Advanced Bioinformatics

Biostatistics & Medical Informatics 776
Computer Sciences 776
Spring 2015

Colin Dewey

Dept. of Biostatistics & Medical Informatics

Dept. of Computer Sciences

cdewey@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
- readings
- tentative schedule
- lecture slides in PDF/PPT
- homework
- project information
- link to Piazza discussion board
- etc.

Your Instructor: Colin Dewey

- email:
  cdewey@biostat.wisc.edu

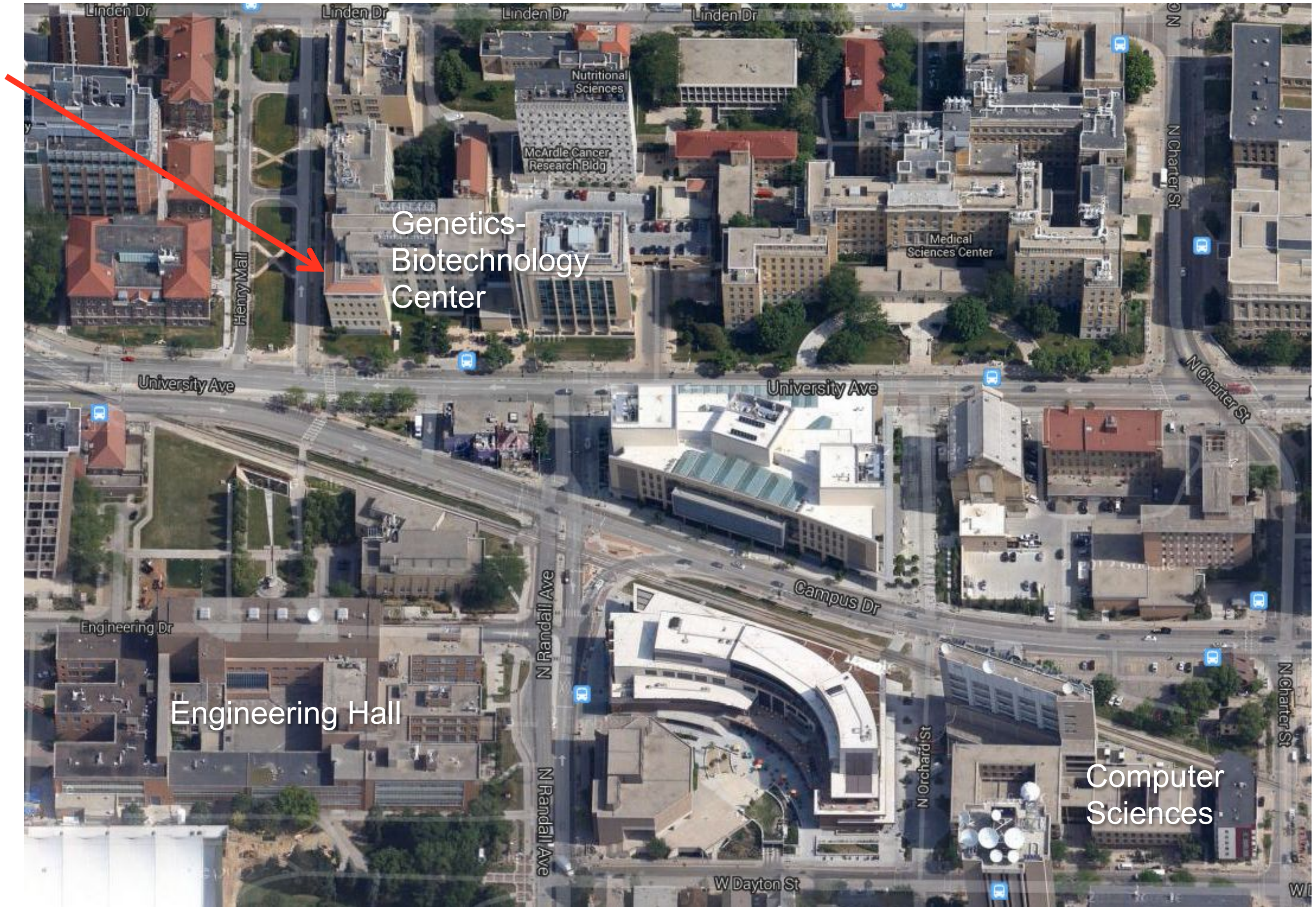
- office hours: Mon 2-3pm, Thu 11am-12pm, room 2128, Genetics-Biotechnology Center

- Associate professor in the department of Biostatistics & Medical Informatics with an affiliate appointment in Computer Sciences

- research interests: probabilistic modeling, biological sequence evolution, analysis of “next-generation” sequencing data (RNA-Seq in particular), whole-genome alignment

Finding My Office: 2128 Genetics-Biotechnology Center

my office



- slightly confusing building(s)
- best bet: use Henry Mall main entrance

Y’all!

- 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 in this class
  - favorite take-out/delivery restaurant in town

Course Requirements

- 5 or so homework assignments: ~40%
  - written exercises
  - programming (in Java, C++, C, Perl, Python) + computational experiments (e.g. measure the effect of varying parameter *x* in algorithm *y*)
  - paper critiques
    - major strength of approach
    - major weakness
    - what would you do next
- project: ~25%
- midterm: ~15%
- final exam: ~15%
- class participation: ~5%

Exams

- Midterm: March 10th, in class
- Final: May 10th, 10:05am-12:10pm
- Please let me know *immediately* if you have a conflict with either of these exam times

Project

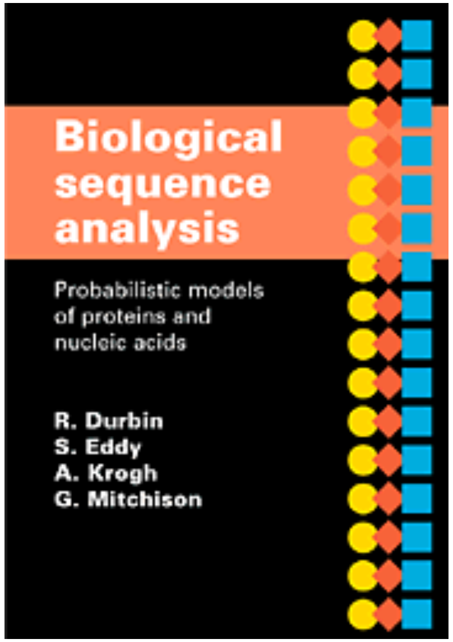
- Design and implement a new computational method for a task in molecular biology
- Or, perform an evaluation of several existing methods
- Run on real biological data
- Project proposal due on March 17th
- Some project suggestions listed on website

Participation

- take advantage of the small class size!
- do the assigned readings
- show up to class
- don’t be afraid to ask questions

Course Readings

- mostly articles from the primary literature (scientific journals, etc.)
- must be using a UW 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/

Computing Resources for the Class

- Linux workstations in Dept. of Biostatistics & Medical Informatics
  - no “lab”, must log in remotely (use WiscVPN)
  - most of you have accounts?
  - two machines

    mi1.biostat.wisc.edu

    mi2.biostat.wisc.edu
- CS department usually offers UNIX orientation sessions at beginning of semester

Piazza Discussion Forum

- Instead of a mailing list
- http://piazza.com/wisc/spring2015/bmics776/home
- Please consider posting your questions to Piazza first, before emailing the instructor
- Also consider answering your classmates' questions!
- Quick announcements will also be posted to Piazza

What you should get out of this course

- An understanding of the major problems in computational molecular biology
- Familiarity with the algorithms and statistical techniques for addressing these problems
- 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 task perspective)

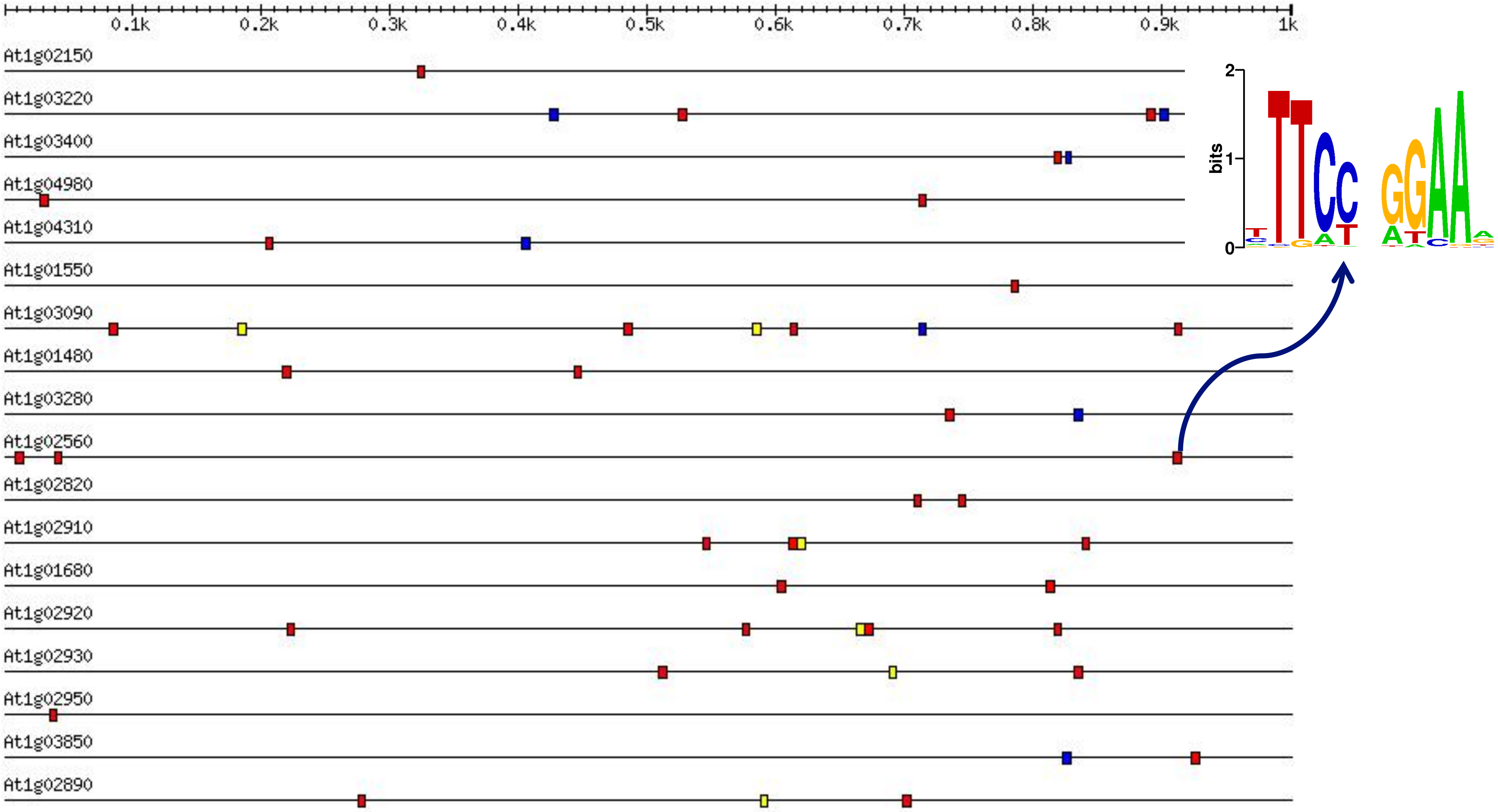
- modeling of motifs and *cis*-regulatory modules
- identification of transcription factor binding sites
- gene finding
- transcriptome quantification and assembly
- RNA sequence and structure modeling
- modeling biological sequence evolution
- large-scale and whole-genome sequence alignment
- modeling the evolution of cellular networks
- genotype analysis and association studies
- protein structure prediction

Major Topics to be Covered (the algorithms perspective)

- Gibbs sampling and EM
- HMM structure search
- duration modeling and semi-Markov models
- pairwise HMMs
- interpolated Markov models and back-off methods
- tries and suffix trees
- sparse dynamic programming
- Markov random fields
- stochastic context free grammars
- Bayesian networks
- branch and bound search
- conditional random fields
- etc.

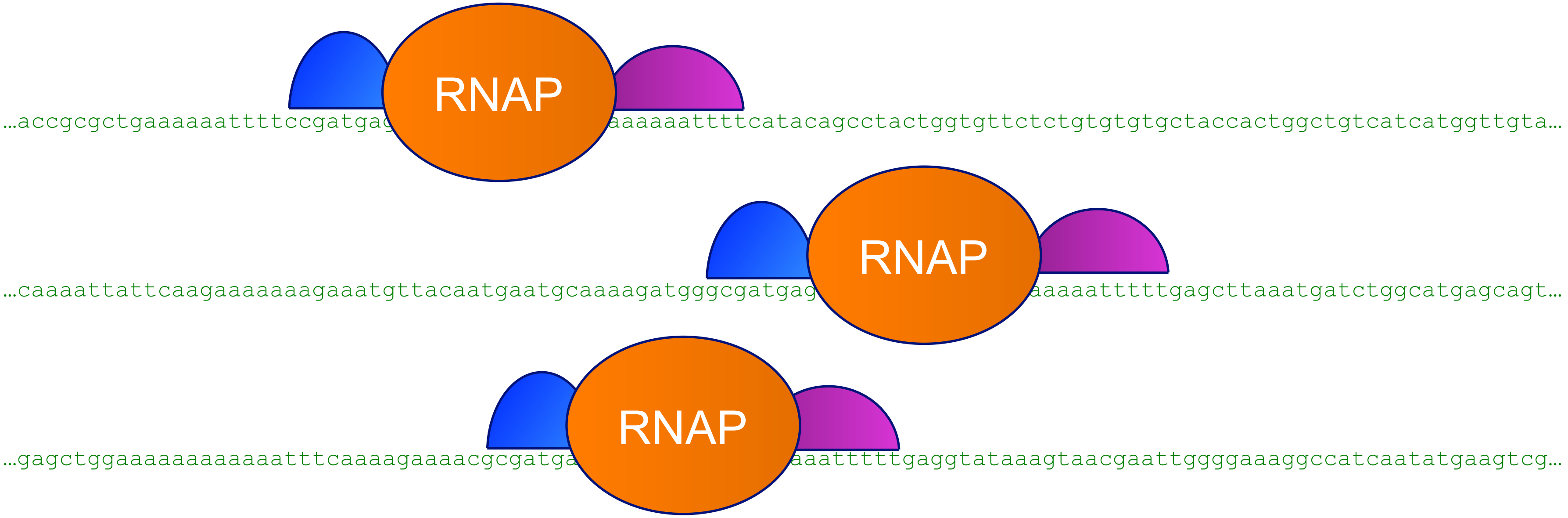
Motif Modeling

What sequence motif do these promoter regions have in common?



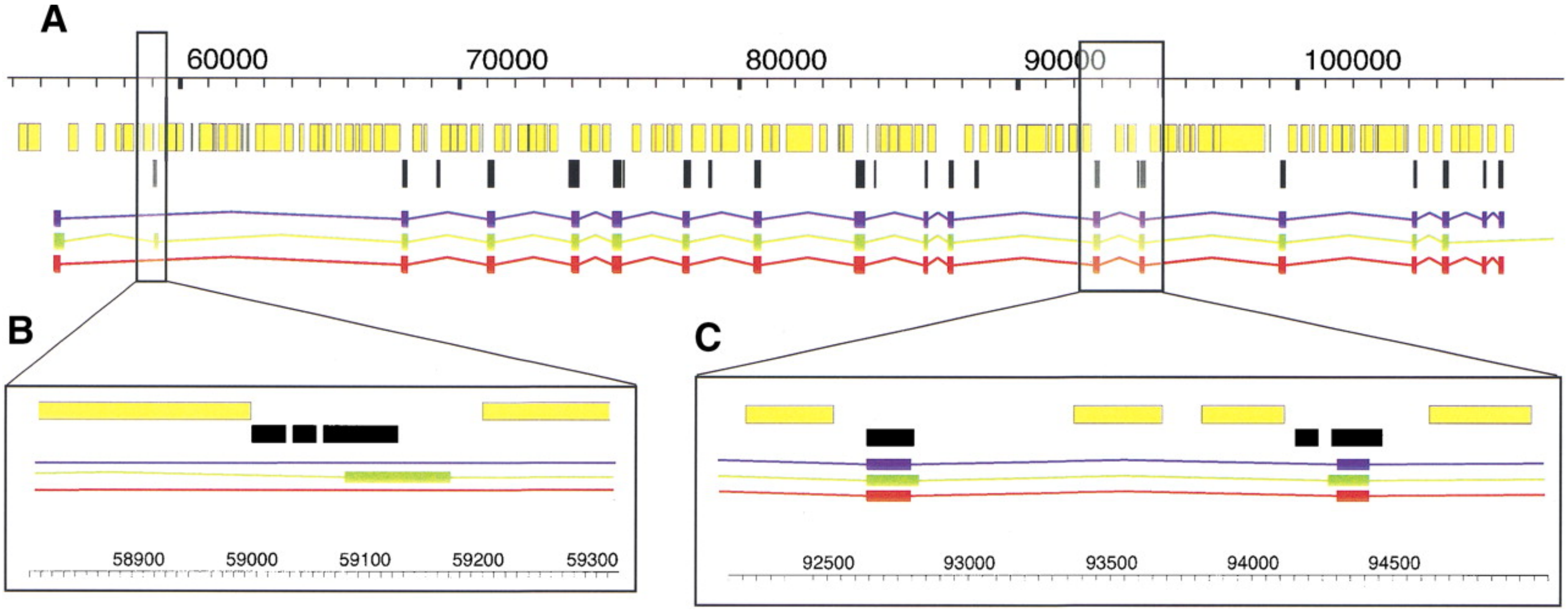
cis-Regulatory Modules (CRMs)

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

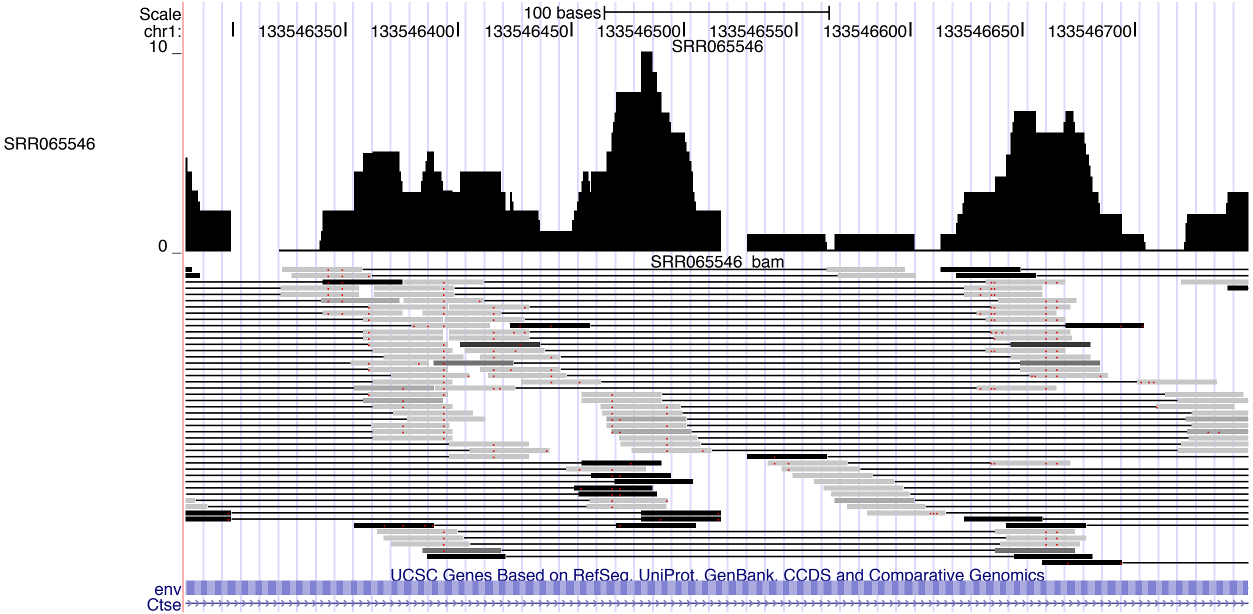

RNAP
…accgcgctgaaaaaattttccgatgag
aaaaaattttcatacagcctactggtgttctctgtgtgtgctaccactggctgtcatcatggttgta…
RNAP
…caaaattattcaagaaaaaaagaaatgttacaatgaatgcaaaagatgggcgatgag
aaaaattttttgagcttaaatgatctggcatgagcagt…
RNAP
…gagctggaaaaaaaaaaaatttcaaaagaaaacgcgatga
aaatttttgaggtataaagtaacgaattggggaaaggccatcaatatgaagtcg…

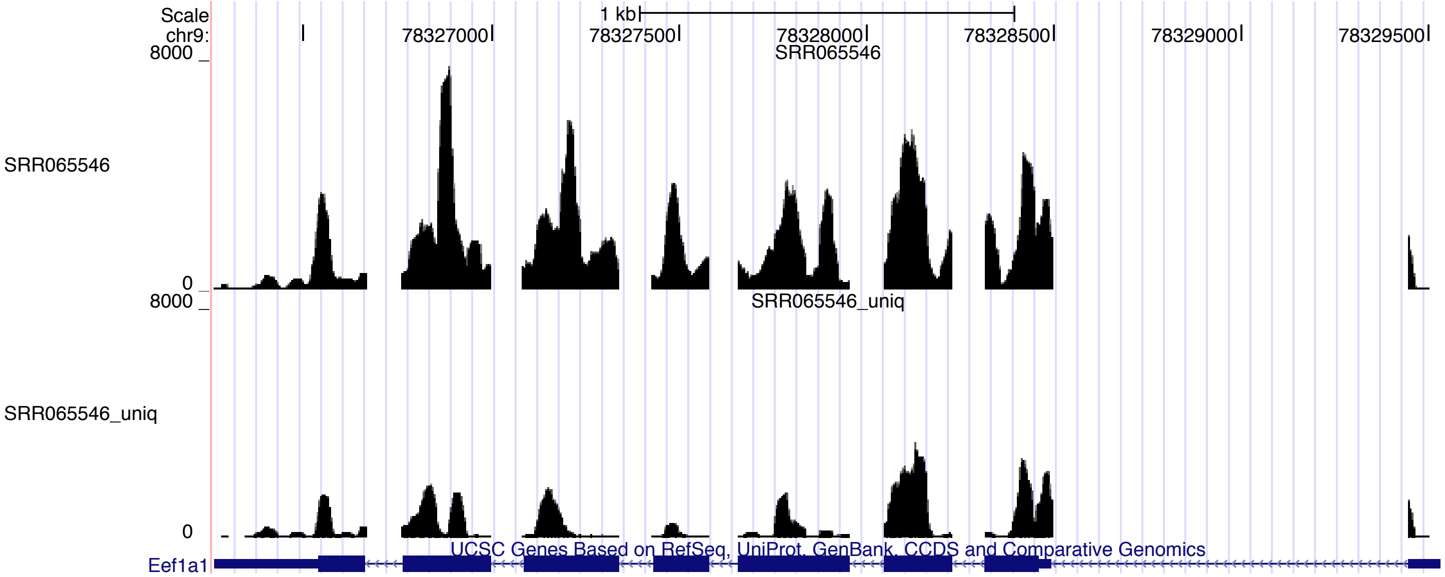
Gene Finding

Where are the genes in this genome, and what is the structure of each gene?



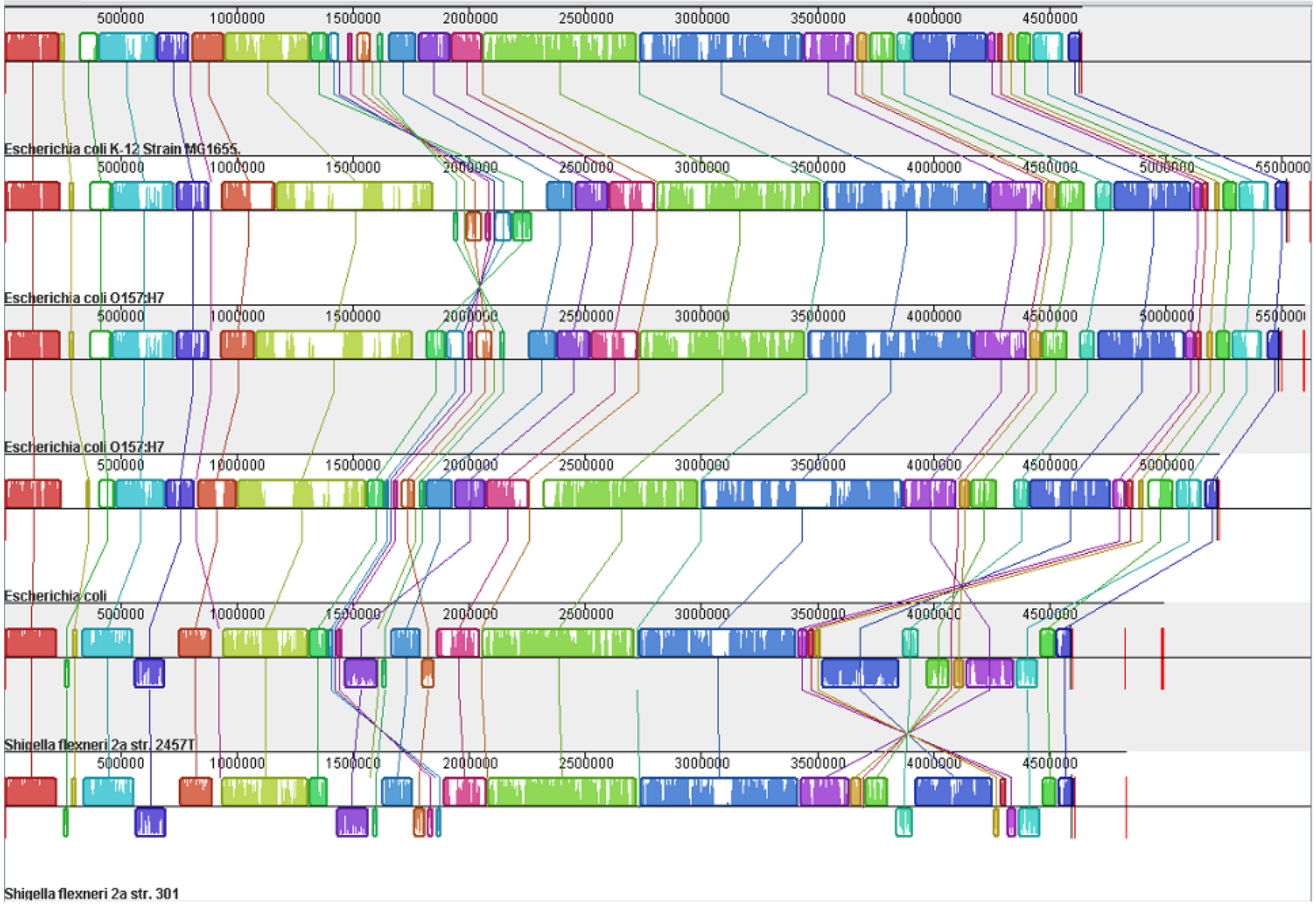
Transcriptome analysis with RNA-Seq





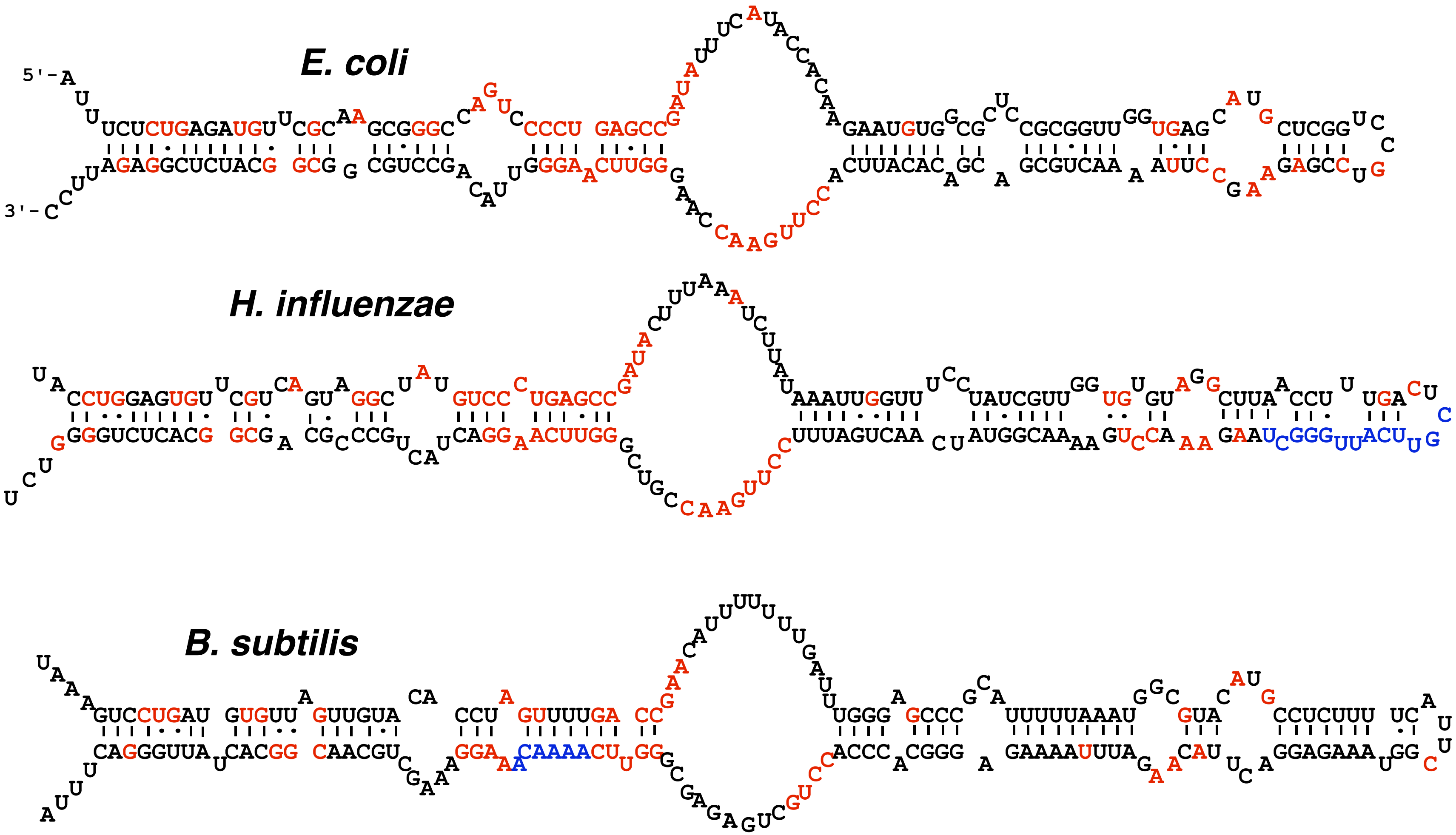
Large Scale Sequence Alignment

What is the best alignment of these 5 genomes?

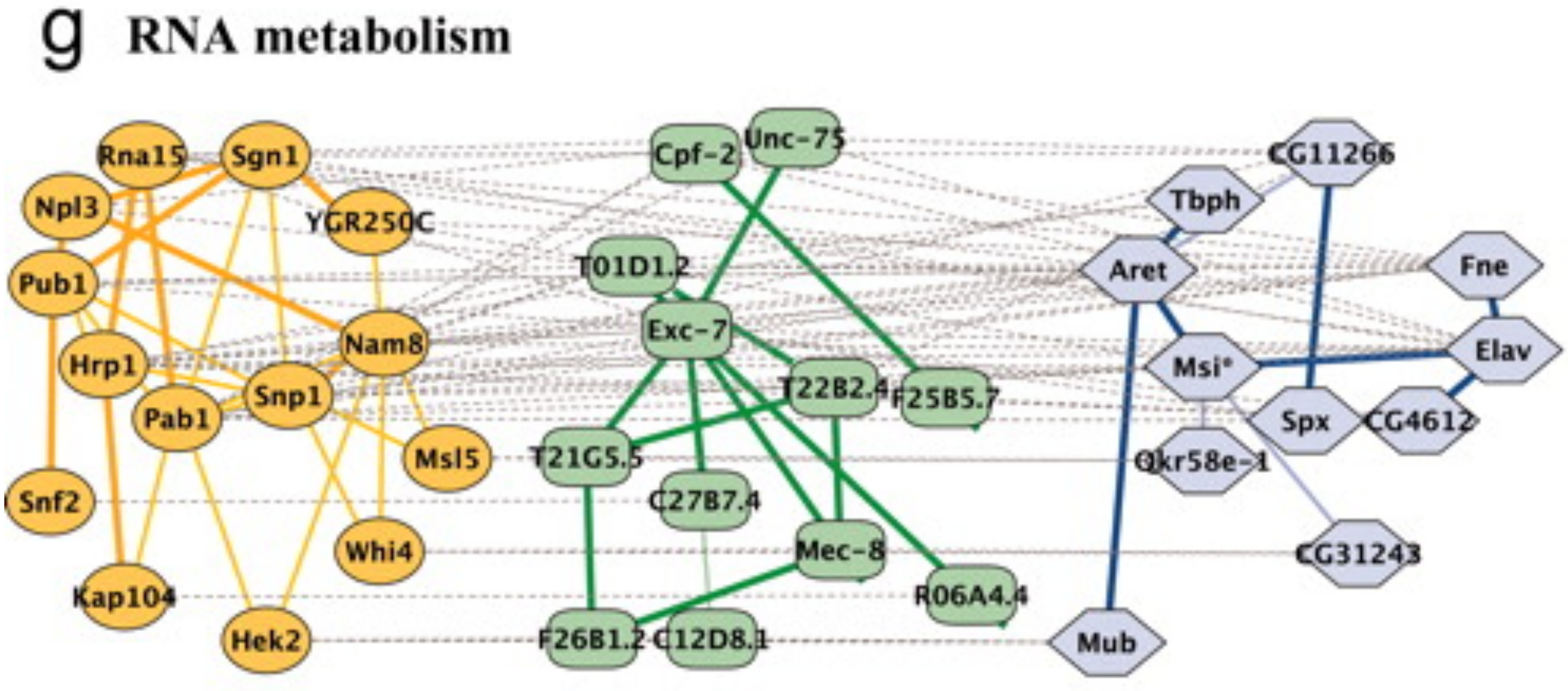


RNA Sequence and Structure Modeling

Given a genome, how can we identify sequences that encode this RNA structure?



Modeling cellular network evolution

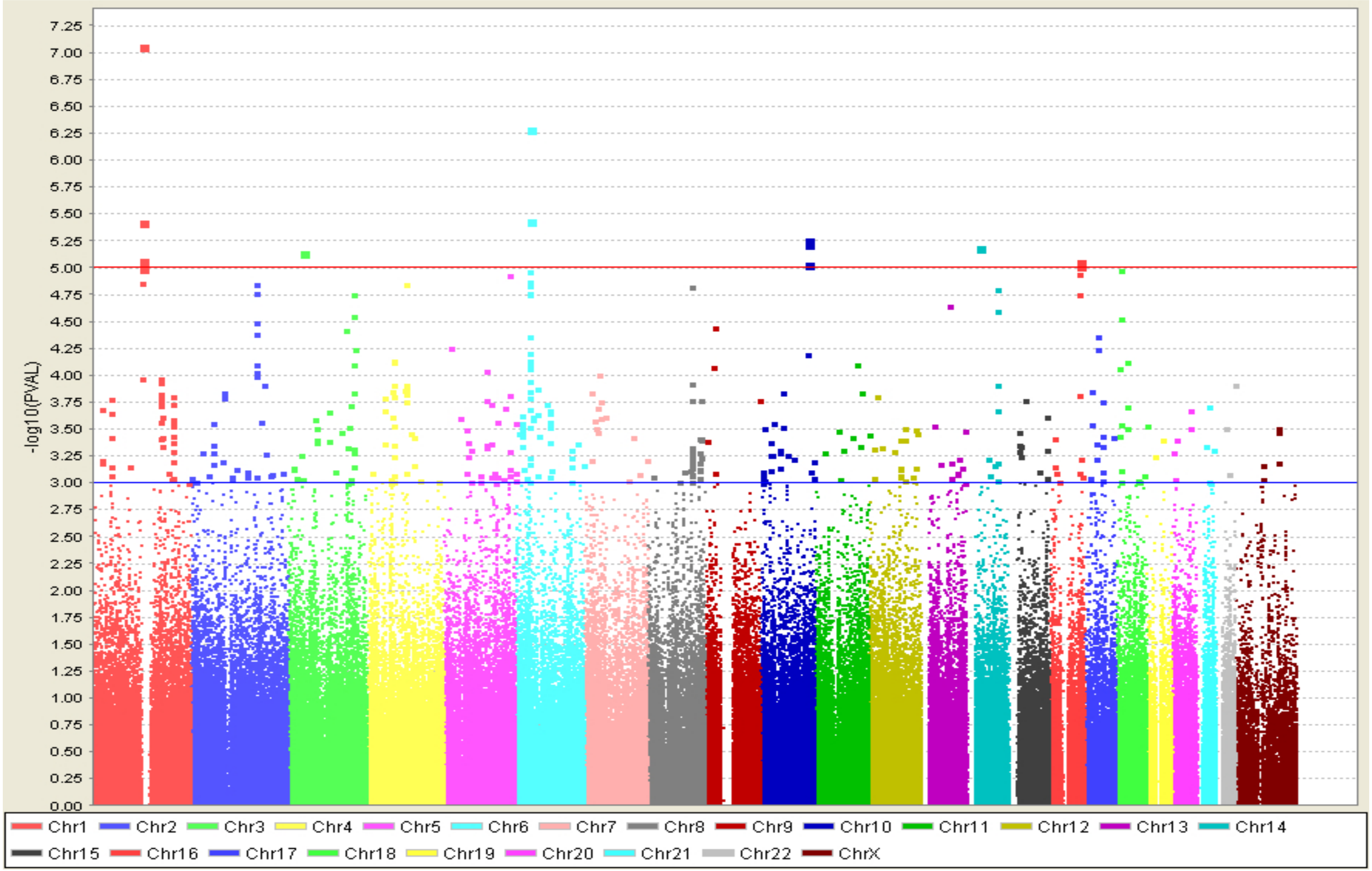


yeast worm fly

Sharan et al. PNAS 2005;102:1974-1979

Genome-wide Association Studies

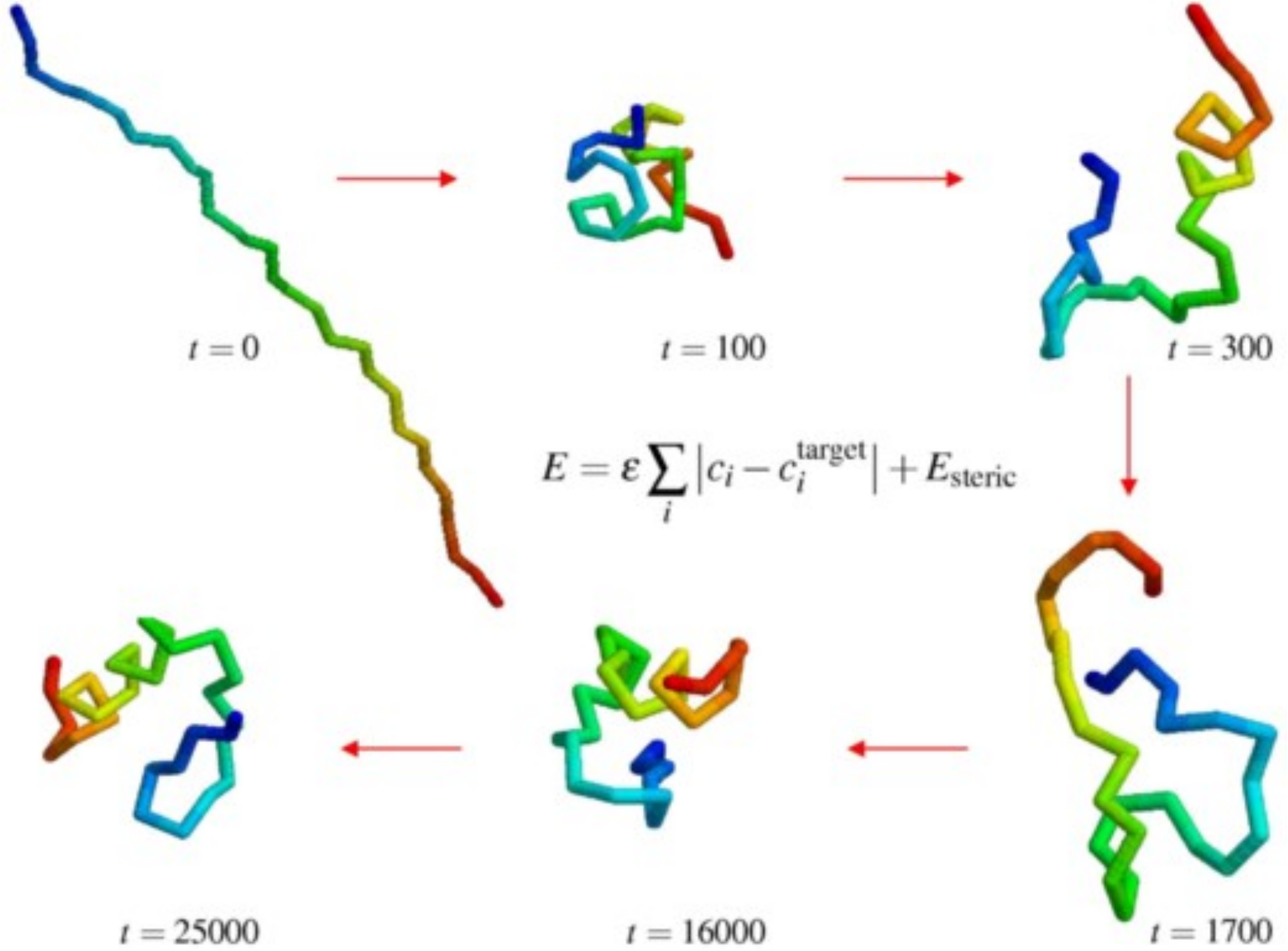
Which genes are involved in diabetes?



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.

Protein Structure Prediction

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



$$E = \varepsilon \sum_i \left| c_i - c_i^{\text{target}} \right| + E_{\text{steric}}$$

Other topics that I am considering adding

- Genome assembly
- Alignment of next-generation sequencing data
- Analysis of proteomics data
  - mass spectrometry
- Gene Ontology (GO) analysis
  - functional analysis of sets of genes

Reading Assignment

- Bailey and Elkan, *1995*
- available on the course web site

Seminars of interest

- Nils Gehlenborg (Broad Institute) – “Driving Biomedical Discovery with Exploratory Data Visualization”
  - Today, 4:00pm, WID Forum

Courses of interest

- Cancer Bioinformatics (BMI 826/CS 838)
  - Prof. Tony Gitter
  - http://www.biostat.wisc.edu/~gitter/BMI826-S15
- Tools for Reproducible Research (BMI 826)
  - Prof. Karl Broman
  - http://kbroman.org/Tools4RR/
- Graphical Models (Stat 992)
  - Prof. Garvesh Raskutti
- Online Machine Learning (ECE 901)
  - Prof. Rebecca Willett

Reading groups

- Computational Systems Biology Reading Group
  - http://lists.discovery.wisc.edu/mailman/listinfo/compsysbiojc
  - Wed 2-3pm every other week
- AI Reading Group
  - http://lists.cs.wisc.edu/mailman/listinfo/airg
  - Wed 4pm