

# BioGrid Texas

Computer Science and Engineering, Physics, and the  
College of Science at the University of Texas at  
Arlington, International Business Machines, and the  
Texas Workforce Commission  
(Presented by: Dave Levine)

# BioGrid at UTA

- BioGrid Texas began in the Spring of 2004
- Funded by the State of Texas Workforce Commission and supported by IBM (Health Care and Life Sciences)
- Conceived of and managed initially by Associate Dean of Science, Paul Medley
- Collaboration between Colleges of Science and Engineering at UTA, UT Southwestern Medical School, UNT Health Sciences and IBM (and others)





# What is BioGrid Texas?

- Collaborative research, development and information-sharing system for Life Sciences and Health Care
  - Virtual Research Park (VRP)
  - Healthcare Collaborative Network (HCN)
- Universal Web interface and open development platform created by IBM Healthcare and Life Sciences
- Very large scale computing network
  - Utilizes high-performance grid computing technology infrastructure at The University of Texas at Arlington
- Allows geographically dispersed R&D teams and health care organizations to collaborate like never before

# Goal: The Virtual Collaboratory

- IBM Life Sciences Virtual Research Park incorporates a sweeping new concept called the “**Collaboratory**”
  - Allows users to solve problems using community resources and knowledge distributed across a grid computing infrastructure

# Experiment Details

Experiment Details	
<b>Owner</b>	Don Majors
<b>Created On</b>	Wed Aug 27 21:55:11 IST 2003
<b>Last Modified</b>	Wed Aug 27 21:55:11 IST 2003
<b>Modified By</b>	Don Majors
<b>ID</b>	CPSE_003
<b>Name</b>	Secondary Structure I
<b>Abstract</b>	

Protocol	Context	Team	Resources	QoS	Status
<b>Name</b>	SECONDARY STRUCTURE PREDICTION				
<b>Description</b>	Secondary structure prediction for a novel DNA sequence				
<b>Inputs</b>	DNA sequence; E.coli dna sequences database; E.coli protein sequences database;				
<b>Outputs</b>	Secondary structure information mapping of substrings of corresponding protein sequence to the secondary structures they belong to .				

## Experiment Details

- Protocol
- Context
- Team
- Resources
- Quality of Service (QoS)
- Status

# Project Results

Protocol Context Team Resources QoS Status

Protein Structure Predictor for CASP - v1.0 or a give sequence homolo...

Job ID : Identify and Substructure

Start Time :

Duration :

Task

- dnasequencesearch
- proteinidentification
- proteinhomologysearch
- proteinsequencesearch
- secondarystructurepredictio
- colorcodedrepresentation

Output : [colorCodedSec.StruInfo](#)

Address <http://gridsrv4.in.ibm.com:8081/cpse/files/colorcodedsecondarystructureinformation.html>

MNKTFEYIDAMPLAASEKAALPKTDIRAVHQALDAEHRTWARREDDSPQGS  
VKARLEQAWPDSLADGQLIKDDEGRDQLKAMPEAKRSSMFPDPWRINPVG  
RFWDRLRGRDVTPLYLARLTKEEQESEQKWRTVGTIRRYILLILTLAQT  
VATWYMKTILPYQGWALINPMDMVGQDLWVSFMQLLPYMLQTGILLFAV  
LFCWVSAGFWTALMGFLQLLIGRDKYSISASTVGDEPLNPEHRTALIMPI  
CNEDVNRVFAGLRATWESVKATGNAKHFDVYILSDSYNPDICVAEQAWM  
ELIAEVGGEGQIFYRRRRRRVVKRKS GNIDDFCRRWGSQYSYMVVLDADSV  
MTGDCLCGLVRLMEANPNAGIIQSSPKASGMDTLYARCQQFATRVYGPLF  
TAGLHFWQLGESHYWGHNAIIRVKPFIEHCALAPLPGEESFAGSILSHDF  
VEAALMRRAGWG VWIAYDLPGSYEELPPNLLDELKRDRRWCHGNLMNFRL  
FLVKGMHPVHRAVFLTGVM SYLSAPLWFMFLALSTALQVHALTEPQYFL  
QPRQLFPVWPQWRPELAIALFASTMVLLFLPKLLSILLIWCKGTEYGGF  
WRVTL SLLLEVLSVLLAPVRMLFHTVFVVS AFLGWEVWVNSPQRDDSDT  
SWG EAFKRHGSQ LLLGLVWAVGMAWLDLRF LFWLAPIVESLILSPFVSVI  
SSRATVGLRTKRWKFLIPEEYSPQVLVDTDRFLEMNRQRSLDDGFMHA  
VFNP SFNALATAMATARHRASKVLEIARDRHVEQALNETPEKLNRRRLV  
LLSDPVTMARLHFRVWNSPERYSSWVSYYE GIKLNPLALRKPDAASQ

#-----  
#



# Application Sharing

Customizable Application Windows with embedded, project-specific applications

Microsoft Internet Explorer  
Address: http://ticblr.in.ibm.com/wps/myportal/.cmd/cs/.ce/155/.s/227/.s.155/227

Virtual Research Park

Medikeng Corp White Pages | Help Now | Feedback

March 4, 2004, Hello Don Majors

Applications | Emboss Tools | IBM LS Tools | Web Tools | GxPharma | PDB Search | xTerm

**Shared Applications Portlet**

Name	Synopsis	Host Member
<a href="#">State U Gene Mining</a>	Gene Mining	State U
<a href="#">State U Regression Analysis</a>	Regression Analysis	State U
<a href="#">Medikeng Interphasse Analysis</a>	Interphasse Analysis	Medikeng/All
<a href="#">RNA desequencer</a>	Searches by sub-sequences for matching library sequences	State U
<a href="#">Compound Thermal Analyzer</a>	Extrapolative analysis based on compound similarities	Medikeng
<a href="#">Graphical Protein Builder</a>	Creates visual interpretations of protein strings	BioCo

Add Remove Change

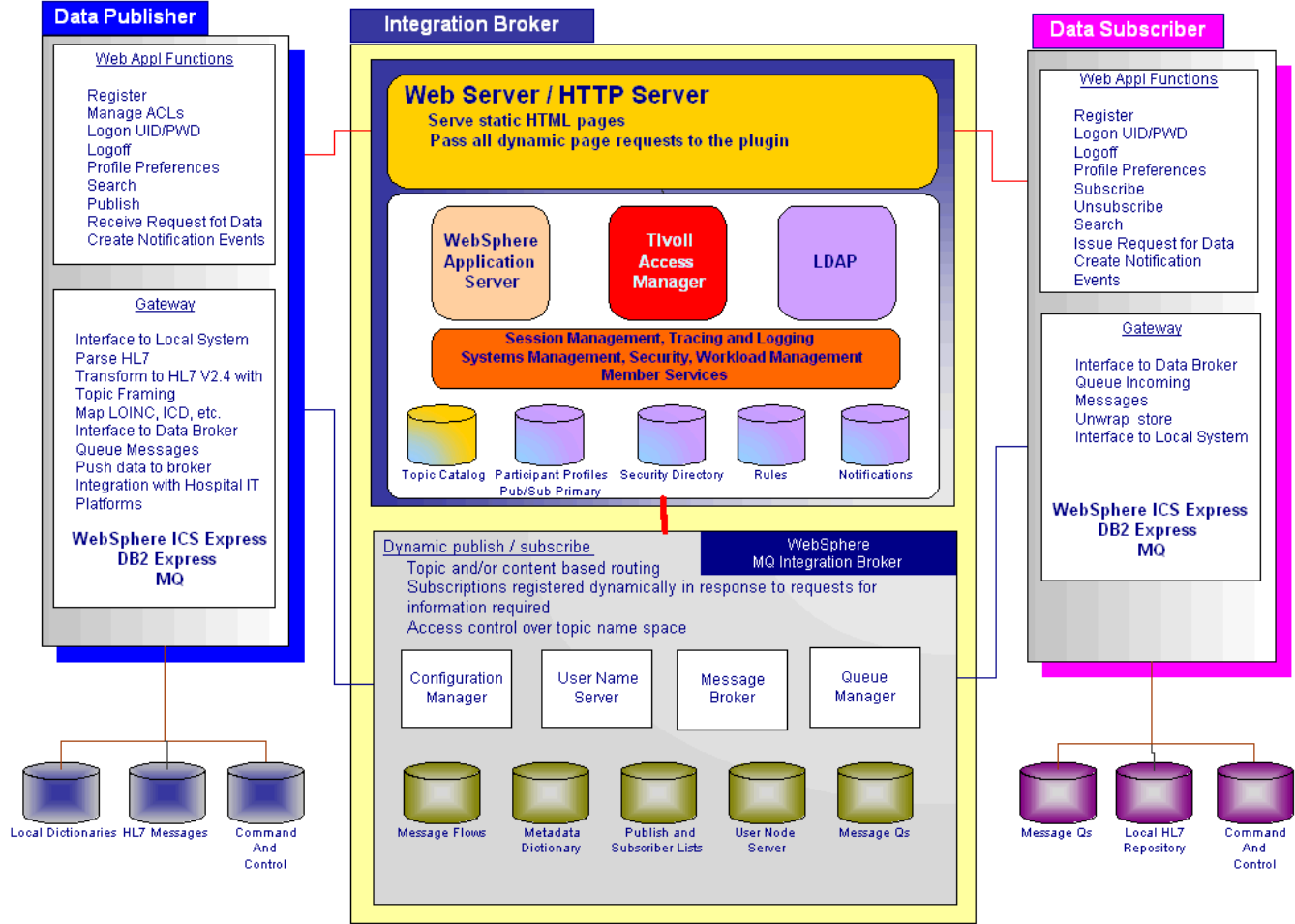
Alert Viewer  
From Message Sent  
No alerts found

Sametime Contact List  
PeopleOptions  
Work  
Bots  
Agents

Alert Generator  
New alert

Done | Local intranet

# Healthcare Collaborative Network



# Focus Groups

- **Bioinformatics, Computational Biology, Cell and Molecular Biology, Genomics and Proteomics**
  - Bioinformatics requires a higher level of competence in math and computer science
  - BioGrid system will bring together people with skills to interface with other life science professionals
    - Communicate with bioinformatics experts and biostatisticians for analysis of genetic and genomic data
    - Help train future generations of bioinformaticians
  - Access to the latest bioinformatics applications

# Three BioGrid Projects

- Skin Cancer (lesion) detection  
(UTSW medical and CSE)
- De novo TE repeat discovery (DNA)  
(Biology and CSE)
- Mosquito and Malaria gene search  
(Biology, CSE, Pharmacology)

# BioGrid Projects – Skin Cancer

- Skin Cancer (lesion) detection

Project allowed dermatologists to upload and annotate digital photographs of skin lesions, some cancerous

A portal into BioGrid allowed a new image to be uploaded and compared to knowledge base

# BioGrid Projects – DNA TEs

- New species are being sequenced weekly
- Part of “understanding” the sequence, and indeed finding genes is to compare sequences to other, similar species
- Much of the sequence are the same “strings” appearing over and over again
- These may be within a gene or located between genes (most DNA is “junk”)

---

# Quick Introduction

---

- 🧬 Biologists are interested in these long DNA sequences of nucleotides composing genes
- 🧬 Many of these sequences (a gene, part of a gene, or “junk”) are repetitive, the same sequence (or nearly the same) appearing over and over again in a chromosome or whole genome
- 🧬 But the genomic data is huge, and genes and TEs don't stand out

---

# Introduction – Some Results

---

C. Elegans –

we found 90% of the ones that had been already identified (by other methods), those were almost all correct, (there are 263),

we found 22 previously unidentified TEs (some don't really exist, but some do),

On one processor it took 24 hours, on our cluster less than 0.5 hour (previously a few days)



---

# Introduction

---

Humans – (only the X chromosome)  
we found 70% of the ones that had been  
already identified (by other methods), those  
were all correct, (there are 682),  
we found “a few” previously unidentified TEs  
(some don't really exist, but some do),  
On one processor it took 2 weeks, on our  
cluster 10 hours (previously a few months)

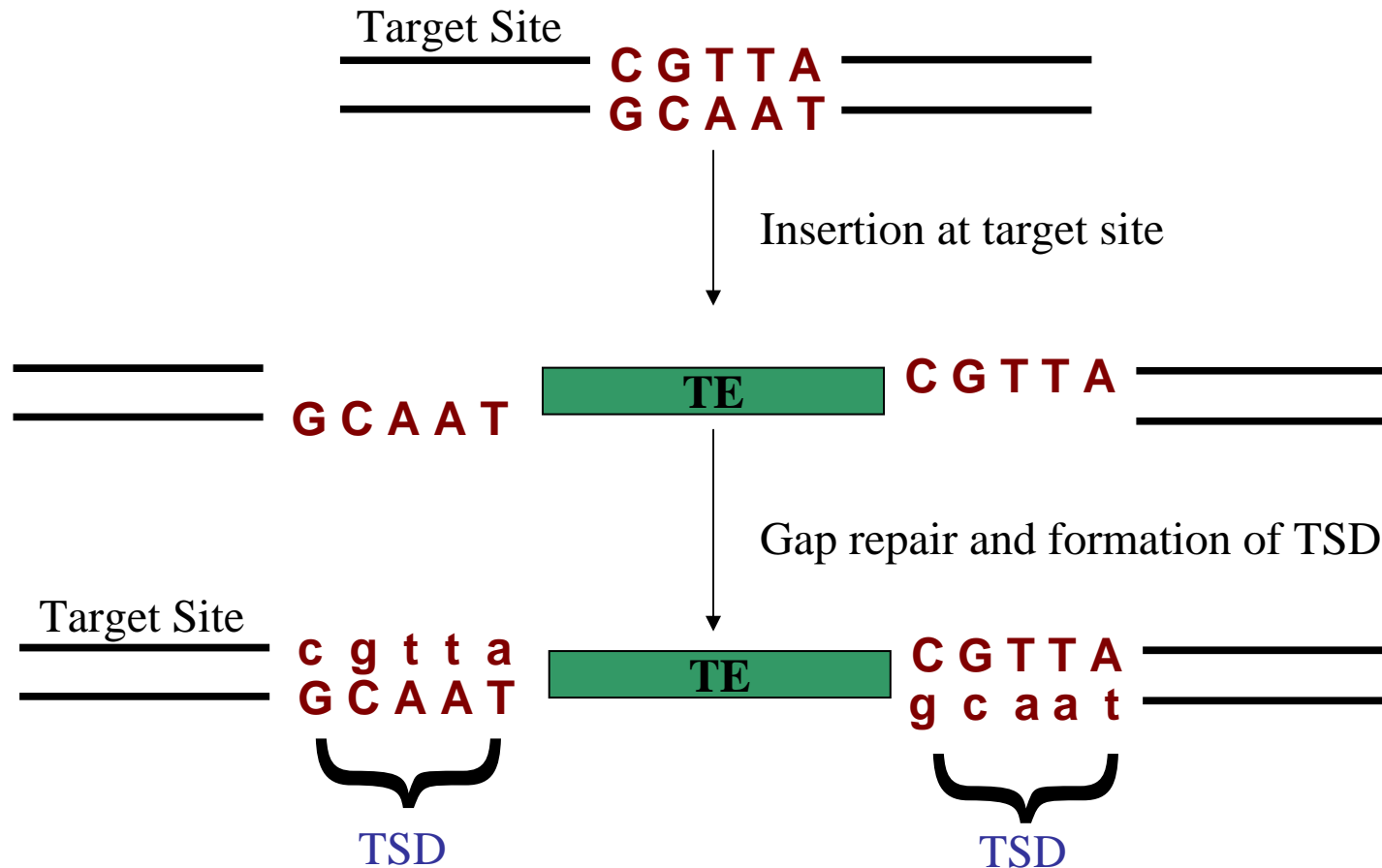
---

# Rationale

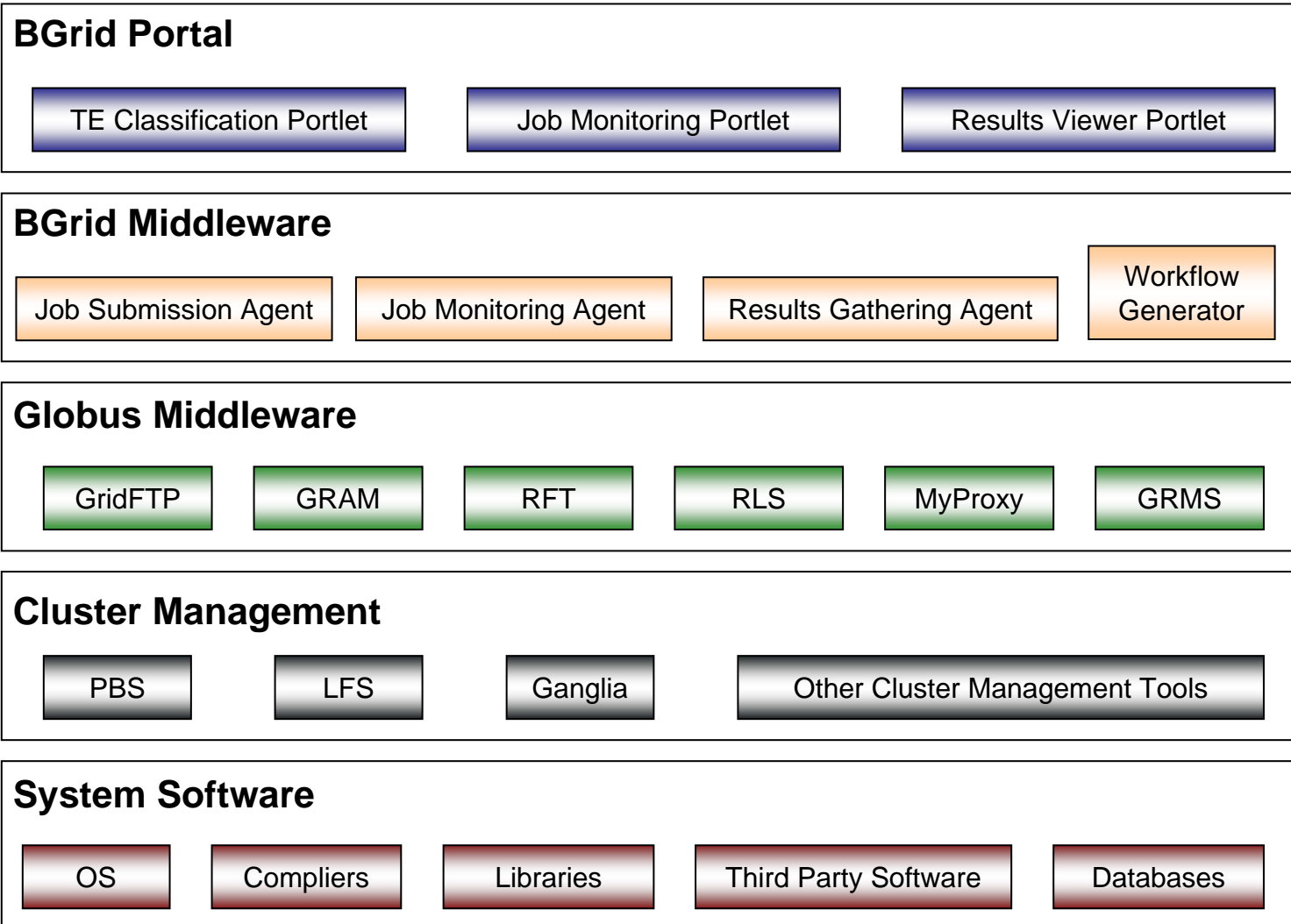
---

- Identifying and classifying TEs can help in genome assembly
- TE annotation is an integral part of genome annotation as they comprise a significant fraction of the genome
- Rebase is a database of annotated repeats
- No tool exists for automatic classification of TEs

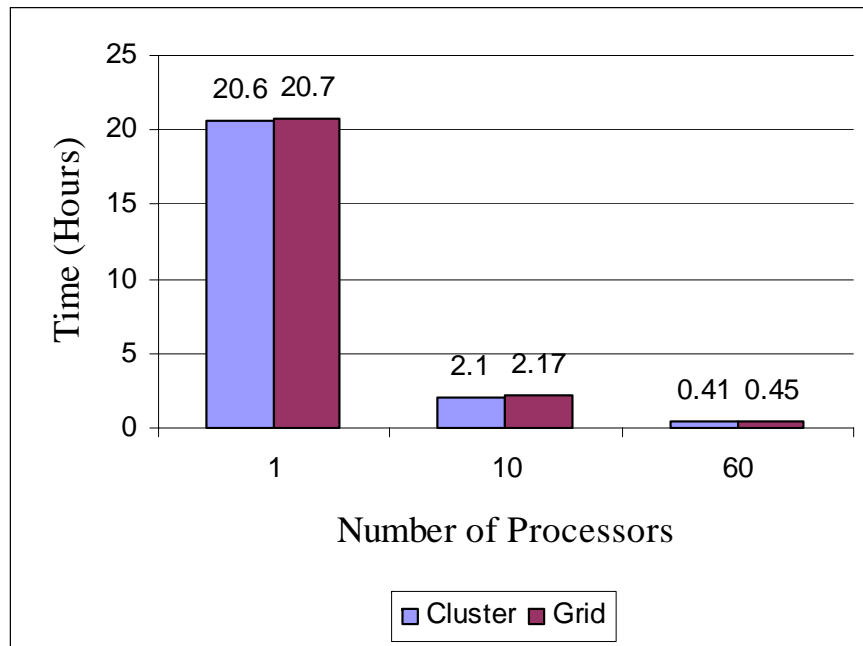
# Formation of TSDs



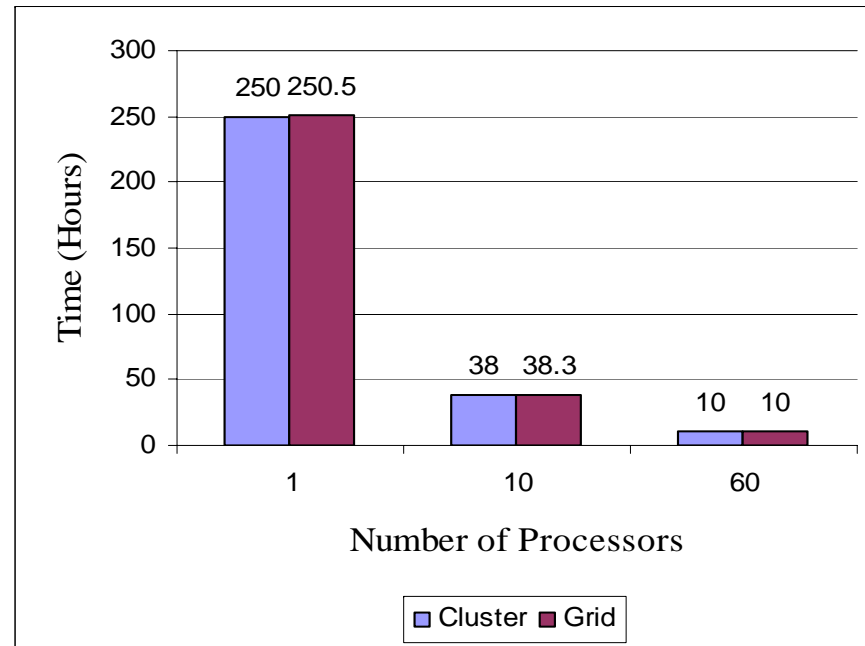
# TE Classification on a Grid



# Results – Turnaround Time



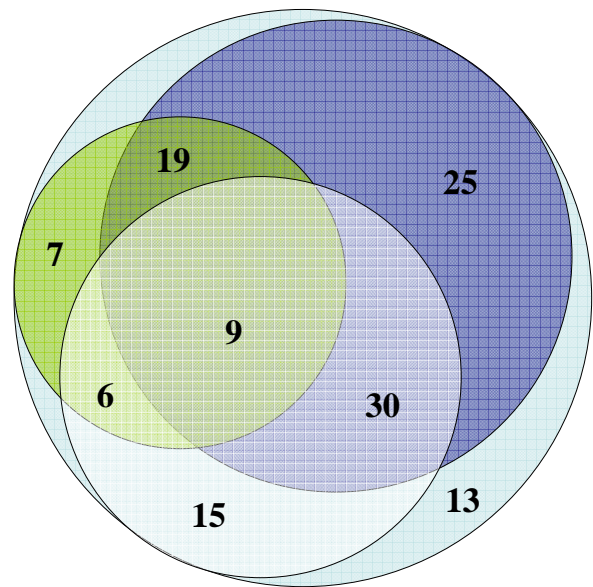
C. Elegans Genome



Human X Chromosome

# Comparison of REPCLASS classification with those of Repbase

Caenorhabditis elegans (worm)



Percentage classified	89.5%
Accuracy of classification	100%

○ Repbase

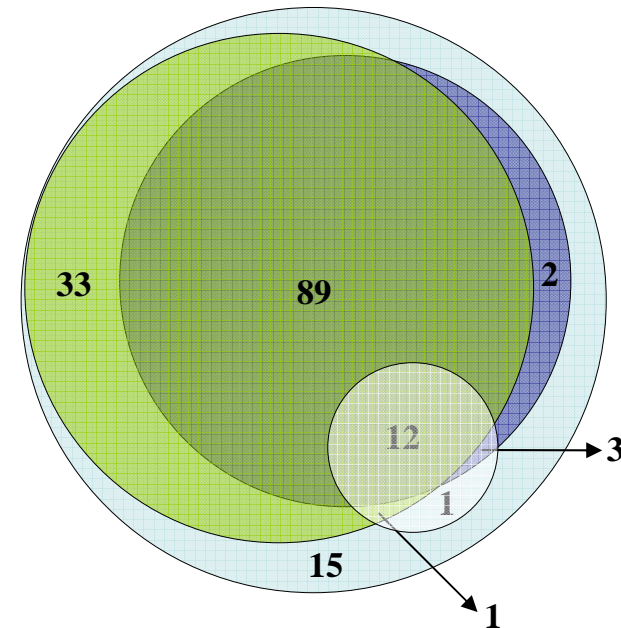
REPCLASS

● Structural

● Homology

○ TSD

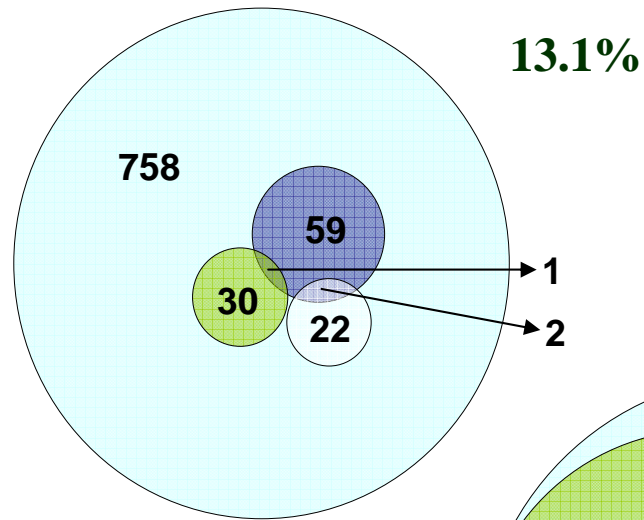
Drosophila Melanogaster (fruit fly)



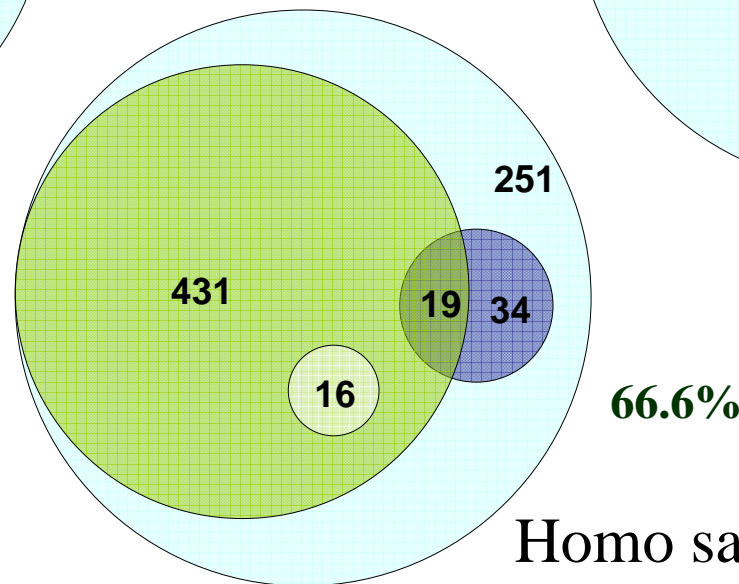
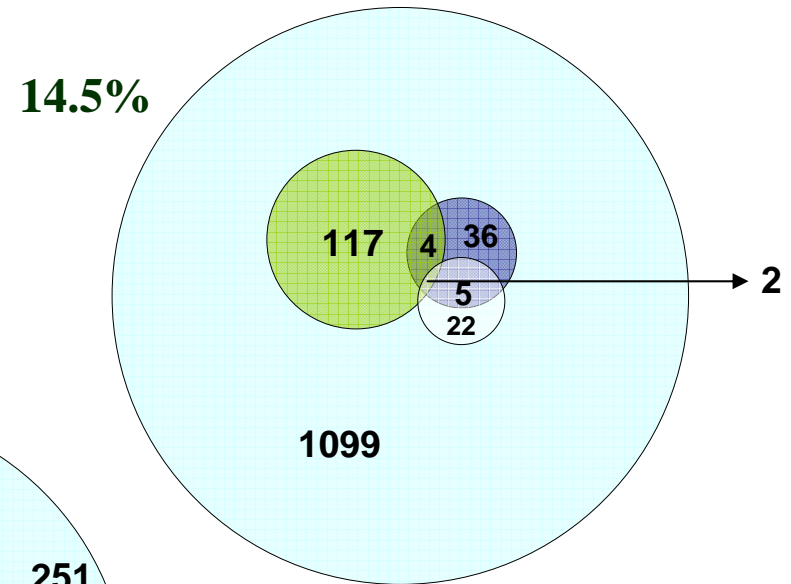
Percentage classified	90.38%
Accuracy of classification	100%

# Classification of new genomes

*S. purpuratus* (sea urchin)



*Ciona intestinalis* (sea squirt)



Homo sapiens X Chromosome

- Unclassified
- REPCCLASS
- Structural
- Homology
- TSD

# Malaria and Mosquitoes

- Malaria is a really nasty disease
- 300 to 500 million people/year get it, more than one million/year die from it
- *Anopheles Gambiae* (and similar) carry it
- Many efforts are based on vaccine, killing mosquitoes, treatment
- We don't want a better mosquito trap, we want a better mosquito



# Mosquitoes (Anopheles)

- There are about 15,000 genes predicted on 5 chromosomes
- Some areas on chromosomes poorly covered (mapped)
- Looking for a gene or a few genes that can be engineered so that Anopheles can't carry malaria

# Mosquitoes (Anopheles)

- May have found some (so far) yet-undiscovered genes
- Currently verifying some results in lab with actual mosquitoes
- (Very creepy)

# Discoveries

- Biologists love talking about their science  
– in detail
- Like physicists they are very computer knowledgeable
- They will explain biology/genetics in as much detail as one needs or wants
- Some parallel applications exist, most very rudimentary

# Thank You

Thank You for your time

I'll be happy to answer questions now  
or later, off-line