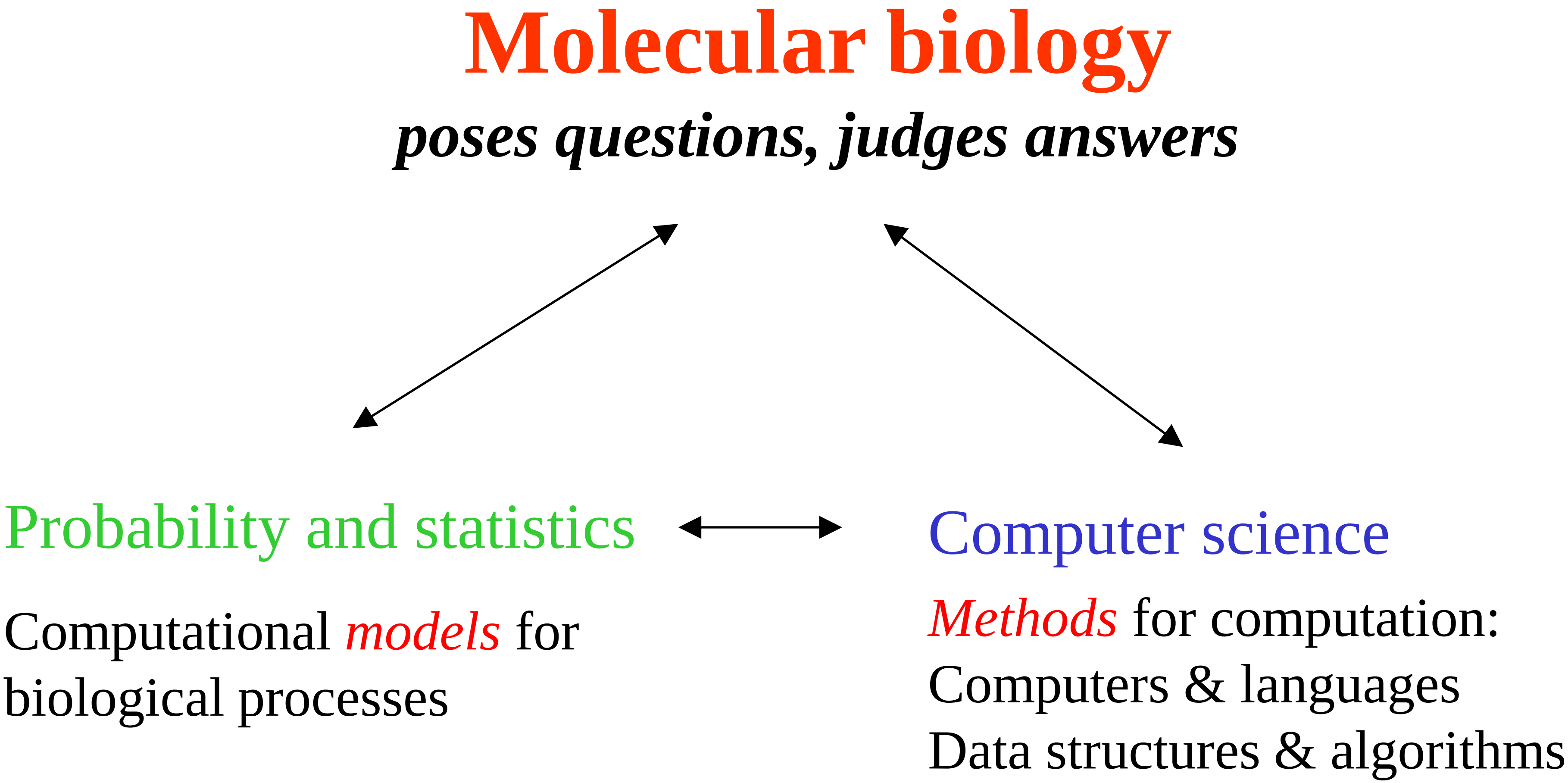
Lecture 1.1: Overviews

- Computational molecular biology
- Genome biology
- Genomicists’ tasks
- Role of computation
- This course

Computational Molecular Biology



Molecular biology

poses questions, judges answers

Probability and statistics

Computational *models* for
biological processes

Computer science

Methods for computation:
Computers & languages
Data structures & algorithms

- This course is about ***sequence-based*** CMB
  - i.e. methods (& models) for obtaining & analyzing the information encoded in the genome

- We do not cover ‘non-linear’ (non-sequence based) computational biology, such as:
  - Most proteomics, metabolic & signalling pathways, models for interacting molecules …

Genome biology overview

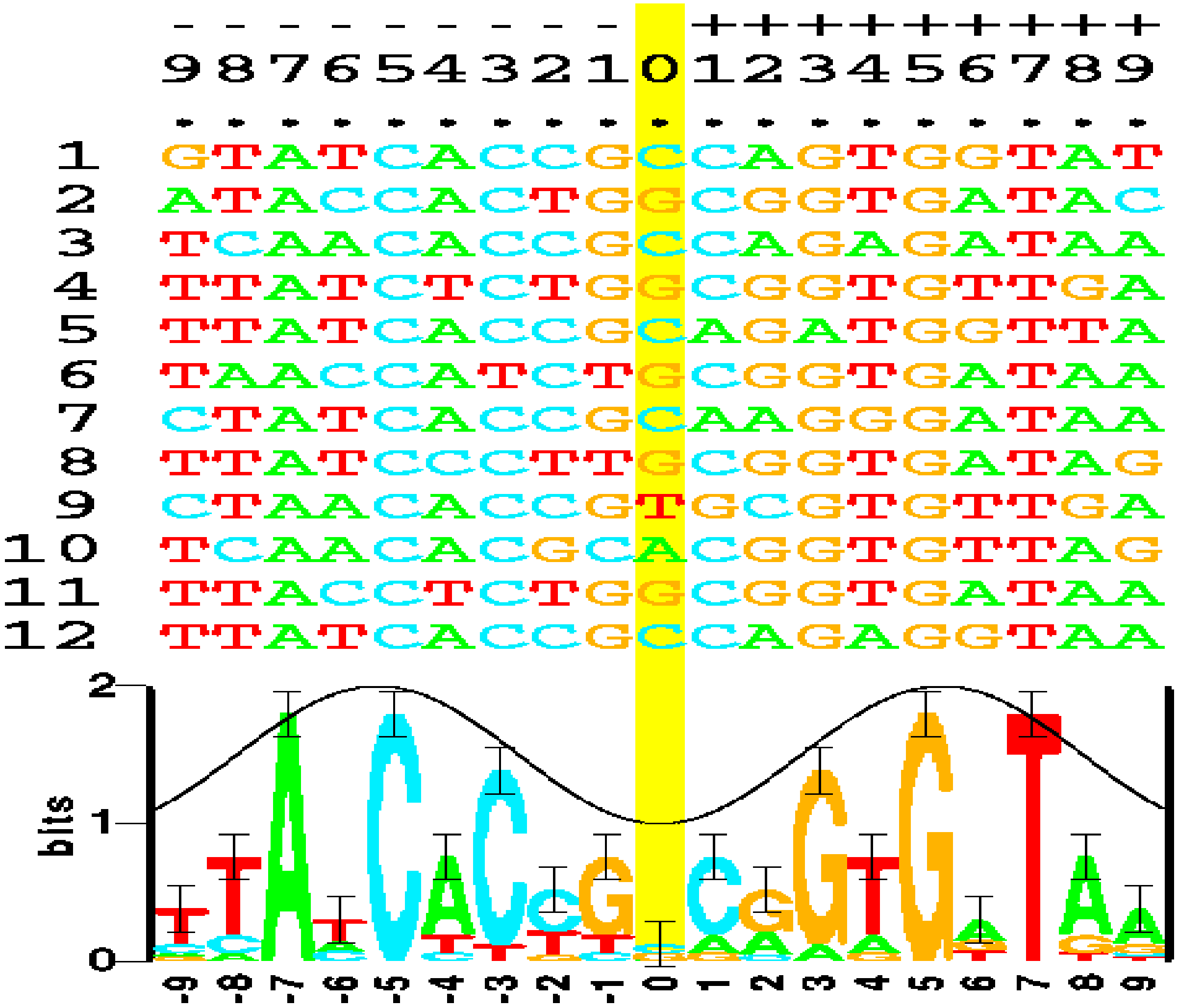
- Genomes undergo two fundamental processes (both involve copying!):
  - Replication
  - Transcription
- Genomic functional information is in the form of *sites:*
  - Short (~2 – ~15 base) sequence segments that bind to an *RNA* or *protein* molecule (the *reader*) to help mediate some function

Two broad classes of sites:

1. Sites acting ***at the DNA level*** (usually via *protein* readers) to help carry out or regulate a fundamental process

   - Replication
     - Replication origins, centromeres, telomeres (each usually having *multiple* sites)
   - Transcription
     - Promoters, enhancers, suppressors (each usually having *multiple* sites, with readers being *transcription factors*)

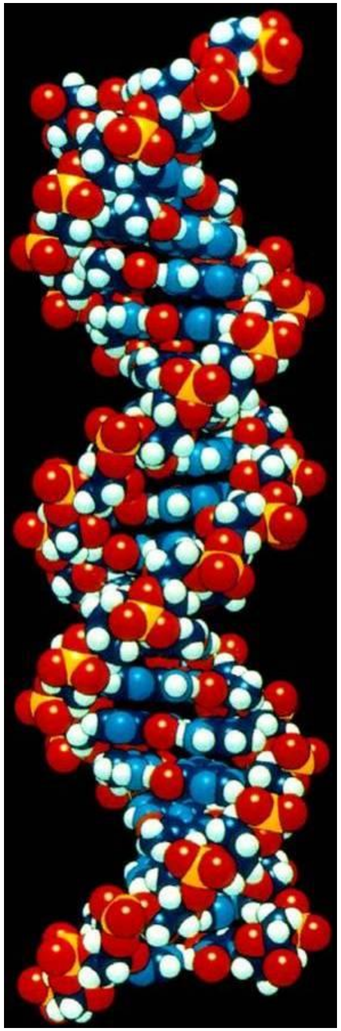
From http://www-lmmb.ncifcrf.gov/~toms/sequencelogo.html

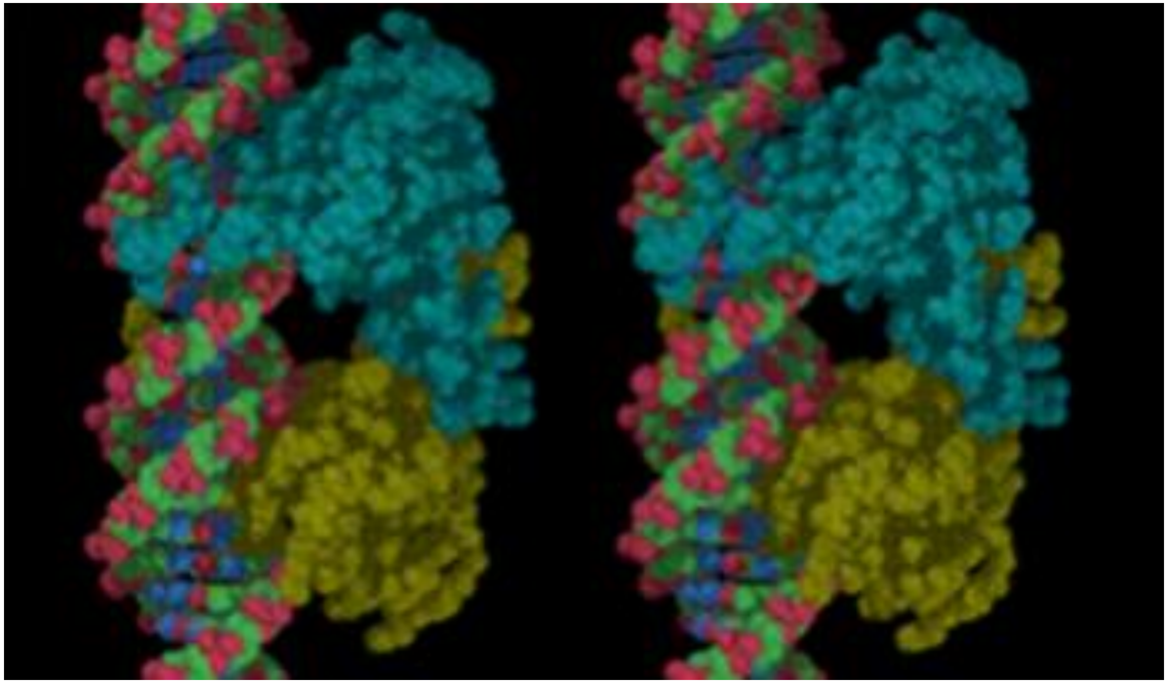


```
   ---------+++++++++
   9876543210123456789
 1 GTATCACCGCCAGTGGTAT
 2 ATACCACTGGCGGTGATAC
 3 TCAACACCGCCAGAGATAA
 4 TTATCTCTGGCGGTGTTGA
 5 TTATCACCGCAGATGGTTA
 6 TAACCATCTGCGGTGATAA
 7 CTATCACCGCAAGGGATAA
 8 TTATCCCTTGCGGTGATAG
 9 CTAACACCGTGCGTGTTGA
10 TCAACACGCACGGTGTTAG
11 TTACCTCTGGCGGTGATAA
12 TTATCACCGCCAGAGGTAA
```

12 Lambda cI and cro binding sites

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].



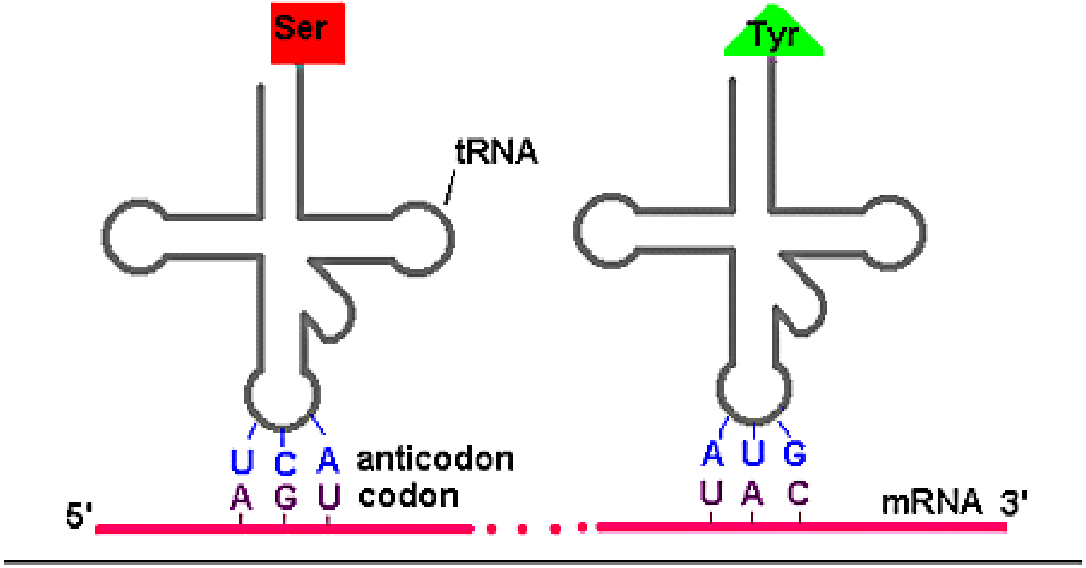


from http://gibk26.bse.kyutech.ac.jp

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

2. Sites acting ***within a transcript*** (often via RNA readers) to 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!), …



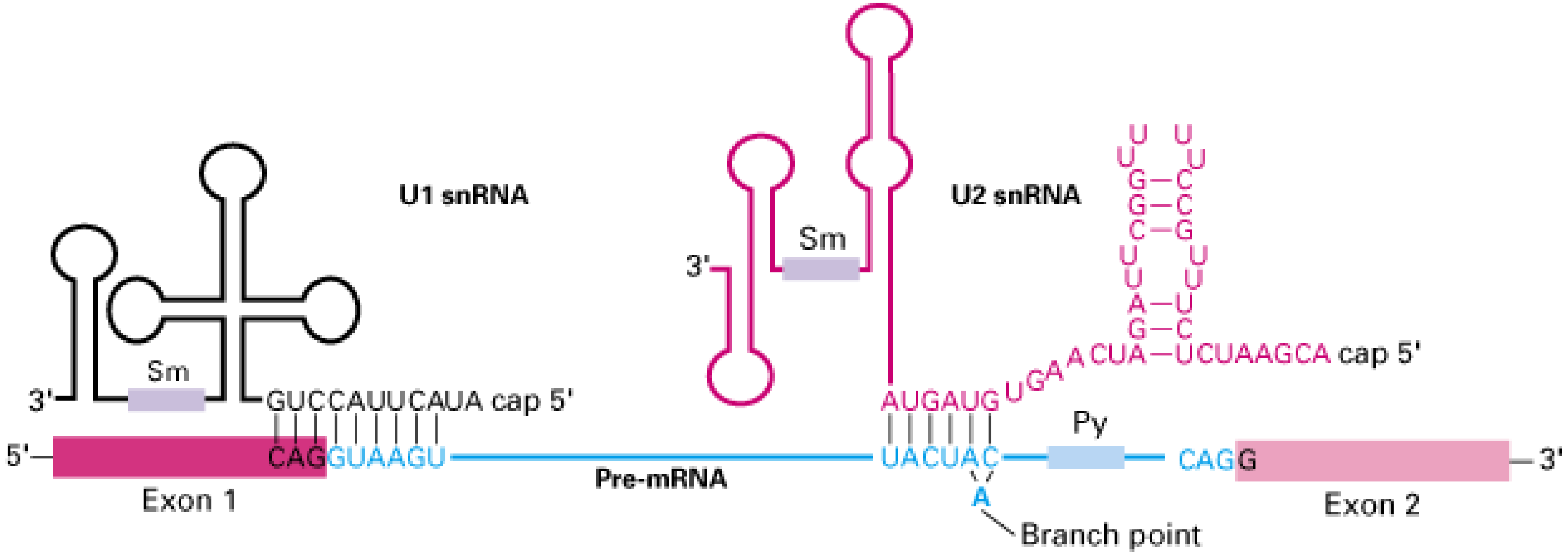
2nd base in codon

1st base in codon

	U	C	A	G	
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

3rd base in codon

The Genetic Code



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

(Jonathon Stillman, Grace Fisher-Adams)

Sites: some key properties

- 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 *sequence logo*
  - *Cluster:* several sites, with the same or different readers, may act collectively to carry out some function
    - Positional constraints within clusters vary in stringency
    - A *gene* is a cluster of sites involved in *expressing* a particular transcript

- Average site density (i.e. 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
- Whether or not a site is active in a given cell may depend on
  - reader status (local concentration, whether modified, etc)
    - Also interaction partners of the reader
  - chromatin & methylation status
  - whether nearby (or overlapping) sites are bound by their readers

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
- *Assemble* to infer transcript sequence
- *Align* to genome sequence
- Main challenges:
  - – A given transcript may not be present in all cells
  - – 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!
- Both methods are error-prone, and neither illuminates site ***function.*** So we also …

Compare genomes of …

- a lab organism & a singly mutated variant with an altered phenotype
  - the mutation must then alter (or create!) a site, 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”)
  - but generally can’t conclude anything about function
    - too many phenotype differences
  - also, many sites may have been *lost*, and *created*, in each lineage

Major computational tasks

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

Appropriate alignment method depends on how similar the sequences are!

- Developing probability models of
  - Genome sequences (sites, and “background”)
  - Sequence evolution
  - Other types of ‘linear’ data associated to the genome (e.g. read depth)

and using them to find genomic features.

This course (approximately!):

- ***Suffix arrays*** 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 regions of biased composition (“HMMs lite”)
- Edit graphs & ***gapped-alignment algorithms***
- ***Hidden Markov models*** and applications
  - Parsing genomes (into sites & non-sites)
  - Finding conserved regions

- We do ***not*** cover (but see Genome 541!)
  - motif-finding methods
  - sequence evolution models
  - more complex machine learning models (e.g. deep neural nets)