TACC, its Natural History Collections and iPlant

AIM-UP
Santa Fe, NM

October 16, 2010

TACC - Mission

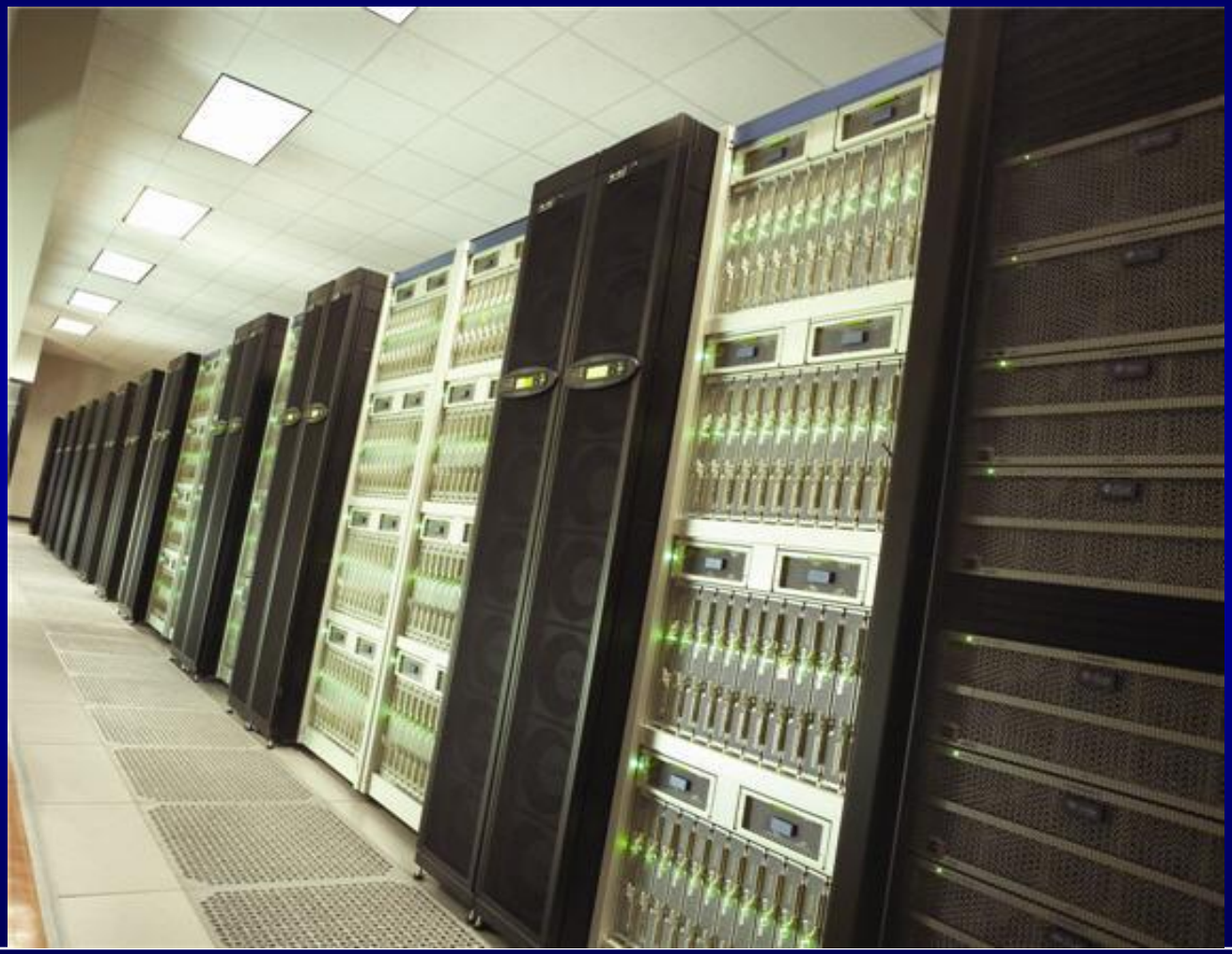
To enable discoveries that advance science and society through the application of advanced computing technologies.

TACC

TACC Overview – Brief History

- **1986 (Classical Era):** UT System Center for High Performance Computing established
- **1992-1997 (Dark Ages):** CHPC budget cut, center moved to UT Austin IT division
- **1997-2001 (Renaissance):** UT Austin supercomputing center a mid-range partner in National Partnership for Advanced Computational Infrastructure (NPACI)
- **1999-2001 (Enlightenment):** internal and external reports advocate need for strong HPC program at UT Austin
- **2001-2010 (Modern Era):** TACC established June 1, 2001, with new name, mission. Now reporting in VP for Research Portfolio

Ranger: World-Class Supercomputing Capability



NSF

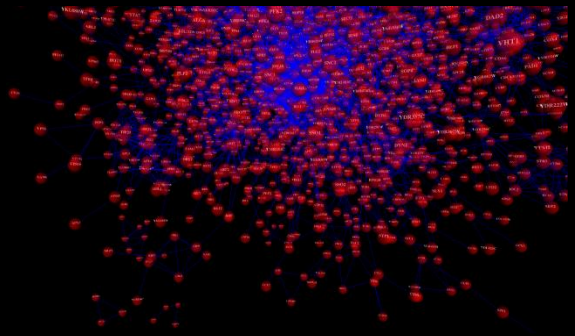
Sun microsystems

AMD Smarter Choice

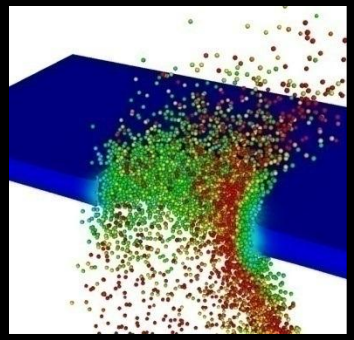
THE UNIVERSITY OF TEXAS AT AUSTIN

TACC

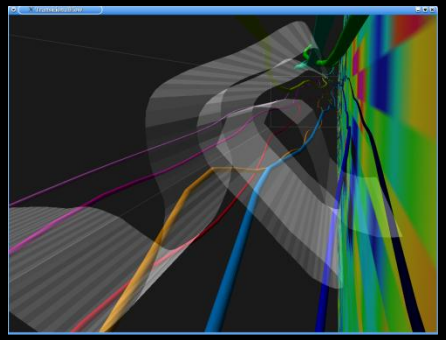
TACC Is a World Leader in Visualization, Too!



Bioinformatics

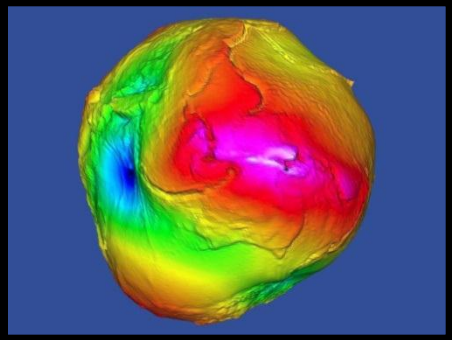


Orbital Debris

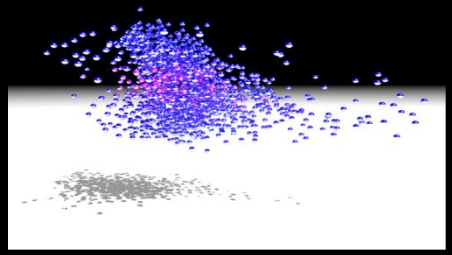


Turbulent Flow

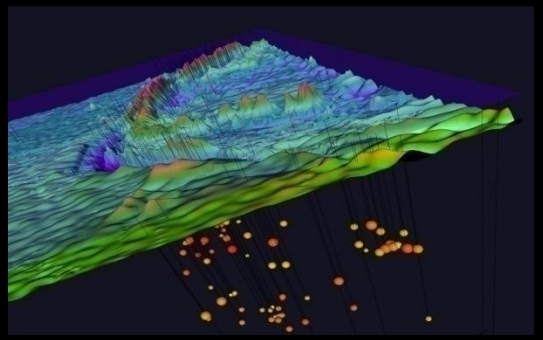
CT Models



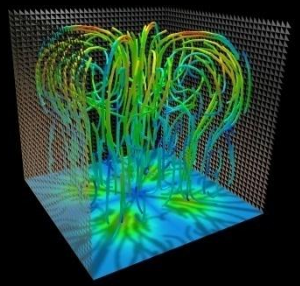
Gravity Map



Quantum Chemistry



GeoSciences



Natural Convection

TACC

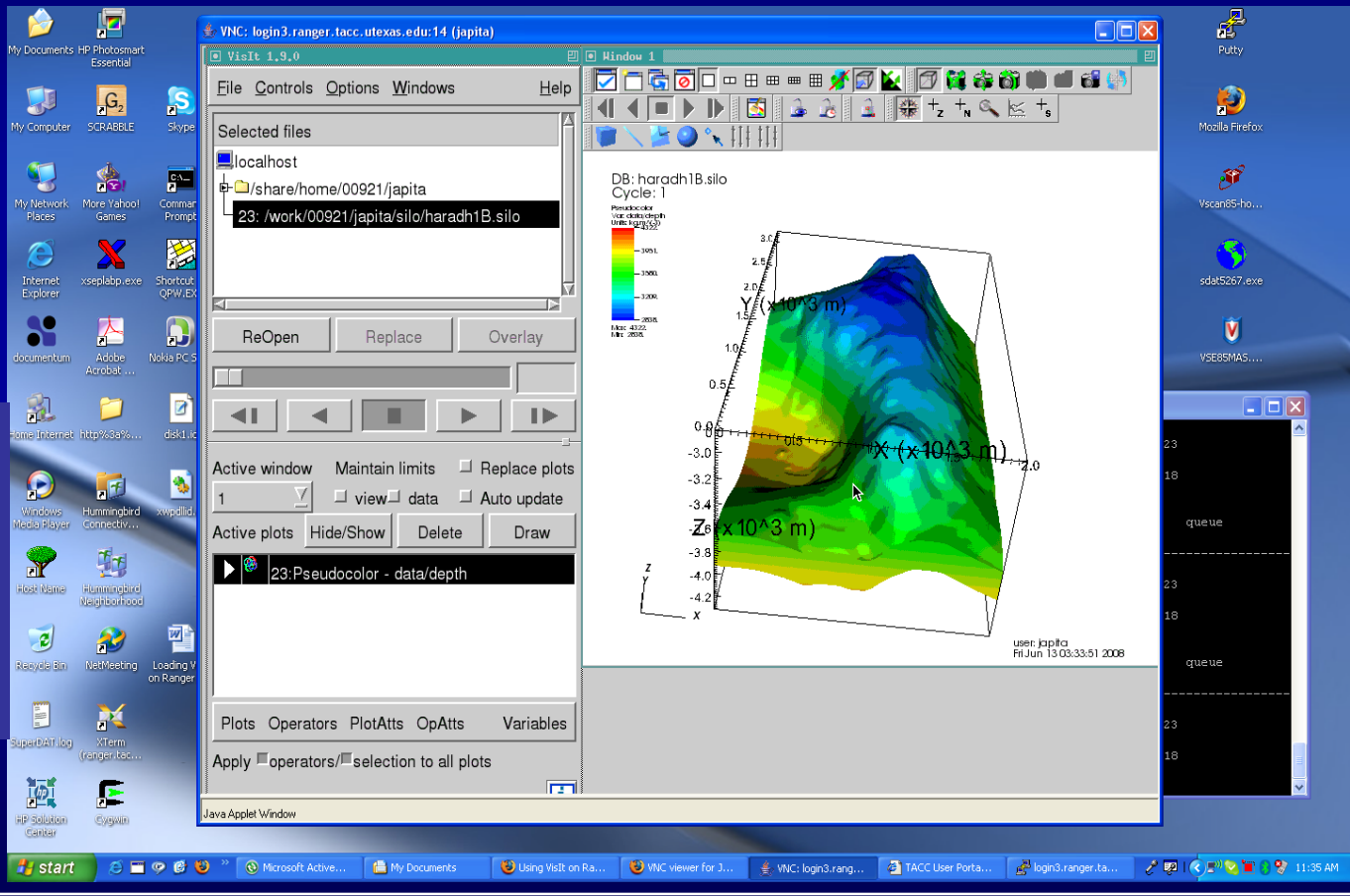
Stallion - Highest Resolution Display Environment in the World

TACC

Remote Visualization

STAR Partner Aramco Services Company is running VisIt software from Saudi Arabia, using seismic data computed on Ranger.

"Visualizing the results right where the data is generated speeds up research considerably."



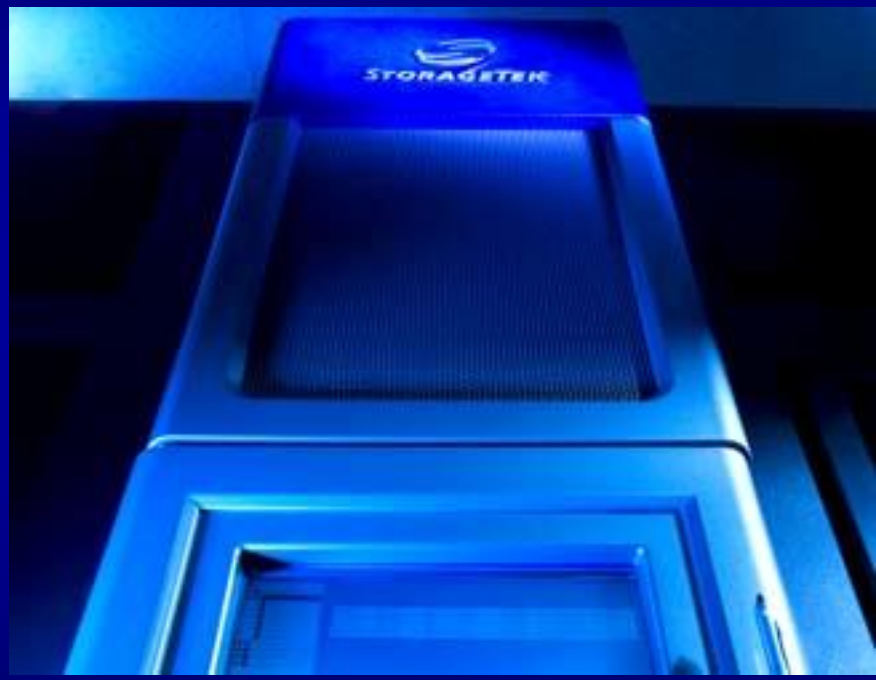
Massive Computing Requires Massive Data Storage: ***Corral***

- 1.2 PB DataDirect Networks online disk storage
- 16 Dell Servers
  - 8 Dell 1950s
  - 8 Dell 2950s
- Mulitple access mechanisms
  - MySQL, PostgreSQL, and SQL Server databases
  - ArcSDE, ArcGIS Server
  - Lustre parallel filesystem
  - iRODS
  - Web-based access

TACC

Ranch Archival System

- Sun StorageTek Silo
  - 10,000 tapes
  - 10 PB capacity
  - Used for long-term storage
  - Not currently ‘allocated,’ but access provided to user of other resources



DMC Overview

- The mission of the data management & collections group (DMC) is to:
  - preserve and make accessible both static and evolving collections of digital data;
  - promote research through the analysis and use of digital data;
  - encourage researchers to organize and make available their digital data;
  - and help integrate diverse collections of digital data into larger and more useful collections.

DMC – Bio Activities

- UT Herbarium
- Texas Natural Sciences Center (TNSC)
- Alaska Herbarium
- Museum of Vertebrate Zoology
- rRNA
- iPlant

PRC

- Plant Resources Center (UT Herbarium)
  - Comprised of The University of Texas (TEX) and Lundell (LL) herbaria.
  - Working to provide online access to the ca. 1,000,000 specimens in the PRC.
  - >400,000 specimens (Texas, Mexico, and type) are databased
  - Data are divided amongst four separate databases with custom schema.
  - TACC is integrating all specimen data to a common transitional schema

PRC (con’t)

- Plant Resources Center (UT Herbarium)
  - High quality scans of all 7500 types, resulting from a Latin American Plants Initiative (LAPI) project supported by the Mellon Foundation, will be integrated.
  - TACC will provide permanent archiving of all data and image files
  - Extensive need to georeference many 10ks of records.

TNSC

- Texas Natural Sciences Center (TNSC)
  - Hosting data and databases for:
    - Ichthyology - Fishes of Texas project, including hydrology data from the Center for Research in Water Resources
    - Non-vertebrate paleontology
    - Odonates
    - Working on forming collaborations with other groups at TNSC including Herpetology, Protists and Vertebrate Paleontology.

Alaska Herbarium

- Host just under 600K image files
- These represent about 200K separate specimens
- ~ 3TB of total storage plus tape backup and geo-plexing at SDSC

Museum of Vertebrate Zoology

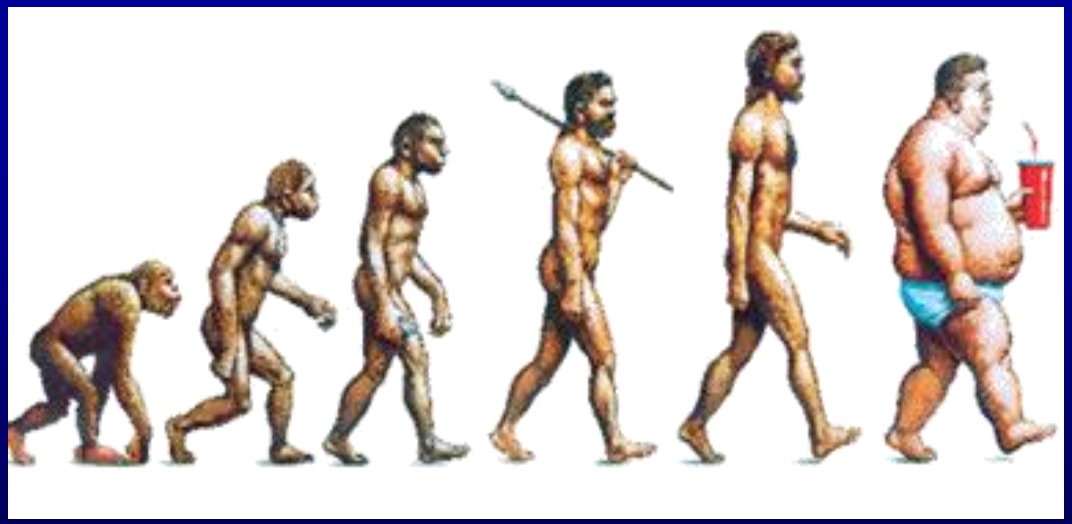
- About 175K files, representing a mix of image and sound data
- Just under 2TB
- Just under a terabyte of 30k killer whale audio recordings

iPlant

- iPlant's mission is to build the CI to support plant biology's Grand Challenge solutions
- Grand Challenges were not defined in advance, but identified through engagement with the community
- A virtual organization with Grand Challenge teams relying on national cyberinfrastructure
- Long term focus on sustainable food supply, climate change, biofuels, ecological stability, etc.
- Hundreds of participants globally… Working group members at >50 US institutions, USDA, DOE, etc.

Why iPlant?

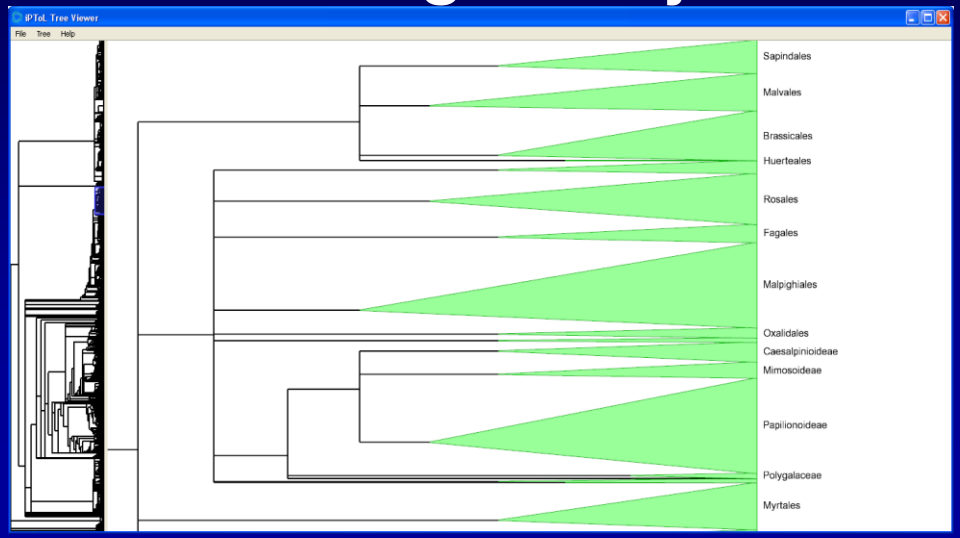
- Activity is being engineered out of life
- Adults eat the equivalent of 4 ½ meals a day
- US Food production @ 3,800 calories /person/day
- Need ~2,000 calories/day at current activity level
- American diet being adopted across the globe



TACC

Recommended Grand Challenge Projects

- Two grand challenges:
- iPlant Tree of Life (IPTOL):
  - Build a single tree showing the evolutionary relationships of all green plant species on Earth
- iPlant Genotype-to-Phenotype (IPG2P)
  - Construct a methodology whereby an investigator, given the genomic and environmental information about a given plant, can predict it's characteristics.



Taken together, these challenges are the key to unlocking many "holy grails" of plant biology, such as the creation of drought resistant or pest resistant crops, or breaking reliance on fossil fuel based fertilizer

TACC

iPlant Cyber Infrastructure

- IPTOL CI:
  - Five areas: Data assembly and integration, visualization, scalable algorithms for large trees, trait evolution, tree reconciliation
- IPG2P CI:
  - Five areas: Data Integration, Visualization, Modeling, Statistical Inference, Next Gen Sequencing Tools

In both, a combination of applying compute resources, developing or enhancing new tools, and creating web-based "discovery environments" to integrate tools and facilitate collaboration.

iPlant: Projects

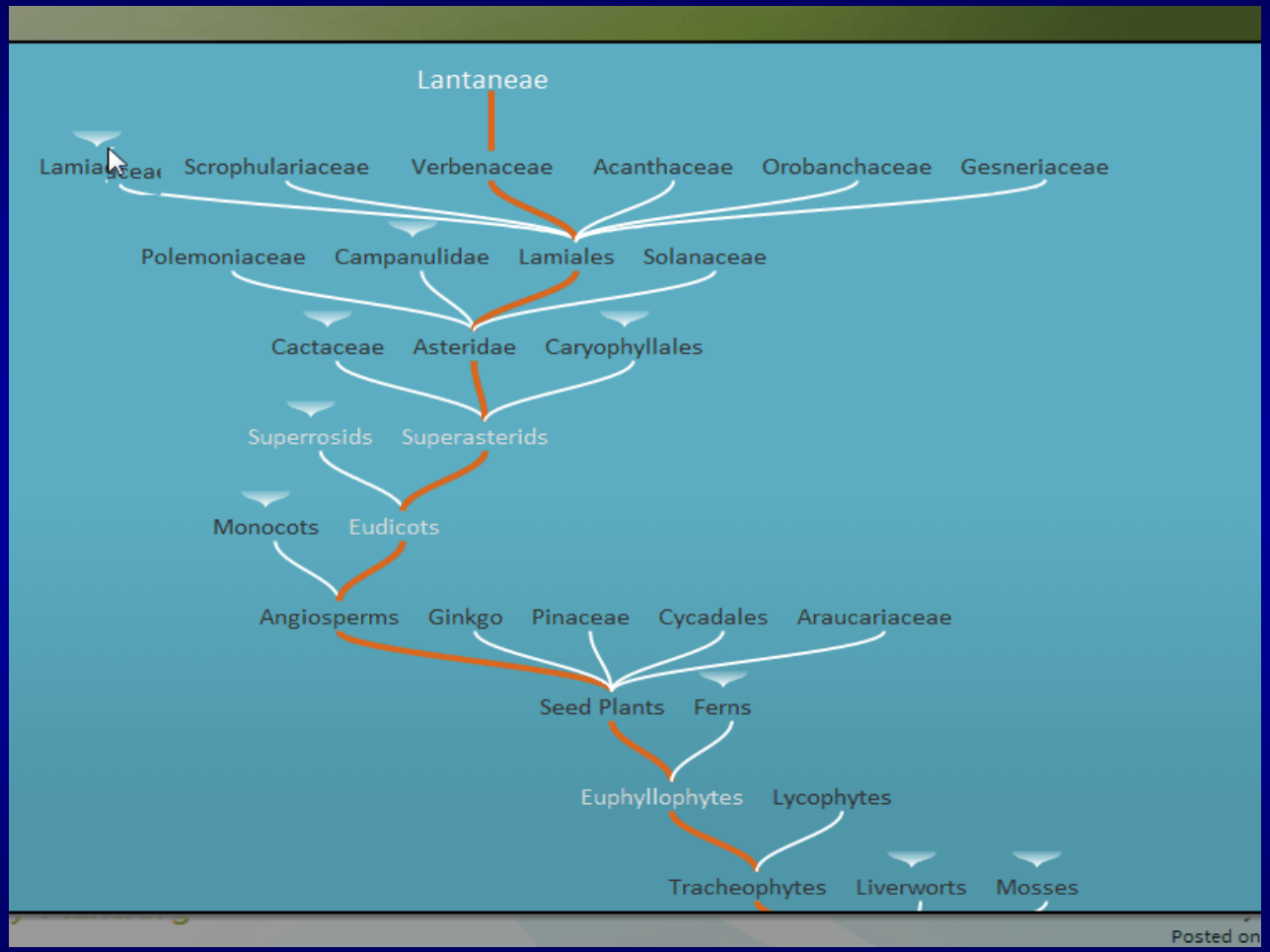
- iPlant includes many projects, two however may be of particular interest:
  - MyPlant
  - DNA Subway

iPlant: Projects: MyPlant

- MyPlant
  - Social networking for plant biologists, but is open to the general public
  - Organized by clade
  - Used to organize data collection for the “big tree”

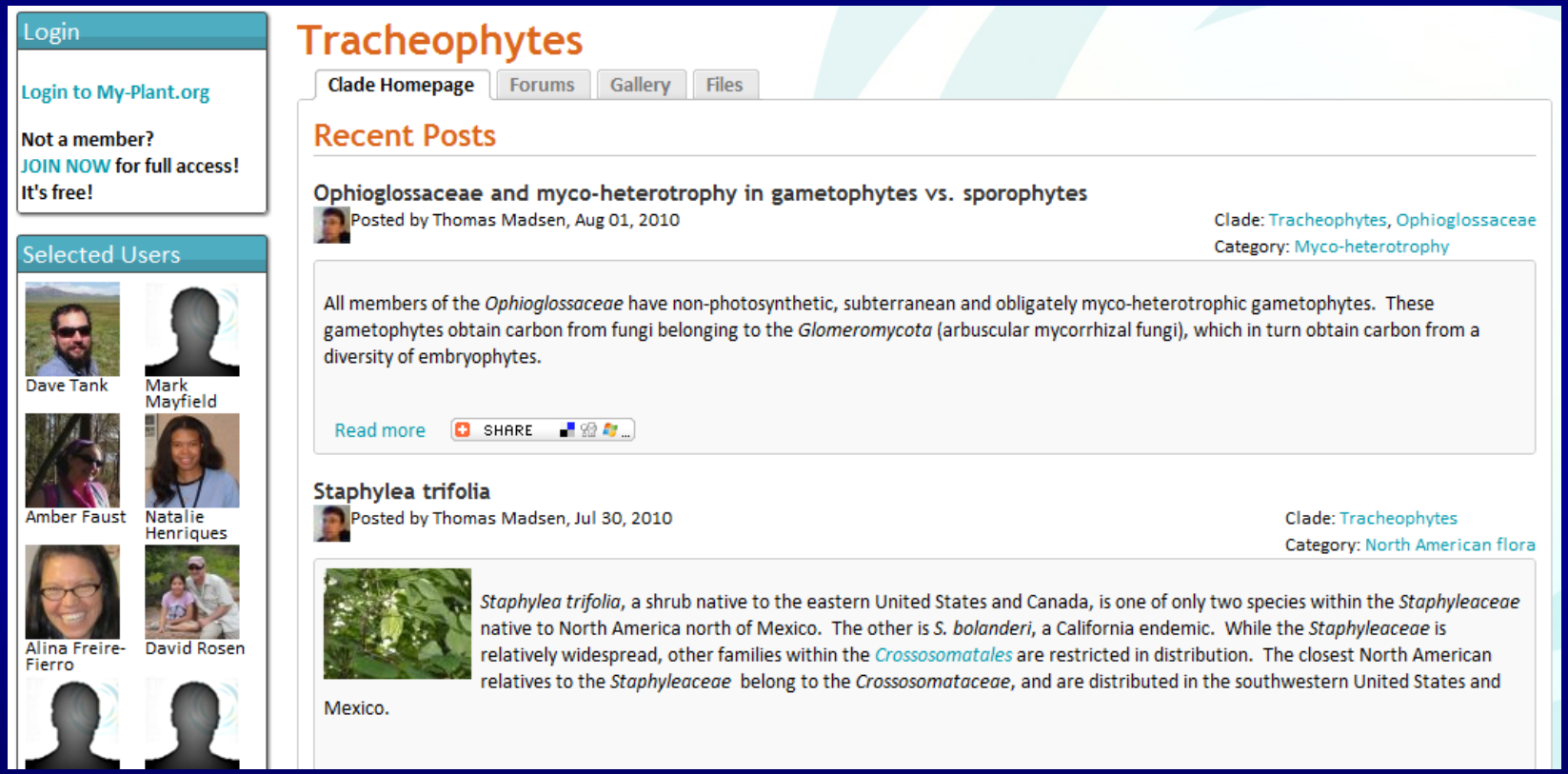
MyPlant

Tree Browser



MyPlant

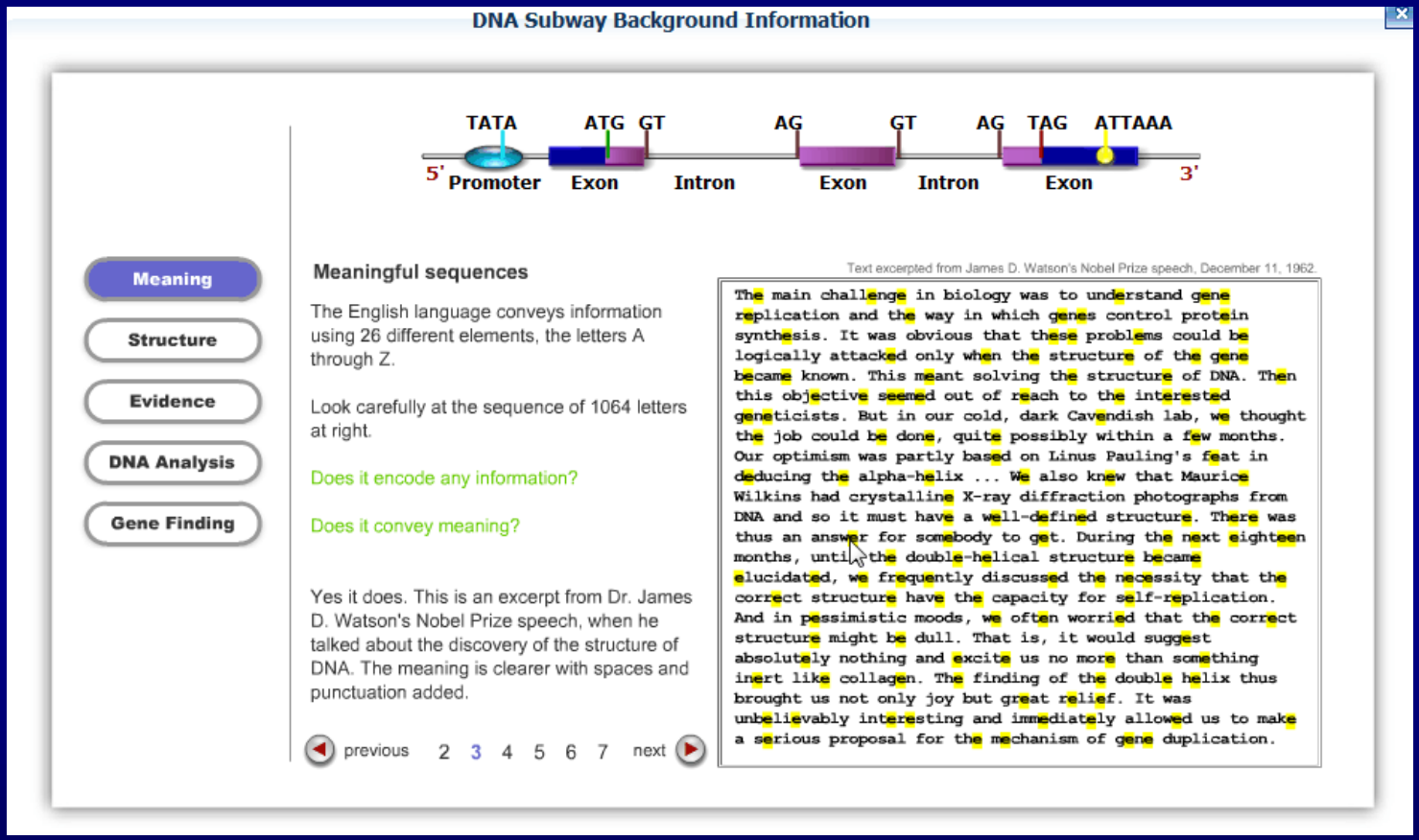
Clade Page



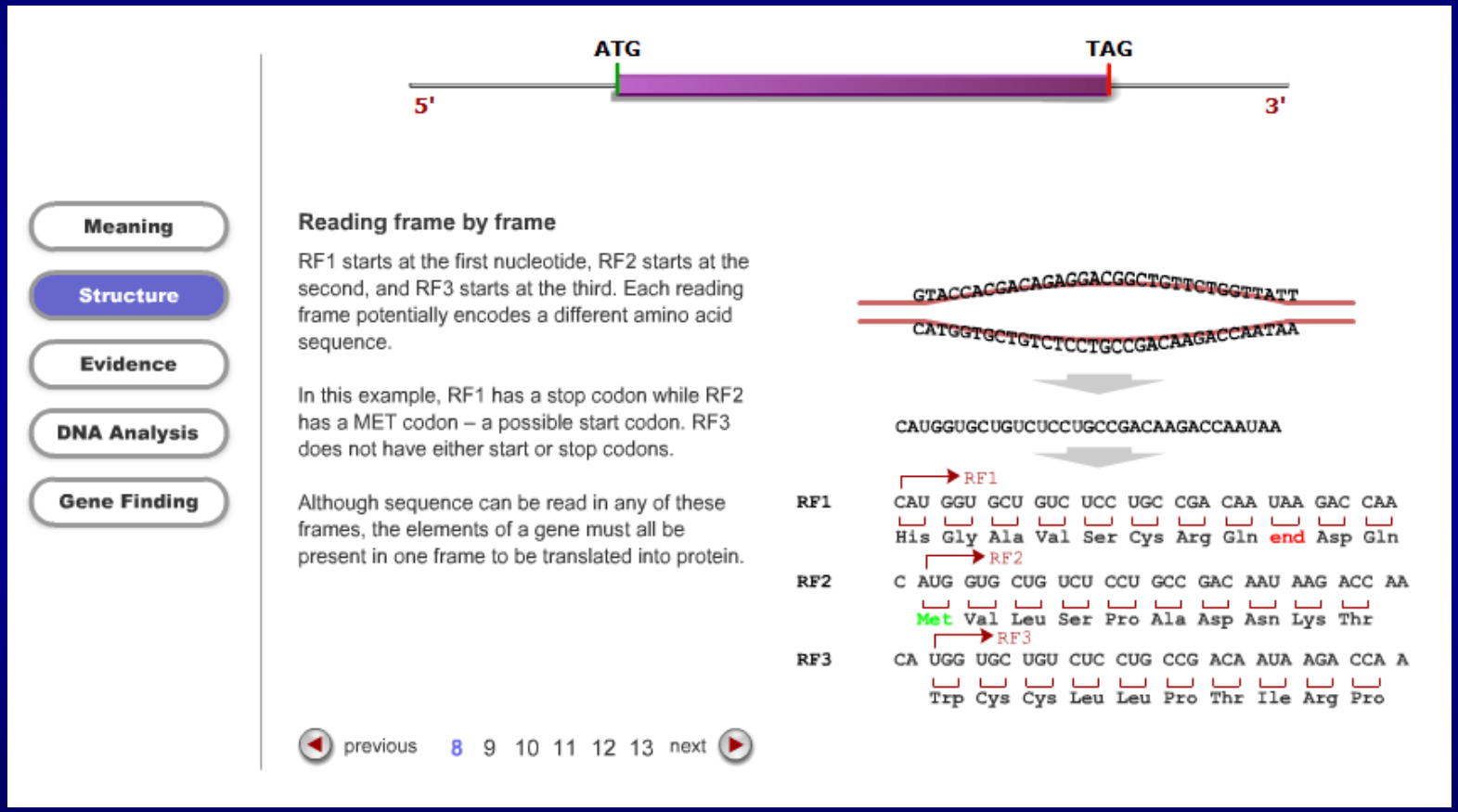
iPlant: Projects: DNA Subway

- DNA Subway
  - *…is a bioinformatics workspace that wraps high-level analysis tools in an intuitive and appealing interface. By "riding" different lines (workflows) you can predict and annotate genes in up to 100,000 base pairs of DNA sequence (Red Line), and prospect entire plant genomes for related genes and sequences (Yellow Line).*
  - High School seen as main target audience

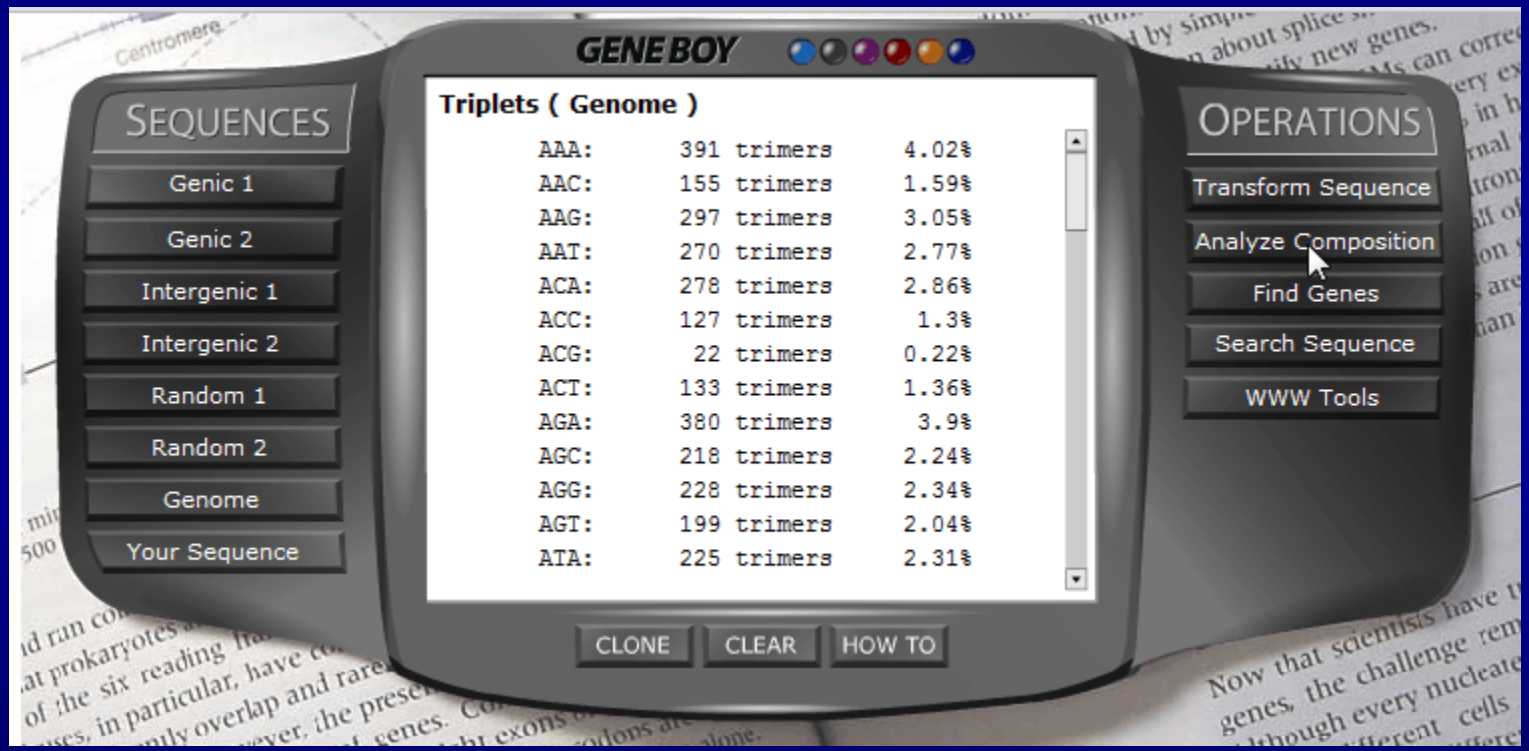
DNA Subway – Background Material



DNA Subway (con't)



DNA Subway Background (con't)



DNA Subway - Tour

- http://dnasubway.iplantcollaborative.org