DALI

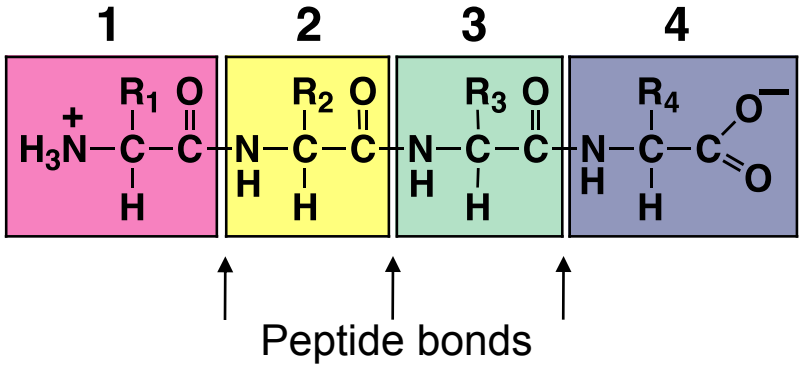
Structural Comparisons by use of Distance Matrix Alignment

Mark Maciejewski
markm@uchc.edu
Nov 15th, 2016

Protein Structure - Primary Sequence

DNA ---> RNA ---> Protein

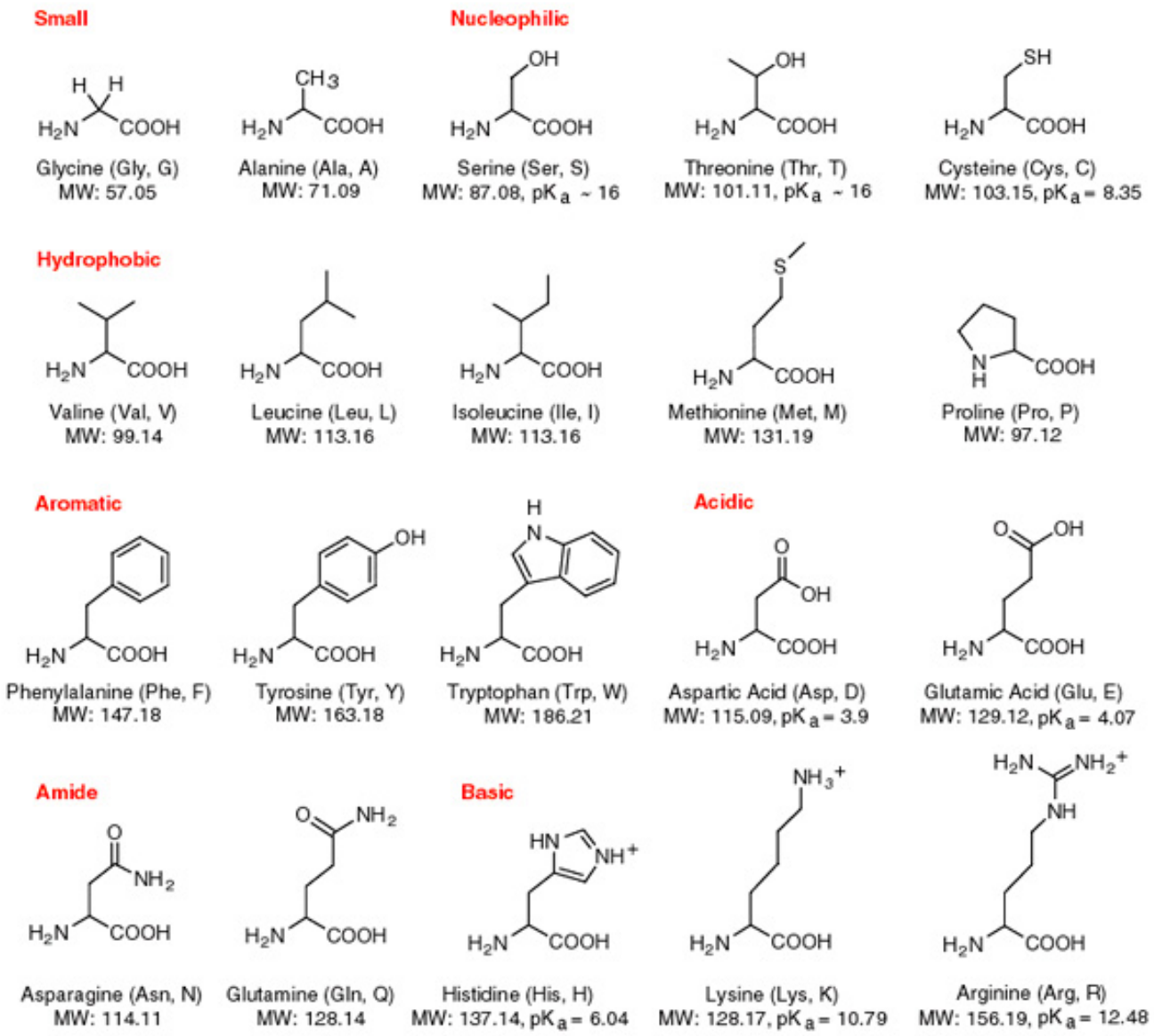
Proteins are composed of long linear chains of amino acids.



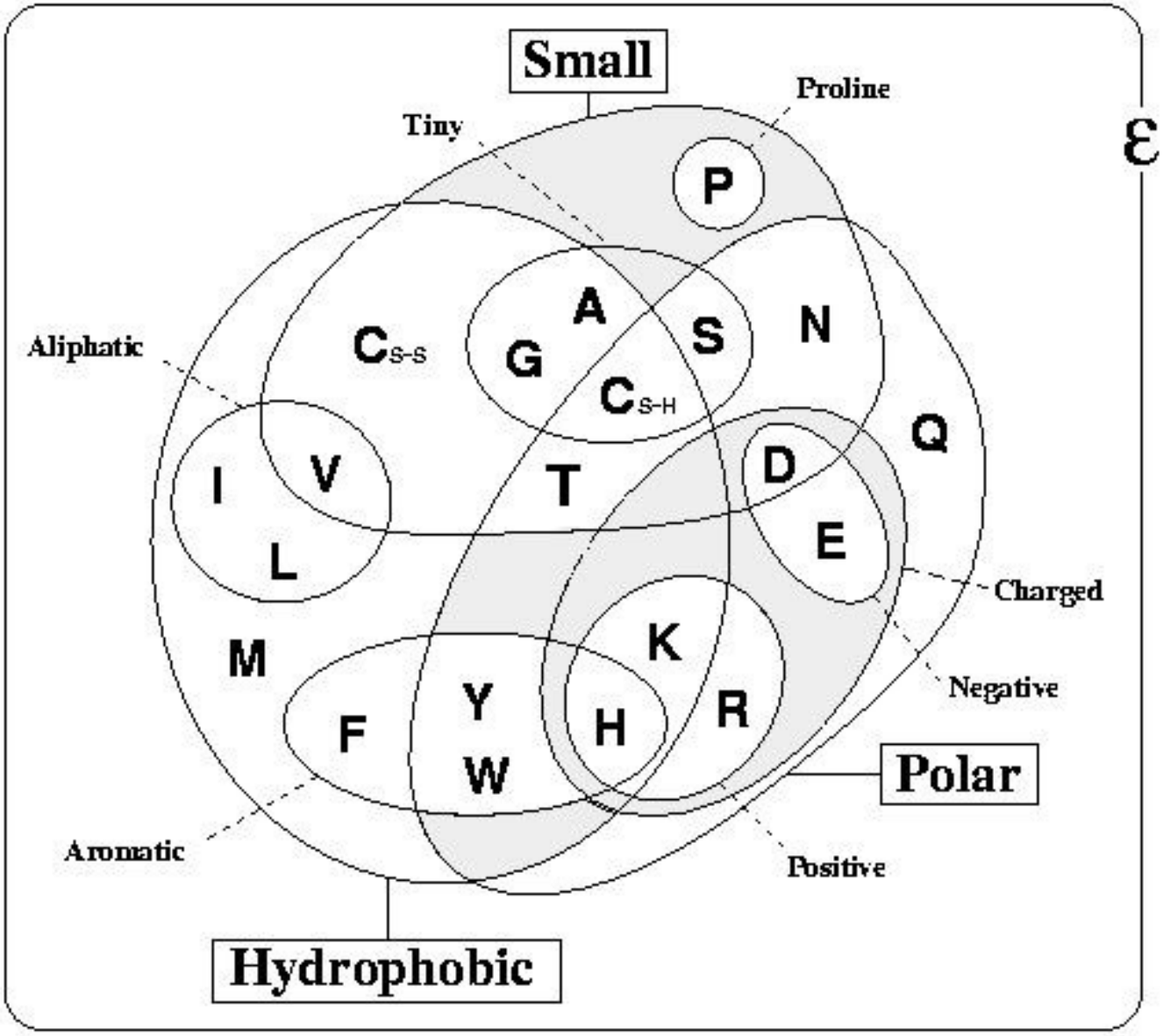
where R = any of 20 amino acids.

Proteins are the workhorse of cells

Protein Structure - Amino Acids



Protein Structure - Amino Acids



Four Levels of Protein Structure

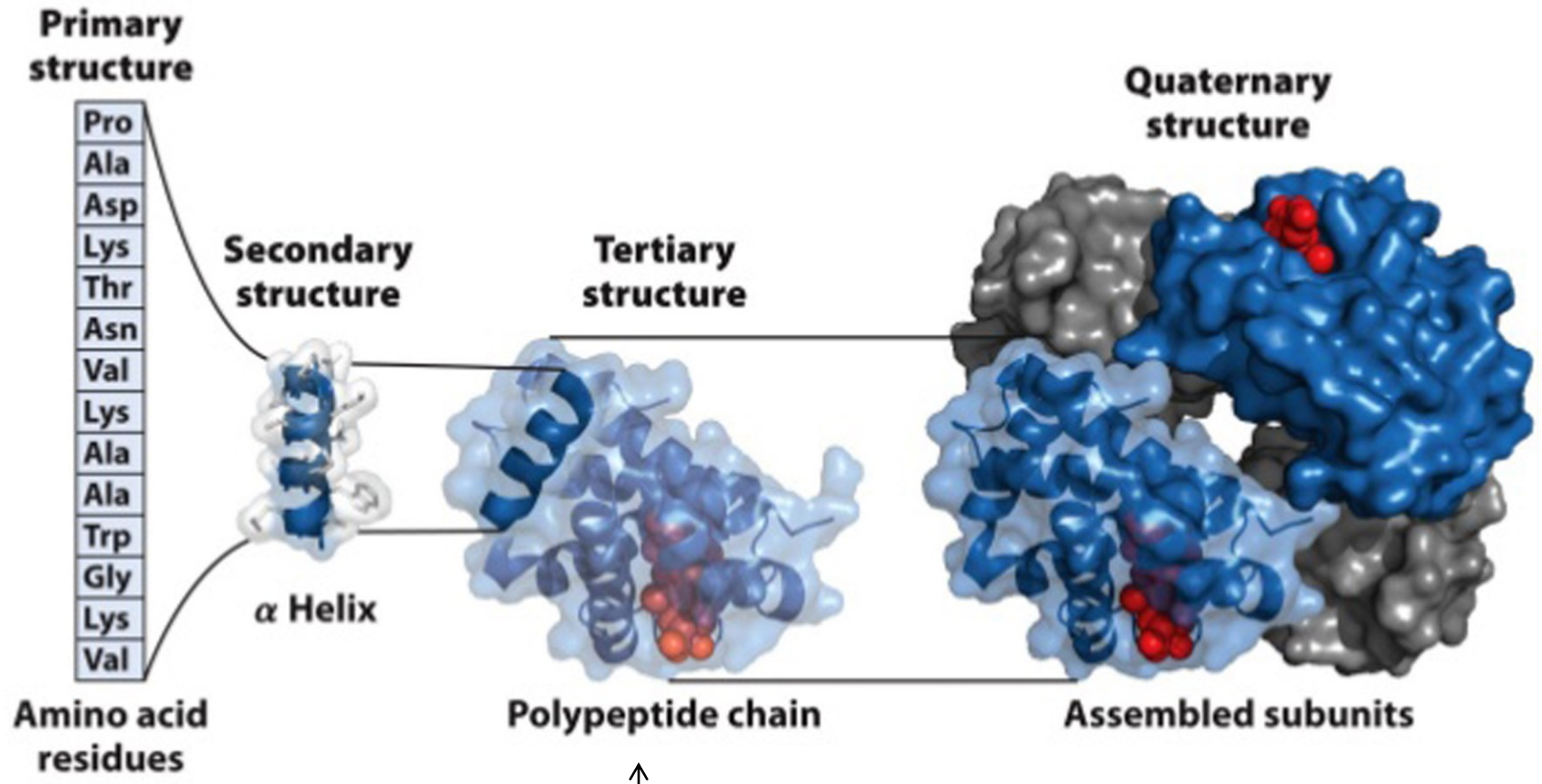
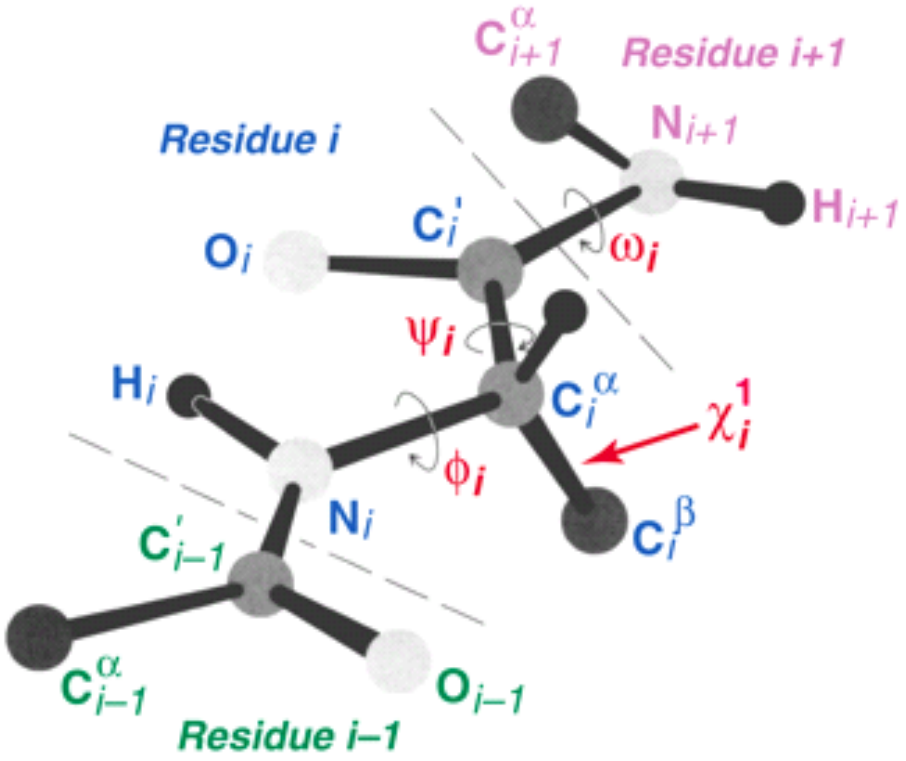


Figure 3-23
Lehninger Principles of Biochemistry, Sixth Edition

>90% Eukaryotic proteins are multi-domain

Rotatable Protein Angles

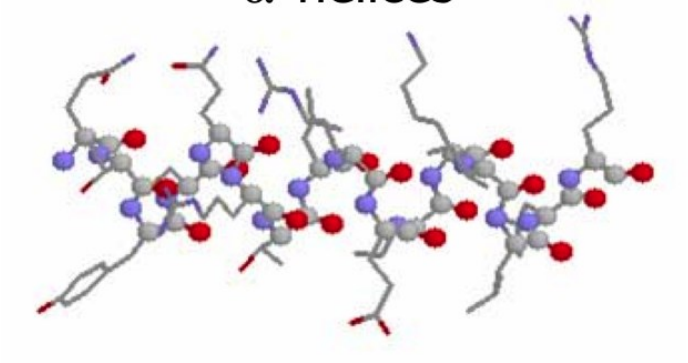


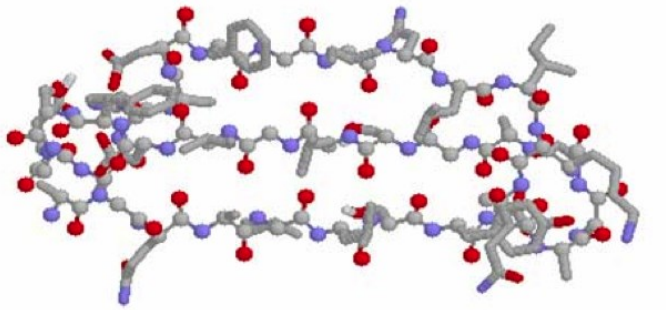
- phi (Φ) and psi (Ψ) angles contain all the information needed to define the backbone chain of the protein.
- chi (χ^n) angles are needed to define side chain orientation.
- The omega (ω)angle is locked at 0 or 180 degrees due to double bond character of the bond (non-rotatable)
- Only phi / psi angles that prevent the side chains from avoiding steric clashes are allowed
- Carbon alpha residue at the center

Secondary Structure

α-helices

β-strands





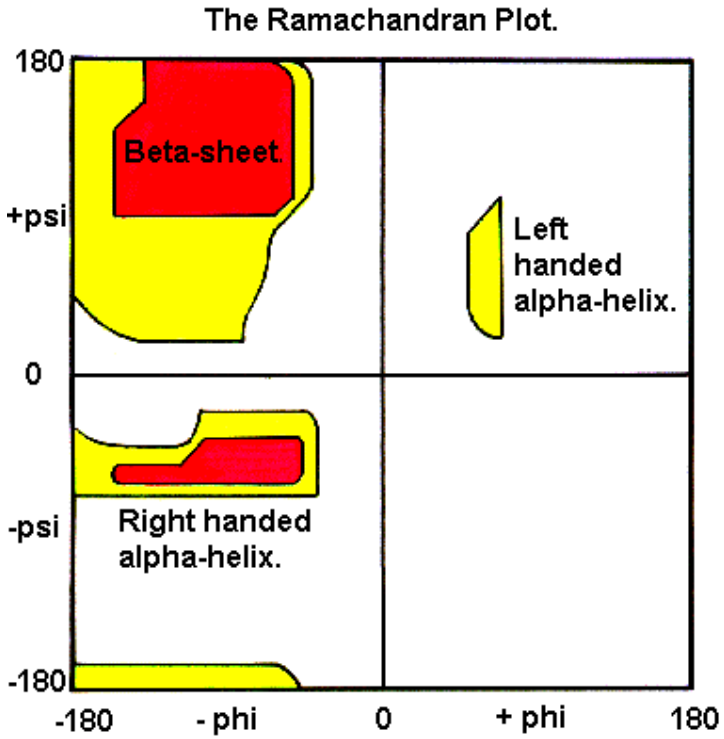
Protein structures are composed of α-helices, β sheets, turns, and loops.

Secondary structure elements occur because their *phi* and *psi* angles, which define the backbone conformation, position side-chains to reduce steric clashes.

Helices have 3.6 residues per turn and have side chains pointing outward.

Strands have side chains alternating in and out of beta sheet plane.

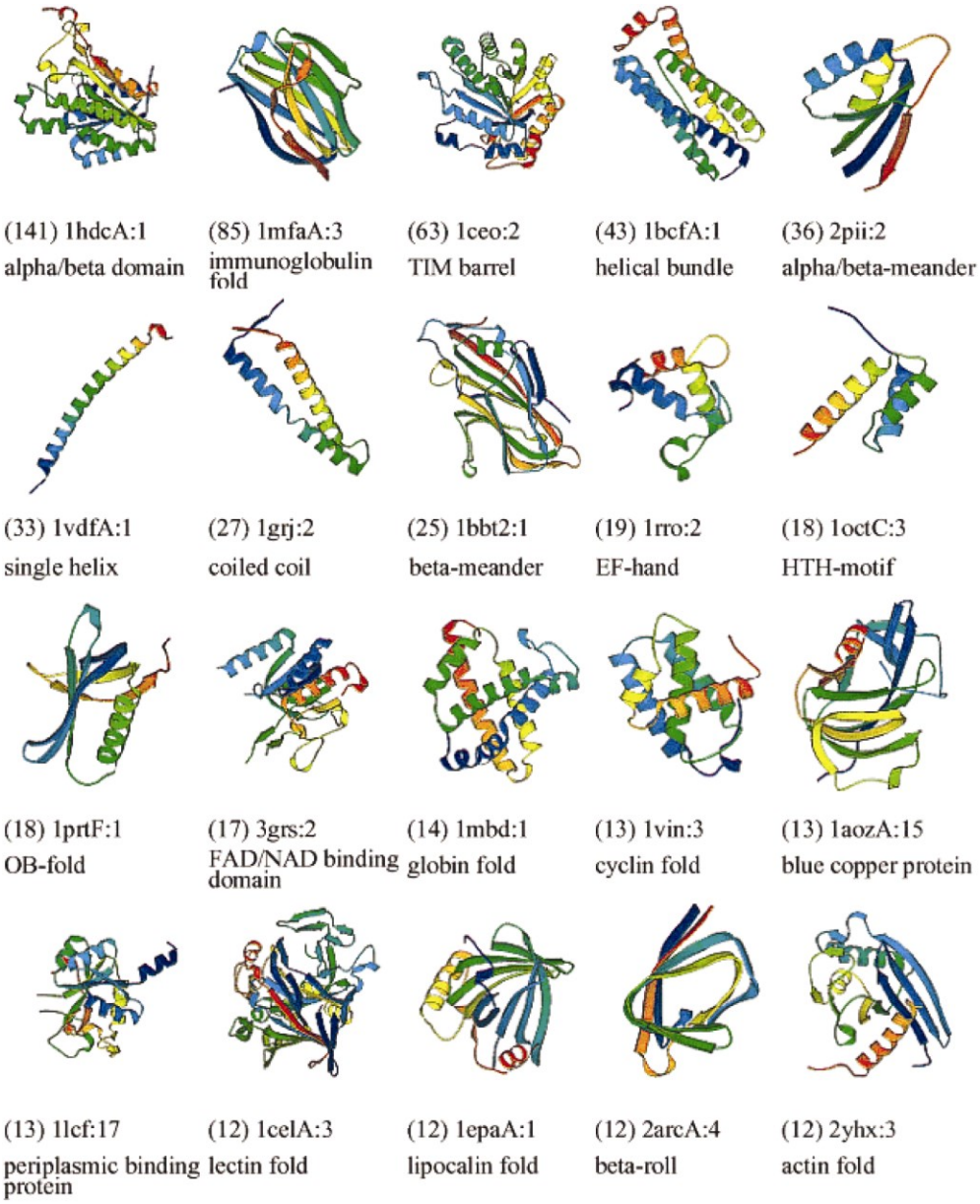
Ramachandran Plot



Ramachandran plot showing the allowed (yellow) and most favorable (red) combinations of *phi* and *psi* angles. The white region is mostly disallowed due to steric clashes between side-chain residues.

Large bulky side-chains such as Leucine have a smaller allowed region of Ramachandran space, whereas small residues such as Glycine have a much larger region of allowed space.

Proteins can fold into a variety of compact 3D shapes



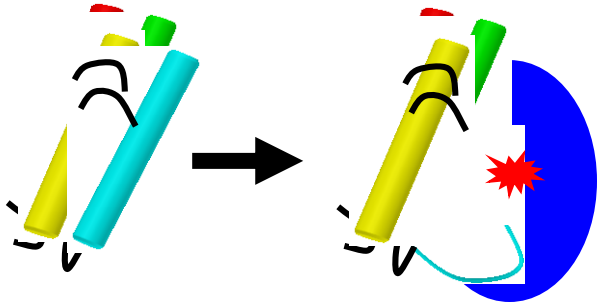
3D structure is required for function

Dynamics also important

Proteins adopt the most stable structure possible (not considering molecular assemblies and aggregates)

Protein structures are only marginally stable

Even a single mutation can cause loss of structure/function

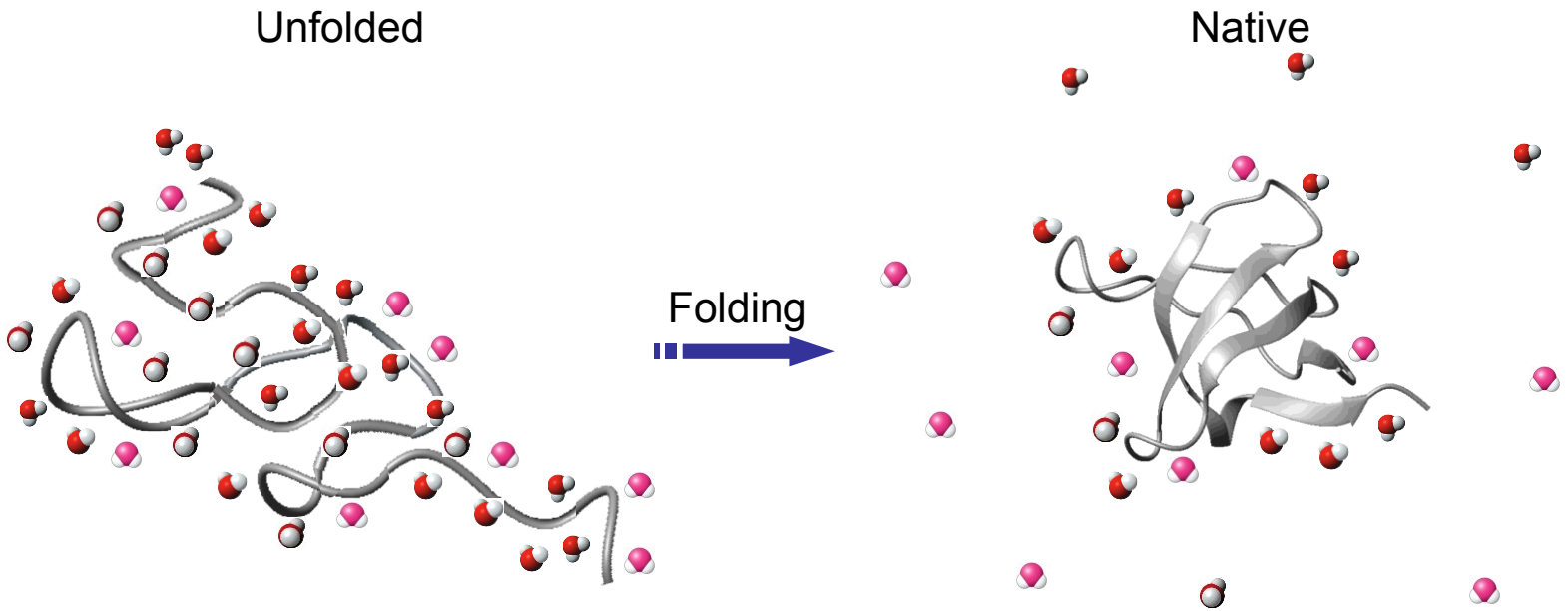


Protein Folding

Consider a protein that exists in two states: Native (**N**) and Unfolded (**U**)

"Driving force" of protein folding is the free energy difference between the **N** and **U** states

$$\Delta G_{N-U} = G_N - G_U < 0$$



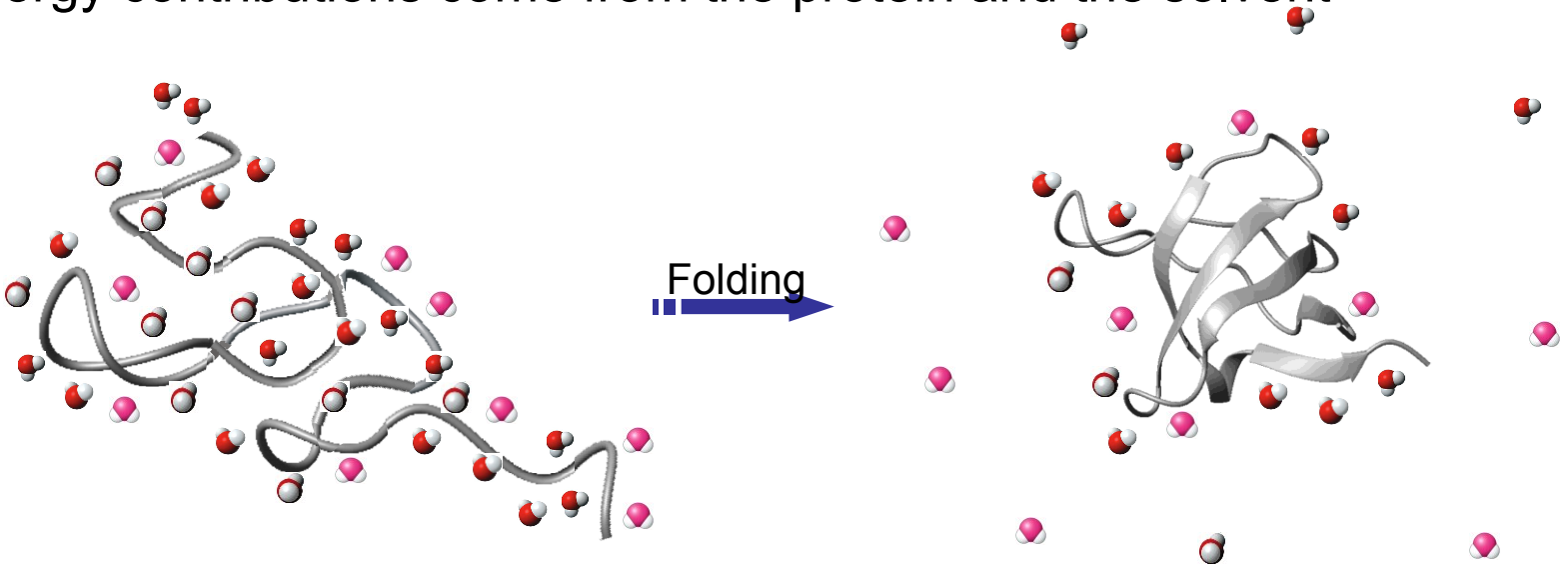
$$\Delta G_{N-U} = -5 \text{ to } -15 \text{ kcal/mol}$$

ΔG_{N-U} consists of large mutually compensating contributions

Proteins are only marginally stable!

Protein Folding

Energy contributions come from the protein and the solvent



Unfolded (significantly hydrated)

- Favorable entropy (protein)
  - Protein is flexible and highly dynamic
- Unfavorable enthalpy (protein)
  - Dynamic nature does not allow significant electrostatic, van der Waals, or hydrogen bonds interactions to form
- Unfavorable entropy (solvent)
  - Solvent is immobilized in clathrate cages around hydrophobic residues which maximize enthalpy between solvent and protein and maximize water hydrogen bonds at the expense of entropy

Folded (less hydrated)

- Unfavorable entropy (protein)
  - Protein is rigid
- Favorable enthalpy (protein)
  - Significant electrostatic, van der Waals, and hydrogen bonds interactions form
- Favorable entropy (solvent)
  - Solvent is released from clathrate cages increasing entropy with a slight reduction in enthalpy

Energetic Contributions to Folding

$$\Delta G_{N\text{-}U} = \Delta H_{N\text{-}U} - T\Delta S_{N\text{-}U}$$ Folding Free Energy

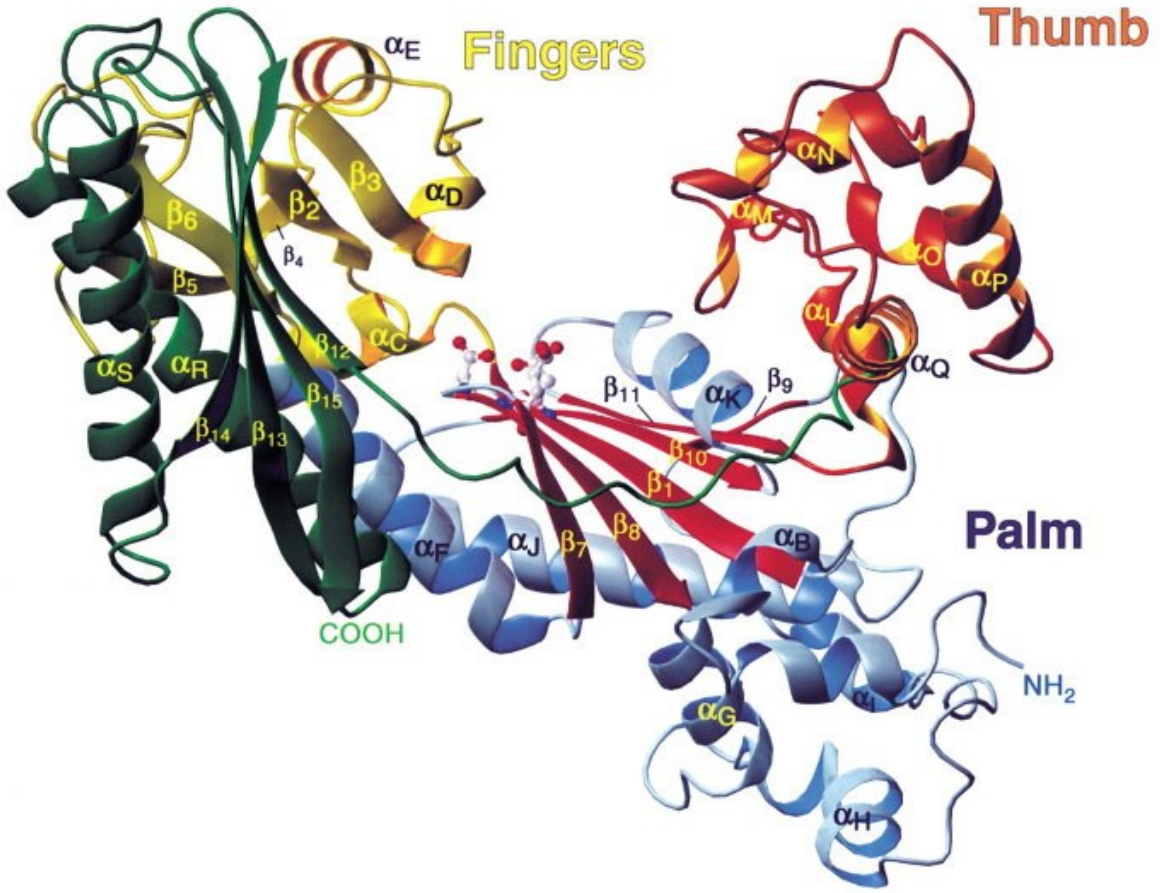
$\Delta H_{N\text{-}U}$ and $\Delta S_{N\text{-}U}$ include contributions from both the protein chain and the solvent:

$$\Delta G_{N\text{-}U} = \{ \Delta H_{N\text{-}U(protein)} - T\Delta S_{N\text{-}U(protein)} \} + \{ \Delta H_{N\text{-}U(solvent)} - T\Delta S_{N\text{-}U(solvent)} \}$$

For a 100 residue protein at room temperature:

$-T\Delta S_{protein}$	+160 kcal/mol	entropy loss upon folding (unfavorable)
$\Delta H_{protein}$	-80 kcal/mol	enthalpic interactions in native state (favorable) (electrostatics, van der Waals, hydrogen bonds)
ΔH_{sovent}-$T\Delta S_{solvent}$	-90 kcal/mol	hydrophobic effect (favorable) Dominated by increase entropy of free water
Total $\Delta G_{N\text{-}U}$	-10 kcal/mol	

Multi Domain Proteins



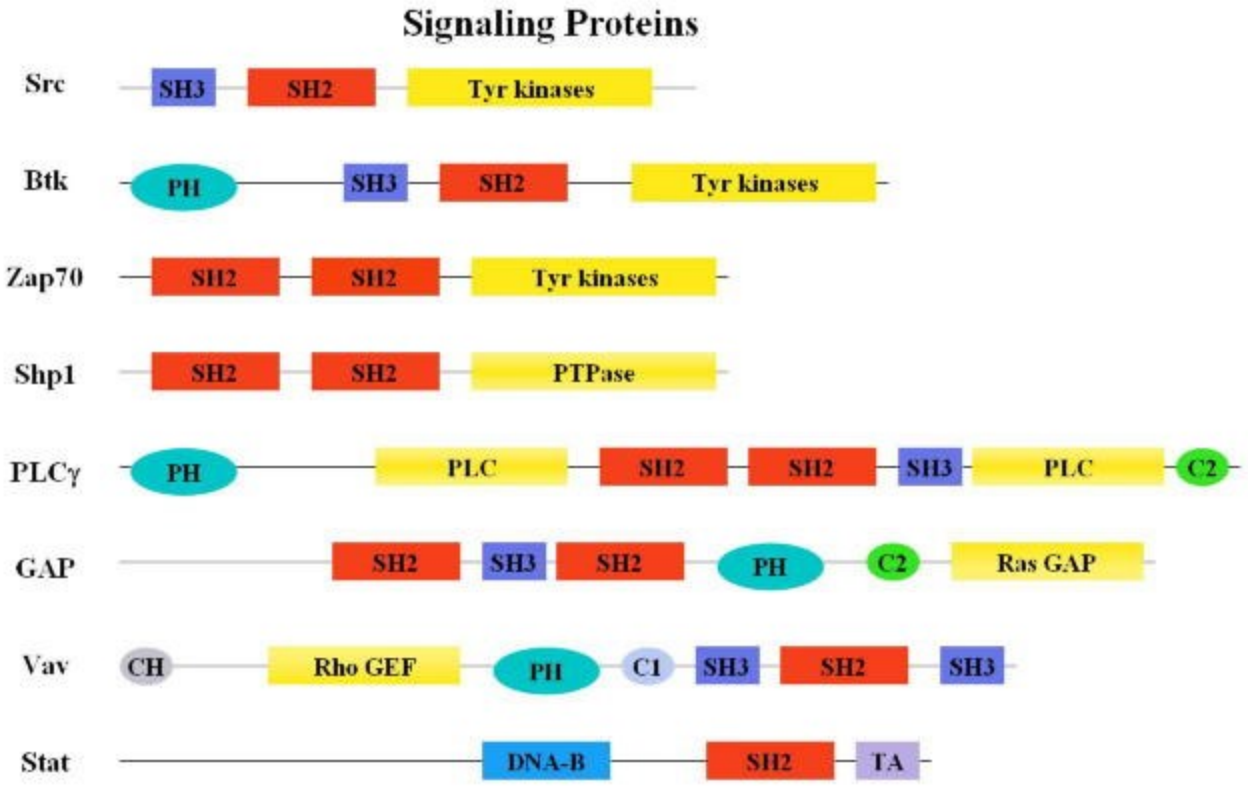
Protein structures are often divided into domains (~17 kDa average domain size).

Inter-domain interactions are often weaker than intra-domain interactions.

Domains will often fold independently

Higher organisms have a larger (> 90%) of multi-domain proteins.

Protein Modules



Domains (Modules) conserved through evolution are “mixed and matched” to form a wide variety of proteins.

The concept of modules is important as modules that are similar in structure are often functionally similar. This means that if we have a structure of a protein, but do not know its function, we can make a guess of its function based on similarity to other protein structures (modules) whose function is known.

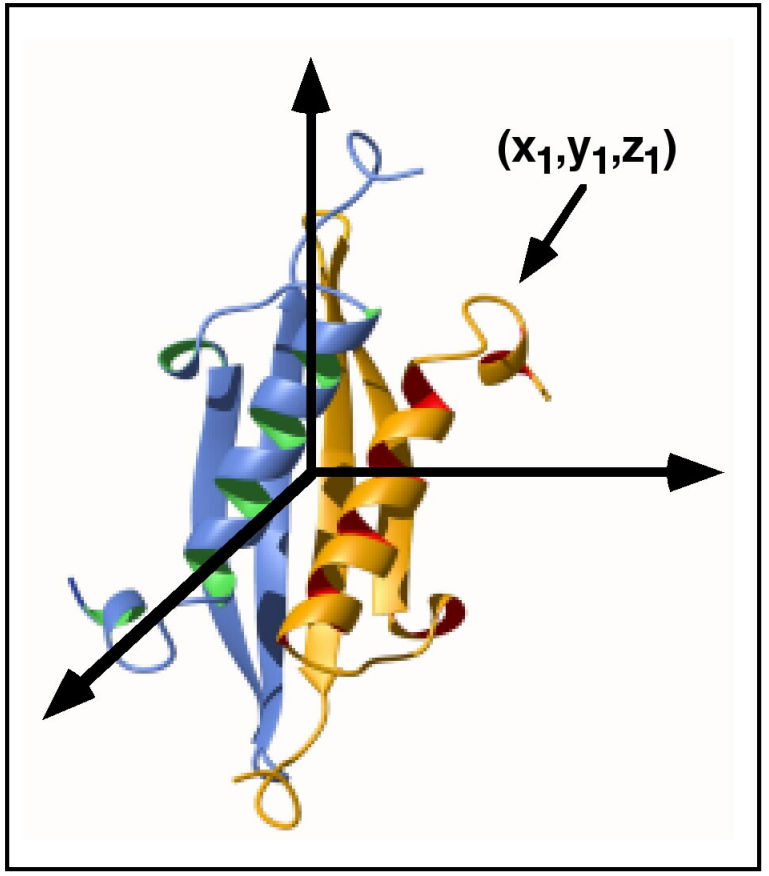
Why do Structural Comparison?

- To compare the same molecule under different conditions to find regions which are likely functionally important.
- Understand how protein structures have evolved.
- To help understand how different primary sequences can give rise to similar folds, which may be helpful in protein structure prediction.
- To aid in protein engineering.
- **To find proteins with similar folds which may lead to a functional prediction. Proteins with similar function often have high structural conservation.**
    - May validate Blast predictions or may be used where no primary sequence alignment exists.
    - May not divulge full biological activity, but may give insight by predicting binding sites of small molecules or macromolecules.
    - May only find similarities at the domain / module level.

Different ways to represent protein structures

Cartesian coordinates

Arrange protein on an imaginary Cartesian coordinate frame and assign (x,y,z) coordinates to each atom (the method using by the Protein Data Bank)



Different ways to represent protein structures

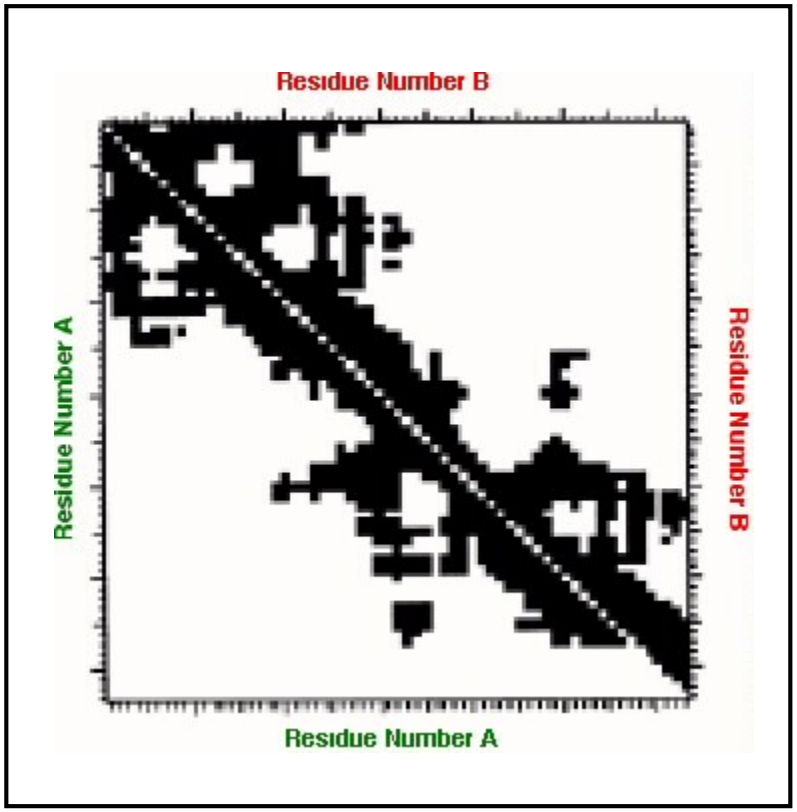
Two-dimensional distance matrices

Construct a matrix of interatomic distances (e.g., Cα–Cα distances) in the protein.

Two proteins are shown and are not the same length

Cα–Cα distances are useful as they are the center of the AA backbone and are invariant for the 20 amino acids.

In this example Cα-Cα distances that are close in space are highlighted in black



The Goal…

Develop a method that can find optimal matches between the shape of a given protein to all other proteins of known structure.

But how?

- Root Mean Square Deviation? (Cartesian coordinates)
- Through the use of distance matrices?

Two important criteria

- We need a quantitative metric for comparisons
- Need to deal with different size proteins, variable loop lengths, different topologies, and multi-domain proteins where one domain may be structurally similar but other domains are not.

Root Mean Square Deviation (RMSD)

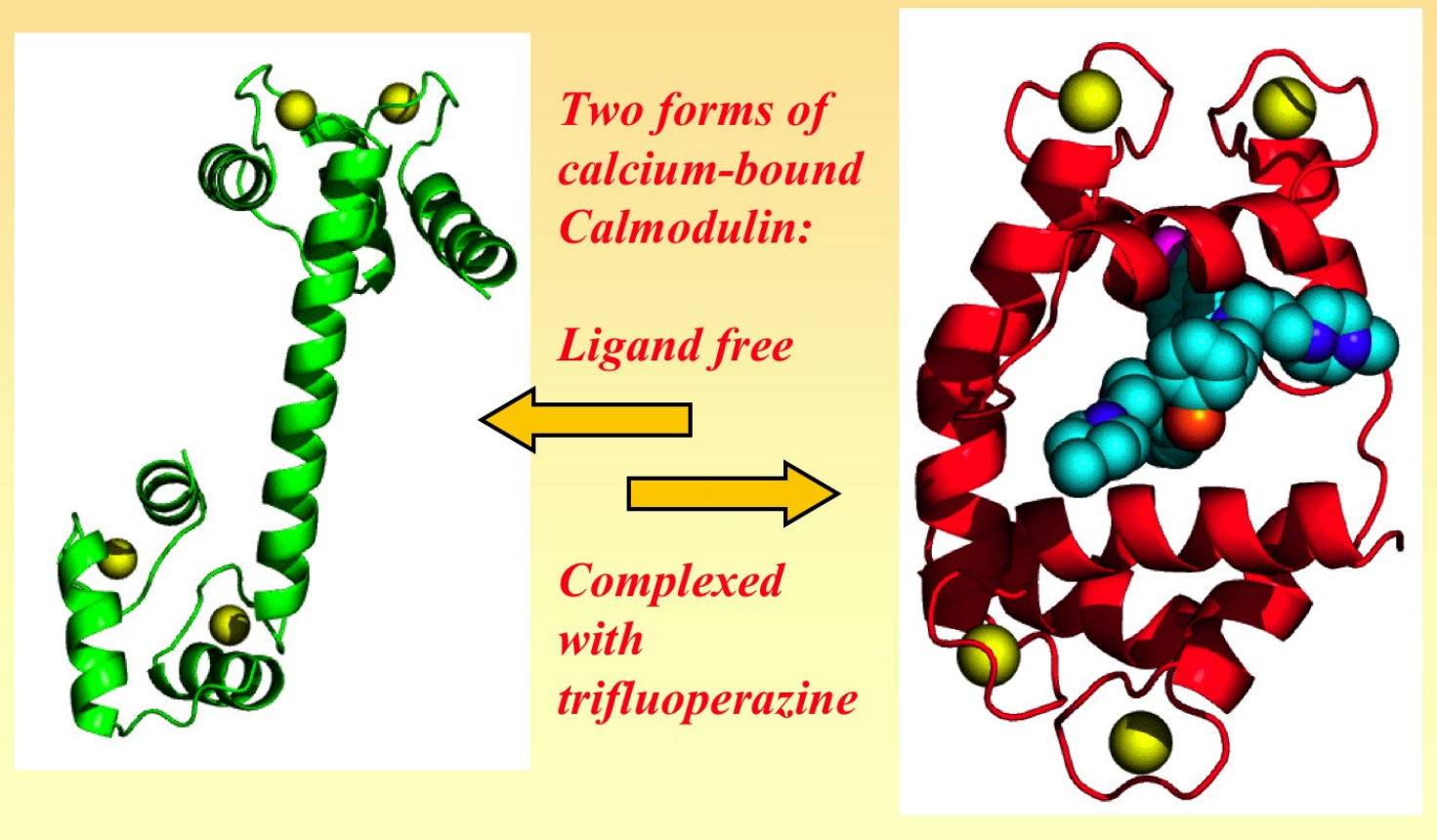
Proteins in the PDB are stored as Cartesian coordinates *(xyz)*. For any two proteins you can find the translation and rotation which will give the best overall alignment of the two molecules.

The RMSD can then be easily calculated by the following equation:

$$RMSD(A,B) = \sqrt{\frac{1}{N}\sum_{i=1}^{N} d(a_i,b_i)^2}$$

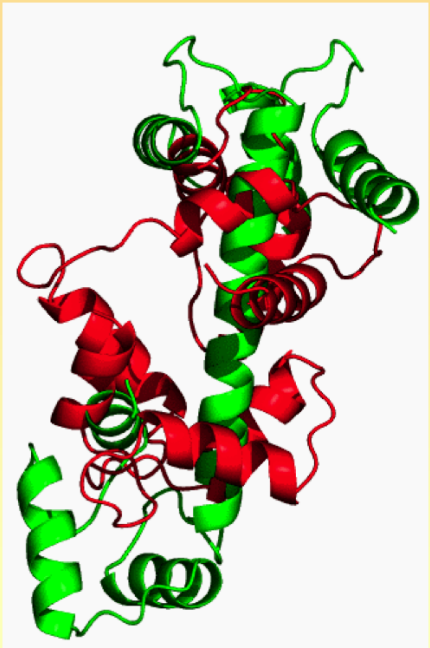
where *N* is the number of atoms, *A* and *B* are the two protein structures, and $d(a_i,b_i)$ is the difference in atomic positions between each of the atoms from the two proteins after they have been rotated and translated to give the best overall alignment.

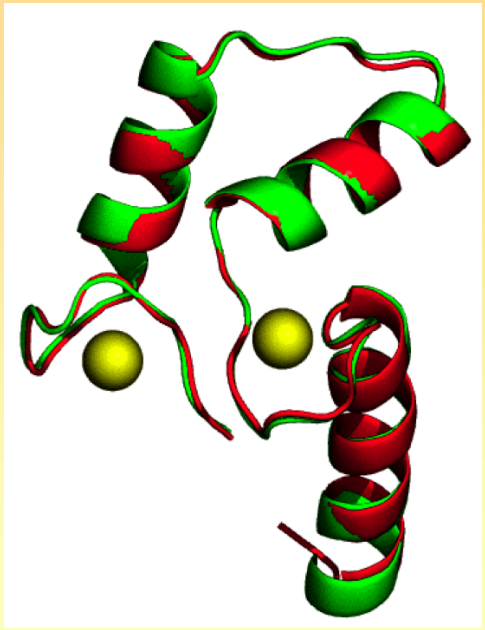
RMSD - Global Versus Local (1)



Calmodulin undergoes a dramatic conformational change of its central helix upon binding to its ligand.

RMSD - Global versus Local (2)





Global alignment:
RMSD =15 Å /143 residues

Local alignment:
RMSD = 0.9 Å/ 62 residues

Global alignment of calmodulin in its bound and free state gives rise to a 15 Å RMSD.

The RMSD is 0.9 Å when only the ends of calmodulin are aligned.

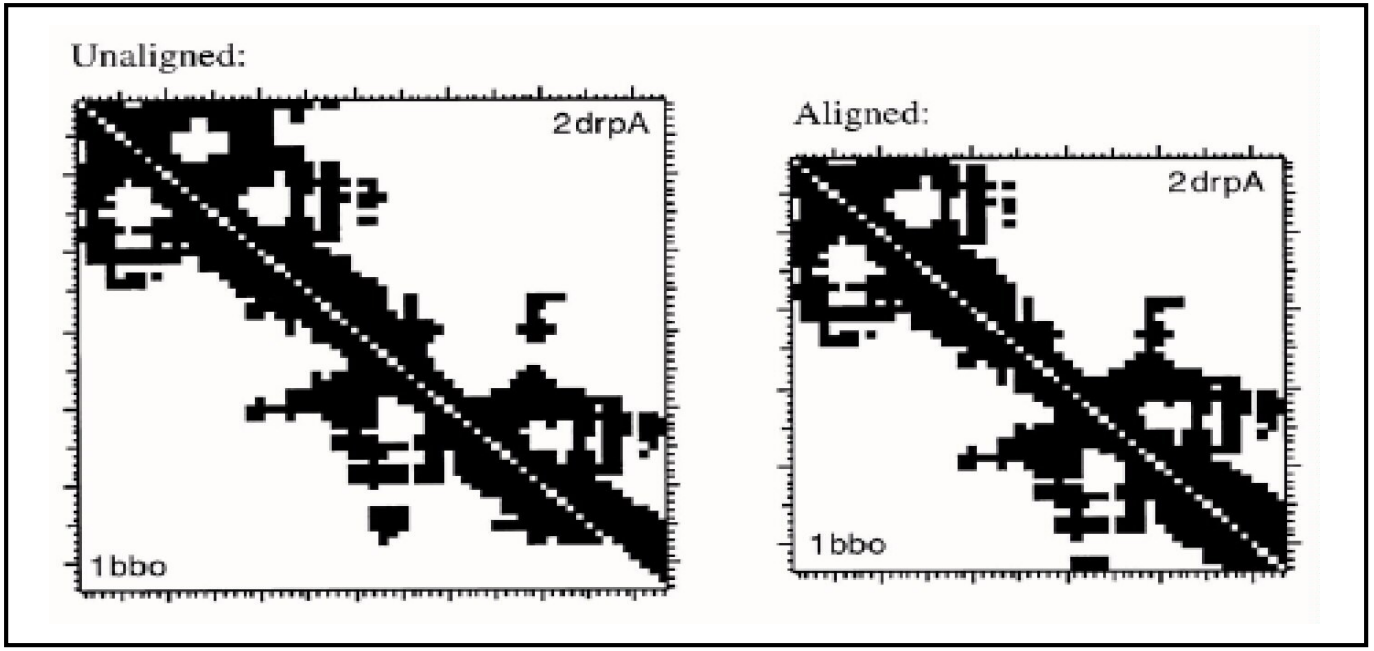
RMSD Not Best Choice

Finding the matching residues to compare to utilize RMSD as a structural comparison is a daunting task and NOT computationally feasible.

RMSD value is not a good metric to define the similarity between two proteins.

RMSD is best suited for comparing “identical” structures such as an NMR bundle or in cases where the residues to use in the alignment are clear.

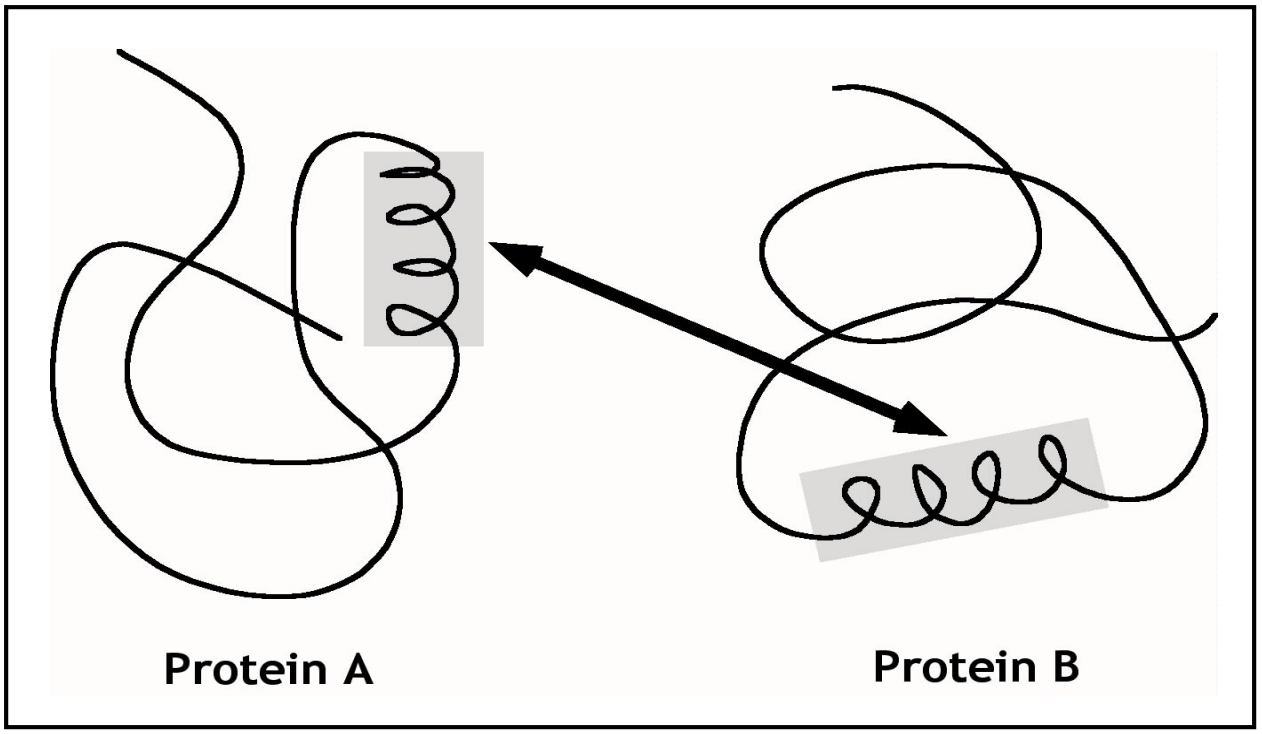
Distance Matrix - Comparison in 2D Space



A Distance Matrix has four advantages as a structural comparison tool:

1. Invariant with respect to rotation and translation.
2. Represents both local and long range structure.
3. Easily adapts to insertions and deletions.
4. Generates a scoring metric of structural similarity.

The problem …

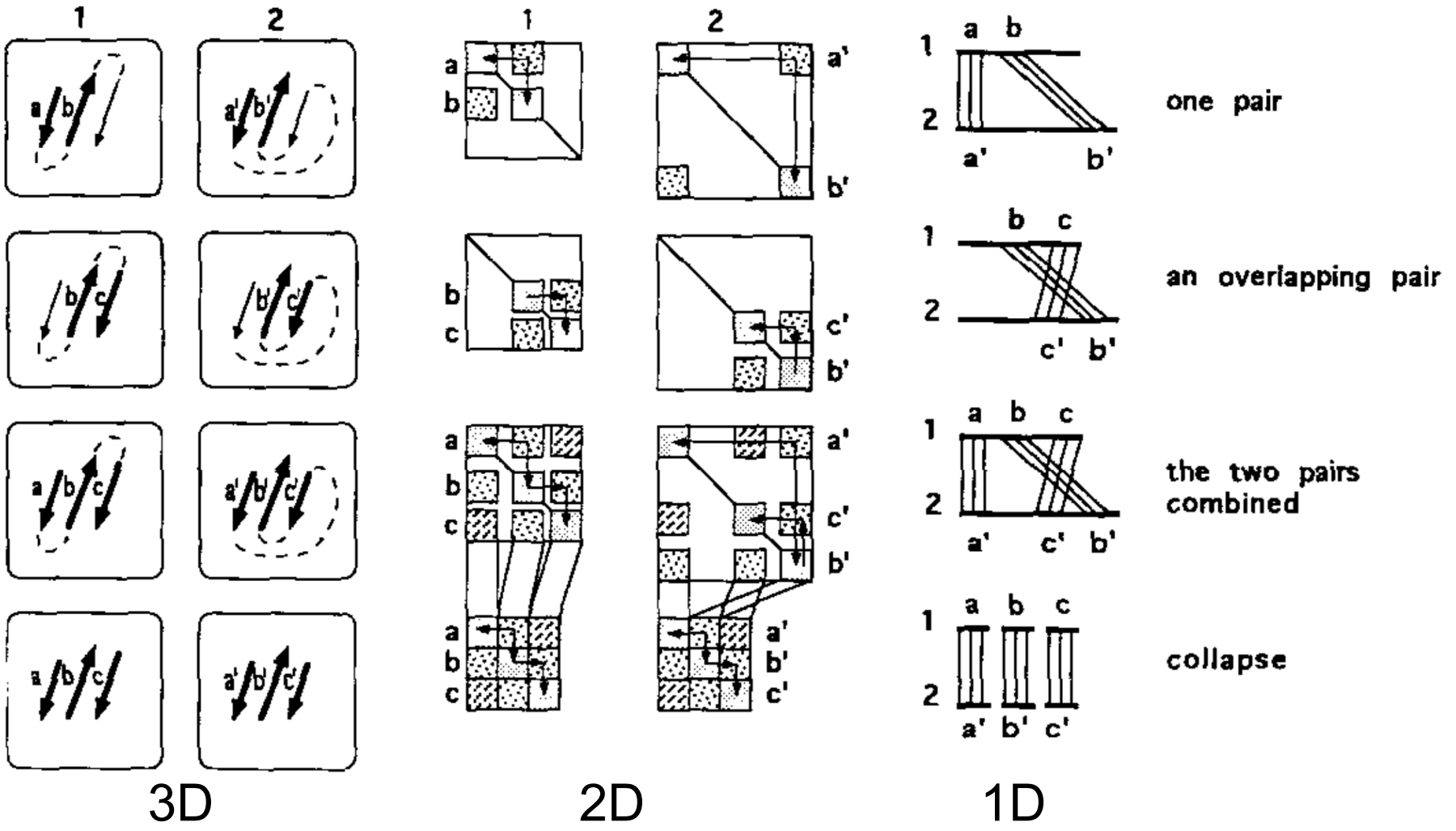


How do we utilize a distance matrix to determine if there is a 3D match between two substructures in the proteins being compared?

DALI answer is to:

Set up distance matrices to describe each protein and then formulate a method of quantitatively comparing the matrices (i.e., devise some kind of similarity score).

Schematic Representation of DALI



Schematic representation of how two topologically different three beta strands can be found to be similar with distance matrices.

The Answer According to DALI - Distance Matrix Alignment

Formulation of the problem

Consider two proteins **A** and **B**. The match of two substructures within each of these proteins can be evaluated using an additive similarity score of the form:

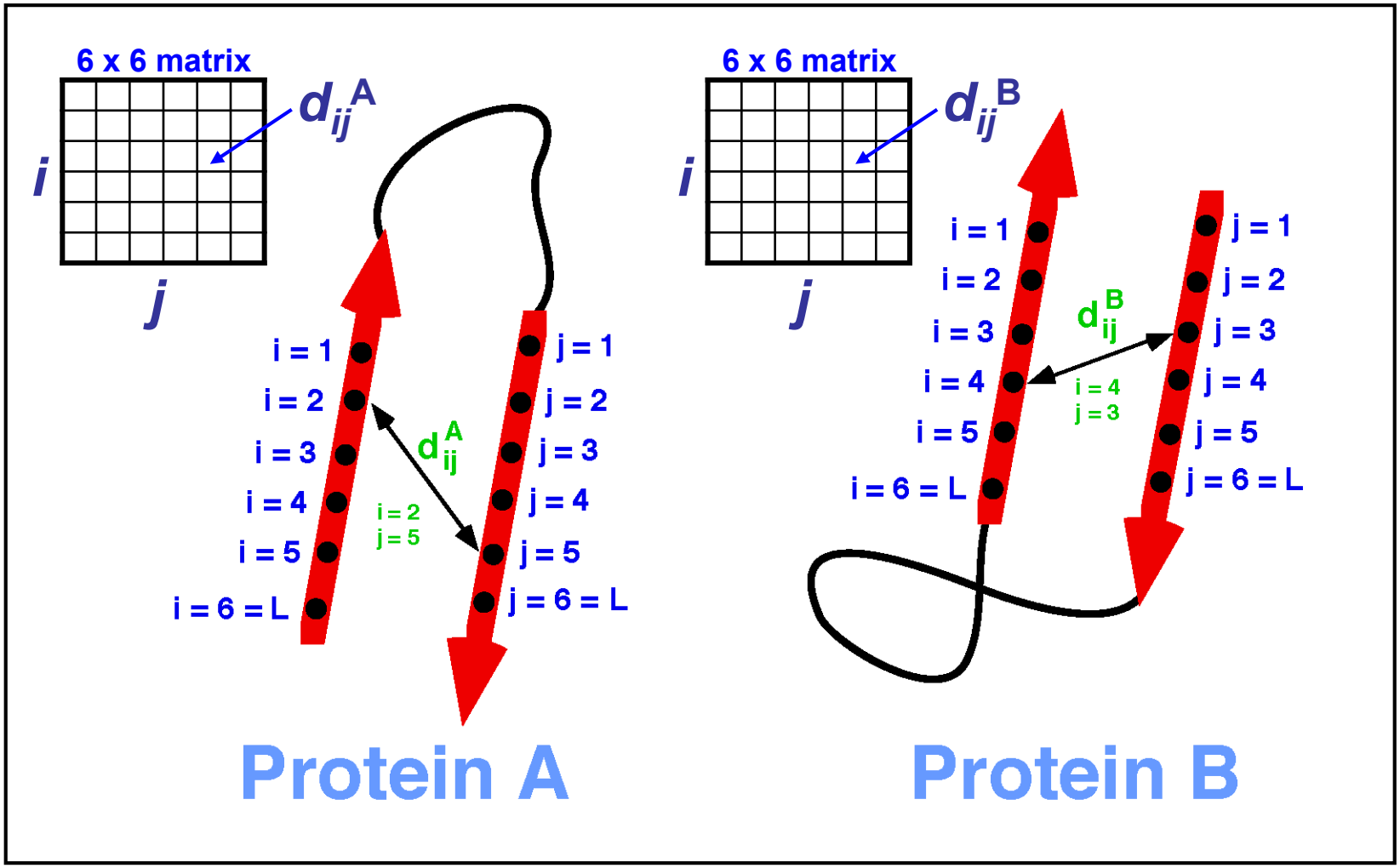
$$S = \sum_{i=1}^{L} \sum_{j=1}^{L} \phi(i,j)$$

where **i** and **j** label pairs of equivalent (matched) residues

L is the number of such pairs (i.e., the size of the substructure)

ϕ is a similarity measure based on the C_α–C_α distances d_{ij}^{A} and d_{ij}^{B}

The Answer According to DALI - Distance Matrix Alignment



Note that each 6x6 distance matrix is for distances inside the same protein.

6x6 Carbon α Distance Matrix (1)

Protein A

98	6.2	8.2	10.8	14.2	15.8	19.5
99	5	5.5	7.3	10.5	12	15.7
100	6.2	4.4	6.1	8.5	10.3	13.7
101	8.6	5.5	4.5	5.4	6.7	9.9
102	11.6	8	6.4	4.9	5.9	7.9
103	14.9	11.4	9	6.5	5.1	5.7
	50	**51**	**52**	**53**	**54**	**55**

Difference in (*i,j*) distances (d_{ij}^{A}) between region 1 and region 2 for protein A

Protein B

86	12.7	10.6	7.3	5.8	4.9	5.3
87	10.6	8.4	6.1	5	6.6	7.4
88	6.8	5.3	4.2	5.6	8.4	10.4
89	5.2	5.3	6.8	8.8	12	14.1
90	3.8	6.5	8.7	11.7	14.8	17.3
91	5.4	8.4	11.5	14.5	17.9	20.5
	76	**77**	**78**	**79**	**80**	**81**

Difference in (*i,j*) distances (d_{ij}^{B}) between region 1 and region 2 for protein B

	-6.5	-2.4	3.5	8.4	10.9	14.2
	-5.6	-2.9	1.2	5.5	5.4	8.3
	-0.6	-0.9	1.9	2.9	1.9	3.3
	3.4	0.2	-2.3	-3.4	-5.3	-4.2
	7.8	1.5	-2.3	-6.8	-8.9	-9.4
	9.5	3	-2.5	-8	-12.8	-14.8

The bottom is the comparison of the two 6x6 distance matrices for proteins A and B.

If the two sub-structures were similar then the differences should be small.

6x6 Carbon α Distance Matrix (2)

Step 1: For each protein every 6 residue contiguous region is used to generate a 6x6 distance matrix to every other 6 residue contiguous region of the same protein.

Thus for each comparison 12 residues from the same protein are used to generate a distance matrix.

Example: Two proteins, one 129 residues and the other 164 residues.

	Protein A	Protein B
Residues	129	164
6-residue contiguous regions	129 – 5 = 124	164 – 5 = 159
Intra-protein distance matrices	(124*123)/2 = 7,626	(159*158)/2 = 12,561

6x6 Carbon α Distance Matrix (3)

Step 2: Every intra-protein distance matrix from protein A will be compared to every intra-protein distance matrix from protein B.

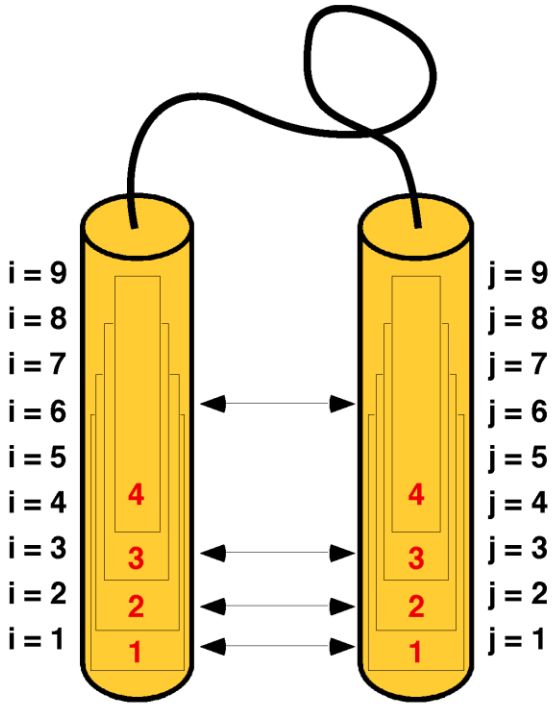
	Protein A	Protein B
Residues	129	164
6-residue contiguous regions	129 – 5 = 124	164 – 5 = 159
Intra-protein distance matrices	(124*123)/2 = 7,626	(159*158)/2 = 12,561
Total Inter-protein distance matrix comparisons between proteins A and B	7,626 x 12,561 = 95,790,186	

And that is just two proteins, how about the whole PDB?

DALI will find some efficiencies to speed the calculation.

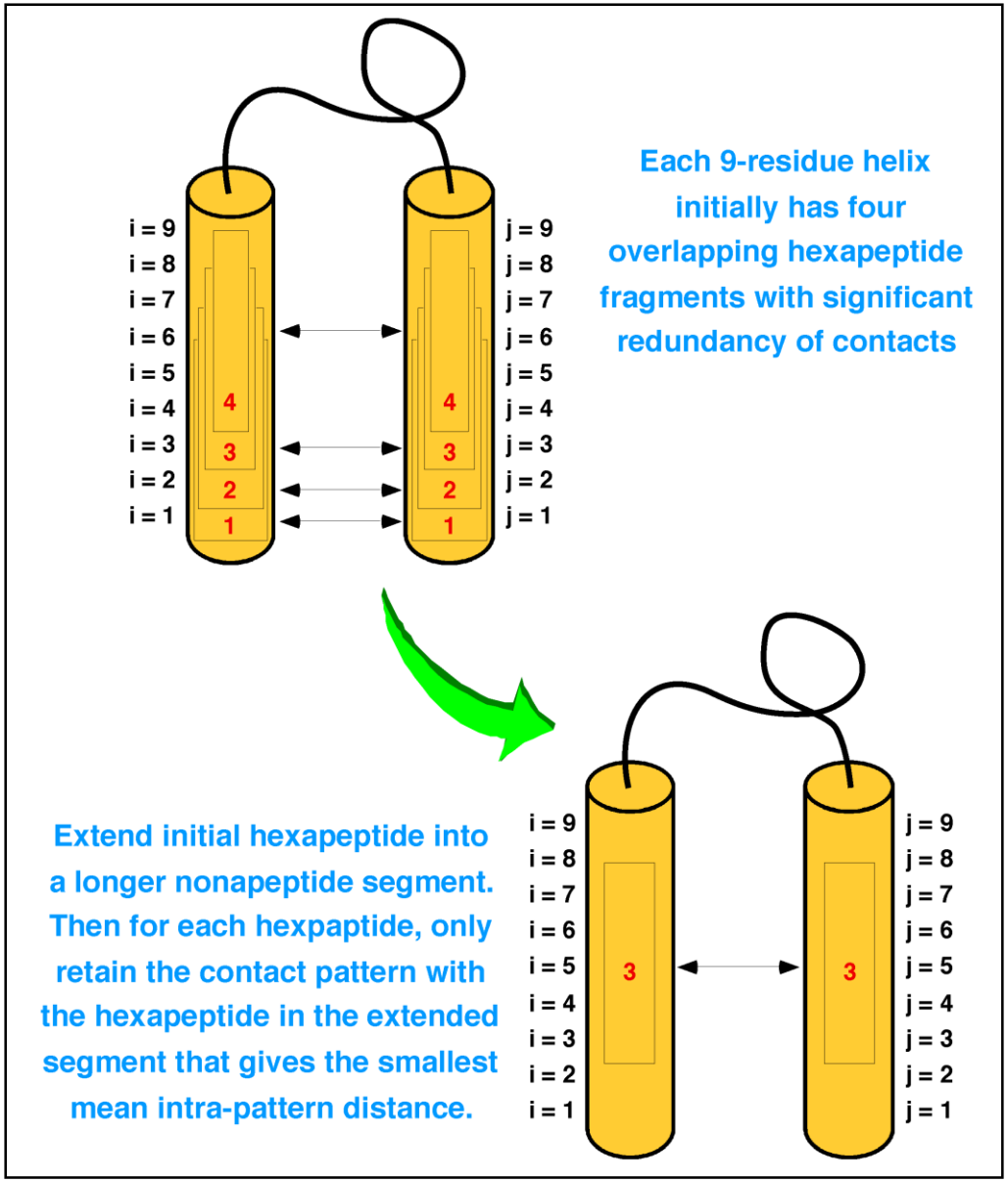
Reducing the Number of Matrices (1)

Neighbouring contact patterns may overlap by as much as 11 of 12 residues as shown below. To reduce this redundant information, successive hexapeptide segments (starting at i, i + 1 ....) that repeat a strongly similar contact pattern along the main diagonal of the distance matrix are **merged** into longer segments.



Each 9-residue helix initially has four overlapping hexapeptide fragments with significant redundancy of contacts

Reducing the Number of Matrices (2)



In the previous example of the 129 and 164 residue proteins the number of distance matrices drops from 96 million to 71 million.

The Scoring Function again (1)

$$S = \sum_{i=1}^{L} \sum_{j=1}^{L} \phi(i,j)$$

where i and j label pairs of equivalent (matched) residues

L is the number of such pairs (i.e., the size of the substructure)

ϕ is a similarity measure based on the C_α–C_α distances d_{ij}^{A} and d_{ij}^{B}

The scoring function S scores the similarity between a distance matrix from a hexapeptide-hexapeptide pair from one protein with the distance matrix from a hexapeptide-hexapeptide pair from a second protein. (24 total residues involved)

Millions of such comparisons will exist.

The scoring function will be used later in an alignment stage where the best possible alignment between two proteins can be made.

The alignment will likely not be across the whole protein, but rather only over regions of the proteins that have structural similarities.

The Scoring Function again (2)

$$S = \sum_{i=1}^{L} \sum_{j=1}^{L} \phi(i,j)$$

where i and j label pairs of equivalent (matched) residues

L is the number of such pairs (i.e., the size of the substructure)

ϕ is a similarity measure based on the C_α–C_α distances d_{ij}^A and d_{ij}^B

In defining the similarity measure ϕ we need to balance two contradictory requirements.

1. Maximizing the number of equivalenced residues in the two proteins
2. Minimizing structural deviations.

If the criteria are so tough that minor structural deviations are not allowed then the equivalenced substructures are likely to be very small, but need to be stringent enough that reasonably similar structures are found.

Rigid Residue-Pair Similarity Score

$$\phi^R(i,j) = \theta^R - \left| d_{ij}^A - d_{ij}^B \right|$$

R stands for rigid, d_{ij}^A and d_{ij}^B are the Cα-Cα distances matrices of proteins *A* and *B*.

θ^R = 1.5 Å is the zero level of similarity. Any equivalenced elements that differ by more than 1.5 Å will count against the score and those less than 1.5 Å will count toward the score. The higher the score the better the similarity.

d_{ij}^A (Å)	d_{ij}^B (Å)	% difference	Score
4.0	4.5	11.8%	1.0
14.0	14.5	3.5%	1.0
5.6	4.0	33.3%	-0.1
15.6	14.0	10.8%	-0.1

This score puts large penalties on relatively small differences in large distances.

NOTE: This example is only for a single distance comparison from the distance matrix. In actuality the score would be the sum of 36 values.

The Elastic Similarity Score (1)

$$\phi E = \left(\theta E - \frac{d_{ij}^{A} - d_{ij}^{B}}{d_{ij}^{av.}} \right) \cdot w(d_{ij}^{av.})$$

Where the *E* stands for elastic, d_{ij}^{A} and d_{ij}^{B} are the Cα-Cα distance matrices of proteins A and B and $d_{ij}^{av.}$ is the average of d_{ij}^{A} and d_{ij}^{B}.

$\theta^{E} = 0.20$

By dividing by the average of the two difference distances ($d_{ij}^{av.}$) and by applying $\theta^{E} = 0.20$, larger differences are tolerated for longer range contacts.

d_{ij}^{A} (Å)	d_{ij}^{B} (Å)	% difference	Score
4.0	4.5	11.8%	0.08
14.0	14.5	3.5%	0.16
5.6	4.0	33.3%	-0.13
15.6	14.0	10.8%	0.09

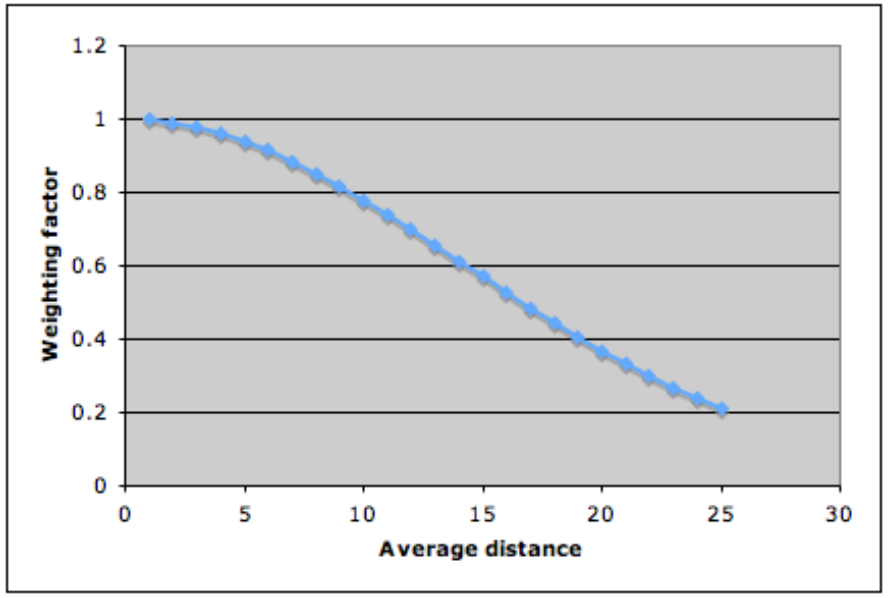
NOTE: This example is only for a single distance comparison from the distance matrix. In actuality the score would be the sum of 36 values.

The Elastic Similarity Score (2)

$$\phi E = \left(\theta E - \frac{d_{ij}^{A} - d_{ij}^{B}}{d_{ij}^{av.}} \right) \cdot w(d_{ij}^{av.})$$

Since pairs in the long distance range are abundant but less discriminative, their contribution is weighted down by the envelope function:

$$w(d_{ij}^{av.}) = exp\{-[(d_{ij}^{av.})^2/\alpha^2)]\}$$



α was calibrated to 20 Å, based on the size of a typical domain. This 20 Å distances will act to reduce domain-domain interactions.

The Elastic Similarity Score (3)

$$\phi E = \left(\theta E - \frac{\left| d_{ij}^{A} - d_{ij}^{B} \right|}{d_{ij}^{av.}} \right) \cdot w(d_{ij}^{av.})$$

$$S = \sum_{i=1}^{L} \sum_{j=1}^{L} \left(0.2 - \frac{\left| d_{ij}^{A} - d_{ij}^{B} \right|}{d_{ij}^{av.}} \right) \cdot e^{-\left(d_{ij}^{av.} / 20 \text{ Å} \right)^2}$$

Distance difference (d_{ij}^{A}-d_{ij}^{B}) (Å)

Average distance (Å)	0.5	1.0	2.0	3.0	4.0
1	-0.30	-0.80	-1.80	-2.79	-3.79
2	-0.05	-0.30	-0.79	-1.29	-1.78
3	0.03	-0.13	-0.46	-0.78	-1.11
4	0.07	-0.05	-0.29	-0.53	-0.77
5	0.09	0.00	-0.19	-0.38	-0.56
6	0.11	0.03	-0.12	-0.27	-0.43
7	0.11	0.05	-0.08	-0.20	-0.33
8	0.12	0.06	-0.04	-0.15	-0.26
9	0.12	0.07	-0.02	-0.11	-0.20
10	0.12	0.08	0.00	-0.08	-0.16
11	0.11	0.08	0.01	-0.05	-0.12
12	0.11	0.08	0.02	-0.03	-0.09
13	0.11	0.08	0.03	-0.02	-0.07
14	0.10	0.08	0.04	-0.01	-0.05
15	0.09	0.08	0.04	0.00	-0.04
16	0.09	0.07	0.04	0.01	-0.03
17	0.08	0.07	0.04	0.01	-0.02
18	0.08	0.06	0.04	0.01	-0.01
19	0.07	0.06	0.04	0.02	0.00
20	0.06	0.06	0.04	0.02	0.00

Table shows the elastic similarity score (ϕE) for average distances ($d_{ij}^{av.}$) ranging from 1 Å to 20 Å for distance differences from 0.5 to 5.0 Å

Larger distance differences are better tolerated at longer average distances, while longer range contributions are weighted down by the weighting factor.

Summary of where we are …

From steps 1 and 2 we have distance matrices for two proteins …

	Protein A	Protein B
Residues	129	164
6-residue contiguous regions	124	159
Intra-protein distance matrices	7,626	12,561
Total Inter-protein distance matrix comparisons between proteins A and B	96 million	
Total after merging	71 million	

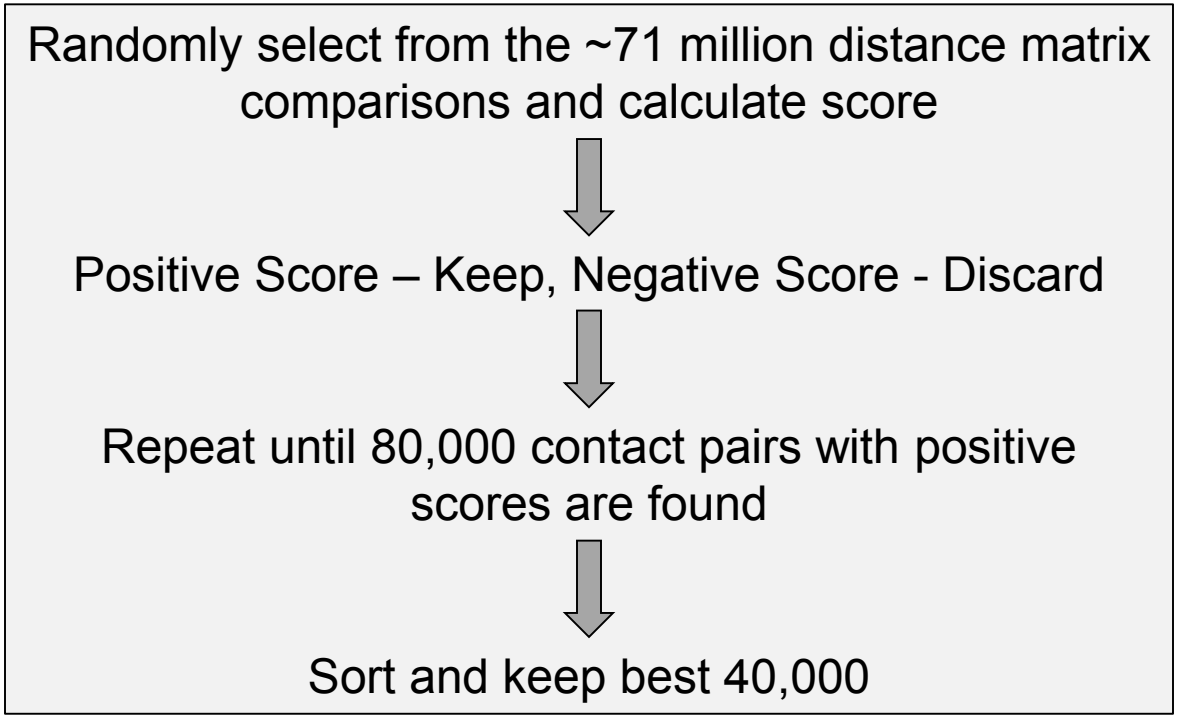
We have a similarity scoring metric…

$$S = \sum_{i=1}^{L} \sum_{j=1}^{L} \left(0.2 - \frac{\left| d_{ij}^{A} - d_{ij}^{B} \right|}{d_{ij}^{av.}} \right) \cdot e^{-\left(d_{ij}^{av.} / 20 \text{ Å} \right)^2}$$

Now what?

Initial Comparisons for Alignment

Step 3: Compare the distance matrices



What do we have at this point?

A list of 40,000 contact patterns that are similar (i.e. have positive scores)

What is similar is the distance pattern between a hexapeptide-hexapeptide pair from Protein A and a hexapeptide-hexapeptide pair from Protein B. (24 total residues)

Alignment Procedure

Step 4: Align the contact patterns

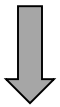
Problem: Find optimal alignment of 40,000 contact patterns such that the alignment occurs over as many residues as possible while improving the overall score.

A Markov Chain Monte Carlo Optimization (MCMC) will be used.

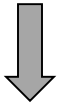
Alignment Procedure

Step 4: Align the contact patterns

- Start from a high dimensional, 40,000 contact pattern, space
- Pick one at random
- Take a “walk” to another contact partner which has some overlap with the initial contact pattern
    - All possible overlapping contact partners are tried at each step in parallel – DALI calls these trajectories.
- Rescore across the now larger contact pattern
    - If the score is better always keep it
    - If the score is worse we may or may not keep it depending on a probability
- Repeat the “walk” until the score does not improve.



Repeat the process

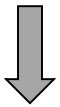


Keep the BEST alignment

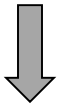
Alignment Procedure

Step 4: Align the contact patterns

- Start from a high dimensional, 40,000 contact pattern, space
- Pick one at random
- Take a “walk” to another contact partner which has some overlap with the initial contact pattern
  - All possible overlapping contact partners are tried at each step in parallel – DALI calls these trajectories.
- Rescore across the now larger contact pattern (alignment)
  - If the score is better always keep it
  - If the score is worse we may or may not keep it depending on a probability
- Repeat the “walk” until the score does not improve.



Repeat the process 100 times



Keep the BEST alignment

Scoring and the MCMC

The MCMC simply defines the probability of accepting any given move through space for any given trajectory.

$$Probability_{(accepting\ move)} = e^{(\beta*(S'-S))}$$

where S = old score, S′ = new score, and β = inverse temperature of system

Higher values of β increase the probability of accepting a move that decreases the overall score.

S (old)	S' (new)	S'-S	Temp (B)	Probability
1000	1010	10	100	> 100%
1000	1000	0	100	100%
1000	990	-10	100	90%
1000	990	-10	50	82%
1000	990	-10	10	37%
1000	900	-100	100	37%
1000	900	-100	50	14%
1000	900	-100	10	0.0%

Net result: Always keep scores that are better (or the same) and possibly keep scores that gets worse, based on a probability, allowing to search past a local minimum.

Overview of Trajectories

Seeds: Typically 100 random contact patterns are used as starting points.

Trajectory: Starting from any given seed contact pattern other overlapping contact patterns are added as long as the score improves or is randomly added based on a probability.

- All possible contact patterns which overlap with alignment are tried
- At any step there may be multiple contact patterns may be added
  - New trajectories are spawned for each of the contact patterns being added.
- Thus, from a single contact pattern many trajectories will be computed in parallel.

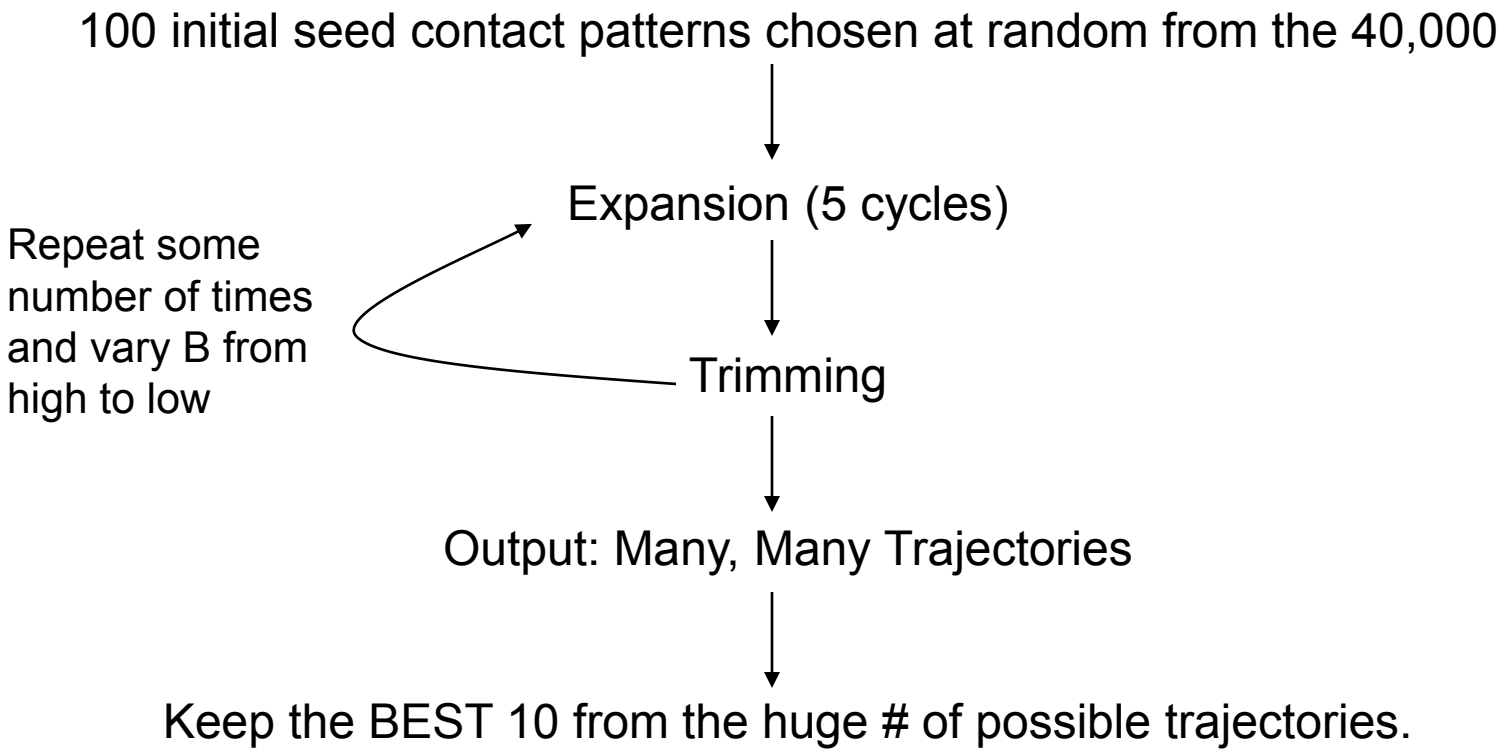
Expansion & Trimming: A trajectory will have both expansion cycles (times when new overlapping contact patterns are being added) and trimming cycles (times when contact patterns are removed if their removal improves the overall score).

Computational load: The computational load is reduced by:

- Killing trajectories that are no longer improving after some number of rounds.
- Killing trajectories which fall off the pace from other trajectories.

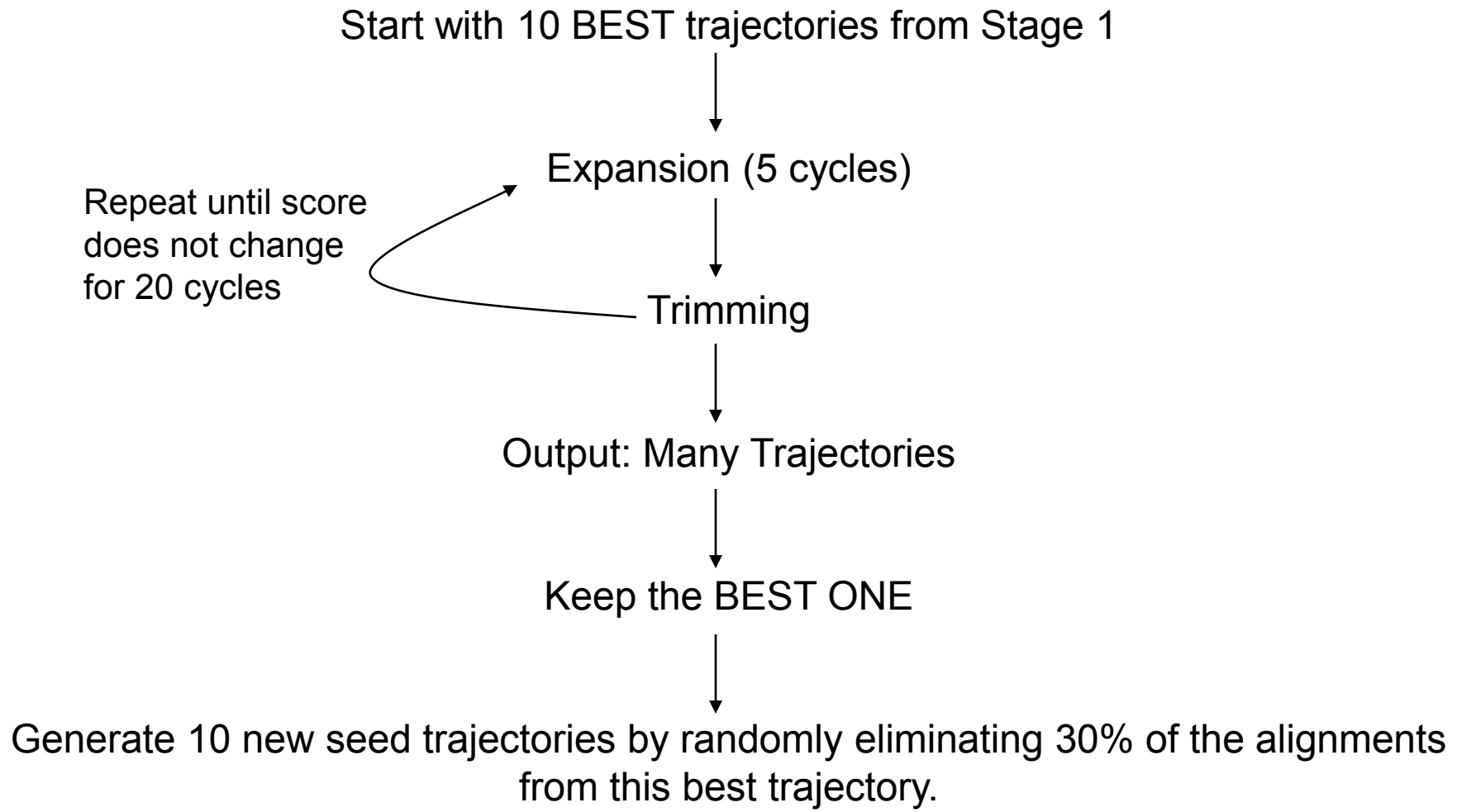
Details of the Alignment Phase (Stage 1)

Stage 1:



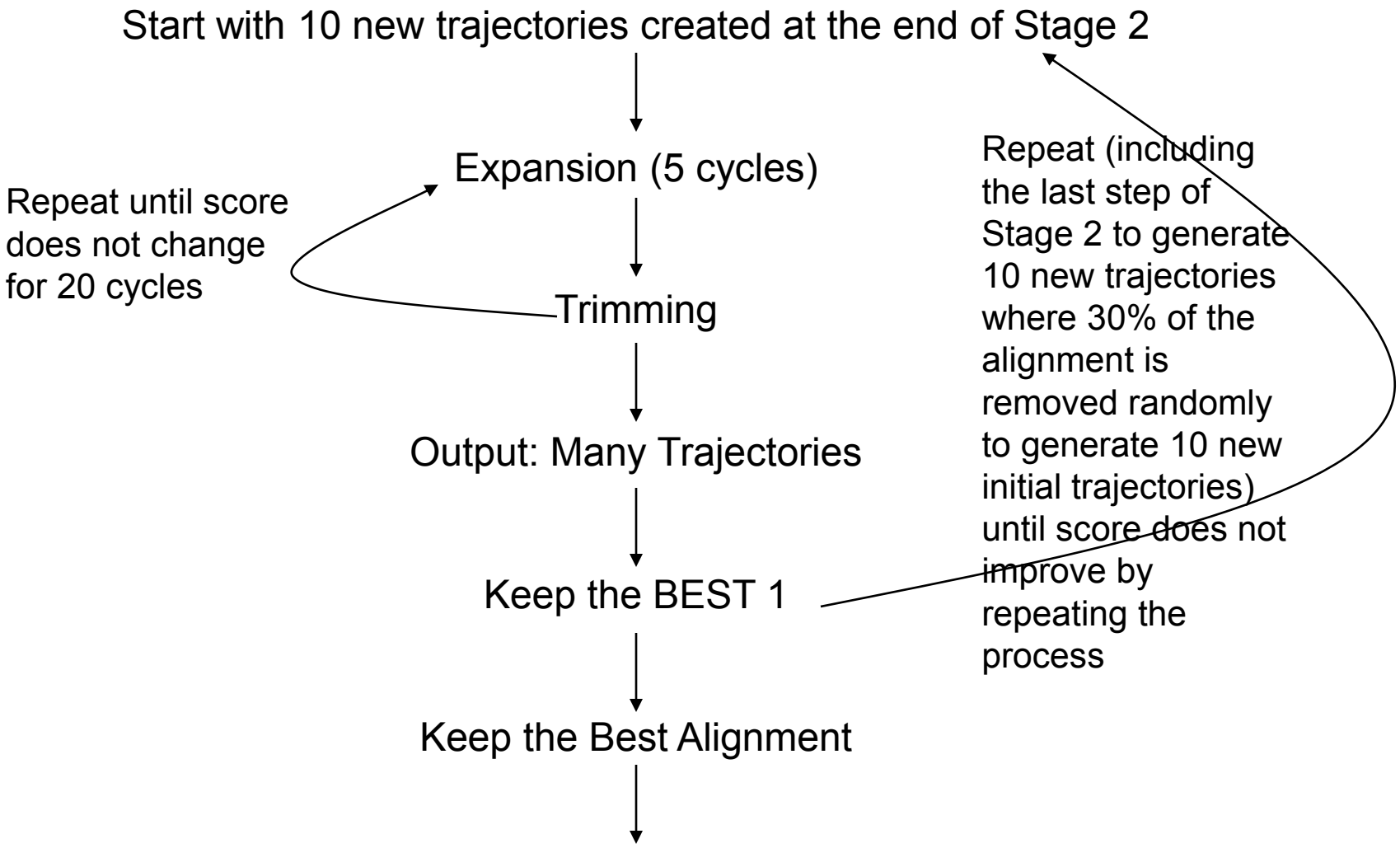
Details of the Alignment Phase (Stage 2)

Stage 2:



Details of the Alignment Phase (Stage 3)

Stage 3:



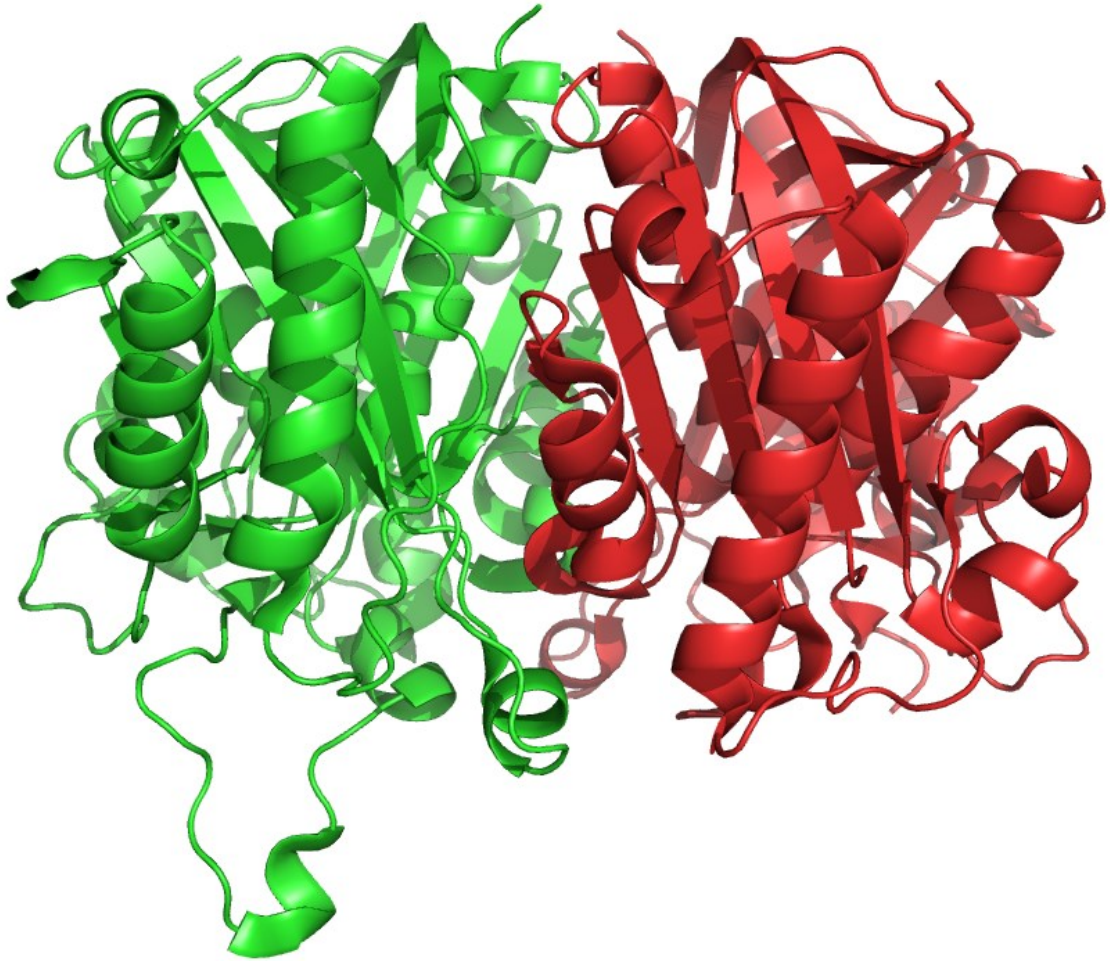
DALI Output

- Z – Alignment score normalized to compare Z scores across different alignments
- rmsd – RMSD across ONLY the aligned portions
- lali – The number of residues aligned
- nres – Total number of residues
- %id – The percent identity

No:	Chain	Z	rmsd	lali	nres	%id	PDB	Description
1:	3lm6-A	66.3	0.0	334	334	100	PDB	MOLECULE: STAGE V SPORULATION PROTEIN AD;
2:	3lm6-B	62.6	0.2	322	322	100	PDB	MOLECULE: STAGE V SPORULATION PROTEIN AD;
3:	3lma-C	52.3	2.4	324	327	80	PDB	MOLECULE: STAGE V SPORULATION PROTEIN AD (SPOVAD);
4:	3lma-D	52.3	1.3	311	311	82	PDB	MOLECULE: STAGE V SPORULATION PROTEIN AD (SPOVAD);
5:	3lma-A	51.7	1.1	309	309	81	PDB	MOLECULE: STAGE V SPORULATION PROTEIN AD (SPOVAD);
6:	3lma-B	51.6	1.2	310	311	83	PDB	MOLECULE: STAGE V SPORULATION PROTEIN AD (SPOVAD);
7:	3h78-A	24.9	3.2	258	329	17	PDB	MOLECULE: PQS BIOSYNTHETIC ENZYME;
8:	3h76-B	24.8	3.3	261	329	17	PDB	MOLECULE: PQS BIOSYNTHETIC ENZYME;
9:	2qyo-B	24.7	3.3	254	317	15	PDB	MOLECULE: 3-OXOACYL-[ACYL-CARRIER-PROTEIN] SYNTHASE 3;
10:	3h78-B	24.7	3.4	263	336	16	PDB	MOLECULE: PQS BIOSYNTHETIC ENZYME;
11:	3h77-B	24.5	3.2	259	338	17	PDB	MOLECULE: PQS BIOSYNTHETIC ENZYME;
12:	3h76-A	24.5	3.3	261	329	17	PDB	MOLECULE: PQS BIOSYNTHETIC ENZYME;
13:	3h77-A	24.4	3.4	263	331	16	PDB	MOLECULE: PQS BIOSYNTHETIC ENZYME;
14:	2ebd-A	24.4	3.2	257	309	16	PDB	MOLECULE: 3-OXOACYL-[ACYL-CARRIER-PROTEIN] SYNTHASE 3;
15:	3lrf-A	24.2	3.0	259	407	16	PDB	MOLECULE: BETA-KETOACYL SYNTHASE;
16:	3oyt-A	24.2	3.1	263	407	14	PDB	MOLECULE: 3-OXOACYL-[ACYL-CARRIER-PROTEIN] SYNTHASE I;
17:	1wl4-A	24.1	2.8	258	394	13	PDB	MOLECULE: ACETYL-COENZYME A ACETYLTRANSFERASE 2;
18:	1f91-C	24.1	2.8	262	406	13	PDB	MOLECULE: BETA-KETOACYL-[ACYL-CARRIER-PROTEIN] SYNTHASE I;
19:	1f91-D	24.1	2.9	262	406	13	PDB	MOLECULE: BETA-KETOACYL-[ACYL-CARRIER-PROTEIN] SYNTHASE I;
20:	1f91-A	24.1	2.9	262	406	13	PDB	MOLECULE: BETA-KETOACYL-[ACYL-CARRIER-PROTEIN] SYNTHASE I;
21:	1dd8-A	24.0	2.9	262	406	13	PDB	MOLECULE: BETA-KETOACYL [ACYL CARRIER PROTEIN] SYNTHASE I;
22:	1mlt-A	24.0	2.9	257	392	13	PDB	MOLECULE: ACETYL-COA ACETYLTRANSFERASE;
23:	1ebl-A	24.0	3.4	260	317	15	PDB	MOLECULE: BETA-KETOACYL-ACP SYNTHASE III;
24:	1ek4-C	24.0	2.9	262	406	13	PDB	MOLECULE: BETA-KETOACYL [ACYL CARRIER PROTEIN] SYNTHASE I;
25:	1dd8-D	24.0	2.9	262	406	13	PDB	MOLECULE: BETA-KETOACYL [ACYL CARRIER PROTEIN] SYNTHASE I;
26:	1dd8-C	24.0	2.9	262	406	13	PDB	MOLECULE: BETA-KETOACYL [ACYL CARRIER PROTEIN] SYNTHASE I;
27:	1dd8-B	24.0	2.9	262	406	13	PDB	MOLECULE: BETA-KETOACYL [ACYL CARRIER PROTEIN] SYNTHASE I;
28:	1f91-B	24.0	2.9	262	406	13	PDB	MOLECULE: BETA-KETOACYL-[ACYL-CARRIER-PROTEIN] SYNTHASE I;
29:	1ek4-B	24.0	2.9	262	406	13	PDB	MOLECULE: BETA-KETOACYL [ACYL CARRIER PROTEIN] SYNTHASE I;

Example: Structure of SpoVA-D with Unknown Function

SpoVA-D is a protein involved in spore germination from *Bacillius subtilis*

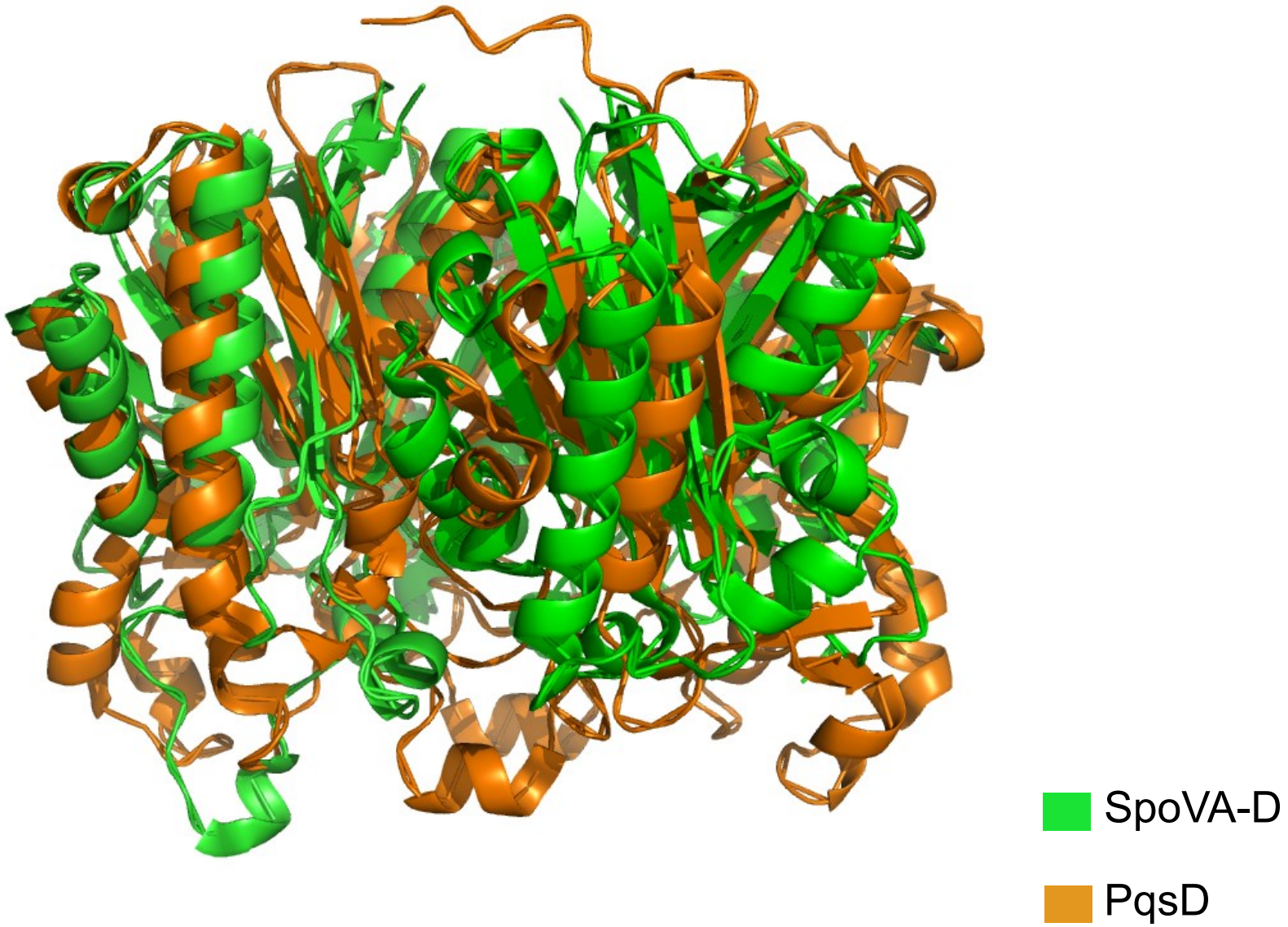


Structure homologs from Dali search

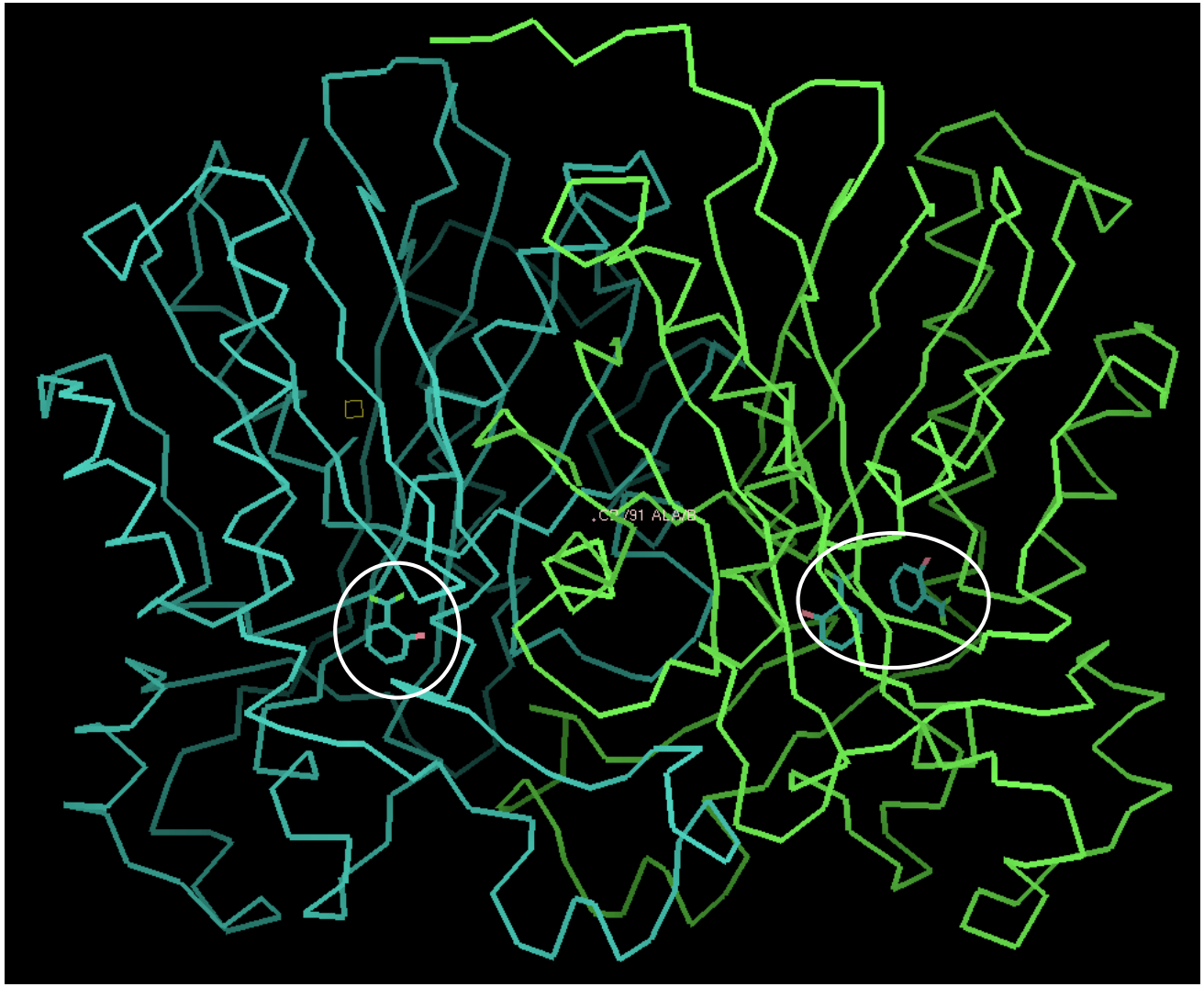
Polyketide synthase superfamily

No:	Chain	Z	rmsd	lali	nres	%id	PDB	Description
1:	3lm6-A	66.3	0.0	334	334	100	PDB	MOLECULE: STAGE V SPORULATION PROTEIN AD;
2:	3lm6-B	62.6	0.2	322	322	100	PDB	MOLECULE: STAGE V SPORULATION PROTEIN AD;
3:	3lma-C	52.3	2.4	324	327	80	PDB	MOLECULE: STAGE V SPORULATION PROTEIN AD (SPOVAD);
4:	3lma-D	52.3	1.3	311	311	82	PDB	MOLECULE: STAGE V SPORULATION PROTEIN AD (SPOVAD);
5:	3lma-A	51.7	1.1	309	309	81	PDB	MOLECULE: STAGE V SPORULATION PROTEIN AD (SPOVAD);
6:	3lma-B	51.6	1.2	310	311	83	PDB	MOLECULE: STAGE V SPORULATION PROTEIN AD (SPOVAD);
7:	3h78-A	24.9	3.2	258	329	17	PDB	MOLECULE: PQS BIOSYNTHETIC ENZYME;
8:	3h76-B	24.8	3.3	261	329	17	PDB	MOLECULE: PQS BIOSYNTHETIC ENZYME;
9:	2qyo-B	24.7	3.3	254	317	15	PDB	MOLECULE: 3-OXOACYL-[ACYL-CARRIER-PROTEIN] SYNTHASE 3;
10:	3h78-B	24.7	3.4	263	336	16	PDB	MOLECULE: PQS BIOSYNTHETIC ENZYME;
11:	3h77-B	24.5	3.2	259	338	17	PDB	MOLECULE: PQS BIOSYNTHETIC ENZYME;
12:	3h76-A	24.5	3.3	261	329	17	PDB	MOLECULE: PQS BIOSYNTHETIC ENZYME;
13:	3h77-A	24.4	3.4	263	331	16	PDB	MOLECULE: PQS BIOSYNTHETIC ENZYME;
14:	2ebd-A	24.4	3.2	257	309	16	PDB	MOLECULE: 3-OXOACYL-[ACYL-CARRIER-PROTEIN] SYNTHASE 3;
15:	3lrf-A	24.2	3.0	259	407	16	PDB	MOLECULE: BETA-KETOACYL SYNTHASE;
16:	3oyt-A	24.2	3.1	263	407	14	PDB	MOLECULE: 3-OXOACYL-[ACYL-CARRIER-PROTEIN] SYNTHASE I;
17:	1wl4-A	24.1	2.8	258	394	13	PDB	MOLECULE: ACETYL-COENZYME A ACETYLTRANSFERASE 2;
18:	1f91-C	24.1	2.8	262	406	13	PDB	MOLECULE: BETA-KETOACYL-[ACYL-CARRIER-PROTEIN] SYNTHASE I;
19:	1f91-D	24.1	2.9	262	406	13	PDB	MOLECULE: BETA-KETOACYL-[ACYL-CARRIER-PROTEIN] SYNTHASE I;
20:	1f91-A	24.1	2.9	262	406	13	PDB	MOLECULE: BETA-KETOACYL-[ACYL-CARRIER-PROTEIN] SYNTHASE I;
21:	1dd8-A	24.0	2.9	262	406	13	PDB	MOLECULE: BETA-KETOACYL [ACYL CARRIER PROTEIN] SYNTHASE I;
22:	1m1t-A	24.0	2.9	257	392	13	PDB	MOLECULE: ACETYL-COA ACETYLTRANSFERASE;
23:	1ebl-A	24.0	3.4	260	317	15	PDB	MOLECULE: BETA-KETOACYL-ACP SYNTHASE III;
24:	1ek4-C	24.0	2.9	262	406	13	PDB	MOLECULE: BETA-KETOACYL [ACYL CARRIER PROTEIN] SYNTHASE I;
25:	1dd8-D	24.0	2.9	262	406	13	PDB	MOLECULE: BETA-KETOACYL [ACYL CARRIER PROTEIN] SYNTHASE I;
26:	1dd8-C	24.0	2.9	262	406	13	PDB	MOLECULE: BETA-KETOACYL [ACYL CARRIER PROTEIN] SYNTHASE I;
27:	1dd8-B	24.0	2.9	262	406	13	PDB	MOLECULE: BETA-KETOACYL [ACYL CARRIER PROTEIN] SYNTHASE I;
28:	1f91-B	24.0	2.9	262	406	13	PDB	MOLECULE: BETA-KETOACYL-[ACYL-CARRIER-PROTEIN] SYNTHASE I;
29:	1ek4-B	24.0	2.9	262	406	13	PDB	MOLECULE: BETA-KETOACYL [ACYL CARRIER PROTEIN] SYNTHASE I;
30:	1ek4-D	24.0	2.9	262	406	13	PDB	MOLECULE: BETA-KETOACYL [ACYL CARRIER PROTEIN] SYNTHASE I;
31:	1wl5-A	24.0	2.7	257	394	13	PDB	MOLECULE: ACETYL-COENZYME A ACETYLTRANSFERASE 2;
32:	2ibu-C	24.0	2.8	254	393	12	PDB	MOLECULE: ACETYL-COA ACETYLTRANSFERASE;
33:	2ib8-C	24.0	2.7	254	392	12	PDB	MOLECULE: ACETYL-COA ACETYLTRANSFERASE;
34:	2ibu-A	24.0	2.8	254	391	12	PDB	MOLECULE: ACETYL-COA ACETYLTRANSFERASE;
35:	2ib7-C	24.0	2.8	254	392	12	PDB	MOLECULE: ACETYL-COA ACETYLTRANSFERASE;
36:	3mqd-A	24.0	2.9	261	407	15	PDB	MOLECULE: BETA-KETOACYL SYNTHASE;
37:	1ek4-A	24.0	2.9	262	406	13	PDB	MOLECULE: BETA-KETOACYL [ACYL CARRIER PROTEIN] SYNTHASE I;
38:	3il9-A	24.0	3.4	259	317	14	PDB	MOLECULE: 3-OXOACYL-[ACYL-CARRIER-PROTEIN] SYNTHASE 3;
39:	2ib7-D	23.9	2.7	254	393	12	PDB	MOLECULE: ACETYL-COA ACETYLTRANSFERASE;
40:	2ibu-D	23.9	2.7	254	393	11	PDB	MOLECULE: ACETYL-COA ACETYLTRANSFERASE;
41:	2iby-D	23.9	2.8	254	393	12	PDB	MOLECULE: ACETYL-COA ACETYLTRANSFERASE;
42:	2ib8-D	23.9	2.7	254	393	12	PDB	MOLECULE: ACETYL-COA ACETYLTRANSFERASE;
43:	2iby-B	23.9	2.8	254	391	12	PDB	MOLECULE: ACETYL-COA ACETYLTRANSFERASE;
44:	2ib7-A	23.9	2.8	254	391	12	PDB	MOLECULE: ACETYL-COA ACETYLTRANSFERASE;
45:	2iby-C	23.9	2.8	254	393	12	PDB	MOLECULE: ACETYL-COA ACETYLTRANSFERASE;
46:	2ibu-B	23.9	2.8	254	391	12	PDB	MOLECULE: ACETYL-COA ACETYLTRANSFERASE;
47:	2ib9-C	23.9	2.7	254	392	12	PDB	MOLECULE: ACETYL-COA ACETYLTRANSFERASE;
48:	2eft-B	23.9	3.4	257	317	14	PDB	MOLECULE: 3-OXOACYL-[ACYL-CARRIER-PROTEIN] SYNTHASE 3;
49:	2eft-A	23.9	3.4	252	317	15	PDB	MOLECULE: 3-OXOACYL-[ACYL-CARRIER-PROTEIN] SYNTHASE 3;

Structure overlay of SpoVA-D and PqsD



Structural homologs of SpoVAD can bind small molecule!



PqsD/anthranilate

DPA has structural similarities to anthranilic acid

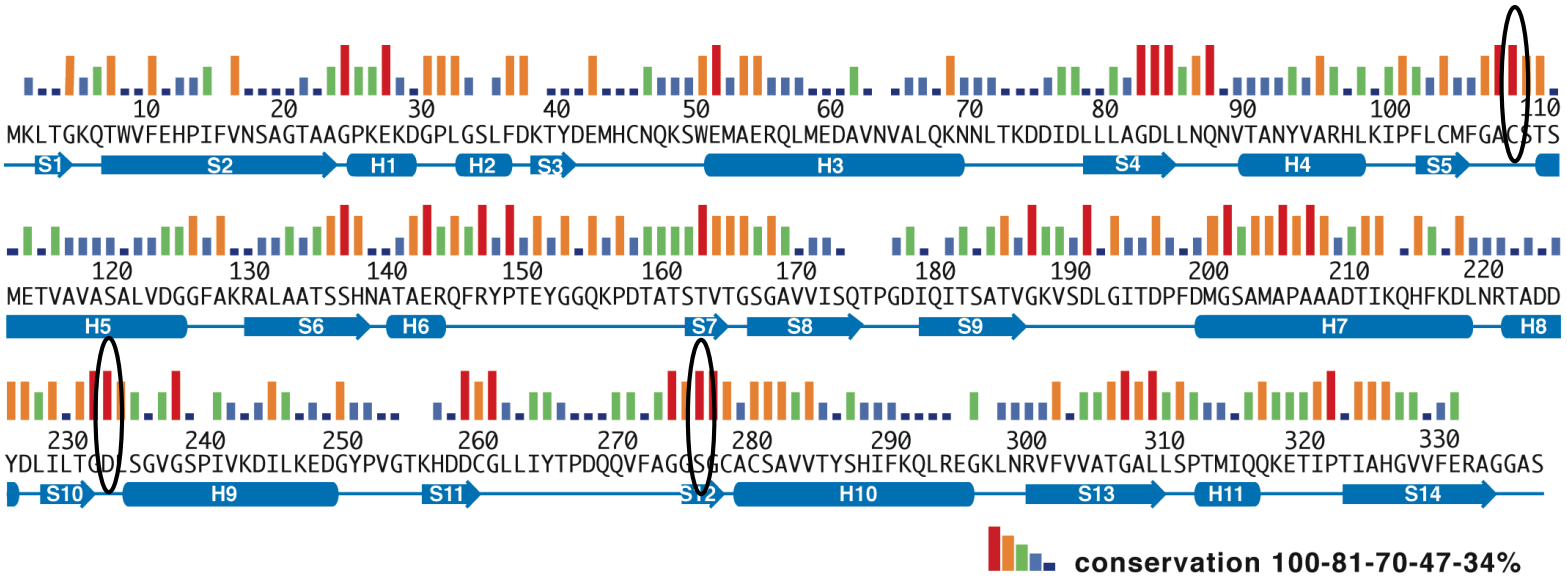
DPA is stored in very high concentrations in spores of *Bacillus subtilis*. Hypothesized, based on structural alignments, that SpoVA-D may bind DPA.

anthranilic acid

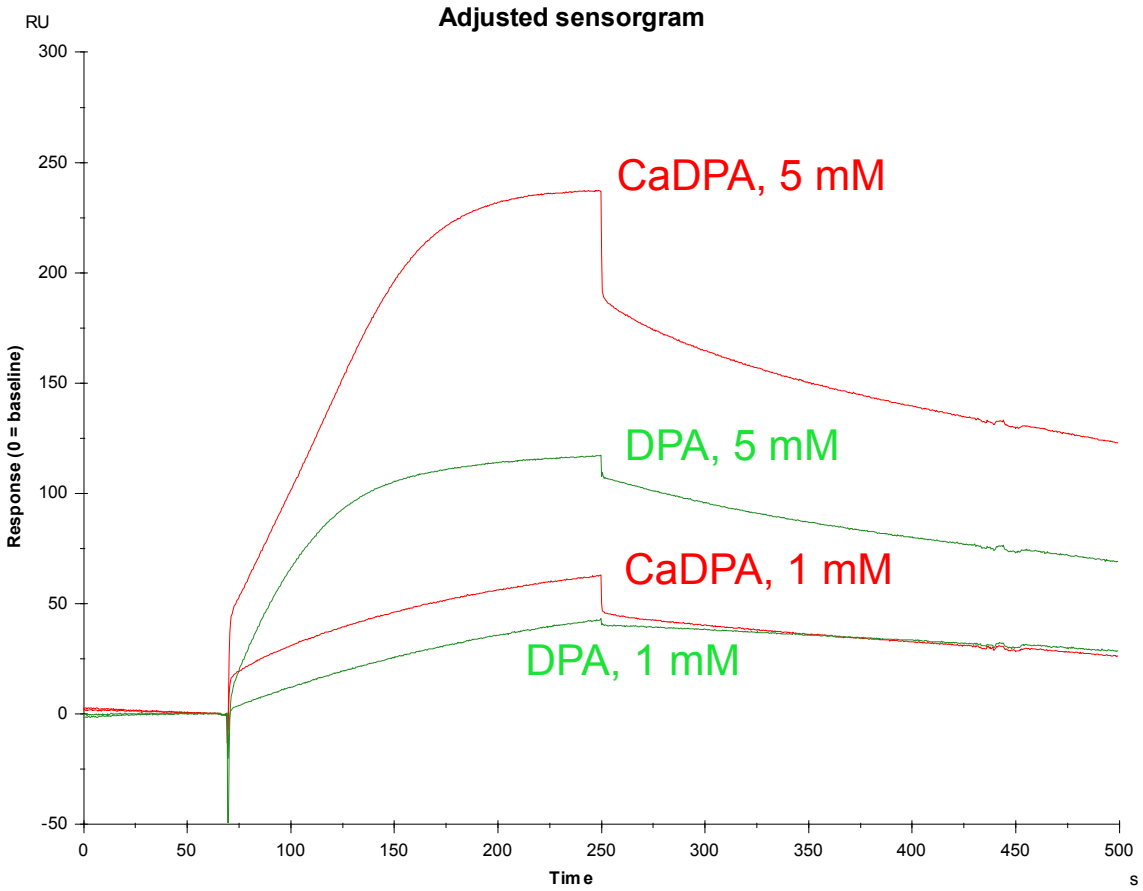
2,6-pyridinedicarboxylic acid
(dipicolinic acid, DPA)

Sequence conservation among 86 SpoVA-D homologs

Residues predicted to bind DPA have very high sequence homology



Does DPA bind SpoVA-D?



Both DPA and Ca^{2+}-DPA bind to SpoVA-D. Calcium bound DPA has faster on/off rates, but K_D is the same.

DALI Web Sites and References

Reference:

Liisa Holm and Chris Sander, Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.* (1993) **233**, 123-138.

DALI server:

http://ekhidna.biocenter.helsinki.fi/dali_server/start

DALI database:

http://ekhidna.biocenter.helsinki.fi/dali/start

DALI and FSSP Web Sites and References

Dali

Reference: Liisa Holm & Chris Sander (1993) Protein structure comparison by alignment of distance matrices. J. Mol. Biol. **233**, 123–138.

DALI can be found at http://www.ebi.ac.uk/dali/

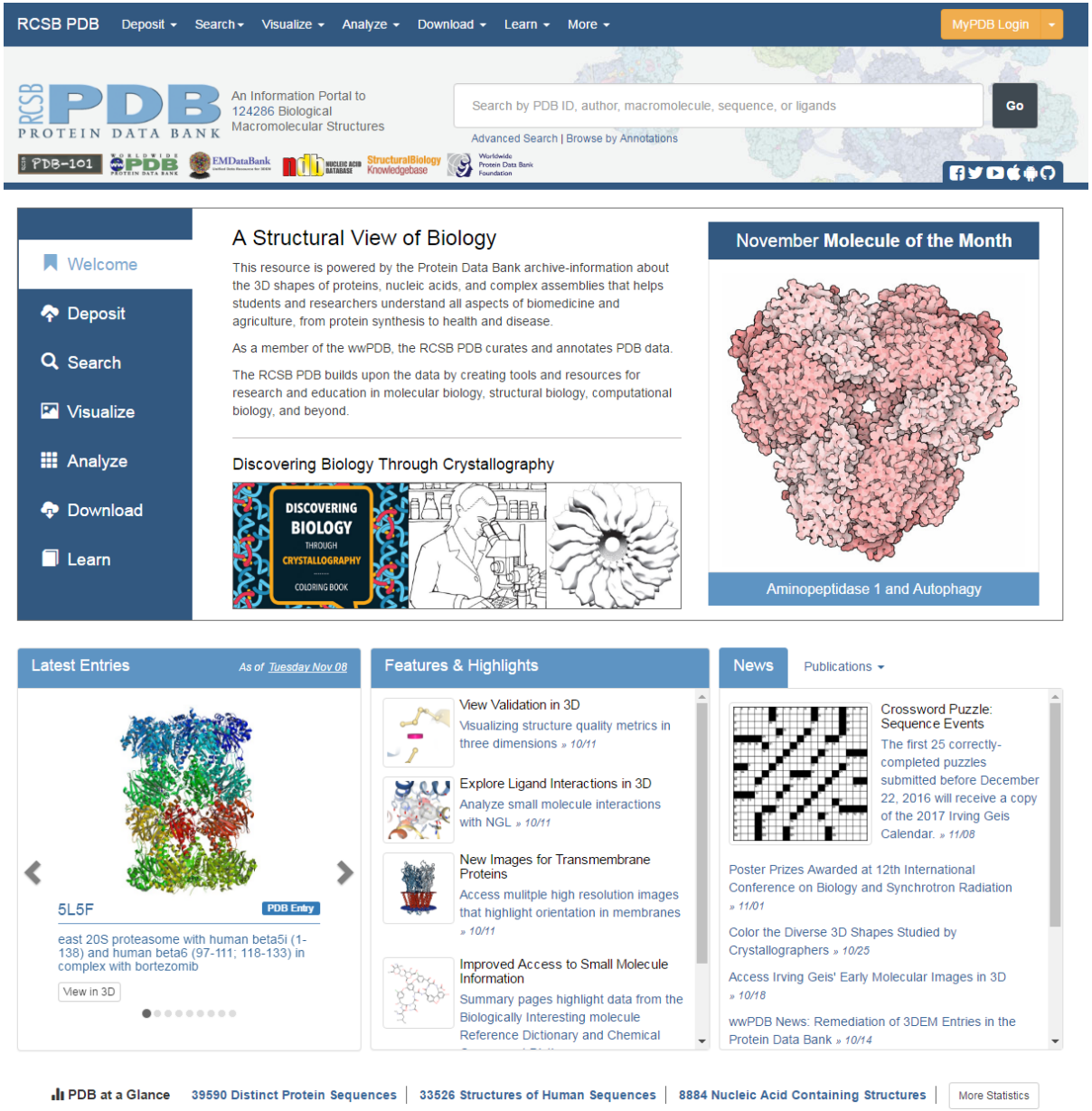
FSSP

The FSSP database is based on exhaustive all-against-all 3D structure comparison of protein structures currently in the Protein Data Bank (PDB). The classification and alignments are automatically maintained and continuously updated using the Dali search engine. [Last update Tues Oct 30 03:04:02 GMT 2001: 2860 sequence families representing 27181 protein structures].

Reference: Liisa Holm & Chris Sander (1996) Mapping the protein universe. Science **273**, 595–602.

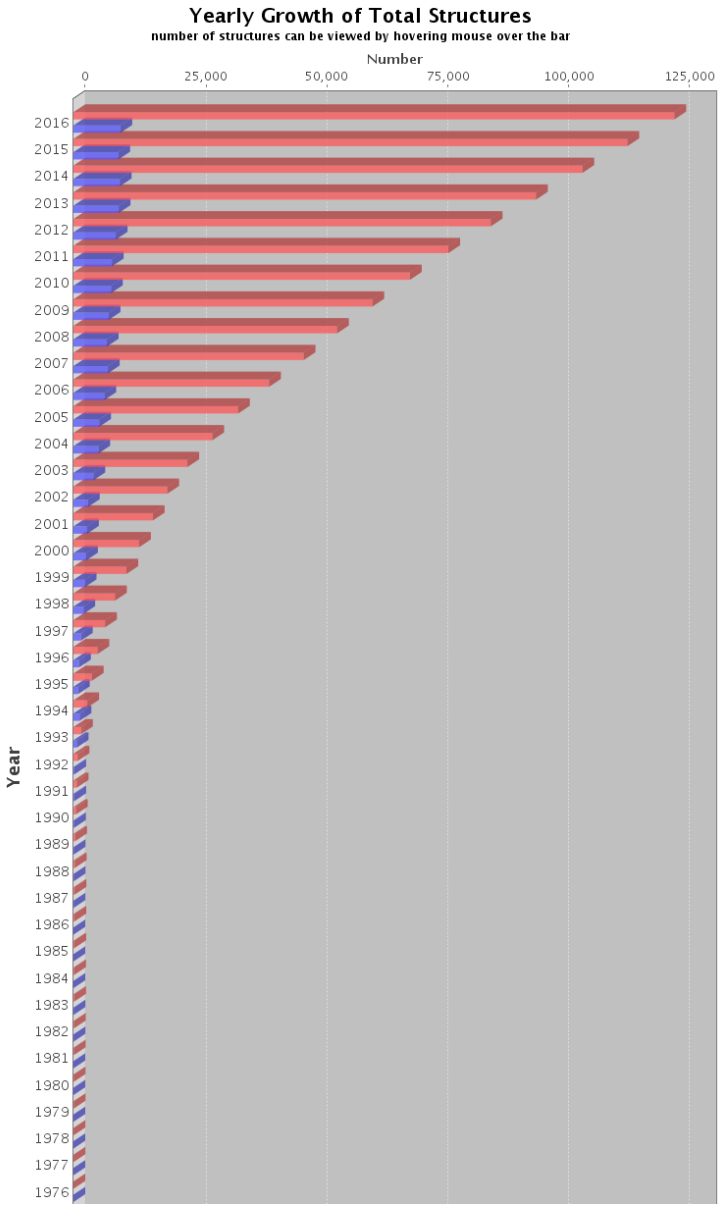
FSSP can be found at http://www.ebi.ac.uk/dali/fssp/fssp.html

Protein Data Bank (PDB)



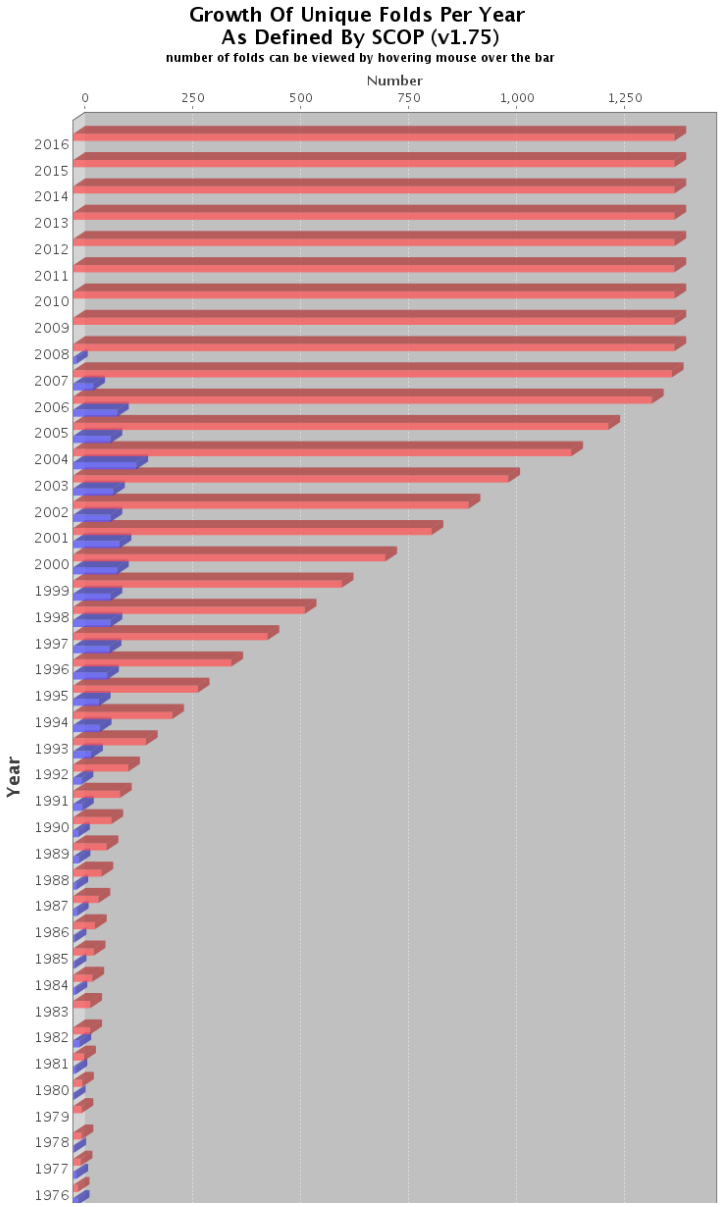
The PDB is a repository where protein structures determined by X-ray crystallography, NMR, EM, and homology are stored. (http://www.rcsb.org/pdb)

PDB Statistics



- To date there have been ~125,000 protein structures deposited to the PDB
- 40% of solved structures are multi-domain
  - In prokaryotes 60% of proteins are multi-domain and in eukaryotes 91%
  - PDB skewed to easy proteins
- Techniques
  - X-ray – 104,000
  - NMR – 10,205
  - Electron microscopy – 900
  - Hybrid – 92
  - Other = 179

Fold Statistics



- According to SCOP there are 1393 unique folds.
- Zero new folds since 2008.
- A fold is defined as having the same secondary structure elements, in the same order, with the same connectivity.
- Nature only allows a finite number of global folds to be energetically favorable.
- However, “the devil is in the details”

A whole unexplored world of intrinsically disordered proteins and invisible states awaits