6.096
Algorithms for Computational Biology

Prof. Manolis Kellis

Today's Goals

- Introduction
  - Class introduction
  - Challenges in Computational Biology

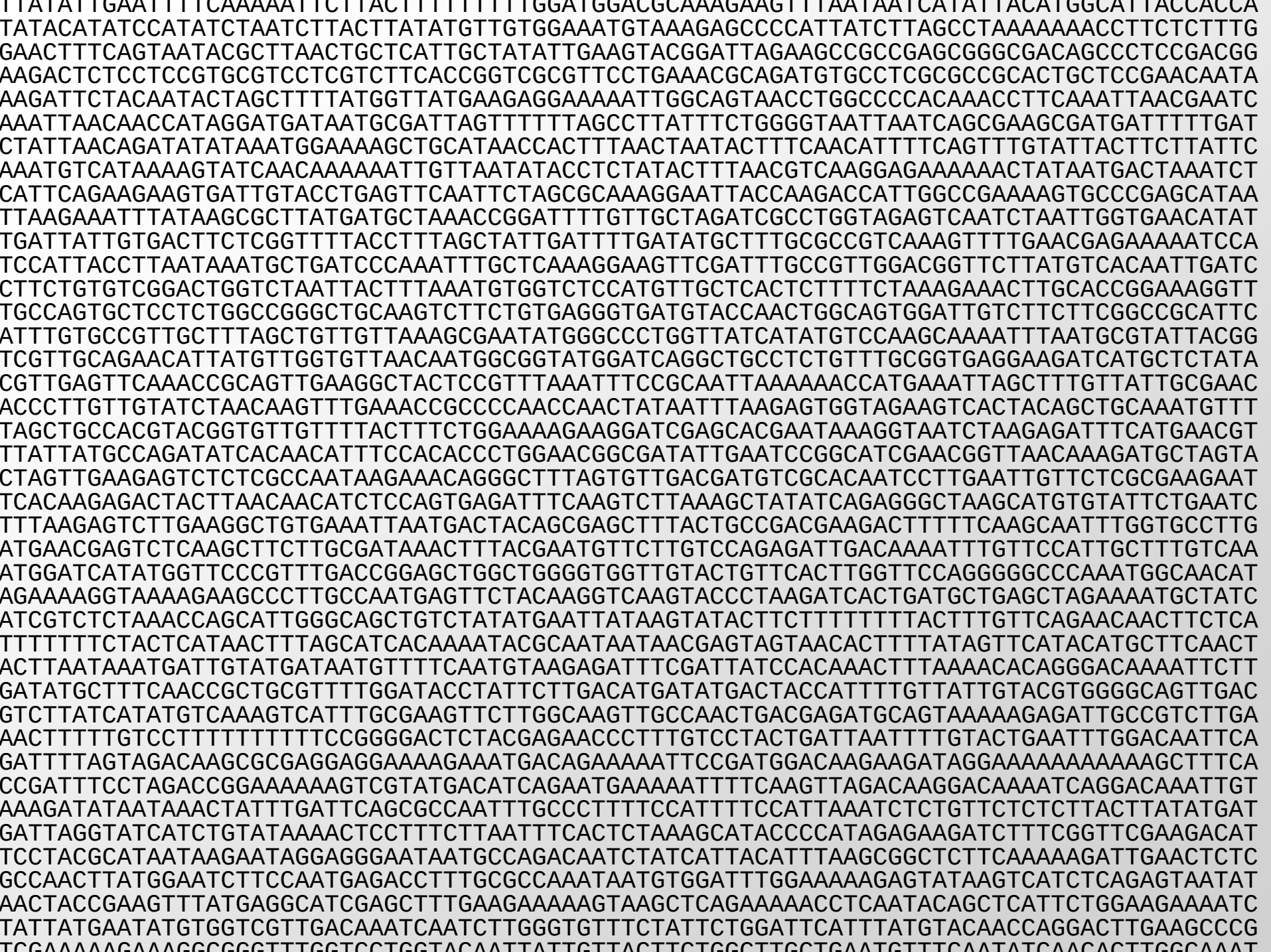
- Gene Regulation: Regulatory Motif Discovery
  - Exhaustive search
  - Content-based indexing
  - Greedy optimization

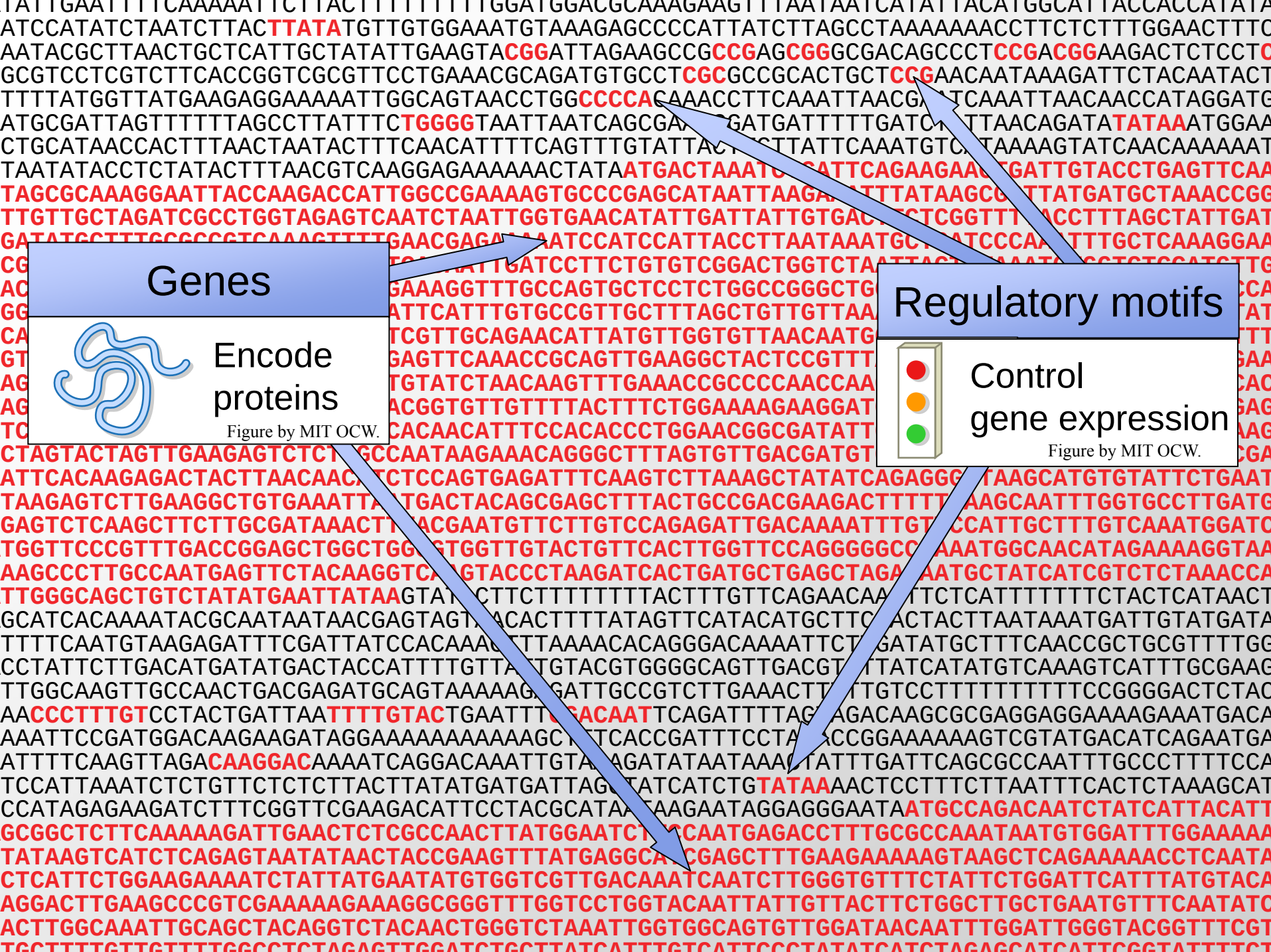
Course Administrivia

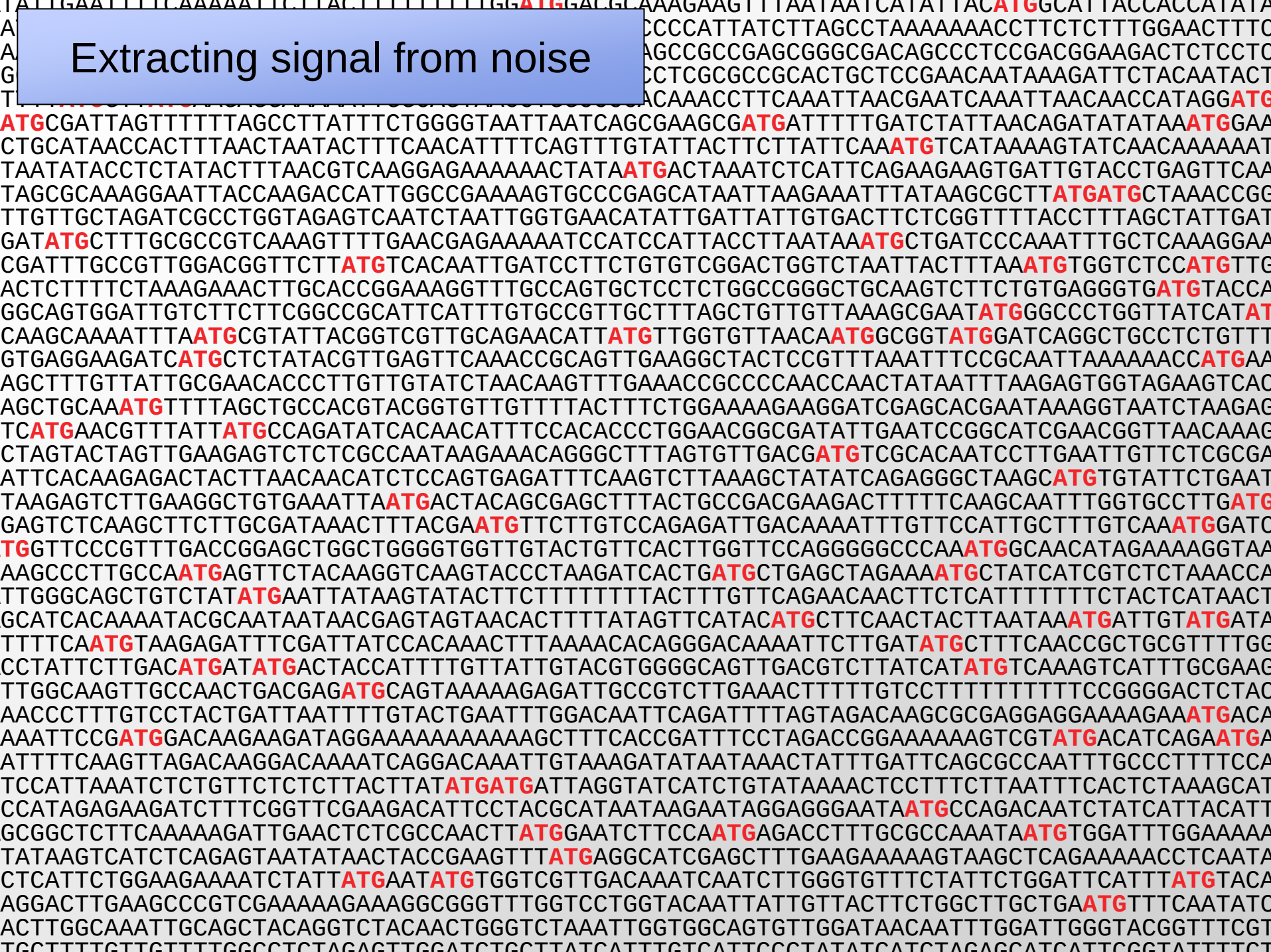
- 6.096 – Algorithms for Computational Biology
  - Taught jointly with 6.046, Introduction to Algorithms
  - Explores specific application area of algorithms
  - Algorithmic challenges in Computational Biology
  - Design principles to address them
- Lectures
  - Grading: 4 problem sets = 60%. Final: 30%. Attendance: 10%

Book references

- Gusfield, Dan. *Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology*. Cambridge, UK: Cambridge University Press, 1997. ISBN: 0521585198.
- Waterman, Michael. *Introduction to Computational Biology: Maps, Sequences, and Genomes*. Boca Raton, FL: CRC Press, 1995. ISBN: 0412993910.
- Durbin, Richard, Graeme Mitchison, S. Eddy, A. Krogh, and G. Mitchison. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge, UK: Cambridge University Press, 1997. ISBN: 0521629713.
- Jones, Neil, and Pavel Pevzner. *An Introduction to Bioinformatics Algorithms*. Cambridge, MA: MIT Press, 2004. ISBN: 0262101068.

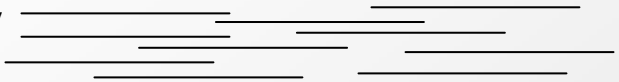



Genes
Encode proteins
Figure by MIT OCW.
Regulatory motifs
Control gene expression
Figure by MIT OCW.


Extracting signal from noise

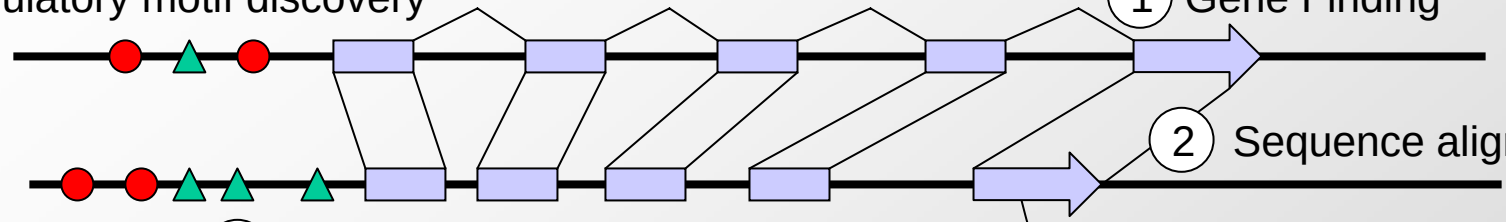
Challenges in Computational Biology

4 Genome Assembly



5 Regulatory motif discovery

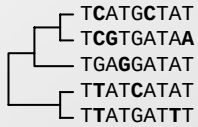
1 Gene Finding

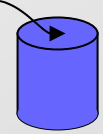


2 Sequence alignment

6 Comparative Genomics

7 Evolutionary Theory

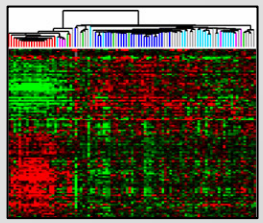




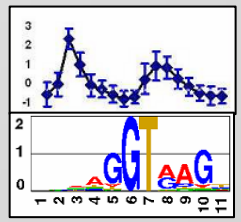
3 Database lookup

8 Gene expression analysis

RNA transcript

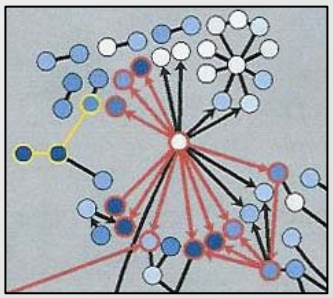


9 Cluster discovery



10 Gibbs sampling

11 Protein network analysis



12 Regulatory network inference

13 Emerging network properties

Algorithms and techniques covered

- Enumeration approaches
  - Exhaustive search, pruning, greedy algorithms, iterative refinement
- Content-based indexing
  - Hashing, database lookup, pre-processing
- Iterative methods
  - Combining sub-problems, memorization, dynamic programming
- Statistical methods
  - Hypothesis testing, maximum likelihood, Bayes' Law, HMMs
- Machine learning techniques
  - Supervised and unsupervised learning, classification

Genomic Scales

	Base pairs	Genes	Notes
Phi-X 174	5,386	10	virus of E. coli
Human mitochondrion	16,569	37	Energy production for human cells
Epstein-Barr virus (EBV)	172,282	80	causes mononucleosis
nucleomorph of Guillardia thet	551,264	511	Remains of the nuclear genome of a red alga (eukaryote) engulfed long ago by another eukaryote
Mycoplasma genitalium	580,073	483	One of the smallest true organisms
Treponema pallidum	1,138,011	1,039	bacterium that causes syphilis
Mimivirus	1,181,404	1,262	A virus (of an amoeba) with a genome larger than several cellular organisms above
Helicobacter pylori	1,667,867	1,589	chief cause of stomach ulcers (not stress and diet)
Methanococcus jannaschii	1,664,970	1,783	Classified in a third kingdom: Archaea.
Haemophilus influenzae	1,830,138	1,738	bacterium that causes middle ear infections
Streptococcus pneumoniae	2,160,837	2,236	the pneumococcus
Propionibacterium acnes	2,560,265	2,333	causes acne
E. coli	4,639,221	4,377	Most well-studied bacterium
Saccharomyces cerevisiae	12,495,682	5,770	Budding yeast. A eukaryote.
Neurospora crassa	38,639,769	10,082	Green mold fungus.
Caenorhabditis elegans	100,258,171	19,000	The first multi-cellular eukaryote to be sequenced.
Arabidopsis thaliana	115,409,949	25,498	a flowering plant (angiosperm) See note.
Drosophila melanogaster	122,653,977	13,379	the fruit fly
Anopheles gambiae	278,244,063	13,683	Mosquito vector of malaria.
Humans	3,000,000,000	22,000	Sequenced in 1999, completed in 2004.
Tetraodon nigroviridis	342,000,000	27,918	Much less repetitive DNA, but slightly more genes.
Rice	4,300,000,000	60,000	Extremely repetitive. Genes show GC gradient
Amphibians	109,000,000,000	?	

- Importance of algorithm design for efficiency
  - Compare human vs. mouse (blocks of 1,000 nucleotides)
    - 3,000,000*3,000,000 comparisons, each 1,000*1,000 operations (w/dynamic progr.)
    - At 1 trillion operations per second, it would take 104 days
  - Search all regulatory motifs of length 20 (11^20) in the human genome
    - 426 years

Today:
Gene Regulation and Motif Discovery

Gene regulation: The process by which genes are turned on or off, in response to environmental stimuli

Regulatory motifs: sequences that control gene usage; short sequence patterns, ~6-12 letters long, possibly degenerate

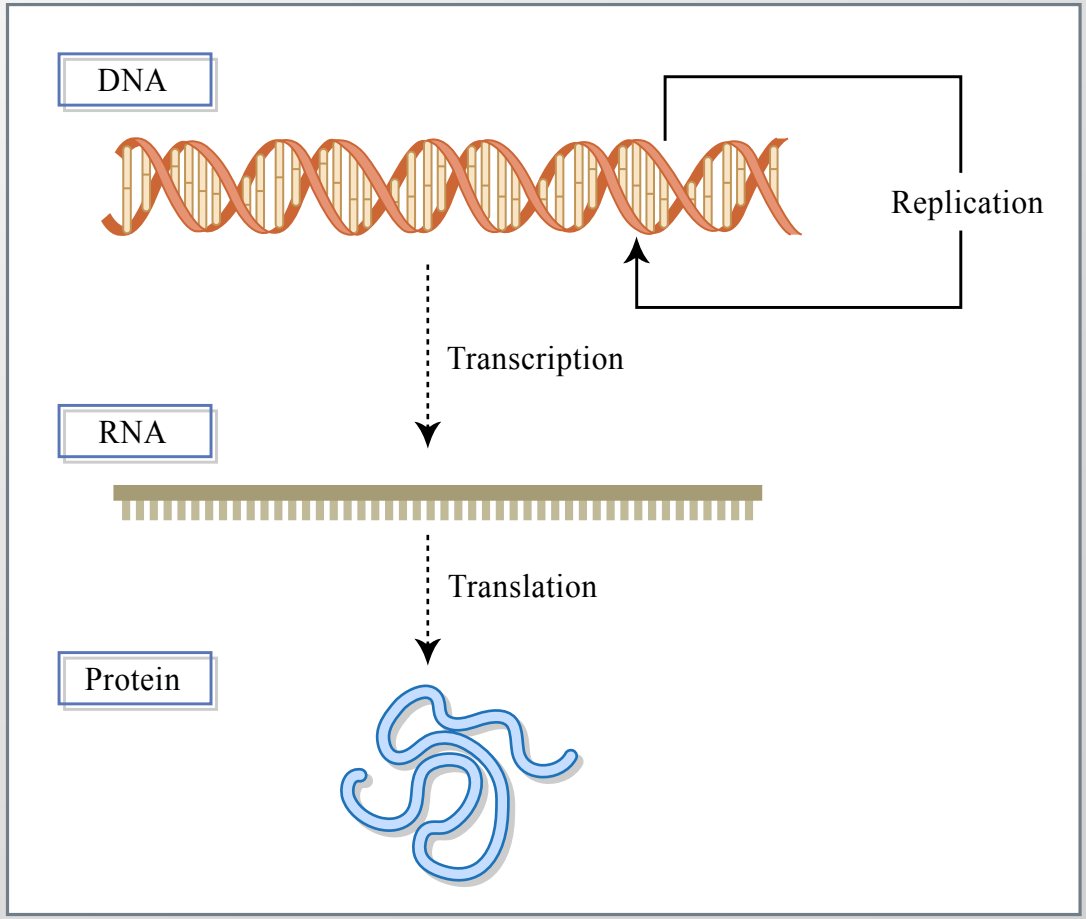
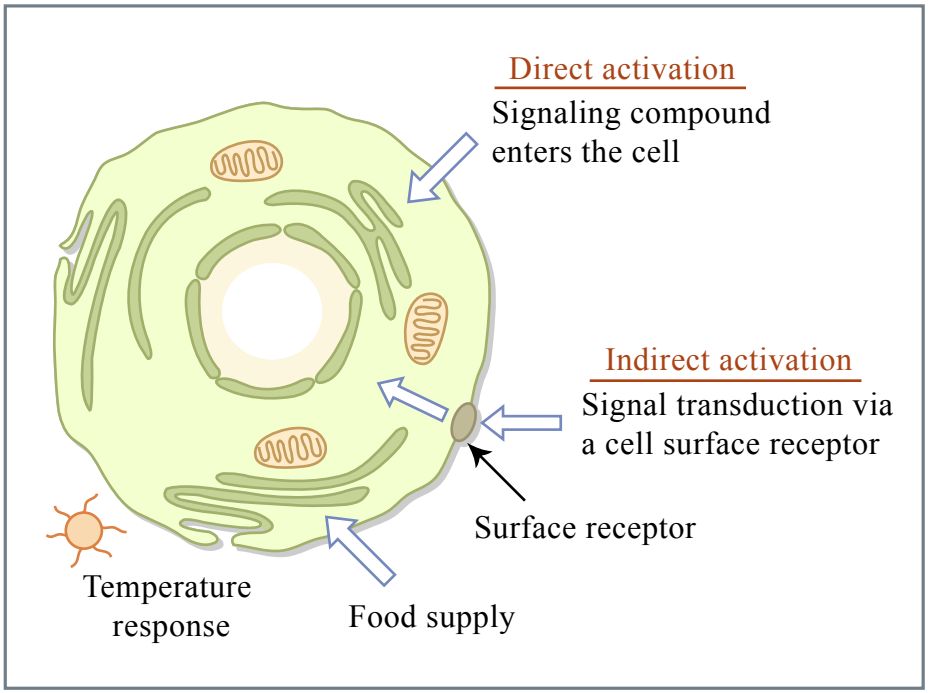


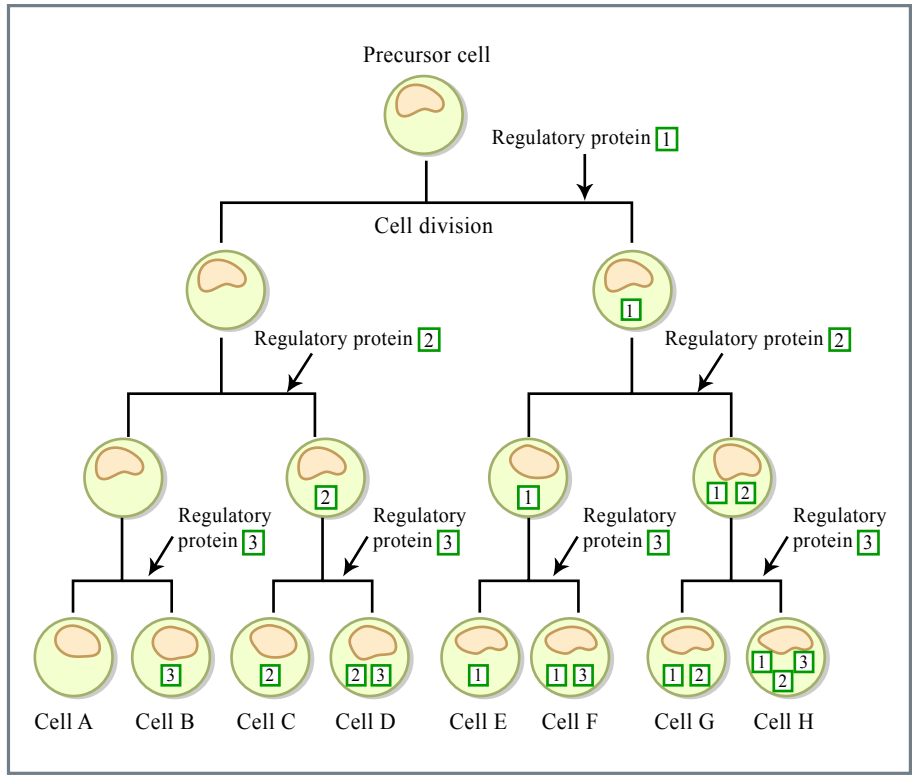
Figure by MIT OCW.

Why cellular programs change

- Environmental Response



- Cell differentiation



– Cells adapt to their environment, carry out different molecular processes, depending on their environment

– Produce same nutrients in entirely different pathways

– Cells have distinct functions: hair, nail, skin, heart, eye, brain, muscle, bone

– Cells differentiate, by using different parts of the same genome

– These morphological changes are due to expression levels

- Genome Remains Unchanged!

Figures by MIT OCW.

How cellular programs change

Regulatory knobs

- DNA level: gene dosage
  - How many copies of a particular gene
  - How many homologs, how many pathways
  - Accessibility of gene within chromatin
- mRNA: Transcription initiation
  - Regulatory motifs recognized by transcription factors
  - Transcription factors recruit transcription machinery
  - Dictates number of messages sent to cytoplasm
- mRNA: Post-transcriptional control
  - How long messages stay active
  - How fast messages they degraded
- Protein: Translation level
  - How many times is each message translated to protein
  - How stable are protein products, how long before degraded
- Protein: Post-translational modifications
  - Some proteins only perform their functions when phosphorylated
  - Some are only active as a hetero-dimer, can regulate only one.

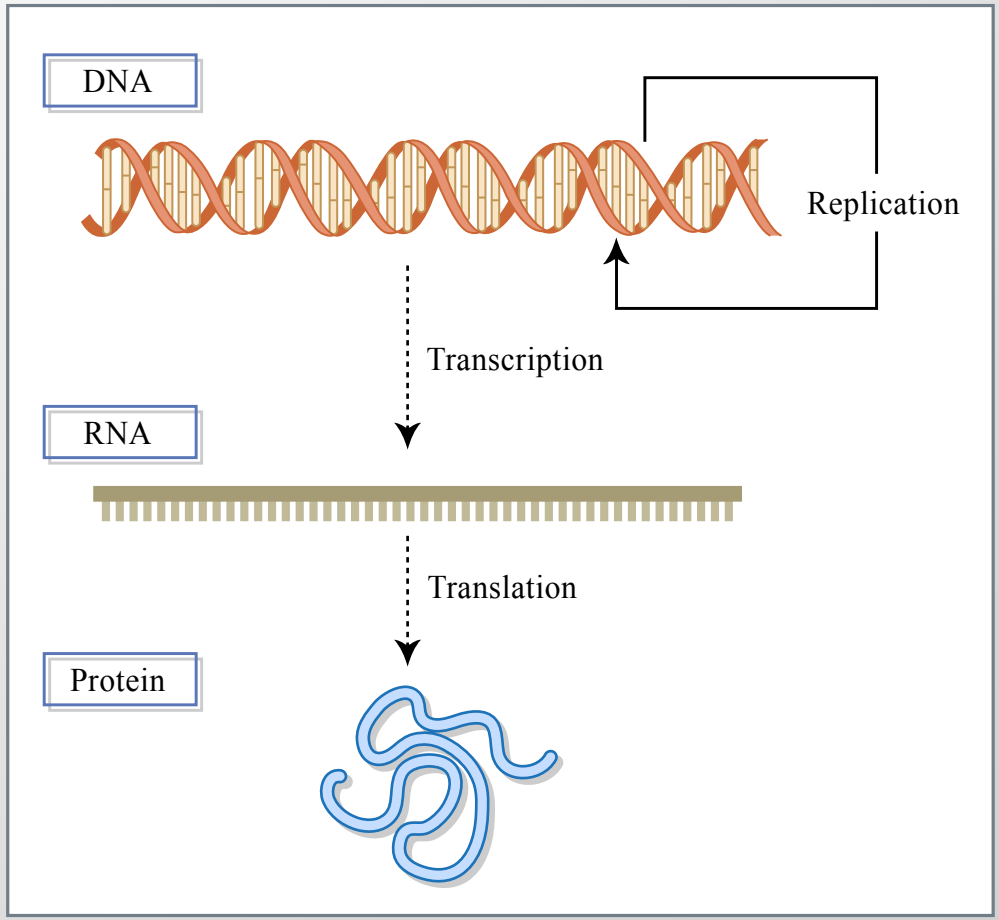
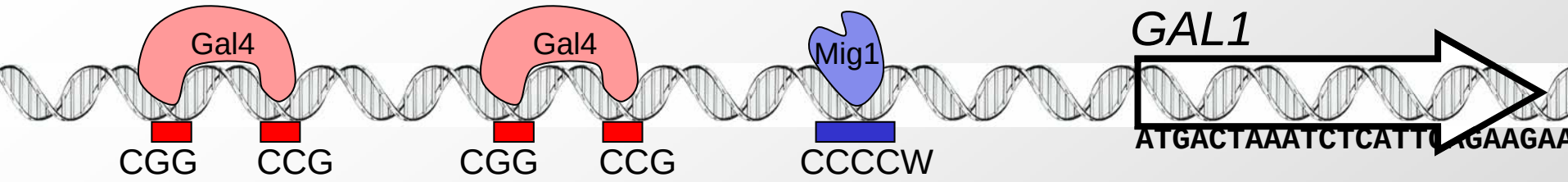


Figure by MIT OCW.

Regulatory motif discovery



- Regulatory motifs
  - Genes are turned on / off in response to changing environments
  - No direct addressing: subroutines (genes) contain sequence tags (motifs)
  - Specialized proteins (transcription factors) recognize these tags

- What makes motif discovery hard?
  - Motifs are short (6-8 bp), sometimes degenerate
  - Can contain any set of nucleotides (no ATG or other rules)
  - Act at variable distances upstream (or downstream) of target gene

Protein / DNA contact dictates regulatory motifs

- Sequence specificity
  - Topology of 3D contact dictates sequence specificity of binding
  - Some positions are fully constrained; other positions are degenerate
- Protein-DNA interactions
  - Proteins read DNA by "feeling" the chemical properties of the bases
  - Without opening DNA (not by base complementarity)

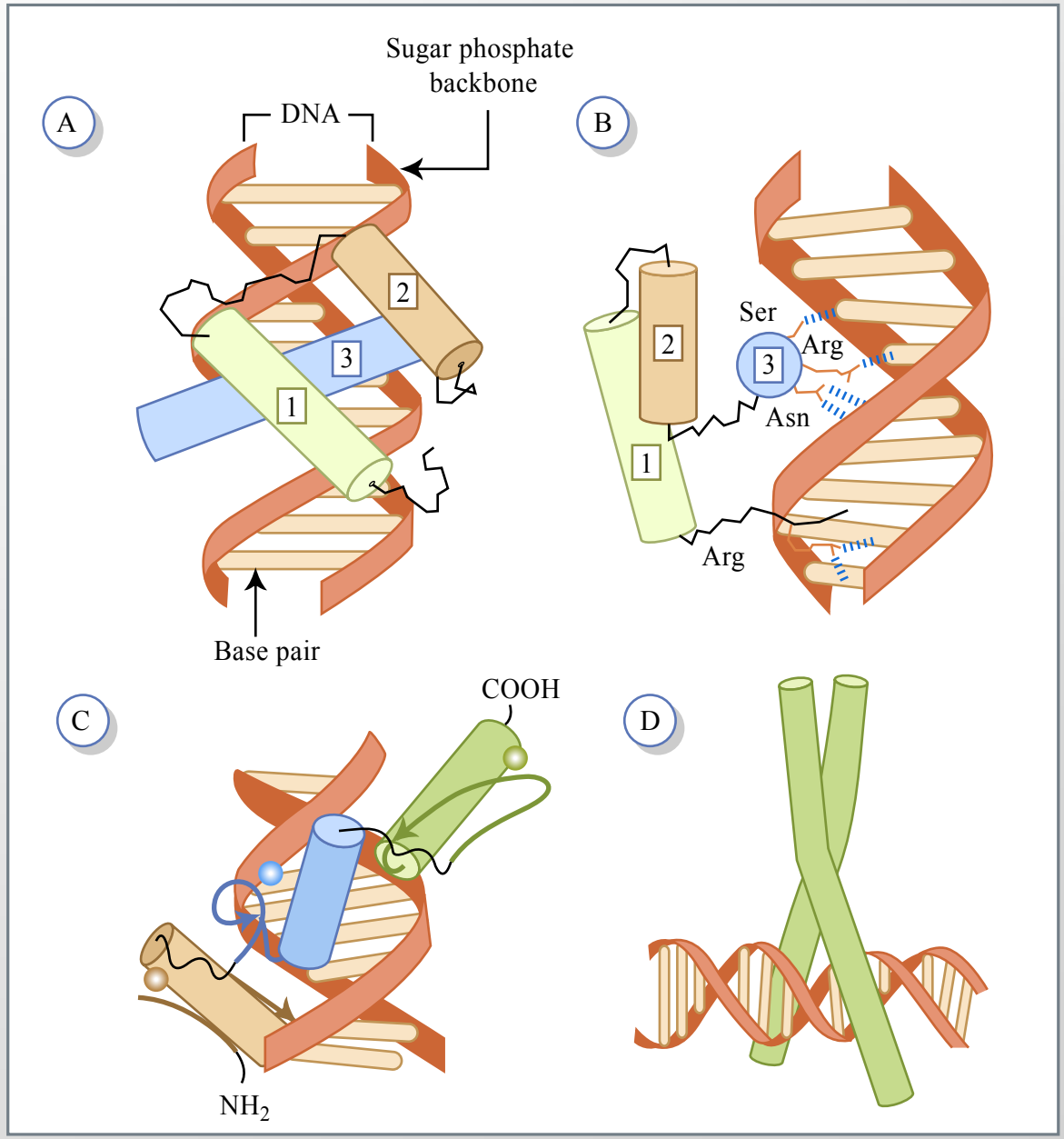
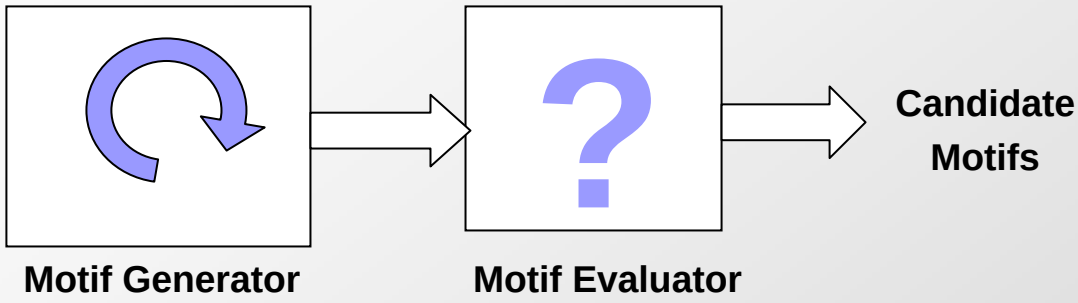


Figure by MIT OCW.

Computational approaches

- Method #1: Enumerate all motifs
- Method #2: Randomly sample the genome
- Method #3: Enumerate motif seeds + refinement
- Method #4: Content-based addressing

Need: Evaluation method



- To test whether a motif is meaningful:
  - Evaluate its conservation rate

Lecture continued on the blackboard