

TACC, its Natural History Collections and iPlant

AIM-UP
Santa Fe, NM

October 16, 2010

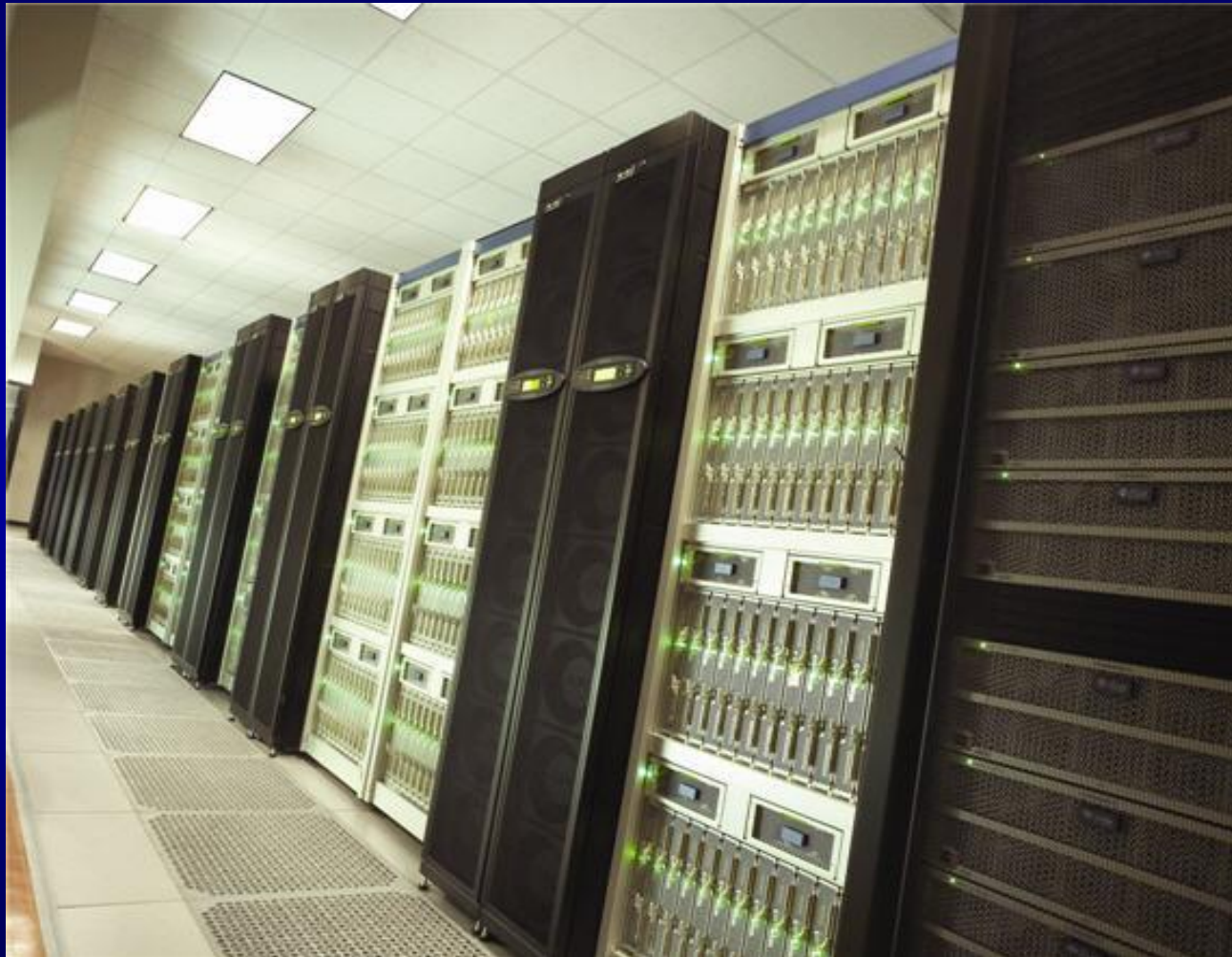
TACC - Mission

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

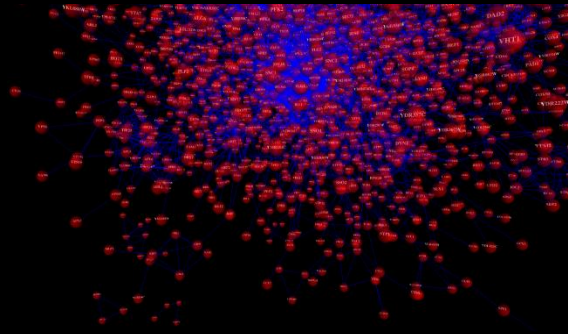
TACC Overview – Brief History

- **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

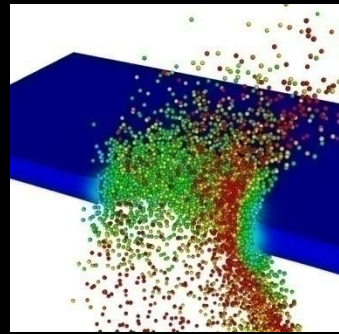
Ranger: World-Class Supercomputing Capability



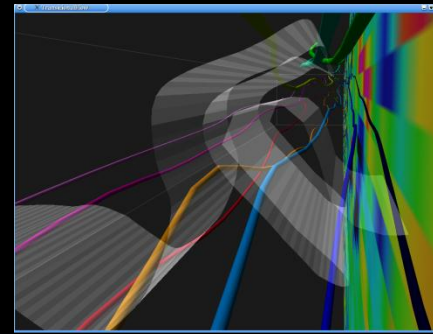
TACC Is a World Leader in Visualization, Too!



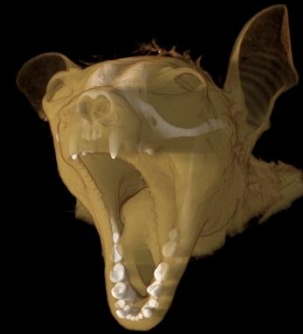
Bioinformatics



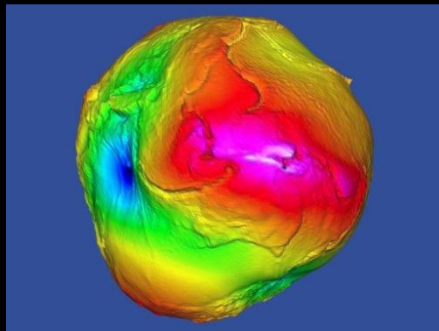
Orbital Debris



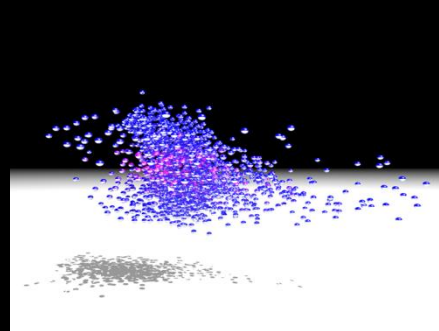
Turbulent Flow



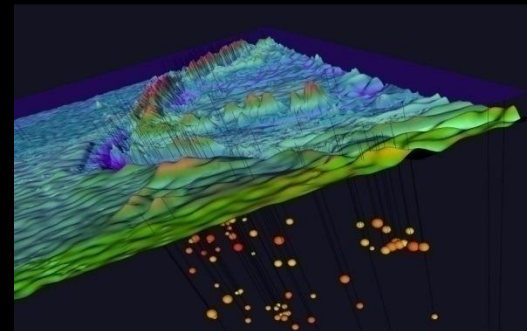
CT Models



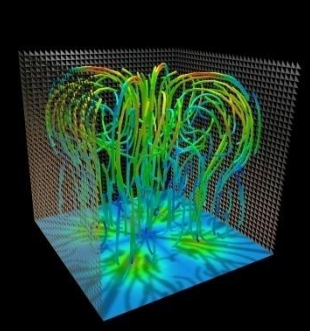
Gravity Map



Quantum Chemistry



GeoSciences



Natural Convection

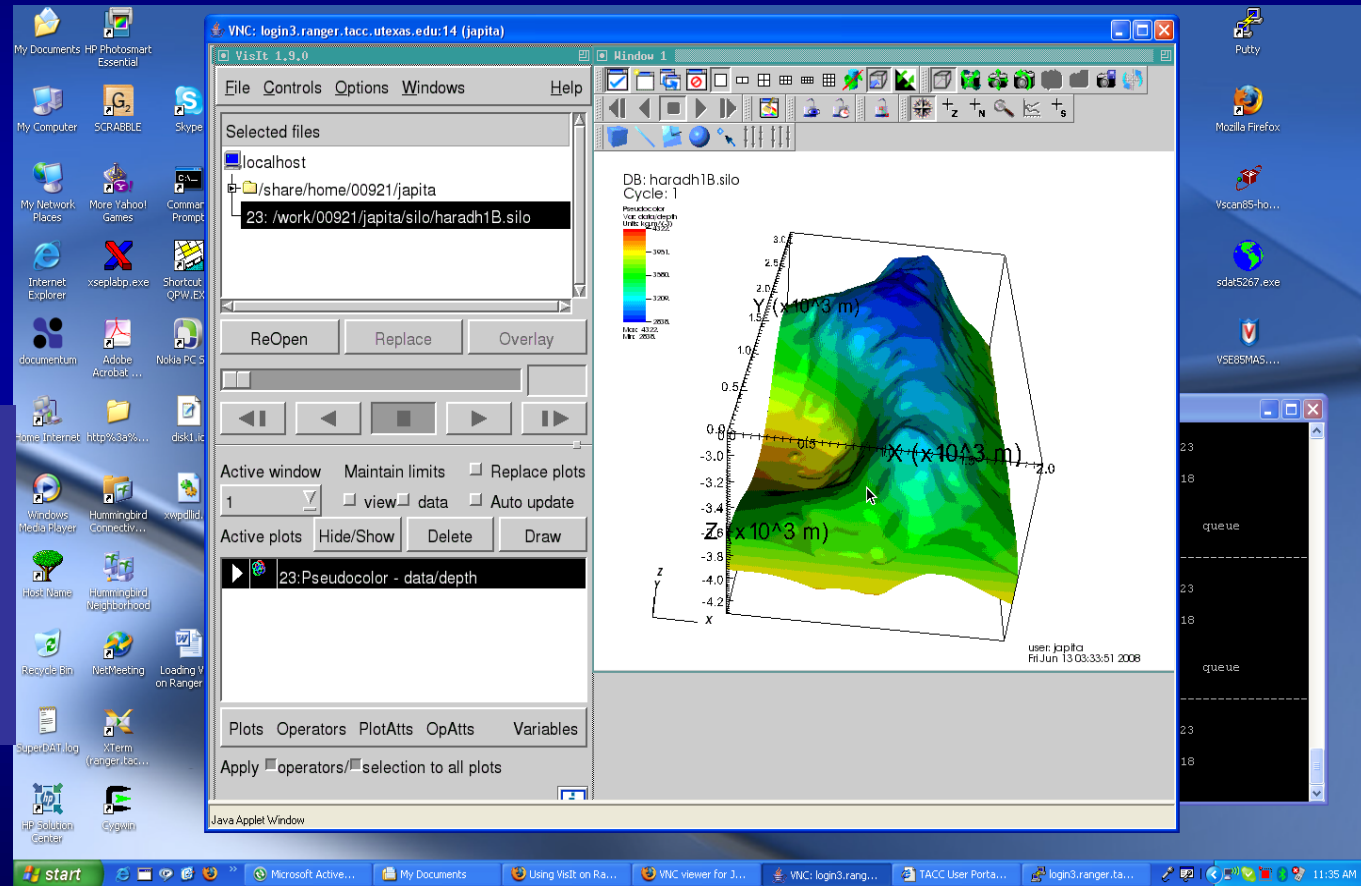
Stallion - Highest Resolution Display Environment in the World



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.”



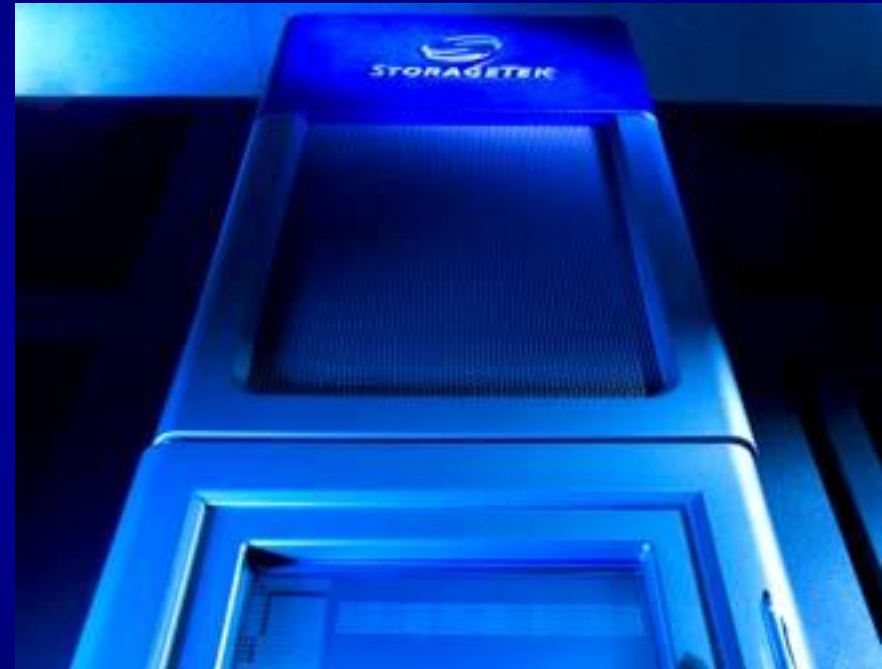
Massive Computing Requires Massive Data Storage: *Corral*

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



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

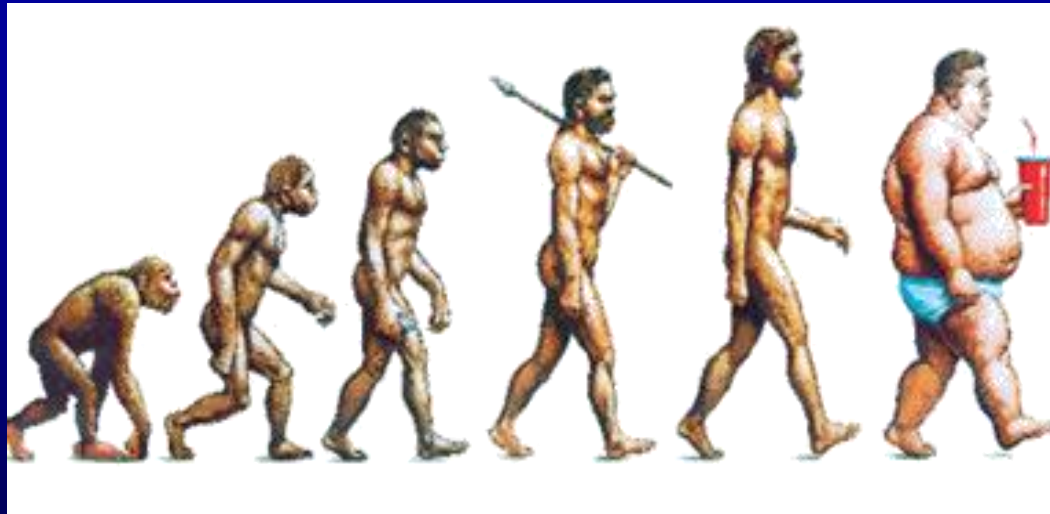
- About 175K files, representing a mix of image and sound data
- Just under 2TB
- Just under a terabyte of 30k killer whale audio recordings

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.

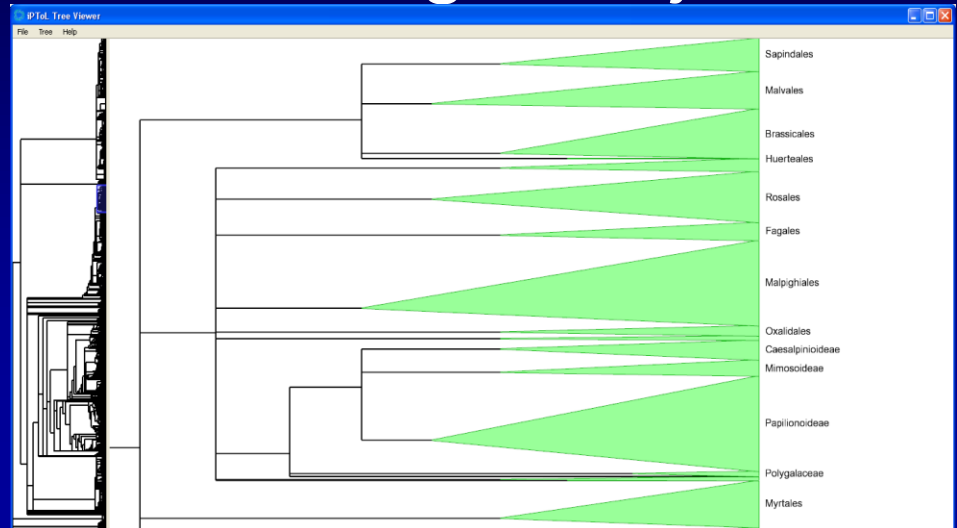
Why iPlant?

- 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



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

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

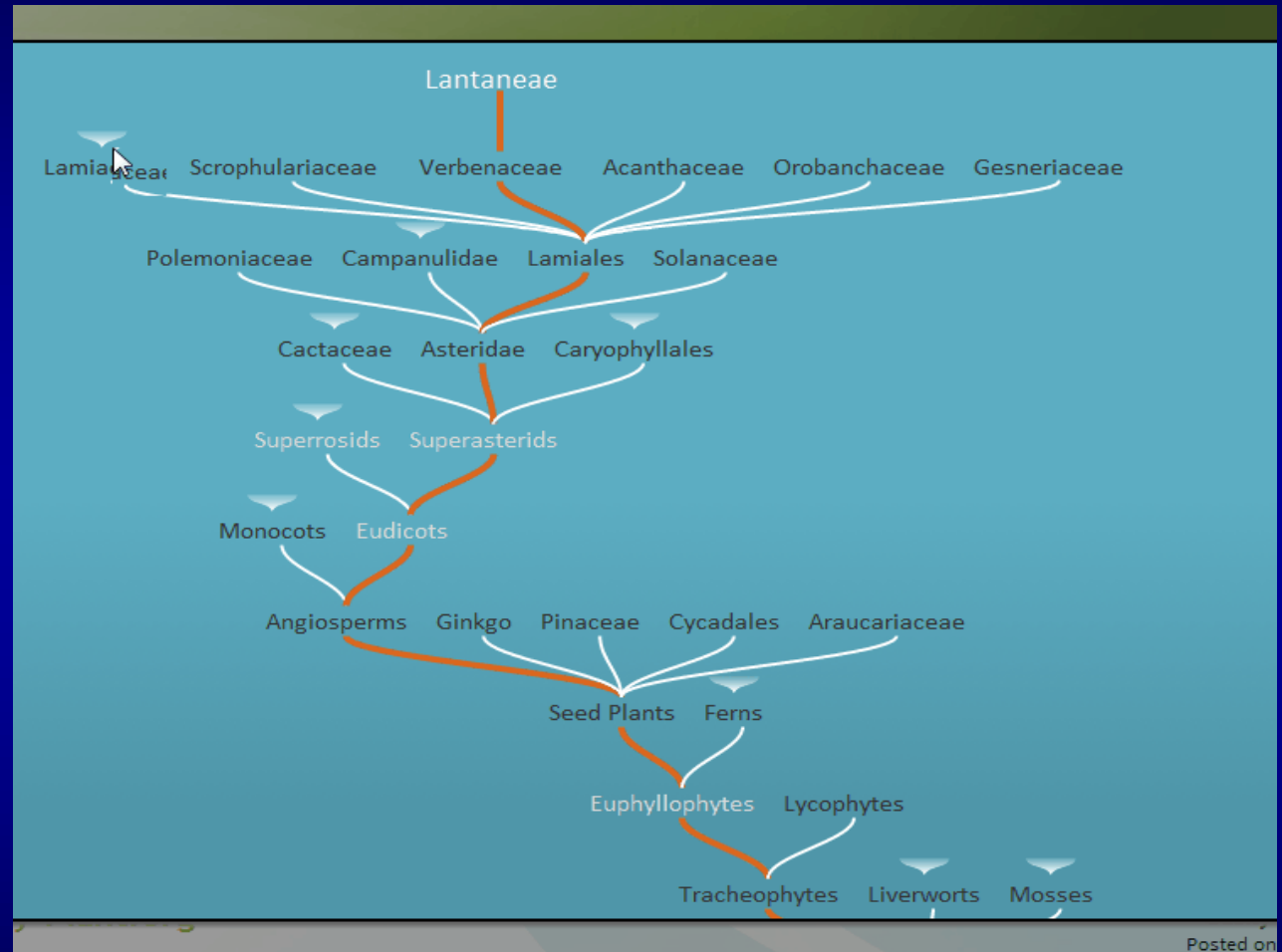
- iPlant includes many projects, two however may be of particular interest:
 - MyPlant
 - DNA Subway

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



Posted on

MyPlant Clade Page

Login

[Login to My-Plant.org](#)

Not a member?
[JOIN NOW](#) for full access!
It's free!

Selected Users

 Dave Tank	 Mark Mayfield
 Amber Faust	 Natalie Henriques
 Alina Freire-Fierro	 David Rosen
	

Tracheophytes

[Clade Homepage](#) [Forums](#) [Gallery](#) [Files](#)

Recent Posts

Ophioglossaceae and myco-heterotrophy in gametophytes vs. sporophytes

 Posted by Thomas Madsen, Aug 01, 2010

Clade: [Tracheophytes](#), [Ophioglossaceae](#)
Category: [Myco-heterotrophy](#)


All members of the *Ophioglossaceae* have non-photosynthetic, subterranean and obligately myco-heterotrophic gametophytes. These gametophytes obtain carbon from fungi belonging to the *Glomeromycota* (arbuscular mycorrhizal fungi), which in turn obtain carbon from a diversity of embryophytes.

[Read more](#) [SHARE](#) 

Staphylea trifolia

 Posted by Thomas Madsen, Jul 30, 2010

Clade: [Tracheophytes](#)
Category: [North American flora](#)



Staphylea trifolia, a shrub native to the eastern United States and Canada, is one of only two species within the *Staphyleaceae* native to North America north of Mexico. The other is *S. bolanderi*, a California endemic. While the *Staphyleaceae* is relatively widespread, other families within the *Crossosomatales* are restricted in distribution. The closest North American relatives to the *Staphyleaceae* belong to the *Crossosomataceae*, and are distributed in the southwestern United States and Mexico.

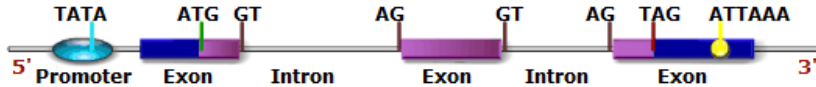


iPlant: Projects: DNA Subway

- 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 Background Information



Meaningful sequences

The English language conveys information using 26 different elements, the letters A through Z.

Look carefully at the sequence of 1064 letters at right.

Does it encode any information?

Does it convey meaning?

Yes it does. This is an excerpt from Dr. James D. Watson's Nobel Prize speech, when he talked about the discovery of the structure of DNA. The meaning is clearer with spaces and punctuation added.

◀ previous 2 3 4 5 6 7 next ▶

Text excerpted from James D. Watson's Nobel Prize speech, December 11, 1962.

The main challenge in biology was to understand gene replication and the way in which genes control protein synthesis. It was obvious that these problems could be logically attacked only when the structure of the gene became known. This meant solving the structure of DNA. Then this objective seemed out of reach to the interested geneticists. But in our cold, dark Cavendish lab, we thought the job could be done, quite possibly within a few months. Our optimism was partly based on Linus Pauling's feat in deducing the alpha-helix ... We also knew that Maurice Wilkins had crystalline X-ray diffraction photographs from DNA and so it must have a well-defined structure. There was thus an answer for somebody to get. During the next eighteen months, until the double-helical structure became elucidated, we frequently discussed the necessity that the correct structure have the capacity for self-replication. And in pessimistic moods, we often worried that the correct structure might be dull. That is, it would suggest absolutely nothing and excite us no more than something inert like collagen. The finding of the double helix thus brought us not only joy but great relief. It was unbelievably interesting and immediately allowed us to make a serious proposal for the mechanism of gene duplication.

DNA Subway (con't)

Meaning

Structure

Evidence

DNA Analysis

Gene Finding

Reading frame by frame

RF1 starts at the first nucleotide, RF2 starts at the second, and RF3 starts at the third. Each reading frame potentially encodes a different amino acid sequence.

In this example, RF1 has a stop codon while RF2 has a MET codon – a possible start codon. RF3 does not have either start or stop codons.

Although sequence can be read in any of these frames, the elements of a gene must all be present in one frame to be translated into protein.



GTACCAGCACAGAGGACGGCTCTTCTGGCTATT
CATGGTGCCTGCTCTCCTGCCGACAAGACCAATAA

CAUGGUGCUGUCUCCUGCCGACAAGACCAUAA

RF1
CAU GGU GCU GUC UCC UGC CGA CAA UAA GAC CAA
His Gly Ala Val Ser Cys Arg Gln end Asp Gln

RF2
C AUG GUG CUG UCU CCU GCC GAC AAU AAG ACC AA
Met Val Leu Ser Pro Ala Asp Asn Lys Thr

RF3
CA UGG UGC UGU CUC CUG CCG ACA AUA AGA CCA A
Trp Cys Cys Leu Leu Pro Thr Ile Arg Pro

◀ previous 8 9 10 11 12 13 next ▶

DNA Subway Background (con't)

The screenshot shows the GENEBOY software interface. The central window displays a table titled "Triplets (Genome)". The table lists 20 different triplets, the number of trimers for each, and their percentage of the total. The triplets are sorted by their percentage in descending order.

Triplet	Number of trimers	Percentage
AAA:	391 trimers	4.02%
AAC:	155 trimers	1.59%
AAG:	297 trimers	3.05%
AAT:	270 trimers	2.77%
ACA:	278 trimers	2.86%
ACC:	127 trimers	1.3%
ACG:	22 trimers	0.22%
ACT:	133 trimers	1.36%
AGA:	380 trimers	3.9%
AGC:	218 trimers	2.24%
AGG:	228 trimers	2.34%
AGT:	199 trimers	2.04%
ATA:	225 trimers	2.31%

The interface also includes a "SEQUENCES" panel on the left with buttons for "Genic 1", "Genic 2", "Intergenic 1", "Intergenic 2", "Random 1", "Random 2", "Genome", and "Your Sequence". On the right, the "OPERATIONS" panel has buttons for "Transform Sequence", "Analyze Composition", "Find Genes", "Search Sequence", and "WWW Tools". At the bottom, there are "CLONE", "CLEAR", and "HOW TO" buttons.

DNA Subway - Tour

- <http://dnasubway.iplantcollaborative.org>