



WP 1

SCHEDULING AND DATA SERVICES

Demonstrations

Tevfik Kosar, Sumeet Dua, Nate Brenner et al.



CCT: Center for Computation & Technology @ LSU

WP1 in a Nutshell

Motivation: Enable domain scientists to focus on their primary research problem, assured that the underlying infrastructure will manage the low-level cpu scheduling and data handling issues.

Use Case: A domain scientist should be able to do:

- Submit a simulation with a single click
 - Which may run on hundreds of processors across the state & access distributed data
- Get informed when results are ready

All **low level details** should be transparent to the domain scientist

- site selection, scheduling, data movement, fault tolerance, automation ..etc

WP1 Team

Senior Personnel: Allen, Brenner, Katz, Kosar (LSU), Box, Dua (Tech)

WP-1 Funded Personnel:

Graduate Students: Esma, Jagadish, Mehmet, Zhiefeng (LSU),
Thanadech (Tech)

Postdocs: TBD

WP-1 Supporting Personnel:

Staff: Prats, Honggao (LONI), Archit, Andrei (LSU)

Students: Vinay, Ibrahim, Jack, Ismail, Emir, Sirish (LSU),
Pradeep, Harpreep (Tech)

WP1 Progress

Basic Grid services deployed across LONI

- Lustre, Globus, Condor, GridFTP

Distributed storage (PetaShare) deployed across six LONI sites

- 170 TB usable (220 TB raw), unified name space

User friendly PetaShare client tools developed

- petashell, petafs, pcommands, petasearch

Stork data scheduler enhanced

- Whole datasets, parallel streams, checksums

End-to-end workflow management of several science driver applications enabled

New site selection algorithms developed

New data mining algorithms developed

WP1 Demonstrations

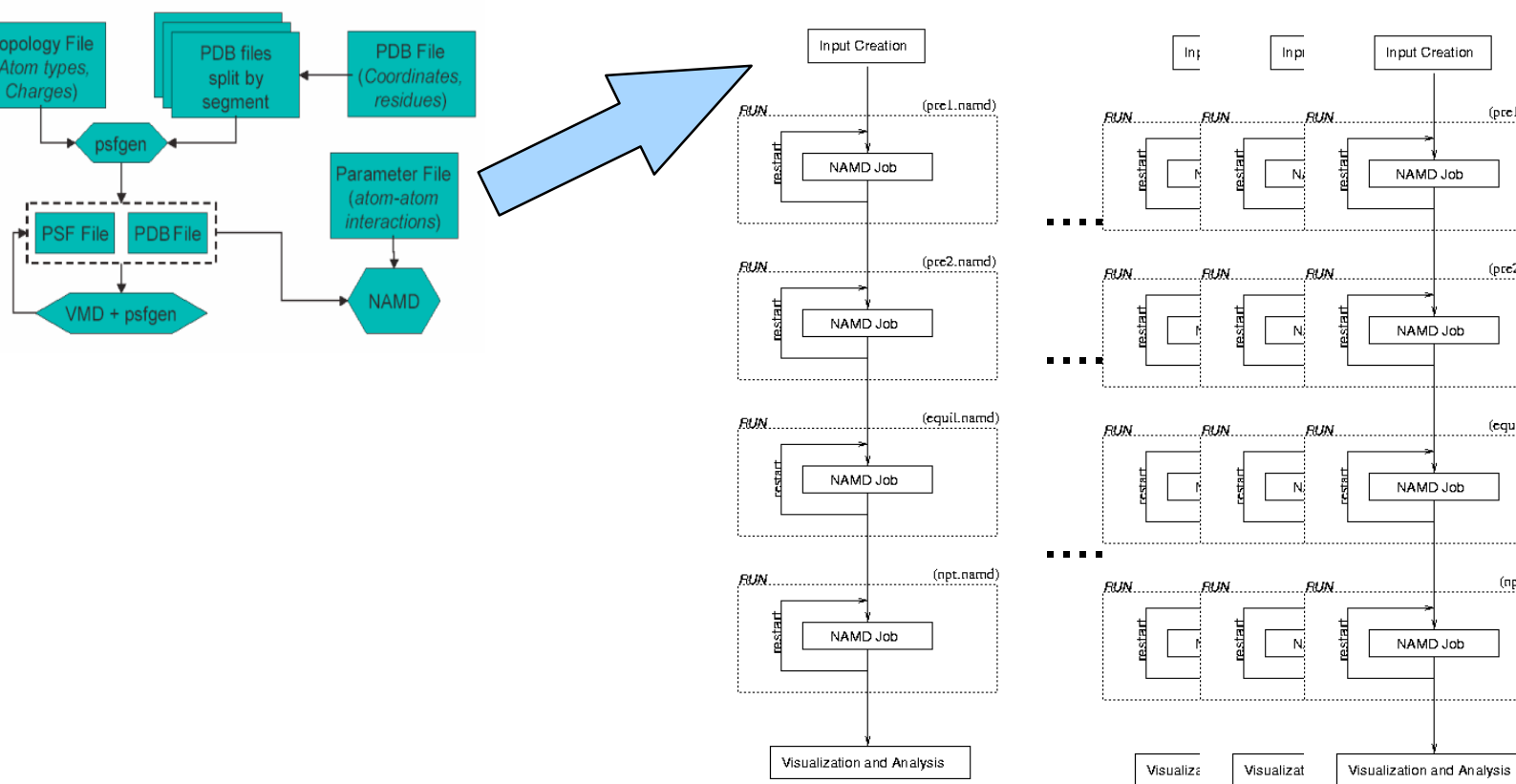
- . End-to-end workflow management
- . Dynamic site selection
- . Distributed data access & retrieval
- . Protein structure classification tools
- . Medical Image classification tool
- . Discovery of DNA folding units

DEMO - 1:
End-to-end Workflow Management
for DNA folding

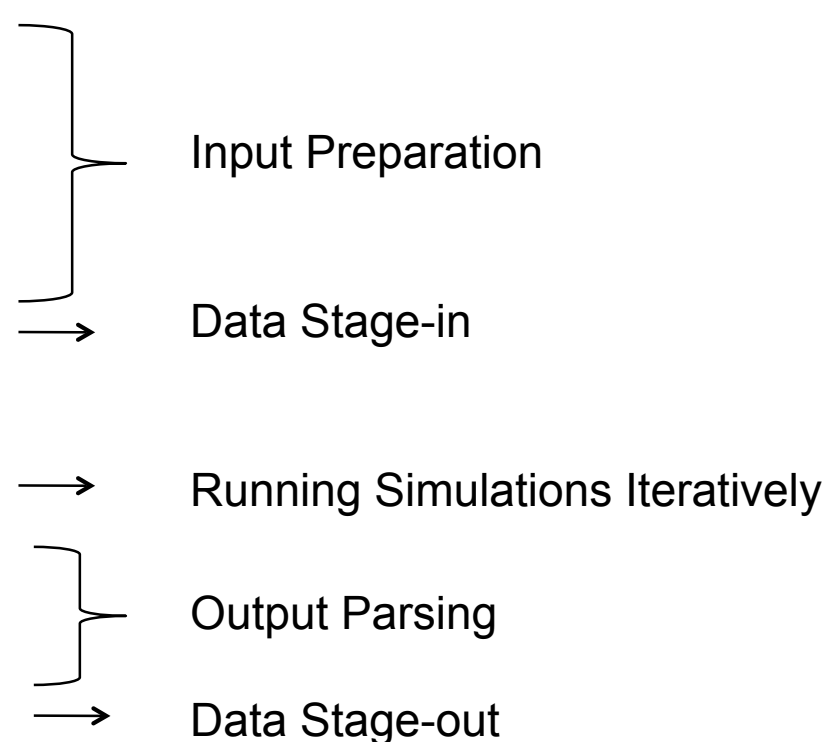
E. Bahsi, T. Kosar (LSU), T. Bishop (Tulane)

Biosensors: MD Fast Track Study

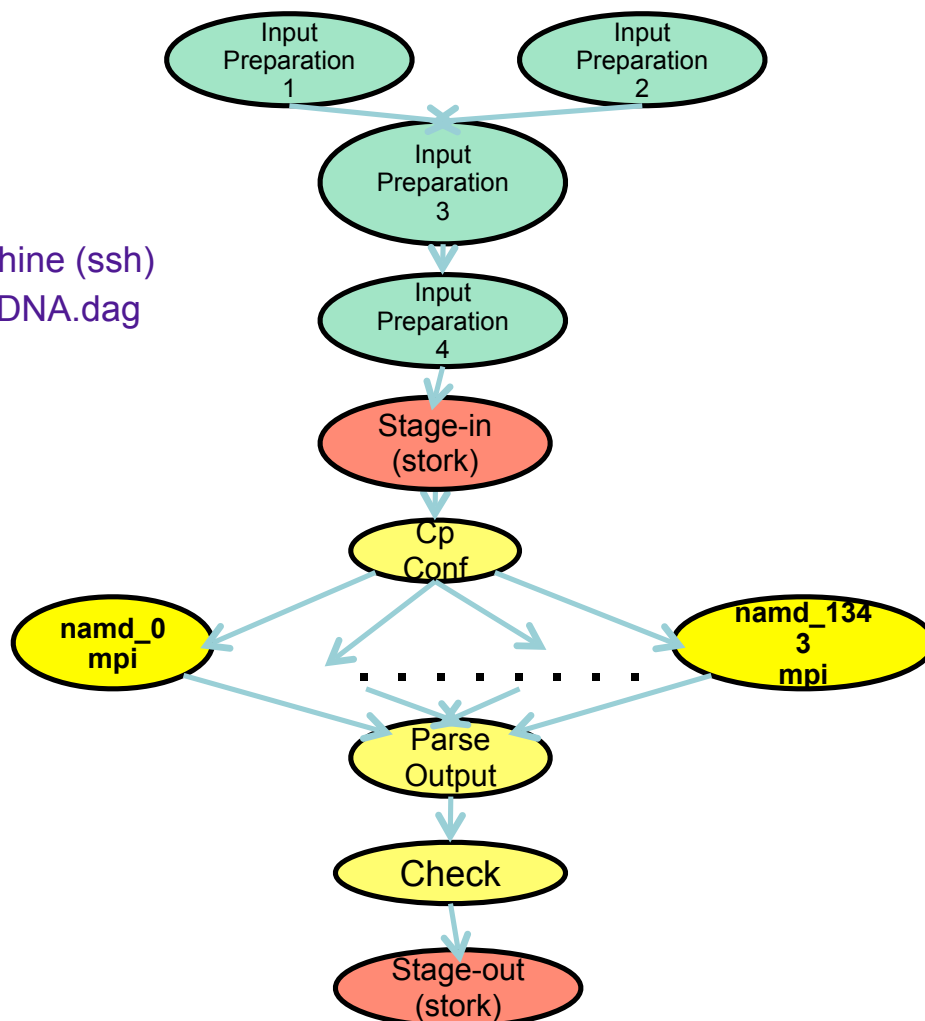
high throughput simulation workflow:



Running DNA Folding Application step-by-step (Before)

1. Connect to local machine (ssh)
 2. Run 01-setup.tcsh
 3. Run 02-mk-dna.awk
 4. Run 03-setup-amber.tcsh
 5. Run 04-setup-sims.tcsh
 6. Run 05-rsync
 7. Connect to cluster (ssh)
 8. Run 06-namd
 9. Run 07-min1.analysis
 10. Run 08-check.sims
 11. Run 09-rsync
 12. Connect to local machine (ssh)
- 
- Input Preparation
- Data Stage-in
- Running Simulations Iteratively
- Output Parsing
- Data Stage-out

Workflow-enabled Application (After)



Connect to local machine (ssh)
Condor_submit_dag DNA.dag

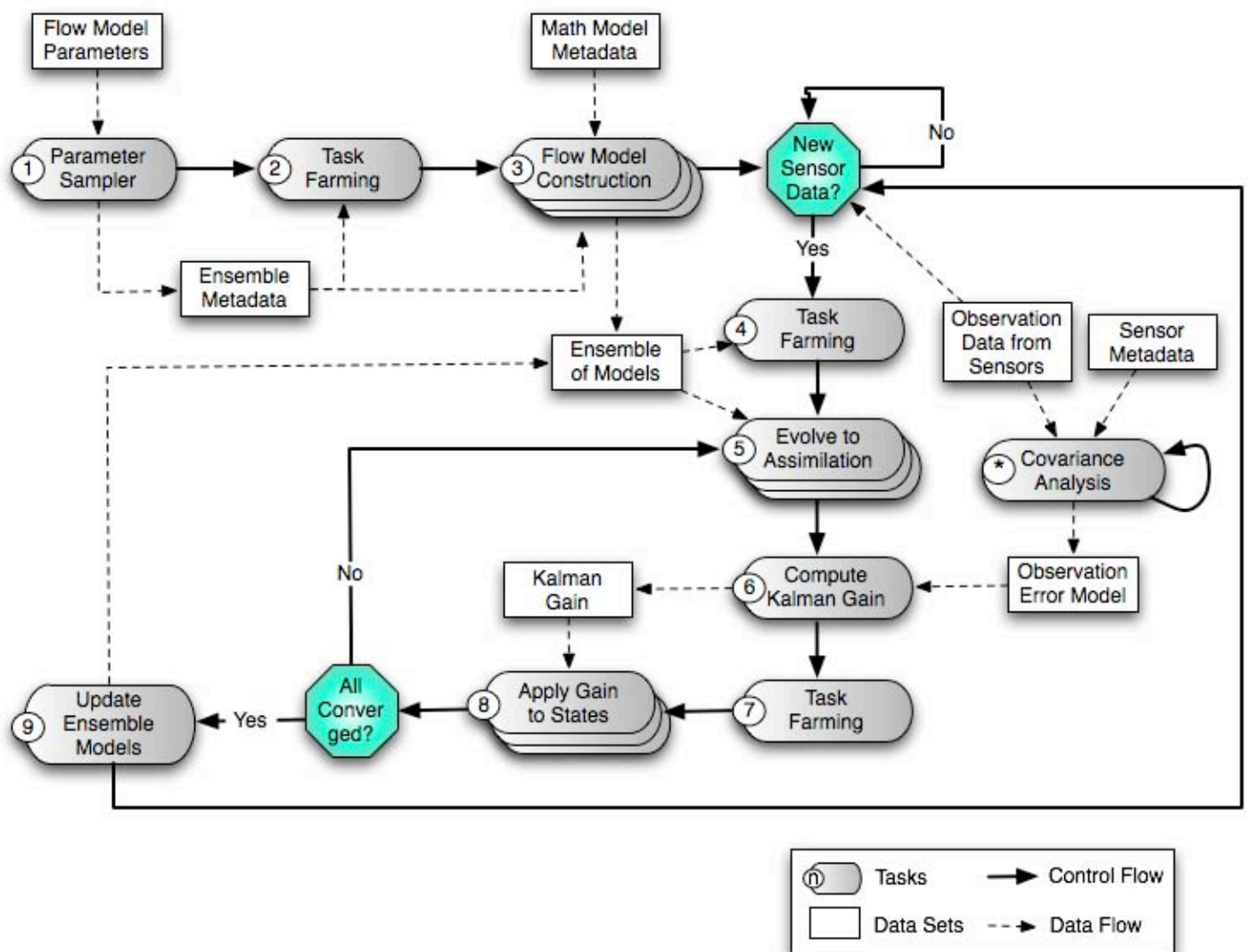
- Advantages**
- Babysitting for workflow
 - Stork for Data Transfer
 - Parallelization of MPI
 - Submit file Generation

DEMO - 2:

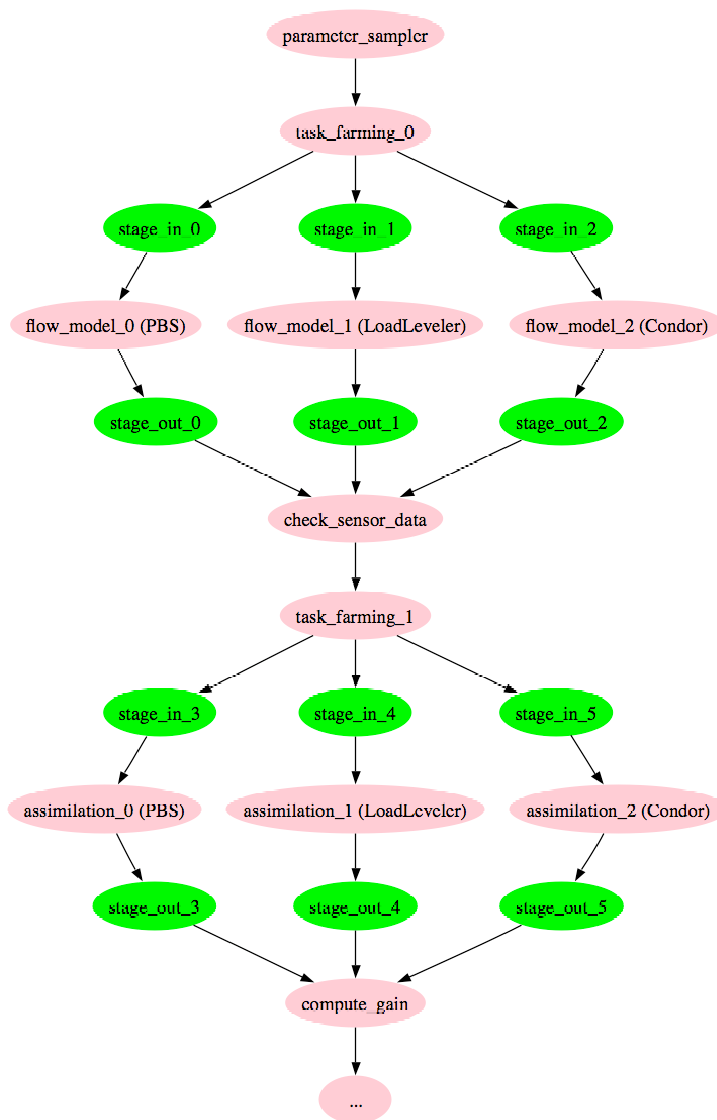
Dynamic Site Selection for
Reservoir Modeling

E. Bahsi, T. Kosar, G. Allen, M. Tyagi, C. White (LSU)

Reservoir Modeling Workflow

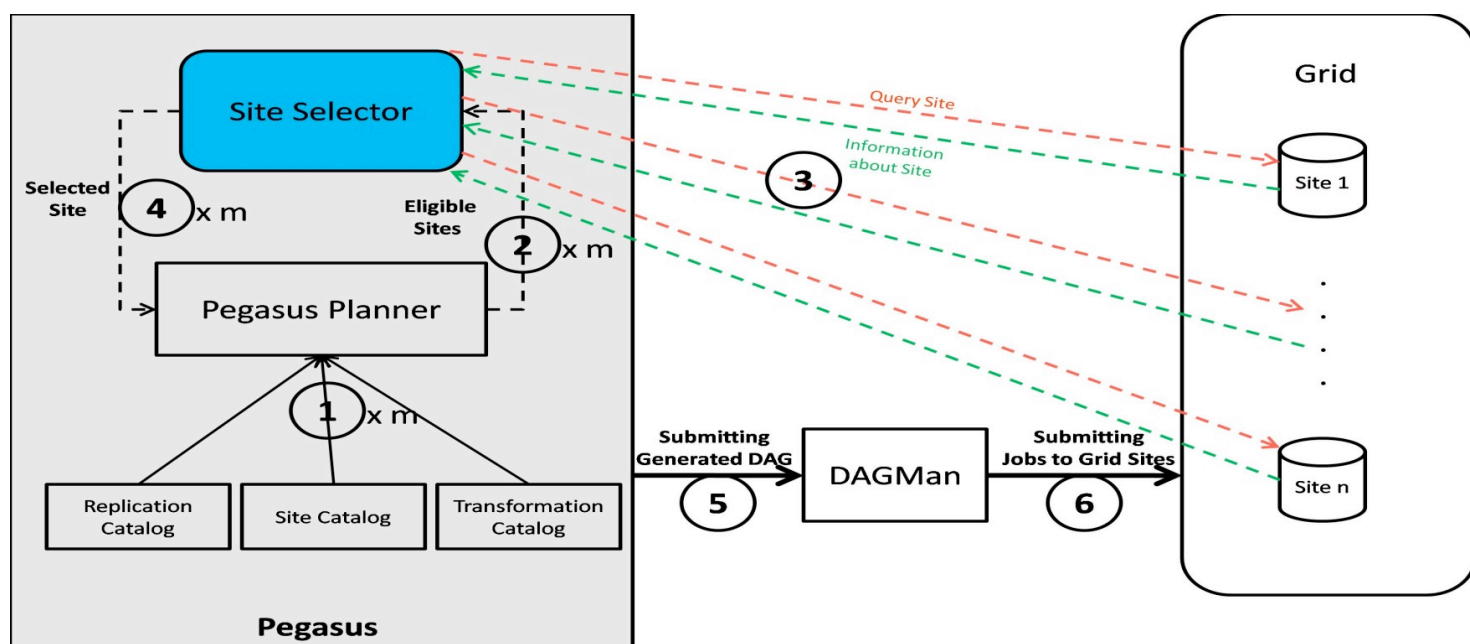


Concrete Workflow Mapping



Site Selection Mechanism

- Two Site Selectors are implemented
- Querying Sites for information about jobs and queue (# of free nodes, total # of nodes, # of jobs in the queue)



DEMO - 3:

Distributed Data Access & Retrieval

I. Akturk, T. Kosar, X. Wang (LSU) et al.

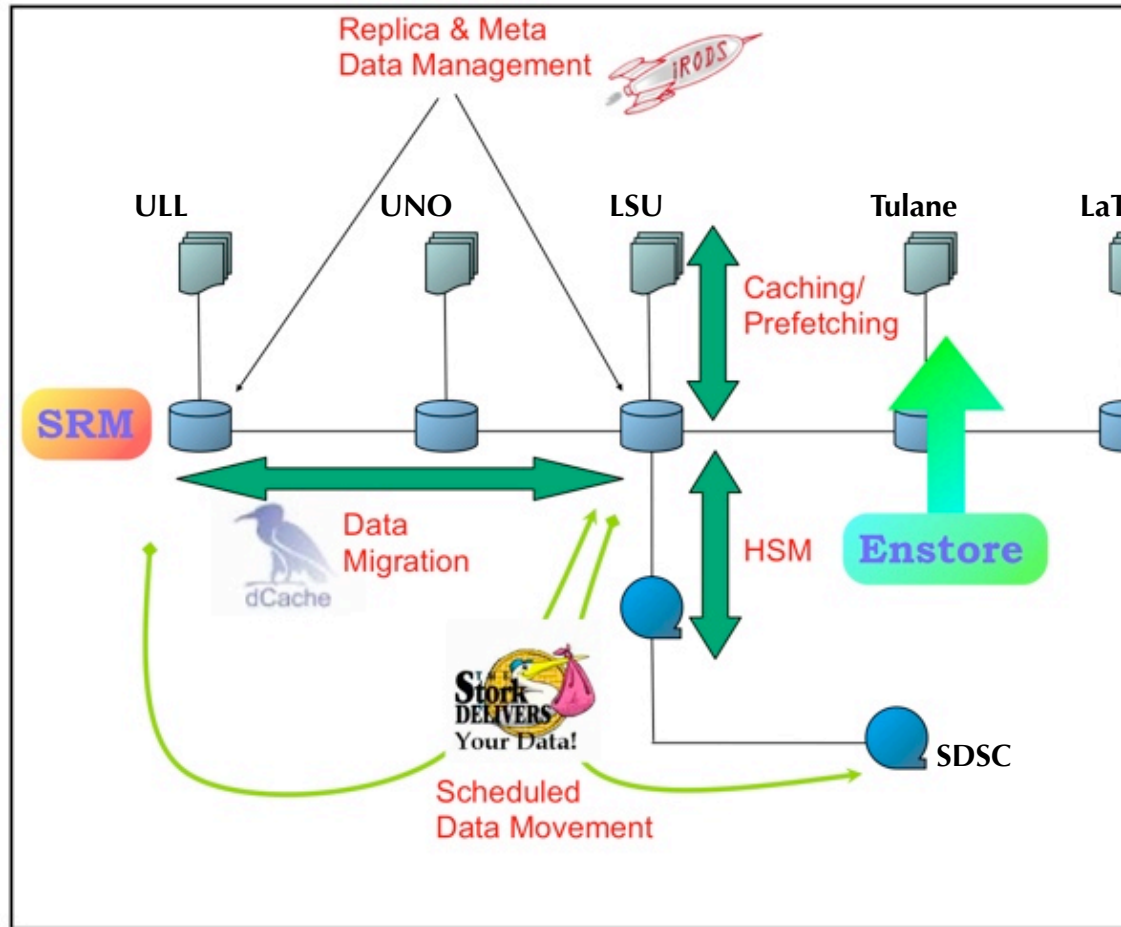
petafs

Level Virtual File System
need to change OS/kernel
need to change code
relinking
recompiling

petashell

SIX Shell Interface
all of the above
without privileged access

PetaShare Core



Web interface:

PetaSearch

DEMO - 4:

Protein Structure Classification Tool

P. Chowriappa, S. Dua (LaTech), H. Thompson (LSUHSC)

Synopsis of Cybertools Efforts

S. Dua et al. @ LA Tech, H. Thompson et al. @

- Information fusion algorithms (automated metadata extraction and information retrieval for data mining)
 - Fusion of stereochemical properties for automated protein core discovery and classification
 - Fusion of synchronization experiments in gene expression analysis (and gene ranking)
- Medical Image Classifier systems
 - Patient classification for Diabetic Retinopathy images

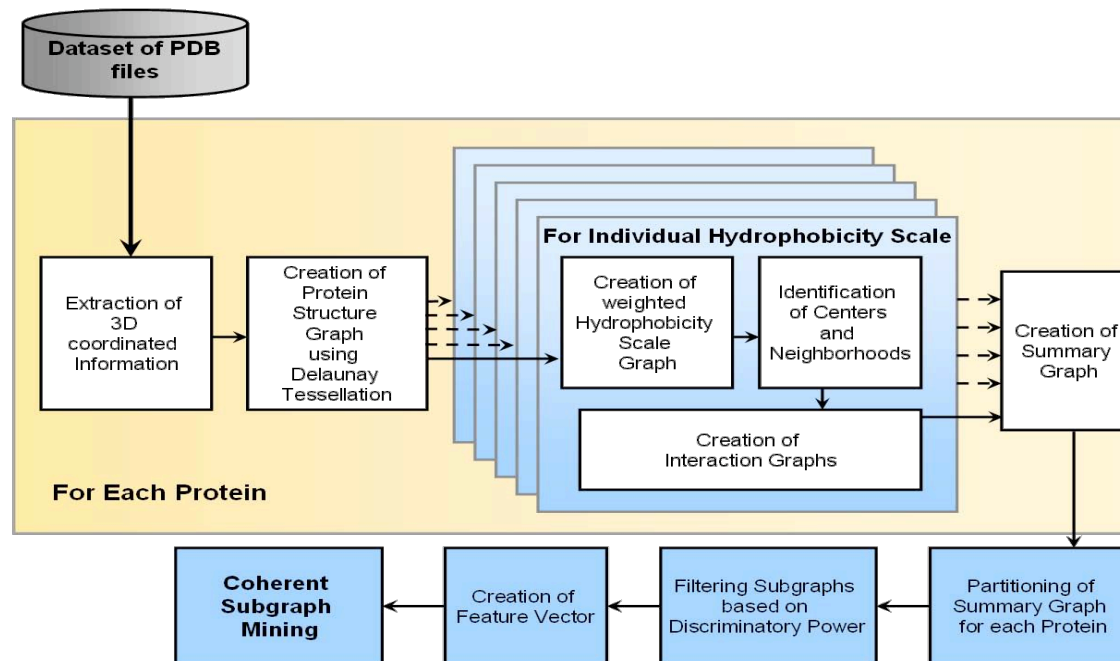


Information fusion: Integration of protein stereochemical properties for

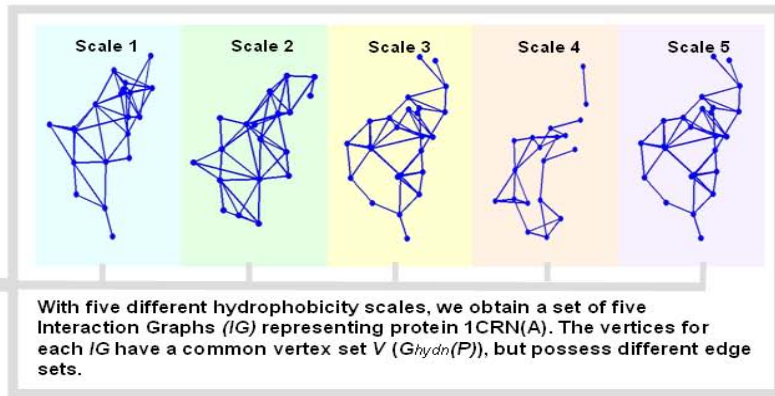
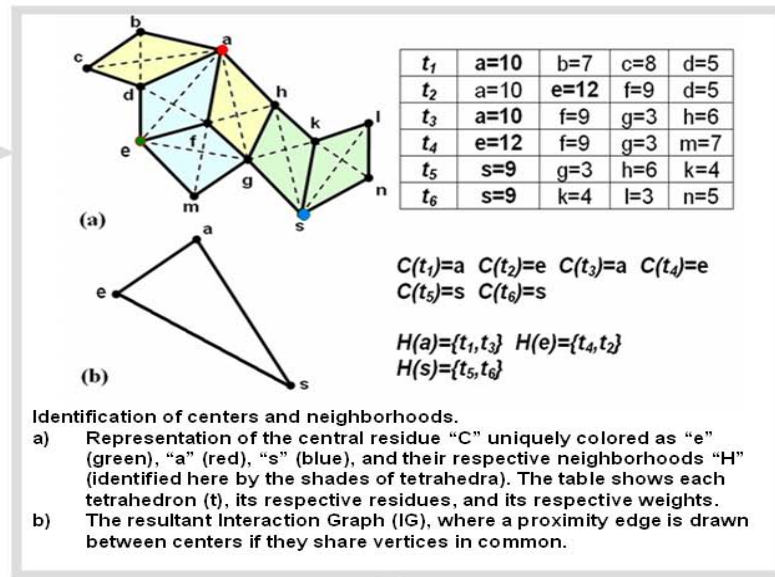
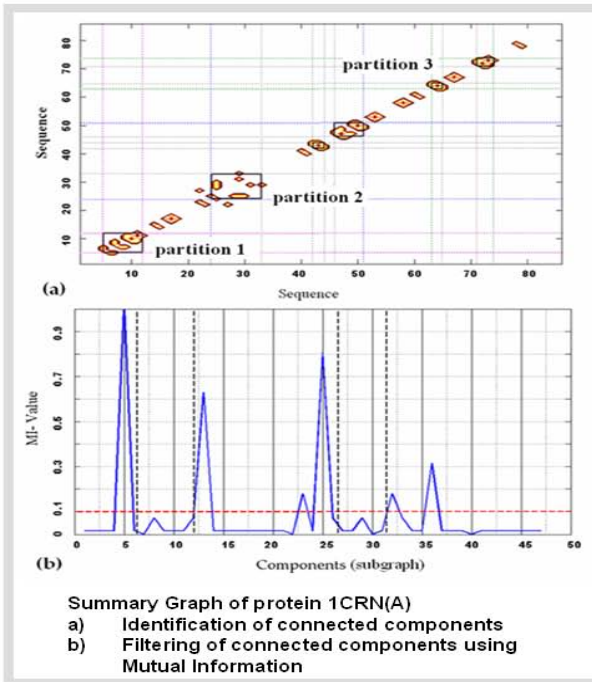
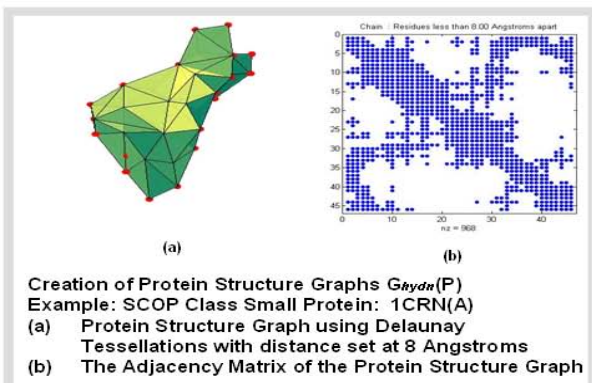
Protein sequence based tools are not sensitive enough to discover similarity between proteins because of the exponential growth in diversity of sequences.

We have developed a Graph Theory based Data Mining Framework to extract and isolate protein structural features that sustain invariance in evolutionary proteins.

We have hypothesized that proteins of the same homology contain conserved hydrophobic residues that exhibit analogous residue interaction patterns in the folded state.



Methodology



Protein Mining (snapshot of results)

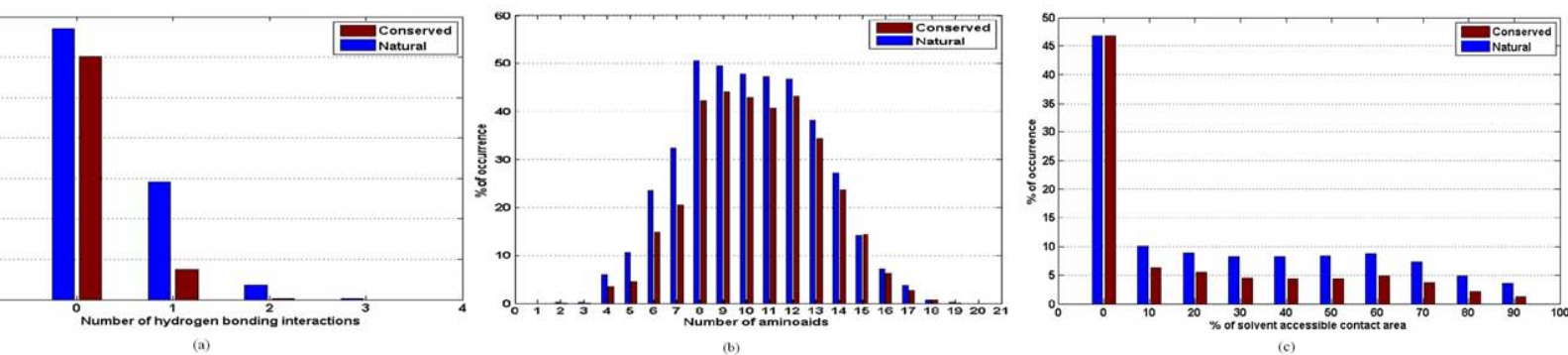


Fig. 5. Composition of amino acids in conserved residues of the summary graphs compared with the entire protein representative set. On the Y-axis is the percentage of amino acids and on the X-axis: a. hydrogen bonding interactions, b. Ooi number in an 8 Å radius around the amino acid and c. solvent accessible contact area as a percentage of residue accessibility.

Ref.: P. Chowriappa, S. Dua, J. Kanno and H. Thompson, "Protein Structure Classification Based on Conserved Hydrophobic Residues", to appear in the *IEEE/ACM Transactions on Computational Biology and Bioinformatics*.

Ref.: S. Dua, P. Chowriappa and R. Rajagopalan, "Spectral Coherence Feature Extraction from Stereochemical Scales for Protein Classification", under review for *IEEE/ACM Transactions on Computational Biology and Bioinformatics*.

Tool Features

Frontend_GUI

Protein Structure Classification Based on Conserved Hydrophobic Residues

Data Preparation

Load Datasets

C1>Select C2>Select Independent Protein

Independent Proteins

Training Set: 1nq7A

C1_Select C2_Select

Load Independent Protein

Description

ASTRAL ASTRAL-version: 1.73
ASTRAL SCOP-sid: d1nq7a_
ASTRAL SCOP-sum: 92047
ASTRAL SCOP-sccs: a.123.1.1
ASTRAL Source-PDB: 1nq7
ASTRAL Source-PDB-REVDAT: 23-SEP-03
ASTRAL Region: a:
ASTRAL ASTRAL-SPAC: 0.63
ASTRAL ASTRAL-AEROSPAC: 0.63
ASTRAL Data-updated-release: 1.67

Process Complete

Classification

Choose Classifier Random Forest Naive Bayse

RandomForest Settings

Number of Trees: 10

Number of Seeds: 1

Number of Features: 0

Training Set

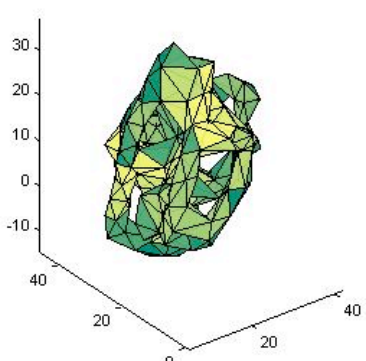
10 Fold CV

Supply Test Protein

CLEAR

CONFUSION MATIX
a b <-- classified as
1 0 | a = all-alpha
0 0 | b = all-beta

=====DETAILED ACCURACIES=====
TP Rate FP Rate Precision Recall F-Measure ROC Area
Class
1 0 1 1 1 ? all-alpha
0 0 0 0 0 ? all-beta

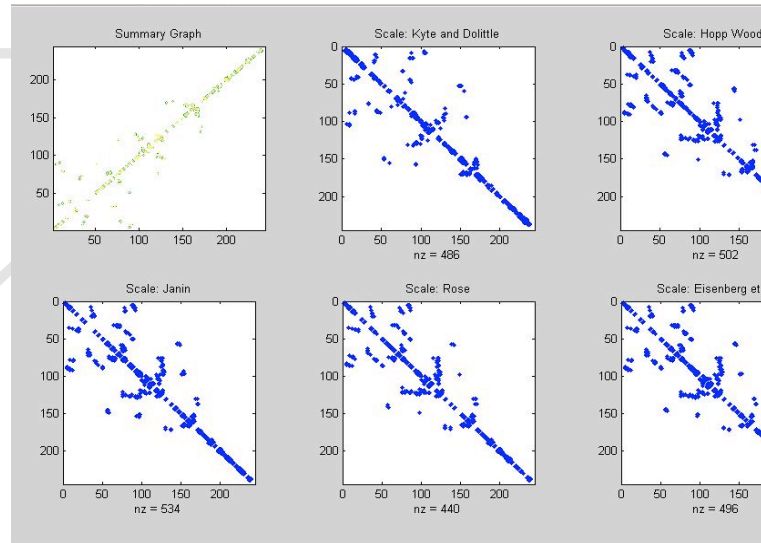
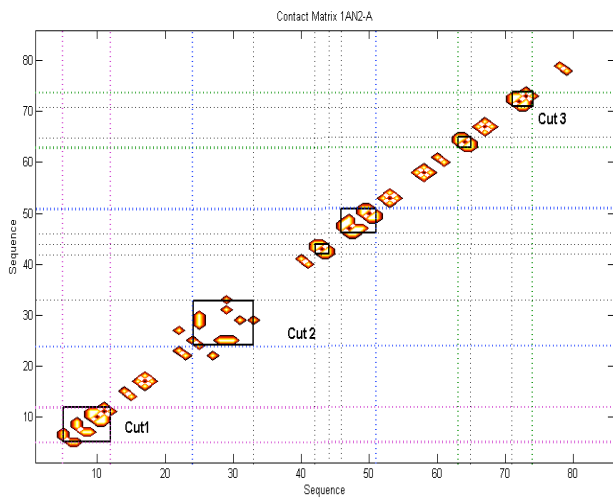


PDBid :1nq7

(DMRL) Data Mining Research Laboratory
 College of Engineering and Science
 Louisiana Tech University
 Ruston, LA - 71270

- Provides for the identification of conserved regions within proteins of the same family
- Integration of five physico-chemical properties
- Classification using Random Forest and Naïve Bayes classifier
- Provides for classification of independent proteins into specific classes

In Depth Analysis



Provides a graphical representation of the Summary Graph for better viewing of conserved hydrophobic residues

Gauge the classification performance using standard measures of calibration

Classification

Choose Classifier Random Forest Naive Bayse

RandomForest Settings

Number of Trees

Number of Seeds

Number of Features

Training Set

10 Fold CV

Supply Test Protein

CLEAR

CONFUSION MATRIX

a b <- classified as

16 0 | a = all-alpha

0 10 | b = all-beta

=====DETAILED ACCURACIES=====

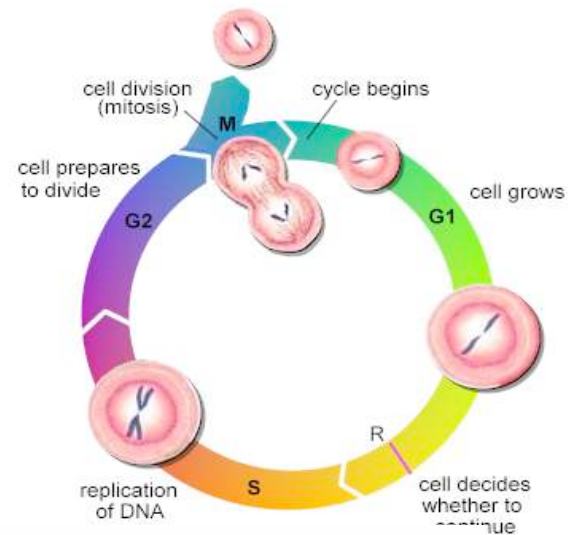
TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
1	0	1	1	1	1	all-alpha
1	0	1	1	1	1	all-beta

Information fusion: Gene Ranking through fusion of Synchronization

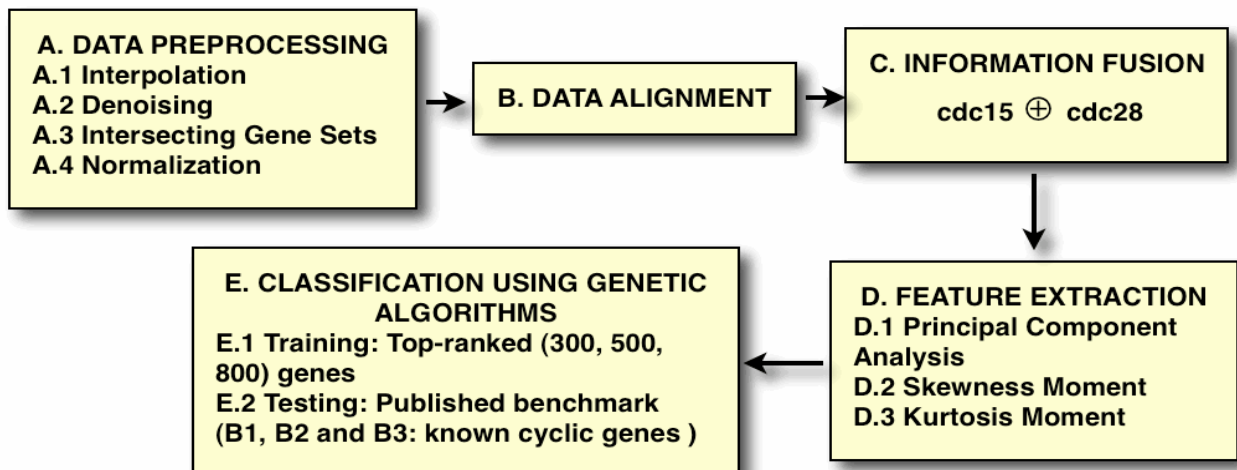
The *cell cycle*, or *cell-division cycle*, is the series of events that take place in a cell leading to its replication.

The cell-division cycle is one of the most fundamental processes of life, allowing cells to multiply and faithfully pass on their genetic information to future generations.

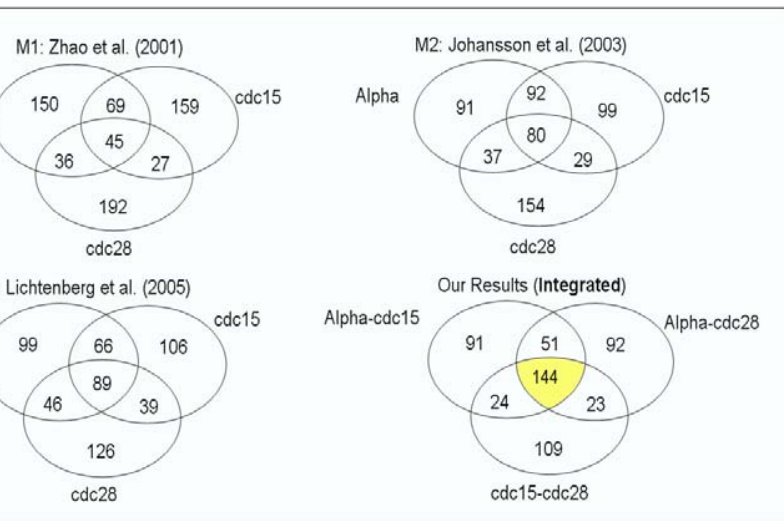
The first critical task in understanding such cyclic systems is to identify the genes that are periodically expressed during the cell cycle – focus of our work.



Our Approach



Gene ranking (snapshot of



Agreement across experiments. Venn diagram based on the top 300 genes from each experiment are shown for the methods that provide ranked lists for the individual and integrated experiments.

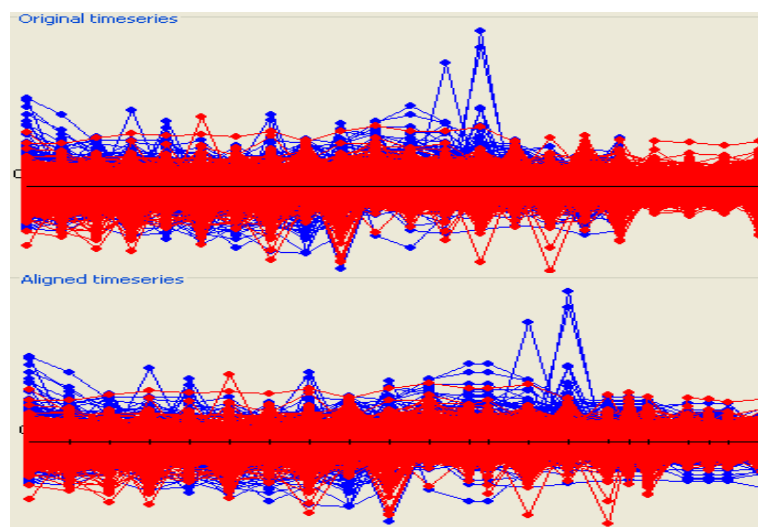


Fig. Data alignment for alpha and cdc15 datasets.

References: A. Alex, S. Dua, P. Chowriappa, "Gene Ranking through the Integration of Synchronization Experiments", to appear in the Proceedings of 2008 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (IEEE-CIBCB08).
 S. Dua, P. Chowriappa and A. E. Alex; "Ranking through Integration of Protein-similarity for Identification of Cell-cyclic Genes", to appear in the Proceedings of the Biotechnology and Bioinformatics Symposium (BIOT-2008).

Conclusion and Directions

Information Fusion and Data Mining

In conclusion, the work has demonstrated evaluation studies on independent sets of protein classes for performance benchmarking purposes.

- Other uses: hypothesis generation, protein model verification, and classification.
- 1 IEEE-TCBB, 1- IEEE-CIBCB and 1-BioT publication.

The work is a result of collaboration between investigators from:

- Louisiana Tech University
- Louisiana State University Health Sciences Center at New Orleans.

Have an independent tool to share with biologists (available through our website).

- Port tool for specific protein biotechnologist from LSUHSC (April-09, thanks to H. Thompson)

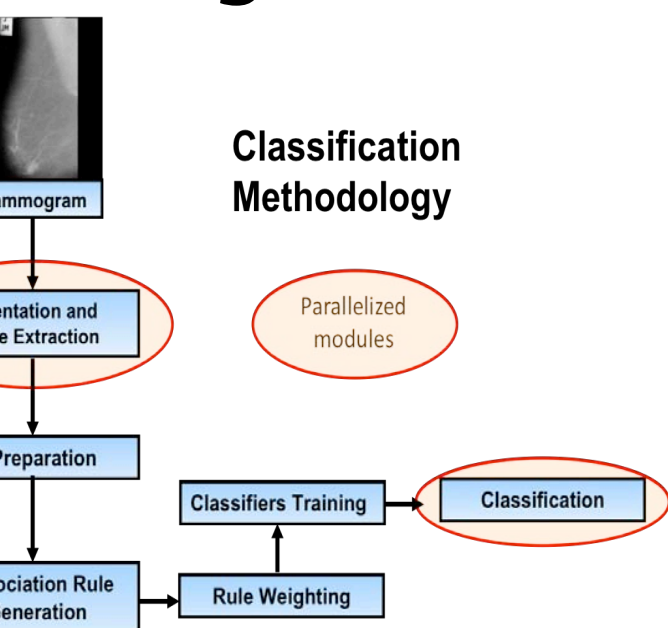
Current effort: We are developing an efficient parallelized version of the algorithm for analyzing entire PDB (Oct. 2008).

DEMO - 5:

Medical Image Classification Tool

S. Dua, H. Singh (LaTech), H. Thompson (LSUHSC)

Mammogram Classification using Weighted Rules based Classification

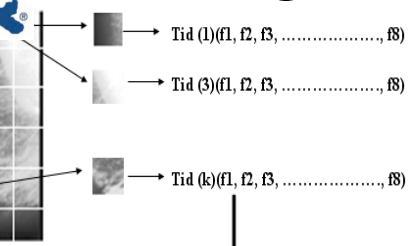


- We have developed a novel method for the classification of medical images (mammogram) using a unique weighted association rule based classification.
- Isomorphic association rules are derived between various textural components extracted from segments of images,
- These discriminatory rules are then used for the classification through exploitation of their intra- and inter-class

Rigorous experimentation has been performed to evaluate the rules' efficacy under different classification scenarios.

The algorithm delivers accuracies as high as 89%, which far surpasses the accuracy rates of other rule based classification techniques.

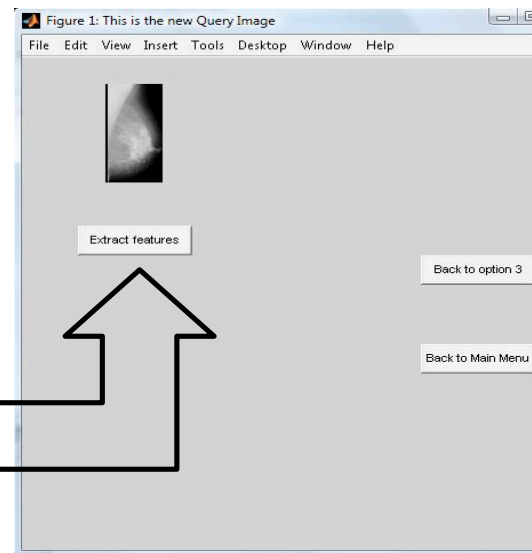
Segmentation and Feature Extraction



- Each image is divided into NxN segments
- Eight texture features extracted from each segment

	F1	F2	...	F7	F8
1	1134	2124	...	7094	8074
2	1134	2124	...	7094	8074
k
n	1120	2167	...	7104	8079

Figure 2. Segmentation and feature extraction



Click feature extraction

Feature	Feature	Calculation
1	Energy	$\sum_{i=0}^n \sum_{j=0}^n \{ p(i, j) \}^2$
2	Contrast	$\sum_{i=0}^n \sum_{j=0}^n (i-j)^2 p(i, j)$
3	Local Homogeneity	$\sum_{i=0}^n \sum_{j=0}^n \frac{p(i, j)}{1 + (i-j)^2}$
4	Correlation	$\sum_{i=0}^n \sum_{j=0}^n (\hat{g}f)p(i, j) - \mu_x \mu_y / \sigma_x \sigma_y$
5	Entropy	$-\sum_{i=0}^n \sum_{j=0}^n p(i, j) \log p(i, j)$
6	Cluster Shade	$\sum_{i=0}^n \sum_{j=0}^n (i-M_x + j-M_y)^3 p(i, j)$
7	Information measure of correlation	$H_{XX} - H_{XY} / \max \{H_X, H_Y\}$
8	Maximum Probability	$\max_{i, j} P(i, j)$

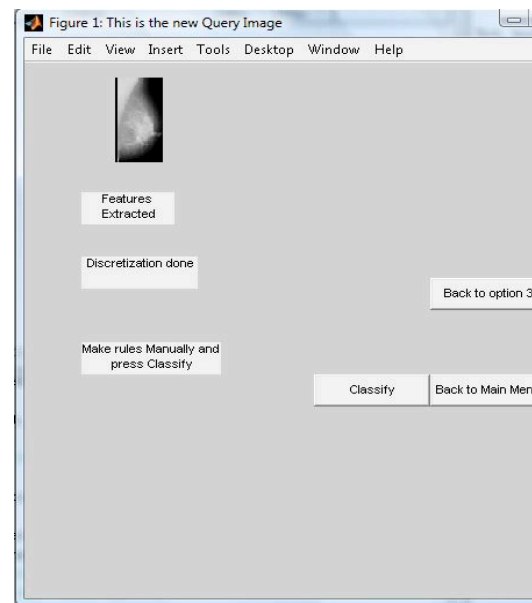
$$\mu_x = \sum_{i=0}^n i p(i, j) \quad \mu_y = \sum_{j=0}^n j p(i, j)$$

$$\sigma_x^2 = \sum_{i=0}^n i^2 p(i, j) - \mu_x^2 \quad \sigma_y^2 = \sum_{j=0}^n j^2 p(i, j) - \mu_y^2$$

$$H_X = -\sum_{i=0}^n P_x(i) \log P_x(i) \quad H_Y = -\sum_{j=0}^n P_y(j) \log P_y(j)$$

$$H_{XY} = -\sum_{i=0}^n \sum_{j=0}^n P(i, j) \log (P_x(i) P_y(j))$$

Table 1. Texture features



Classification

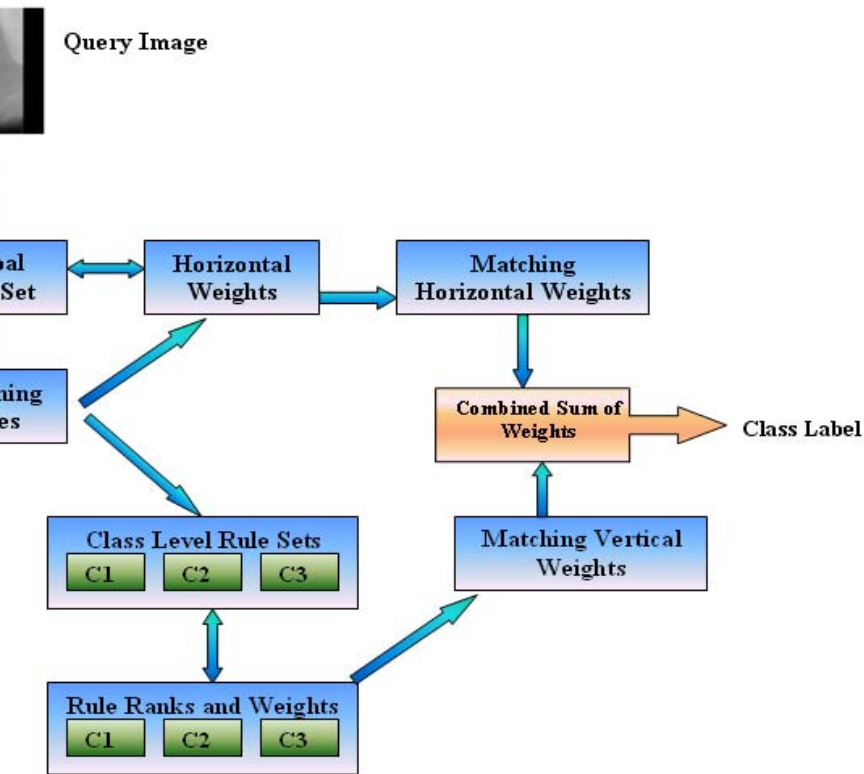
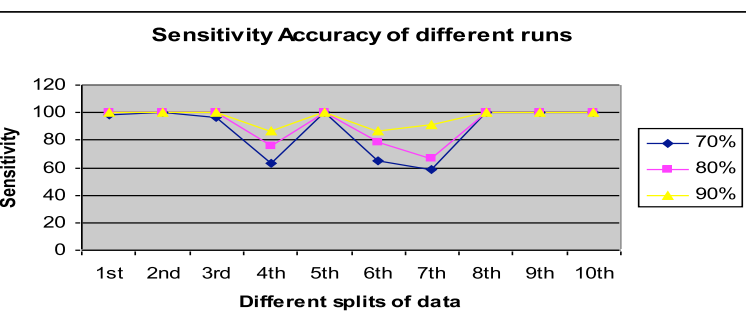


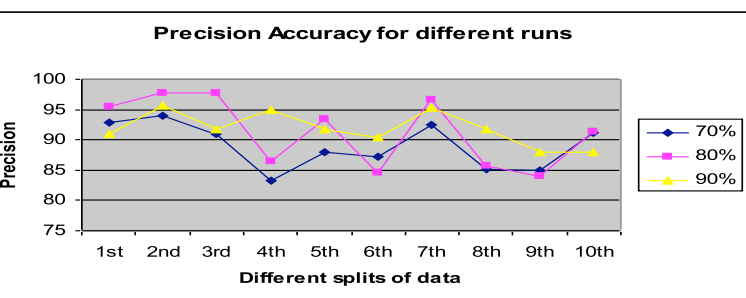
Figure. Classification Mechanism

- Form horizontal weights of rule
- Form vertical weights for rule
- Take query image and find matching rules
- Find corresponding horizontal and vertical weights
- Add these weights to form cumulative sum
- Classify to the class with high weight
- Display images from same class

Mammogram classification (snapshot of results)



(a)



(b)

The change of Precision (a) and Recall (b) with different percentages of training versus testing data.

True Classes	Reported Classes		
	Normal	Benign	Malign
Normal	22	0	0
Benign	1	5	0
Malign	1	0	3

The confusion matrix for three classes considered for classification. The number indicates the number of cases reported.

Reference: S. Dua, H. Singh, H.W Thompson, "Associative Classification of Mammograms using weighted Rules based Classification", under review for Expert Systems and Applications Journal (Elsevier).

Diabetic Retinopathy Patient Classification

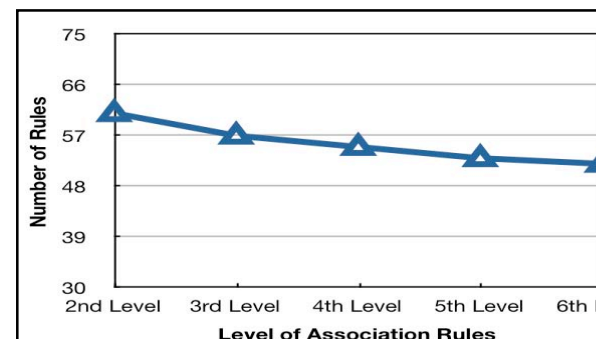
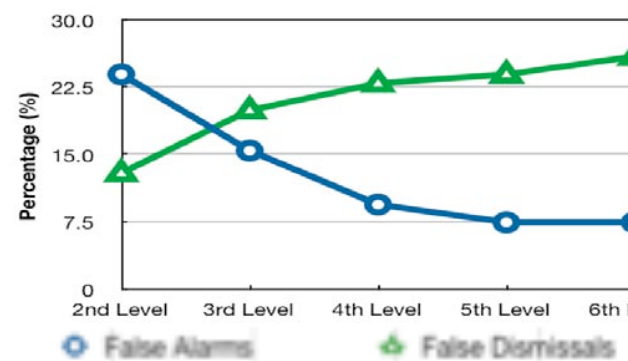
Patient classification in medical imaging has a range of applications spanning both the biomedical and healthcare delivery domains. We have developed a unique classifier for automated integration and classification of images of patients.

Patients were suffering from either Non-proliferative Diabetic Retinopathy (NPDR) or Proliferative Diabetic Retinopathy (PDR).



Diabetic Retinopathy Patient

Patient ID	Common rules	FA (avg.)	FD (%)	FA (%)
	42	455	0	30
	309	409	0.48	24
	4	420	0	33
	15	351	3.6	30
	15	465	0	36
	40	505	15	32
	728	114	0.14	9
	27	457	0.92	29
	671	101	0.4	8



Reference: S. Dua, V. Jain, H.W. Thompson, "Patient Classification using Association Mining of Clinical Images", appeared in the Proceedings of The Fifth IEEE International Symposium on Biomedical Imaging (ISBI '08)

Conclusion and Directions

Image Classification

We can autonomously classify images based on discovered content, rather than user-supplied metadata.

- 1 IEEE-ISBI publication, 1 under review.

The work is a result of collaboration between investigators from:

- Louisiana Tech University
- Louisiana State University Health Sciences Center at New Orleans.

The tool is not specific to mammograms or DR images.

- Can we easily extended (without recoding) to other image domain



DEMO - 6:
DNA Folding Units Discovered by
Data Mining

N. Brenner et al (LSU)

IMAGE FUSION AND DATA MINING

Faculty: Dr. S. Sitharama Iyengar (LSU)
Dr. Nathan E. Brener (LSU)
Dr. Bijaya B. Karki (LSU)
Dr. Hilary Thompson (LSUHSC)

Project Coordinator: Dr. Dimple Juneja

Graduate Students: Dr. Hua Cao
Rathika Natarajan
Archit Kulshrestha
Harsha Bhagawaty
Asim Shrestha
Jagadish Kumar
Gaurav Khanduja
Dipesh Bhattarai

Integration with other investigators: Dr. Allen, Dr. Acharya, Dr. Bishop,
Dr. Blake, Dr. Soper

Collaborators: LSU Health Sciences Center (LSUHSC)
LATech
Air Force Institute of Technology



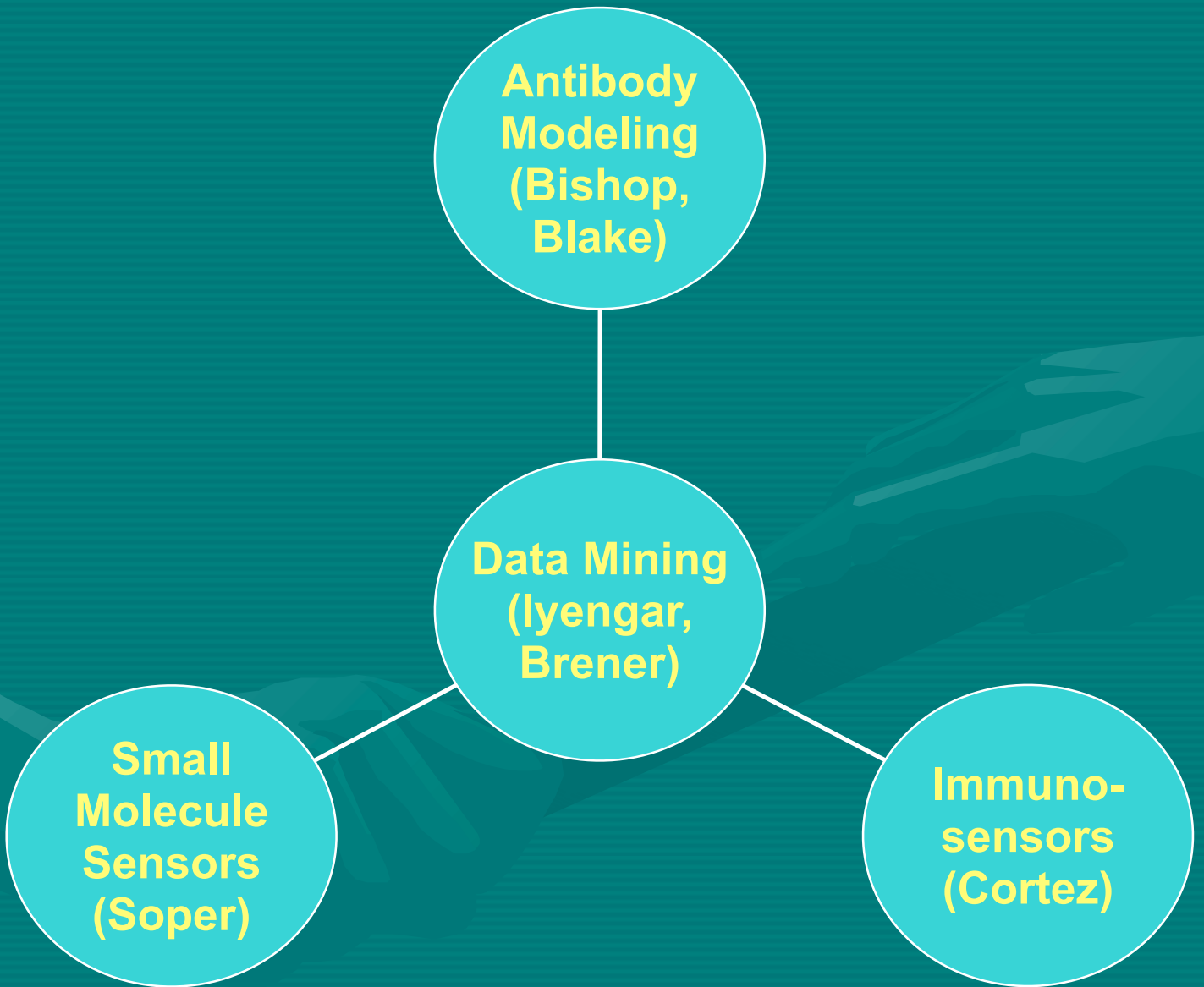
DATA MINING

Antibody
Modeling
(Bishop,
Blake)

Data Mining
(Iyengar,
Brener)

Small
Molecule
Sensors
(Soper)

Immuno-
sensors
(Cortez)



Data Mining Algorithms

Searching for features of interest in large data sets

Potential CyberTools applications:

- Antibody modeling (Bishop, Blake)
- Small molecule sensors (Soper)
- Immunosensors (Cortez)

Test problem

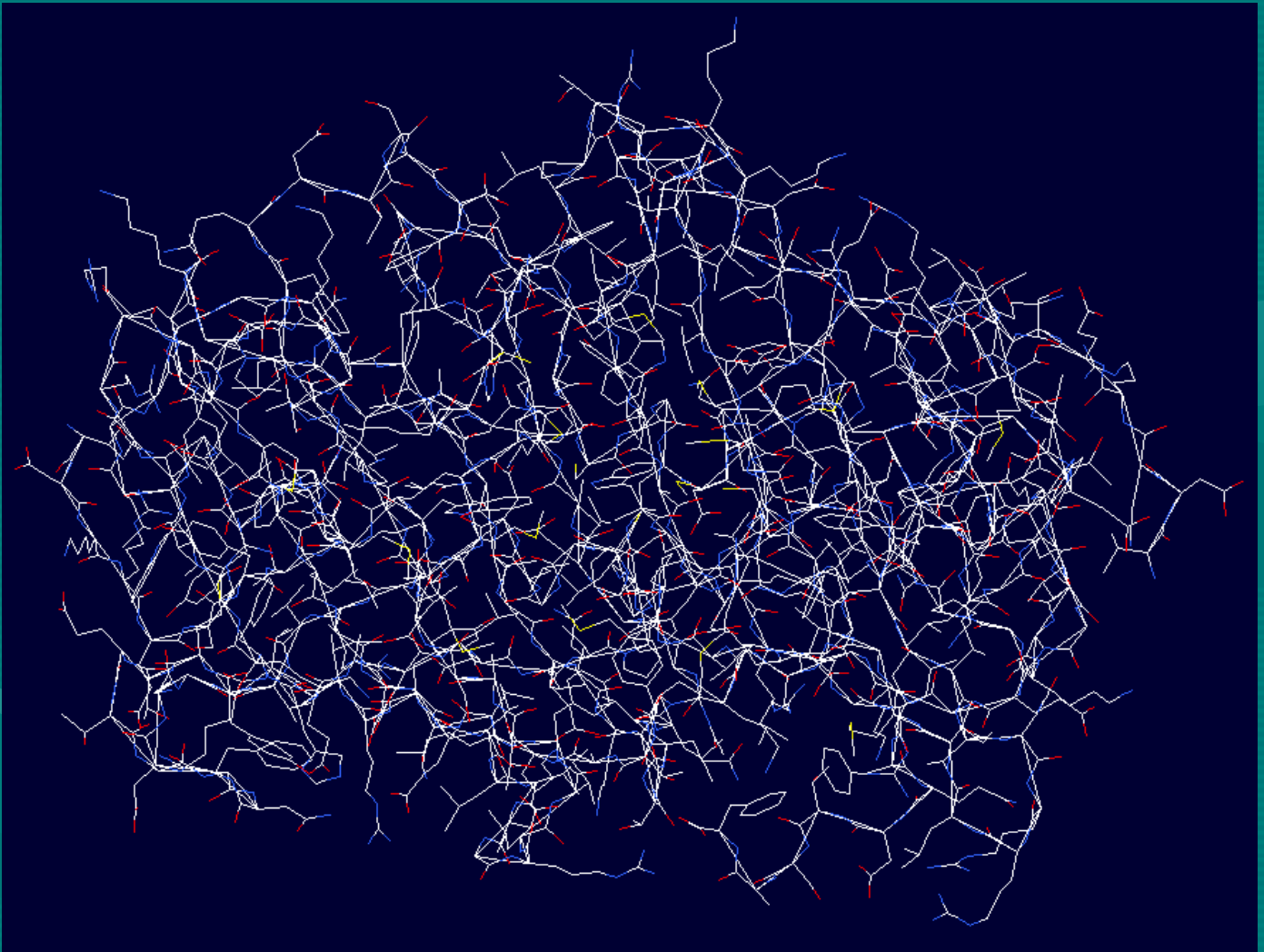
- Protein Databank (PDB). Look for common protein folding units (can be of variable length)

New Data Mining Algorithm

New efficient clustering algorithm to classify proteins according to common folding units. Based on conformational angle representation to reduce parameters.

- Represent the protein structure as a series of conformational angles**
- Partition the proteins into fragments (folding units) of a specified size**
- Cluster the fragments into groups**

Example of Randomly Selected Proteins



1mka

Common Folding Units Discovered by Data Mining

Randomly Selected Proteins

1ash, 1bsr, 1cca, 1cew, 1clm, 1crn, 1cct, 1erb, 1fut, 1hng, 1hoe, 1ibu, 1mka, 1mng, 1pkp, 1udi, 1utg, 1yal, 2vab, 5pti

3698 fragments



From 1mka
 α helix

Group 1 514 fragments

Amino

<u>Acid</u>	<u>phi</u>	<u>psi</u>
GLN	-60.078	-41.741
LEU	-69.310	-35.875
VAL	-65.116	-46.320
GLY	-67.025	-36.399
PHE	-62.244	-39.936
TYR	-66.128	-38.417
LEU	-64.114	-37.476
GLY	-70.167	-32.912

Common Folding Units Discovered by Data Mining

Randomly Selected Proteins

1ash, 1bsr, 1cca, 1cew, 1clm, 1crn, 1cct, 1erb, 1fut, 1hng, 1hoe, 1lbu, 1mka, 1mng, 1pkp, 1udi, 1utg, 1yal, 2vab, 5pti

3698 fragments



Group 2
188 fragments
From 1erb
 β pleated sheet



Group 3
79 fragments
From 1bsr



Group 4
61 fragments
From 1lbu

Milestones and Future Work

Oct 2007- Jan 2008

- Designed new data mining algorithm

Jan 2008- Aug 2008

- Implemented new algorithm for large data sets
- Tested algorithm on Protein Data Bank
- Verified that algorithm finds features of interest (common protein folding units)
- This data mining tool runs fast and handles large data sets

Future Work

- Apply this software tool to the data used by the science drivers (Bishop, Blake, Soper, Cortez)

Thank You!

