

# Combinatorics Inspired by Biology

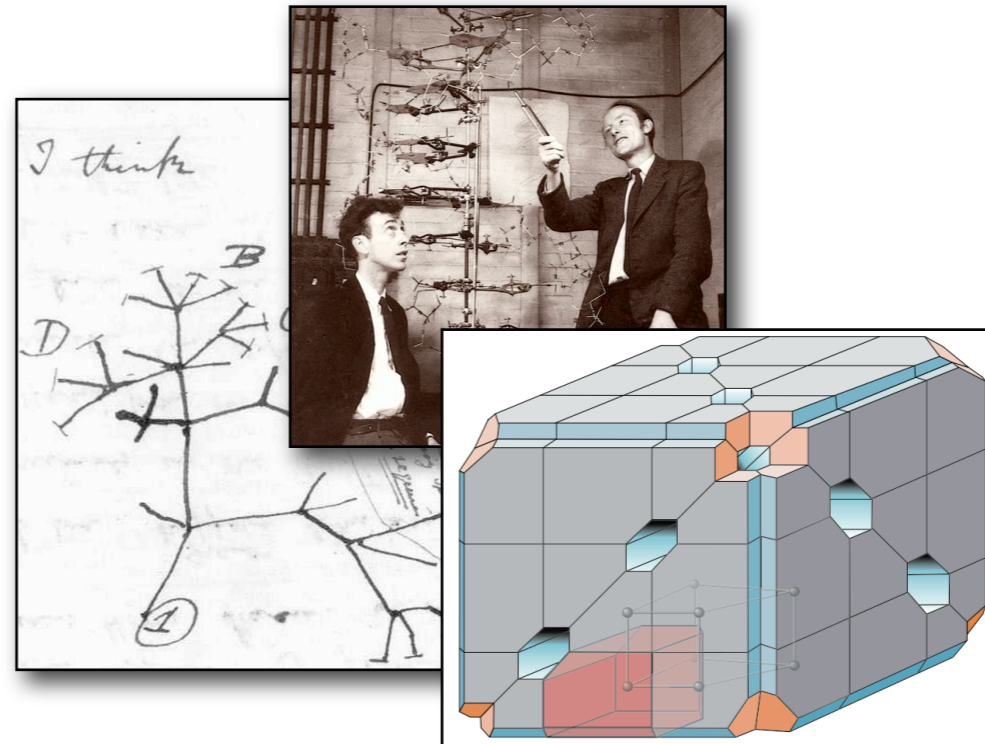
SIAM AN09

---

Lior Pachter

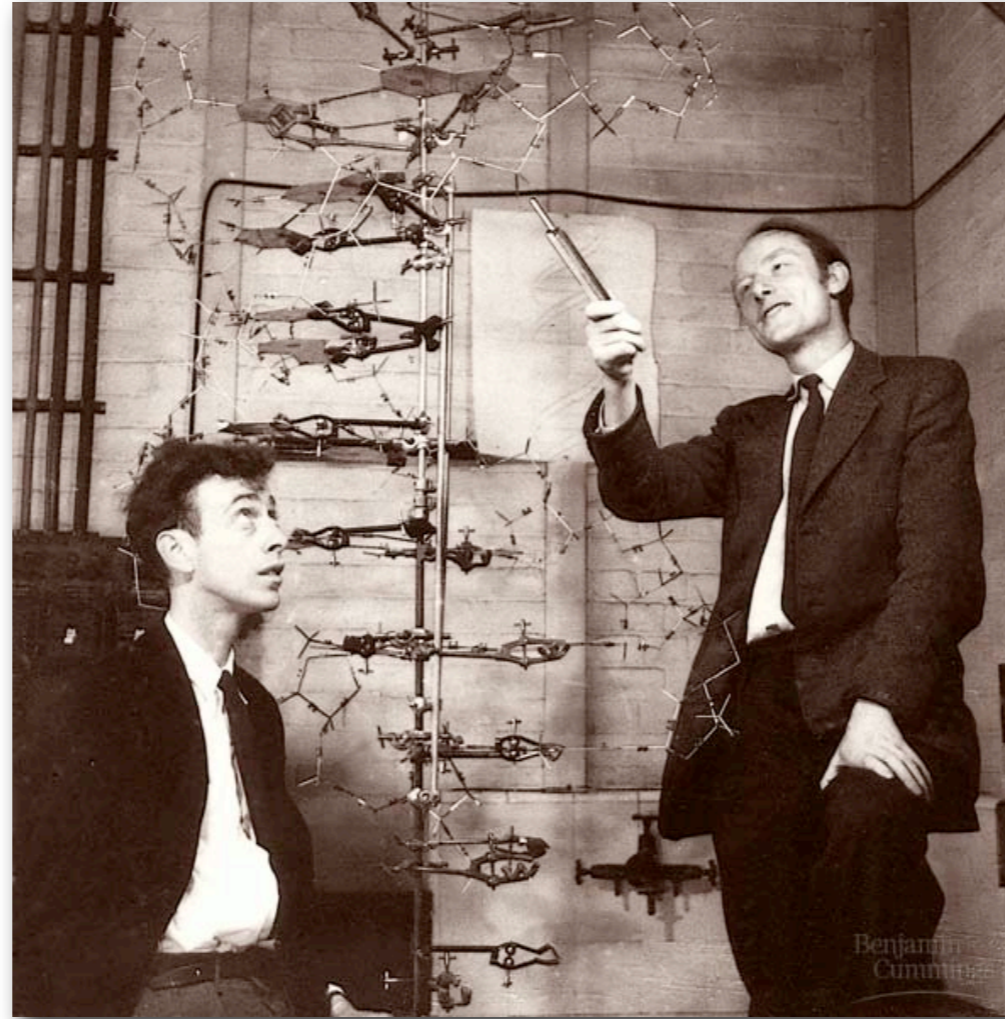
Departments of Mathematics,  
Molecular & Cellular Biology and  
Computer Science

University of California, Berkeley



# One of the unifying principles of biology

---

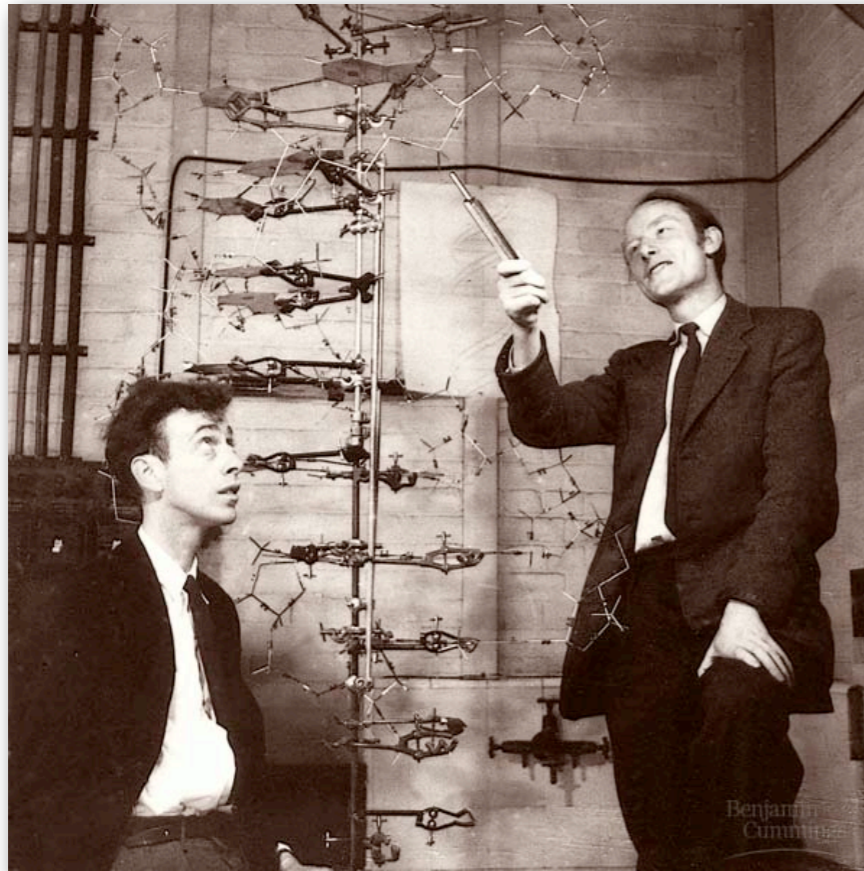


“It has not escaped our attention that the specific pairing we have postulated immediately suggests a copying mechanism for the genetic material”

# One of the unifying principles of biology

---

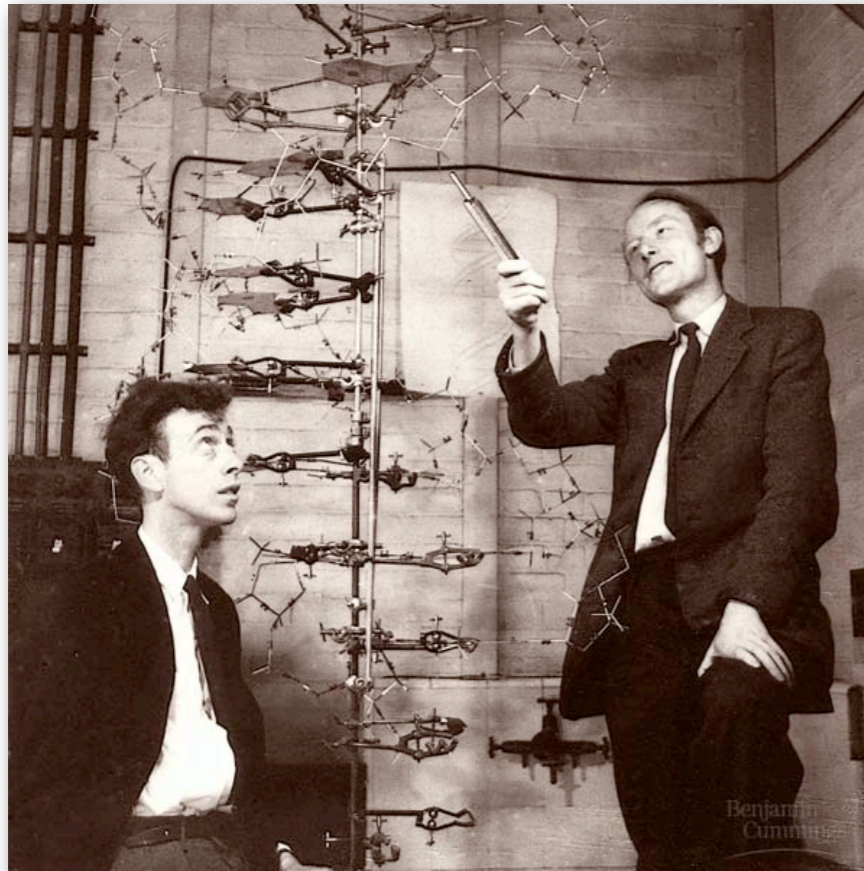
- **1953:** Watson, Crick and Franklin's work leads to the proposed double helix structure for DNA



“It has not escaped our attention that the specific pairing we have postulated immediately suggests a copying mechanism for the genetic material”

# One of the unifying principles of biology

---

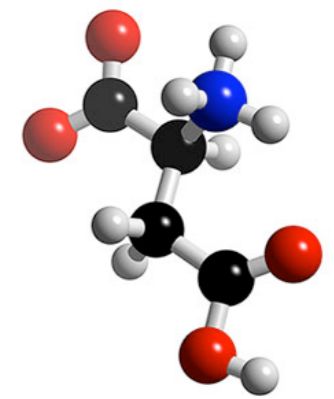


- **1953:** Watson, Crick and Franklin's work leads to the proposed double helix structure for DNA
- **1954:** George Gamow founds the RNA tie club. The club consists of 20 regular members and 4 honorary members.

"It has not escaped our attention that the specific pairing we have postulated immediately suggests a copying mechanism for the genetic material"

# History of the discovery of the 20 universal amino acids

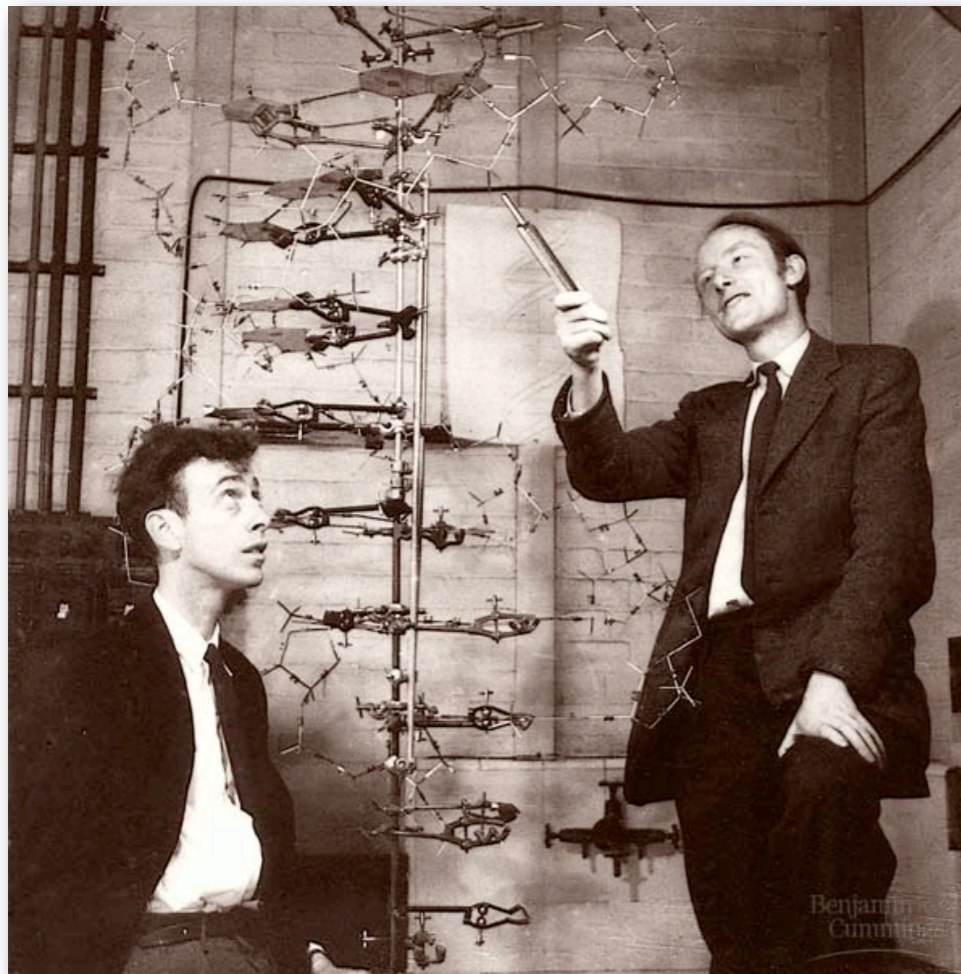
---



- **1806:** Louis-Nicolas Vauquelin and Pierre-Jean Robiquet isolate the first amino acid in asparagus (asparagine).
- **1868:** Friedrich Miescher discovers nucleic acids.
- **1935:** William Rose discovers the 20th (and final) universal amino acid (threonine).
- **1939:** Caspersson and Schultz propose a connection between RNA and protein synthesis.
- **1942:** Starvation experiments in graduate students establish that eight amino acids are essential.

# Deciphering the “genetic code”

---

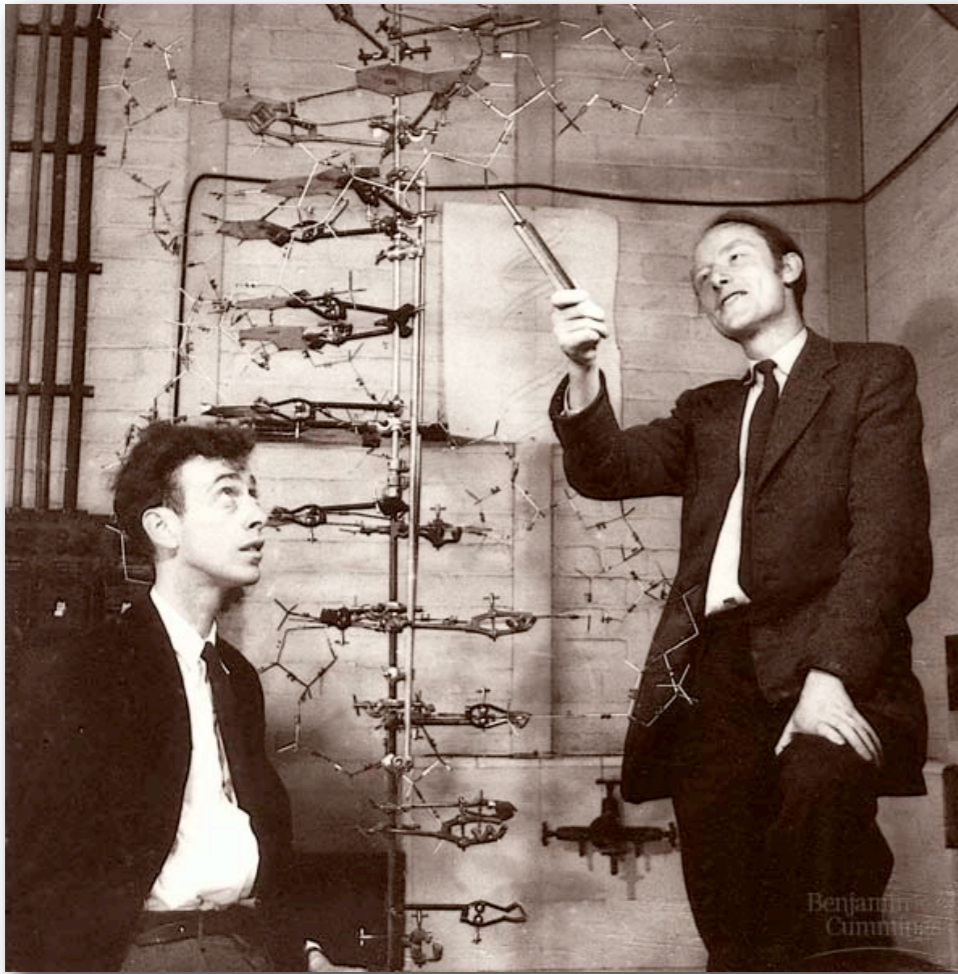


- **1953:** Watson, Crick and Franklin’s work leads to the proposed double helix structure for DNA
- **1954:** George Gamow founds the RNA tie club. The club consists of 20 regular members and 4 honorary members.

“It has not escaped our attention that the specific pairing we have postulated immediately suggests a copying mechanism for the genetic material”

# Deciphering the “genetic code”

---

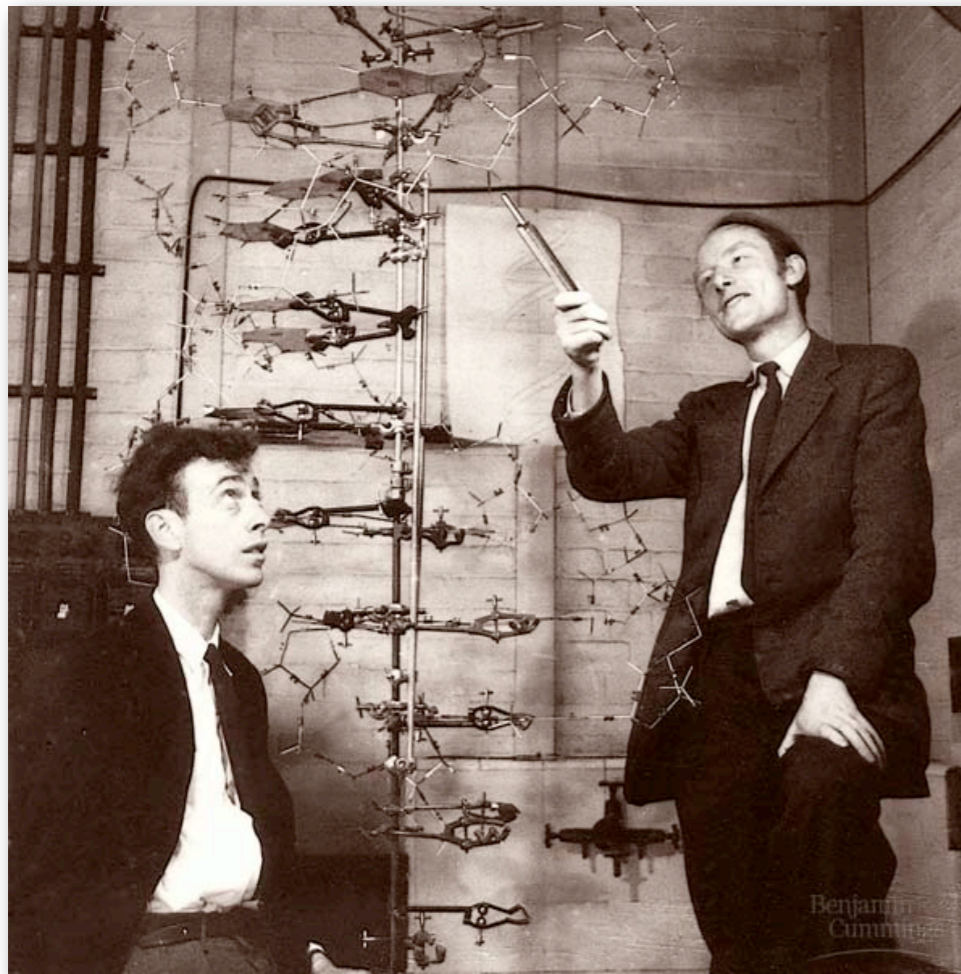


- **1953:** Watson, Crick and Franklin’s work leads to the proposed double helix structure for DNA
- **1954:** George Gamow founds the RNA tie club. The club consists of 20 regular members and 4 honorary members.
- **1955:** Francis Crick suggests the “adaptor hypothesis”.

“It has not escaped our attention that the specific pairing we have postulated immediately suggests a copying mechanism for the genetic material”

# Deciphering the “genetic code”

---



- **1953:** Watson, Crick and Franklin’s work leads to the proposed double helix structure for DNA
- **1954:** George Gamow founds the RNA tie club. The club consists of 20 regular members and 4 honorary members.
- **1955:** Francis Crick suggests the “adaptor hypothesis”.
- **1956** Gamow thinks mathematically about the number of nucleotides necessary to make one amino acid.

“It has not escaped our attention that the specific pairing we have postulated immediately suggests a copying mechanism for the genetic material”



$$4^2 < 20 < 4^3$$

---

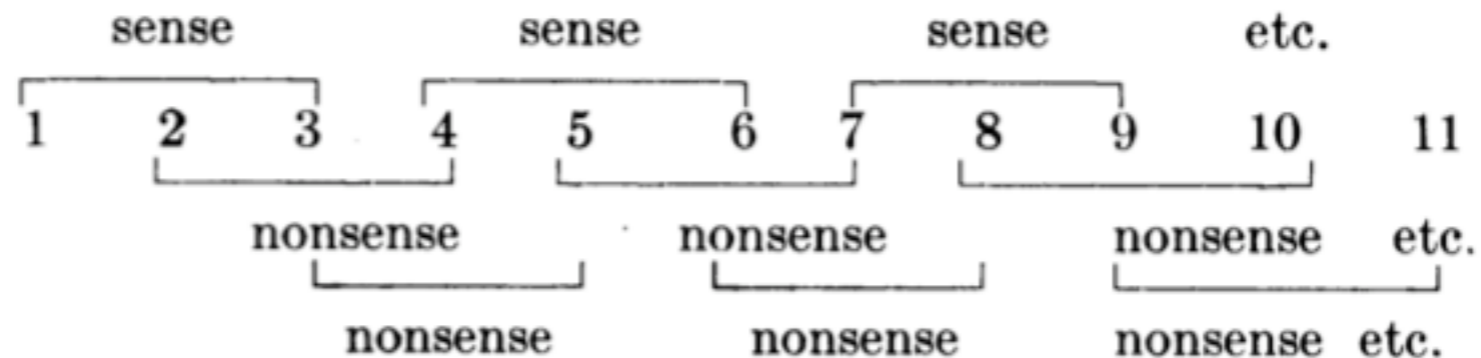
It is assumed in one of the more popular theories of protein synthesis that amino acids are ordered on a nucleic acid strand (see, for example, Dounce [1]) and that the order of the amino acids is determined by the order of the nucleotides of the nucleic acid. There are some twenty naturally occurring amino acids commonly found in proteins, but (usually) only four different nucleotides. The problem of how a sequence of four things (nucleotides) can determine a sequence of twenty things (amino acids) is known as the "coding" problem.

Crick, Griffith and Orgel, *PNAS* 1957.

# A sense and nonsense proposal

---

- There must be a mechanism for determining “frame”:



- Crick, Griffith and Orgel guess that this may be accomplished via a “comma free code”. That is, an assignment of a subset of nucleotide triplets to sense codons such that any sequence of successive sense codons only has nonsense codons in the shifted positions.

# Two simple observations and a question

---

- Repeats cannot be sense triplets. For example, if AAA is nonsense because the frame cannot be determined in the sequence AAAAAA.
- Shifts of sense triplets must be nonsense. For example, if AGC is a sense triplet, then GCA and CAG cannot be. This is because the frame must be recognizable in the sequence AGCAGC.
- **Question:** what is the maximum size of a comma free code on a four letter alphabet?

---

20

# Example of a comma free code

---

AGA	AGG	ACA	ACG	ACC
GCA	GCG	GCC	ATA	ATG
ATC	ATT	GTA	GTG	GTC
GTT	CTA	CTG	CTC	CTT

# The resulting math

---

- Let  $W(k,n)$  be the maximum number of words in a comma free dictionary where each word has size  $k$  and the size of the alphabet is  $n$ .

- Theorem: 
$$W(k, n) \leq \frac{1}{k} \sum_{d|k} \mu(d) n^{\frac{k}{d}}$$

- Proved by Golomb, Gordon and Welch in 1958 (Canadian Journal of Mathematics).

- Showed that the bound is tight for  $k=1,3,5,7,9,11,13$  and  $15$ .

- Showed that the bound is tight for  $k=2$ ,  $W(2, n) = \lfloor \frac{1}{3}n^2 \rfloor$

- Crick, Griffith and Orgel's result is the special case  $W(3,4)=20$ .

# The math

---

- Eastman (1965) showed that the bound is tight for all odd  $k$ .
- The problem of determining whether the bound is tight remains **open for even  $k$** . The current best result is due to Tang, Golomb and Graham from 1987.

- The bound 
$$W(k, n) \leq \frac{1}{k} \sum_{d|k} \mu(d) n^{\frac{k}{d}}$$

is based on counting the number of necklaces of period  $n$  with  $k$  types of beads. This connects the problem to numerous questions in enumerative combinatorics.

# My first (biological?) paper

---

- Motivated by a paper of Kleitman and Füredi on zero sum sequences we asked how many subsets of a finite abelian group sum to to the identity element.
- A special case of this is the number of subsets of  $\{1, \dots, n\}$  that sum to  $k \pmod n$

$$N_n^k = \frac{1}{n} \sum_{\substack{s|n \\ s \text{ odd}}} 2^{\frac{n}{s}} \frac{\varphi(s)}{\varphi\left(\frac{s}{(k,s)}\right)} \mu\left(\frac{s}{(k,s)}\right).$$

[N. Kitchloo and P., 1994]

- The Crick et al. bound agrees with our formula ( $k=1$ ), and it is an **open problem** to find an explicit bijection when  $n$  is not prime [see also Stanley EC1].



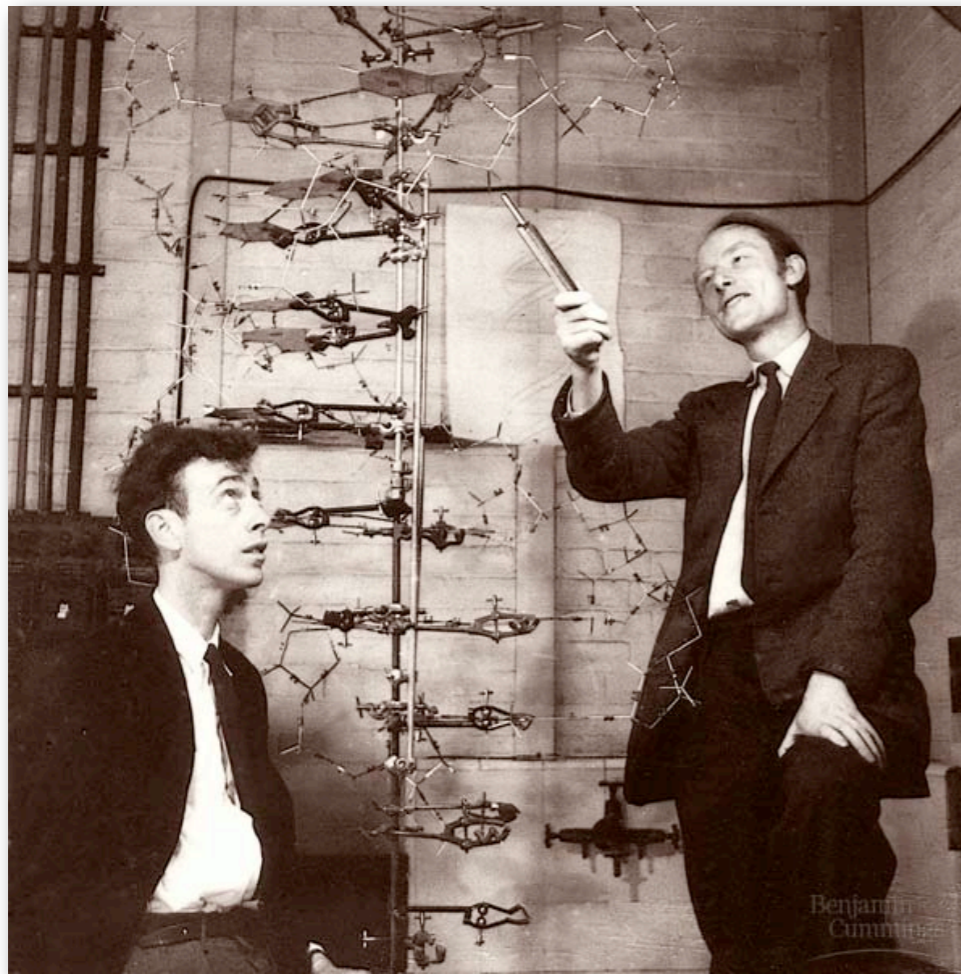
# Genetic chaos and a philosophical comment...

---

- In their paper, in a section titled “Applications”, Golomb, Gordon and Welch write:  
  
“ The reasonableness of this [comma free] condition can be seen if we think of the sequence of nucleotides as an infinite message, written without punctuation, from which any finite portion must be decodeable into a sequence of amino acids by suitable insertion of commas. If the manner of inserting commas were not unique, **genetic chaos would result.**”
  
- In fact, the theory is (completely) false. However it is an example of **interesting mathematics inspired by biology.**

# Deciphering the “genetic code”

---



“It has not escaped our attention that the specific pairing we have postulated immediately suggests a copying mechanism for the genetic material”

- **1953:** Watson, Crick and Franklin’s work leads to the proposed double helix structure for DNA
- **1954:** George Gamow founds the RNA tie club. The club consists of 20 regular members and 4 honorary members.
- **1955:** Francis Crick suggests the “adaptor hypothesis”.
- **1956:** Gamow thinks mathematically about the number of nucleotides necessary to make one amino acid.
- **1963:** Pauling and Zuckerkandl compare amino acid sequences thereby ushering in the era of comparative genomics.

# Pauling's (biological) idea [1963]

**TABLE II**  
SUBSTITUTION FREQUENCIES IN GLOBINS<sup>a</sup>

→ Substituent residue. Percentage of total residue sites of which the substituent occurs

	ALA	ARG	ASN	ASP	CYS	GLM	GLU	GLY	HIS	ILU	LEU	LYS	MET	PHE	PRO	SER	THR	TRY	TYR	VAL
ALA	•		28			31	33									31				
ARG		•						50				58				25				
ASN	33		•	47					33			33				33	33			
ASP	44		22	•			47	34	22			28				25				
CYS	(66)				•															
GLM				56		•	30	40				70								
GLU	50			44			•	38				47			24					
GLY	51			33			30	•				27			36					
HIS				28					•		26	30			22	22				
ILU	39									•	58									46
LEU	21									23	•	23		28						30
LYS	23	21		28			31	23			21	•			21					
MET	22									22	89		•	22						45
PHE									22		61			•						
PRO	50			43			57	43				21			•					
SER	49			24			24	36				24				•		40		
THR	32						28	24				24				52	•			
TRY	(40)										(40)			(60)				•		
TYR									(33)					(50)					•	
VAL	36									21	43	21								•

Residue site

Function is related to conservation

# Pauling's (biological) idea [1963]

**TABLE II**  
SUBSTITUTION FREQUENCIES IN GLOBINS<sup>a</sup>

→ Substituent residue. Percentage of total residue sites of which the substituent occurs

	ALA	ARG	ASN	ASP	CYS	GLM	GLU	GLY	HIS	ILU	LEU	LYS	MET	PHE	PRO	SER	THR	TRY	TYR	VAL
ALA	•		28			31	33									31				
ARG		•						50				58				25				
ASN	33		•	47					33			33				33	33			
ASP	44		22	•			47	34	22			28				25				
CYS	(66)				•															
GLM				56		•	30	40				70								
GLU	50			44			•	38				47			24					
GLY	51			33			30	•				27			36					
HIS				28					•		26	30			22	22				
ILU	39									•	58									46
LEU	21									23	•	23		28						30
LYS	23	21		28			31	23			21	•			21					
MET	22									22	89		•	22						45
PHE									22		61			•						
PRO	50			43			57	43				21			•					
SER	49			24			24	36				24				•		40		
THR	32						28	24				24					52			
TRY	(40)										(40)			(60)				•		
TYR									(33)					(50)						
VAL	36									21	43	21								•

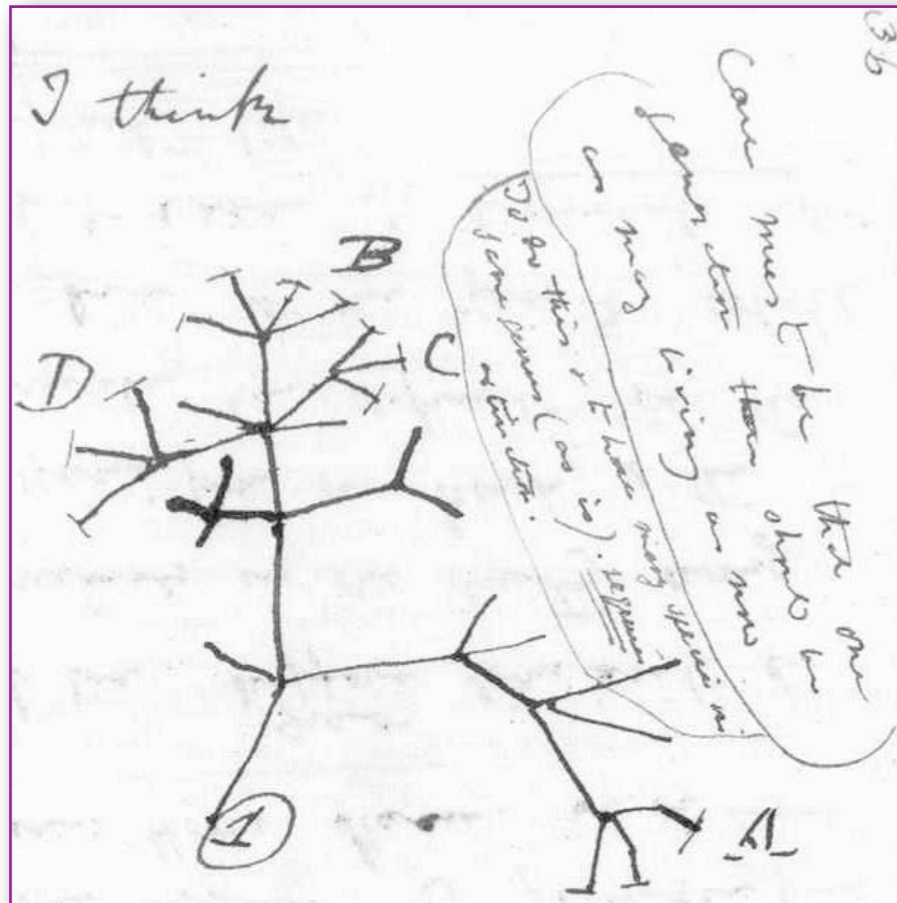
Residue site

Function is related to conservation

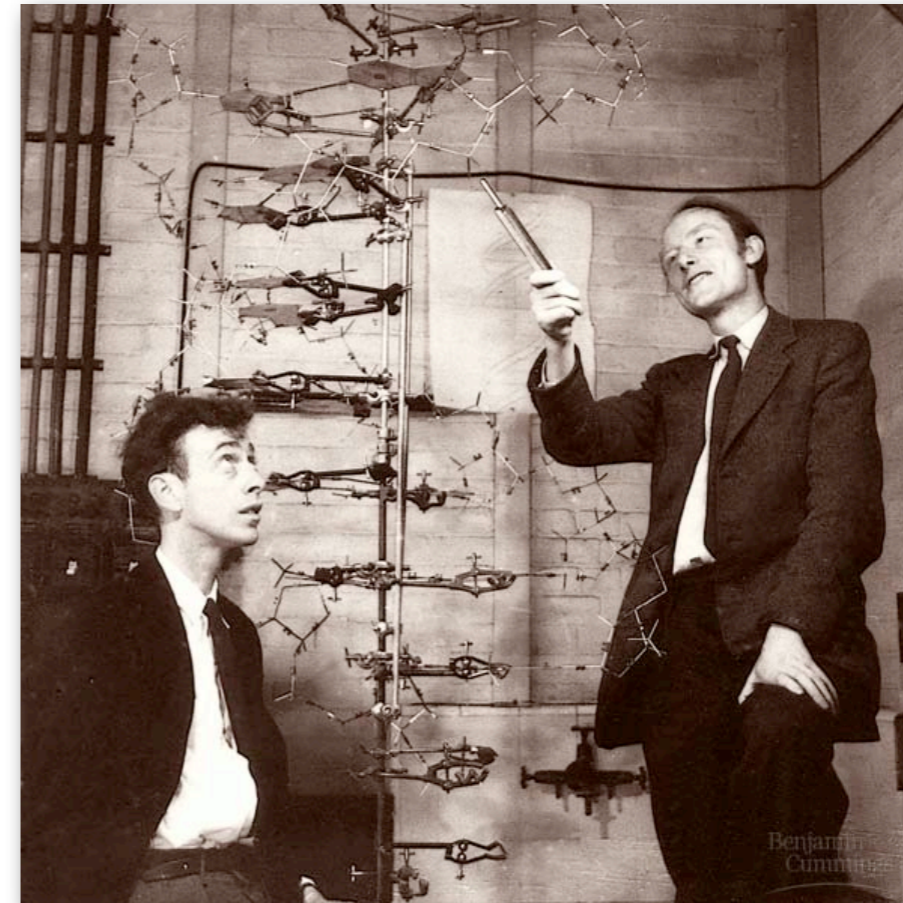
“Nothing in Biology Makes Sense Except in the Light of Evolution”

- Theodosius Dobzhansky (1973)

# Two unifying principles of biology

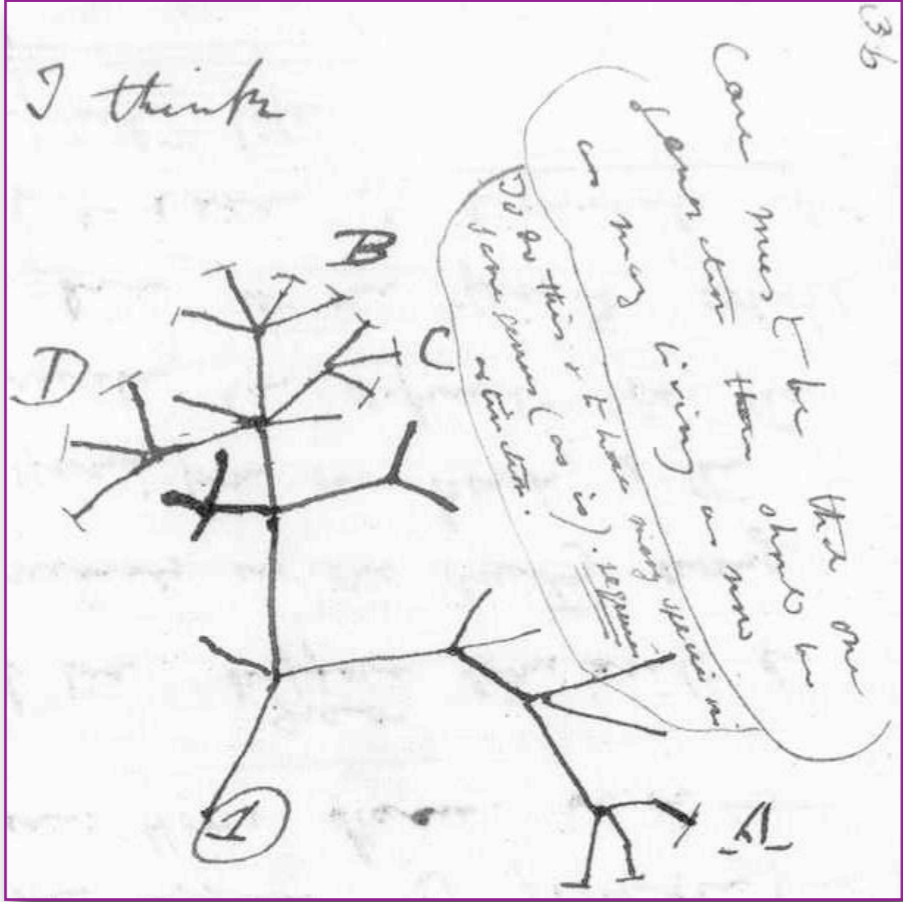


“Nothing in biology makes sense except in the light of evolution”



“It has not escaped our attention that the specific pairing we have postulated immediately suggests a copying mechanism for the genetic material”

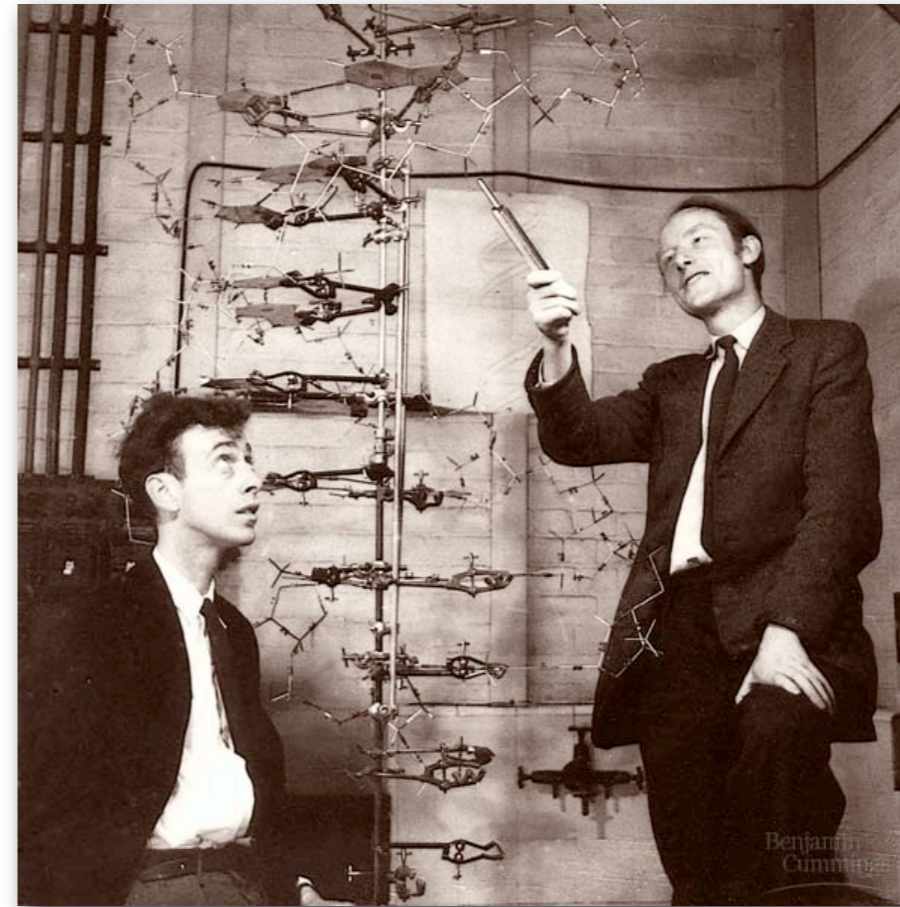
# Two unifying principles of biology



A handwritten sketch of a phylogenetic tree on a piece of paper. The tree has a central vertical trunk with several horizontal branches extending to the left and right. The branches are labeled with letters: 'A' at the bottom left, 'B' at the top left, 'C' at the top right, and 'D' at the middle left. To the right of the tree, there is a large, roughly drawn oval containing handwritten text. Above the tree, the words 'I think' are written. In the top right corner of the paper, the number '36' is written. The entire sketch is enclosed in a red rectangular border.

"Nothing in biology makes sense except in the light of evolution"

**An algorithm for a biological problem that reveals a solution to a math problem**

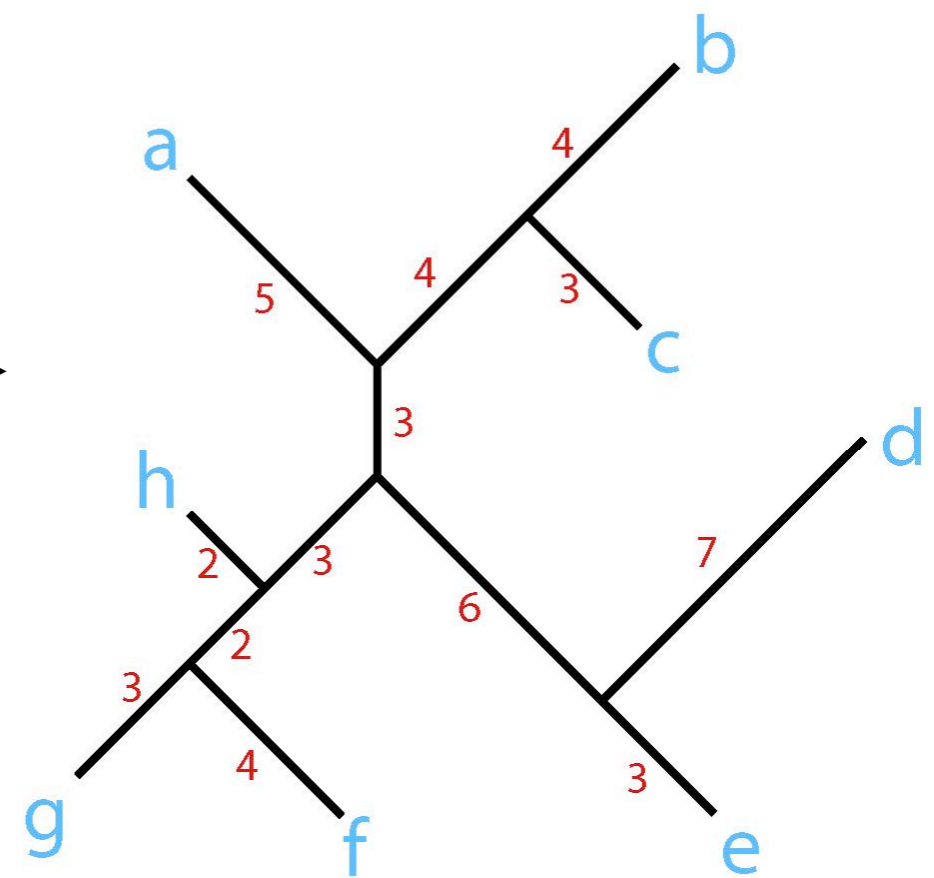


"It has not escaped our attention that the specific pairing we have postulated immediately suggests a copying mechanism for the genetic material"

# The neighbor-joining algorithm [Saitou & Nei, 1987]

- Input: dissimilarity map  $D = (d_{ij})_{i,j=1}^n$ .
- Output: tree  $T$  together with  $w : E(T) \rightarrow \mathbf{R}$ .

	a	b	c	d	e	f	g	h
a	0	13	12	21	17	17	16	13
b	13	0	7	24	20	20	19	16
c	12	7	0	23	19	19	18	15
d	21	24	23	0	10	22	21	18
e	17	20	19	10	0	18	17	14
f	17	20	19	22	18	0	7	8
g	16	19	18	21	17	7	0	7
h	13	16	15	18	14	8	7	0



# The neighbor-joining algorithm [Saitou & Nei, 1987]

- Given dissimilarity map  $D = (d_{ij})_{i,j=1}^n$ .

- Choose  $Q_{ab}$  that minimizes

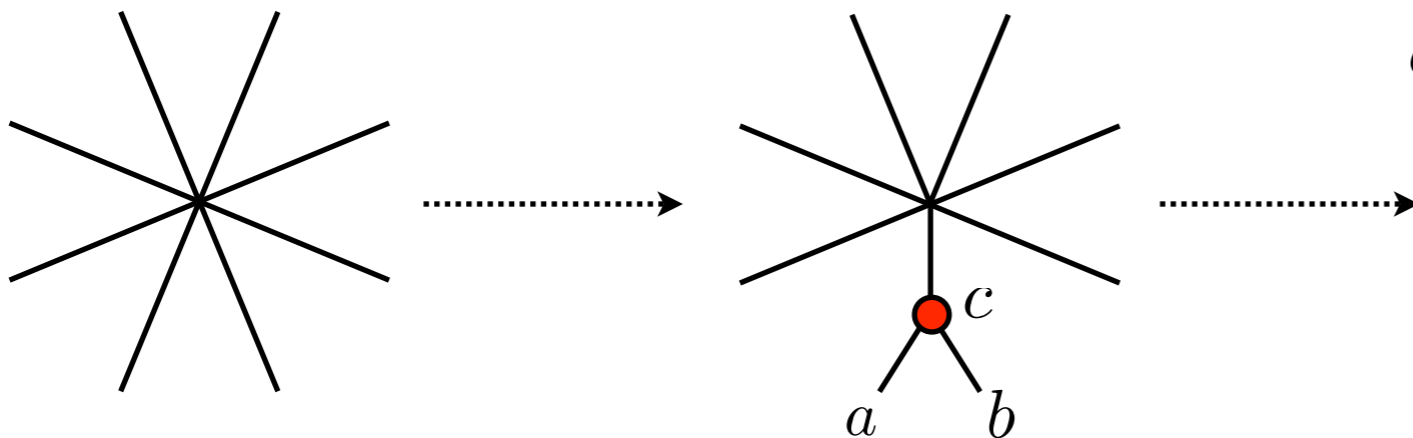
$$Q_{ij} = d_{ij} - \frac{1}{n-2} \left( \sum_k (d_{ik} + d_{jk}) \right).$$

- Replace a and b with a new leaf c and create a new dissimilarity map with

- $$d_{cx} = \frac{1}{2} (d_{ax} + d_{bx} - d_{ab}),$$

$$d_{ac} = \frac{1}{n-2} \sum_{k \neq b} \frac{1}{2} (d_{ak} + d_{ab} - d_{bk}),$$

$$d_{bc} = \frac{1}{n-2} \sum_{k \neq a} \frac{1}{2} (d_{bk} + d_{ab} - d_{ak}).$$





# The neighbor-joining algorithm [Saitou & Nei, 1987]

---

- Given dissimilarity map  $D = (d_{ij})_{i,j=1}^n$ .

- Choose  $Q_{ab}$  that minimizes

$$Q_{ij} = d_{ij} - \frac{1}{n-2} \left( \sum_k (d_{ik} + d_{jk}) \right).$$

- Replace a and b with a new leaf c and create a new dissimilarity map with



[Web](#) [Images](#) [Video](#) [News](#) [Maps](#) [more »](#)

neighbor joining

Search

[Advanced Scho](#)  
[Scholar Preferen](#)  
[Scholar Help](#)

**Scholar** All articles - [Recent articles](#)

[The neighbor-joining method: a new method for reconstructing phylogenetic trees -](#)

N Saitou - Molecular Biology and Evolution, 1987 - SMBE

Page 1. The **Neighbor-joining** Method: A New Method for Reconstructing Phylogenetic Trees' Naruya Saitou2 and Masatoshi Nei ... Page 2. **Neighbor-joining** Method 407 ...

[Cited by 16248](#) - [Related articles](#) - [Web Search](#) - [All 19 versions](#)

# The traveling salesman problem

---

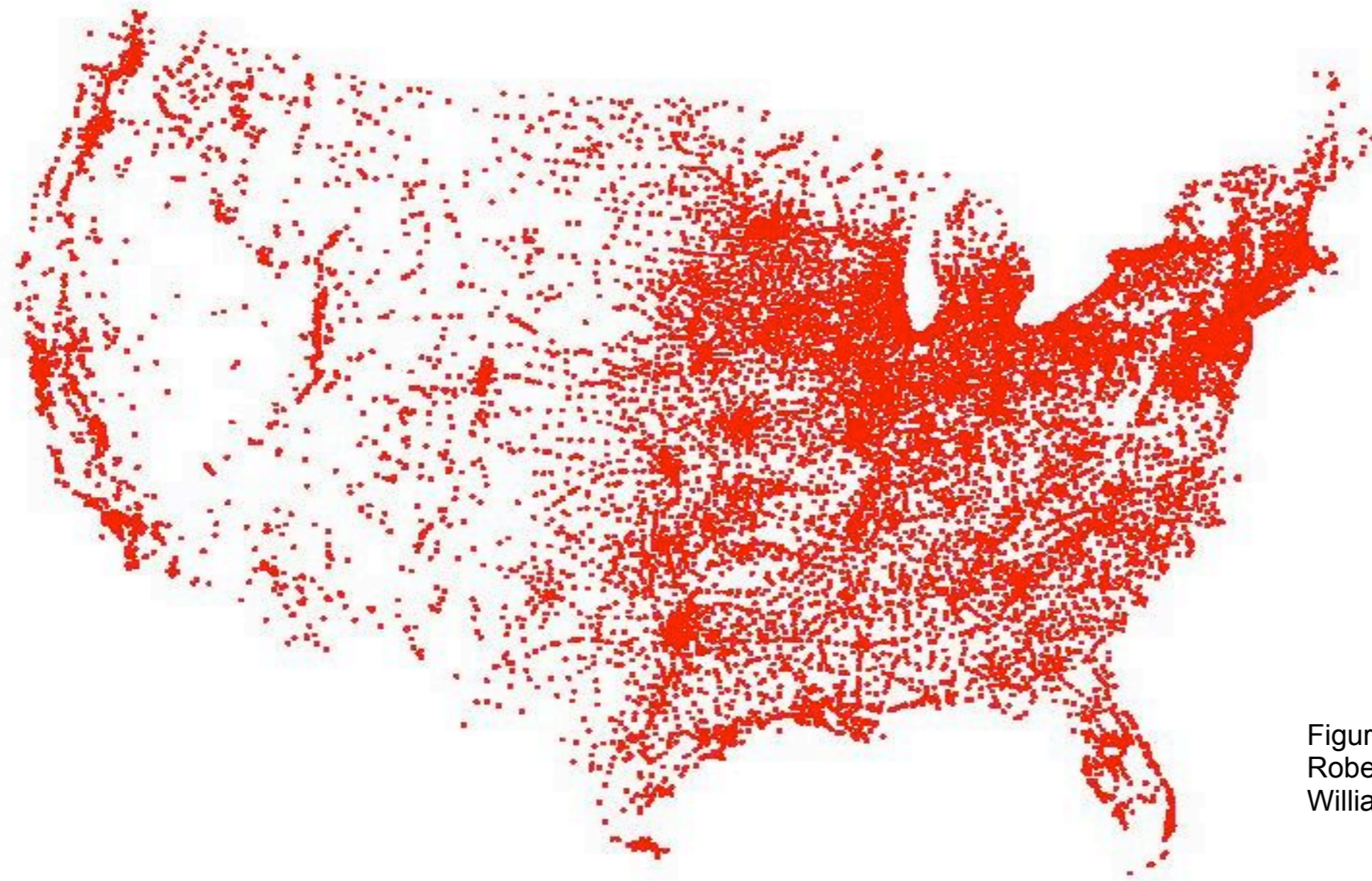


Figure credit: David Applegate,  
Robert Bixby, Vasek Chvatal and  
William Cook, 2002.

# The traveling salesman problem

---

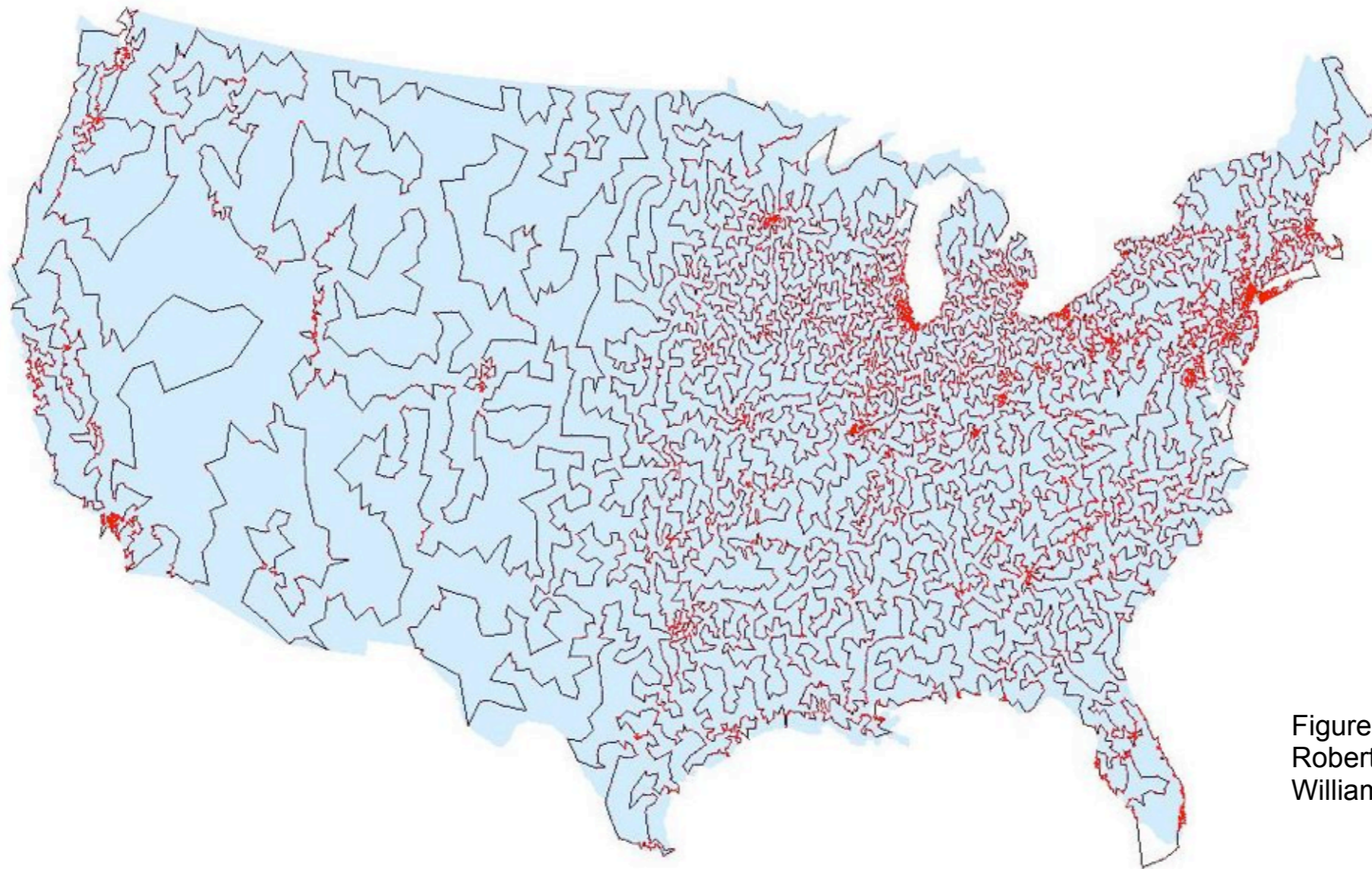
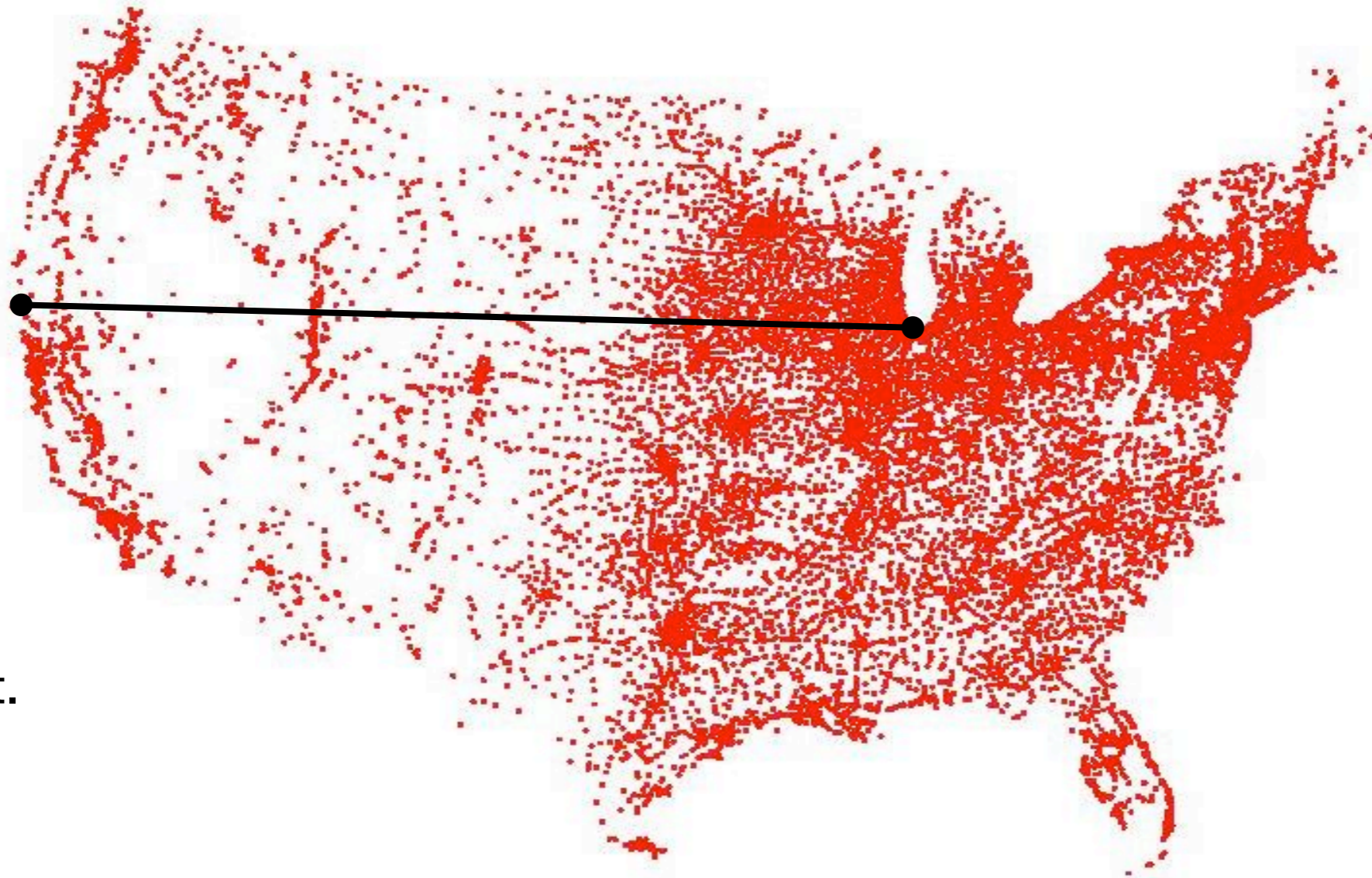


Figure credit: David Applegate, Robert Bixby, Vasek Chvatal and William Cook, 2002.

# A greedy algorithm

---

- Pick a pair of cities for which the average length of tours with those cities as neighbors is minimized:



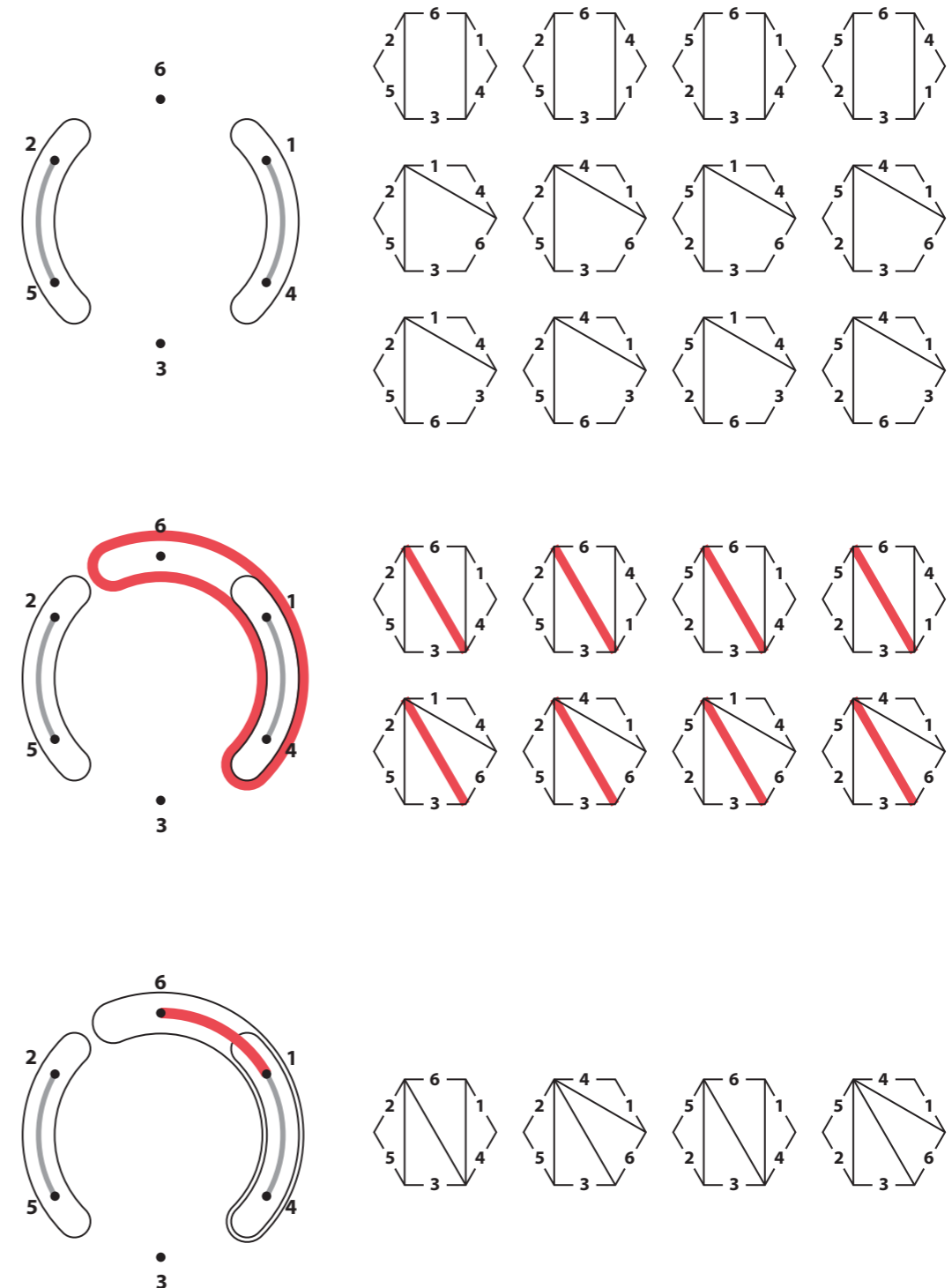
- Repeat.

# A greedy algorithm

- Fix the neighbors and repeat, each time minimizing the *average length* of the remaining possible tours.

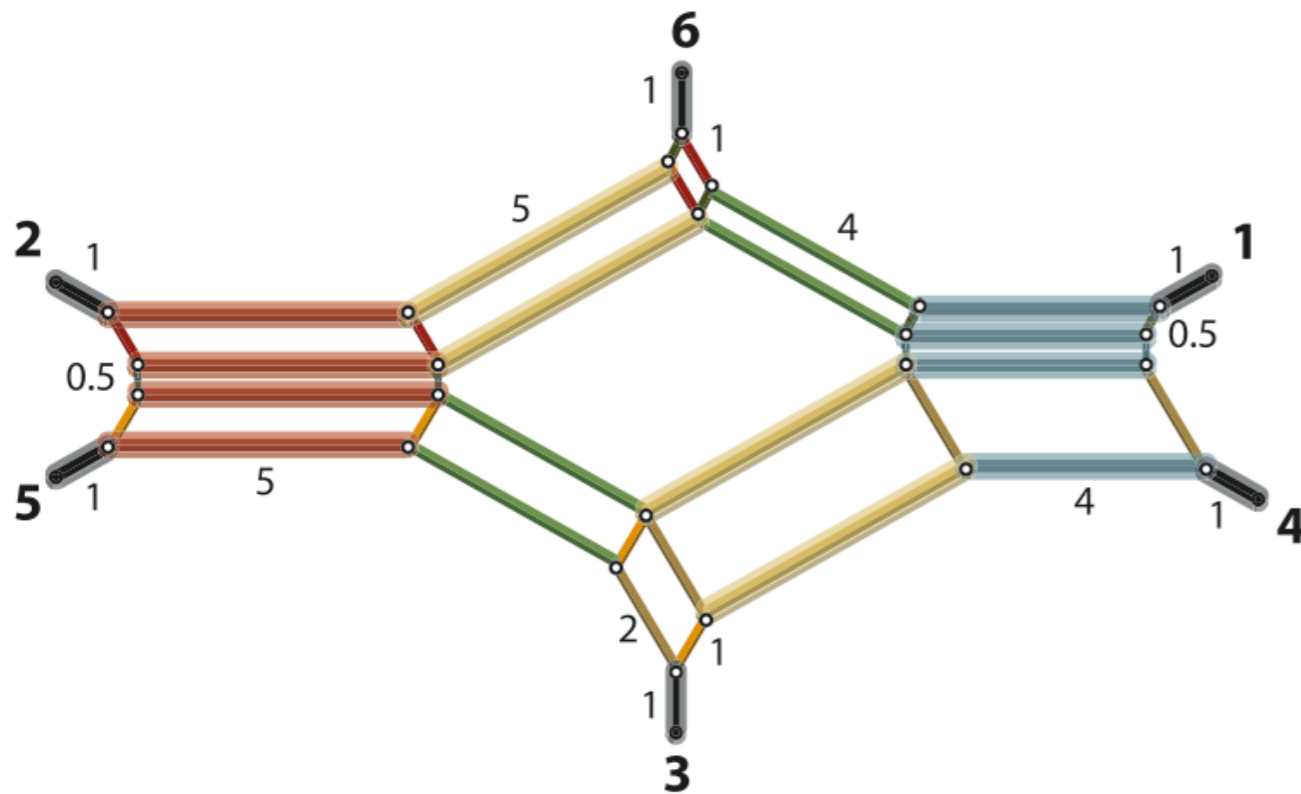
$$l(\delta, \mathcal{C}) := \frac{1}{|o(\mathcal{C})|} \sum_{(x_1, \dots, x_n) \in o(\mathcal{C})} \left[ \frac{1}{2} \sum_{i=1}^n \delta(x_i, x_{i+1}) \right]$$

- This algorithm is called the *neighbor-net algorithm*.



# Why neighbor-*net*?

Once a circular ordering (tour) has been established, weights can be assigned to the circular splits so that if the distance between cities is computed as the sum of the split weights that separate them, then these distances approximate the original distances.



	1	2	3	4	5	6
1	0	21.5	15	5	22	11
2	21.5	0	15.5	23.5	4.5	12.5
3	15	15.5	0	12	13	16
4	5	23.5	12	0	23	14
5	22	4.5	13	23	0	15
6	11	12.5	16	14	15	0

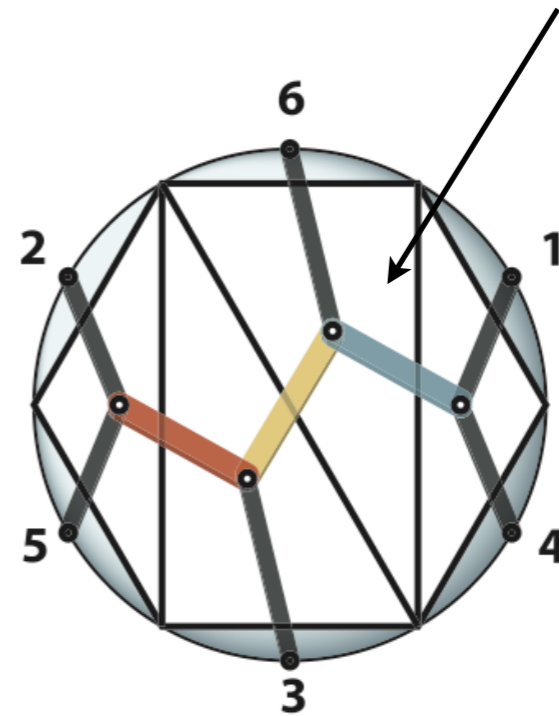
# Why **neighbor-net**?

A record of the steps with at which cities are joined produces a triangulation of the  $n$ -gon. This triangulation is in bijection with a tree. **This tree is the neighbor-joining tree.**

- Input: dissimilarity map.

	1	2	3	4	5	6
1	0	21.5	15	5	22	11
2	21.5	0	15.5	23.5	4.5	12.5
3	15	15.5	0	12	13	16
4	5	23.5	12	0	23	14
5	22	4.5	13	23	0	15
6	11	12.5	16	14	15	0

- Output: cyclic permutation and a tree .



# Neighbor-net can be implemented in time $O(n^3)$

---

- 1987: Saitou and Nei present the neighbor-joining algorithm for building phylogenetic trees that runs in  $O(n^3)$ .
- 1988: Studier and Keppler complete a proof that it is consistent.
- 2002: Bryant and Moulton adapt this algorithm to produce the  $O(n^3)$  of the neighbor-net algorithm just explained.
- 2007: Levy and P. explain that the algorithm of Bryant and Moulton is a greedy algorithm for the traveling Salesman problem.

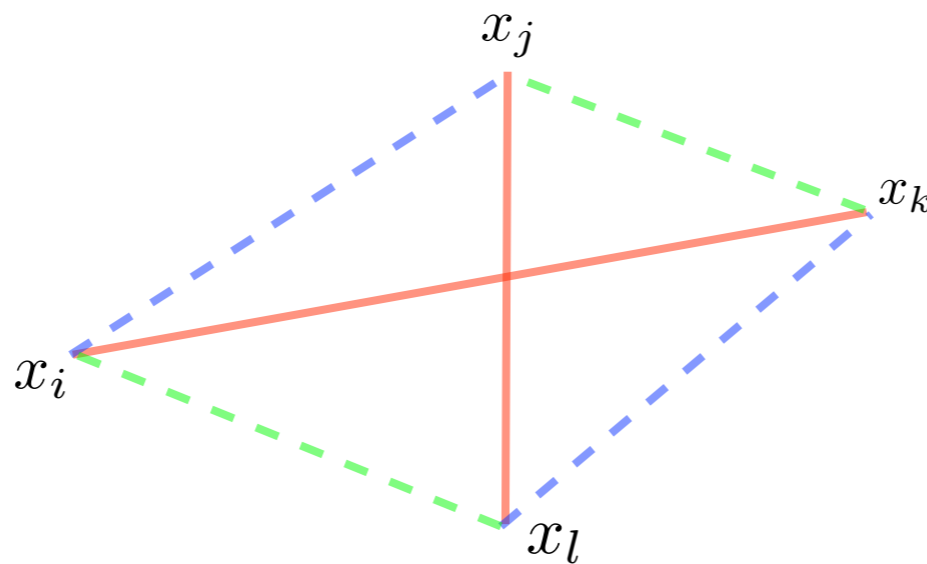


# The Kalmanson conditions

---

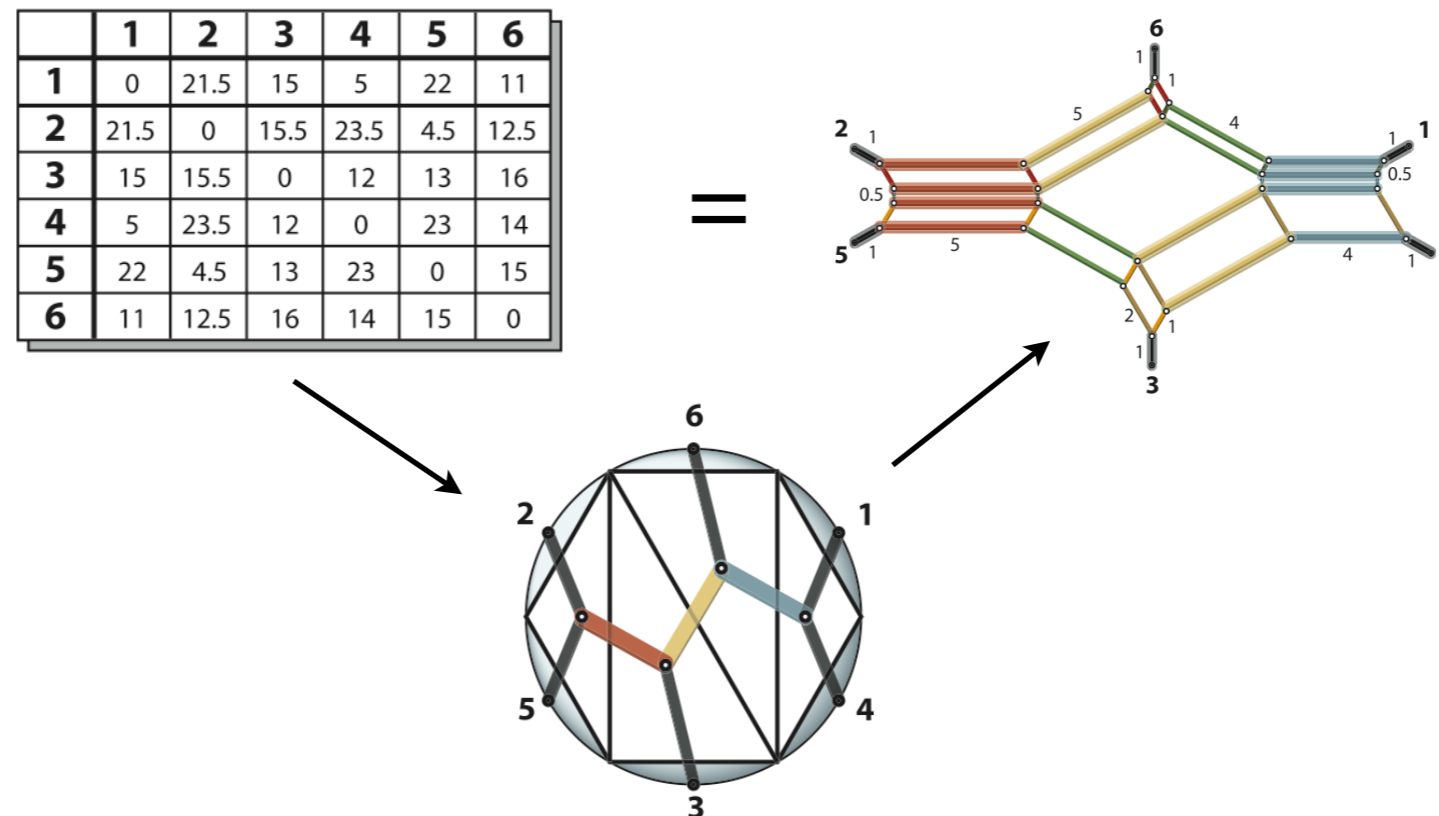
- A metric satisfies the Kalmanson conditions if for some circular permutation  $x_1, \dots, x_n$  it satisfies the parallelogram law:

$$\begin{aligned}\delta(x_i, x_j) + \delta(x_k, x_l) &\leq \delta(x_i, x_k) + \delta(x_j, x_l) \\ \delta(x_i, x_l) + \delta(x_j, x_k) &\leq \delta(x_i, x_k) + \delta(x_j, x_l).\end{aligned}$$



# Property of the neighbor-net algorithm

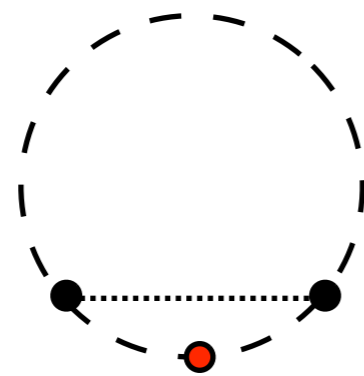
- **Theorem:** If a dissimilarity map satisfies the Kalmanson conditions with respect to some circular ordering then neighbor-net outputs the circular ordering.
- **Corollary:** neighbor-net is statistically consistent. Equivalently, neighbor-net provides optimal (and **robust!**) TSP solutions for Kalmanson matrices.



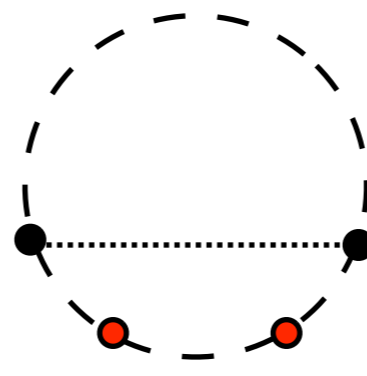
# Optimality of the neighbor-net algorithm

- It suffices to show that for Kalmanson matrices, at any step only neighbors in the circular ordering will be joined.

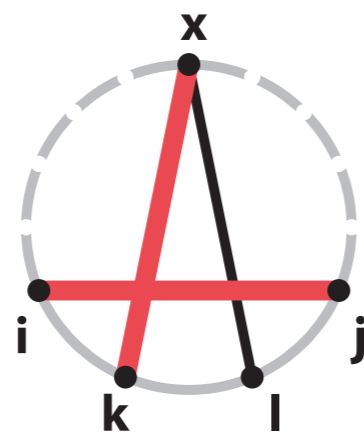
- Cases:



1 point in between



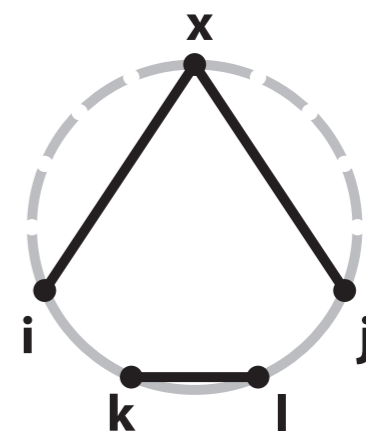
2 points in between



$\cong$

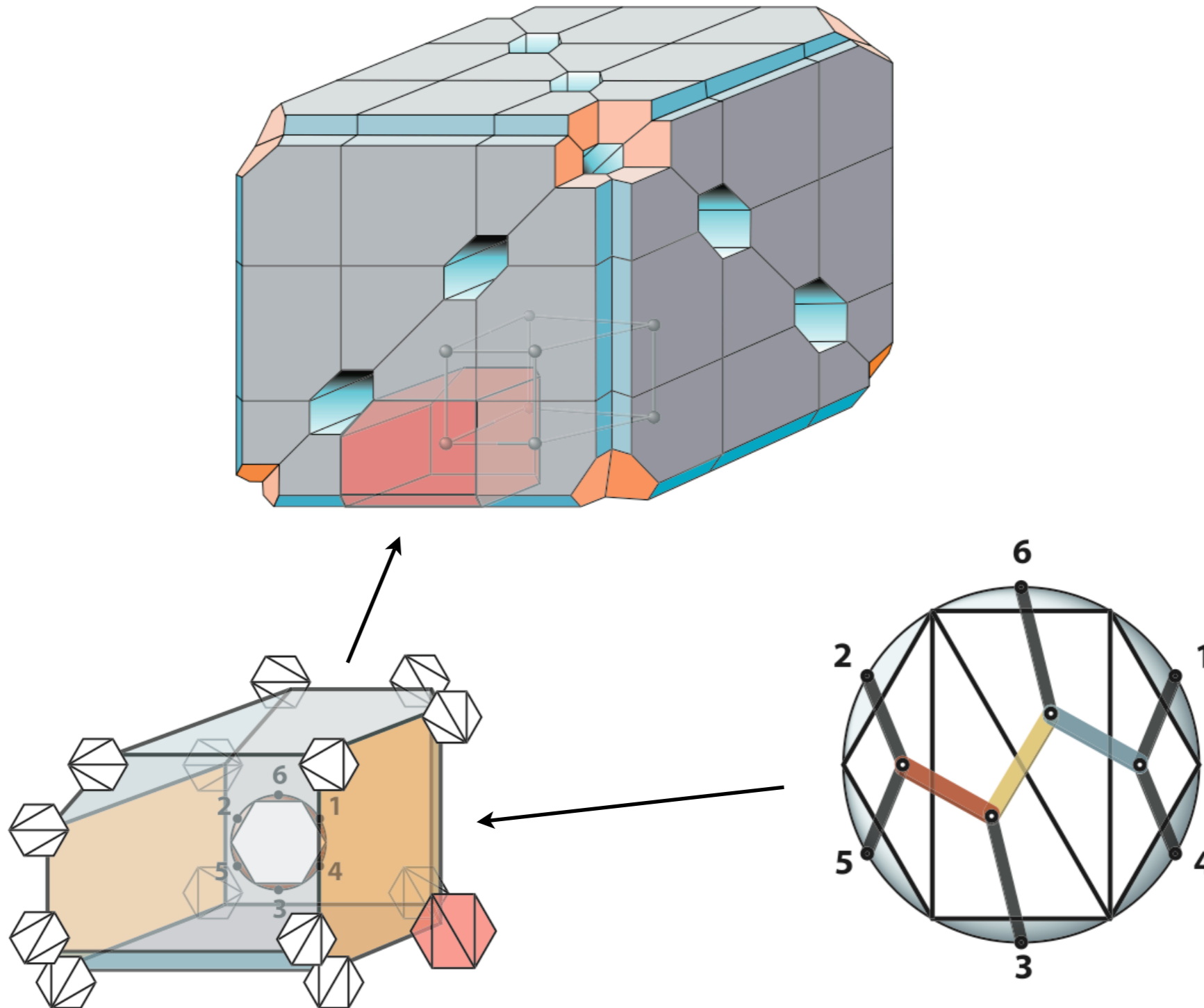


$\cong$



- If there are more than 3 points in between, the proof that there are two adjacent points that will be joined first is non-constructive.

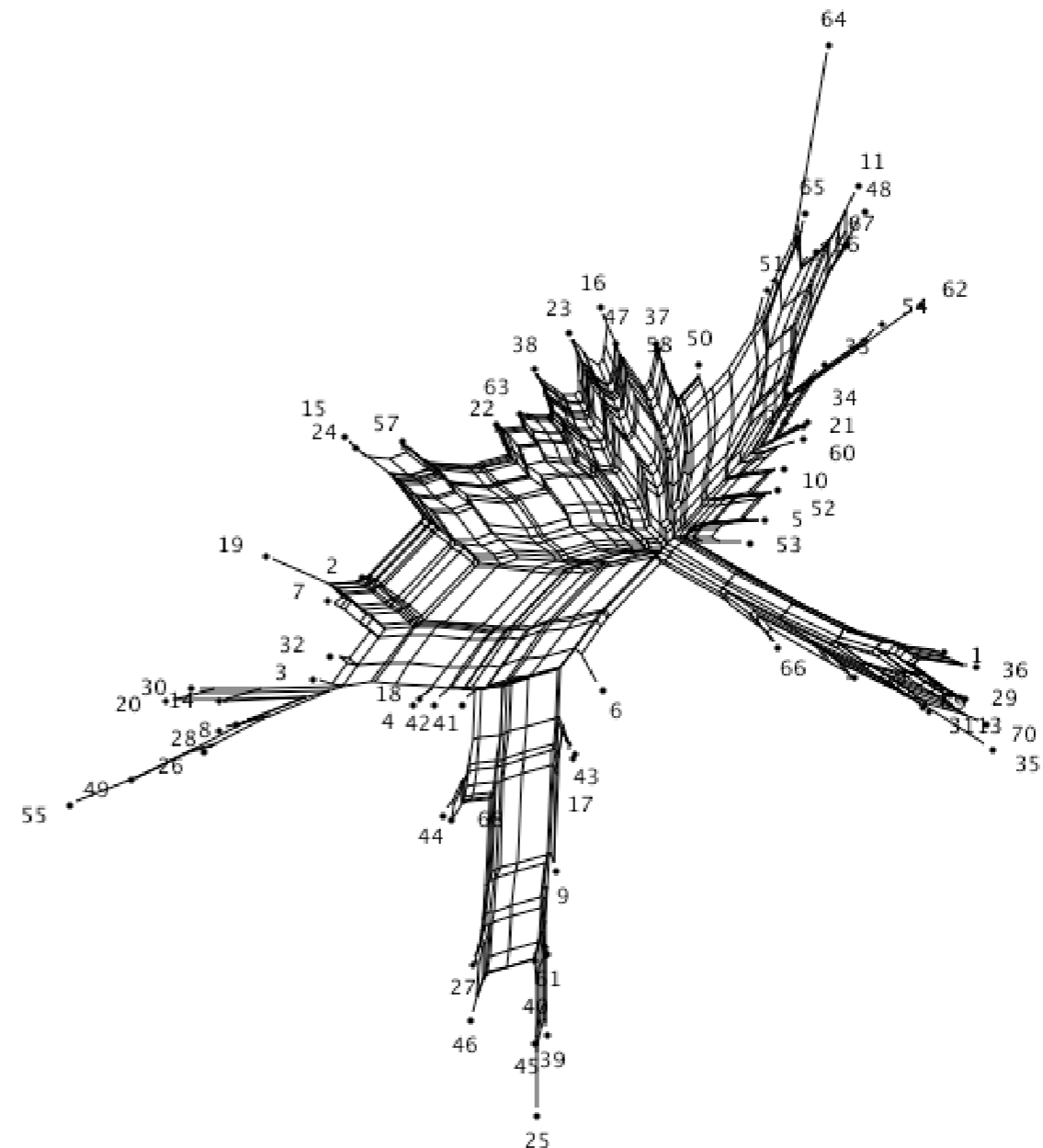
# $\overline{\mathcal{M}}_0^n(\mathbf{R})$ and the space of phylogenetic networks



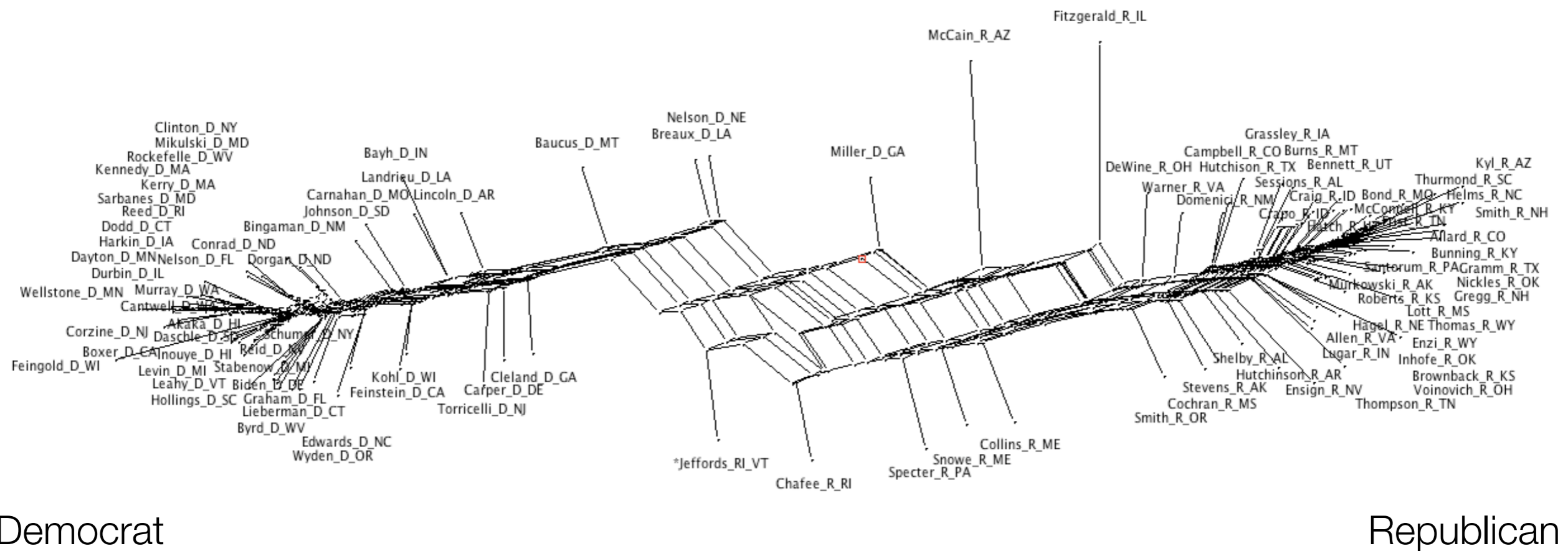
# Experiment: the traveling salesman problem

---

- Experimented with the 70 city st70.tsp problem from TSPLIB.
- Optimal tour has length 678.598.
- Neighbor-net tour has length 759.801.



# Example: votes of the United States senate



# Some conclusions

---

- Mathematics, computer science and statistics have emerged as essential sciences for biology. **At the same time, biological problems can offer interesting opportunities for mathematics.**
- In this talk I have focused on aspects of biology which touch on combinatorics, but there are many similar stories from other branches of biology and in other areas of mathematics.
- Biologists may not always have formal math training (on average), yet they respect mathematics and have on many occasions discovered beautiful mathematics. In his autobiography, **Darwin wrote “I have deeply regretted that I did not proceed far enough at least to understand something of the great leading principles of mathematics; for men thus endowed seem to have an extra sense.**