HIP: Hackathons, Interoperability, Phylogeny

SHORT TITLE: HIP

Rutger Vos, PhD; University of Reading; r.a.vos@reading.ac.uk
Enrico Pontelli, PhD; New Mexico State University/Computer Science; epontell@cs.nmsu.edu
Arlin Stolzfus, PhD; University of Maryland/NIST; arlin@umd.edu

PROJECT SUMMARY

The potential for synthetic research based on aggregating, integrating, and re-using data is enormous, yet most resources remain interoperable. To realize this potential, software and databases that handle evolutionary trees (and their associated annotations) must be interoperable. Interoperability, in turn, requires tools based on common standards. In the past few years, evolutionary informaticists, with help from NESCent, have been building a software toolbox for solving interoperability problems, based on the EvoIO "stack" of NeXML, CDAO and PhyloWS. This toolbox makes it possible to begin building a worldwide network of interoperable evolutionary resources. The HIP (Hackathons, Interoperability, Phylogenies) aims to use the hackathon mechanism (which we have helped to develop at NESCent) to grow this network directly, by adding links to it, and indirectly, by creating examples for others to follow. To support this project within a working-group budget, we leverage support from strategic partners. Each of the planned series of 3 hackathons will bring together scientific programmers with related challenges. The hackathons target early-career scientists, who often have the most technical expertise and the most potential to pass along their skills and enthusiasm.

PUBLIC SUMMARY

The Internet increases the potential for scientists to share information, find new patterns, and test ideas. Yet, this potential often is not realized, due to software components' inability to work together (commonly referred to as "interoperability"). To realize the potential of synthetic evolutionary science software and databases that handle evolutionary trees (and their associated annotations) must work together, i.e., they must be "interoperable". Interoperability, in turn, requires tools based on common standards. In the past few years, evolutionary researchers, with help from NESCent, have been building a software toolbox for solving interoperability problems. This toolbox makes it possible to begin building a worldwide network of interoperable evolutionary resources. Growing this network is the goal of the HIP (Hackathons, Interoperability, Phylogenies) working group. We aim to grow the network directly, by adding links to it, and indirectly, by creating examples for others to follow. To achieve this, we will stage "hackathons": small, intense meetings that bring together scientific programmers to bring down interoperability barriers. The hackathons target early-career scientists, who often have the most technical expertise and the most potential to pass along their skills and enthusiasm. The hackathons will have different themes, and will focus significantly on the needs of strategic partners.

INTRODUCTION AND GOALS

Sidlauskas, et al (2010) make a strong case for the value of synthetic research defined as "the extraction of otherwise unobtainable insight from a combination of disparate elements". They distinguish several modes of synthesis: data aggregation, methodological integration, conceptual synthesis and reuse of results. Synthesis as a problem in knowledge engineering requires semantic integrity across boundaries (e.g., disciplinary or methodological). That is, in order to re-use, aggregate, or integrate data (in the sense of "information"), the meaning of data mean must be clear to an information consumer in a context other than its original. This requires agreements about data representation. Furthermore, to support large-scale integration, data must not require "hands-on" interpretation, but must be accessible to automatic processing by software.

Thus, synthetic research depends on development and adoption of standards for knowledge representation, and the technology to support these standards, so as to make data accessible, searchable and combinable. Examples of this were shown at the 2010 iEvoBio meeting, where the adoption of this approach by the TreeBASE project made its data combinable with the Tree of Life and the UniProt project,

searchable in new ways, and accessible in new visualizations such as superimposition of taxa on Google maps (Vos, et al., 2010). The innovations that make TreeBASE interoperable were enabled by NESCent's forward-looking support for behind-the-scenes technology development. From 2006 to 2009, NESCent supported an "evolutionary informatics" working group that spawned a trio of projects:

- **NeXML**, an XML format that allows phylogenetic data to be expressed in a predictable and validatable way (Vos, et al., in prep.; http://www.nexml.org). Data expressed in NeXML can be annotated with terms from the CDAO and other knowledge representations, thereby allowing metadata stored by community resources to be "unlocked" and available for clients, strategic partners and third party projects to be integrated in novel ways.
- **CDAO**, the Comparative Data Analysis Ontology (Prosdocimi, et al., 2009 http://www.evolutionaryontology.org). The CDAO provides a framework for the explicit, computable representation of core concepts in comparative evolutionary analysis. It allows for the import of other vocabularies for data and metadata: this is essential to keep pace with synthetic use of data from other domains of knowledge.
- **PhyloWS**, a web services standard (http://evoinfo.nescent.org/PhyloWS) that provides a common application programming interface for phylogenetic resources. This allows phylogenetic data to be queried and identified in a way that is agnostic of the underlying implementation of the resource.

Together, these form the **EvoIO stack**, designed to enable a global network of interoperable resources. Using this stack, TreeBASE can now represent its contents using CDAO and serialize it using NeXML, and data can be searched using the PhyloWS standard. Thus, NESCent's past efforts have increased the potential for a global network of interoperable phylogenetic resources to emerge. However, the EvoIO stack is in its early stages, and it is not easy for novices to deploy. Most researchers do not know that it exists or are unaware of its benefits. The sociology and infrastructure of science (its system of hiring, promotion and funding) discourage forward-looking technology changes that would benefit the entire community without providing direct and tangible benefits to a resource-provider.

We propose to grow the network of interoperable evolutionary resources by forming a working group to organize a series of hackathons. Such meetings have been successfully organized for evolutionary informatics at NESCent before (Lapp, et al., 2007, Lapp, et al., 2009). The proposed hackathons promote synthesis by the development of standards-compliant software. Achieving this in parallel projects will increase connectivity of a network of interoperable community resources, which in turn will increase the capacity for end-users to conduct synthetic research based on aggregating data, integrating resources, and automating analyses on a large scale.

The working group promotes the adoption of common standards in phylogenetics by raising awareness of their existence and utility and by sharing knowledge on how to deploy them. Young scientists will be trained in the application of best practices to implement standards and stack technologies. This will go beyond the hackathons in initiatives such as Google Summer of Code projects or graduate fellowships. Leaders of community resources will learn to leverage software engineering concepts, programming techniques and communication tools to manage software development teams.

PROPOSED ACTIVITIES

The proposed activities will be carried out by the working group and the hackathon participants. The working group remains relatively constant in membership and provides planning, follow-up and evaluation for all aspects of the project, while hackathons will have different sets of participants.

Planning and partnering by the working group - We will recruit additional working group members (up to a total of 8 to 10), from key projects, NESCent staff, and under-represented projects. The working group will meet at NESCent in mid-2011 to refine its strategic vision, and to begin planning the first hackathon. Subsequently, the working group will hold quarterly teleconferences and communicate on-line. From past experience, the proposers are adept at this mode of collaboration. The working group is responsible for engaging strategic partners and pursuing further financial support.

Hackathons - Organization of hackathons will follow the scheme established by previous NESCent hackathons, which typically host 20 participants for 5 days. The first day will begin with talks on best practices and common resources, and will end with participants self-organizing into project groups (3 to 7 members), based on OpenSpace principles. Subsequent days will be spent on projects. Participants will be instructed to develop links that directly build the network of interoperable resources available to end-users, or to work on reference implementations and proofs-of-concept that will inspire such links.

Participants are roughly a 2:1 mixture of invitees and applicants responding to an advertised call for participation. By inviting participants, we take advantage of the fact that real networks (social, biological, computer) have critical nodes with a high degree of connectivity. Key nodes in an interoperable network of tree-related resources might include EoL, iPlant, NCBI, TreeBASE, ToLWeb and others. By issuing an open call, we expand our network of connections, encouraging participation from under-represented groups and developers whose resources are not well known or widely used. Applicants are selected on technical ability, collaborativeness, strategic opportunities, and diversity.

The proposers do not envision the development of novel resources or databases; rather, the scope of the hackathons is to improve interoperability between existing resources. For three hackathons, we identify areas of interest and projected key participants. The hackathons are ordered such that one builds on the previous, and so, in addition to key participants and interoperability experts, there are return visitors to enable the transition from one hackathon to the next.

Hackathon 1: Data resources. NESCent (Durham, NC) Winter 2011 – will focus on key data providers. Hackathon projects will be in the area of supporting import, querying and export of richly annotated data. External projects targeted for hackathon participation will include TreeBASE, Dryad, the Tree of Life web project, PhenoScape, MorphBank, MorphoBank, TimeTree and PhylomeDB.

Hackathon 2: Data integration environments. Field Museum (Chicago, IL) Summer 2012 - will focus on data integration and exploration environments. Hackathon projects will take advantage of the accomplishments of the first hackathon and will include aggregating data via phyloreferencing and taxon referencing, and integrating data exploration environments with image repositories. External projects targeted for hackathon participation will include BioSync, TOLKIN, PhyLoTa, pPOD and the iPlant discovery environment.

Hackathon 3: Visualization tools. University of Arizona (Tucson, AZ) Winter 2012 - will focus on visualization of rich phylogenetic data as is available from data providers and data integration environments. Hackathon projects will take advantage of preceding hackathons to enable visualization of semantically annotated phylogenies (e.g., ones that incorporate character state changes and other biological events such as speciations, extinctions, gene duplications and metadata). External projects targeted for hackathon participation will include iPlant, Mesquite, PhyloBox, jsPhyloSVG, PhyloWidget and Archeopteryx.

Follow-up - Hackathons produce tangible outcomes such as proof-of-concept software and code revisions, and intangible outcomes such as agreement on best practices, awareness of available resources, and opportunities for collaboration. From experience, we know that these intangible outcomes bear fruit after hackathons end. However, tangible outcomes (though publicly available often do not bear fruit for scientific end-users. To address this we plan to put more emphasis on project follow-ups. First, we will stress that each team aim to produce: a stable software deliverable; a proof-of-concept used to gain further support; a technical publication; or a plan for a Google Summer of Code project, graduate fellowship or visiting scientist visit. Second, we will keep a list of projects and their current status on the working group web site. Each project will be assigned a working group member, who tracks the status of the project and advises participants on how to bring the project to fruition. Working group teleconferences will include, as a fixed agenda item, reports on project status.

Participating Fields and Partial List of Proposed Participants

The working group will include the 3 authors of the present proposal, Rutger A. Vos (post-doc, phyloinformatics), Arlin Stoltzfus (senior researcher, bioinformatics & evolution) and Enrico Pontelli

(professor, knowledge representation & reasoning, as well as Dr. Mark Westneat, and a set of 4 to 6 other individuals chosen from key projects, NESCent staff, and under-represented projects. Hackathon participants are not known in advance. We will be careful in extending invitations and in reviewing applicants to an open call. Experience indicates that it is not sufficient merely to pick a representative of a targeted resource: the applicant's ability to collaborate and their technical expertise are crucial. These events naturally attract early-career scientists.

RATIONALE FOR NESCENT SUPPORT

Many potential hackathon participants belong to NESCent's in-house community, and several of the projects developed by them make excellent targets for hackathon projects (see collaborations with other NESCent activities, below). The proposers note the excellent informatics support provided by NESCent staff and whose help in meeting the IT needs for the hackathons will be invaluable. NESCent has world-class IT and logistic resources and the know-how to host hackathons. The proposers recognize NESCent's unique culture of institutional support for initiatives such as ours. In contrast, other granting agencies usually do not support sustainable, ongoing development of infrastructure that serves community needs. Lastly, the proposers have been instrumental in developing the hackathon strategy deployed at NESCent, whereas other agencies might underestimate the value of this approach. Indeed, NSF declined a proposal for a phylogenetic Data Interoperability Network (Stoltzfus, et al, 2009) that included many of the ideas proposed here. However, NESCent understands the strengths of this approach, as well as its weaknesses (which we aim to address in our project).

Note on budget considerations

We do not expect NESCent to commit more funds to this project than it would to a typical working group. We estimate the cost of a typical hackathon at 2 to 2.5 times that of a working group meeting (allowing that hackathons outside of NESCent may entail extra costs). Therefore, the potential of our proposal depends on our ability to secure external funds, which so far include two major commitments and the following additional commitments to fund the travel of personnel:

- $15K from iPlant to co-sponsor a hackathon at their Arizona location (letter, Dr. S. Goff)
- $10K from the EoL BioSynC to co-sponsor a hackathon (letter, Dr. M. Westneat)
- 3 person-trips for TOLKIN project personnel (letter, Dr. N Cellinese)
- 3 person-trips for working group leader (Dr. A Stoltzfus)
- 2 person-trips for Biodiversity Synthesis Center of EoL personnel (letter, Dr. M. Westneat)
- 1 person-trip for working group leader (Enrico Pontelli)

We estimate that these commitments are sufficient to stretch a normal working group budget to cover 1 working group meeting and 2 hackathons. We intend to secure further support to allow a third hackathon as our plans mature over the next year. Options to obtain further support include:

- Applying to organizations such as NCBO and the Phenotype RCN for meeting support
- Submitting an NSF workshop proposal to fund a full hackathon
- Requesting travel support from individual grant-funded projects

To support hackathon followups, we have some of the same options, in addition to applying for NESCent short-term visiting scientist funds.

COLLABORATIONS WITH OTHER NESCENT ACTIVITIES

There are many researchers at NESCent who we'd like to invite to the hackathons. We note especially the following individuals and the areas of their expertise relevant to the proposed hackathons: Jim Balhoff's implementation of NeXML support with EQ annotation of character states in Phenex/PhenoScape; Vladimir Gapeyev's contributions to TreeBASE; Ryan Scherle's expertise with Dryad; Hilmar Lapp's development of the PhyloWS standard; Jeet Sukumaran's contributions to the design of the NeXML standard and of DendroPy.

ANTICIPATED IT NEEDS

The working group does not expect long-term maintenance by NESCent of a public resource. In addition to communication tools we will supply ourselves (mailing list, wiki at http://www.evoio.org, live channels such as friendfeed or twitter) we envision the following IT needs:

- **Conferencing facilities** for conference calls when organizing the hackathons, and video-conferencing during the events. We would like to use NESCent's infrastructure for this.
- **LCD projectors** for group programming and wiki review, ideally for each hackathon team.
- **WiFi access** for all participants at the hackathons.

PROPOSED TIMETABLE

- **Throughout** - The working group has quarterly teleconferences throughout its mandate period.
- **Summer, 2011** - The group meets at NESCent to develop a strategic vision, and to develop themes for the first hackathon. Planning for the hackathon begins immediately thereafter.
- **Winter 2011 or Spring, 2012** - Over the past 3 months, the working group has selected applicants. First hackathon takes place, probably at NESCent.
- **Summer or Fall, 2012** - Over the past 3 months, the working group has selected applicants. Second hackathon takes place, probably at the Field Museum in Chicago.
- **Winter 2012 or Spring, 2013** - Over the past 3 months, the working group has selected applicants. Third hackathon takes place, probably at the University of Arizona at Tucson.

ANTICIPATED OUTCOMES

The working group will develop a wiki publicizing its strategic vision for a network of interoperable resources. It will maintain a publicly accessible spreadsheet with the current status of hackathon projects. By mid-2012, it will produce a report for publication on progress in achieving its strategic vision. In addition to describing tangible outcomes, this report will serve as a guide for others wishing to organize scientific hackathons.

Hackathons produce intangible outcomes on an individual level, such as awareness of resources, training, and connections. On a community level, hackathons increase appreciation of the benefits of interoperability, and its connection to standards. This includes appreciation for emerging standards (NeXML, PhyloWS, CDAO) and undeveloped or under-supported standards (e.g. MIAPA, LSIDs).

Computer code produced from hackathon projects is open-source and publicly available by the end of the hackathon. The specific nature of these outcomes cannot be predicted reliably. However, the following likely outcomes indicate what we mean by "growing the network of interoperable resources":

- **Increased utilization of next-generation data formats** - Participants working on tree visualization tools and data resources (listed earlier) will increase their support for NeXML as format for phylogenetic data exchange. For end-users, this means increased interoperability of software and resources that exchange phylogeny data. Ultimately, users will be able to choose software for strictly scientific reasons, instead of limiting themselves to those compatible with their existing workflow.
- **Increased use of web services to import or export phylogenies** - For a tree visualization tool that uses NeXML it is a short step to implement a PhyloWS search interface to access trees directly from TreeBASE or other resources. If cutting-edge tree viewers provide access to several such resources, this will stimulate other projects to export phylogenies via PhyloWS to leverage their cutting-edge visualization capabilities. Ultimately, the end-user will not be limited to locally saved trees; users with special visualization needs will choose a preferred tool, rather than the one chosen by the data provider.
- **Scientific use-cases driving expanded vocabulary support** - Participating projects will present use-cases that drive improvements in language support for representing data and metadata, expanding the scope of artefacts such as CDAO and NeXML. For end-users, this means that interoperable resources will cover more of the kinds of information important for their research. In our discussions with stakeholders, it is clear that there is an urgent need for language to annotate methods and phenotypes.

REFERENCES

Lapp H, Bala S, Balhoff J, Bouck A, Goto N, Holder M, Holland R, Holloway A, Katayama T, Lewis P, et al. 2007. The 2006 NESCent Phyloinformatics Hackathon: A Field Report. Evolutionary Bioinformatics, 3:287-296.

Lapp H, Stoltzfus A, Vision T, Vos R. 2009. Evolutionary Data Leaping to Web 3.0: Some Highlights From NESCent's Third Hackathon. ASN/SSB/SSE meeting.

Prosdocimi F, Chisham B, Pontelli E, Thompson J, Stoltzfus A. 2009. Initial Implementation of a Comparative Data Analysis Ontology. Evolutionary Bioinformatics:47-66.

Sidlauskas B, Ganapathy G, Hazkani-Covo E, Jenkins K, Lapp H, McCall L, Price S, Scherle R, Spaeth P, Kidd D. 2010. Linking Big: The Continuing Promise of Evolutionary Synthesis. Evolution, 64:871-880.

Vos R, Lapp H, Piel W, Tannen V. 2010. TreeBASE2: Rise of the Machines. Nature Precedings doi:10.1038/npre.2010.4600.1.

SUMMARY

Name	Rutger Aldo Vos
Date of Birth	5 October 1975
Nationality	Dutch
Address	22 Projection East, Merchants Place, Reading, RG11EG, UK
E-mail	R.A.Vos@reading.ac.uk
Website	http://rutgervos.blogspot.com
Phone	+44-7540-986655
Education	MSc., University of Amsterdam, 2000
	PhD., Simon Fraser University, 2006

RECENT FUNDING SOURCES

Funding for my research comes from a disparate number of sources. In the last years, these included:

- Marie Curie research fellowship 2009-2011
- NESCent working group funding from 2007-2009
- Google Summer of Code grants to hire student developers, 2007-present
- CIPRES Postdoctoral fellowship, 2006-2009
- President's Research Stipend, 2005
- SSB Systematic Biology Graduate Research Award, 2003
- SFU Travel Award, 2003
- SFU Graduate Fellowships, 2002, 2003

EMPLOYMENT HISTORY

Marie Curie Research Fellow

Starting in November 2009 to the present, I have been working at the University of Reading with professor Mark Pagel as a Marie Curie research fellow working on comparative genomics of Primates.

Consultant

Starting from Fall 2008 to the present, I have been working on the TreeBASE project as an outside consultant for the University of Pennsylvania.

Visiting scientist

From January to August 2007 I worked with professor Wayne Maddison as a visiting scientist at the Berlin Institute for Advanced Study (Wissenschaftskolleg zu Berlin).

Research assistant, postdoctoral fellow

Starting from January 2004 to the present, I have been working at the University of British Columbia, as an international collaborator on the CyberInfrastructure for Phylogenetic Research, contributing to architecture and database design.

Research assistant

From January 2001 through August 2006 I have worked intermittently as a research assistant with professor Arne Mooers at Simon Fraser University on a number of projects involving phylogenetic inference.

Educational software developer

From June through December 2000 I worked at BioMedia, in web development and software design for life sciences education.

Educational software developer

From June through September 1999 I worked at the Amsterdam Science and Technology Education Laboratory (AMSTEL institute) where I developed an educational CD-ROM on mitosis and meiosis visualized using 3D microscope photography.

Research assistant

From September 1998 through June 1999 I worked at the University of Amsterdam performing lab experiments (RAPD-PCR) tracking population divergence in laboratory strains of the Western Flower Thrips *Frankliniella occidentalis*.

SOFTWARE

I am an active contributor of open source bioinformatics software. In the past, I have developed educational software (for BioMedia, the Zoological Museum of Amsterdam, Simon Fraser University and the AMSTEL institute). The projects listed below – for which I am the primary author – are current and ongoing. Their combined worth is estimated by an independent open source code analysis group at $8.5 million. My programming skills include Perl, Java, C, web standards (XML/XSLT/XSD/HTML/CSS), ontologies (OWL, protege), databases (MySQL, PostgreSQL, dbxml), collaborative development best practices (svn, test-driven development) and design patterns.

TreeBASE

TreeBASE is a relational database of phylogenetic information hosted by the San Diego Supercomputing Center. In the Fall of 2008 I joined the development project to redesign TreeBASE from the ground up.

Bio::Phylo

Bio::Phylo is a collection of modules for phylogenetic analysis and biodiversity assessments. Primary author: Rutger Vos, with contributions from Aki Mimoto, Klaas Hartmann and Jason Caravas.

NeXML

NEXML is an emerging data exchange standard for phylogenetics and bioinformatics.

CIPRES

The CyberInfrastructure for Phylogenetic Research is an NSF-sponsored project to develop an architecture for data exchange and distributed analysis. I am the designer of the perl5 branch of the infrastructure.

COPE

An implementation of OMG-compliant CORBA ORB and related architecture, COPE was originally conceived by Bart Schuller, I am the current maintainer.

Exception::Class::TCF

A port of Java/C++ style exception handling.

Enrico Pontelli, Ph.D.

Professional Preparation

- University of Udine (Italy), Computer Science, Laurea, March 1991.
- University of Houston, Computer Science, Master of Science, August 1992.
- New Mexico State University, Computer Science, Ph.D., August 1997.

Appointments

- Department Head, Computer Science Dept., NMSU, 2/2009-present.
- Professor, Computer Science Dept., NMSU, 8/2005-present.
- Associate Professor, Computer Science Dept., NMSU, 8/02–07/05.
- Assistant Professor, Computer Science Dept., NMSU, 08/97–07/02.
- Lecturer, Computer Science Dept., University of Texas at El Paso, 01/96–06/96.
- Consultant, EniData and Esprit Project AXL, 1991.

Selected Publications

1. Dovier, A. Formisano, **E. Pontelli.** "Multi-valued Action Languages with Constraints in CLP(FD)," Theory and Practice of Logic Programming, (to appear), 2010.
2. Dal Palu, A., Dovier, A., Fogolari, F., **Pontelli, E.**, "Constraint-based Protein Fragment Assembly," In Bio-Logical, 2009.
3. Liu, L., **Pontelli, E.**, T. Son, and M. Truszczynski. "Logic programs with abstract constraint atoms", Artificial Intelligence Journal, (to appear), 2010.
4. Tu, P., **Pontelli, E.**, T. Son, S. To. "Applications of Parallel Processing Technologies in Heuristic Search Planning", Concurrency and Computation, 21(15):1928-1960, 2009.
5. Chisham, **E. Pontelli**, F. Prosdocimi, A. Stoltzfus, J. Thompson. "Initial Implementation of a Comparative Data Analysis Ontology", Evolutionary Bioinformatics, 5:47-66, 2009.
6. Dal Palu, A. Dovier, **E. Pontelli**. "A Constraint Solver for Discrete Lattices, its Parallelization, and Application to Protein Structure Prediction", Software Practice & Experience, 37(13):1405-1449, 2007.
7. Dal Palu, A., Dovier, A., and Pontelli, E., "Computing Approximate Solutions of the Protein Structure Determination Problem using Global Constraints on Discrete Crystal Lattices." Int. Journal of Data Mining and Bioinformatics, 4(1):1-20, 2010.
8. Son, T., and **Pontelli, E.**. "Planning with Preferences in Logic Programming", Theory and Practice of Logic Programming, 6(5):559-608, 2006.
9. Son, T., **Pontelli, E.**, C. Sakama. "Logic Programming for Multi-agent Planning with Negotiation", International Conference on Logic Programming, Springer, 2009.
10. **Pontelli, E.**, H. Le, T.C. Son. ``An Investigation in Parallel Execution of Answer Set Programs on Distributed Memory Platforms: Task Sharing, and Dynamic Scheduling'', Computer Languages, Systems, and Structures, 36:158-202, 2010.

Synergistic Activities

- I am the director of the Young Women in Computing program, funded by a NSF BPC grant; since 2006, the program has engaged cohorts of high school women in summer programs and academic-year activities. The work is in collaboration with high schools from the Las Cruces and the Gadsden School Districts. I am also the Director of two summer outreach programs, one involving high school students in 2-year exposure to bioinformatics, and one providing training in computer science and mathematics to students from tribal colleges and rural community colleges.
- I am serving as director of the Knowledge representation, Logic, and Advanced Programming laboratory; the lab conducts research in the field of logic programming, knowledge representation and reasoning, autonomous agents, and parallel processing. The lab is also dedicated to sustain participation of Hispanic students into research and education.
- I am serving as Director of the CREST Center for Research Excellence in Bioinformatics and Computational Biology. The Center, funded by a NSF CREST grant, supports research and educational activities in the field of bioinformatics, and provides a variety of outreach activities (especially for high-school and college students).

- I have published over 140 peer-reviewed publications in the areas of logic and constraint programming, Internet computing, assistive technologies, parallel processing, and bioinformatics.
- I have created and directed for the first three years the Doctoral Consortium program associated to the International Conference on Logic Programming.
- I have chaired/co-chaired Doctoral Consortia associated to the ASSETS conference (Computers and Accessibility) and the ICCHP conference (Computers and Handicaps).
- I have organized and directed the 3rd International Summer School on Computational Logic, funded by CRAW/CDC, that took place in Las Cruces in July 2008. 33 graduate students attended the school.
- I have served as Program Chair of the ACM International Conference on Computers and Accessibility (ASSETS 2005); I have served as General Chair of the same conference for 2007.
- I have served as Program Chair of the 2008 International Conference on Logic Programming.
- I am the Program Chair of the 2010 Symposium on Declarative Aspects of Multicore Programming.
- I am the Editor-in-Chief of the quarterly newsletter of the Association for Logic Programming (ALP) and member elect of the ALP Executive Committee.
- I am member of the editorial board of the ACM Transactions in Accessible Computing journal.
- I have served as Guest Editor of the Journal of Functional & Logic Programming, the Journal of Behavior and Information Technology, and Theory and Practice of Logic Programming.
- I have developed a support program, called Pathways in Computer Science, which provides assistance to undergraduate CS students through their gateway CS courses. The program has enabled increased retention and graduation, and it has contributed to doubling the number of undergraduate students transitioning to the graduate program.

Awards and Honors

- 2008, First Prize, Non-deterministic Track, International Planning Competition
- 2006 Arts&Sciences Faculty Outstanding Achievement Award
- 2005 NMSU University Research Council Award for Creative Scholarship
- 2003 Best Paper Award, ACM International Conference on Universal Usability.
- 2002 D. Roush Award for Excellence in Teaching, NMSU.
- 1999 NSF Career Award

Selected Grants

- NSF, Minority Institution Infrastructure (co-PI), 2002–2008.
- Department of Education, NIDDR program, 2000-2005.
- Department of Education, Graduate Assistants in Areas of National Need (co-PI), 2003–2006.
- NSF, Research to Aid Persons with Disabilities (PI), 2008–2011.
- NSF, Career Award (PI), 1999–2004.

Recent Collaborators

Faculty: S. Tran (NMSU), A. Stoltzfus (NIST), J. Thompson (U. Strasbourg), A.I. Karshmer (U. San Francisco), A. Dovier (U. of Udine), A. Dal Palu (U. Parma), C. Baral (Arizona State U.), M. Truszczynski (U. Kentucky), P. Bonatti (U. Napoli), C. Sakama (Wakayama U.), A. Formisano (U. Perugia).
Doctoral Students: O. El-Khatib (2007), I. Elkabani (2006), H. Le Viet (2007), E. Saad (2005), Y. Pan (2007), K. Villaverde (2002), C. Liu (2008), I. Abu Doush (2009), A. Arredondo, A. Alqaddoumi, B. Chisham.
MS Students: 20 MS students graduated since 2006
Graduate Advisors: G. Gupta (U.T. Dallas), L. Slothouber (U. Houston)

Recent Courses:

Parallel Programming, Formal Language & Automata, Programming Languages Structure, Semantics of Programming Languages, Semantic Web, Constraint Programming, Computational Logic.

Arlin Stoltzfus, Ph.D.
Research Biologist, National Institute of Standards and Technology
Adjunct Professor, UMCP; Fellow, IBBR
240 314 6208 (arlin@umd.edu)

Professional preparation

INSTITUTION AND LOCATION	DEGREE	YEAR(s)	FIELD OF STUDY
Grinnell College, Iowa, USA	B.A., *c.laude*	1985	English
University of Iowa, Iowa, USA	Ph.D.	1991	Biology
Dalhousie Univ., Halifax, Canada	Post-Doctoral	1999	Molecular Evolution

Appointments

2010 institutional realignment (CARB is now IBBR; UMD affiliation is now UMCP)
2006-present Associate Professor (UMBI)
1999-present CARB Fellow, Research Biologist (NIST), Adjunct Asst Professor (UMBI)
1991-1999 Post-doctoral fellow, Dalhousie University, Halifax, Nova Scotia, Canada
1985-1991 Research Assistant, Department of Biology, University of Iowa, Iowa City, Iowa

Selected publications (in chronological order)

1. **Stoltzfus, A**., Spencer, D.F., Zuker, M., Logsdon, J.M., Jr., and Doolittle, W.F. 1994. Testing the exon theory of genes: the evidence from protein structure. *Science* **265:** 202-207.
2. **Stoltzfus, A**. 1999. On the possibility of constructive neutral evolution. *J Mol Evol* **49:** 169-181.
3. Yampolsky, L.Y., and **Stoltzfus, A**. 2001. Bias in the introduction of variation as an orienting factor in evolution. *Evol Dev* **3:** 73-83.
4. Qiu, W.G., Schisler, N., and **Stoltzfus, A**. 2004. The evolutionary gain of spliceosomal introns: sequence and phase preferences. *Mol Biol Evol* **21:** 1252-1263.
5. Yampolsky, L.Y., and **Stoltzfus, A**. 2005. The Exchangeability of Amino Acids in Proteins. *Genetics* **170:** 1459-1472.
6. Gopalan, V., Qiu, W.G., Chen, M.Z., and **Stoltzfus, A**. 2006. Nexplorer: Phylogeny-based exploration of sequence family data. *Bioinformatics, 22:120-121*.
7. T. Hladish, V. Gopalan, C. L. Liang, W. G. Qiu, P. J. Yang, and **A. Stoltzfus**. 2007. Bio::NEXUS: a Perl API for the NEXUS format for comparative biological data. *BMC Bioinformatics* **8**:191.
8. Lapp, H., Lapp, H., Bala, S., Balhoff, J.P., Bouck, A., Goto, N., Holder, M., Holland, R., Holloway, A., Katayama, T., Lewis, P. O., Mackey, A., Osborne, B. I., Piel, W. H., Kosakovsky Pond, S. L., Poon, A., Qiu, W.G., Stajich, J. E.. **Stoltzfus, A**., Thierer, T., Vilella, A.J., Vos, R., Zmasek, C.M., Zwickl, D., Vision, T.J., The 2006 NESCent Phyloinformatics Hackathon: A field report. *Evolutionary Bioinformatics*, 2007. **3**: p. 357-366.
9. **Stoltzfus, A** and Yampolsky, Y. 2009. Climbing Mount Probable: Mutation as a cause of non-randomness in evolution. *J. Heredity* **100** (5): 637-47.
10. F. Prosdocimi, B. Chisham, E. Pontelli, J. D. Thompson, A. Stoltzfus, 2009. Initial Implementation of a Comparative Data Analysis Ontology (CDAO). *Evolutionary Bioinformatics* 5, 47-66.

C. Synergistic Activities

Developer and project leader, Bio::NEXUS, an open-source Perl API for the NEXUS file format.
Developer and project leader, Comparative Data Analysis Ontology (CDAO, www.evolutionaryontology.org)
Developer and maintainer of Nexplorer, a publicly available web-based phylogenetic browser and editor for comparative data (www.molevol.org/nexplorer).

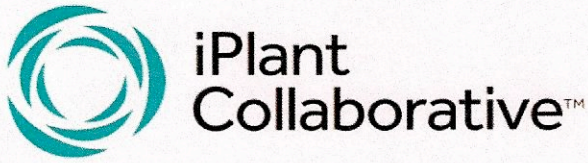
Co-leader of the NESCent Evolutionary Informatics working group, 2006 to 2009, focusing on improving interoperability through standards and technology (evoinfo.nescent.org).
Served as co-organizer of NESCent hackathons in 2006 and 2009, empowering early-career researchers with the skills and connections to improve interoperability.

D. Collaborations and Other Affiliations

Collaborators: Jim Balhoff (NESCent), Amy Bouck (Univ North Carolina), Brandon Chisham (New Mexico State Univ), Jonathon Eisen (UC Davis Genome Center, UC Davis, CA), Joe Felsenstein (University of Washington, Seattle,WA), Gupta Gopal (University of Texas, Dallas, TX), Vivek Gopalan (Bioinformatics, Lockheed Martin), Naohisa Goto (Osaka University). Tom Hladish (Univ Texas, Austin), Mark Holder (Florida State University), Richard Holland (European Bioinformatics Institute), Alisha Holloway (UC Davis), John Huelsenbeck (University of California, San Diego CA), Toshiaki Katayama (University of Tokyo), Sergei Kosakovsky Pond (UC San Diego), Sudhir Kumar (Center for Evolutionary Functional Genomics, The Biodesign Institute, Arizona State University, Tempe, AZ), Hilmar Lapp (NESCent), Paul O. Lewis (University of Connecticutt, Storrs, CT), Chengzhi Liang (Cold Spring Harbor Laboratory), Aaron Mackey (Glaxo (GSK)), David Maddison (University of Arizona, Tucson, AZ), Wayne Maddison (University of British Columbia, Vancouver, BC (Canada)), Holder Mark (Dept. Ecology and Evo. Biol., Univ. Kansas, Lawrence, KS), Eric Nawrocki (Howard Hughes Institute at Vanelia Farms), Brian Osborne (NESCent), Bill Piel (Yale University), Enrico Pontelli (New Mexico State Univ), Art Poon (UC San Diego), Francisco Prosdocimi (Univ. Strasbourg, France), Wei-Gang Qiu (Hunter College CUNY), Nick Schisler, (Furman University, Greenville, SC), Jason Stajich (UC Berkeley), David L. Swofford (Florida State University, Tallahassee, FL), Tobias Thierer (Biomatters Ltd.), Julie Thompson (Univ. Strasbourg, France), Albert Vilella (European Bioinformatics Institute), Todd Vision (UNC, NESCent), Rutger Vos (University of British Columbia, Vancouver, BC (Canada)), Xuhua Xia (University of Ottawa, Ottawa, ON (Canada)), Lev Yampolsky (East Tennessee State Univ), Christian Zmasek (Burnham Institute for Medical Research, La Jolla, CA), Derrick Zwickl (Univ of Kansas).

Graduate and Post-Doctoral Advisors: W. Ford Doolittle (Dalhousie University, semi-retired), Roger Milkman (Woods Hole Marine Biological Laboratory, retired).

Thesis Advisor and Post-graduate-Scholar Sponsor: Weigang Qiu (Hunter College, CUNY), Chengzhi Liang (CSHL), Danny DeKee (no scientific affiliation), Vivek Gopalan (Lockheed Martin NIAID Bioinformatics Support)



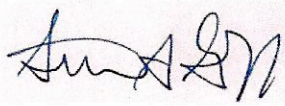
November 29, 2010

Dear Drs. Vos, Stoltzfus, and Pontelli,

I am writing to confirm institutional support from the iPlant Collaborative/University of Arizona for your National Evolutionary Synthesis Center (NESCent) working group proposal to organize hackathons for evolutionary data interoperability. We see this as an excellent opportunity for testing and expanding iPlant's Application Programming Interfaces (APIs) as well as for iPlant to engage with the larger community of evolutionary science resource developers.

For this project, iPlant commits to co-sponsoring one hackathon at the University of Arizona (up to $15,000) and also commits funds to send two iPlant representatives to the other hackathons that will be organized as part of the proposal.

Best regards,



Stephen A. Goff

PI and Project Director
The iPlant Collaborative

Adjunct Professor
University of Arizona, Plant Sciences Dept.

Thomas J. Keating Bioresearch Building, 1657 East Helen Street, Tucson, Arizona 85721

Biodiversity Synthesis Center
1400 S. Lakeshore Drive
Chicago, IL 60605
mwestneat@fieldmuseum.org

November 30, 2010

Dr. Arlin Stoltzfus
University of Maryland

Dear Arlin:

This letter expresses my interest in collaborating with you on the proposed NESCent working group entitled HIP: Hackathons, Interop, Phylogenies. I am interested for a number of reasons: interoperability is a key to future success of EOL and other biodiversity projects and phylogenetics programs that I am involved with, phylogenies are the key to being able to synthesize evolutionary information, and hackathons are one of the best ways to actually get things done in this area. Also, I have a particular interest in tree visualization and end user tools that depend on these elements.

I am able to commit to the project in several ways. First, I am able to serve on the working group itself to help with organization and hackathon design. Second, I can commit to paying for attendance at one or more hackathons for programmers working with our group, at least two person-trips, possibly more. I am able to cosponsor a hackathon and commit up to $10,000 for one, with a particular but not exclusive interest in hacking on visualization or interoperability with EOL. Finally, we can offer our meeting venue and facilities in Chicago free of charge if that would be useful.

I look forward to working with you and the group.

Sincerely,

M. Westneat

Mark W. Westneat
Curator of Zoology, Field Museum of Natural History
Director, Biodiversity Synthesis Center of the Encyclopedia of Life

UF UNIVERSITY of FLORIDA

Florida Museum of Natural History
Herbarium & Informatics, Department of Natural History

354 Dickinson Hall
PO Box 117800
Gainesville, FL 32611-7800
Tel. 352-273-1979
Fax 352-846-1861
ncellinese@flmnh.ufl.edu

Dr. Arlin Stoltzfus
National Institute of Standards and Technology
9600 Gudelsky Drive
Rockville, MD

Dr. Rutger Vos
School of Biological Sciences
University of Reading
Reading, RG6 6BX
United Kingdom

November, 27th 2010

Dear Arlin, Enrico and Rutger,

I am very excited to know that you are proposing a NESCent Working Group that includes a series of hackathons with the goal of promoting interoperability through a deployment of the foundational technologies provided by CDAO, PhyloWS and NeXML as key resources.

As you know, I am very interested in data integration and interoperability, and as part of my own development with TOLKIN (www.tolkin.org), I would be able to offer a test-bed that serves as reference implementation for your development. Reference implementations, especially when viewed as an essential component of standards development, help to identify problems early on, and optimize usability for the phyloinformatics community.

Given the diverse data types and associated metadata stored in TOLKIN, it serves as an ideal test-bed for reference implementation of standards and specifications developed through the Working Group. In addition, current TOLKIN workflow development adds considerable significance for testing the applicability and benefits of the foundational technologies.

I am willing to sponsor one TOLKIN programmer to participate to one Hackathon during the first year, and one postdoc from my lab to participate in coding activities during years 2 and 3, that is, a total of 3 person-trips worth of support.

I wish you the best for the success of your proposal.

Sincerely,

Nico Cellinese
Assistant Curator, Herbarium & Informatics

Assistant Professor, Department of Biology

www.phylodata.org

December 3, 2010

Dr. Rutger Vos
Marie Curie Research Fellow
School of Biological Sciences
University of Reading
United Kingdom

Dear Rutger:

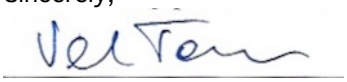
On behalf of the pPOD (Processing Phylodata) project, I am very pleased to support your group's efforts in proposing a working group to make evolutionary data accessible, searchable, and combinable, thereby promoting synthesis at a practical, technology-oriented level. As you point out, the working group will organize "hackathons" -at NESCent and elsewhere. I believe that our project can highly benefit from participation in these activities.

The pPOD project aims to develop and provide a reference implementation for a core set of technologies that will enable interoperability, i.e., both data and tool integration, for the database and workflow infrastructure used by AToL projects, following a three-pronged approach: (1) develop an extensible core data model for phylogenetic data; (2) develop schema mappings for peer-to-peer data integration and exchange, where a project can join existing integration groups by providing mappings between the schema of their data and the core data model or one of its extensions; (3) develop a scientific workflow system (lab notebook) that will allow research groups to put together the data integration components with the local database access components and with the analysis tools. Thus, pPOD, although focused on AToL efforts, shares many interoperability goals with your vision.

Therefore, I very much hope that your proposal is funded so that pPOD members can participate in the envisioned hackathons, in particular in the first one, focused on key exchanges of richly annotated data, involving TreeBASE (in which I am involved), Dryad, the Tree of Life web project, PhenoScape, MorphBank, MorphoBank, TimeTree and PhylomeDB, but especially in the second one which will focus on data integration and exploration environments, involving BioSync, TOLKIN, PhyLoTa, our pPOD and the iPlant Discovery Environment (in which I am also involved). We will support the travel of our members from pPOD funds.

I wish you the best with your application.

Sincerely,



Val Tannen
Professor of Computer and Information Science
University of Pennsylvania