

Minutes of the First Annual NECC meeting held in Burlington, VT March 12, 2010

We were pleased to have Dr. Jennifer Schopf from Office of Cyberinfrastructure at NSF as a visitor for our meeting. Dr. Michael Sayre of NCRP participated in part of the meeting by conference phone. Participant List attached

Executive Committee Meeting and Final Group Discussion

Next meetings:

We decided that the next two meetings will be in Maine in March 2011 and in Delaware in 2012.

In the next rounds of C2 and Track-2

NECC will make proposals for regional fiber enabled research.

Current Track-2 and Supplement Cyber-enabled Research

The Skate genome project is "under control." The genome data will be available to the NEBC soon.

The first annotation jamboree will be on May 24-28 in Delaware. There is funding to support 6 Jamboree participants from each state. The NEBC group needs to get the word out and select participants, but work with the PD or PI from each state.

Metagenome project: We do not yet have a person recruited into the Vermont Bioinformatics Core to organize the algal metagenome project. When that person is hired, we will organize the sampling and sample preparation for the sequencing which will be supported by a supplement to Vermont.

The funding of the project is discussed under the Racer Awards below. The scientists in each of the four NECC states participating in the metagenome project should identify the scientists who will do the sampling and preparation.

- ME - sampling, contact Vicki Nemeth e-mail to Vicki, Ben and Carolyn
- RI- Clinton Chichester will arrange in RI for water and sample preparation
- NH - contact Jeff Schloss and keep Michelle Gregoire in the loop
- VT - contact Jim Vincent

We reviewed the functions of the Water Research Working Group.

The group will review Racer proposals. We should solicit proposals for Year 2 RACER projects, which must be collaborative cyber-enabled research among our NECC states. We should set up the Working Group and make the awards for Year 2 in October 2010. The Racer awards are a means by which we can branch out into marine research in addition to our fresh water algal research.

We will revisit the table in the Track-2 proposal and make sure that the names are still appropriate. Judy Van Houten will solicit nominations from each NECC state:

- Three from each state
- Disciplinary expertise water and bioinformatics
- Agree to meet face to face at least one per state at the review
- The more cyber savvy the better

There will be two RACER awards per year. Judy will send a draft of the RFP to the executive committee to review. Wording will come from the proposal.

- \$10k each
- RFP - to be sent out by May 1
- Short preproposals will be posted on the web by June 1 (science speed dating) to facilitate finding collaborators
- Web site
- Tyler will upload
- 3 page proposals by July 15
- We will provide a Template for the budget
- Review by August 15
- Questions will go to the PI for each state

Review Criteria will include:

- Intellectual Merit, broader impact (collaboration;)
- Interdisciplinary; benefit of collaboration; how they use cyberinfrastructure; expertise of collaborators; next steps toward sustainability
- Final reports to us within one month of finishing the 12 month awards
- Jobs and workforce development are important

RACER funding:

First year RACER is for metagenomics. This will be Vermont's contribution to RACER because the cost of the sequencing will be more than 3 contributions from this state for the RACER program. This may cause a short fall in RACER awards in year 3 as each state will contribute \$4k per year for the RACER awards and there will be two \$10k awards in each of years 2 and 3. We decided to wait until the Year 3 solicitation to determine whether we do, indeed, have a short fall.

For the second year, we need a solicitation (above)

Water Research Working Group:

Supposed to organize a regional water research conference; Maine will host next year (March) by leveraging the Regional Water Research conference that they will hold anyway.

Watershed Project:

We discussed a facebook type site for the project. Liza Ray said that ideally there will be a blog. We thought about NING.

Cybertool Training:

Karl Steiner will work with Steve Borkeske in Delaware on the team that will hold the cybertools seminars.

Assessment:

Before the end of year 2, each state will include Track-2 in their AAAS or other external assessment or evaluation process and provide that section of the evaluation/assessment report to the NECC Executive Committee. At the end of the Track 2 funding period each state will submit their formative evaluation or assessment to the NECC executive committee.

Reporting issues?

There appear to be no problems with the quarterly reporting for Track-2 or the supplements.

TV shows

These are planned for future years through Maine and Vermont.

Find future sources of funding

We should seek new sources of funding for NECC. We should focus on the Bioinformatics workforce development beyond Track-2 and IDeA while we also focus on science drivers.

Michelle Gregoire and others subscribe to NSF and NIH alert services and will share information about upcoming solicitations with the NECC. An important possibility for future NECC funding is the NSF Cyberenabled Discovery Initiative (CDI).

Common Understanding

The Executive Committee will initiate a process that insures that each NECC state has the same understanding of the purpose and scope of work of NECC. Draft wording was discussed and will be distributed.

Individual committee reports follow:

Technical committee

- Last meeting 1 month ago, video conference VT has responses to RFP, it is likely that at least 1 is good. Could have 12 waves to Albany by September.
- RI- Didn't get BTOP in first round, trying again. Building a triangle. If BTOP doesn't happen, maintenance will be a problem. Plan is to move forward on two links and hold the third. Working with Kingston to use municipal space on pole. Also working with a disaster preparedness grant of about \$700k this will help connect URI to Kingston police and possibly hospital. No pushback yet from incumbent carriers.
- NH-Started calling R&E network I-Beam NH. Build from Oxford, lease from Lighttower. UMaine, UNH lit 10G wave to NOX. Major node in Manchester NH. Working on route Manchester-Plymouth-Dartmouth. INBRE Plymouth-Keane. Executed RFP, picked vendor, will have 12 strands. DWDM gear Plymouth and Hanover. Looking at different technologies, price vs capability. Will send map and some notes. Got three credible responses to RFP. Might reach Dartmouth in a year. Still working on what the endpoints look like.

Data below:

NH R&E fiber network is referred to as I-BEAM NH. Infrastructure to Broaden Education Access and Mindshare in New Hampshire UMS invited UNH to join on their fiber from Portland ME to the NoX, also connecting UNH Durham to NoX, and creating a major DWDM hub node in Manchester NH. This is now a live DWDM optical network (since November) and UMS and USNH each have a 10G pipe into the NoX.

UNH and Dartmouth collaborated in the NH component of RII Track 2. We will be acquiring an IRU on 12-strands of fiber from Manchester north to Plymouth State College and Dartmouth. This is the NH NECC network component funded by NSF EPSCoR RII Track 2.

The Vermont NECC segment will meet the NH network at Dartmouth. NIH INBRE supplement will fund a 12-strand IRU from Manchester to Keene State College. We have completed an RFP for both the NSF (Manchester-Plymouth-Dartmouth) and the NIH (Manchester-Keene) IRUs, have selected a vendor, and are in final contract negotiations. The work will require fiber construction over the next year. We will purchase optical equipment next year to light the fiber. We expect to complete the network within the NSF and NIH time limits. UMS operates the network from Portsmouth through Manchester to the NoX. UNH will operate the I-BEAM network north and west of Manchester, with Dartmouth operating their node in Hanover

Northern route may be good for economic development. Still in the talking phase. May have a common vendor in NH and VT. This may be good or bad.

ME- Having trouble with local telephone. Network to Cambridge, 10G, 2 waves to start, 40G capable. Has been working very well. Expanding K-12 network this summer. There are about 900 schools and libraries. Previously, the fastest was about 10 MB. Next year the slowest will be 10MB. Working with Fairpoint for faster tail circuits. Fairpoint is not seeing the positive side and has been pushing back. Trying to pull together RFP for the grant. BTOP funding may allow us to reach further. Similar to NH, ME will put out an RFP with some options. Fairpoint is trying to block Maine Fiber Company from putting fiber on poles. The involvement of multiple branches of government is interesting to watch. Should have RFP out before the end of the month, unclear what the result will be. DE- Infrastructure upgrades going in to add disk capacity for storage and backup. Visual look at the map we need to be sure to consider both sides of the borders when we plan.

There will be transitions where we will have high capacity but not all the way to Boston. We need to continue to be aware that this is an ongoing process that will require cooperation across multiple funding opportunities. We need to stick to the atlantica vision and continue to build it piece by piece.

Wave division from UNH to Plymouth, 12 strands, 6 Dartmouth, 6 UNH system. UVM, to Dartmouth, still figuring out the end connection, considering 10G ethernet. Services to end users? High speed IP, dynamic circuit, dynamic VLAN, file services? It seems that 10Gig ethernet might be a choice. There may be short term and long term answers.

What does the faculty see? Is the traffic separate for data center traffic vs regular? Do we have special peer connections or go through NOX. Campus connectivity (within a campus) will be an issue that we all will be facing. Faster connectivity is a first priority, special peer connections can be added as needed for a relatively low incremental cost. Creating a secondary peering point outside the NOX may be a possible future direction. Priority traffic may be desired, but it may be that priority only matters if congestion is a problem. It is more likely that on-campus networking will be more of a concern than priority traffic across the network.

Albany to Boston, there are numerous possibilities but is not settled.

Federated identity management: Maine has a test Shibboleth server up and running. This is a first step to federated. Vermont has done this at a campus level and is willing and able to help. NSF is pushing InCommon NIH has been pushing InCommon for years, NSF is behind. We reported on a test done live successfully where data in DE was backed up and restored from Maine. Database has been backed up to Maine and Vermont can switch over to Maine. There is a software upgrade going on now that prevents a demo. We are looking at a live backup. Delaware is a folder on Jim's desktop, data stays in Delaware; it is not copied to Vermont. Basic upload and download. Sequencing in Delaware, goes into database, researcher accesses. Fewer headaches in moving data around. Multiple researchers accessing a single point. Authenticating is a logical step. It should be easy for the end user, but also secure. Again, internal campus upgrades may be as much of a concern as the broader network. UNH is working on identity management (as are other groups). They are in a tough situation. Shibboleth has a single sign on regardless of what resource you are accessing. This allows researchers to access resources. NIH uses federated identity management for InCommon. InCommon builds on top of Shibboleth and makes cross institution authentication possible. VT has Shibboleth. DE participates in InCommon. Maine is experimenting. NH and RI do not have Shibboleth. TeraGrid uses InCommon authentication. Separates authentication from authorization. It identifies who you are. Credentials stay at your home site.

Applications have roles that have certain sets of privileges, certain users are set up with roles, and authentication verifies that a person is who they say they are. Is there an InCommon ssh client? TeraGrid is moving toward InCommon Search for Jim Basney at NCSA for more info. What will the data center provide? Web based access, simple, too many features adds complications. Keep it focused down. Jim is working on use cases based on researchers. The answers run the gamut. Shared data center usage should (must?) be multiple institutions working in collaboration; it is not a resource for a single researcher at a single institution. **Mike Sayre is strongly supportive of mandatory collaboration.** Vision: Samples, sequencing, raw data in data center, multiple analyses, adding data products in data center, results shared collaboratively, and annotated data, all will be made public. Another VT project, analysis of Twitter streams, pulls in other data, share data set, allow other researchers to add public data to Twitter data. It is outside of bioinformatics but could fit a shared model. Database is multiuser read and write. Raw data is essentially read only. Concurrent readers and writers have MySQL back end. Snapshots stored offline. Databases are on the order of Gigabytes. Simplified access, data looks local, analysis runs locally; data gets put back in database. Almost all data lives in MySQL. One data repository, but multiple applications can use. Front end talks to database, user sees files. running CLC Bio Genomics on desktop. Genomics pipeline being built open source. TeraGrid is nine sites and will soon be XD. The 9 sites are now two competing teams. There is no common mounted file system for TeraGrid.

Use cases from end users. The technical and the researchers need to work together. Technical people will ask different questions than researchers. It is hard to be user driven without knowing both the needs and the technical capabilities.

Experiences with video conferencing. Recording capability is very good for training. Ease of use is a concern. Portable is very nice. Tanberg. Polycom is great for Polycom to Polycom. Mike Sayer mentioned skype. Audio can conference call. Video is point to point. There are security concerns with Skype, many potential leakage points. Skype is

ok with government. EVO has been tried with some success. Not open source, sometimes it crashes. There is a commercial package Mirial, is 169 euros. Small software development projects. Possibly REU supplement. R01 and R21 from NIH would support software development. Take aways: Continue building, continue talking, and continue thinking about endpoints. Test federated identity management. Shiboleth enabled demonstration should happen in the near future. Talk to George Loftus about OSHEAN plan as this will affect our plans.

North East Bioinformatics Cores (NEBC) Report

Metagenomics

The collaborative project to survey freshwater algal blooms using next generation sequencing requires some coordination among the states. Water researchers in each state who have agreed to perform the sampling need to be contacted and coordinated. Collection, preparation and delivery of the samples for sequencing needs to be assigned to specific people and coordinated in time such that samples can be sequenced together. Bioinformatics analysis will be performed initially by the UVM VGN Bioinformatics Core. All raw data will be made available through the shared data center as soon as it is acquired. Other researchers may perform different analyses or use the data for other projects. James Vincent will follow up with a reminder to all states to begin coordination of efforts.

Little Skate Genome

The Little Skate genome sequencing project is well underway. Libraries have been prepared for sequencing and are queued now at the University of Delaware sequencing center. We expect the arrival of the first data in late March to early April. Assembly of the draft genome will be undertaken by Ben King and James Vincent. De novo assembly software is in place and has been tested with the shared data center. Specialized hardware purchased by the UVM Bioinformatics Core for this project is being shared with Ben King for the analysis. As drafts of the assembly are produced they will be deposited back to the shared data center. Three bioinformatics groups within NEBC had independently assessed software suites for hosting the genome and associated annotations. All came to the same conclusion that GMOD would be most appropriate. The BioIT specialist from Vermont had already been working on developing proficiency installing and maintaining this large suite of programs. Implementation in the shared data center will be a next step.

Little Skate Annotation Jamboree

The first Little Skate genome annotation workshop has been scheduled for the week of May 24 at the University of Delaware under the leadership of Cathy Wu. Workshop presenters have been selected and topics for each day have been scheduled. The committee agreed to try to record the workshop sessions for use as teaching materials by other NECC members. Active recruitment of participants in the workshop will target researchers who are interested in developing new collaborative projects around the little skate genome sequence. Students who are interested in attending future workshops and contributing to the annotation process will also be sought.

Shared Data Center

The NEBC has tested the use of the shared data center for storage and retrieval of data directly from a bioinformatics desktop application. In addition, a test of redundancy and failover capacity was successfully conducted. The shared data center is ready for use in both the metagenomics project and the Little Skate genome project.

Watershed Project Committee -

In attendance: Annette Schloss (NH), Amy Broadhurst (DE), Lillian Gamache (VT), Meredith Simard (VT), Lydia Pitkin (VT), and Liza Ray (moderator- VT).

The Committee reviewed the structure of both the high school and undergraduate components of the program. The undergraduate and high school students from each state will convene for training in Vermont. High school training will be held June 28 - July 2 and undergraduate training will be held May 31 - June 4. States either have or are in the process of identifying mentors with whom they will match their students. Students will be under the direction of the mentors once they have completed training in Vermont and until their final presentation at the annual symposium. All participants will reconvene in April for a symposium.

Mentors and coordinators have access to existing resources that originated with the VT EPSCoR Streams Project, such as training manual, outreach materials, and laboratory methods and protocols. Liza Ray, VT EPSCoR Streams Project Coordinator and NECC Watershed Project Committee Chair, will work with mentors from each state to ensure that they understand what material will be covered at the training week and she will address content issues related to students' independent projects where appropriate during the training week.

Students from DE, ME, NH, and RI will not be required to collect data in the same format or using the same protocols as the VT EPSCoR Streams Project. However, all projects should maintain a focus on watershed issues and align with the Streams Project data where feasible and appropriate. The specific focus of each student or group's project will be shaped around the interests and data collection capabilities of specific mentors and their associated facilities.

The methods, quality control procedures, and raw data that each group collects should be well-documented and submitted to the VT EPSCoR Streams Project coordinator. The coordinator will post the data and methodologies used to collect them to the Watershed Project Web site. Because the Watershed Project is mainly a workforce development program, the data do not need to contribute to a common database that is of sufficient quality to use in peer-reviewed publications. This strategy emphasizes our goal of workforce development.

The Watershed Project will include links to Picture Post (<http://picturepost.unh.edu/>), a program hosted by UNH and led by Annette Schloss. We will also develop interactive components on the Watershed Project Web site, such as videos, and additional training resources, such as data analysis tutorials. For the 2010-11 year, the Watershed Project will host a non-interactive "Facebook"-like page that features participant bios. Over the course of the 2010-11 year, the Watershed Project will deploy an interactive Web environment in which participants can communicate with each other.