

Harvard-MIT Division of Health Sciences and Technology
HST.508: Quantitative Genomics, Fall 2005
Instructors: Leonid Mirny, Robert Berwick, Alvin Kho, Isaac Kohane

Welcome to HST.508/Biophysics 170

Our emphasis

- Evolution
- Quantitative
- Medical applications

Syllabus

1. Evolutionary and population genetics
2. Comparative genomics
3. Structural genomics and proteomics
4. Functional genomics and networks

Module 1

1. Evolutionary and population genetics

- The basic forces of evolution: mutation, recombination, mating, migration. Neutral evolution and [drift](#), effective population size, coalescent theory.
- Selection, fitness, and diffusion models. Selection at genetic and higher levels
- [Phylogenetic analysis](#). Models of nucleotide evolution: [Jukes-Cantor](#), Kimura, maximum likelihood models; Human/mouse/rate examples.
- **Measuring selection:** from 'classical' methods to [maximum likelihood](#) (with applications to disease evolution, HIV and influenza)
- Medical Lecture:
Genetic diversity and evolution of hepatitis C virus

Drift, mutations, selection

Consider a deleterious recessive allele A_1 of frequency p in a randomly mating human population with mutations, selection and drift. Steady state distribution of p is given by

$$P(p) = C \exp(-2Nsp^2)(1-p)^{4Nu_1-1} p^{4Nu_2-1} \quad (1)$$

where N is an effective population size, s is selection coefficient, u_1 and u_2 are mutation rates to A_1 and from A_1 respectively, and C is a normalization coefficient.

Module 2

2. Comparative genomics

- [Sequence comparison](#), [substitution matrices](#), alignment methods, alignment statistics. [Multiple alignments](#), profiles and PSSMs.
- Genome comparison and genome evolution: duplication, recombination, insertions, repeats. Orthologs, paralog, in/out-paralog. Algorithms of genome alignment.
Conserved non-coding, positive selection. Motif discovery.
- Prediction of gene function using: homology, context, structure, networks.
- [SNPs](#): microevolution, history of population, markers medical applications ([example](#))
- [Medical Lecture](#) Finding the keys to human heart disease in the genomes of other animals.

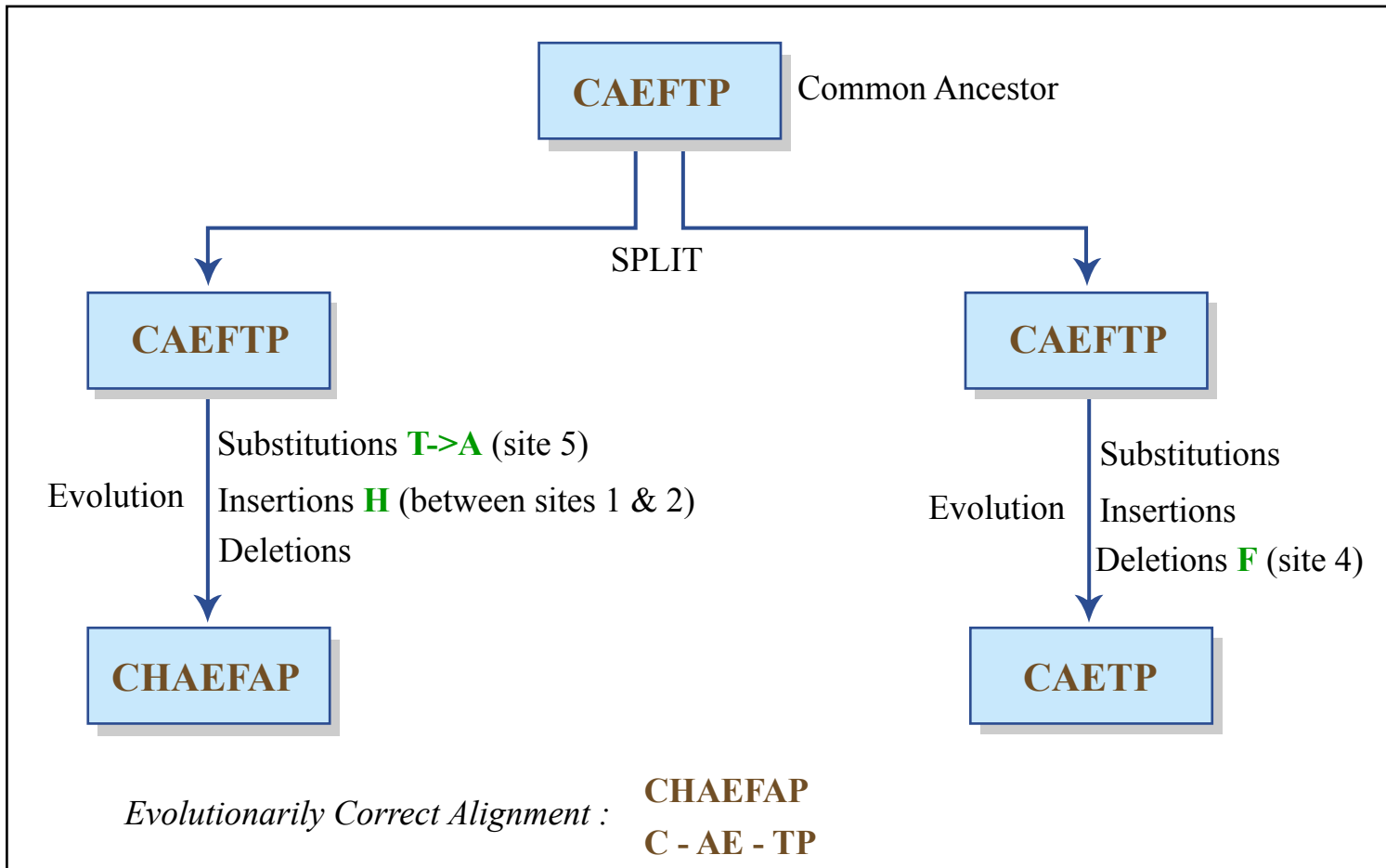


Figure by MIT OCW.

Model for Sequence evolution (DNA): Each site of the DNA sequence evolves according to a Markov Chain with state space {A,C,G,T}.

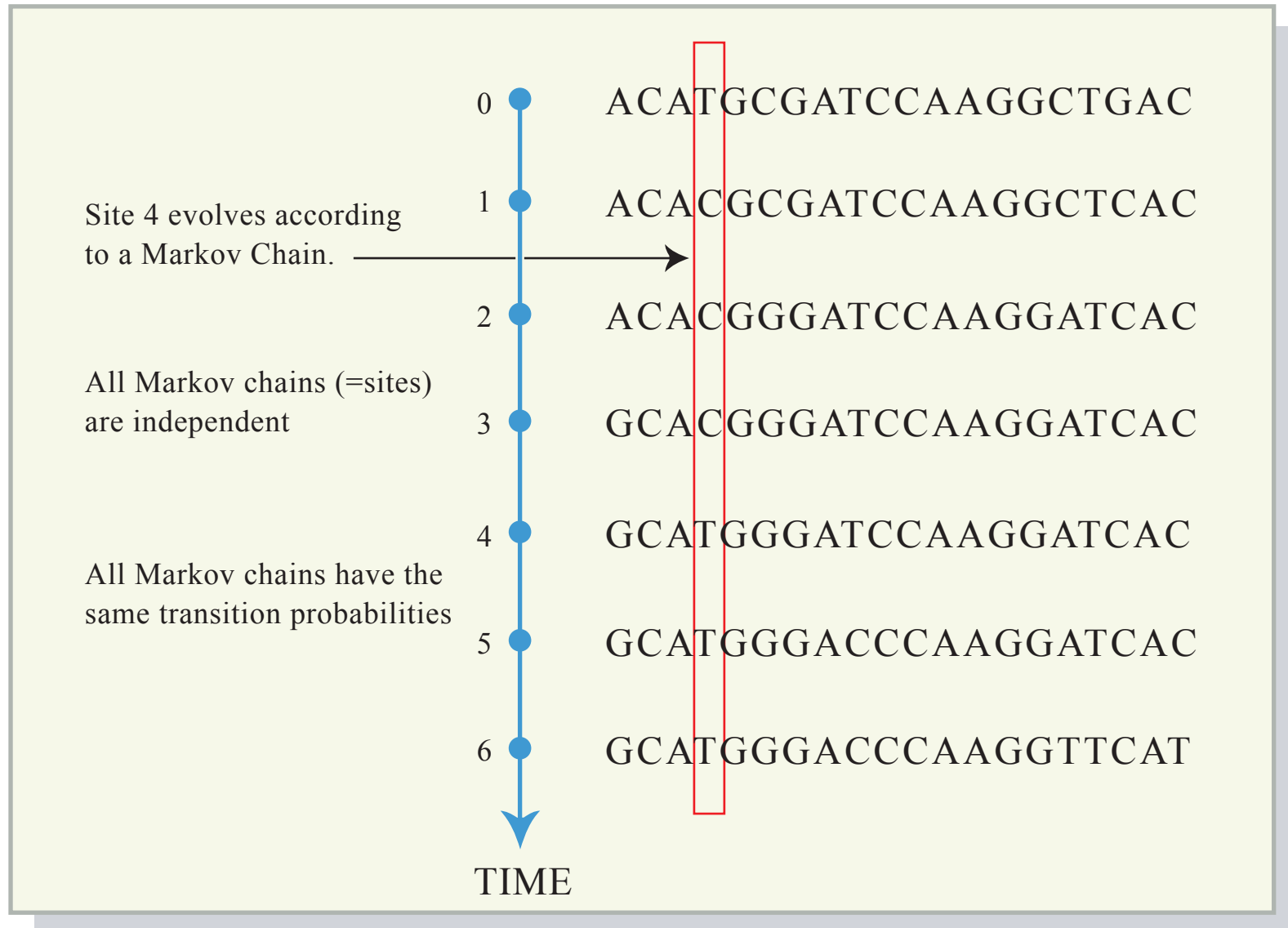


Figure by MIT OCW.

MARKOV CHAIN

Let $X_0, X_1, X_2, X_3, \dots$ be a Markov chain with **state space** S , for example $S = \{a, c, g, t\}$.

TRANSITION MATRIX

$$P = \begin{pmatrix} p_{a,a} & p_{a,c} & p_{a,g} & p_{a,t} \\ p_{c,a} & p_{c,c} & p_{c,g} & p_{c,t} \\ p_{g,a} & p_{g,c} & p_{g,g} & p_{g,t} \\ p_{t,a} & p_{t,c} & p_{t,g} & p_{t,t} \end{pmatrix}.$$

Here

$$p_{i,j} = \mathbf{P}(X_{n+1} = j | X_n = i)$$

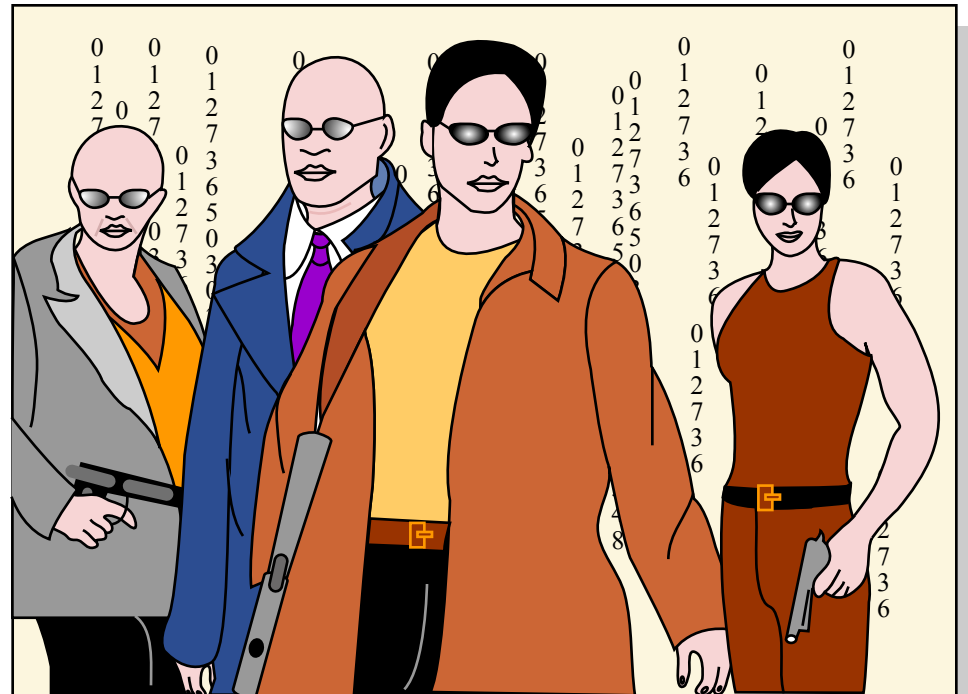
for $n \geq 0$, where $i, j \in \{a, c, g, t\}$.

MARKOV CHAIN

Let $X_0, X_1, X_2, X_3, \dots$ be a Markov chain with state space S , for example $S = \{a, c, g, t\}$.

TRANSITION MATRIX

$P =$



Here

$$p_{i,j} = \mathbf{P}(X_{n+1} = j | X_n = i)$$

Figure by MIT OCW.

for $n \geq 0$, where $i, j \in \{a, c, g, t\}$.

Simplest model for sequence evolution: Jukes-Cantor

$$\begin{pmatrix} p_{a,a} & p_{a,c} & p_{a,g} & p_{a,t} \\ p_{c,a} & p_{c,c} & p_{c,g} & p_{c,t} \\ p_{g,a} & p_{g,c} & p_{g,g} & p_{g,t} \\ p_{t,a} & p_{t,c} & p_{t,g} & p_{t,t} \end{pmatrix} = \begin{pmatrix} 1 - 3\alpha & \alpha & \alpha & \alpha \\ \alpha & 1 - 3\alpha & \alpha & \alpha \\ \alpha & \alpha & 1 - 3\alpha & \alpha \\ \alpha & \alpha & \alpha & 1 - 3\alpha \end{pmatrix}$$

The stationary distribution is $\vec{\pi} = (0.25, 0.25, 0.25, 0.25)$.

The parameter α depends on the time scale

(if the unit time is 100.000 generations, α would take a smaller value than if the unit time were chosen as 200.000 generations).

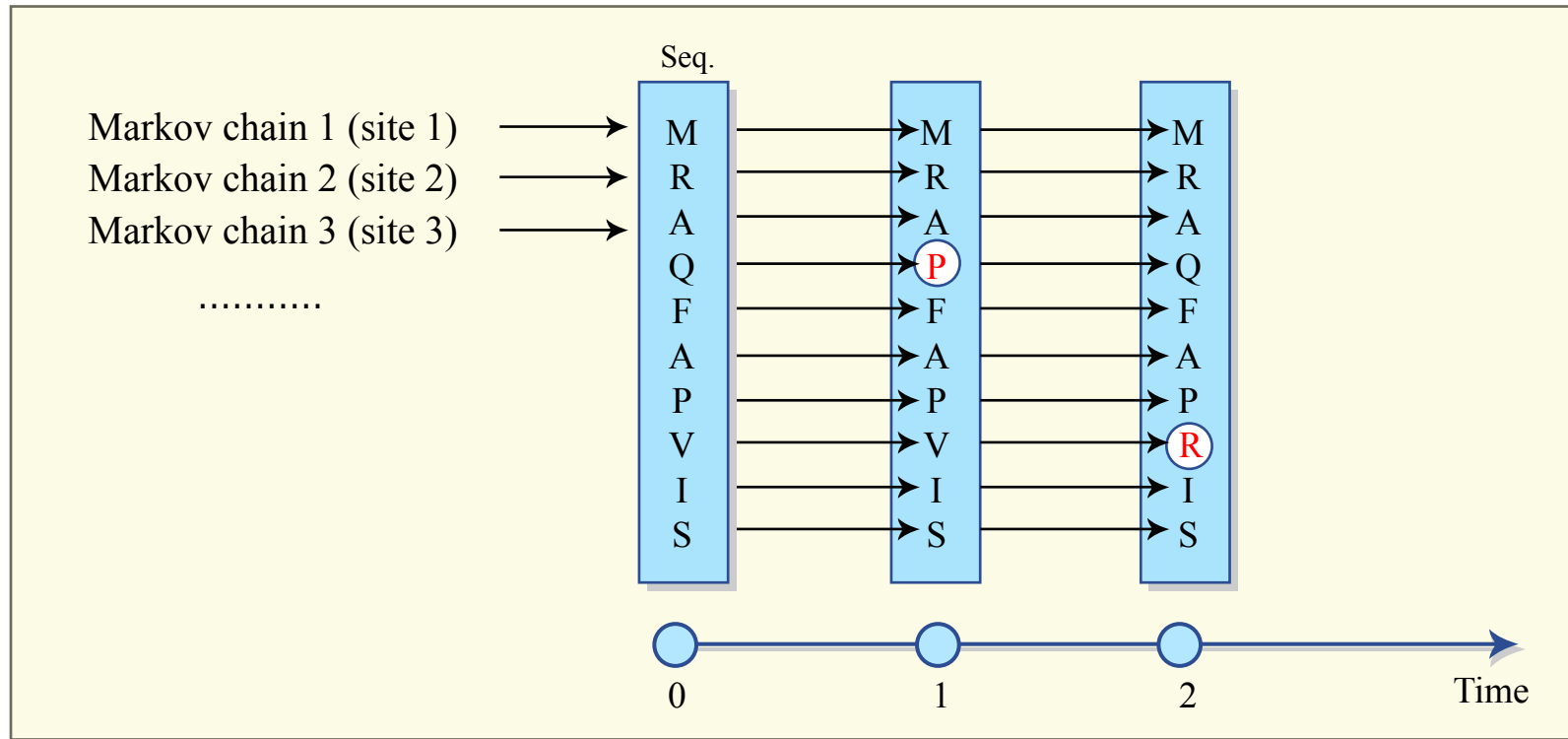
Necessary: $\alpha < 1/3$.

The n -step transition probabilities can be computed:

$$\mathbf{P}(X_n = i | X_0 = i) = 0.25 + 0.75 \cdot (1 - 4\alpha)^n, \text{ for } i \in \{a, c, g, t\}.$$

$$\mathbf{P}(X_n = j | X_0 = i) = 0.25 - 0.25 \cdot (1 - 4\alpha)^n, \text{ for } i, j \in \{a, c, g, t\}, i \neq j.$$

Underlying Model: Each site in the sequence evolves according to a Markov chain, and independently of the other sites.



All the Markov chains have the same transition matrix P (matrix with dimension 20×20).

Figure by MIT OCW.

FROM TRANSITION MATRIX TO ALIGNMENT SCORES

Two hypothesis:

1. Sequences S1 and S2 are unrelated (=random matching)
2. Sequences S1 and S2 have a common ancestor.

$$\text{Score} = \text{Log} (P1/P2)$$

P1 - probability of observed alignment given model 1

P2 - probability of observed alignment given model 2

Homology

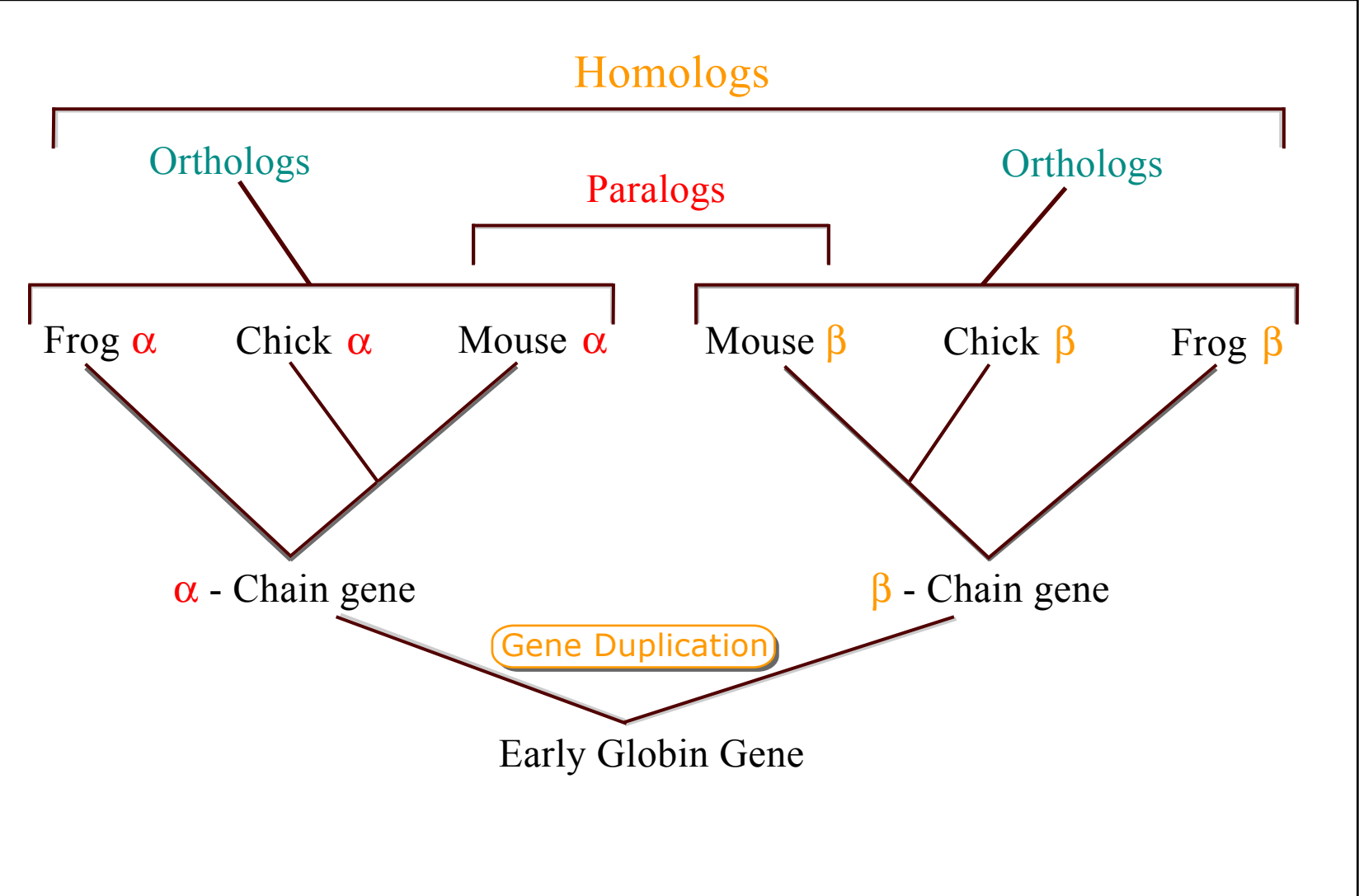


Figure by MIT OCW.

Guest lecture 1

- **Richard Lewontin**

- evolutionary geneticist

- philosopher of science

- social critic

- numerous publications including, "The Spandrels of San Marco", "The Genetic Basis of Evolutionary Change", "Biology as Ideology", "The Triple Helix: Gene, Organism, and Environment" ...

- <http://hrst.mit.edu/hrs/evolution/public/profiles/lewontin.html>

- <http://www.nybooks.com/authors/4463>

(Compilation of information by MIT OCW.)

Module 3

3. Structural genomics and proteomics

- Overview of [protein structures](#), [domain architecture](#). Sequence-structure mapping, protein folding, forces and interactions.
- Structure-based substitution matrices. Protein structure prediction. Threading.
- Protein function: binding and kinetics. Michaelis-Menten kinetics, inhibition. Protein-DNA recognition: models and algorithms.
- Proteomics: networks of protein-protein interactions, complexes, modules. Power-law distributions, clustering coefficient. Evolution of networks.
- Medical Lecture Hemoglobin and the anemias.

Module 3

3. Structural genomics and proteomics

- Overview of [protein structures](#), [domain architecture](#). Sequence-structure mapping, protein folding, **forces and interactions**.
- Structure-based substitution matrices. Protein structure prediction. Threading.
- Protein function: binding and kinetics. Michaelis-Menten kinetics, inhibition. Protein-DNA recognition: models and algorithms.
- Proteomics: networks of protein-protein interactions, complexes, modules. Power-law distributions, clustering coefficient. Evolution of networks.
- Medical Lecture Hemoglobin and the anemias.

ELECTRO + SOLVENT :

Dielectric effect


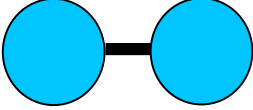
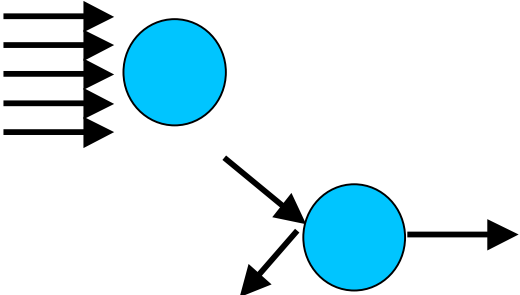
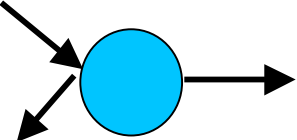
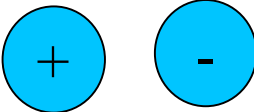
$$V = \frac{q_i q_j}{4\pi \epsilon r_{ij}}; \quad \epsilon = 80$$

Figure removed due to copyright considerations.

2-3 Kcal/mol
 $\Delta G \sim T$

Linear in T
 \Rightarrow entropic!

More forces:

Elastic		1-100 pN
Covalent		10^5 pN
Viscous		1-1000 pN
Collisional		10^{-12} - 10^{-9} pN for 1 collision/s
Thermal		100-1000 pN
Gravity		10^{-9} pN
Electrostatic and VdW		1-1000 pN
Magnetic		$\ll 10^{-6}$ pN

Hydrophobic effect

Frank & Evans 1945

- Water molecules form hydrogen bonds
- Polar groups do not disturb the network of water-water interactions.
- Non-polar (hydrophobic) groups disrupt the network leading to formation of “local ordering” of water.
- Local ordering **reduces the entropy**

Figure removed due to copyright reasons.

Please see Figure 2 in:
Laidig, Keith E., and Valerie Daggett. "Testing the Modified Hydration-Shell Hydrogen-Bond Model of Hydrophobic Effects Using Molecular Dynamics Simulation." *J Phys Chem* 100 (1996): 5616-5619.

Module 3

3. Structural genomics and proteomics

- Overview of [protein structures](#), [domain architecture](#). Sequence-structure mapping, protein folding, forces and interactions.
- Structure-based substitution matrices. Protein structure prediction. Threading.
- Protein function: binding and kinetics. Michaelis-Menten kinetics, inhibition. **Protein-DNA recognition: models and algorithms.**
- Proteomics: networks of protein-protein interactions, complexes, modules. Power-law distributions, clustering coefficient. Evolution of networks.
- Medical Lecture Hemoglobin and the anemias.

RECOGNITION OF BINDING MOTIFS IN DNA

Figures removed due to copyright reasons.

RECOGNITION OF BINDING MOTIFS IN DNA 2

HOMEODOMAIN

Figures removed due to copyright reasons.

Module 3

3. Structural genomics and proteomics

- Overview of [protein structures](#), [domain architecture](#). Sequence-structure mapping, protein folding, forces and interactions.
- Structure-based substitution matrices. Protein structure prediction. Threading.
- Protein function: binding and kinetics. Michaelis-Menten kinetics, inhibition. Protein-DNA recognition: models and algorithms.
- **Proteomics: networks of protein-protein interactions, complexes, modules.** Power-law distributions, clustering coefficient. Evolution of networks.
- Medical Lecture Hemoglobin and the anemias.

Examples of biological networks

- Protein-protein interactions
proteins ~5000
interactions ~7000

Figures removed due to copyright reasons.

Databases:

BindDB

MIPS

DIP

A comprehensive analysis of protein-protein interactions in *S.cerevisiae*.

Nature. 2000, 623-7. Uetz P et al

Examples of biological networks

- Protein-DNA interactions (TF-upstream binding)

Figures removed due to copyright reasons.

Please see Figure 5 in T. I., Lee, et al. "Transcriptional regulatory networks in *Saccharomyces cerevisiae*." *Science* 298, no. 5594 (Oct 25, 2002): 799-804.

Metabolic Pathways

Figure removed due to copyright considerations.

Module 3

3. Structural genomics and proteomics

- Overview of [protein structures](#), [domain architecture](#). Sequence-structure mapping, protein folding, forces and interactions.
- Structure-based substitution matrices. Protein structure prediction. Threading.
- Protein function: binding and kinetics. Michaelis-Menten kinetics, inhibition. Protein-DNA recognition: models and algorithms.
- Proteomics: networks of protein-protein interactions, complexes, modules. **Power-law distributions**, clustering coefficient. Evolution of networks.
- Medical Lecture Hemoglobin and the anemias.

it's not a random graph!

Figures removed due to copyright reasons.

Please see figures 1e and 2 in Jeong, H., B. Tombor, R. Albert, Z. N. Oltvai, and A. L. Barabasi.
"The large-scale organization of metabolic networks." *Nature* 407, no. 6804 (Oct 5, 2000): 651-4.

IT'S ALMOST SCALE-FREE (=POWER-LAW) GRAPH

Figures removed due to copyright reasons. Please see:

Liljeros, F., et al. "Distributions of number of sexual partnerships have power law decaying tails and finite variance." eprint ARXIV, (<http://arxiv.org/>). (May 2003).

and

Liljeros F., et al. "The web of human sexual contacts." *Nature* 411, no. 6840 (Jun 21, 2001): 907-8.

Module 3

3. Structural genomics and proteomics

- Overview of [protein structures](#), [domain architecture](#). Sequence-structure mapping, protein folding, forces and interactions.
- Structure-based substitution matrices. Protein structure prediction. Threading.
- Protein function: binding and kinetics. Michaelis-Menten kinetics, inhibition. Protein-DNA recognition: models and algorithms.
- Proteomics: networks of protein-protein interactions, complexes, modules. **Power-law distributions**, clustering coefficient. Evolution of networks.
- Medical Lecture Hemoglobin and the anemias.

Module 4

4. Functional Genomics and Networks

- Gene regulation and function, conservation, detecting regulatory elements.
- RNA expression: clustering and classification.
- RNA expression: classification, 2-way clustering, regulatory modules. Integration of expression and proteomic data.
- **Dynamics of biological networks metabolic, regulatory.** FBA, signaling, regulation of gene expression.
- Medical Lecture: Two examples: phenylketonuria (monogenic) and diabetes type 2 (multigenic+). “Disease” genes vs. “susceptibility” genes. “Environmental” vs. “Developmental” regulation of gene expression.

cDNA microarray expt

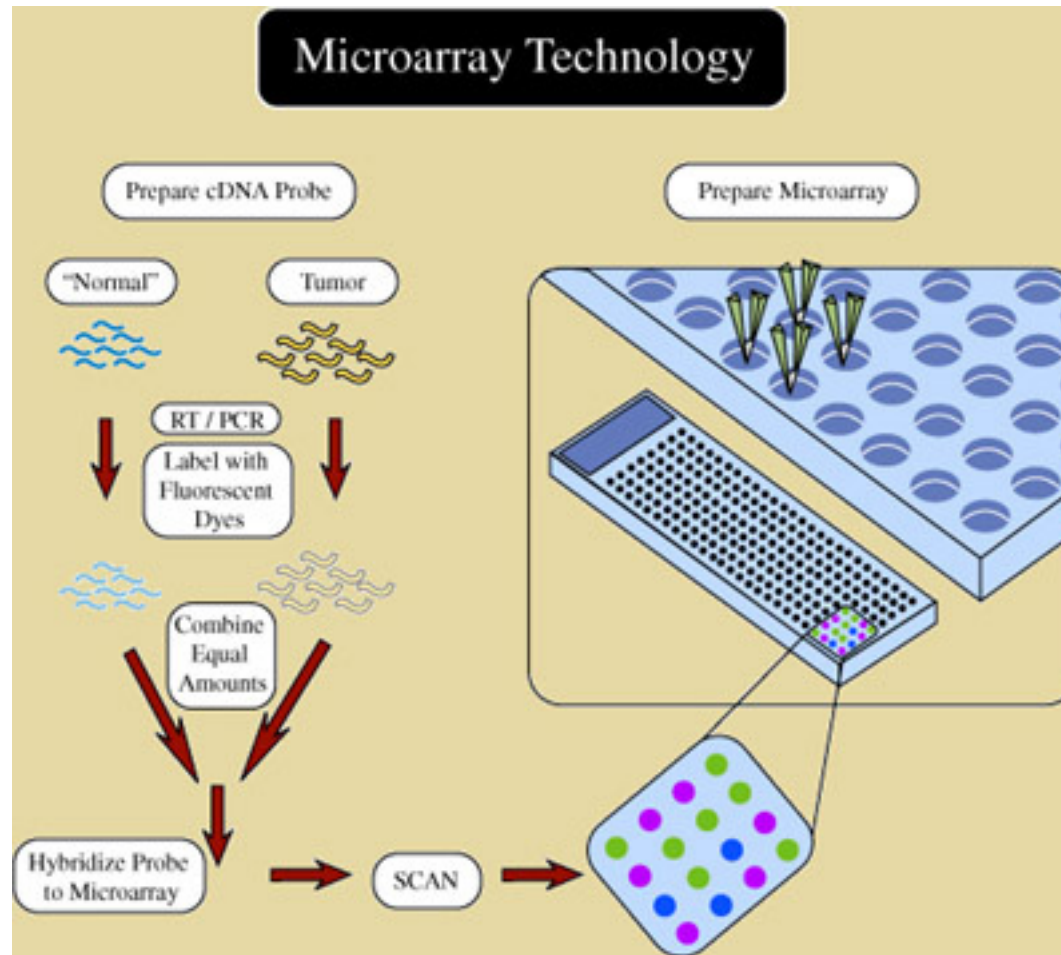


Figure by MIT OCW.

HYBRIDIZATION, STAINING

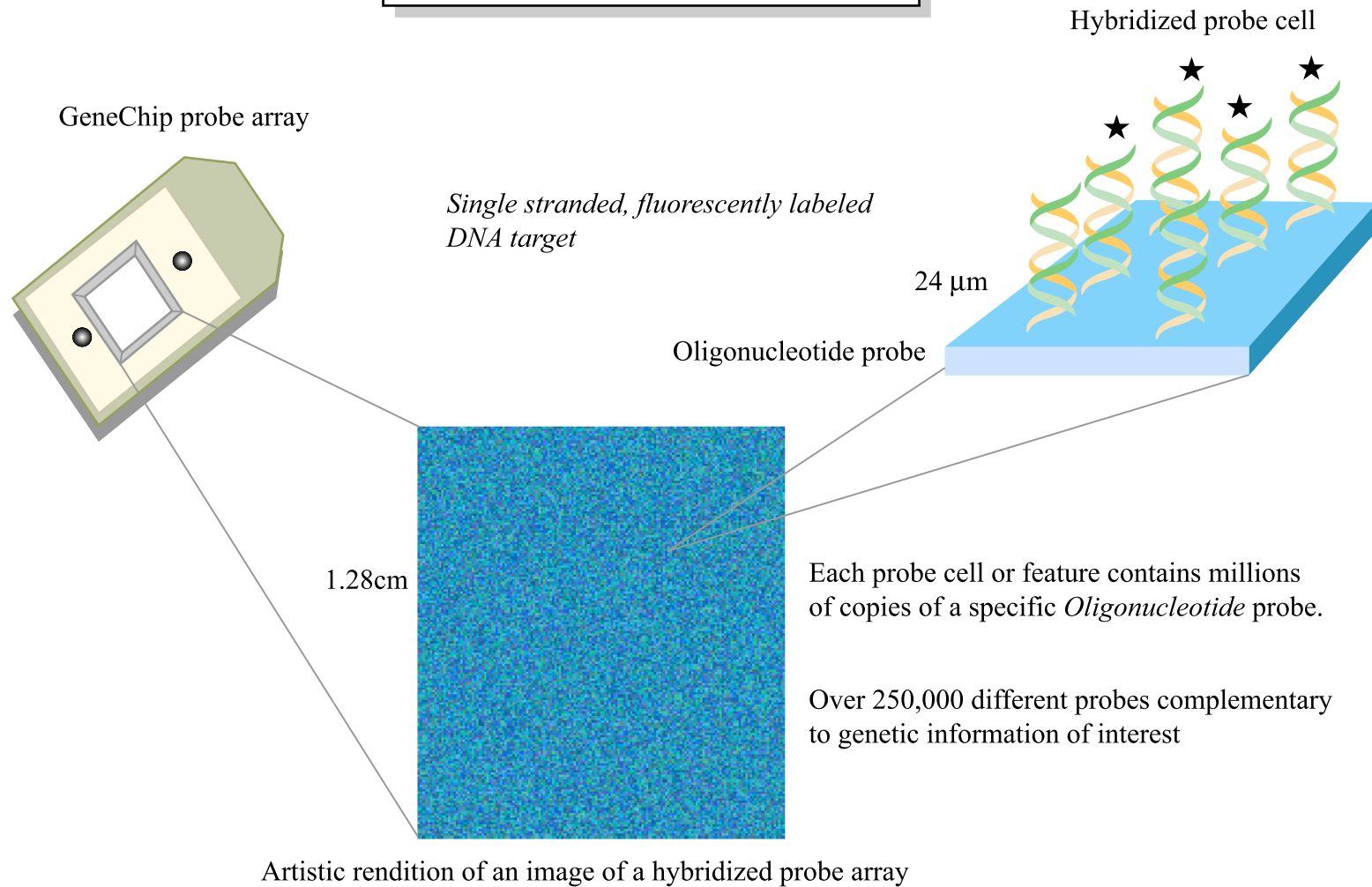


Figure by MIT OCW.

Figure removed due to copyright reasons.

DEFINITION OF THE CLUSTERING PROBLEM

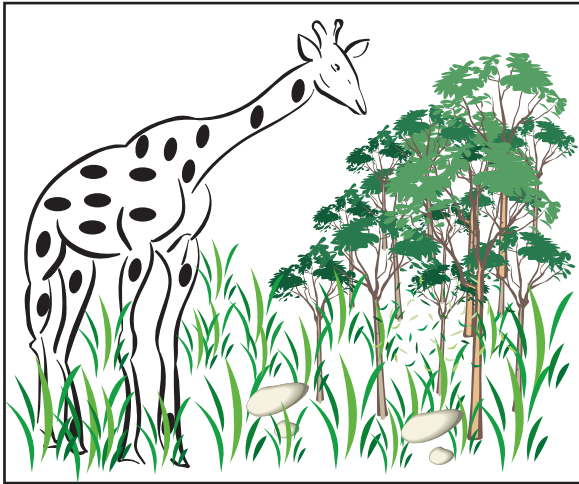


Figure by MIT OCW.

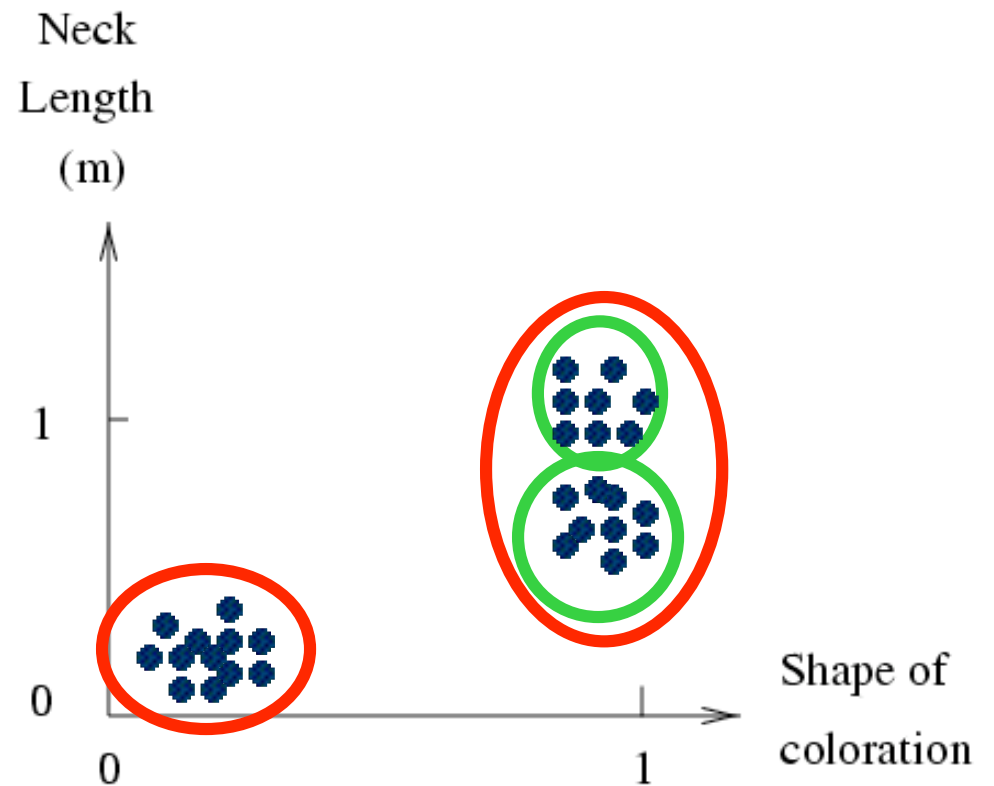
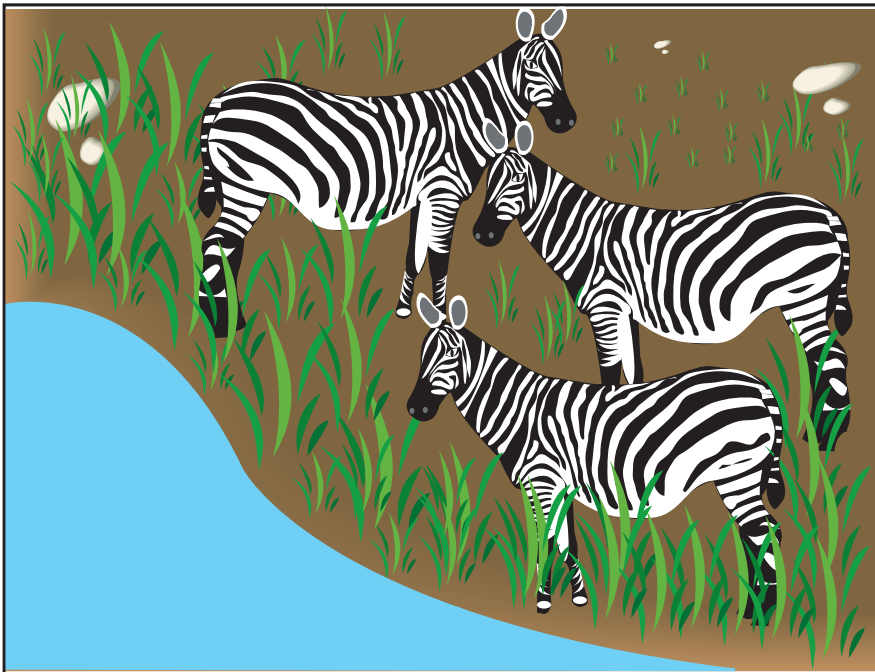
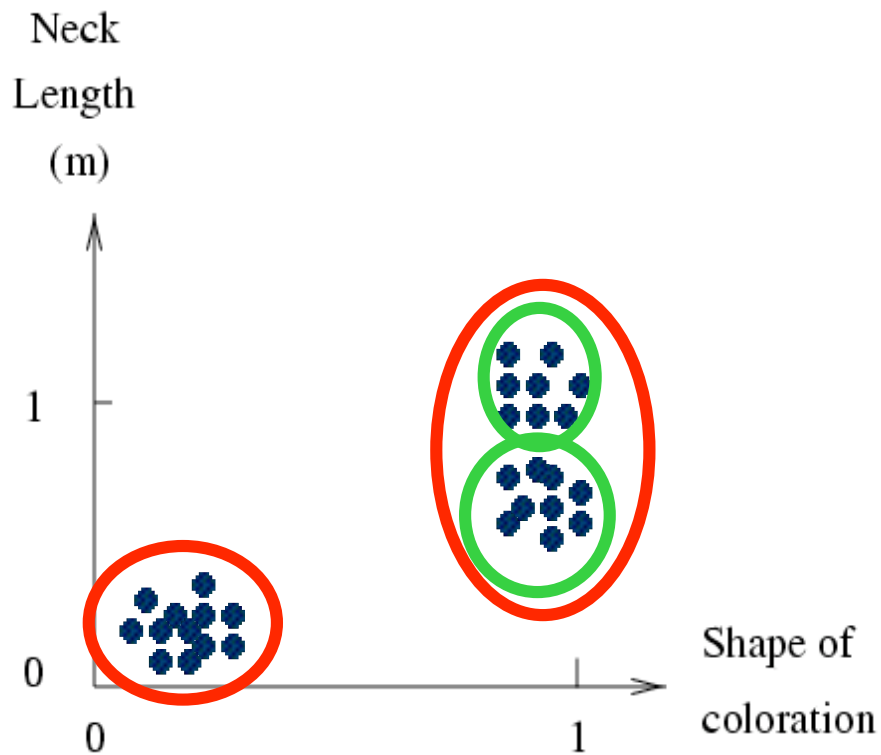
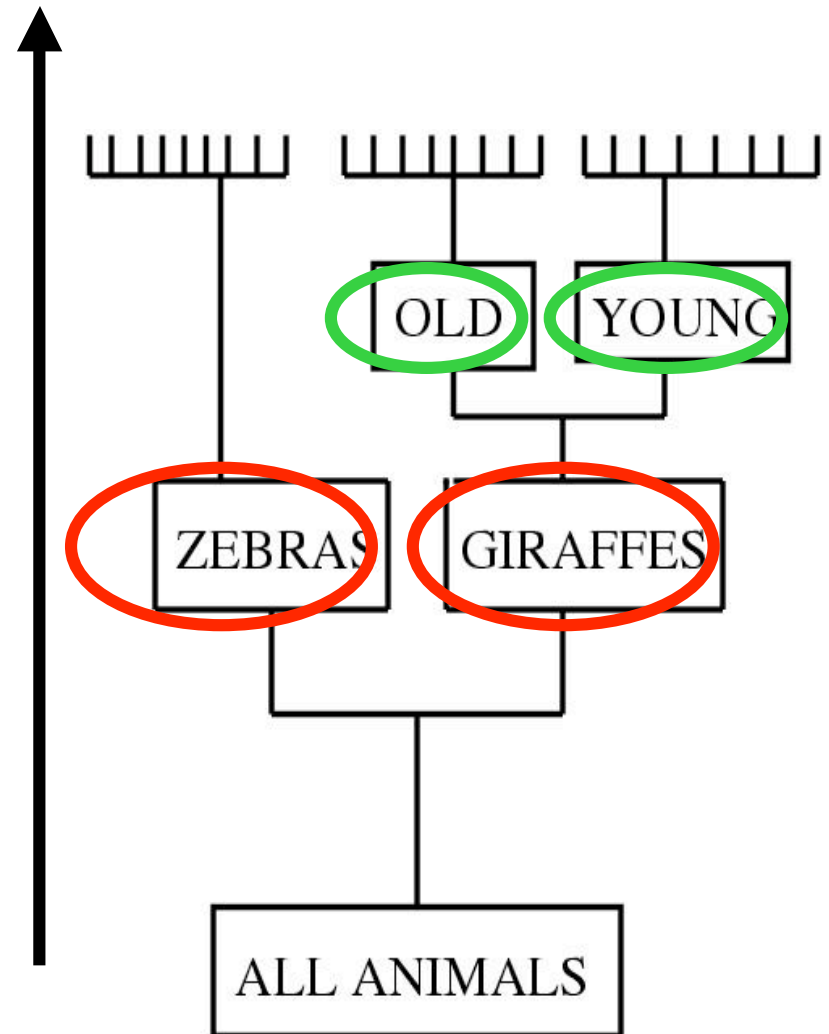


Figure by MIT OCW.

CLUSTER ANALYSIS YIELDS DENDROGRAM



T (RESOLUTION)



BUT WHAT ABOUT THE OKAPI?

Figure by MIT OCW.

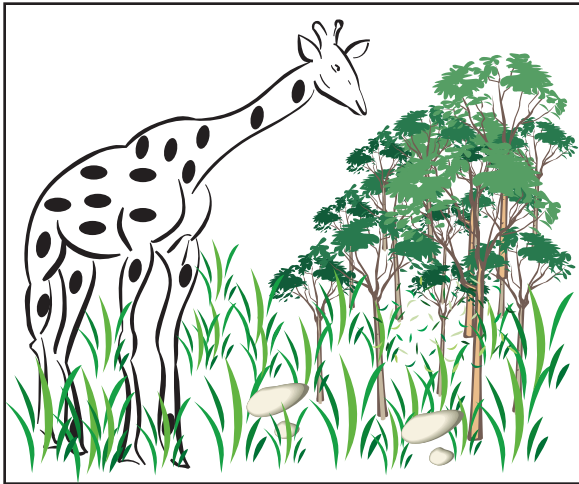


Figure by MIT OCW.

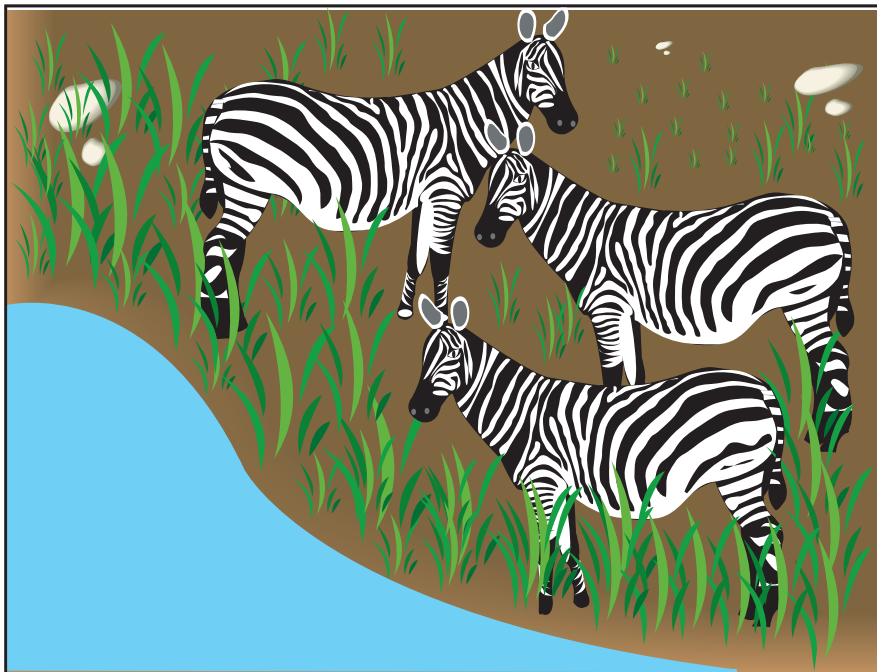
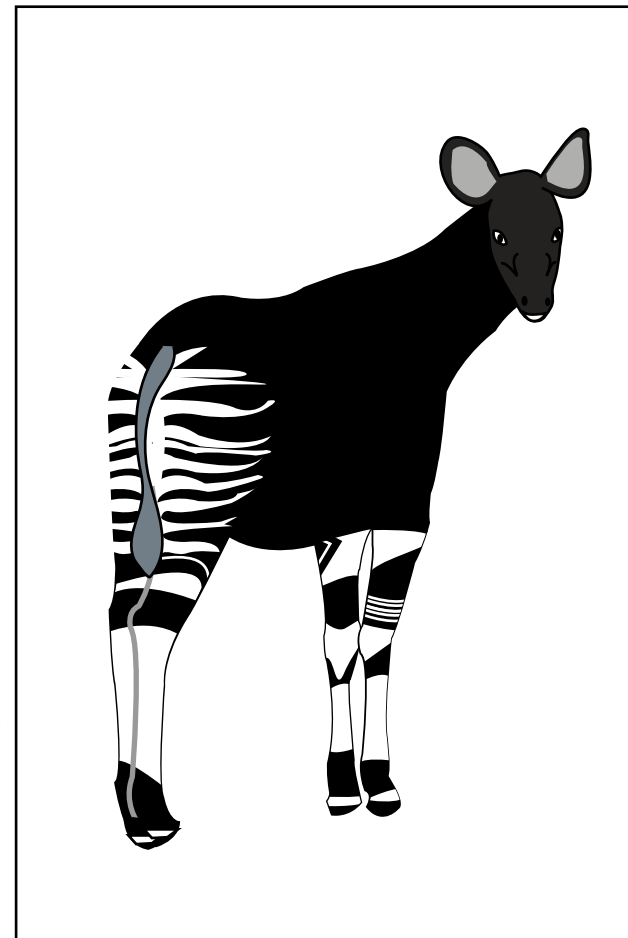
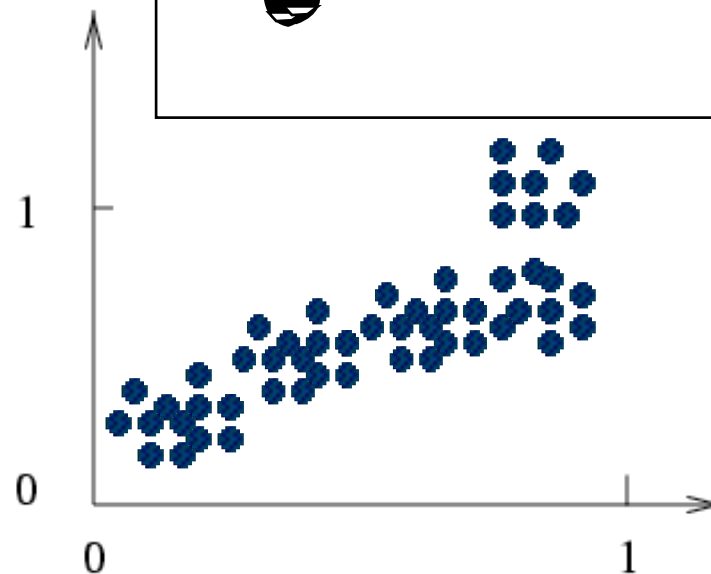


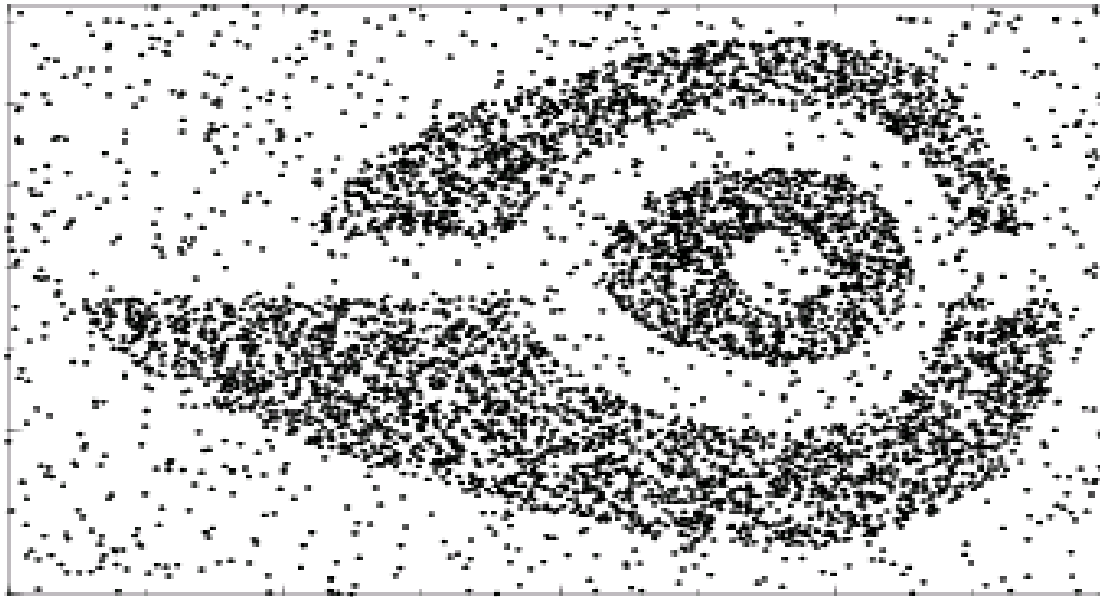
Figure by MIT OCW.



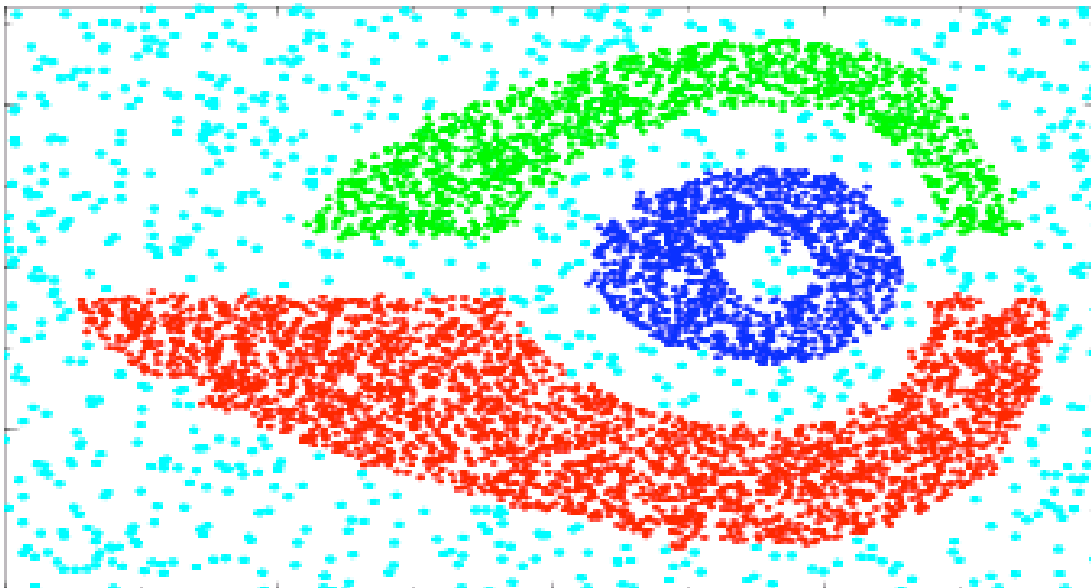
Neck
Length
(m)



Shape of
coloration



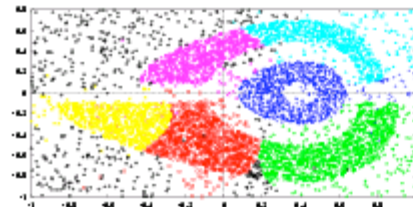
how many clusters?



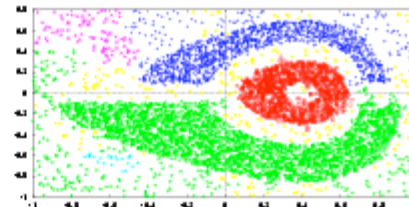
3 **LARGE**
MANY small
(SPC)

Figures derived from Blatt, M., S. Wiseman, and E. Domany. "Superparamagnetic Clustering of Data." *Phys Rev Lett* 76, no. 18 (1996): 3251.

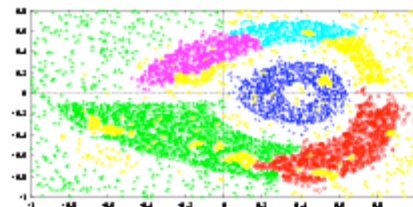
OTHER METHODS



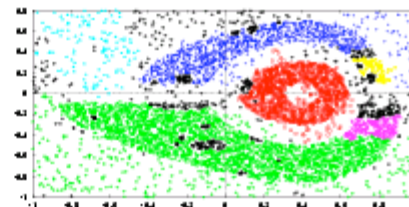
Valley seeking (Fukunaga)



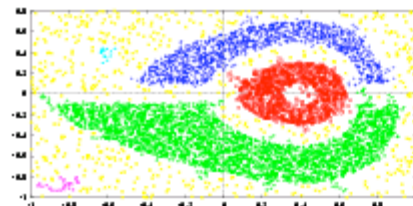
Minimal spanning tree (Zhan)



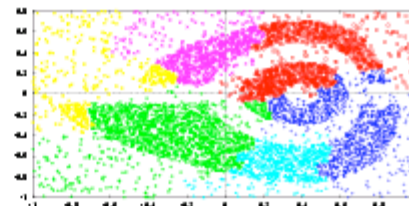
K shared neighbors (Jarvis)



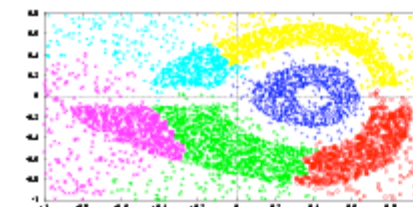
Mutual neighborhood (Gowda)



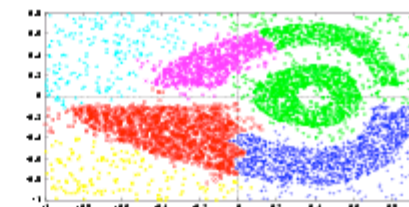
Single linkage method (Milligan, Dekel)



Complete linkage method (Milligan)



minimum variance (Ward)



arithmetic averages (Sokal)

Figures derived from Blatt, M., S. Wiseman, and E. Domany. "Superparamagnetic Clustering of Data." *Phys Rev Lett* 76, no. 18 (1996): 3251.

Functionally related genes

Image removed due to copyright reasons.

CENTRAL DOGMA

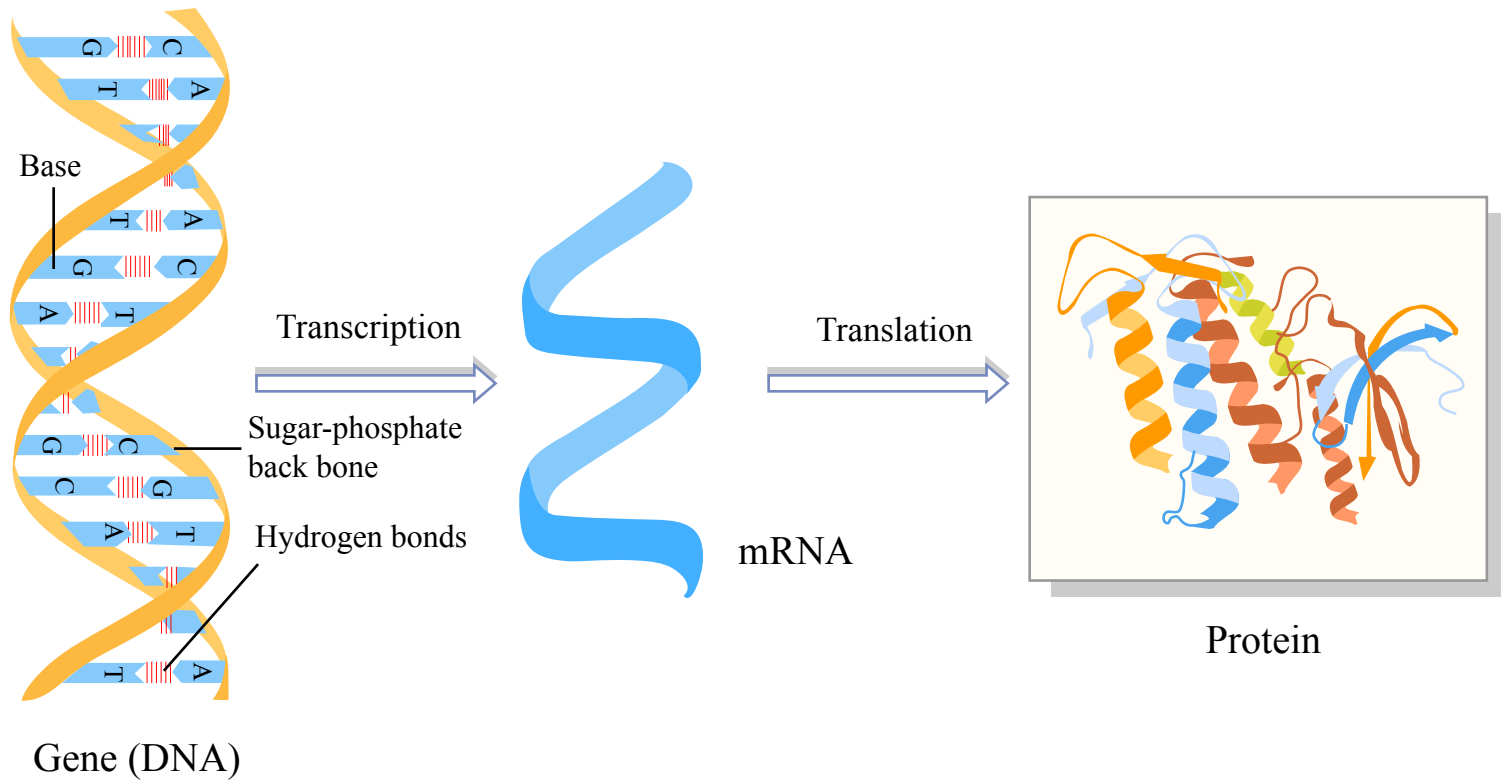


Figure by MIT OCW.

HST.508 Cases

Case 1: Dr. Gail Genomous

- Took HST.508 in 2005
- Since 2008 works at R&D of Shrek Pharmaceuticals
- She works on target identification for mysterious green-rash syndrome.

- Pathways involved are know.
But candidate drugs
although binding to their targets
have been inefficient
on model organisms.
- Her goals are
 - Suggest better target proteins
 - Assist drug design/organic syntheses group in drug development

Case 1: Dr. Gail Genomous

- She found that involved drug targets are enzymes of well-known metabolic pathway. She sets up flux-balance simulations to study this pathways (**Lec F4**).
- Simulations suggest that inhibition of targeted enzymes does not shutoff the pathway. Dr.Genomous identifies other enzymes that need to be inhibited to shutoff the pathway. She suggests that down-regulating these enzymes is the most efficient intervention strategy.

Case 1: Dr. Gail Genomous

- She finds out that these enzymes are co-expressed (**Lec F1**), but transcription factor is unknown.
- However, available protein-protein interactions suggest that these enzymes are also activated by a kinase pi314 (**Lec S4**).
- Dr.Genomous model the structure of pi314 by homology to another kinase (**Lec S1**).
- Comparison of pi314 with its orthologs from related species (*P.Winnie*, *D.Scooby* etc) suggests functional region of the structure (**Lec C2,S2**).
- Dr. Genomous proposes drug design group the new drug target and a specific functional region to be targeted.

Case 2: Dr. Pete Proteomson

- Took HST.508 in 2005
- Works on his dissertation at the Whitetail institute
- He is interested in tail discoloration syndrome (TDS).
- In collaboration with the hospital, his lab has performed expression profiling of patients and normal individuals.
- His goals are
 - Identify genes involved
 - Suggest and test mechanisms of the disease

Case 2: Dr. Pete Proteomson

- Dr Proteomson has detected differentially expressed genes (**Lecture F2**)
- Mapped differentially expressed genes on the network of protein-protein interactions (**Lec. S3**) and network of synthetically lethal genes. Results suggested that most of these genes belong to the same pathway (**Lec S4**).
- Using homology, genomic and network location, he predicted function for some of these genes as protein kinases involved in cell signaling (**Lec S4**).

Case 2: Dr. Pete Proteomson

- Mapping selected gene onto a database of SNPs revealed rare polymorphisms in some of these genes (**Lec C4**).
- Population analysis suggests that these mutations are likely to be deleterious (**Lec E3**). Mapping mutations on known structures of protein kinases and cross-species comparison strongly supported deleterious nature of some SNPs (**Lec S2,C4**).
- Finally, Dr. Proteomson demonstrated that patients carry more deleterious SNPs in the identified pathway than normal individuals. This result supports the role of identified pathways in development of the green rash.