CS 68: BIOINFORMATICS

Prof. Sara Mathieson
Swarthmore College
Spring 2018

Outline: May 4

May the fourth be with you…

- Groups 5&6 from last time
- Disease biology beyond GWAS
- Secondary structure prediction

Notes:

-Final presentation guidelines posted
-Let me know if you would like to meet today or next week
-Evaluations will be online
-If I get 100% response rate before the presentations, I will bring snacks

Project Presentation Notes

- Date: Thursday, May 17, 2-5pm (in our classroom)
- Each person will have 4 minutes to present (including questions, so aim for around 3:30)
- Email me your slides by 1pm on May 17! (PowerPoint or PDF okay)
- I will have a laser pointer / slide advancer clicker

Project Presentation Notes

Your presentation should include

- Motivation and Scientific Question
- Data and Methods
- Results and Interpretation
- Conclusions and Future Work

Project Presentation Notes

Your presentation should include

- Motivation and Scientific Question
- Data and Methods
- Results and Interpretation
- Conclusions and Future Work

Presentation Tips

- Speak loudly (to the back of the class)
- Avoid text-heavy slides, use images/diagrams
- Include citations for any figures you did not make
- Ask at least one question to another group

Project Presentation Notes

Your presentation should include

- Motivation and Scientific Question
- Data and Methods
- Results and Interpretation
- Conclusions and Future Work

Presentation Tips

- Speak loudly (to the back of the class)
- Avoid text-heavy slides, use images/diagrams
- Include citations for any figures you did not make
- Ask at least one question to another group

Submit by 10pm on May 17

- Lab Notebook (include references)
- All project code
- Presentation slides

Project Presentation Notes

Your presentation should include

- Motivation and Scientific Question
- Data and Methods
- Results and Interpretation
- Conclusions and Future Work

Presentation Tips

- Speak loudly (to the back of the class)
- Avoid text-heavy slides, use images/diagrams
- Include citations for any figures you did not make
- Ask at least one question to another group

Submit by 10pm on May 17

- Lab Notebook (include references)
- All project code
- Presentation slides

Think about reproducibility!

Groups 5&6 from last time…

Other areas of Computational Biology

- **Biological Modeling**
  - *Drug entering the body*
  - *Tissue and surgical modeling*
  - *Gene networks*
  - *Intersects with computer vision, computer graphics, and graph theory*

- **Secondary and Tertiary Structure**
  - RNA secondary structure prediction
  - Protein folding

- **Neuroscience**
  - *Modeling the brain*

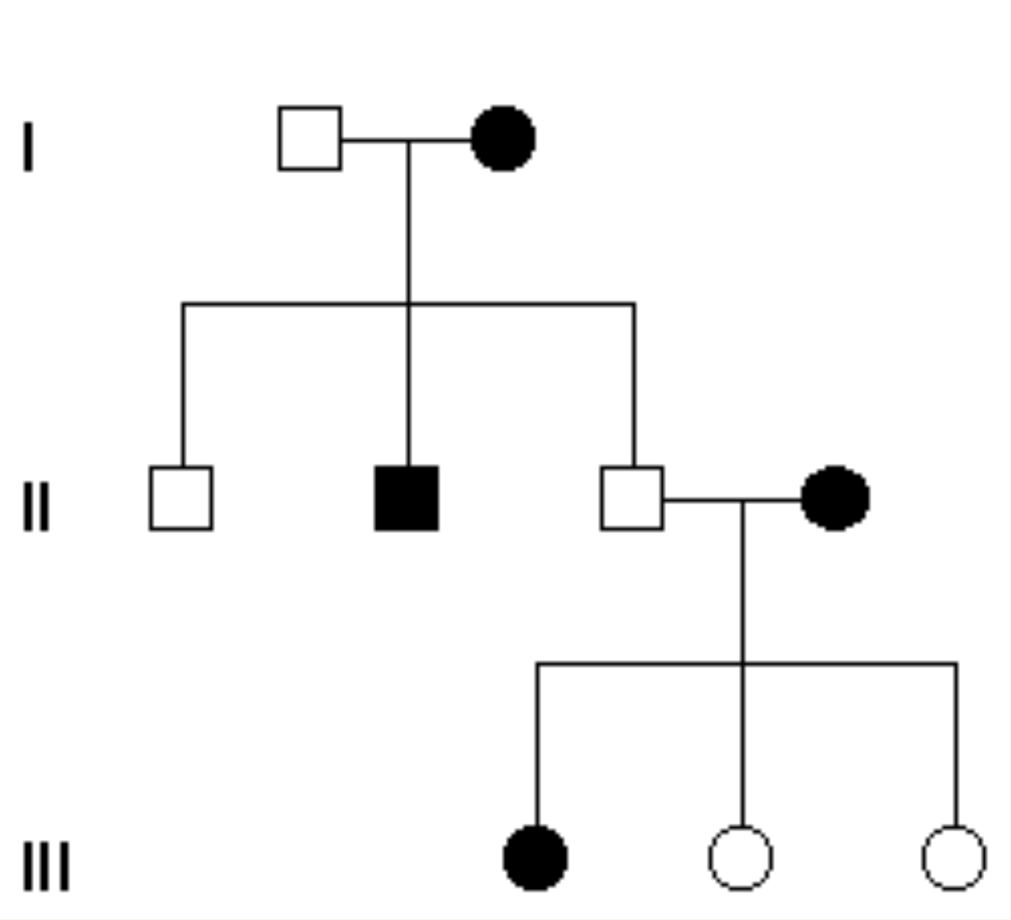
- **Disease biology**
  - *Pedigree analysis*
  - *Infectious disease models*
  - *Cancer biology*

Computational disease biology beyond GWAS

Pedigree Analysis

Beyond GWAS: pedigree analysis

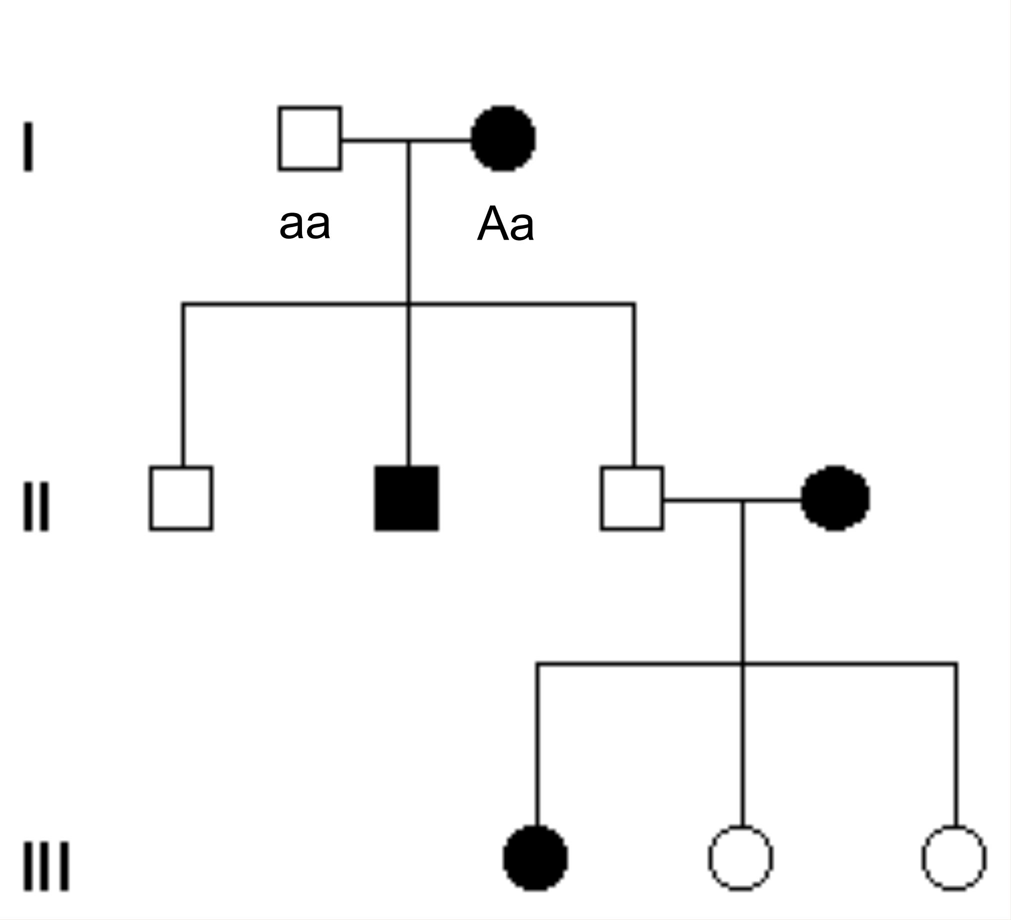
Dominant



Copyright © 2000. Phillip McClean

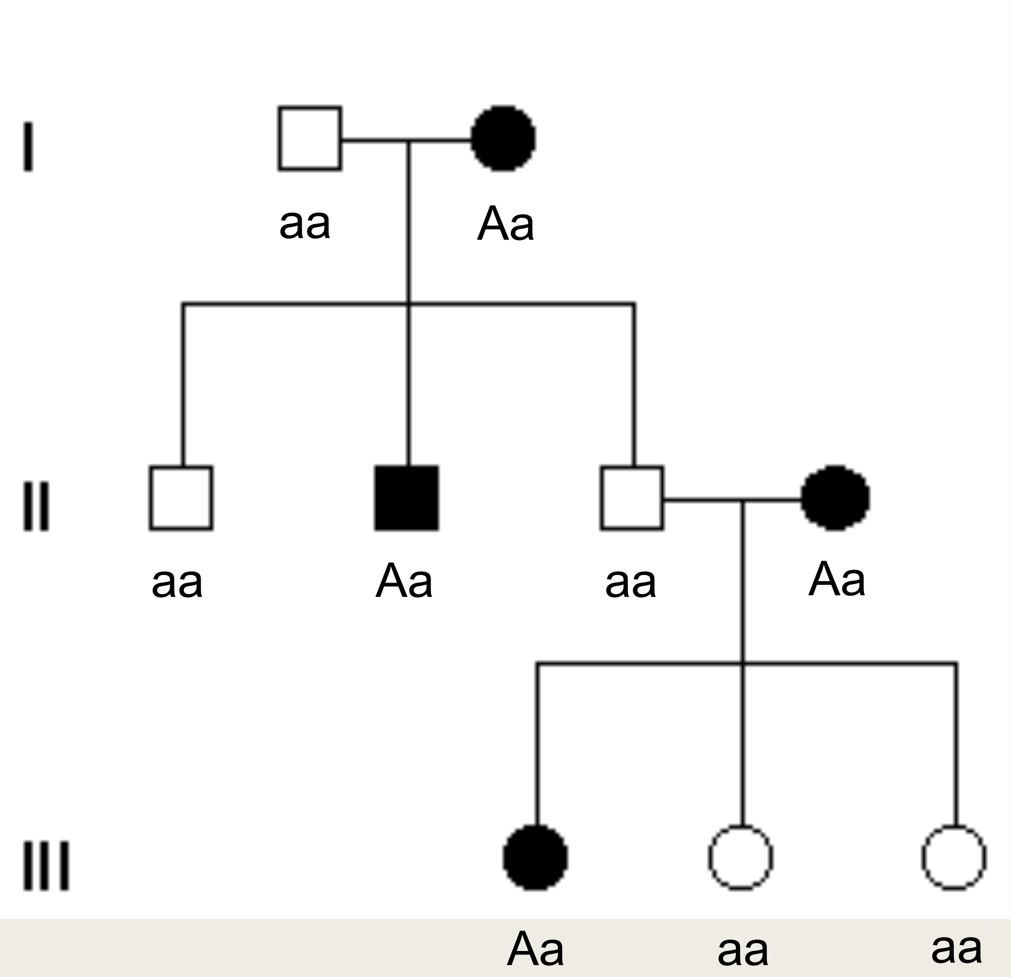
Beyond GWAS: pedigree analysis

Dominant



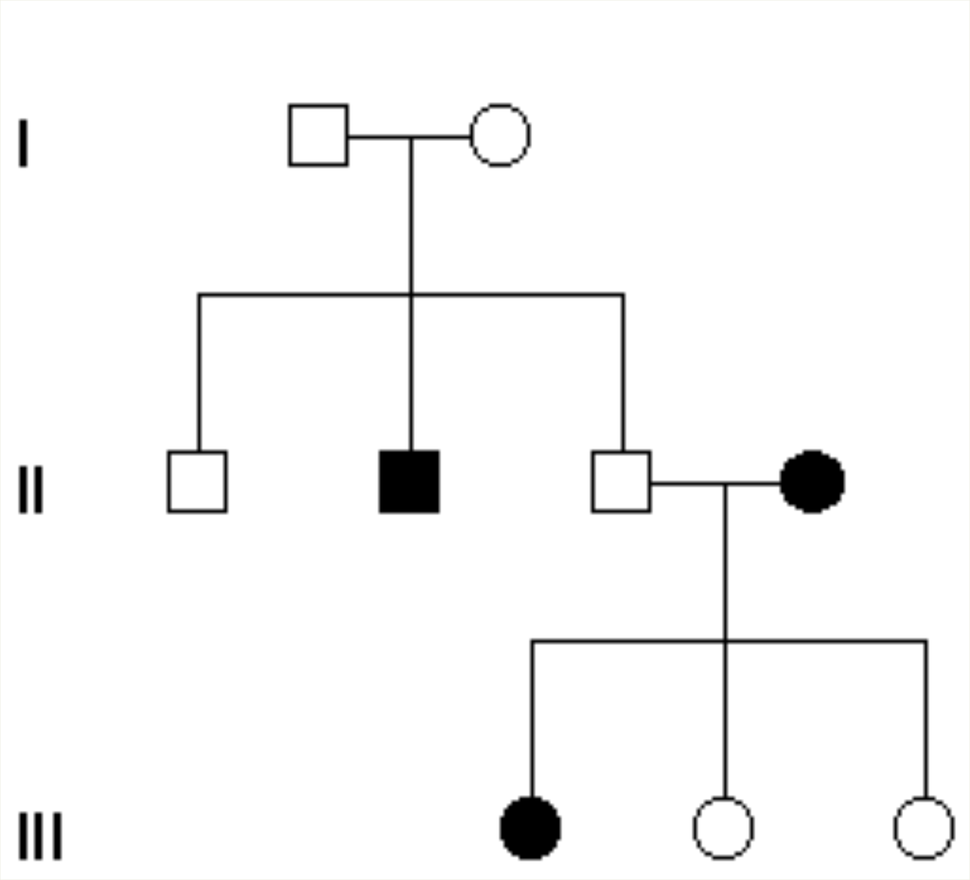
Beyond GWAS: pedigree analysis

Dominant



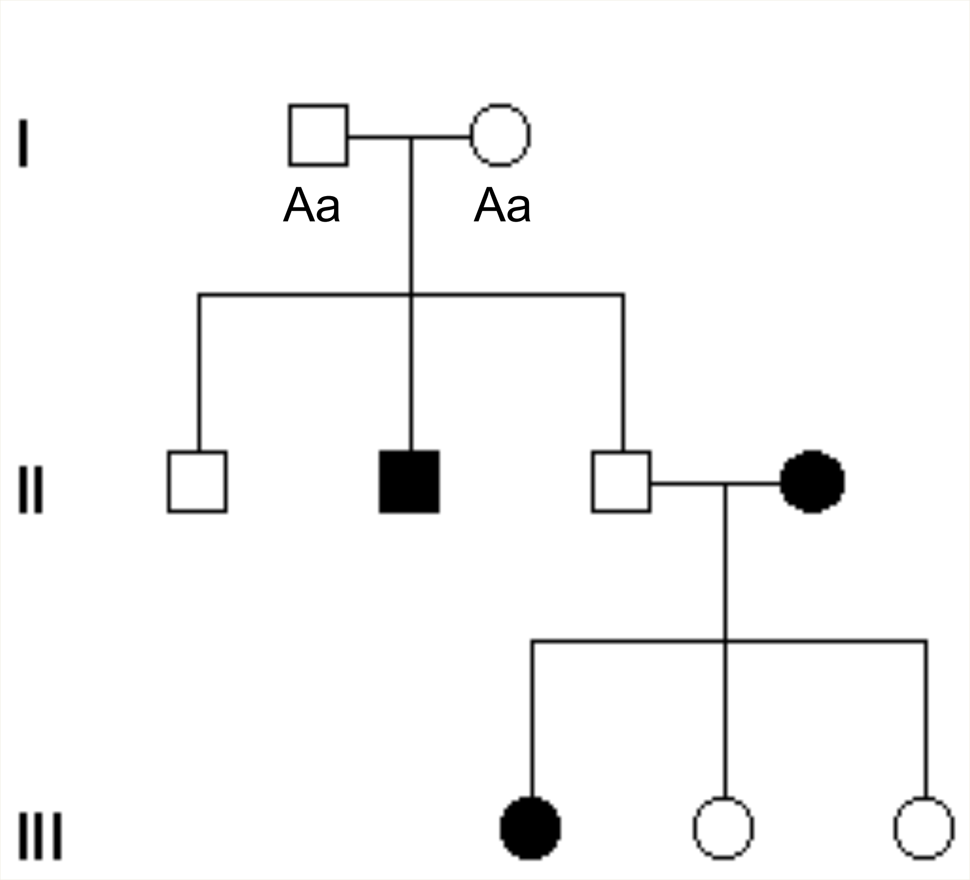
Beyond GWAS: pedigree analysis

Recessive



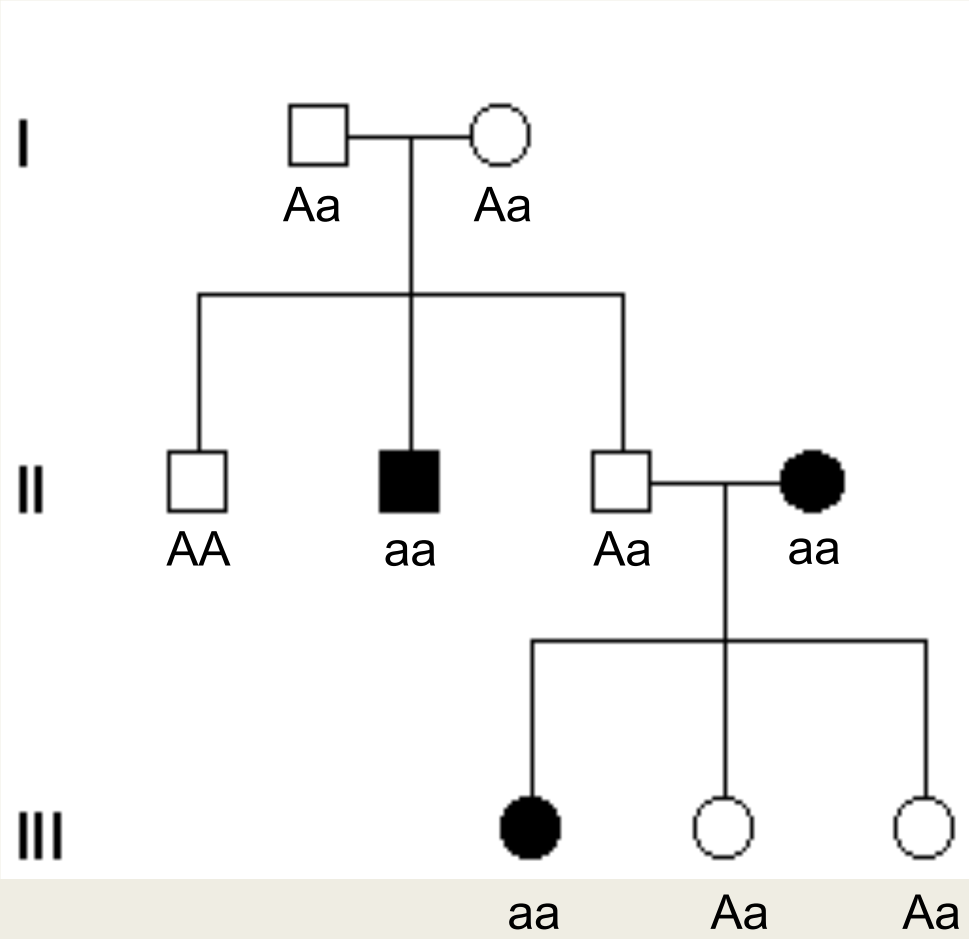
Beyond GWAS: pedigree analysis

Recessive



Beyond GWAS: pedigree analysis

Recessive



Ability to taste the chemical PTC

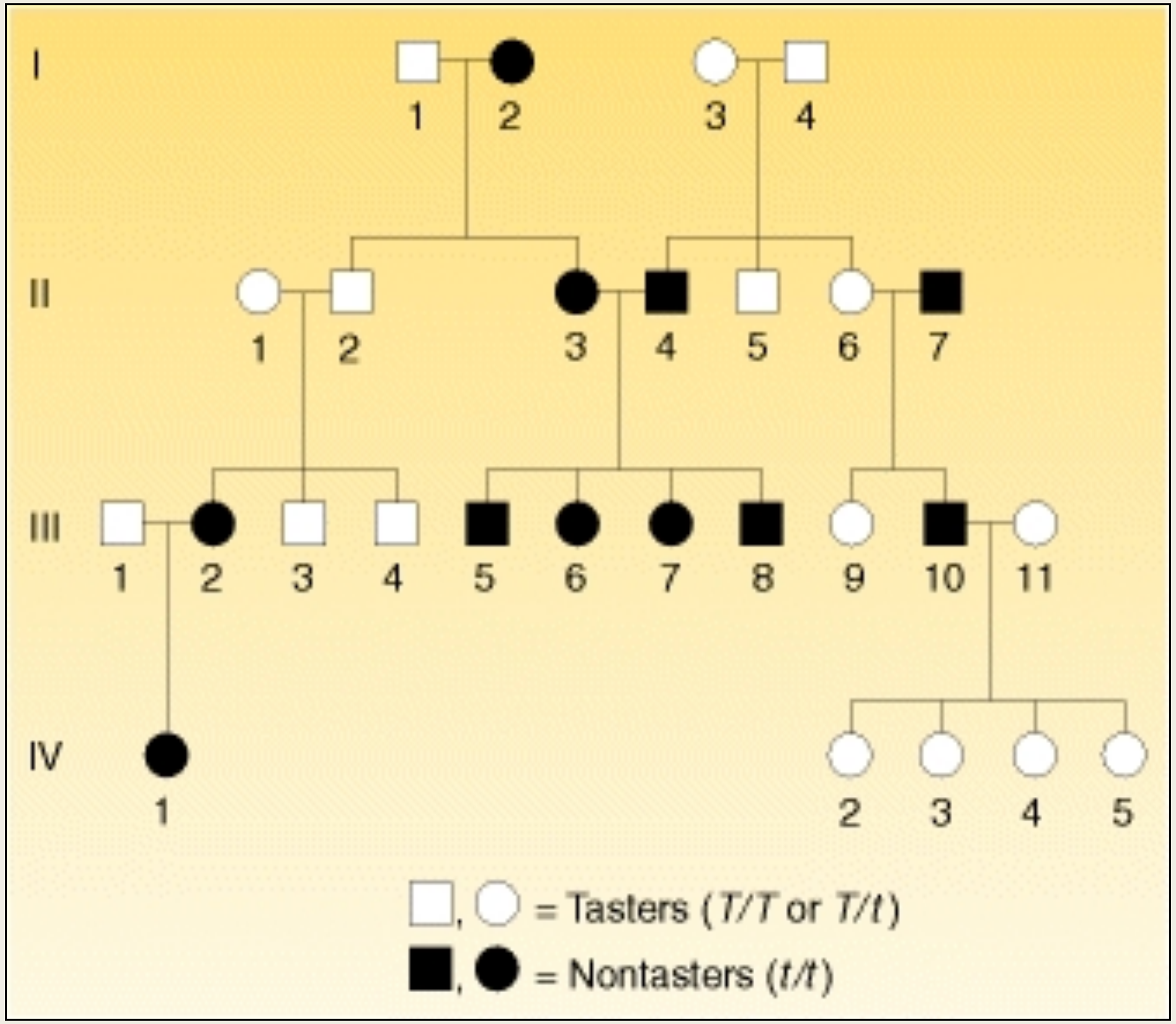

I
1
2
3
4
II
1
2
3
4
5
6
7
III
1
2
3
4
5
6
7
8
9
10
11
IV
1
2
3
4
5
□, ○ = Tasters (*T*/*T* or *T*/*t*)
■, ● = Nontasters (*t*/*t*)

Image: *Human Pedigree Analysis* (1999)

Hemophilia in the Royal Family – X linked

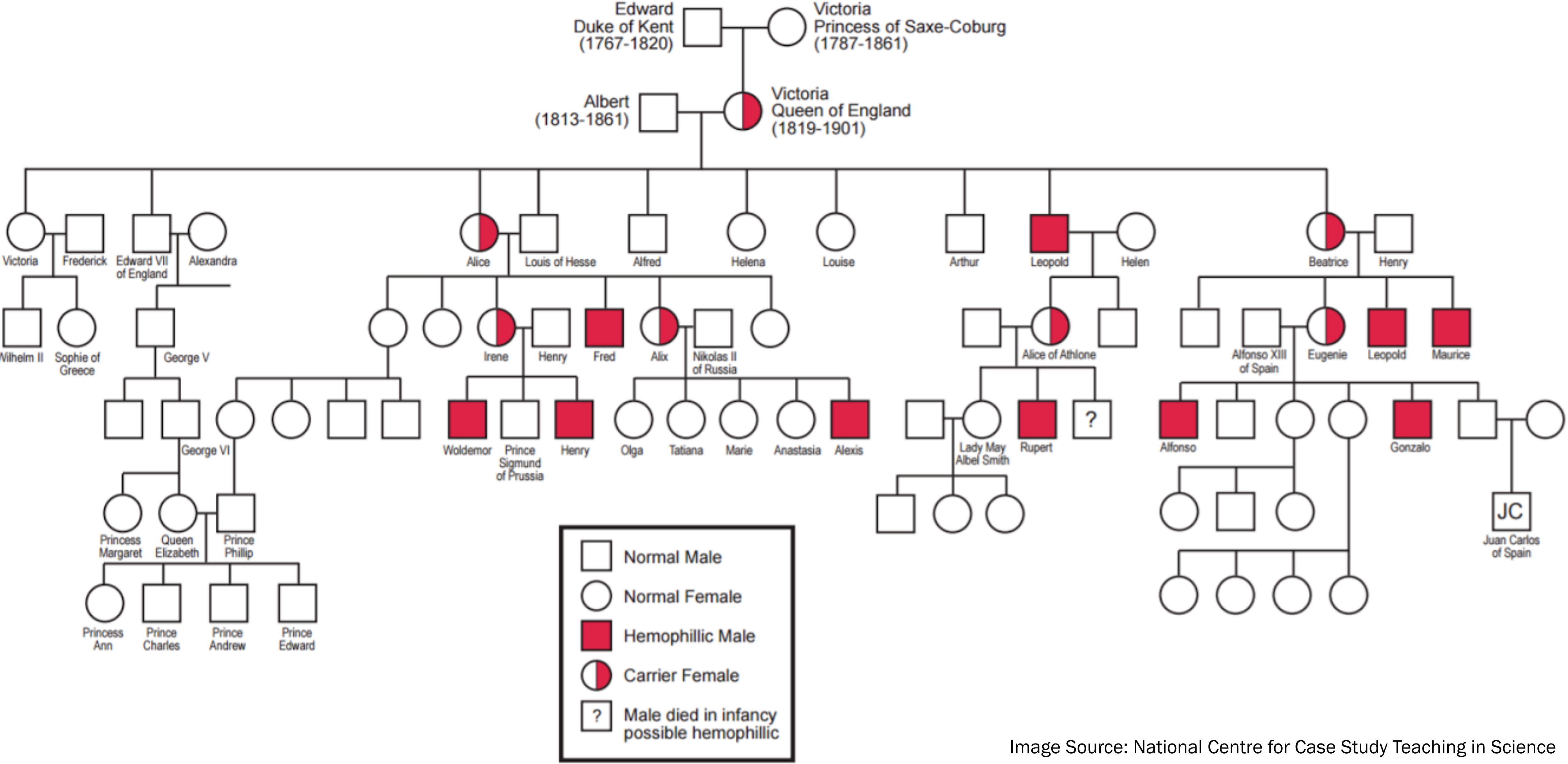


Image Source: National Centre for Case Study Teaching in Science

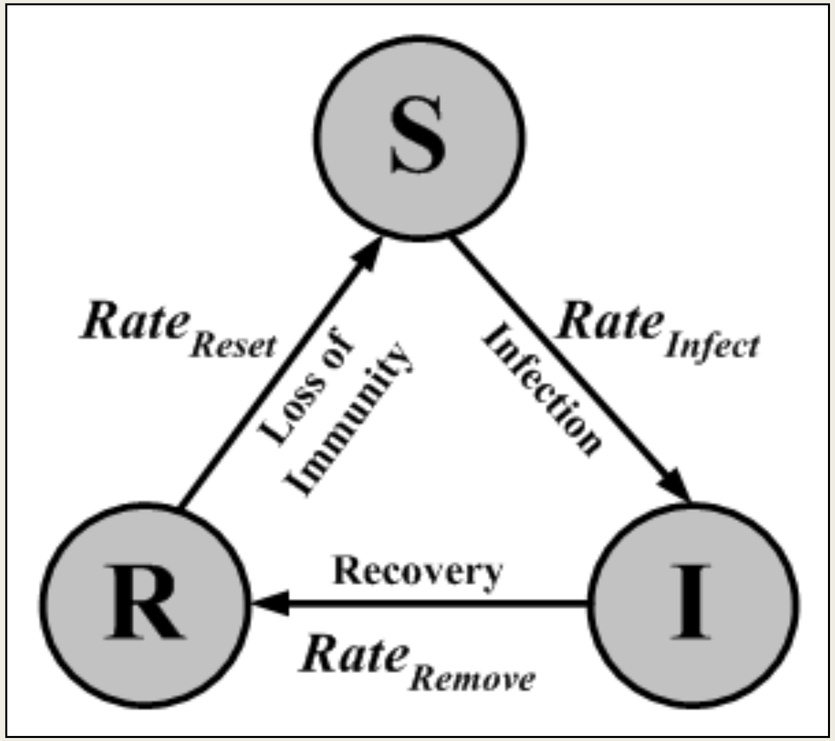
Infectious disease modeling

SIR models for infectious disease

- Recent applications:

  - *H1N1, “swine flu”, 2009*
  - *Ebola, 2015*

Susceptible



Recovered

Infected

“Influence of Local Information on Social Simulations in Small-World Network Models” (2005)

SIR models for infectious disease

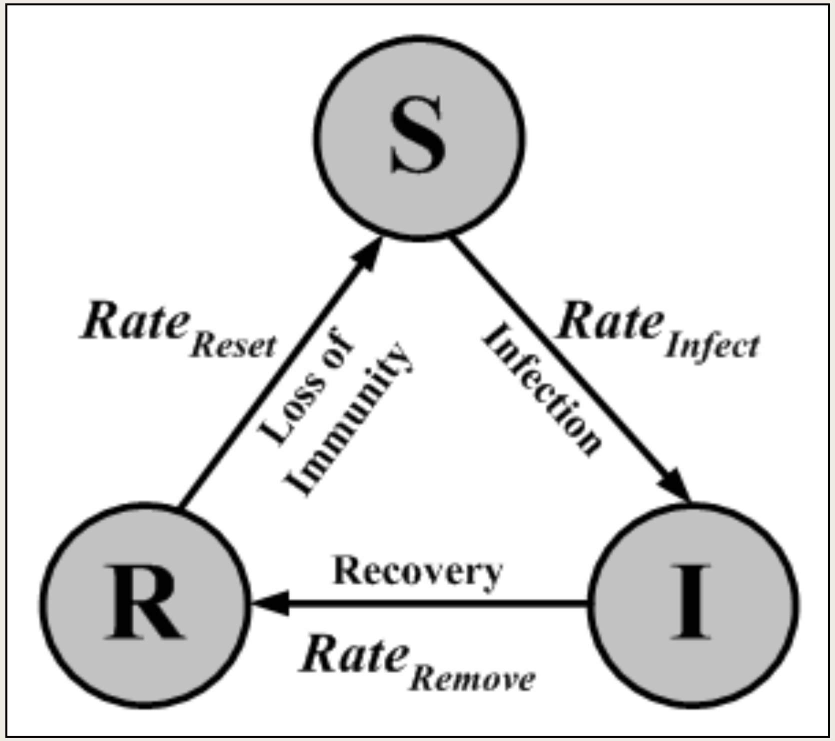
- Recent applications:
  - *H1N1, “swine flu”, 2009*
  - *Ebola, 2015*

$$\frac{dS}{dt} = -\frac{\beta I S}{N},$$

$$\frac{dI}{dt} = \frac{\beta I S}{N} - \gamma I,$$

$$\frac{dR}{dt} = \gamma I.$$

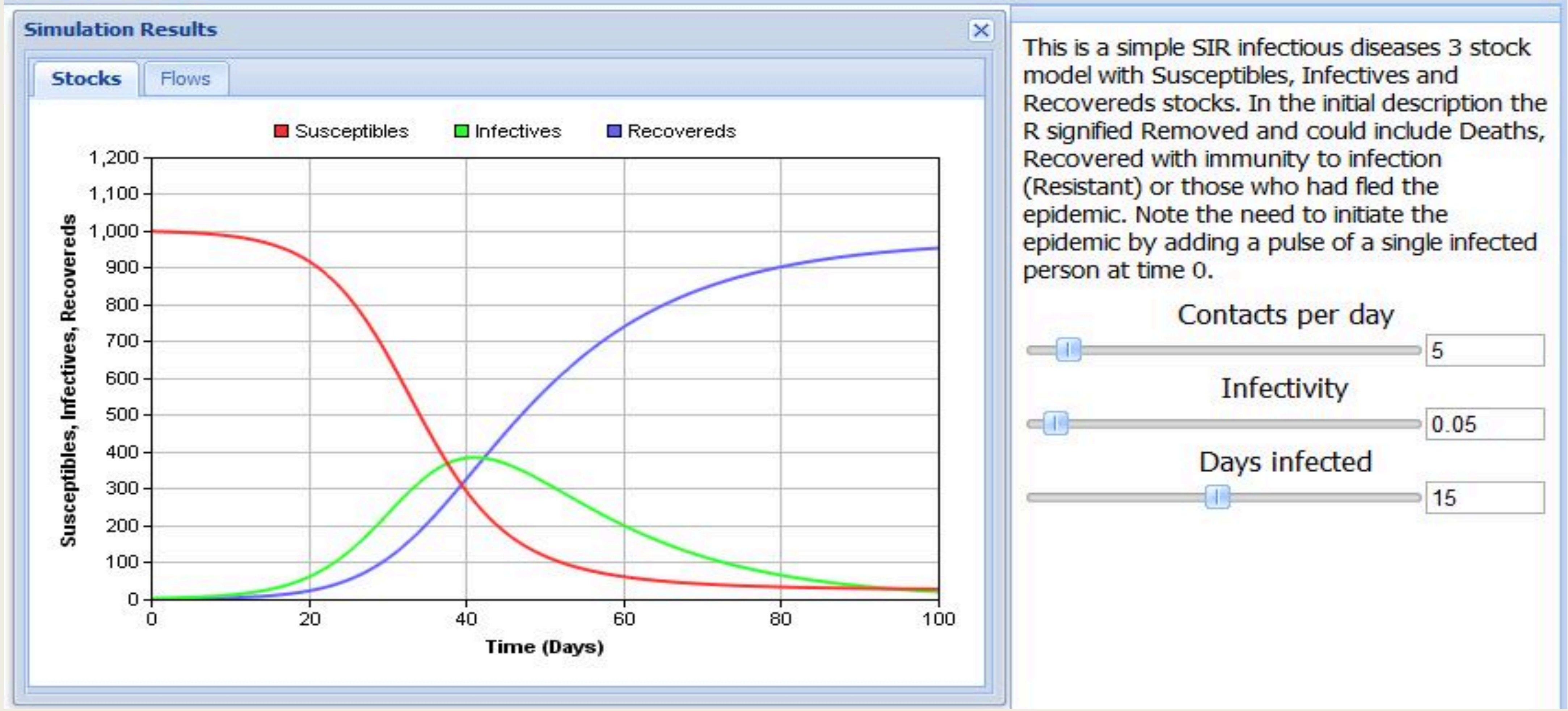
Modeled through differential equations



“Influence of Local Information on Social Simulations in Small-World Network Models” (2005)

Hethcote H (2000). "The Mathematics of Infectious Diseases"

SIR models for infectious disease



This is a simple SIR infectious diseases 3 stock model with Susceptibles, Infectives and Recovereds stocks. In the initial description the R signified Removed and could include Deaths, Recovered with immunity to infection (Resistant) or those who had fled the epidemic. Note the need to initiate the epidemic by adding a pulse of a single infected person at time 0.

Contacts per day: 5

Infectivity: 0.05

Days infected: 15

www.systemswiki.org

Cancer biology

Evolution of a cancerous tumor

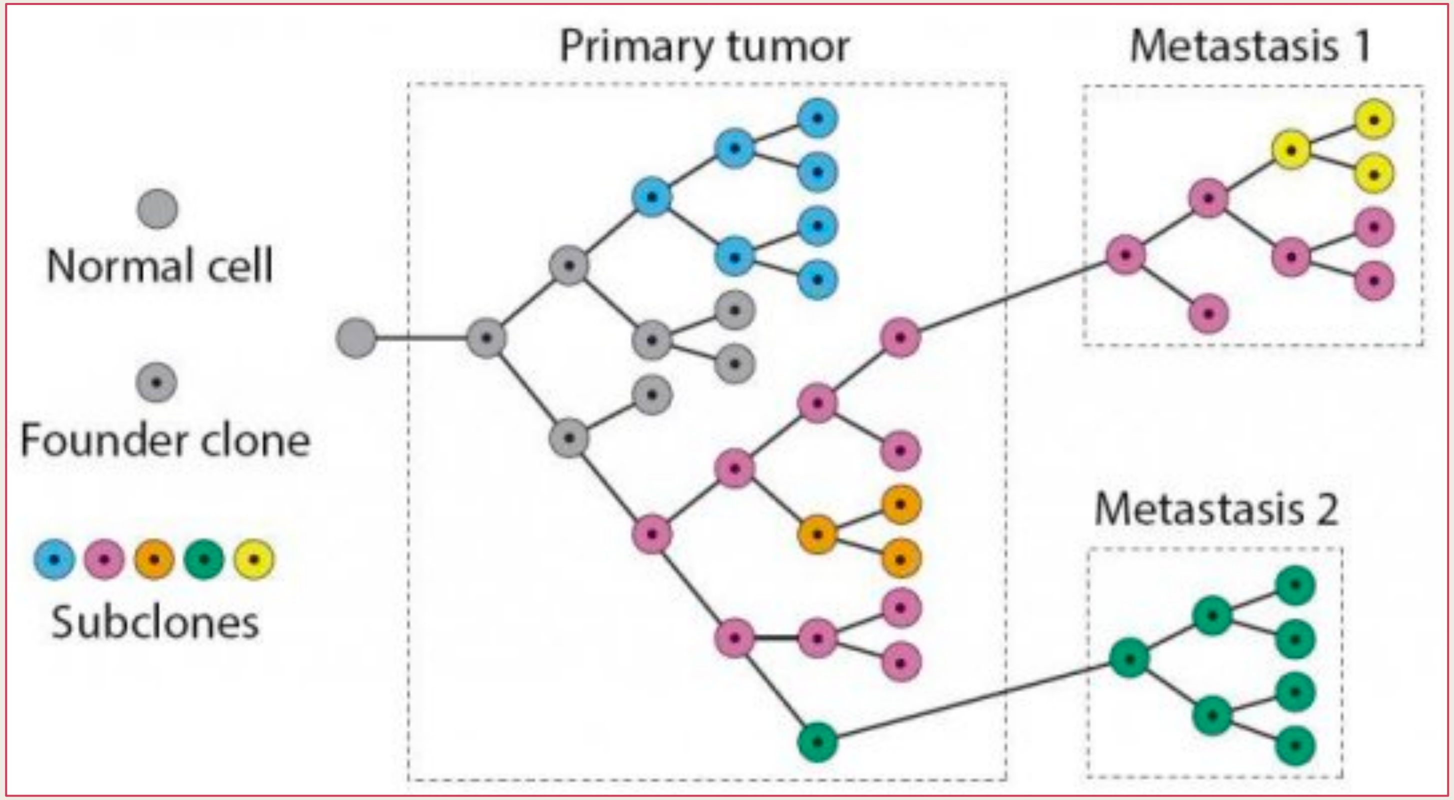


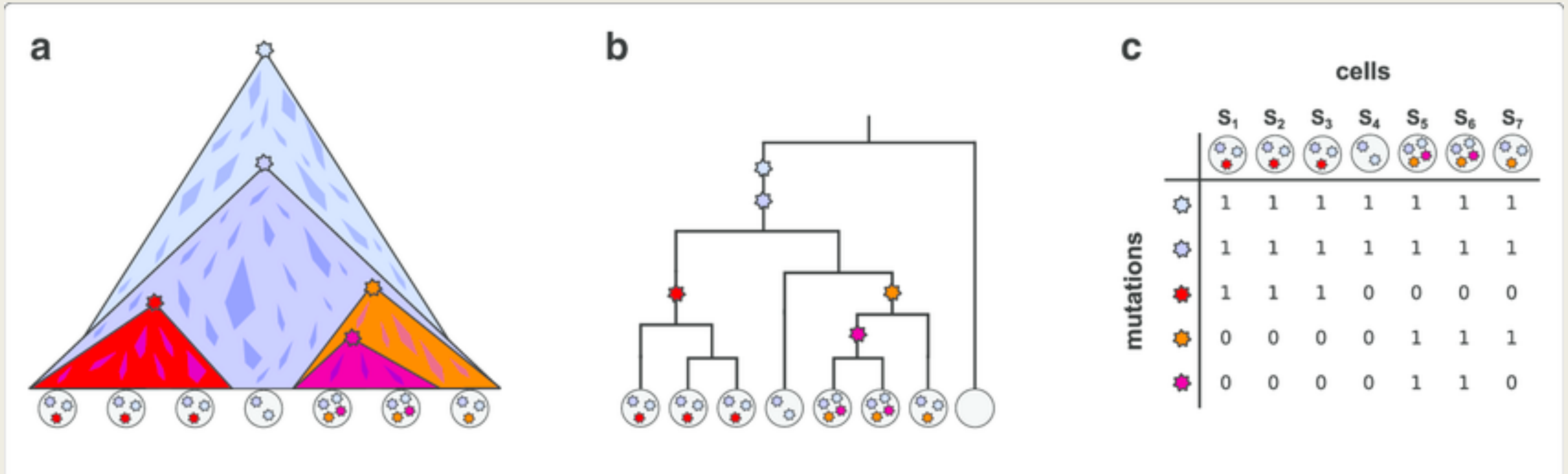
Image: ScienceDaily

Phylogenetic analysis of cancer cells

- Cancerous tumors often contain many different types of cells

- Once one mutation happens that causes the initial issue, mutations accumulate

- We can try to reconstruct the “ancestral” state to figure out what first went wrong

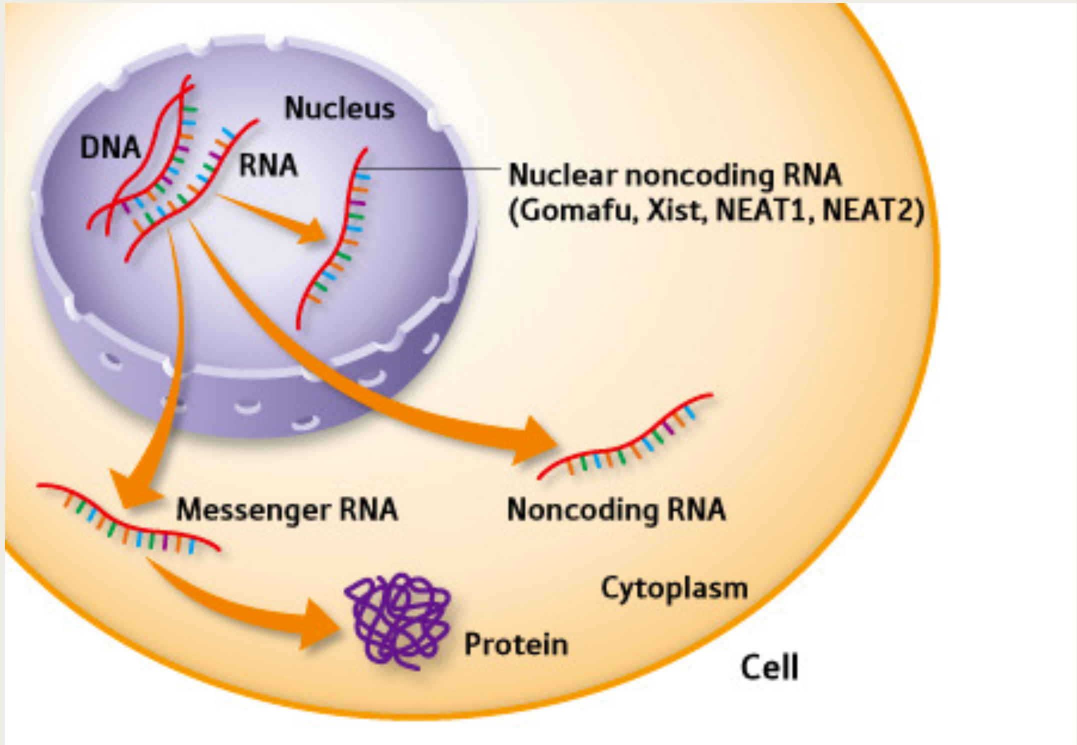


	S_1	S_2	S_3	S_4	S_5	S_6	S_7
✹	1	1	1	1	1	1	1
✹	1	1	1	1	1	1	1
✹	1	1	1	0	0	0	0
✹	0	0	0	0	1	1	1
✹	0	0	0	0	1	1	0

“Tree inference for single-cell data”, Genome Biology, 2016

Beyond a linear sequence…

RNA folding



- RNA does not stay as a linear sequence
- It folds into a secondary structure that minimizes energy

https://www.youtube.com/watch?v=KBI69y2ziXw

Image: genius.com

RNA secondary structure: larger example

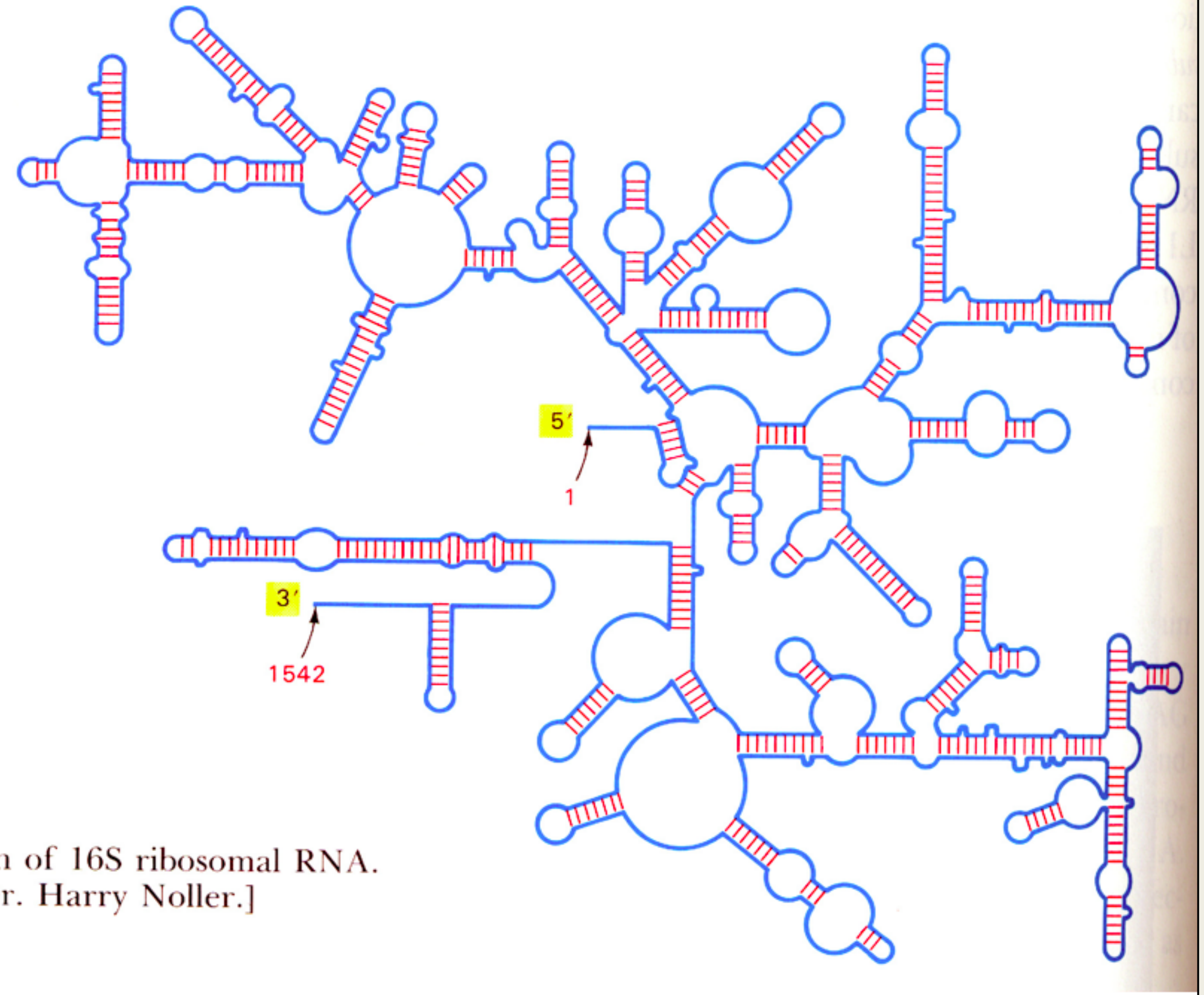
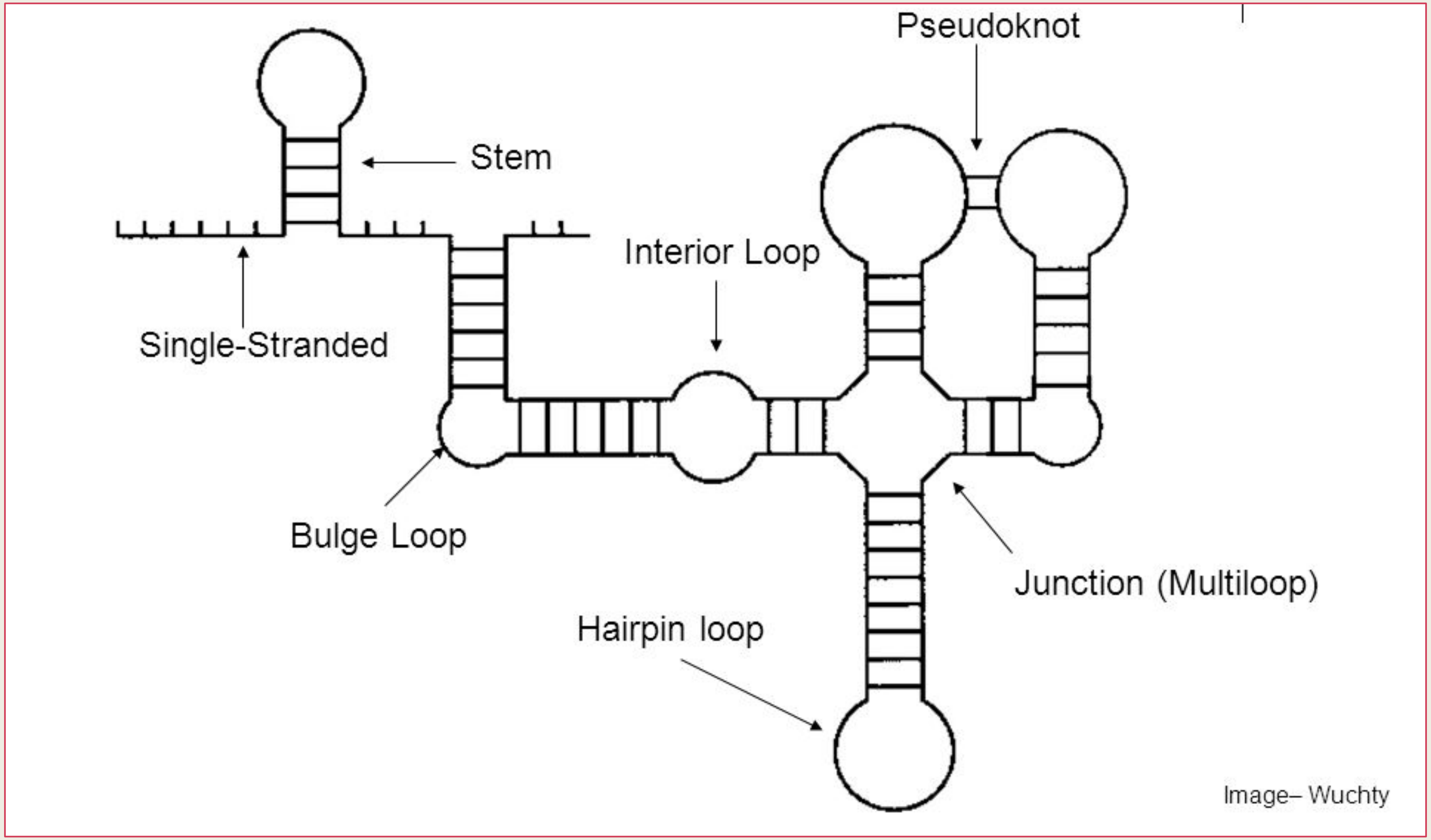
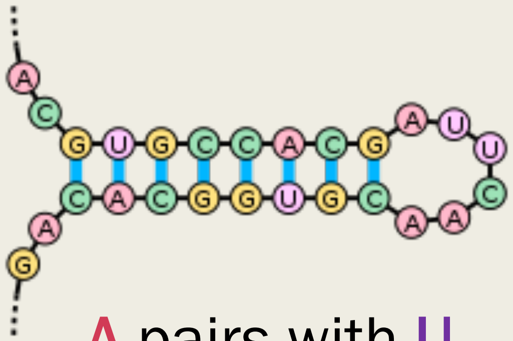


Figure 30-18
Folding pattern of 16S ribosomal RNA.
[Courtesy of Dr. Harry Noller.]

Features of RNA secondary structure



Image– Wuchty



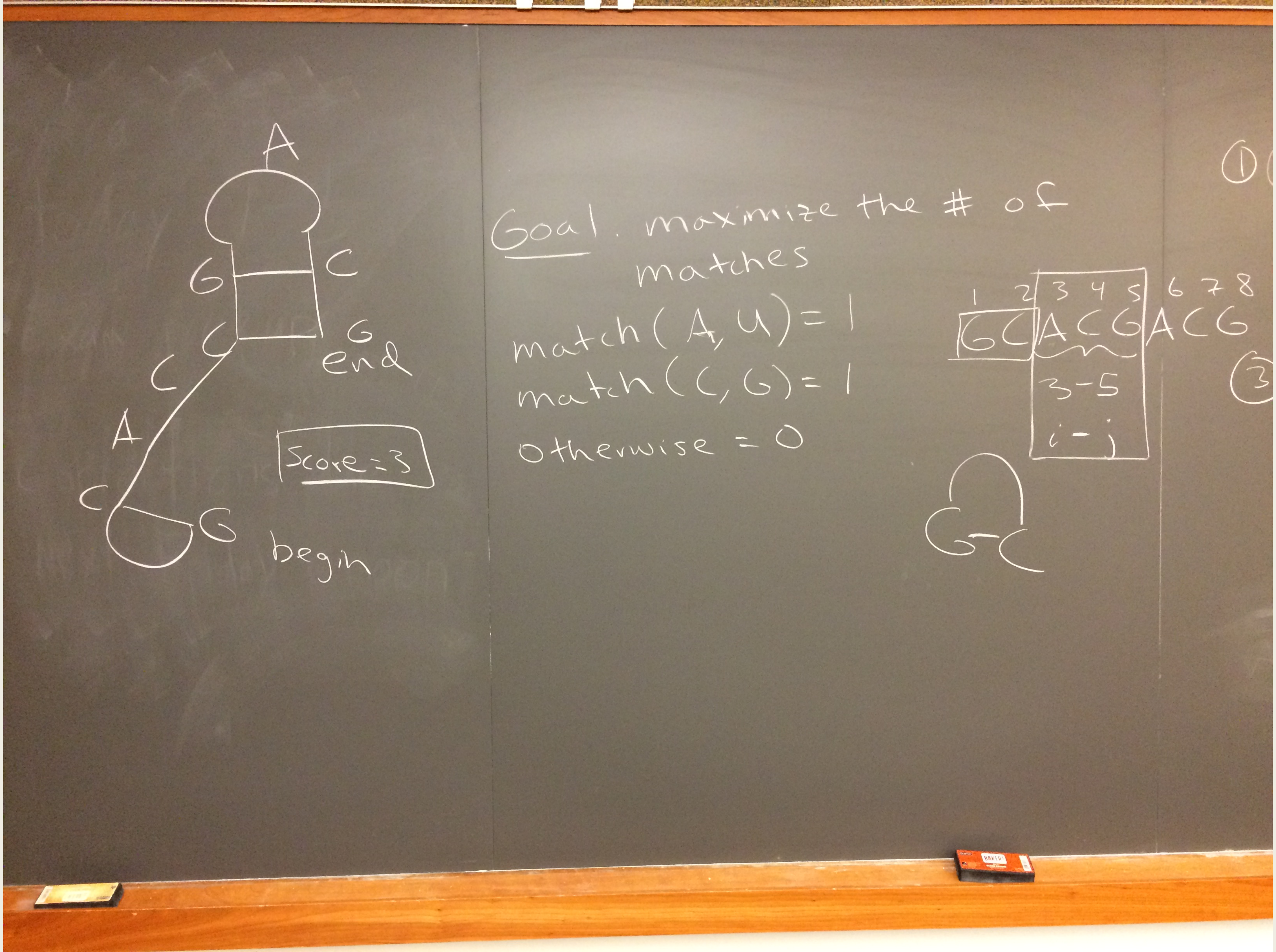
A pairs with U

C pairs with G

Image: wikipedia

Enter: computational biology

- Goal: how could we predict RNA secondary structure?
- Inspiration: sequence alignment
- Answer: dynamic programming (Nussinov’s algorithm)

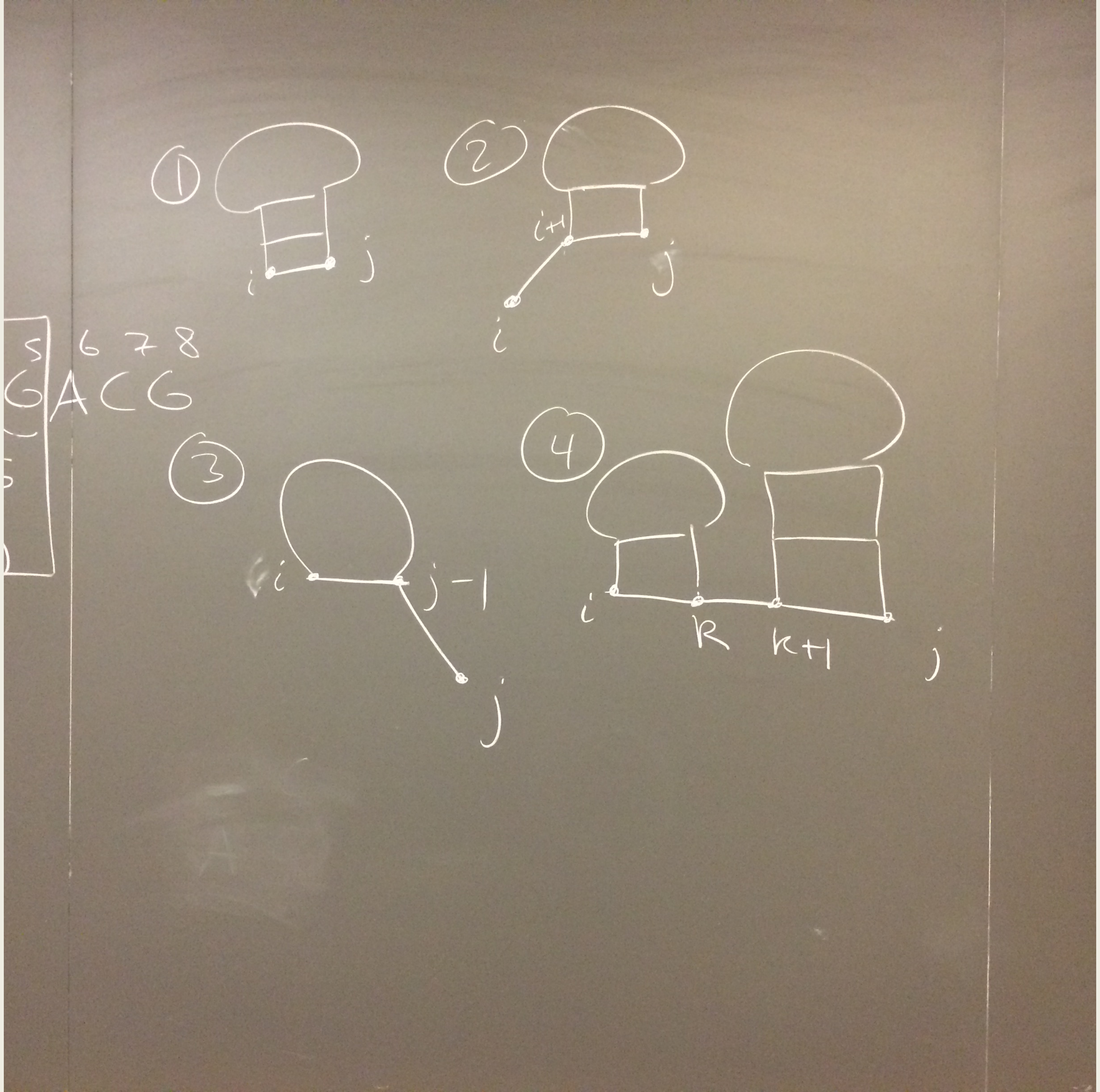


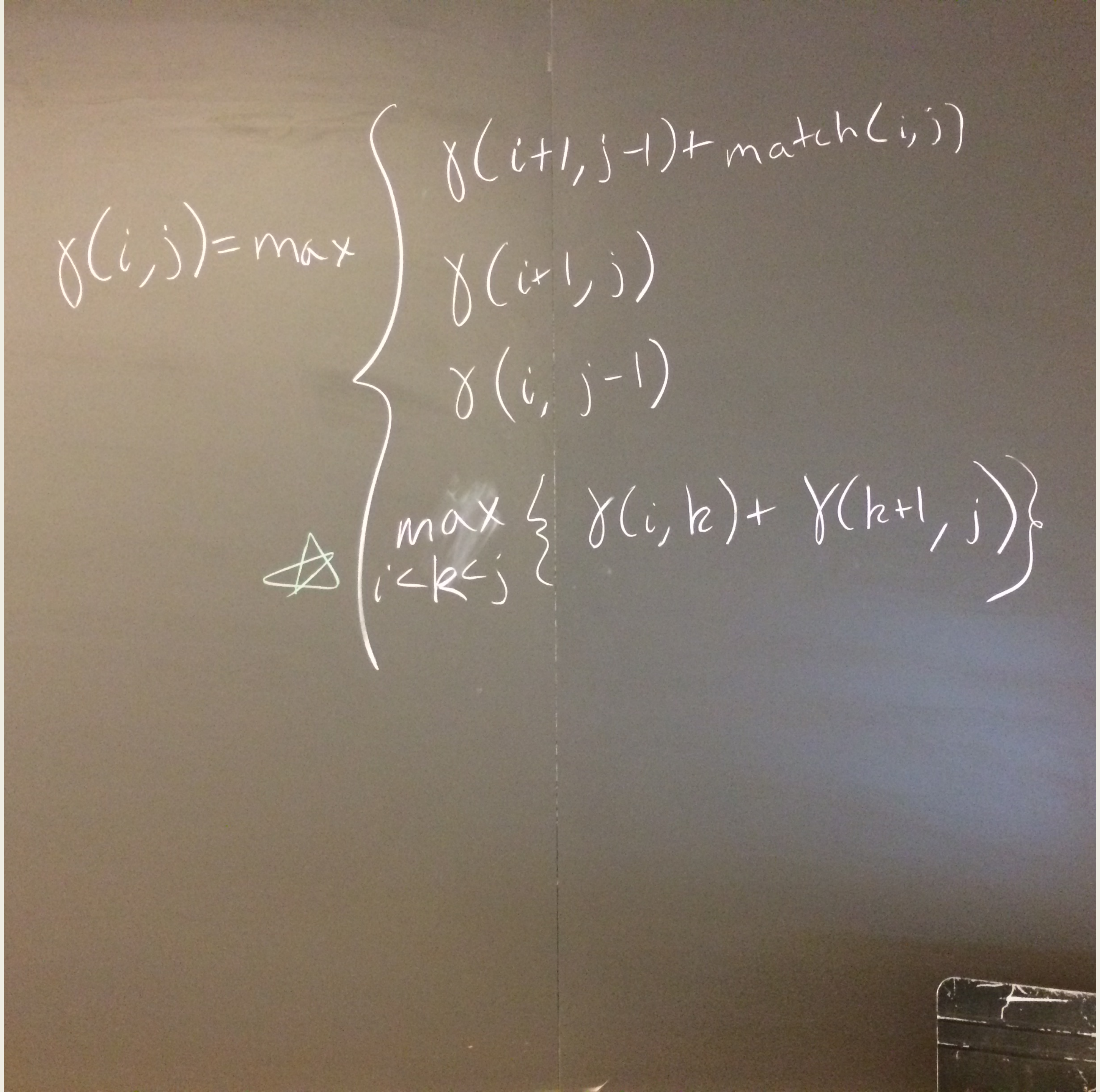
Goal: maximize the # of matches

match(A,U)=1

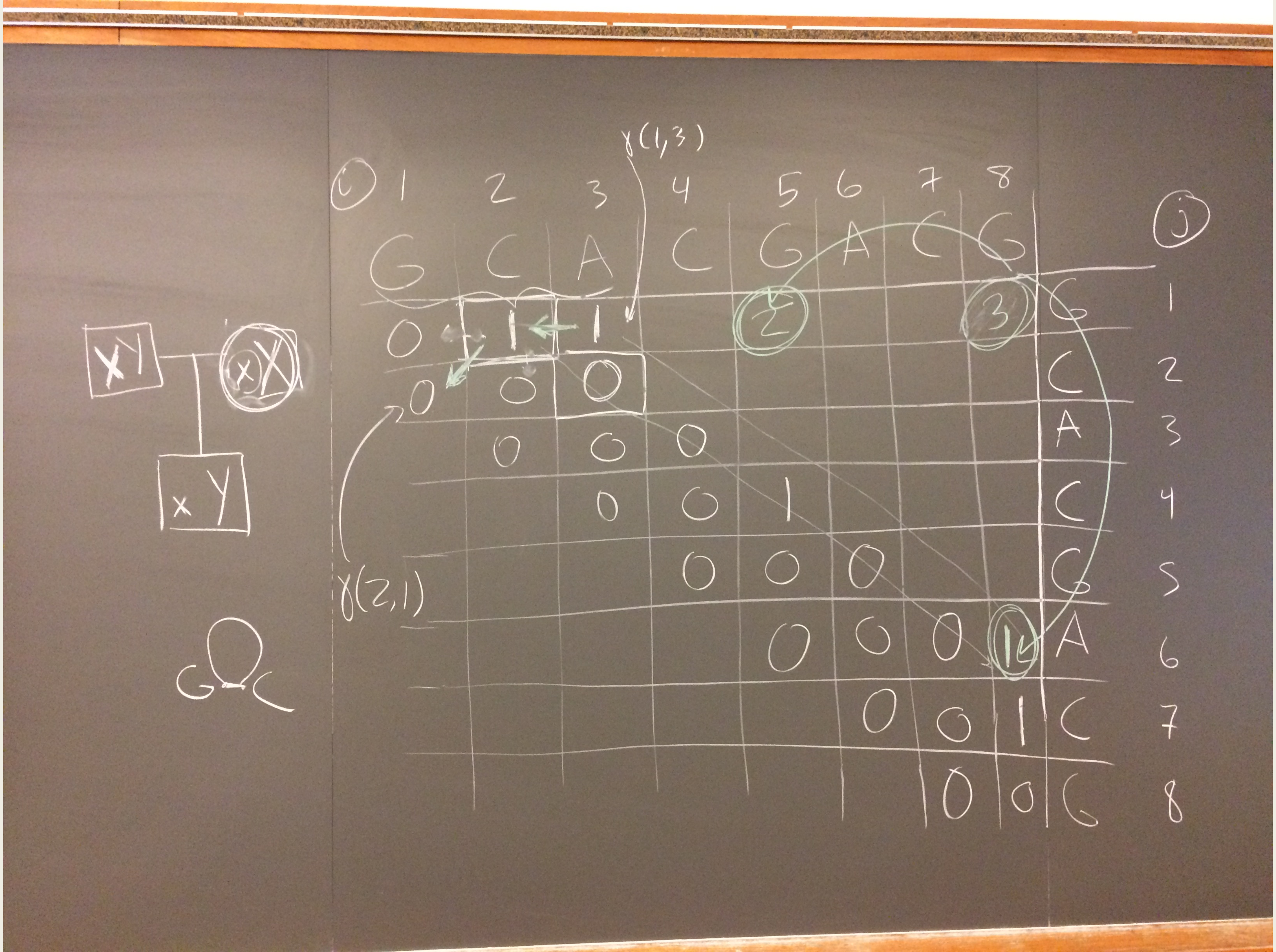
match(C,G)=1

otherwise = 0


1
i
j
2
i+1
i
j
5 6 7 8
GACG
3
i
j-1
j
4
i
k
k+1
j

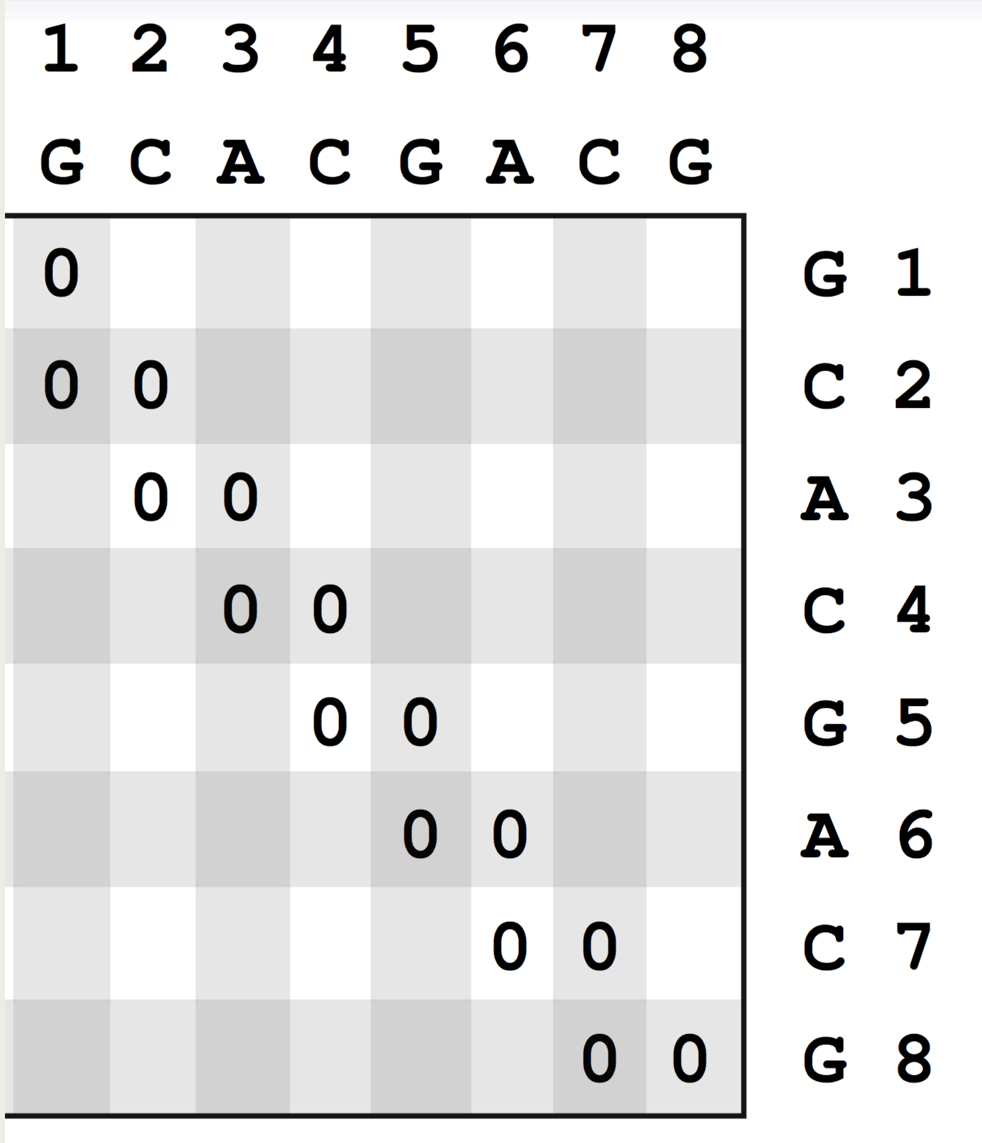


$$\gamma(i,j) = \max \begin{cases} \gamma(i+1, j-1) + \text{match}(i,j) \\ \gamma(i+1, j) \\ \gamma(i, j-1) \\ ☆ \quad \max\limits_{i<k<j} \left\{ \gamma(i,k) + \gamma(k+1, j) \right\} \end{cases}$$



i	1	2	3	4	5	6	7	8		j
	G	C	A	C	G	A	C	G		
	0	1	1		2			3	G	1
	0	0	0						C	2
		0	0	0					A	3
			0	0	1				C	4
				0	0	0			G	5
					0	0	0	1	A	6
						0	0	1	C	7
							0	0	G	8

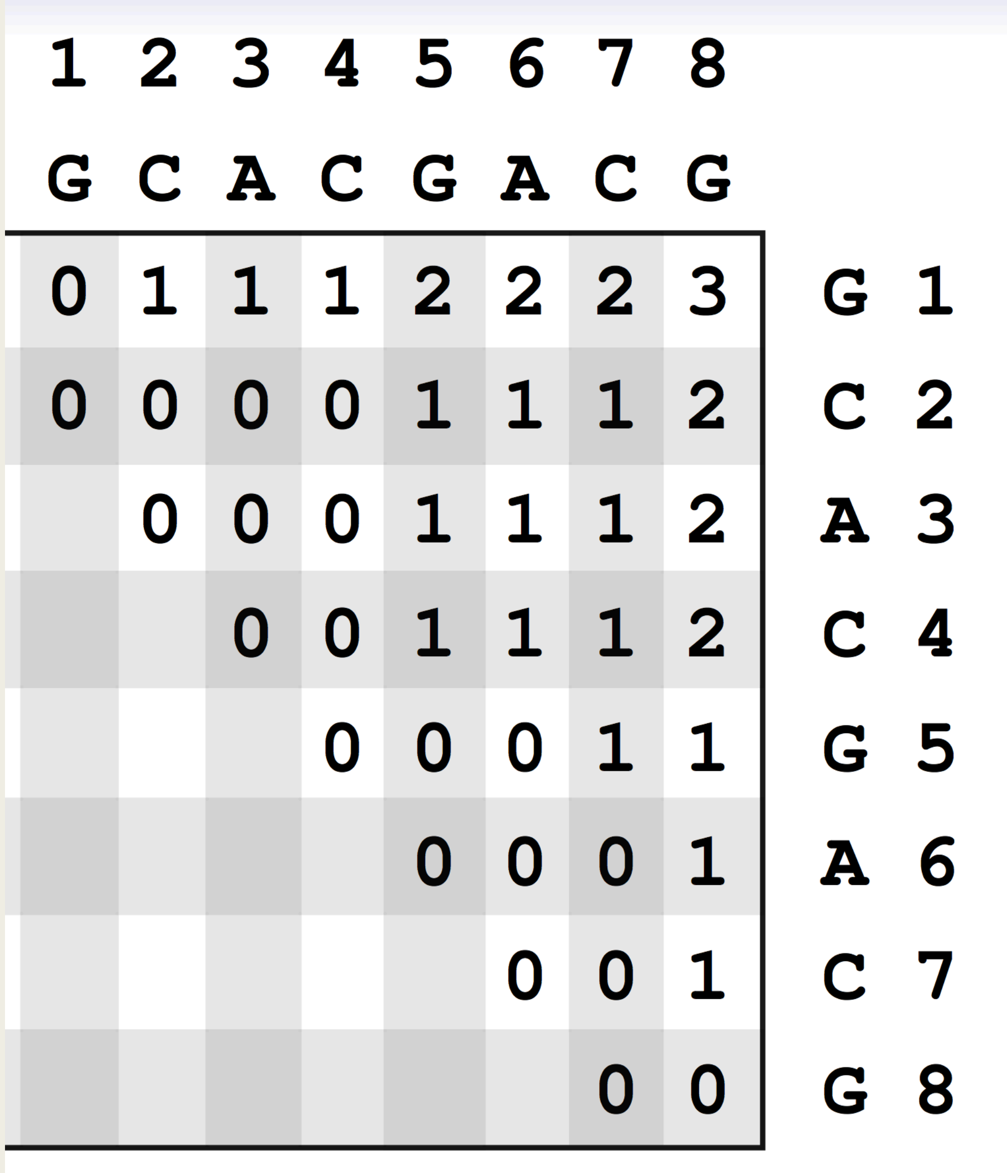
Example



1	2	3	4	5	6	7	8		
G	C	A	C	G	A	C	G		
0								G	1
0	0							C	2
	0	0						A	3
		0	0					C	4
			0	0				G	5
				0	0			A	6
					0	0		C	7
						0	0	G	8

S.Will, MIT 18.417, Fall 2011

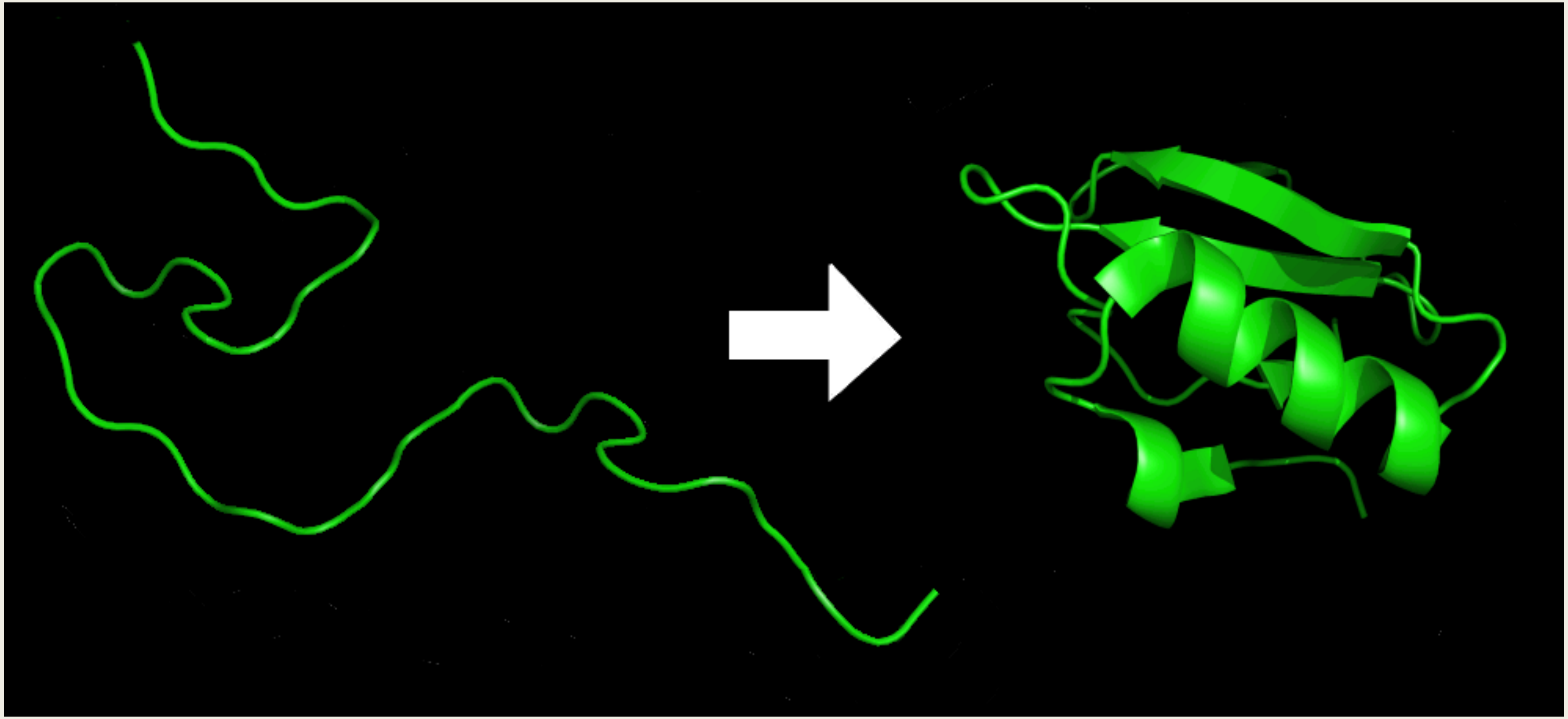
Example solution. Exercise: back-tracing



1	2	3	4	5	6	7	8		
G	C	A	C	G	A	C	G		
0	1	1	1	2	2	2	3	G	1
0	0	0	0	1	1	1	2	C	2
	0	0	0	1	1	1	2	A	3
		0	0	1	1	1	2	C	4
			0	0	0	1	1	G	5
				0	0	0	1	A	6
					0	0	1	C	7
						0	0	G	8

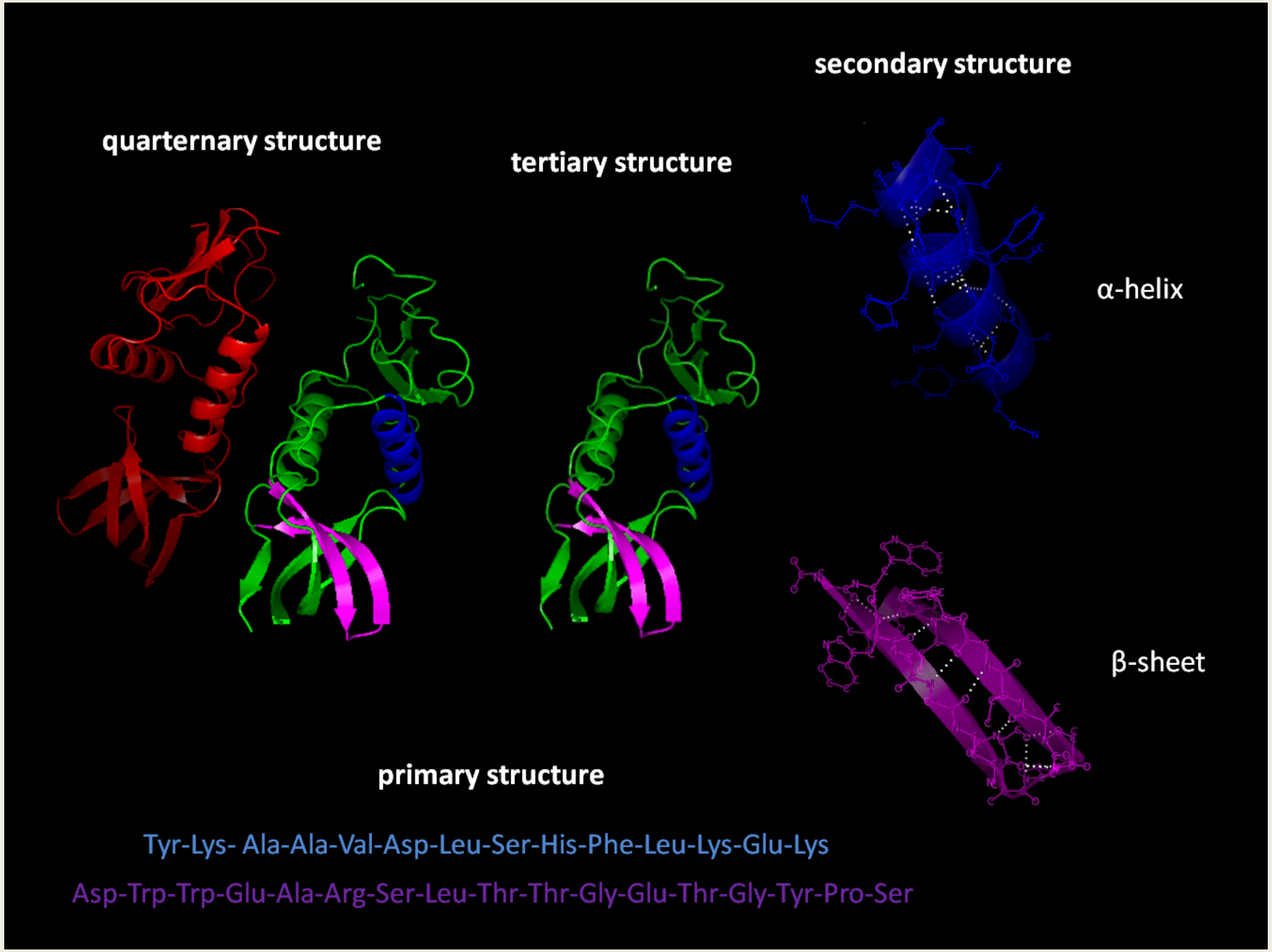
S.Will, MIT 18.417, Fall 2011

Protein folding: from sequence to structure



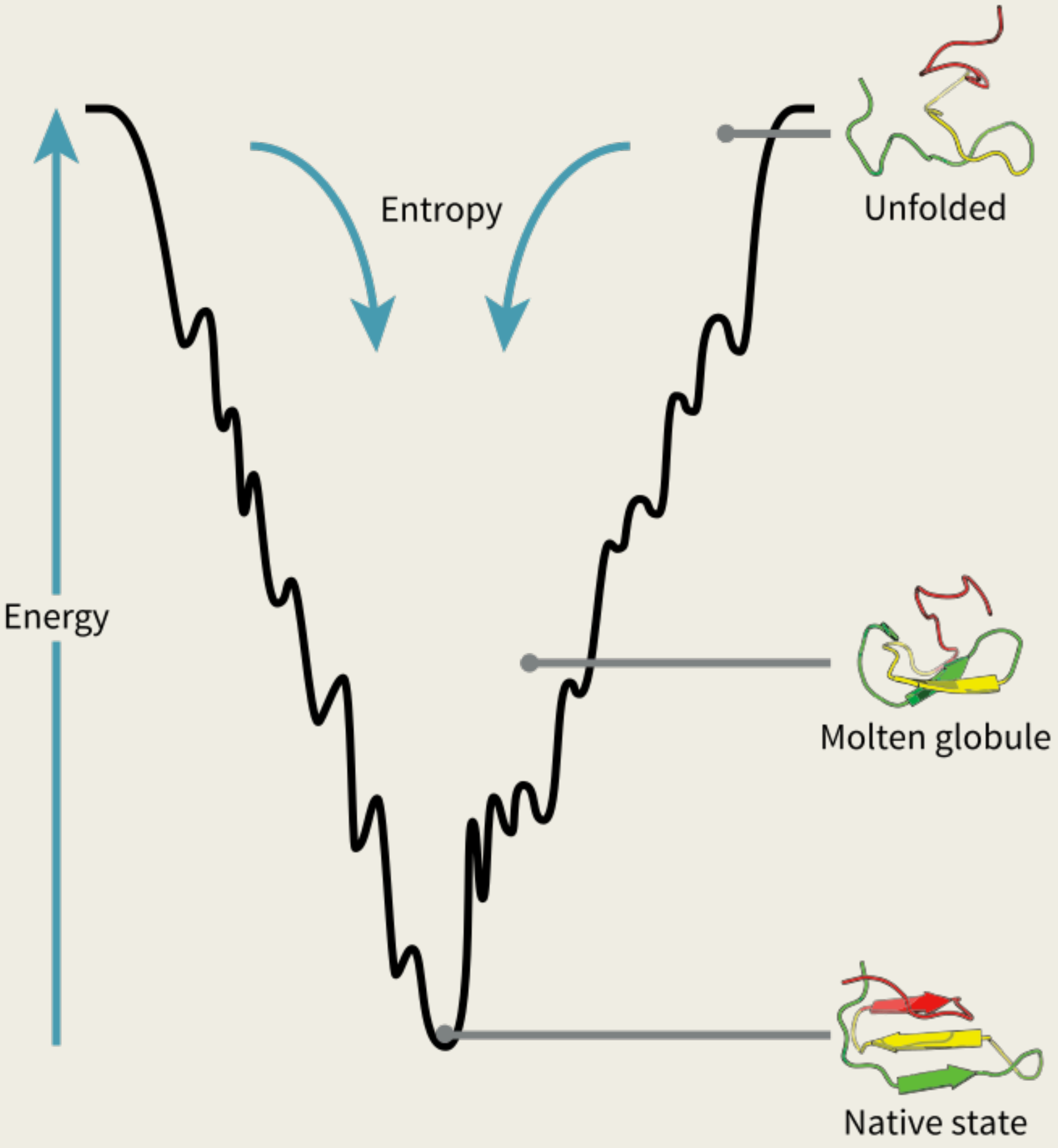
By: DrKjaergaard, Wikipedia

Protein structure beyond the sequence


secondary structure
quarternary structure
tertiary structure
α-helix
β-sheet
primary structure
Tyr-Lys- Ala-Ala-Val-Asp-Leu-Ser-His-Phe-Leu-Lys-Glu-Lys
Asp-Trp-Trp-Glu-Ala-Arg-Ser-Leu-Thr-Thr-Gly-Glu-Thr-Gly-Tyr-Pro-Ser

By: Holger87, Wikipedia

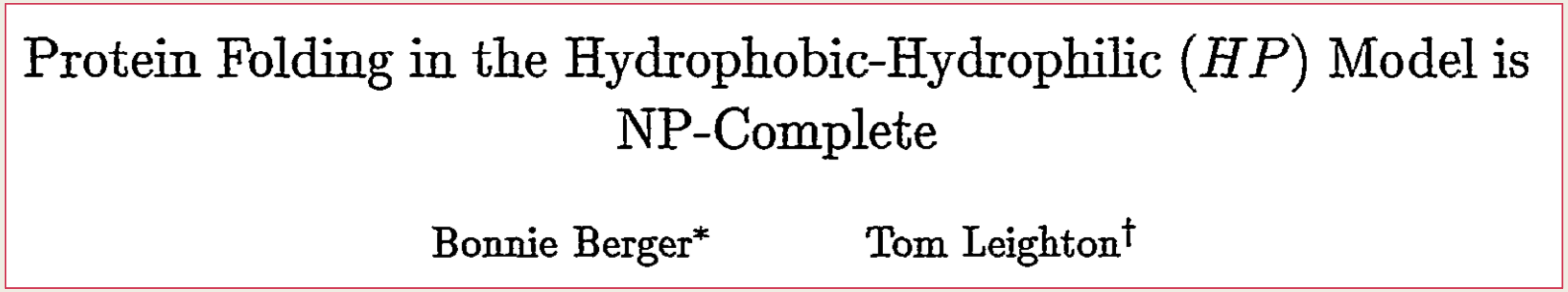
Proteins seek a low-energy configuration



By: Thomas Splettstoesser, Wikipedia

Breakthrough in protein folding

- Bonnie Berger and Tom Leighton prove protein folding is NP-Complete (1998)
- Helped pave the way for approximation algorithms

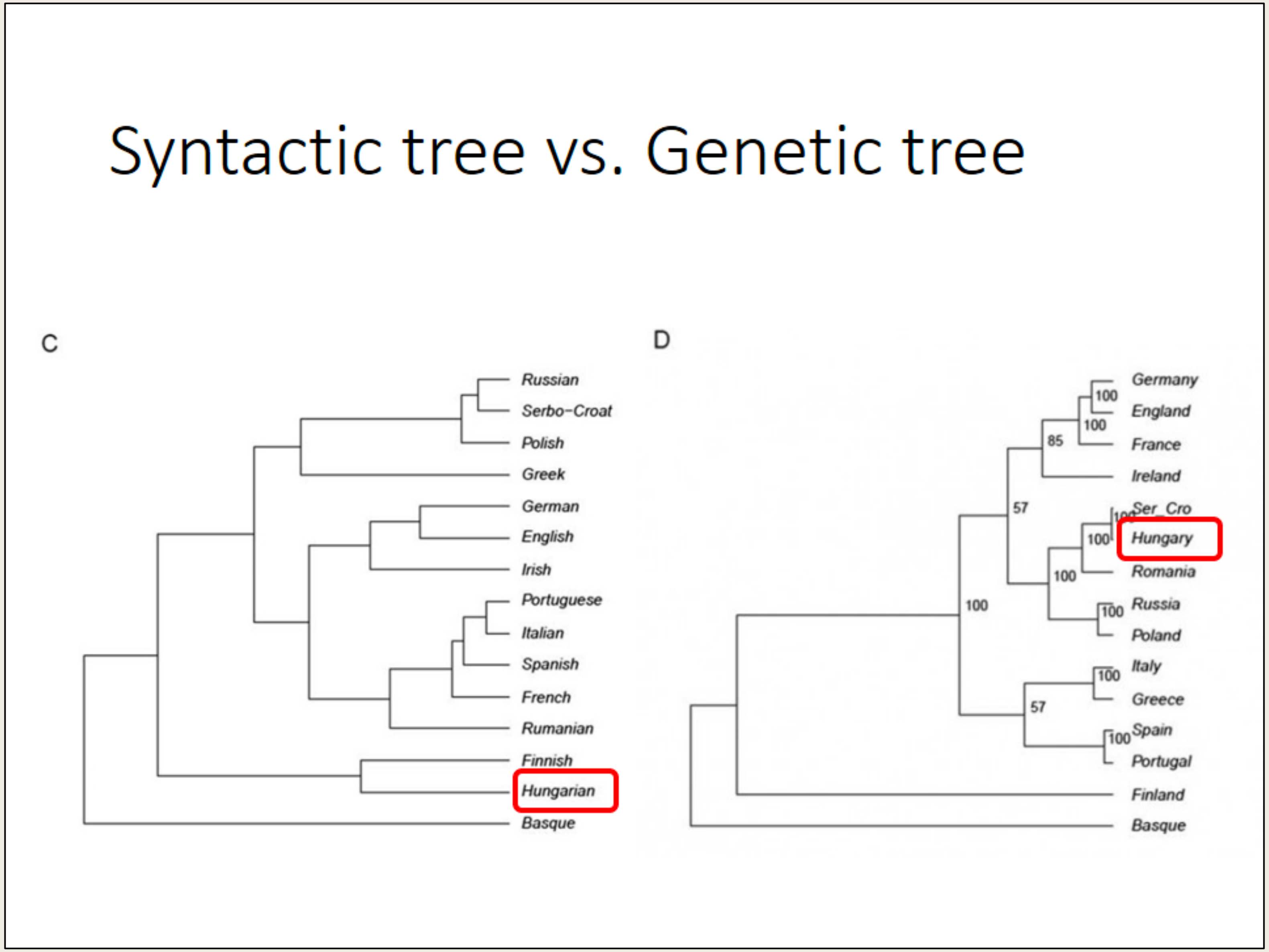


Protein Folding in the Hydrophobic-Hydrophilic (HP) Model is NP-Complete

Bonnie Berger* Tom Leighton†

Final thoughts

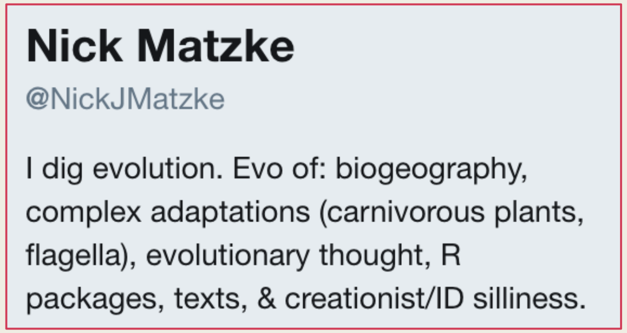
Combining linguistics and genetics



Syntactic tree vs. Genetic tree

Other thoughts

- Interested in evolution vs. creationism debate? Recommend following Nick Matzke



Nick Matzke

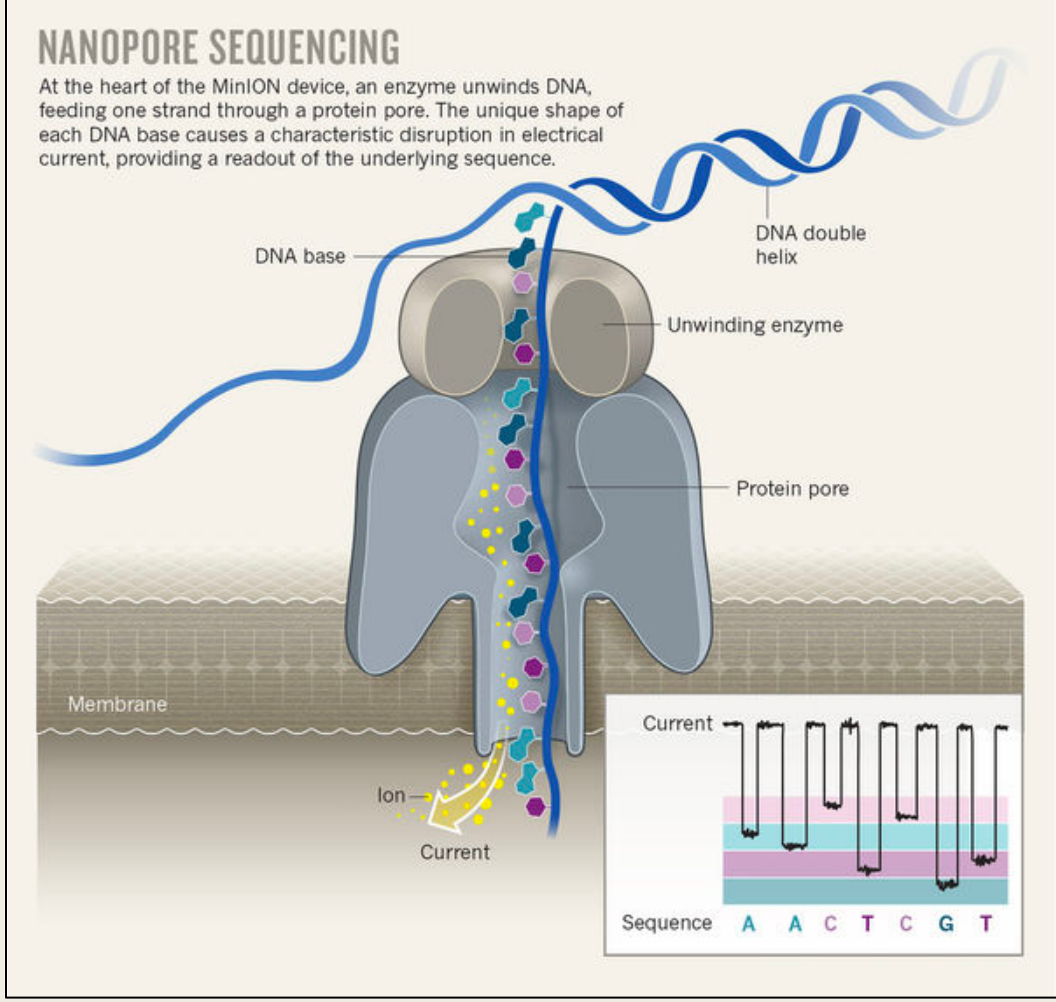
@NickJMatzke

I dig evolution. Evo of: biogeography, complex adaptations (carnivorous plants, flagella), evolutionary thought, R packages, texts, & creationist/ID silliness.

- Industry:
  - *Genentech*
  - *Illumina*
  - *23andMe*
  - *Ancestry.com*
  - *Invitae*
  - *Google Genomics*

Areas of Opportunity

- *Managing and analyzing data quickly and in a more automated way*
- *Intersecting with biochemistry to make sequencing better*
- *Sequencing more species, especially to assist conservation efforts*
- *Microbiome sequencing and understanding*


NANOPORE SEQUENCING
At the heart of the MinION device, an enzyme unwinds DNA, feeding one strand through a protein pore. The unique shape of each DNA base causes a characteristic disruption in electrical current, providing a readout of the underlying sequence.
DNA base
DNA double helix
Unwinding enzyme
Protein pore
Membrane
Ion
Current
Current
Sequence
A A C T C G T

Example: Oxford Nanopore

Image: blogs.nature.com