are there here? (consider both strands).

5′ CTAGTATGACATTACGTTAATTAATTA 3′

ATG = start codon
TGA, TAG, TAA = Stop codons
How many orfs are there here? (consider both strands).

5’ CTAGTATGACATTACGT TAATTA 3’
3’ GATCATACTGTAATGCAATTAAT 5’

ATG = start codon
TGA, TAG, TAA = Stop codons
def restOfORF(DNA):

def oneFrame(DNA,start,minOrfSize):

allOrfs(DNA,minOrfSize):
Topics for today

• Short summary of protein production in bacteria
• Finding open reading frames (ORFs)
• Determining which ORFs are genes, take 1
Not all ORFs are genes...
The “scenario”

• Imagine we’ve just sequenced Vibrio cholerae, but have no annotations
• We do have good annotations for the closely related E. coli

http://www.flickr.com/photos/54976525@N08/7748440602
Length based gene finding

>gi|15640032|ref|NC_002505.1| Vibrio cholerae O1 biovar El Tor str. N16961 chromosome I, complete sequence
AGGGTCATTAAATATATATAAAGATCTATATAGAGATCTTTTTATTAGATCTACTATTAGGAGCAGGAT
CTTTTGAGATAAGTAAGTTATACAAAGTACATCGGCAGATCGAGATCGTGTGATCAACCA
CTGATGCTTCAAGGATTAGCTGGGATCAAAAACCTATGTCTATACACGGACACCTTTGGGATCTAAAAG
AGTGTGAATTTTGTATTAGGAAGATCACCGGATAATCGTATAGTTATCCACATGAGATTTGATTAAAAA
GCATCAATCTAATTTTTCTACTCGGTTAAATTTTTACCAACATCCAAAAAGAAGCGCCATTAGGGCGCT
CTGATGGAATAGGGCATTATTTTAGAAGCGATTGATGACGCGTTTGAGCCAAGCGTCAAGCTCAGCAGCTA
AGGTAGGTGCTCTTGTACATCGATGAGATTTGATTAAAAAAGCATCAATCTAATTTTTCTACTCGGTTAA

...
E. coli

Length of ORF (in nucleotides)

ORF count

Red: true genes
Blue: noncoding ORFs

Lengths of ecoli coding and noncoding ORFs
True positive and false discovery rates

Our predictions

Gene

Not gene

ORFs

a
b
c
d
e
f
g
h
True positive and false discovery rates

ORFs

Gene

Not gene

Our predictions
True positive and false discovery rates

Our predictions

<table>
<thead>
<tr>
<th>Gene</th>
<th>ORFs</th>
<th>Not gene</th>
</tr>
</thead>
<tbody>
<tr>
<td>b c e f</td>
<td>a</td>
<td>g h</td>
</tr>
<tr>
<td>a c e f g</td>
<td>d</td>
<td>h</td>
</tr>
<tr>
<td>a c e f</td>
<td>b</td>
<td>d</td>
</tr>
<tr>
<td>a c e f g</td>
<td>d</td>
<td>h</td>
</tr>
</tbody>
</table>

Reality

<table>
<thead>
<tr>
<th>Gene</th>
<th>ORFs</th>
<th>Not gene</th>
</tr>
</thead>
<tbody>
<tr>
<td>b c e f</td>
<td>a</td>
<td>g h</td>
</tr>
<tr>
<td>a c e f g</td>
<td>d</td>
<td>h</td>
</tr>
<tr>
<td>a c e f</td>
<td>b</td>
<td>d</td>
</tr>
<tr>
<td>a c e f g</td>
<td>d</td>
<td>h</td>
</tr>
</tbody>
</table>
Our predictions | Reality
---|---
**Gene**
| | |
| a | a |
| c | b |
| e | c |
| f | d |
| g | d |

**Not gene**
| | |
| b | e |
| d | f |
| h | g |
| | h |

True positive rate = \( \frac{\text{Num. true positives (a and c)}}{\text{Real positives (a,b,c,d)}} \) = \( \frac{2}{4} \)

False discovery rate = \( \frac{\text{Num false positives (e,f,g)}}{\text{Num. positive predictions (a,c,e,f,g)}} \) = \( \frac{3}{5} \)
Where do the known annotations come from?

Red: true genes
Blue: noncoding ORFs

LOCUS U00096 4641652 bp DNA circular BCT 01-AUG-2014
DEFINITION Escherichia coli str. K-12 substr. MG1655, complete genome.
ACCESSION U00096
VERSION U00096.3 GI:545778205
DBLINK BioProject: PRJNA225
BioSample: SAMN02604091
KEYWORDS .
SOURCE Escherichia coli str. K-12 substr. MG1655
ORGANISM Escherichia coli str. K-12 substr. MG1655
Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacterales; Enterobacteriaceae; Escherichia.
Isolate RNA → Sequence → CTGATACGGTGAG...
Hand in your worksheet please!

(and be sure you put your name on it)