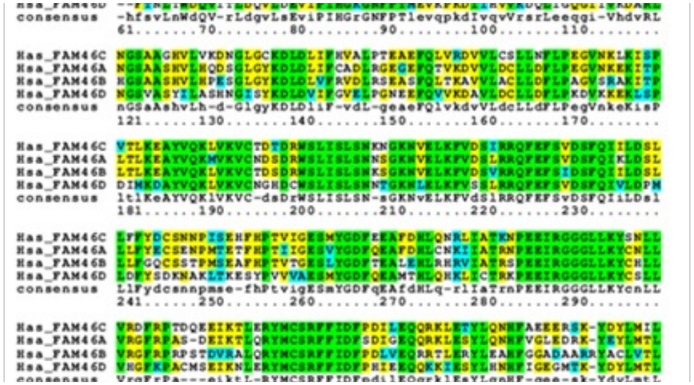
Prof. Sorin Istrail
CSCI 1820/2820:
An overview

January 25, 2024

- **Ch. 1 The BLAST Algorithm and Karlin-Altschul Statistics**
- **Ch. 2 Genome Assembly Algorithms and Haplotype Assembly Algorithms**
- **Ch. 3 Hidden-Markov Models: The Learning Problem**
- **Ch. 4 Recombination and Ancestral Recombination Graphs (ARGs) Algorithms**
- **Ch.5 Rigorous clustering: Spectral Graph Theory Algorithms**
- **Ch. 6 Algorithms for Constructing Suffix Trees in Linear Time**
- **Ch. 7 Protein Folding (An Introduction)**

Ch. 1: BLAST Algorithm



Given a biomolecular query sequence Q
and a database DB of biomolecular sequences

Find all the biomolecular sequences in DB that have high alignment scores to the query

Biomolecular: DNA, RNA, protein

Problems we need to solve along the way:

Problem 1. General scoring schemes as hypotheses testing frameworks
The Karlin-Altschul Statistics and the max scoring subsequence

Problem 2. Random Walks Theory and The Gambler's Ruin's Problem

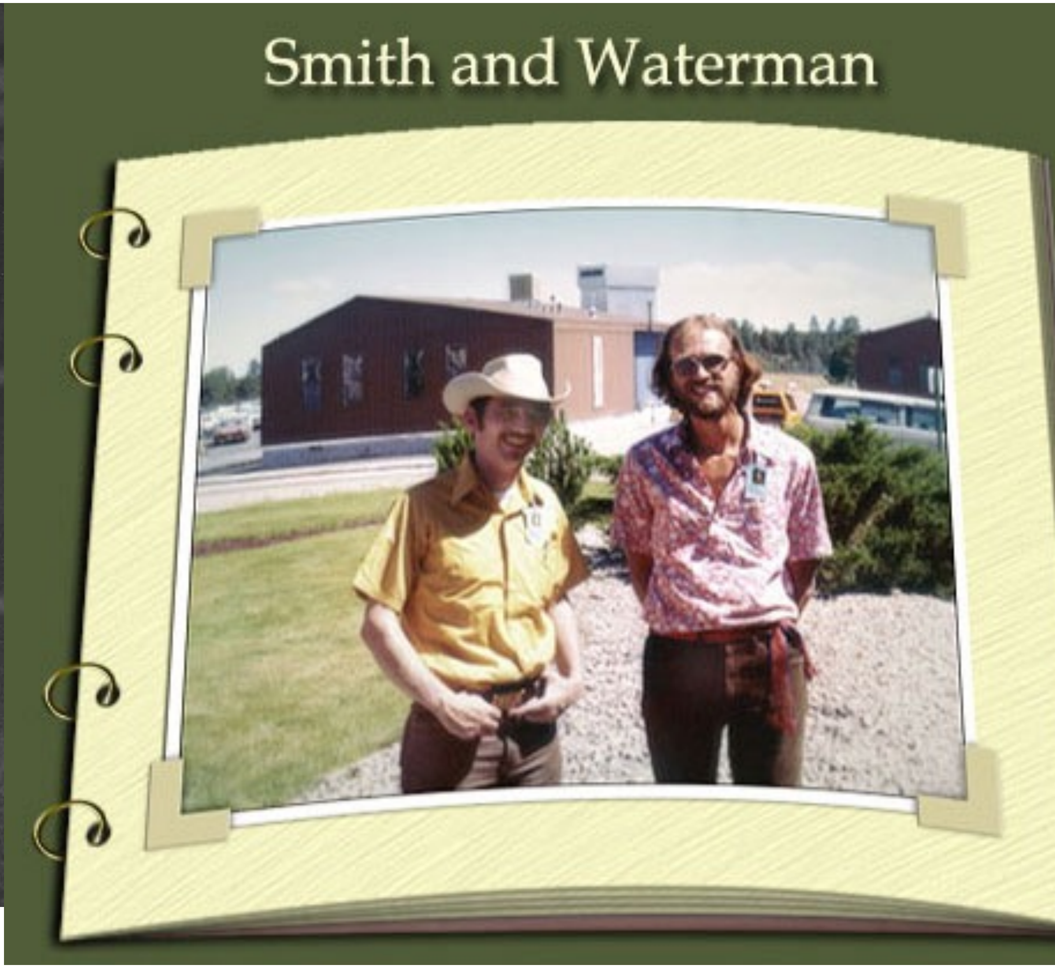
Problem 3. De-noising: how long an alignment needs to be non-random?

Problem 4. Information Theory and the theory of scoring matrices for alignment

Dr. Margaret Oakley Dayhoff
The Mother & Father of Bioinformatics

Temple Smith and Michael Waterman
at Los Alamos, New Mexico

Photo by David Lipman, Taken Summer of 1980



Sir Ronald Aylmer Fisher

The Lady Tasting Tea Problem

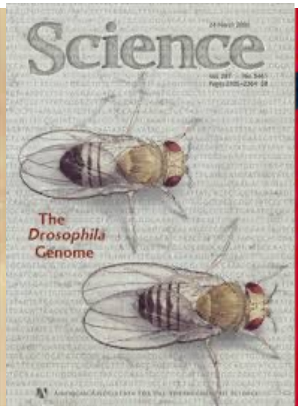
	the
	Null Hypothesis
Born	Ronald Aylmer Fisher 17 February 1890 London, England
Died	29 July 1962 (aged 72) Adelaide, South Australia, Australia

- Linear discriminant analysis is a generalization of Fisher's li discriminant[47][83]
- Fisher information, see also scoring algorithm also known a scoring, and Minimum Fisher information, a variational prin which, when applied with the proper constraints needed to empirically known expectation values, determines the best distribution that characterizes the system.[84]
- *F*-distribution, arises frequently as the null distribution of a statistic, most notably in the analysis of variance
- Fisher–Tippett–Gnedenko theorem : Fisher's contribution t made in 1927
- Fisher–Tippett distribution
- Fisher-Yates shuffle algorithm
- Von Mises–Fisher distribution[85]
- Inverse probability, a term Fisher used in 1922, referring to fundamental paradox of inverse probability" as the source o confusion between statistical terms which refer to the true v estimated, with the actual value arrived at by estimation, w subject to error.[86]
- Fisher's permutation test
- Fisher's inequality[87]
- Sufficient statistic, when a statistic is *sufficient* with respect a statistical model and its associated unknown parameter if statistic that can be calculated from the same sample provic additional information as to the value of the parameter".[88]
- Fisher's noncentral hypergeometric distribution, a generali the hypergeometric distribution, where sampling probabiliti modified by weight factors.
- Student's *t*-distribution, widely used in statistics.[89][90]
- The concept of an ancillary statistic and the notion (the anc principle) that one should condition on ancillary statistics.

The BLAST algorithm

- Detect all *word hits* (exact, or nearly identical matches) of a given length between the two sequences
  - k=10 for nucleotide sequences (exact word matches)
  - k=3 for protein sequences (nearly identical word matches)
- Extend the word hits in both directions to high-scoring *gap-free* segment pairs (HSPs)
  - retain only HSPs that score above a threshold
  - start from the center of the HSP (original BLAST, 1990), or from the center of a pair of HSPs located close to each other on the same diagonal (gapped BLAST, 1997)
- Extend the HSPs in both directions allowing for gaps
  - use dynamic programming, and stop when the alignment score falls more than a threshold X below the best score yet seen
- Report all statistically significant local alignments
  - E-value (starting with BLAST 2.0) is used to measure the statistical significance
  - *E-value* = the number of alignments with score equal to or higher than *s* one would expect to find by chance when searching the database

Ch. 2: Genome Assembly Algorithms



Questions: What algorithms to use to assemble DNA pieces into contigs and scafolds?
How long are the contigs?
How much the DNA target region is covered by the contigs?
How to measure the success of a genome assembly?

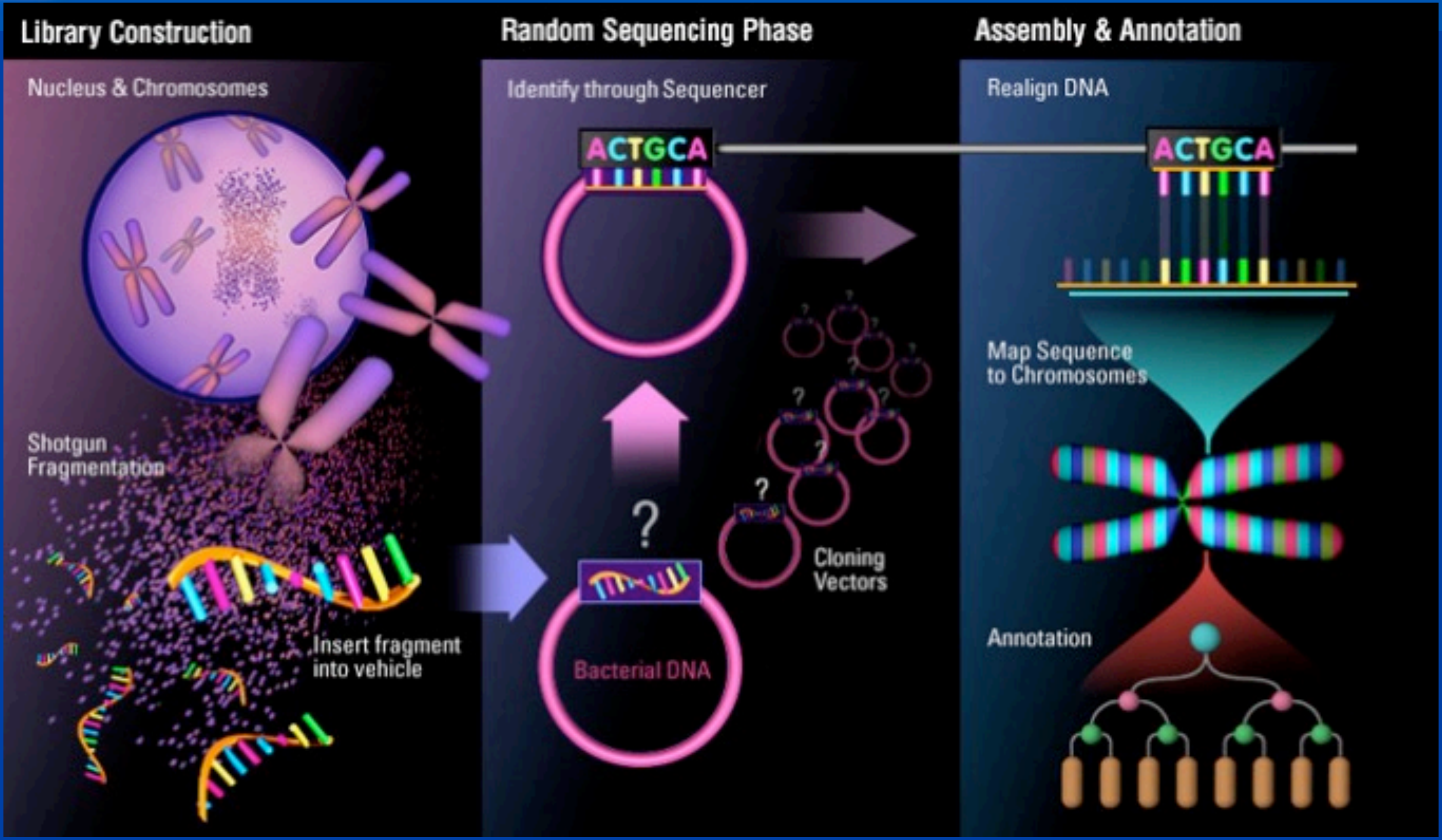
Problems we need to solve along the way

Problem 1. Genome Assembly Algorithms

Problem 2. Poisson statistics for DNA and Genome Assembly

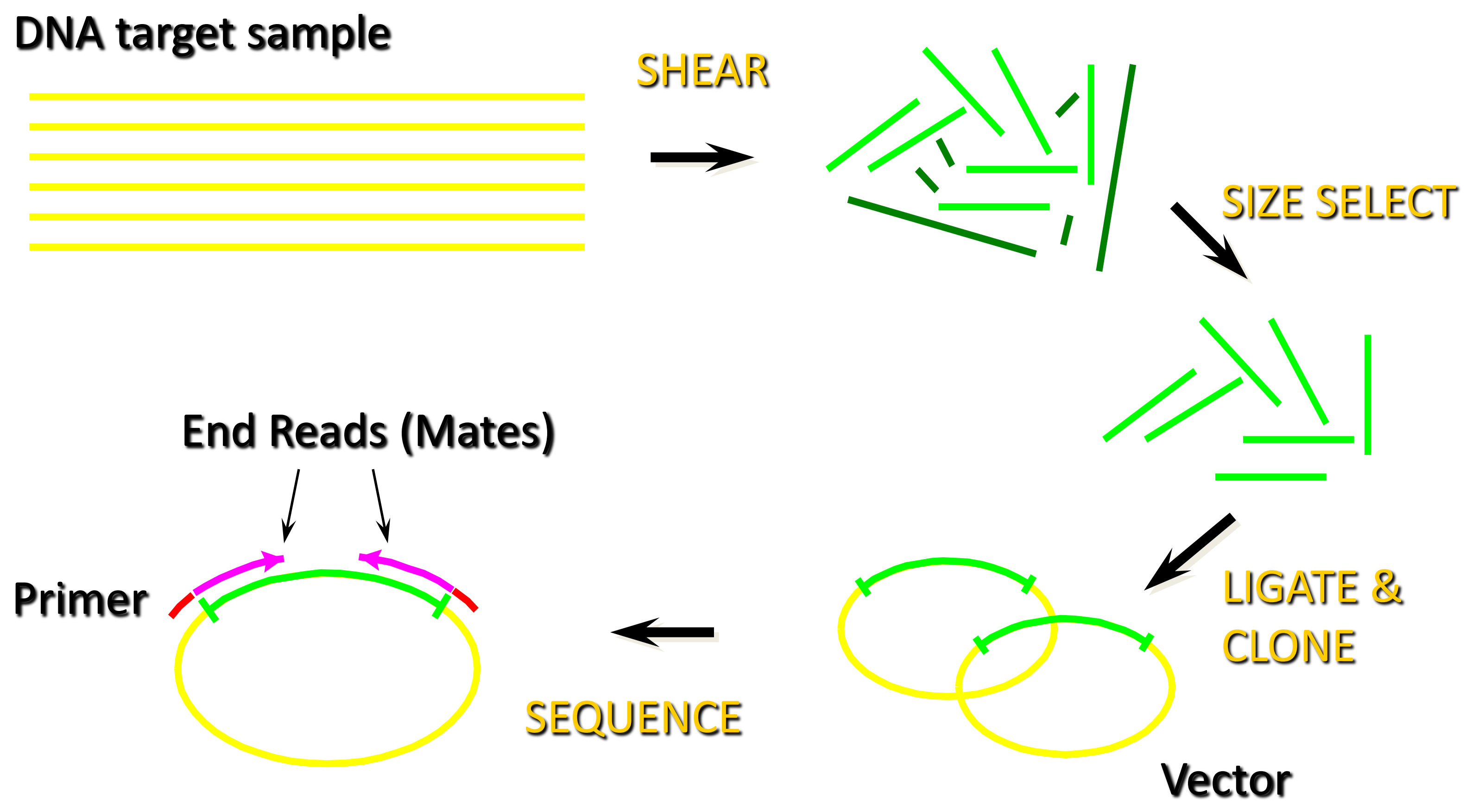
Problem 3. Ham Smith’s DNA Lab with no windows

Whole Genome Shotgun Sequencing

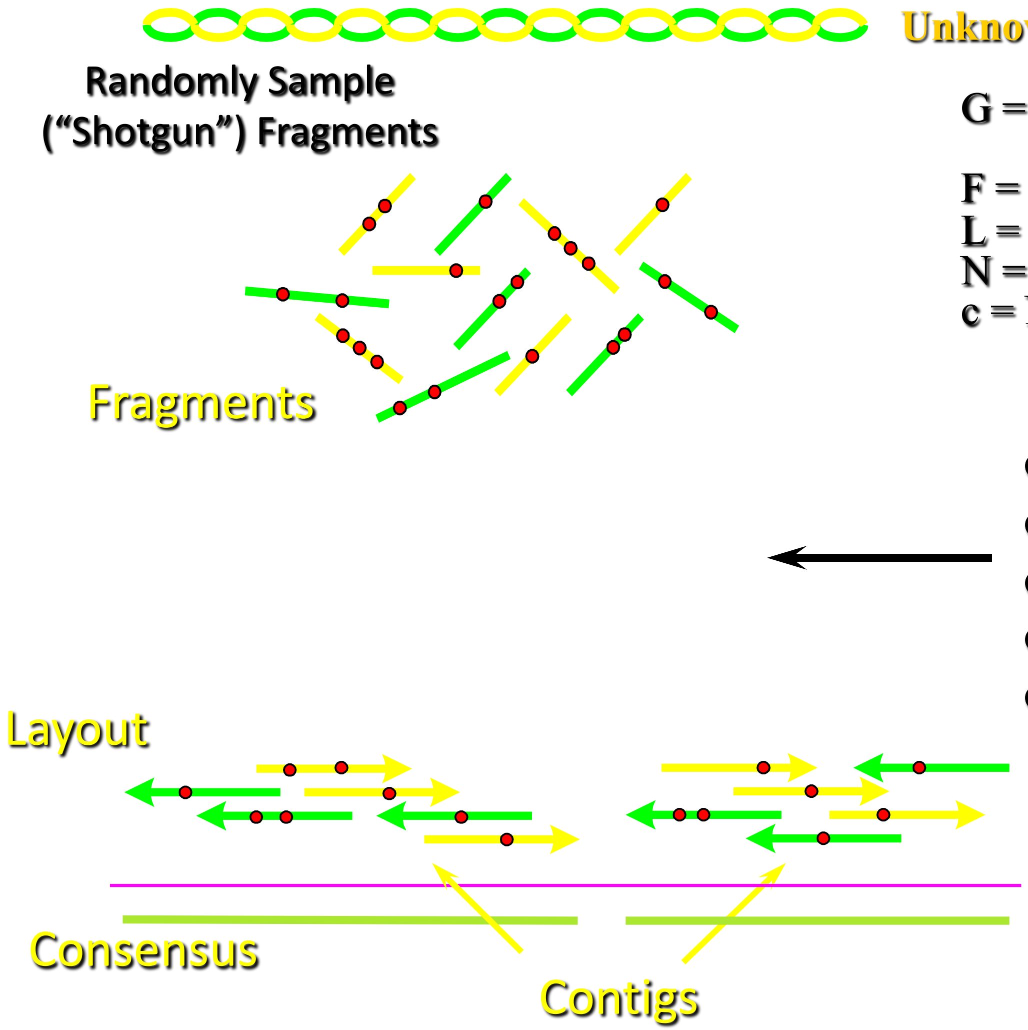


CELERA
A PE Corporation Business

Shotgun DNA Sequencing (Technology)



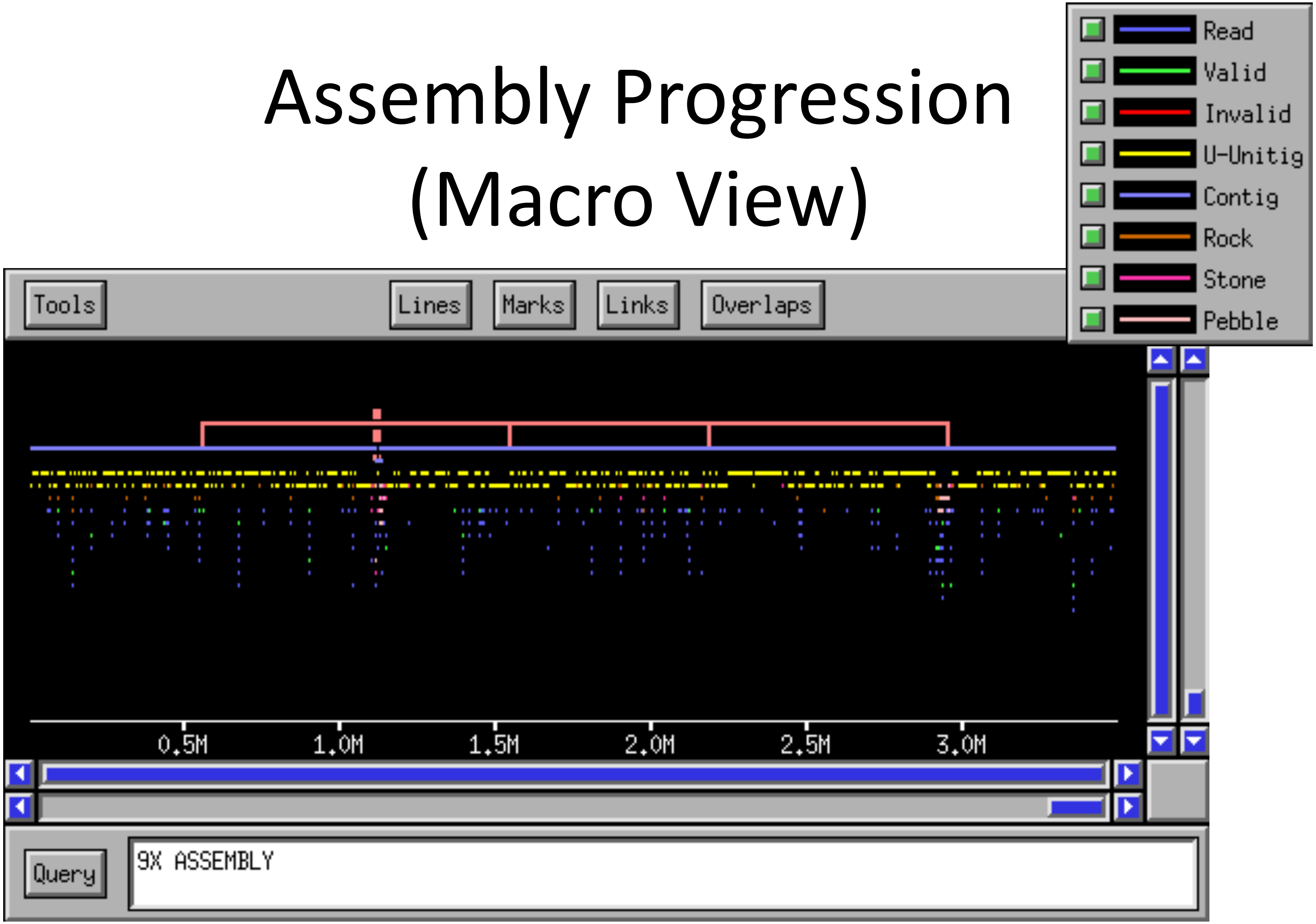
Shotgun DNA Sequencing (Computation)


Unknown "Target" DNA Sequence
Randomly Sample ("Shotgun") Fragments
Fragments
Layout
Consensus
Contigs

G = 100Kbp	Target Length (e.g., BAC, P1, PAC)
F = 1600	# of Fragments
L = 500	Avg. Fragment Length
N = FL = 800Kbp	Total Bases Sequenced
c = N/G = 8	Avg. Coverage

- ● UNKNOWN ORIENTATION
- ● SEQUENCING ERRORS
- ● INCOMPLETE COVERAGE
- ● CONSTRAINTS (MATES)
- ● REPEATS

Assembly Progression (Macro View)

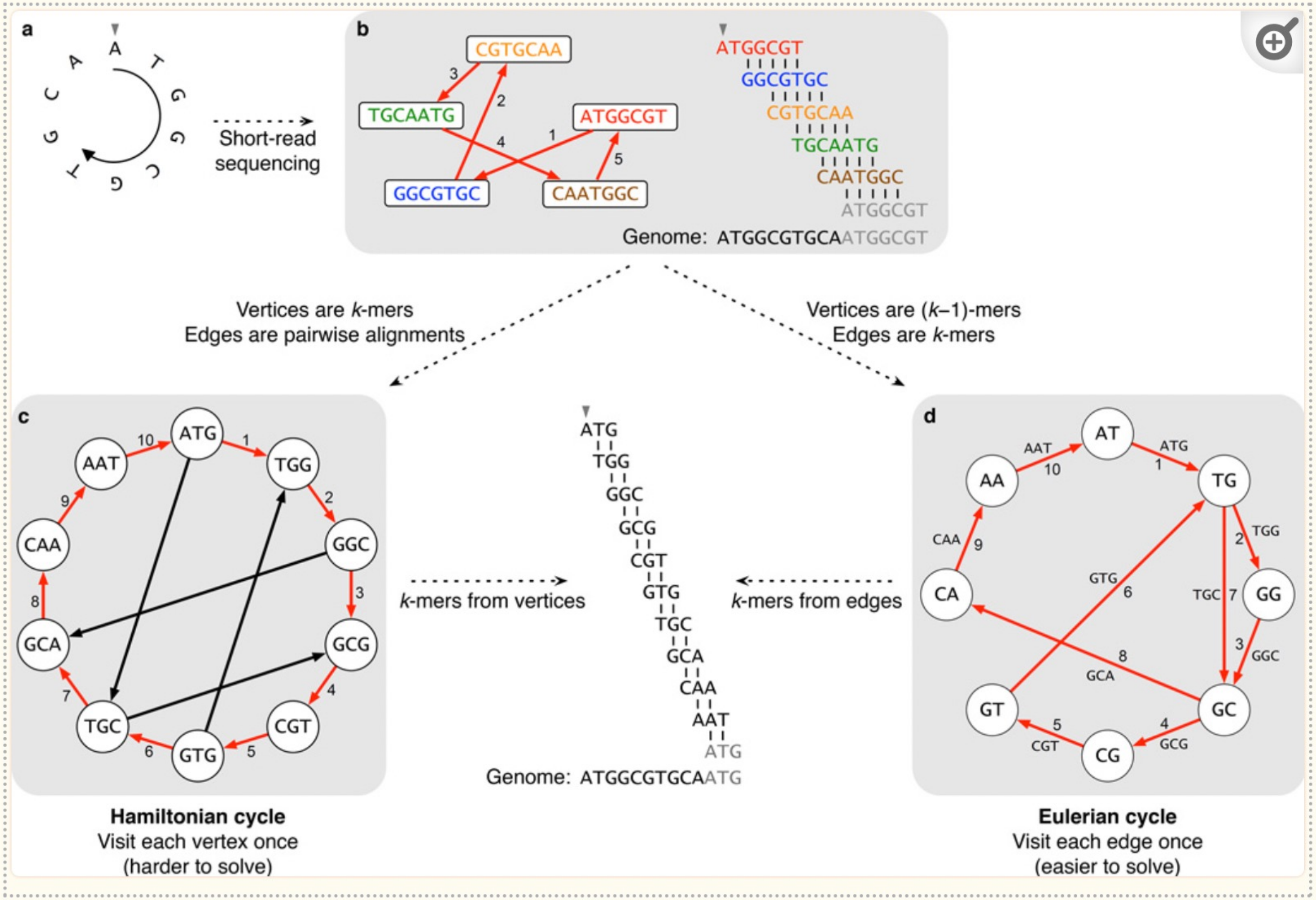


9X ASSEMBLY

Siméon Denis Poisson

Born	21 June 1781 Pithiviers, Kingdom of France (present-day Loiret)
Died	25 April 1840 (aged 58) Sceaux, Hauts-de-Seine, Kingdom of France
Alma mater	École Polytechnique
Known for	Poisson process Poisson equation Poisson kernel Poisson distribution Poisson limit theorem Poisson bracket Poisson algebra Poisson regression Poisson summation formula Poisson's spot Poisson's ratio Poisson zeros Conway–Maxwell–Poisson distribution Euler–Poisson–Darboux equation
Scientific career	
Fields	Mathematics and physics
Institutions	École Polytechnique Bureau des Longitudes Faculté des sciences de Paris École de Saint-Cyr
Academic advisors	Joseph-Louis Lagrange Pierre-Simon Laplace
Doctoral students	Michel Chasles Joseph Liouville
Other notable students	Nicolas Léonard Sadi Carnot Peter Gustav Lejeune Dirichlet

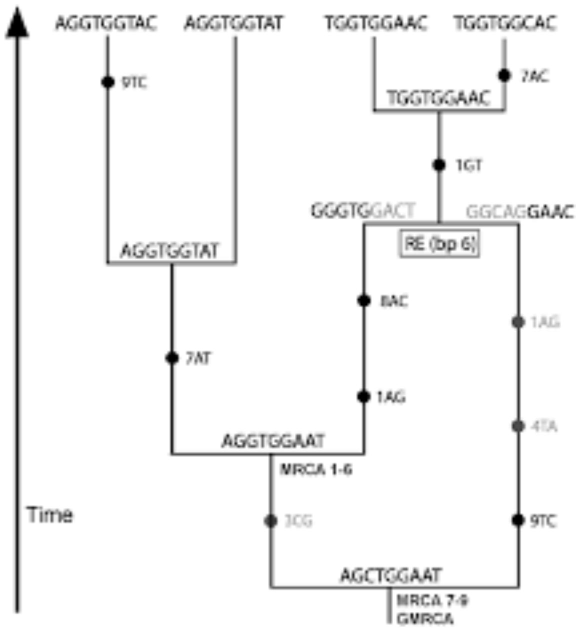
de Bruijn Genome Assembly

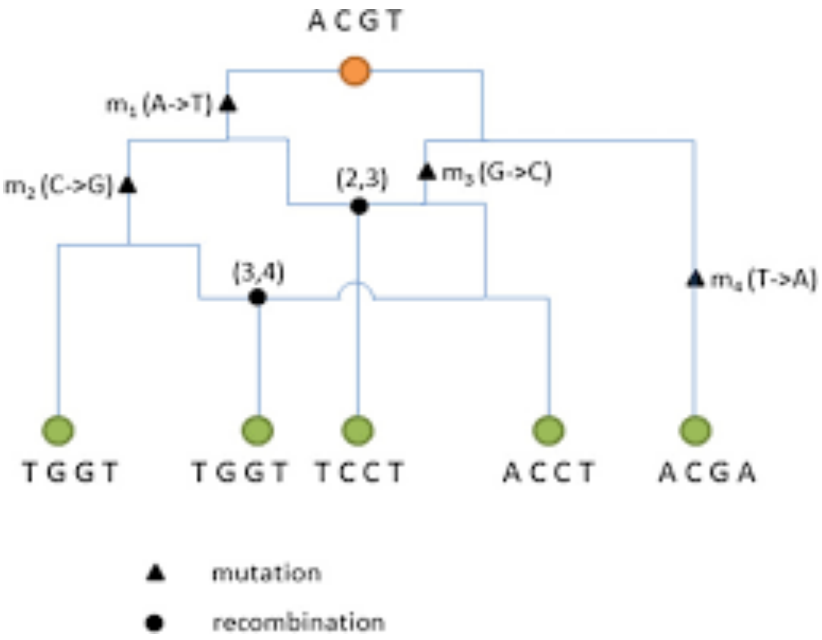


Hamiltonian cycle
Visit each vertex once
(harder to solve)

Eulerian cycle
Visit each edge once
(easier to solve)

Ch. 3 Recombination and Ancestral Recombination Graphs (ARG) Algorithms

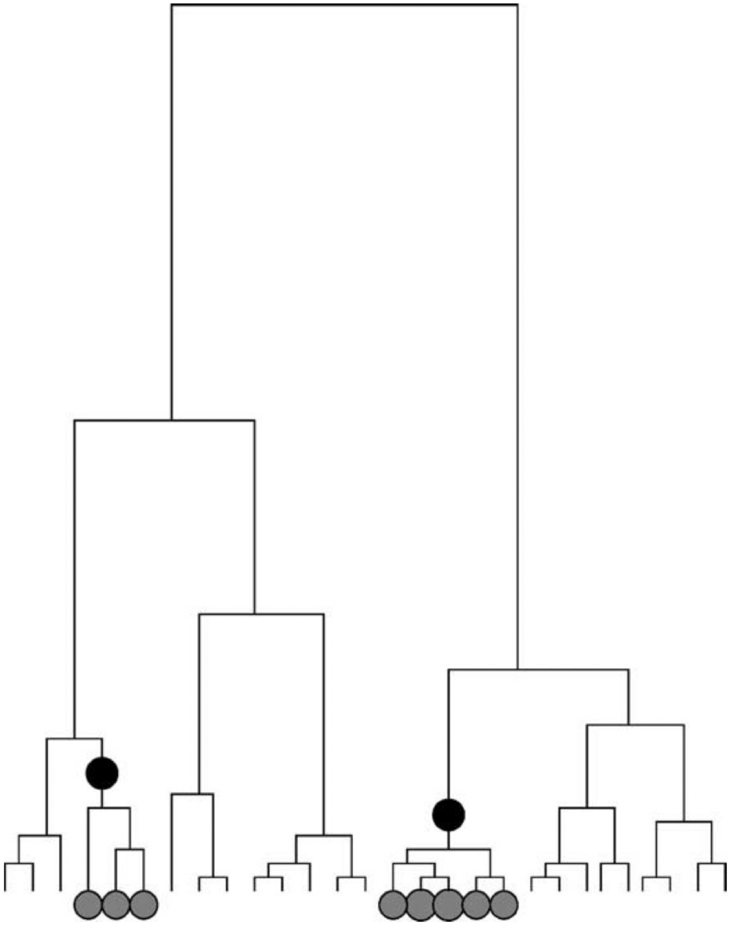


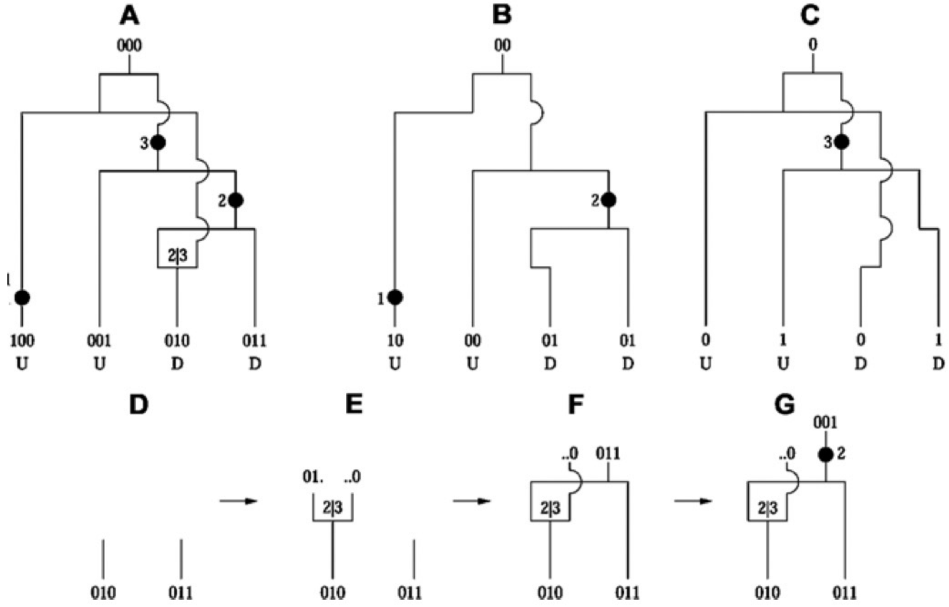


How do we reconstruct genealogies of a sample of individuals incorporating past mutations and recombinations?

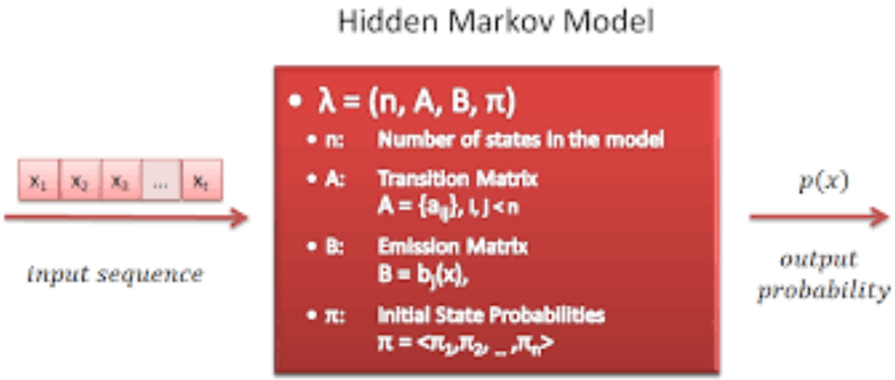
Recombination + Phylogenetic Trees = ARG

Ancestral Recombination Graph and Marginal Trees





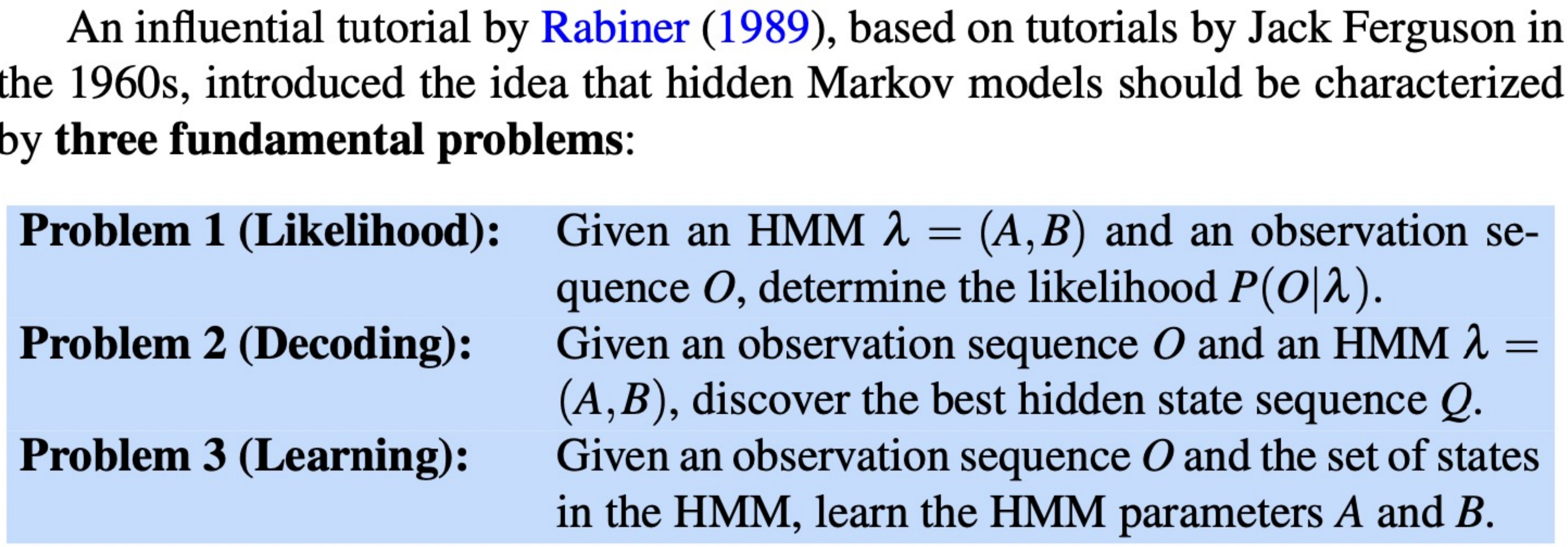
Ch. 4: HMM - the Learning Problem



- $\lambda = (n, A, B, \pi)$
  - n: Number of states in the model
  - A: Transition Matrix $A = \{a_{ij}\}$, $i, j < n$
  - B: Emission Matrix $B = b_j(x)$,
  - π: Initial State Probabilities $\pi = <\pi_1, \pi_2, \dots, \pi_n>$

Maximum Likelihood and the Expectation-Maximization problem

Three Fundamental HHM Problems



An influential tutorial by Rabiner (1989), based on tutorials by Jack Ferguson in the 1960s, introduced the idea that hidden Markov models should be characterized by **three fundamental problems**:

Problem 1 (Likelihood): Given an HMM $\lambda = (A,B)$ and an observation sequence O, determine the likelihood $P(O|\lambda)$.

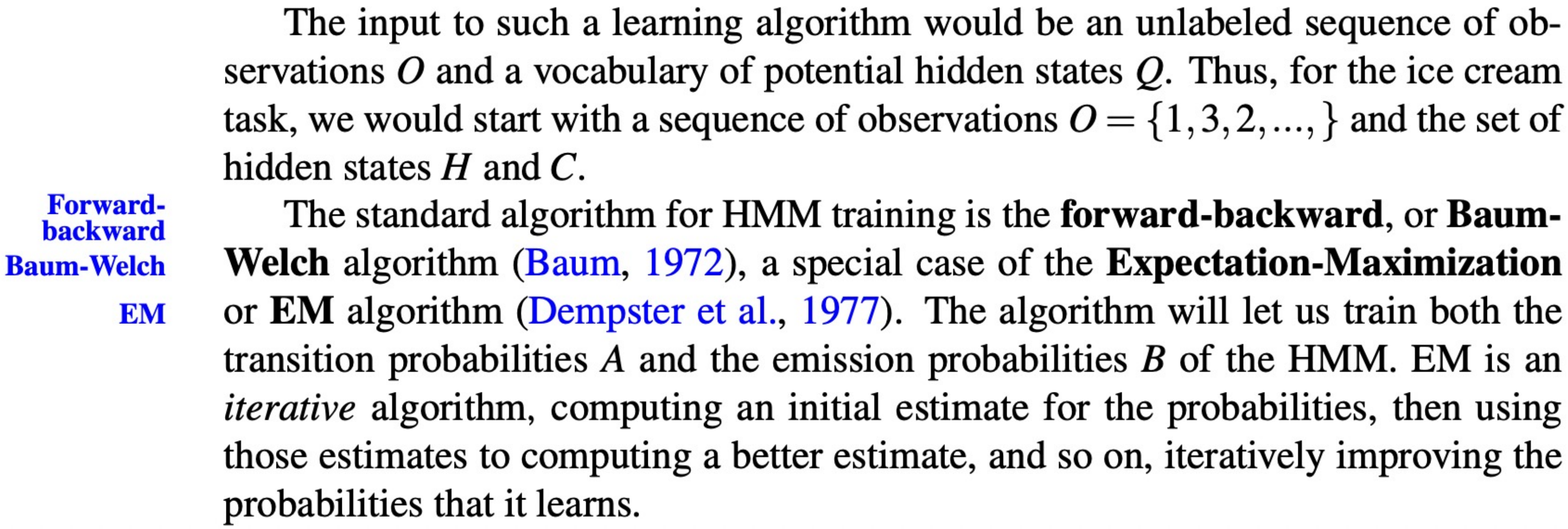
Problem 2 (Decoding): Given an observation sequence O and an HMM $\lambda = (A,B)$, discover the best hidden state sequence Q.

Problem 3 (Learning): Given an observation sequence O and the set of states in the HMM, learn the HMM parameters A and B.

HMM Training: The Forward-Backward Algorithm

We turn to the third problem for HMMs: learning the parameters of an HMM, that is, the A and B matrices. Formally,

> **Learning:** Given an observation sequence O and the set of possible states in the HMM, learn the HMM parameters A and B.



The input to such a learning algorithm would be an unlabeled sequence of observations O and a vocabulary of potential hidden states Q. Thus, for the ice cream task, we would start with a sequence of observations $O = \{1,3,2,...,\}$ and the set of hidden states H and C.

Forward-backward
Baum-Welch
EM

The standard algorithm for HMM training is the **forward-backward**, or **Baum-Welch** algorithm (Baum, 1972), a special case of the **Expectation-Maximization** or **EM** algorithm (Dempster et al., 1977). The algorithm will let us train both the transition probabilities A and the emission probabilities B of the HMM. EM is an *iterative* algorithm, computing an initial estimate for the probabilities, then using those estimates to computing a better estimate, and so on, iteratively improving the probabilities that it learns.

Ch. 5 Rigorous Clustering Algorithms Spectral Graph Theory Algorithms

Algorithms and Statistical Theory

- An introduction to Linear Algebra foundations for graph theory
- Principles of Clustering Theory
- Graph Laplacians
- Graph cuts and random walks intuitions for Spectral Clustering
- Unnormalized Spectral Clustering Algorithms
- Normalized Spectral Clustering Algorithms
- Algorithmic Fairness and Clustering

Algorithmic Fairness and Clustering

Fairness Notion 2: Exactly Balanced Clustering

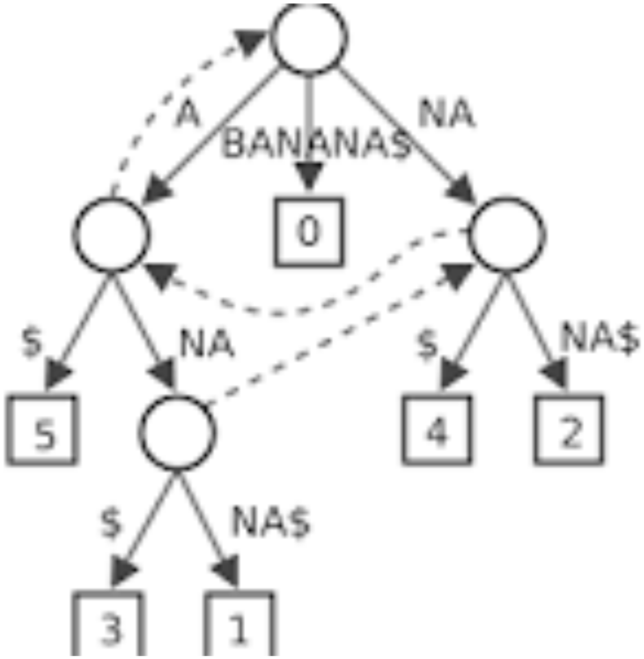
- **For multiple protected groups** (Rösner and Schmidt, ICALP 2018)
  - The data points are partitioned into ℓ many groups, say $\mathrm{C} = C_1, C_2, \ldots, C_\ell$.
  - A cluster S is called ***exactly balanced*** if $\frac{|C_i \cap S|}{|S|} = \frac{|C_i|}{|C|}$ for all i.
  - A clustering is called ***fair*** if every cluster is ***exactly balanced***.

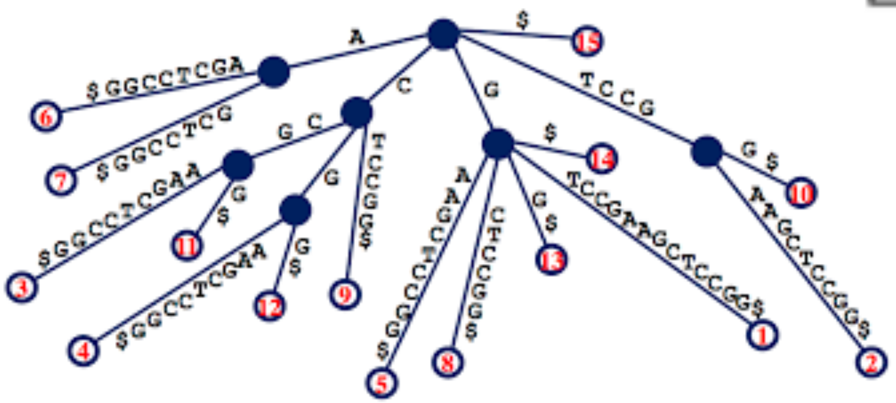
Pierre-Simon Laplace

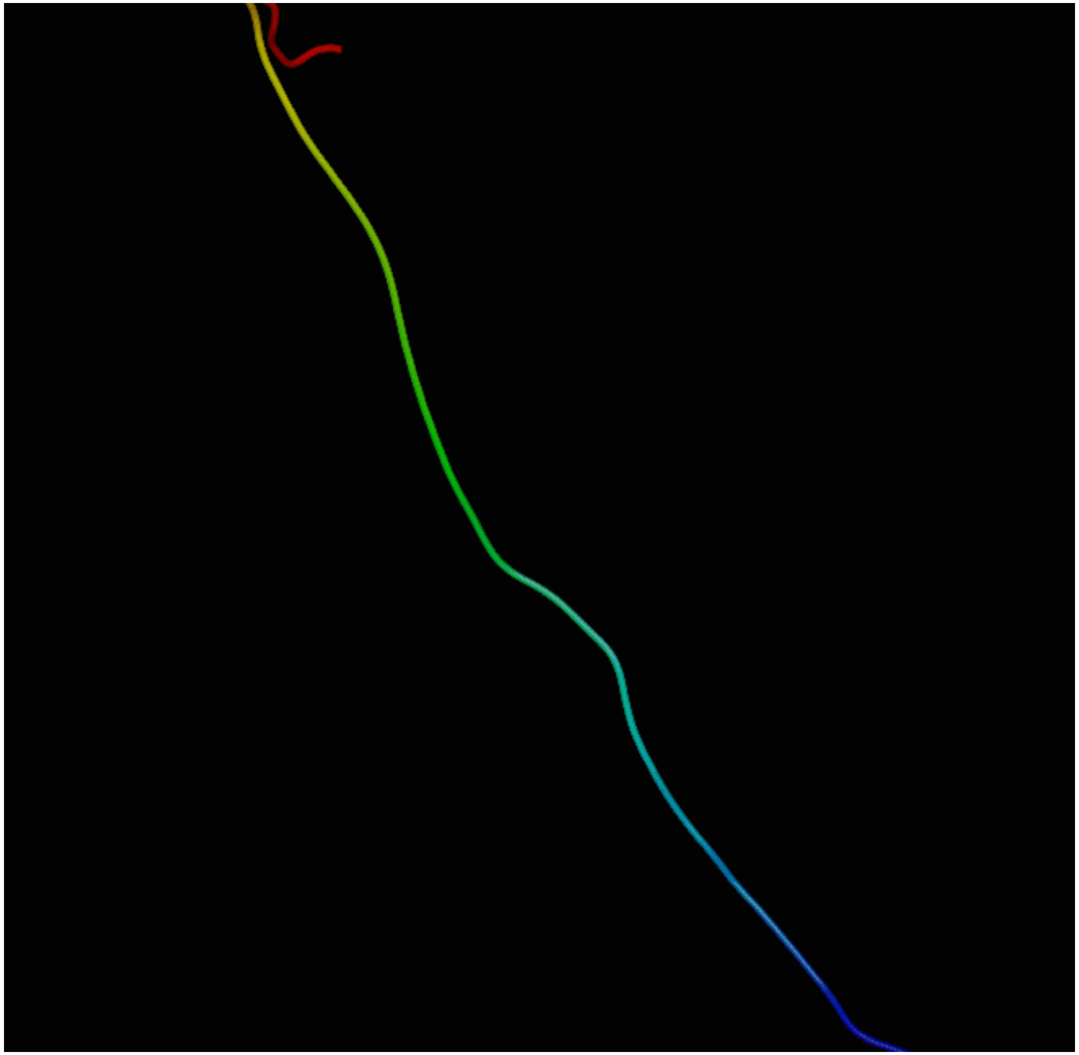
Born	23 March 1749 Beaumont-en-Auge, Normandy, Kingdom of France
Died	5 March 1827 (aged 77) Paris, Kingdom of France
Alma mater	University of Caen
Known for	show
Scientific career	
Fields	Astronomy and Mathematics

Notable students	Siméon Denis Poisson Napoleon Bonaparte

Ch. 6 Suffix Trees in Linear Time







Ch. 7

The Protein Folding Problem

Statistical Mechanics
models

Mixed character of the problem :

continuous mathematics -- geometry of surfaces &
discrete mathematics -- combinatorics of folds