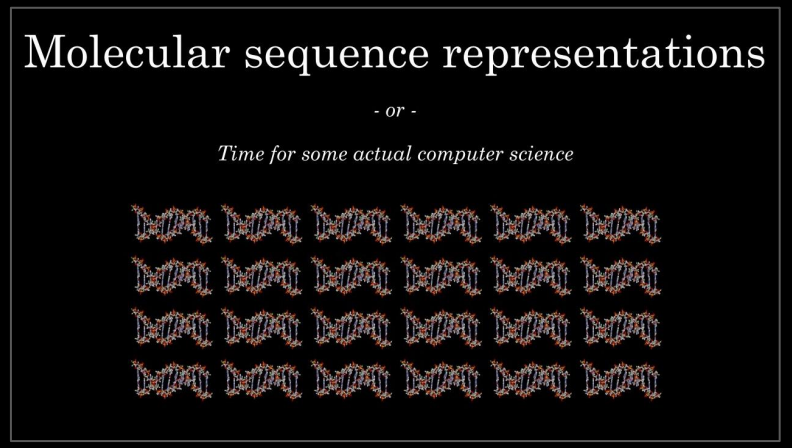
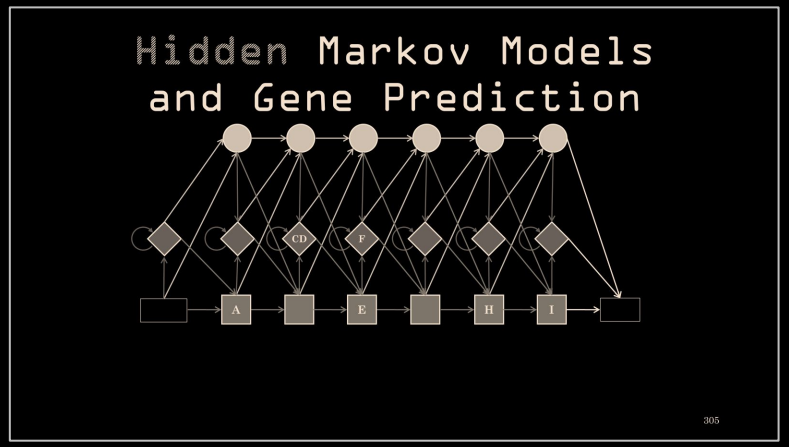


Module 1



Homology and Sequence Alignment

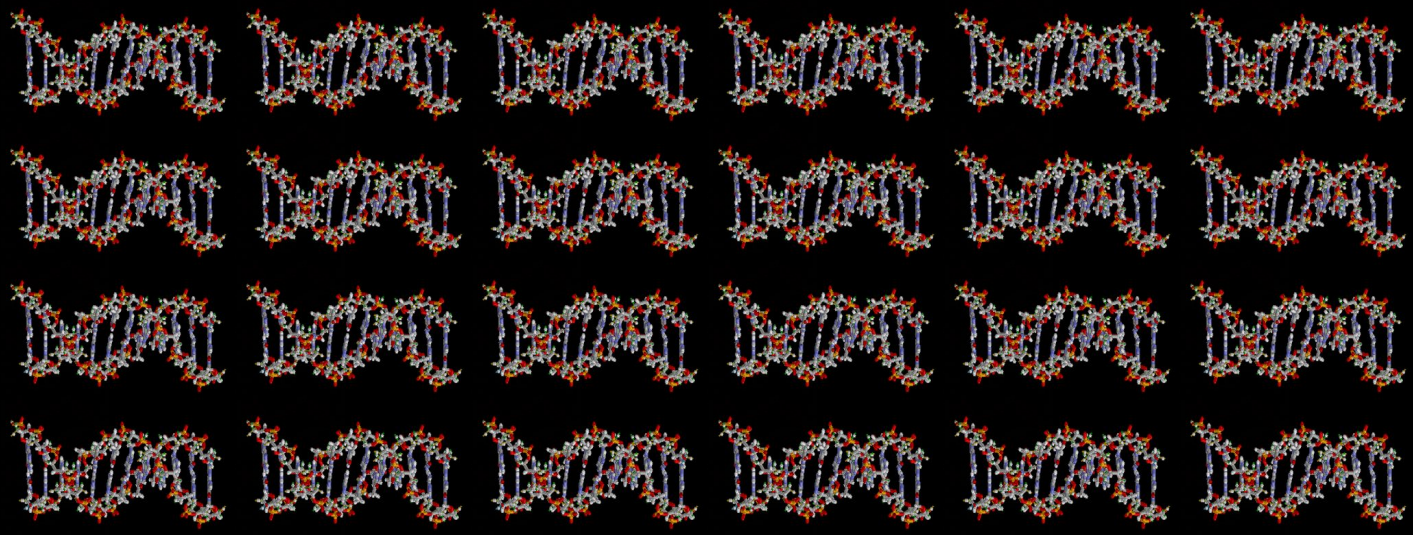
Module Summary

- There are many ways to represent sequences…there is no universal best choice
- Point, insertion and deletion mutations make the alignment problem non-trivial (with exponential complexity!) We need efficient algorithms and appropriate statistics
- How can we efficiently do:
  - Fast database searches?
  - VERY fast database searches?
  - Multiple sequence alignment?

Molecular sequence representations

- or -

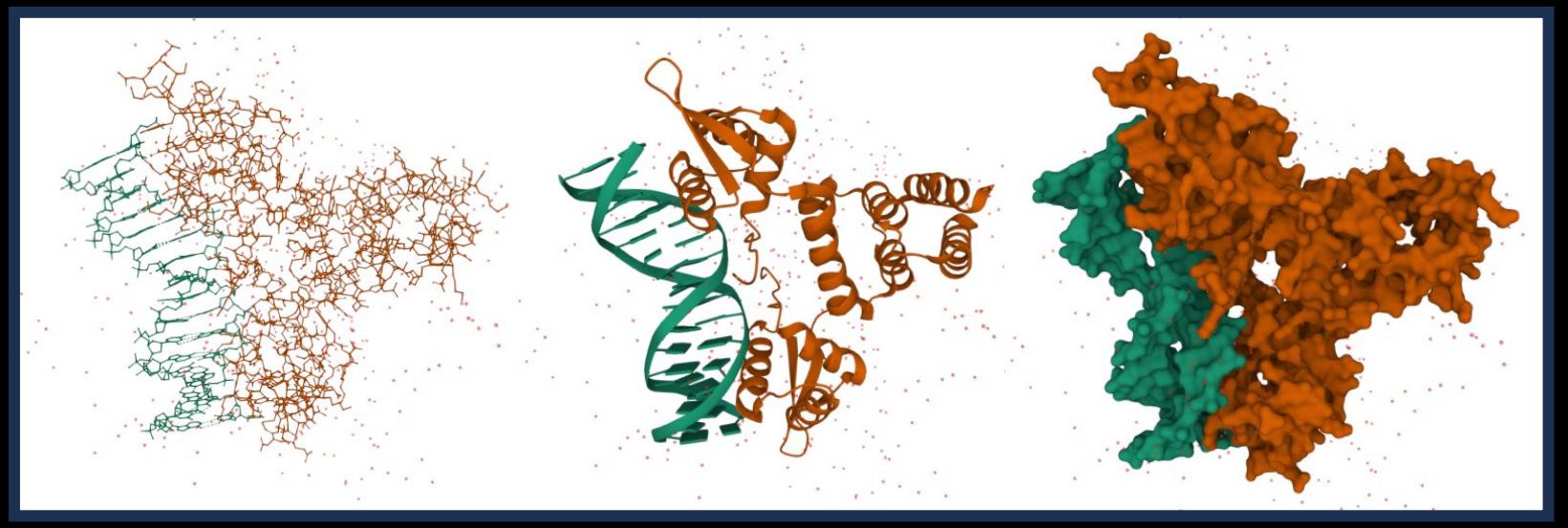
Time for some actual computer science



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

Overview

1. Goals of sequence representation
2. Text string-based sequence representations
3. Advanced sequence representations: probabilities, data structures, models
4. Structural representations



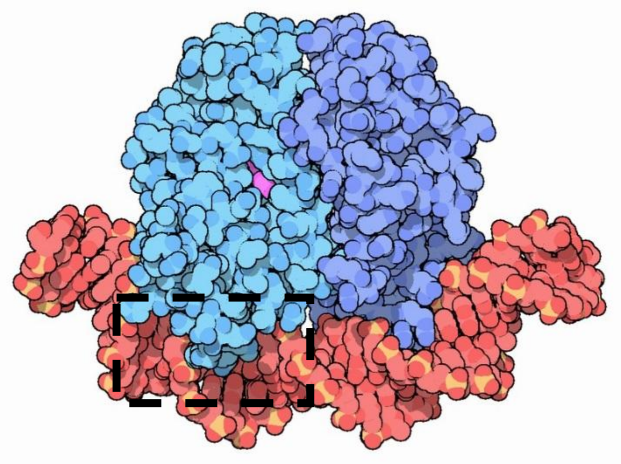
???

Comparison / classification / analysis

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

Goals of representation

(1) Identify functional patterns (e.g., sequence *motifs* or functional *domains*) in DNA or protein sequences



Catabolite activator protein (CAP or CRP) bound to DNA

CAP-lacking sequences

```
blahblahblahblahblahblahblahblahblahblahbla
h
blahblahblahblahblahblahblahblahblahblahbla
h
blahblahblahblahblahblahblahblahblahblahbla
h
blahblahblahblahblahblahblahblahblahblahbla
```

```
atatgcctga cggagttcacacttgtaagttt tcaactacg
t
attcagtaca aaacgtgatcaacccctcaatt ttcccttgc
t
tcgctttgtc agctgtgacaagctccgcaaat cgtgacaat
a
aaaaacattt tagagtgatatgtataacatta tggcgttta
t
a
cgcacattgg gtataacgtgatcatatcaaca
a
```

Experimentally validated CAP binding sites

What's the difference?

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

Goals of representation

Distinguish phenotypes based on sequence or structural variation

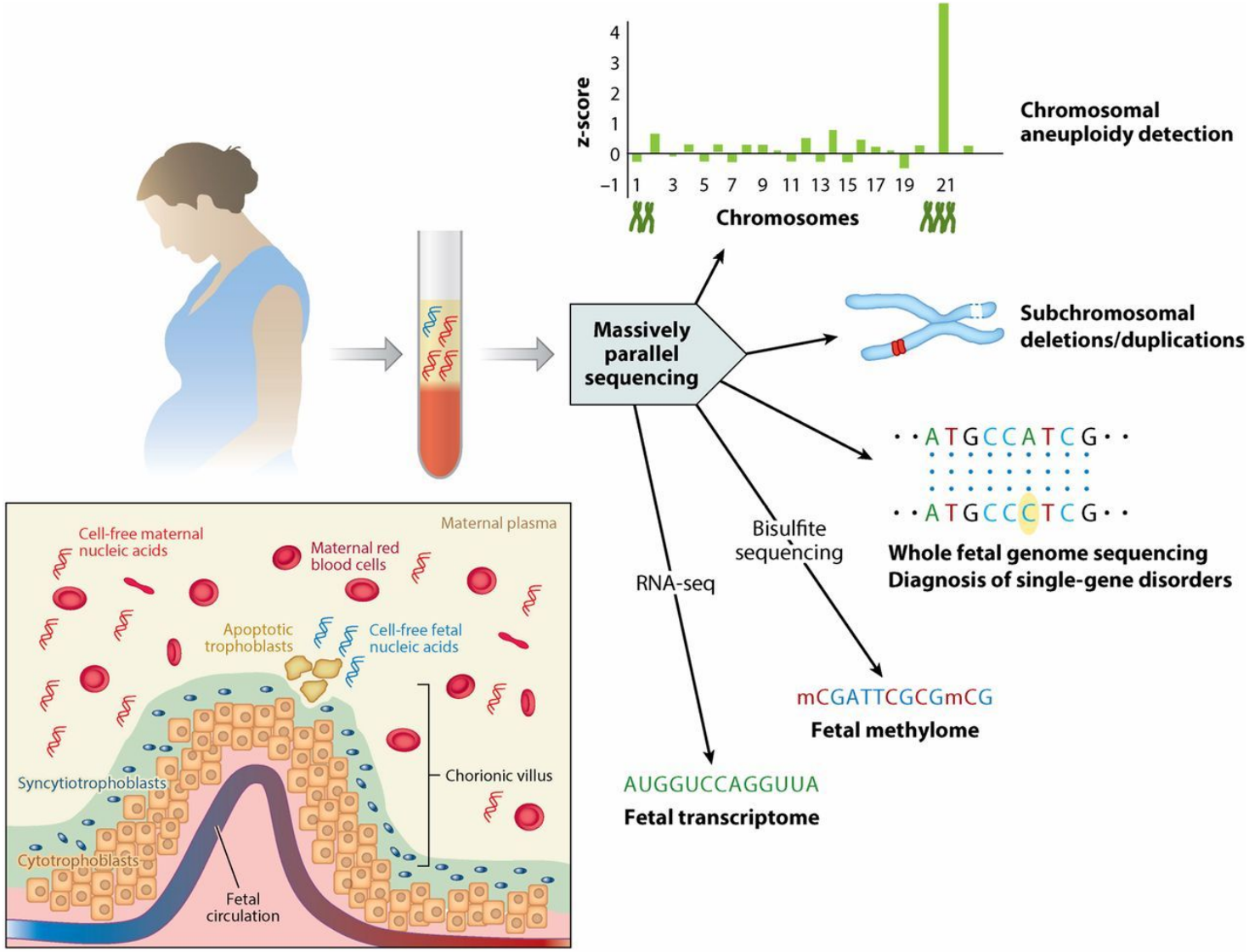
e.g., huntingtin gene, responsible for Huntington's disease

GCTGCCGGGACGGGTCCAAGATGGACGGCCGCTCAGGTTCTGCTTTTACCTGCGGCCCAGAGCCCCATTCA
TTGCCCCGGTGCTGAGCGGCGCCGCGAGTCGGCCCGAGGCCTCCGGGGACTGCCGTGCCGGGCGGGAGACC
GCCATGGCGACCCTGGAAAAGCTGATGAAGGCCTTCGAGTCCCTCAAGTCCTTC**CAGCAGCAGCAGCAGCAGCA**
GCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAACAG

(About 10,000 more nucleotides in the gene)

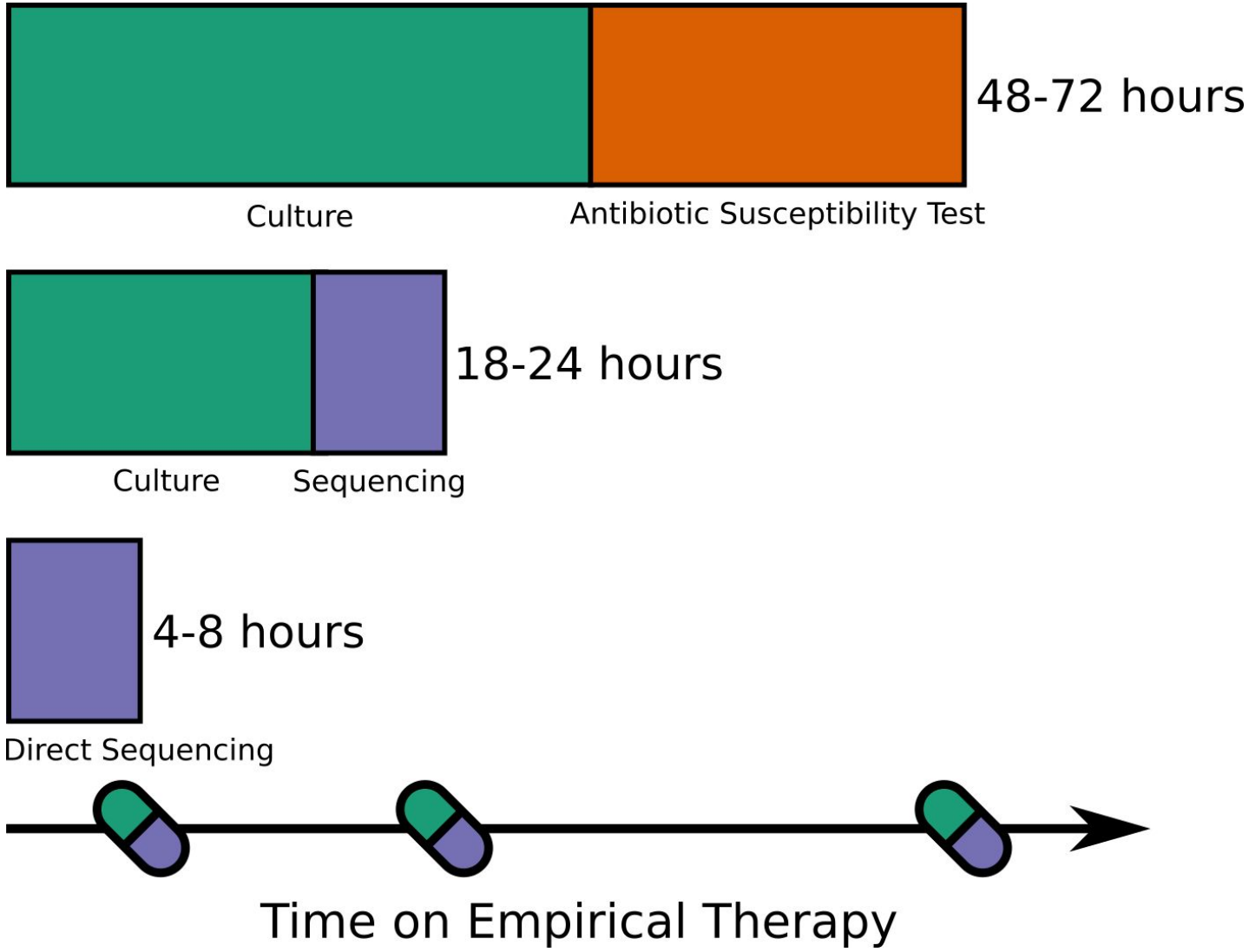
# of CAG repeats	Effect
< 27	Healthy
27-35	Intermediate
36-39	Disease (reduced penetrance)
> 39	Full disease effects

Non-invasive Pre-Natal Testing of Cell-Free DNA

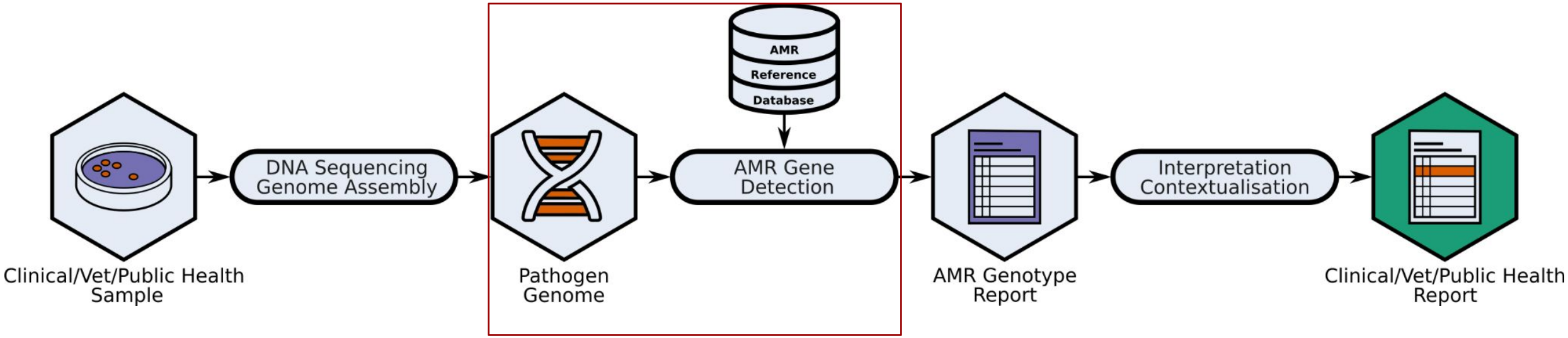


Wong FCK, Lo YMD. 2016. Annu. Rev. Med. 67:419–32

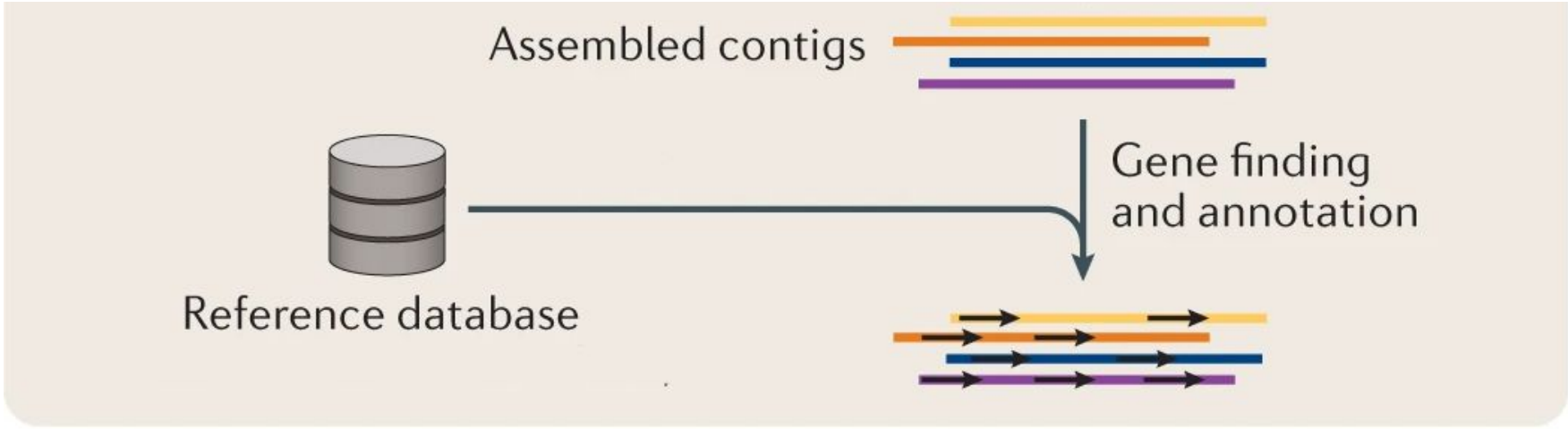
Using genomics for rapid clinical diagnostics



Identifying AMR genes requires comparing sequences



Identifying AMR genes requires comparing sequences



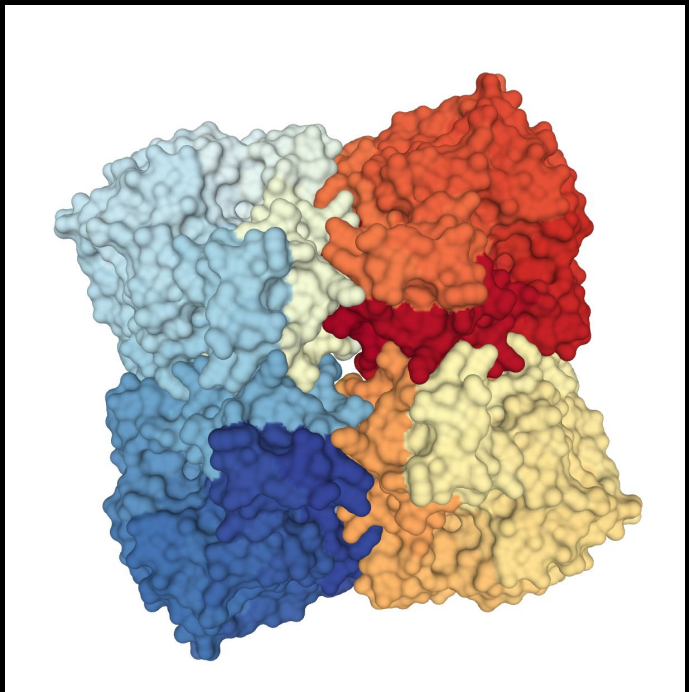
Modified from: Boolchandani, Manish, Alaric W. D'Souza, and Gautam Dantas. "Sequencing-based methods and resources to study antimicrobial resistance." Nature Reviews Genetics 20.6 (2019): 356-370.

Goals of representation

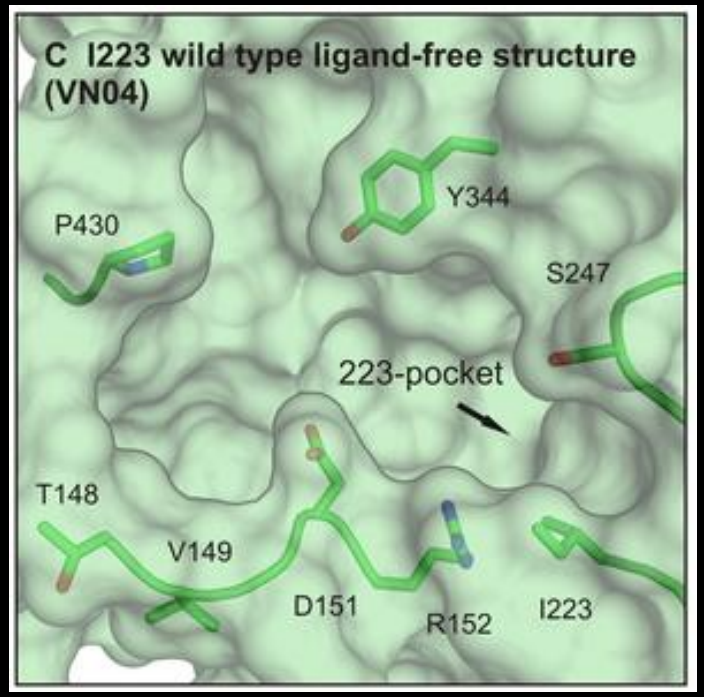
Identify important changes at the sequence and structural level

e.g. oseltamivir resistance in influenza H1N1 Neuraminidase

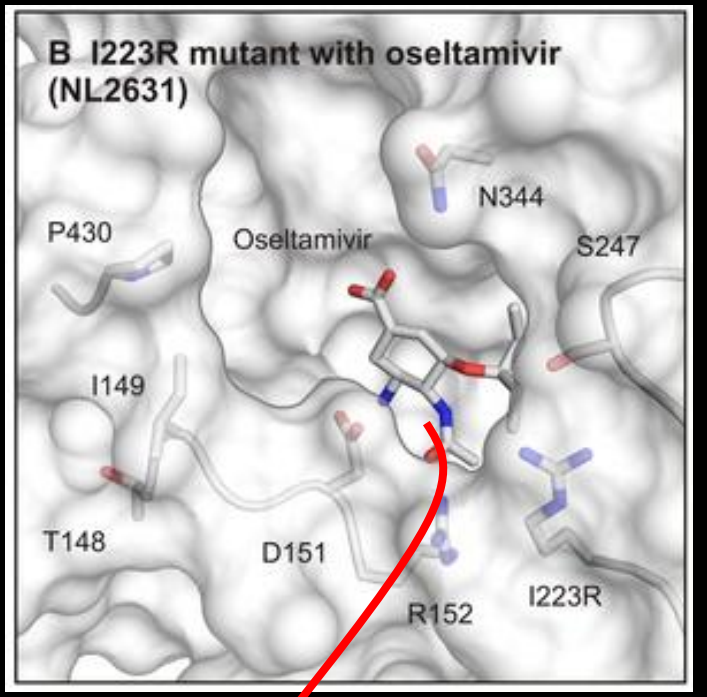
Neuraminidase structure



Normal structure



Mutated version

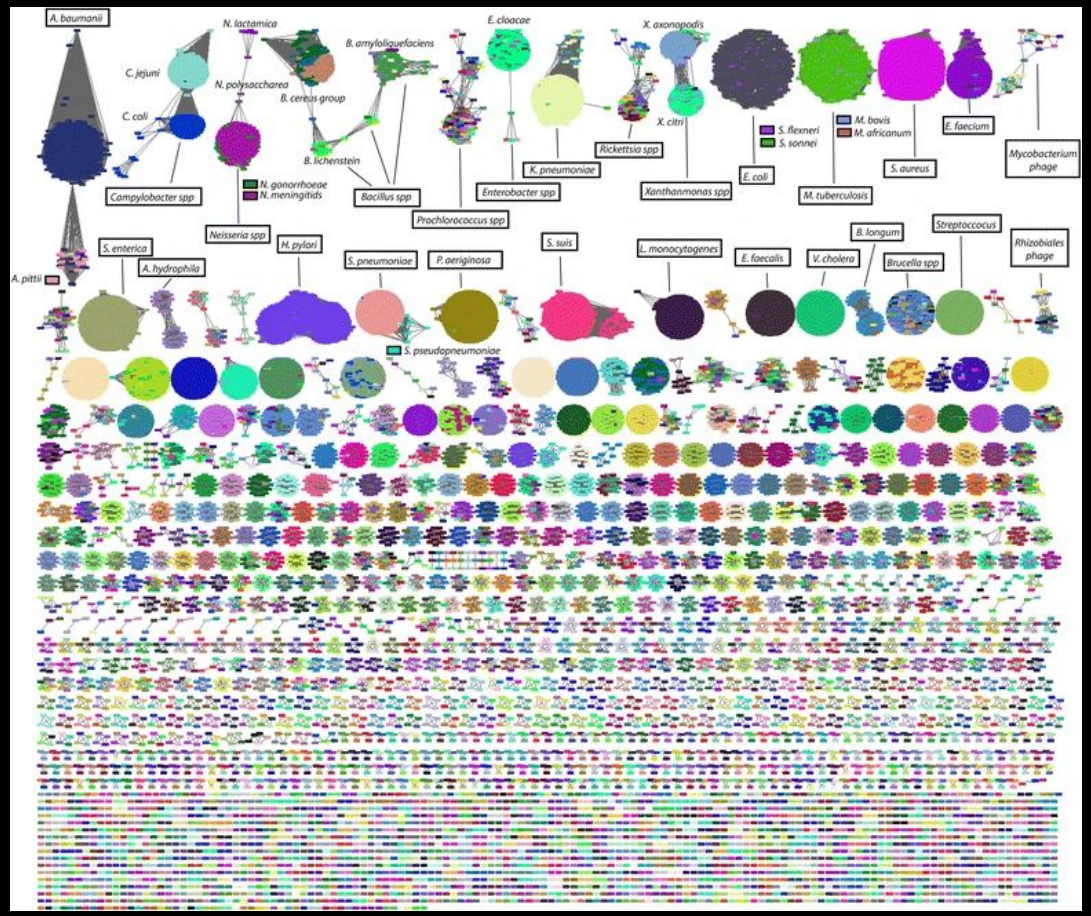


Doesn't fit!

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

Goals of representation

Compute global dissimilarity between sequences in large datasets



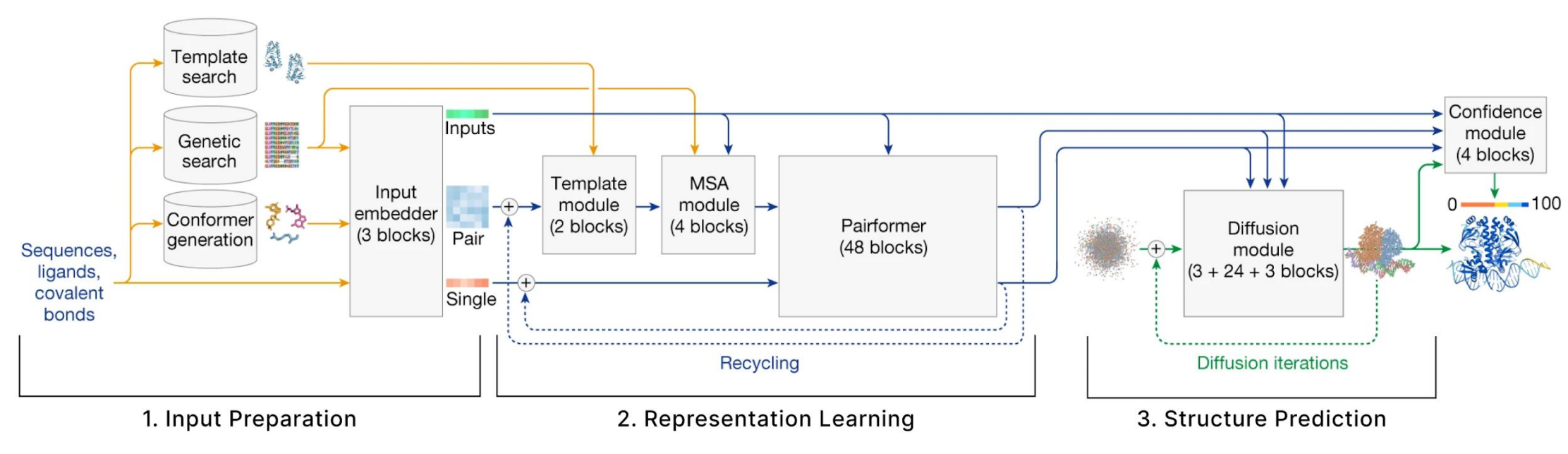
>54,000 genomes clustered using MinHash (MASH) distance

Ondov et al (2016) *Genome Biol*

Goals of representation

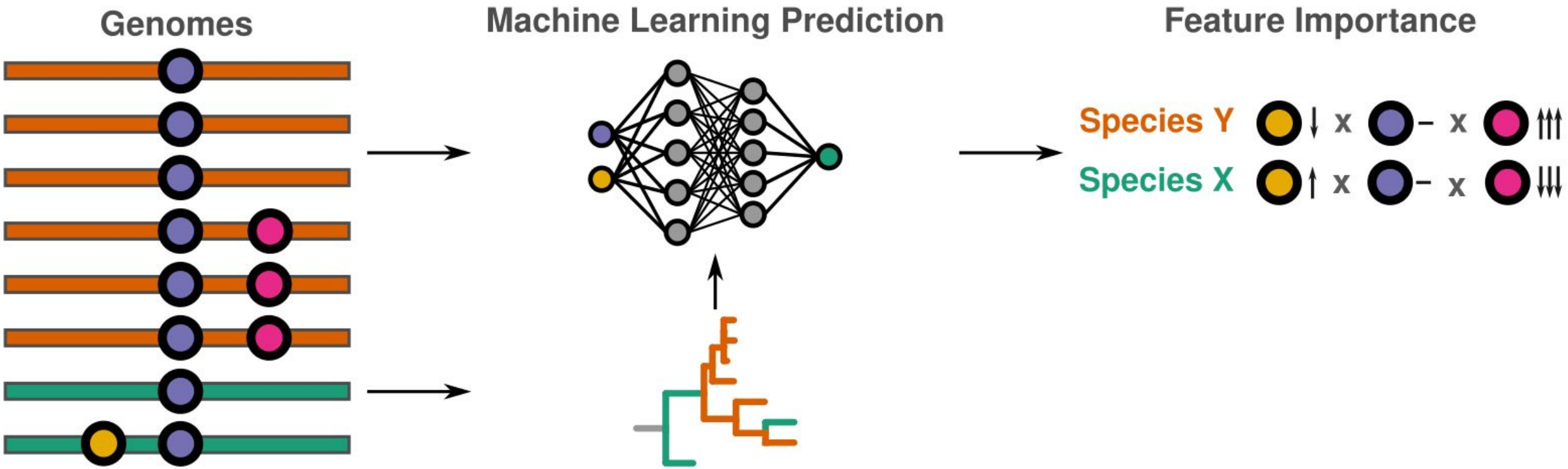
Find best encoding for machine-learning classification

Sequence / structure encoding for AlphaFold



Simon and Silberg (2024), "The Illustrated AlphaFold"
https://elanapearl.github.io/blog/2024/the-illustrated-alphafold/

Predicting ability to infect certain species from genome



Sequence representations

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

Biological Sequences

Primary structure is just the sequence

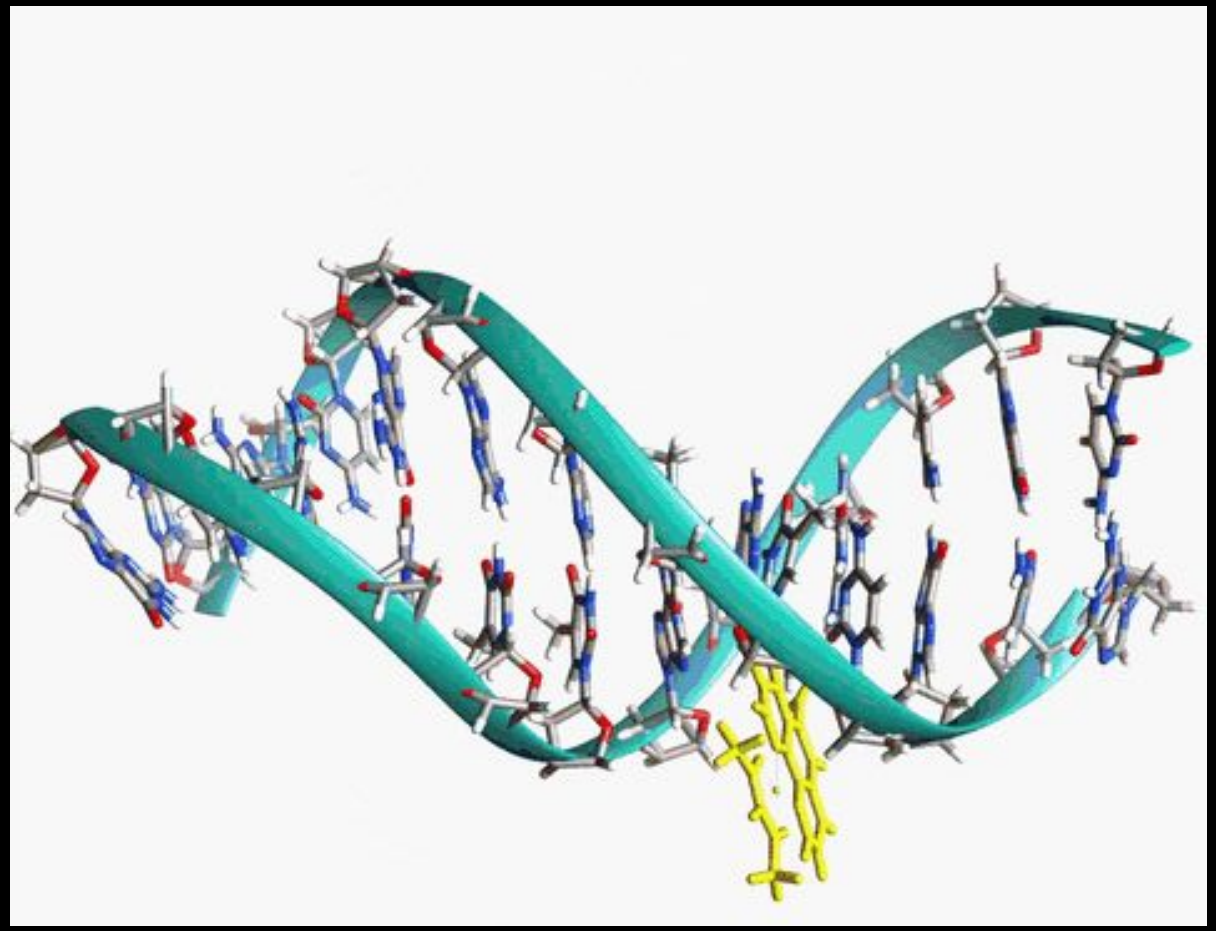
Higher levels of "structure" describe the three-dimensional features of molecules

DNA **…ACCGAATTTACGATACATG…**

Protein **…MLQELIVNEW…**

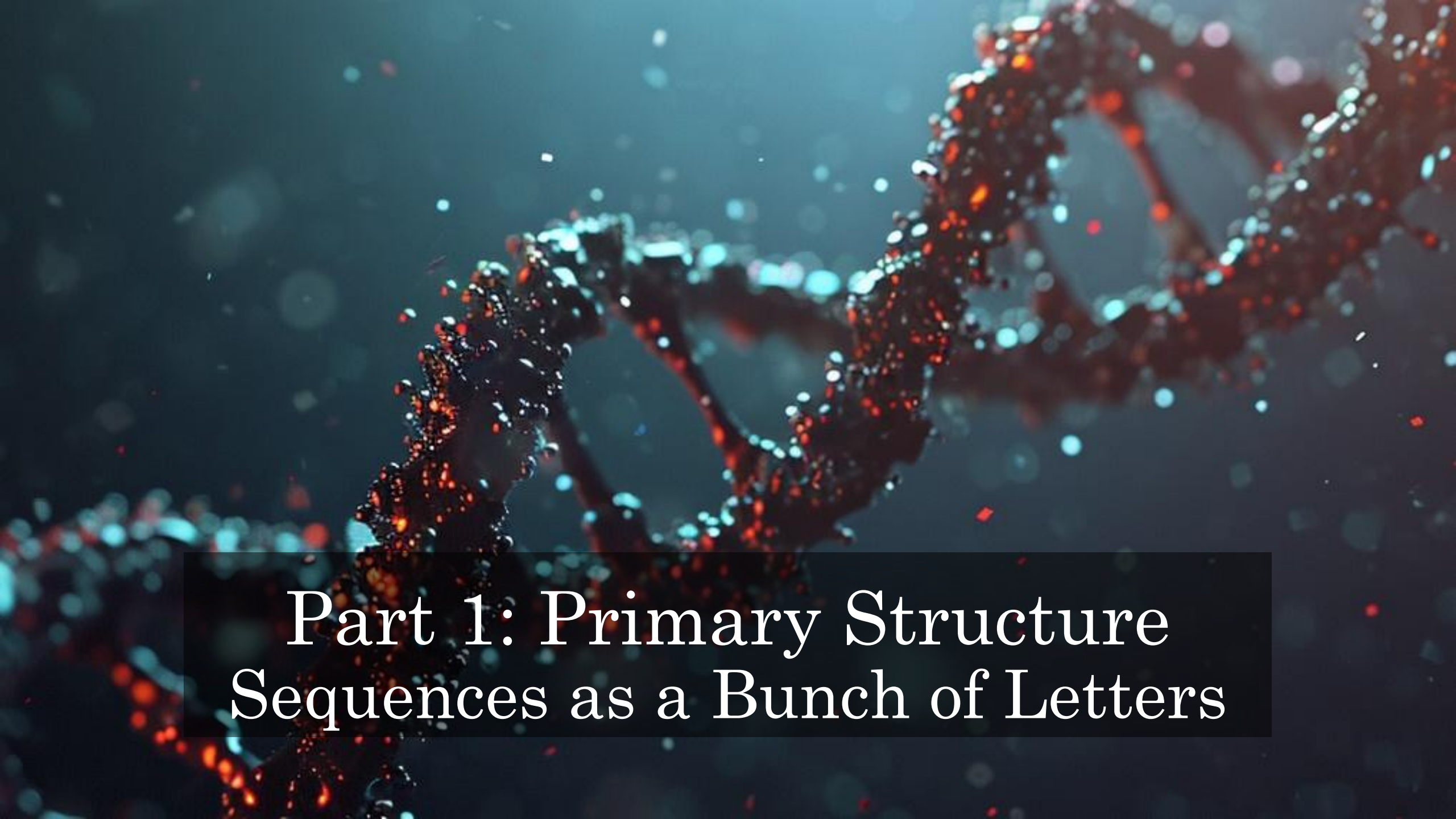
Sequence Representations of DNA

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



?

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



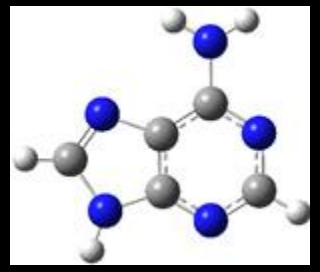
Part 1: Primary Structure
Sequences as a Bunch of Letters

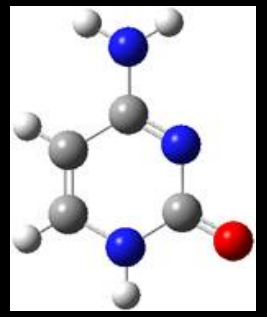
The most common representation is (as you have already seen) a STRING representation with an alphabet of four letters

{A,C,G,T}

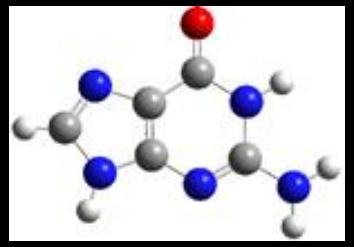
But there is a lot more we can do.

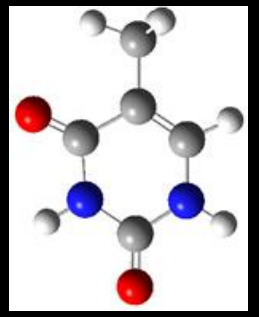
Meet the nucleotides!





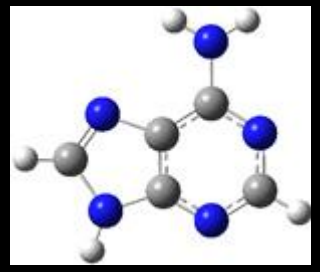
A	G
C	T

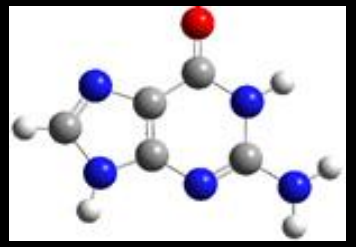




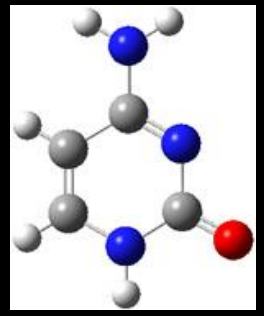
Degenerate characters

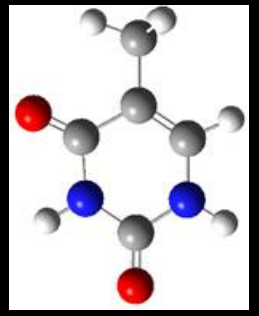
Every pair of nucleotides has something in common

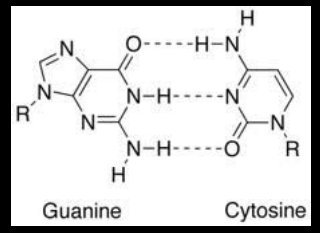




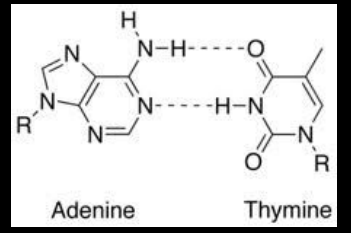
A	G
C	T



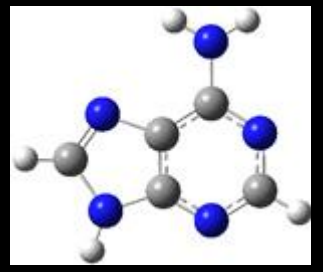


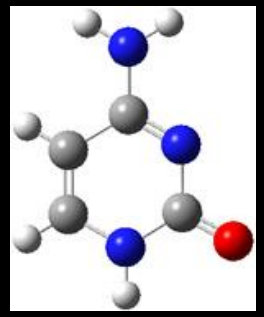


STRONG vs. WEAK base pairing

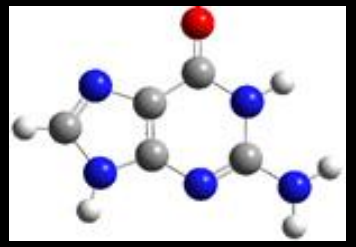


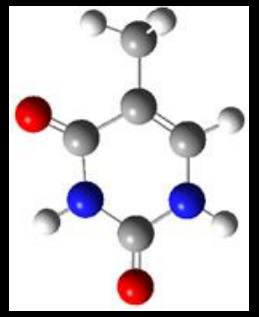
Degenerate characters





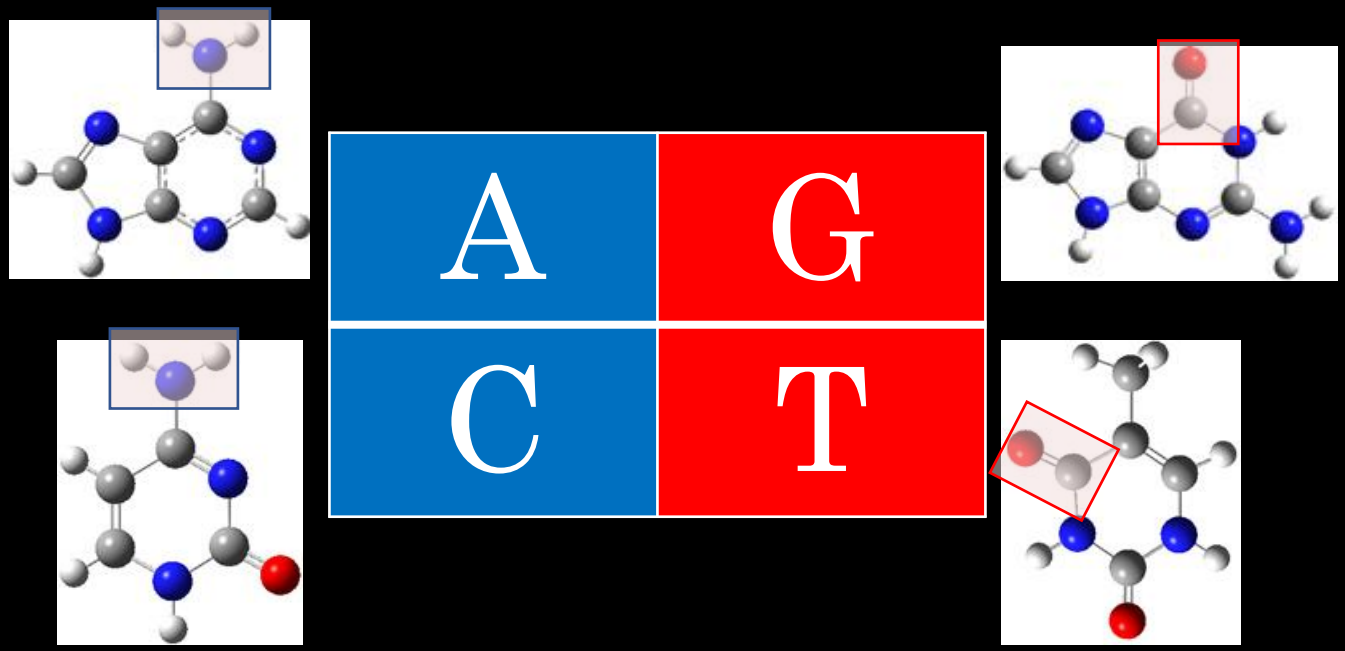
A	G
C	T





PURINE (large) vs. PYRIMIDINE (small)

Degenerate characters



AMINO vs. KETO functional groups
(rarely useful)

IUPAC nomenclature – All possible sets

A	C	G	T	Code	¬Code
■				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 more frequent {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

R/Y recoding

GTCTAAAAAGTTCAAGGTTT
AACAAAGAAAATGAAGGTAT

Original gene sequences
(distance = 8/20)

RYYYRRRRRRRYYYRRRRYYY
RRYRRRRRRRRRYRRRRRYRY

Recoded gene sequences
(distance = 5/20)

Highlights the rarer (on average) changes

Word frequencies: *k*-mers

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

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

Nucleotides (k=1): {A,C,G,T}
Dinucleotides (k=2): {AA,AC,…,TT}
Trinucleotides (k=3): {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%

Total number of k-mers in a sequence of length l = l – k + 1

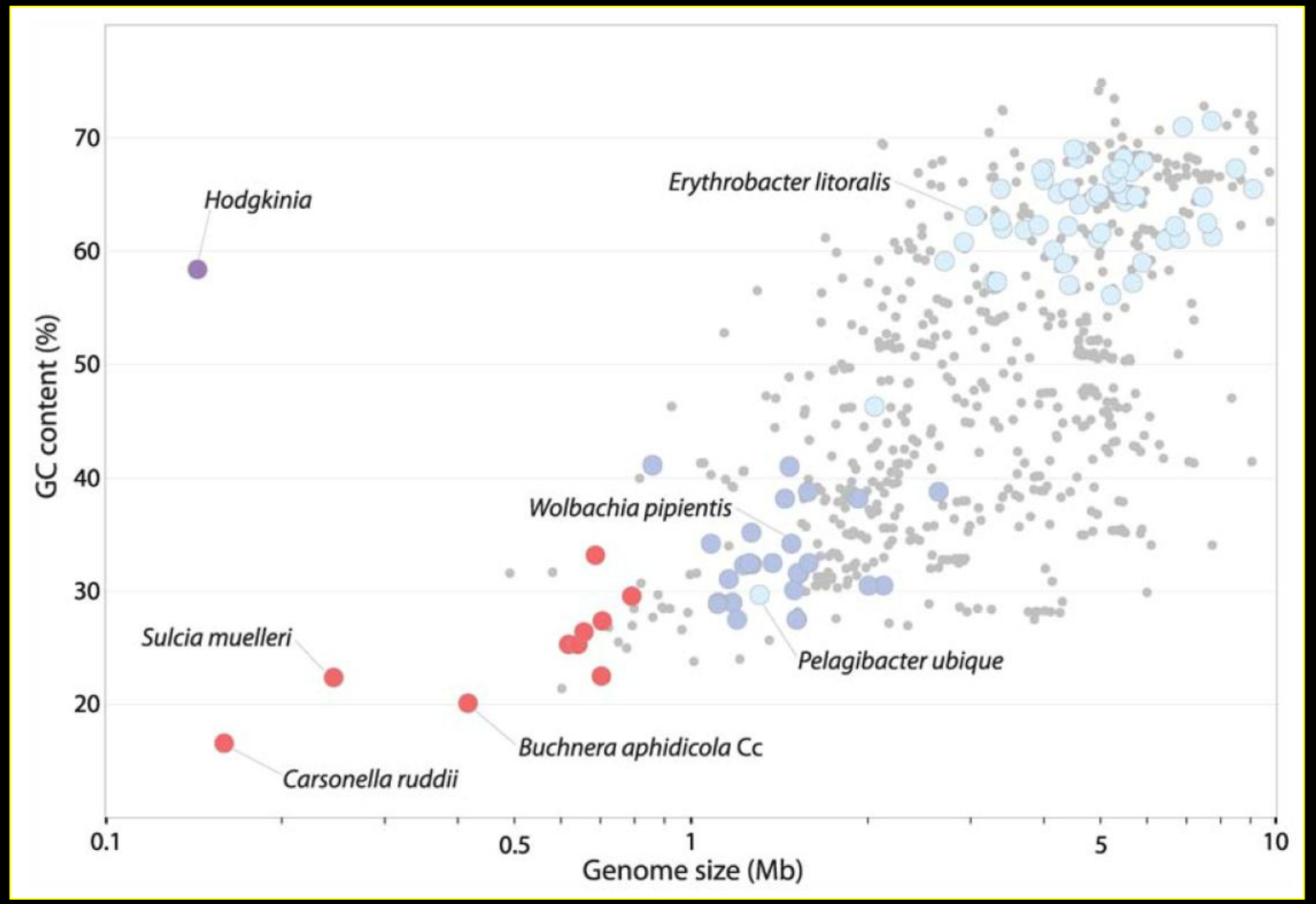
k = 1		k = 2		k = 3	
A	**6/20** = 0.30	AA	**1/19** = 0.053	AAA	**0/18** = 0.00
C	0.25	AC	0.053	ACC	1/18 = 0.056
G	0.35	…	…	…	…
T	0.10	GC	0.158	TTT	0

Sequence composition (k>1)

k-mers are usually an *overlapping representation*

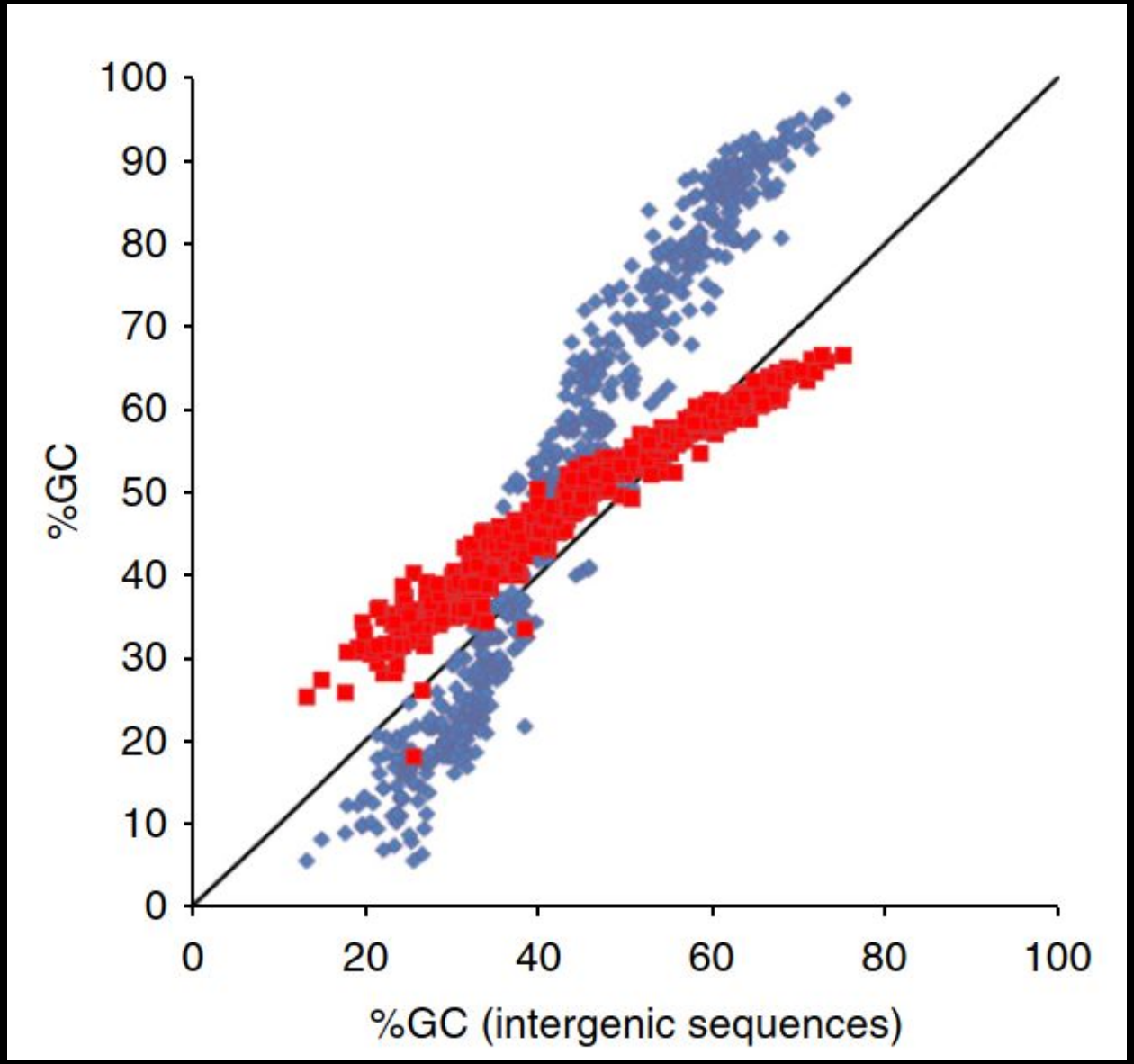
AC	1
CC	1
CG	1
GG	1
GC	1

G+C content of bacterial genomes



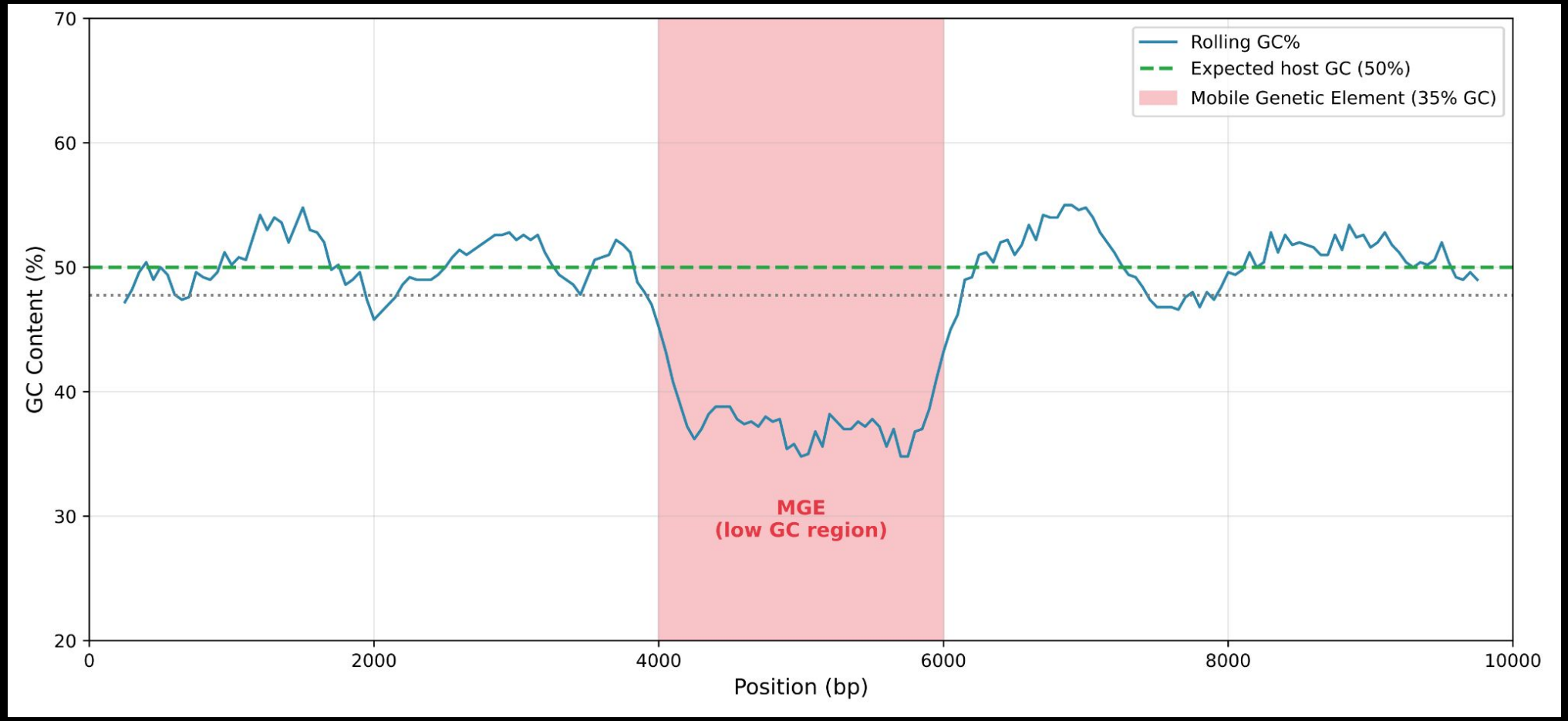
McCutcheon et al. (2009) *PLoS Genet*

G+C content varies *within* genomes



Hershberg (2016) *Encyclopedia of Genes and Genomes*

Mobile genetic elements can (sometimes) be found using GC%



k-mer variants: 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

Can also tailor literal / wildcard combinations to specific expected patterns

k-mer variants: Degeneracy

Length k

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

k=2: { AA, AB, AC, AD, AG, …, VV }

15 letters in IUPAC alphabet, 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…

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

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

Hmmm.

This is a problem we will return to

Tokenization: Escaping the tyranny of fixed length

- Why should we rely on a fixed k?
- Tokenization builds variable-length representations based on abundance in a given training set
- Frequently used in transformers and other "deep learning" architectures.

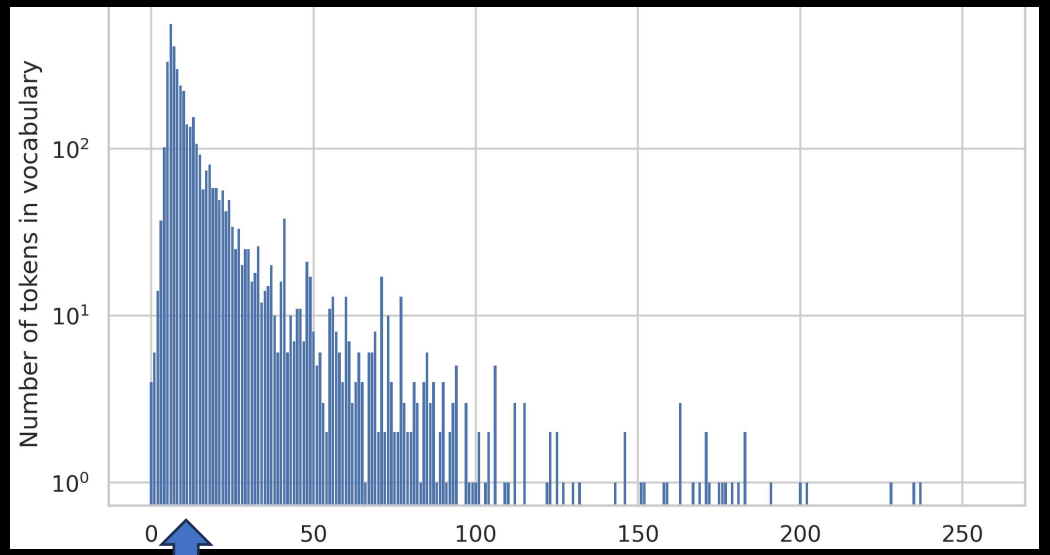
Tokenization

Example: complete 6-mer decomposition - 4^6 k-mers = 4096

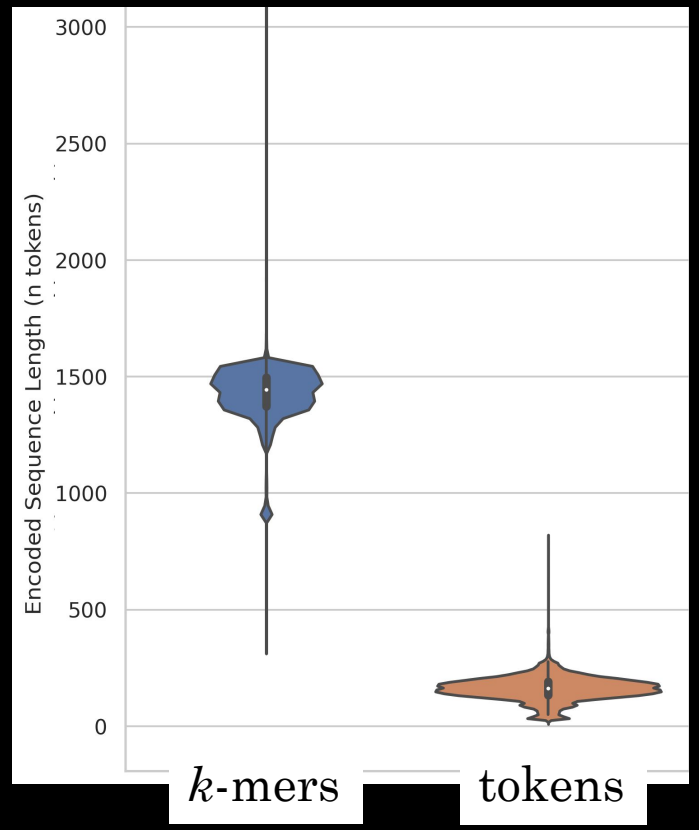
16S ribosomal RNA gene from bacteria (~1500 bp in length)

What if we try tokenization?

Much wider range of features



Sequence representations are more compact

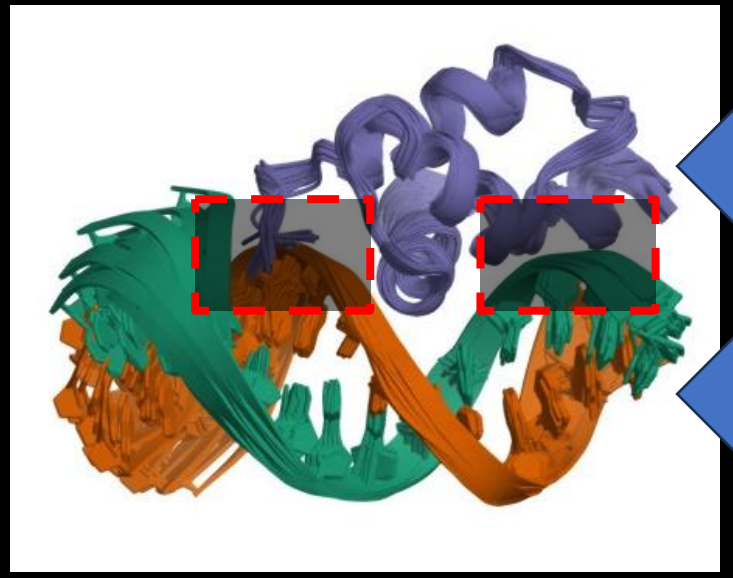


Alex Manuele, MCS thesis

Advantages of Word-Based Representations

- SPEED: Instead of computationally demanding sequence alignment (coming soon), *k*-mer and token counts can be placed in a lookup table that can be rapidly searched

- CUSTOMIZATION: You can do an exhaustive count of all words of a given *k*, but you can also tailor the set of representations based on knowledge of the problem

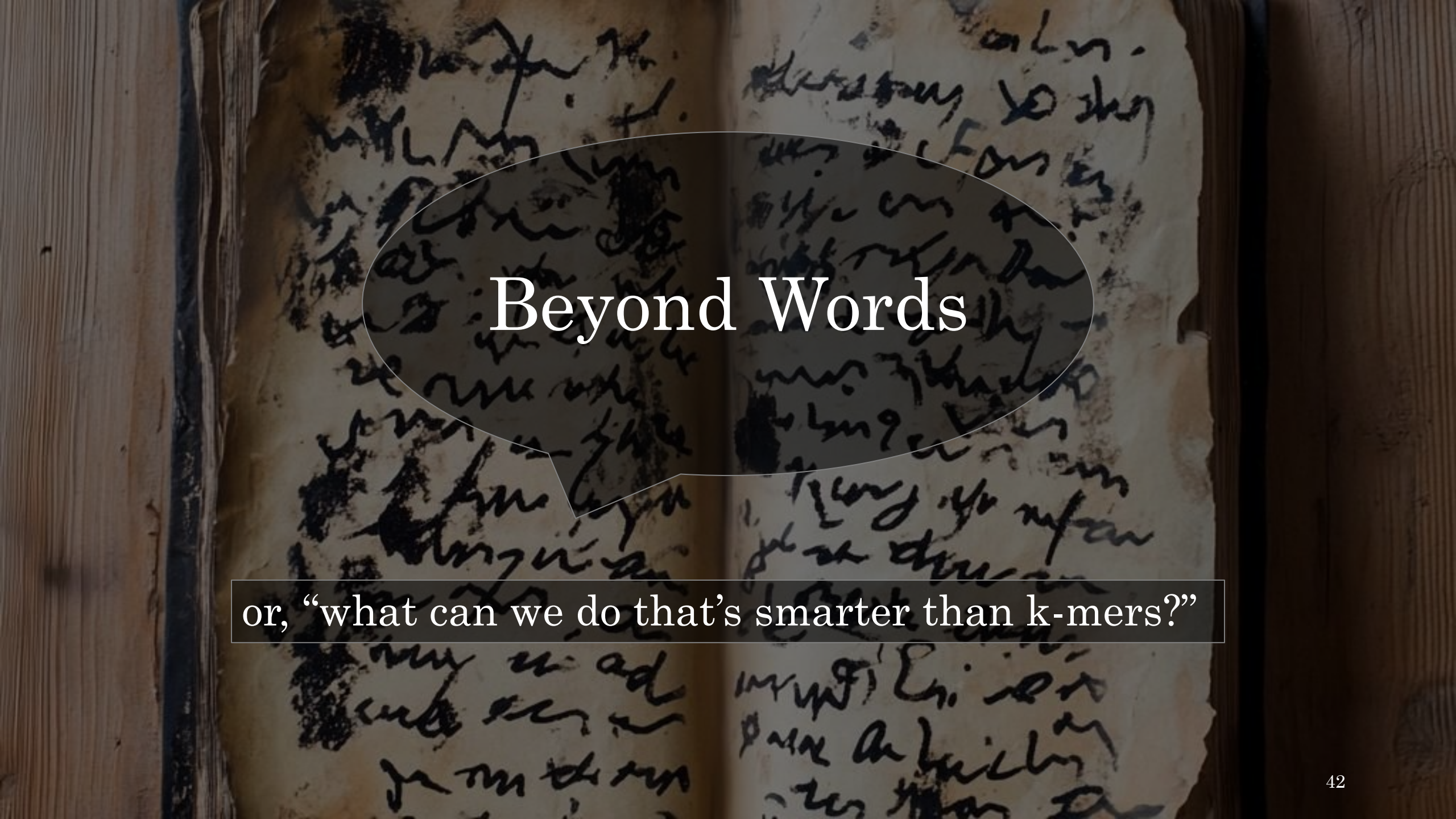


https://www.rcsb.org/structure/2O9L

Disadvantages of Word-Based Representations

- LOSS OF CONTEXT: You get a compositional summary, but you lose all information about which patterns are close to each other in the sequence.

- REDUNDANCY among overlapping words – can be inefficient and introduce big correlations in your data
- NOVEL DATA can struggle to generalise to new data with very different word distributions (e.g., a new taxa!)



Beyond Words

or, “what can we do that’s smarter than k-mers?”

Randomly sampling *k*-mers

- The idea: *k*-mers are great, but there sure are a lot of them.
  - If *k*-mers are too short, we can lose key information that differentiate genomes
  - If *k*-mers are too long, there will be an overwhelming number of them and many will be unique

- We can instead define an "appropriate" *k* and sample randomly from the resulting *k*-mers

{ AA, AC, AG, AT, CA, CC, CG, CT, GA, GC, GG, GT, TA, TC, TG, TT }

Ondov et al. (2016) *Genome Biol*

minHash

- Problems with random *k*-mer sampling:
  - Potential correlation among sampled *k*-mers
  - Are you always going to choose the same *k*-mers? Will they always make sense?

- Hashing: a potential solution
  - Use a hash function to map k-mers to some value
  - Hashing is deterministic and can map very similar k-mers to very different values

```
AAAAAAAAAA → 583250
AAAAAAAAAT → 385325
```

Ondov et al. (2016) *Genome Biol*

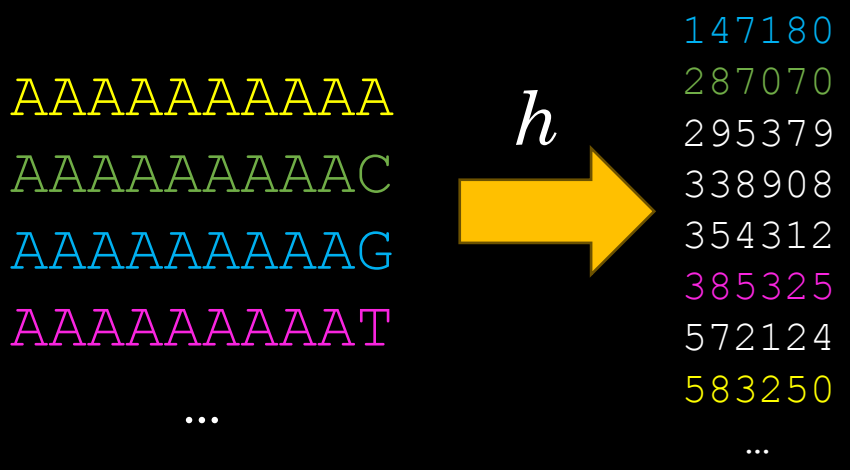
Sketching

- Compute the set of k-mers

AAAAAAAAAA
AAAAAAAAAC
AAAAAAAAAG
AAAAAAAAAT
…

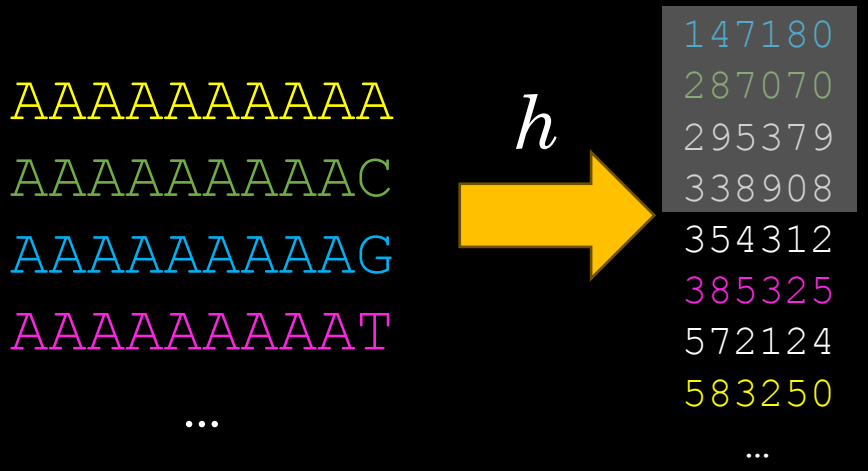
Sketching

- Apply a hash function that changes their order



Sketching

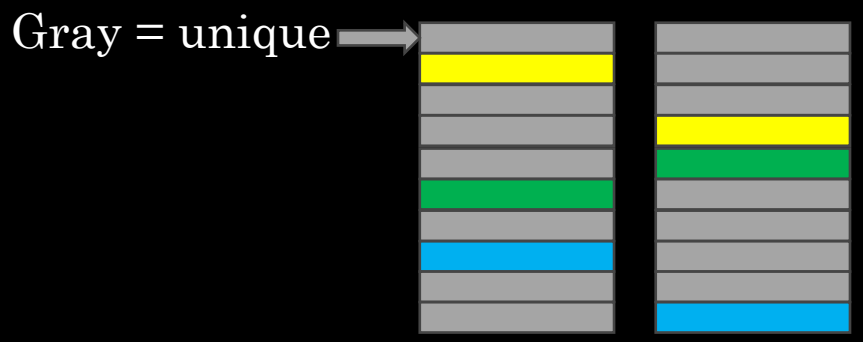
- Take the smallest hash function values (the *minHash*)



- This is the new sequence representation! It is a random and representative sample of the original sequence

Similarity Calculation

- Compare the minHashes of two genomes, and determine their Jaccard similarity index *J(A,B)*



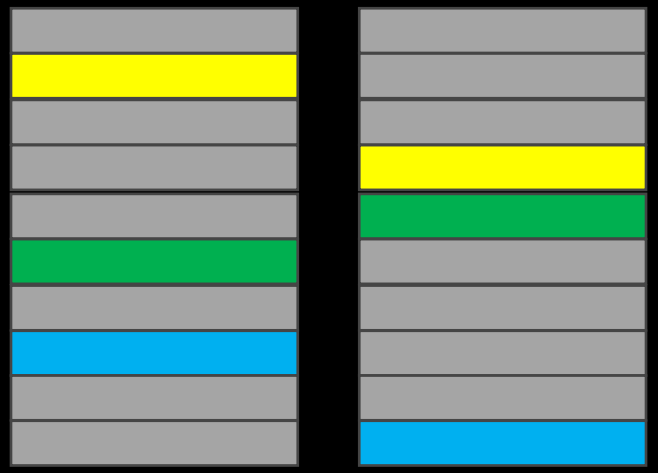
$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \approx \frac{|S(A \cup B) \cap S(A) \cap S(B)|}{|S(A \cup B)|}$$

True distance
(all k-mers / hashed values)

Jaccard index
(distance based on sketches)

Similarity Calculation

- Compare the minHashes of two genomes, and determine their Jaccard similarity index *J(A,B)*



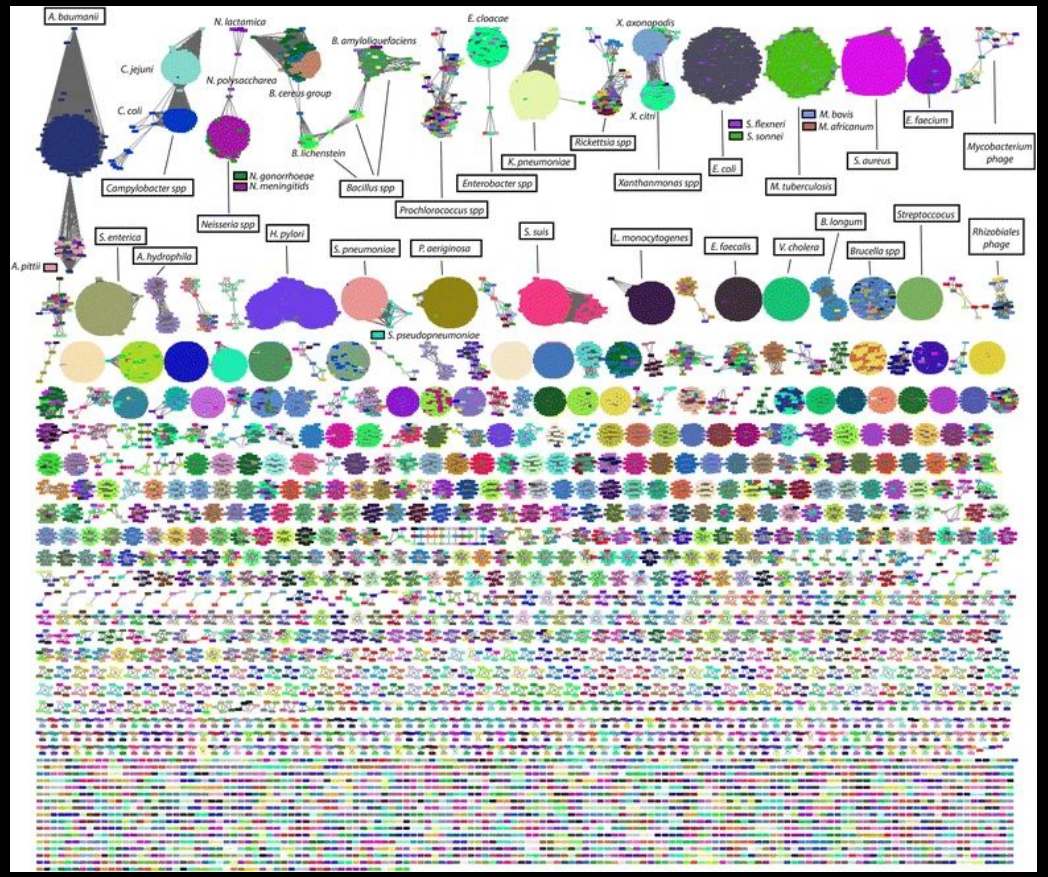
Intersection of both hashed sets: 3 elements

$$\frac{|S(A \cup B) \cap S(A) \cap S(B)|}{|S(A \cup B)|}$$

Total number of unique values in both hashes = 17

Jaccard index = 3/17 = **0.176** (not especially similar)

Mash Distance



$D \leq 0.05$, p-value $\leq 10^{-10}$

Jaccard index

$$D = -\frac{1}{k} \ln \frac{2j}{1+j}$$

k-mer length

Why is it non-linear?

Because distances between genomes do not increase linearly with number of mutations!

Ondov et al (2016) *Genome Biol*

DNA2Vec

- Associations among DNA words based on neighbourhood similarity

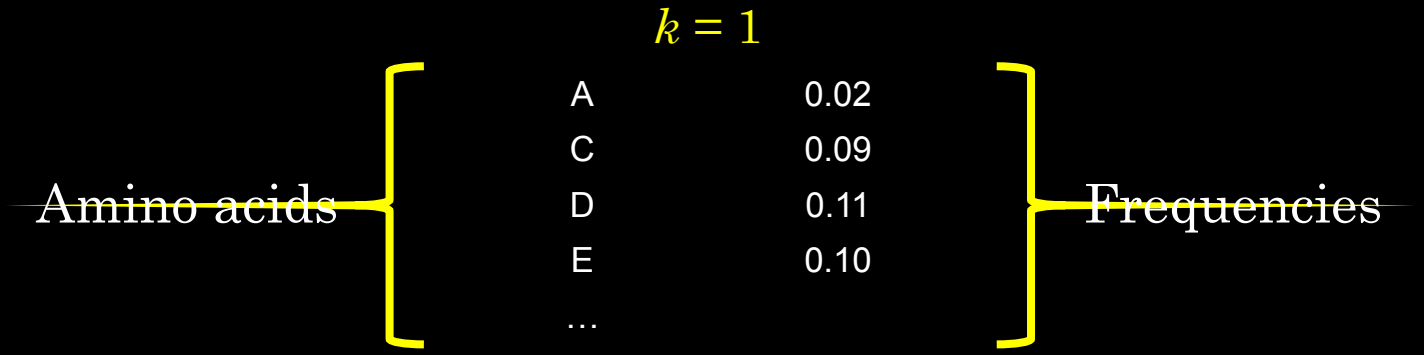
1. Do a *k*-mer decomposition of the sequence
2. Each *k*-mer Z has a neighbourhood of adjacent *k*-mers
3. Train a machine-learning classifier to predict the adjacent *k*-mers, given Z

- More detail coming in classification module

Ng (2017) *arXiv*

Protein sequences: amino acid k-mers

Naïvely: 20^k

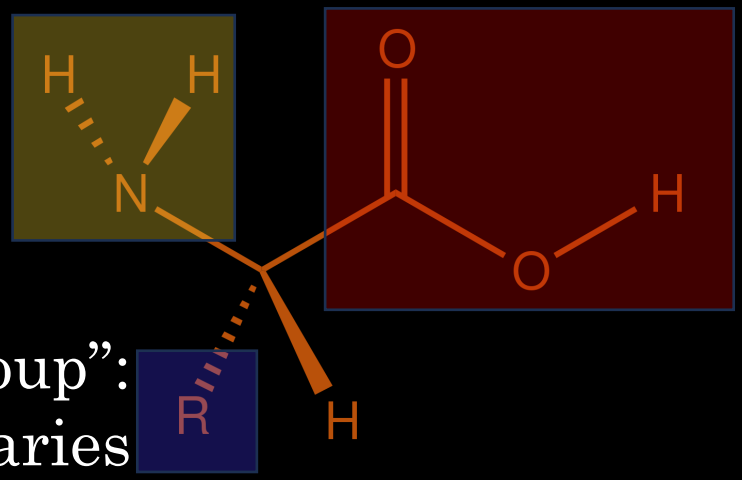


There is no complete degenerate alphabet for amino acids (although there could be – we would just need 2^{20} characters)

We can consider STRUCTURAL and FUNCTIONAL categories instead

General Amino Acid Structure

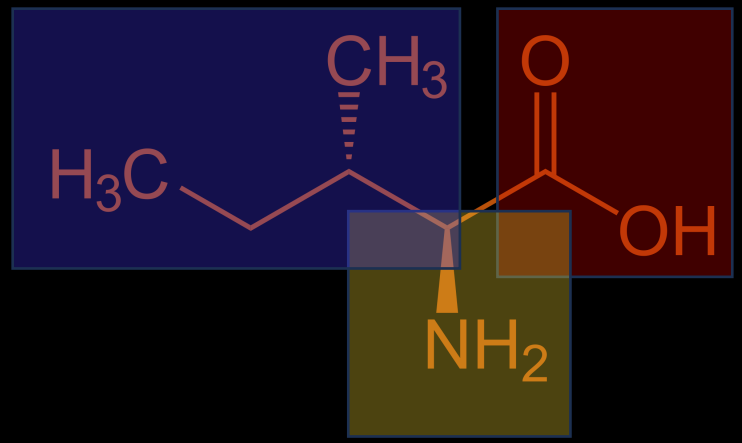
Amino group
(the "front")



Carboxyl group
(the "back")

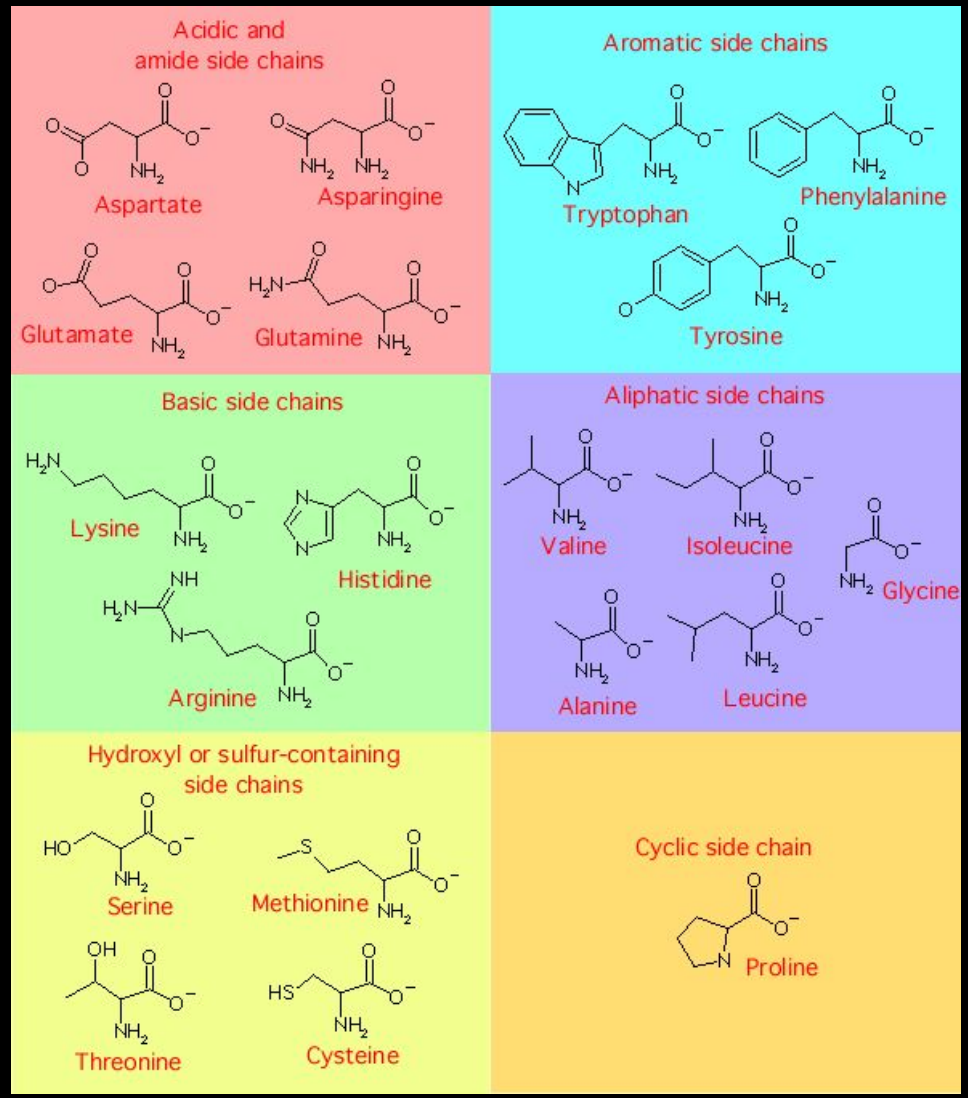
"R group":
varies

Isoleucine



Again, from Wikipedia

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
A C D E F G H IV K L M N P Q R S T W Y

n = 19

This is one of many possible sets of groupings!

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

Correlation representations

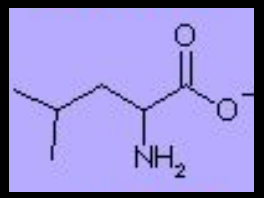
- e.g., pseudo-amino acid composition
- Look at global correlations θ_i of chemical / structural features at a series of distances λ_i

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

protein length

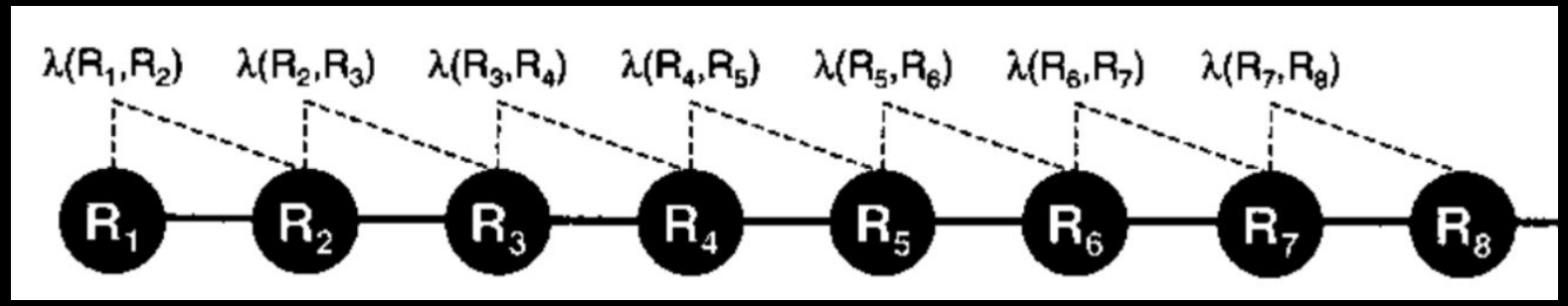
correlation of adjacent amino acids ($\lambda = 1$)

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



Leucine
Super-hydrophobic

Example: hydrophobicity (amino acid aversion to water)

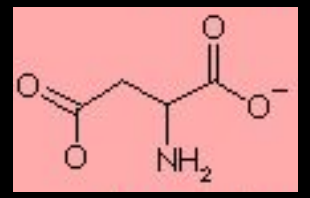


M A Q D Q K E K

74 41 -10 -55 -10 -23 -31 -23

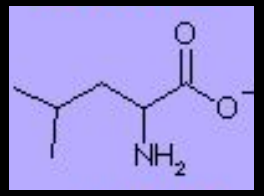
\+ - + + + + +

High average correlation



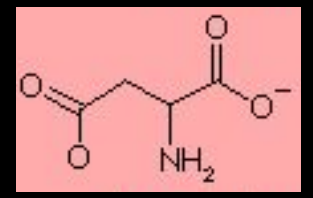
Aspartic aciD
Super-hydrophilic

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

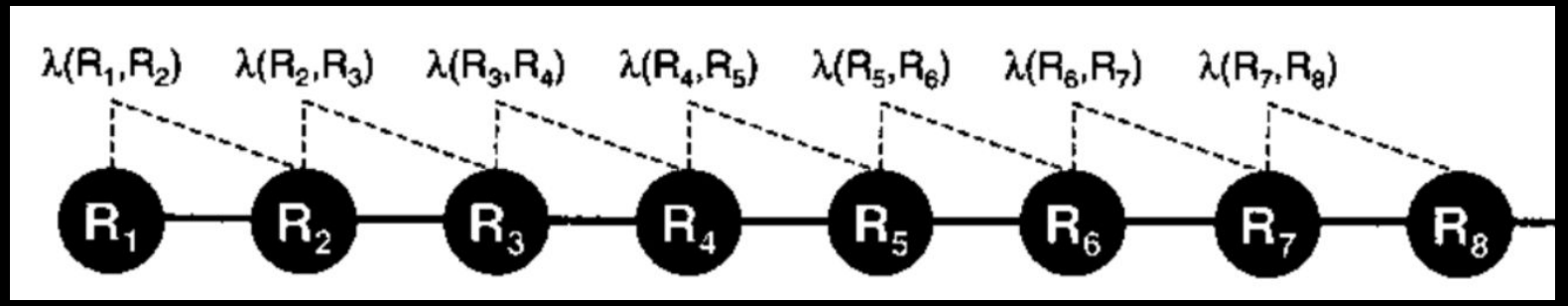


Leucine
Super-hydrophobic

Example: hydrophobicity (amino acid aversion to water)



Aspartic aciD
Super-hydrophilic



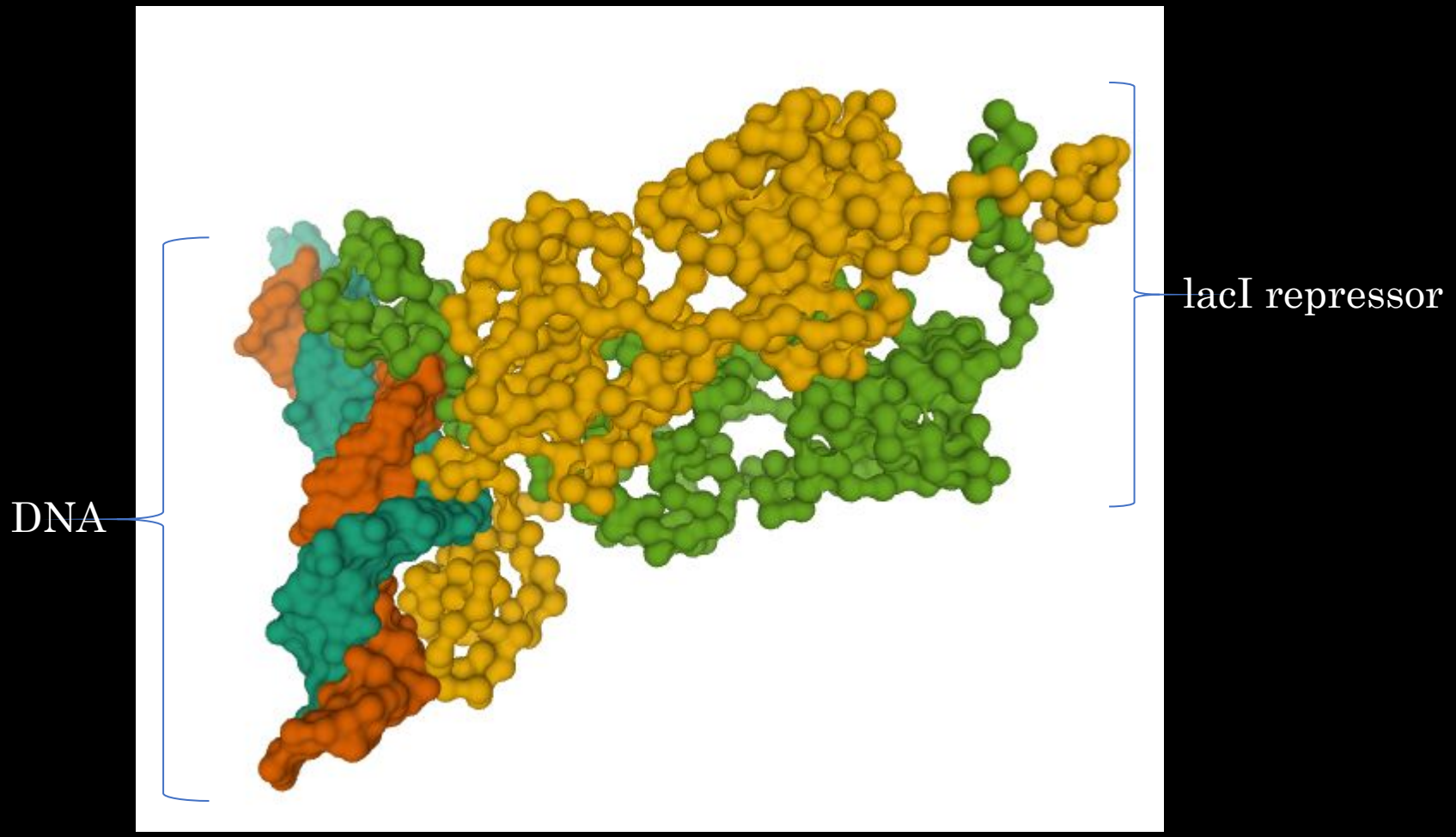
Q A Q D A K A K

-10 41 -10 -55 41 -23 41 -23

- - + - - - -

Poor average correlation

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

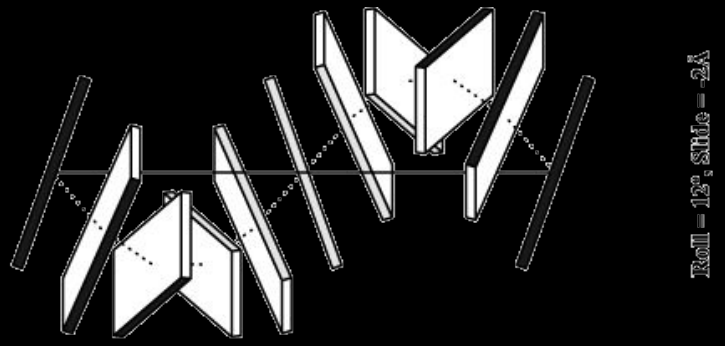


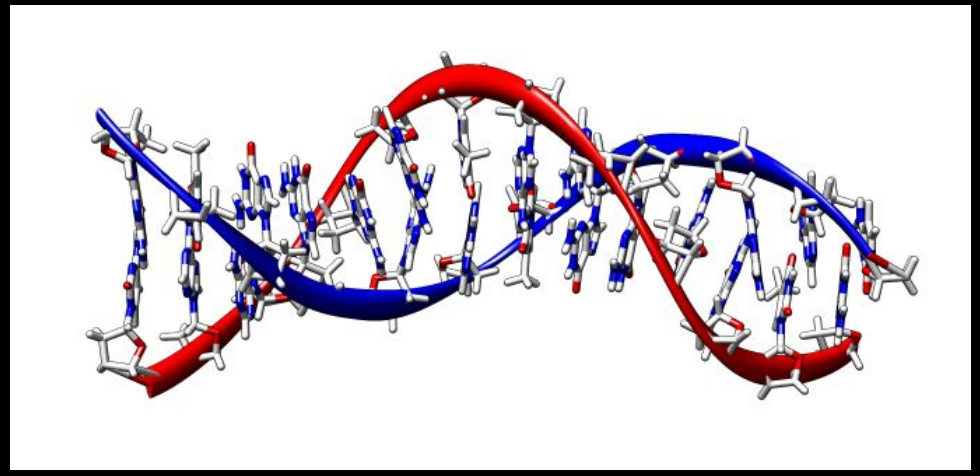
Part 2: Structural representations

Structural representations of DNA

- Two ways to think about DNA structure:

Geometric parameters



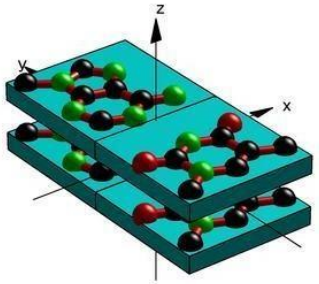


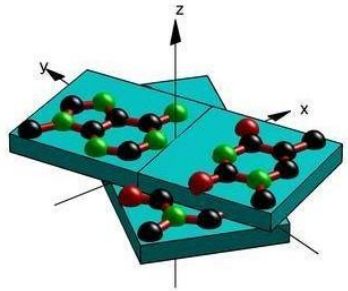
Atomic coordinates

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

We can compute the…

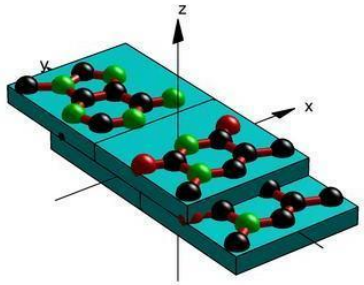
Rise(dz)

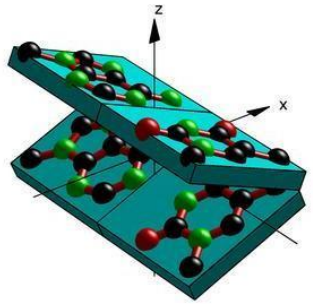




Twist(rz)

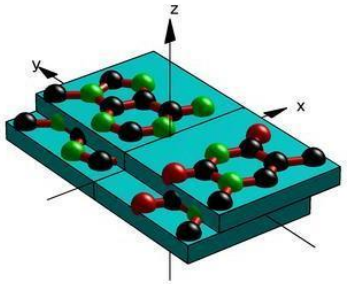
Slide(dy)

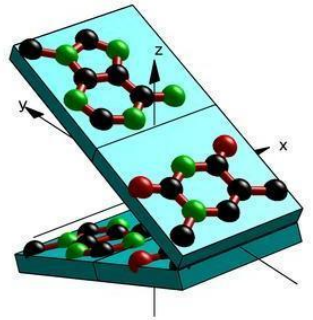




Roll(ry)

Shift(dx)





Tilt(rx)

Between consecutive nucleotide pairs in the helix

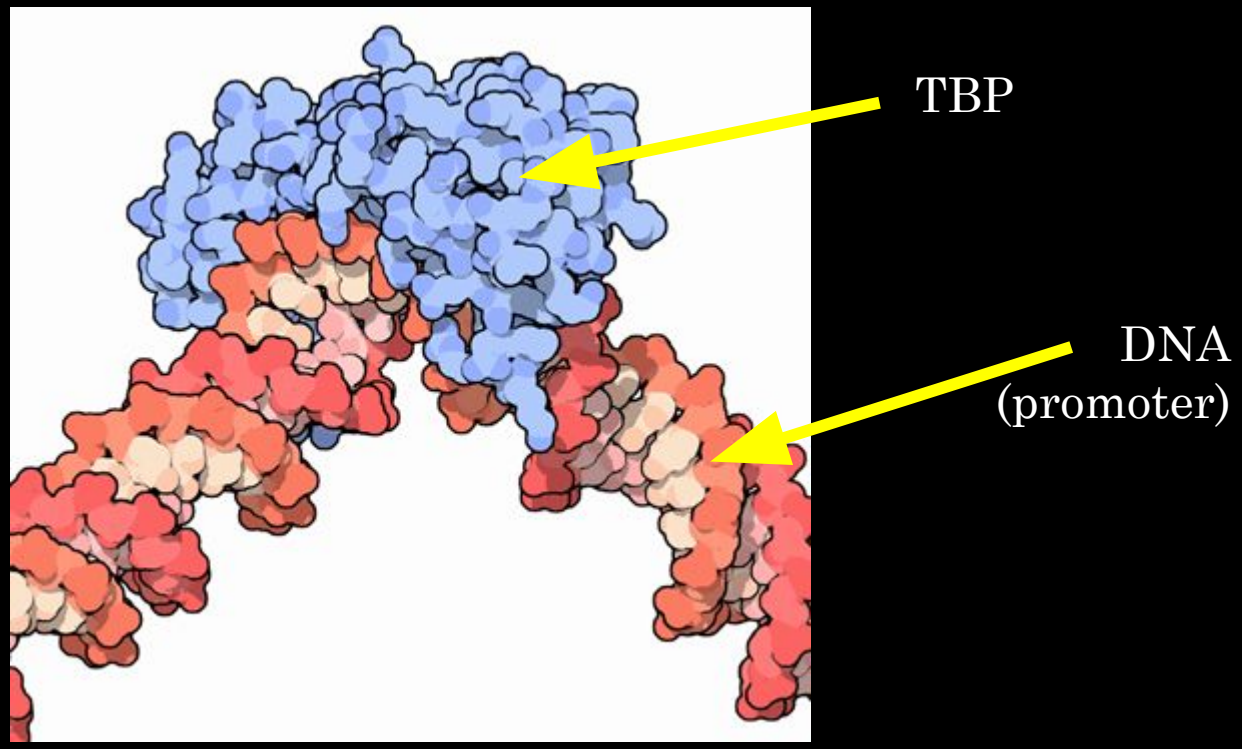
gibk26.bse.kyutech.ac.jp/jouhou/readout/help.html

Static parameters (twist, roll)

- AND -

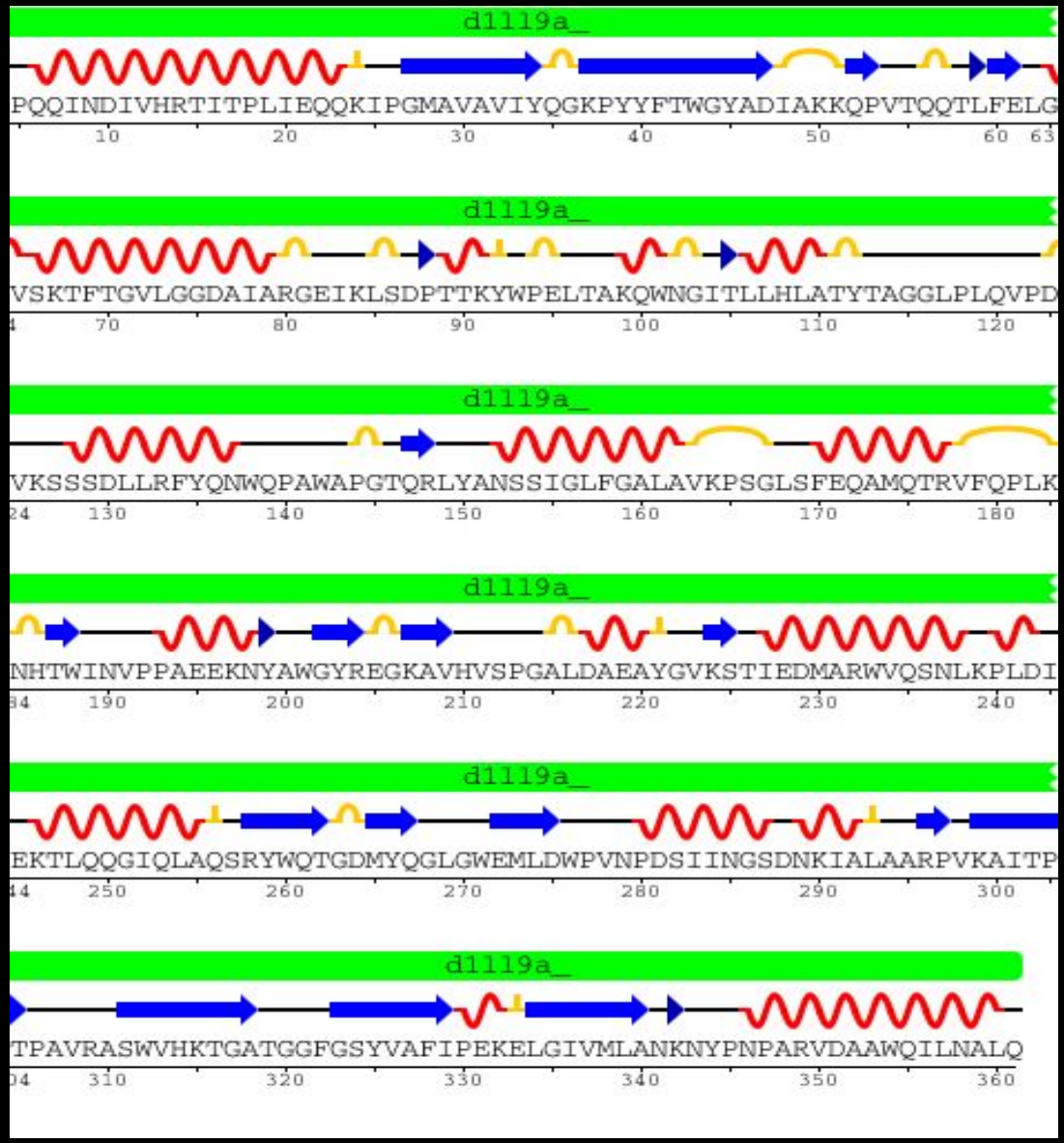
Dynamic parameters (flexibility/deformability)

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



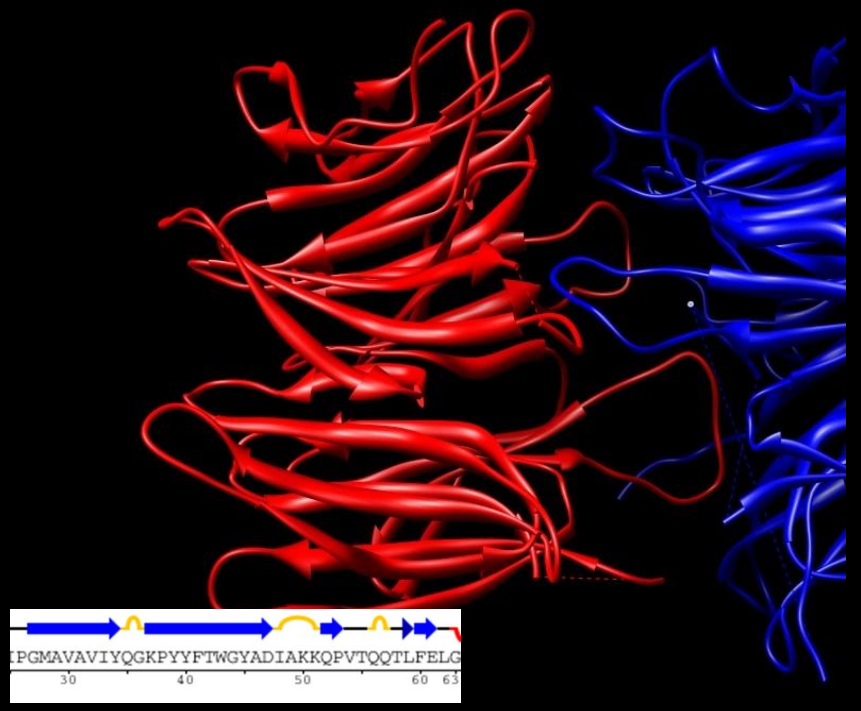
Proteins tend to have more “interesting” structures that govern their behaviour, so structural methods are more frequently applied to proteins than to DNA

Secondary structure



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


QINDIVHRTITPLIEQQ
10
20


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*

Less excruciating options: choose a subset of atoms (e.g., carbon atoms in the backbone)

Summary

1. There are many different applications of DNA, protein, and genome representations

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