Case study: unknown RNA

You are performing a **metagenomic** analysis of Yellowstone National Park hot springs. After RT-PCR and sequencing, you ended up with the following information:

**Metagenomics:**
Genetic material collected directly from environmental samples. Create a profile of the biological diversity of a specific environment.
Reads:
Short stretches of sequenced bases

Contig:
Set of overlapping reads, assembled into a contiguous sequence

Coverage:
Number of reads that overlapped to create the contig

Case study: unknown RNA

How would you find the identity of one of these contigs?

>gi|380750588|gb|JQ756122.1| Uncultured clone contig00002 (partial)
ACCGCTCGAAATCATCGTGTCTTGAGAGTGTCTAAAGCTTCCG

Directionality of the unknown nucleotide sequence?

- + Strand
- Complement: ATGGCGTG
- Reverse: TACCGCAC
- Reverse complement: GTGCGGTA
- Reverse complement: CACGCCAT
**Nucleotide ambiguity codes are used for:**

*When more than 1 type of base is permitted within a recognition sequence.*

1. restriction enzyme sites \((\text{Accl} = \text{GT'mk}_{-}\text{AC})\)
2. recognition sequences \((\text{AYYAUGR})\)
3. genetic code \((\text{GGN} = \text{glycine})\)
4. consensus sequences \((\text{for alignments})\)

*Or when we don’t know the real sequence (SNPs)*

---

**Internationally accepted nucleotide codes:**

- \(\text{A, C, G, T, U, I (inosine)}\) are *obvious* codes
- \(\text{N code for any base (or unknown residue)}\)

**Double base codes:**

- \(\text{A or G} = \text{puRine}\)
- \(\text{C or T} = \text{pYrimidine}\)
- \(\text{G or T} = \text{Keto}\)
- \(\text{A or C} = \text{aMino}\)
- \(\text{G or C} = \text{Strong base pair}\)
- \(\text{A or T} = \text{Weak base pair}\)
More internationally accepted nucleotide codes:

B, D, H, V are the triple base codes

- \( B = \text{not A} \) (C or G or T)
- \( D = \text{not C} \) (A or G or T)
- \( H = \text{not G} \) (A or C or T)
- \( V = \text{not T} \) (A or C or G)

. (dot) = missing base or gap in sequence

E, F, J, L, O, P, Q, Z have no base codes

Ambiguity codes

Represent many sequences with a single sequence…

Consensus sequence:

\[
\begin{align*}
\text{A} & \quad \text{T} & \quad \text{G} & \quad \text{G} & \quad \text{C} & \quad \text{G} & \quad \text{T} & \quad \text{A} & \quad \text{C} & \quad \text{C} & \quad \text{G} & \quad \text{T} & \quad \text{A} \\
\text{A} & \quad \text{C} & \quad \text{G} & \quad \text{G} & \quad \text{C} & \quad \text{G} & \quad \text{C} & \quad \text{A} & \quad \text{C} & \quad \text{C} & \quad \text{A} & \quad \text{T} & \quad \text{A} \\
\text{A} & \quad \text{C} & \quad \text{T} & \quad \text{G} & \quad \text{C} & \quad \text{G} & \quad \text{A} & \quad \text{A} & \quad \text{C} & \quad \text{C} & \quad \text{G} & \quad \text{T} & \quad \text{A}
\end{align*}
\]
Find related sequences

<table>
<thead>
<tr>
<th>Accession</th>
<th>Description</th>
<th>Max score</th>
<th>Total score</th>
<th>Query coverage</th>
<th>E value</th>
<th>Gen id</th>
</tr>
</thead>
<tbody>
<tr>
<td>J017195.1</td>
<td>Homo sapiens DSM 2003, complete genome</td>
<td>52.0</td>
<td>52.0</td>
<td>1%</td>
<td>0.008</td>
<td>76%</td>
</tr>
<tr>
<td>YG7335.1</td>
<td>Geobacillus stearothermophilus GL1, complete genome</td>
<td>50.0</td>
<td>50.0</td>
<td>0%</td>
<td>0.21</td>
<td>94%</td>
</tr>
<tr>
<td>YG302596.1</td>
<td>Nectria haematococca strain 77-13-4, complete genome</td>
<td>49.4</td>
<td>49.4</td>
<td>0%</td>
<td>0.3</td>
<td>92%</td>
</tr>
<tr>
<td>YG302596.1</td>
<td>Mycoplasma genitalium 3535, complete genome</td>
<td>49.4</td>
<td>49.4</td>
<td>0%</td>
<td>0.3</td>
<td>100%</td>
</tr>
<tr>
<td>YG302596.1</td>
<td>Cryptococcus neoformans var. grubii H99 chromosome 1, complete genome</td>
<td>44.5</td>
<td>44.5</td>
<td>0%</td>
<td>0.1</td>
<td>93%</td>
</tr>
<tr>
<td>YG302596.1</td>
<td>Aspergillus fumigatus A3, complete genome</td>
<td>44.5</td>
<td>44.5</td>
<td>0%</td>
<td>0.1</td>
<td>94%</td>
</tr>
<tr>
<td>YG302596.1</td>
<td>Rhodococcus erythropolis strain H1, complete genome</td>
<td>44.5</td>
<td>44.5</td>
<td>0%</td>
<td>0.1</td>
<td>94%</td>
</tr>
<tr>
<td>YG302596.1</td>
<td>Desulfovibrionales strain J2, complete genome</td>
<td>44.5</td>
<td>44.5</td>
<td>0%</td>
<td>0.1</td>
<td>94%</td>
</tr>
<tr>
<td>YG302596.1</td>
<td>Streptomyces alcaligenes strain subsp. alcaligenes</td>
<td>44.5</td>
<td>44.5</td>
<td>0%</td>
<td>0.1</td>
<td>94%</td>
</tr>
<tr>
<td>YG302596.1</td>
<td>Streptomyces griseus strain ATCC 14940, complete genome</td>
<td>44.5</td>
<td>44.5</td>
<td>0%</td>
<td>0.1</td>
<td>94%</td>
</tr>
<tr>
<td>YG302596.1</td>
<td>Nuss medium BAC clone BP21-SIR10, complete genome</td>
<td>44.5</td>
<td>44.5</td>
<td>0%</td>
<td>0.1</td>
<td>94%</td>
</tr>
<tr>
<td>YG302596.1</td>
<td>Nuss medium BAC clone BP21-SIR10, complete genome</td>
<td>44.5</td>
<td>44.5</td>
<td>0%</td>
<td>0.1</td>
<td>94%</td>
</tr>
<tr>
<td>YG302596.1</td>
<td>Nuss medium BAC clone BP21-25B14, complete genome</td>
<td>44.5</td>
<td>44.5</td>
<td>0%</td>
<td>0.1</td>
<td>94%</td>
</tr>
<tr>
<td>YG302596.1</td>
<td>Nuss medium BAC clone BP21-25B14, complete genome</td>
<td>44.5</td>
<td>44.5</td>
<td>0%</td>
<td>0.1</td>
<td>94%</td>
</tr>
<tr>
<td>YG302596.1</td>
<td>Nuss medium BAC clone BP21-1771, complete genome</td>
<td>44.5</td>
<td>44.5</td>
<td>0%</td>
<td>0.1</td>
<td>94%</td>
</tr>
</tbody>
</table>


What can you do if all hits are low scoring?

Translate to amino acid sequence

Every sequence has 6 potential open reading frames (ORFs)

ATGGCGTGCAAGCAGGATATCCGTTTTTG...

1  M A C K H G Y P F L
2  W R A S T D I R F C
3  G V Q A R I S V F V

TTGTGGTCCATCAACACATCTTGAATCAA...
-1  L W F H Q H I L N Q
-2  C G S I N T S * I N
-3  V V P S T H L E S M
Algorithms that search for genes look for 3rd base GC bias or codon frequency matches.

Codons are not randomly distributed: certain codons occur more frequently in coding regions.
The amino acids also have standard single letter codes

Obvious AA codes:

<table>
<thead>
<tr>
<th>A</th>
<th>Ala</th>
<th>Alanine</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>Cys</td>
<td>Cysteine (not Cystine)</td>
</tr>
<tr>
<td>G</td>
<td>Gly</td>
<td>Glycine</td>
</tr>
<tr>
<td>H</td>
<td>His</td>
<td>Histidine</td>
</tr>
<tr>
<td>I</td>
<td>Ile</td>
<td>Isoleucine</td>
</tr>
<tr>
<td>L</td>
<td>Leu</td>
<td>Leucine</td>
</tr>
<tr>
<td>M</td>
<td>Met</td>
<td>Methionine</td>
</tr>
<tr>
<td>P</td>
<td>Pro</td>
<td>Proline</td>
</tr>
<tr>
<td>S</td>
<td>Ser</td>
<td>Serine</td>
</tr>
<tr>
<td>T</td>
<td>Thr</td>
<td>Threonine</td>
</tr>
<tr>
<td>V</td>
<td>Val</td>
<td>Valine</td>
</tr>
</tbody>
</table>

AA phonetic codes:

<table>
<thead>
<tr>
<th>F</th>
<th>Phe</th>
<th>PHenylalanine (fffffffenylalanine)</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>Asn</td>
<td>AsparagiNe (asparaginnnnnne)</td>
</tr>
<tr>
<td>R</td>
<td>Arg</td>
<td>ARginine (arrrrrrrginine)</td>
</tr>
<tr>
<td>Y</td>
<td>Tyr</td>
<td>TYrosine (tyyyyyyyyyrosine)</td>
</tr>
</tbody>
</table>

AA non-obvious codes:

<table>
<thead>
<tr>
<th>D</th>
<th>Asp</th>
<th>Aspartic acid</th>
</tr>
</thead>
<tbody>
<tr>
<td>E</td>
<td>Glu</td>
<td>Glutamic acid</td>
</tr>
<tr>
<td>K</td>
<td>Lys</td>
<td>Lysine</td>
</tr>
<tr>
<td>Q</td>
<td>Gln</td>
<td>Glutamine</td>
</tr>
<tr>
<td>W</td>
<td>Trp</td>
<td>Tryptophan</td>
</tr>
</tbody>
</table>
More amino acids codes:

AA ambiguity codes:
- B = Asx = Aspartic acid or Asparagine
- Z = Glx = Glutamic acid or Glutamine
- X = Any amino acid
- . (dot) = deletion or gap in sequence
- * (star) = End or translation terminator

J, O, U have no amino acid codes

Attributes of the real genetic code

The code is mostly universal
mitochondria, protozoa nuclei, and some mycoplasma use slight variations

Of the 64 codons, 3 are stop signals, the rest code for AAs
the code is highly degenerate, but has NO ambiguity**

# of codons encoding an AA correlates with its frequency
- 6 codons = L, R, S
- 4 codons = A, G, P, T, V
- 3 codons = I
- 2 codons = C, D, E, F, H, K, N, Q, Y
- 1 codon = M, W

** UGA is opal terminator or may encode selenocysteine, if in right 2D RNA context
Addition of Arg is relatively recent

**Arg synthesis gives off urea.**

Can't metabolize this before urea cycle of N₂ fixation is established in evolution
The original code probably included Orn instead.

**Orn** functions are now **Lys** functions, and only the most "robust" basic AA needs are given to **Arg.**

---

Addition of Arg is relatively recent

![Graph showing observed and expected AA frequency](http://ars.els-cdn.com/content/image/1-s2.0-S0753332202002846-fx1.gif)

But 6 codons are too many for present **Arg** needs, so there is less **Arg** than predicted by its codon number.
### 2nd codon base:

<table>
<thead>
<tr>
<th>Base</th>
<th>Kcal/mole</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gly G</td>
<td>+2.39 (hydrophobic)</td>
</tr>
<tr>
<td>Leu U</td>
<td>+2.28</td>
</tr>
<tr>
<td>Ile U</td>
<td>+2.15</td>
</tr>
<tr>
<td>Val U</td>
<td>+1.99</td>
</tr>
<tr>
<td>Ala C</td>
<td>+1.94</td>
</tr>
<tr>
<td>Phe U</td>
<td>-0.76</td>
</tr>
<tr>
<td>Cys G</td>
<td>-1.24</td>
</tr>
<tr>
<td>Met U</td>
<td>-1.48</td>
</tr>
<tr>
<td>Thr C</td>
<td>-4.88</td>
</tr>
<tr>
<td>Ser C</td>
<td>-5.06</td>
</tr>
<tr>
<td>Trp G</td>
<td>-5.89</td>
</tr>
<tr>
<td>Tyr A</td>
<td>-6.11</td>
</tr>
<tr>
<td>Glu A</td>
<td>-9.38</td>
</tr>
<tr>
<td>Lys A</td>
<td>-9.52</td>
</tr>
<tr>
<td>Asn A</td>
<td>-9.68</td>
</tr>
<tr>
<td>Glu A</td>
<td>-10.19</td>
</tr>
<tr>
<td>His A</td>
<td>-10.23</td>
</tr>
<tr>
<td>Asp A</td>
<td>-10.92</td>
</tr>
<tr>
<td>Arg G</td>
<td>-15</td>
</tr>
</tbody>
</table>

**Predominantly:**
- inside = C or U
- outside = A (or G)

### 1st codon base:

<table>
<thead>
<tr>
<th>Base</th>
<th>Å</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gly G</td>
<td>3 (small)</td>
</tr>
<tr>
<td>Ala G</td>
<td>14</td>
</tr>
<tr>
<td>Ser A</td>
<td>21</td>
</tr>
<tr>
<td>Cys U</td>
<td>30</td>
</tr>
<tr>
<td>Asp G</td>
<td>30</td>
</tr>
<tr>
<td>Thr A</td>
<td>32</td>
</tr>
<tr>
<td>Val G</td>
<td>36</td>
</tr>
<tr>
<td>Asn A</td>
<td>36</td>
</tr>
<tr>
<td>Glu G</td>
<td>41</td>
</tr>
<tr>
<td>Ile A</td>
<td>46</td>
</tr>
<tr>
<td>Leu U</td>
<td>46</td>
</tr>
<tr>
<td>Glu C</td>
<td>47</td>
</tr>
<tr>
<td>His C</td>
<td>50</td>
</tr>
<tr>
<td>Met A</td>
<td>52</td>
</tr>
<tr>
<td>Lys A</td>
<td>58</td>
</tr>
<tr>
<td>Phe U</td>
<td>62</td>
</tr>
<tr>
<td>Tyr U</td>
<td>69</td>
</tr>
<tr>
<td>Arg C</td>
<td>70</td>
</tr>
<tr>
<td>Trp U</td>
<td>83 (large)</td>
</tr>
</tbody>
</table>

**Predominantly:**
- small = A or G
- large = C or U
**3rd codon base: most degenerate position**

The clustering of AAs by 1st and 2nd codon bases probably reflects original 2-base code, with 3rd base spacer.

Original 16 AAs were “large vs small” and “inside” vs “outside”

Remnants of original code are still evident.

**Multiple alignments**

Considerations:
- Sequences must share some identity
- Amino acid vs. nucleotide alignments
- Substitution matrices / gap penalties
- Cannot make an accurate phylogenetic tree without a high-quality alignment

Example substitution matrix

Phylogenetic trees

Considerations
- Branch lengths
- Boot strap values
- Sampling method
- Replicates
- Outgroups

Multiple alignments

**Motif:** an amino acid sequence pattern that has biological significance
Motifs for structure, post translational modifications, cleavage site, etc.

If you have a protein sequence and you want to predict its function, start by looking for known motifs or domains


Prosite motif database

(http://prosite.expasy.org/)

Sequence vs. motif database or motif vs. database
Links to many other databases (SwissProt, ExPASy, SwissModel, etc.)
Prosite regular expression rules
(http://prosite.expasy.org/)

[AC] encloses: 1 or more alternative symbols (e.g. A or C)
A(3) means: 3 A’s (or whatever) in a row
X(a,b) designates: the lowest (a) and highest (b) possible
number of repeats of the previous (x) symbol
{AC} means: NOT these amino acids (e.g. not A or C)
< means: pattern is valid only if at N-terminus
> means: pattern valid only if found at COOH-terminus

Regular expression language may vary between programs

Prosite motifs use regular expressions

FMDol1k  KQGYCGGAVLAK.DGADTFIVGTHSAG
FMDsat2k  KAGYCGGAVLAK.DGAETFIVGTHSAG
EMCr    RNGWCGSALLADLGGSKK.ILGIHSAG
MengoM  RKGWCGSAILADLGGSKK.ILGFHSAG
TMEGd7   RFGWCGSAIICNVNG.KKAVYGMHSAG
TMEDa    RSGWCGSAIICNVNG.NKAVYGMHSAG

**N-{P}-[ST]-{P}**

Asn glycosylation site

- **N** This pattern looks for **N**
- **{P}** Followed by anything **except P**
- **[ST]** Followed by either **S or T**
- **{P}** Followed by anything **except P**

**CAUTION**: motif searches have >20% false positives because many motifs are too short for good statistics.

[Image: http://www.macalester.edu/psychology/what hap/UBNRP/tse10/levles%20of%20protein.jpg]
2D structure prediction

Helices, sheets, turns, random coils

Why is this a difficult problem?
Structure is context dependent
  alpha helices are local
  beta sheets are long-range

2D ab initio methods
2D homology methods

Stereochemical ab initio modeling:

Amphipathic
(means: hates both)

Helices or sheets have a hydrophobic and a hydrophylic side

These sequences have predictable periodicity
(hydrophobic residues in 1-2-5)

Example: HelicalWheel (in Lasergene Protean)
http://www.biochem.arizona.edu/classes/bioc462/462a/NOTES/LIPIDS/transport.html
**Ab initio statistical modeling**

Observe that certain AA have propensities to be in certain types of structures.

**Example:**

- **Helix formers:** E, M, A, L, K, F, Q, W, I, V
- **Indifferent formers:** D, H, R, T, S, C
- **Helix breakers:** Y, N, P, G
- **Sheet formers:** V, I, Y, F, W, L, C, T, Q, M
- **Indifferent formers:** R, N, H
- **Sheet breakers:** S, G, P, D, E

**Chou & Fasman Rules (generalized)**

**Helices:**

Cluster of 4 helix-formers within length of 6 AA - propagate helix in both directions until at least 4 helix breakers are found.

**β-sheets:**

3/5 β-formers needed to nucleate sheet. In the case of a tie, helix usually wins.

**Turns:**

4 out of 4 AA that prefer turns.

Caution: >50 pgm programs that do this type of prediction. If you get similar answers from several of them, especially for helices, you can probably believe it.
Programs plot the **propensity** of each AA to take a given conformation

Chau & Fasman calculations

Sheet
Helix

3D Structure prediction methods

1. Homology modeling
2. Protein threading
3. *Ab initio (de novo)* approaches

*Internal scoring methods are important*
3D Homology modeling

Compares unknown sequence : sequence with solved structure
Depends on sequence similarity

```
MATTMEQETCAHSLTFEECPKCSALQYRNGFYLKYDEEYYPEELLTDGEDDVFDPELDMEVVFE
```
vs.
```
MATTMEQETCAHSLTFEECPKCSALQYRNGFYLKYDEEYYPEELLTDGEDDVFDPELDMEVVFE
```

3D protein threading

Compares protein:structure template
The unknown protein does not necessarily need to have sequence similarity, only structural
3D ab initio (de novo) methods

Rely entirely on physics, with no structural information from previously solved structures

What’s the difference?

<table>
<thead>
<tr>
<th>Method</th>
<th>Requirements</th>
<th>Computational difficulty</th>
<th>Speed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Homology modeling</td>
<td>Clear homology (&gt;30% id) to a template fold of known structure within the PDB</td>
<td>Easy</td>
<td>Fast</td>
</tr>
<tr>
<td>Threading</td>
<td>A template fold of known structure within the PDB</td>
<td>Medium</td>
<td>Medium</td>
</tr>
<tr>
<td>Ab initio</td>
<td>Target sequence and/or fragment library</td>
<td>Hard</td>
<td>Slow</td>
</tr>
</tbody>
</table>

Adapted from “Protein structure prediction using threading” Xu, J., Jiao, F., Yu, L.. pp. 91-119. Protein structure prediction. 2nd Ed. 2008. Humana Press. pg 63
3D structure prediction

Calicivirus RdRp, contig00002
Homology modeling: MODELLER
Visualized in PyMol

Experimental confirmation of predictions

Things we’ve learned:
Coding sequence
Close relatives and their functions
Potential post-translational modifications / motifs
Potential 2D/3D structure

Now test your predictions!
Algorithms that search for genes look for 3rd base GC-like bias

More GCs at the 3rd position in ORFs
Streptomyces segment encoding ORFs

3D structure prediction programs

<table>
<thead>
<tr>
<th>Program</th>
<th>Freely available</th>
<th>Website</th>
<th>Computing time (days)</th>
<th>Methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zhang Server (p-TASER)</td>
<td>Yes, web</td>
<td><a href="http://zhanglab.cmb.med.unc.edu/p-pTASER.php">http://zhanglab.cmb.med.unc.edu/p-pTASER.php</a></td>
<td>1-2</td>
<td>Threading, GOMETS, all info covering for unannotated regions</td>
</tr>
<tr>
<td>RAPTOR</td>
<td>No, $$$</td>
<td><a href="http://www.cbs.dtu.dk/services/RAPTOR/">http://www.cbs.dtu.dk/services/RAPTOR/</a></td>
<td>???</td>
<td>Combines 3 different threading methods</td>
</tr>
<tr>
<td>BAKER_ROBETTA</td>
<td>Hardware licensed</td>
<td><a href="http://robetta.bakerlab.org/">http://robetta.bakerlab.org/</a></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pm_sp_TASER</td>
<td>Yes, web</td>
<td><a href="http://cash.biology.gatech.edu/">http://cash.biology.gatech.edu/</a></td>
<td>3-4</td>
<td>Single threading method with multiple scoring to identify templates</td>
</tr>
<tr>
<td>METASERVER</td>
<td>Yes, web</td>
<td><a href="http://metaserver.cbi.pku.edu/">http://metaserver.cbi.pku.edu/</a></td>
<td>1-2</td>
<td>Threading template from SPARKS, SP &amp; PROSPECTOR, 3-he input into TASER</td>
</tr>
<tr>
<td>Phyre_de_novo</td>
<td>Yes, web</td>
<td><a href="http://www.sbg.bio.ic.ac.uk/phyre2/">http://www.sbg.bio.ic.ac.uk/phyre2/</a></td>
<td>&lt;1</td>
<td>Homology modeling</td>
</tr>
<tr>
<td>HHpred6</td>
<td>Yes, web</td>
<td><a href="http://groups.sbg.bio.ic.ac.uk/hhpred6.html">http://groups.sbg.bio.ic.ac.uk/hhpred6.html</a></td>
<td>&lt;1</td>
<td>Homology modeling with HHM</td>
</tr>
<tr>
<td>MULTICOM-CLUSTER</td>
<td>Yes, web</td>
<td><a href="http://metaserver.cbi.pku.edu/multicom_3c.html">http://metaserver.cbi.pku.edu/multicom_3c.html</a></td>
<td>&lt;1</td>
<td>Threading (template-based modeling)</td>
</tr>
<tr>
<td>SAM_Ti3-server</td>
<td>Yes, web</td>
<td><a href="http://protein.bio.univmpa.de/SAM_Ti3/Ti3-pygro.html">http://protein.bio.univmpa.de/SAM_Ti3/Ti3-pygro.html</a></td>
<td>&lt;1</td>
<td>Threading with HHM</td>
</tr>
</tbody>
</table>

*Most of the best scoring CASP programs are *metaservers*

Case study: unknown protein

Virgen-Slane et al 2012 PNAS.