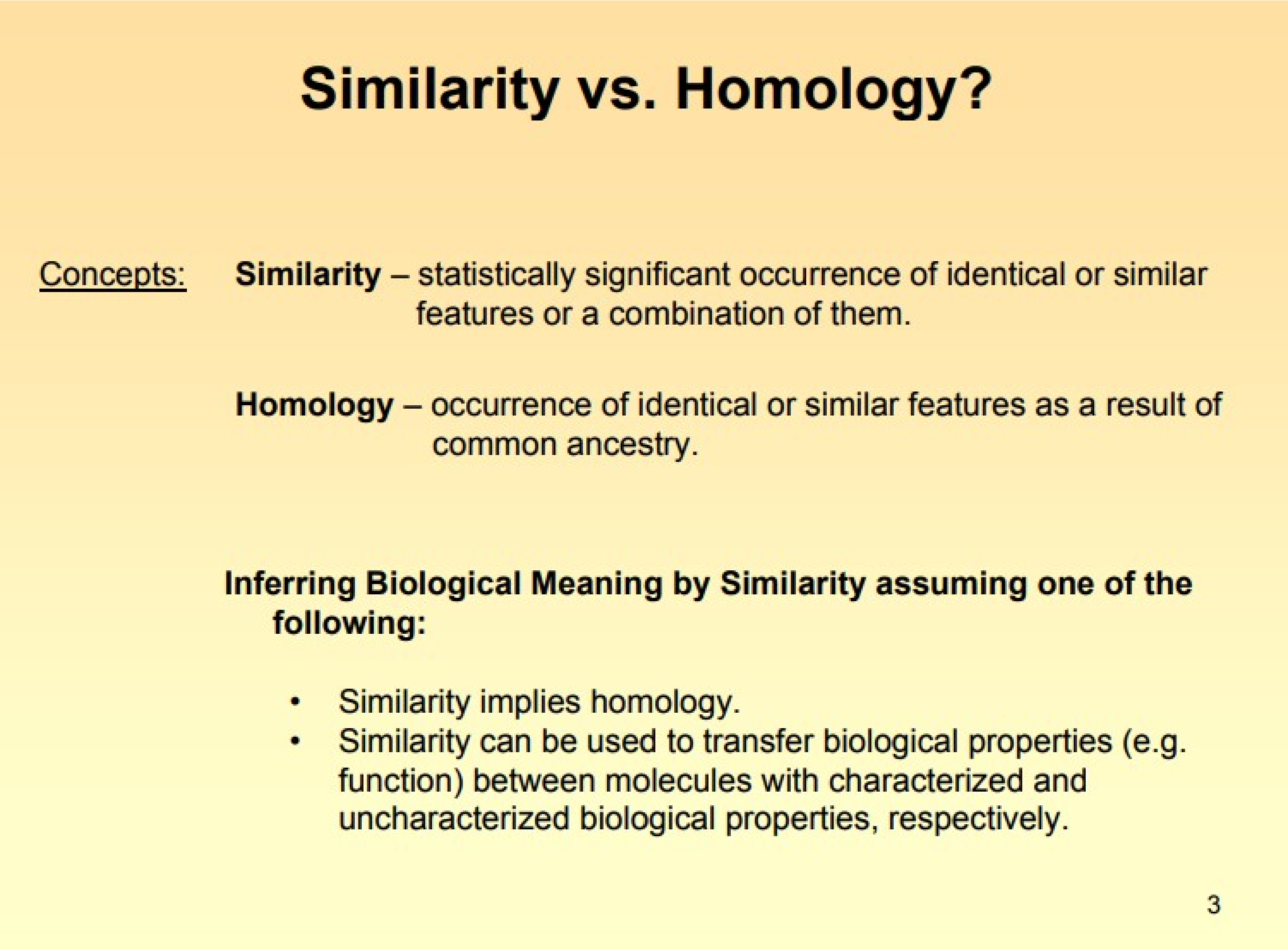
Alineamiento rigido de Proteinas



Similarity vs. Homology?

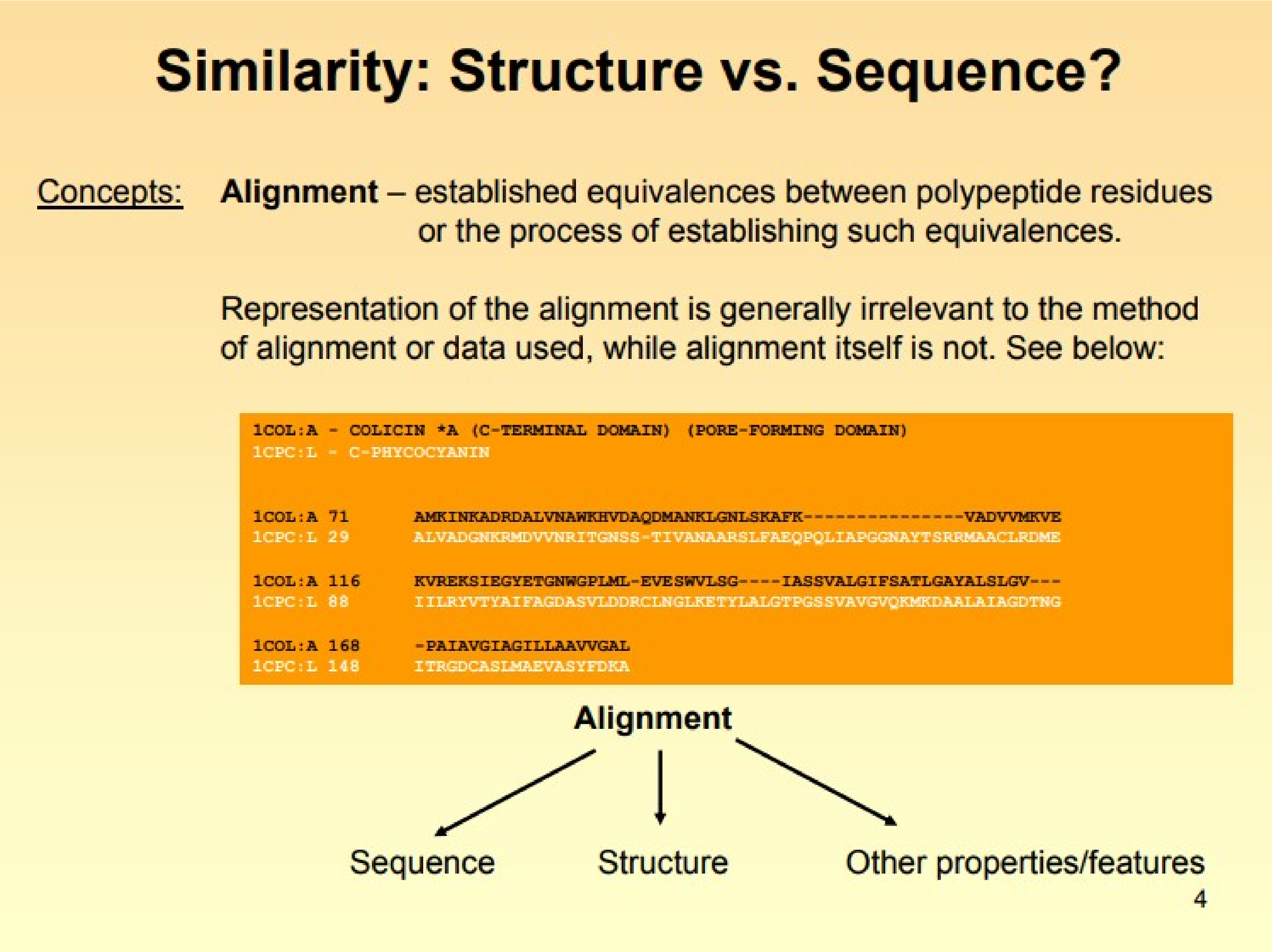
Concepts: **Similarity** – statistically significant occurrence of identical or similar features or a combination of them.

Homology – occurrence of identical or similar features as a result of common ancestry.

Inferring Biological Meaning by Similarity assuming one of the following:

- Similarity implies homology.
- Similarity can be used to transfer biological properties (e.g. function) between molecules with characterized and uncharacterized biological properties, respectively.

Alineamiento rigido de Proteinas



Similarity: Structure vs. Sequence?

<u>Concepts:</u> **Alignment** – established equivalences between polypeptide residues or the process of establishing such equivalences.

Representation of the alignment is generally irrelevant to the method of alignment or data used, while alignment itself is not. See below:

```
1COL:A - COLICIN *A (C-TERMINAL DOMAIN) (PORE-FORMING DOMAIN)
1CPC:L - C-PHYCOCYANIN


1COL:A 71      AMKINKADRDALVNAWKHVDAQDMANKLGNLSKAFK---------------VADVVMKVE
1CPC:L 29      ALVADGNKRMDVVNRITGNSS-TIVANAARSLFAEQPQLIAPGGNAYTSRRMAACLRDME

1COL:A 116     KVREKSIEGYETGNWGPLML-EVESWVLSG----IASSVALGIFSATLGAYALSLGV---
1CPC:L 88      IILRYVTYAIFAGDASVLDDRCLNGLKETYLALGTPGSSVAVGVQKMKDAALAIAGDTNG

1COL:A 168     -PAIAVGIAGILLAAVVGAL
1CPC:L 148     ITRGDCASLMAEVASYFDKA
```

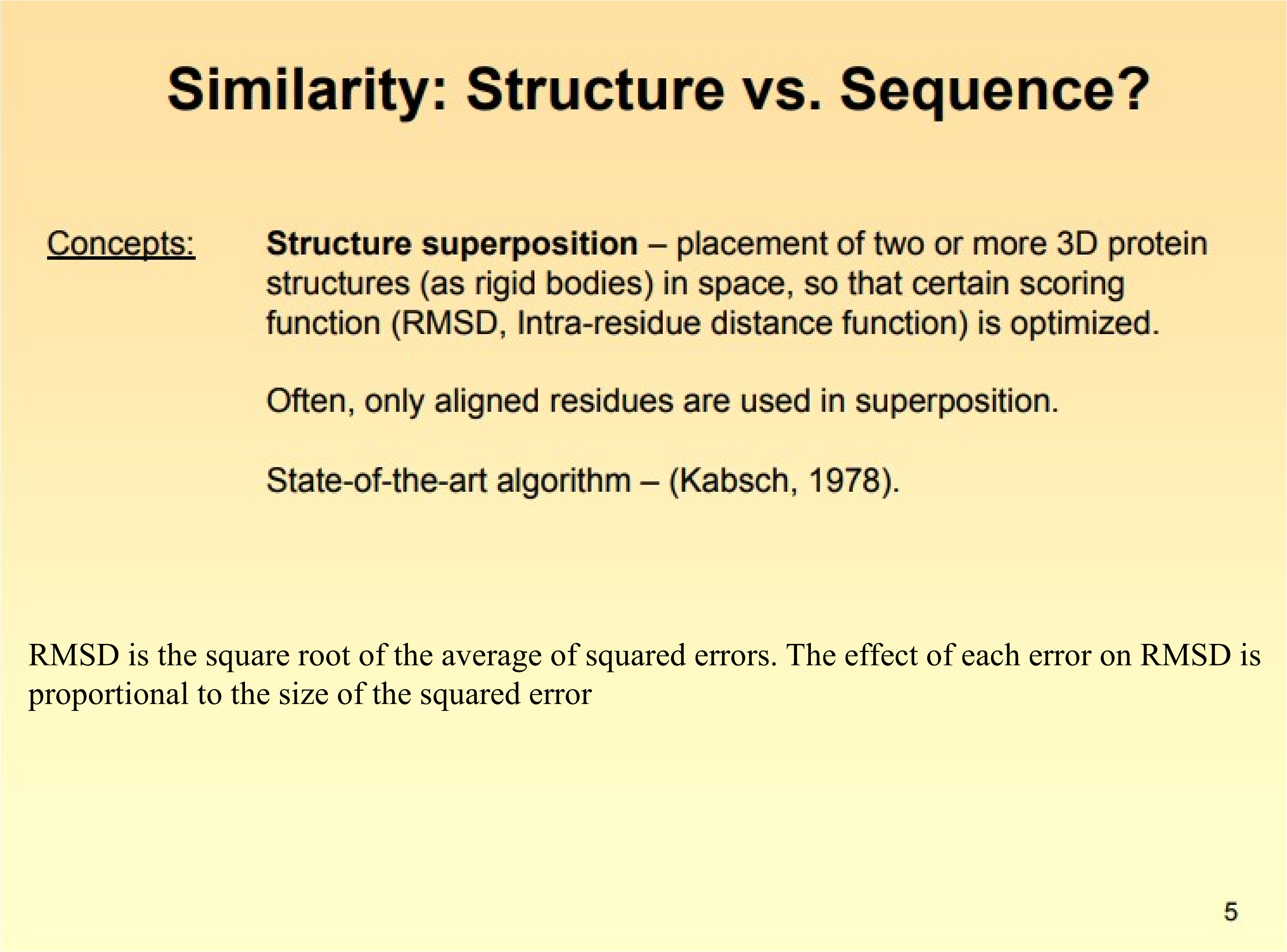
Alignment

Sequence

Structure

Other properties/features

Alineamiento rigido de Proteinas



Similarity: Structure vs. Sequence?

Concepts:

Structure superposition – placement of two or more 3D protein structures (as rigid bodies) in space, so that certain scoring function (RMSD, Intra-residue distance function) is optimized.

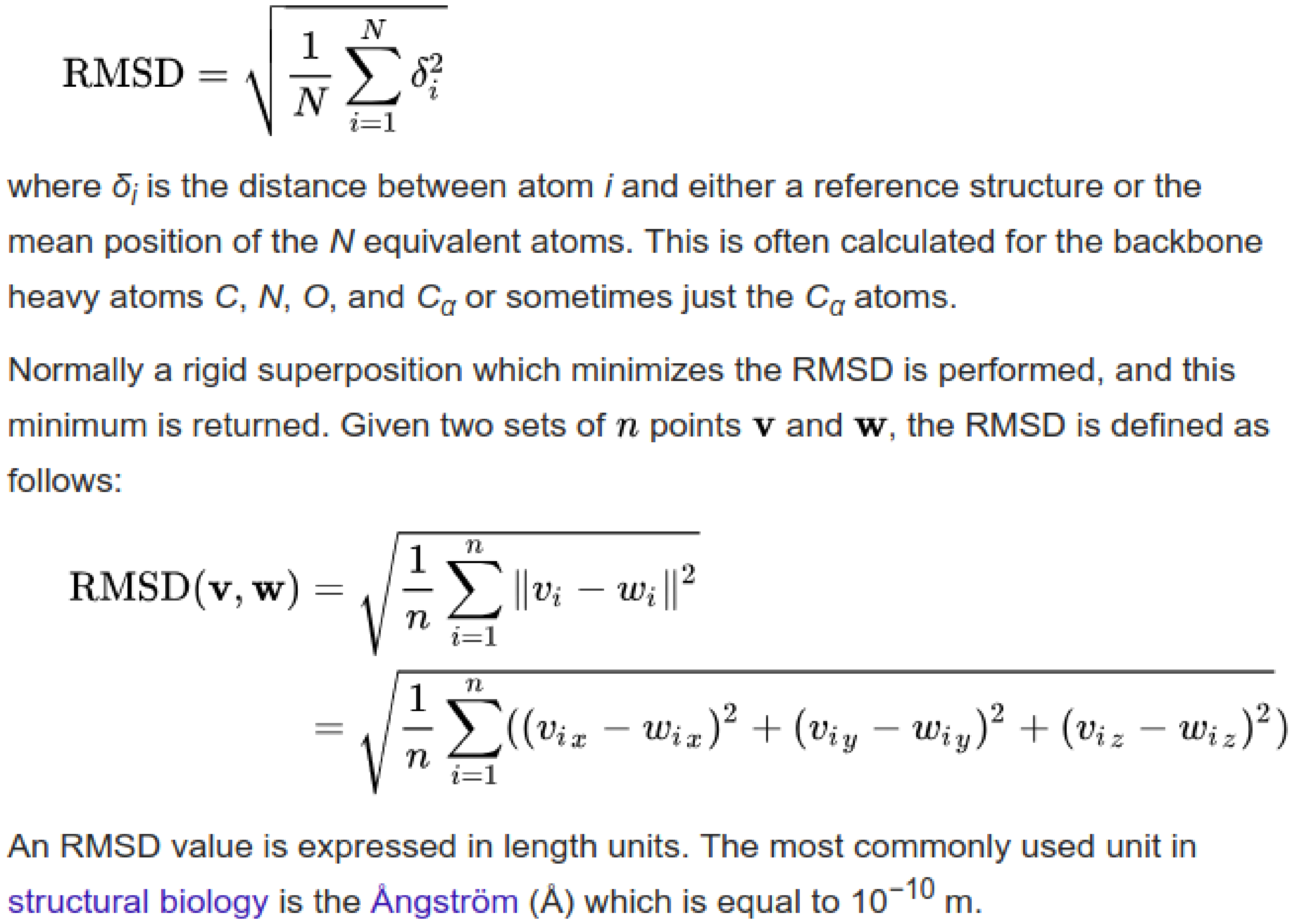
Often, only aligned residues are used in superposition.

State-of-the-art algorithm – (Kabsch, 1978).

RMSD is the square root of the average of squared errors. The effect of each error on RMSD is proportional to the size of the squared error

Alineamiento rigido de Proteinas

Root-Mean-Square Deviation of atomic positions



$$\mathrm{RMSD} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\delta_i^2}$$

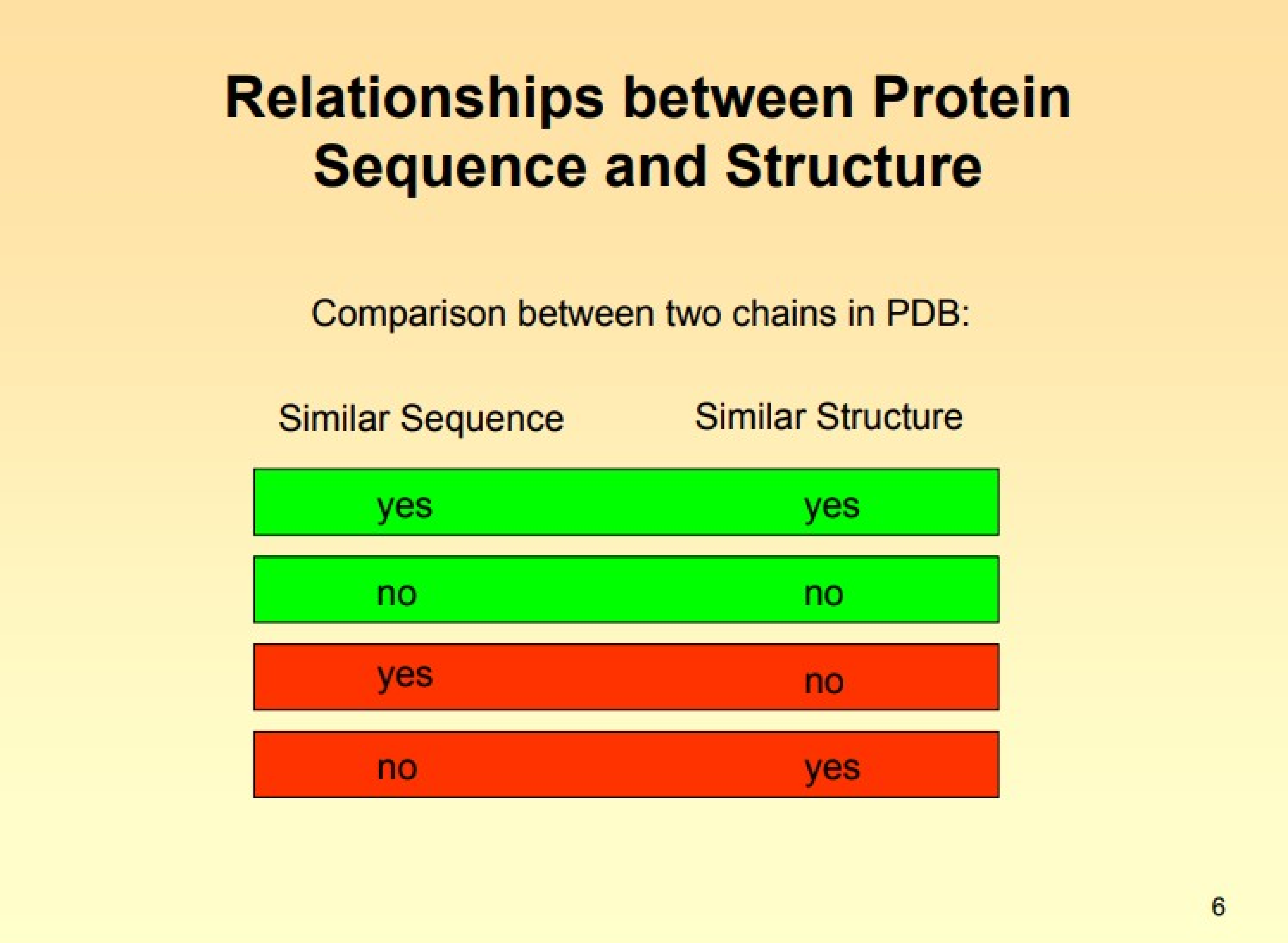
where δ_i is the distance between atom *i* and either a reference structure or the mean position of the *N* equivalent atoms. This is often calculated for the backbone heavy atoms *C*, *N*, *O*, and C_α or sometimes just the C_α atoms.

Normally a rigid superposition which minimizes the RMSD is performed, and this minimum is returned. Given two sets of n points $\mathbf{v}$ and $\mathbf{w}$, the RMSD is defined as follows:

$$\begin{aligned}\mathrm{RMSD}(\mathbf{v},\mathbf{w}) &= \sqrt{\frac{1}{n}\sum_{i=1}^{n}\|v_i - w_i\|^2}\\ &= \sqrt{\frac{1}{n}\sum_{i=1}^{n}((v_{i\,x} - w_{i\,x})^2 + (v_{i\,y} - w_{i\,y})^2 + (v_{i\,z} - w_{i\,z})^2)}\end{aligned}$$

An RMSD value is expressed in length units. The most commonly used unit in structural biology is the Ångström (Å) which is equal to 10^{-10} m.

Alineamiento rigido de Proteinas

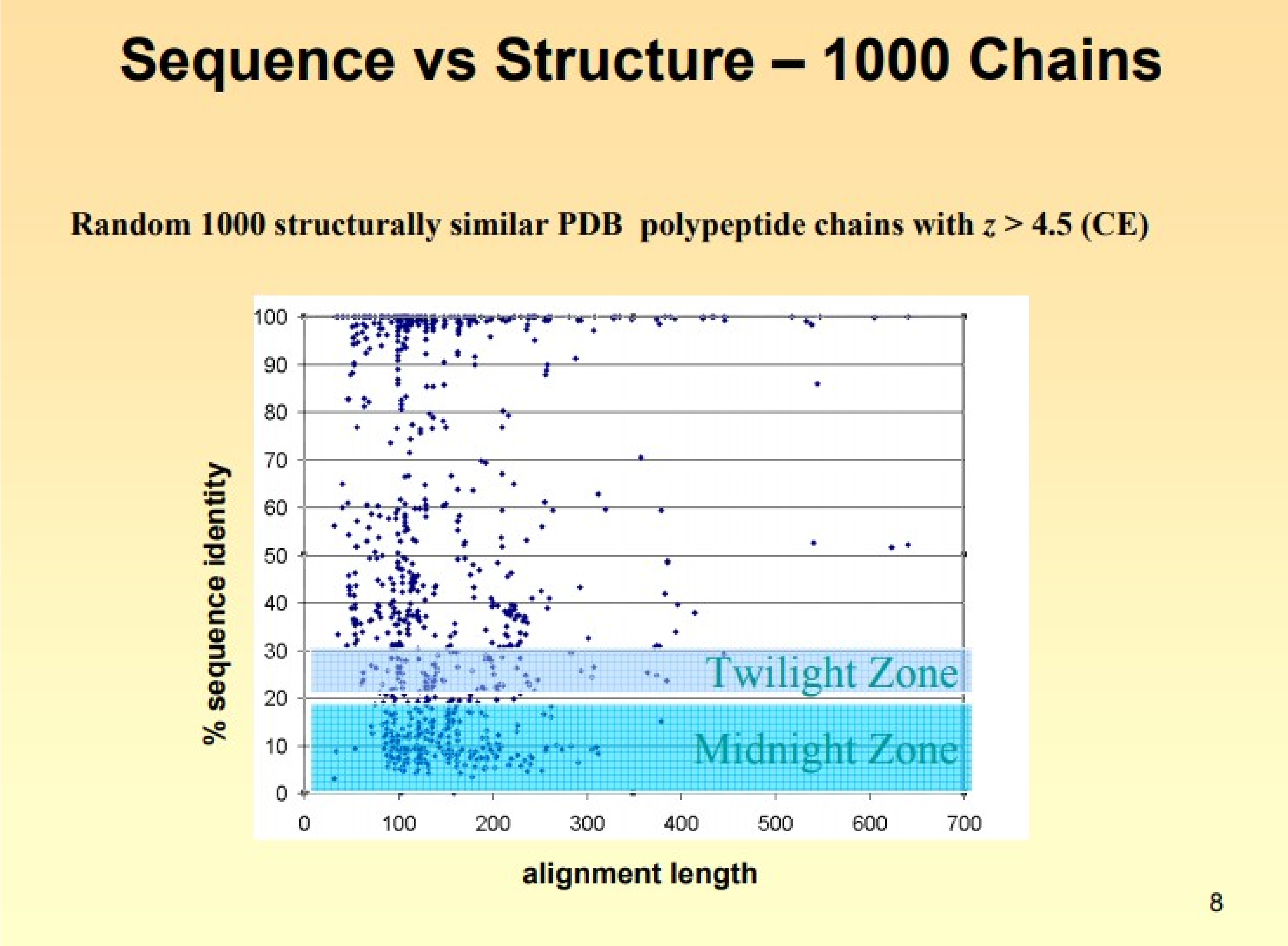


Relationships between Protein Sequence and Structure

Comparison between two chains in PDB:

Similar Sequence	Similar Structure
yes	yes
no	no
yes	no
no	yes

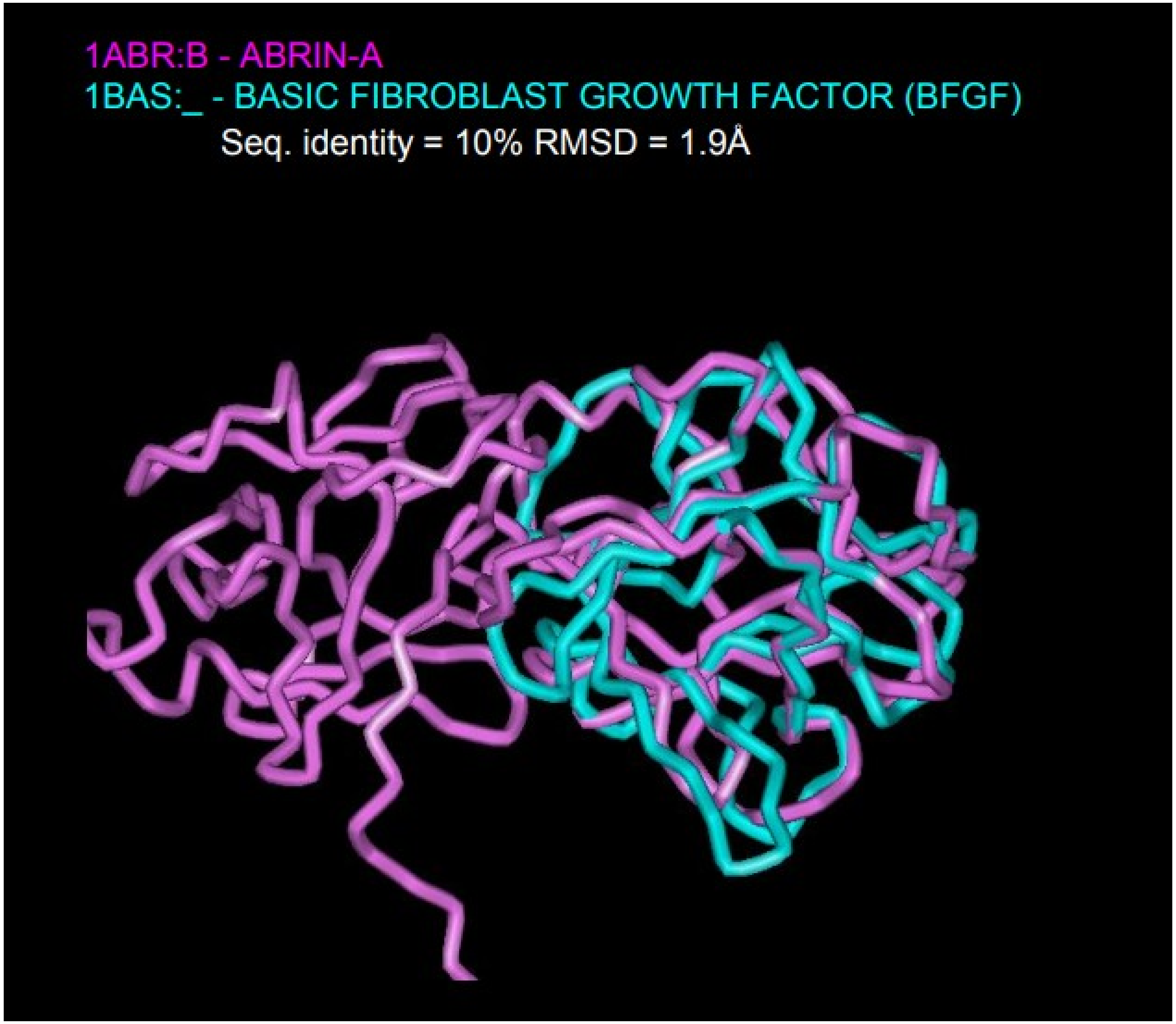
Alineamiento rigido de Proteinas



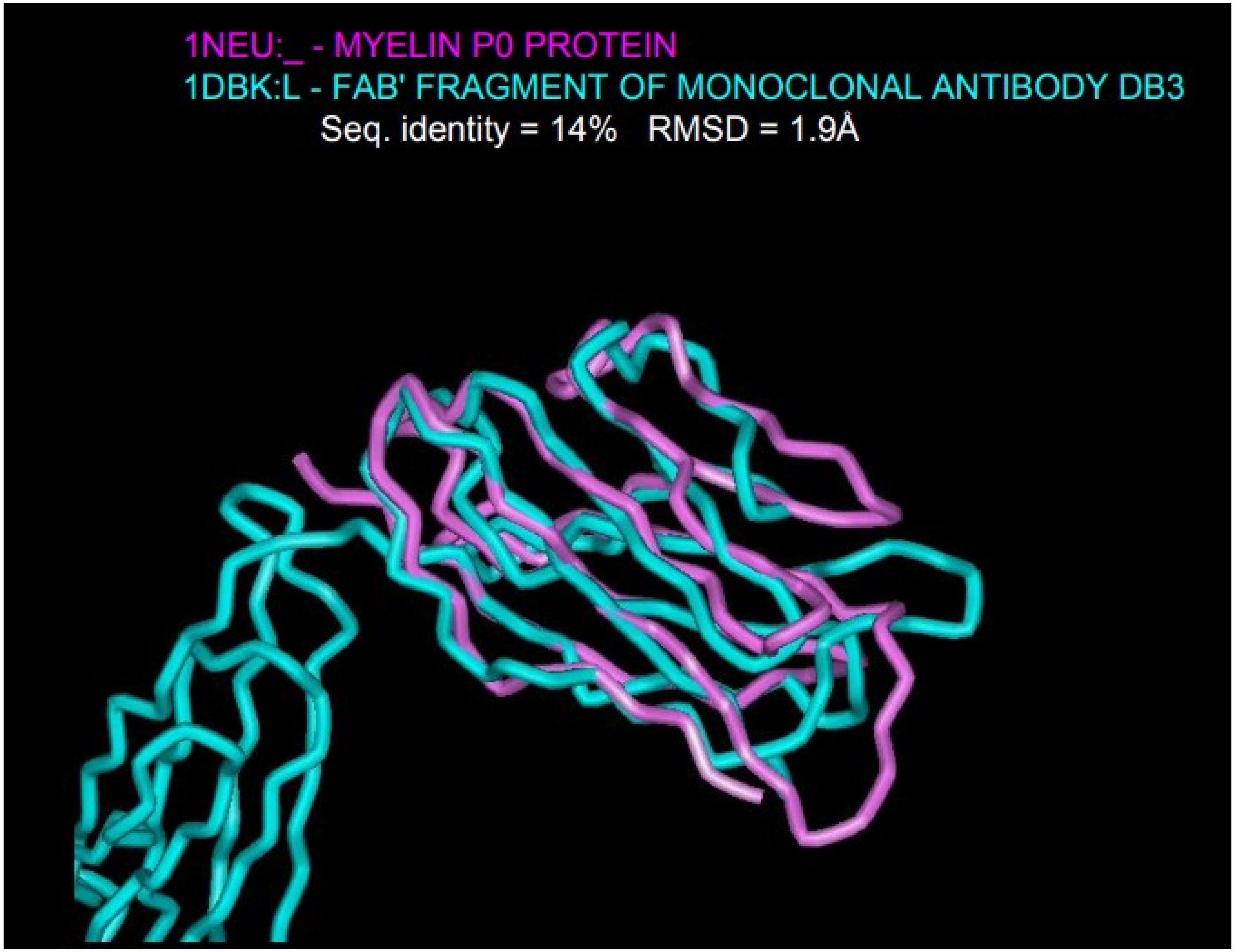
Sequence vs Structure – 1000 Chains

Random 1000 structurally similar PDB polypeptide chains with *z* > 4.5 (CE)

Alineamiento rigido de Proteinas

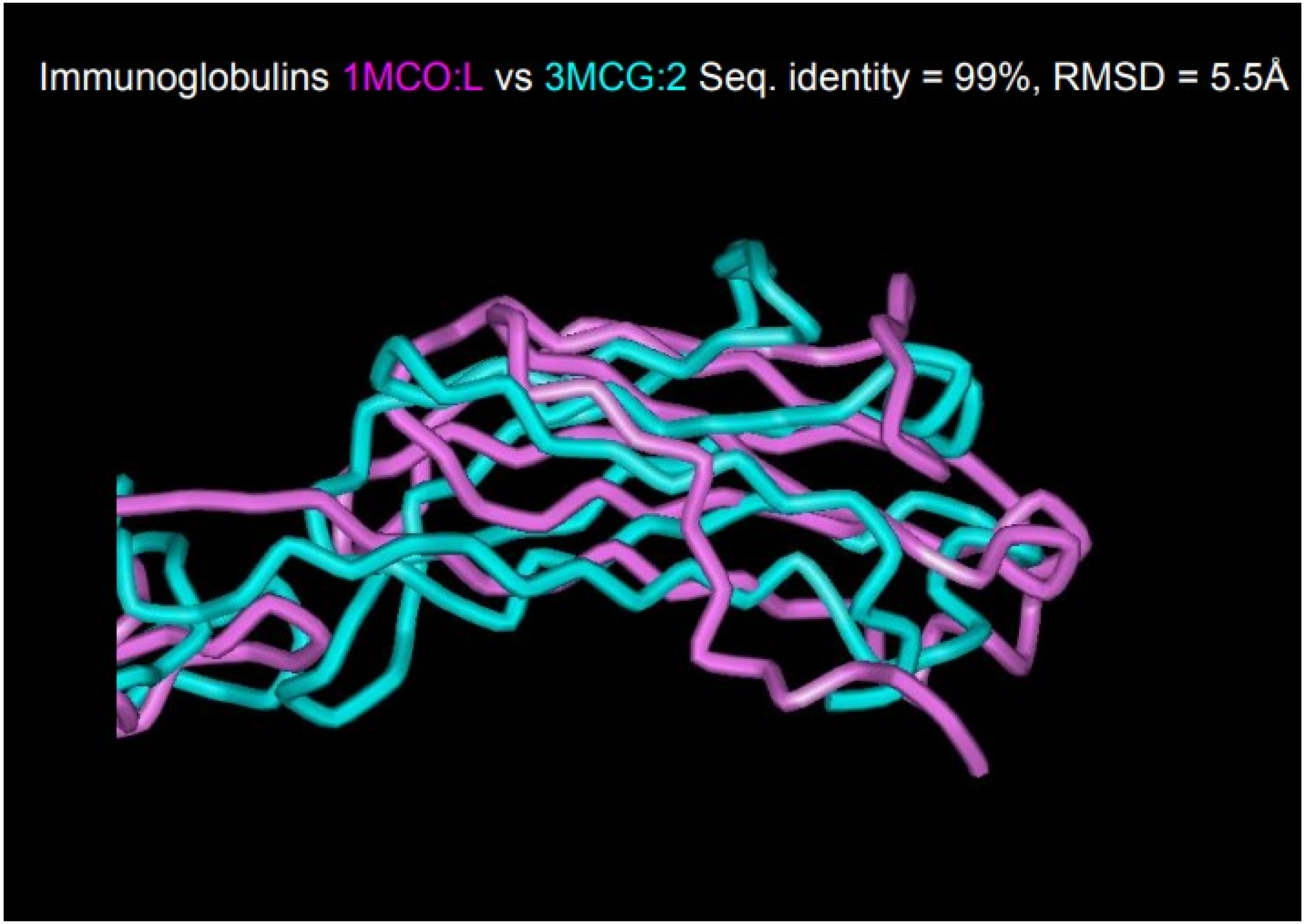

1ABR:B - ABRIN-A
1BAS:_ - BASIC FIBROBLAST GROWTH FACTOR (BFGF)
Seq. identity = 10% RMSD = 1.9Å

Alineamiento rigido de Proteinas



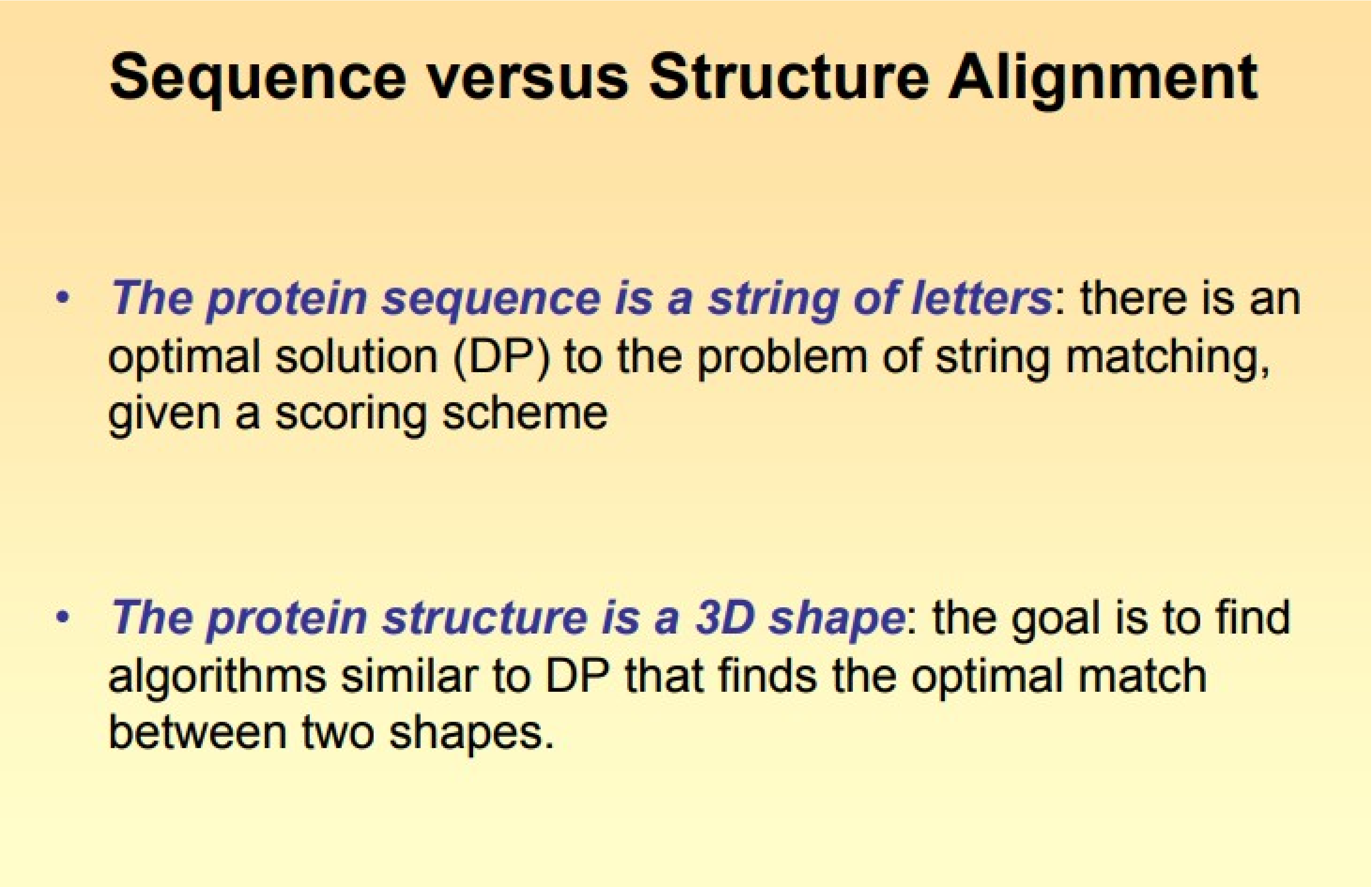
1NEU:_ - MYELIN P0 PROTEIN
1DBK:L - FAB' FRAGMENT OF MONOCLONAL ANTIBODY DB3
Seq. identity = 14%   RMSD = 1.9Å

Alineamiento rigido de Proteinas



Immunoglobulins 1MCO:L vs 3MCG:2 Seq. identity = 99%, RMSD = 5.5Å

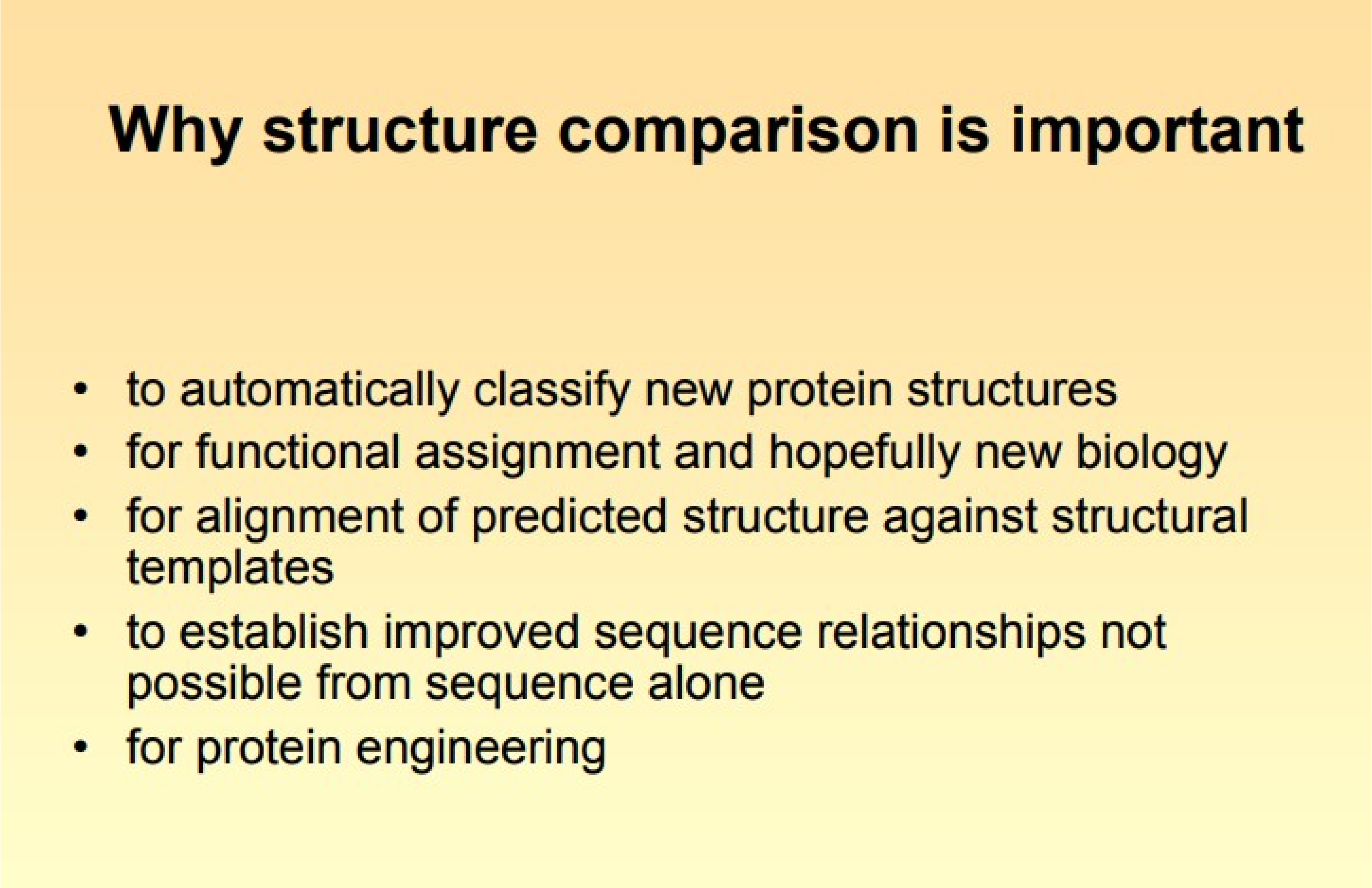
Alineamiento rigido de Proteinas



Sequence versus Structure Alignment

- ***The protein sequence is a string of letters***: there is an optimal solution (DP) to the problem of string matching, given a scoring scheme
- ***The protein structure is a 3D shape***: the goal is to find algorithms similar to DP that finds the optimal match between two shapes.

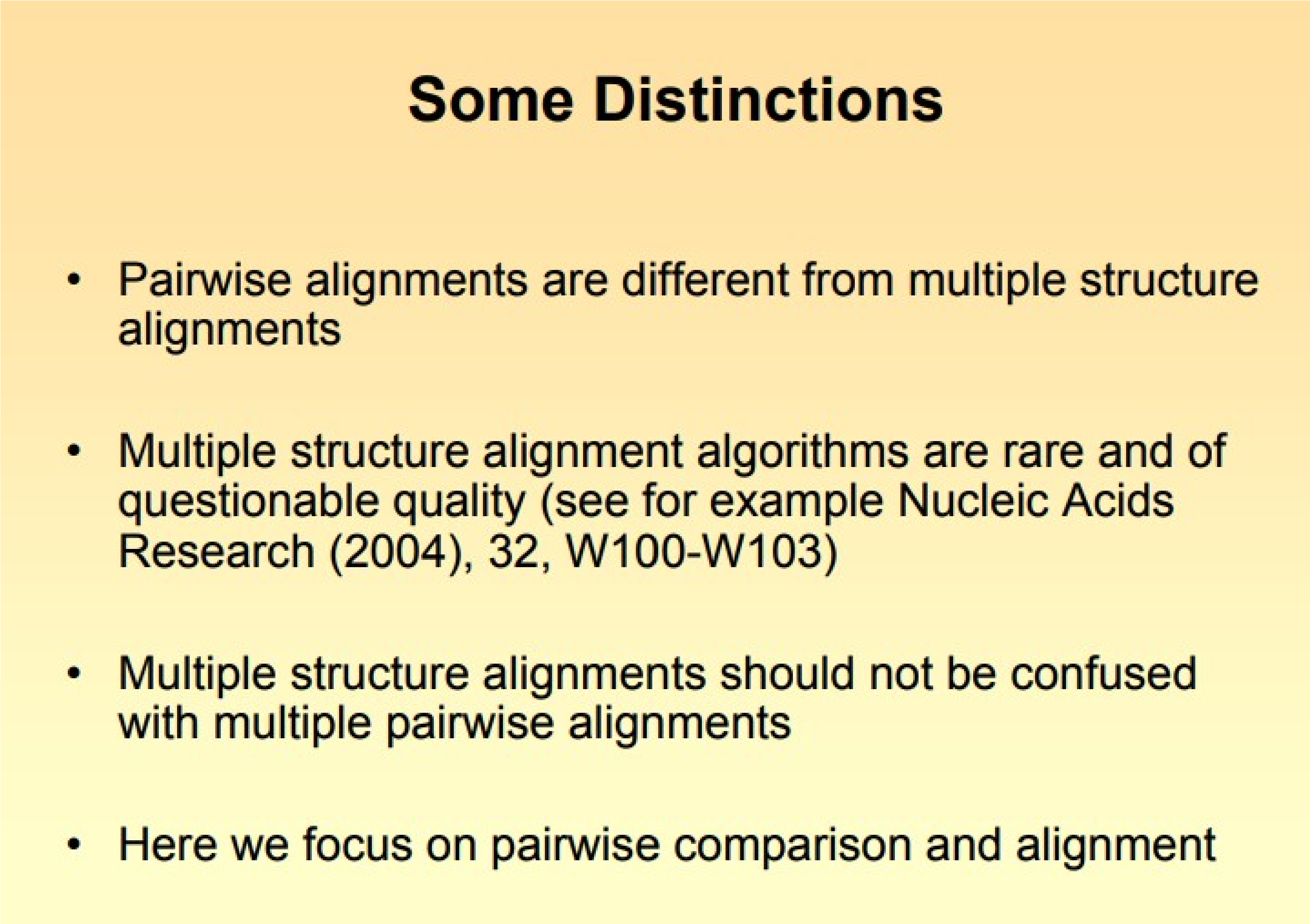
Alineamiento rigido de Proteinas



Why structure comparison is important

- to automatically classify new protein structures
- for functional assignment and hopefully new biology
- for alignment of predicted structure against structural templates
- to establish improved sequence relationships not possible from sequence alone
- for protein engineering

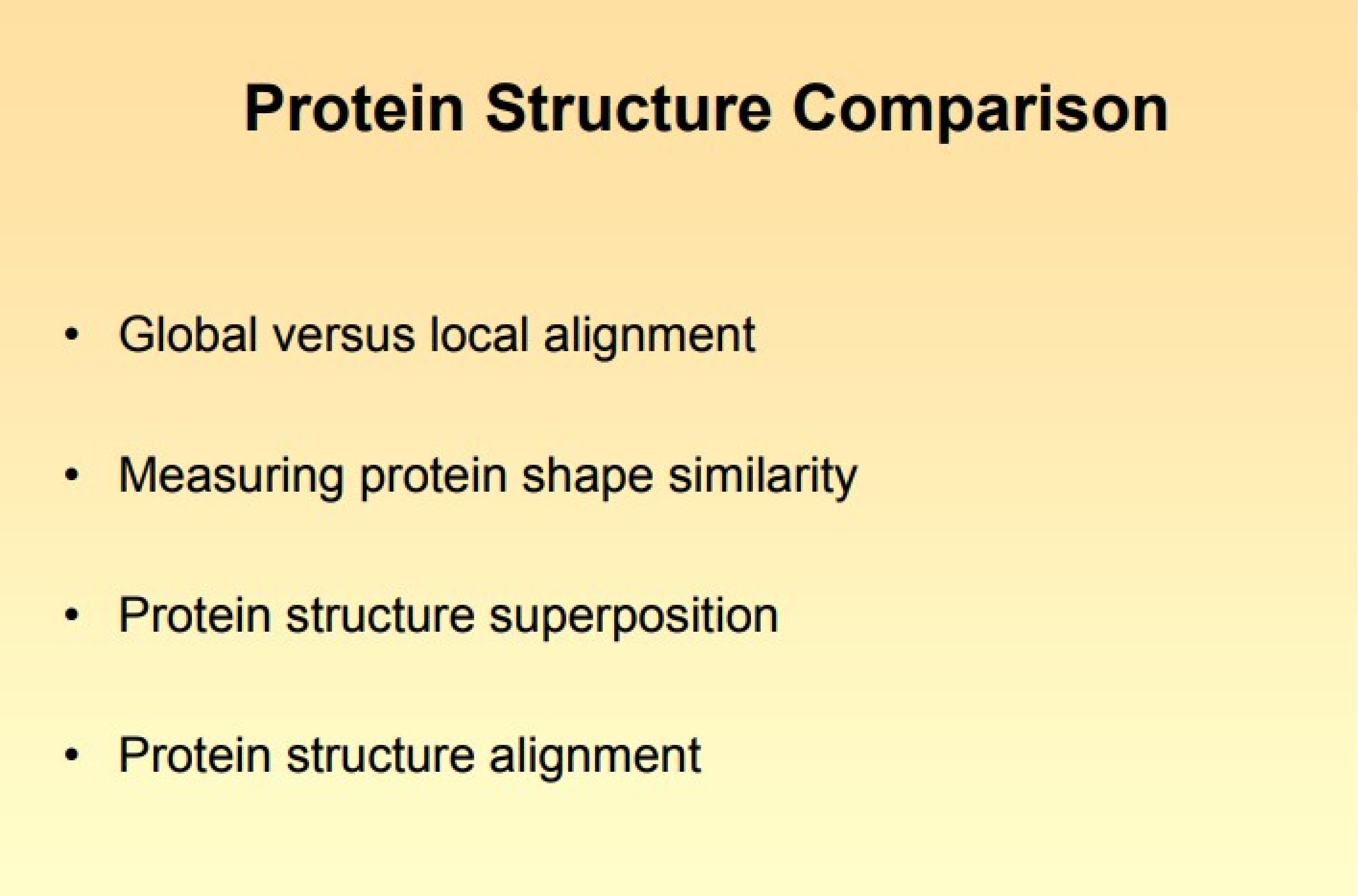
Alineamiento rigido de Proteinas



Some Distinctions

- Pairwise alignments are different from multiple structure alignments
- Multiple structure alignment algorithms are rare and of questionable quality (see for example Nucleic Acids Research (2004), 32, W100-W103)
- Multiple structure alignments should not be confused with multiple pairwise alignments
- Here we focus on pairwise comparison and alignment

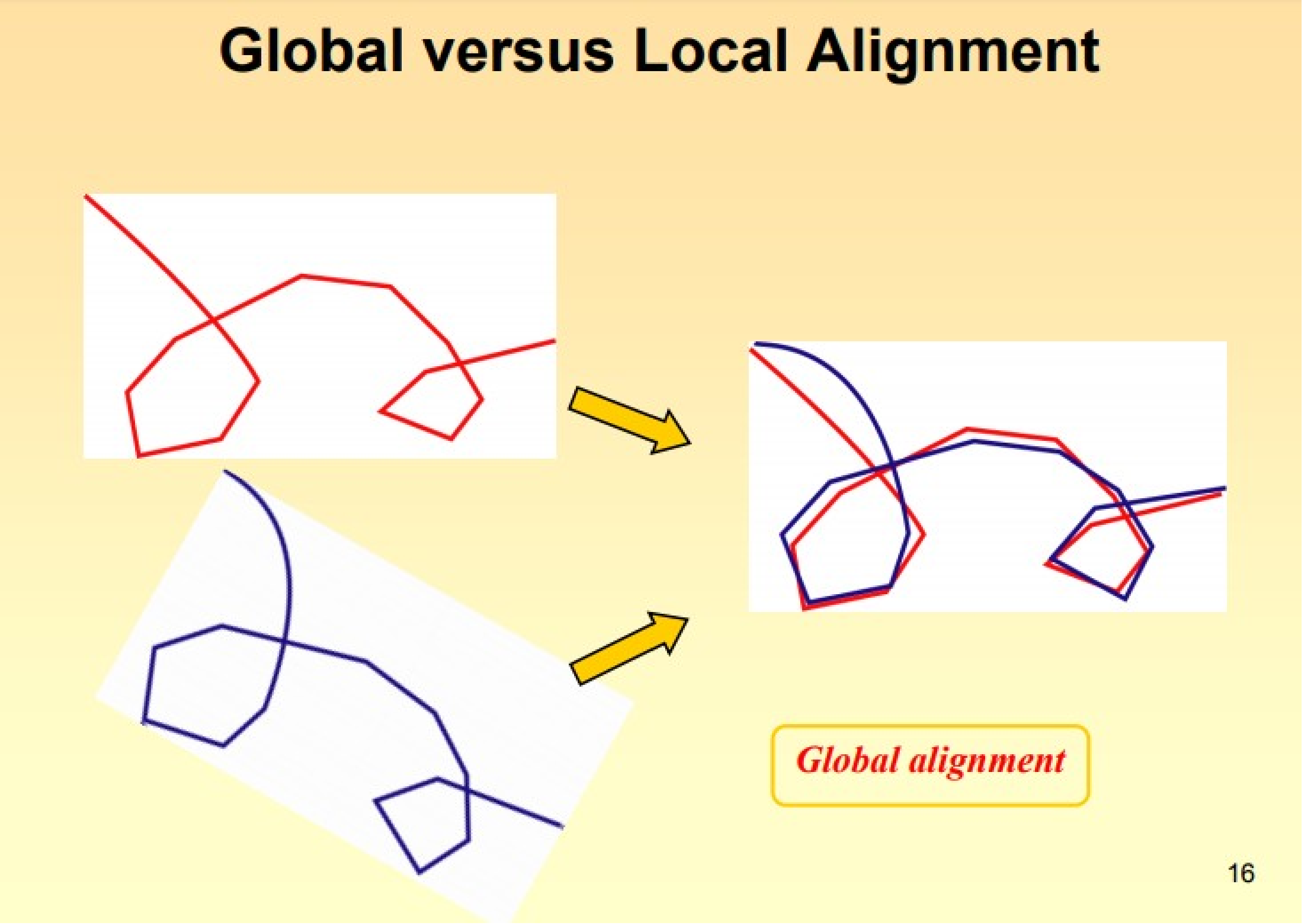
Alineamiento rigido de Proteinas



Protein Structure Comparison

- Global versus local alignment
- Measuring protein shape similarity
- Protein structure superposition
- Protein structure alignment

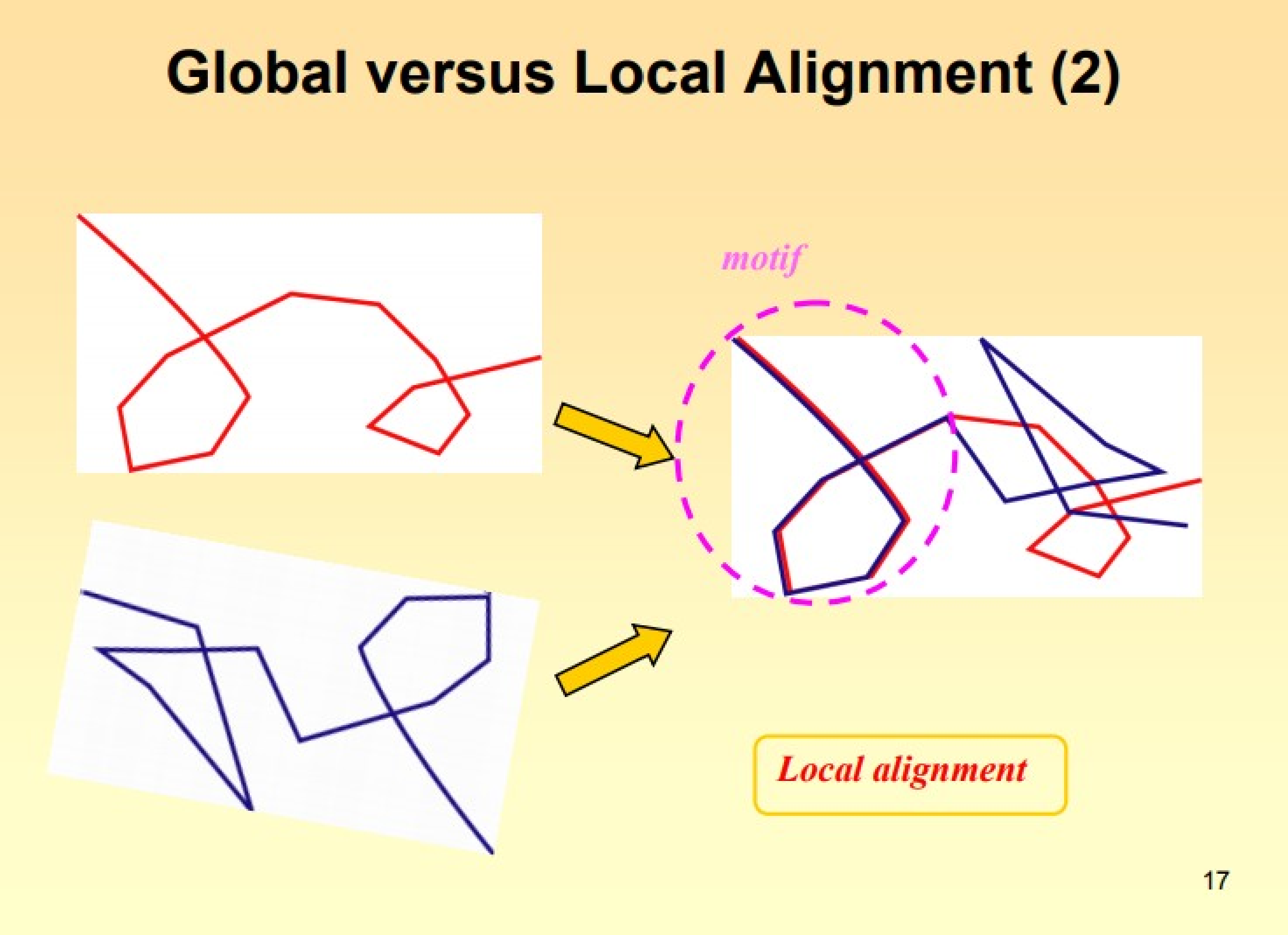
Alineamiento rigido de Proteinas



Global versus Local Alignment

Global alignment

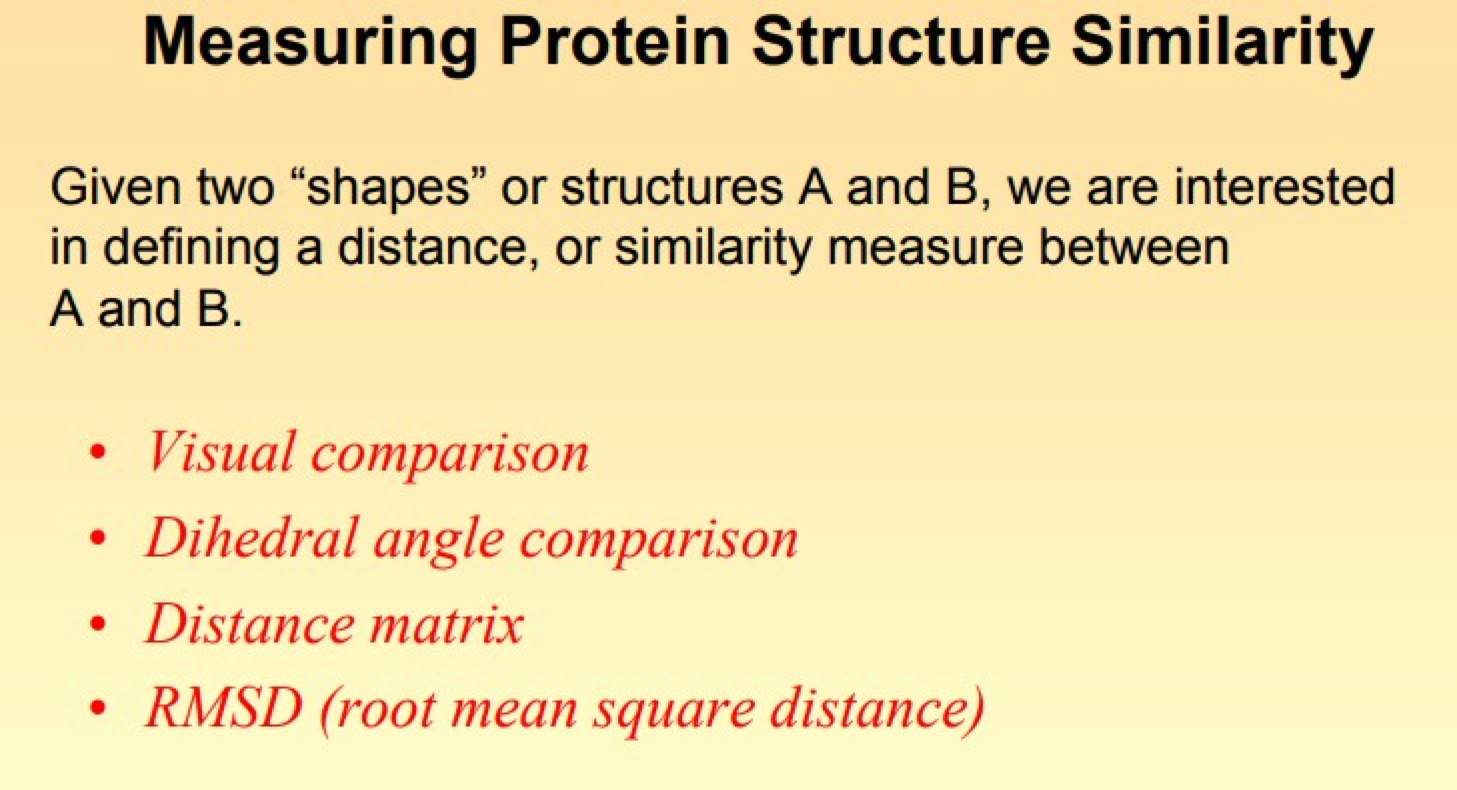
Alineamiento rigido de Proteinas



Global versus Local Alignment (2)

Local alignment

Alineamiento rigido de Proteinas

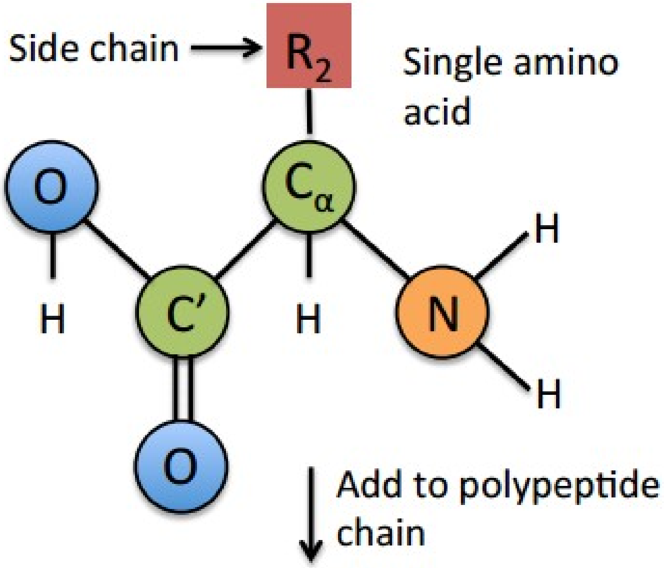


Measuring Protein Structure Similarity

Given two “shapes” or structures A and B, we are interested in defining a distance, or similarity measure between A and B.

- *Visual comparison*
- *Dihedral angle comparison*
- *Distance matrix*
- *RMSD (root mean square distance)*

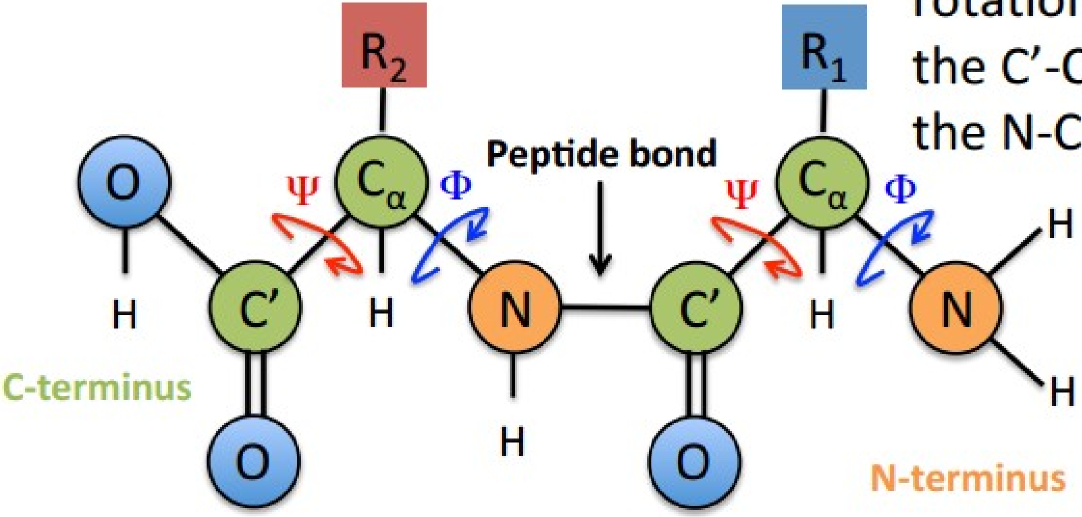
Amino acids and the peptide bond



Amino acids are connected by **peptide bonds**

Polypeptide chains are extended from the N-terminus to the C-terminus

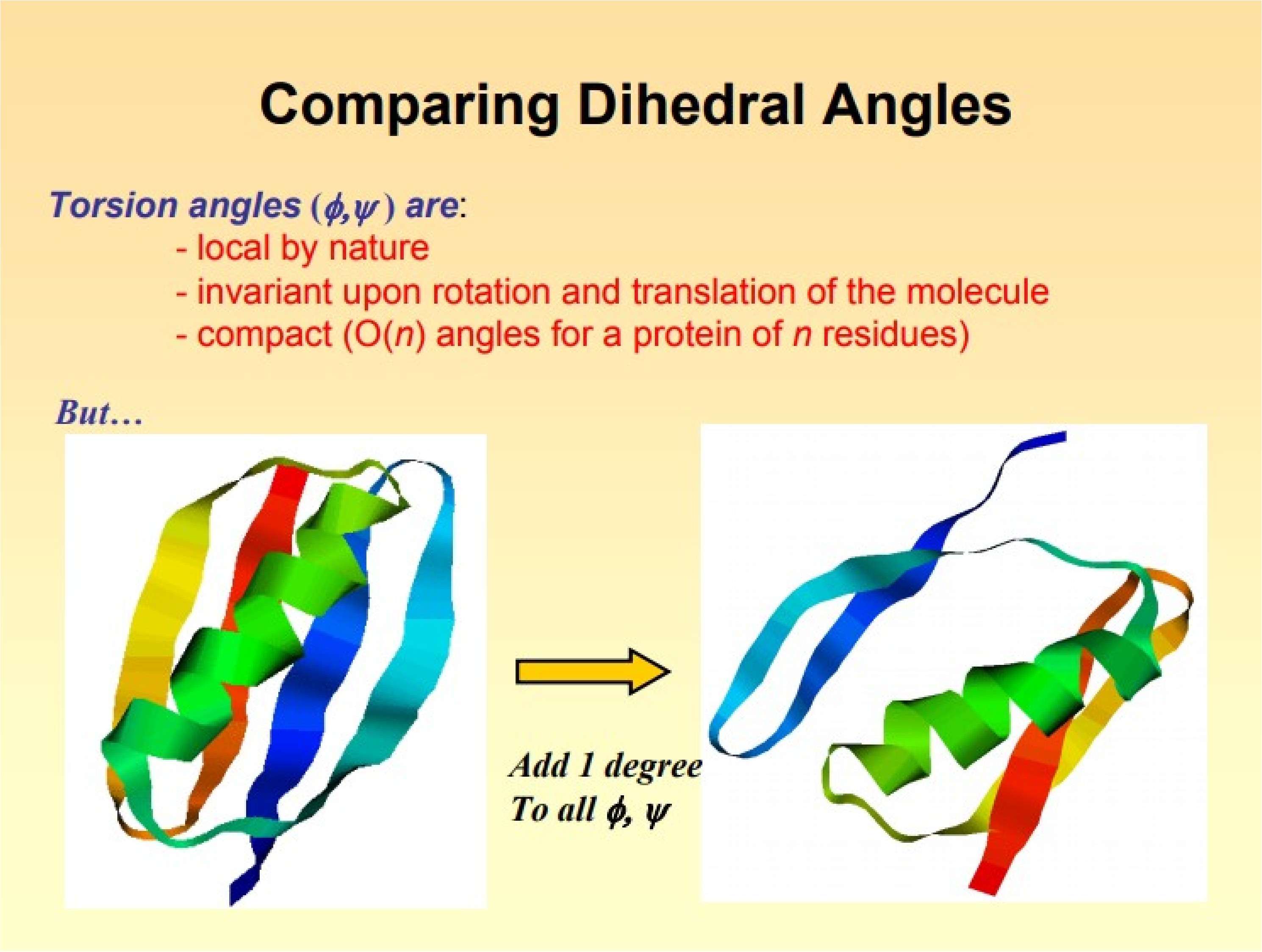
Due to resonance, the peptide bond is rigid, meaning that rotation can only occur along the C′-C_α and C_α-N bonds, not the N-C′ bonds



C′-C_α : Ψ (psi)

C_α-N : Φ (phi)

Alineamiento rigido de Proteinas



Comparing Dihedral Angles

Torsion angles (ϕ,ψ) are:

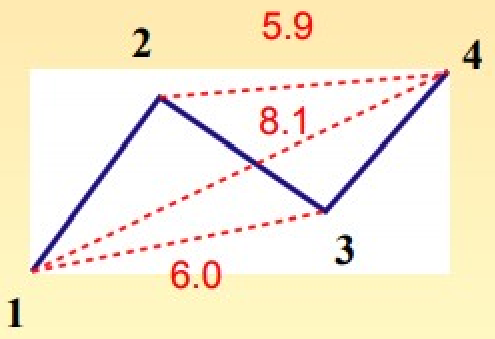
- local by nature
- invariant upon rotation and translation of the molecule
- compact (O(*n*) angles for a protein of *n* residues)

But…

Add 1 degree
To all ϕ, ψ

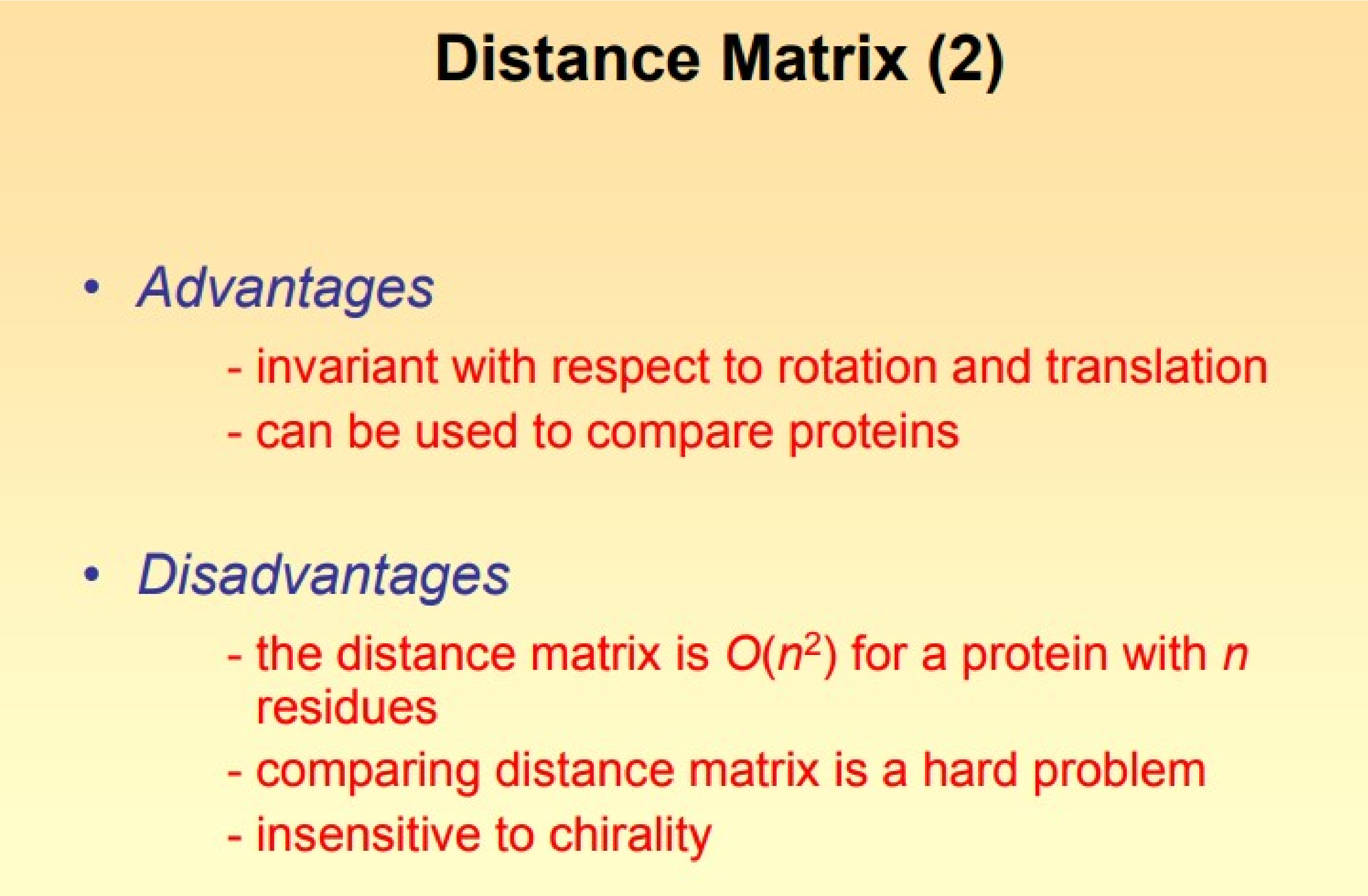
Alineamiento rigido de Proteinas

Por Matriz de Distancias



	1	2	3	4
1	0	3.8	6.0	8.1
2	3.8	0	3.8	5.9
3	6.0	3.8	0	3.8
4	8.1	5.9	3.8	0

Alineamiento rigido de Proteinas

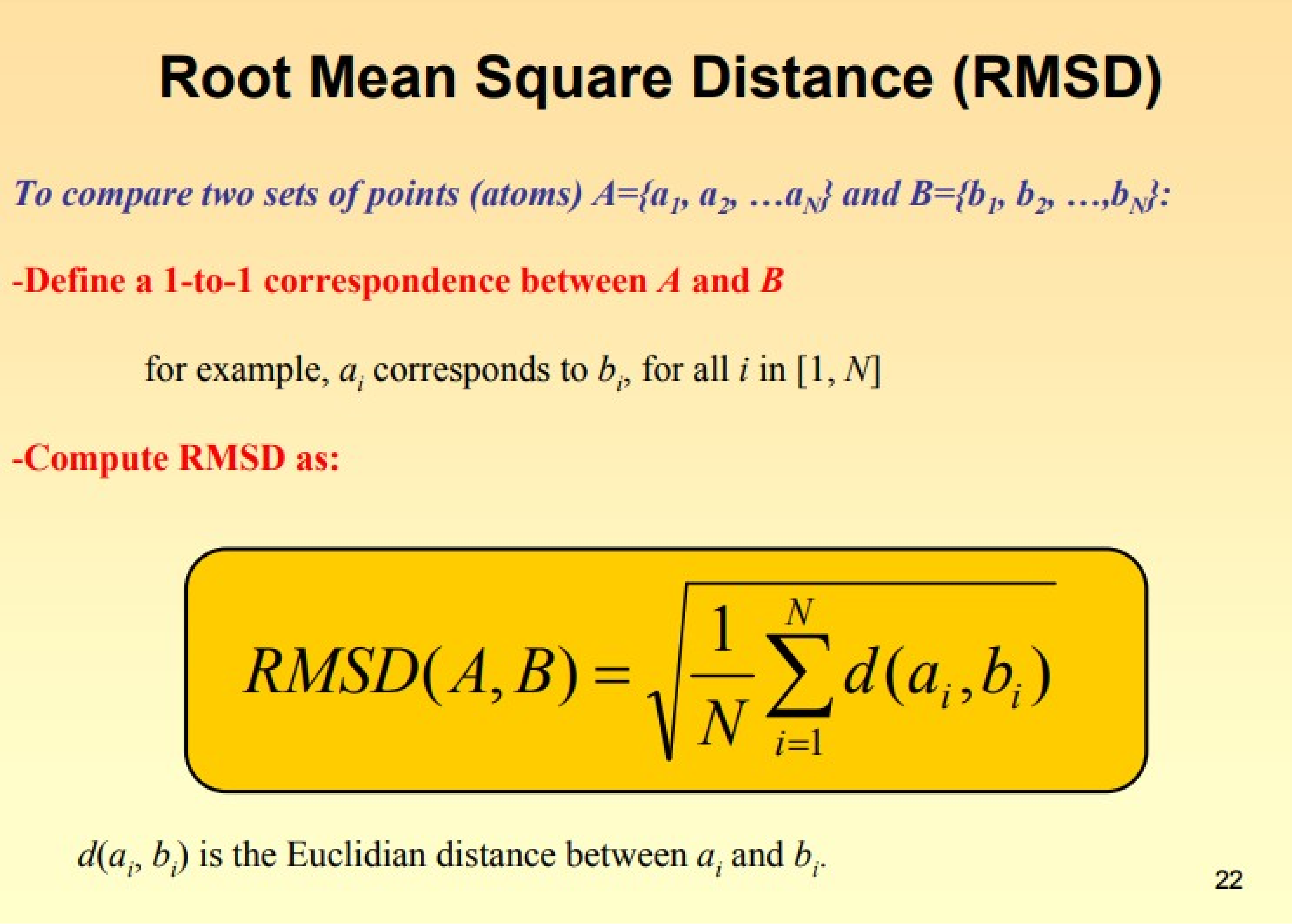


Distance Matrix (2)

- *Advantages*
  - invariant with respect to rotation and translation
  - can be used to compare proteins
- *Disadvantages*
  - the distance matrix is $O(n^2)$ for a protein with n residues
  - comparing distance matrix is a hard problem
  - insensitive to chirality

A chiral molecule is non-superposable on its mirror image. The presence of an asymmetric carbon center is one of several structural features that induce chirality in organic and inorganic molecules.

Alineamiento rigido de Proteinas



Root Mean Square Distance (RMSD)

To compare two sets of points (atoms) $A=\{a_1, a_2, \ldots a_N\}$ and $B=\{b_1, b_2, \ldots, b_N\}$:

-Define a 1-to-1 correspondence between *A* and *B*

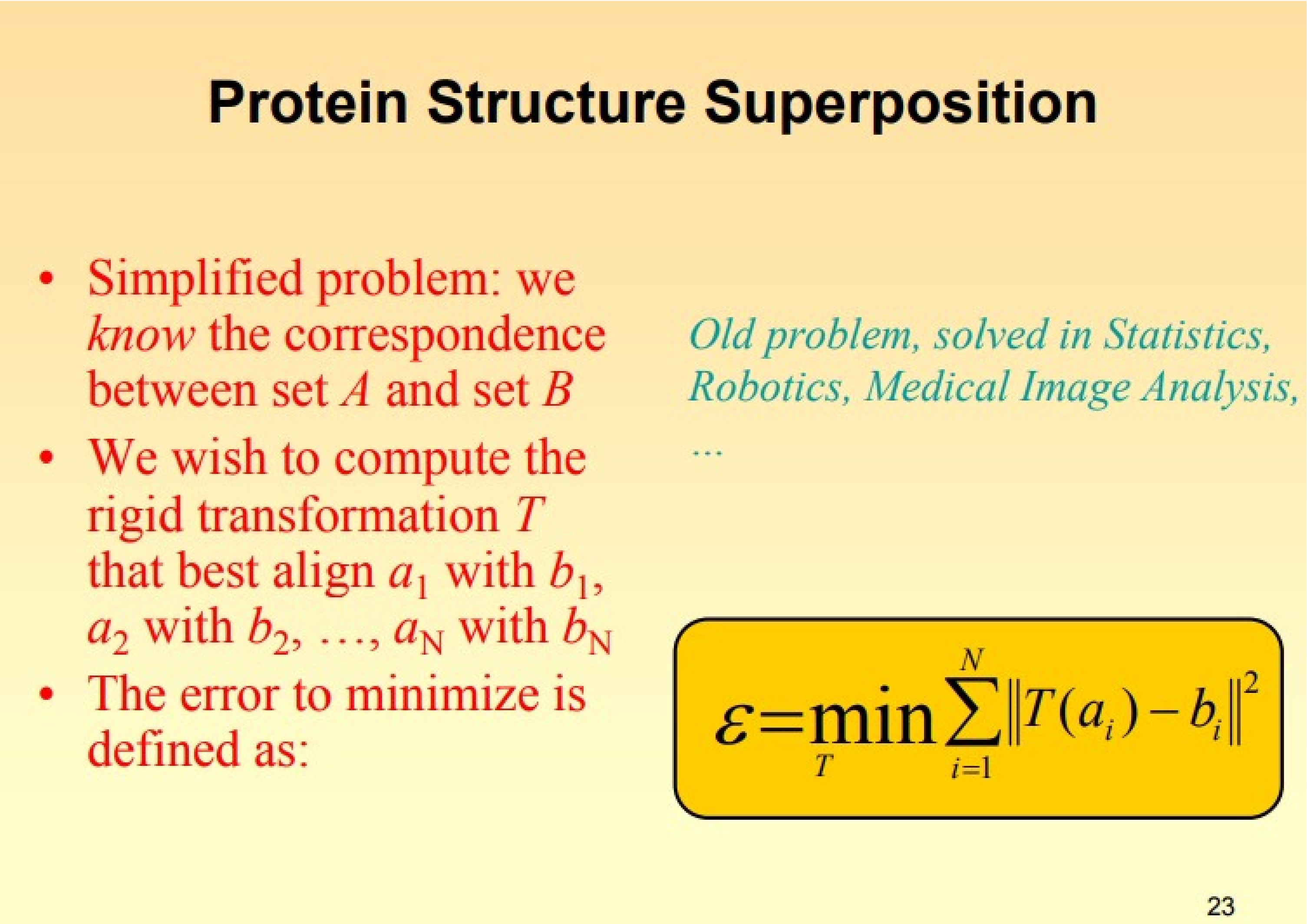
for example, a_i corresponds to b_i, for all i in $[1, N]$

-Compute RMSD as:

$$\boxed{RMSD(A,B) = \sqrt{\frac{1}{N}\sum_{i=1}^{N} d(a_i, b_i)}}$$

$d(a_i, b_i)$ is the Euclidian distance between a_i and b_i.

Alineamiento rigido de Proteinas



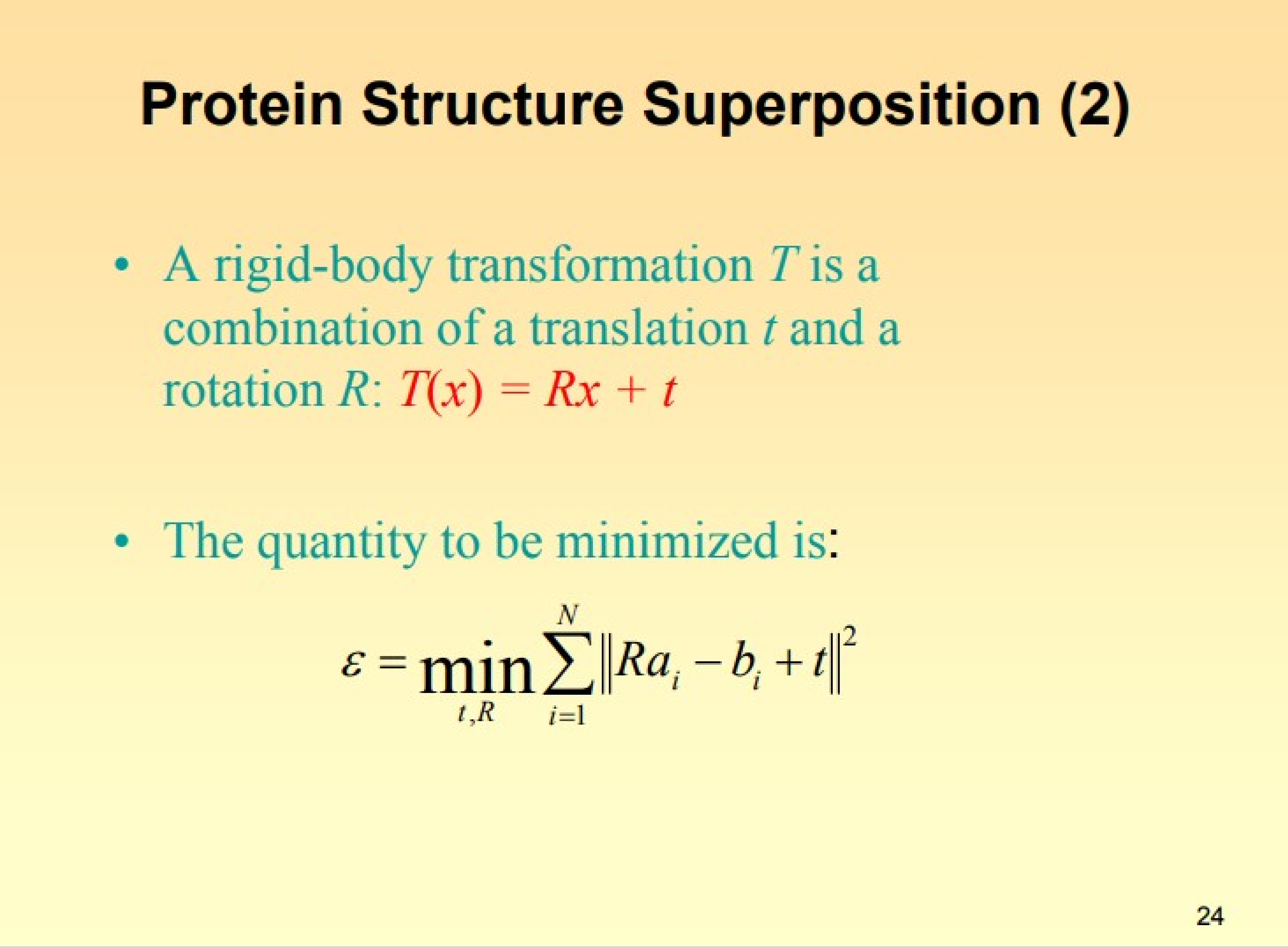
Protein Structure Superposition

- Simplified problem: we *know* the correspondence between set A and set B
- We wish to compute the rigid transformation T that best align a_1 with b_1, a_2 with b_2, …, a_N with b_N
- The error to minimize is defined as:

Old problem, solved in Statistics, Robotics, Medical Image Analysis, …

$$\boxed{\varepsilon = \min_{T} \sum_{i=1}^{N} \left\| T(a_i) - b_i \right\|^2}$$

Alineamiento rigido de Proteinas

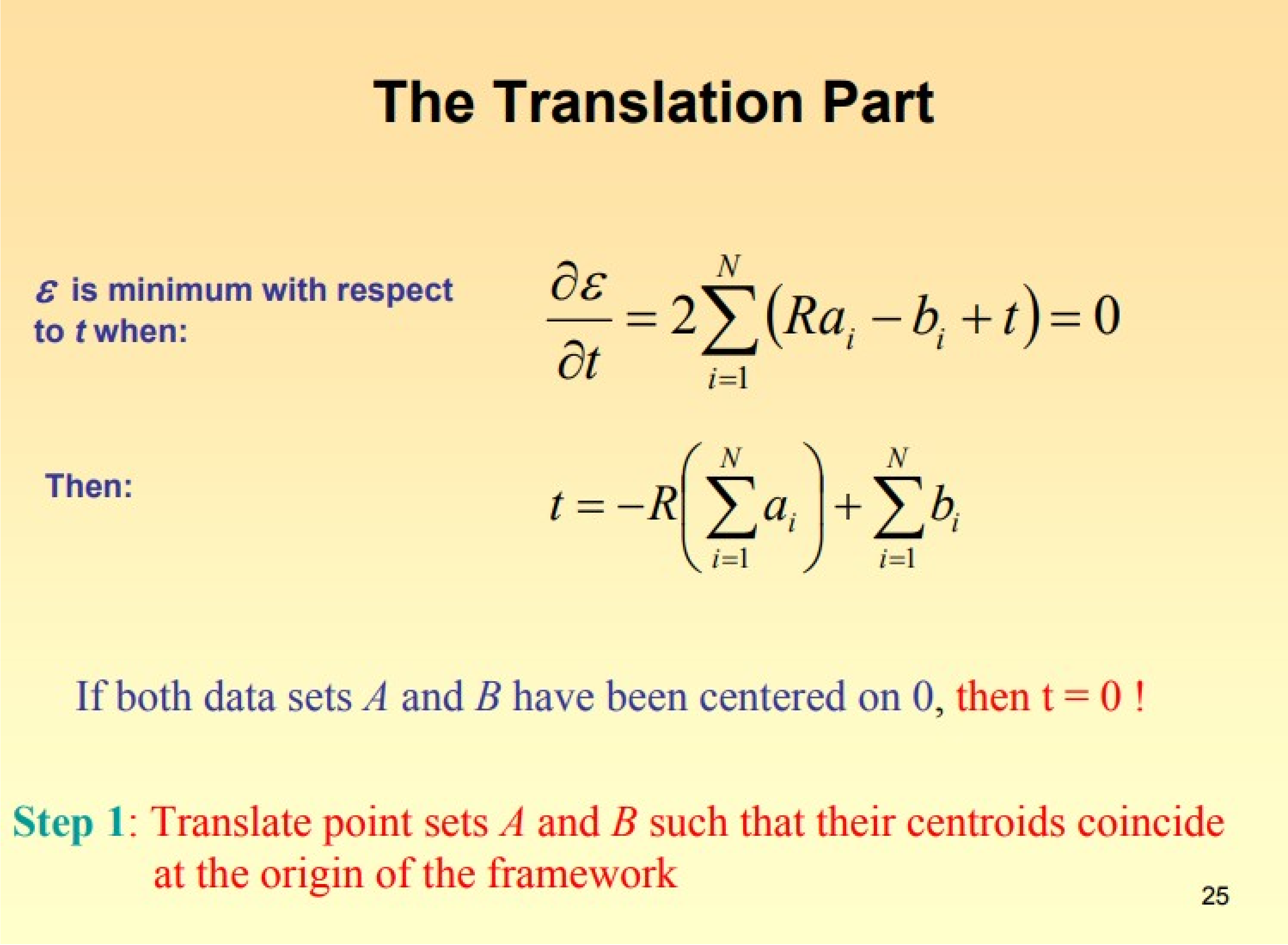


Protein Structure Superposition (2)

- A rigid-body transformation *T* is a combination of a translation *t* and a rotation *R*: $T(x) = Rx + t$
- The quantity to be minimized is:

$$\varepsilon = \min_{t,R} \sum_{i=1}^{N} \left\| Ra_i - b_i + t \right\|^2$$

Alineamiento rigido de Proteinas



The Translation Part

ε is minimum with respect to *t* when:

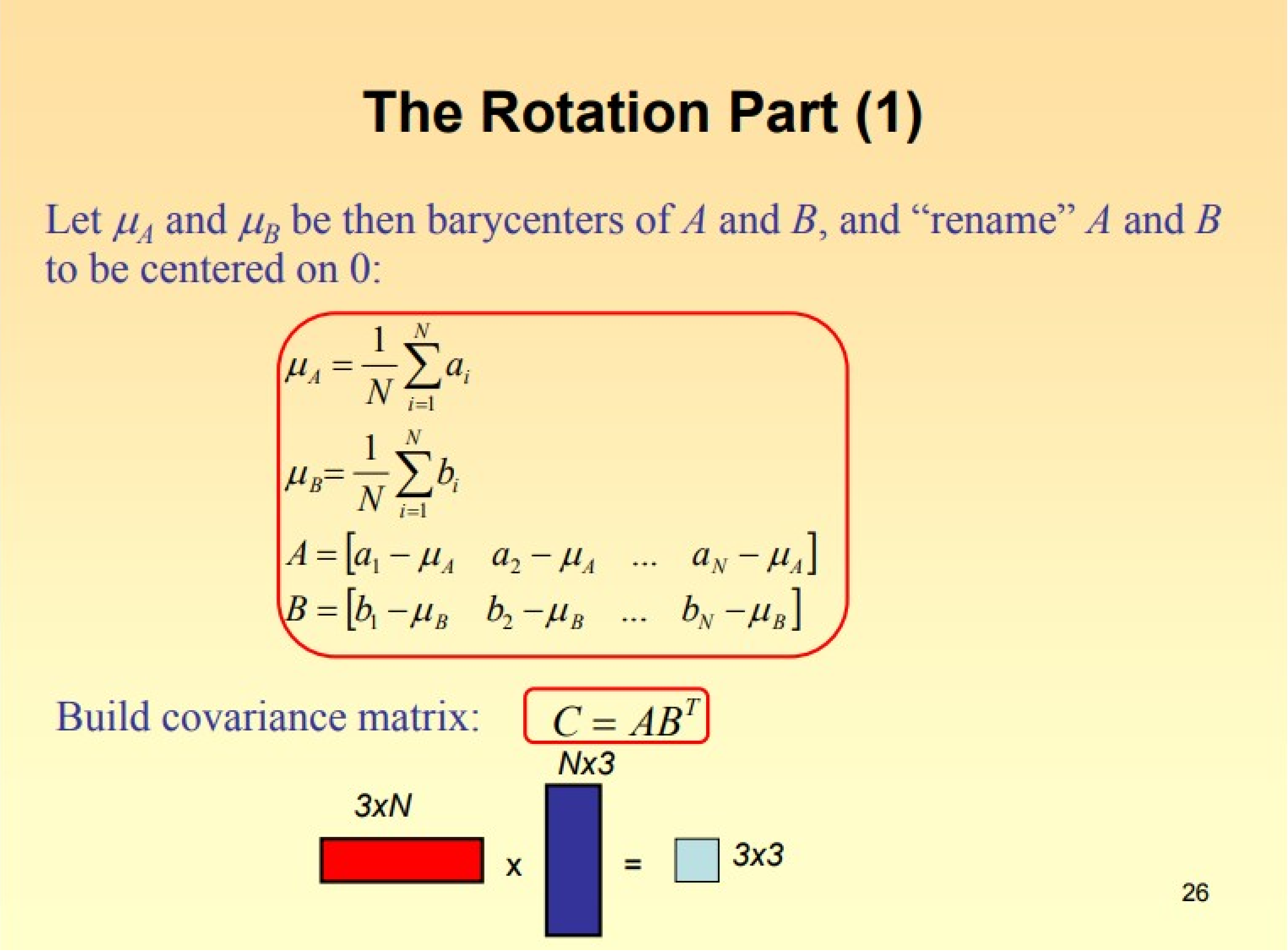
$$\frac{\partial \varepsilon}{\partial t} = 2\sum_{i=1}^{N}\left(Ra_i - b_i + t\right) = 0$$

Then:

$$t = -R\left(\sum_{i=1}^{N} a_i\right) + \sum_{i=1}^{N} b_i$$

If both data sets *A* and *B* have been centered on 0, then t = 0 !

Step 1: Translate point sets *A* and *B* such that their centroids coincide at the origin of the framework



The Rotation Part (1)

Let μ_A and μ_B be then barycenters of A and B, and “rename” A and B to be centered on 0:

$$\boxed{\begin{array}{l}\mu_A = \dfrac{1}{N}\sum_{i=1}^{N} a_i \\ \mu_B = \dfrac{1}{N}\sum_{i=1}^{N} b_i \\ A = \left[a_1 - \mu_A \quad a_2 - \mu_A \quad \ldots \quad a_N - \mu_A\right] \\ B = \left[b_1 - \mu_B \quad b_2 - \mu_B \quad \ldots \quad b_N - \mu_B\right]\end{array}}$$

Build covariance matrix: $\boxed{C = AB^T}$

Nx3

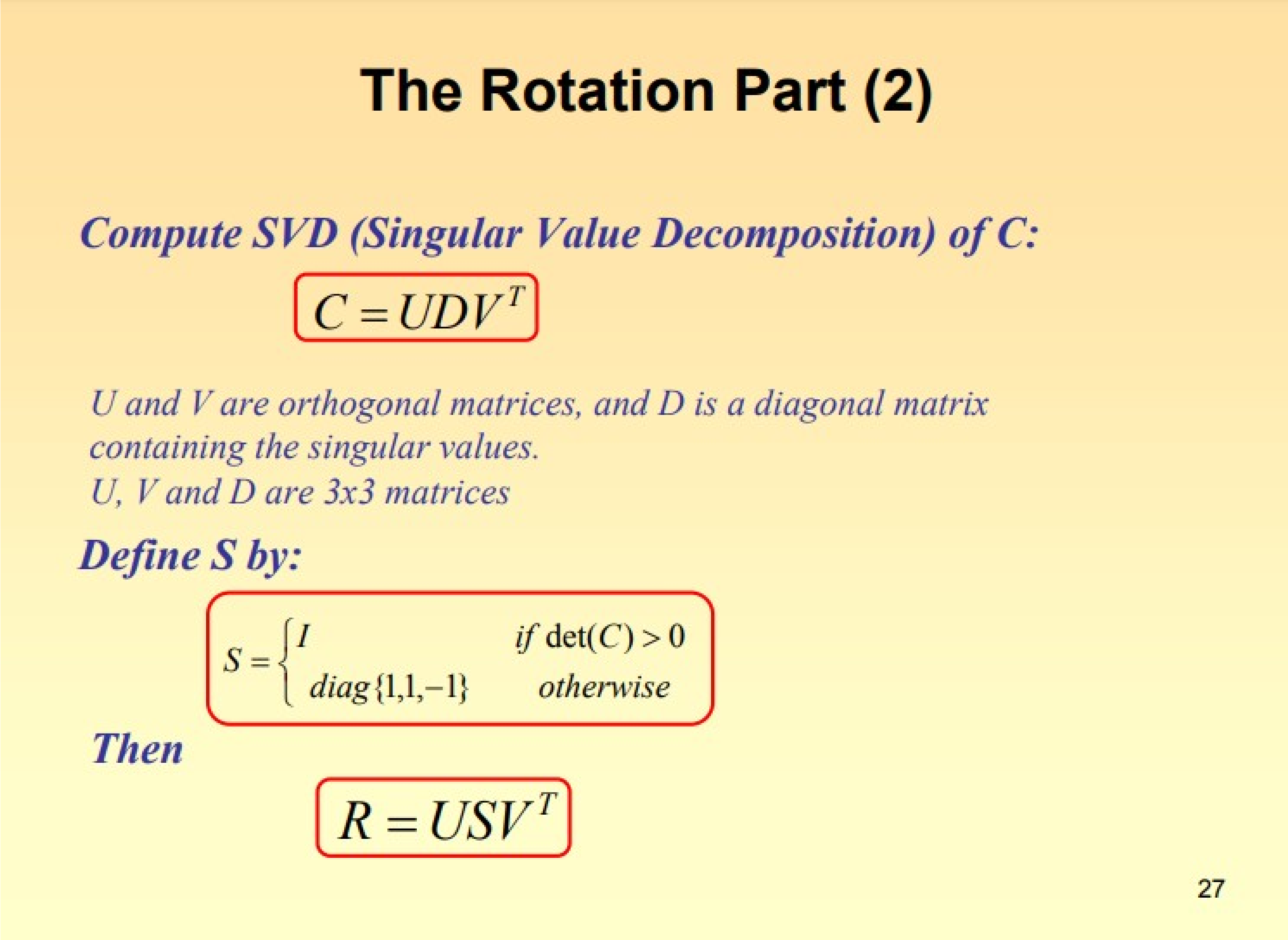
3xN

x

=

3x3

Alineamiento rigido de Proteinas



The Rotation Part (2)

Compute SVD (Singular Value Decomposition) of C:

$$\boxed{C = UDV^T}$$

U and V are orthogonal matrices, and D is a diagonal matrix containing the singular values.
U, V and D are 3x3 matrices

Define S by:

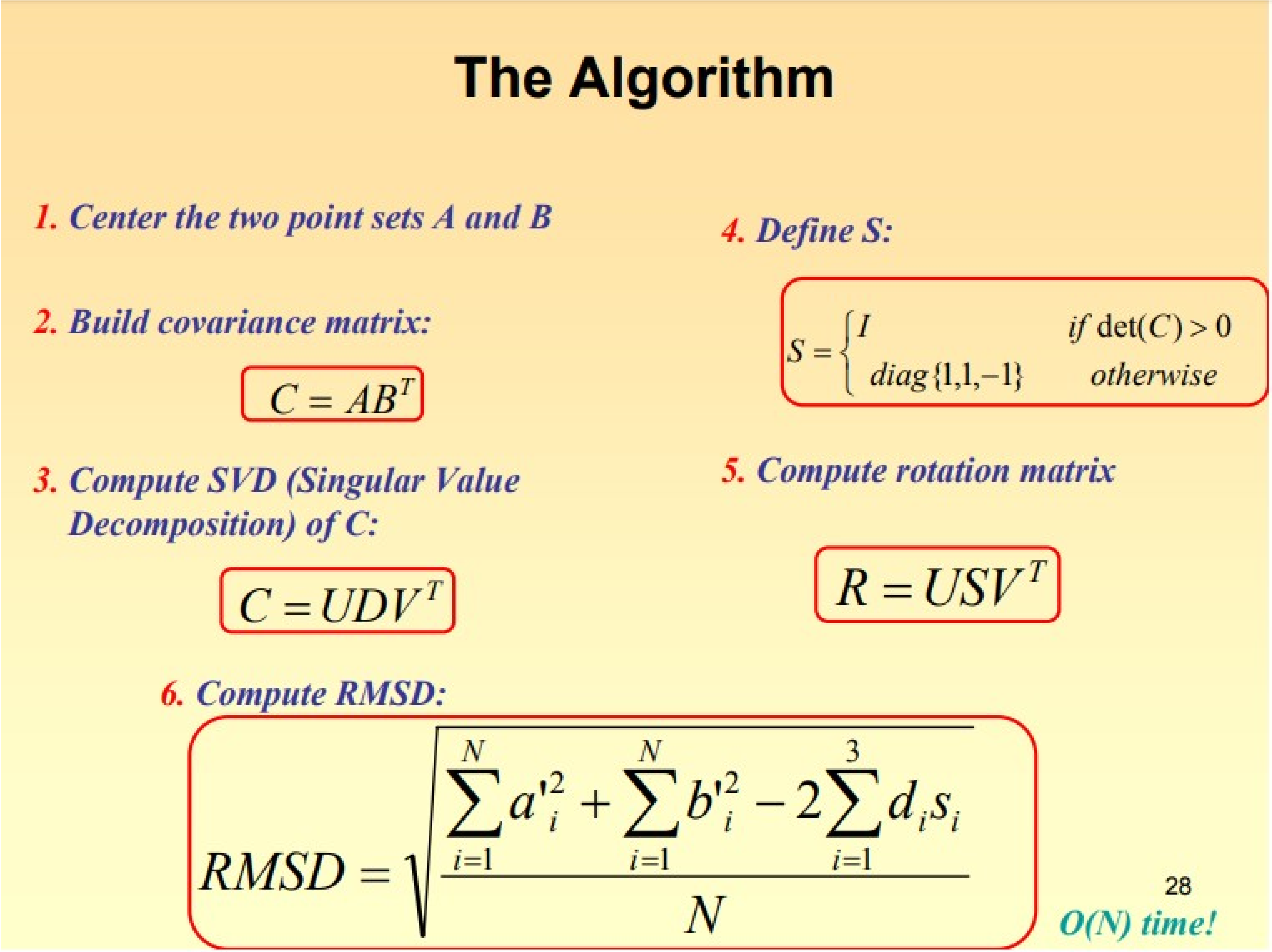
$$\boxed{S = \begin{cases} I & \text{if } \det(C) > 0 \\ diag\{1,1,-1\} & \text{otherwise} \end{cases}}$$

Then

$$\boxed{R = USV^T}$$

Una matriz ortogonal es una matriz cuadrada cuya matriz inversa coincide con su matriz traspuesta

Alineamiento rigido de Proteinas



The Algorithm

1. *Center the two point sets A and B*

2. *Build covariance matrix:*

$$\boxed{C = AB^T}$$

3. *Compute SVD (Singular Value Decomposition) of C:*

$$\boxed{C = UDV^T}$$

4. *Define S:*

$$\boxed{S = \begin{cases} I & \text{if } \det(C) > 0 \\ diag\{1,1,-1\} & otherwise \end{cases}}$$

5. *Compute rotation matrix*

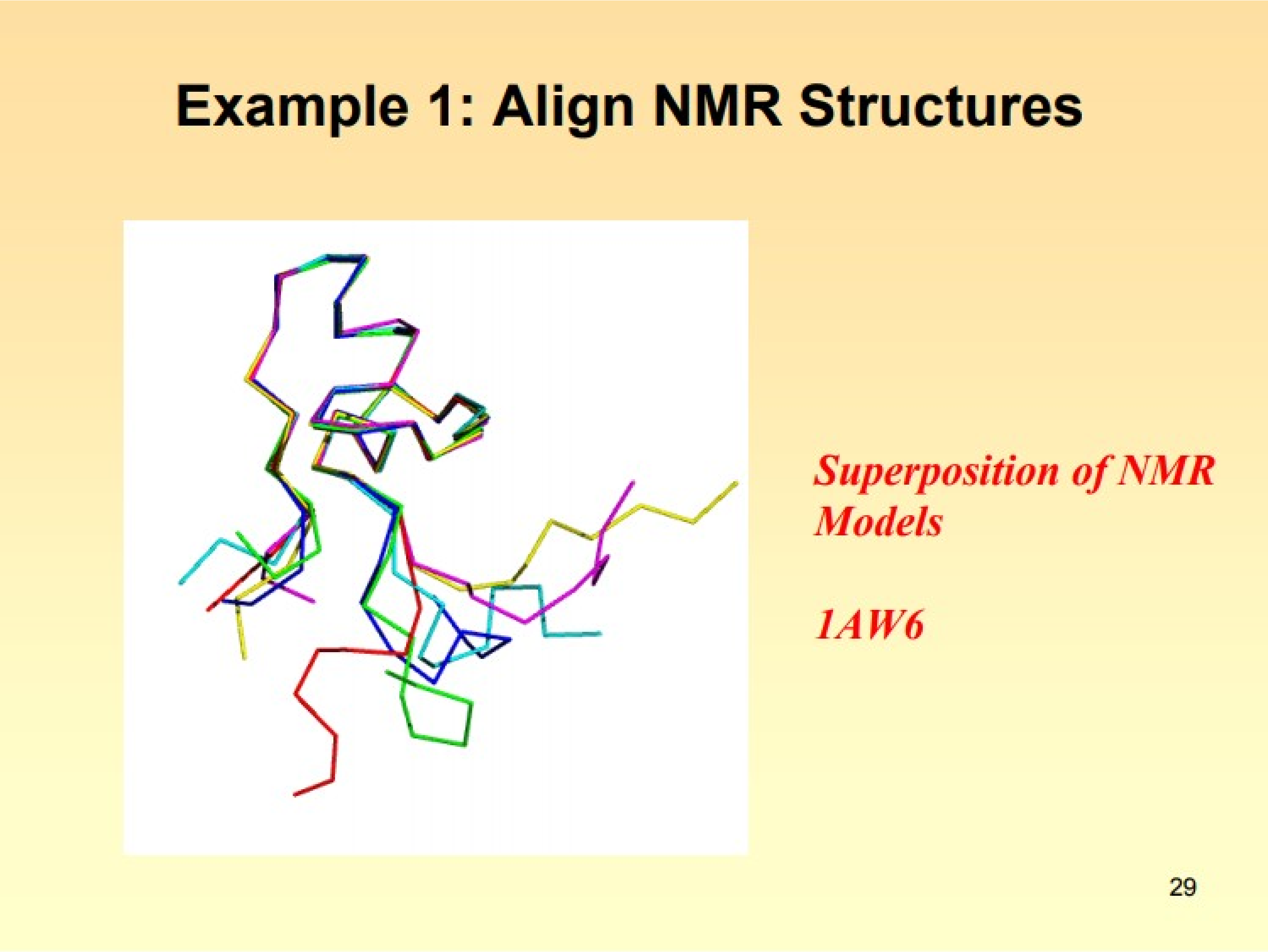
$$\boxed{R = USV^T}$$

6. *Compute RMSD:*

$$\boxed{RMSD = \sqrt{\frac{\sum_{i=1}^{N} a'^2_i + \sum_{i=1}^{N} b'^2_i - 2\sum_{i=1}^{3} d_i s_i}{N}}}$$

O(N) time!

Alineamiento rigido de Proteinas



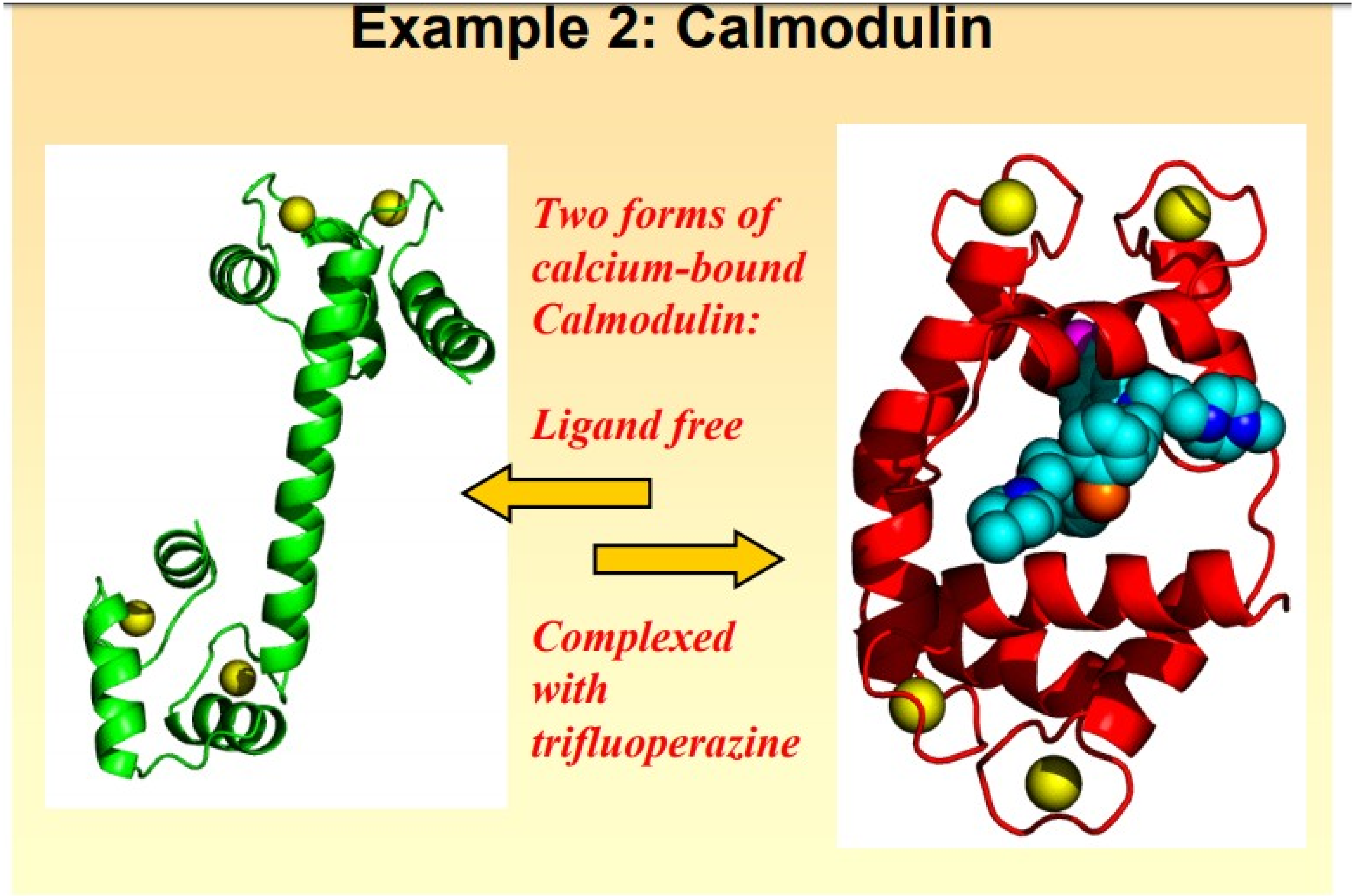
Example 1: Align NMR Structures

Superposition of NMR Models

1AW6

NMR – Nuclear magnetic resonance

Alineamiento rigido de Proteinas



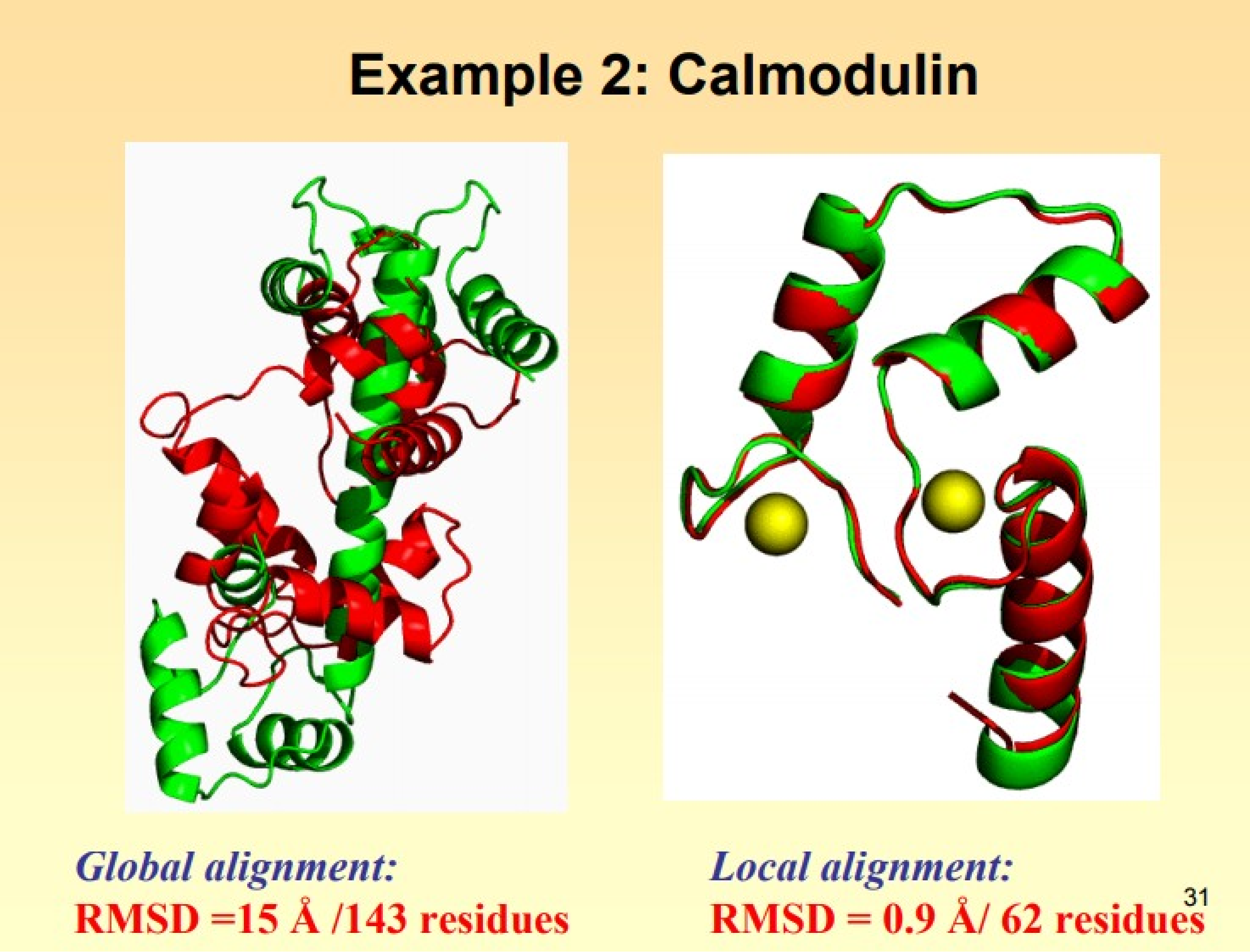
Example 2: Calmodulin

Two forms of calcium-bound Calmodulin:

Ligand free

Complexed with trifluoperazine

Alineamiento rigido de Proteinas

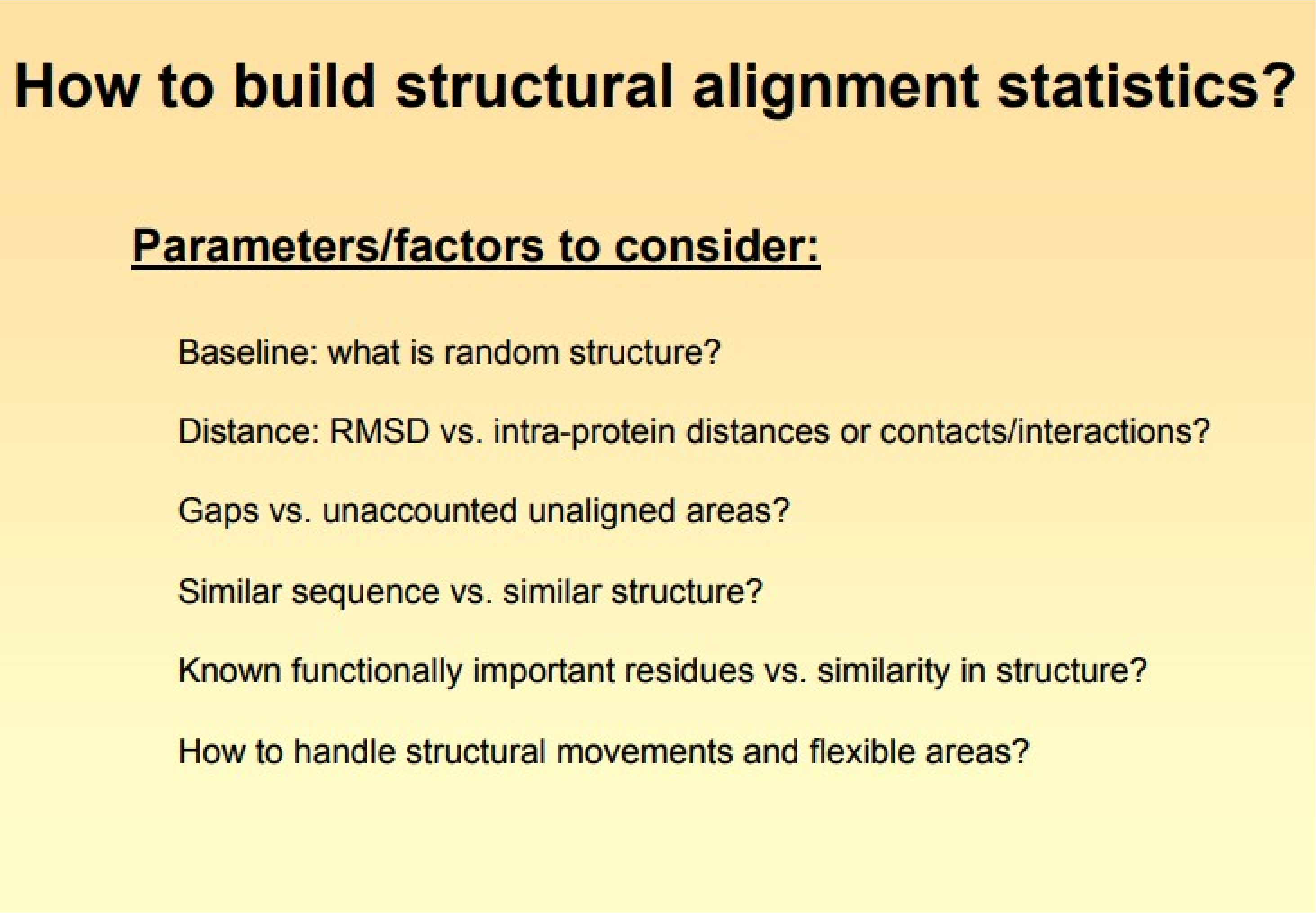


Example 2: Calmodulin

Global alignment:
RMSD =15 Å /143 residues

Local alignment:
RMSD = 0.9 Å/ 62 residues

Alineamiento rigido de Proteinas



How to build structural alignment statistics?

Parameters/factors to consider:

Baseline: what is random structure?

Distance: RMSD vs. intra-protein distances or contacts/interactions?

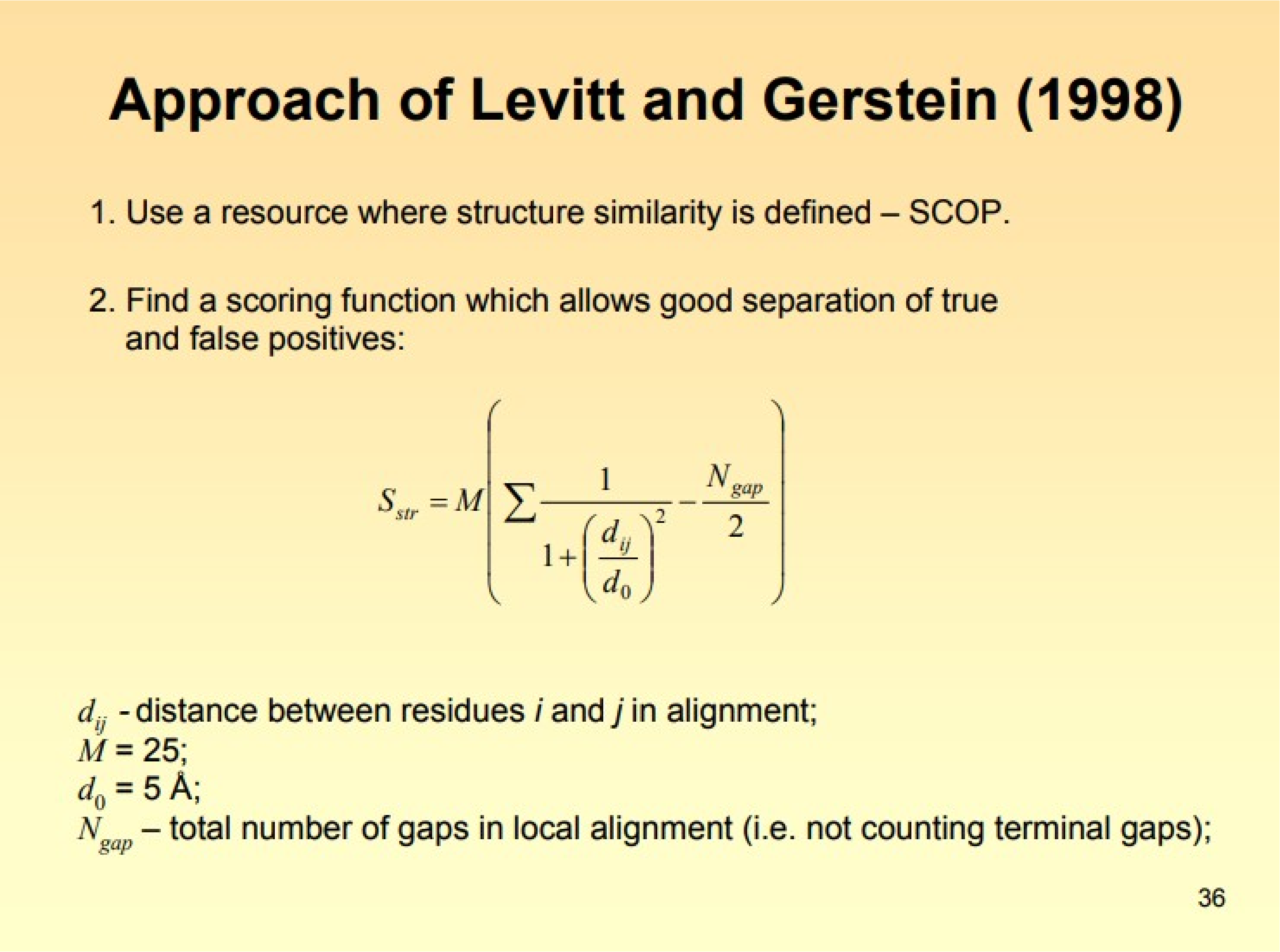
Gaps vs. unaccounted unaligned areas?

Similar sequence vs. similar structure?

Known functionally important residues vs. similarity in structure?

How to handle structural movements and flexible areas?

Alineamiento rigido de Proteinas



Approach of Levitt and Gerstein (1998)

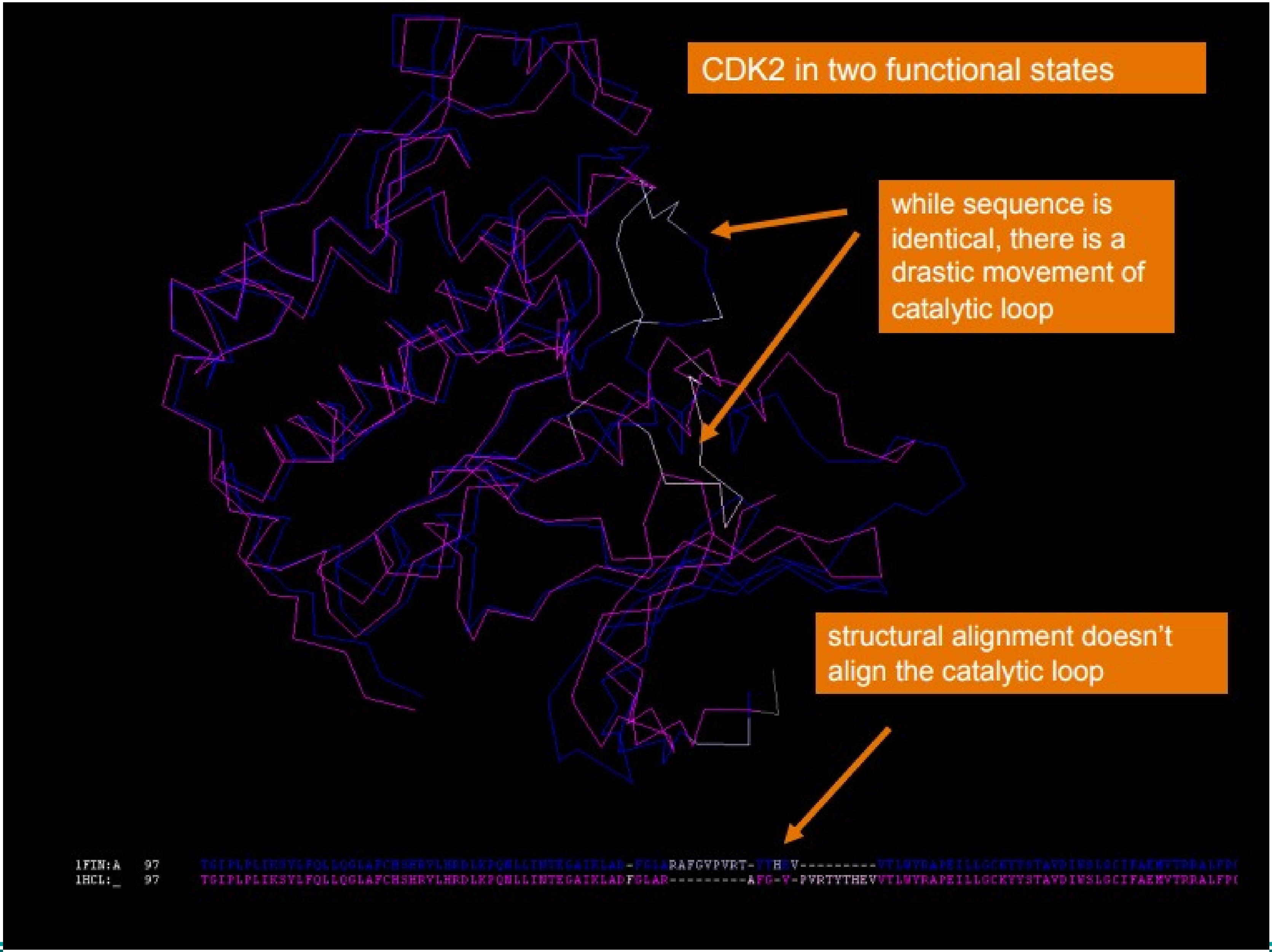
1. Use a resource where structure similarity is defined – SCOP.

2. Find a scoring function which allows good separation of true and false positives:

$$S_{str} = M\left(\sum \frac{1}{1+\left(\frac{d_{ij}}{d_0}\right)^2} - \frac{N_{gap}}{2}\right)$$

d_{ij} - distance between residues *i* and *j* in alignment;
M = 25;
d_0 = 5 Å;
N_{gap} – total number of gaps in local alignment (i.e. not counting terminal gaps);

Alineamiento rigido de Proteinas



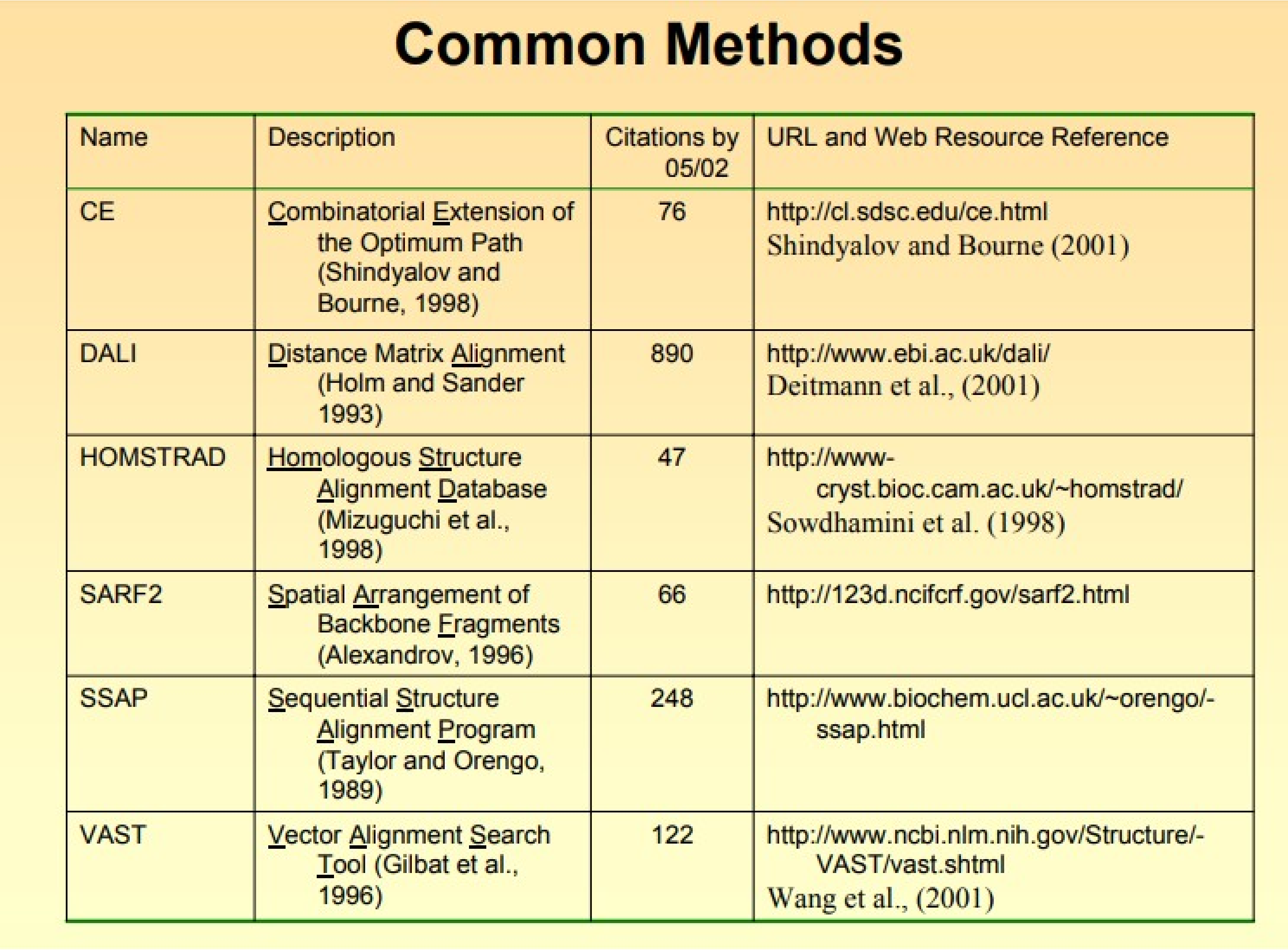
CDK2 in two functional states

while sequence is identical, there is a drastic movement of catalytic loop

structural alignment doesn't align the catalytic loop

```
1FIN:A   97    TGIPLPLIKSYLFQLLQGLAFCHSHRVLHRDLKPQNLLINTEGAIKLAD-FGLARAFGVPVRT-YTHEV---------VTLWYRAPEILLGCKYYSTAVDIWSLGCIFAEMVTRRALFP(
1HCL:_   97    TGIPLPLIKSYLFQLLQGLAFCHSHRVLHRDLKPQNLLINTEGAIKLADFGLAR---------AFG-V-PVRTYTHEVVTLWYRAPEILLGCKYYSTAVDIWSLGCIFAEMVTRRALFP(
```

Alineamiento rigido de Proteinas



Common Methods

Name	Description	Citations by 05/02	URL and Web Resource Reference
CE	Combinatorial Extension of the Optimum Path (Shindyalov and Bourne, 1998)	76	http://cl.sdsc.edu/ce.html Shindyalov and Bourne (2001)
DALI	Distance Matrix Alignment (Holm and Sander 1993)	890	http://www.ebi.ac.uk/dali/ Deitmann et al., (2001)
HOMSTRAD	Homologous Structure Alignment Database (Mizuguchi et al., 1998)	47	http://www-cryst.bioc.cam.ac.uk/~homstrad/ Sowdhamini et al. (1998)
SARF2	Spatial Arrangement of Backbone Fragments (Alexandrov, 1996)	66	http://123d.ncifcrf.gov/sarf2.html
SSAP	Sequential Structure Alignment Program (Taylor and Orengo, 1989)	248	http://www.biochem.ucl.ac.uk/~orengo/-ssap.html
VAST	Vector Alignment Search Tool (Gilbat et al., 1996)	122	http://www.ncbi.nlm.nih.gov/Structure/-VAST/vast.shtml Wang et al., (2001)

Alineamiento rigido de Proteinas

The accuracy of protein structure alignment servers

Naeem Aslama, Asif Nadeema, et.al.

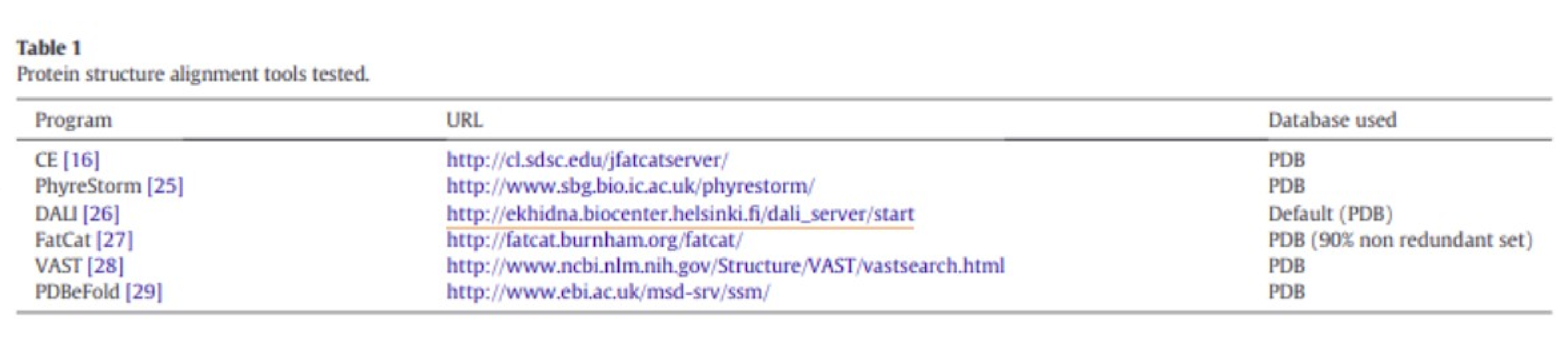


Table 1
Protein structure alignment tools tested.

Program	URL	Database used
CE [16]	http://cl.sdsc.edu/jfatcatserver/	PDB
PhyreStorm [25]	http://www.sbg.bio.ic.ac.uk/phyrestorm/	PDB
DALI [26]	http://ekhidna.biocenter.helsinki.fi/dali_server/start	Default (PDB)
FatCat [27]	http://fatcat.burnham.org/fatcat/	PDB (90% non redundant set)
VAST [28]	http://www.ncbi.nlm.nih.gov/Structure/VAST/vastsearch.html	PDB
PDBeFold [29]	http://www.ebi.ac.uk/msd-srv/ssm/	PDB

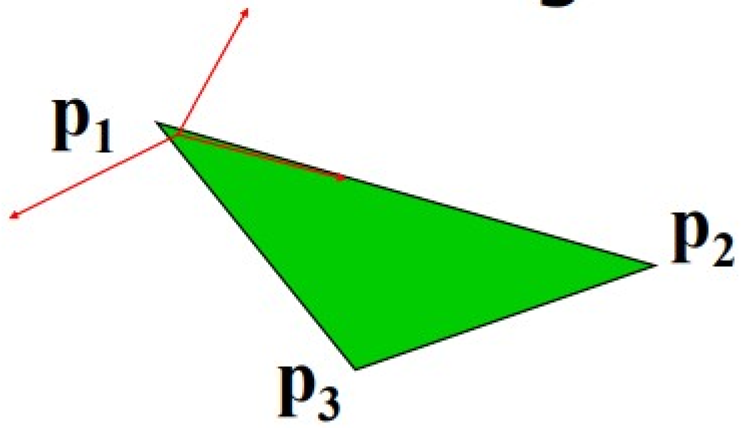
Alineamiento rigido de Proteinas

Geometric Hashing

- Developed for object recognition in Computer Vision (Lamdan, Schwartz, Wolfson, 1988 - rigid, Wolfson ,1991 -flexible).
- Adapted to Molecular Biology (Nussinov, Wolfson , 1989).
- Motivated by *associative memory* ideas and efficient *hashing* techniques.

Alineamiento rigido de Proteinas

A 3-D reference frame can be uniquely defined by the ordered vertices of a non-degenerate triangle



The lengths of the triangle sides are rigid motion *invariant.*

Geometric Hashing Technique

Geometric Hashing is a technique for matching a target molecule (or a collection of such) against a set of one or more models(e.g. database). The models are assumed to be known in advance.

Geometric Hashing Technique

For each model object (first molecule in our case or each molecule in database) do:

1.Pick a reference frame.
2.Compute the 3D orthonormal basis associated with this reference frame and its shape signature (e.g. triangle sides length).
3.Compute the coordinates of all the other points (in a pre-specified neighborhood) in this reference frame.
4.Use each coordinate as an address to the hash (look-up) table. Store the entry [protein id, ref. frame, shape sign., point] at the hash table address.
5.Repeat above steps for each model reference frame (noncollinear triplet of model points).

Alineamiento rigido de Proteinas

Geometric Hashing Technique

Recognition stage of the algorithm uses the hash table, prepared in the preprocessing step.

Matching of a target:

1.For each reference frame of the target:

2.Compute the 3D orthonormal basis and the shape signature associated with it.

3.Compute the coordinates of all other points in the current reference frame.

4.Use each coordinate to access the hash-table and retrieve all the records [protein id, ref. frame, shape sign.,point].

5.For records with matching shape signature ``vote'' for the pair [protein, ref. frame].

6.Compute the transformations of the ``high scoring'' hypotheses. For each hypothesis one can also register the pairs of matching points. This match list along with the transformation comprise a seed match.

Alineamiento rigido de Proteinas

Geometric Hashing Technique

The alignment algorithm that uses Geometric Hashing:

We first define local neighbours of residues. Note that if we use all points for all possible triplets, each atom will have a redundant representation. It will appear in the hash table in all possible reference frames. (In practise, since we are not interested in very closed nor very distant atoms, we pick atoms in an annulus defined by min and max radii).

The Geometric Hashing technique is applied next using only neighboring points to detect seed matches defined by a transformation and a match-list. Many of the matches obtained before represent the same transformation, i.e. different match lists may share the same transformation. We cluster seed matches and merge match-lists that were found.

The last step of the algorithm is the extending step. The seed matches are extended to contain additional matching pairs and best RMSD transformation is detected. For this purpose a heuristic iterative matching algorithm which minimizes the sum of the distances between the newly matched pairs is applied.

Alineamiento rigido de Proteinas

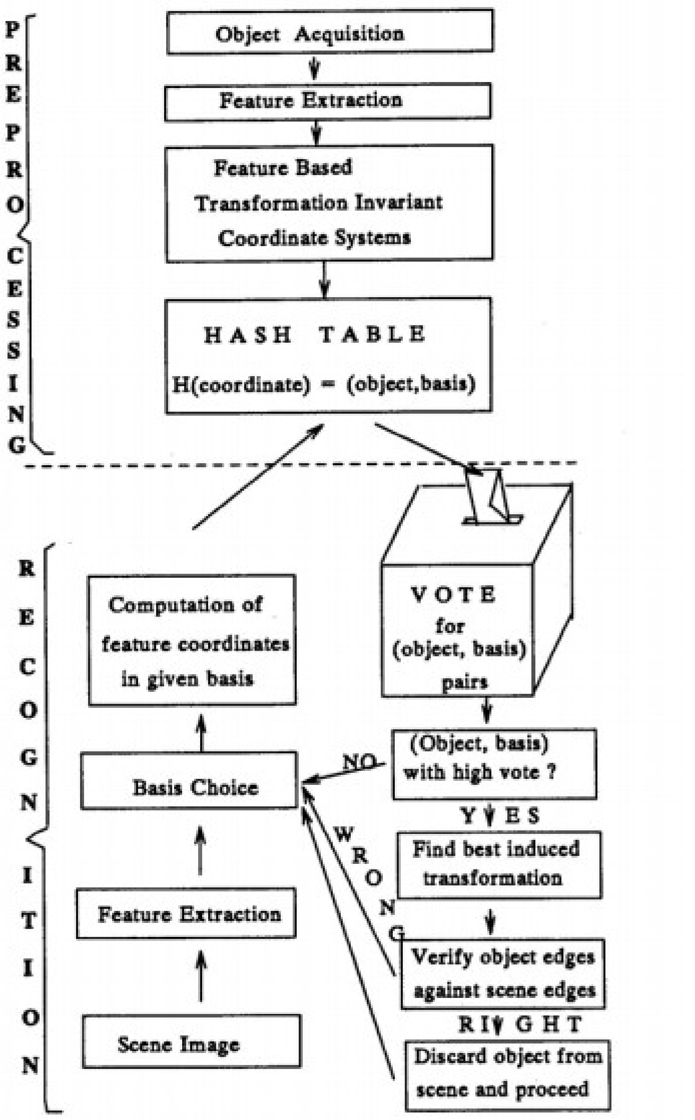
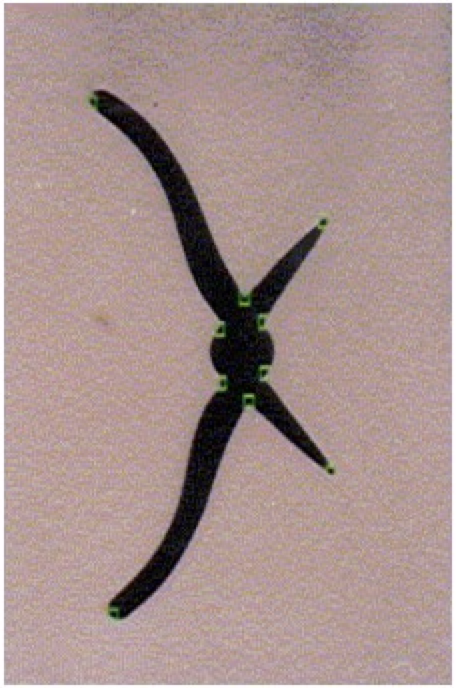

PREPROCESSING
Object Acquisition
Feature Extraction
Feature Based Transformation Invariant Coordinate Systems
HASH TABLE
H(coordinate) = (object,basis)
RECOGNITION
Computation of feature coordinates in given basis
VOTE for (object, basis) pairs
(Object, basis) with high vote ?
NO
Basis Choice
YES
Find best induced transformation
WRONG
Feature Extraction
Verify object edges against scene edges
RIGHT
Scene Image
Discard object from scene and proceed

Figure 1 : The general scheme of the object recognition algorithm.

Lamdan & Wolfson, Geometric Hashing, ICCV'88

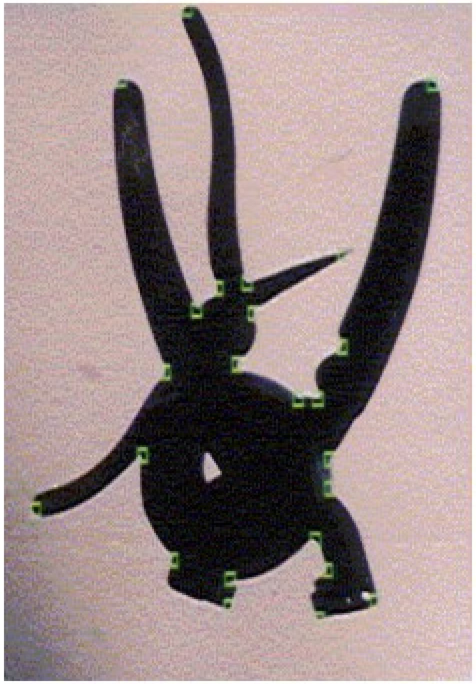
Alineamiento rigido de Proteinas

Model Database



Alineamiento rigido de Proteinas

Scene

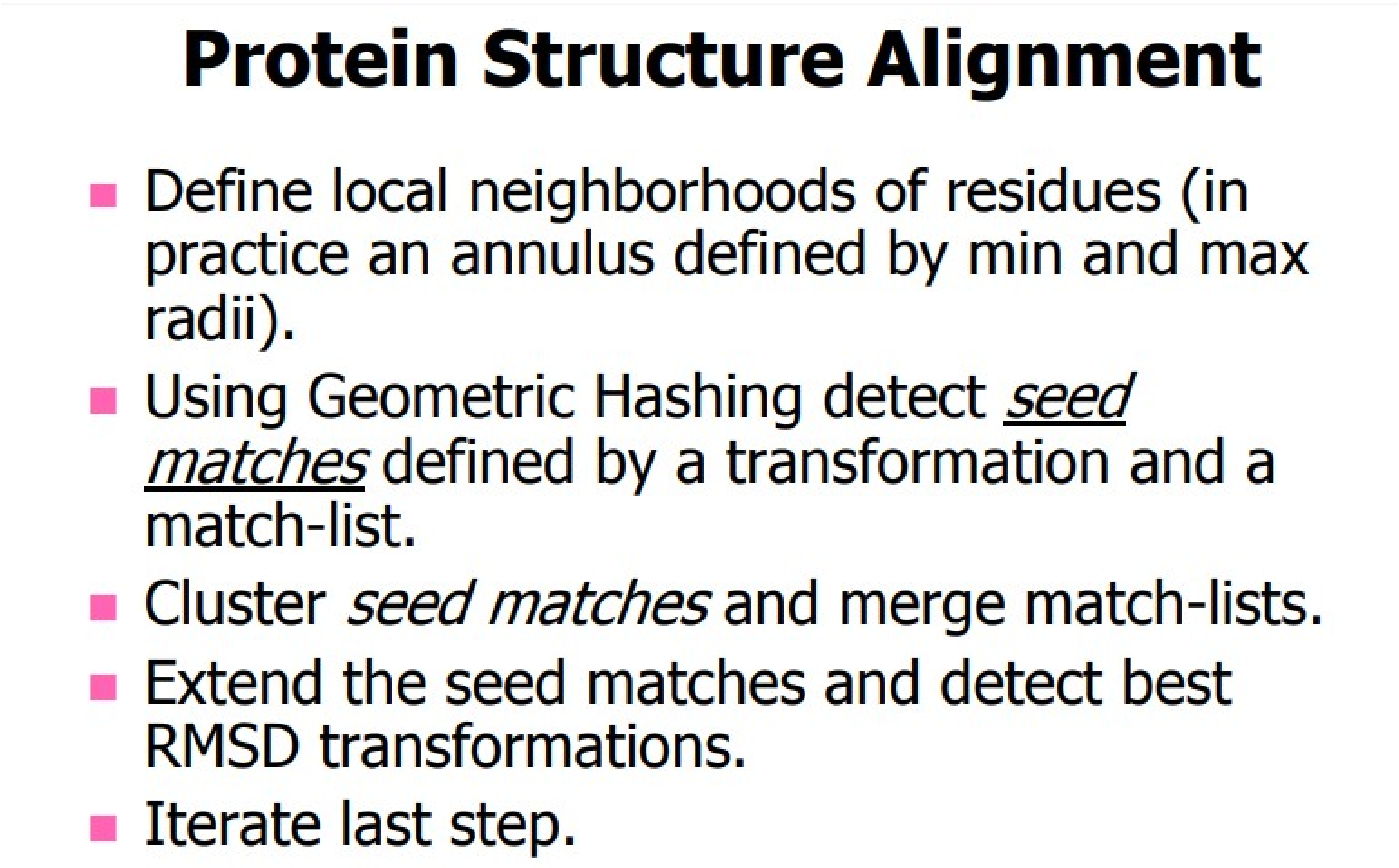


Alineamiento rigido de Proteinas

Recognition

Lamdan, Schwartz, Wolfson, "Geometric Hashing",1988.

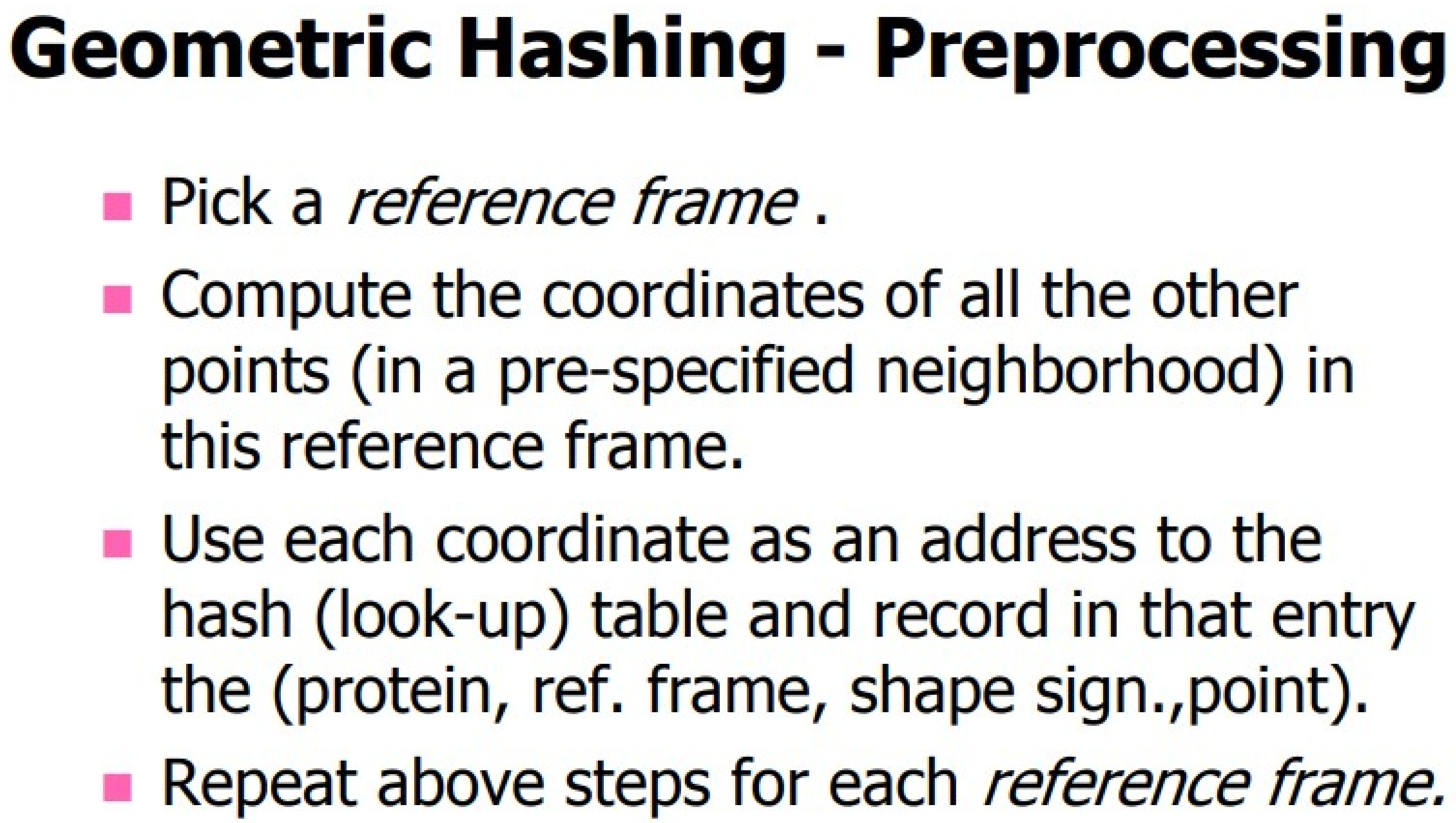
Alineamiento rigido de Proteinas



Protein Structure Alignment

- Define local neighborhoods of residues (in practice an annulus defined by min and max radii).
- Using Geometric Hashing detect *seed matches* defined by a transformation and a match-list.
- Cluster *seed matches* and merge match-lists.
- Extend the seed matches and detect best RMSD transformations.
- Iterate last step.

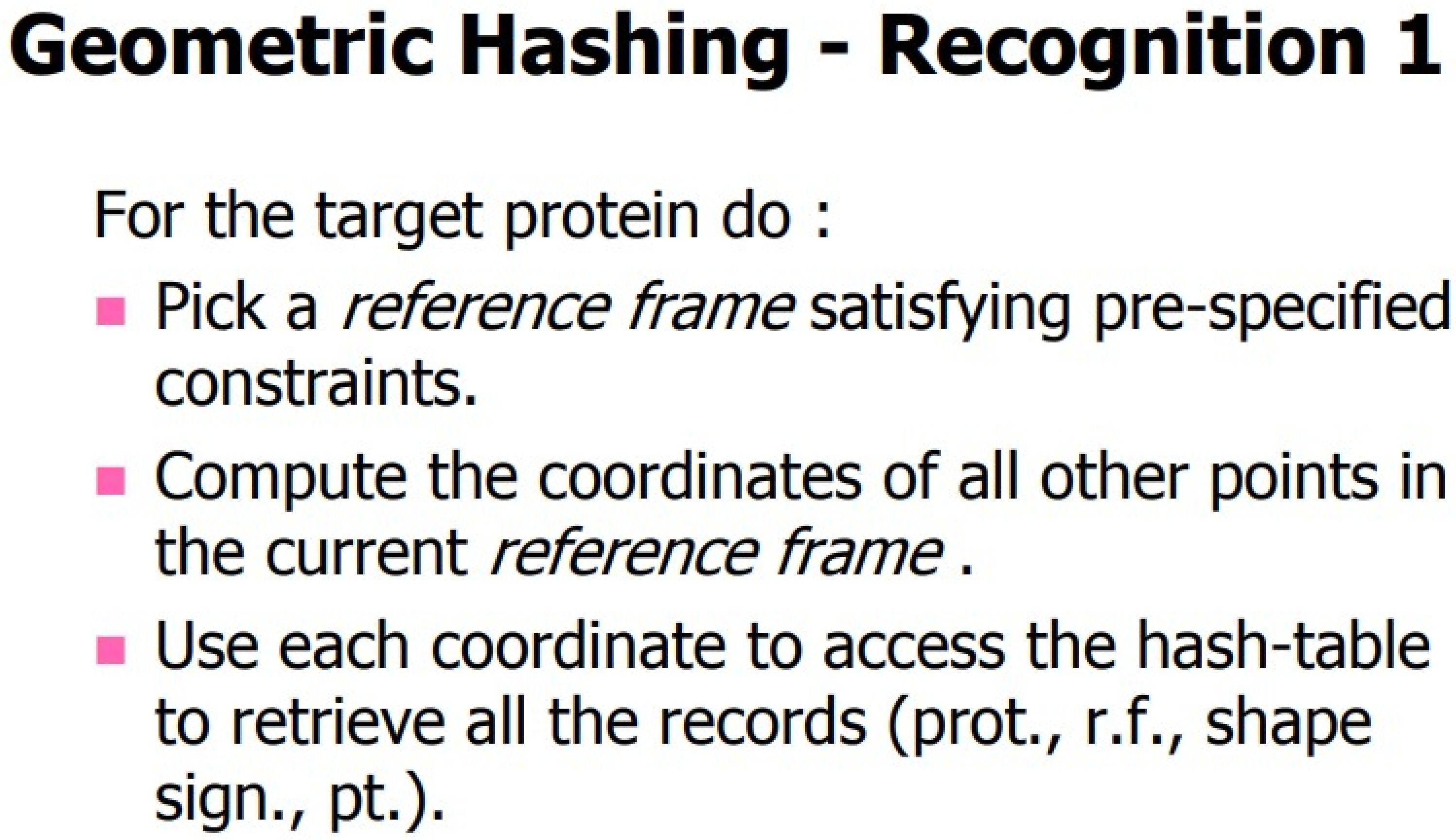
Alineamiento rigido de Proteinas



Geometric Hashing - Preprocessing

- Pick a *reference frame* .
- Compute the coordinates of all the other points (in a pre-specified neighborhood) in this reference frame.
- Use each coordinate as an address to the hash (look-up) table and record in that entry the (protein, ref. frame, shape sign.,point).
- Repeat above steps for each *reference frame.*

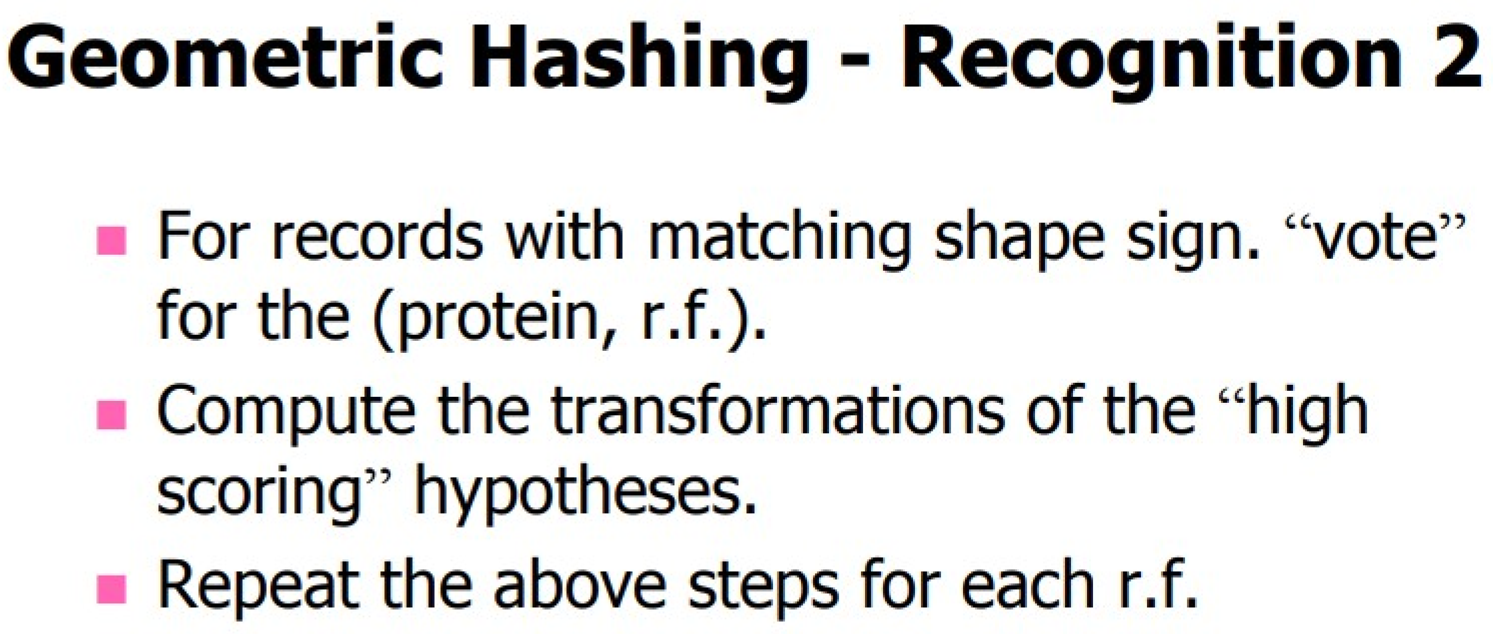
Alineamiento rigido de Proteinas



Geometric Hashing - Recognition 1

For the target protein do :

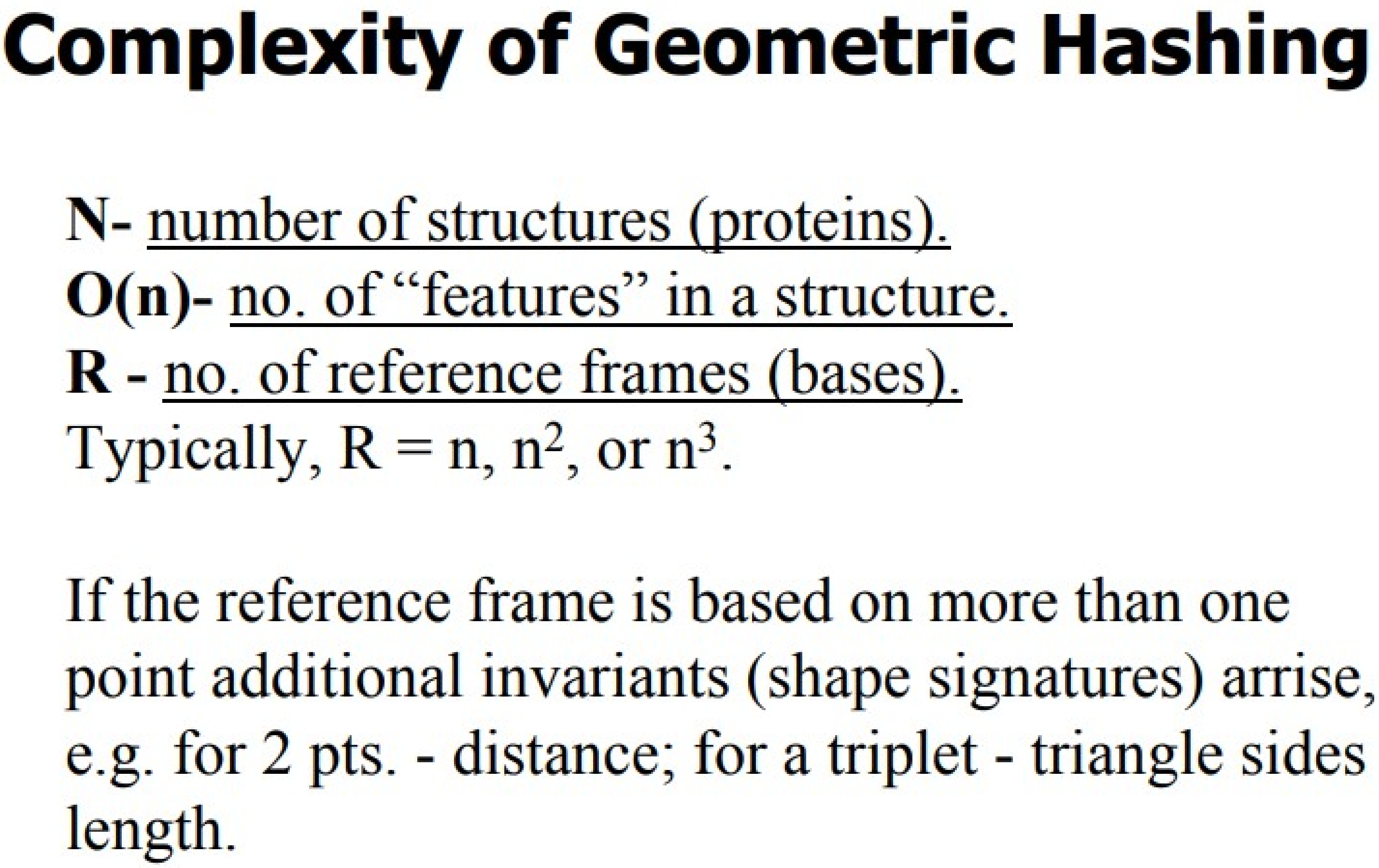
- Pick a *reference frame* satisfying pre-specified constraints.
- Compute the coordinates of all other points in the current *reference frame* .
- Use each coordinate to access the hash-table to retrieve all the records (prot., r.f., shape sign., pt.).



Geometric Hashing - Recognition 2

- For records with matching shape sign. “vote” for the (protein, r.f.).
- Compute the transformations of the “high scoring” hypotheses.
- Repeat the above steps for each r.f.

Alineamiento rigido de Proteinas

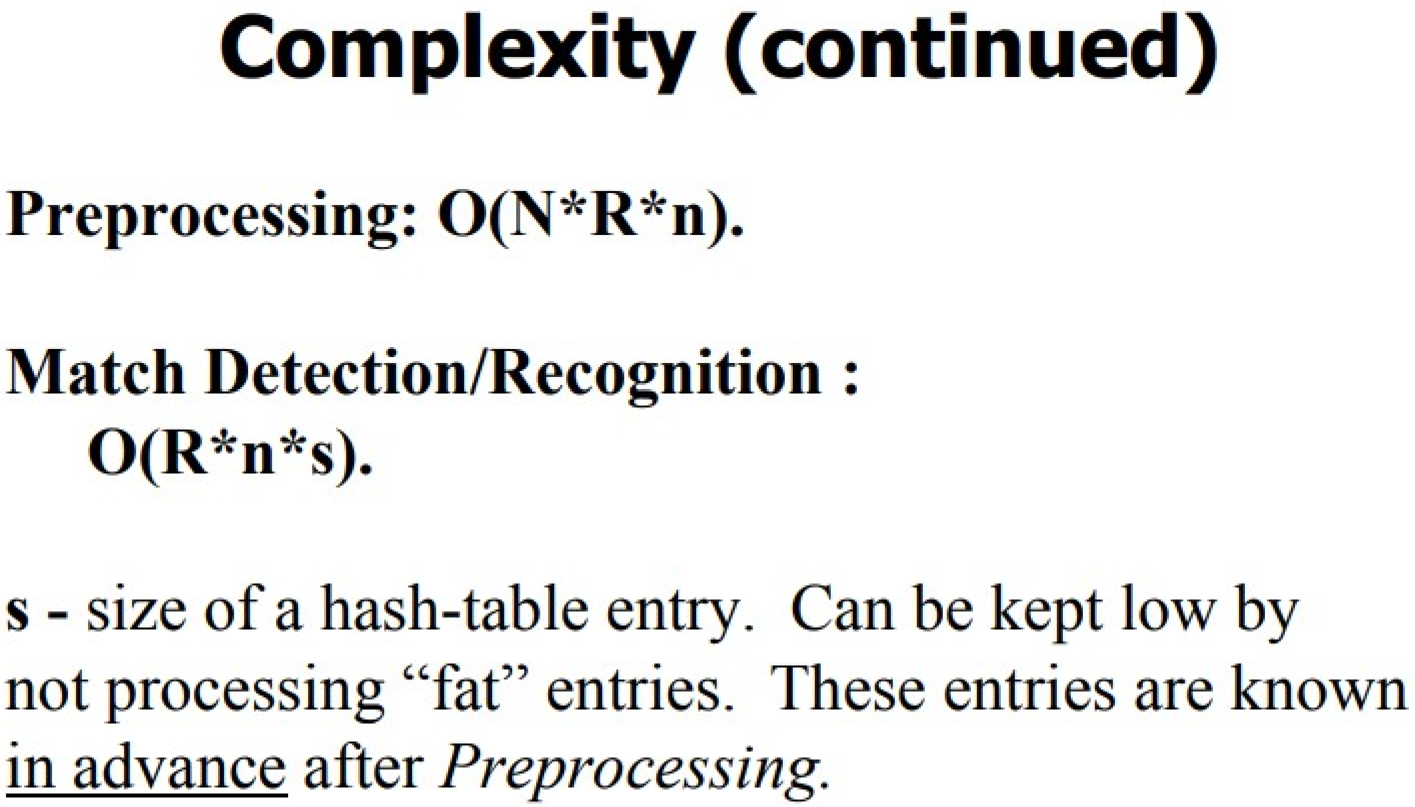


Complexity of Geometric Hashing

N- number of structures (proteins).
O(n)- no. of "features" in a structure.
R - no. of reference frames (bases).
Typically, R = n, n^2, or n^3.

If the reference frame is based on more than one point additional invariants (shape signatures) arrise, e.g. for 2 pts. - distance; for a triplet - triangle sides length.

Alineamiento rigido de Proteinas



Complexity (continued)

Preprocessing: O(N*R*n).

Match Detection/Recognition :
O(R*n*s).

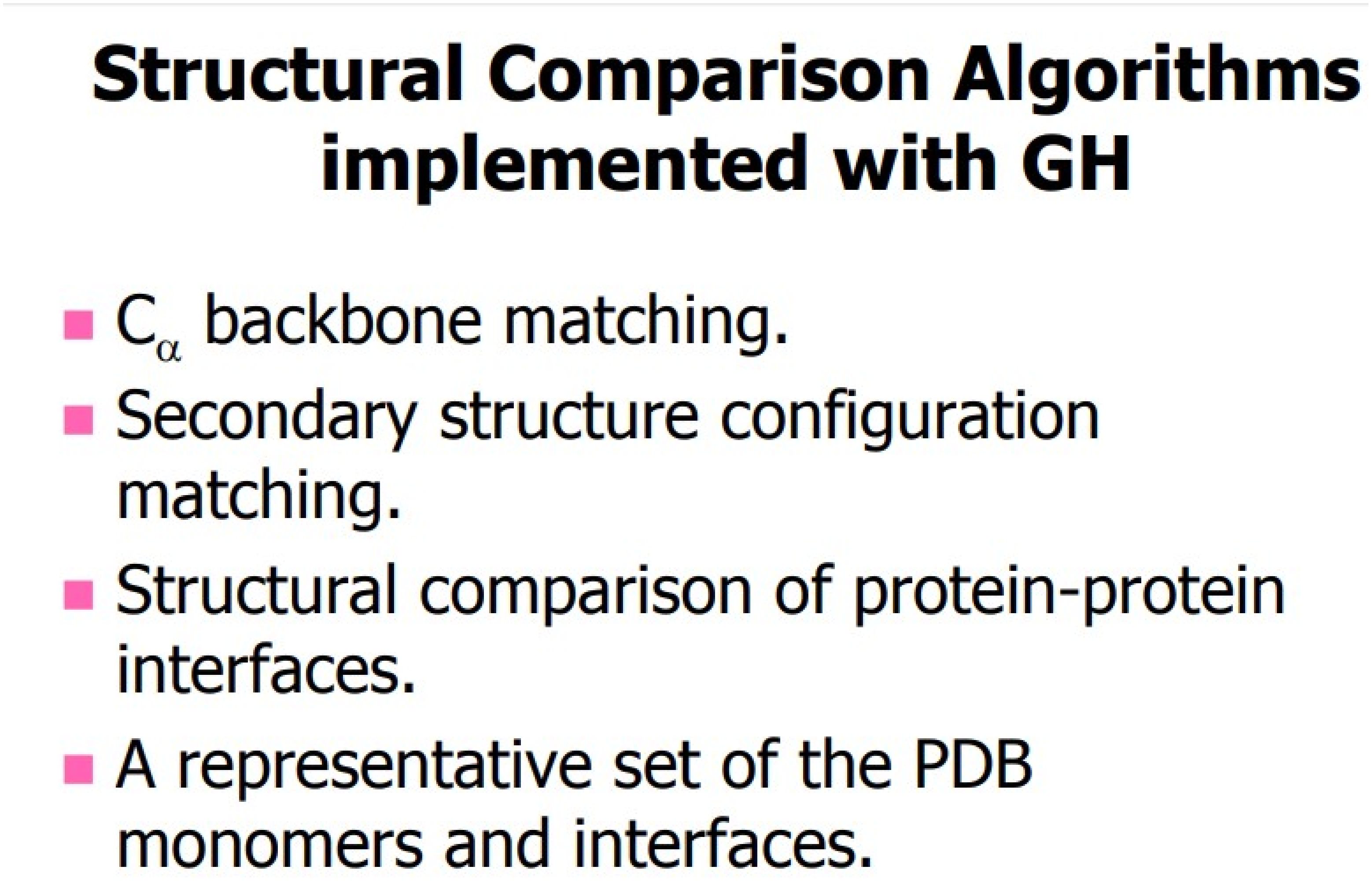
s - size of a hash-table entry. Can be kept low by not processing “fat” entries. These entries are known in advance after *Preprocessing*.

Alineamiento rigido de Proteinas

Advantages :

- Sequence order independent.
- Can match partial disconnected substructures.
- Pattern detection and recognition.
- Highly efficient.
- Can be applied to protein-protein interfaces, surface motif detection, docking.

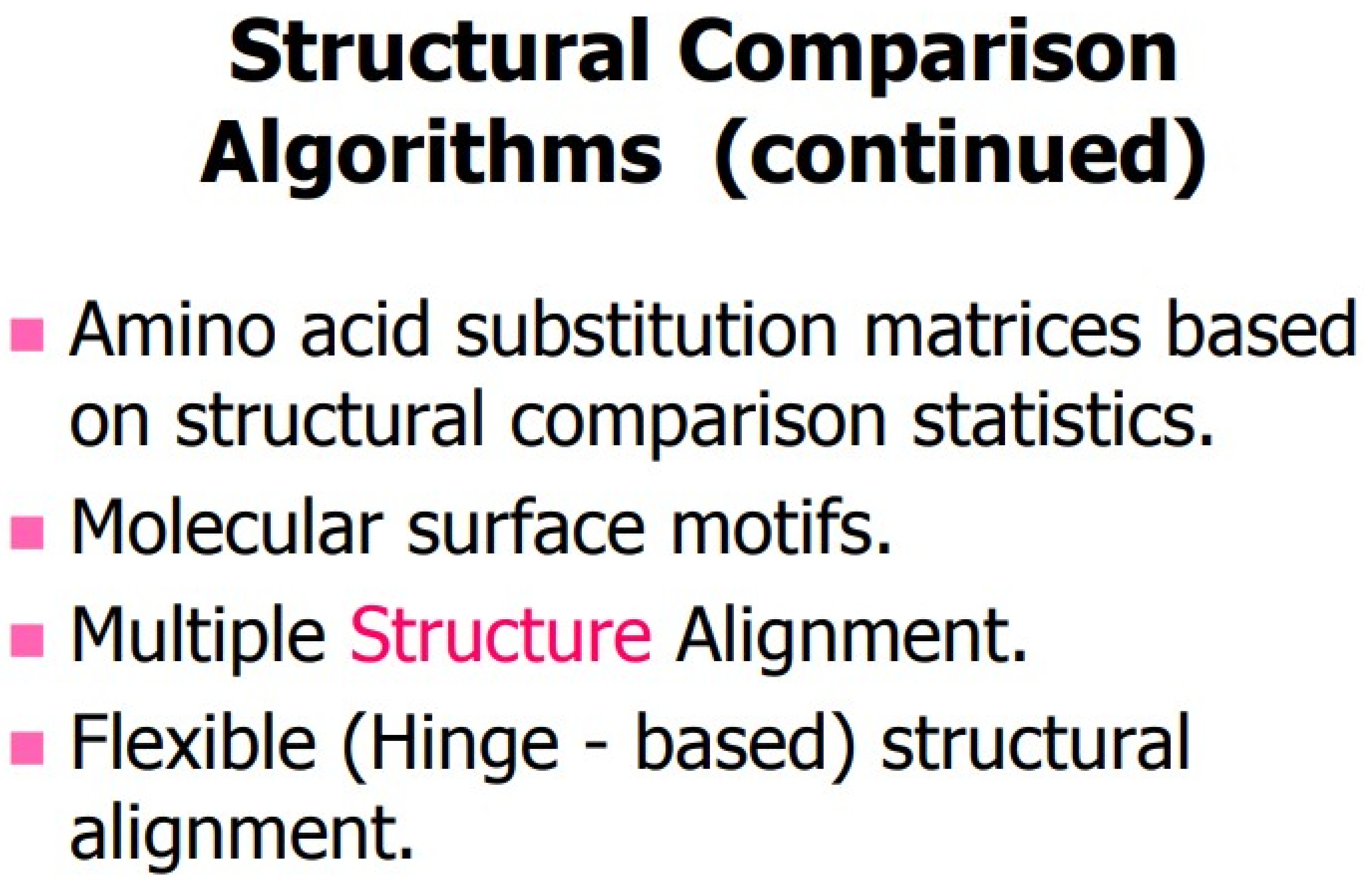
Alineamiento rigido de Proteinas



Structural Comparison Algorithms implemented with GH

- C_α backbone matching.
- Secondary structure configuration matching.
- Structural comparison of protein-protein interfaces.
- A representative set of the PDB monomers and interfaces.

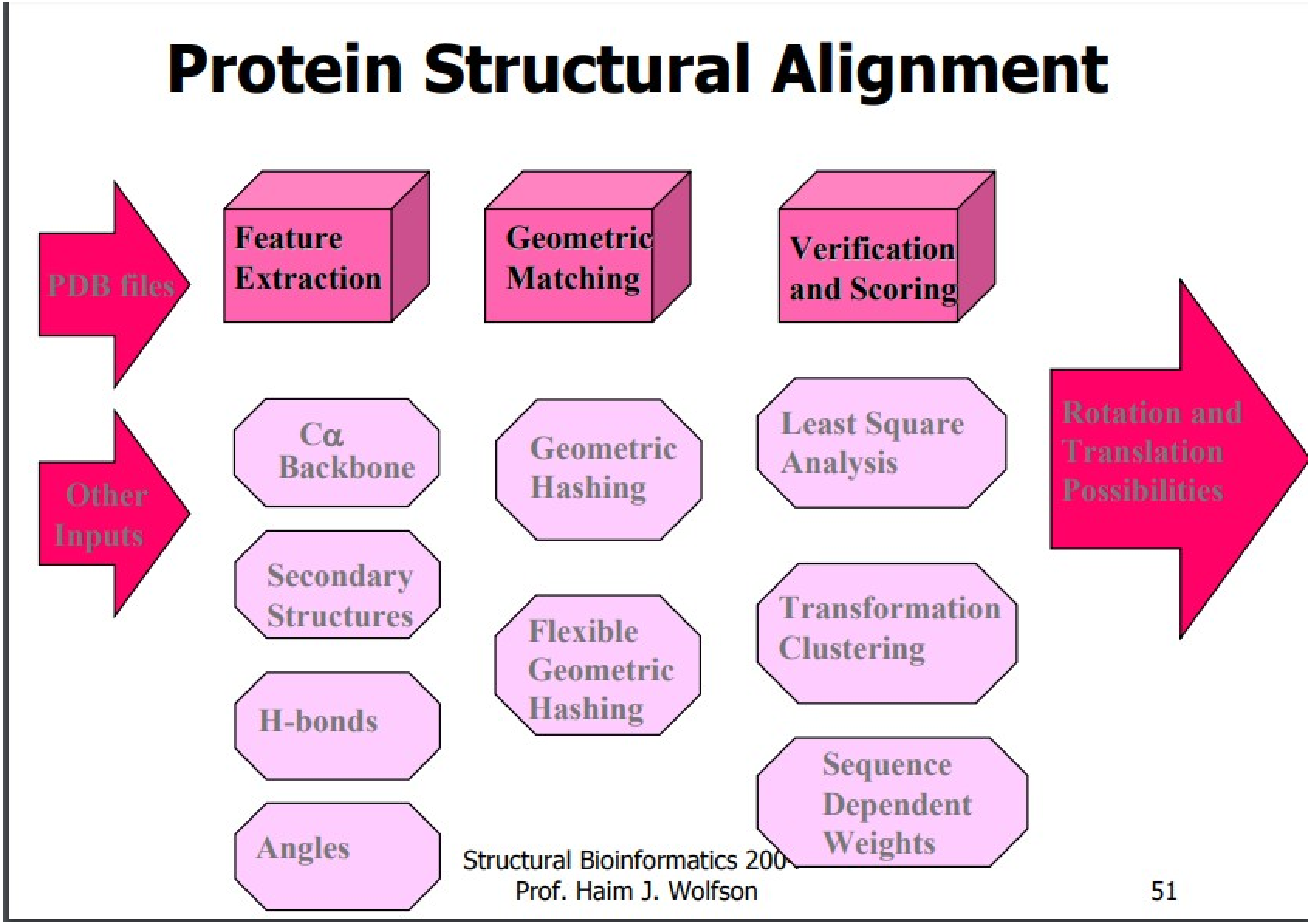
Alineamiento rigido de Proteinas



Structural Comparison Algorithms (continued)

- Amino acid substitution matrices based on structural comparison statistics.
- Molecular surface motifs.
- Multiple Structure Alignment.
- Flexible (Hinge - based) structural alignment.

Alineamiento rigido de Proteinas



Protein Structural Alignment

PDB files

Other Inputs

Feature Extraction

Cα Backbone

Secondary Structures

H-bonds

Angles

Geometric Matching

Geometric Hashing

Flexible Geometric Hashing

Verification and Scoring

Least Square Analysis

Transformation Clustering

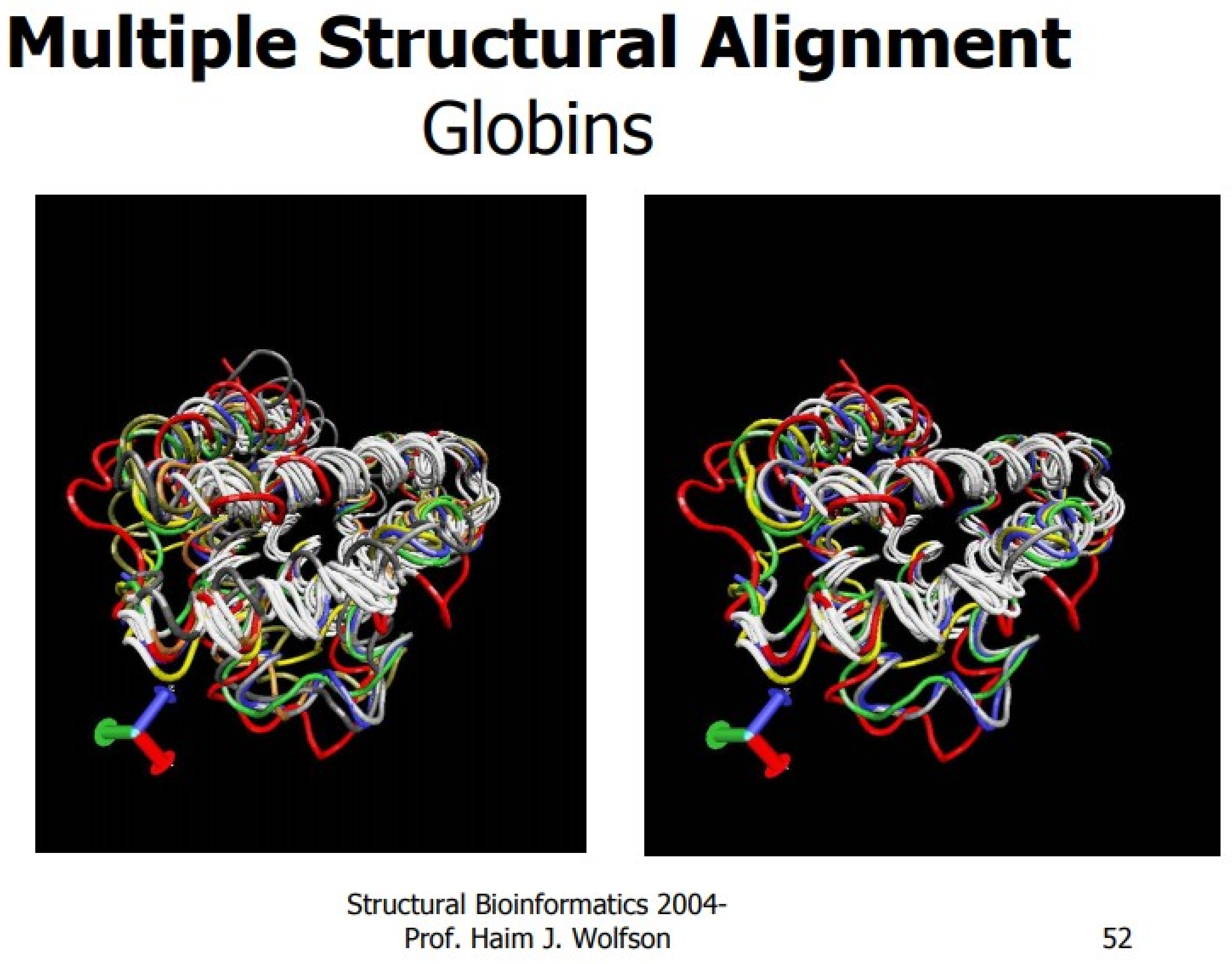
Sequence Dependent Weights

Rotation and Translation Possibilities

Structural Bioinformatics 200
Prof. Haim J. Wolfson

51

Alineamiento rigido de Proteinas



Multiple Structural Alignment

Globins

Structural Bioinformatics 2004-
Prof. Haim J. Wolfson

3TA6
7TIM