Lecture 1: Overviews

- Computation in biology
- Computational molecular biology
  - Probabilities
- Interpreting genomes
  - Genome biology
    - Sites
  - Genomicists' tasks
  - Computational tasks
- This course

Computation and Biology

- Computation is ‘junior author’

- Computation is ***technology***
  - Technology helps ***drive*** science
  - … but should not ***displace*** science
    - not an end in itself
    - ***novelty*** & ***aesthetics*** should not override ***utility***

- Computational analysis generates *hypotheses*
  - which must ultimately be tested by experiment.
  - *But* hypotheses should
    - have some reasonable chance of being correct, and
    - carry indication of reliability.
  - Some computational findings may not be testable in lab
    - Evolution is a much more sensitive experimentalist

Computational Molecular Biology

Molecular biology

poses questions, judges answers

Statistics

Probability models for biological processes

Computer science

Deterministic methods for computing:
Computers & languages
Data structures & algorithms

Biology involves *probabilities*, at several levels:

- Fundamental physical laws governing molecular systems

- Evolutionary processes

Probabilistic Physical Laws

- Structure & pairwise interactions of atoms & molecules:
  - quantum mechanics & quantum electrodynamics

- Systems of interacting molecules:
  - statistical mechanics & thermodynamics

"*The true logic of this world is in the calculus of probabilities*" – James Clerk Maxwell

"*I cannot believe that God plays dice with the cosmos*" – Albert Einstein

  - but two of his four great 1905 papers dealt with statistical aspects of nature (photoelectric effect & Brownian motion)!

Probabilistic Evolutionary Processes

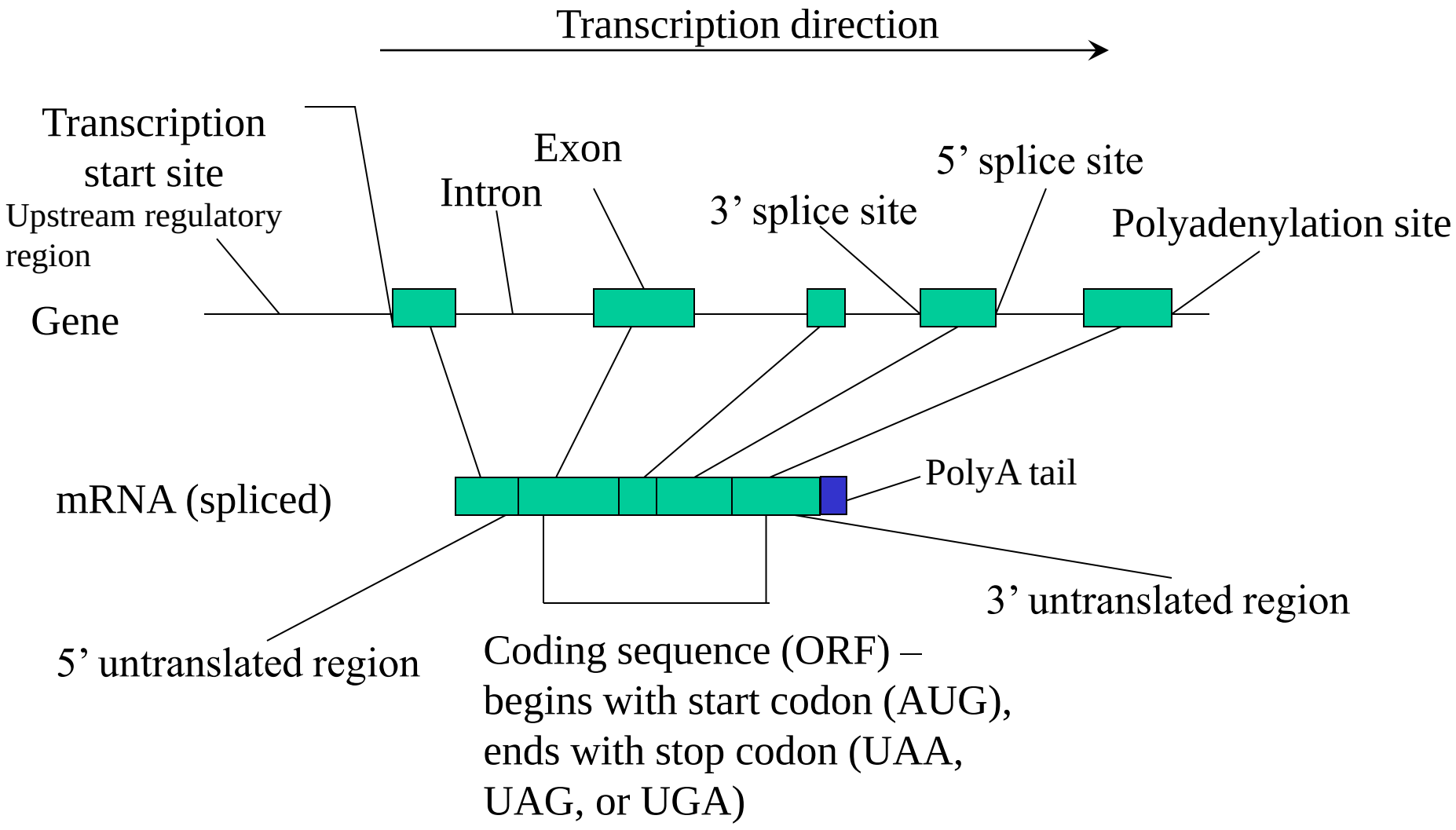
- Mutations (imperfect replication)
- Transmission of DNA from parent to offspring in populations of individuals
- Random aspects of environment

Probabilities have shaped the genome!

Genome biology overview

- Genomes undergo two fundamental processes (both involve copying!):
  - Replication
  - Transcription
- Genomic functional information is in the form of *sites:*
  - Short (~3 – ~15 base) sequence segments that bind to an *RNA* or *protein* molecule (the *reader*) to help mediate some function
- Sites may *act* (= be read) at the DNA or RNA transcript level

(Protein-coding) Gene Structure in Eukaryotes



Sites

- *Binding ≠ reading*
  - chance non-functional occurrences of site-like sequence may be transiently bound
    - inefficient, but evolutionarily significant!
- A site may be inactive in some cells
  - Reader may be absent, inactivated, or obstructed from binding (sites can overlap!)
- *Background* (= non-site) sequence carries information:
  - site spacing
  - mutations

Sites: genomic distribution

- Sites typically *recur:*
  - multiple sites within a genome, with possibly varying sequences, may be recognized by the same reader
    - Sequence variation may be represented by a *motif* or (better!) a *sequence logo*
- Sites typically *cluster* (into '*features*'):
  - several sites, with the same or different readers, acting collectively to carry out a function
    - site *ordering, orientation* and *spacing* may be important
  - *gene* = cluster of sites involved in ***expressing*** a particular transcript

- Average site density (= the fraction of the genome that is functional) may be quite small!
  - < 10% of human genome
    - remaining > 90% mostly transposon relics, ‘dead’ genes & processed pseudogenes
  - strength of selection for ‘genome efficiency’ is expected to depend on
    - Population size
    - Reproductive life span
    - Genome size

DNA sites

- Readers are usually *proteins*
- Help carry out or regulate a fundamental process
  - Replication
    - Replication origins, centromeres, telomeres (each having *multiple* sites)
  - Transcription
    - Promoters, enhancers, suppressors (each usually having *multiple* sites, with readers being ***transcription factors***)

From http://www-lmmb.ncifcrf.gov/~toms/sequencelogo.html

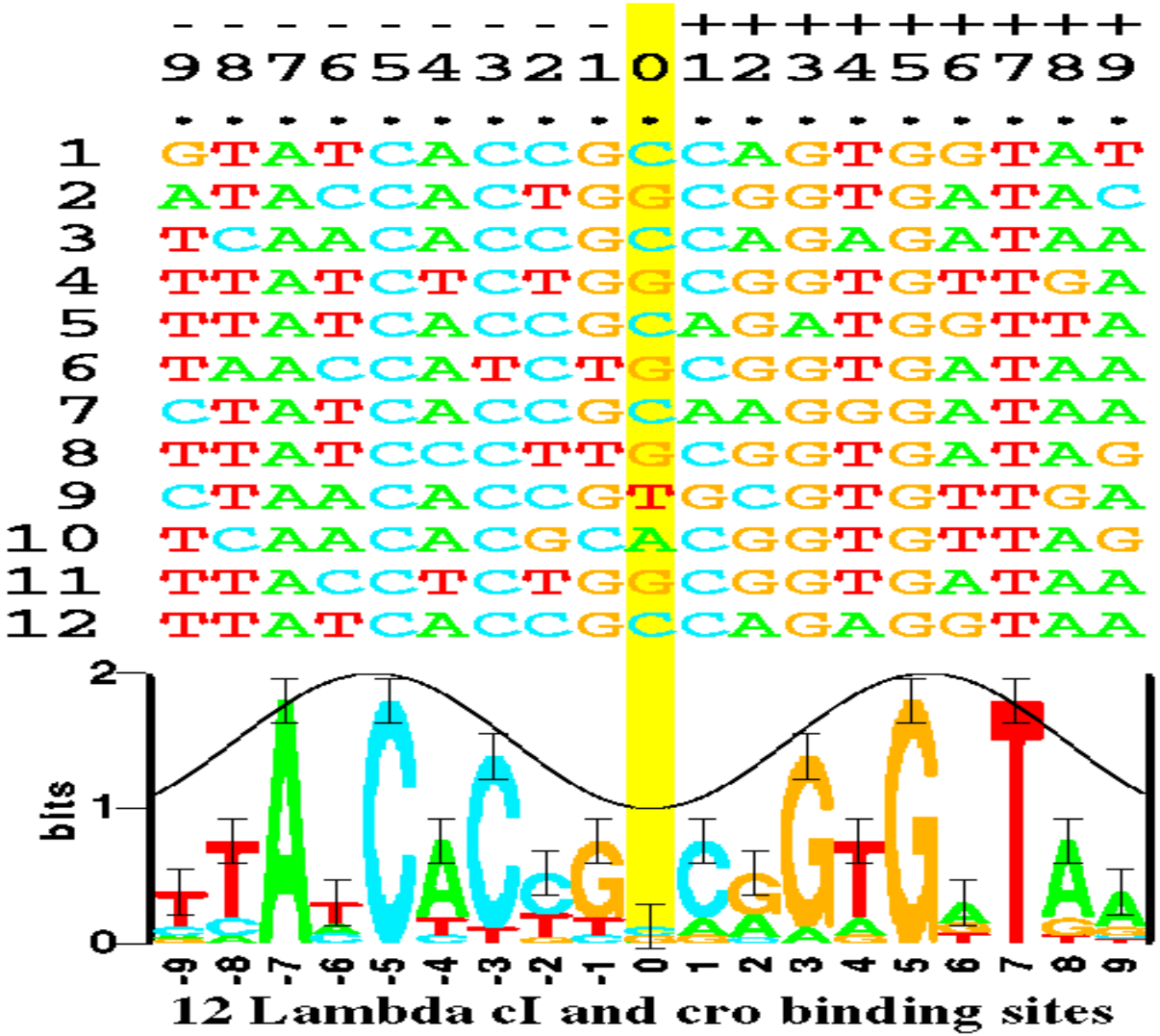
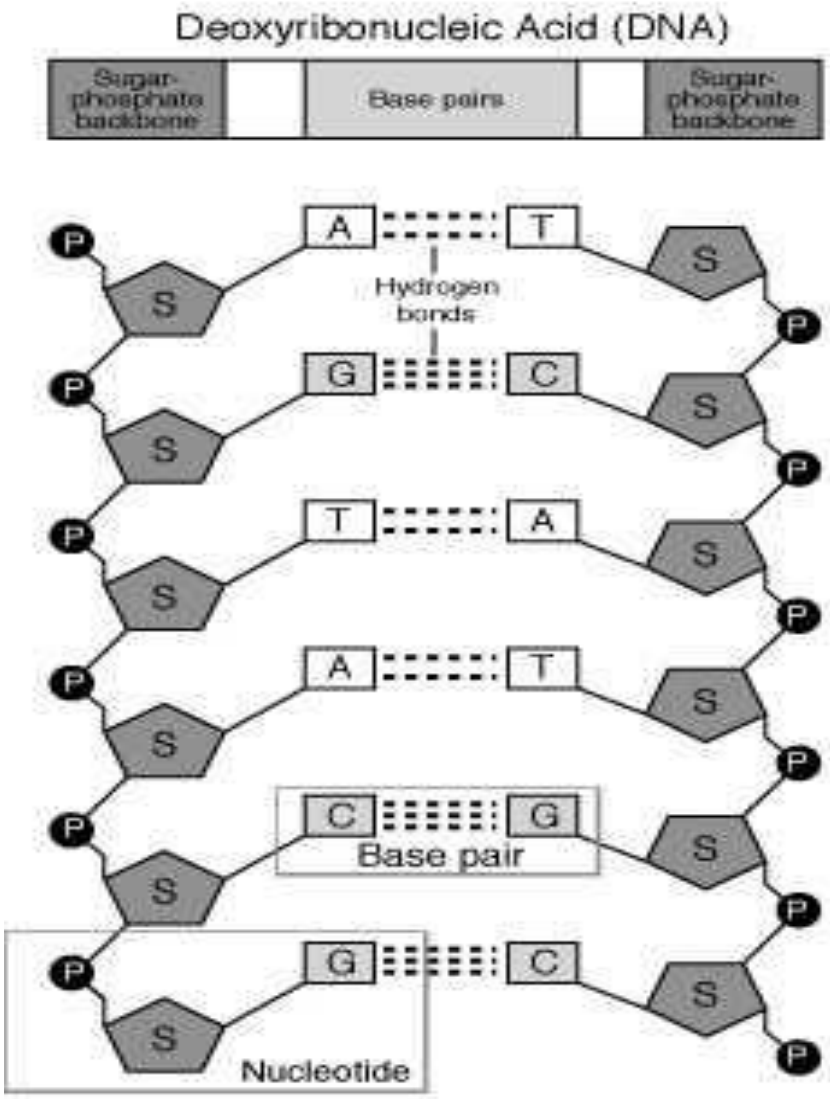
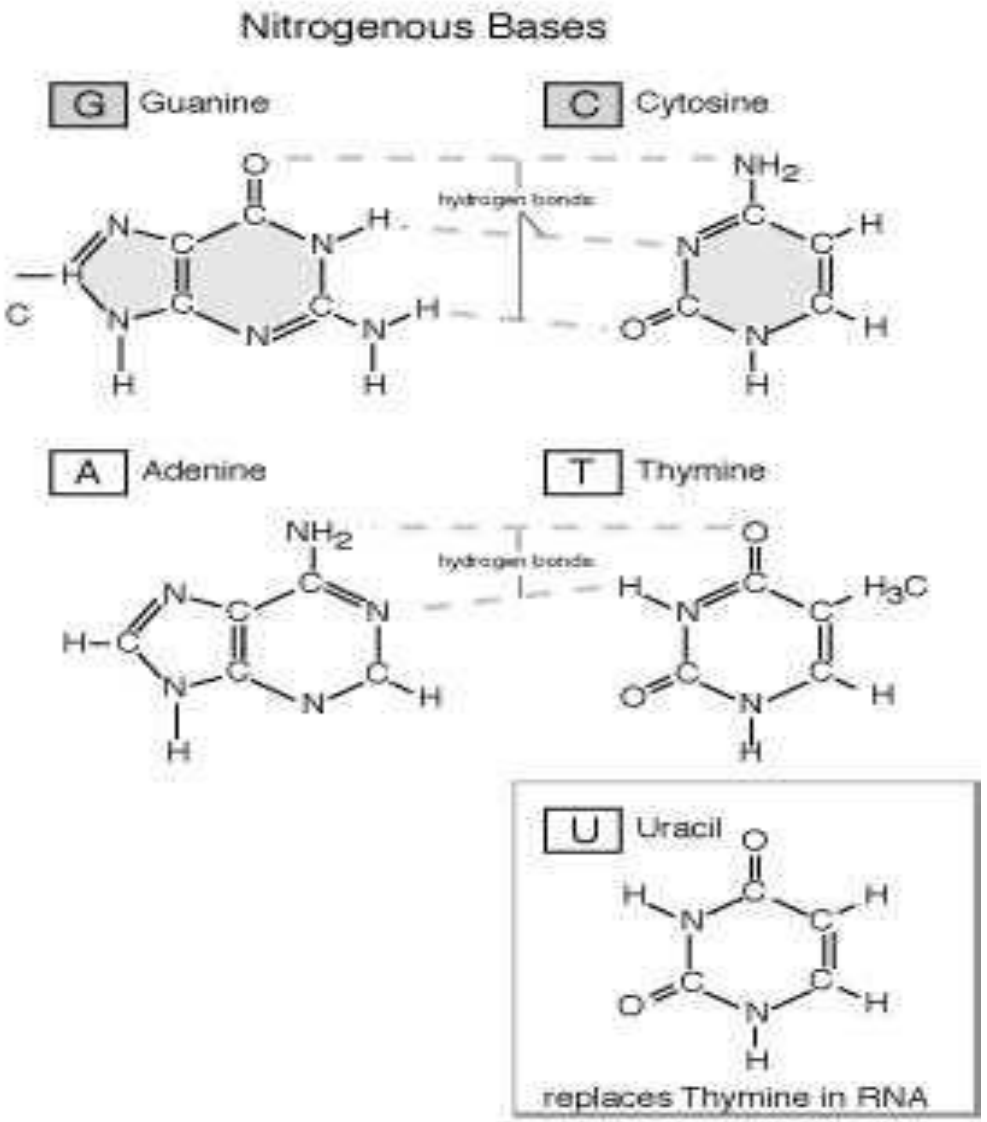
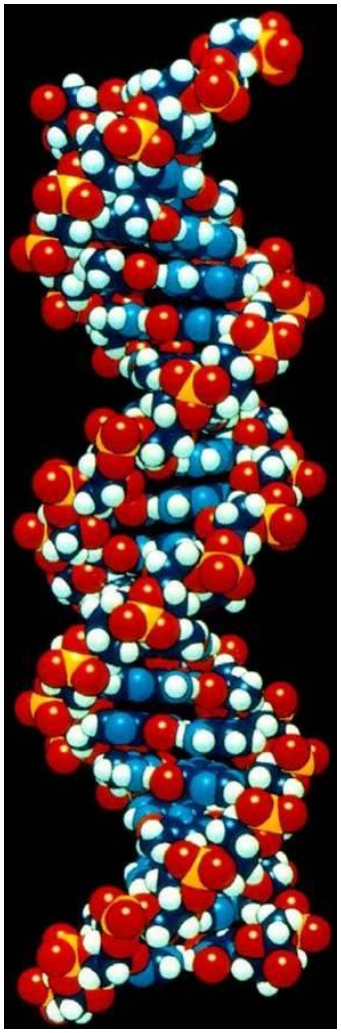


Fig. 1. Some aligned sequences and their sequence logo. At the top of the figure are listed the 12 DNA sequences from the P_L and P_R control regions in bacteriophage lambda. These are bound by both the cI and cro proteins [16]. Each even numbered sequence is the complement of the preceding odd numbered sequence. The sequence logo, described in detail in the text, is at the bottom of the figure. The cosine wave is positioned to indicate that a minor groove faces the center of each symmetrical protein. Data which support this assignment are given in reference [17].

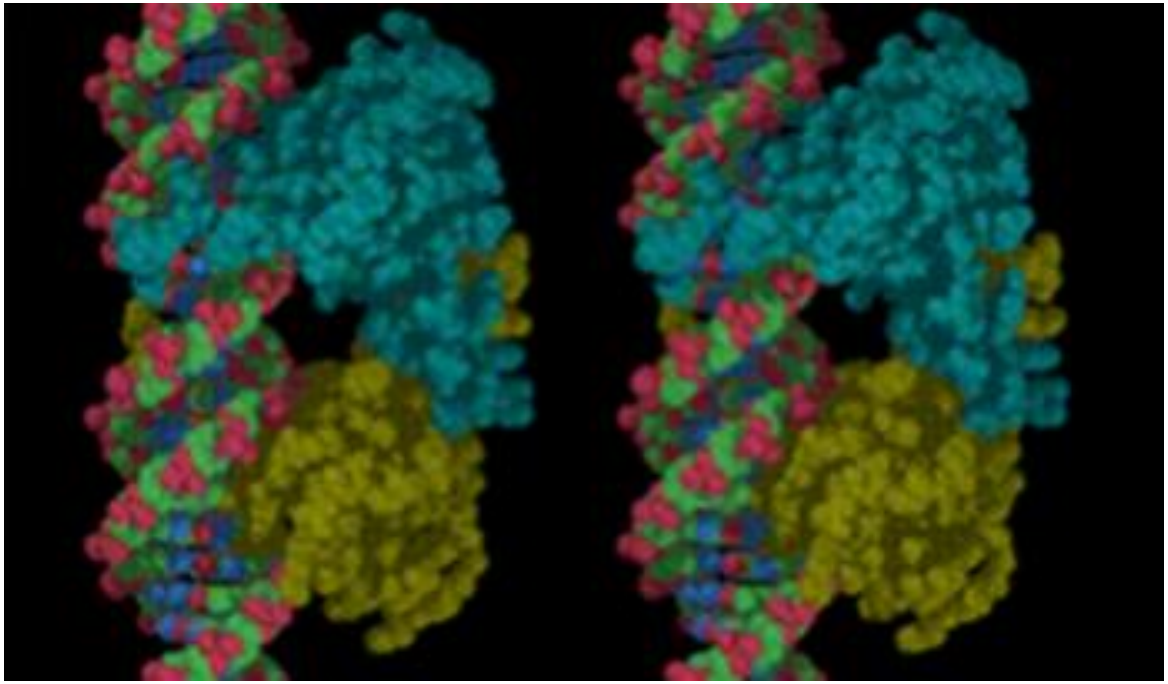
National Human Genome Research Institute (NHGRI) http://www.nhgri.nih.gov/DIR/VIP/ by artist Darryl Leja







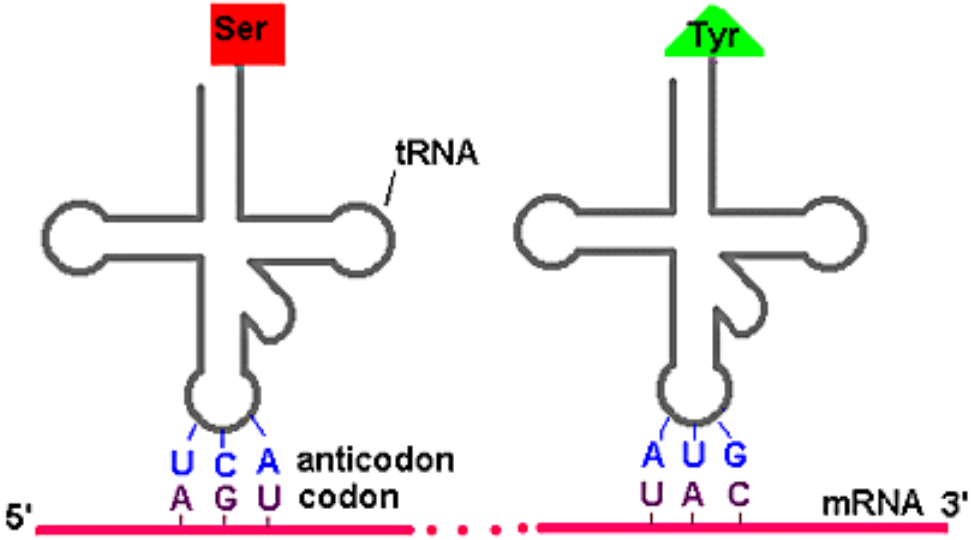
from http://www.dna-dna.net/



from http://gibk26.bse.kyutech.ac.jp

RNA transcript sites

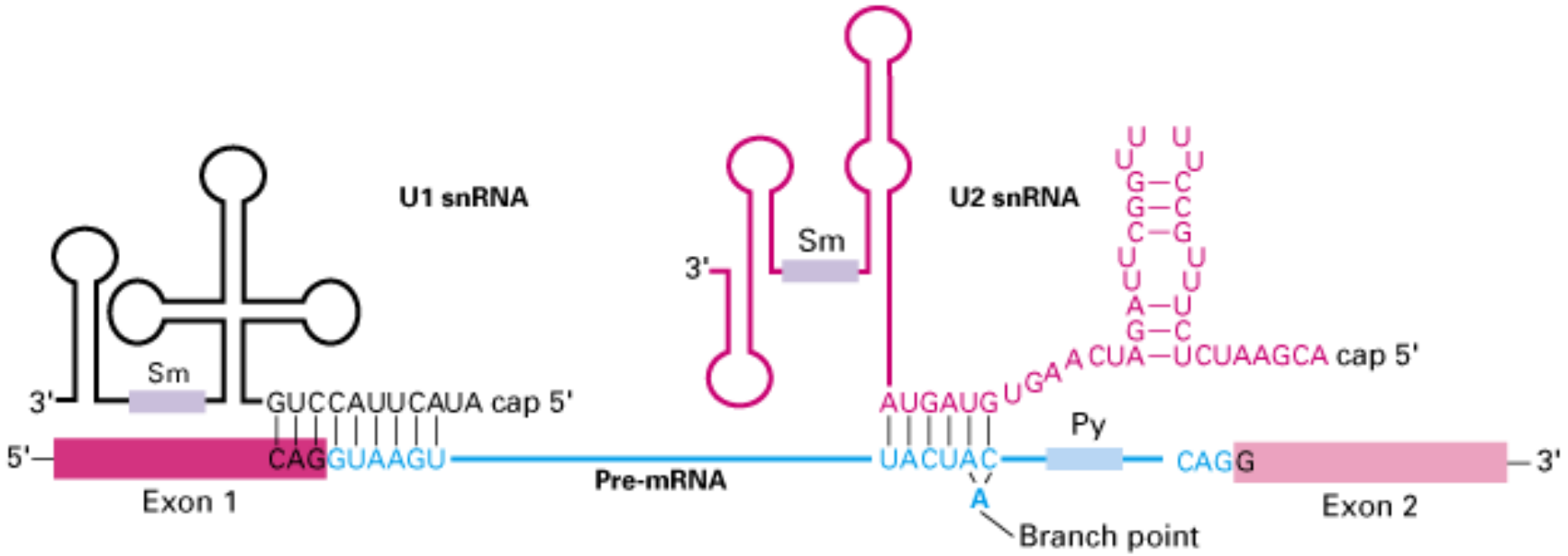
- Readers are *often* RNA
- Help carry out the transcript's function
  - in *protein coding* transcripts:
    - Translation start sites, codons (reader = charged tRNA), splice sites, microRNA binding sites, polyadenylation sites, …
  - in *functional RNA* transcripts:
    - Stem structures (the transcript reads itself!), …



2nd base in codon

1st base in codon	U	C	A	G	3rd base in codon
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	STOP	STOP	A
	Leu	Ser	STOP	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

The Genetic Code



from http://departments.oxy.edu/biology/Stillman/bi221/111300/processing_of_hnrnas.htm

(Jonathon Stillman, Grace Fisher-Adams)

Genomicists' tasks

- Find the ***genome sequence***
- Find the ***transcripts***
- Find the ***sites*** ...
- … and their ***functions*** …

Finding the genome sequence

- Get *reads* (short, overlapping, error-prone pieces of the sequence)
- *Assemble* : identify read overlaps, infer underlying sequence
- Main challenge:
  - (Near-)duplicate sequences

Finding transcripts ("RNASeq")

- Get *reads* from cDNA copies of the processed (spliced + edited) transcripts
- *Align* to genome sequence
- *Assemble* to infer transcript sequence
- Main challenges:
  - Expression bandwidth
  - Transcripts may be processed in more than one way (isoforms)
  - A transcript may be non-functional!

Finding sites

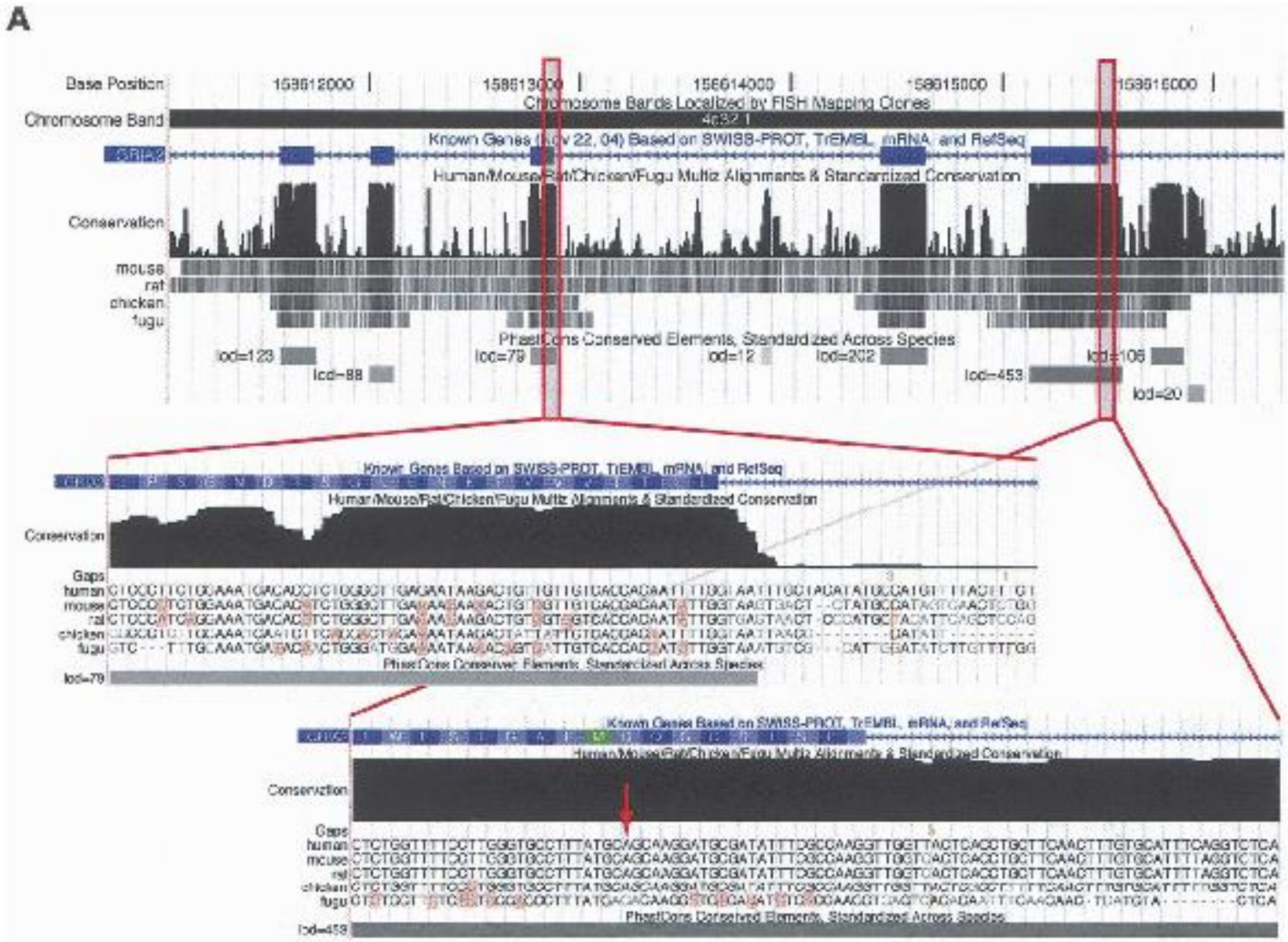
- Direct detection of binding events (e.g. ChIPSeq)
  - ***but*** binding may be non-functional!

- Computational search for clusters of recurring motifs
  - ***but*** motifs occur frequently by chance, in any large genome!

Compare genomes of …

- a lab organism & a singly mutated variant with an altered phenotype
  - the mutation must then alter (or create!) a site
    - or alter site spacing
  - and the phenotypic change illuminates its function
  - but remember that cells with identical genomes can sometimes have different phenotypes!
    - Tissues in multicellular organisms
- members of a natural population
  - Usually *multiple* genomic and phenotypic differences
  - find correlations (of *recurring* differences) to identify sites that affect a particular phenotype.

- different species
  - *Many* differences
  - ***atypically similar* (= "conserved")** regions likely represent site clusters in which mutations have been selected against ("purifying selection")
    - and likely have similar functions in the two species
  - But many sites may have been *lost*, and *created*, in each lineage



from Siepel A. *et al.* (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* 15:1034-50.

Some major computational tasks

- Comparing & aligning sequences
  - Reads to reads
    - assembly
  - Reads to genomes
    - variant detection
    - transcript assembly
  - Genomes to genomes (or portions thereof)
    - Evolutionary conservation

Appropriate alignment method depends on how similar the sequences are!

- Computational ***models*** of
  - Genome sequences
    - sites, site clusters, and “background”
  - Sequence evolution
    - Evolutionarily related sequences
      - Alignment scoring
    - Conserved vs neutrally evolving regions
  - Other types of ‘linear’ data associated to the genome (e.g. read depth)

Probability Models

- We use *probability models* for this purpose:
  - genomes are products of a probabilistic process (evolution)
  - detecting biological “signal” against “noise” of background sequence or mutations
  - measure of reliability

Models: simplicity vs complexity

- "*All models are wrong; some models are useful.*" – George Box

- "*What is simple is always wrong. What is not is unusable.*" – Paul Valery

- "*Everything should be made as simple as possible, but not simpler.*" – Albert Einstein (?)

Some disadvantages of complexity

- Computational challenge
- Overfitting
- (Lack of) interpretability

This course

- The focus is ***sequence-based*** CMB
  - i.e. methods (& models) for obtaining & analyzing the information encoded in the genome
- We emphasize the underlying ***biology***
- ***Simple / interpretable*** computational models are favored
- ***Proofs*** are often only intuitive sketches, omitting details

Main topics

- ***Suffix arrays*** (& hash tables) for finding exact matches
- ***Background sequence models***
- ***Site models,*** weight matrices & sequence logos
- Highest weight paths on weighted directed acyclic graphs: ***dynamic programming algorithm***
- Finding non-background-like regions ("HMMs lite")
- Edit graphs & ***gapped-alignment algorithms***
- ***Hidden Markov models*** and applications
  - Parsing genomes (into sites & non-sites)
  - Finding conserved regions
- Simple molecular evolution models

We do ***not*** cover:

- Other motif-finding methods
- Sequence evolution models (in depth)
- Statistical genetics
- Deep neural nets & other complex machine-learning models
- 'Non-linear' (non-sequence based) computational biology, such as:
  - Most proteomics, metabolic & signalling pathways, models for interacting molecules …

(See Genome 541, & courses in CSE, Stat, Biostat)