



The University of
Nottingham

Protein Structure Prediction and Analysis

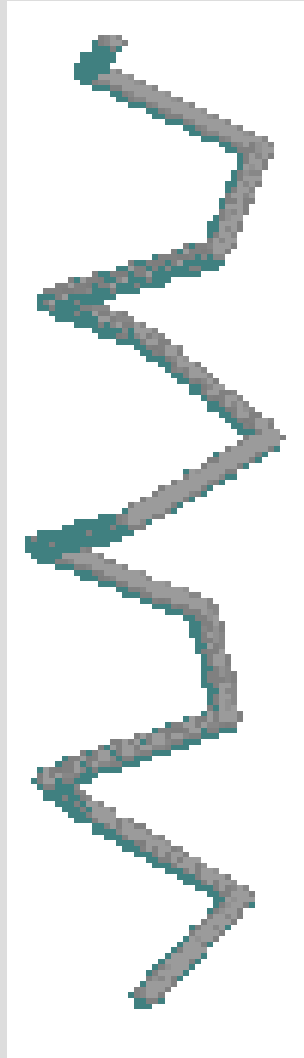
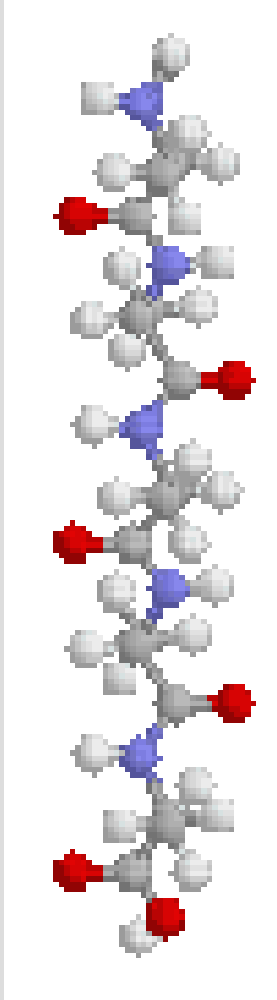
Pooja Jain

19 September 2006

Overview

- Protein Structure – Background
- Protein Structure – Prediction
- Participation in CASP 7
- Analysis of Protein SSEs Contacts
- Results
- Conclusions
- Future Work

Levels of Protein Structure



Primary Structure

Amino acid residue sequence

Secondary Structure

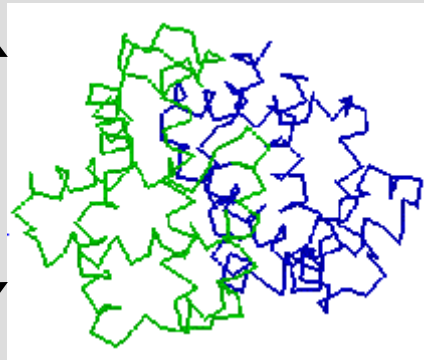
Local arrangement of residues
such as – Helix, Beta Sheets
and
loop

Levels of Protein Structure



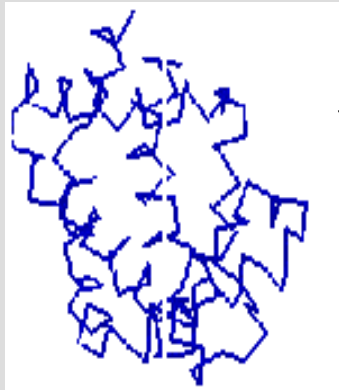
Tertiary Structure

Spatial conformation of local structures

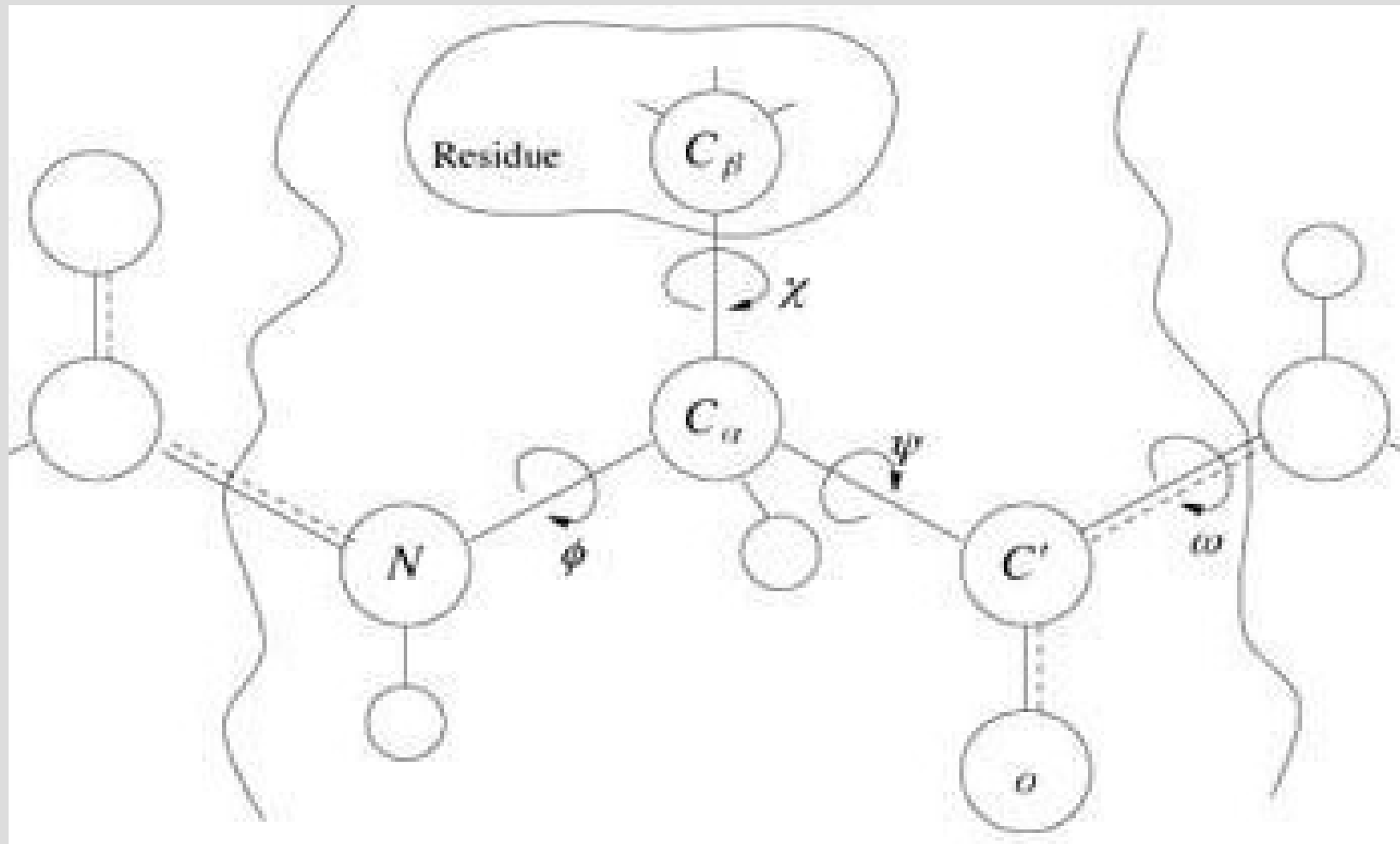


Quaternary Structure

Spatial relationship among multiple chains



Protein Backbone



What Determines Structure ?

- Hydrophobic effect – Strongest determinants of protein structures
- Hydrogen bonds – Essential in stabilising the basic secondary structures
- Van der Waal forces – Stabilises the hydrophobic core
- Electrostatic forces – Stabilises oppositely charged residue side chains

Problem

Problem : Given the amino acid sequence of a protein
what's its shape in 3D ?

Subproblems :

- Secondary structure prediction
- Residue-residue contact prediction
- Dihedral angles prediction

Why is Prediction Needed ?

- Function of a protein is influenced by its structure
- Structure is more conserved than sequence
- Time consuming and costly experimental methods
- Large gap between available protein sequences and structure

Growth of Protein Sequences & Structures

The latest update of PDB shows **35480** resolved structures of proteins (12-Sep-06)

The latest release of Swiss-Prot shows **195954** sequence entries (05-Sep-06)

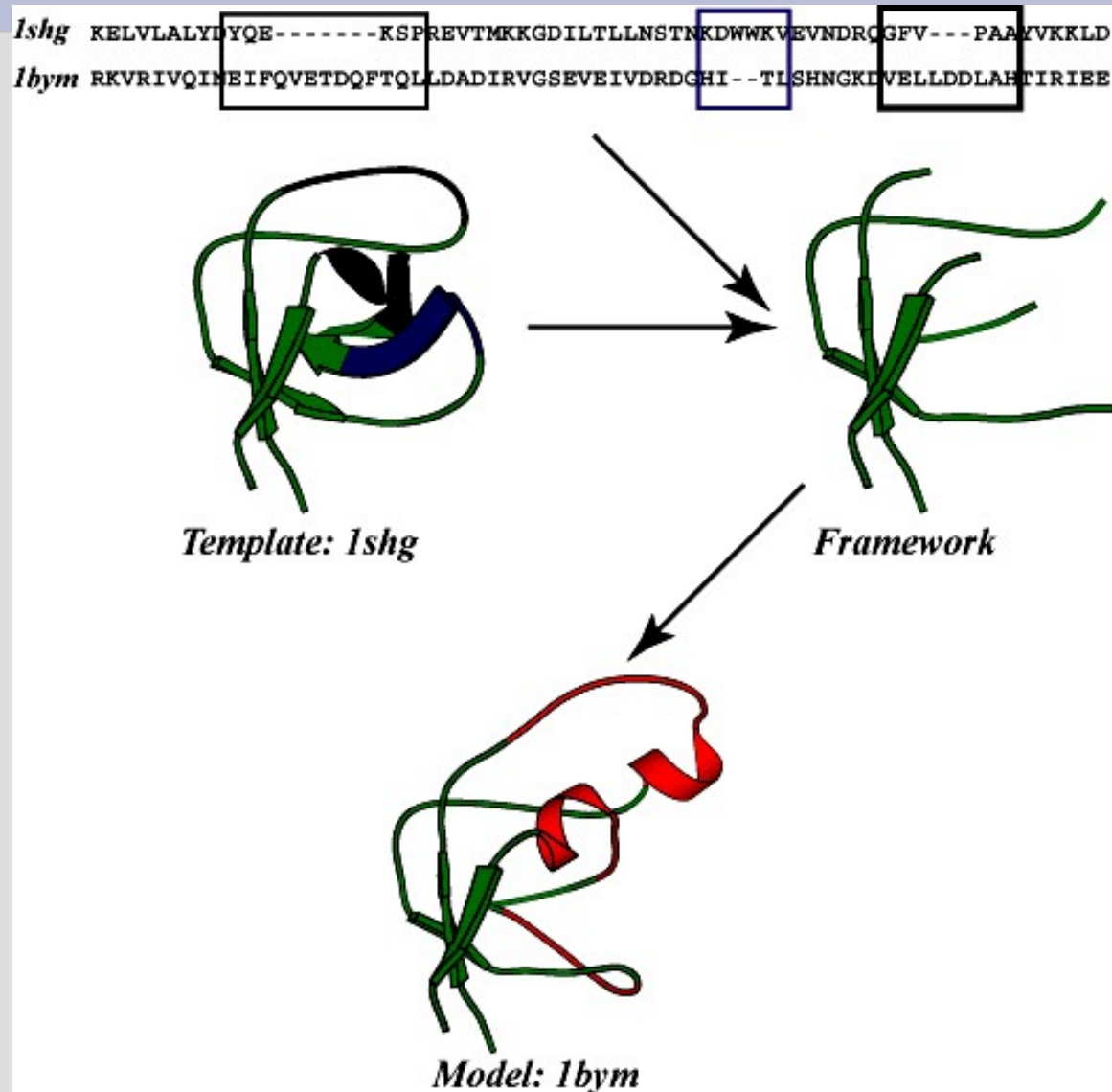
Target Sequence Category

- **Homology Modelling Targets (HM – Targets)**
 - Easy targets – Close homologous protein sequences (>75% Similarity)
 - Hard targets – Distantly related protein sequences (>30% Similarity)
- **Fold Recognition Target (FR – Targets)**
 - A target for which a protein with known structure, sharing the same fold exists (20-30% Similarity)
- **New Fold Targets (NF – Targets)**
 - No homologous or protein with the same fold exists.

Prediction Methods

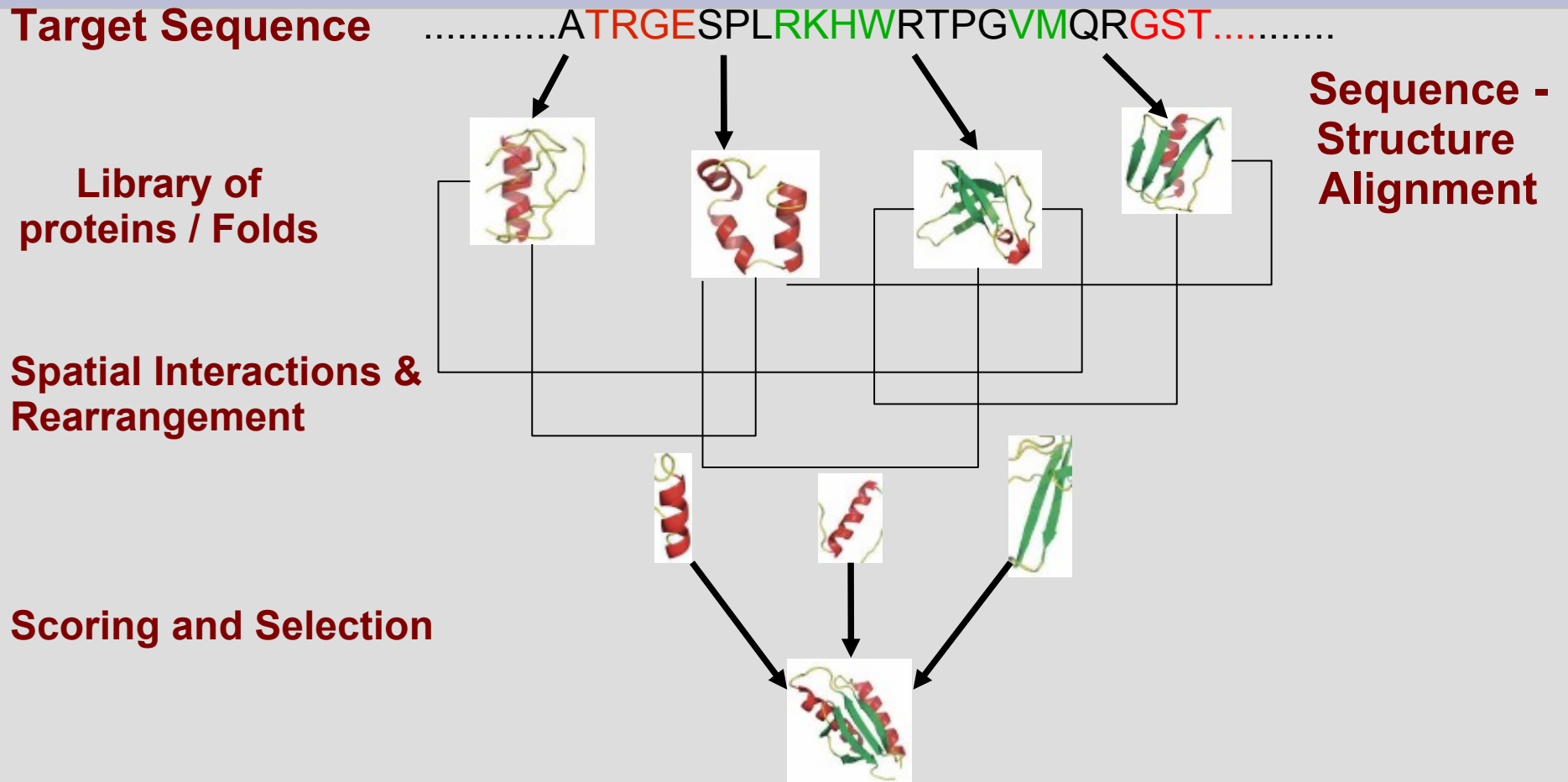
- Homology Modelling – Sequence based methods for HM targets
- Protein Threading – Sequence-structure alignment methods for FR targets
- Consensus Methods – Vote a prediction from some candidates generated by different prediction programs
- De- Novo – Build a structure without any prior knowledge i.e., without referring to an existing structure

Homology Modeling: How it works?



1. Find template
2. Align target sequence with template
3. Generate model:
add loops
add side-chains
4. Refine model

Protein Threading: How it works?



De novo Prediction

Given the amino acid sequence predict the structure based on the energetic or statistical principles.

- Assumption 1: All the information about the structure of a protein is contained in its sequence.
(Anfinsen 1960s)
- Assumption 2: Native structure is at the global free energy minimum. Therefore, search is limited to the conformation space with low global free energy
- Finding native-like conformation requires
 - A scoring function (Potential)
 - A search strategy (optimisation algorithms)

C
A
S
P
7



CASP 7 : Participation

Goal : Evaluation of Methods in the area of protein structure prediction where algorithms are tested against resolved, but unpublished proteins.

- 100 targets have been released
- 30 targets < 150 residues in length were considered

3D atomic coordinates (Tertiary Structure) prediction

- Predict 3D structure from 1D sequence data (***de-novo***)
- Simulating the protein folding process (MD)

C
A
S
P
7



CASP 7 : De-novo prediction

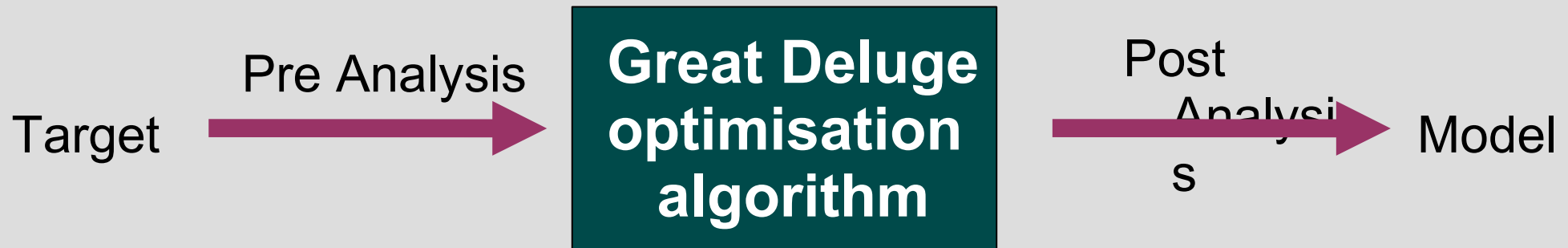
- Great Deluge local search optimisation algorithm applied to all(extended)-atom model
 - Energy function - a weighted sum of Lennard-Jones, Electrostatic, Hydrogen-Bond and Hydrophobic terms
 - Parameters are taken (mainly) from CHARMM 22 and modified during the verification of the algorithm.

C
A
S
P
7



CASP 7 : Strategy

Great Deluge optimisation algorithm applied to all (extended) - atom model



Pre Analysis : Assumptions

- Distantly related proteins can have similar structure
- Solvent accessibility of amino acids guides protein folding
- Similar sequence yields similar secondary structure

Pre Analysis -Tools

- PSI-BLAST- Position-Specific Iterative BLAST
- DESTRUCT – Predicts Protein secondary structure with dihedral angles
- DSSP- Define Secondary Structure Of Proteins
- SABLE - Prediction of Solvent AccessiBiLitiEs

Pre Analysis -Results

Length : 76 residues

Target T0309: MASKKVHQINVKGF-FDMDVMEVTEQTKAEYTYDFKEILSEFNGKNVSITVKEENELP--VKGVEMAGDPLEHHHHHHH

DESTRUCT: CCCCCCHCCCCC-EHHCHHHHHCCCCCCECCCCHEEEEEHCCCCCHEHEHCCCC-CHCCECCCCCCECCCCC

PSI-/1TF7|F: SRAINVFKMRGSHWHDKAIREFMISDKGPDIKDSFRNFERIISGSPTRITVDEKSELSRIVRGVQ-EKGPESHHHHHH

Alignment: S+++ ++G+D+E K+F+ +G ITVE++EL V+GV+ P HHHHHH

DEST/1TF7|F: CCCCCCHHCCCCCCHHEEEEECCCCCCECCCHHHHEEEEECCCCCECCCHHCHHHHHHHCCC-CCCCCCCCC

DSSP/1TF7|F: EEEEEEEEESSS---B-EEEE-SS-EEE---TTBS--TTSS--B-

Solv Acc: 4444513413142-11323124235545534232414411452444313132544452--4253142244414334544

0 -> fully Buried

9 -> fully Exposed

Row 1 – CASP 7 Target Length

Row 2 – CASP7 Target Sequence

Row 3 – Target Secondary structure prediction by DESTRUCT

Row 4 – PSI BLAST Hit , target against homologous protein

Row 5 – PSI BLAST Alignment or sequence profile

Row 6 – Homologous protein secondary structure prediction by DESTRUCT

Row 7 – Homologous protein secondary structure assignments by DSSP

Row 8 – Target Solvent Accessibility by SABLE

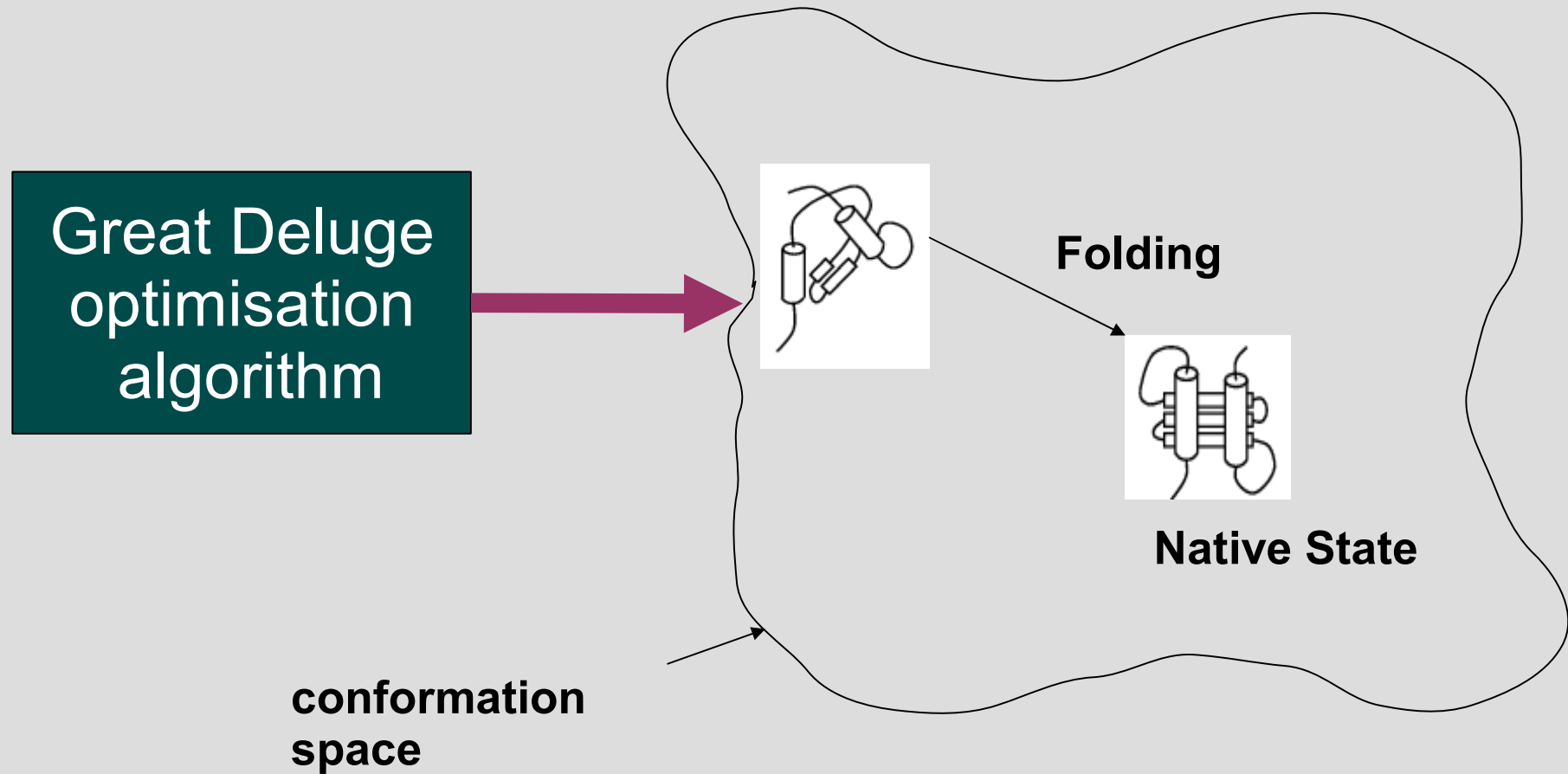
Results were used to select an appropriate model from Great Deluge Algorithm

Post Analysis: Assumptions

The large number of degrees of freedom in an unfolded polypeptide chain results in the astronomical number of possible conformations for it. (*The Levinthal paradox*)

A particular polypeptide chain may be able to assume multiple conformations depending on its environment, and the biologically active conformation *may not* be the most thermodynamically favorable.

Post Analysis: Assumptions



Post Analysis

Mainly involves:

- Simulation of the protein folding process using CHARMM (Chemistry at HARvard Macromolecular Mechanics)
- Analysis of the simulation results
 - Surface Accessibility RMSD
 - Secondary Structure Similarity
 - Radius of Gyration

Post Analysis : MD Protocol

- Simulations were performed with CHARMM28 using the CHARMM19 all-atom potential energy parameter set with Generalized Born Implicit Solvent system at 298K.
- The simulations were run for 5 ns (in blocks of 50 ps) and the conformational snapshots were saved at every 2 ps.

Post Analysis – Trajectory Analysis

For all of the sampled conformations:

- Obtained DSSP Secondary Structure Assignments
- Obtained DSSP Surface Accessibility values
- Calculated geometric Radius of Gyration

Compared against the values from pre analysis:

- Secondary Structure prediction from DESTRUCT
- Surface Accessibility values from SABLE
- Radius of Gyration calculated using the following formula

$$R_{\text{gyr}} = 2.83 \times N^{0.34}$$

Post Analysis : Results

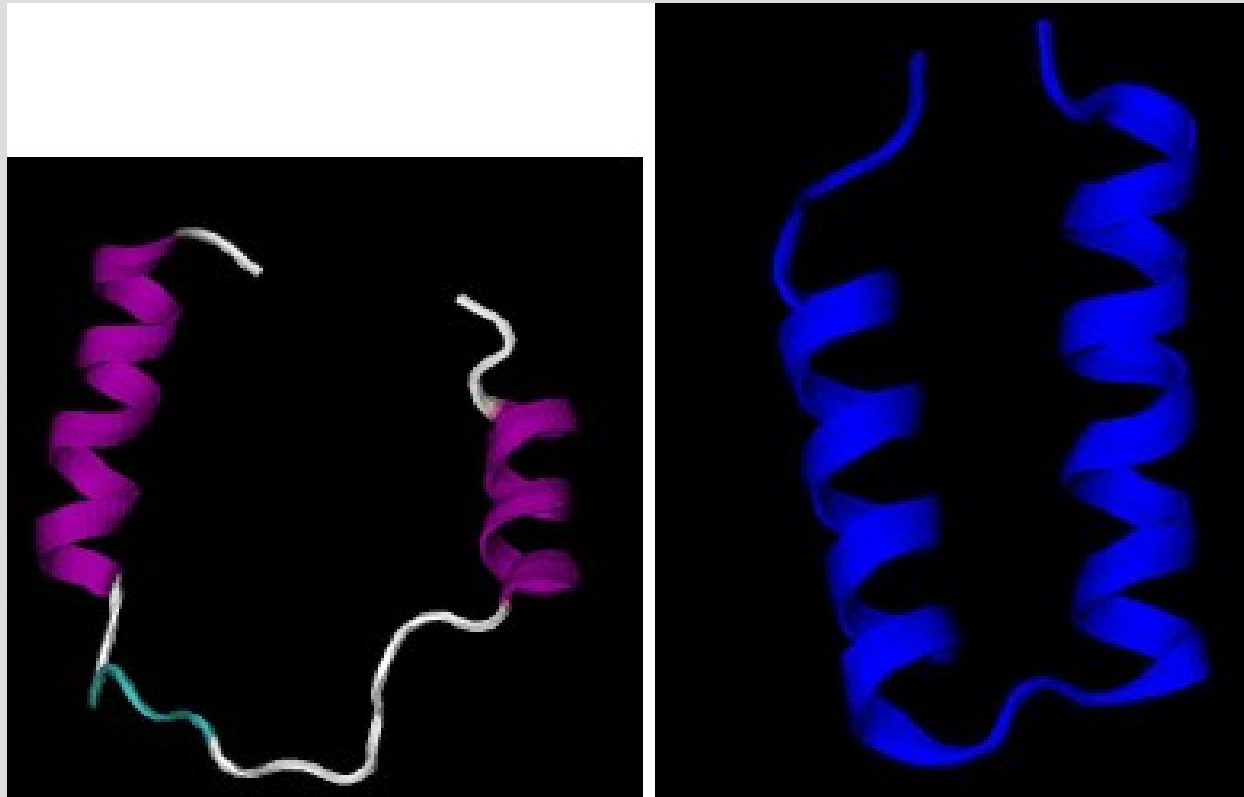
The Final conformation with:

- Minimum Surface Accessibility
- Minimum Radius of Gyration
- Acceptable secondary structure similarity

Post Analysis : Results

Target	Published Structure	Q_3	SOV	RGY_G^S	RGY_G^M	$RMSDC_\alpha$
T0288	2GZV	71.4	77.4	12.8	15.5	14.4
T0283	2HH6	71.4	75.4	16.7	17.4	17.5
T0300	2H3R	83.3	79.5	19.4	15.5	15.4
T0306	2HD3	57.9	60.4	13.9	12.6	13.5
T0309	2H4O	50.0	43.1	16.1	12.1	13.5
T0311	n/a	n/a	n/a	n/a	n/a	n/a
T0314	2HG6	47.2	44.8	15.2	13.4	15.3
T0327	2HGC	42.3	37.6	11.8	13.5	7.8
T0335	2HEP	71.4	61.8	11.0	12.7	6.8
T0348	2HF1	42.6	37.7	11.8	11.6	11.0
T0349	2HFV	56.0	47.8	17.9	12.2	12.5
T0350	2HC5	71.6	72.3	14.3	14.5	16.3
T0351	2HG7	58.3	45.3	11.8	14.6	10.3
T0352	n/a	n/a	n/a	n/a	n/a	n/a
T0353	n/a	n/a	n/a	n/a	n/a	n/a
T0363	n/a	n/a	n/a	n/a	n/a	n/a
T0366	2IWO	64.1	66.5	12.3	14.5	15.7
T0382	n/a	n/a	n/a	n/a	n/a	n/a
T0383	2HNG	44.8	37.5	17.5	14.6	11.3

CASP7 : Results

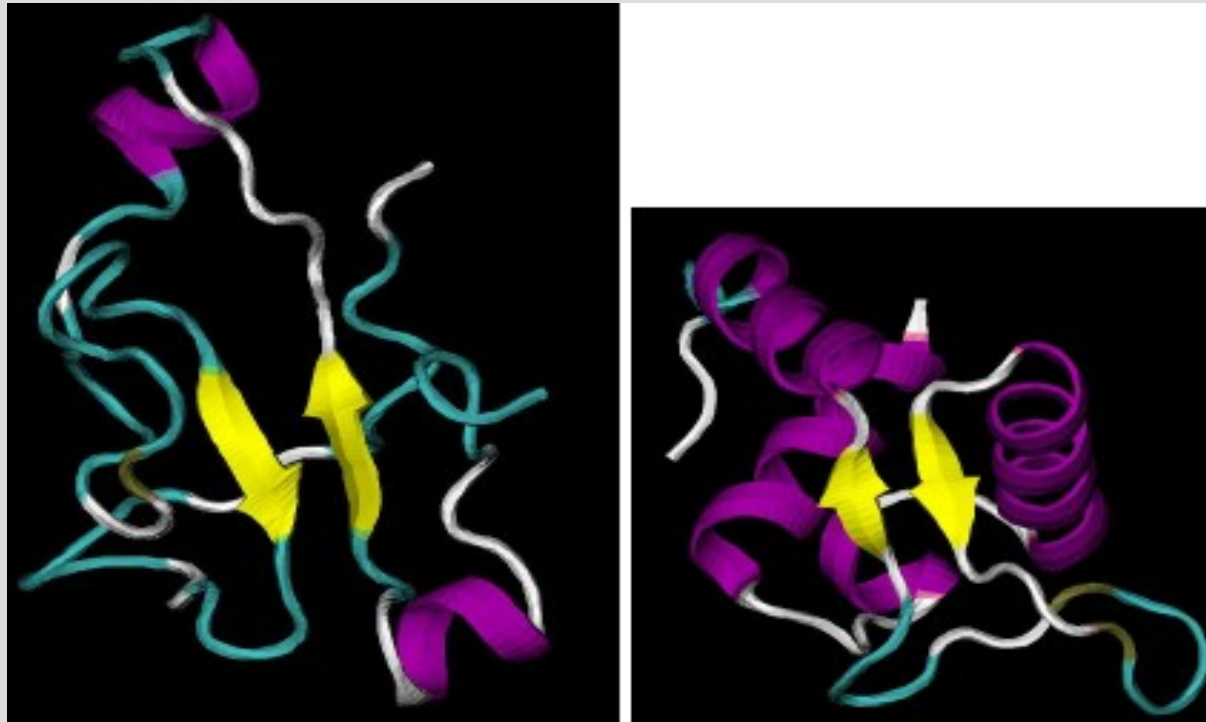


Target Structure
T0327

Published Structure

RMSDC-alpha : 7.8A°

CASP7 : Results



Target Structure
T0335

Published Structure

RMSDC-alpha : 6.8A°

Lessons from CASP7

- In-adequate preprocessing of target sequence
- Optimisation of search by considering radius of gyration
- Correct secondary structure prediction
- Inclusion of secondary structure packing rules
- Search should follow the hierarchical folding pathway

Exploring Protein Contacts

“ Is it possible to study the similarity among two proteins based on the mutual contacts between its secondary structure elements (SSEs) ?”

Well known facts :

- Arrangement of SSEs (architecture) and connectivity (topology) governs the overall structure of proteins.
- Dissimilar sequence can have similar SSEs and same fold and overall 3D structure

Similar proteins, similar contacts between their SSEs

Contact Matrix

2D representation of 3D protein structure. Three categories were considered

- **Distance Matrix**
 - Pairwise distances between SSEs from COM
- **Orientation Matrix**
 - Relative orientation of SSEs in 3D space
- **Length Matrix**
 - Length of the SSEs in a protein
- Dataset of 77 proteins (3-helical family (Class:All Alpha)) consisting of only 3 helices was used

Correlation of Similarity

- 9 Structural descriptors for each of the proteins were derived from the 3 Matrices.
- RMS differences between respective category of structural descriptors were calculated.
- Against MAX-CMO similarity measure calculated from ProCKSi server.

ProCKSi-Server <http://www.procksi.net/>

Regression Study

- Regression Methods
 - Multilinear Regression (MLR) using R.
 - SVM (SMOreg) using WEKA (*Waikato Environment for Knowledge Analysis*)
- Population – 2926 pairs of proteins described by 21 structural descriptors

Regression Results : MLR

Multilinear Regression gave a percentage correlation coefficient (R²) of 66.2 with :

- Residual standard error: 0.07692 and
- p-value: $< 2.2e-16$

Regression Results : SVM

Complexity Parameter (C)	Kernel Parameter (E)					
	1.0		2.0		3.0	
	Train	Test	Train	Test	Train	Test
0.05	60.6	60.5	78.8	77.2	92.3	89.1
0.5	60.0	59.9	88.7	86.6	97.8	96.6
1.0	60.0	59.9	90.2	88.6	98.8	98.2
1.5	60.0	59.9	90.8	89.4	-	-
2.0	60.0	59.9	91.0	89.8	-	-
3.0	60.0	59.9	90.8	89.8	-	-

Table 1.1: Percentage Regression Coefficients (R^2) for different combinations of complexity parameter, C and kernel parameter, E. The E values of 1, 2 and 3 represent the linear, polynomial and RBF kernel respectively

Conclusions

Support Vector Regression has shown high correlation among MAX-CMO similarity measure and the structural descriptors derived

SSEs contacts can be a useful measure to identify structural similarities for proteins

Future Work

- Apply the methodology to a larger set
- Study the structural descriptors for proteins containing strands and a combination of strands and helices
- Establish a method to predict the structural descriptors from the protein sequence

Acknowledgments

Supervisor
Prof. Jonathan Hirst

Research Fellow
Dr. Yuri Bykov

Research fellows and Ph.D. students at
Computational Chemistry Group

BIOPTRAIN Project

THANK YOU

QUESTIONS ?