Advanced Bioinformatics

Biostatistics & Medical Informatics 776
Computer Sciences 776
Spring 2017

Anthony Gitter

gitter@biostat.wisc.edu

www.biostat.wisc.edu/bmi776/

Agenda Today

- Introductions
- Course information
- Overview of topics

Course Web Site

- www.biostat.wisc.edu/bmi776/
- Syllabus and policies
- Readings
- Tentative schedule
- Lecture slides (draft posted before lecture)
- Announcements
- Homework
- Project information
- Link to Piazza discussion board

Your Instructor: Anthony Gitter

- Email: gitter@biostat.wisc.edu

- Office: room 3268, Discovery Building

- Assistant professor, Biostatistics & Medical Informatics
- Affiliate faculty, Computer Sciences
- Investigator, Morgridge Institute for Research

- Research interests: biological networks, time series analysis, computational problems related to cancer and virology

Your TA: Li Liu

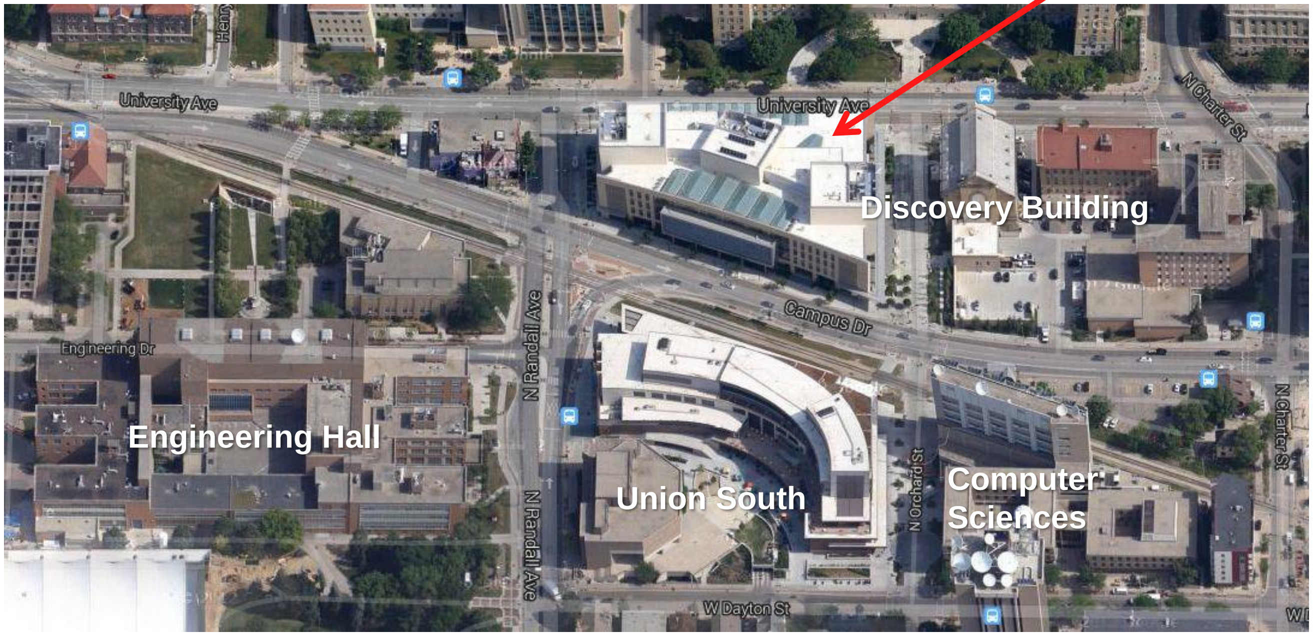
- Email: lliu262@wisc.edu
- Office: CS building, room to be announced
- Graduate student, Computer Sciences

Office Hours

- Instructor: Tuesday and Thursday, 2:30-3:30 PM
  - Immediately after class
- TA: Will be announced soon

Finding My Office: Discovery Building

my office



- 3^{rd} floor has restricted access
- See Piazza if you need building access
- Stop at visitor desk to call my office if card does not work

You

- So that we can all get to know each other better, please tell us your
  - name
  - major or graduate program
  - research interests and/or topics you’re especially interested in learning about
  - favorite programming language

Course Requirements

- 4 or 5 homework assignments: ~40%
  - Written exercises
  - Programming (Python)
  - Computational experiments (e.g. measure the effect of varying parameter *x* in algorithm *y*)
  - Five late days permitted

- Project: ~25%
- Midterm: ~15%
- Final exam: ~15%
- Class participation: ~5%

Exams

- Midterm: March 7, in class
- Final: Sunday May 7, 2:45-4:45 PM

- Let me know *immediately* if you have a conflict with either of these exam times

Computing Resources for the Class

- Linux workstations in Dept. of Biostatistics & Medical Informatics
  - No “lab”, must log in remotely (use WiscVPN)
  - Will create accounts for everyone on course roster
  - Two machines

    mi1.biostat.wisc.edu

    mi2.biostat.wisc.edu
  - HW0 tests your access to these machines
  - Homework must be able to run on these machines

- CS department usually offers Unix orientation sessions at beginning of semester

Programming Assignments

- Piazza poll supports using Python
  - 4 have used Python
  - 8 willing to learn
  - 0 strongly opposed
  - Many abstained, any more input?

- Will set up Python environment on biostat servers
  - Debating Python 2 vs. Python 3
  - Will install permitted packages and post the list

- HW0 will be Python introduction

- Use Piazza for Python discussion

Project

- Design and implement a new computational method for a task in molecular biology
- Improve an existing method
- Perform an evaluation of several existing methods
- Run on real biological data
- Suggestions will be provided
- Each student works individually
- Not directly related to your existing research

Participation

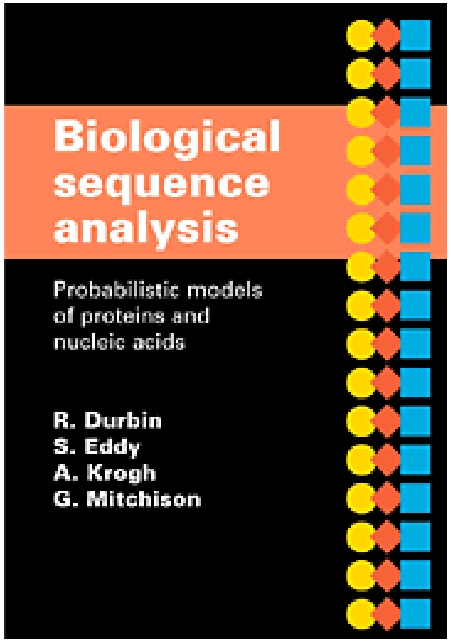
- Do the assigned readings
- Show up to class
- No one will have the perfect background
  - Ask questions about computational or biological concepts
- Correct me when I am wrong
- Piazza discussion board

Piazza Discussion Board

- Instead of a mailing list
- http://piazza.com/wisc/spring2017/bmics776/home
- Post your questions to Piazza instead of emailing the instructor or TA
  - Unless it is a private issue
- Answer your classmates' questions
- Announcements will also be posted to Piazza
- Supplementary material for lecture topics

Course Readings

- Mostly articles from the primary literature
- Must be using a campus IP address to download some of the articles (can use WiscVPN from off campus)
- *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. Cambridge University Press, 1998.



Biological sequence analysis

Probabilistic models of proteins and nucleic acids

R. Durbin
S. Eddy
A. Krogh
G. Mitchison

Prerequisites

- BMI/CS 576 or equivalent
- Knowledge of basic biology and methods from that course will be assumed
- May want to go over the material on the 576 website to refresh
- http://www.biostat.wisc.edu/bmi576/

What you should get out of this course

- An understanding of some of the major problems in computational molecular biology
- Familiarity with the algorithms and statistical techniques for addressing these problems
- How to think about different data types
- At the end you should be able to
  - Read the bioinformatics literature
  - Apply the methods you have learned to other problems both within and outside of bioinformatics

Major Topics to be Covered (the algorithms perspective)

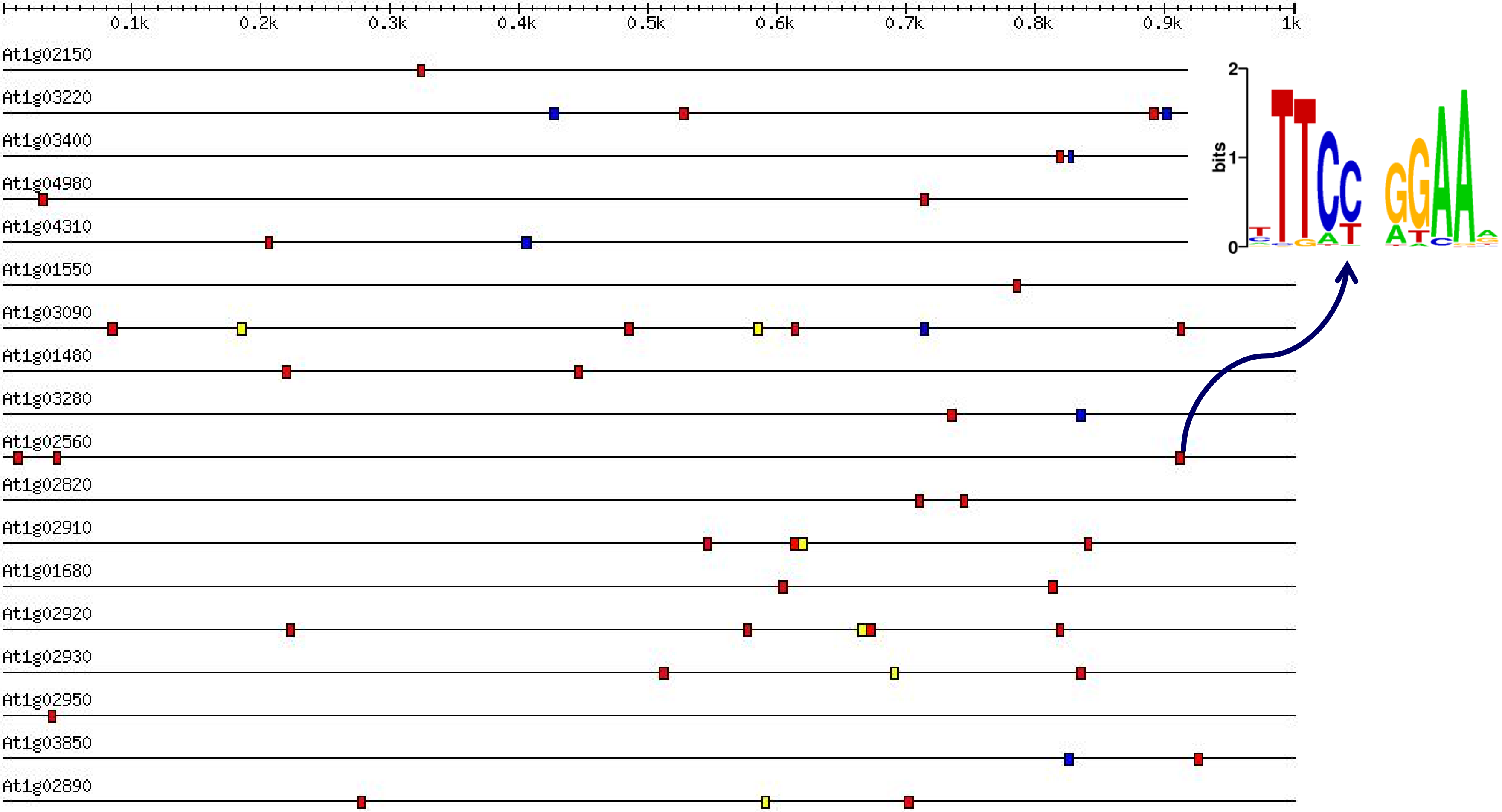
- Expectation Maximization
- Gibbs sampling
- Mutual information
- Multiple hypothesis testing correction
- Convolutional neural networks
- Linear programming
- Interpolated Markov models
- Duration modeling and semi-Markov models
- Tries and suffix trees
- Markov random fields
- Stochastic context free grammars
- Branch and bound search

Major Topics to be Covered (the task perspective)

- Modeling of motifs and *cis*-regulatory modules
- Identification of transcription factor binding sites
- Genotype analysis and association studies
- Regulatory information in epigenomic data
- Transcriptome quantification
- Mass spectrometry peptide and protein identification
- Pathways in cellular networks
- Gene finding
- Large-scale sequence alignment
- RNA sequence and structure modeling
- Protein structure prediction

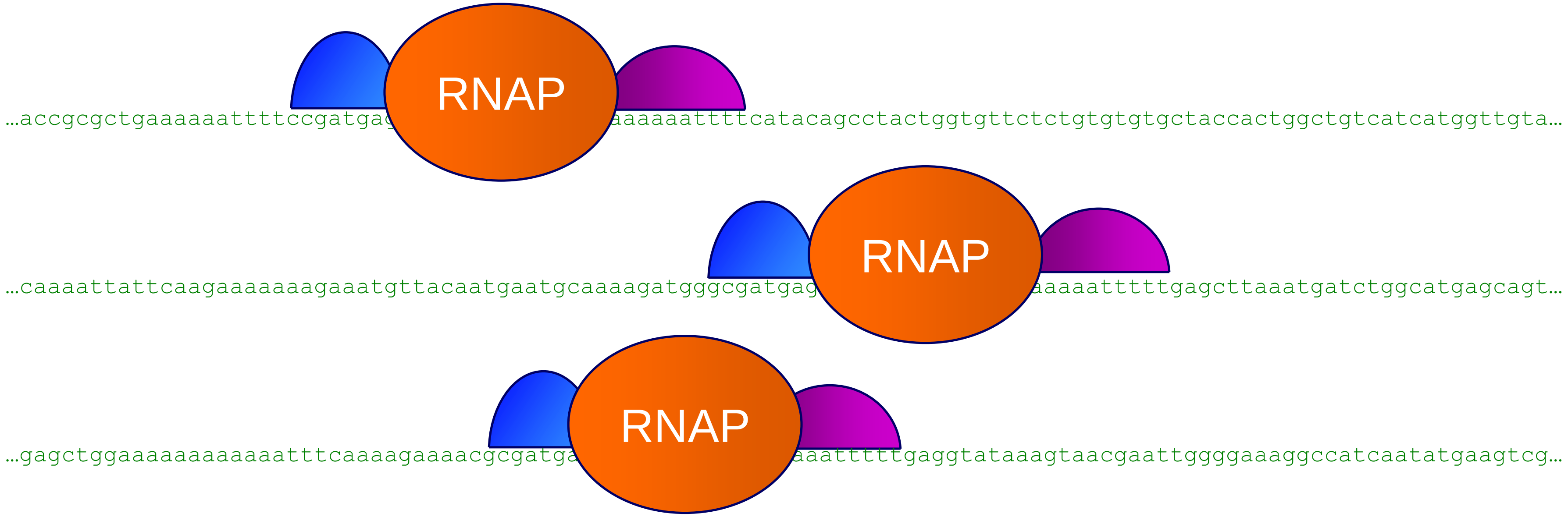
Motif Modeling

What sequence motif do these promoter regions have in common?



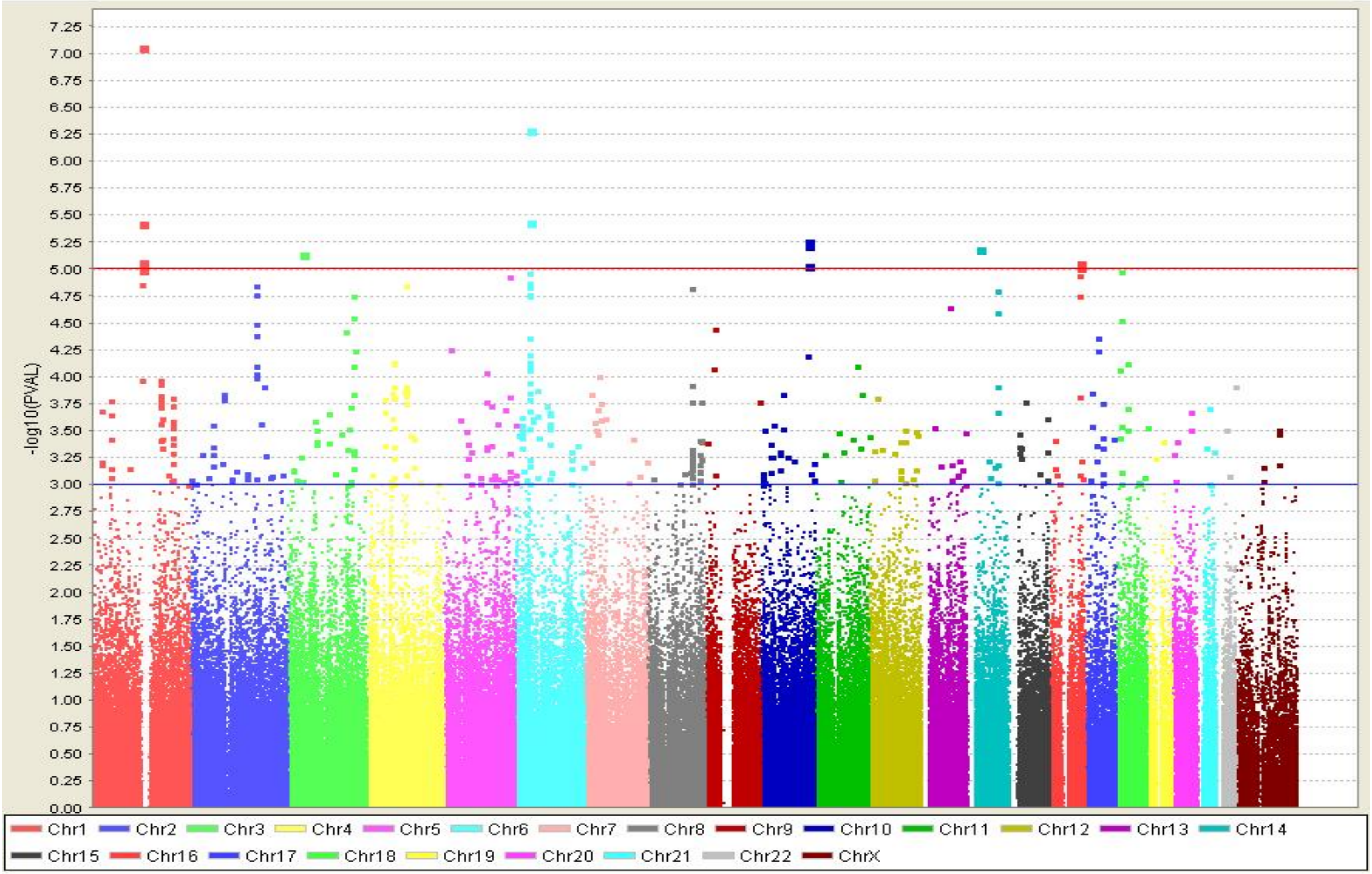
cis-Regulatory Modules

What configuration of sequence motifs do these promoter regions have in common?


RNAP
…accgcgctgaaaaaattttccgatgag
aaaaaattttcatacagcctactggtgttctctgtgtgtgctaccactggctgtcatcatggttgta…
RNAP
…caaaattattcaagaaaaaaagaaatgttacaatgaatgcaaaagatgggcgatgag
aaaaattttttgagcttaaatgatctggcatgagcagt…
RNAP
…gagctggaaaaaaaaaaaatttcaaaagaaaacgcgatga
aaatttttgaggtataaagtaacgaattggggaaaggccatcaatatgaagtcg…

Genome-wide Association Studies

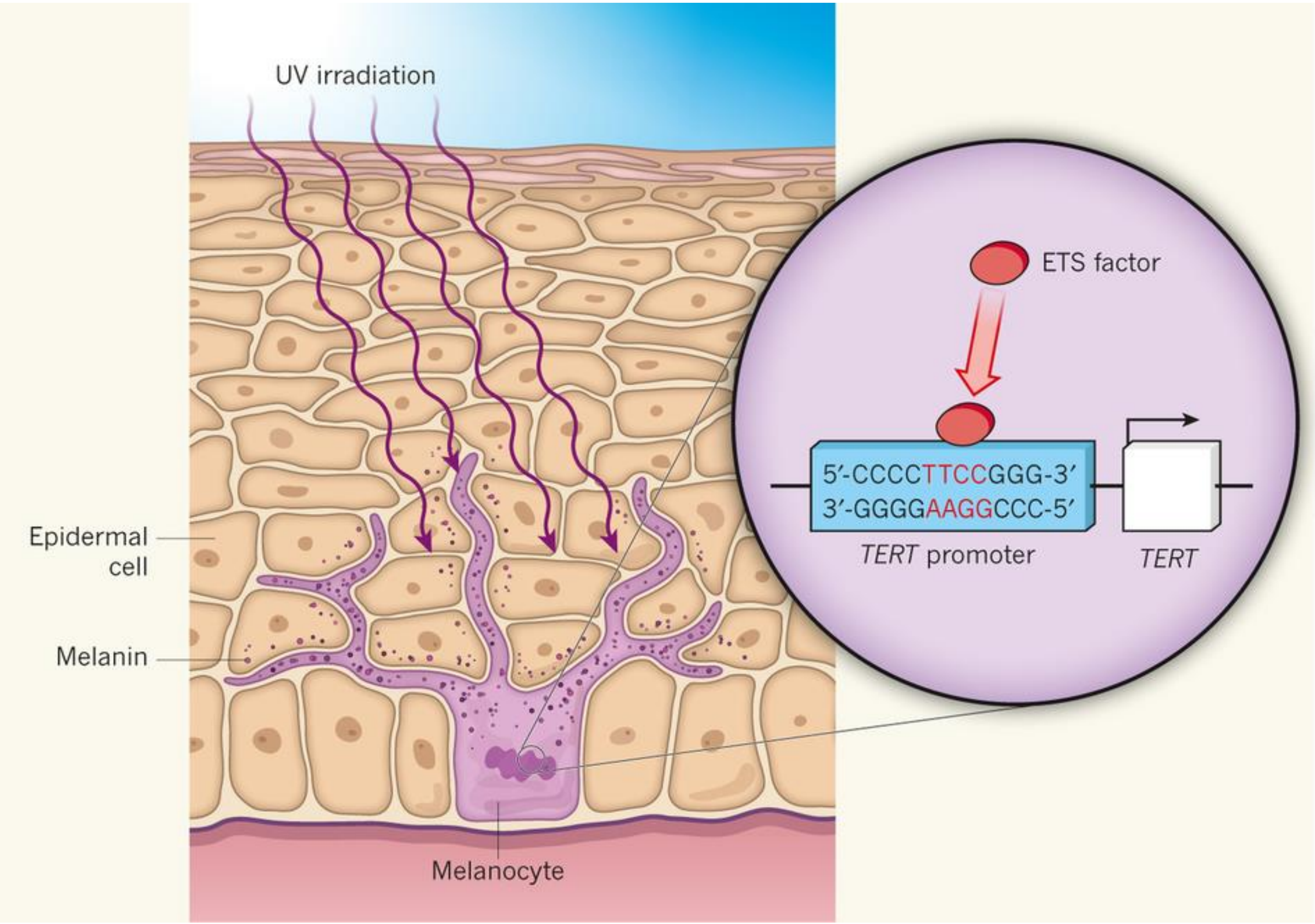
Which genes are involved in diabetes?



Type 2 diabetes association P values by chromosome (386,731 markers). The x-axis is the genomic position by chromosome 1-22 and X (by color), and the y-axis is the negative base 10 logarithm of the P value.

Noncoding Genetic Variants

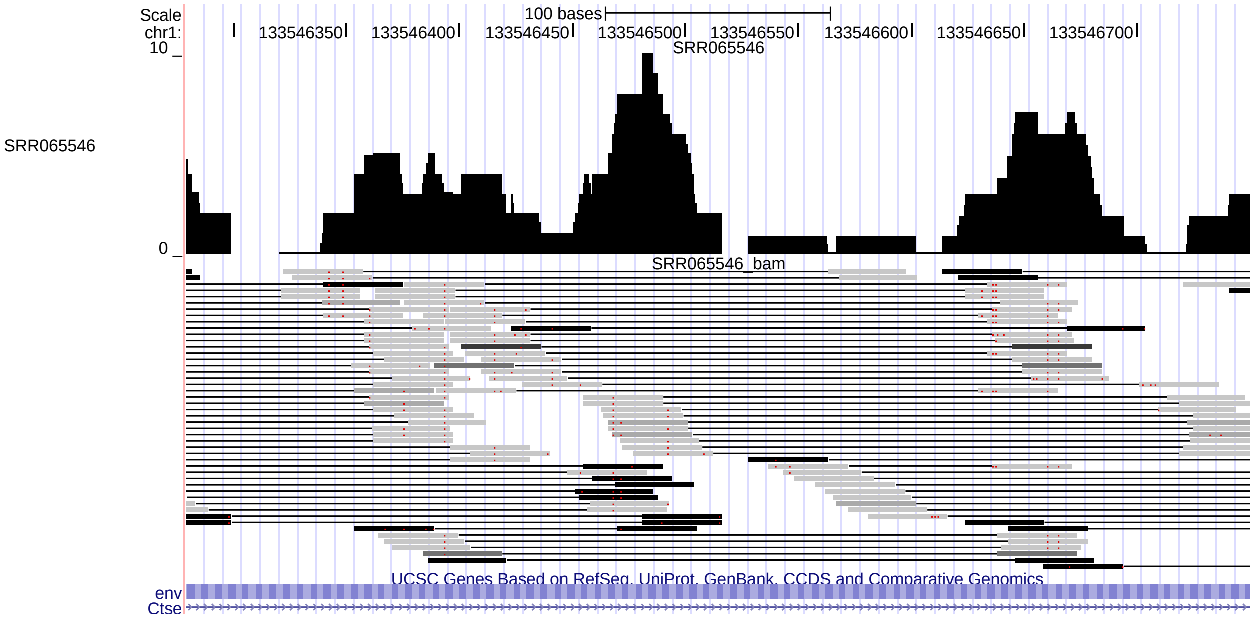
How do genetic variants outside protein coding regions impact phenotypes?



Patton and Harrington, *Nature*, 2013

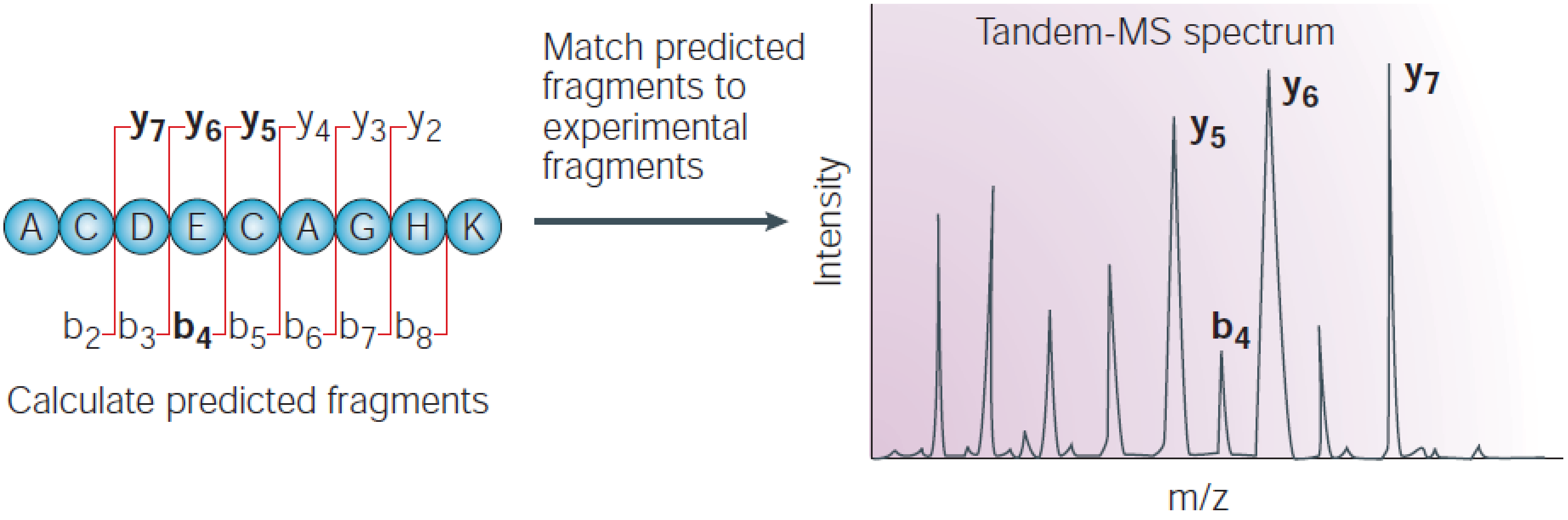
Transcriptome Analysis with RNA-Seq

What genes are expressed and at what levels?



Proteomic Analysis with Mass Spectrometry

What proteins are expressed and at what levels?

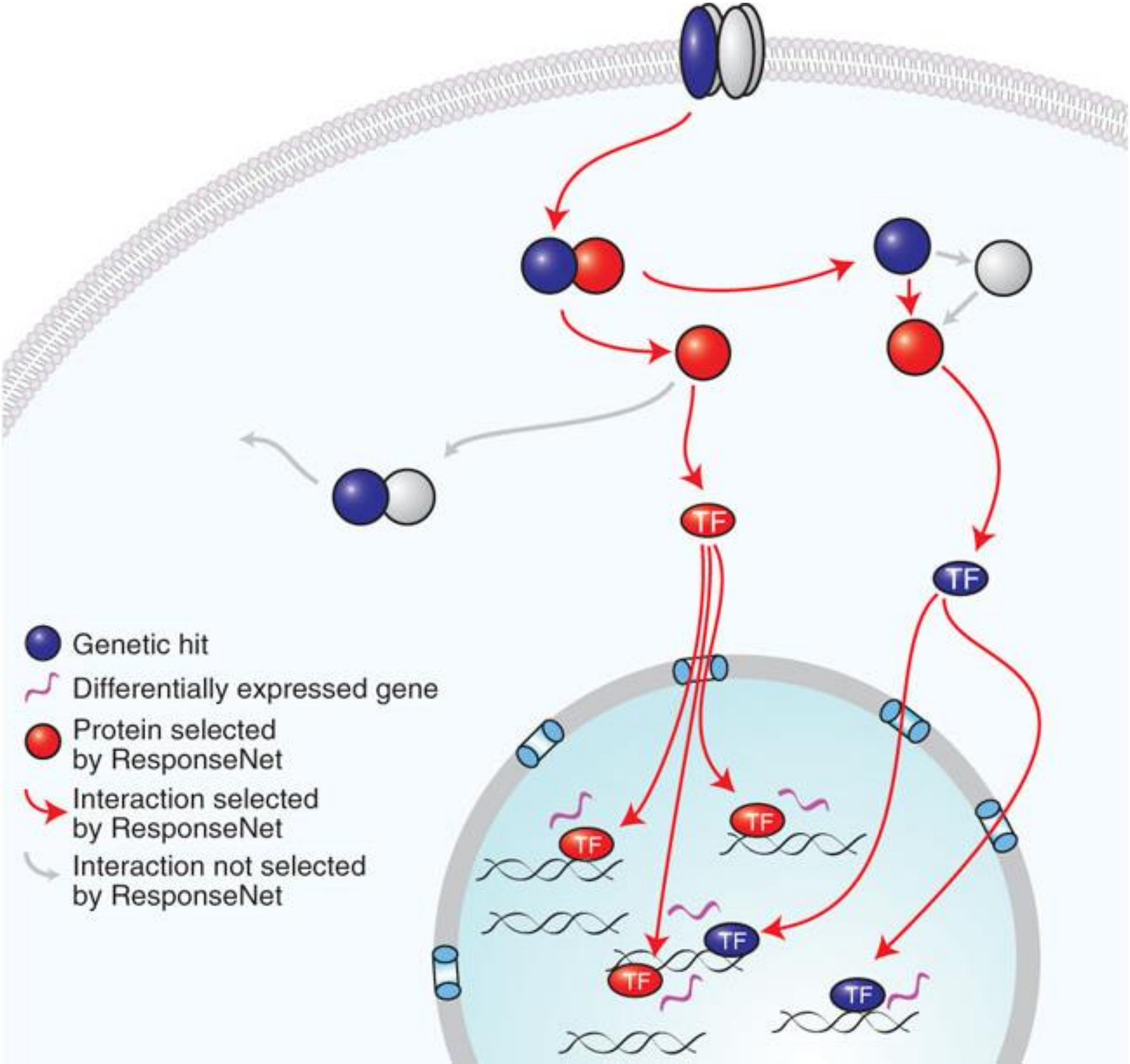


y_7 y_6 y_5 y_4 y_3 y_2
A C D E C A G H K
b_2 b_3 b_4 b_5 b_6 b_7 b_8
Calculate predicted fragments
Match predicted fragments to experimental fragments
Tandem-MS spectrum
Intensity
y_5
y_6
y_7
b_4
m/z

Steen and Mann, *Nature Reviews Molecular Cell Biology*, 2004

Identifying Signaling Pathways

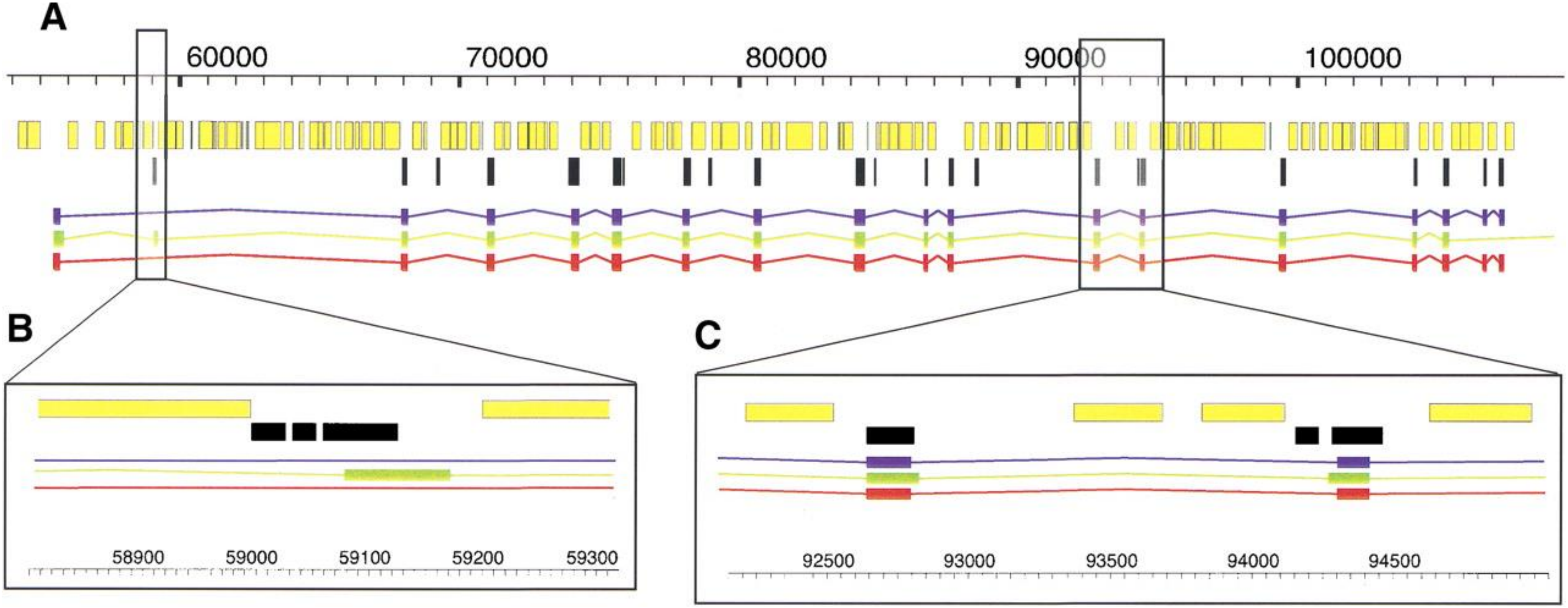
How do proteins coordinate to transmit information?


Genetic hit
Differentially expressed gene
Protein selected by ResponseNet
Interaction selected by ResponseNet
Interaction not selected by ResponseNet
TF
TF
TF
TF
TF
TF
TF

Yeger-Lotem et al., *Nature Genetics,* 2009

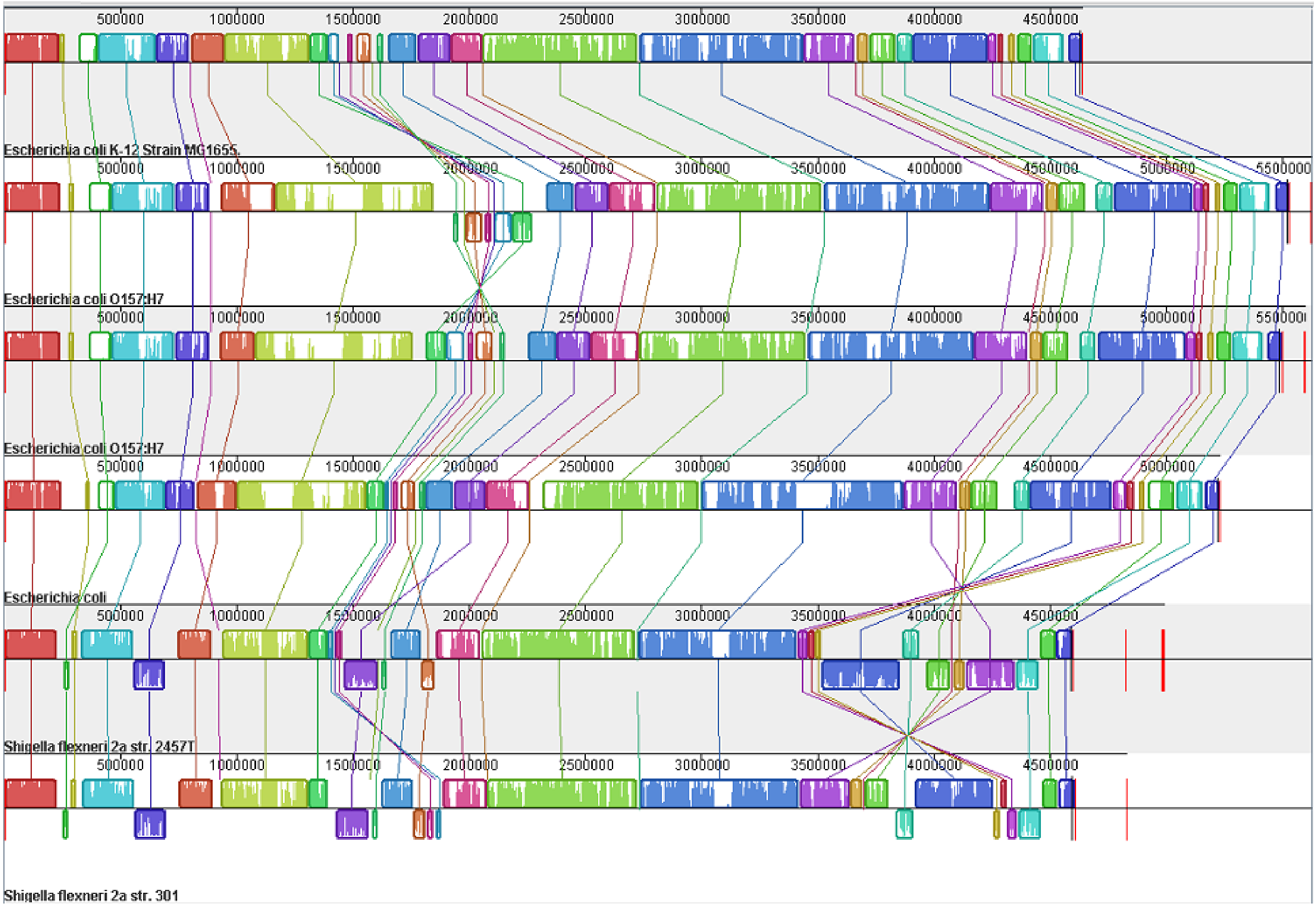
Gene Finding

Where are the genes and functional elements in a genome?



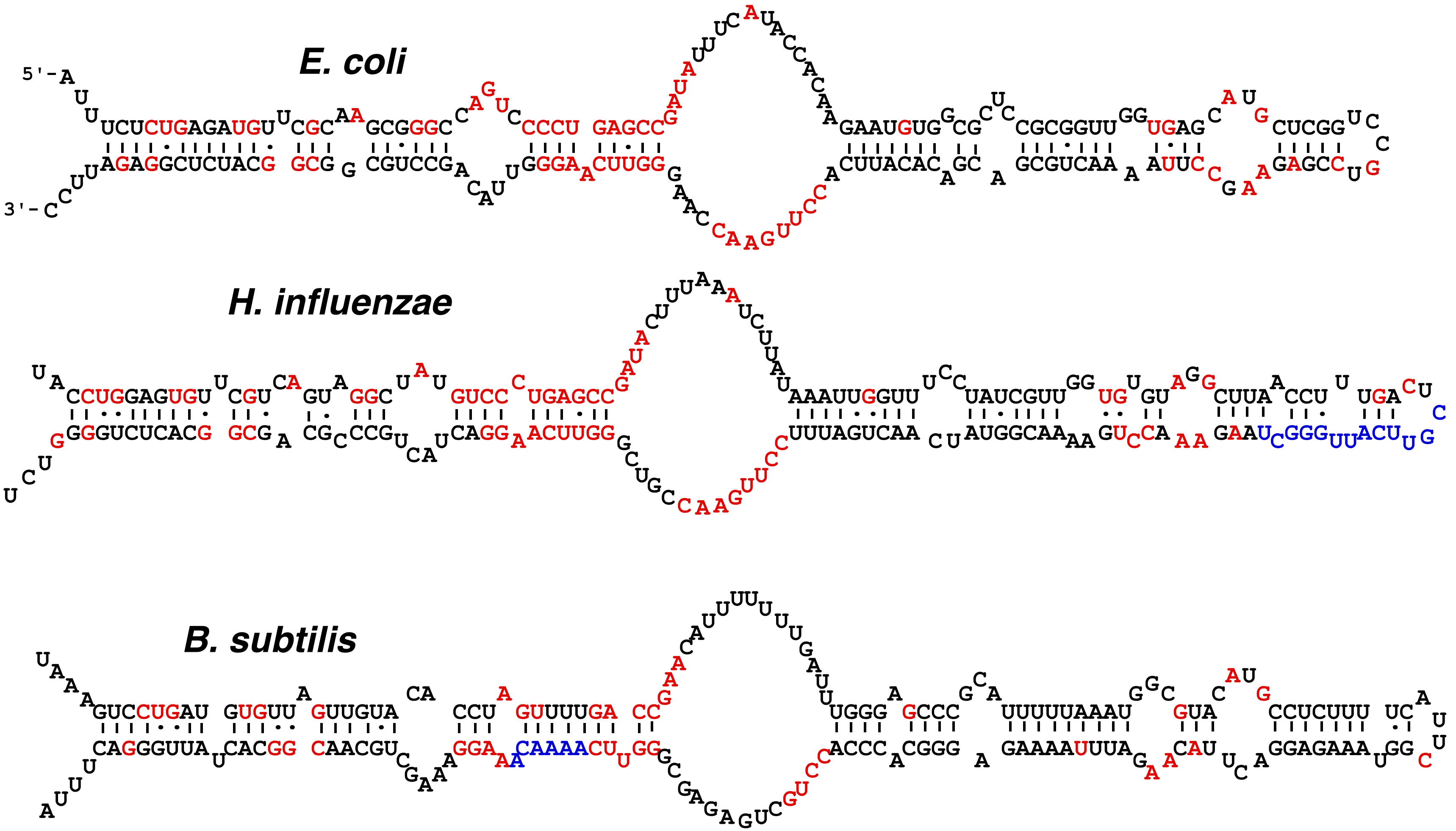
Large Scale Sequence Alignment

What is the best alignment of these 6 genomes?



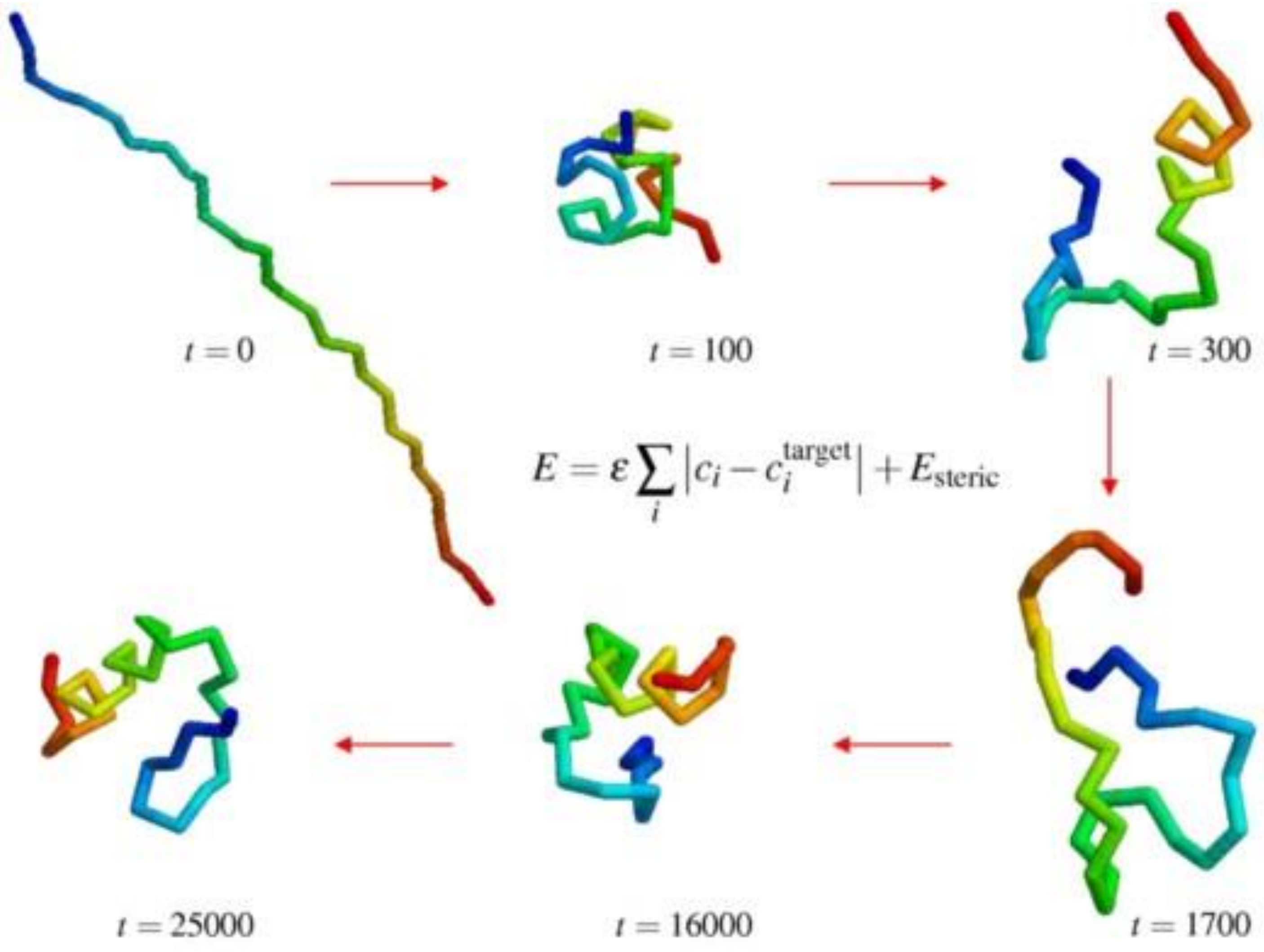
RNA Sequence and Structure Modeling

How can we identify sequences that encode this RNA structure?



Protein Structure Prediction

Can we predict the 3D shape of a protein from its sequence?


$t = 0$
$t = 100$
$t = 300$
$E = \varepsilon \sum_i \left| c_i - c_i^{\mathrm{target}} \right| + E_{\mathrm{steric}}$
$t = 1700$
$t = 16000$
$t = 25000$

Other Topics

- Many topics we aren’t covering
  - Protein function annotation
  - Modeling long reads
  - Metagenomics
  - Metabolomics
  - Sequence compression
  - Graph genomes
  - Single-cell sequencing
  - Pseudo- and quasi-alignment
  - Text mining
  - Others?

Reading Groups

- Computational Systems Biology Reading Group
  - http://lists.discovery.wisc.edu/mailman/listinfo/compsysbiojc
- AI Reading Group
  - http://lists.cs.wisc.edu/mailman/listinfo/airg

- Will also announce relevant seminars