Discovering of Protein Folding Pathways and Folding Intermediates by the Analysis of Protein Folding Simulations

Luis Garreta

Doctoral Student in Engineering emph. in Computer Science

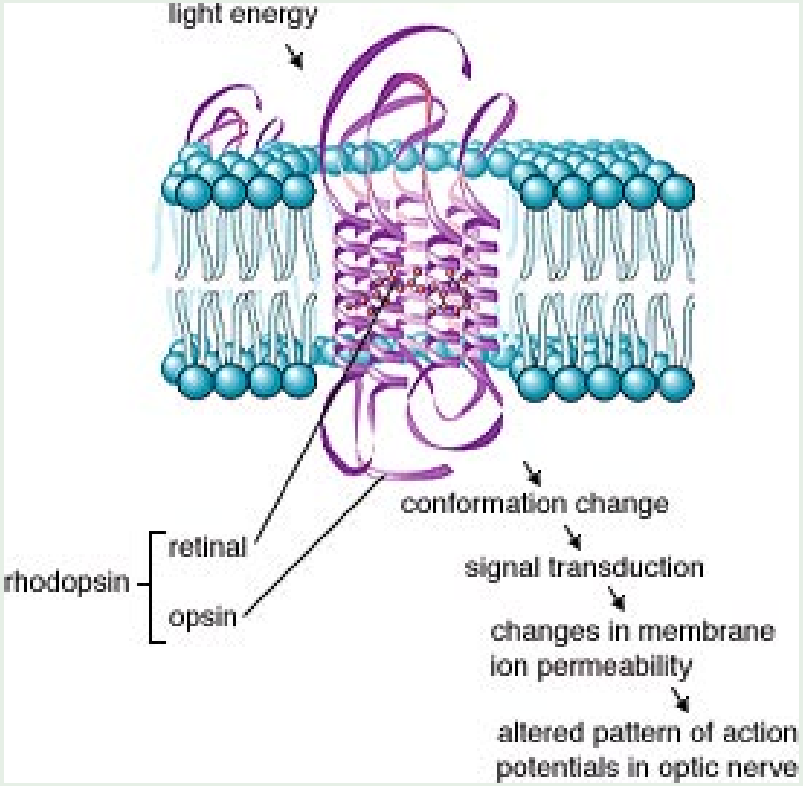
School of System Engineering and Computation
Bioinformatics Research Group
Universidad del Valle
Cali-Colombia

Agenda

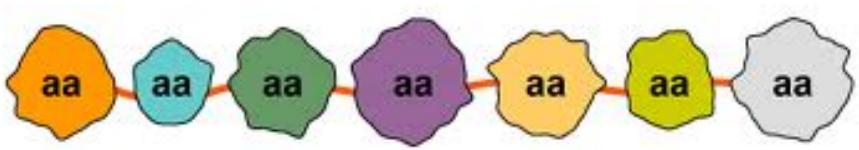
Importance of Proteins

- Most important biomolecules
- Key roles in all living systems.
- They are part of
  - vision,
  - immune system,
  - muscles,
  - tissues,
  - ...
- They are in all of our body.
- They are fundamental for life.

Rhodopsin Protein



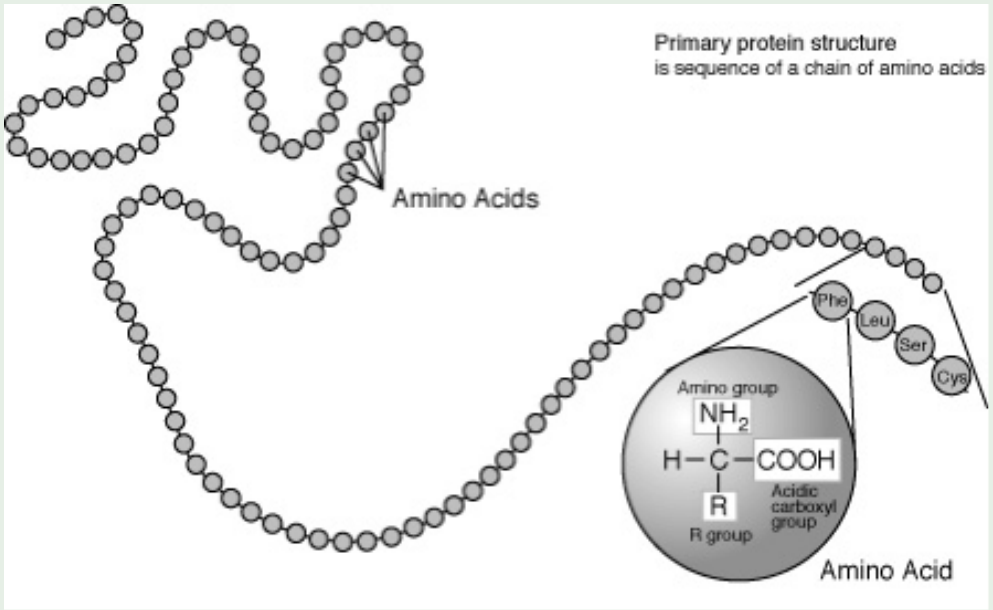
Wh



.

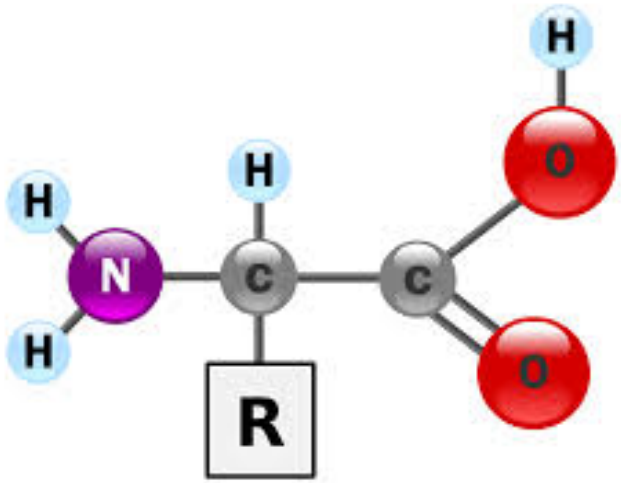
A necklace of beads

- Each bead is an amino acid
- Amino acids have an structure

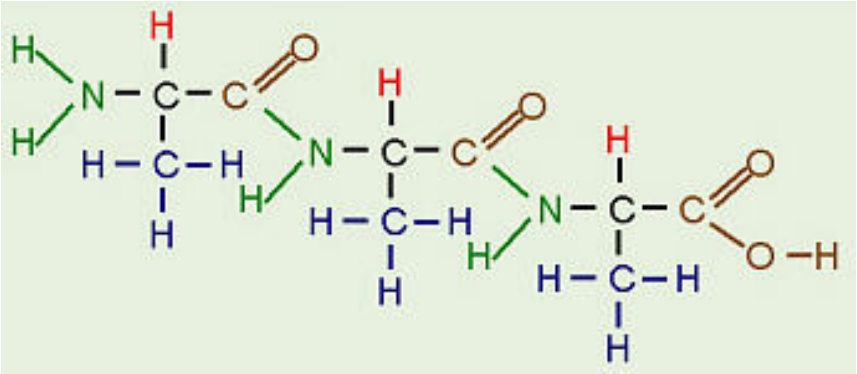


Amino Acid Strcuture

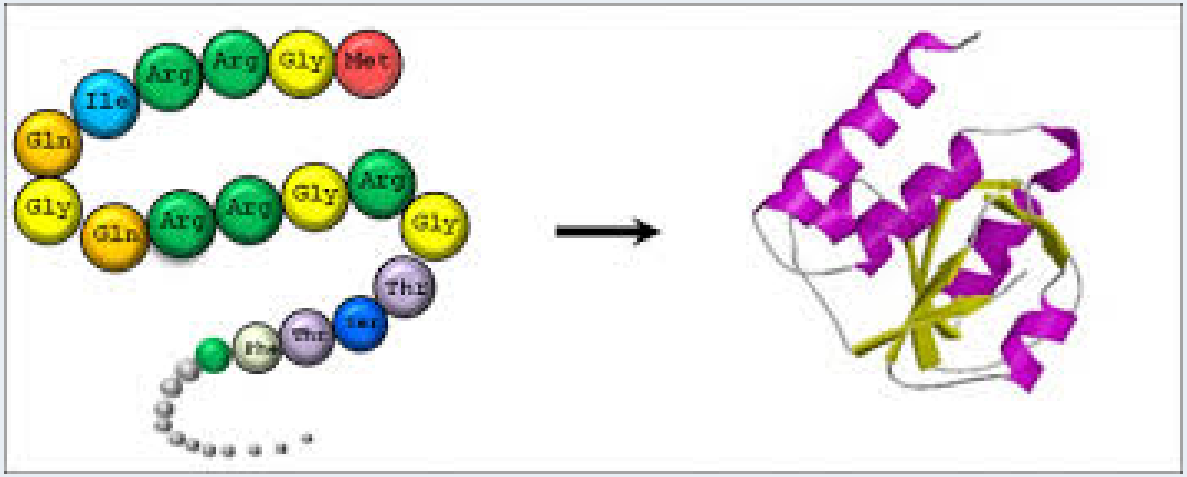
- Composed by atoms
  - Carbon,
  - Oxygen,
  - Hydrogen, and
  - Nitrogen



Chains of Amino Acids Form 3D Structures

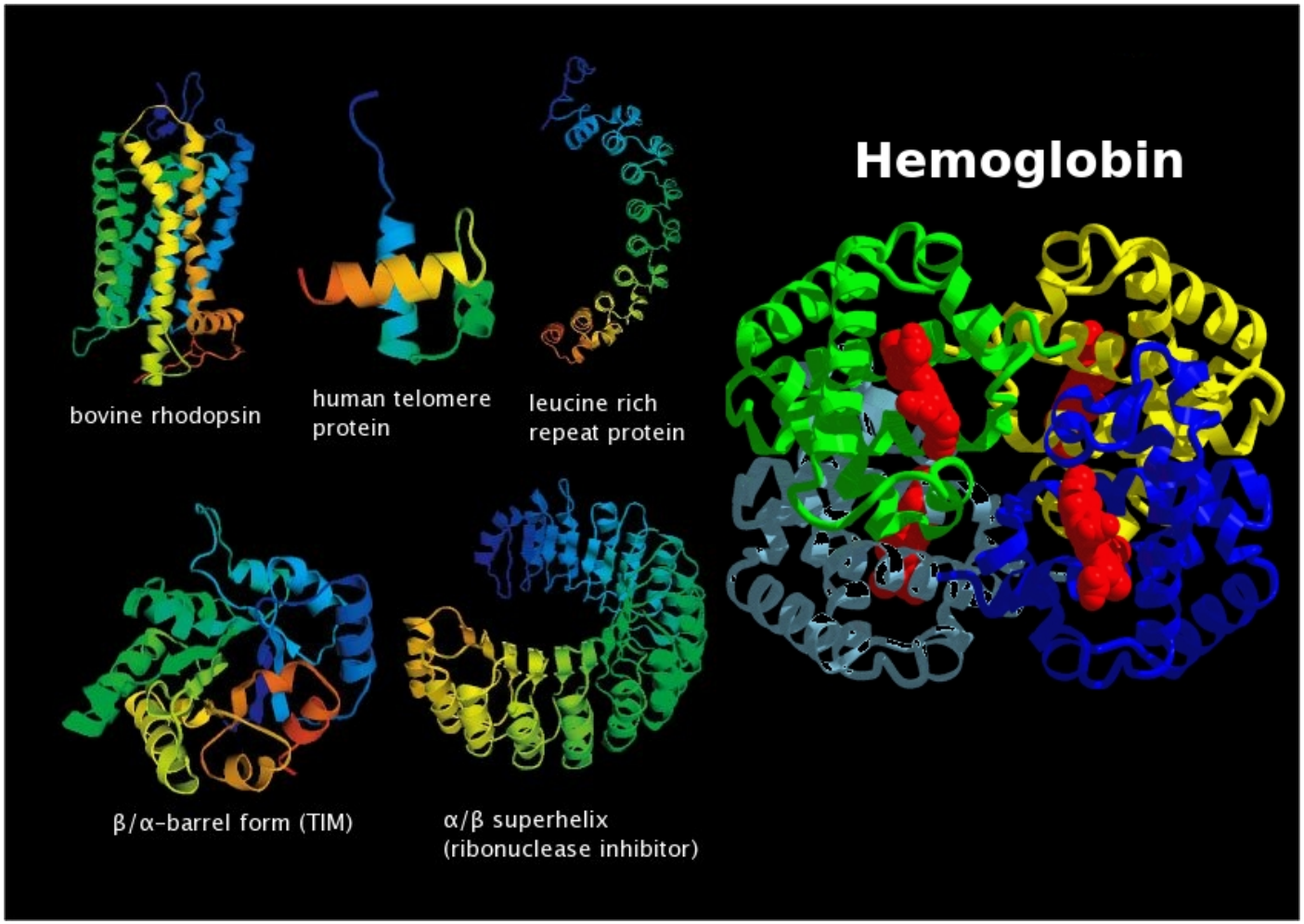


Proteins form Tridimensional Shapes



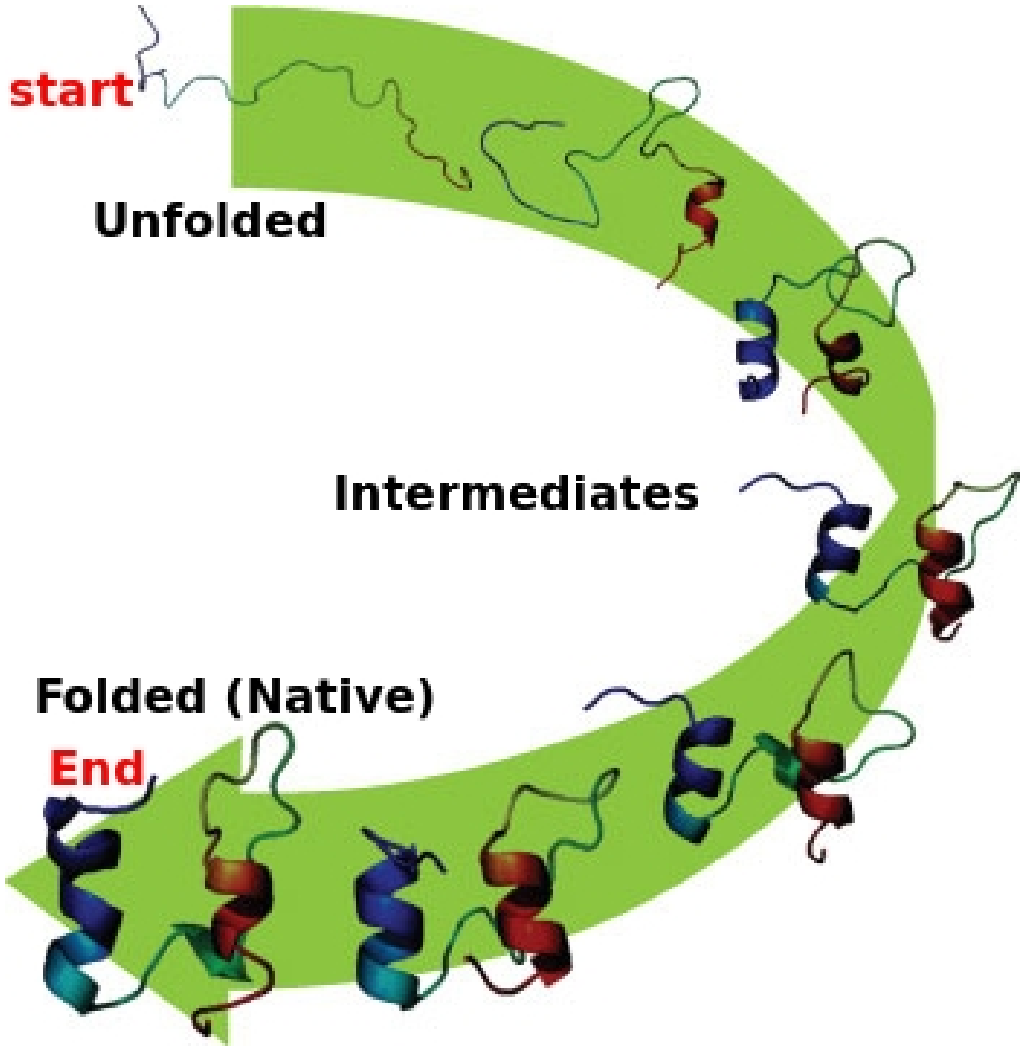
Examples of Proteins

- Proteins take many 3D structures (shapes)
- The shape is associated with the function
- Hemoglobin transports oxygen



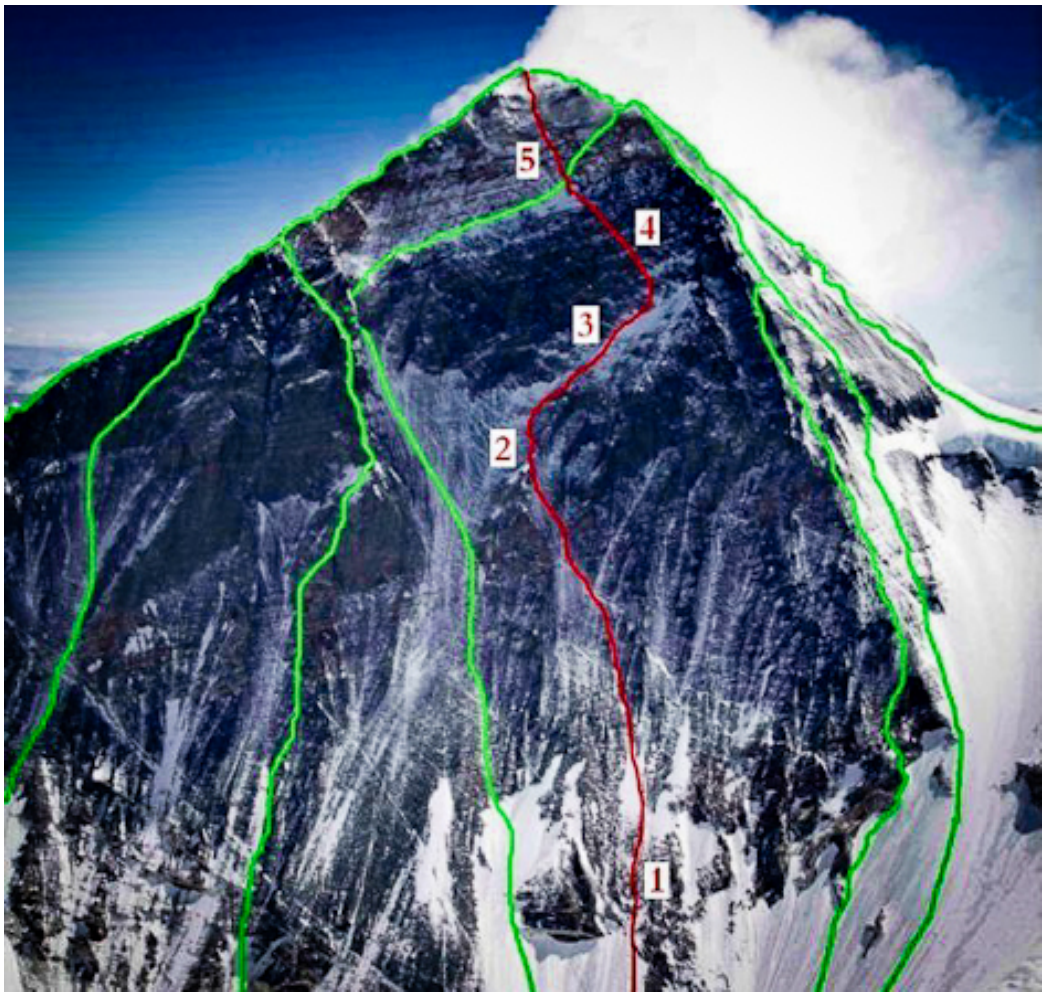
Protein Folding Process

- Complex process in biology
- Hard to understand by scientist
- Challenge for biologists, biophysics, and computer scientists
- Very Important:
  - Medicine
  - Drugs discovery
  - Treatment of diseases
  - Among others...

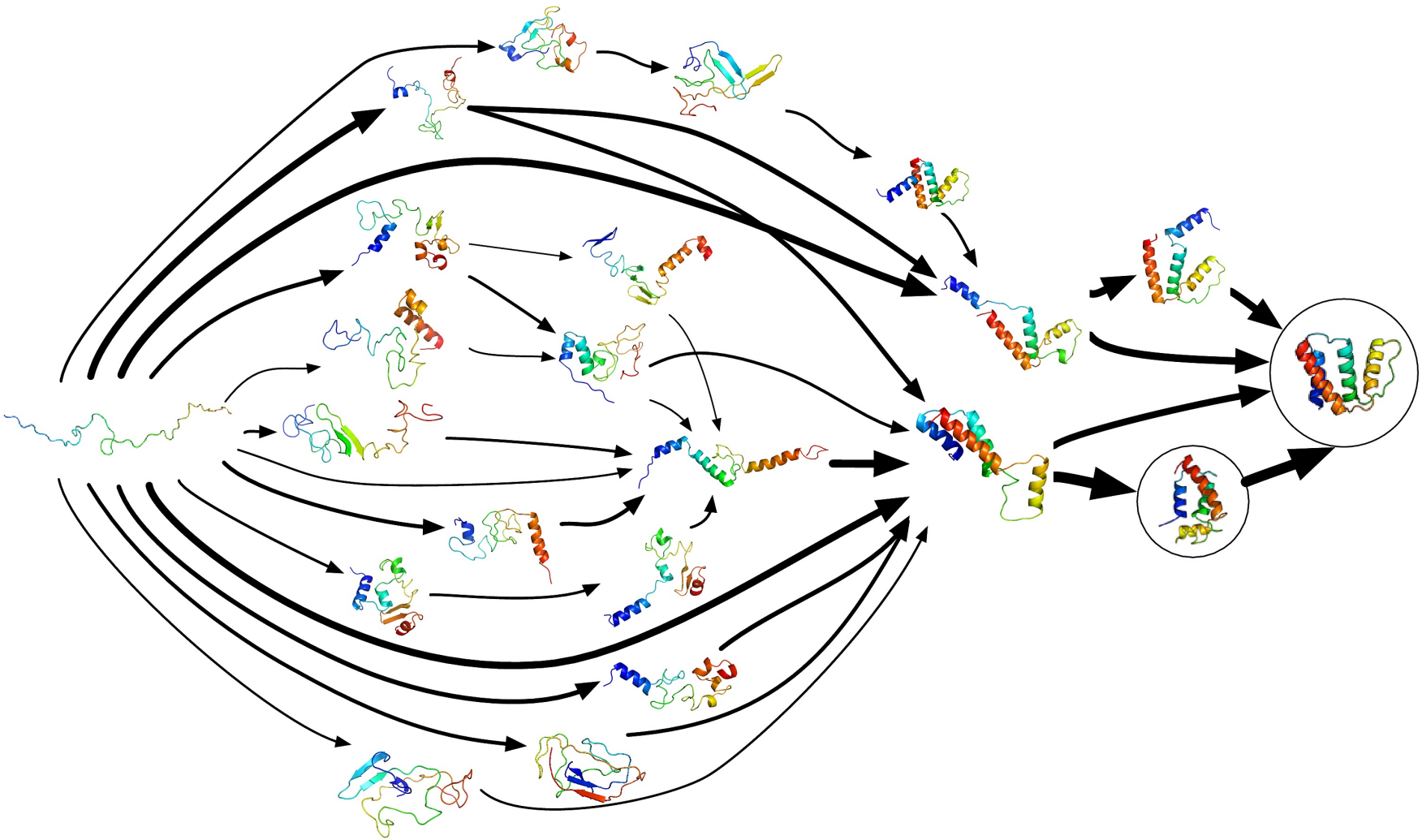


Protein Folding as Mountain Climbing

- Bottom: unfolded state
- Points : intermediate states
- Top : folded or native state



Protein Folding Pathways



The Problem

Protein folding pathways and folding intermediates have not been observed experimentally.

Our Hypothesis

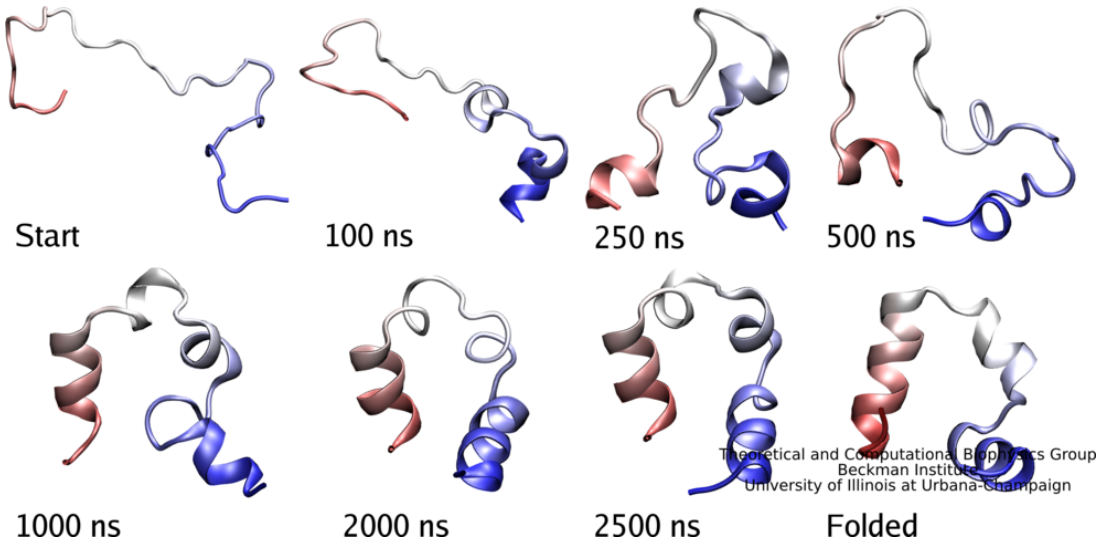
If we can measure features related with folding process of a protein, we can determine the status of a protein during its folding, and so we can observe if the protein follows a pathway and if this pathway has intermediates.

Data

Protein Folding Simulations

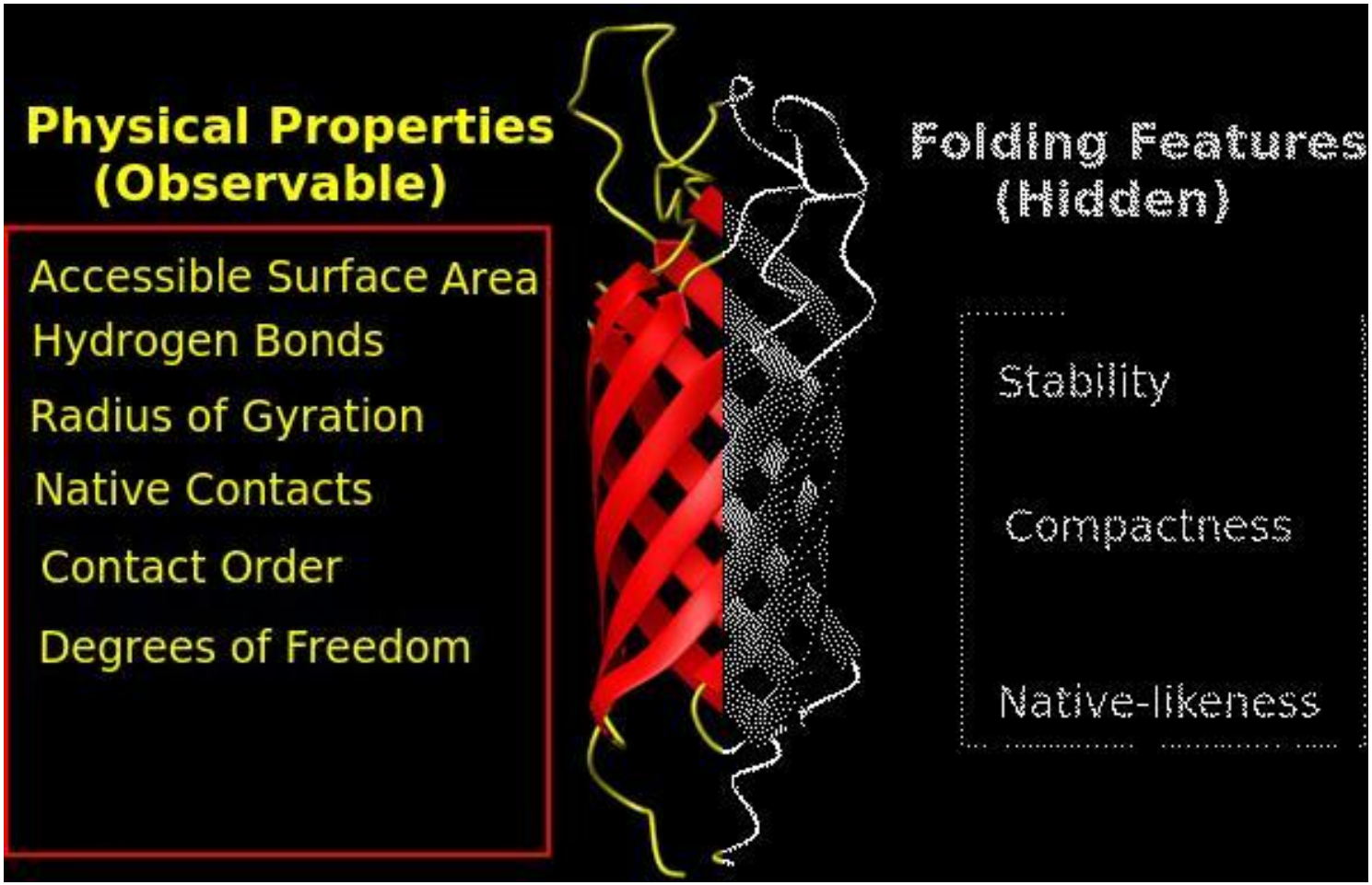
Two types of Simulation Techniques:

- Molecular Dynamics
- Probabilistic Roadmap Method



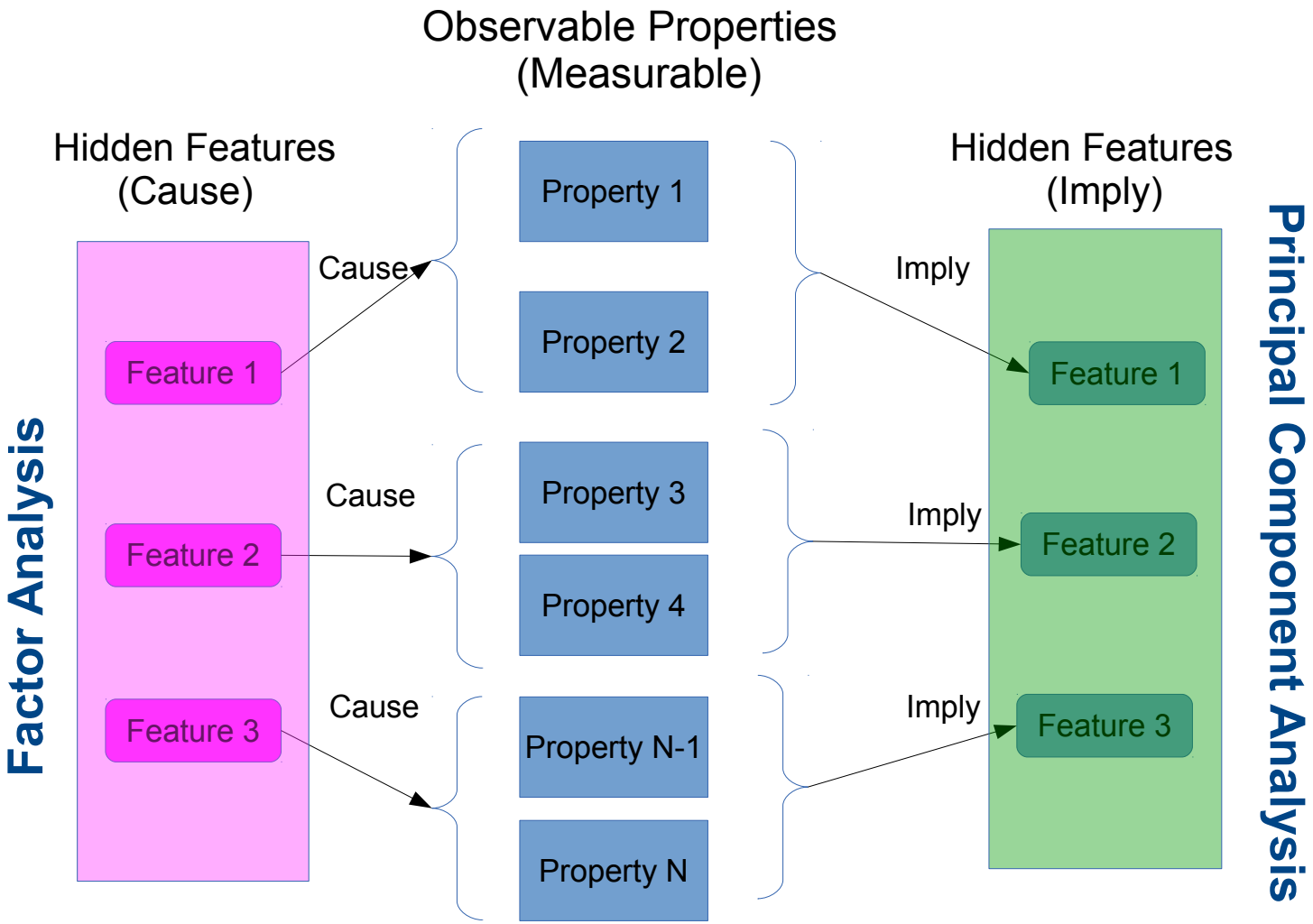
Methodology

Discovering hidden features from observable physical properties



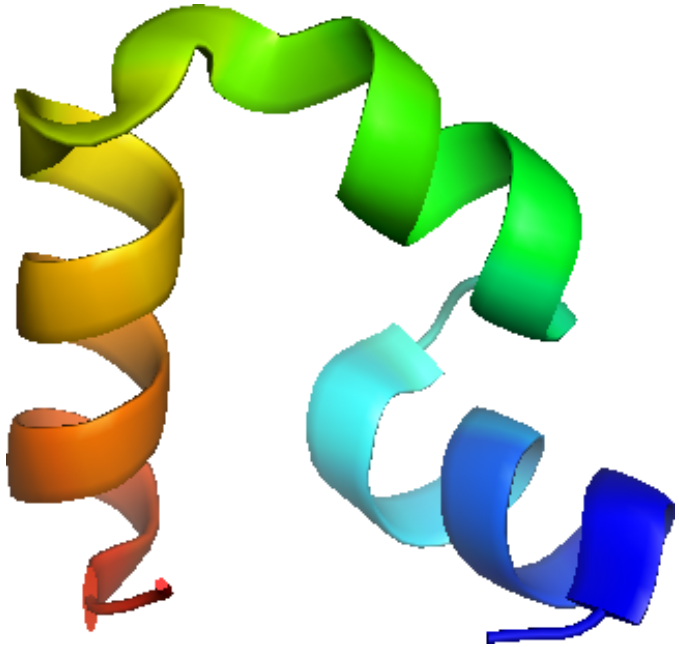
Methods

Multivariate Analysis



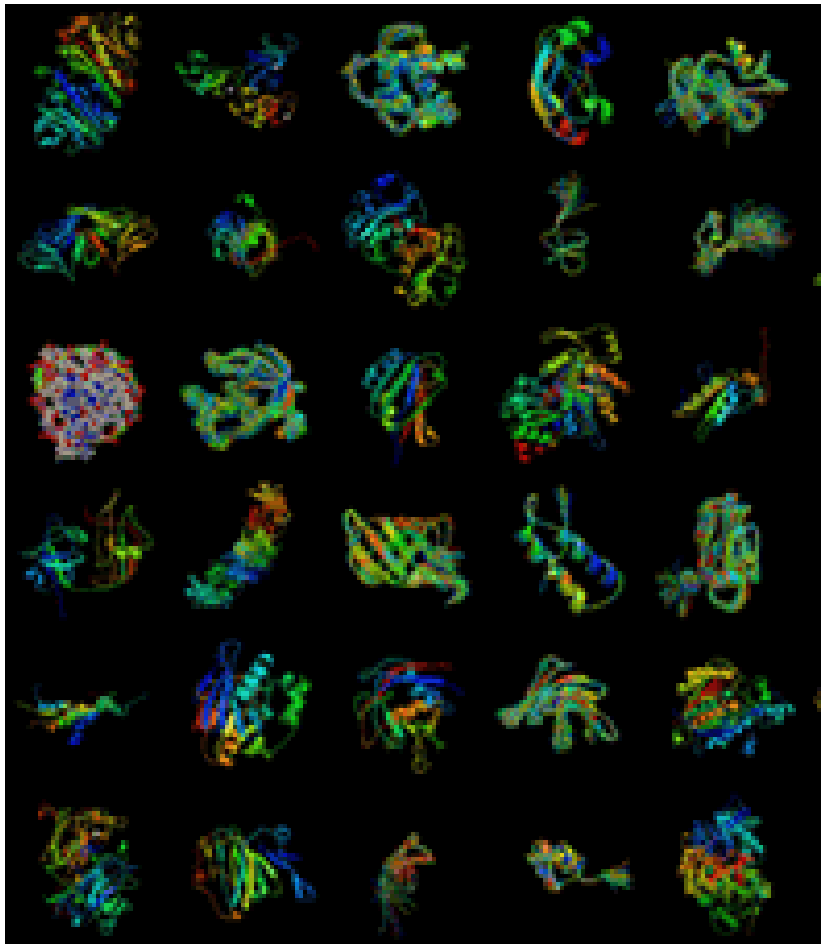
A Preliminary Analysis with The Villin-headpiece Protein

- Molecular Dynamics
- 32 Amino Acids
- 9 physical properties
- Factor Analysis: Failed!!
- Principal Component Analysis: Succeeded



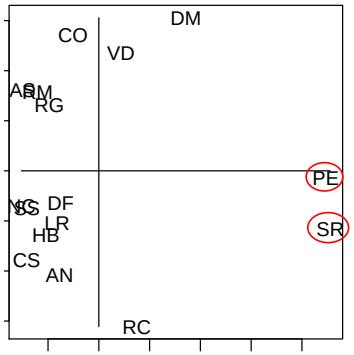
A Full Analysis with 27 Proteins

- Probabilistic Roadmap Method
- Different topologies and sizes
- 16 physical properties
- Principal Component Analysis among other methods for data reduction

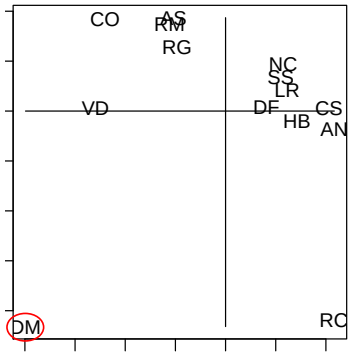


A Visual Analysis

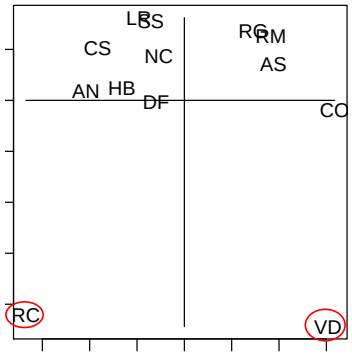
Four Analysis using Multidimensional Scaling (MDS)



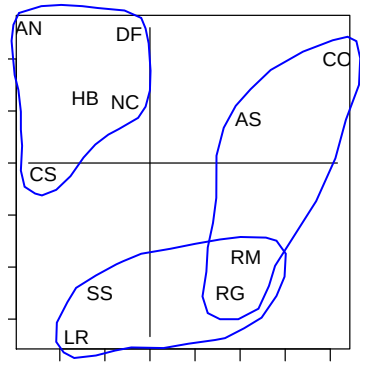
(a) 16 properties



(a) 14 properties



(a) 13 properties

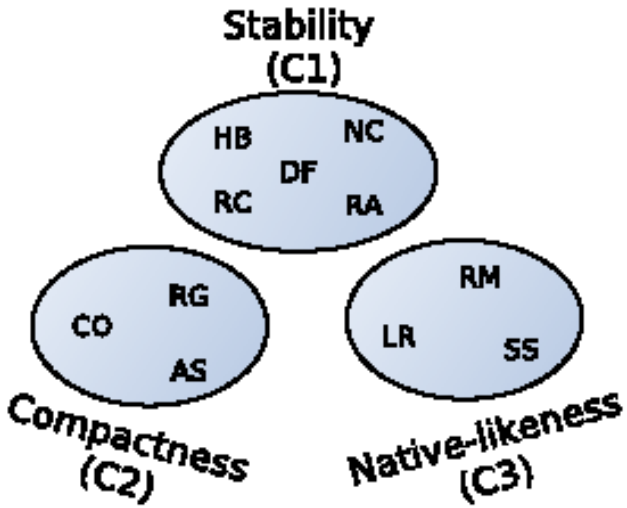


(a) 11 properties

Principal Component Analysis

		Components								
Property	ID	C1	C2	C3	C4	C5	C6	C7	C8	C9..C16
Native Contacts	NC	.72								
Contact Order	CO		-.93							
Radius of Gyration	RG		.64							
Hydrogen Bonds	HB	.72								
Access. Surface Area	AS		.82							
Root Mean Square Dev.	RM			-.61						
Local Root Mean Sq. Dev.	LR			-.84						
Residues in Correct SSEs	RC	.73								
Residues in Any SSEs	RA	.89								
Structural Score	SS			.81						
Degree of Freedom	DF	-.73								
Potential Energy	*PE				1.0					
Dipole Moment	*DM						.94			
Voids	*VD							.78		
Rigid Cluster	*CL					.83				
Stressed Regions	*SR								.99	
Proportion of Variance		.63	.09	.06	.06	.05	.03	.02	.03	
Cumulative Variance		.63	.72	.78	.84	.89	.92	.94	.97	

(a) Matrix of Loadings



(b) Folding Features

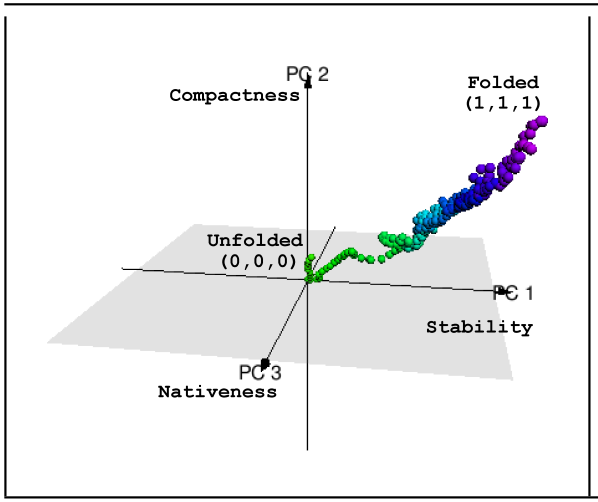
The Folding Status of A Protein Conformation

$$
\begin{aligned}
F_1 &= p_1 * C_{1,1} + \ldots + p_k * C_{1,k} + \ldots + p_{11} * C_{1,11} \\
F_2 &= -(p_1 * C_{2,1} + \ldots + p_k * C_{2,k} + \ldots + p_{11} * C_{2,11}) \\
F_3 &= p_1 * C_{3,1} + \ldots + p_k * C_{3,k} + \ldots + p_{11} * C_{3,11}
\end{aligned}
\tag{1}
$$

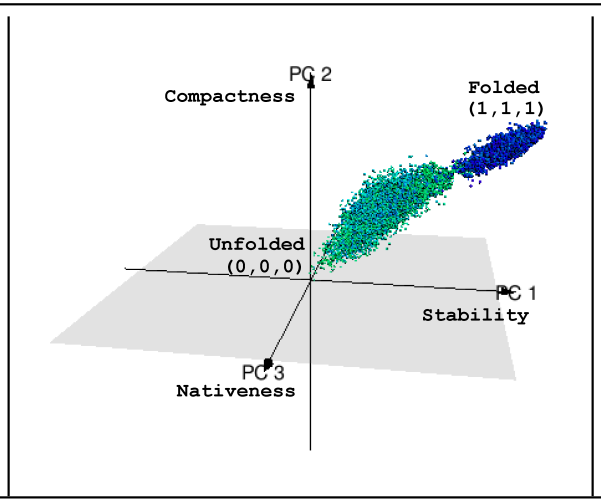
The Folding Status

The vector of three features: $[F_1, F_2, F_3]$,
we called as the Folding Status of a Protein conformation

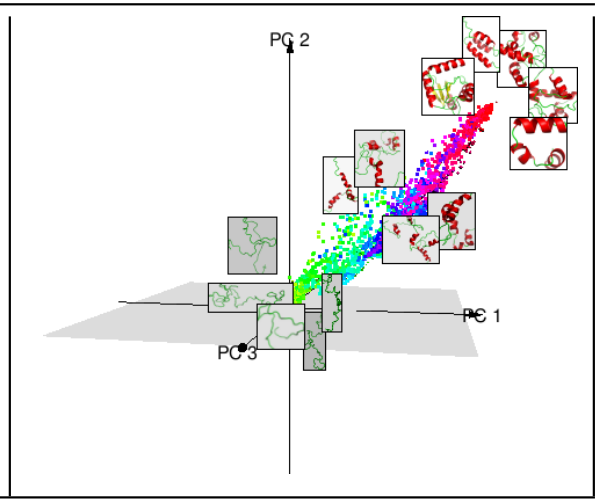
3D Space of Features For Protein



(a) Protein with visible intermediates

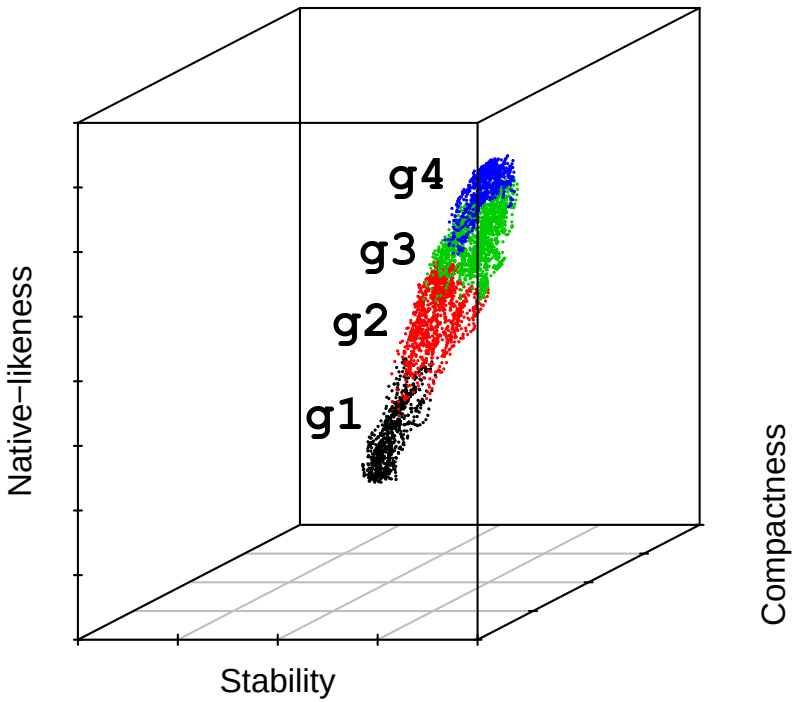


(b) Protein without intermediates

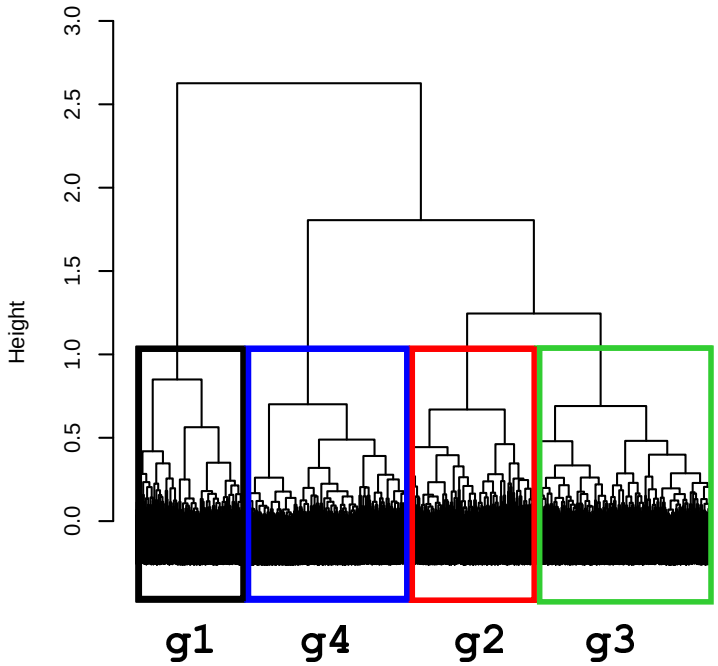


(c) Multiples proteins

Organization of Protein Conformations

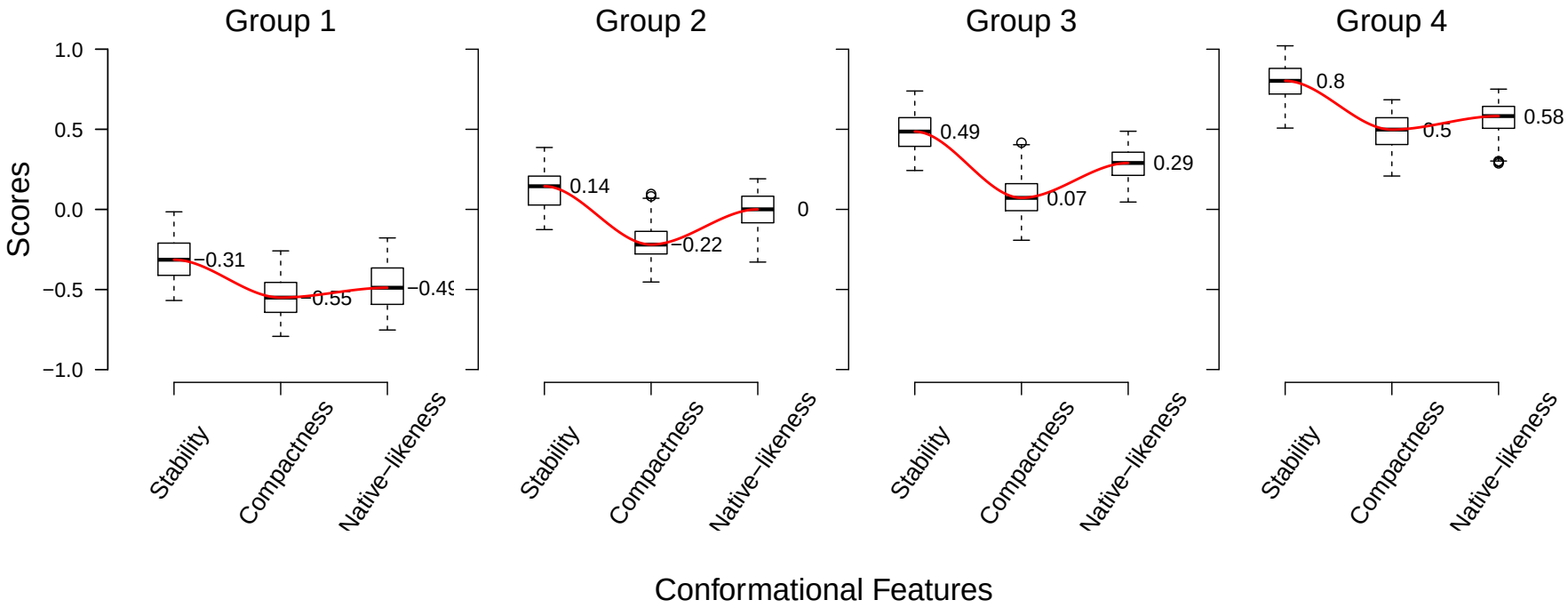


(a) 3D Visualization



(b) Cluster dendrogram

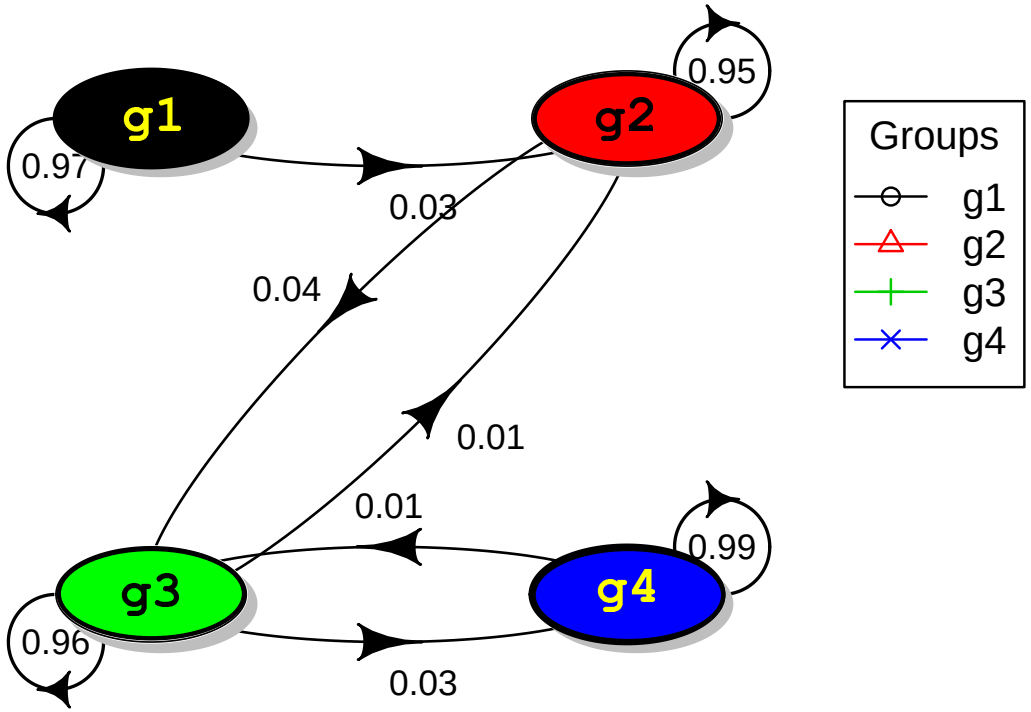
Distribution of the Features on the Groups



Dynamic Behavior of Groups

Dynamic assignation of groups as a Markov chain

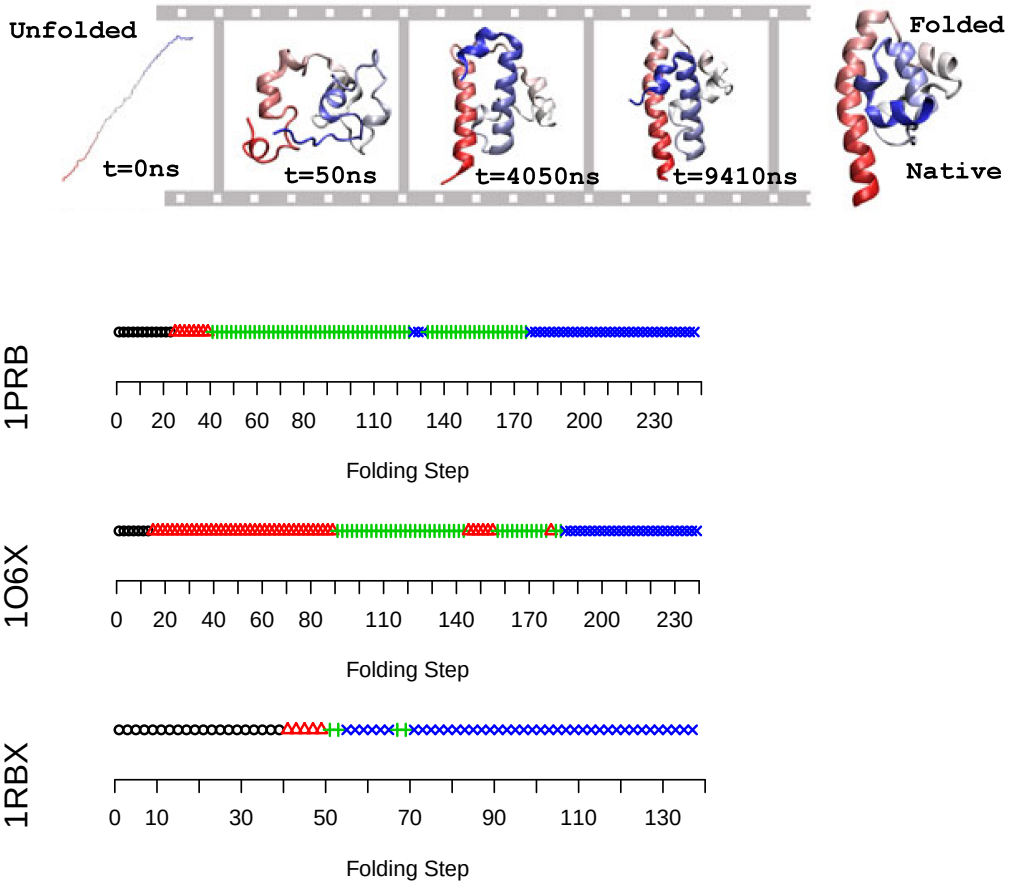
- An initial state: Group 1
- Two intermediates: Groups 2 and 3
- And a final state: Group 4



Groups on Pathways

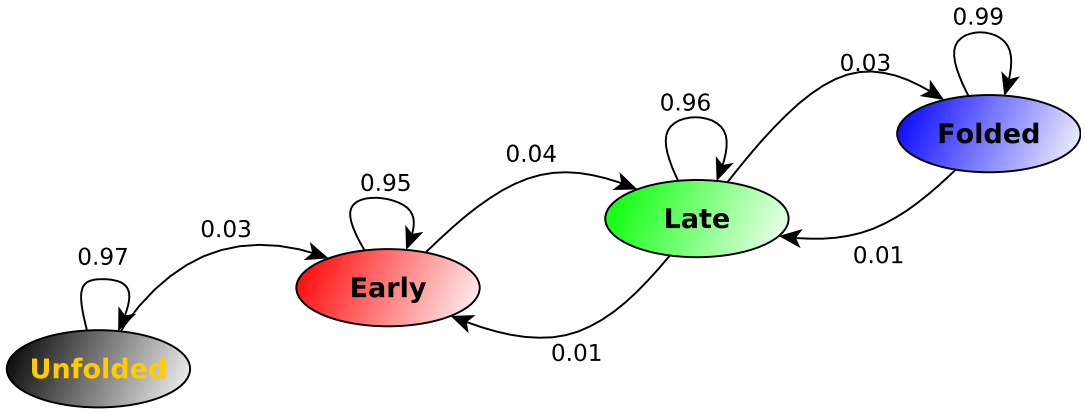
A sample of three pathways

- Step 0: unfolded structure
- Last step: native (folded) structure

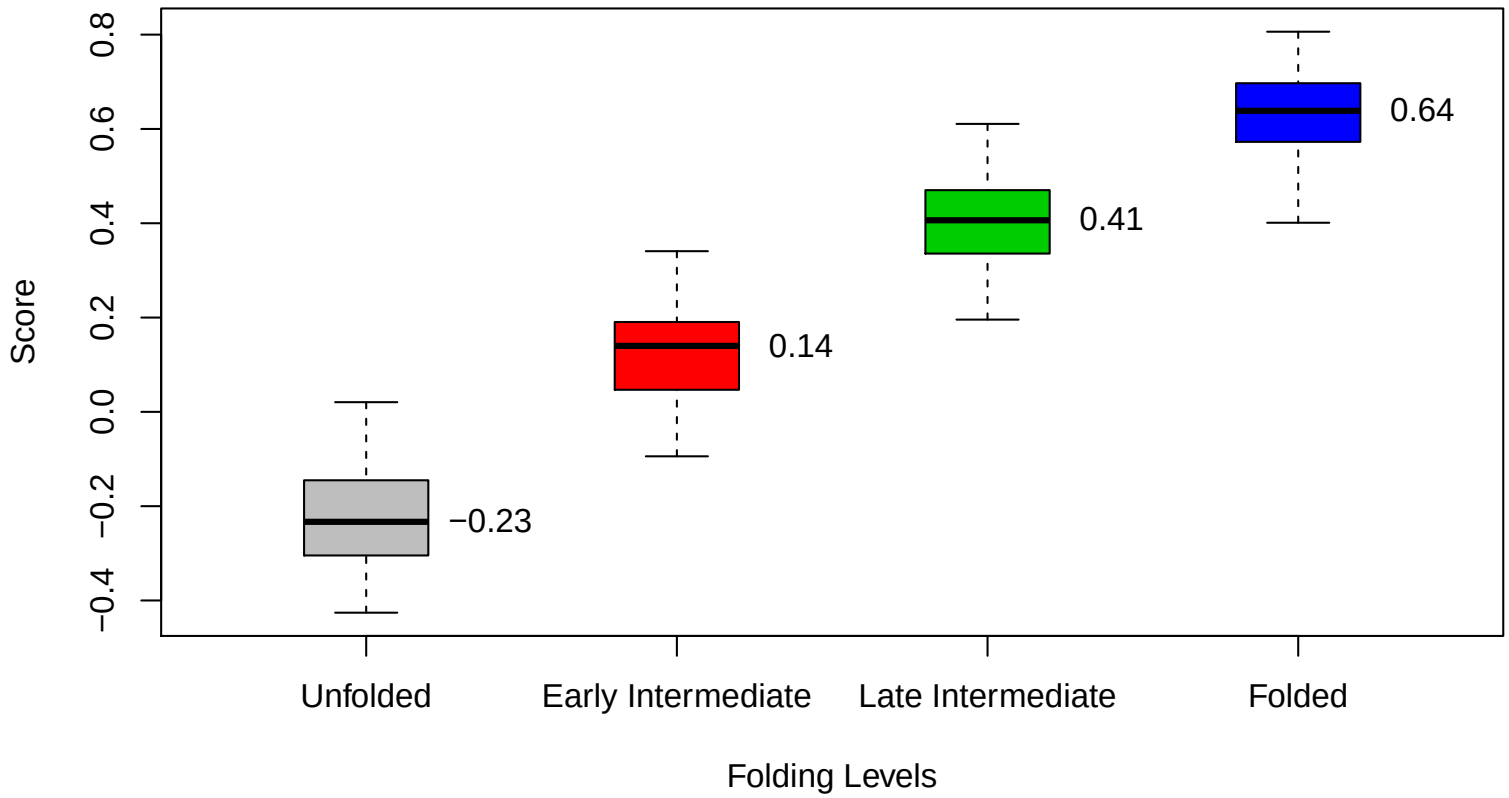


Protein Folding Levels

- Four folding levels
  - Unfolded
  - Early intermediate
  - Late intermediate
  - Folded
- A global pathway:
  - Sequence of folding states
- Folding Intermediates:
  - Early and late



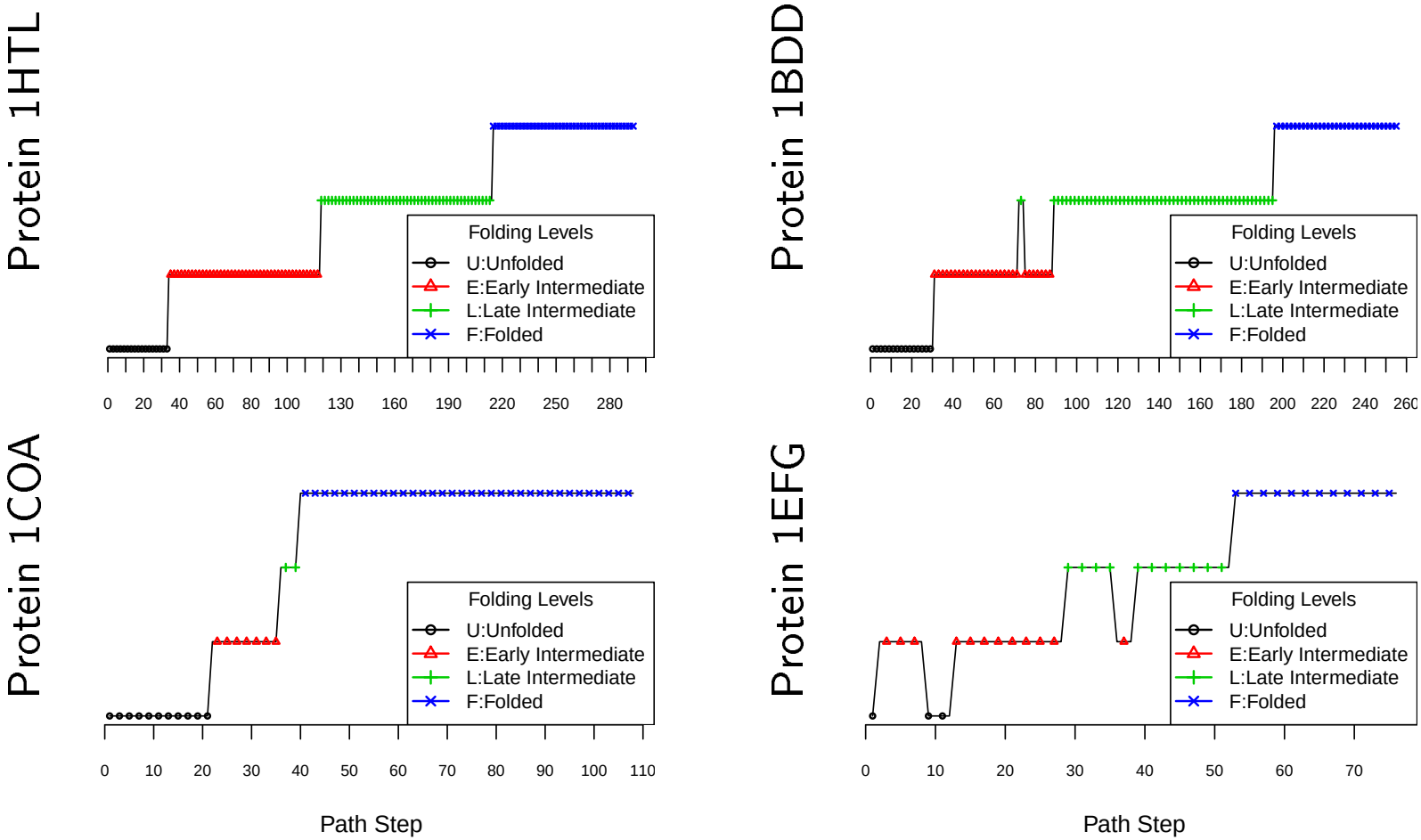
Folding Level and the ICF Score



A Protein Folding Classificator

$$lev(c) = \underset{g_k}{argmin}\, \delta(c, g_k) | k = 1, ..., 4$$

Predicted Folding Levels for Conformations



Conclusions

- We can measure hidden folding features
- Theoretical evidence of Protein Folding Pathways
- Theoretical evidence of Folding Intermediates
- We provided a set of theoretical tools to analyze protein folding
- Further work:
  - Verify our results with Molecular Dynamics simulations
  - Verify experimentally our results