

AMAT 583 Lecture 12, 10/3/19

Today: Edit distance examples

Another example of a metric from biology/chem

Review: Let S be the set of DNA sequences, (i.e., strings of letters A, T, C, G of any length).

Def: The edit distance $d_{\text{edit}}: S \times S \rightarrow [0, \infty)$, is given by

$d_{\text{edit}}(x, y) =$ minimum # el. ops. required to transform x into y .

Elementary Operations:

- change one letter
- insert one letter
- remove one letter

Example: $x = \text{ATA}$
 $y = \text{TAT}$

$x = \text{ATA} \rightarrow \text{AT} \rightarrow \text{TAT} = y$.

There's no single elementary operation transforming x into y , so $d_{\text{edit}}(x, y) = 2$.

Exercise : $x = ATCG$ $d_{edit}(x,y) = ?$
 $y = GGTCG$ Ans : 2.

$ATCG \rightarrow GTCG \rightarrow GGTCG$

Let's verify that edit is a metric:

- Property 1) is clearly satisfied
- An elementary operation can always be undone by an elementary operation, so $d_{edit}(x,y) = d_{edit}(y,x)$.
- To establish triangle inequality, need to show that

$$\forall x,y,z \in S, d(x,z) \leq d(x,y) + d(y,z).$$

Let $d(x,y) = m$, $d(y,z) = n$.

Then there's a sequence α of elementary ops. transforming x to y , and a sequence β of elt. ops. transforming y to z . Then α followed by β is a sequence of $m+n$ elt. ops. transforming x to z . Thus $d(x,z) \leq m+n$.

Another example of a metric space from biology

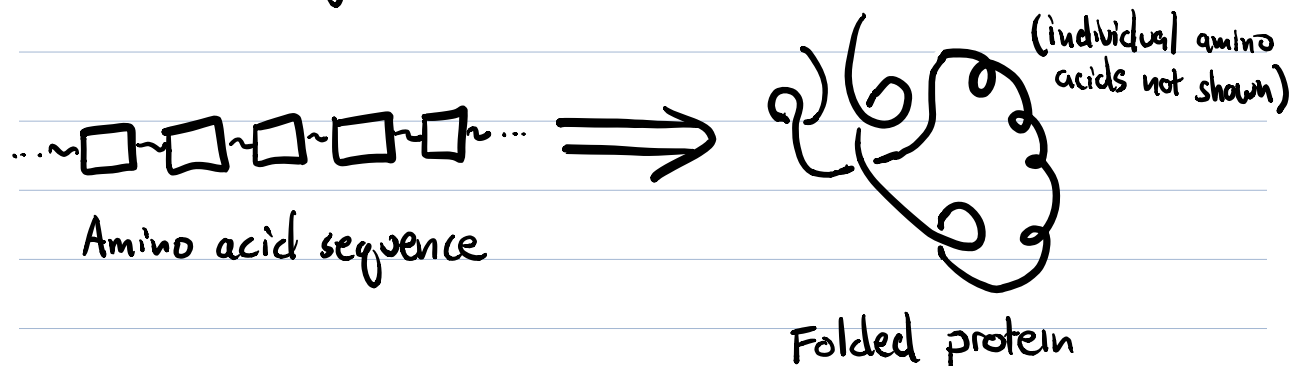
Background: The primary function of DNA is to serve as a blue-print from which proteins are constructed.

Simplified definition of a protein:

A protein is a string of subunits called amino acids connected by covalent bonds.

There are 20 different amino acids, with names like "arginine", "lysine," and "tryptophan."

Proteins fold into complex 3-D structures, with essential biological function (e.g. enzymes, neurotransmitters)



DNA sequences called Genes specify the amino acid sequence of protein.

Rough explanation: Three nucleotides specify one amino acid.

Ex: CGA TTT ACC



Alanine ~ Lysine ~ Tryptophan

Determining the amino acid sequence from the DNA sequence is very easy.

But accurately determining the 3-D structure of the protein from the amino acid sequence is challenging.

This is called the "protein structure prediction problem."

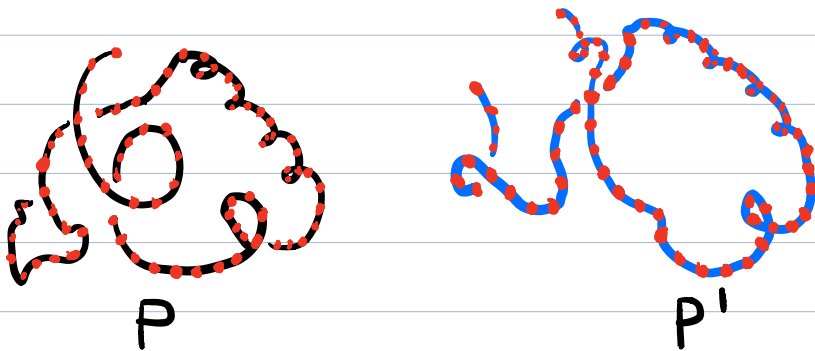
- one of the fundamental problems of computational biology
- applications to drug discovery
- biannual competitions on this problem called CASP
- lots of software available.

Note: In favorable cases, the structure can be determined by experiment, e.g., by a technique called x-ray crystallography. But this is expensive, time consuming.

~~and~~ requires a lot of skill.

Computers are used to get fast solutions.

Question: Suppose I know the folded structure P of a protein. How do I measure the accuracy of a predicted structure P' ?



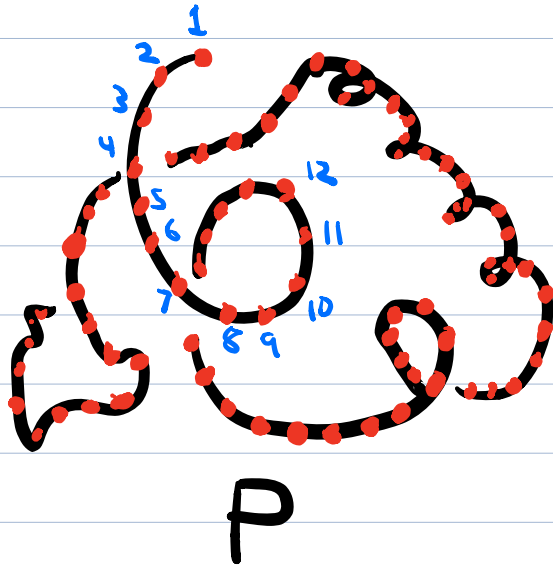
To assess the performance of a structure prediction method, e.g. in a competition like CASP, we need an answer.

Standard Answer: Compute a metric called RMSD (root mean squared deviation) between P and P' .

RMSD is a fundamental tool in the study of molecules.

How to represent the 3-D structure of a protein P mathematically

- Order the atoms of the protein (choice of order doesn't matter).

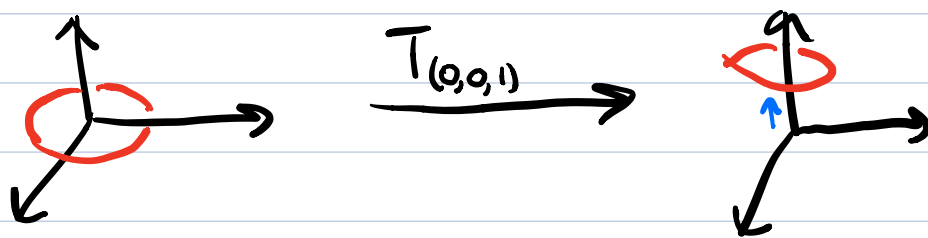


- Let O^n denote the set of all ordered subsets of \mathbb{R}^3 of size n . We think of P as an element of O^n .
- For $P \in O^n$, denote the i^{th} point in P by (x_i, y_i, z_i) .
- Define a function $V: O^n \rightarrow \mathbb{R}^{3n}$ by V is invertible!
 $V(P) = (x_1, y_1, z_1, x_2, y_2, z_2, \dots, x_n, y_n, z_n)$.
This represents the protein's 3-D structure as a single point in a high-dimensional space!

Note: This representation throws away a lot of info about the protein (atom type, bond info), but for many applications, that is ok.

Rigid motions

- A translation in \mathbb{R}^3 is a function $T_{\vec{v}}: \mathbb{R}^3 \rightarrow \mathbb{R}^3$ given by $T_{\vec{v}}(\vec{x}) = \vec{x} + \vec{v}$ for some fixed $\vec{v} \in \mathbb{R}^3$

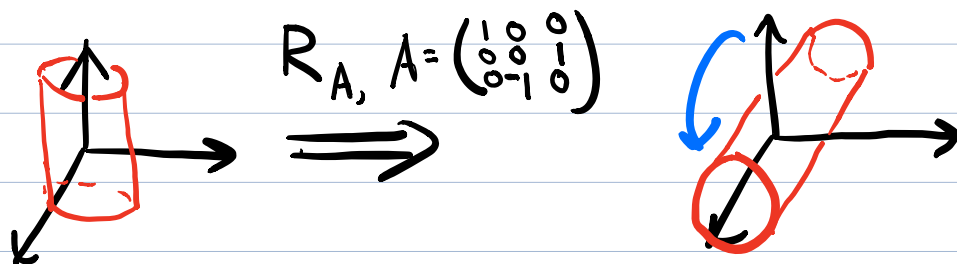


Interpretation: $T_{\vec{v}}$ shifts a geometric object in the direction \vec{v} without rotating.

- A rotation in \mathbb{R}^3 is a function $R_A: \mathbb{R}^3 \rightarrow \mathbb{R}^3$ of the form

$$R_A(\vec{x}) = A\vec{x} \text{ where } A \text{ is a } 3 \times 3 \text{ matrix with determinant } 1$$

Interpretation: R_A rotates a geometric object about the origin in \mathbb{R}^3 .



A rigid motion in \mathbb{R}^3 is a translation followed by a rotation, i.e., a function

$\varphi: \mathbb{R}^3 \rightarrow \mathbb{R}^3$ of the form

$$\varphi = R_A \circ T_{\vec{v}}$$

↑ rotation
↑ translation

Let E be the set of all rigid motions in \mathbb{R}^3 .

Definition: Let P, P' be 3-D structures for a given protein with n atoms, regarded as subsets of \mathbb{R}^3 of size n .

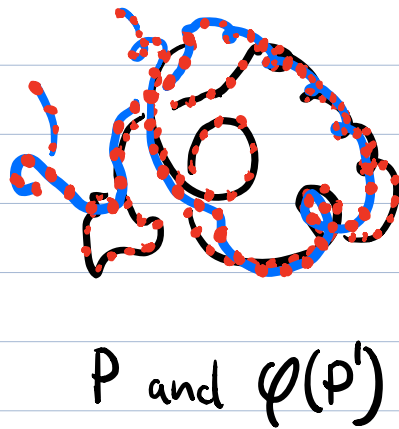
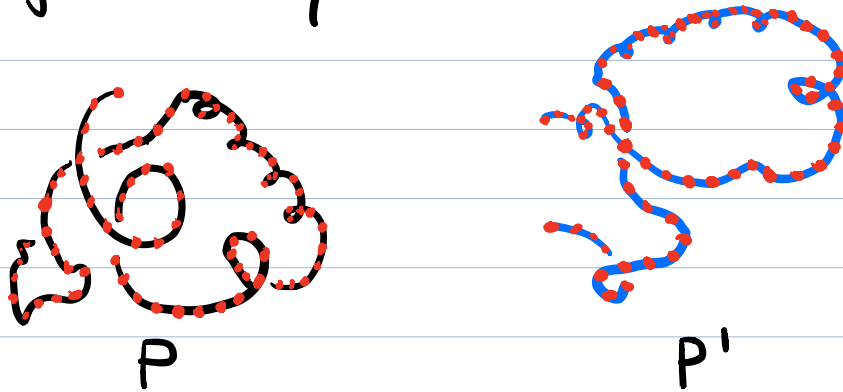
$$\text{RMSD}(P, P') = \min_{\varphi \in E} \frac{1}{\sqrt{n}} d_2(V(P), V(\varphi(P'))).$$

↑
ordinary
Euclidean
distance

↑
rigid motion of P'

Interpretation: To compute $\text{RMSD}(P, P')$,

1) Align P and P' as well as possible via a rigid motion φ



2) Represent P and $\varphi(P')$ as points $V(P), V(\varphi(P'))$ in \mathbb{R}^{3n} .

3) RMSD is the Euclidean distance between these points, normalized so that RMSD doesn't tend to grow as # of atoms grows.

Formally, we regard this as a function

$$\text{RMSD}: O^n \times O^n \rightarrow [0, \infty).$$

This function is symmetric and satisfies the triangle inequality, but we can have

$$\text{RMSD}(P, P') = 0 \text{ if } P \neq P' \text{ but } \varphi(P) = P' \text{ for some rigid motion } \varphi.$$

Here's how we get a genuine metric here:

Define an equivalence relation \sim on O^n by

$$P \sim Q \text{ iff } \exists \text{ a rigid motion } \varphi: \mathbb{R}^3 \rightarrow \mathbb{R}^3 \text{ with } \varphi(P) = Q.$$

Fact: $\text{RMSD}(P, Q) = \text{RMSD}(P', Q')$ if $P \sim P'$ and $Q \sim Q'$

(Exercise: Prove this).

As a consequence, $\text{RMSD}: O^n \times O^n \rightarrow [0, \infty)$ descends to a genuine metric on O^n / \sim .