CS4054
Bioinformatics

Spring 2025

Rushda Muneer

What do we mean by "structure"?

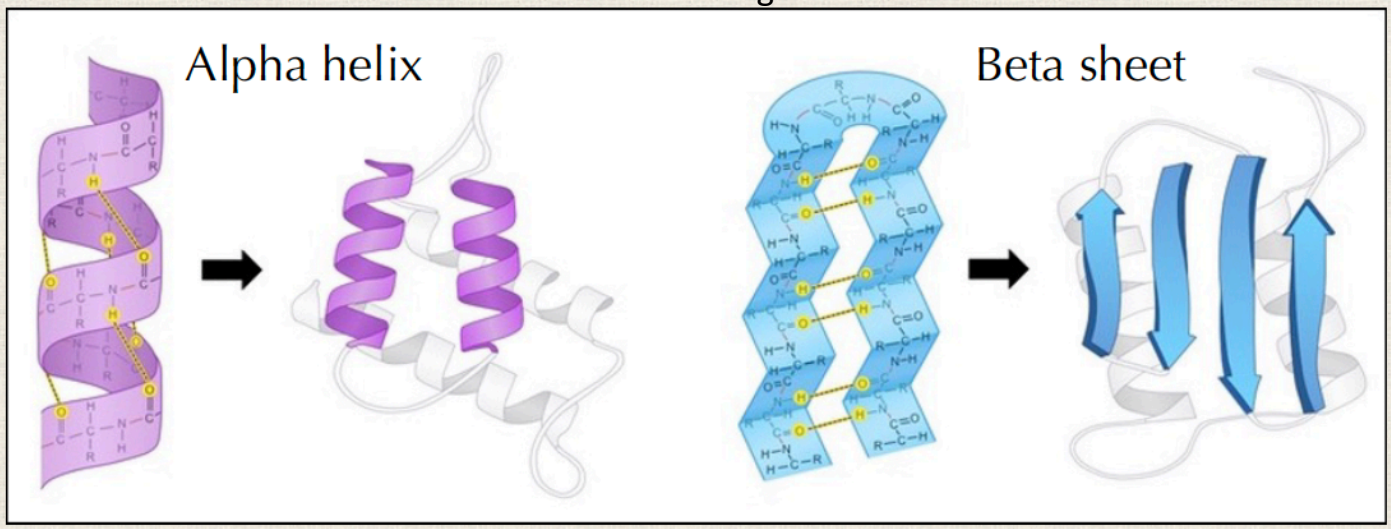
A protein's **primary structure** refers to the amino acid sequence of its **polypeptide** chain.

```
MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFDNPVLPFNDGVYFA
STEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVIKVCEFQFCNDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNF
KNLREFVFKNIDGYFKIYSKHTPINLVRDLPQGFSALEPLVDLPIGINITRFQTLLALHRSYLTPGDSSSGWTAGAAAYYVGYLQPRTFLLKY
NENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSA
SFSTFKCYGVSPTKLNDLCFTNVYADSFVIRGDEVRQIAPGQTGKIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFE
RDISTEIYQAGSTPCNGVEGFNCYFPLQSYGFQPTNGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCVNFNFNGLTGTGVLTESNKK
FLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQDVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLI
GAEHVNNSYECDIPIGAGICASYQTQTNSPRRARSVASQSIIAYTMSLGAENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICG
DSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDPSKPSKRSFIEDLLFNKVTLADAGFIKQY
GDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSA
IGKIQDSLSSTASALGKLQDVVNQNAQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASAN
LAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICHDGKAHFPREGVFVSNGTHWFVTQRNFYEPQIIT
TDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQY
IKWPWYIWLGFIAGLIAIVMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
```

What do we mean by "structure"?

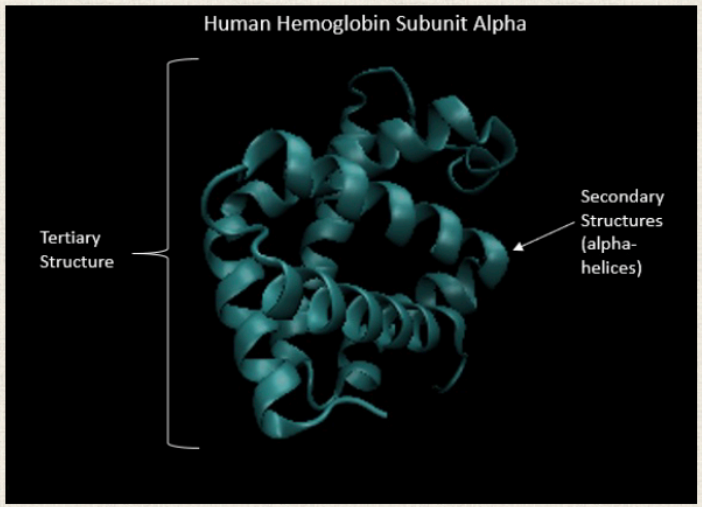
A **secondary structure** is a repeating substructure that forms as a substructure of the overall folded protein.

Forms of Folding



What do we mean by "structure"?

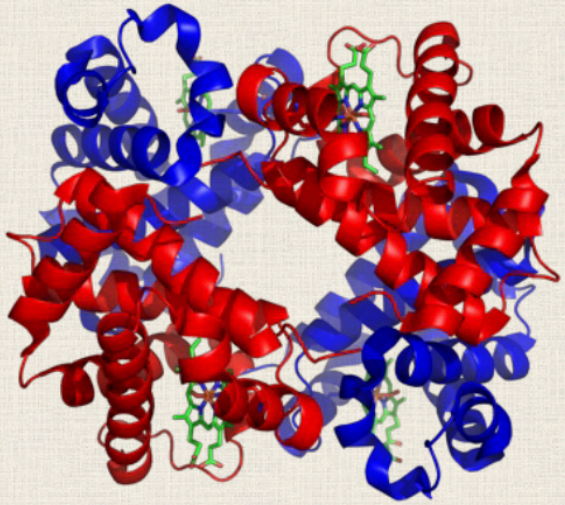
A protein's **tertiary structure** describes its final 3D shape after the polypeptide chain has folded and is chemically stable. This is what we most commonly refer to as the "structure" of a protein.



What do we mean by "structure"?

Some proteins have a **quaternary structure**, which describes the protein's interaction with other copies of itself to form a single functional unit, or a **multimer**.

Hemoglobin is a multimer consisting of two alpha subunits and two beta subunits.



Levels of Protein Structure

Primary structure

Amino acid sequence of the peptide chains.

Secondary structure

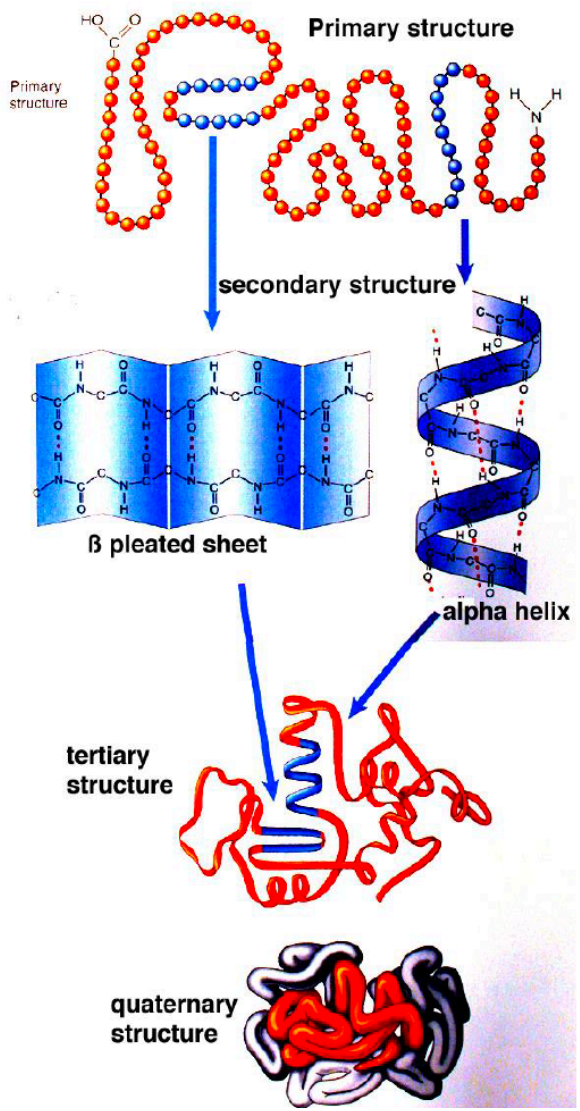
Highly regular sub-structures (*alpha helix* and *strands of beta sheet*)

Tertiary structure

Three-dimensional structure of a single protein molecule

Quaternary structure

Complex of several protein molecules or polypeptide chains.



Homology modeling

This idea serves as the foundation of **homology modeling** for protein structure prediction (a.k.a. **comparative modeling**). By using the known protein structure of a homologous protein as a template, we can in theory improve both the accuracy and speed of protein structure prediction.

Homology modeling

This idea serves as the foundation of **homology modeling** for protein structure prediction (a.k.a. **comparative modeling**). By using the known protein structure of a homologous protein as a template, we can in theory improve both the accuracy and speed of protein structure prediction.

STOP: If we do not know which template to use before we begin, how could we find a suitable template?

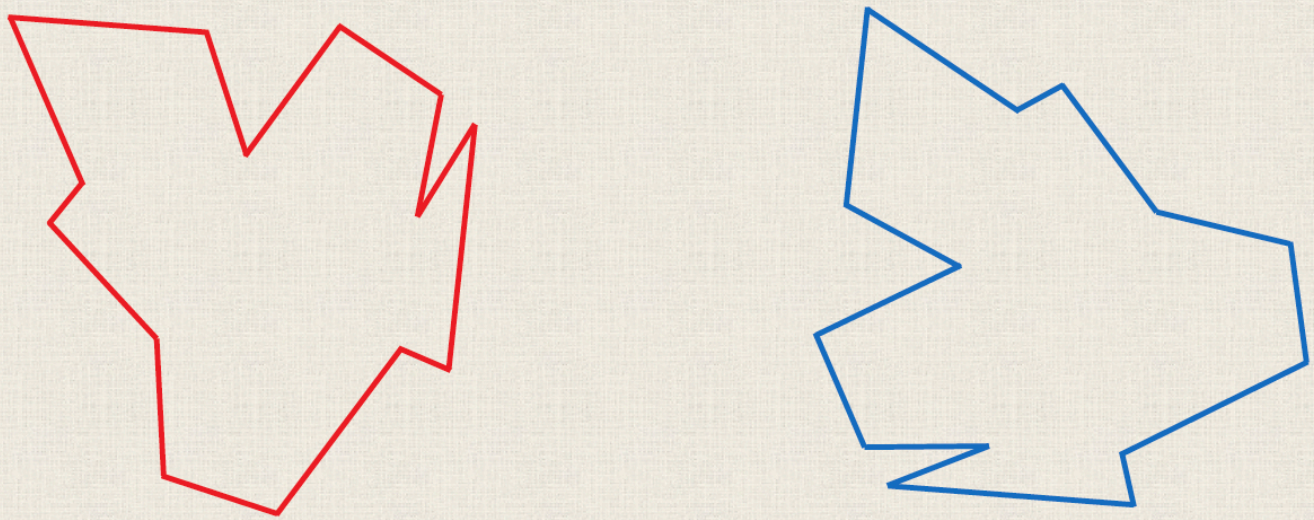
Homology modeling

This idea serves as the foundation of **homology modeling** for protein structure prediction (a.k.a. **comparative modeling**). By using the known protein structure of a homologous protein as a template, we can in theory improve both the accuracy and speed of protein structure prediction.

Answer: One natural thing to do would be to search for similar *sequences* for our novel protein in a database using an algorithm like BLAST.

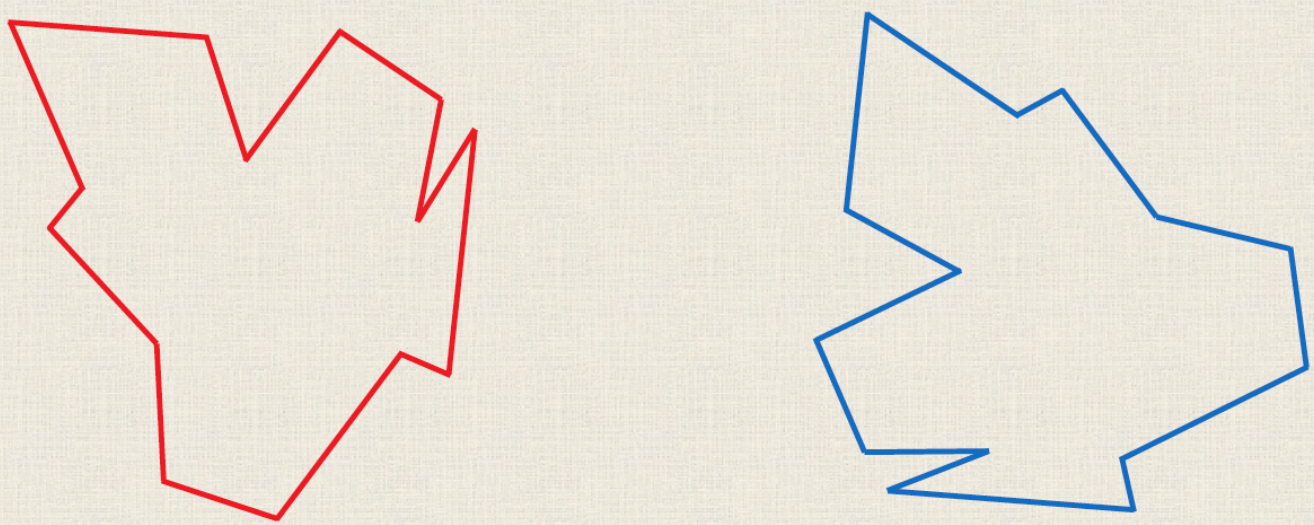
Comparing protein structures is analogous to comparing shapes

Goal: Develop a "distance function $d(S, T)$ that quantifies how different shapes S and T are.



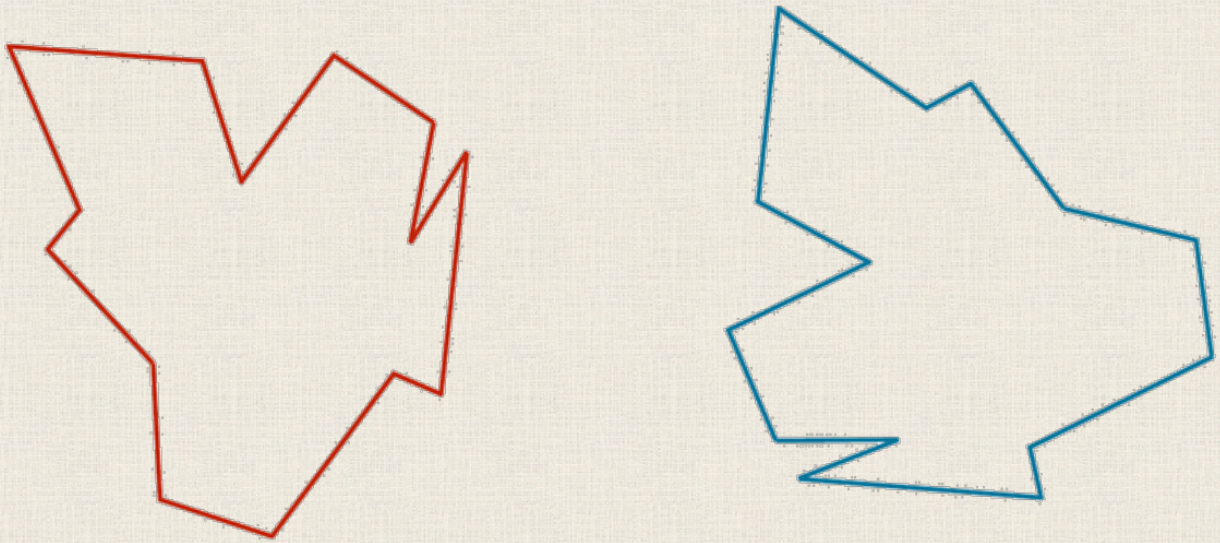
Comparing protein structures is analogous to comparing shapes

STOP: Consider the two shapes in the figure below. How similar are they?



Comparing protein structures is analogous to comparing shapes

Note: The two shapes are in fact the same! We can superimpose/flip/rotate the red shape to see why.



Determining Similarity of Aligned Shapes with RMSD

Sample n points along the boundary of S and T, converting S and T into *vectors* $s = (s_1, \ldots, s_n)$ and $t = (t_1, \ldots, t_n)$.

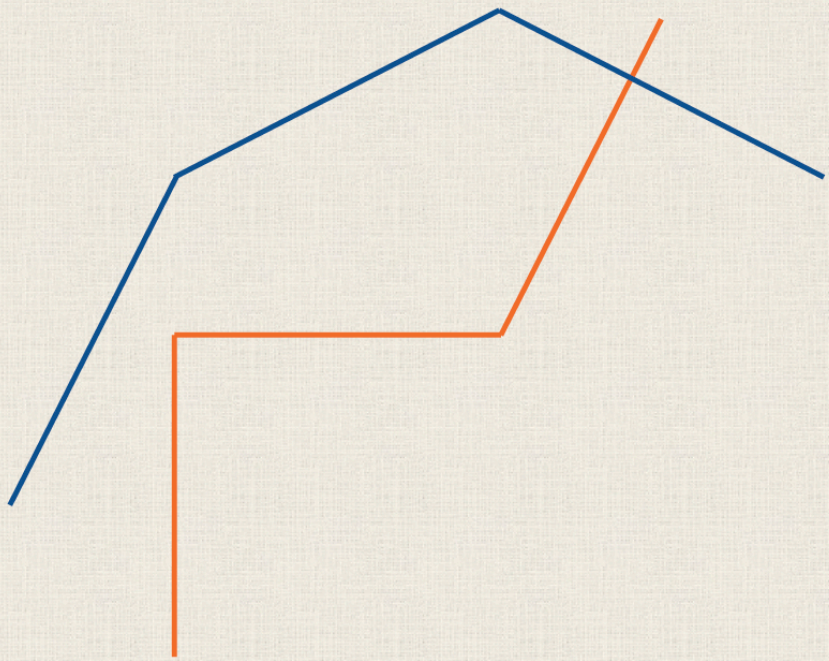
We then compute the **root mean square deviation** (**RMSD**) between the two shapes,

$$\mathrm{RMSD}(s, t) = \sqrt{\frac{1}{n} \cdot (d(s_1, t_1)^2 + d(s_2, t_2)^2 + \cdots + d(s_n, t_n)^2)}$$

the square root of the average squared distance between corresponding points in the vectors.

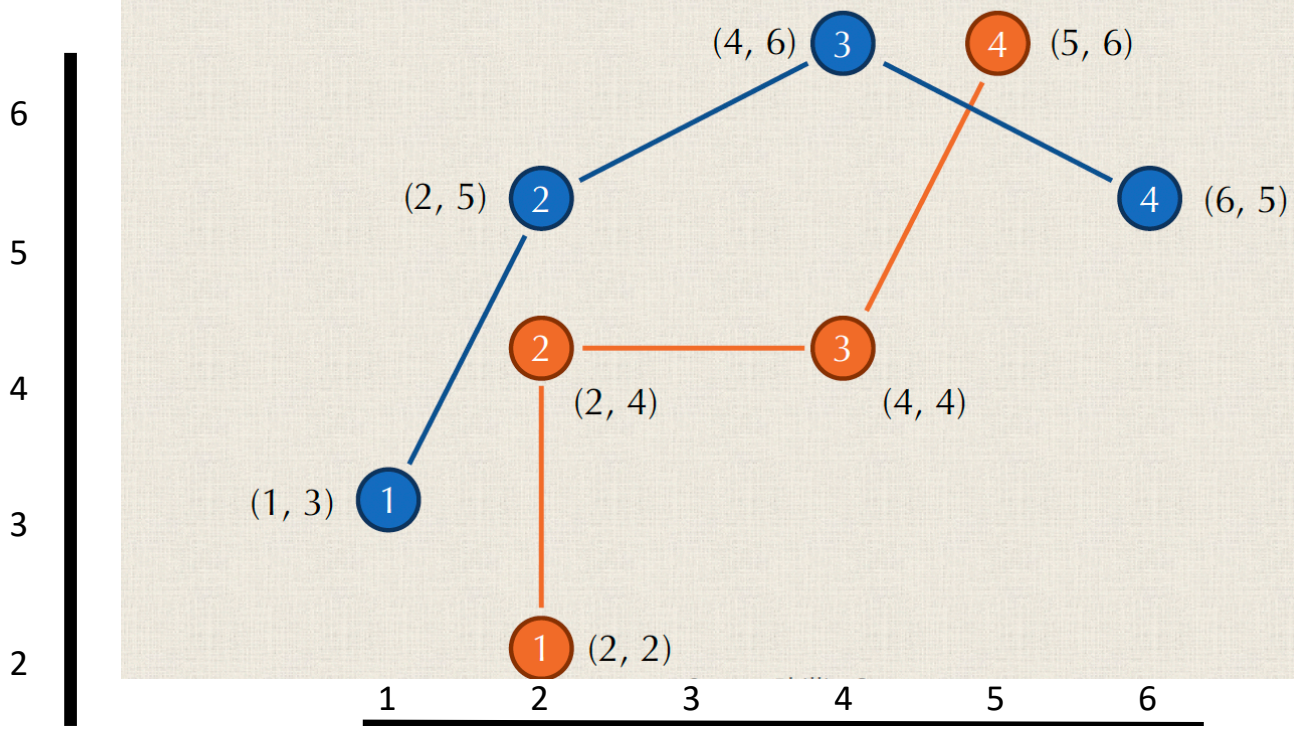
An example of computing RMSD

Consider the two shapes shown below.



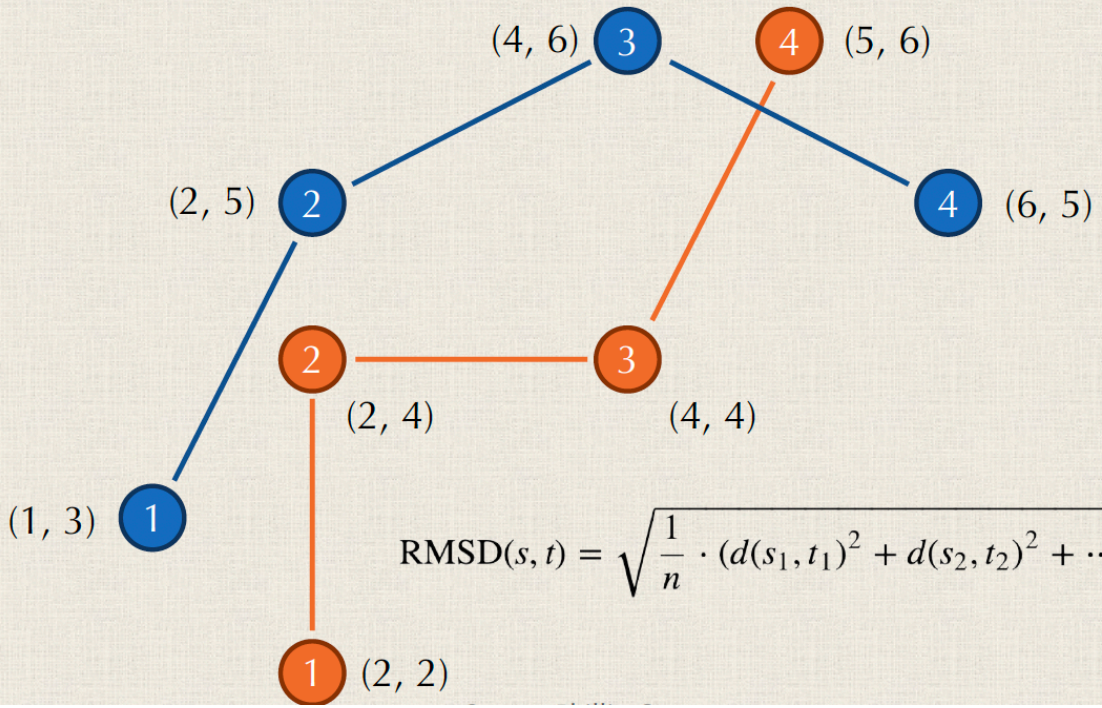
An example of computing RMSD

We vectorize by sampling $n = 4$ points from each.



An example of computing RMSD

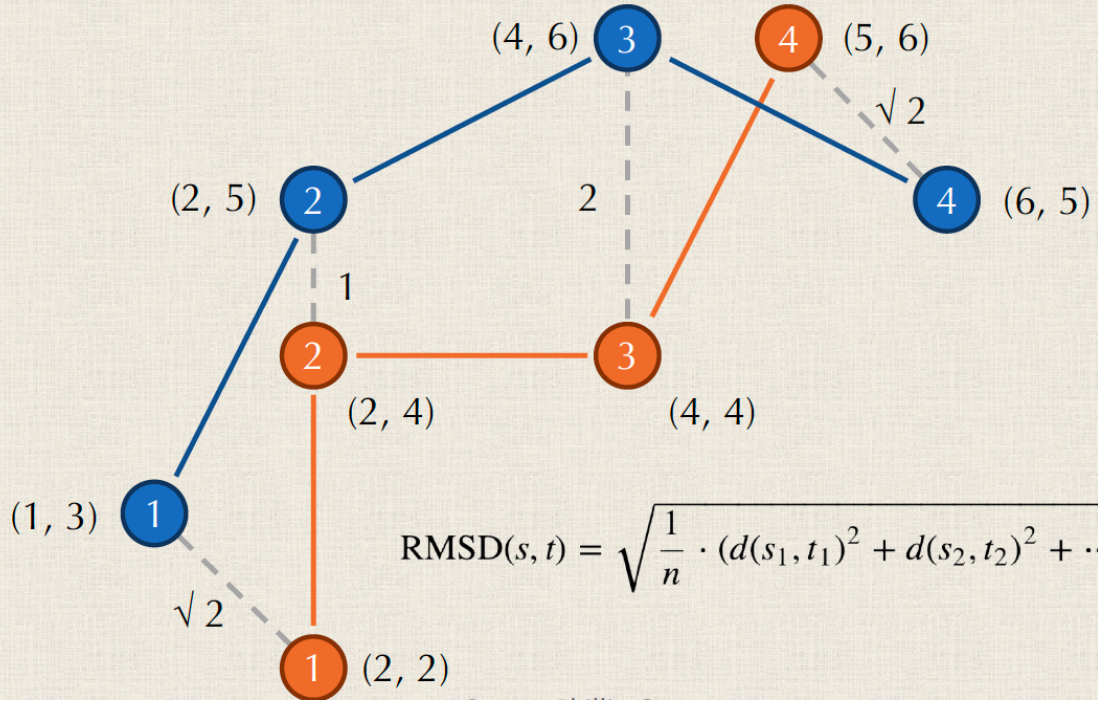
Exercise: Compute the RMSD for this example.



$$\mathrm{RMSD}(s,t) = \sqrt{\frac{1}{n} \cdot (d(s_1,t_1)^2 + d(s_2,t_2)^2 + \cdots + d(s_n,t_n)^2)}$$

An example of computing RMSD

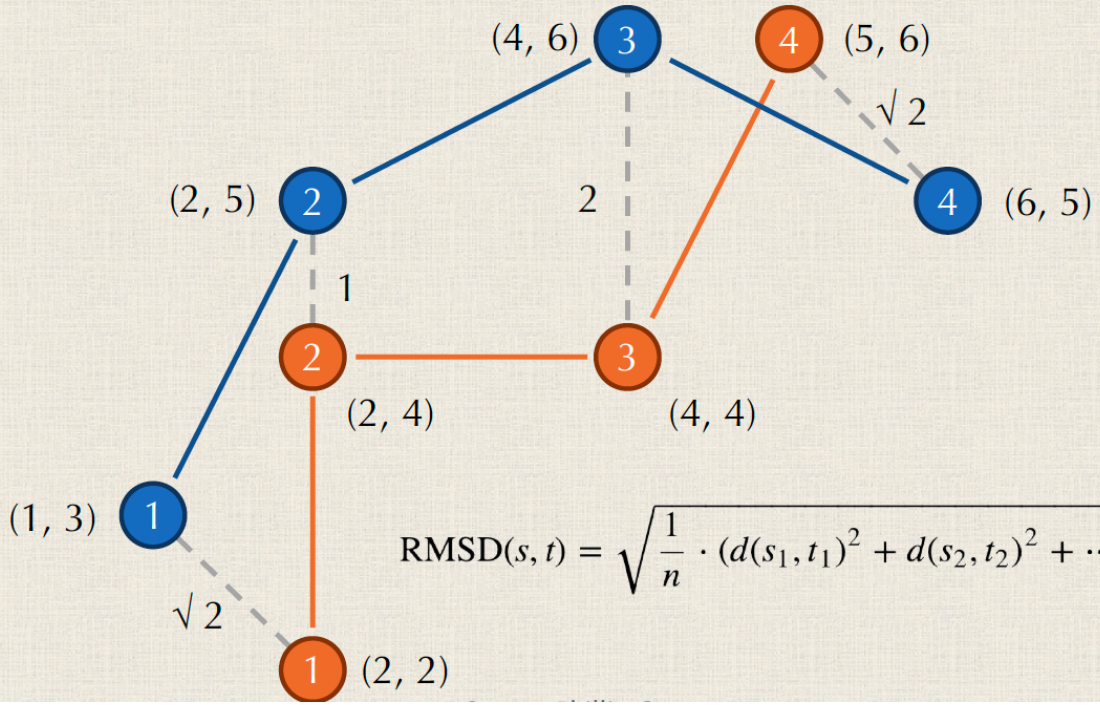
We first find the distances between corresponding points.



$$\mathrm{RMSD}(s,t) = \sqrt{\frac{1}{n} \cdot (d(s_1,t_1)^2 + d(s_2,t_2)^2 + \cdots + d(s_n,t_n)^2)}$$

An example of computing RMSD

RMSD = √((1/4) * (2 + 1 + 4 + 2)) = √(9/4) = 3/2.



$$\mathrm{RMSD}(s,t) = \sqrt{\frac{1}{n} \cdot (d(s_1,t_1)^2 + d(s_2,t_2)^2 + \cdots + d(s_n,t_n)^2)}$$