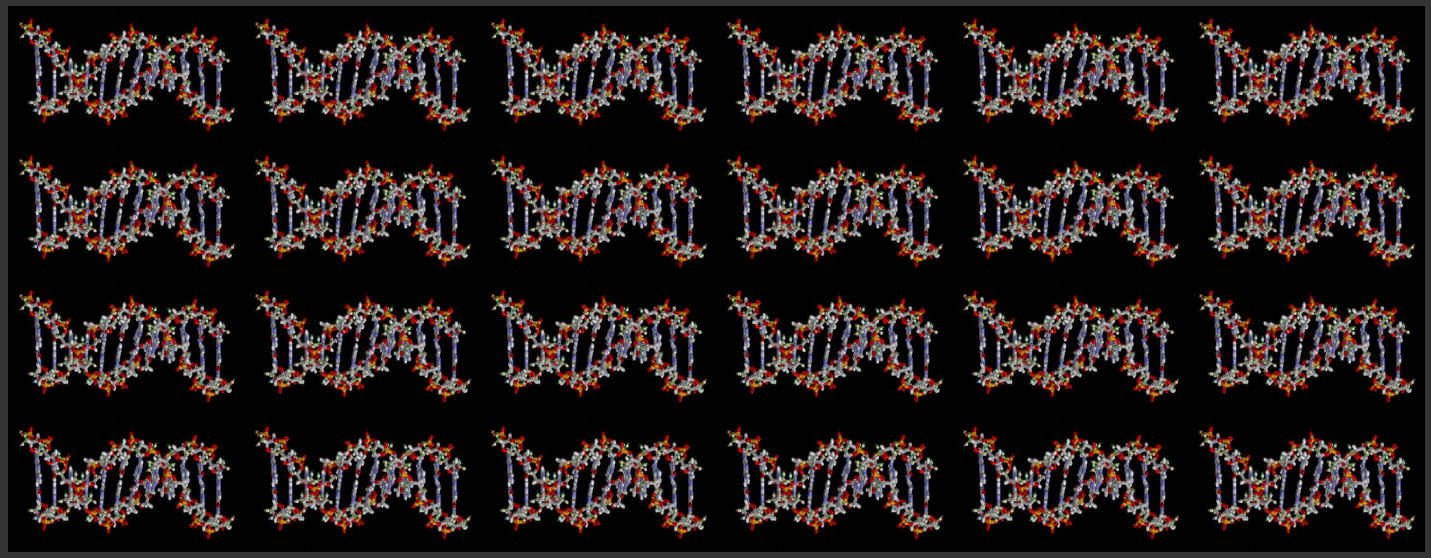
Molecular sequence representations

- or -

Time for some actual computer science



CSCI 4181 / 6802

https://giphy.com/gifs/dna-l1fWtMmQbuGvm

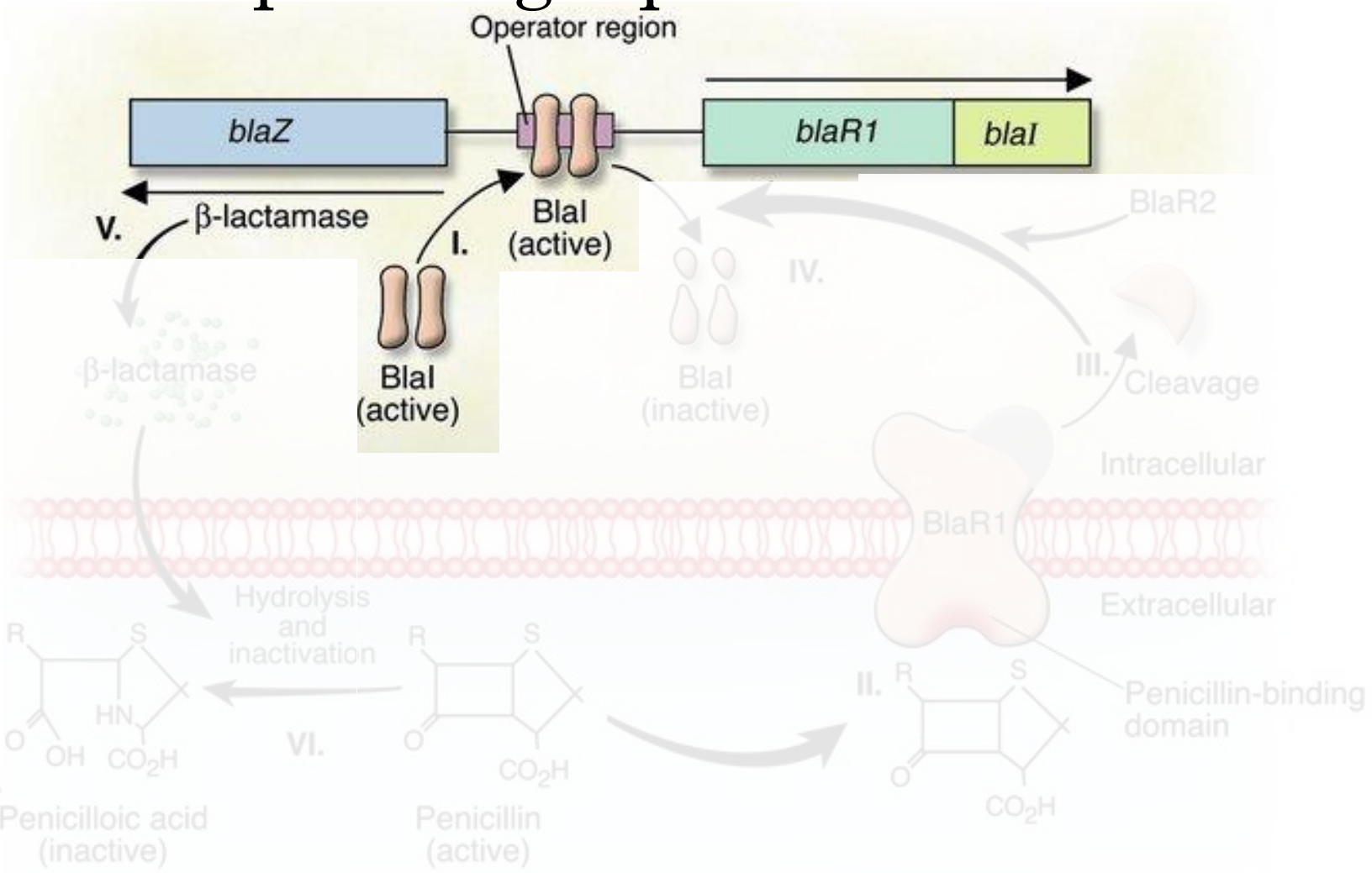
Overview

1. Representing biological information is HARD
2. Sequence representations
3. Structure representations
4. Even more representations (e.g., gene presence / absence)

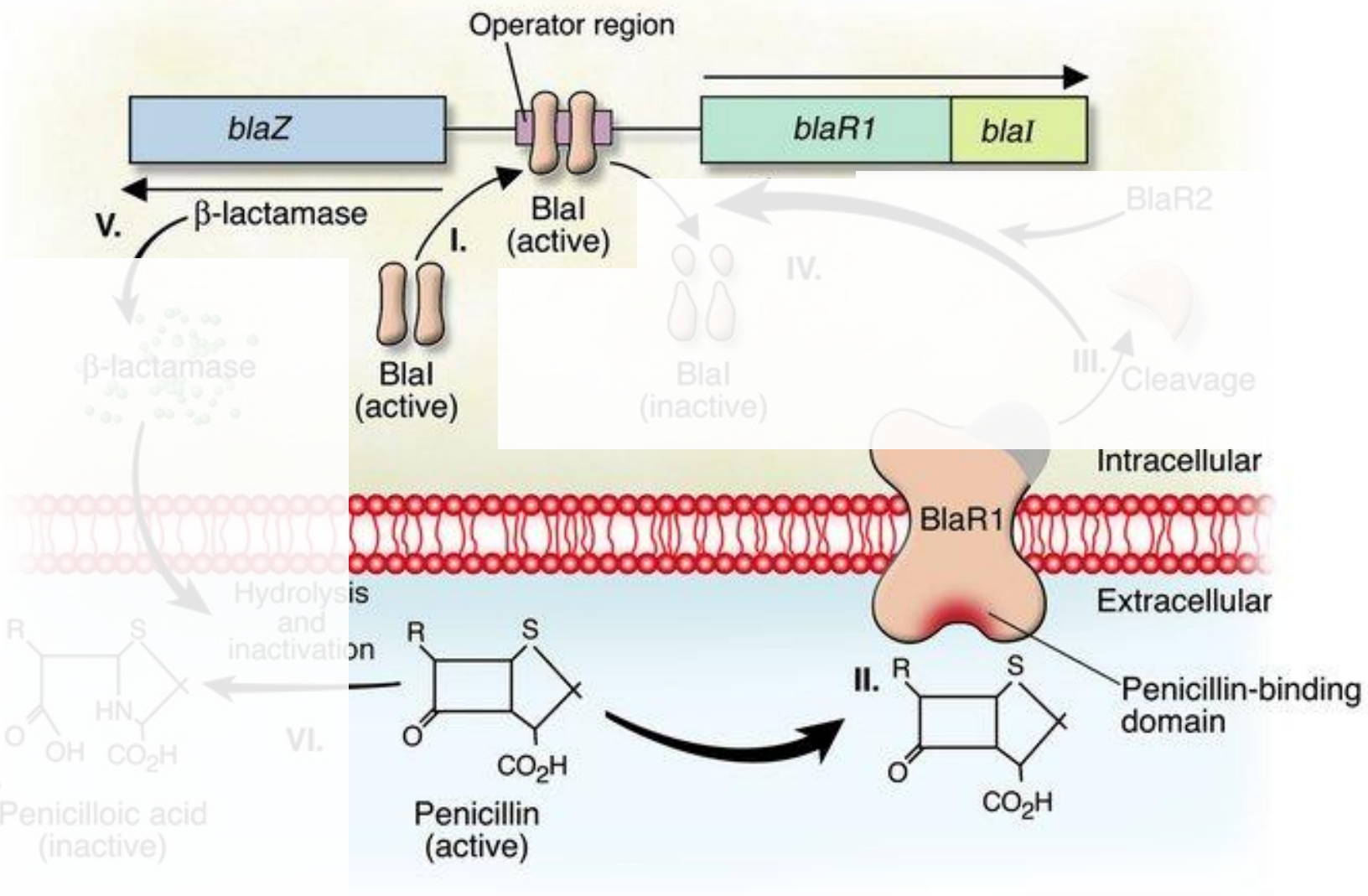
Self-defence in several easy steps

Lowy (2013) *J Clin Invest*

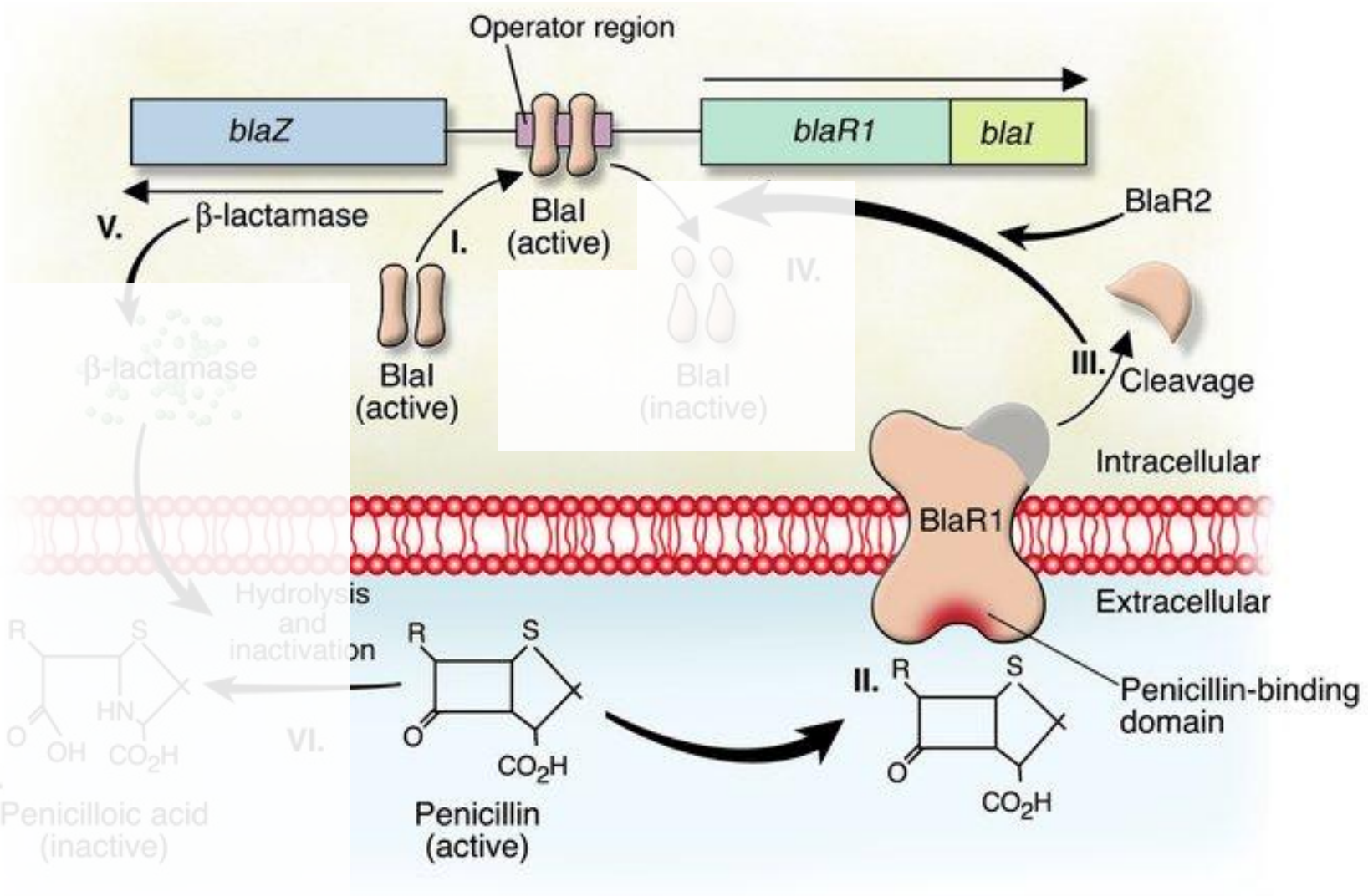
Safe and sound – BlaI keeps things quiet



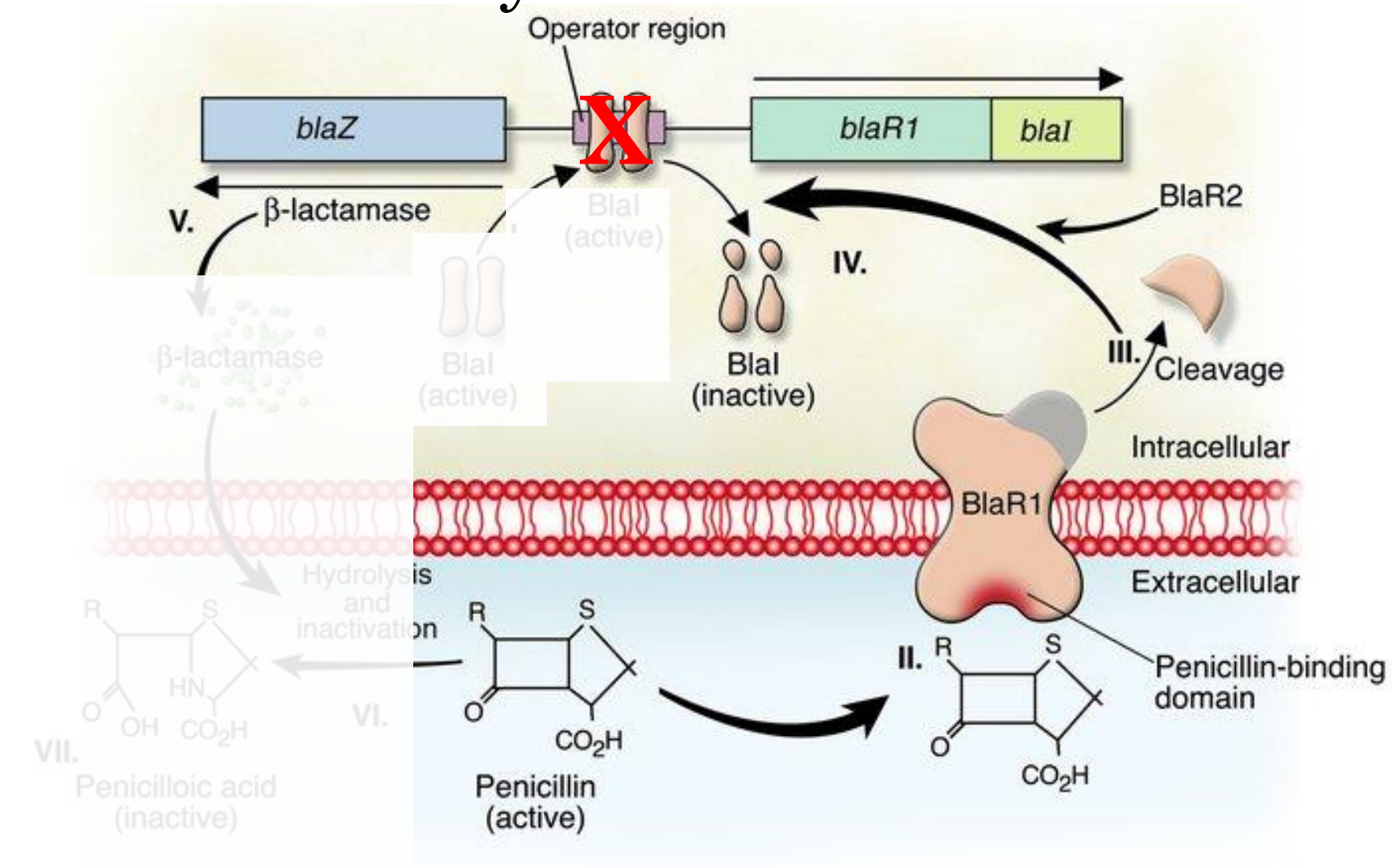
Danger!
Penicillin is in the air



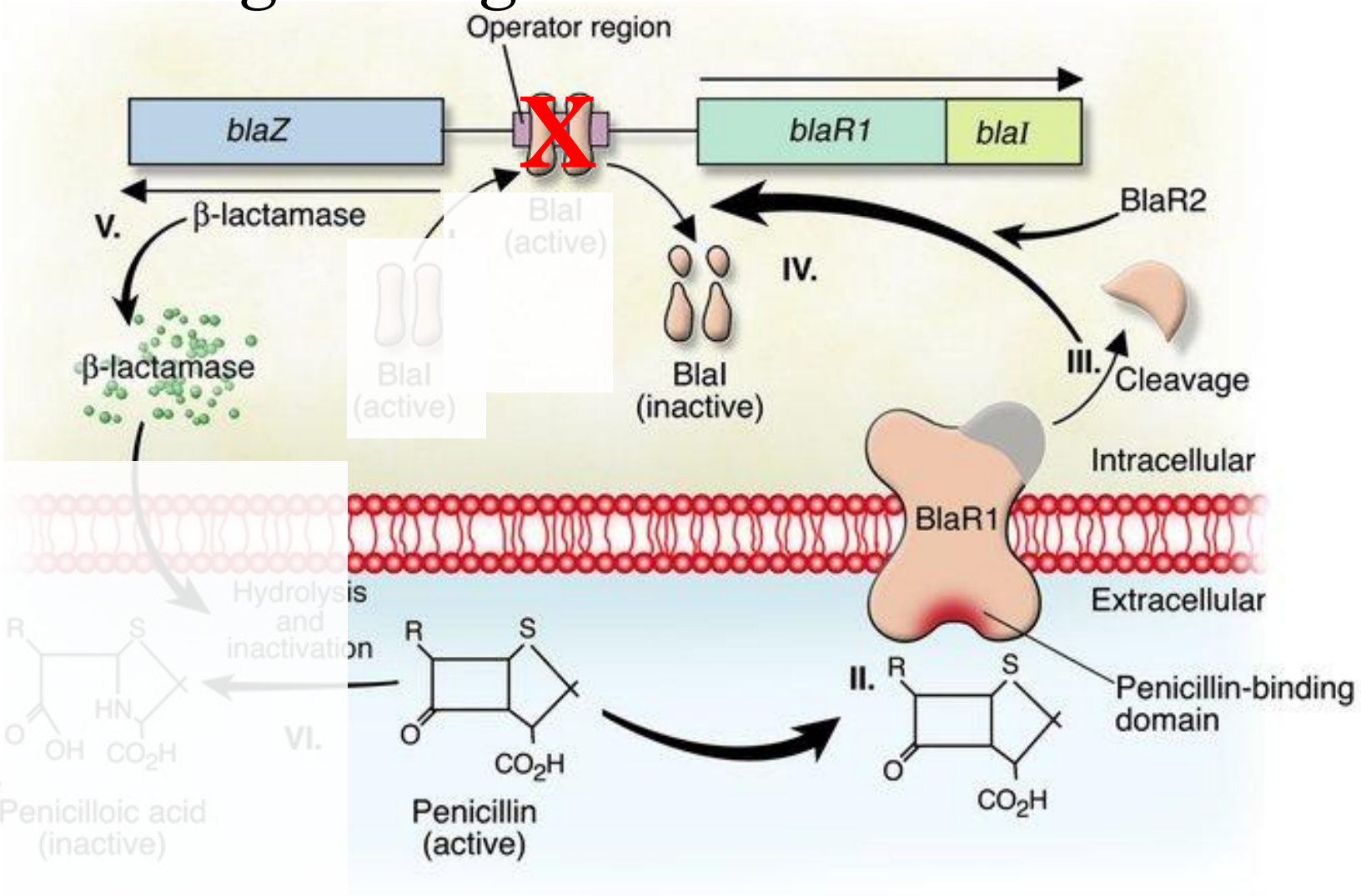
Oh snap!


Operator region
blaZ
blaR1
blaI
β-lactamase
V.
I.
BlaI (active)
BlaR2
III.
Cleavage
BlaI (active)
BlaI (inactive)
IV.
Intracellular
BlaR1
Extracellular
Penicillin-binding domain
II.
Penicillin (active)
Penicilloic acid (inactive)
VI.
VII.

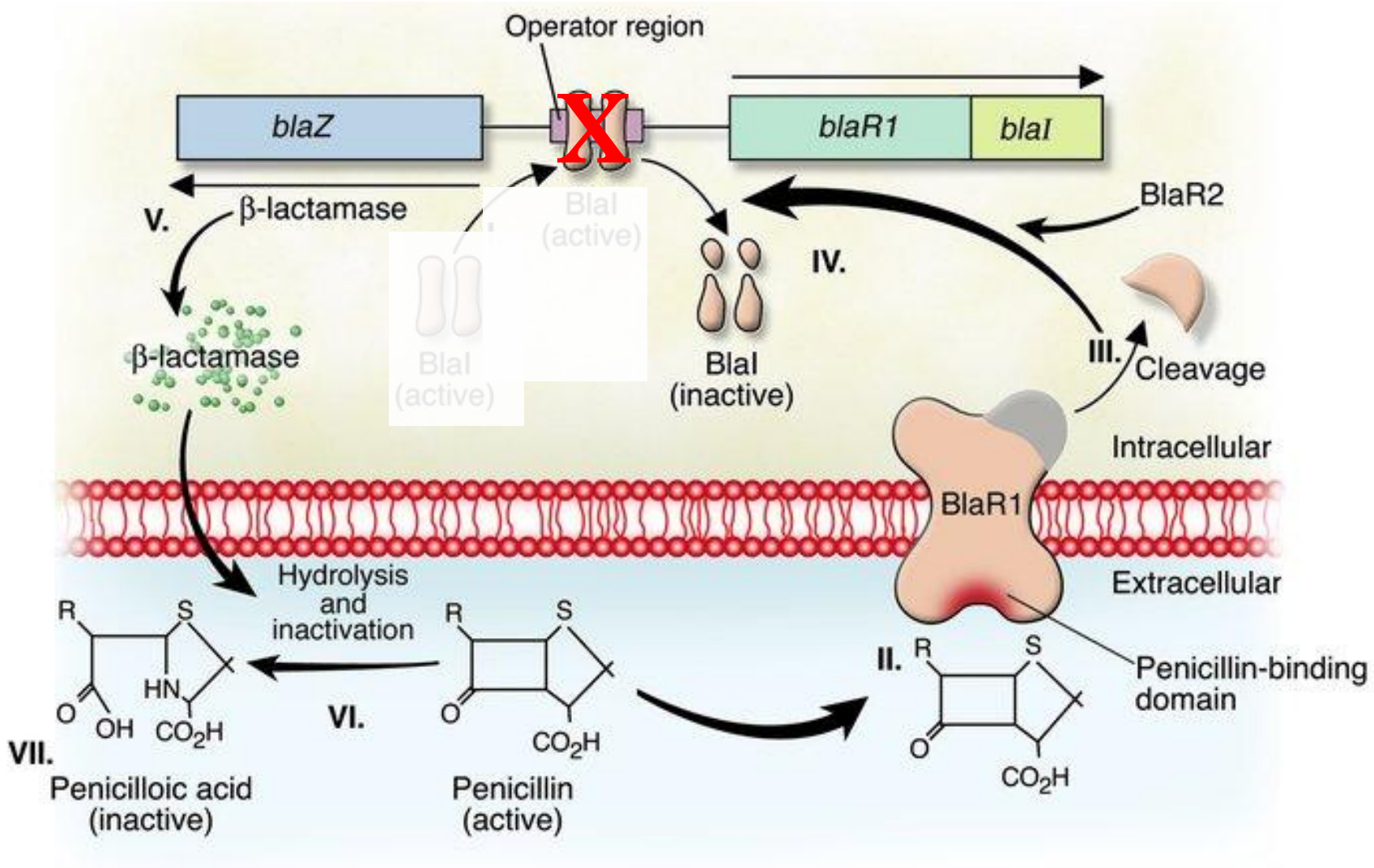
Destroy the repressor, activate the system



Transcribe and translate
- The beginning of the end



Export and destroy



???

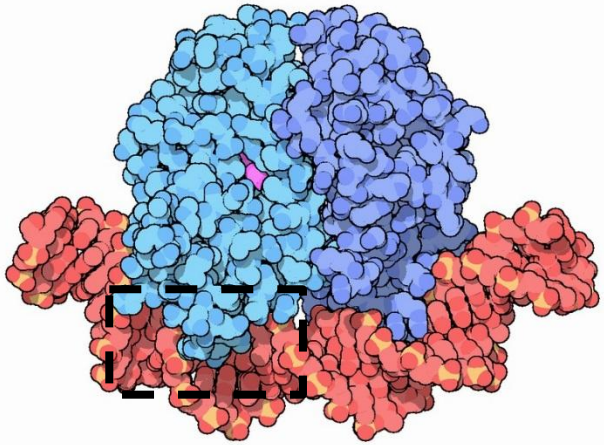
Comparison / classification / analysis

BlaI repressor bound to DNA
https://www.rcsb.org/3d-view/1XSD/1

Goals of representation

Identify functional patterns in DNA or protein sequences

Catabolite activator protein (CAP or CRP)



atatgcctgacggagttcacacttgtaagtttcaactacg
t
attcagtacaaaacgtgatcaacccctcaatttccccttgc
t
tcgctttgtcagctgtgacaagctccgcaaatcgtgacaat
a
aaaaacatttagagtgatatgtataacattatggcgttta
t
caatctccgcgagcgtgccagttttcacattcttcagttgc
a
cgcacattggtataacgtgatcatatcaacagaatcaata
a
tgggcagcttcttcgtcaaatttatcatgtggggcatcctt
a
tttcactaaaaagtgtgatcggggacaatatatttacgcac
g
taggtgctttttgtggcctgcttcaaacgtttcgcccctcc
t
tgaatgcgccaaactgtgatagtgtcatcatttcaaaagcgt

CAP-lacking sequences

ahblahblahblahblahblahblahbla
ahblahblahblahblahblahblahbla
ahblahblahblahblahblahblahbla
h
blahblahblahblahblahblahblahblahblahblahbla
h
blahblahblahblahblahblahblahblahblahblahbla
h
blahblahblahblahblahblahblahblahblahblahbla
h
blahblahblahblahblahblahblahblahblahblahbla
h

CAP binding sites

http://en.wikipedia.org/wiki/File:48-CataboliteActivatorProtein-1cgp.tif

Goals of representation

Distinguish phenotypes based on sequence or structure variation

e.g., huntingtin

GCTGCCGGGACGGGTCCAAGATGGACGGCCGCTCAGGTTCTGCTTTTACCTGCGGCCCAGAGC
CCCATTCATTGCCCCGGTGCTGAGCGGCGCCGCGAGTCGGCCCGAGGCCTCCGGGGACTGCCG
TGCCGGGCGGGAGACCGCCATGGCGACCCTGGAAAAGCTGATGAAGGCCTTCGAGTCCCTCAA
GTCCTTC**CAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCA**
GCAGCAGCAACAG

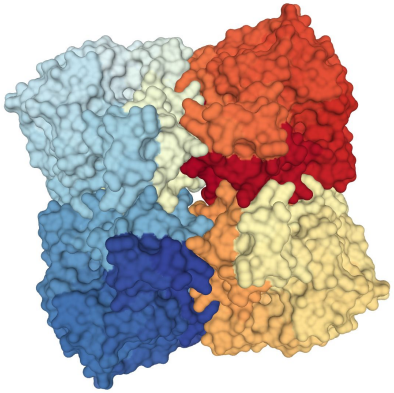
(About 10,000 more nucleotides in the gene)

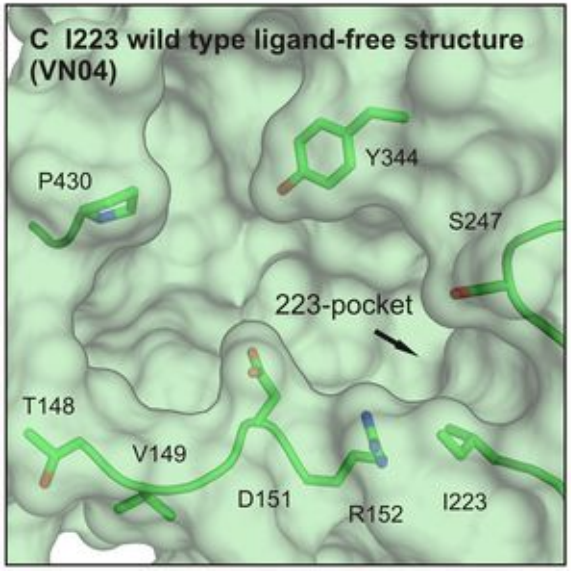
# of CAG repeats	Effect
< 27	Healthy
27-35	Intermediate
36-39	Disease (reduced penetrance)
> 39	Full disease effects

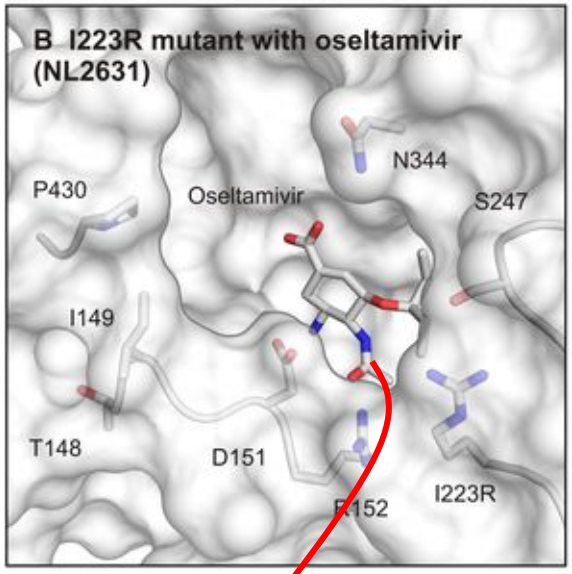
Goals of representation

Identify important changes at the sequence and structural level

e.g. oseltamivir resistance in influenza H1N1 Neuraminidase







Doesn't fit!

van der Vries et al (2012) *PLoS Pathogens*

Sequence representations

https://www.nist.gov/industry-impacts/accuracy-genetic-testing-technologies

Biological Sequences

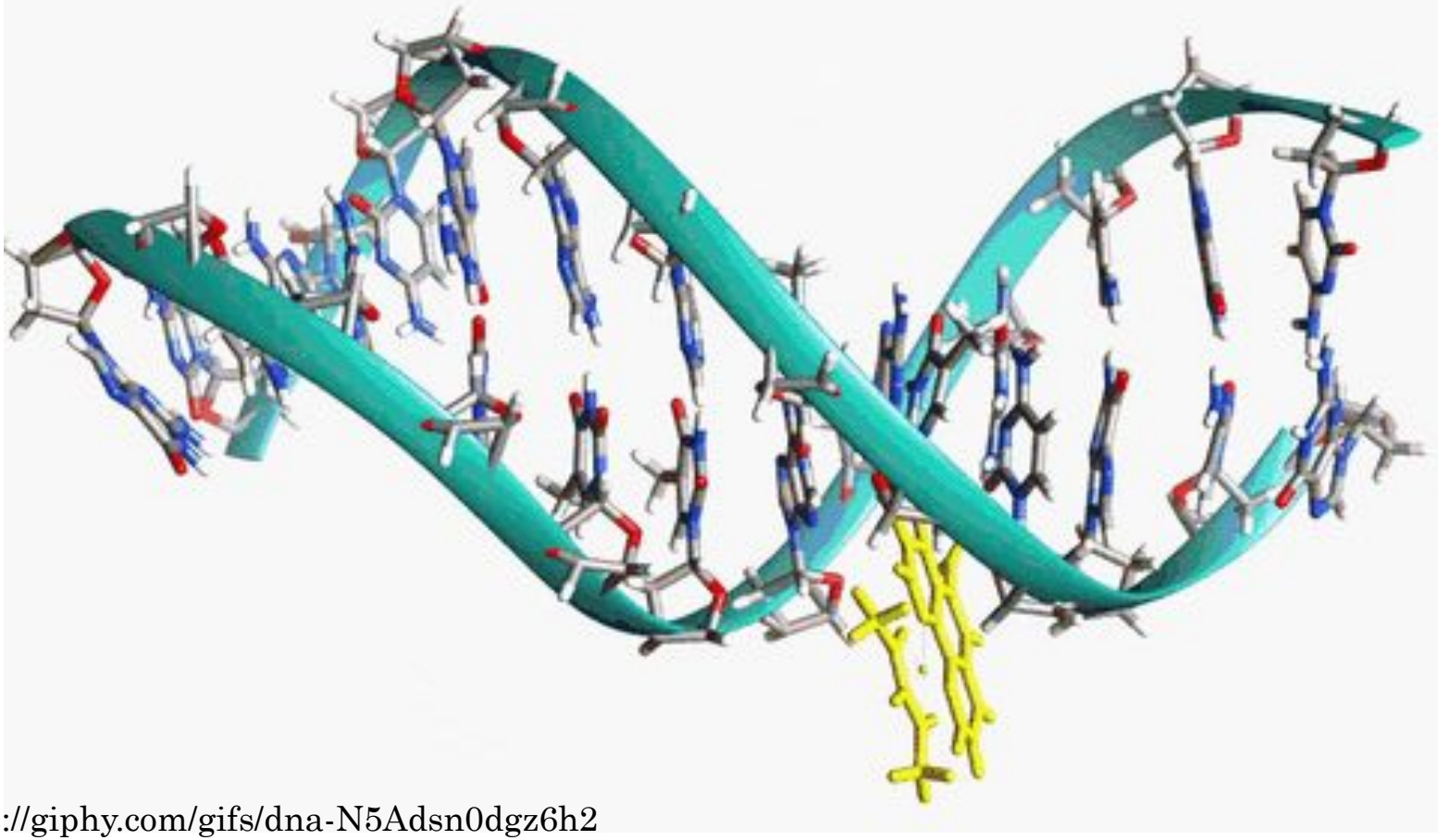
Primary structure (higher structures describe the three-dimensional features of molecules)

DNA **…ACCGAATTTACGATACATG…**

Protein **…MLQELIVNEW…**

Sequence Representations of DNA

Convert linear, double-stranded DNA into representation(s) that comprise a *feature set*



https://giphy.com/gifs/dna-N5Adsn0dgz6h2

The most common representation is (as you have already seen) a STRING representation with an alphabet of four letters

{A,C,G,T}

Sparse or ‘one-hot’ encodings are typically used

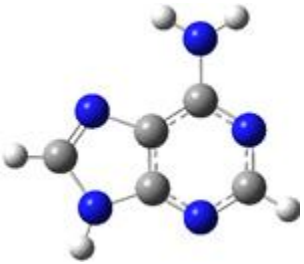
A = {1,0,0,0}

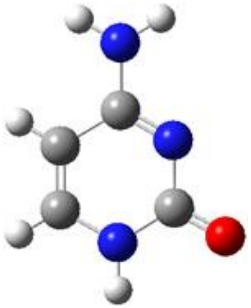
Why should we prefer this to the more compact:

A = {1}, C = {2}

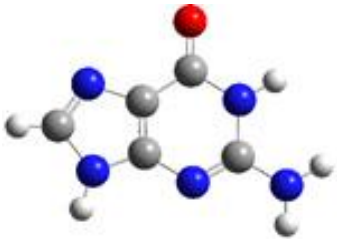
Degenerate characters

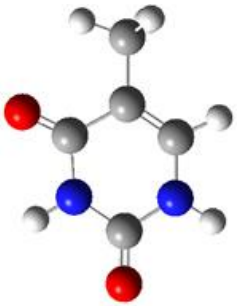
Every pair of nucleotides has something in common





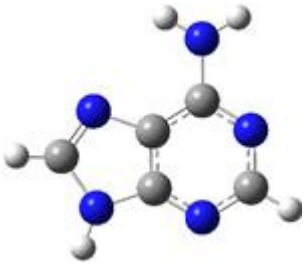
A	G
C	T

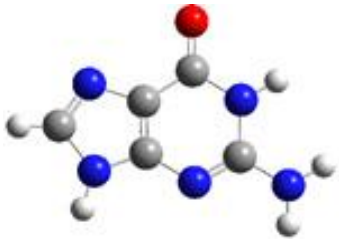


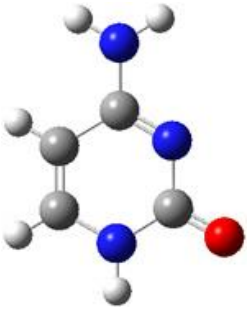


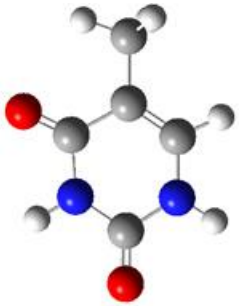
Degenerate characters

Every pair of nucleotides has something in common

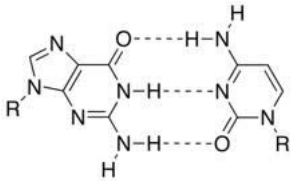




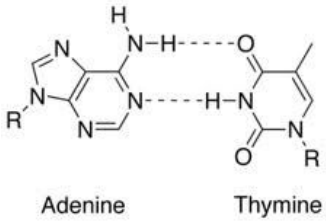




A	G
C	T

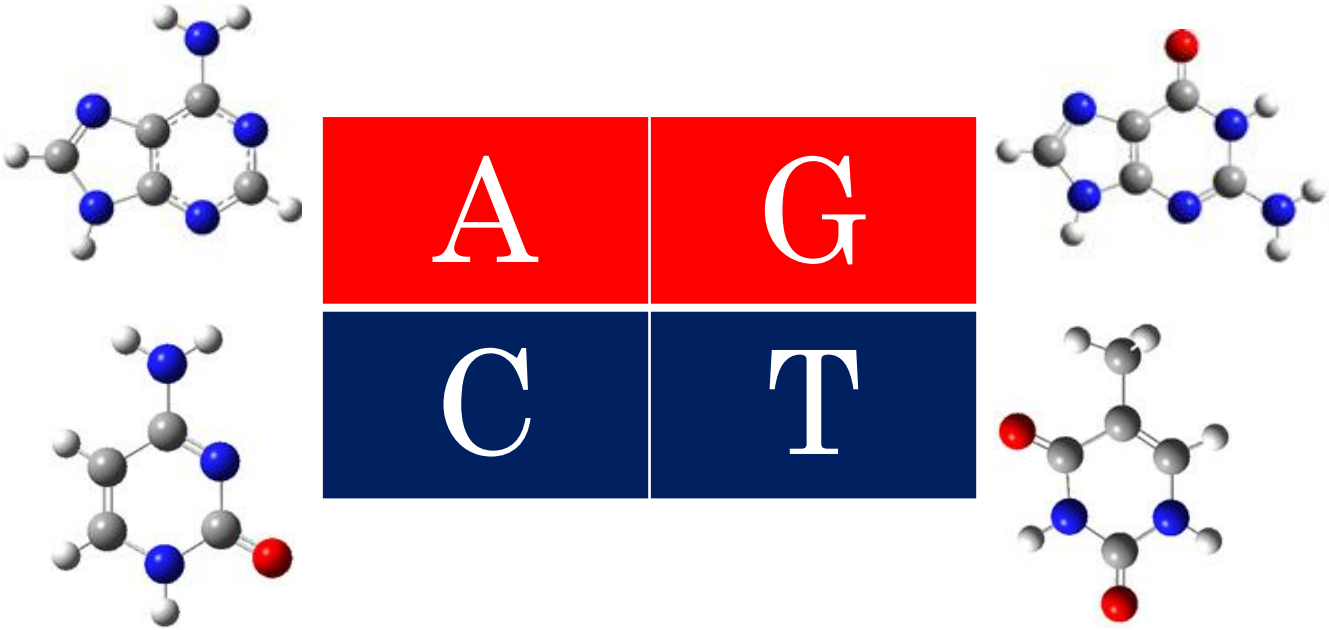


STRONG vs. WEAK base pairing



Degenerate characters

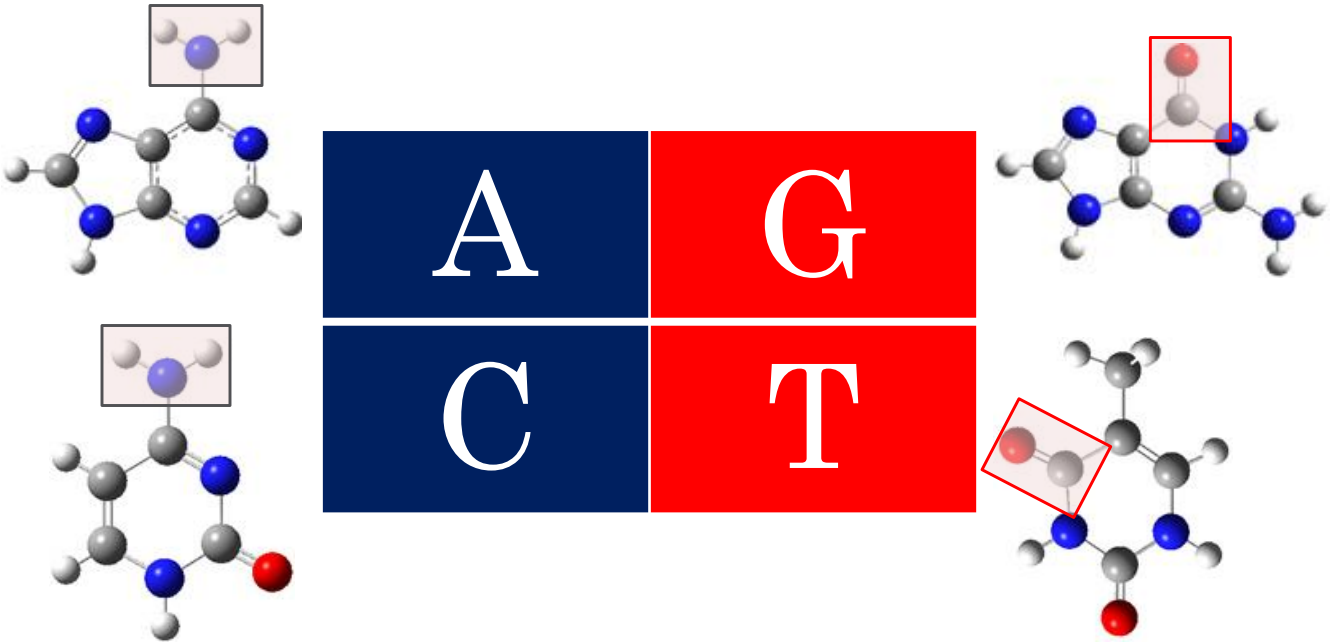
Every pair of nucleotides has something in common



PURINE (large) vs. PYRIMIDINE (small)

Degenerate characters

Every pair of nucleotides has something in common



AMINO vs. KETO functional groups
(rarely useful)

IUPAC nomenclature

A	C	G	T	
■				A/B
	■			C/D
		■		G/H
			■	T/V
■	■			M/K
■		■		R/Y
	■	■		W/S
■	■	■	■	N/X

Example: recoding

- Transitions: replace one nucleotide with the other of the same size
- Transversions: replace one nucleotide with one of a different size

(C↔T) and (A↔G) generally > {A,G}↔{C,T}

R/Y recoding hides transitions (since C,T→Y and A,G→R)

Good for dissimilar sequences as it reduces the number of differences

GTCTAAAAAGTTCAAGGTT	RYYYRRRRRRYYYRRRRYY
AACAAAGAAAATGAAGGTA	RRYRRRRRRRRYRRRRRYR
Original gene sequences	Recoded gene sequences

Word frequencies

Decompose a sequence into a set of words of a given length

k-mers: the collection of words of a given length *k*

Nucleotides:{A,C,G,T}

Dinucleotides:{AA,AC,…,TT}

Trinucleotides:{AAA,AAC,…,TTT}

etc…

$$N(k)=4^k$$

Sequence composition (k=1)

Most common: (G+C) content

ACCGGCGCTTAGCAGGAAGA
TGGCCGCGAATCGTCCTTCT

12 G-C pairs, 8 A-T pairs, so (G+C)% = 60%

$k = 1$

A	**6/20** = 0.30
C	0.25
G	0.35
T	0.10

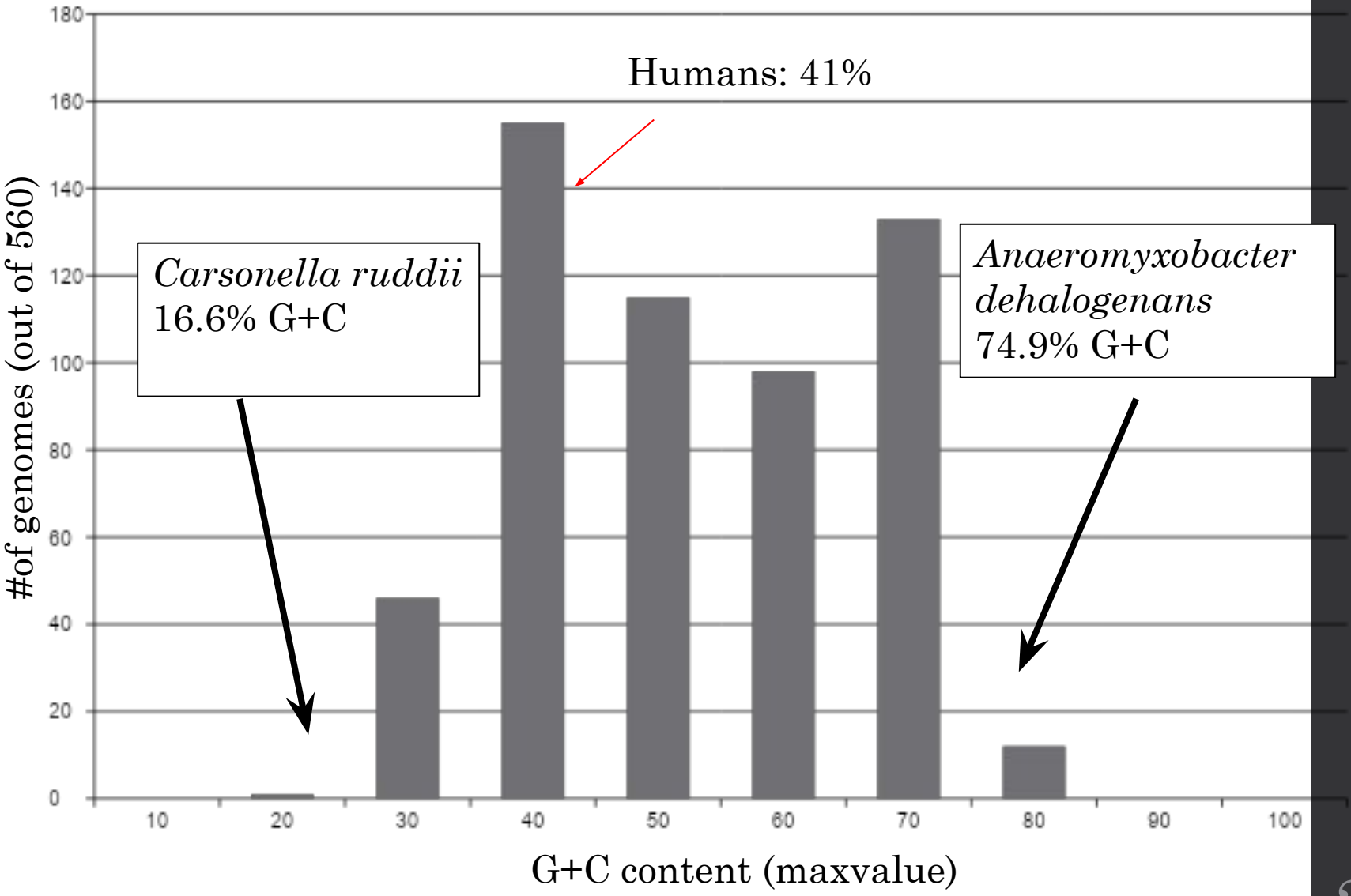
$k = 2$

AA	**1/19** = 0.053
AC	0.053
…	…
GC	0.158

$k = 3$

AAA	**0/18** = 0.00
ACC	1/18 = 0.056
…	…
TTT	0

G+C content distribution of bacterial genomes



Gap spectra

Like *k*-mers, but include internal wildcards

Length = *k*

of 'literals' = L

k=4, L=2:{ ANNA, ANNC, …, TNNT }

Can model higher-order relationships without exhaustive enumeration

Massive Degeneracy

Generalize previous ideas about composition

Length k

Any IUPAC character (except X) can be used at any position

k=2: { AA, AB, AC, AD, AG, …, VV }

15 possible letters, therefore $N(k) = 15^k$

All possible degenerate characters of length 1 to (say) 10

{ A, B, C, …, V }

{ AA, AB, …, VV }

…

{ AAAAAAAAAA, AAAAAAAAAC, …, VVVVVVVVVV }

So…

$$15^1 + 15^2 + 15^3 + 15^4 + 15^5 + 15^6 + 15^7 + 15^8 + 15^9 + 15^{10}$$

$$\cong 15^{10}$$

$$\cong 5.8 \times 10^{11}$$

Hmmm.

Markov models of composition

Sequences as *k*th-order Markov chains:

The next state in a series of random variables is dependent only on the previous *k* states.

$$\Pr(X_{n+1}=x \mid \mathbf{X_n}=y)$$

Order = 1

$$x, y \in S = \{ A, C, G, T \}$$

Zeroth-order Markov model:

$$\Pr(X_{n+1} = x \mid X_n = y) = \Pr(x)$$

Therefore

$$\Pr(xy) = \Pr(x) \times \Pr(y) \ \forall \ (x,y)$$

(Independent events)

First-order Markov model (human genome)

X_n \ X_{n+1}	A	C	G	T
A	0.300	0.205	0.285	0.210
C	0.322	0.298	0.078	0.302
G	0.248	0.246	0.298	0.208
T	0.177	0.239	0.292	0.292

$$\Pr(X_{n+1} = A \mid X_n = G) = 0.248$$

Durbin,Chapter 3

Second-order Markov model
(numbers lazily copied)

X_n \ X_{n+1}	A	C	G	T
AA	0.300	0.205	0.285	0.210
AC	0.322	0.298	0.078	0.302
AG	0.248	0.246	0.298	0.208
AT	0.177	0.239	0.292	0.292
…				

$$\Pr(X_{n+1} = A \mid X_n = G, X_{n-1} = A) = 0.248$$

DNA2Vec/BioBERT etc.

- Associations among DNA words based on neighbourhood similarity
- Coming in classification module

Ng (2017) *arXiv*

What about proteins?

Protein sequences

Naïvely: 20^k

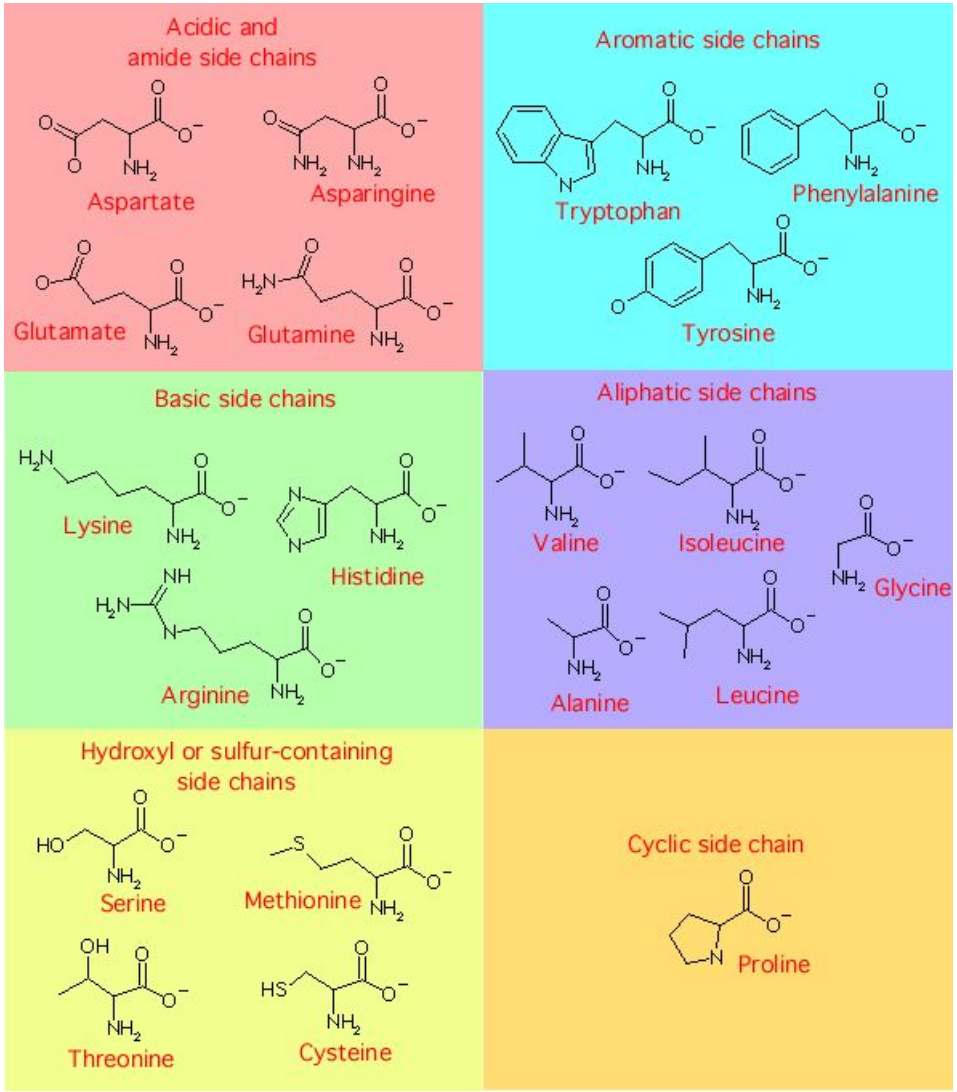
$k = 1$

A	0.02
C	0.09
D	0.11
E	0.10
…	

There is no complete degenerate alphabet for amino acids (although there could be – we would just need 2^{20} characters)

We can consider STRUCTURAL and FUNCTIONAL categories instead

Structural and functional attributes



http://www.bact.wisc.edu/Microtextbook/images/book_4/chapter_2/2-10.gif

Reduced amino acid alphabets

n = 2	ADEGKNPQRST CFHILMVWY
	ADEGNPST CHKQRW FILMVY
	AGNPST CHWY DEKQR FILMV
	AGPST CFWY DEN HKQR ILMV
	APST CW DEGN FHY ILMV KQR
	AGST CW DEN FY HP ILMV KQR
	AST CG DEN FY HP ILV KQR MW
	AST CW DE FY GN HQ ILV KR MP
	AST CW DE FY GN HQ IV KR LM P
	AST C DE FY GN HQ IV KR LM P W
	AST C DE FY G HQ IV KR LM N P W
	AST C DE FY G H IV KR LM N P Q W
	AST C DE FL G H IV KR M N P Q W Y
	AST C DE F G H IV KR L M N P Q W Y
	AT C DE F G H IV KR L M N P Q S W Y
	AT C DE F G H IV K L M N P Q R S W Y
	A C DE F G H IV K L M N P Q R S T W Y
n = 19	A C D E F G H IV K L M N P Q R S T W Y

This is one of many possible examples!

Susko and Roger (2007) *PROTEINS: Mol Biol Evol*

Correlation representations

- e.g., pseudo-amino acid composition
- Look at global correlations θ_i of chemical / structural features at a series of distances λ_i

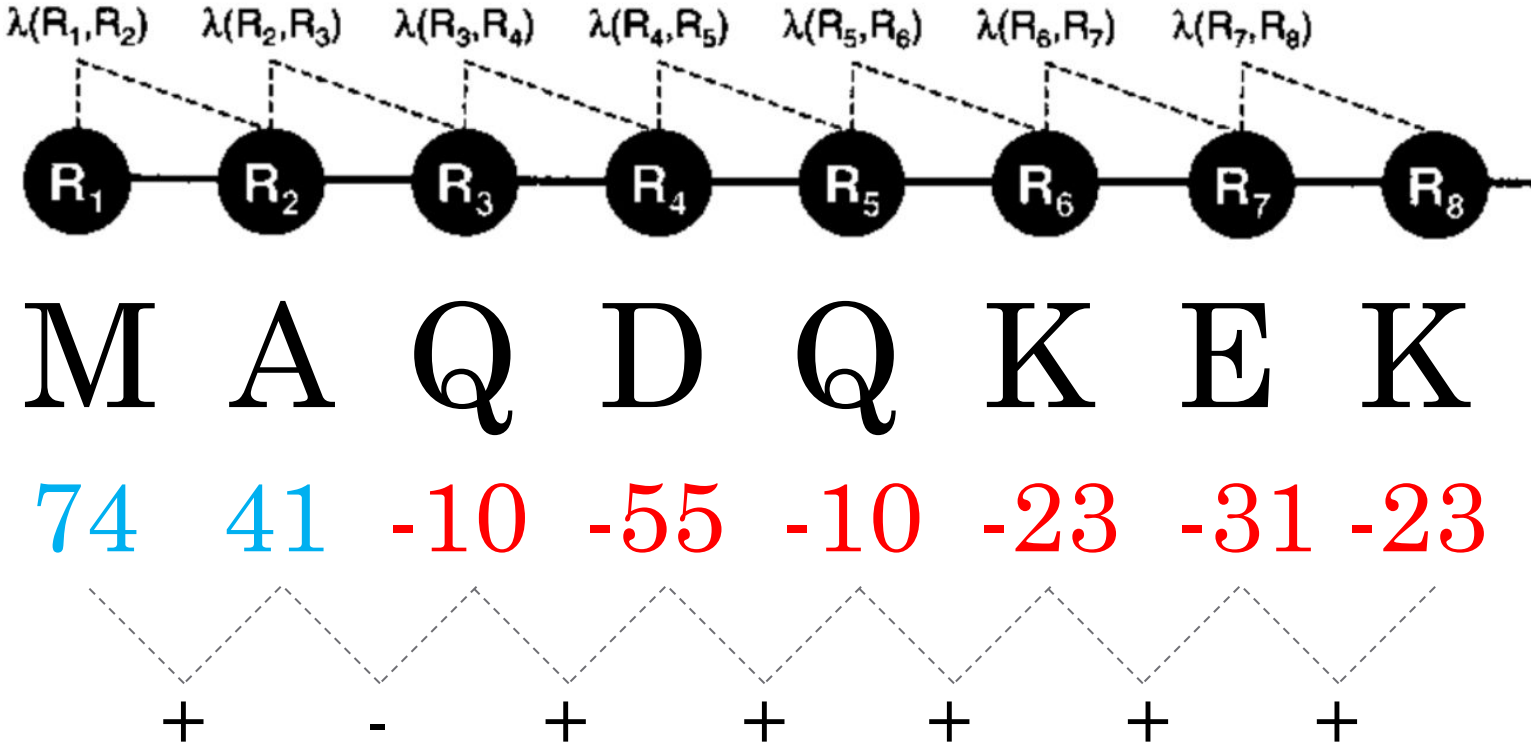
$$\theta_1 = \frac{1}{L-1} \sum_{i=1}^{L-1} \Theta(R_i, R_{i+1})$$

protein length

correlation of adjacent amino acids ($\lambda = 1$)

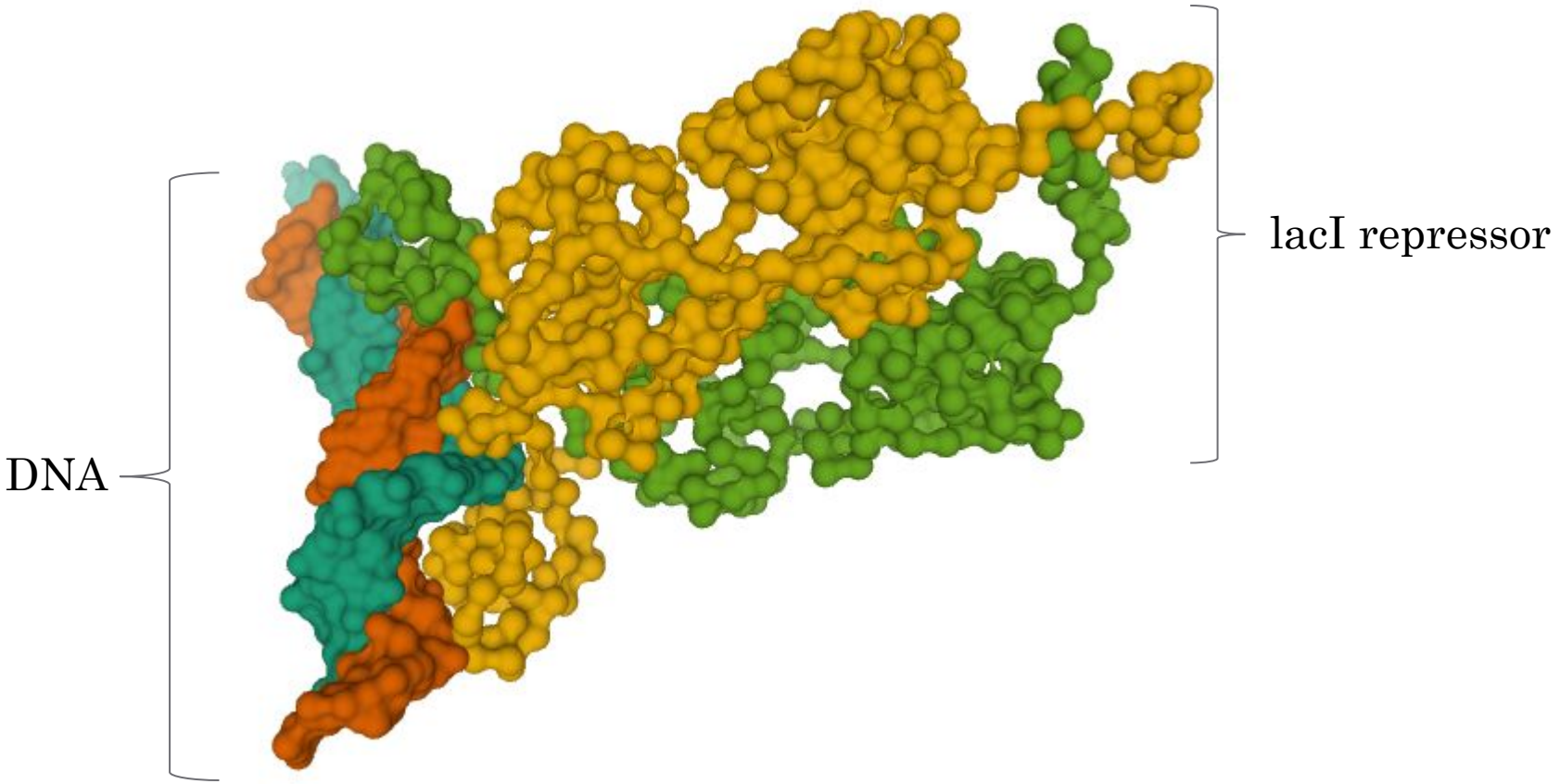
Chou (2001) *PROTEINS: Structure, Function, and Genetics*

Example: hydrophobicity



High average correlation

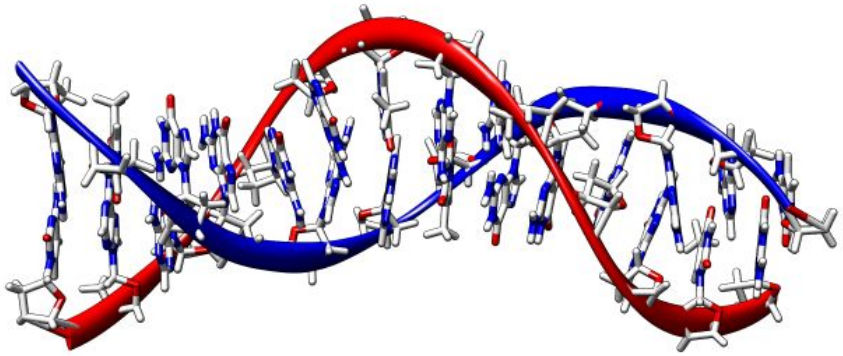
Chou (2001) *PROTEINS: Structure, Function, and Genetics*



Structural representations

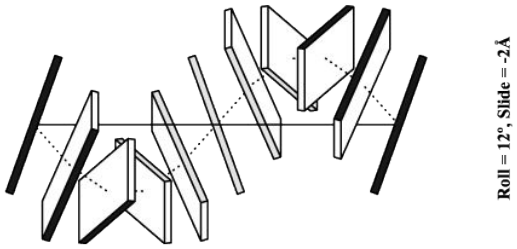
Structural representations of DNA

- Two ways to think about DNA structure:

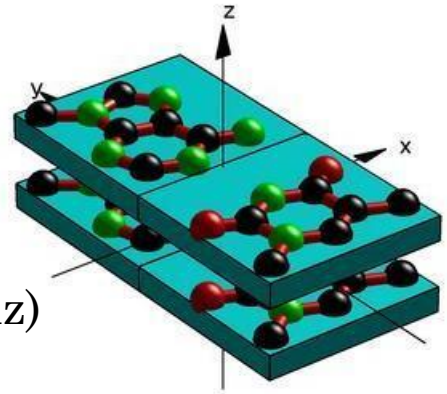


Atomic coordinates

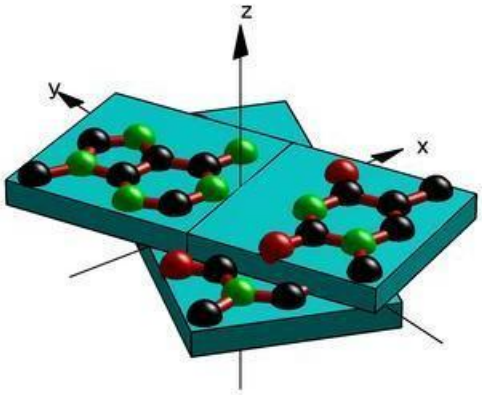
Geometric parameters



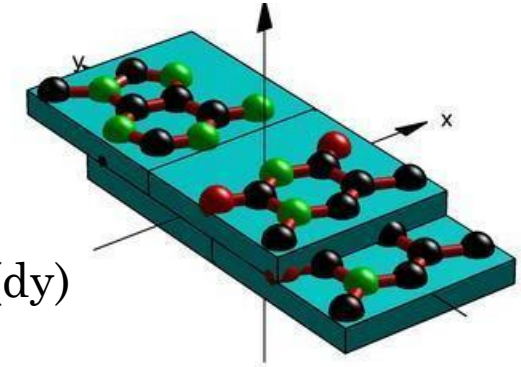
http://rutchem.rutgers.edu/~xiangjun/3DNA/images/regular_dna.gif



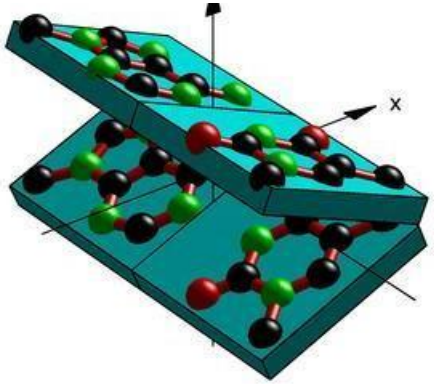
Rise(dz)



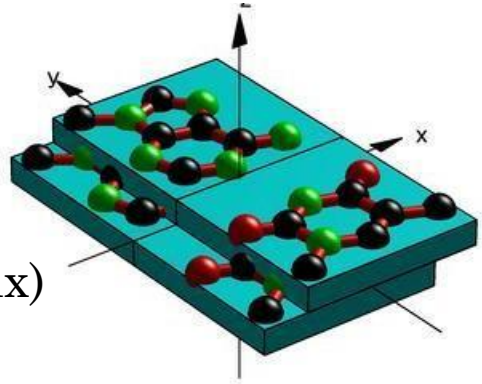
Twist(rz)



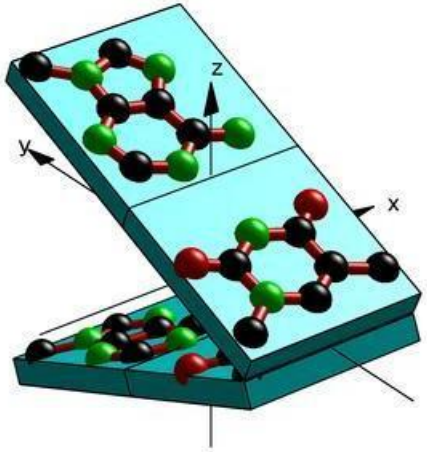
Slide(dy)



Roll(ry)



Shift(dx)



Tilt(rx)

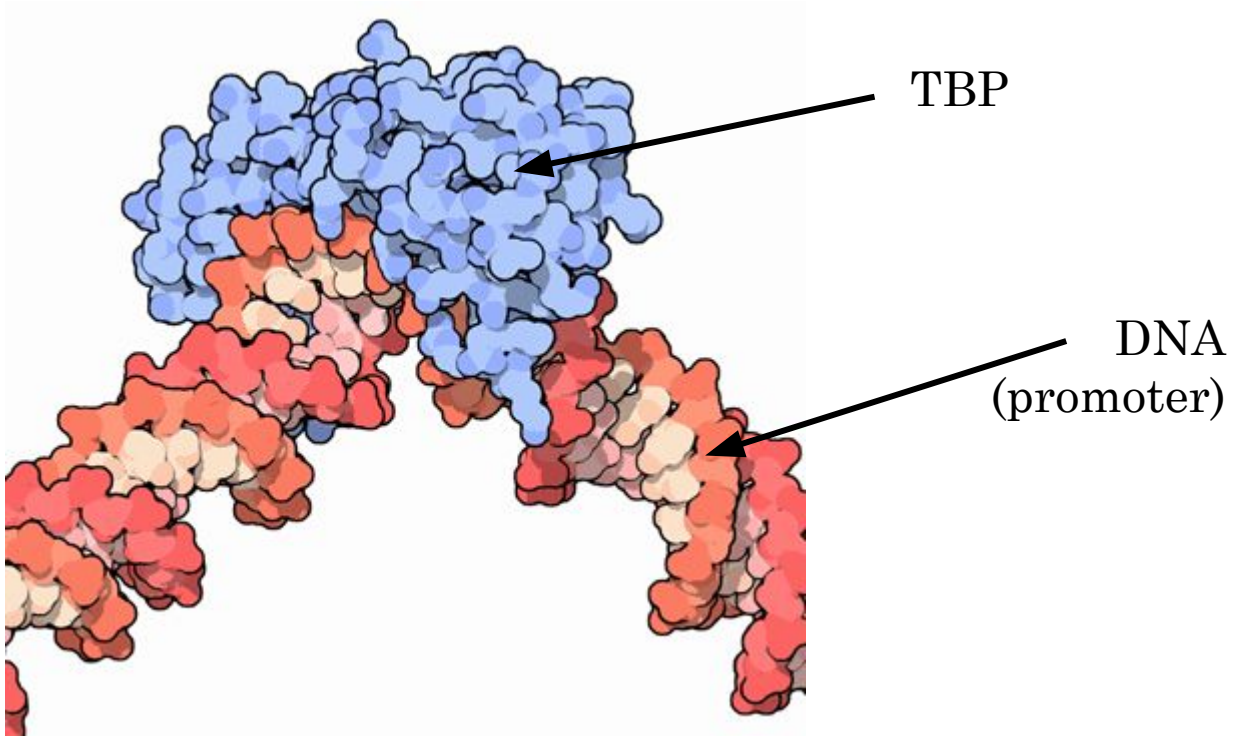
gibk26.bse.kyutech.ac.jp/jouhou/readout/help.html

Static parameters (twist, roll)

- AND -

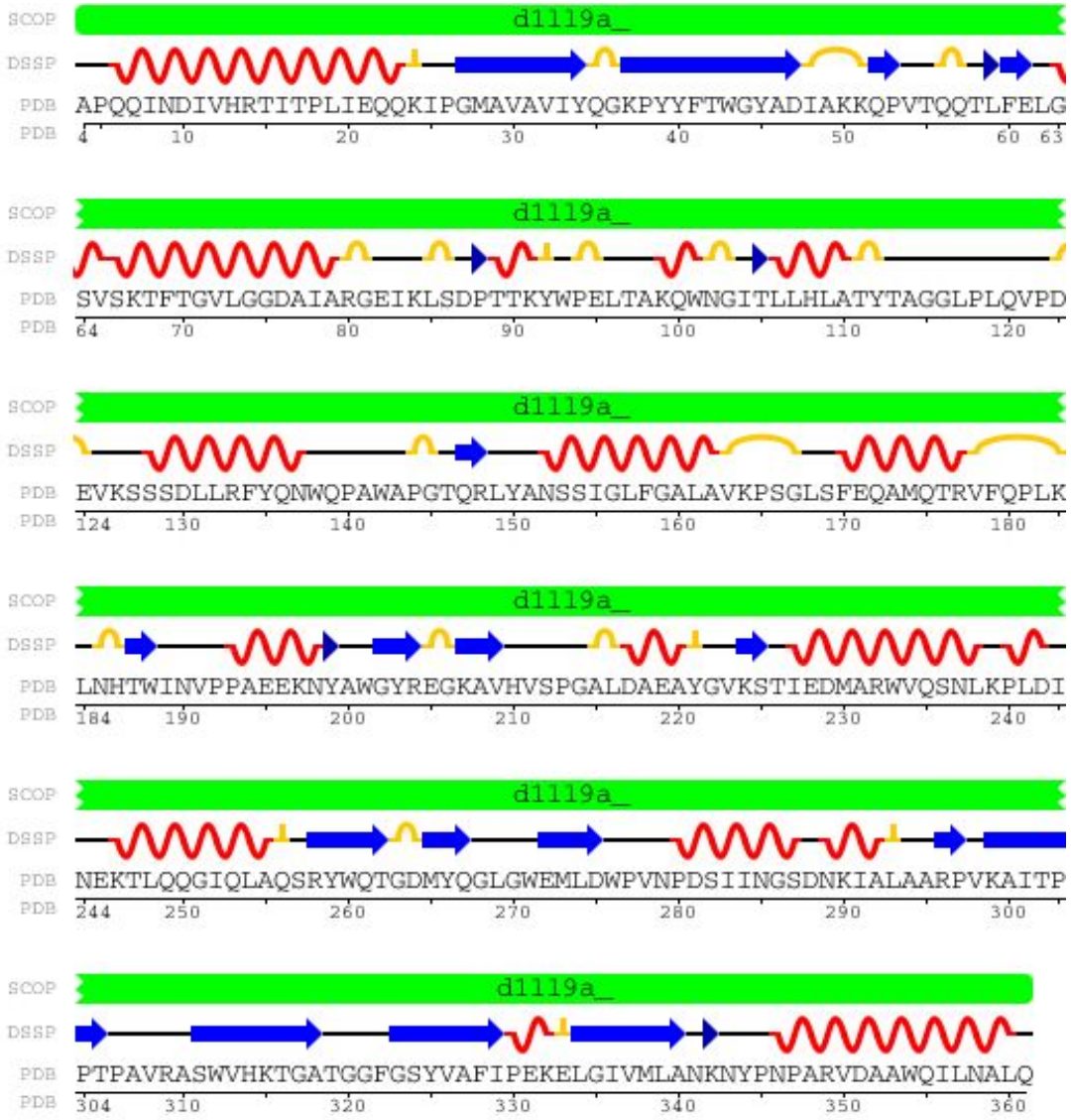
Dynamic parameters (flexibility/deformability)

e.g., DNA complex with TATA-box binding protein (TBP)

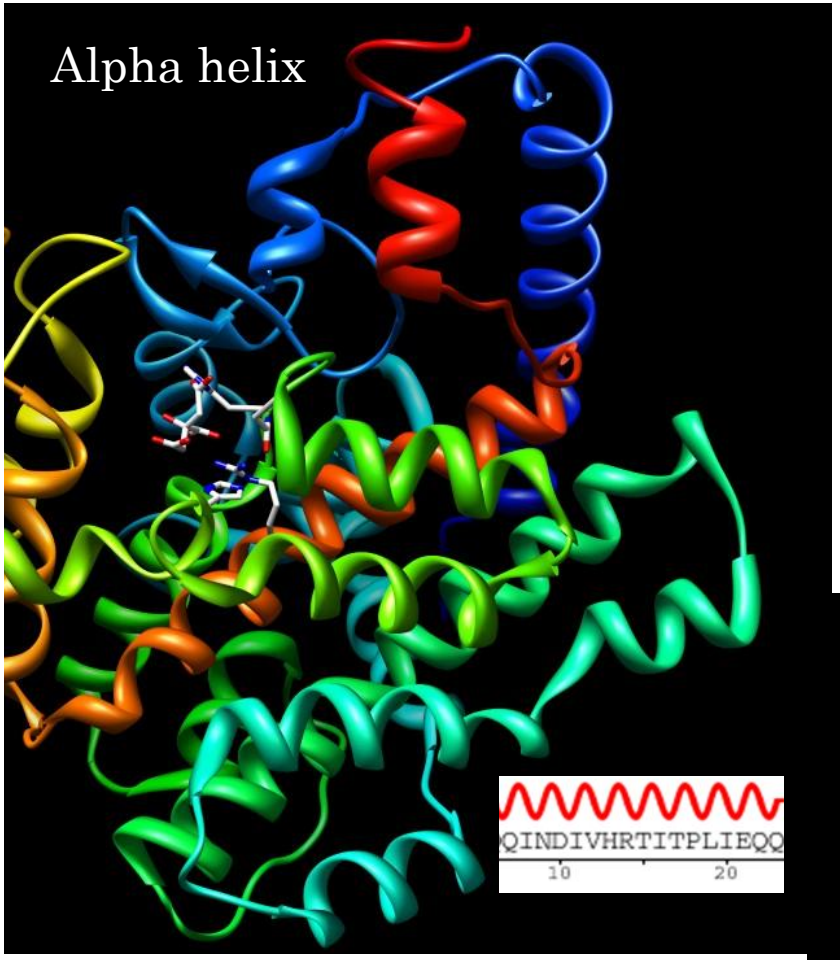


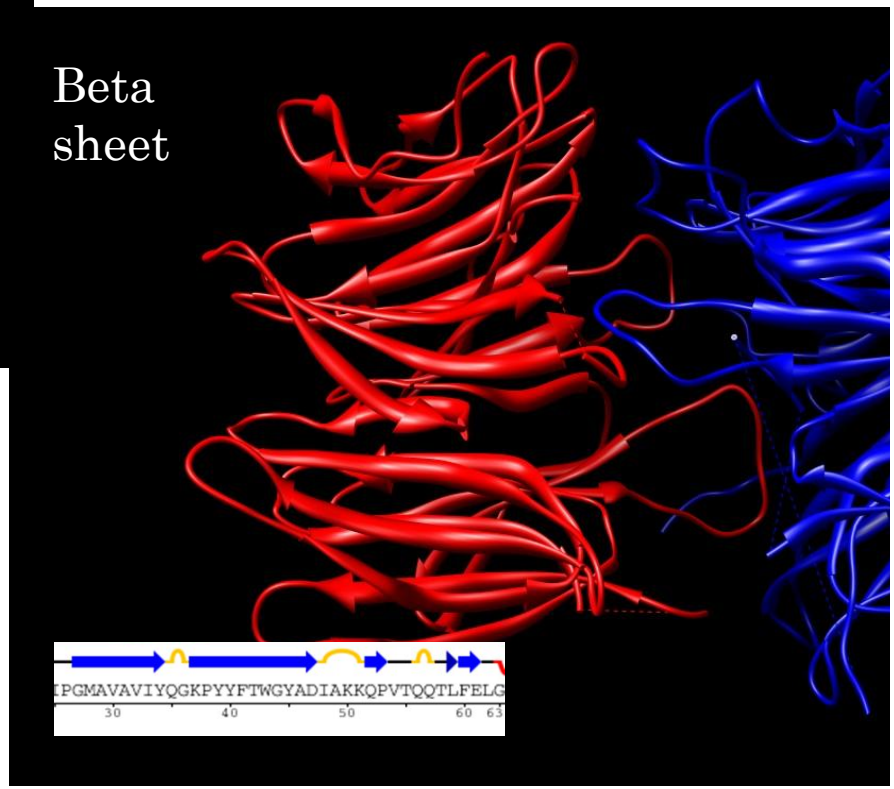
Proteins tend to have more “interesting” structures that govern their behaviour, so structural methods are more frequently applied to proteins than to DNA

Secondary structure



http://www.rcsb.org/pdb/explore/remediatedSequence.do?structureId=1LL9

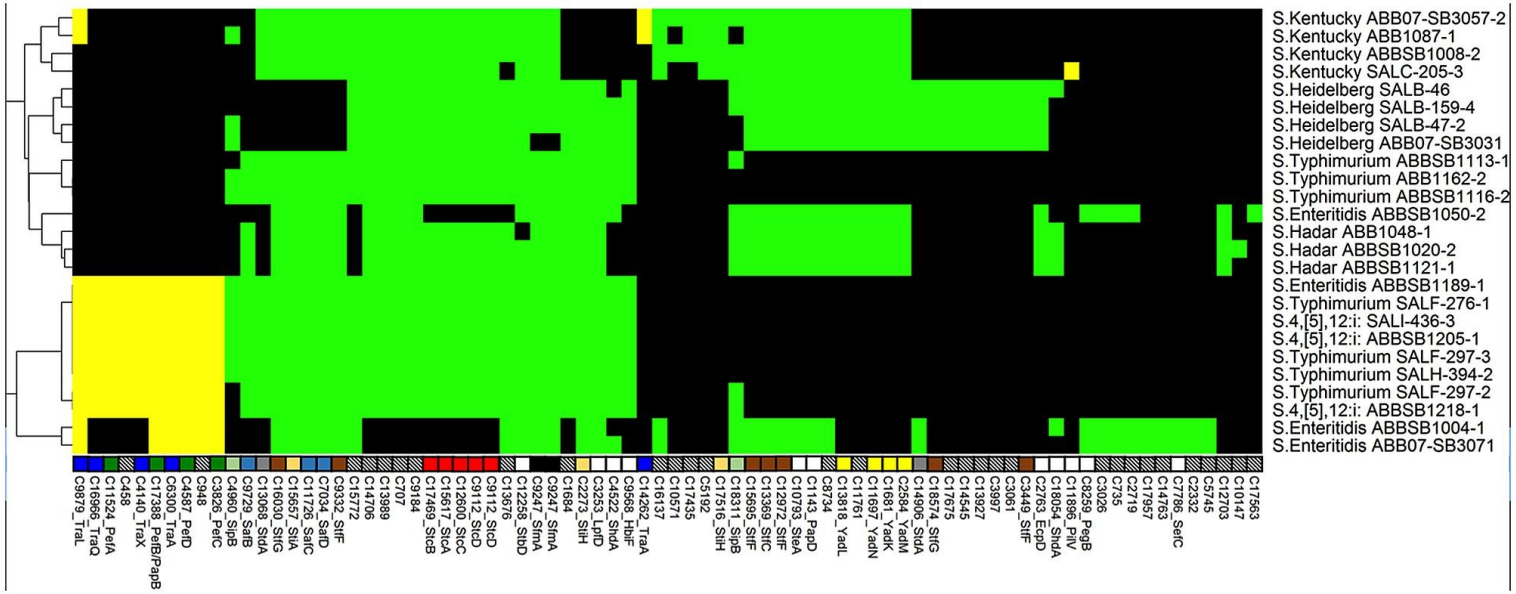

Alpha helix
QINDIVHRTITPLIEQQ
10
20


Beta sheet
IPGMAVAVIYQGKPYYFTWGYADIAKKQPVTQQTLFELG
30
40
50
60
63

Tertiary structure (e.g., atomic coordinates)

ATOM	3	C	PRO	A	1	63.886	41.846	3.646	1.00 22.65	C
ATOM	4	O	PRO	A	1	64.467	41.039	2.948	1.00 22.51	O
ATOM	5	CB	PRO	A	1	61.985	43.079	2.551	1.00 22.54	C
ATOM	6	CG	PRO	A	1	61.974	43.966	1.334	1.00 23.59	C
ATOM	7	CD	PRO	A	1	63.440	44.213	0.951	1.00 24.08	C
ATOM	8	N	GLN	A	2	63.711	41.737	4.969	1.00 23.06	N
ATOM	9	CA	GLN	A	2	64.116	40.581	5.732	1.00 20.94	C
ATOM	10	C	GLN	A	2	63.002	40.196	6.653	1.00 18.99	C
ATOM	11	O	GLN	A	2	62.479	41.045	7.339	1.00 21.48	O
ATOM	12	CB	GLN	A	2	65.410	40.873	6.513	1.00 18.89	C
ATOM	13	CG	GLN	A	2	65.904	39.624	7.267	1.00 21.48	C
ATOM	14	CD	GLN	A	2	67.379	39.737	7.626	1.00 27.58	C
ATOM	15	OE1	GLN	A	2	67.863	39.075	8.566	1.00 30.78	O
ATOM	16	NE2	GLN	A	2	68.080	40.643	6.939	1.00 26.63	N
ATOM	17	N	PHE	A	3	62.612	38.932	6.659	1.00 18.87	N
ATOM	18	CA	PHE	A	3	61.548	38.503	7.542	1.00 19.11	C
ATOM	19	C	PHE	A	3	62.096	37.578	8.572	1.00 18.63	C
ATOM	20	O	PHE	A	3	62.597	36.517	8.167	1.00 13.98	O
ATOM	21	CB	PHE	A	3	60.413	37.726	6.820	1.00 16.68	C
ATOM	22	CG	PHE	A	3	59.665	38.563	5.831	1.00 19.69	C

x y z

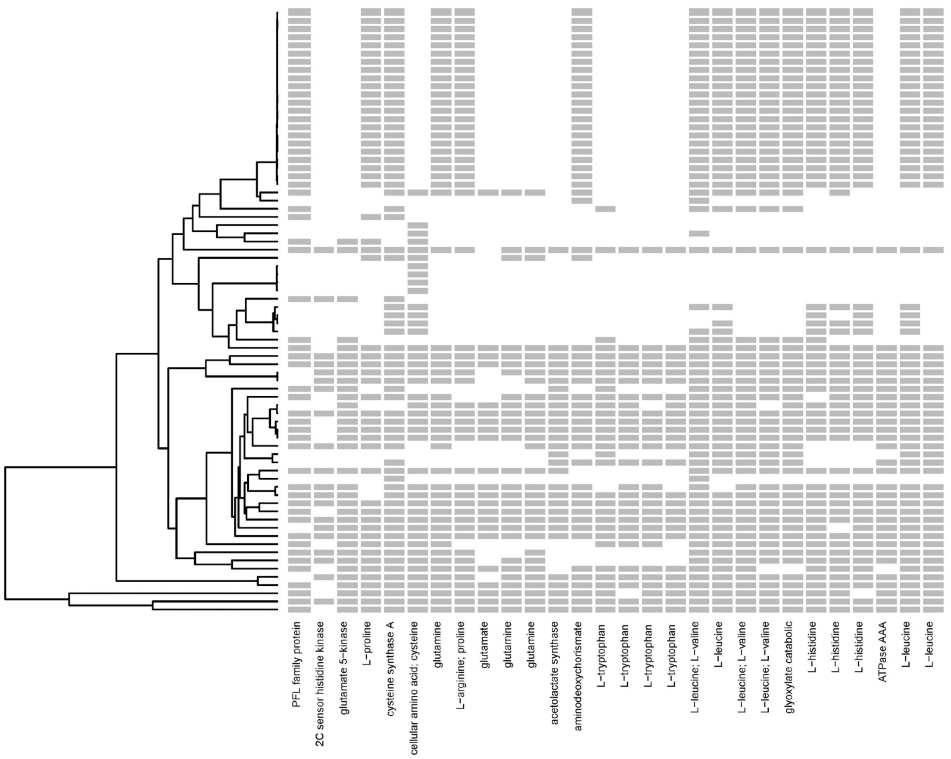


Phylogenetic profiles

Presence / absence patterns of genes across genomes

Bit strings!

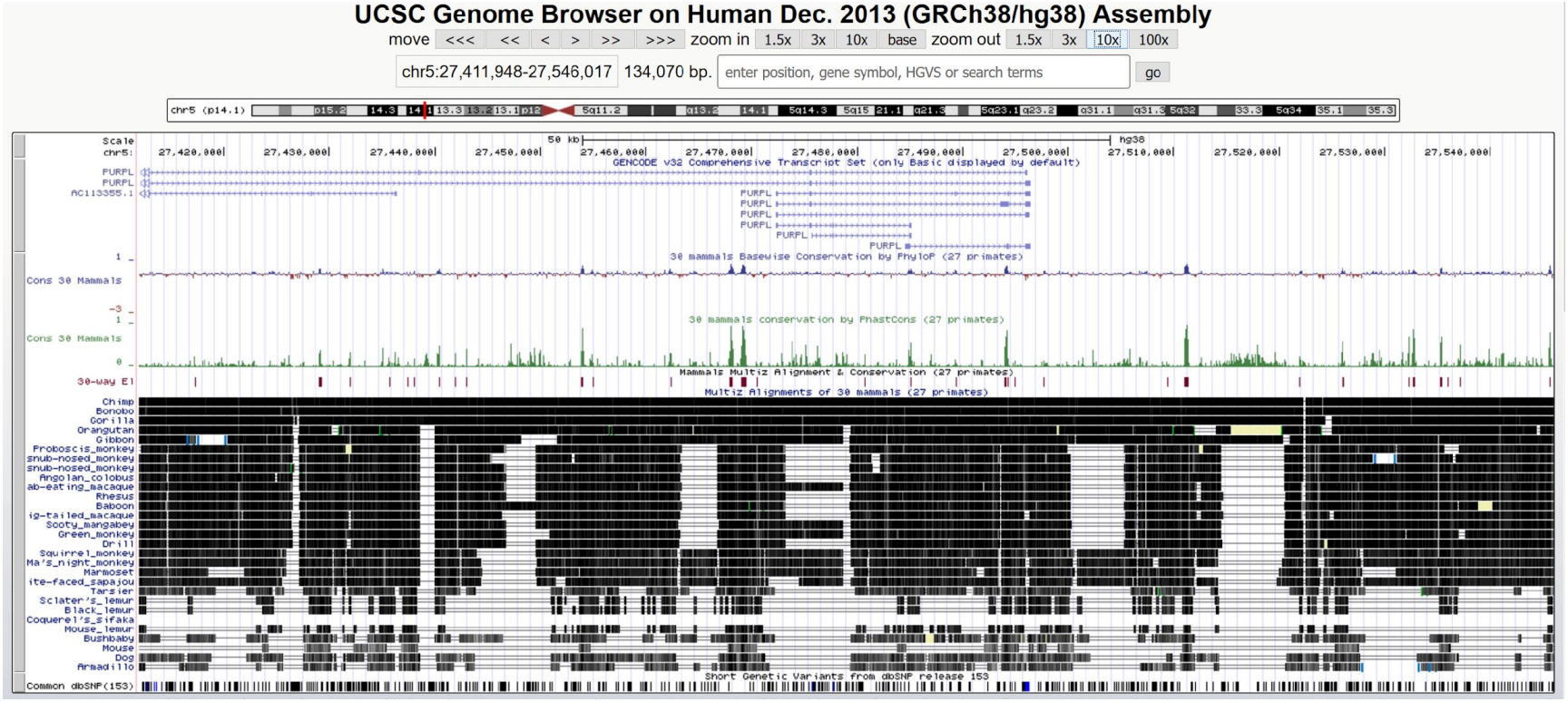
Genomes
(related by phylogenetic tree)



Individual genes

Liu et al. Genome Biol Evol, Volume 10, Issue 9, September 2018, Pages 2255–2265, https://doi.org/10.1093/gbe/evy178

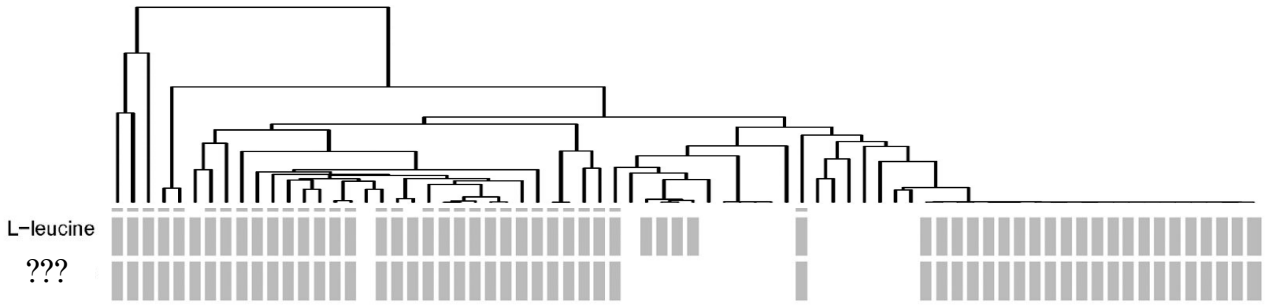
Comparative genomics of mammals



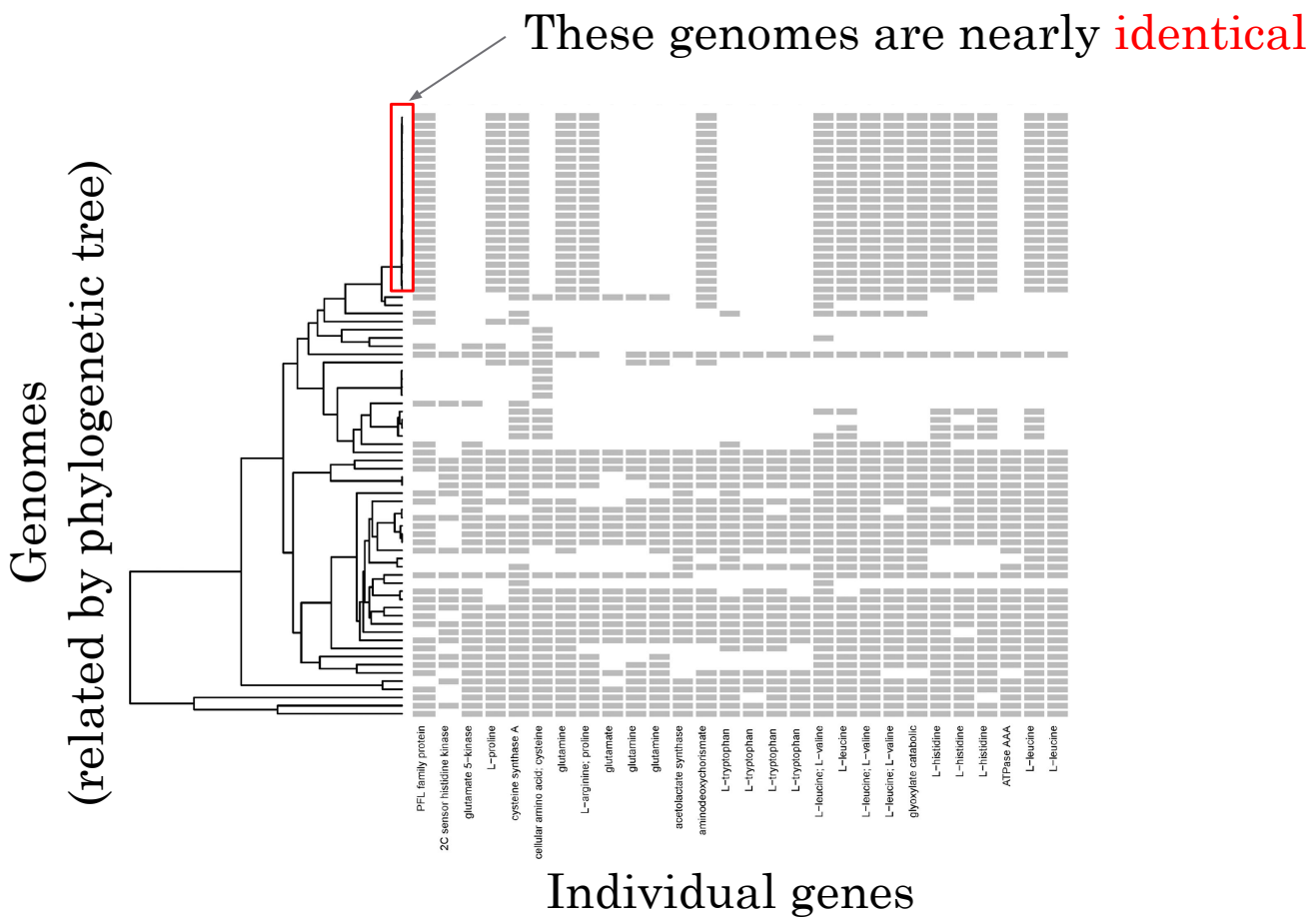
Why?

The distribution of **genes** across **genomes** can tell us about:

- The capabilities of those genomes (can genome *x* make amino acid *y*?)
- The roles of specific genes (e.g., found only in organisms that live at high temperatures, etc)
- Guilt by association!



One issue with phylogenetic profiles



Liu et al. Genome Biol Evol, Volume 10, Issue 9, September 2018, Pages 2255–2265, https://doi.org/10.1093/gbe/evy178

Summary

1. There are many different applications of DNA. Protein, and genome representation

2. No single representation is ideal for every task

3. DNA and protein have fundamentally different structures, and some types of representation make sense for one but not the other