6.047 / 6.878 Computational Biology: Genomes, Networks, Evolution
Fall 2008

6.047 / 6.878
Computational Biology:
Genomes, Networks, Evolution

Manolis Kellis

James Galagan

Goals for the term

- Introduction to computational biology
  - Fundamental problems in computational biology
  - Algorithmic/machine learning techniques for data analysis
  - Research directions for active participation in the field
- Ability to tackle research
  - Problem set questions: algorithmic rigorous thinking
  - Programming assignments: hands-on experience w/ real datasets
- Final project:
  - Research initiative to propose an innovative project
  - Ability to carry out project's goals, produce deliverables
  - Write-up goals, approach, and findings in conference format
  - Present your project to your peers in conference setting

Course outline

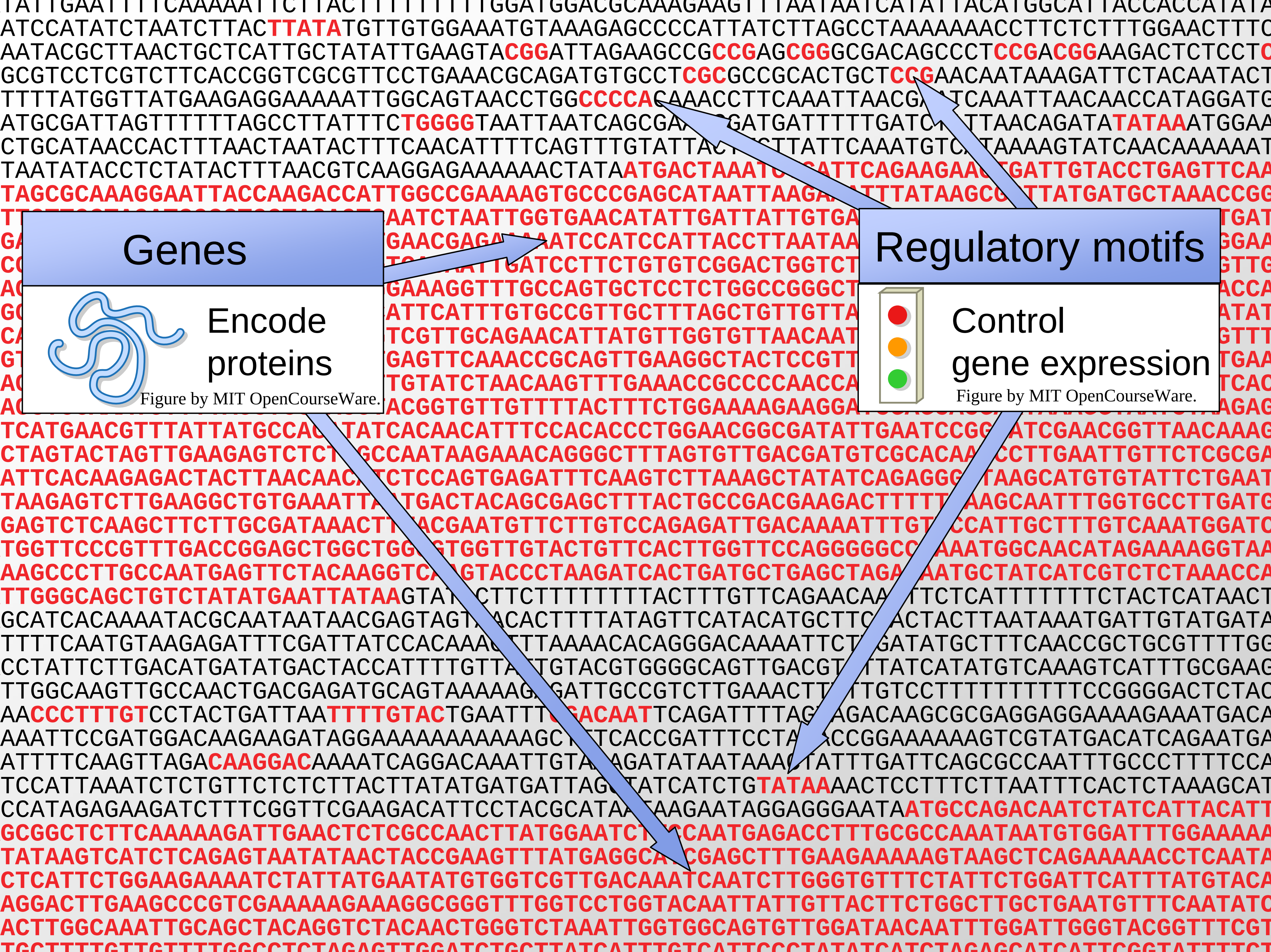
- Organization
  - Duality: Computation and Biology
    - Important biological problems
    - Fundamental computational techniques
  - Foundations and Frontiers
    - First half: well-defined problems and general methodologies
    - Second half: in-depth look at complex problems, combine techniques learned, opens to projects, research directions
- Topics covered
  - First half: the foundations
    - String matching, genome analysis, expression clustering/classification, regulatory motifs, biological networks, evolutionary theory, populations
  - Second half: the frontiers
    - Comparative genomics, Bayesian networks, systems biology, genome assembly, metabolic modeling, miRNA, genome evolution

Why Computational Biology ?

Why Computational Biology: Last year's answers

- Lots of data (* lots of data)
- There are rules
- Pattern finding
- It's *all* about data
- Ability to visualize
- Simulations
- Guess + verify (generate hypotheses for testing)
- Propose mechanisms / theory to explain observations
- Networks / combinations of variables
- Efficiency (reduce experimental space to cover)
- Informatics infrastructure (ability to combine datasets)
- Correlations
- Life itself is digital. Understand cellular instruction set

```
TATTGAATTTTCAAAAATTCTTACTTTTTTTTTGGATGGACGCAAAGAAGTTTAATAATCATATTACATGGCATTACCACCATATA
ATCCATATCTAATCTTACTTATATGTTGTGGAAATGTAAAGAGCCCATTATCTTAGCCTAAAAAAACCTTCTCTTTGGAACTTTC
AATACGCTTAACTGCTCATTGCTATATTGAAGTACGGATTAGAAGCCGCCGAGCGGGCGACAGCCCTCCGACGGAAGACTCTCCTC
GCGTCCTCGTCTTCACCGGTCGCGTTCCTGAAACGCAGATGTGCCTCGCGCCGCACTGCTCCGAACAATAAAGATTCTACAATACT
TTTTATGGTTATGAAGAGGAAAAATTGGCAGTAACCTGGCCCCACAAACCTTCAAATTAACGAATCAAATTAACAACCATAGGATG
ATGCGATTAGTTTTTAGCCTTATTTCTGGGGTAATTAATCAGCGAAGCGATGATTTTTGATCTATTAACAGATATATAAATGGAA
CTGCATAACCACTTTAACTAATACTTTCAACATTTTCAGTTTGTATTACTTCTTATTCAAATGTCATAAAAGTATCAACAAAAAAT
TAATATACCTCTATACTTTAACGTCAAGGAGAAAAAACTATAATGACTAAATCTCATTCAGAAGAAGTGATTGTACCTGAGTTCAA
TAGCGCAAAGGAATTACCAAGACCATTGGCCGAAAAGTGCCCGAGCATAATTAAGAAATTTATAAGCGCTTATGATGCTAAACCGG
TTGTTGCTAGATCGCCTGGTAGAGTCAATCTAATTGGTGAACATATTGATTATTGTGACTTCTCGGTTTTACCTTTAGCTATTGAT
GATATGCTTTGCGCCGTCAAAGTTTTGAACGAGAAAAATCCATCCATTACCTTAATAAATGCTGATCCCAAATTTGCTCAAAGGAA
CGATTTGCCGTTGGACGGTTCTTATGTCACAATTGATCCTTCTGTGTCGGACTGGTCTAATTACTTTAAATGTGGTCTCCATGTTG
ACTCTTTTCTAAAGAAACTTGCACCGGAAAGGTTTGCCAGTGCTCCTCTGGCCGGGCTGCAAGTCTTCTGTGAGGGTGATGTACCA
GGCAGTGGATTGTCTTCTTCGGCCGCATTCATTTGTGCCGTTGCTTTAGCTGTTGTTAAAGCGAATATGGGCCCTGGTTATCATAT
CAAGCAAAATTTAATGCGTATTACGGTCGTTGCAGAACATTATGTTGGTGTTAACAATGGCGGTATGGATCAGGCTGCCTCTGTTT
GTGAGGAAGATCATGCTCTATACGTTGAGTTCAAACCGCAGTTGAAGGCTACTCCGTTTAAATTTCCGCAATTAAAAAACCATGAA
AGCTTTGTTATTGCGAACACCCTTGTTGTATCTAACAAGTTTGAAACCGCCCCAACCAACTATAATTTAAGAGTGGTAGAAGTCAC
AGCTGCAAATGTTTTAGCTGCCACGTACGGTGTTGTTTTACTTTCTGGAAAAGAAGGATCGAGCACGAATAAAGGTAATCTAAGAG
TCATGAACGTTTATTATGCCAGATATCACAACATTTCCACACCCTGGAACGGCGATATTGAATCCGGCATCGAACGGTTAACAAAG
CTAGTACTAGTTGAAGAGTCTCTCGCCAATAAGAAACAGGGCTTTAGTGTTGACGATGTCGCACAATCCTTGAATTGTTCTCGCGA
ATTCACAAGAGACTACTTAACAACATCTCCAGTGAGATTTCAAGTCTTAAAGCTATATCAGAGGGCTAAGCATGTGTATTCTGAAT
TAAGAGTCTTGAAGGCTGTGAAATTAATGACTACAGCGAGCTTTACTGCCGACGAAGACTTTTTCAAGCAATTTGGTGCCTTGATG
GAGTCTCAAGCTTCTTGCGATAAACTTTACGAATGTTCTTGTCCAGAGATTGACAAAATTTGTTCCATTGCTTTGTCAAATGGATC
TGGTTCCCGTTTGACCGGAGCTGGCTGGGGTGGTTGTACTGTTCACTTGGTTCCAGGGGGCCCAAATGGCAACATAGAAAAGGTAA
AAGCCCTTGCCAATGAGTTCTACAAGGTCAAGTACCCTAAGATCACTGATGCTGAGCTAGAAAATGCTATCATCGTCTCTAAACCA
TTGGGCAGCTGTCTATATGAATTATAAGTATACTTCTTTTTTTTTACTTTGTTCAGAACAACTTCTCATTTTTTTCTACTCATAACT
GCATCACAAAATACGCAATAATAACGAGTAGTAACACTTTTATAGTTCATACATGCTTCAACTACTTAATAAATGATTGTATGATA
TTTTCAATGTAAGAGATTTCGATTATCCACAAACTTTAAAACACAGGGACAAAATTCTTGATATGCTTTCAACCGCTGCGTTTTGG
CCTATTCTTGACATGATATGACTACCATTTTGTTATTGTACGTGGGGCAGTTGACGTCTTATCATATGTCAAAGTCATTTGCGAAG
TTGGCAAGTTGCCAACTGACGAGATGCAGTAAAAAGAGATTGCCGTCTTGAAACTTTTTGTCCTTTTTTTTTTCCGGGGACTCTAC
AACCCTTTGTCCTACTGATTAATTTTGTACTGAATTTGGACAATTCAGATTTTAGTAGACAAGCGCGAGGAGGAAAAGAAATGACA
AAATTCCGATGGACAAGAAGATAGGAAAAAAAAAAAGCTTTCACCGATTTCCTAGACCGGAAAAAAAGTCGTATGACATCAGAATGA
ATTTTCAAGTTAGACAAGGACAAAATCAGGACAAATTGTAAAGATATAATAAACTATTTGATTCAGCGCCAATTTGCCCTTTTCCA
TCCATTAAATCTCTGTTCTCTCTTACTTATATGATGATTAGGTATCATCTGTATAAAACTCCTTTCTTAATTTCACTCTAAAGCAT
CCATAGAGAAGATCTTTCGGTTCGAAGACATTCCTACGCATAATAAGAATAGGAGGGAATAATGCCAGACAATCTATCATTACATT
GCGGCTCTTCAAAAAGATTGAACTCTCGCCAACTTATGGAATCTTCCAATGAGACCTTTGCGCCAAATAATGTGGATTTGGAAAAA
TATAAGTCATCTCAGAGTAATATAACTACCGAAGTTTATGAGGCATCGAGCTTTGAAGAAAAAGTAAGCTCAGAAAAACCTCAATA
CTCATTCTGGAAGAAAATCTATTATGAATATGTGGTCGTTGACAAATCAATCTTGGGTGTTTCTATTCTGGATTCATTTATGTACA
AGGACTTGAAGCCCGTCGAAAAAGAAAGGCGGGTTTGGTCCTGGTACAATTATTGTTACTTCTGGCTTGCTGAATGTTTCAATATC
ACTTGGCAAATTGCAGCTACAGGTCTACAACTGGGTCTAAATTGGTGGCAGTGTTGGATAACAATTTGGATTGGGTACGGTTTCGT
TGCTTTTGTTGTTTTGGCCTCTAGAGTTGGATCTGCTTATCATTTGTCATTCCCTATATCATCTAGAGCATCATTCGGTATTTTCT
```



Genes

Encode proteins

Figure by MIT OpenCourseWare.

Regulatory motifs

Control gene expression

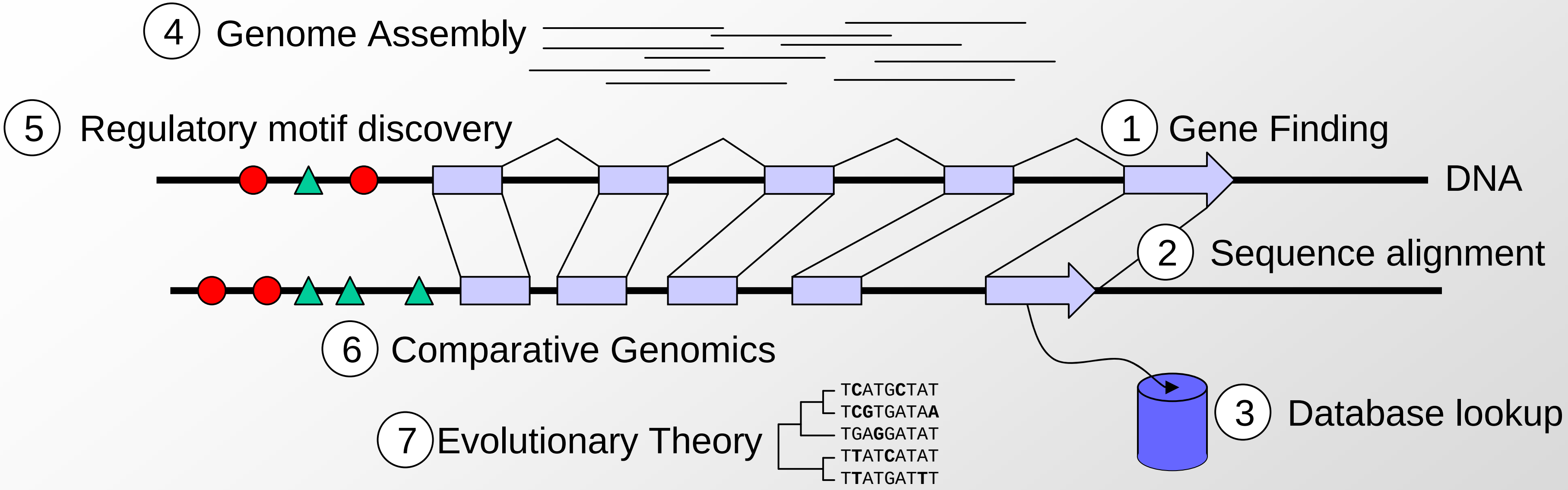
Figure by MIT OpenCourseWare.

TATTGAATTTTCAAAAATTCTTACTTTTTTTTTTGGATGGACGCAAAGAAGTTTAATAATCATATTACATGGCATTACCACCATATA
ATCCATATCTAATCTTAC**TTATA**TGTTGTGGAAATGTAAAGAGCCCATTATCTTAGCCTAAAAAAACCTTCTCTTTGGAACTTTC
AATACGCTTAACTGCTCATTGCTATATTGAAGTA**CGG**ATTAGAAGCCG**CCG**AG**CGG**GCGACAGCCCT**CCG**A**CGG**AAGACTCTCCT**C**
GCGTCCTCGTCTTCACCGGTCGCGTTCCTGAAACGCAGATGTGCCT**CGC**GCCGCACTGCT**CCG**AACAATAAAGATTCTACAATACT
TTTTATGGTTATGAAGAGGAAAAATTGGCAGTAACCTGG**CCCCA**CAAACCTTCAAATTAACGAATCAAATTAACAACCATAGGATG
ATGCGATTAGTTTTTTAGCCTTATTTC**TGGGG**TAATTAATCAGCGAAGCGATGATTTTTGATCTATTAACAGATA**TATAA**ATGGAA
CTGCATAACCACTTTAACTAATACTTTCAACATTTTCAGTTTGTATTACTTCTTATTCAAATGTCATAAAAGTATCAACAAAAAAT
TAATATACCTCTATACTTTAACGTCAAGGAGAAAAAACTATA**ATGACTAAATCTCATTCAGAAGAAGTGATTGTACCTGAGTTCAA**
TAGCGCAAAGGAATTACCAAGACCATTGGCCGAAAAGTGCCCGAGCATAATTAAGAAATTTATAAGCGCTTATGATGCTAAACCGG
TTGTTGCTAGATCGCCTGGTAGAGTCAATCTAATTGGTGAACATATTGATTATTGTGACTTCTCGGTTTTACCTTTAGCTATTGAT
GATATGCTTTGCGCCGTCAAAGTTTTGAACGAGAAAAATCCATCCATTACCTTAATAAATGCTGATCCCAAATTTGCTCAAAGGAA
CGATTTGCCGTTGGACGGTTCTTATGTCACAATTGATCCTTCTGTGTCGGACTGGTCTAATTACTTTAAATGTGGTCTCCATGTTG
ACTCTTTTCTAAAGAAACTTGCACCGGAAAGGTTTGCCAGTGCTCCTCTGGCCGGGCTGCAAGTCTTCTGTGAGGGTGATGTACCA
GGCAGTGGATTGTCTTCTTCGGCCGCATTCATTTGTGCCGTTGCTTTAGCTGTTGTTAAAGCGAATATGGGCCCTGGTTATCATAT
CAAGCAAAATTTAATGCGTATTACGGTCGTTGCAGAACATTATGTTGGTGTTAACAATGGCGGTATGGATCAGGCTGCCTCTGTTT
GTGAGGAAGATCATGCTCTATACGTTGAGTTCAAACCGCAGTTGAAGGCTACTCCGTTTAAATTTCCGCAATTAAAAAACCATGAA
AGCTTTGTTATTGCGAACACCCTTGTTGTATCTAACAAGTTTGAAACCGCCCCAACCAACTATAATTTAAGAGTGGTAGAAGTCAC
AGCTGCAAATGTTTTAGCTGCCACGTACGGTGTTGTTTTACTTTCTGGAAAAGAAGGATCGAGCACGAATAAAGGTAATCTAAGAG
TCATGAACGTTTATTATGCCAGATATCACAACATTTCCACACCCTGGAACGGCGATATTGAATCCGGCATCGAACGGTTAACAAAG
CTAGTACTAGTTGAAGAGTCTCTCGCCAATAAGAAACAGGGCTTTAGTGTTGACGATGTCGCACAATCCTTGAATTGTTCTCGCGA
ATTCACAAGAGACTACTTAACAACATCTCCAGTGAGATTTCAAGTCTTAAAGCTATATCAGAGGGCTAAGCATGTGTATTCTGAAT
TAAGAGTCTTGAAGGCTGTGAAATTAATGACTACAGCGAGCTTTACTGCCGACGAAGACTTTTTCAAGCAATTTGGTGCCTTGATG
GAGTCTCAAGCTTCTTGCGATAAACTTTACGAATGTTCTTGTCCAGAGATTGACAAAATTTGTTCCATTGCTTTGTCAAATGGATC
TGGTTCCCGTTTGACCGGAGCTGGCTGGGGTGGTTGTACTGTTCACTTGGTTCCAGGGGGCCCAAATGGCAACATAGAAAAGGTAA
AAGCCCTTGCCAATGAGTTCTACAAGGTCAAGTACCCTAAGATCACTGATGCTGAGCTAGAAAATGCTATCATCGTCTCTAAACCA
TTGGGCAGCTGTCTATATGAATTATAAGTATACTTCTTTTTTTTACTTTGTTCAGAACAACTTCTCATTTTTTTCTACTCATAACT
GCATCACAAAATACGCAATAATAACGAGTAGTAACACTTTTATAGTTCATACATGCTTCAACTACTTAATAAATGATTGTATGATA
TTTTCAATGTAAGAGATTTCGATTATCCACAAACTTTAAAACACAGGGACAAAATTCTTGATATGCTTTCAACCGCTGCGTTTTGG
CCTATTCTTGACATGATATGACTACCATTTTGTTATTGTACGTGGGGCAGTTGACGTCTTATCATATGTCAAAGTCATTTGCGAAG
TTGGCAAGTTGCCAACTGACGAGATGCAGTAAAAAGAGATTGCCGTCTTGAAACTTTTTGTCCTTTTTTTTTTTCCGGGGACTCTAC
AA**CCCTTTGT**CCTACTGATTAA**TTTTGTAC**TGAATTT**GGACAAT**TCAGATTTTAGTAGACAAGCGCGAGGAGGAAAAGAAATGACA
AAATTCCGATGGACAAGAAGATAGGAAAAAAAAAAAGCTTTCACCGATTTCCTAGACCGGAAAAAAGTCGTATGACATCAGAATGA
ATTTTCAAGTTAGA**CAAGGAC**AAAATCAGGACAAATTGTAAAGATATAATAAACTATTTGATTCAGCGCCAATTTGCCCTTTTCCA
TCCATTAAATCTCTGTTCTCTCTTACTTATATGATGATTAGGTATCATCTG**TATAA**AACTCCTTTCTTAATTTCACTCTAAAGCAT
CCATAGAGAAGATCTTTCGGTTCGAAGACATTCCTACGCATAATAAGAATAGGAGGGAATA**ATGCCAGACAATCTATCATTACATT**
GCGGCTCTTCAAAAAGATTGAACTCTCGCCAACTTATGGAATCTTCCAATGAGACCTTTGCGCCAAATAATGTGGATTTGGAAAAA
TATAAGTCATCTCAGAGTAATATAACTACCGAAGTTTATGAGGCATCGAGCTTTGAAGAAAAAGTAAGCTCAGAAAAACCTCAATA
CTCATTCTGGAAGAAAATCTATTATGAATATGTGGTCGTTGACAAATCAATCTTGGGTGTTTCTATTCTGGATTCATTTATGTACA
AGGACTTGAAGCCCGTCGAAAAAGAAAGGCGGGTTTGGTCCTGGTACAATTATTGTTACTTCTGGCTTGCTGAATGTTTCAATATC
ACTTGGCAAATTGCAGCTACAGGTCTACAACTGGGTCTAAATTGGTGGCAGTGTTGGATAACAATTTGGATTGGGTACGGTTTCGT
TGCTTTTGTTGTTTTTGGCCTCTAGAGTTGGATCTGCTTATCATTTGTCATTCCCTATATCATCTAGAGCATCATTCGGTATTTTCT

Extracting signal from noise

```
TATTGAATTTTCAAAAATTCTTACTTTTTTTTTGGATGGACGCAAAGAAGTTTAATAATCATATTACATGGCATTACCACCATATA
A CCCATTATCTTAGCCTAAAAAAACCTTCTCTTTGGAACTTTC
A AGCCGCCGAGCGGGCGACAGCCCTCCGACGGAAGACTCTCCTC
G CCTCGCGCCGCACTGCTCCGAACAATAAAGATTCTACAATACT
T ACAAACCTTCAAATTAACGAATCAAATTAACAACCATAGGATG
ATGCGATTAGTTTTTAGCCTTATTTCTGGGGTAATTAATCAGCGAAGCGATGATTTTTGATCTATTAACAGATATATAAATGGAA
CTGCATAACCACTTTAACTAATACTTTCAACATTTTCAGTTTGTATTACTTCTTATTCAAATGTCATAAAAGTATCAACAAAAAAT
TAATATACCTCTATACTTTAACGTCAAGGAGAAAAAACTATAATGACTAAATCTCATTCAGAAGAAGTGATTGTACCTGAGTTCAA
TAGCGCAAAGGAATTACCAAGACCATTGGCCGAAAAGTGCCCGAGCATAATTAAGAAATTTATAAGCGCTTATGATGCTAAACCGG
TTGTTGCTAGATCGCCTGGTAGAGTCAATCTAATTGGTGAACATATTGATTATTGTGACTTCTCGGTTTTACCTTTAGCTATTGAT
GATATGCTTTGCGCCGTCAAAGTTTTGAACGAGAAAAATCCATCCATTACCTTAATAAATGCTGATCCCAAATTTGCTCAAAGGAA
CGATTTGCCGTTGGACGGTTCTTATGTCACAATTGATCCTTCTGTGTCGGACTGGTCTAATTACTTTAAATGTGGTCTCCATGTTG
ACTCTTTTCTAAAGAAACTTGCACCGGAAAGGTTTGCCAGTGCTCCTCTGGCCGGGCTGCAAGTCTTCTGTGAGGGTGATGTACCA
GGCAGTGGATTGTCTTCTTCGGCCGCATTCATTTGTGCCGTTGCTTTAGCTGTTGTTAAAGCGAATATGGGCCCTGGTTATCATAT
CAAGCAAAATTTAATGCGTATTACGGTCGTTGCAGAACATTATGTTGGTGTTAACAATGGCGGTATGGATCAGGCTGCCTCTGTTT
GTGAGGAAGATCATGCTCTATACGTTGAGTTCAAACCGCAGTTGAAGGCTACTCCGTTTAAATTTCCGCAATTAAAAAACCATGAA
AGCTTTGTTATTGCGAACACCCTTGTTGTATCTAACAAGTTTGAAACCGCCCCAACCAACTATAATTTAAGAGTGGTAGAAGTCAC
AGCTGCAAATGTTTTAGCTGCCACGTACGGTGTTGTTTTACTTTCTGGAAAAGAAGGATCGAGCACGAATAAAGGTAATCTAAGAG
TCATGAACGTTTATTATGCCAGATATCACAACATTTCCACACCCTGGAACGGCGATATTGAATCCGGCATCGAACGGTTAACAAAG
CTAGTACTAGTTGAAGAGTCTCTCGCCAATAAGAAACAGGGCTTTAGTGTTGACGATGTCGCACAATCCTTGAATTGTTCTCGCGA
ATTCACAAGAGACTACTTAACAACATCTCCAGTGAGATTTCAAGTCTTAAAGCTATATCAGAGGGCTAAGCATGTGTATTCTGAAT
TAAGAGTCTTGAAGGCTGTGAAATTAATGACTACAGCGAGCTTTACTGCCGACGAAGACTTTTTCAAGCAATTTGGTGCCTTGATG
GAGTCTCAAGCTTCTTGCGATAAACTTTACGAATGTTCTTGTCCAGAGATTGACAAAATTTGTTCCATTGCTTTGTCAAATGGATC
TGGTTCCCGTTTGACCGGAGCTGGCTGGGGTGGTTGTACTGTTCACTTGGTTCCAGGGGGCCCAAATGGCAACATAGAAAAGGTAA
AAGCCCTTGCCAATGAGTTCTACAAGGTCAAGTACCCTAAGATCACTGATGCTGAGCTAGAAAATGCTATCATCGTCTCTAAACCA
TTGGGCAGCTGTCTATATGAATTATAAGTATACTTCTTTTTTTTACTTTGTTCAGAACAACTTCTCATTTTTTTCTACTCATAACT
GCATCACAAAATACGCAATAATAACGAGTAGTAACACTTTTATAGTTCATACATGCTTCAACTACTTAATAAATGATTGTATGATA
TTTTCAATGTAAGAGATTTCGATTATCCACAAACTTTAAAACACAGGGACAAAATTCTTGATATGCTTTCAACCGCTGCGTTTTGG
CCTATTCTTGACATGATATGACTACCATTTTGTTATTGTACGTGGGGCAGTTGACGTCTTATCATATGTCAAAGTCATTTGCGAAG
TTGGCAAGTTGCCAACTGACGAGATGCAGTAAAAAGAGATTGCCGTCTTGAAACTTTTTGTCCTTTTTTTTTTTCCGGGGACTCTAC
AACCCTTTGTCCTACTGATTAATTTTGTACTGAATTTGGACAATTCAGATTTTAGTAGACAAGCGCGAGGAGGAAAAGAAATGACA
AAATTCCGATGGACAAGAAGATAGGAAAAAAAAAAAGCTTTCACCGATTTCCTAGACCGGAAAAAAGTCGTATGACATCAGAATGA
ATTTTCAAGTTAGACAAGGACAAAATCAGGACAAATTGTAAAGATATAATAAACTATTTGATTCAGCGCCAATTTGCCCTTTTCCA
TCCATTAAATCTCTGTTCTCTCTTACTTATATGATGATTAGGTATCATCTGTATAAAACTCCTTTCTTAATTTCACTCTAAAGCAT
CCATAGAGAAGATCTTTCGGTTCGAAGACATTCCTACGCATAATAAGAATAGGAGGGAATAATGCCAGACAATCTATCATTACATT
GCGGCTCTTCAAAAAGATTGAACTCTCGCCAACTTATGGAATCTTCCAATGAGACCTTTGCGCCAAATAATGTGGATTTGGAAAAA
TATAAGTCATCTCAGAGTAATATAACTACCGAAGTTTATGAGGCATCGAGCTTTGAAGAAAAAGTAAGCTCAGAAAAACCTCAATA
CTCATTCTGGAAGAAAATCTATTATGAATATGTGGTCGTTGACAAATCAATCTTGGGTGTTTCTATTCTGGATTCATTTATGTACA
AGGACTTGAAGCCCGTCGAAAAAGAAAGGCGGGTTTGGTCCTGGTACAATTATTGTTACTTCTGGCTTGCTGAATGTTTCAATATC
ACTTGGCAAATTGCAGCTACAGGTCTACAACTGGGTCTAAATTGGTGGCAGTGTTGGATAACAATTTGGATTGGGTACGGTTTCGT
TGCTTTTGTTGTTTTGGCCTCTAGAGTTGGATCTGCTTATCATTTGTCATTCCCTATATCATCTAGAGCATCATTCGGTATTTTCT
```

Challenges in Computational Biology



4 Genome Assembly

5 Regulatory motif discovery

1 Gene Finding

DNA

2 Sequence alignment

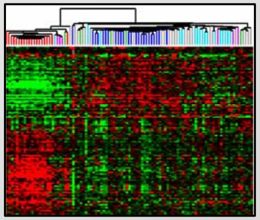
6 Comparative Genomics

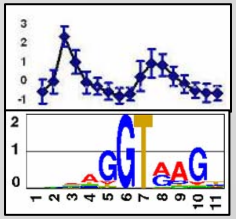
3 Database lookup

7 Evolutionary Theory

8 Gene expression analysis

RNA transcript

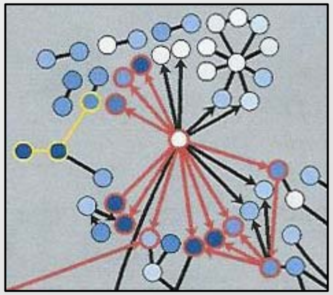




9 Cluster discovery

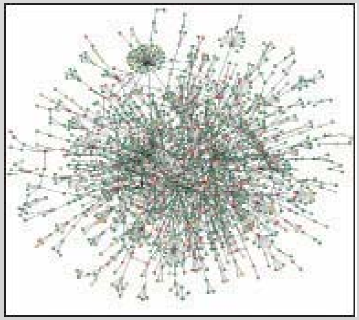
10 Gibbs sampling

11 Protein network analysis



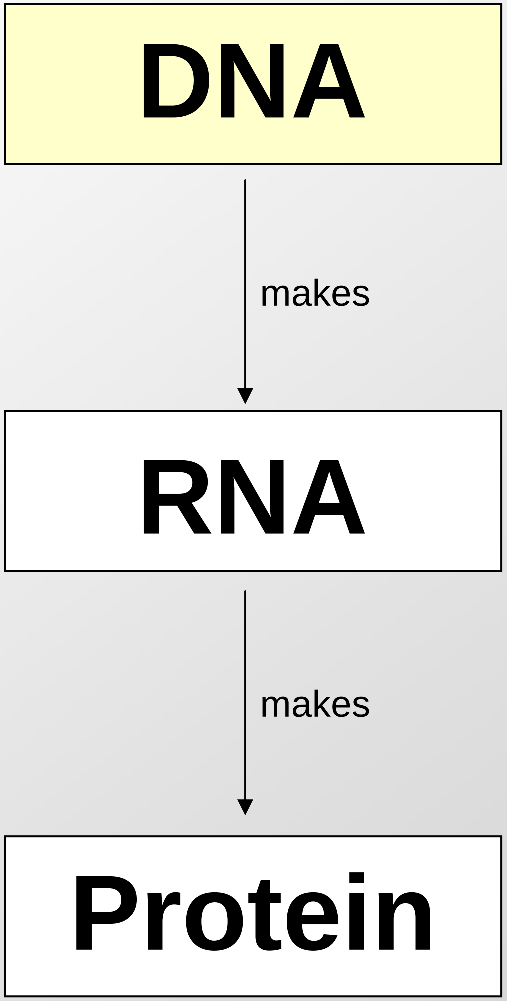
12 Metabolic modelling

13 Emerging network properties



Molecular Biology Primer

“Central dogma” of Molecular Biology



DNA

makes

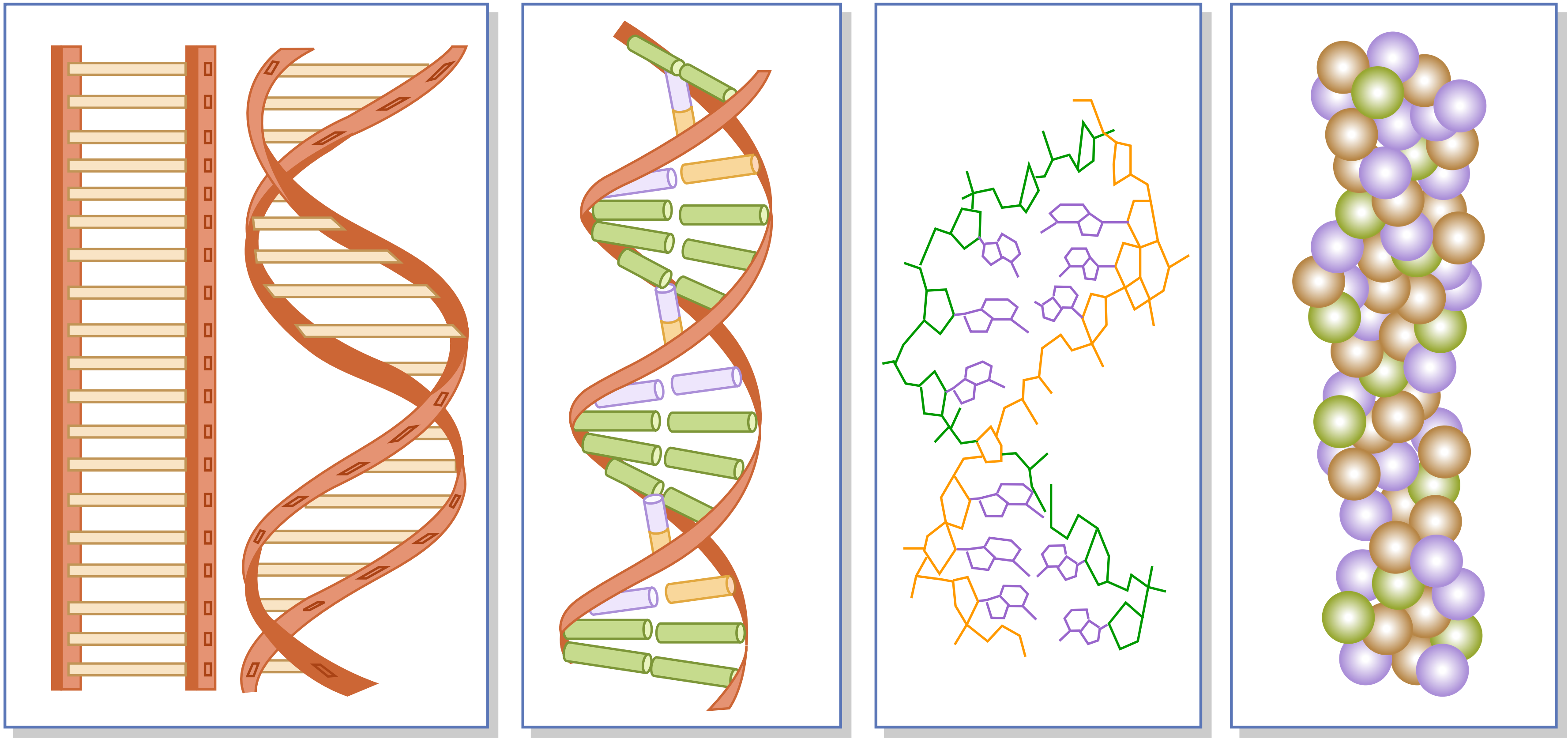
RNA

makes

Protein

DNA: The double helix

- The most noble molecule of our time



In fact, the two DNA strands are twisted around each other to make a double helix.

Traditional

Fancy

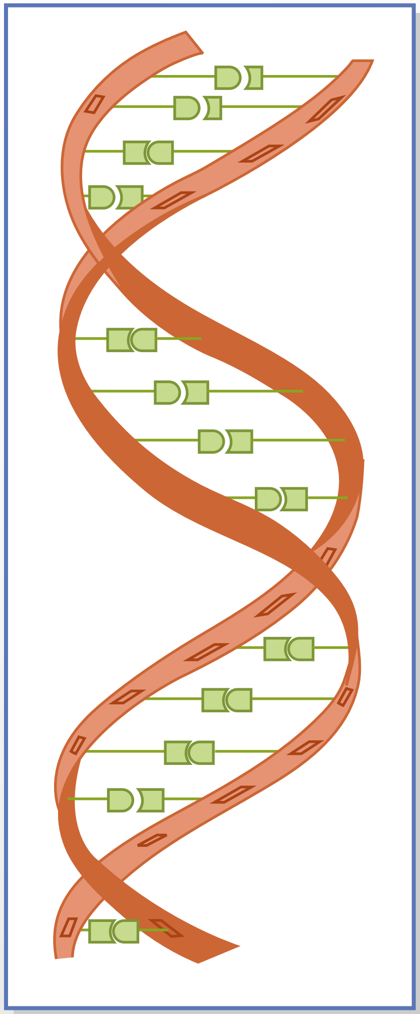
Chemical

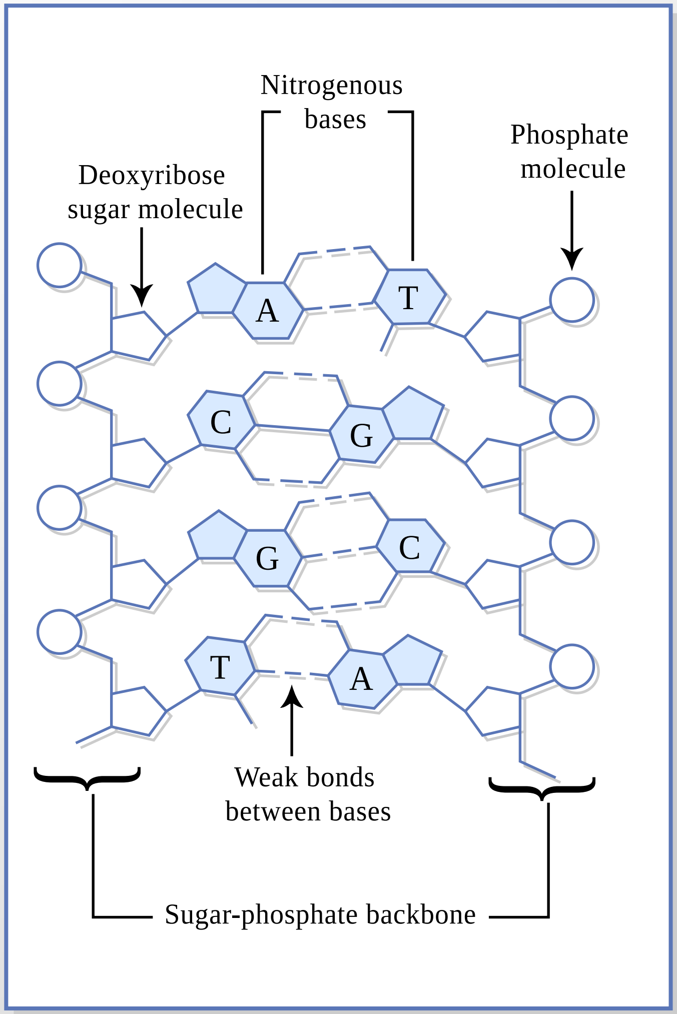
Atomic

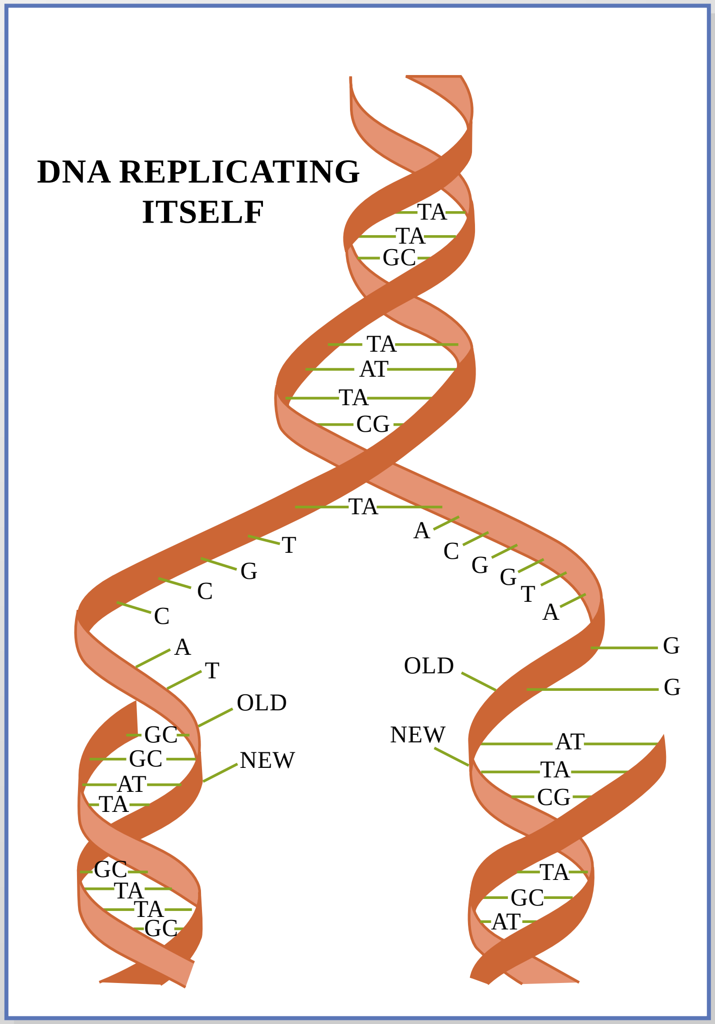
Figures by MIT OpenCourseWare.

DNA: the molecule of heredity

- Self-complementarity sets molecular basis of heredity
  - Knowing one strand, creates a template for the other
  - “It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material.” Watson & Crick, 1953

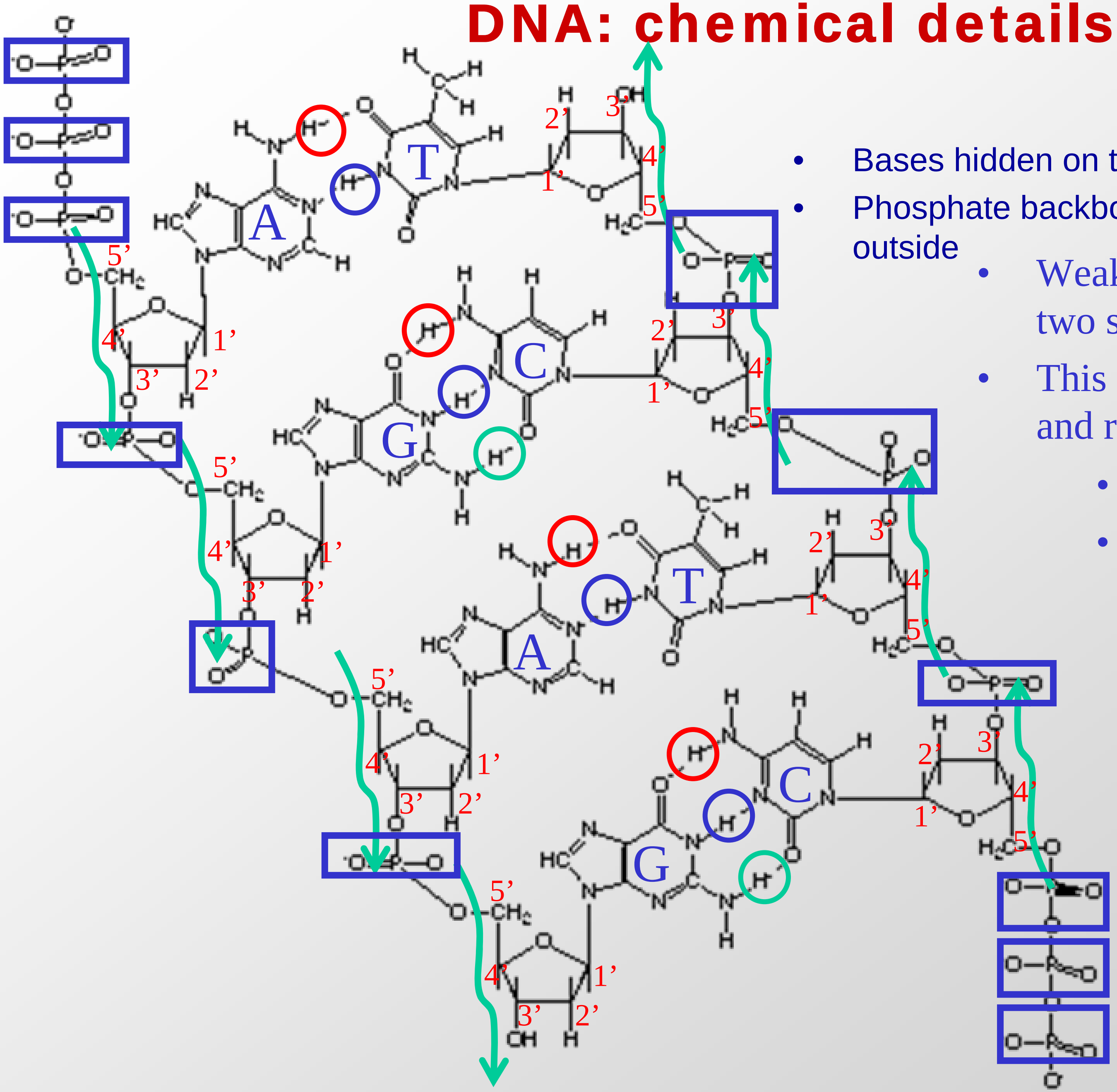






Figures by MIT OpenCourseWare.

DNA: chemical details

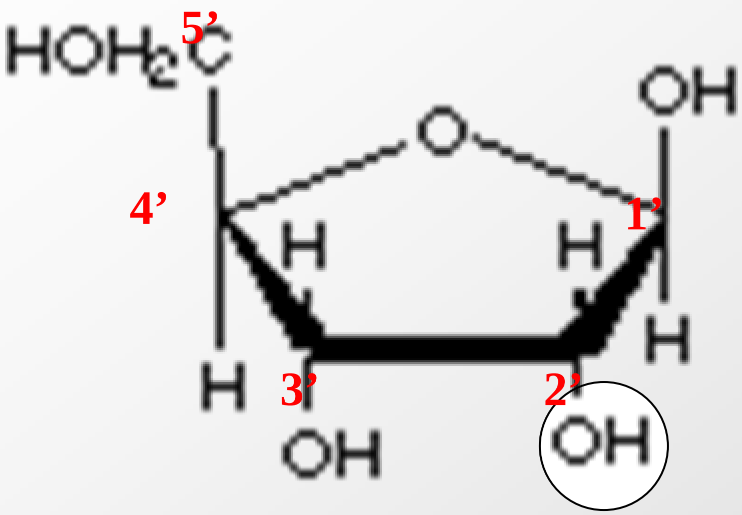


- Bases hidden on the inside
- Phosphate backbone outside
- Weak hydrogen bonds hold the two strands together
- This allows low-energy opening and re-closing of two strands
  - Anti-parallel strands
  - Extension 5’→3’ tri-phosphate coming from newly added nucleotide

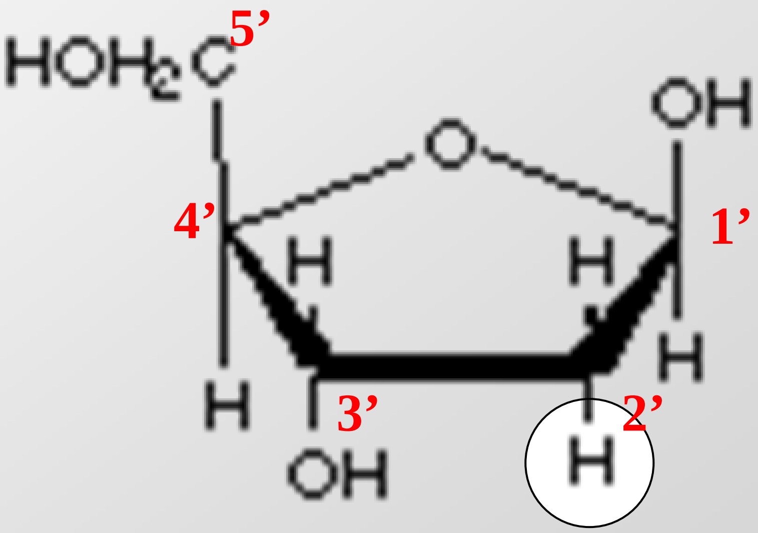
The only parings are:

- A with T
- C with G

DNA: deoxyribose sugar

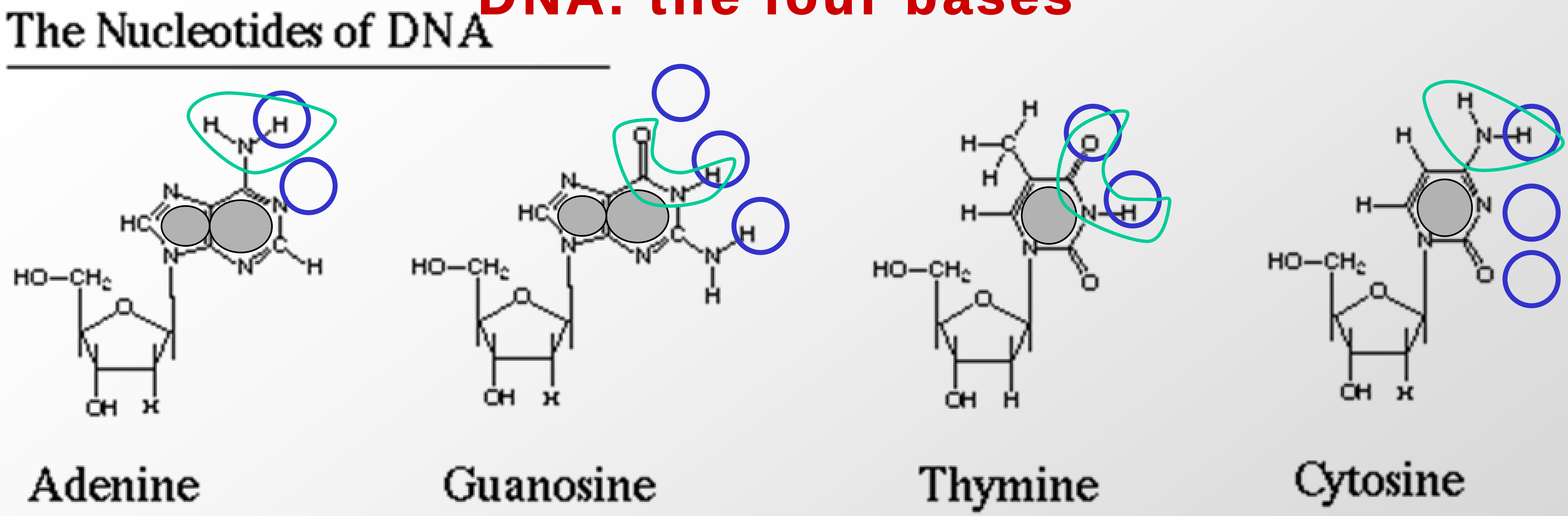


Ribose
(in RNA)



Deoxyribose
(in DNA)

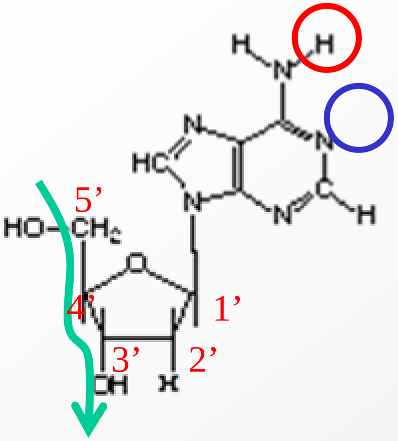
DNA: the four bases



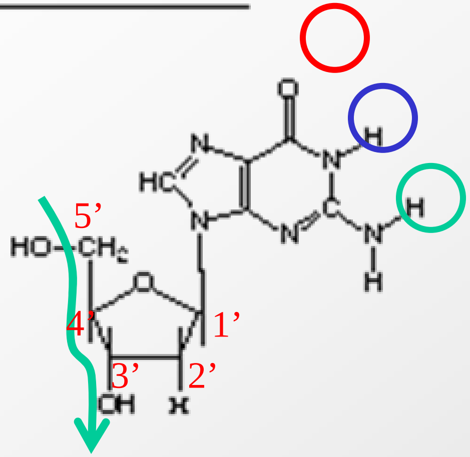
Purine	Purine		
		Pyrimidine	Pyrimidine
Weak		Weak	
	Strong		Strong
Amino			Amino
	Keto	Keto	

DNA: base pairs

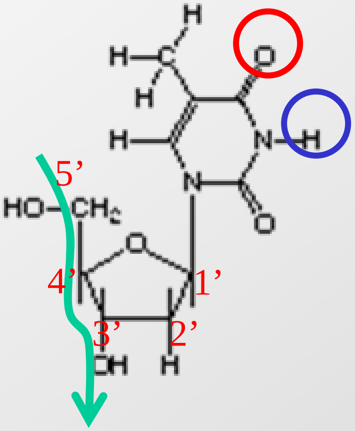
The Nucleotides of DNA



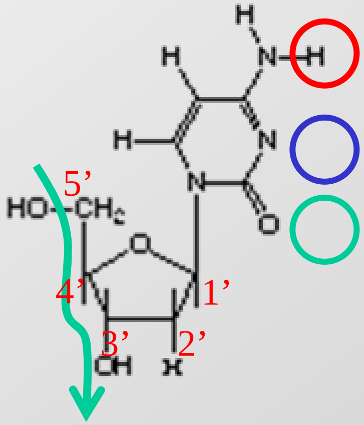
Adenine



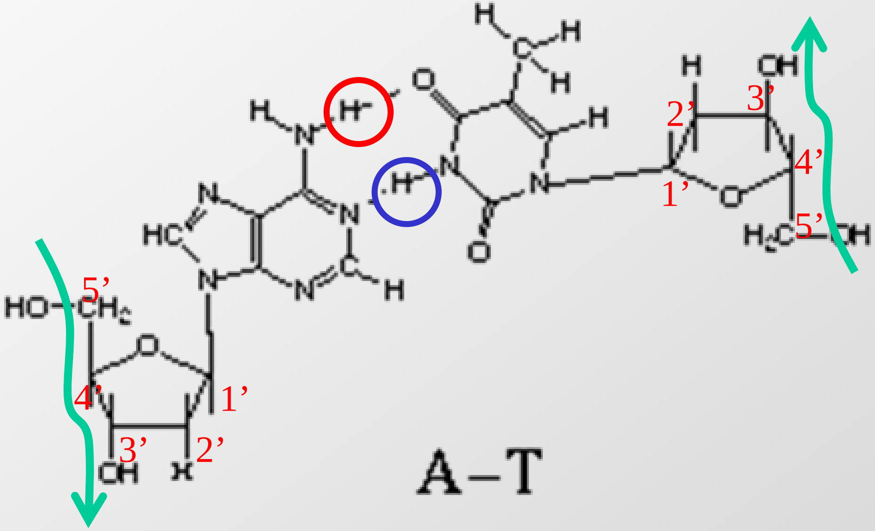
Guanosine



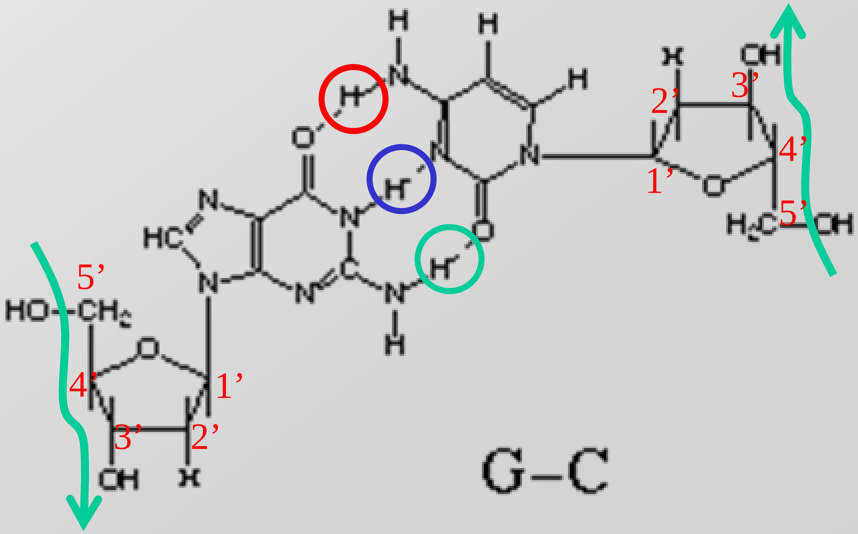
Thymine



Cytosine

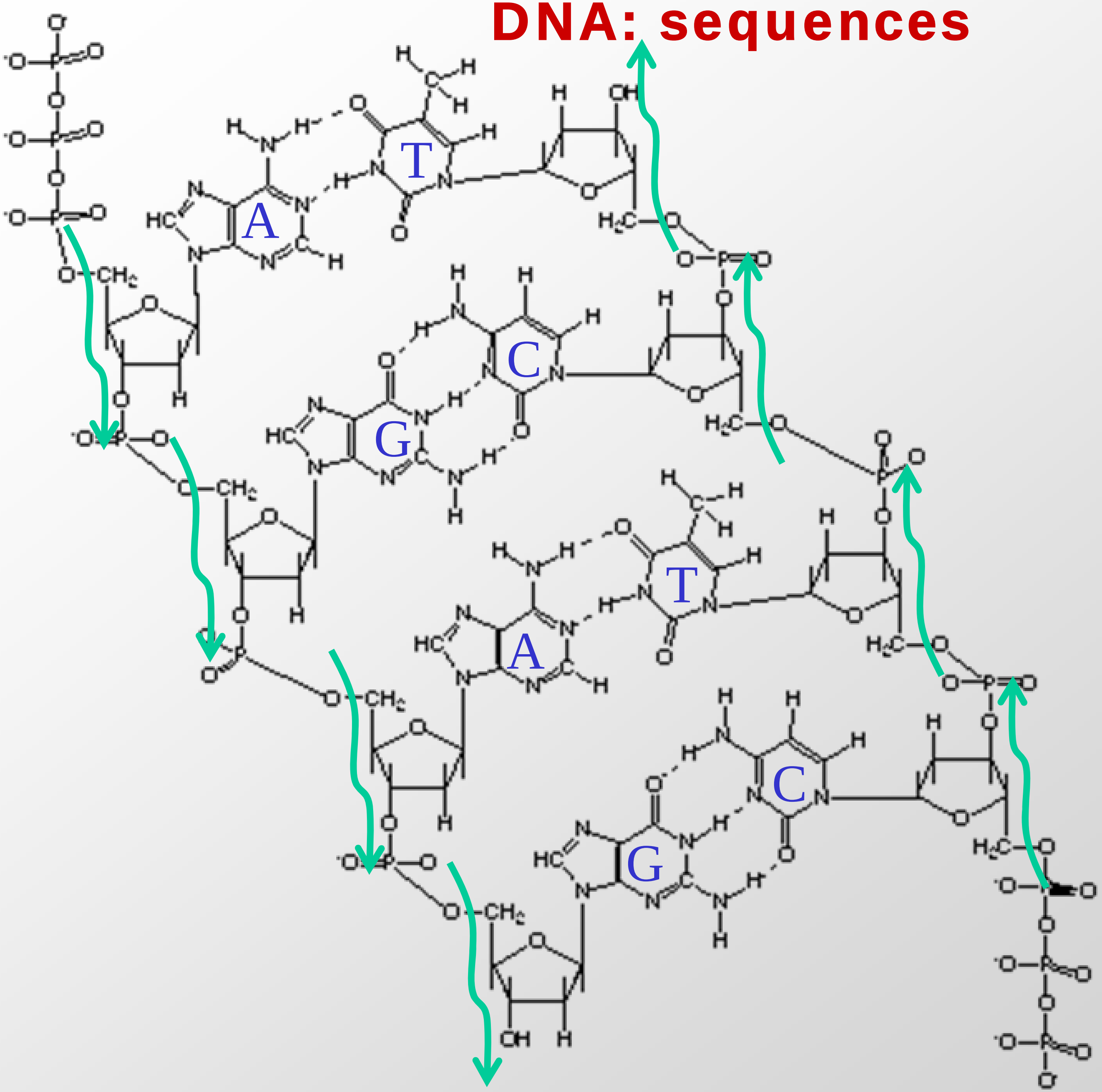


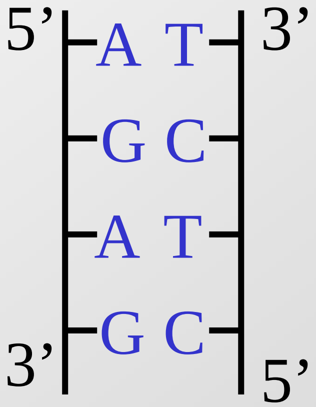
A–T

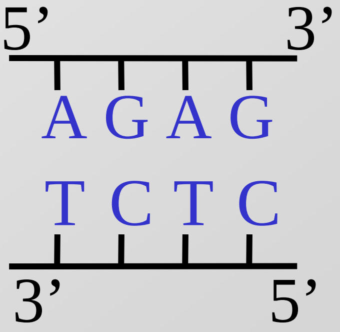


G–C

DNA: sequences







AGAG

or

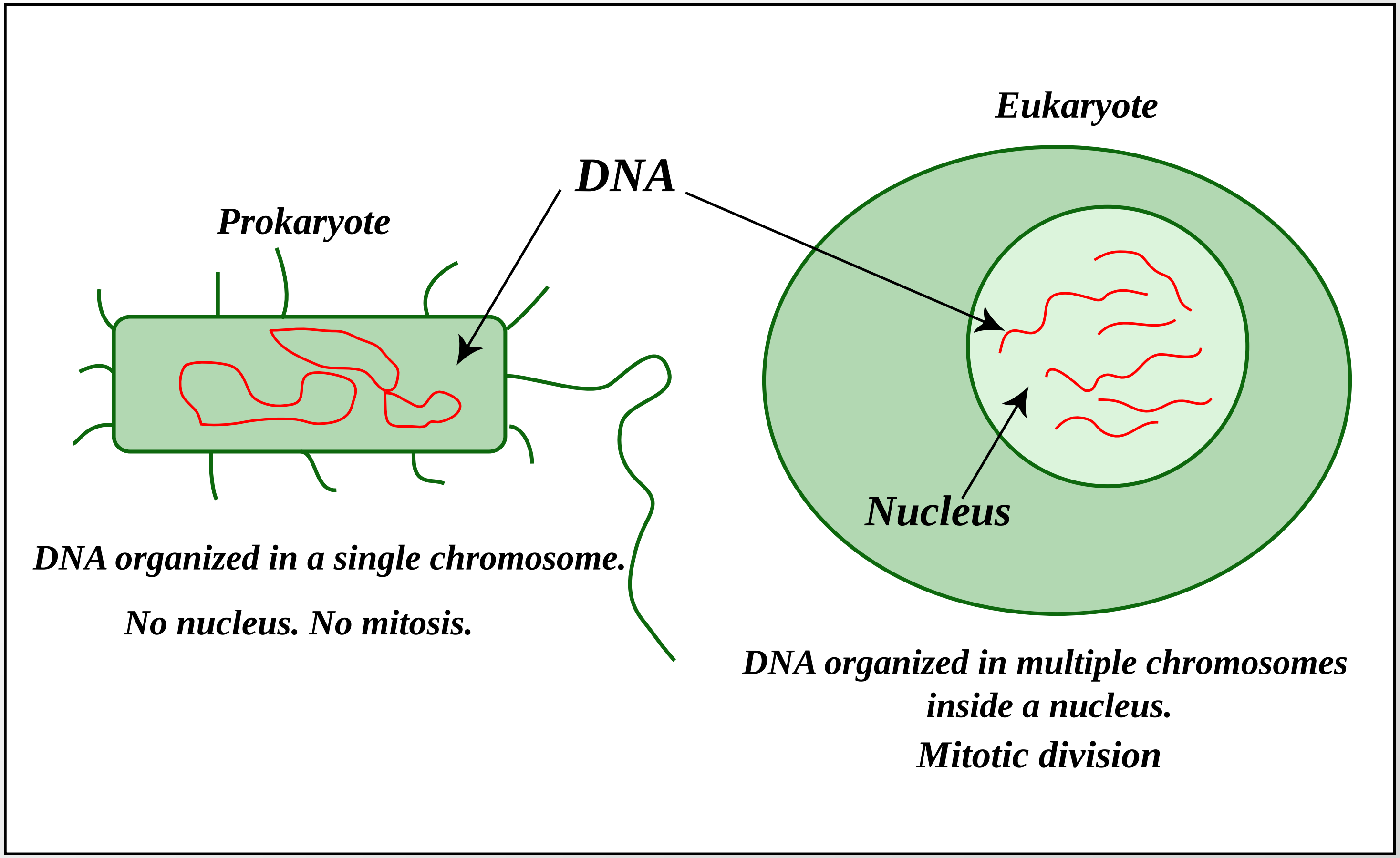
CTCT

DNA packaging

- Why packaging
  - DNA is very long
  - Cell is very small
- Compression
  - Chromosome is 50,000 times shorter than extended DNA
- Using the DNA
  - Before a piece of DNA is used for anything, this compact structure must open locally

Image removed due to copyright restrictions.
Please see: Figure 8-10 from Alberts, Bruce, and
Martin Raff. *Essential Cell Biology*. New York,
NY: Garland Publishing Inc., 1997. ISBN: 0815320450.

Chromosomes inside the cell



DNA organized in a single chromosome.

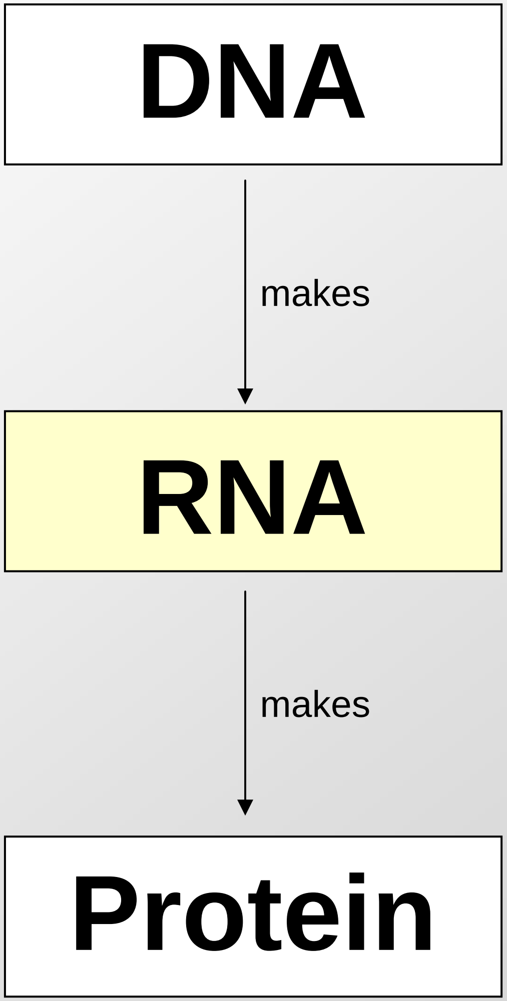
No nucleus. No mitosis.

DNA organized in multiple chromosomes inside a nucleus.

Mitotic division

Figures by MIT OpenCourseWare.

“Central dogma” of Molecular Biology



DNA

makes

RNA

makes

Protein

Genes control the making of cell parts

- The gene is a fundamental unit of inheritance
  - Each DNA molecule ⬄ 10,000+ genes
  - 1 gene ⬄ 1 functional element (one “part” of cell machinery)
  - Every time a “part” is made, the corresponding gene is:
    - Copied into mRNA, transported, used as blueprint to make protein
- RNA is a temporary copy
  - The medium for transporting genetic information from the DNA information repository to the protein-making machinery is an RNA molecule
  - The more parts are needed, the more copies are made
  - Each mRNA only lasts a limited time before degradation

mRNA: The messenger

- Information changes medium
  - single strand vs. double strand
  - ribose vs. deoxyribose sugar

A T T A C G G T A C C G T

|| || || || ||| ||| ||| || || ||| ||| ||| ||

U A A U G C C A U G G C A

  - Compatible base-pairing in hybrid

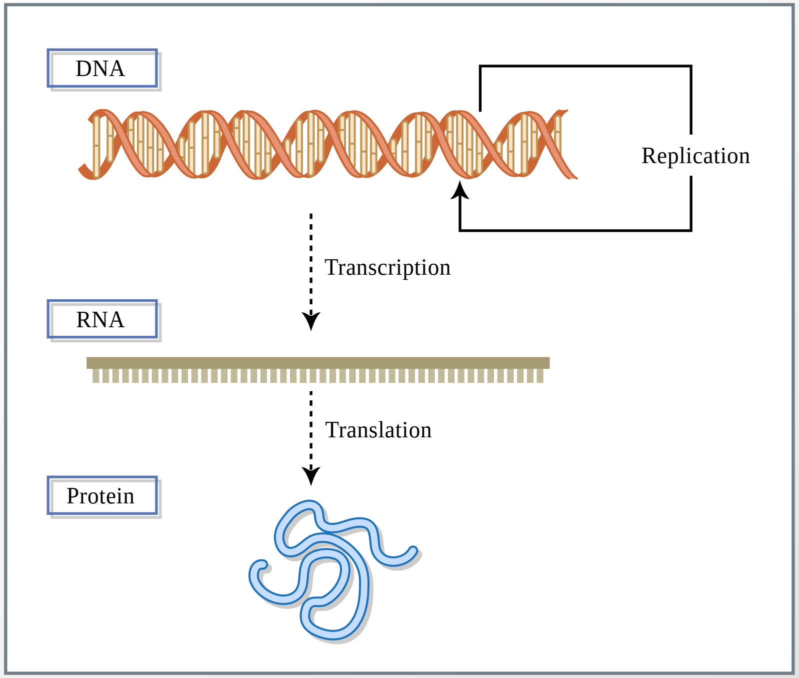
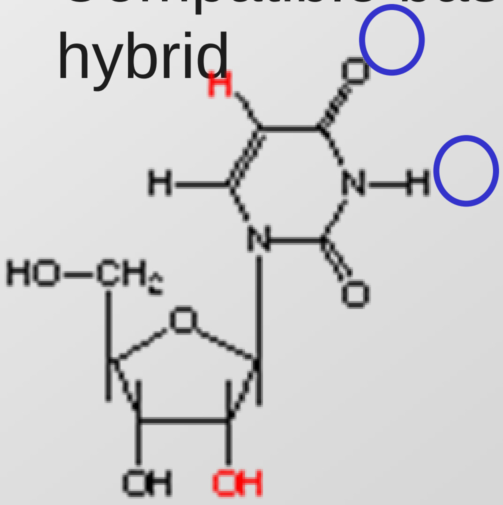
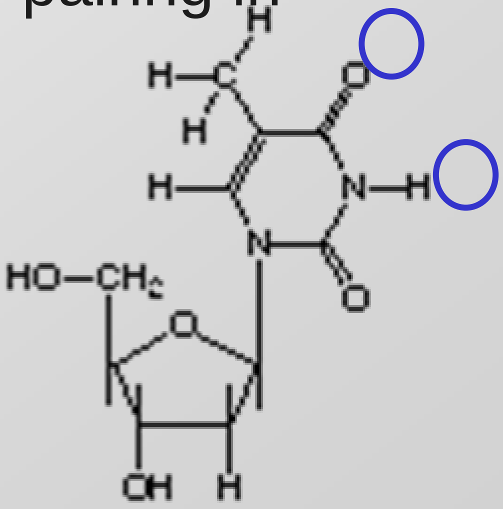


Figure by MIT OpenCourseWare.



uracil (RNA)



thymine (DNA)

From DNA to RNA: Transcription

Image removed due to copyright restrictions. Please see: Figure 7-9 from Alberts, Bruce, and Martin Raff. *Essential Cell Biology*. New York, NY: Garland Publishing Inc., 1997. ISBN: 0815320450.

From pre-mRNA to mRNA: Splicing

- In Eukaryotes, not every part of a gene is coding
  - Functional exons interrupted by non-translated introns
  - During pre-mRNA maturation, introns are spliced out
  - In humans, primary transcript can be 10^6 bp long

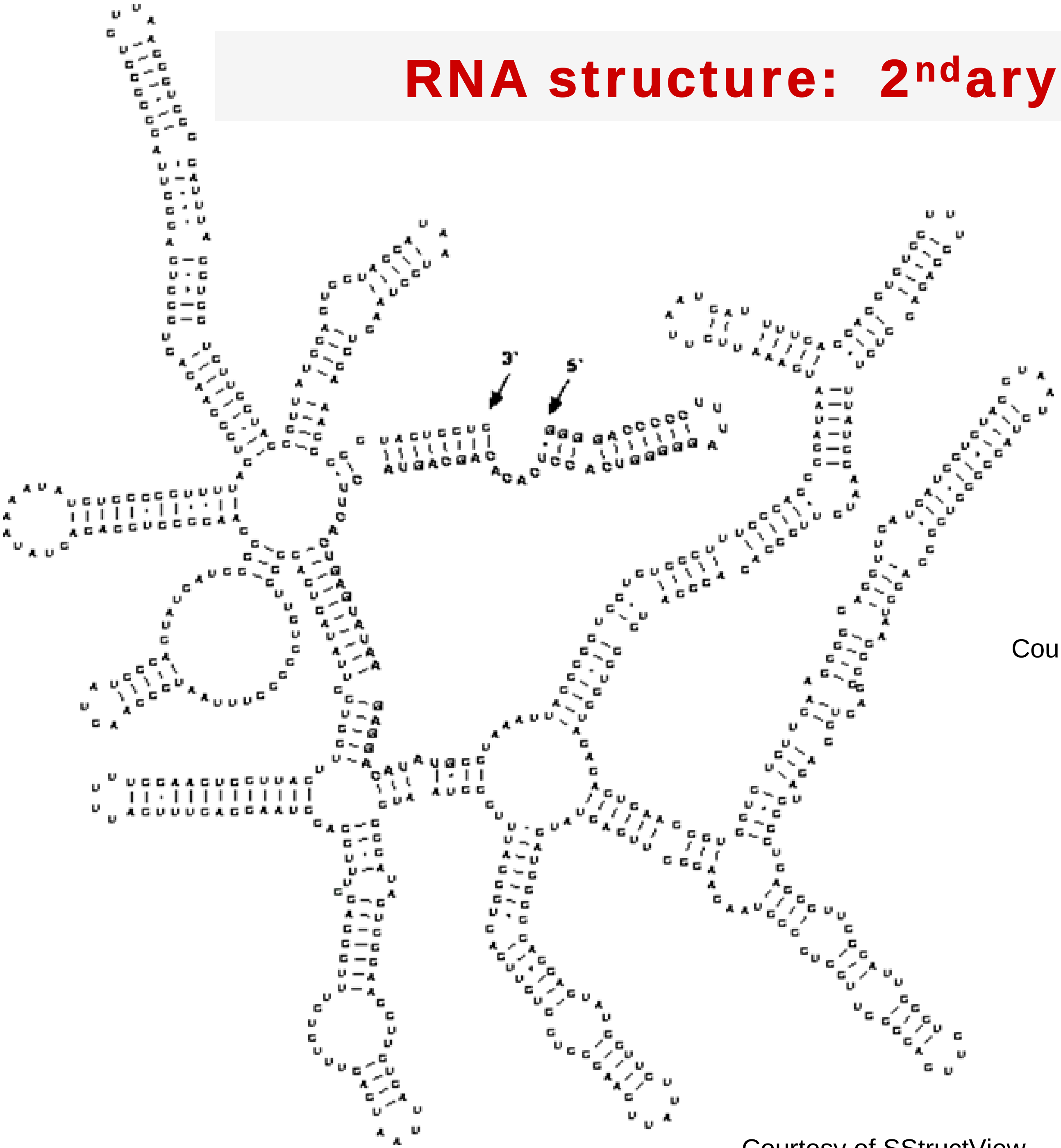
Image removed due to copyright restrictions. Please see: Figure 7-16 from Alberts, Bruce, and Martin Raff. *Essential Cell Biology*. New York, NY: Garland Publishing Inc., 1997. ISBN: 0815320450.

– Alternative splicing can yield different exon subsets for the same gene, and hence different protein products

RNA can be functional

- Single Strand allows complex structure
  - Self-complementary regions form helical stems
  - Three-dimensional structure allows functionality of RNA
- Four types of RNA
  - mRNA: messenger of genetic information
  - tRNA: codon-to-amino acid specificity
  - rRNA: core of the ribosome
  - snRNA: splicing reactions
- To be continued…
  - We'll learn more in a dedicated lecture on RNA world
  - Once upon a time, before DNA and protein, RNA did all

RNA structure: 2ndary and 3rdary



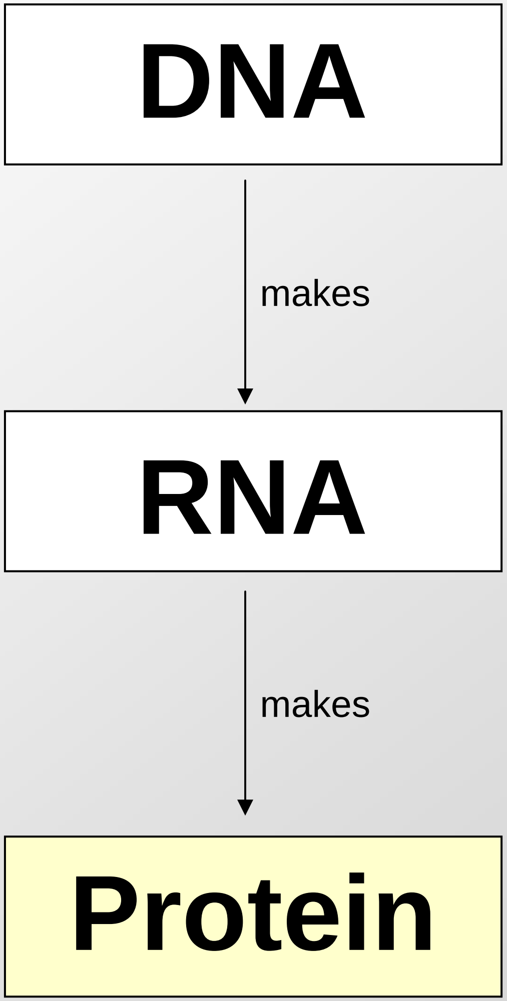
Courtesy of Wikimedia Commons.

Courtesy of SStructView.

Splicing machinery made of RNA

Image removed due to copyright restrictions. Please see: Figure 7-16 from Alberts, Bruce, and Martin Raff. *Essential Cell Biology*. New York, NY: Garland Publishing Inc., 1997. ISBN: 0815320450.

“Central dogma” of Molecular Biology



Proteins carry out the cell's chemistry

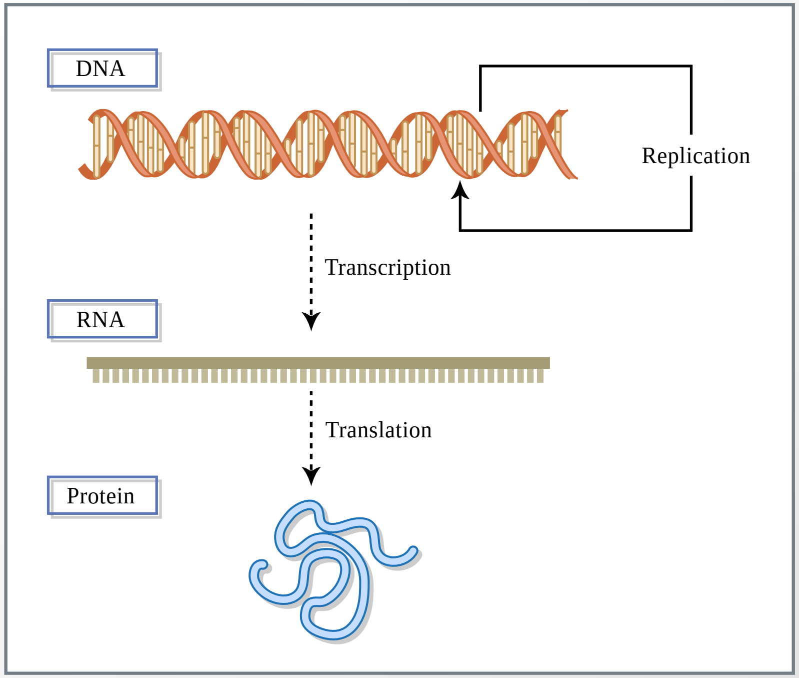


Figure by MIT OpenCourseWare.

- More complex polymer
  - Nucleic Acids have 4 building blocks
  - Proteins have 20. Greater versatility
  - Each amino acid has specific properties

Sequence → Structure → Function

- The amino acid sequence determines the three-dimensional fold of protein
- The protein's function largely depends on the features of the 3D structure

Proteins play diverse roles

- Catalysis, binding, cell structure, signaling, transport, metabolism

Protein structure

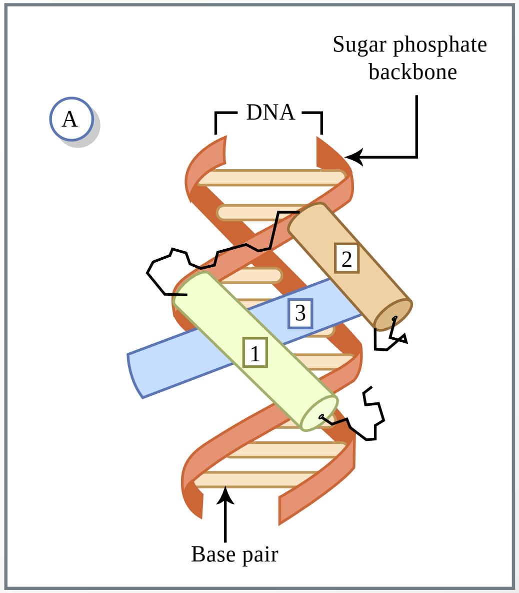


Figure by MIT OpenCourseWare.

Helix-turn-helix

Common motif for DNA-binding proteins that often play a regulatory role as mRNA level transcription factors

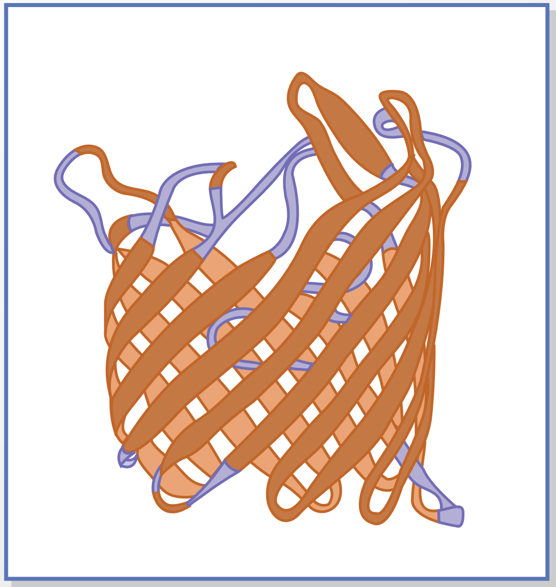


Figure by MIT OpenCourseWare.

Beta-barrel

Some antiparallel b-sheet domains are better described as b-barrels rather than b-sandwiches, for example streptavadin and porin. Note that some structures are

intermediate between the extreme barrel and sandwich arrangements.

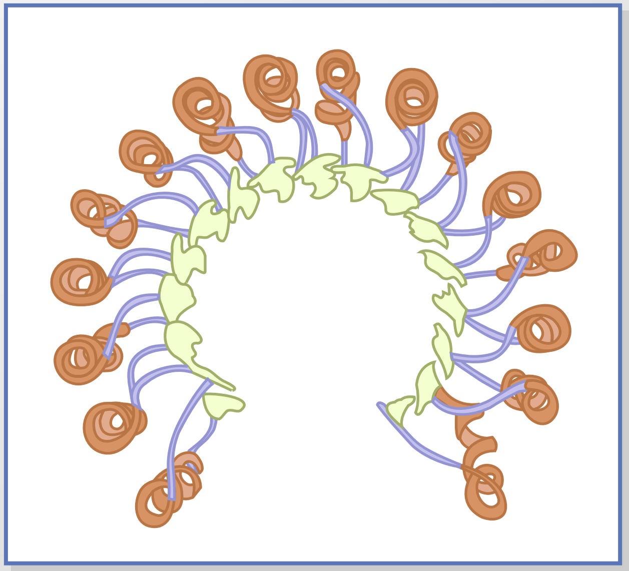


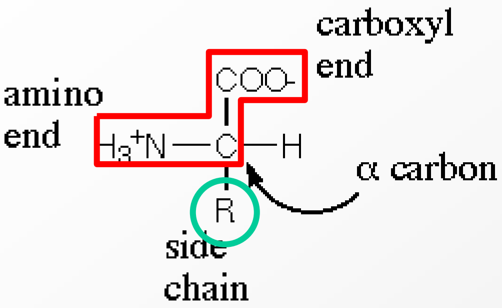
Figure by MIT OpenCourseWare.

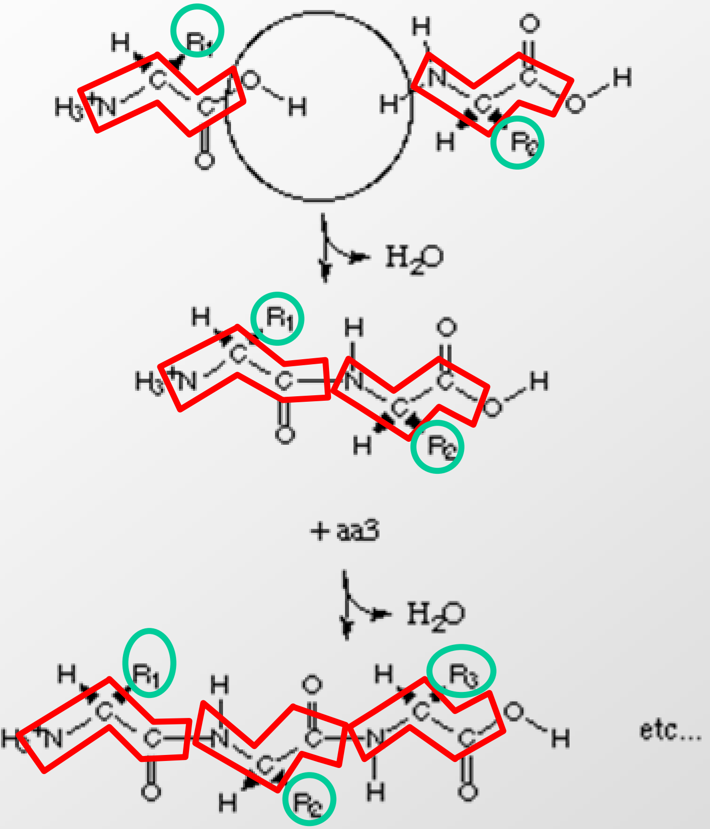
Alpha-beta horseshoe

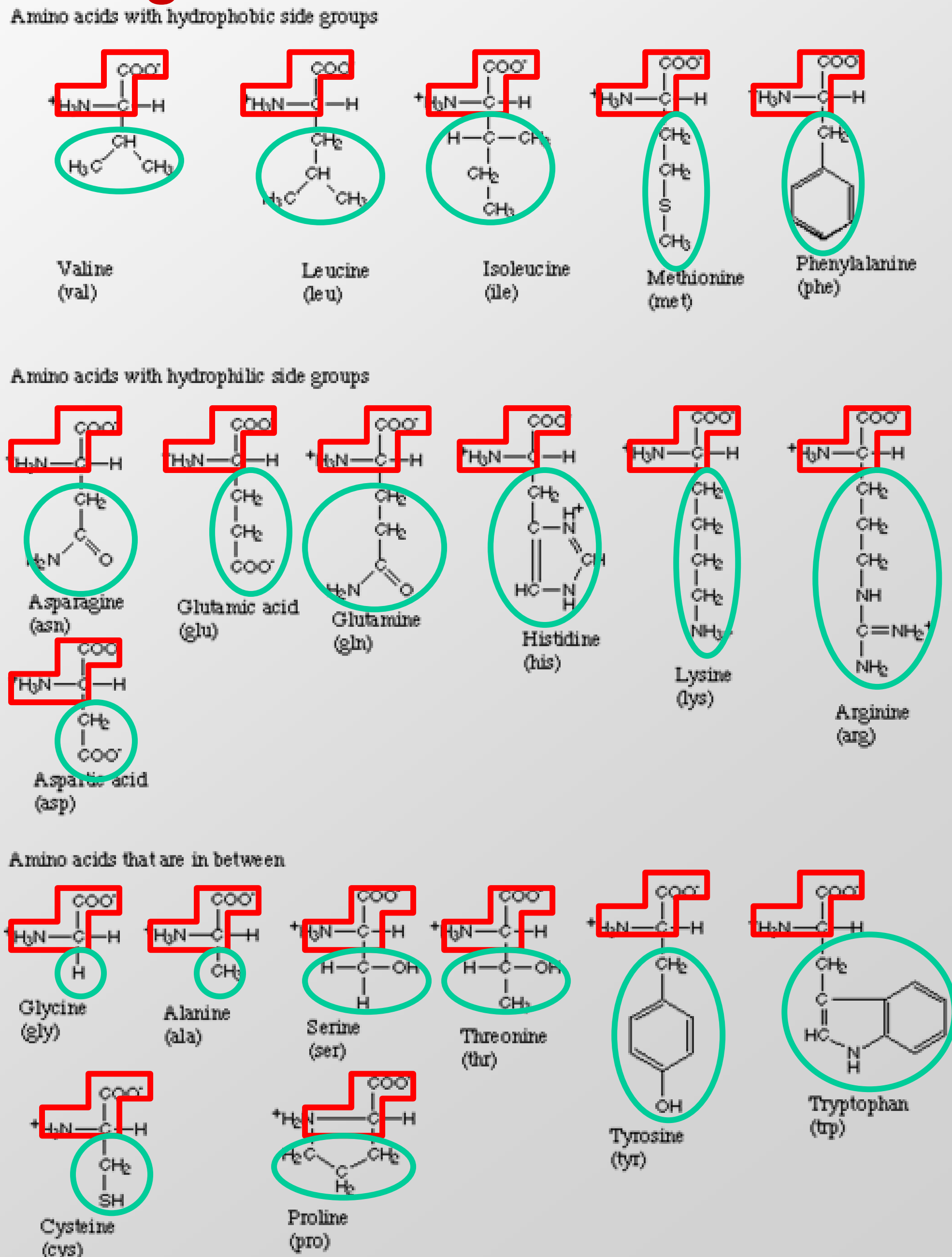
this placental ribonuclease inhibitor is a cytosolic protein that binds extremely strongly to any ribonuclease that may leak into the cytosol. 17-stranded parallel b sheet curved into an open horseshoe shape, with 16 a-helices packed against the outer surface. It doesn't form a barrel although it looks as though it should. The strands are only very slightly slanted, being nearly parallel to the central `axis'.

Protein building blocks

- Amino Acids







From RNA to protein: Translation

- Ribosome

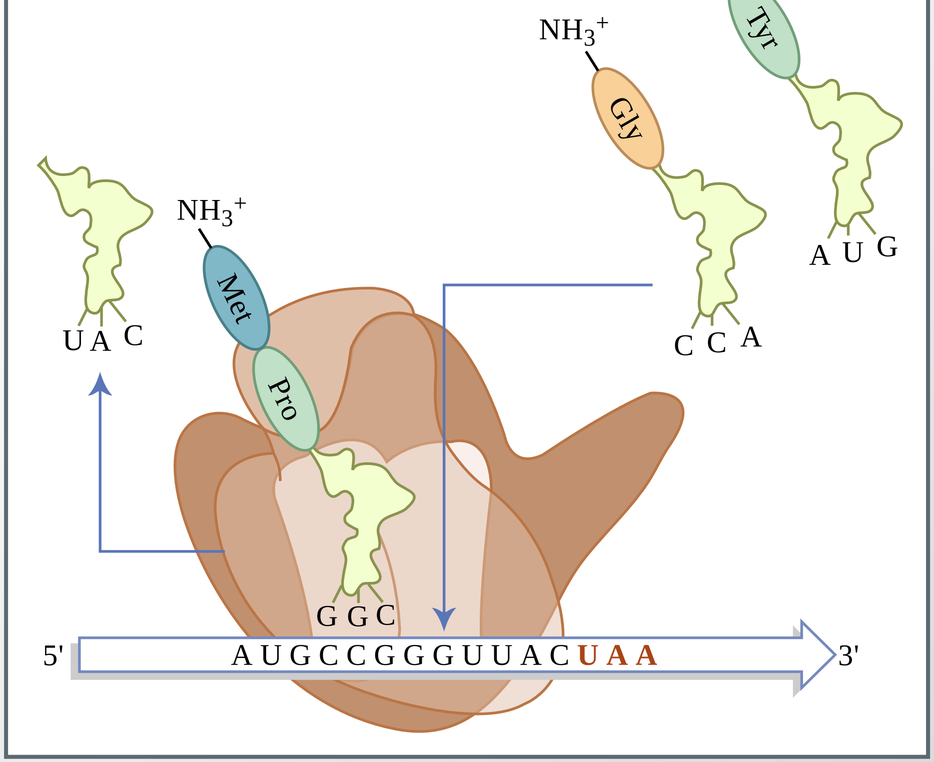
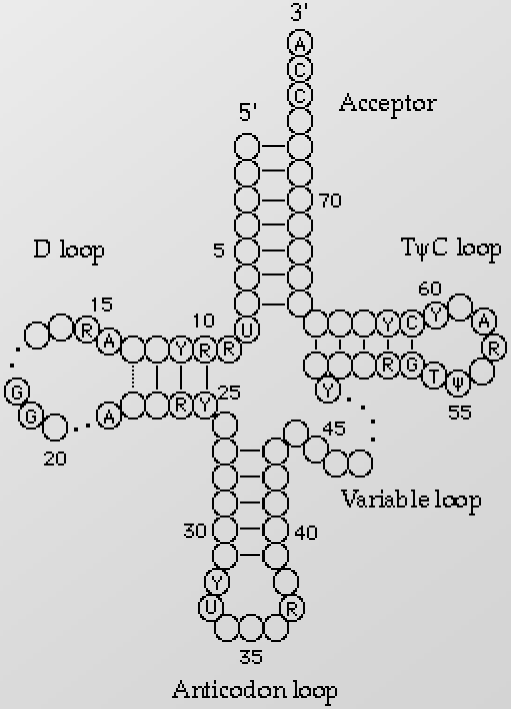
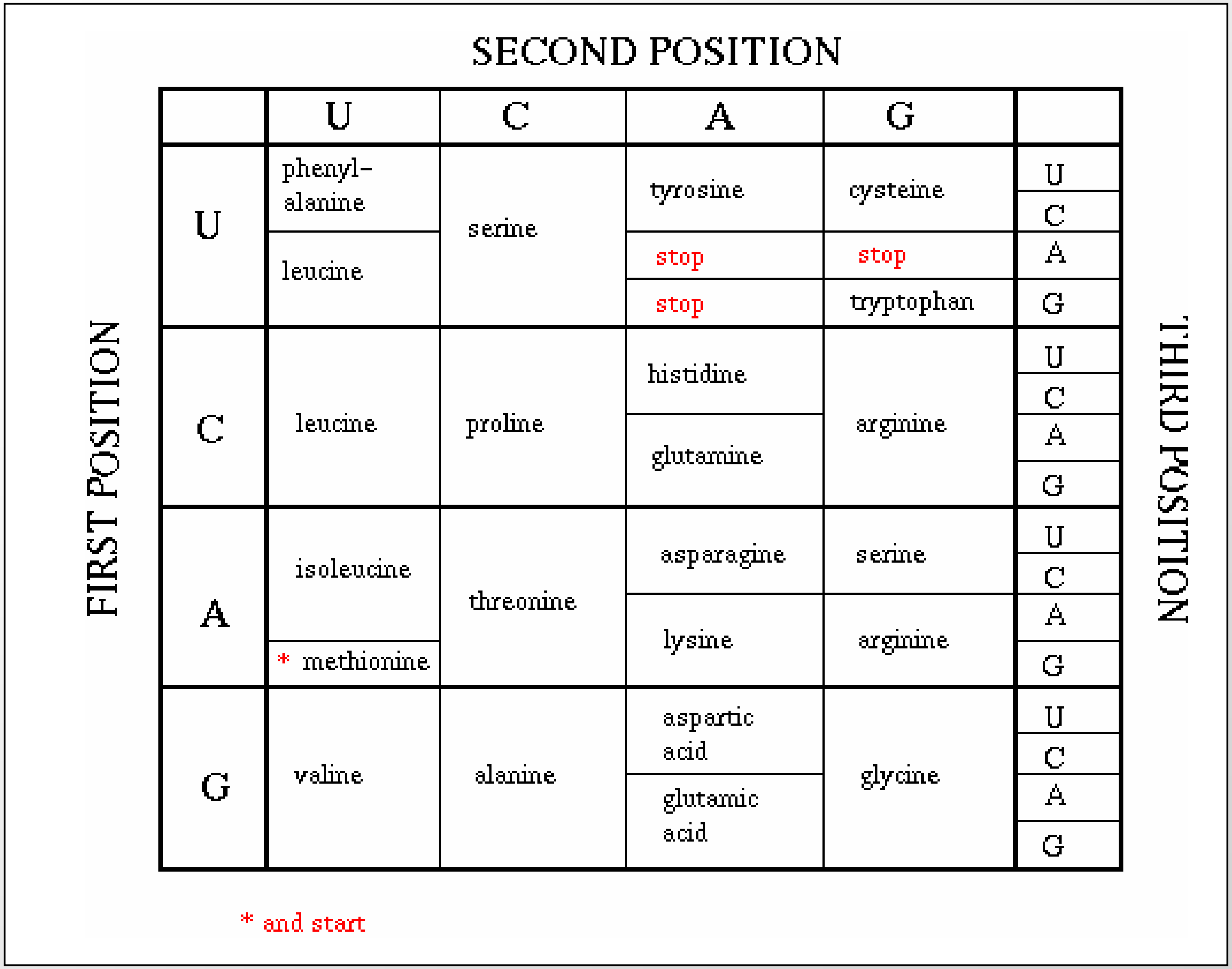


Figure by MIT OpenCourseWare.

- tRNA



The Genetic Code



SECOND POSITION

FIRST POSITION

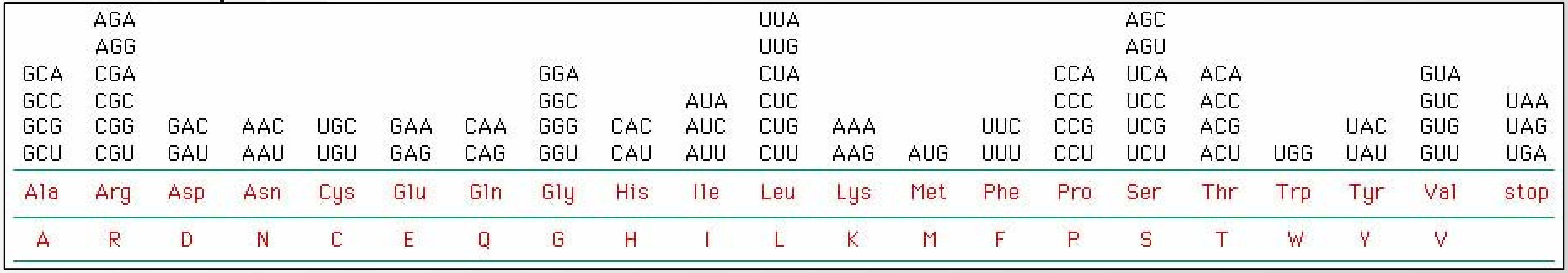
THIRD POSITION

	U	C	A	G	
U	phenylalanine	serine	tyrosine	cysteine	U
					C
	leucine		stop	stop	A
			stop	tryptophan	G
C	leucine	proline	histidine	arginine	U
					C
			glutamine		A
					G
A	isoleucine	threonine	asparagine	serine	U
					C
			lysine	arginine	A
	* methionine				G
G	valine	alanine	aspartic acid	glycine	U
					C
			glutamic acid		A
					G

* and start

The Genetic Code

- Degeneracy of the genetic code
  - To encode 20 amino acids, two nucleotides are not enough (4^2=16). Three nucleotides are too many (4^3=64)
  - The genetic code is degenerate. Same amino acid can be represented by more than one codon. Room for innovation
  - Moreover, amino acids with similar properties can be substituted for each other without changing the structure of the protein



GCA GCC GCG GCU	AGA AGG CGA CGC CGG CGU	GAC GAU	AAC AAU	UGC UGU	GAA GAG	CAA CAG	GGA GGC GGG GGU	CAC CAU	AUA AUC AUU	UUA UUG CUA CUC CUG CUU	AAA AAG	AUG	UUC UUU	CCA CCC CCG CCU	AGC AGU UCA UCC UCG UCU	ACA ACC ACG ACU	UGG	UAC UAU	GUA GUC GUG GUU	UAA UAG UGA
Ala	Arg	Asp	Asn	Cys	Glu	Gln	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val	stop
A	R	D	N	C	E	Q	G	H	I	L	K	M	F	P	S	T	W	Y	V	

- Six possible translation frames for every nucleotide stretch
  - GCU.UGU.UUA.CGA.AUU.A → Ala – Cys – Leu – Arg – Ile -
  - G.CUU.GUU.UAC.GAA.UUA → - Leu – Val – Tyr – Glu - Leu
  - Stop codon every 3/64. Long ORFs are unlikely, probably genes
  - In some viruses as many as four overlapping frames are functional

Summary: The Central Dogma

DNA makes RNA makes Protein

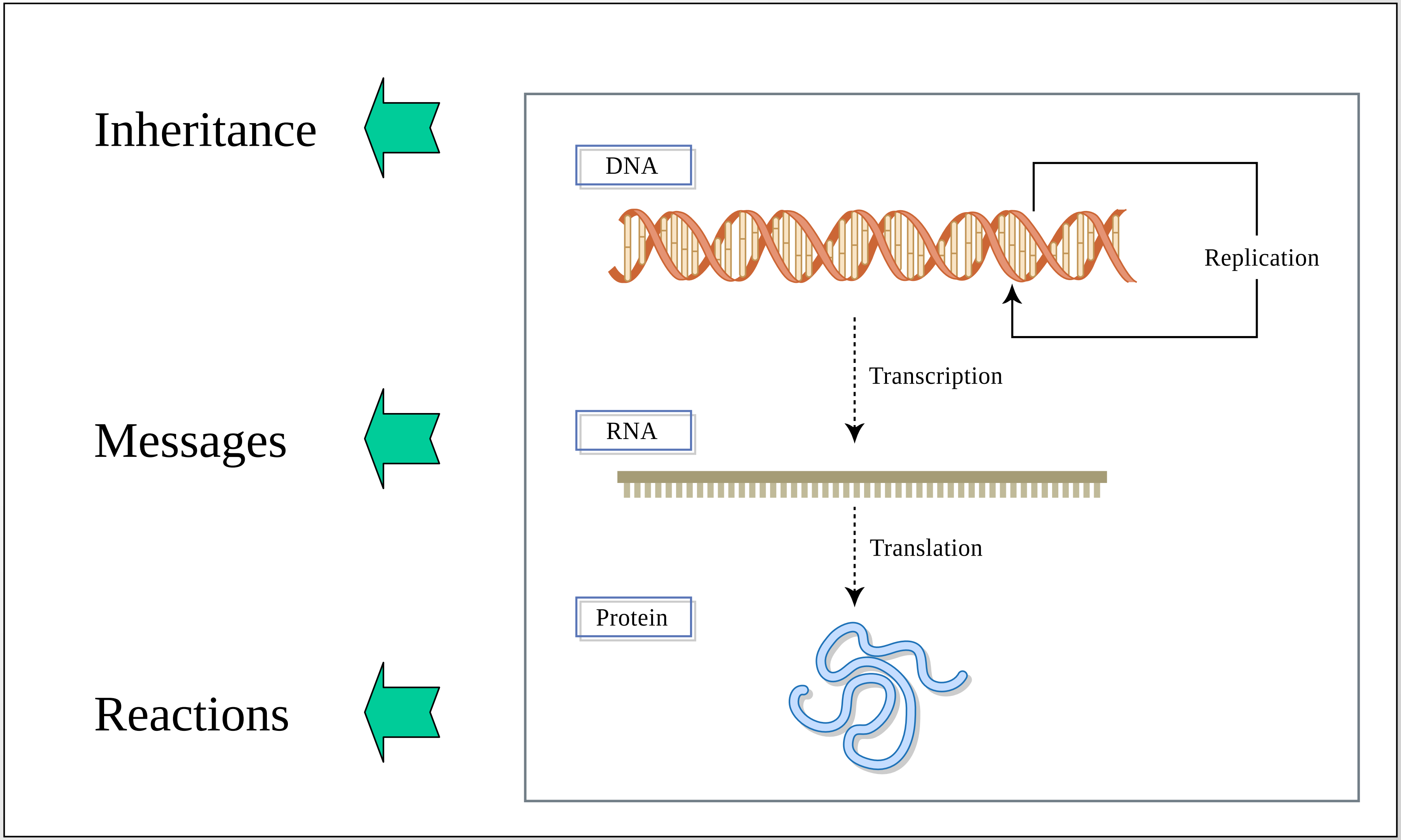
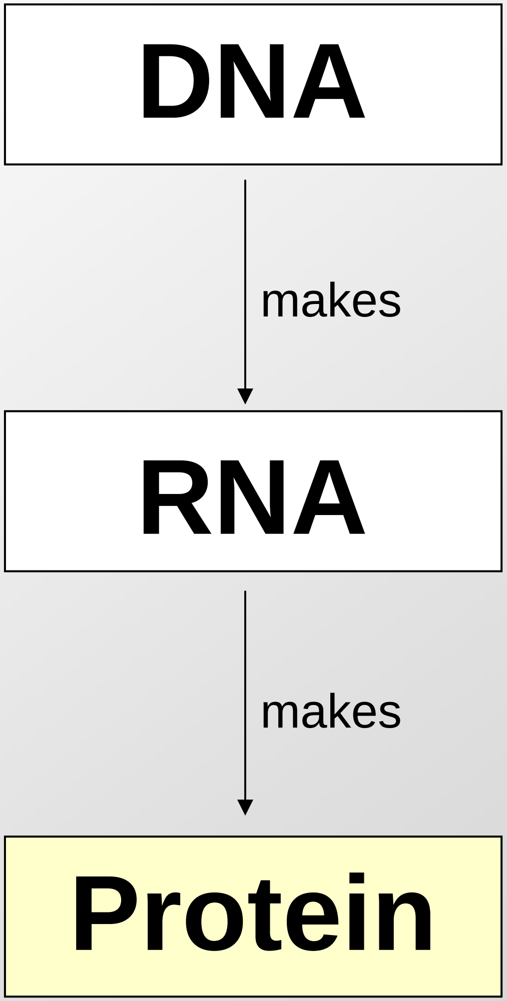


Figure by MIT OpenCourseWare.

Cellular dynamics and regulation

How cells move through this Central Dogma



DNA

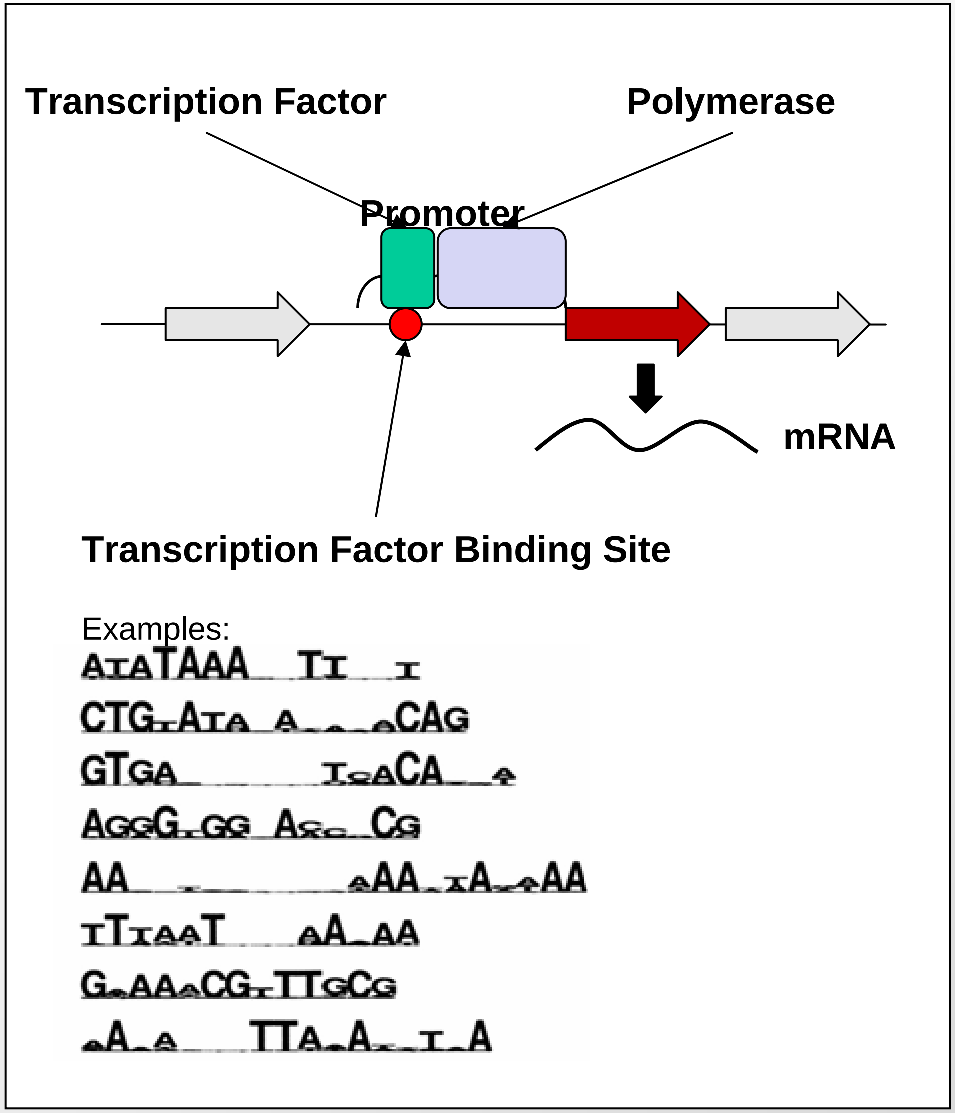
makes

RNA

makes

Protein

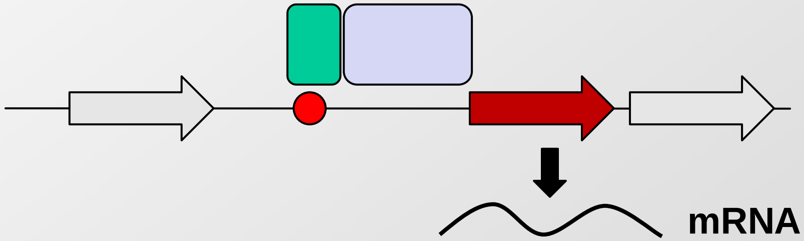
Regulation of Gene Expression



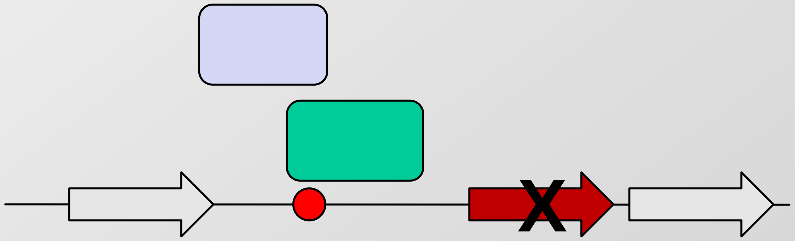
- Upstream of genes are *promoter* regions
- Contain promoter sequences or *motifs*
- *Transcription factors* (TFs) bind to motifs
- TFs recruit *RNA polymerase*
- Gene transcription

Regulatory Interactions

- Gene Activation

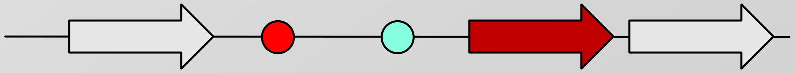


- Gene Repression



- Combinatorial Regulation

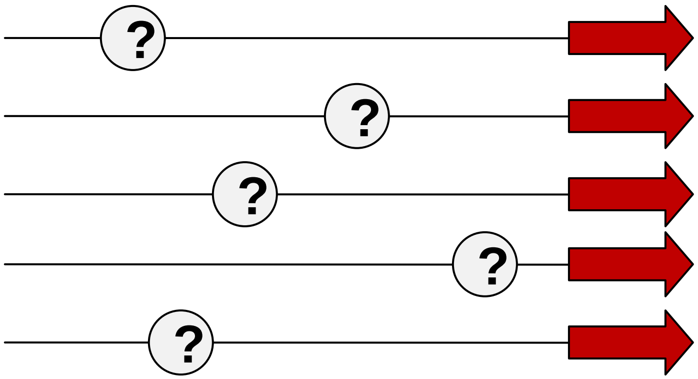
	○	



Computational Motif Prediction

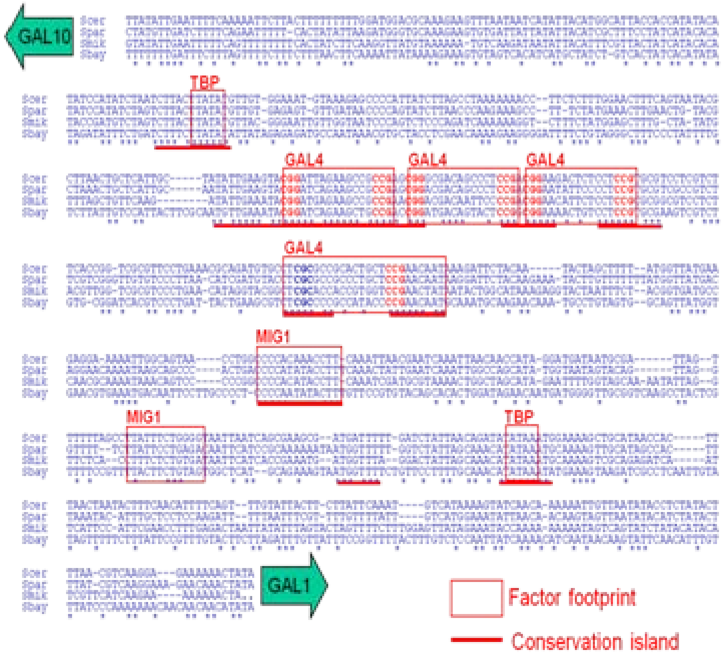
How do we find new transcription factor binding sites?

Gene regulated by same TF



Probabilistic model of promoters

Expectation maximization
Gibbs Sampling



Comparative sequence analysis

Evaluate motif conservation
across several related species

Regulatory Circuits

- Regulation depends on various intracellular and extracellular *signals*

Regulatory Circuits

- Regulation depends on various intracellular and extracellular *signals*

- Transcription factors regulate other factors that in turn regulate others – *regulatory network*

Computational Approaches

- Modeling regulatory networks
  - Bayesian Networks
- Inferring regulatory network models from experimental data
  - Microarray data
  - Guest lecture from Aviv Regev – computation inference of module networks
- Architectural properties of regulatory networks
  - Guest lecture from Uri Alon – modular structure of regulatory networks

Metabolism

- The totality of all chemical reactions in living matter

- Regulates the flow of *mass* and *energy* to perpetuate and replicate a state of low entropy

- **Catabolism**
  - Break down complex molecules to *release energy*

- **Anabolism**
  - Using energy to *assemble complex molecules*

Metabolic Pathways

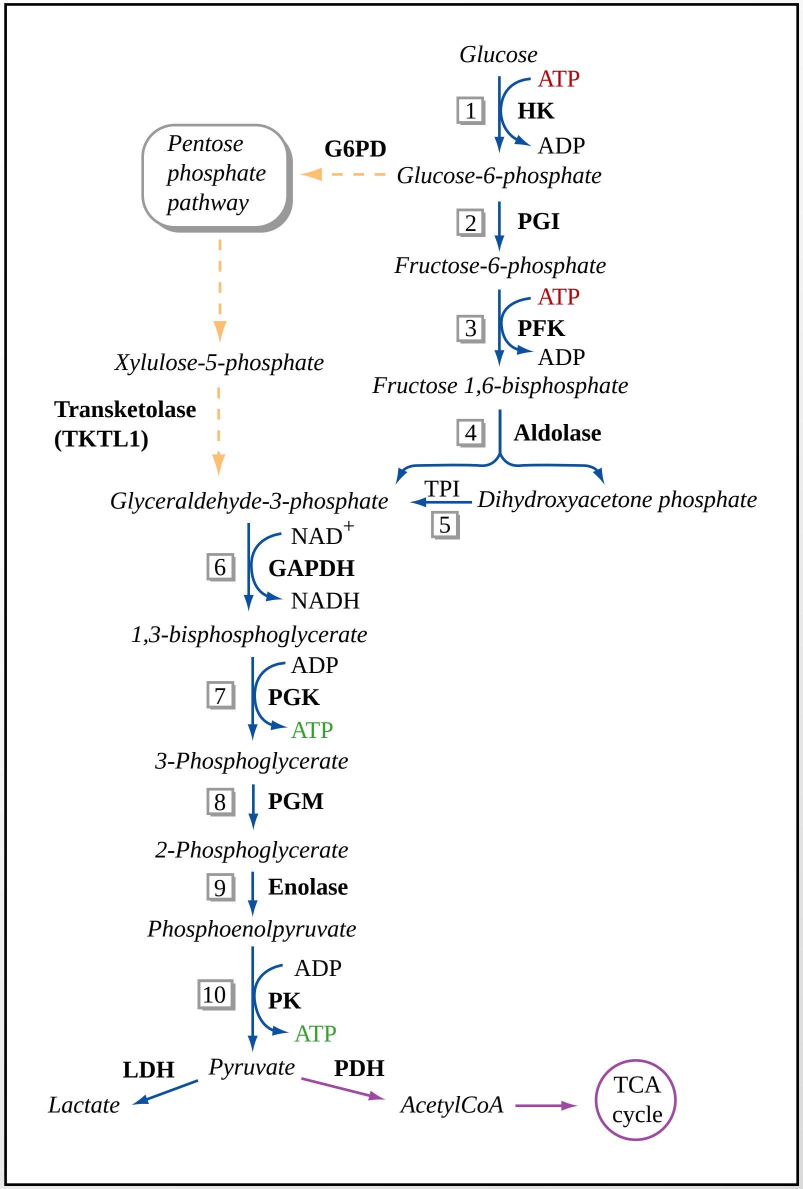
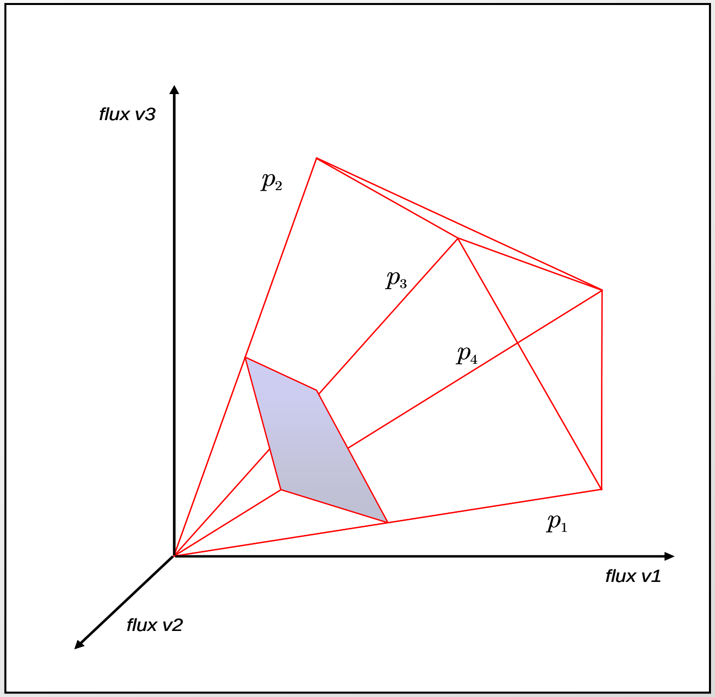


Figure by MIT OpenCourseWare.

In the living cell
reactions are organized into
Metabolic Pathways

1. Links **products** of one reaction to the **substrates** of another
2. Allows **energy** produced by reactions to be **captured** by others
3. **Regulation** of metabolism

Computational Metabolic Modeling

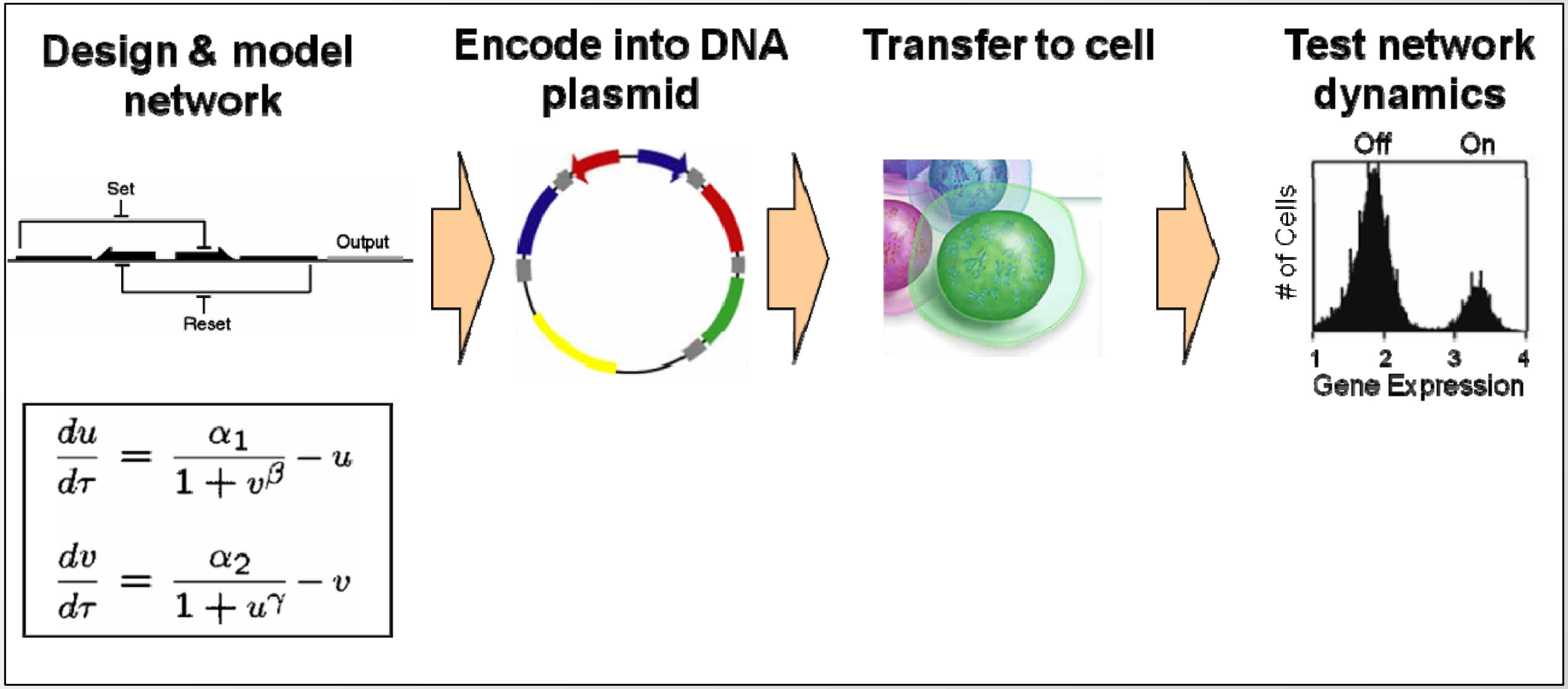


Flux Balance Analysis

- Predict steady-state metabolism
- Predict metabolic time- courses
- Predict mutant phenotypes
- Model gene regulation

Synthetic Biology

Synthetic Regulatory Networks



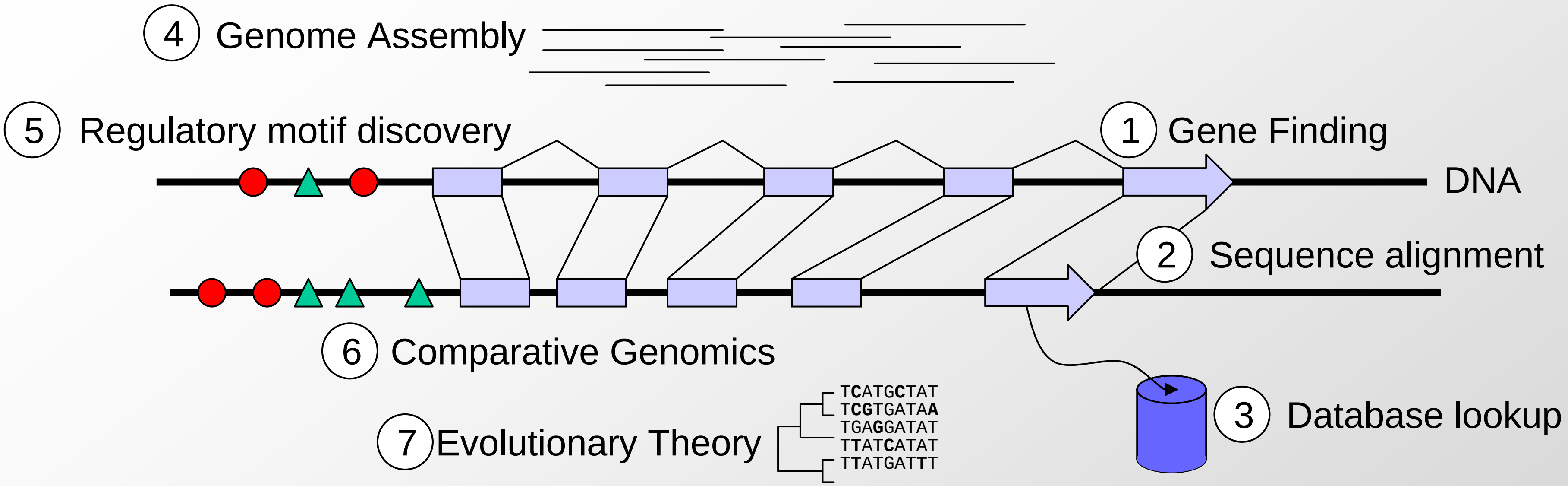
$$\frac{du}{d\tau} = \frac{\alpha_1}{1 + v^\beta} - u$$

$$\frac{dv}{d\tau} = \frac{\alpha_2}{1 + u^\gamma} - v$$

Courtesy of Jim Collins. Used with permission.

Jim Collins, BU

Challenges in Computational Biology



4 Genome Assembly

5 Regulatory motif discovery

1 Gene Finding

DNA

2 Sequence alignment

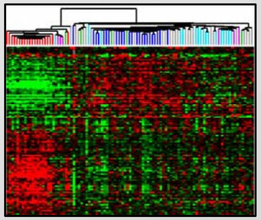
6 Comparative Genomics

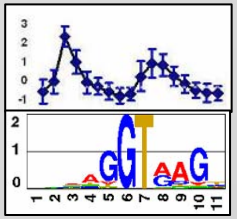
7 Evolutionary Theory

3 Database lookup

8 Gene expression analysis

RNA transcript

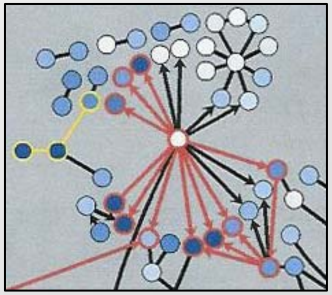




9 Cluster discovery

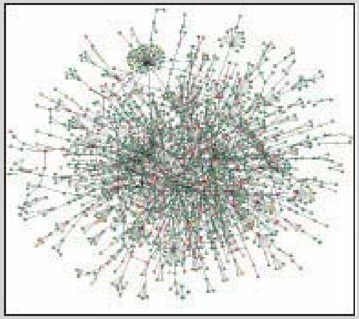
10 Gibbs sampling

11 Protein network analysis



12 Metabolic modelling

13 Emerging network properties



Recitation tomorrow! Room / time TBA

- Intro to python
  - We'll use it for our problem sets, already in PS1
- Introduction to algorithms / running time
  - Searching a genome for all motif occurrences
  - Pattern-based/sample-based enumeration
  - Table lookup for speeding up search
- Introduction to probability / statistics
  - Likelihood ratios and hypothesis testing
- Molecular biology Q&A
  - Central dogma, splicing, genomes
  - Other questions

Today:

Regulatory 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

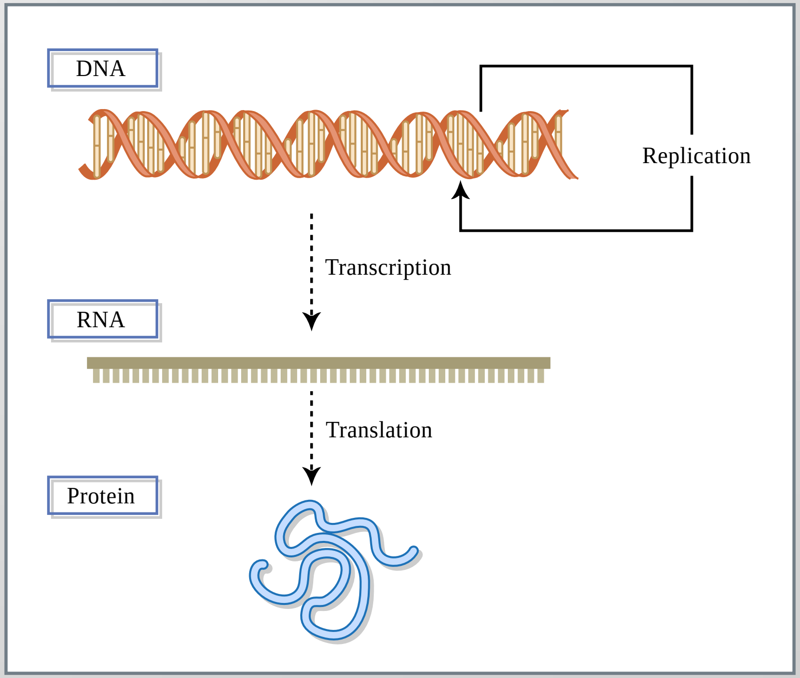
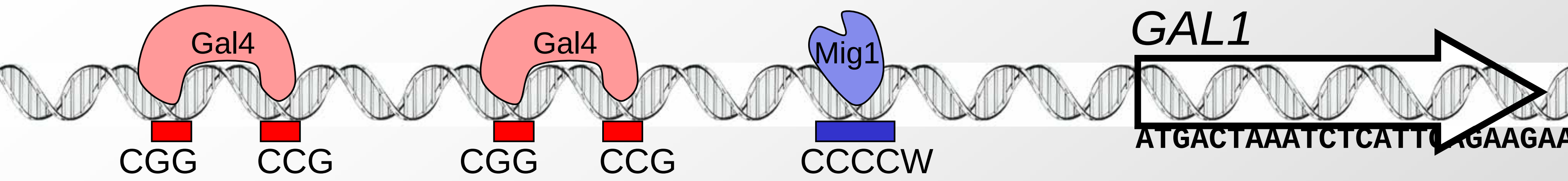


Figure by MIT OpenCourseWare.

Regulatory motif discovery

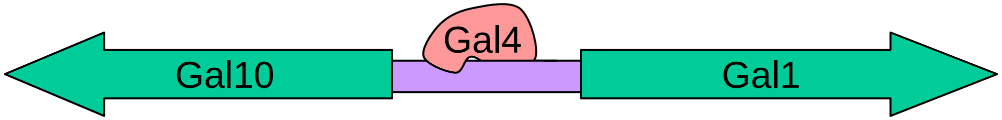


- Regulatory motifs (summary)
  - Genes are turned on / off in response to changing environments
  - No direct addressing: subroutines (genes) contain sequence tags (motifs)
  - Specialized proteins (transcription factors) recognize these tags

- 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

- How can we discover them?

Motifs are preferentially conserved across evolution



```
Scer   TTATATTGAATTTTCAAAAATTCTTACTTTTTTTTTGGATGGACGCAAAGAAGTTTAATAATCATATTACATGGCATTACCACCATATACA
Spar   CTATGTTGATCTTTTCAGAATTTTT-CACTATATTAAGATGGGTGCAAAGAAGTGTGATTATTATATTACATCGCTTTCCTATCATACACA
Smik   GTATATTGAATTTTTCAGTTTTTTTTCACTATCTTCAAGGTTATGTAAAAAA-TGTCAAGATAATATTACATTTCGTTACTATCATACACA
Sbay   TTTTTTTGATTTCTTTAGTTTTCTTTCTTTAACTTCAAAATTATAAAAGAAAGTGTAGTCACATCATGCTATCT-GTCACTATCACATATA
        * * ****  * *  *   ** ** *  *   **           **  ** * *    *    **   **    *  * * ** * * *

                          TBP
Scer   TATCCATATCTAATCTTACTTATATGTTGT-GGAAAT-GTAAAGAGCCCCATTATCTTAGCCTAAAAAAACC--TTCTCTTTGGAACTTTCAGTAATACG
Spar   TATCCATATCTAGTCTTACTTATATGTTGT-GAGAGT-GTTGATAACCCCAGTATCTTAACCCAAGAAAGCC--TT-TCTATGAAACTTGAACTG-TACG
Smik   TACCGATGTCTAGTCTTACTTATATGTTAC-GGGAATTGTTGGTAATCCCAGTCTCCCAGATCAAAAAAGGT--CTTTCTATGGAGCTTTG-CTA-TATG
Sbay   TAGATATTTCTGATCTTTCTTATATATTATAGAGAGATGCCAATAAACGTGCTACCTCGAACAAAAGAAGGGGATTTTCTGTAGGGCTTTCCCTATTTTG
       **   ** ***  **** ******* **      *            *    *          **  **      * *** *    ***    *
                                            *    *         *                                        *  *
                                    *   GAL4               GAL4              GAL4
Scer   CTTAACTGCTCATTGC-----TATATTGAAGTACGGATTAGAAGCCGCCGAGCGGGCGACAGCCCTCCGACGGAAGACTCTCCTCCGTGCGTCCTCGTCT
Spar   CTAAACTGCTCATTGC-----AATATTGAAGTACGGATCAGAAGCCGCCGAGCGGACGACAGCCCTCCGACGGAATATTCCCCTCCGTGCGTCGCCGTCT
Smik   TTTAGCTGTTCAAG--------ATATTGAAATACGGATGAGAAGCCGCCGAACGGACGACAATTCCCCGACGGAACATTCTCCTCCGCGCGGCGTCCTCT
Sbay   TCTTATTGTCCATTACTTCGCAATGTTGAAATACGGATCAGAAGCTGCCGACCGGATGACAGTACTCCGGCGGAAAACTGTCCTCCGTGCGAAGTCGTCT
             **  **          ** ***** ******* ****** ***** ***  ****   * *** ***** * *  ****** ***    * ***

                                        GAL4
Scer   TCACCGG-TCGCGTTCCTGAAACGCAGATGTGCCTCGCGCCGCACTGCTCCGAACAATAAAGATTCTACAA-----TACTAGCTTTT--ATGGTTATGAA
Spar   TCGTCGGGTTGTGTCCCTTAA-CATCGATGTACCTCGCGCCGCCCTGCTCCGAACAATAAGGATTCTACAAGAAA-TACTTGTTTTTTTATGGTTATGAC
Smik   ACGTTGG-TCGCGTCCCTGAA-CATAGGTACGGCTCGCACCACCGTGGTCCGAACTATAATACTGGCATAAAGAGGTACTAATTTCT--ACGGTGATGCC
Sbay   GTG-CGGATCACGTCCCTGAT-TACTGAAGCGTCTCGCCCCGCCATACCCCGAACAATGCAAATGCAAGAACAAA-TGCCTGTAGTG--GCAGTTATGGT
            ** *   ** *** *             ***** ** *  *   ****** **     *   * **
                                                                                  * *
                               *                                                                 ** ***
                                    MIG1
Scer   GAGGA-AAAATTGGCAGTAA----CCTGGCCCCACAAACCTT-CAAATTAACGAATCAAATTAACAACCATA-GGATGATAATGCGA------TTAG--T
Spar   AGGAACAAAATAAGCAGCCC----ACTGACCCCATATACCTTTCAAACTATTGAATCAAATTGGCCAGCATA-TGGTAATAGTACAG------TTAG--G
Smik   CAACGCAAAATAAACAGTCC----CCCGGCCCCACATACCTT-CAAATCGATGCGTAAAACTGGCTAGCATA-GAATTTTGGTAGCAA-AATATTAG--G
Sbay   GAACGTGAAATGACAATTCCTTGCCCCT-CCCCAATATACTTTGTTCCGTGTACAGCACACTGGATAGAACAATGATGGGGTTGCGGTCAAGCCTACTCG
              ****    *         *   *****     ***              * * *    *  * *          *
                                                                                                   **
                MIG1                                                      TBP   *
Scer   TTTTTAGCCTTATTTCTGGGGTAATTAATCAGCGAAGCG--ATGATTTTT-GATCTATTAACAGATATATAAATGGAAAAGCTGCATAACCAC-----TT
Spar   GTTTT--TCTTATTCCTGAGACAATTCATCCGCAAAAAATAATGGTTTTT-GGTCTATTAGCAAACATATAAATGCAAAAGTTGCATAGCCAC-----TT
Smik   TTCTCA--CCTTTCTCTGTGATAATTCATCACCGAAATG--ATGGTTTA--GGACTATTAGCAAACATATAAATGCAAAAGTCGCAGAGATCA-----AT
Sbay   TTTTCCGTTTTACTTCTGTAGTGGCTCAT--GCAGAAAGTAATGGTTTTCTGTTCCTTTTGCAAACATATAAATATGAAAGTAAGATCGCCTCAATTGTA
        * *      *    ***       * **   *        *** ***   *  *  **  ** * ********   ****    *
                                         *

Scer   TAACTAATACTTTCAACATTTTCAGT--TTGTATTACTT-CTTATTCAAAT----GTCATAAAAGTATCAACA-AAAAATTGTTAATATACCTCTATACT
Spar   TAAATAC-ATTTGCTCCTCCAAGATT--TTTAATTTCGT-TTTGTTTTATT----GTCATGGAAATATTAACA-ACAAGTAGTTAATATACATCTATACT
Smik   TCATTCC-ATTCGAACCTTTGAGACTAATTATATTTAGTACTAGTTTTCTTTGGAGTTATAGAAATACCAAAA-AAAAATAGTCAGTATCTATACATACA
```

Increase power by testing conservation in many regions

Conservation island

Framing the problem computationally

- How do we find all instances of a motif in a genome?
  - Naïve algorithm: Search every position
- How do we count all instances of every 6-mer in a genome
  - Naïve algorithm: Scan the genome for each motif
  - Improvement: Scan genome once, filling a table
- How do we count all instances of every 50-mer in a genome
  - Table is no longer feasible, most entries empty
  - Use a hash table
- How do we search a new motif in a known genome
  - Pre-processing of the database
- How do we deal with motif degeneracy and ambiguities
  - Hash in multiple places, increase alphabet size, partial hashing

Computational approaches for motif discovery

- Method #1: Enumerate all motifs
  - Combinatorial search

- Method #2: Randomly sample the genome
  - Statistical approach

- Method #3: Enumerate motif seeds + refinement
  - Hill-climbing

- Method #4: Content-based addressing
  - Hashing