Advanced Bioinformatics

Biostatistics & Medical Informatics 776
Computer Sciences 776
Spring 2012

Colin Dewey
Dept. of Biostatistics & Medical Informatics
Dept. of Computer Sciences
cdewey@biostat.wisc.edu
www.biostat.wisc.edu/bmi776/

Agenda Today

- course information
- overview of topics
- introductions

Course Web Site

- www.biostat.wisc.edu/bmi776/
- syllabus
- readings
- tentative schedule
- lecture slides in PDF
- homework
- mailing list archive
- etc.

Your Instructor: Colin Dewey

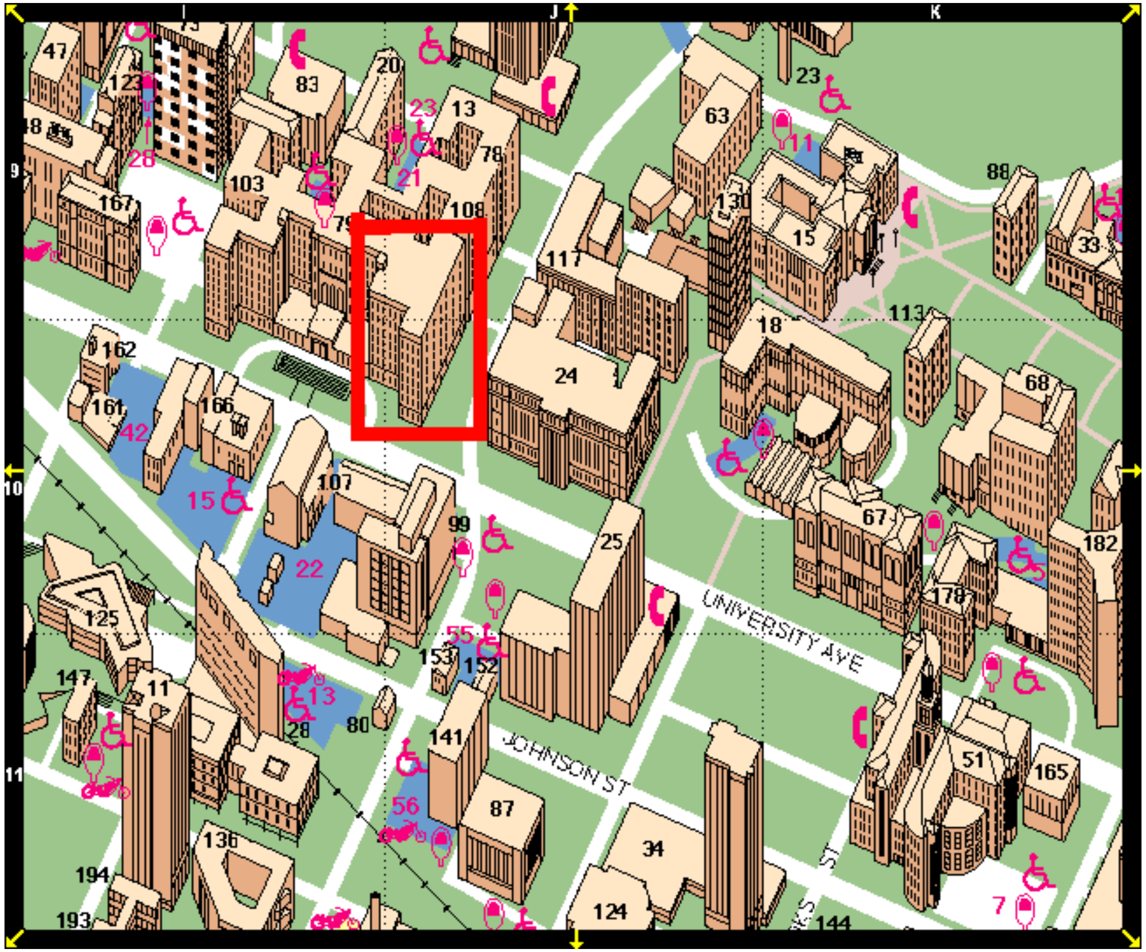
- email:
  cdewey@biostat.wisc.edu

- office hours: Wed 9:30-10:30am, Thu 11:00am-12:00pm
  room 5785, Medical Sciences Center

- my home department is Biostatistics & Medical Informatics, and I have an affiliate appointment in Computer Sciences

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

Finding My Office:
5785 Medical Sciences Center



- confusing building
- best bet: enter at door marked *420 North Charter*

Course Requirements

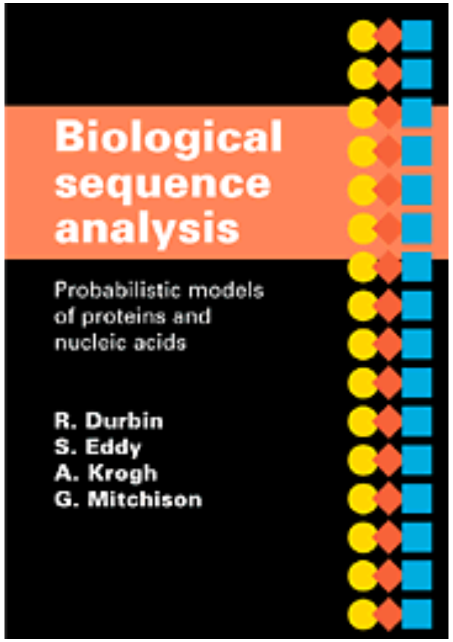
- 4 or so homework assignments: ~20%
  - written exercises
  - programming (in Java, C++, C, Perl, Python) + computational experiments (e.g. measure the effect of varying parameter *x* in algorithm *y*)
- 4 or so paper critiques: ~20%
  - major strength of approach
  - major weakness
  - what would you do next
- project: ~25%
- final exam: ~ 25%
- class participation: ~10%

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 “On Campus” profile)
- *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

Computing Resources for the Class

- Linux workstations in Dept. of Biostatistics & Medical Informatics
  - no “lab”, must log in remotely
  - 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
- the “CS 1000” UNIX tutorial
  - online at http://www.cs.wisc.edu/csl/cs1000/

The Class Mailing List

- bmi776-1-s12@lists.wisc.edu
- you will be automatically subscribed
- check your mail daily or have it forwarded to an account where you do
    - mailing list has your @wisc.edu address

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)

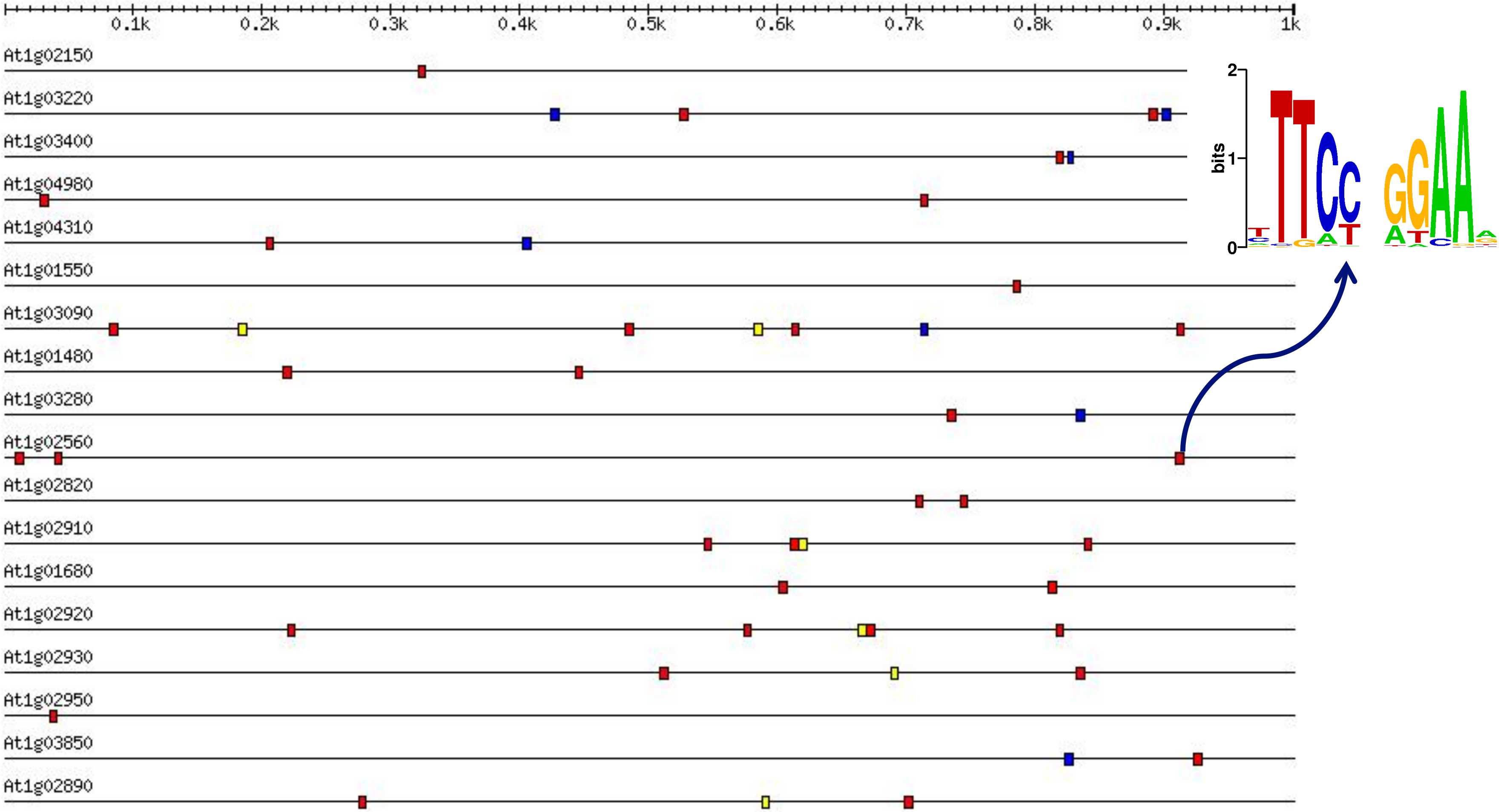
- modeling of motifs and *cis*-regulatory modules
- identification of transcription factor binding sites
- gene finding
- transcriptome quantification and assembly
- RNA sequence and structure modeling
- modeling biological sequence evolution
- large-scale and whole-genome sequence alignment
- modeling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

Major Topics to be Covered (the algorithms perspective)

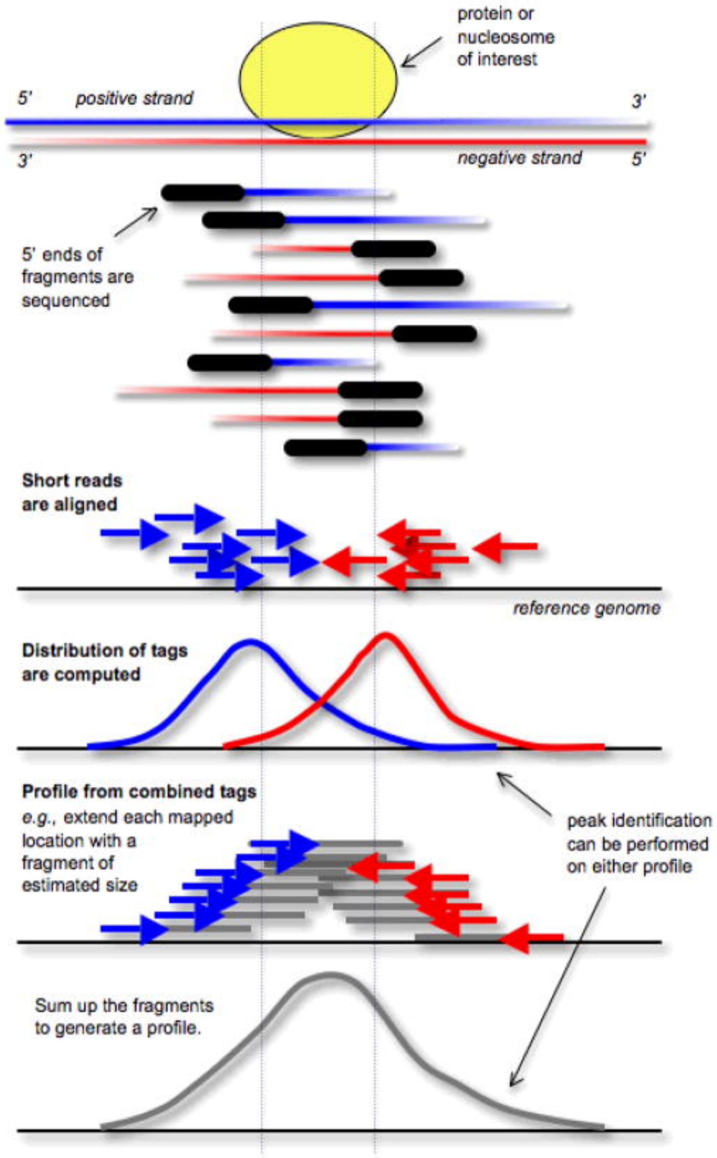
- Gibbs sampling and EM
- HMM structure search
- duration modeling and semi-Markov models
- pairwise HMMs
- interpolated Markov models and back-off methods
- parametric alignment
- tries and suffix trees
- sparse dynamic programming
- Markov random fields
- stochastic context free grammars
- Bayesian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

Motif and CRM Modeling

What sequence motifs do these promoter regions have in common?



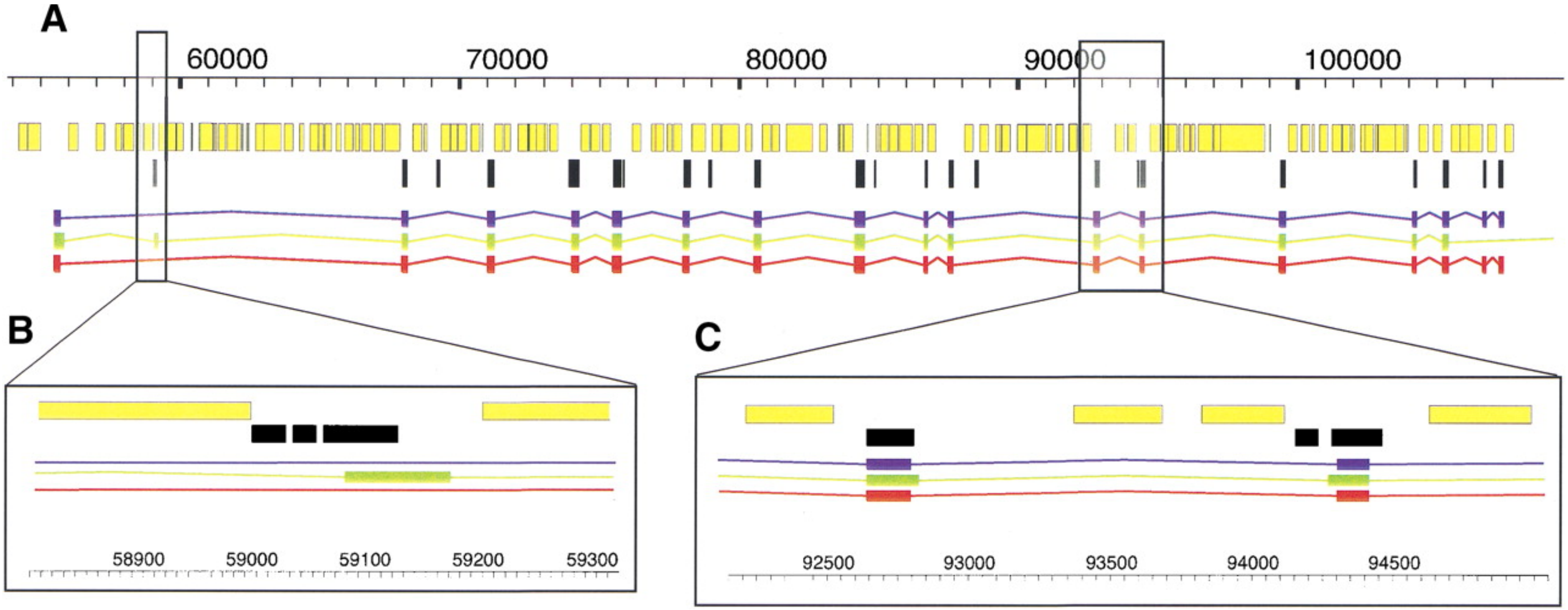
Experimental binding site prediction with second-generation sequencing data (ChIP-Seq)



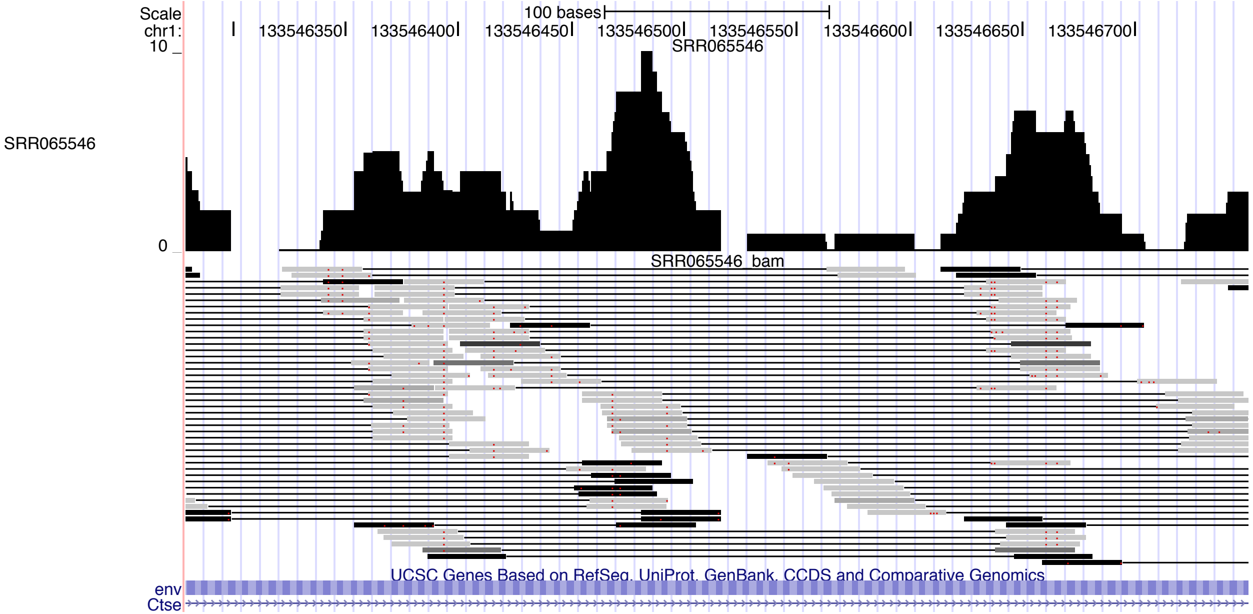
(Park, Nat Rev Genet, 2009)

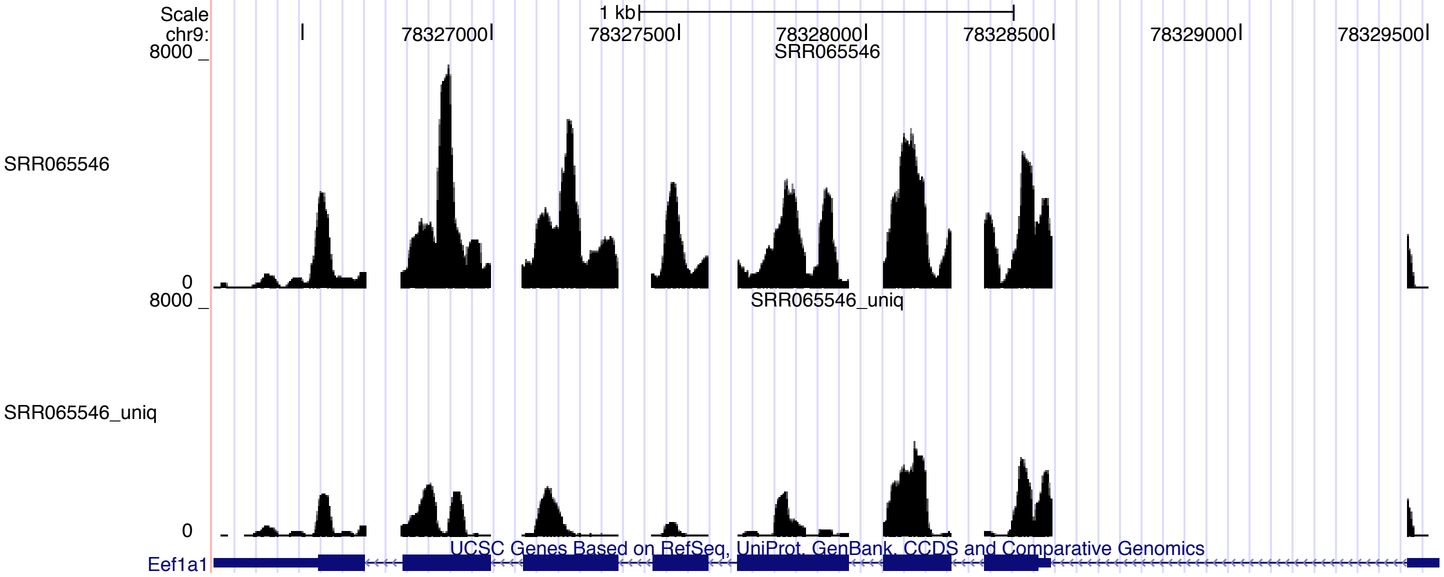
Gene Finding

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



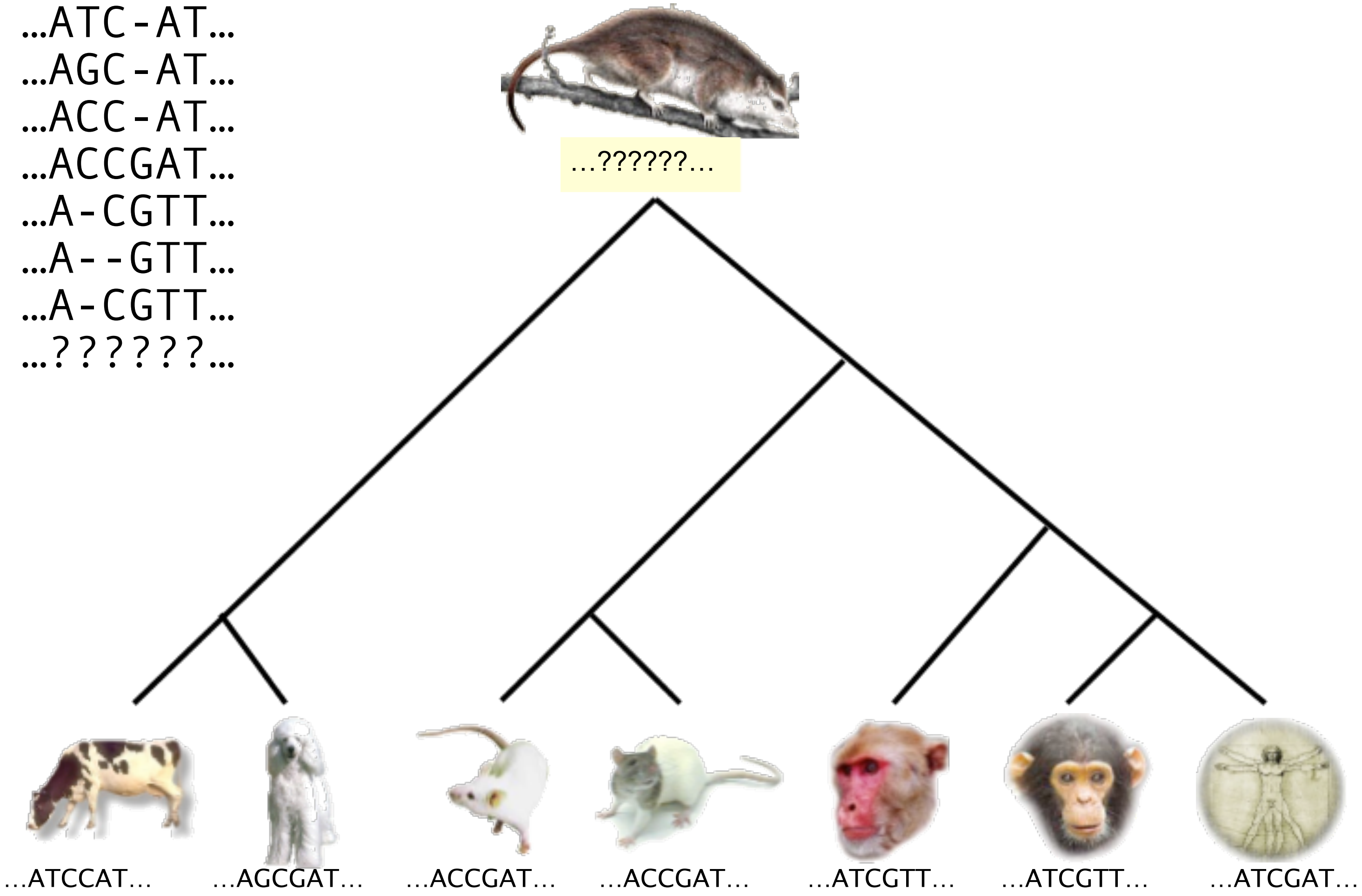
Transcriptome analysis with RNA-Seq





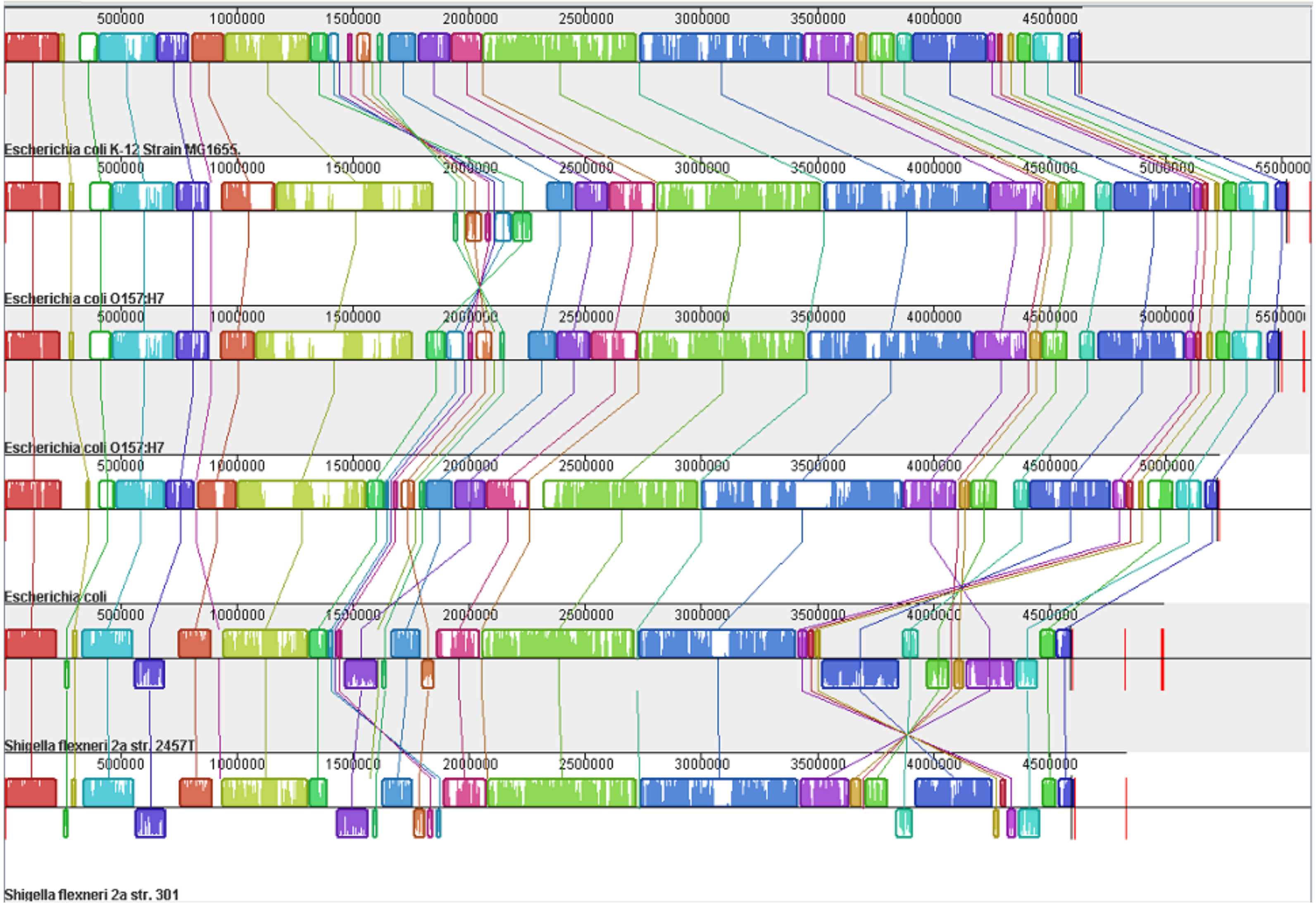
Modeling biological sequence evolution

Cow	…ATC-AT…
Dog	…AGC-AT…
Mouse	…ACC-AT…
Rat	…ACCGAT…
Macaque	…A-CGTT…
Chimp	…A--GTT…
Human	…A-CGTT…
Ancestor	…??????…



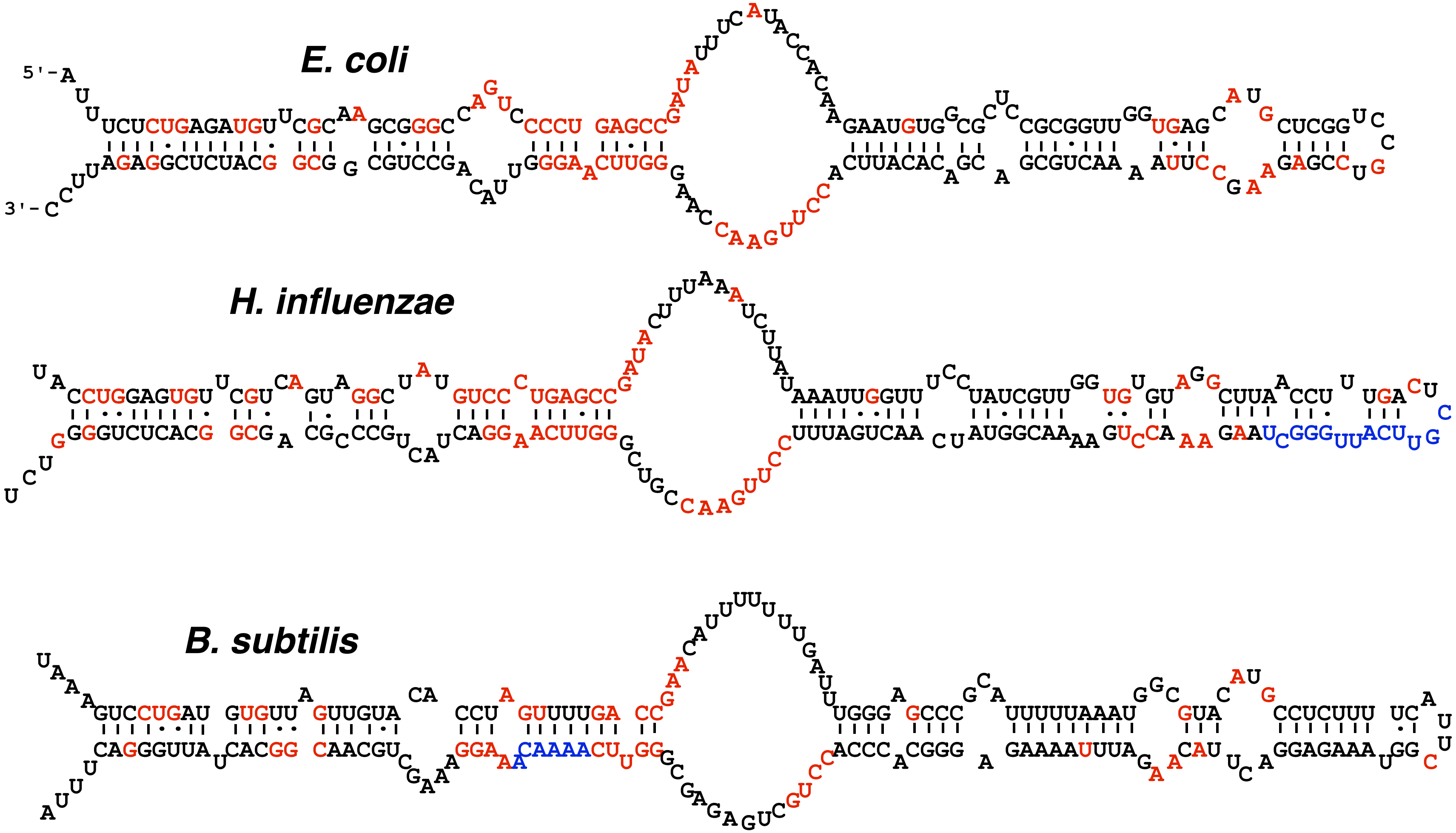
Large Scale Sequence Alignment

What is the best alignment of these 5 genomes?

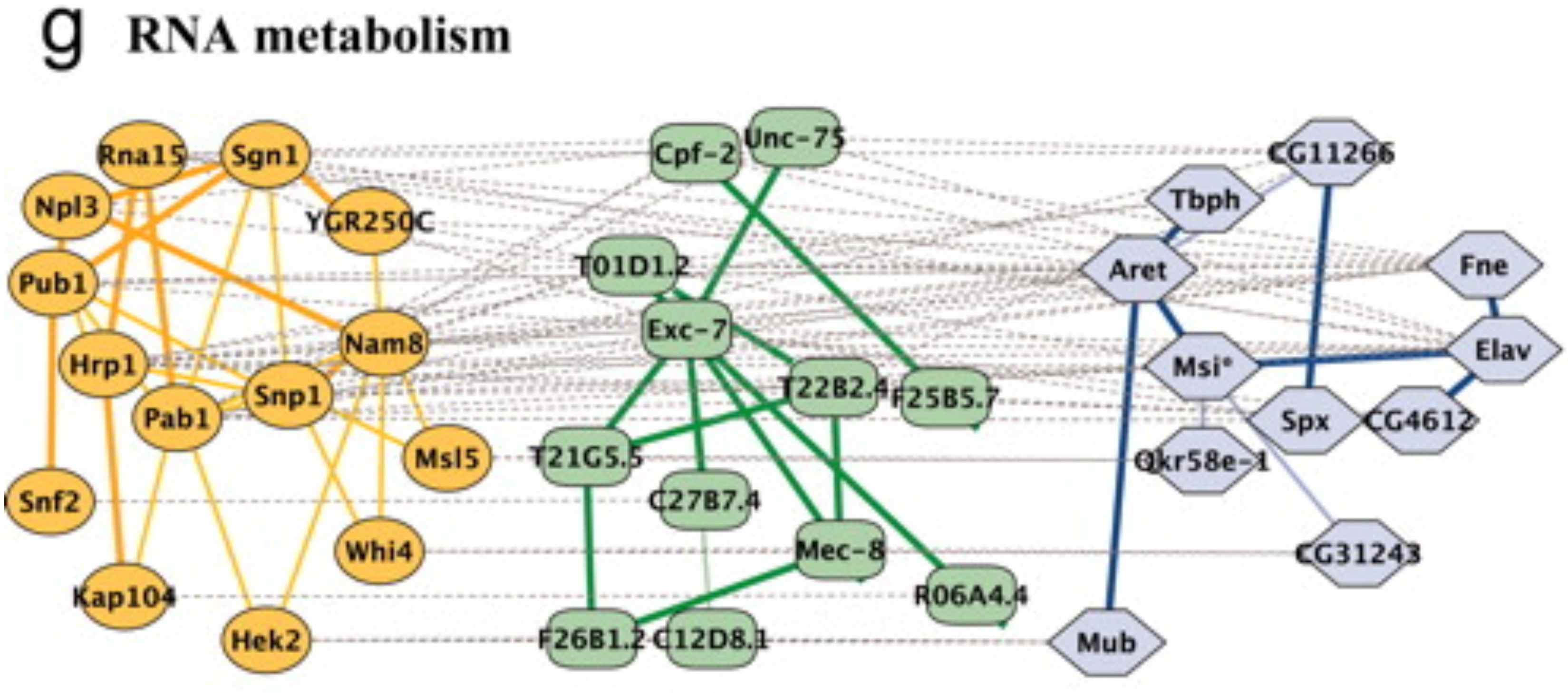


RNA Sequence and Structure Modeling

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



Modeling cellular network evolution

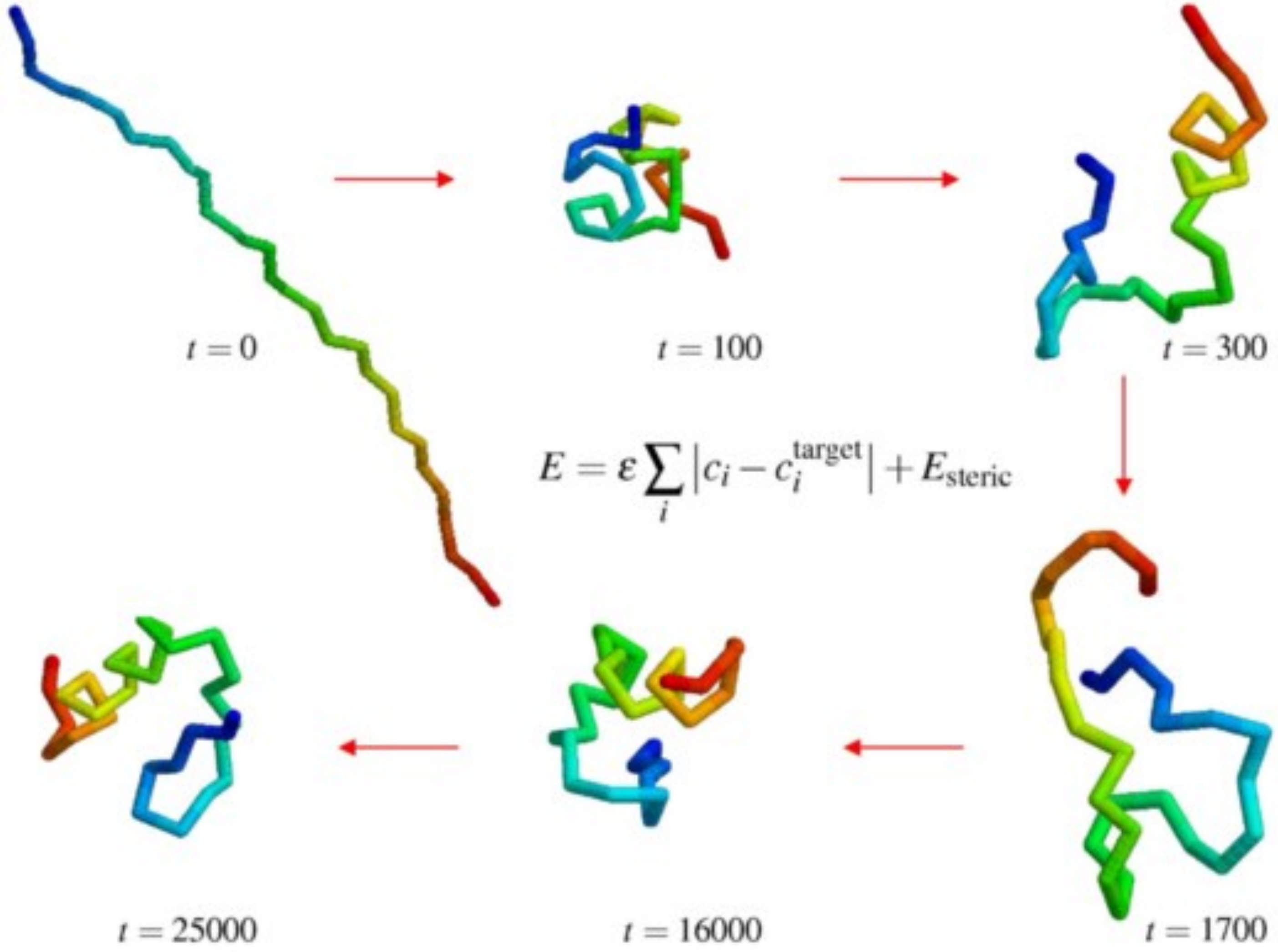


yeast worm fly

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

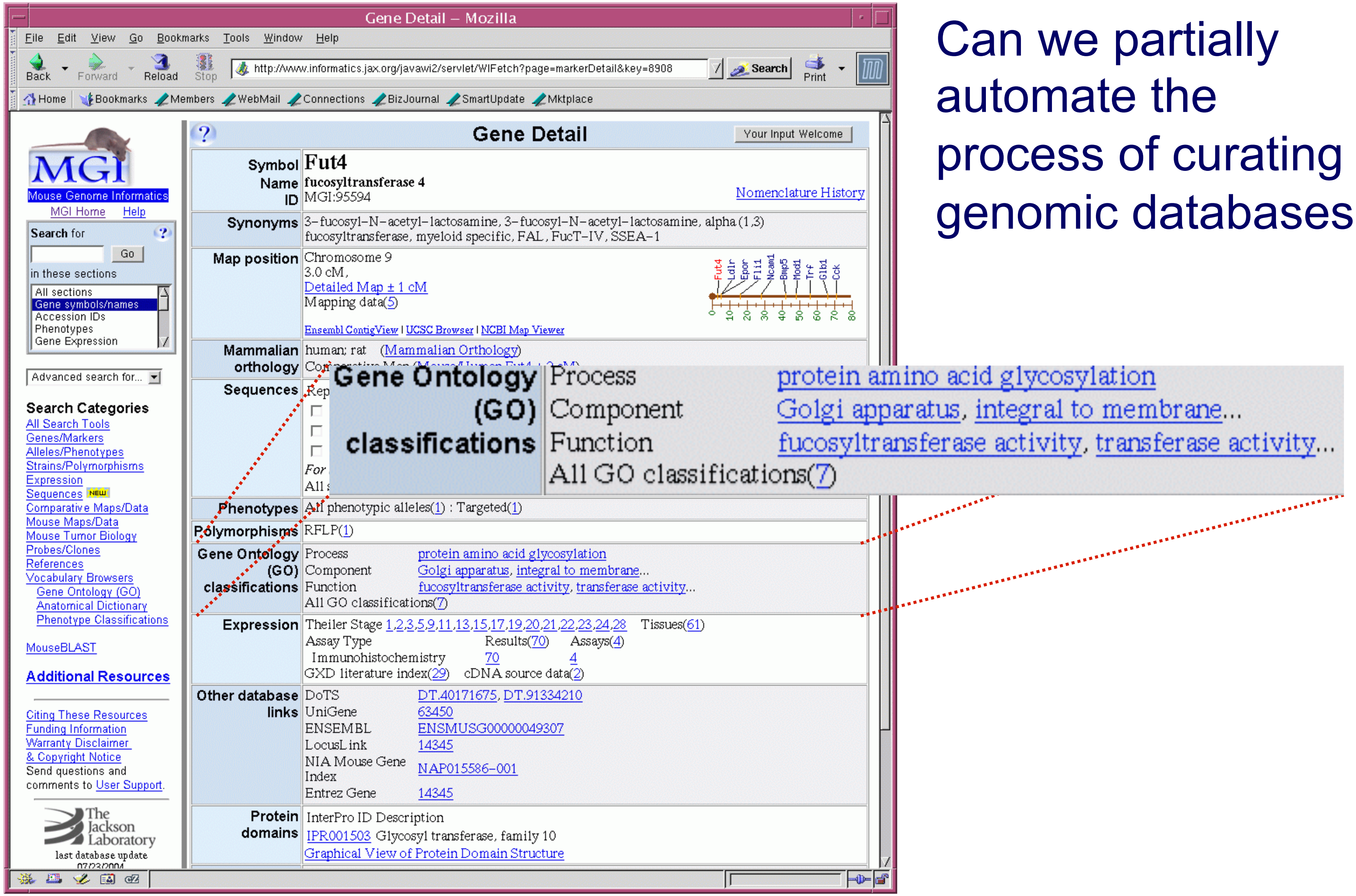
Protein Structure Prediction

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


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

Biomedical Text Mining

Can we partially automate the process of curating genomic databases?



Search Categories

All Search Tools
Genes/Markers
Alleles/Phenotypes
Strains/Polymorphisms
Expression
Sequences
Comparative Maps/Data
Mouse Maps/Data
Mouse Tumor Biology
Probes/Clones
References
Vocabulary Browsers
Gene Ontology (GO)
Anatomical Dictionary
Phenotype Classifications

MouseBLAST

Additional Resources

Citing These Resources
Funding Information
Warranty Disclaimer
& Copyright Notice
Send questions and comments to User Support.

last database update
07/23/2004

Gene Detail

Symbol Name ID	Fut4 **fucosyltransferase 4** MGI:95594 Nomenclature History
Synonyms	3–fucosyl–N–acetyl–lactosamine, 3–fucosyl–N–acetyl–lactosamine, alpha (1,3) fucosyltransferase, myeloid specific, FAL, FucT–IV, SSEA–1
Map position	Chromosome 9 3.0 cM, Detailed Map ± 1 cM Mapping data(5) Ensembl ContigView | UCSC Browser | NCBI Map Viewer
Mammalian orthology	human; rat (Mammalian Orthology) Co
Sequences	Rep □ □ □ *For* All
Phenotypes	All phenotypic alleles(1) : Targeted(1)
Polymorphisms	RFLP(1)
Gene Ontology (GO) classifications	Process protein amino acid glycosylation Component Golgi apparatus, integral to membrane... Function fucosyltransferase activity, transferase activity... All GO classifications(7)
Expression	Theiler Stage 1,2,3,5,9,11,13,15,17,19,20,21,22,23,24,28 Tissues(61) Assay Type Results(70) Assays(4) Immunohistochemistry 70 4 GXD literature index(29) cDNA source data(2)
Other database links	DoTS DT.40171675, DT.91334210 UniGene 63450 ENSEMBL ENSMUSG00000049307 LocusLink 14345 NIA Mouse Gene Index NAP015586–001 Entrez Gene 14345
Protein domains	InterPro ID Description IPR001503 Glycosyl transferase, family 10 Graphical View of Protein Domain Structure

Gene Ontology (GO) classifications	Process	protein amino acid glycosylation
	Component	Golgi apparatus, integral to membrane...
	Function	fucosyltransferase activity, transferase activity...
	All GO classifications(7)	

Biomedical Text Mining

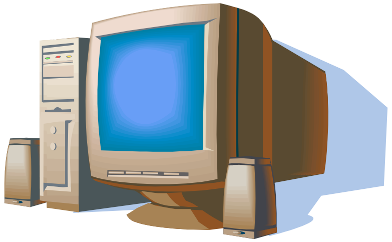
Vol. 273, No. 1, Issue of January 2, pp. 386–391, 1998
Printed in U.S.A.

The CC Chemokine I-309 Inhibits CCR8-dependent Infection by Diverse HIV-1 Strains*

(Received for publication, September 30, 1997)

Richard Horuk‡§, Joseph Hesselgesser‡, Yiqing Zhou‡, Daryl Faulds‡, Meredith Halks-Miller‡, Susan Harvey‡, Dennis Taub¶, Michel Samson‖, Marc Parmentier‖, Joseph Rucker**, Benjamin J. Doranz**§§, Robert W. Doms**

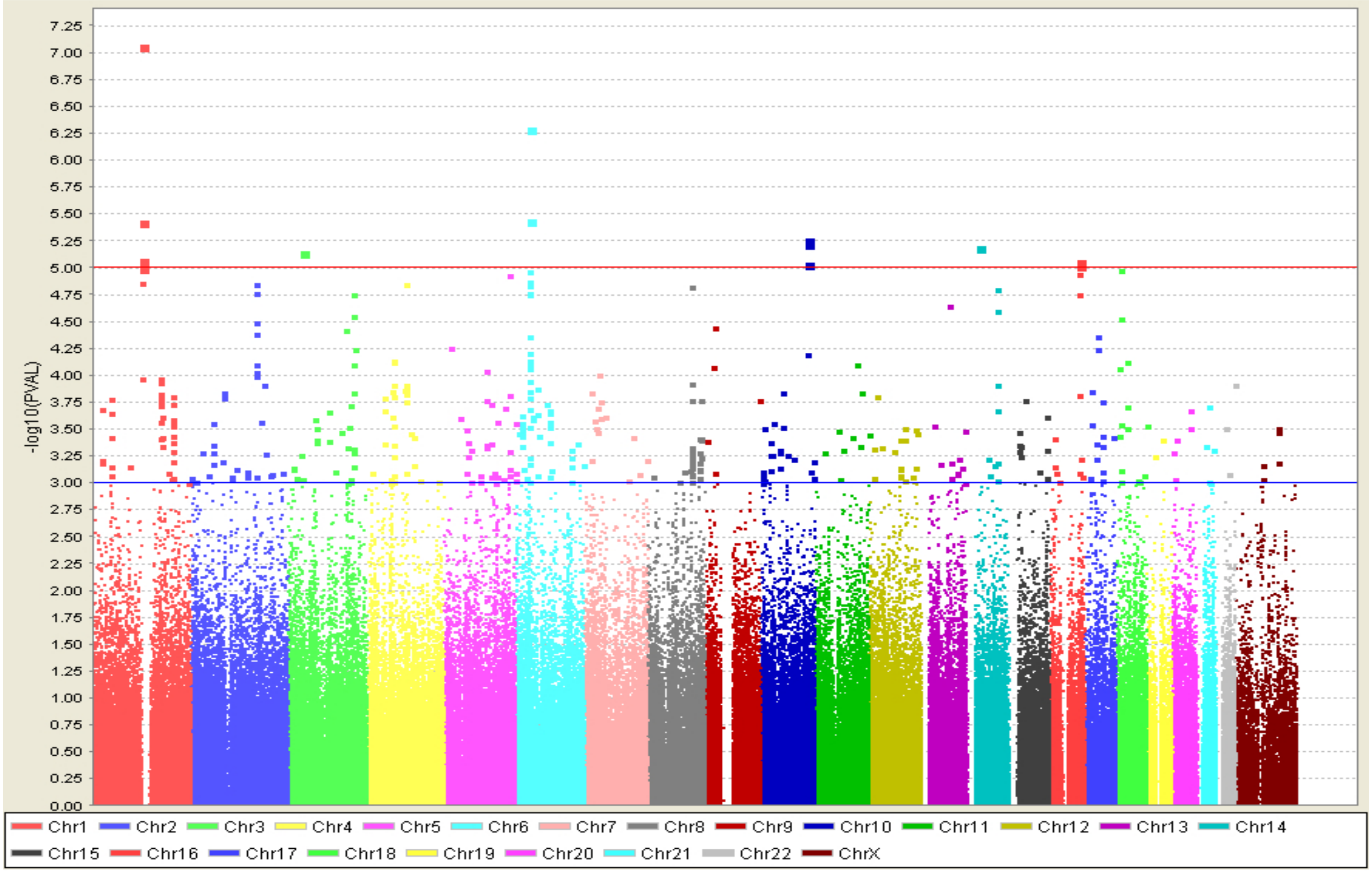
From the ‡Departments of Immunology, Pharmacology, and Cell Biology, Berlex Biosciences, Richmond, California 94804, the ¶Laboratory of Immunology, NIA, National Institutes of Health, Baltimore, Maryland 21224, the ‖IRIBHN, Université libre de Bruxelles, Campus Erasme, 808 route de Lennik, B-1070 Bruxelles, Belgium, and the **Department of Pathology and Laboratory Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104



gene: FUT 4
GO concept : protein amino-acid glycosylation

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.

Reading Assignment

- Bailey and Elkan, *ISMB* ’95
- Lawrence et al., *Science* ’93
- available on the course web site