CSE 428

Spring 2021

https://courses.cs.washington.edu/courses/cse428/21sp/

Overview

Course Web Pages:

https://courses.cs.washington.edu/courses/cse428/21sp/

TA:

Alyssa La Fleur

Group-Project-oriented:

- Typically teams of ~2-4 students
- We will offer some projects ideas
- We are open to student-generated ideas
  - "computers" + "biology"
  - (+ reasonable scope + something we can facilitate)

Project Challenges

Organization & Scheduling

Bio Jargon

Tools from elsewhere

Did I mention Organization & Scheduling?

The #1 challenge identified by 99% of former students

What I hope you will learn

See previous slide!

You'll see real DNA/RNA seq data in all of them, plus

Some mixture of:

- data structures,
- algorithms,
- data analytics,
- statistics,
- biology,
- HCI,
- ML, …

Project Evaluation

Weekly Goals + Progress reports

Some midcourse checkpoint

Final written reports + oral presentations

  Including evaluation of code, test results, etc.

Peer comments

Project Idea: Next Few Slides

Open-ended, underspecified; as you think about them, both let your imagination run free, *and* think carefully about how to scale and stage your project so you can collect low-hanging fruit before potentially getting lost in the open-ended weeds. (Fortunately, I don't think mixing metaphors is a crime in this state–yet.)

Misc. Projects From 428's Past

Just to give you some idea of scope, here are some projects from previous iterations of 428:

- Convenient web interface for "phylogenetic footprinting" in prokaryotes
- Build a genome assembler
- Machine learning applied to cancer genomics
- Convenient web interface for exploring "Foldit" results
- Visualization of technical biases in RNAseq
- Downstream impact of technical biases in RNAseq
- Crossover detection in DNAseq data

Discovery of regulatory non-coding RNA ("ribosomal leaders") in lower Eukaryotes

* You might remember my "L19" example from 427 – in some bacteria, excess L19 down-regulates itself by binding to its own mRNA
* This kind of thing is widespread in Prokaryotes
* Few if any examples are known in Eukaryotes
* But I speculate that one group of single-celled Euks is a strong contender to show this behavior
* Chances are I'm wrong! But should be "fun" to try; you'll see real data, a variety of state-of-the-art algorithms, and methods to evaluate your results

2021 Project Idea #2

Deep learning for non-coding RNA discovery and classification.

* ncRNA is an immense landscape that is radically changing our understanding of molecular biology, esp. regulation
* But really hard problems
* Deep learning (DL) is sweeping the world, on hard problems
* Initial results suggest DL can be faster and better at ncRNA discovery/classification tasks than classical statistical methods
* Do you believe it?
* Initial DL architectures seem pretty naive: can you improve them?
* Again, look at real algorithms, real data, cutting-edge problems

2021 Project Idea #3

You may have an idea that you want to pursue, and preferably recruit a partner or three to help

Again, we're open to this, provided its got a reasonable blend of Comp + Bio, reasonably scoped, and something that we can facilitate.

More Details
Idea #1:
Ribosomal Leaders

Hypothesis

High copy-number, multi-protein complexes may benefit from stoichiometric control of their components

Ribosomes are high copy-number, multi-protein complexes

Widespread use of "ribosomal leader" auto regulation in prokaryotes exemplifies this (but details vary)

Eukaryotes may accomplish this by other means, although a few putative examples are known

My suggestion: Unique biology of ciliates probably greatly exacerbates the stoichiometry problem, making them a prime target for discovery of Eukaryotic ribosomal leader auto-regulation.

NATURE VOL. 227 AUGUST 8 1970

Central Dogma of Molecular Biology

by
FRANCIS CRICK
MRC Laboratory
Hills Road,
Cambridge CB2 2QH

The central dogma of molecular biology deals with the detailed residue-by-residue transfer of sequential information. It states that such information cannot be transferred from protein to either protein or nucleic acid.

"The central dogma, enunciated by Crick in 1958 and the keystone of molecular biology ever since, is likely to prove a considerable over-simplification."

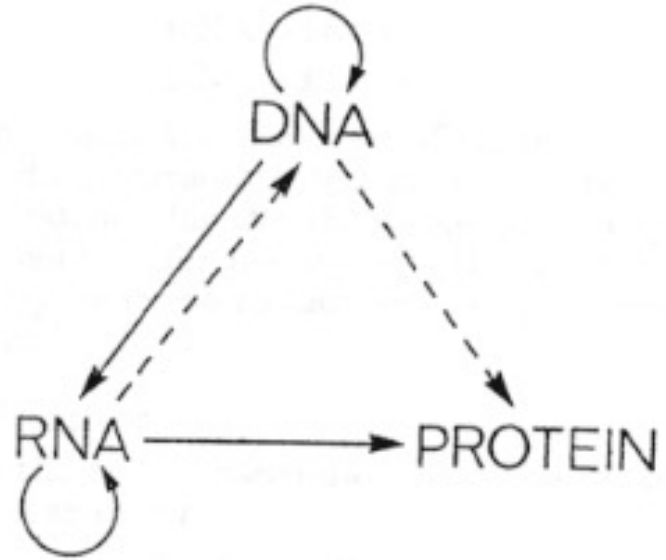
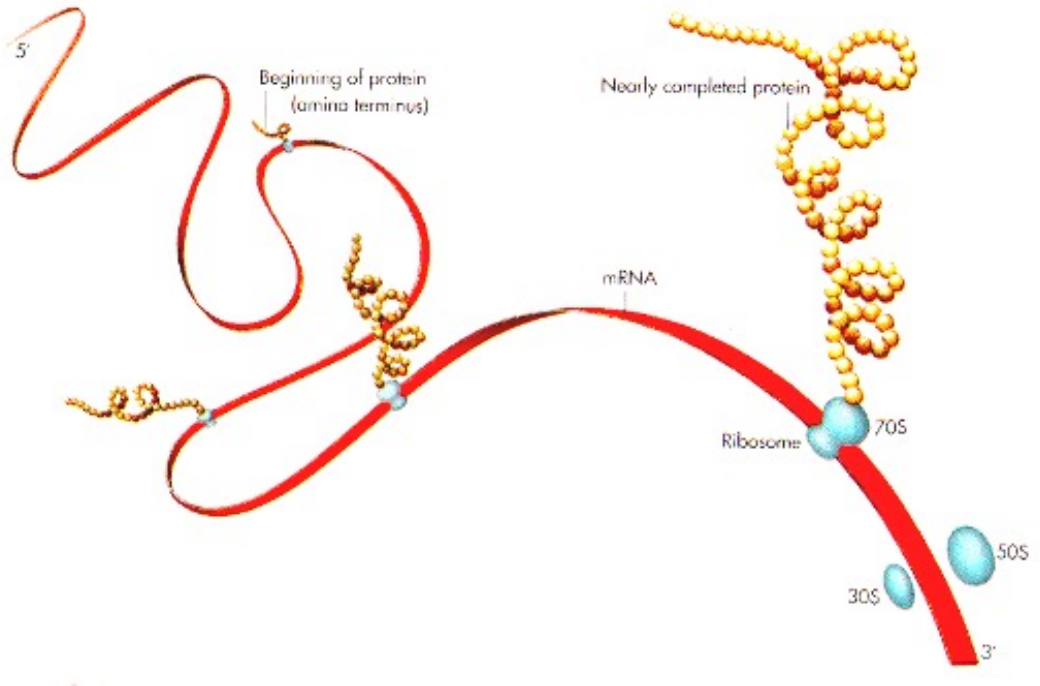


Fig. 2. The arrows show the situation as it seemed in 1958. Solid arrows represent probable transfers, dotted arrows possible transfers. The absent arrows (compare Fig. 1) represent the impossible transfers postulated by the central dogma. They are the three possible arrows starting from protein.

Translation: mRNA → Protein



Watson, Gilman, Witkowski, & Zoller, 1992

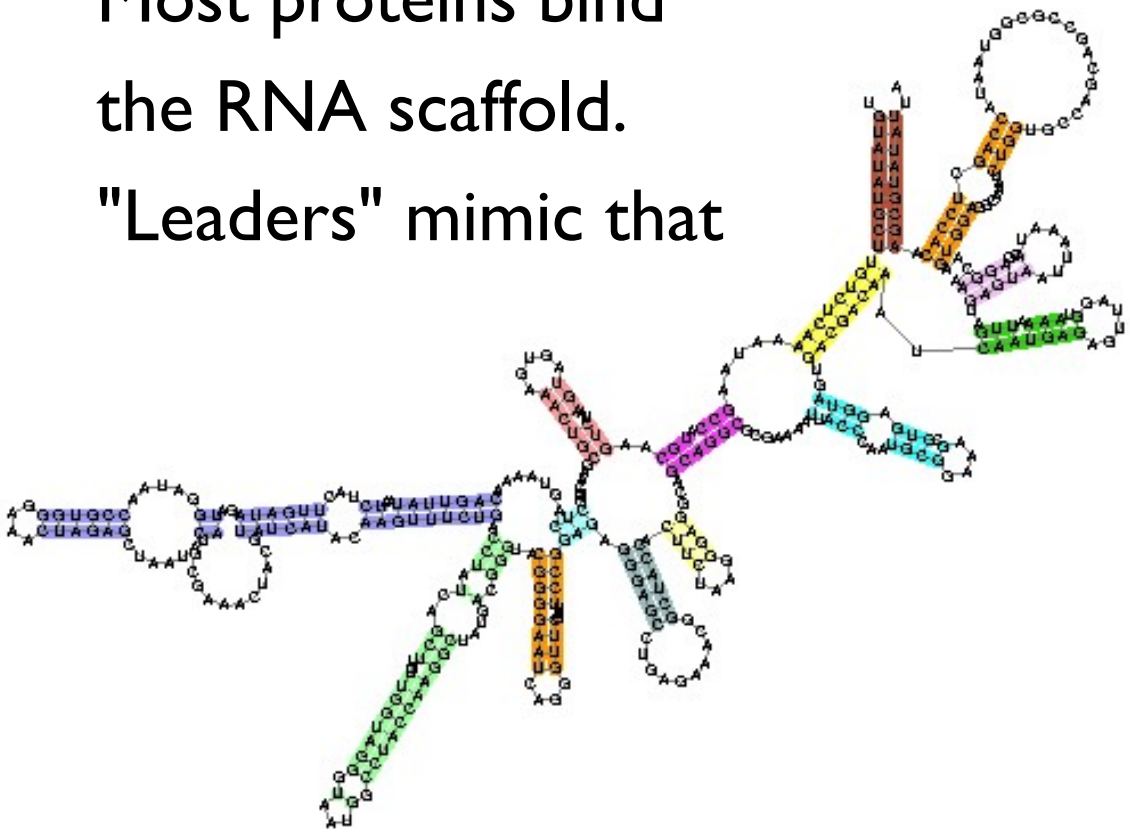
Ribosome is a large complex,

~ 1/2 RNA, 1/2 protein

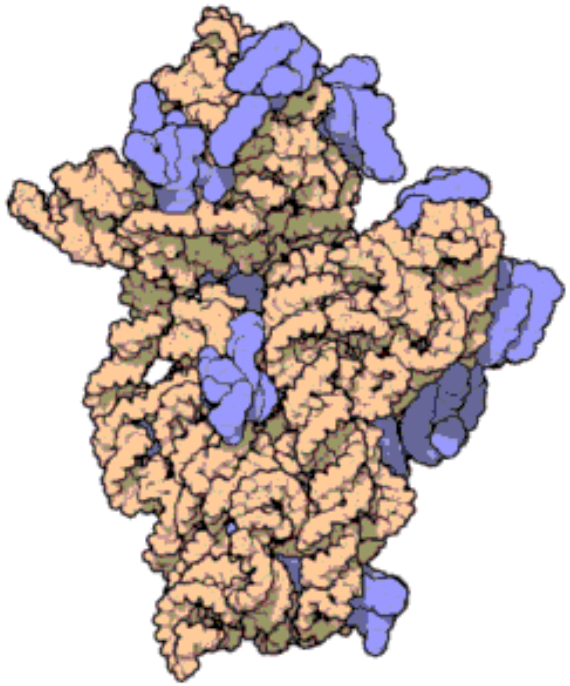
Most proteins bind

the RNA scaffold.

"Leaders" mimic that



Small subunit ribosomal RNA, 5' domain taken from the Rfam database. This example is RF00177, a fragment from an uncultured bacterium.



An example of a fully-assembled small subunit of ribosomal RNA in prokaryotes, specifically *Thermus Thermophilus*. The actual ribosomal RNA (16S) is shown coiled in orange with ribosomal proteins attaching in blue.

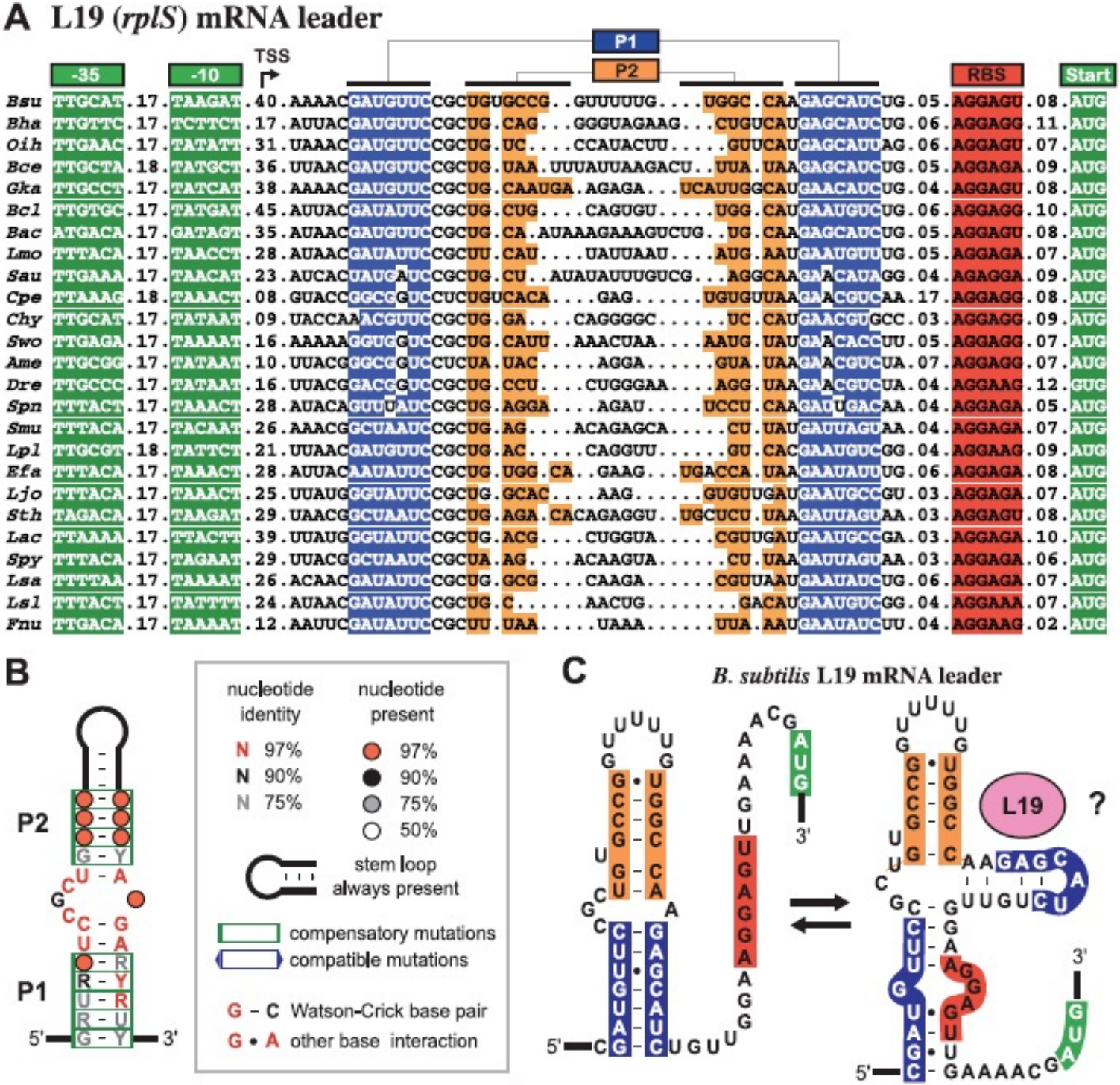


Figure 3. Putative Autoregulatory Structure in L19 mRNA Leaders

Eckert and Weinberg *BMC Microbiology* (2020) 20:130
https://doi.org/10.1186/s12866-020-01823-6

BMC

RESEARCH ARTICLE

Discovery of 20 novel ribosomal leader candidates in bacteria and archaea

Iris Eckert and Zasha Weinberg*

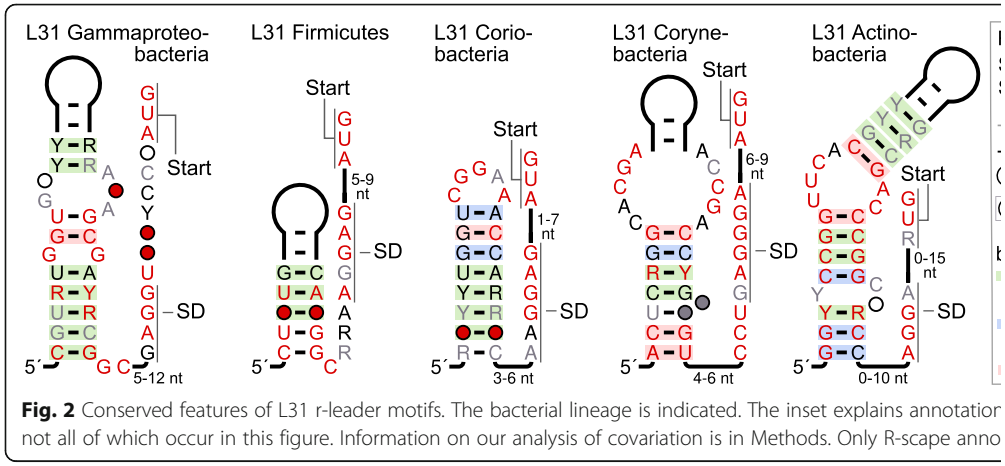
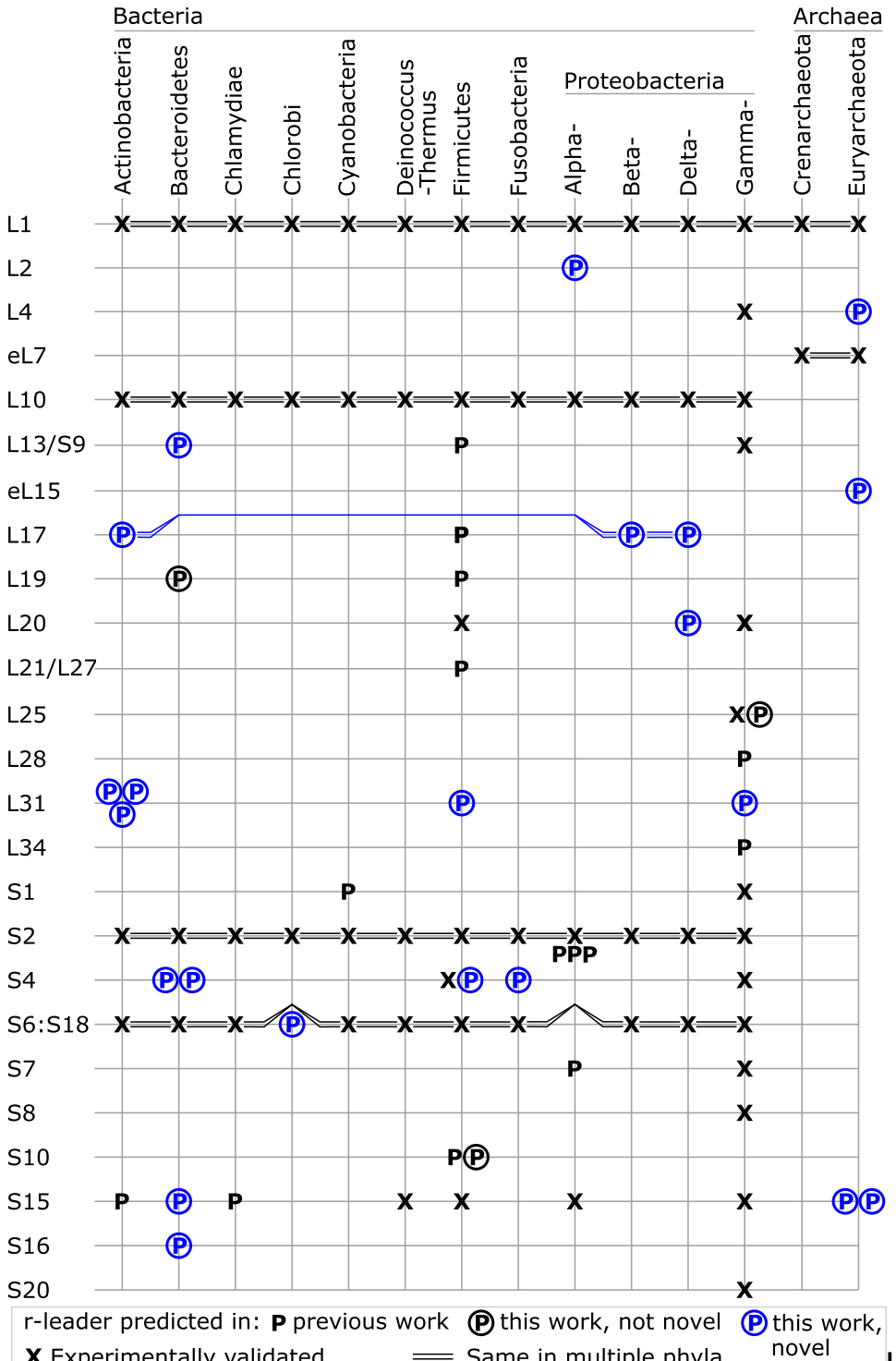


Fig. 2 Conserved features of L31 r-leader motifs. The bacterial lineage is indicated. The inset explains annotation not all of which occur in this figure. Information on our analysis of covariation is in Methods. Only R-scape anno



Ciliates

Large, single-celled Eukaryotes

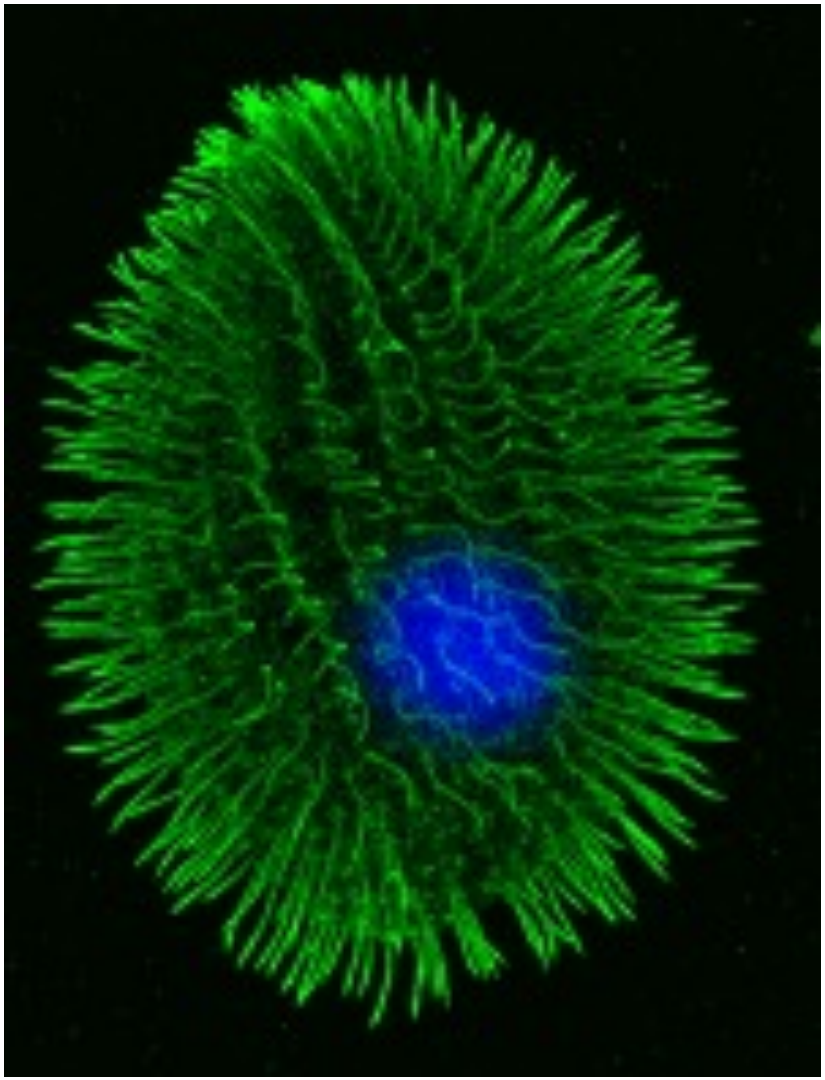
Reproduce by fission, with occasional sexual cycles ("conjugation")

Two nuclei:

MIC: typical diploid nuc with 5-10 pairs of large chromosomes, but *transcriptionally silent*

MAC: copy of MIC, but processed to have many copies of many mini-chromosomes (e.g. in Tt, several hundred, w/ ~45 copies of each, and >10^3 copies of rDNA) *transcriptionally active*

In fission, MIC is usual mitotic div, but MAC is amitotic–pinches in half w/ semi-random segregation of chromosome copies to each daughter cell. (Restored by conjugation.)



Tetrahymena thermophila

Outline/Challenges

Pick genes; for each:

Find genomes

Find chosen gene in multiple genomes

Download sequences (5' UTR, especially)

Validate them?

Prep them for ncRNA "discovery"

Predict:

- CMfinder
- Rscape, etc.

Evaluation!!

- Multiperm, SISSIz, control genes, outgroup, paralogs, visualize, Rscape, …

Other directions: introns, 3' UTR, CDS, Other complexes (spliceosome, maybe? RNA or DNA Pol, telomerase?)

Chances of success:

???

… But as Dr. Laughlin said, nature might have the last laugh. “Given the rule of thumb that 99 percent of one’s own cool ideas tend not to work out,” he said, “I think the smart money [is against us]." …

https://www.nytimes.com/2021/03/23/science/astronomy-oumuamua-comet.html?referringSource=articleShare

More Details
Idea #2:
Deep Learning for ncRNA Classification & Discovery

See separate .pdf file linked from course home page as "Idea #2"

Idea #3:
Yours?

Next steps

review slides

which (if any) appeals?

form groups

skim references on web

talk to/email me/Alyssa

we may have fragments of code for parts of this

(may or may not be useful…)

Form a Group/Form a Plan!

Next Steps

Set up CSE GITLab Repo
Share with ruzzo@cs and lafleur1@cs (and teammates)
create a "group-info.txt" doc with
  team members names & emails, other contact info
  members roles as they become defined
Create a "Progress" doc for *weekly reports:*
  goals for next week
  review of what did/didn't get accomplished last week
Bibliography
Make a "Plan" (what needs to be done, in what order); revise it as you go

New Results

Fifty generations of amitosis: tracing asymmetric allele segregation in polyploid cells with single-cell DNA sequencing

Valerio Vitali, Rebecca Rothering, Francesco Catania

doi: https://doi.org/10.1101/2021.03.29.437473

Posted March 30, 2021.