



The University of
Nottingham

BIOinformatics
OPTimisation
TRAINing
BIOPTRAIN EST

BIOPTRAIN PROGRESS MEETING - I

Pooja Jain
12 July 2006

Main Goals

- Participation in CASP 7 (The community-wide assessment of protein structure prediction methods)
- Optimization of algorithms for Protein structure prediction following the hypothesis that:

the information required by protein to adopt its final conformation is encoded in its primary structure (Anfinsen 1960s)

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: Evaluation of Methods in the area of protein structure prediction where algorithms are tested against resolved, but unpublished proteins.

- 89 targets have been released so far
- 30 targets < 150 residues in length are considered

Prediction season is scheduled to run from April-August 06

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
- DESTSTRUCT – Predicts Protein secondary structure with dihedral angles
- DSSP- Define Secondary Structure Of Proteins
- SABLE - Prediction of Solvent AccessiBiLitiEs
- Pfam – Protein family Database

Pre Analysis -Results

Length : 76 residues

Target T0309: MASKKVHQINVKGF-FDMDVMEVTEQTKEAEYTYDFKEILSEFNGKNVSITVKEENELP--VKGVEMAGDPLEHHHHHH

DESTRUCT: CCCCCCHCCCCC-EHHCHHHHHCCCCCCECCCCHEEEEEHCCCCCHEHEHCCCC-CHCCECCCCCCECCCCC

PSI-/1TF7|F: SRAINVFKMRGSHWHDKAIREFMISDKGPDIKDSFRNFERIISGSPTRITVDEKSELSRIVRGVQ-EKGPESHHHHHH

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

DEST/1TF7|F: CCCCCCHHCCCCCCHHEEEECCECCCECCCHHHHEEEECCECCCECCCHHCHHHHHHCCC-CCCCCCCCC

DSSP/1TF7|F: EEEEEEEESSS---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

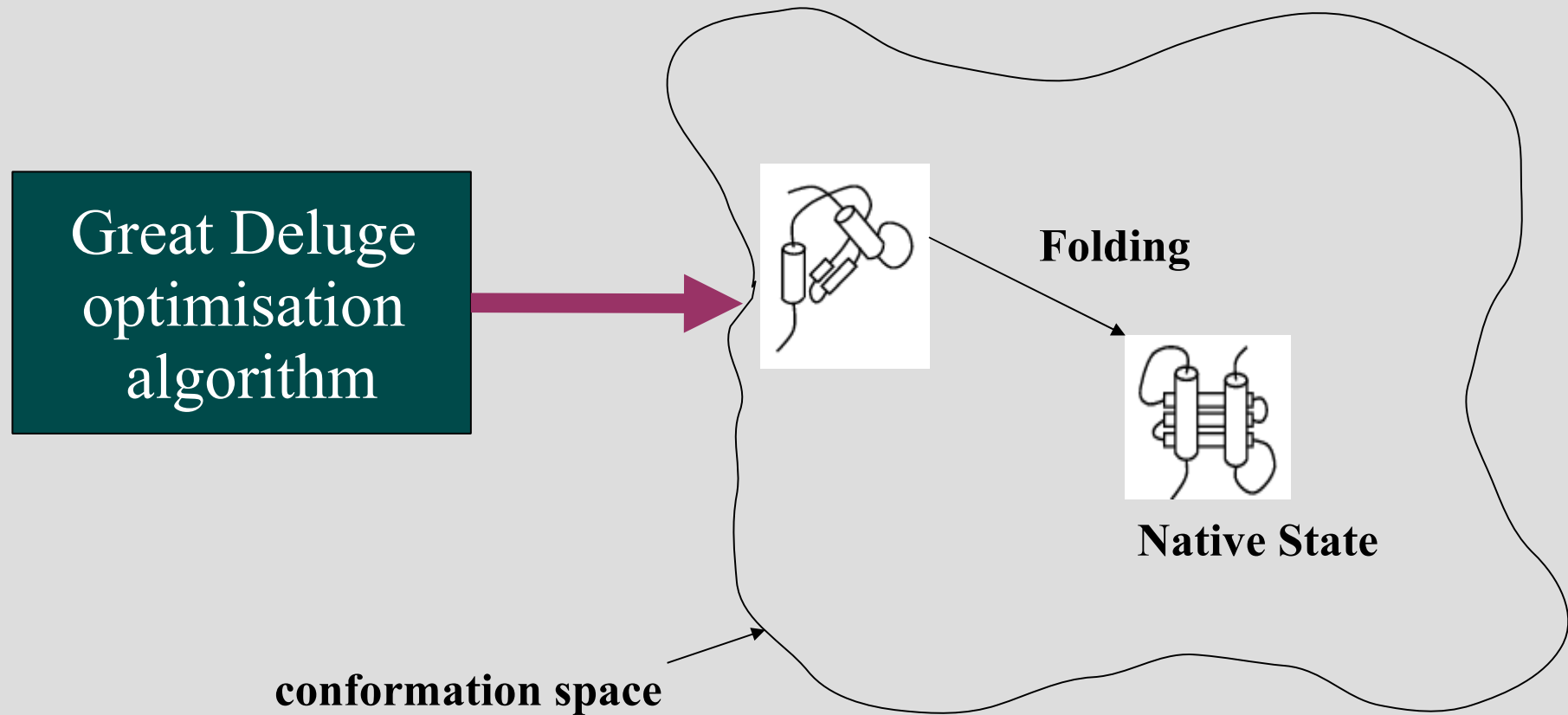
Results were used to select an appropriate model from Great Deluge Algorithm

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 Radius of Gyration

Compared against the values from pre analysis:

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

$$R_{\text{gyr}} = 2.48 \times N^{0.34}$$

Post Analysis : Results

The Final conformation with:

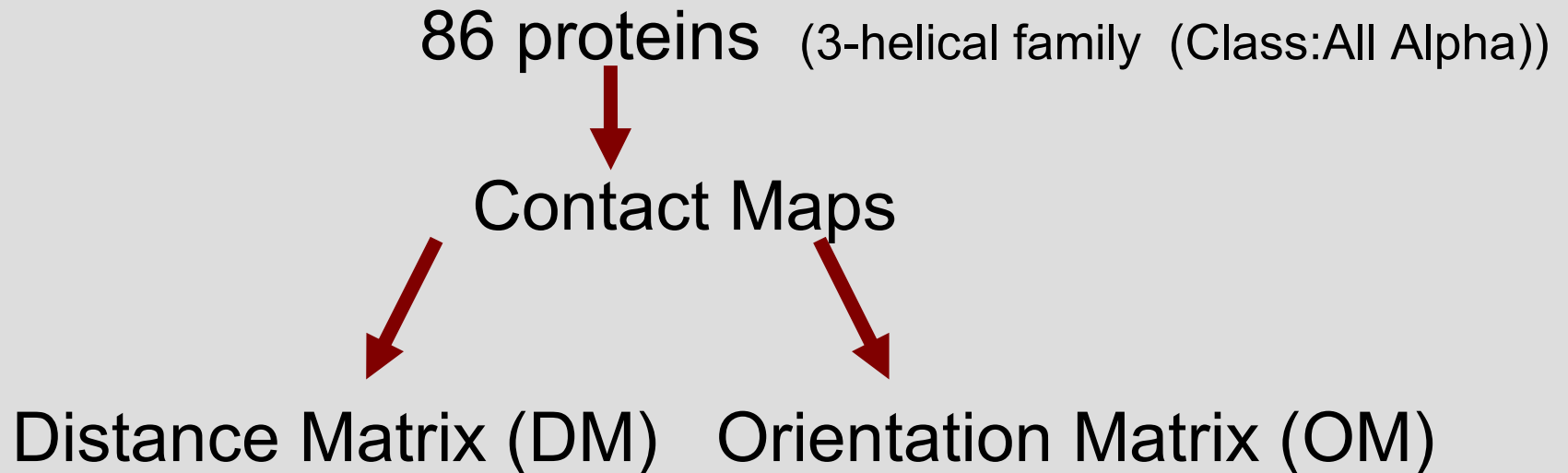
- Minimum Surface Accessibility
- Minimum Radius of Gyration
- Acceptable secondary structure similarity

Main Goals

- Participation in CASP 7 (The community-wide assessment of protein structure prediction methods)
- Optimization of algorithms for Protein structure prediction following the hypothesis that:

the information required by protein to adopt its final conformation is encoded in its primary structure (Anfinsen 1960s)

Protein Structure Prediction



DM ---- pairwise relative distance for SSEs

OM ---- pairwise relative orientation of the SSEs

Acknowledgments

My Supervisor
Dr. Jonathan Hirst

Research Fellow
Dr. Yuri Bykov

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

BIOPTRAIN EST Marie Curie Fellowship

AND

Thank you