PROJECT SUMMARY

Overview:

Because species phylogenies (trees) are so widely useful in the life sciences, there has been a major worldwide effort to determine trees for various groups and assemble them into a grand "Tree of Life" (ToL) - though it is better to think of expert ToL knowledge as a forest of available trees than as a single authoritative tree. While experts continue improving the ToL, our focus is on getting this knowledge into the hands of scientists, educators, and the public, so that they can get online species trees as easily as they get online driving directions. The key to delivering ToL knowledge automatically is to understand the most frequent pattern of use: begin with a list of species (or taxa), identify a larger tree that includes them, and extract the needed parts. By studying how scientists do this, we find that it involves 4 main operations of identifying a source tree, reconciling names, extracting the subtree, and (often) scaling it using fossil calibrations. In preliminary work, we staged a software hackathon, challenging early-career researchers to prototype an open community infrastructure to support workflows for on-the-fly delivery of ToL knowledge. Participants responded with a system based on integration of web services. The distributed design allows experts (in phylogeny, taxonomy, calibration) to maintain independent resources yet be part of an integrated solution. To build this community infrastrucure, we will (1) design a robust architecture for on-the-fly delivery of expert ToL knowledge in a manner that is computable, convenient, and credible; (2) implement a production system that exemplifies this design, using workflow composition by state-of-the-art automated planning algorithms; and (3) cultivate its use in a larger phyloinformatics community. Project staff, together with our partners in tree aggregation (OpenTree) and name resolution (Global Names), will supply expertise. Design and evaluation will be guided by practical use-cases of value to researchers, educators and the public.

Intellectual Merit :

Phylogenetic knowledge is disseminated today via a very large number of idiosyncratic pathways, most of which are not easily traceable. This will change when the system that we develop becomes the largest conduit for computable ToL knowledge, allowing a feedback loop on quality, coverage, interoperability, and other issues. Our plan of work will expose the frequency of various types of naming errors in sources; will reveal practical measures of coverage by name-banks and ToL source trees; will provide a practical context to compare name-alignment strategies; will expose and resolve metadata representation issues; and will provide a foundation for strategies to deal with challenges in resolving conflicts and assessing quality.

Broader Impacts :

This project will benefit a significant segment of the research community supported by the NSF BIO Directorate, as well as educators and the public, by providing a robust cyberinfrastructure that enhances the value of expert phylogenetic knowledge, including the knowledge currently embedded in phylogeny publications, and future knowledge that will emerge from ToL projects. The system we develop will reduce the burden for technical and non-technical users of extracted trees; greatly enlarge the pool of users; and expand the scope and scale of feasible projects, allowing larger, more integrated, and more automated projects. We expect automatic discovery of species trees for reconciliation to become a major use-case, and to increase the quality of genome annotations. Our client application combining automated tree discovery with high-value data sets will facilitate new evolutionary studies, and inspire others to integrate automatic tree discovery in useful tools. Our project to auto-generate interactive tree images for resources such as EOL and Wikipedia will expose millions of users each year to expert ToL knowledge. We will deploy client applications in high school and college classrooms, some with diverse populations. Finally, this project will leverage, integrate with, and energize the phyloinformatics community, and will exemplify open approaches to science and technology. Our 3rd-year hackathon will engage mostly early-career scientists in a community infrastructure project, providing valuable training experiences, exposure to new technologies, and networking experiences. Existing resource-providers (e.g., ToLWeb, TreeBASE, OToL, iPlant, EoL), as well as emerging ones, will have a new way to expose content and become part of a dynamic knowledge delivery system.

1458572

TABLE OF CONTENTS

For font size and page formatting specifications, see GPG section II.B.2.

	Total No. of Pages	Page No.* (Optional)*
Cover Sheet for Proposal to the National Science Foundation		
Project Summary (not to exceed 1 page)	1	
Table of Contents	1	
Project Description (Including Results from Prior NSF Support) (not to exceed 15 pages) **(Exceed only if allowed by a specific program announcement/solicitation or if approved in advance by the appropriate NSF Assistant Director or designee)**	15	
References Cited	7	
Biographical Sketches (Not to exceed 2 pages each)	2	
Budget (Plus up to 3 pages of budget justification)	5	
Current and Pending Support	1	
Facilities, Equipment and Other Resources	1	
Special Information/Supplementary Documents (Data Management Plan, Mentoring Plan and Other Supplementary Documents)	13	
Appendix (List below.) **(Include only if allowed by a specific program announcement/ solicitation or if approved in advance by the appropriate NSF Assistant Director or designee)**		
Appendix Items:		

*Proposers may select any numbering mechanism for the proposal. The entire proposal however, must be paginated. Complete both columns only if the proposal is numbered consecutively.

TABLE OF CONTENTS

For font size and page formatting specifications, see GPG section II.B.2.

	Total No. of Pages	Page No.* (Optional)*
Cover Sheet for Proposal to the National Science Foundation		
Project Summary (not to exceed 1 page)		
Table of Contents	1	
Project Description (Including Results from Prior NSF Support) (not to exceed 15 pages) **(Exceed only if allowed by a specific program announcement/solicitation or if approved in advance by the appropriate NSF Assistant Director or designee)**	0	
References Cited		
Biographical Sketches (Not to exceed 2 pages each)	2	
Budget (Plus up to 3 pages of budget justification)	6	
Current and Pending Support	3	
Facilities, Equipment and Other Resources	1	
Special Information/Supplementary Documents (Data Management Plan, Mentoring Plan and Other Supplementary Documents)	13	
Appendix (List below.) **(Include only if allowed by a specific program announcement/ solicitation or if approved in advance by the appropriate NSF Assistant Director or designee)**		
Appendix Items:		

*Proposers may select any numbering mechanism for the proposal. The entire proposal however, must be paginated. Complete both columns only if the proposal is numbered consecutively.

TABLE OF CONTENTS

For font size and page formatting specifications, see GPG section II.B.2.

	Total No. of Pages	Page No.* (Optional)*
Cover Sheet for Proposal to the National Science Foundation		
Project Summary (not to exceed 1 page)		
Table of Contents	1	
Project Description (Including Results from Prior NSF Support) (not to exceed 15 pages) **(Exceed only if allowed by a specific program announcement/solicitation or if approved in advance by the appropriate NSF Assistant Director or designee)**	0	
References Cited		
Biographical Sketches (Not to exceed 2 pages each)	2	
Budget (Plus up to 3 pages of budget justification)	6	
Current and Pending Support	1	
Facilities, Equipment and Other Resources	1	
Special Information/Supplementary Documents (Data Management Plan, Mentoring Plan and Other Supplementary Documents)	13	
Appendix (List below.) **(Include only if allowed by a specific program announcement/ solicitation or if approved in advance by the appropriate NSF Assistant Director or designee)**		
Appendix Items:		

*Proposers may select any numbering mechanism for the proposal. The entire proposal however, must be paginated. Complete both columns only if the proposal is numbered consecutively.

A. Introduction

A.1 Project Motivation

Phylogenetic trees are useful in all areas of biology, both to organize knowledge by guiding classification (taxonomy), and for process-based models that allow scientists to make robust inferences from comparisons of evolved entities (genes, species, etc). The transformative potential of assembling a *Tree of Life (**ToL**)*, a phylogeny covering 10^7 or more species [1], was articulated in an NSF workshop report [2]. The first draft of a grand phylogenetic synthesis—a single **synthetic tree** with 2.5 X 10^6 species (tree.opentreeoflife.org)—recently emerged from the *Open Tree of Life (**OpenTree**)* project (note: in this proposal, the first use of key terms or acronyms is in bold).

Though useful, neither this tree, nor any other single tree, will be the sole authority on phylogenetic knowledge. *When we refer to "**the ToL**" or "**ToL knowledge**" here, we do not mean any single tree, but the dispersed set of available trees that represent the current state of ToL knowledge*.

While experts continue expanding the ToL, addressing gaps and conflicts, our focus is on dissemination— putting ToL knowledge in the hands of researchers, educators, and the public. In our vision for the future, users get species trees online as easily as they currently get driving directions. To achieve this, we will build an open web-based system that enables flexible on-the-fly delivery of phylogenetic knowledge. Our approach is grounded in (1) our years of experience organizing and working with a "phyloinformatics" community to address interoperability challenges (see A.4.2; [3-8]), (2) recent analyses of the state of annotation, archiving, and re-use of phylogenies by ourselves and others [9-11], (3) a prototype ToL delivery system (www.phylotastic.org) from a software development hackathon we staged [12], and (4) our work as scientific users and creators of phylogenetic methods and phylogenetic knowledge [13-33].

The premise of disseminating knowledge is that it will be re-used. How do trees get re-used? On a per-tree basis, re-use is rare— most trees are inferred *de novo* for a specific study, stored on someone's hard drive, and not used again [11]. Yet, large species trees are re-used in ways that other trees are not. In a sample of 40 phylogeny articles, we found 6 cases in which scientists obtained a desired tree by extraction from a larger species tree [11]. These studies implicate diverse uses: functional analyses of leaf traits by Walls [34], or lactation traits by Riek [35]; phylogenetic diversity of forest patches [36]; analyzing niche-diversity correlations [37], spatial distribution of wood traits [38], and spatial patterns of diversity [39]. The implicated trees include those covering 4,510 extant mammals [40], 55,473 angiosperm species [41], and 1,566 angiosperm taxa [42].

We refer to this documented, quantitatively important pattern of re-use as **subtree extraction** from ToL **source trees**. The frequency of subtree extraction in [11] implicates ~10^3 published studies per year. In addition, we note that the *Encyclopedia of Life (**EOL**)* [43] currently has ~300,000 web pages for higher taxa, each of which might benefit from a species tree. An even larger potential market resides in "tree reconciliation", comparing a gene tree with a species tree to determine which branchings are duplications, and which are speciations [44]. Because this distinction is necessary to diagnose orthology of sequences, a predictor of shared function, tree reconciliation is highly useful in genome annotation— a major use-case in bioinformatics.

Yet, this mode of re-use currently presents technical barriers requiring considerable expertise and effort to overcome. The vast majority of users simply couldn't handle a tree with more than a thousand species, even if they knew how to find and obtain the right tree— a challenge, as only 4% of trees are archived [11]. Tree files generally lack machine-processable metadata on sources and methods, crucial for quality evaluation; common tree formats do not support such metadata (Fig. 2 of [11]); a new format, **NeXML** [7], allows semantic annotations, but is not yet widely used; a proposed annotation standard, **MIAPA** (Minimum Information About a Phylogenetic Analysis) [45], is not fully developed (see [11]). The largest and most valuable species trees often provide a topology without branch lengths, yet users often need branch lengths in downstream analysis steps: to close this gap, proficient users may create crude branch lengths with specialized software (e.g., [46, 47]). Even matching a list of species names with a source tree is problematic, given the proliferation of aberrant

names [48]. For instance, in Riek [35] noted above, the sets of 40 species names in an extracted subtree, in the main data table, and in the original source tree were all different— the author somehow used innate (human) pattern-matching and taxonomic expertise to align names. Though non-matches may reflect genuine differences in taxonomies, the proliferation of names in scientific papers and data is mainly due to spelling errors and lexical variants [49].

Thus, whereas subtree extraction is conceptually simple, real-world uses are surrounded by complications, currently requiring a combination of expert skills, hands-on attention, and specialized software. The typical workflow begins with upstream steps that establish the user's focus on a particular set of species, and proceeds with: (1) discovery and acquisition of an appropriate ToL source tree; (2) negotiating an optimal alignment with the set of query names; (3) subtree extraction and optional grafting; and (4) scaling the extracted subtree.

Tools that facilitate and automate this workflow would be highly beneficial. A tool called "Phylomatic" [46] provides direct access to ToL source trees for plants, extraction and grafting operations, and scaling operations. Remarkably, 4 out of 6 cases of subtree extraction in [11] used Phylomatic to perform grafting and extraction on the framework tree provided by the Angiosperm Phylogeny Group (**APG**). The *Taxonomic Name Resolution Service (**TNRS**)* [49] from iPlant or **GNR** (see B.4.2) breaks name-strings into semantic elements, accesses authoritative taxonomy sources to validate names, corrects spelling errors, identifies synonyms and homonyms, and provides other information such as taxonomic derivations (e.g., Mammalia: Carnivora: Felidae: Felinae: *Felis silvestris*).

Inspired by these examples of lowering barriers, the **HIP** *(Hackathons, Interoperability, Phylogenies)* working group (PIs A. Stoltzfus, E. Pontelli and R. Vos) of **NESCent** (the National Evolutionary Synthesis Center) proposed the idea of a flexible system for on-the-fly delivery of custom trees that would support many kinds of tree re-use, and be open for both users and data providers. We invited a group of largely early-career scientists to prototype such a system using the "hackathon" model, and they responded with a system for composing workflows from web services: by implementing an agreed-upon **API** (applications programming interface), any resource-provider may join the online system, and any developer may write a web client to harvest data [12].

The aim of this proposal is to design and implement this kind of open system in such a way that it becomes a sustainable web of resources, supported by the community without centralization.

We expect that this system will have a major long-term impact on ToL coverage and quality, but our focus here is on **dissemination**. With the exception of expanding the capacity to scale trees by adding annotations to calibrated trees (see B.4.3), we will not generate, annotate, evaluate, edit, or police expert knowledge. Instead, we will build an architecture that is neutral with respect to content. Quality assurance, in this context, means reliably delivering data that traces to expert sources: it does not mean choosing which expert is right. The scope and quality of outputs will depend on the state of inputs. Existing name services (GNR, iPlant, ITIS, uBio, etc. [49]) make expert decisions about which curated name-banks to use, how often to update content, and how to define default behavior. With respect to species trees, 4 projects currently expose content via web services: TreeBASE [50] (an archive of about 10,000 gene and species trees, active for 20 years), ToLWeb [51] (a manually curated hierarchy of ~10^5 species), the Phylomatic web service (supplying the APG tree and derivatives) and OpenTree (~4000 source trees, plus synthetic tree of 2.5 X 10^6 species).

From our perspective, this is all useful knowledge to be disseminated. We anticipate that it will increase rapidly in the next decade. This increase will not come predominantly from mining old studies. The majority of old trees (and underlying alignments and character matrices) are not archived and are not accessible without extraordinary effort [9-11]; yet there was a distinct surge in archiving in 2011 (for reasons discussed in [10]). In the case of trees, the emerging world of networked data will be based on newer results that authors have submitted to archives, as part of a larger trend in science toward reproducibility, openness, and data sharing.

How will the proposed system work, in practice, given the state of resources now and in the near future? Suppose the user asks for a tree for species A to Z, and a client program composes and

executes a workflow that returns an extracted tree with 22 of the 26 species, where the TNRS has interpreted the user's request for species C to refer to species C' due to a misspelling. Though casual users will ignore it, the proposed system will generate a provenance report, including a TNRS report with the taxonomic authority and a confidence score for each matched name [49]. Users who have a reason to prefer one taxonomic authority over another may impose a preference [49].

Whereas names can be validated, and users may set the quality of allowable matches (via a threshold confidence score), there is no way of knowing if an extracted tree is the true tree, and there is no generic quality metric for trees. Experts judge trees by whatever measures of consistency or reliability have been computed; by concordance with (possibly flawed) expectations; and by perceptions of the quality of methods. Because the problem of assessing quality has not been solved by the community, our system will not impose a solution. When multiple trees satisfy a query, the obvious choices are (1) to return all trees, (2) to return the first tree arbitrarily, or (3) to combine the trees using a standard method such as majority-rule consensus.

Yet, such a situation may be rare, due to the multitude of other factors at play, and the fact that clients are free to do what they want. Users with a list of target species typically will want the tree with the best **coverage** (i.e., the most species matched). Clients may have a built-in preference for a single large source tree, e.g., the APG tree for users analyzing plant data. Clients may implement preferences based on metadata, enabled by our use of NeXML (which allows trees, alignments and other character matrices, and annotations) and by the fact that OpenTree and TreeBASE expose metadata (as of this writing, OpenTree has annotations for ~1/10th of its source trees). Thus, a client may choose trees based on recency of publication, or use of Bayesian methods, or availability of source data (currently, only in TreeBASE). An interactive client could show multiple trees, each with metadata, allowing the user to choose.

Clients may differ, but the core architecture will not favor particular sources, adjudicate conflicts, or prevent users from doing valid but foolish things with data. We focus instead on giving users the tools to do smart things easily. Expert knowledge *already* includes errors and conflicts, and users *already* do stupid things, and this has not caused science to grind to a halt, because healthy scientific communities have effective feedback loops. A system to disseminate knowledge becomes part of those feedback loops; a flexible automated system exposes errors and conflicts rapidly, traceably, and systematically, providing opportunities to evaluate and implement remedies. For instance, a fungal systematist might design a client to retrieve all source trees with at least 10 fungal species, and use the results to analyze conflicts between trees, identify gaps, and rate each source tree based on concordance with a trusted set of trees. The fungal systematist can't force data providers to delete poorly rated trees, but because trees will have universal identifiers, the fungal systematist may set up a rating service that clients can use to make choices. Scientists developing a MIAPA standard might assign each tree a rating of silver, gold or platinum based on the level of metadata coverage. A third party might rate services based on their response time and availability.

To summarize, the above arguments establish the two main premises of this proposal. *First,* in cases from genomics to functional biology to community ecology to education, the need for a custom species tree is now satisfied, or could be satisfied, by extraction from a ToL source tree, without tedious and demanding methods of *de novo* phylogeny inference: expert ToL knowledge is an increasingly valuable product that users want. *Second,* getting ToL knowledge into the hands of users requires overcoming interoperability barriers with technology to reduce the needed effort and expertise. We have studied these barriers at length [11], recognized and evaluated technologies that lower these barriers [46, 49], and engaged the community to devise a solution [12].

A.2 Specific aims

Our mission can be accomplished by (1) recognizing the key challenges of dissemination and engaging the community in prototyping a solution; (2) bringing this solution to fruition with a carefully designed reference implementation with the capacity of a production service; (3) cultivating the solution as a community resource. Having accomplished 1, we focus here on 2 and 3. Working with OpenTree, GNR (see B.4.2), other strategic partners, and a broader community, we will:

A.2.1. Design a sustainable architecture for ToL delivery. Working with partners and other domain experts, we will iteratively design and evaluate an open web-services API for components and controllers that (1) supports a range set of use-cases, (2) allows for name-resolution, tree discovery, subtree extraction, and scaling of trees, (3) documents the tree discovery process, and (4) provides for a sustainable distributed open architecture.

A.2.2. Implement a production service and reference implementation. We will implement a production service that (1) is fronted by a client with a graphical interface as well as a programmable web-services interface, (2) includes name-resolution, source-tree-discovery, subtree extraction and scaling, (3) meets anticipated demands for speed and capacity, and (4) is a well documented, installable open-source package that serves as a reference implementation, exemplifying the APIs for others to use as a pattern or a foundation.

A.2.3. Cultivate a community resource. We will cultivate this system as a community resource by (1) involving partners, domain experts, and a broader phyloinformatics community in the design process; (2) partnering with other projects to involve them as service-providers or consumers; (3) developing innovative clients demonstrating quantitatively important use-cases; (4) disseminating knowledge of the system via manuals, screencasts, and conference presentations; (5) staging a hackathon for participants to add services or develop clients.

A.3 Intellectual Merit

Phylogenetic knowledge is disseminated today via a very large number of idiosyncratic pathways, most of which are not easily traceable [11]. This will change when the system that we develop becomes the largest conduit for computable ToL knowledge, allowing a feedback loop on quality, coverage, interoperability, and other issues. Our plan of work will expose the frequency of various types of naming errors in sources; will reveal practical measures of coverage by name-banks and ToL source trees; will provide a practical context to compare name-alignment strategies; will expose and resolve metadata representation issues; and will provide a foundation for strategies to deal with unresolved problems in rooting, conflicts, the validity of substitution operations, and practical methods to treat uncertainty. While the problem of automatic service composition has been explored in the computer science literature (e.g., [52, 53]), fully automated solutions that include discovery (from semantically-rich repositories), automated adaptation, and online monitoring will be novel.

A.4 Broader Impacts

This project will benefit a significant segment of the research community supported by the NSF BIO Directorate, as well as educators and the public, by providing a robust cyberinfrastructure that enhances the value of expert phylogenetic knowledge, including the knowledge currently embedded in phylogeny publications, and future knowledge that will emerge from ToL projects.

The current scale of actual use (by researchers) of custom species trees obtained by extraction is ~10^3 published studies per year, by our crude estimate. This project, if successful, will reduce the burden for these users; greatly enlarge the pool of users by making the process easier; and expand the scope and scale of feasible projects, allowing larger, more integrated, and more automated projects. We expect automatic discovery of custom species trees for reconciliation to become a major use-case, and to increase the quality of genome annotations. Our use-case combining automated tree discovery with high-value data sets will facilitate new evolutionary studies, and will serve as a model for developers to embed this capacity in web sites, workbenches, and other tools, introducing automatic tree discovery to many new users. Our project to auto-generate interactive tree images for resources such as EOL and Wikipedia will expose millions of users each year to up-to-date phylogenetic information. We will deploy client applications in high school and college classroom, some with diverse populations (e.g., southern New Mexico high schools), realizing pedagogical models (e.g., situated learning) that promote STEM engagement and learning.

This project will leverage, integrate with, and energize the phyloinformatics community, including its human and computational resources. Our 3rd-year hackathon will engage mostly early-career scientists in a community infrastructure project, providing valuable training experiences, exposure to

new technologies, networking, and exposure to best practices for scientific software development. We have demonstrated our ability to recruit women and minorities to hackathon events, reaching a level of 40% women and minorities at a recent event [54]. Existing resource-providers (e.g., ToLWeb, TreeBASE, OpenTree, iPlant, EOL) will have new ways to expose content and become part of a dynamic knowledge delivery system. Our approach will leverage, improve and energize emerging community standards (e.g., MIAPA, NeXML, and **CDAO**— the Comparative Data Analysis Ontology [5]), and will exemplify open approaches to science and technology.

A.5. Preliminary results

A.5.1. A proof-of-concept system, with demonstrations. The prototype ToL delivery system [12] is accessible for evaluation via a website (http://www.phylotastic.org), a public code repository, and a server image, exemplifying our approach to dissemination and open development. Key accomplishments are illustrated by live demonstration software and screencasts, including: (1) a meta-TNRS that communicates with 3 core TNRS systems; (2) a subtree extractor implemented via Hadoop; (3) a tree database implemented as a triple-store; (4) a scaling service (DateLife) that adds divergence dates; (5) a reconciliation demo (Reconciliotastic); (6) a phylogeography tool (PhyloGeotastic); and (7) a draft ontology for phylogeny annotation, and its demonstration on a set of 10 high-value ToL source trees.

A.5.2. Interoperability work with the phyloinformatics community. Pontelli and Stoltzfus have served (with Hilmar Lapp, Rutger Vos, and others) as leaders and organizers in various interoperability projects supported by NSF via NESCent (NSF 0423641; K. Smith), including the *Hackathons, Interoperability, Phylogenies (HIP) working group* (2010 to 2013; A. Stoltzfus, R. Vos and E. Pontelli, Co-PIs); the *Evolutionary Informatics Working Group* (R. Vos and A. Stoltzfus, Co-PIs; 2006 to 2009); the *Phyloinformatics Hackathon* (organized in 2006 by H. Lapp, T. Vision, A. Stoltzfus, R. Vos); and the *Evolutionary Database Interoperability Hackathon* (organized in 2009 by H. Lapp, A. Stoltzfus, R. Vos, T. Vision, K. Schulz). We also have received support from iPlant, the Biodiversity Synthesis Center of EOL, and the Biodiversity Information Standards organization (aka **TDWG**). These efforts have built a cohesive phyloinformatics community and led directly to, or contributed importantly to, tangible outcomes such as the NeXML format [7], CDAO [5], Phylotastic [12], various improvements to Bio* libraries described in [4], and the crucial study of data re-use in [11] (which emerged from a TDWG workshop). In the fall of 2014 we will stage a hackathon (with OpenTree and Arbor) to leverage OpenTree's recently deployed suite of web services providing access to its synthetic tree, 4000 source trees, and reference taxonomy.

A.5.3. Results from relevant prior NSF support. Pontelli served as PI of grant HRD-0420407 (Center of Research Excellence in Bioinformatics and Computational Biology; 08/04-07/10).

- Intellectual Merit: the grant enabled the establishment of a bioinformatics research center at NMSU that developed innovative projects in the areas of semantic web, protein structure prediction, ecology, and synthesis of gene regulatory networks, leading to over 200 peer-reviewed publications and leveraging over $4M of additional funding.
- Broader Impact: The Center, which established graduate degree programs in bioinformatics, is now self-sustaining. Over 9 years, it has involved 30 faculty and researchers, graduated >50 students, and led 7 large-scale research projects, now independently funded.

Pontelli is the PI of grant DGE-0947465 (**DISSECT**: Advancing Computational Thinking in the Classroom through Applied Computational Research; 04/10-03/15).

- Intellectual Merit: this project explores a novel model for infusing computational thinking in 6th to 11th grade science courses [55, 56], based on peer-interactions between Computer Science graduate students and K-12 teachers.
- Broader Impact: The proposed activities enhanced the interest and success of students in STEM disciplines, improving computational competence of K-12 students in school districts (Las Cruces and Gadsden) with a diverse student population (70% to 90% Hispanic).

O'Meara is PI on Phylogeographic Inference using Approximate Likelihoods (DEB 1257669: $340,000)

- Intellectual merit: Since the grant was awarded a few months ago, methods have been extended to do exhaustive searches and do present phylogeographic scenarios in interactive figures
- Broader Impact: Presented workshop on the method; invited presentation at national meeting.

O'Meara is Co-PI on Historical naming traditions and cryptic speciation bias biodiversity estimates (DEB 1144974, $141,143)

- Intellectual merit: We tested Initial hypotheses about species delimitation using modern methods
- Broader Impact: None by O'Meara yet

B. Design and implementation plan

B.1 Conceptual Design

B.1.1. Design objectives. The primary design criterion is to deliver expert Tree-of-Life knowledge in a computable, convenient, and credible way. The system must deliver *computable* information, including trees encoded using standard formats, and metadata encoded using available standards (e.g., for citations), based on ontologies and controlled vocabularies. The user interaction must be *convenient*, in the sense of: **(1)** enabling interfaces that align with user expectations from upstream workflow steps, **(2)** returning results in seconds or minutes (not hours, days or weeks, as for *de novo* tree inference), and **(3)** returning the form of result that integrates into downstream steps. The system must provide a *credible* alternative to *de novo* inference for quantitatively important use-cases. In phylogenetics, where the external standard of truth for an inference– actual evolutionary history– is inaccessible, trees are judged by how they are produced, or who produces them, which means that the system must generate a description of sources and methods, sufficient to satisfy users who may wish to include such a description in the *methods* section of a scientific paper.

The second design criterion is to foster a *sustainable* and *adaptable* community infrastructure. The main implication of this criterion, in the context of an ever-changing landscape of resource-providers and funding, and in light of an NSF report on sustainability [57], is that the system must be *distributed*, *flexible* and *open*. We envision a community of practice in which different groups of experts can add resources to the system independently, via modular components that interact through common standards.

B.1.2. Implementing workflows through composition of web services. In our initial design based on [12], a **controller** satisfies the user's query by composing a workflow from available web services. A **client** is any program that uses one or more component **services**, even indirectly. A controller is a client that mediates workflow operations (e.g., linking the output of X to the input of Y). It may have a user interface, or be a library module included in client applications. Our production controller will access a registry of services and invoke planning algorithms for workflow composition, but other controllers may be simpler. In a typical flow of operations (see Figure 1), the controller: **(1)** Receives a query from user (through a client), issues ticket; **(2)** Invokes TNRS to clean up names; **(3)** Invokes treestores to discover trees that satisfy the query with optimal coverage; **(4)** Invokes an extractor to get the relevant subtrees; **(5)** Invokes a scaling service to add branch lengths or dates; **(6)** Provides the resulting subtrees to the user, with metadata including a report on provenance.

B.1.3. Adaptability and sustainability through a distributed modular system. Separate from the development of a production service, a critical contribution of this project will be the development of an abstract *architecture,* composed of standards (e.g., APIs, ontologies), methodologies and libraries to enable the creation of new services and their integration, allowing for:

- **Diverse providers**. This project will provide a complete set of production services for the canonical workflow, but the system is open. Anyone with a web server can implement an API, register their service, and begin processing any requests they receive. Metadata will aid users in the selection of services (e.g., in terms of QoS or "authority" of data sources).

- **Diverse workflows**. There isn't just one way to use ToL knowledge. The canonical workflow (Fig. 1) takes advantage of 4 key steps. However, if the problem is merely to discover all available trees with species A, B, and C, the workflow might have one step (tree discovery) or two (name reconciliation then tree discovery); if the problem is merely to validate names, then only the TNRS is needed; etc. Similarly, the 4-component workflow can be expanded with additional operations (e.g., to enrich trees of additional metadata) or embedded in larger analysis workflows.

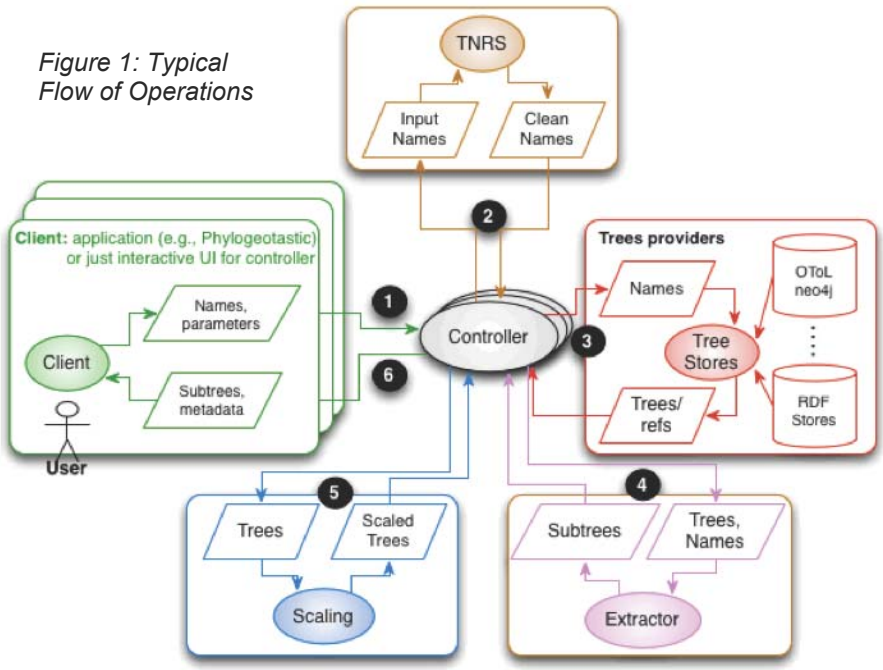


Figure 1: Typical Flow of Operations

- **Diverse clients**. Given many workflows, there may be many controllers. For the same workflow, there may be different clients that cater to different cases, e.g., our client (below) for web resources will be distinct from that for analyzing high-value comparative data sets.
- **Redundant components**. For purposes of testing, we will implement the minimal redundancy (>1 of each component). Eventual community adoption will result in multiple components of each type, e.g., resources like TreeBASE [50] or ToLWeb [51] may offer their content via the TreeStore API (one of the goals of the 3rd-year hackathon).

B.2. Iterative design and evaluation by the design team

The approach used here will rely on (1) iteration; (2) a primary focus on use-cases for evaluation; (3) involvement of the whole design team to tackle integration challenges; (4) community engagement by various means. After an initial requirements-gathering phase, there will be four main design-implement-evaluate cycles (each beginning with a face-to-face meeting), culminating in release of the production version. The design team of 8 to 10 people includes use-case experts and key staff from the Architecture team (NMSU), the Scaling team (UTK), the TreeStore team (our OpenTree partner), the GNR TNRS team (our partner at **MBL**, the Marine Biological Laboratory, Woods Hole, MA), and the Evaluation-Clients team (UMD). Use-case experts will change over time depending on focal use-cases (letters: Walls of iPlant; Parr of EOL; Sidlauskas of PhyloGeoTastic). The design team, led by Pontelli, will meet twice per month by videoconference, and will have a face-to-face meeting every 6 months in years 1 and 2. The face-to-face meetings will begin with a summative evaluation of the current state of the system, based on use cases (C.2).

B.3. Use cases

Design and evaluation will be guided by use-cases described briefly here (see also C.2).

B.3.1. Generate trees from scientific publications. Publications that use a species tree are useful for testing because they exemplify what researchers want, and what qualifies as adequate for publication. For this project we will obtain a set of recently published studies that represent diverse sub-disciplines and use-cases, and use them to generate ad hoc tests. At least 40 studies will be a

random sample of recent literature (as in our previous study [11]). The challenges are to reproduce the trees and, as appropriate, constraints on tree discovery and downstream analysis.

B.3.2. Flexible analysis of high-value comparative data sets. A typical research use of extracted trees, exemplified in Walls [34] or Riek [35], is to conduct an evolutionary analysis of a matrix of compared traits for a small set of species. For this use-case, Dr. Walls of iPlant (letter) will join the design team. Our approach is inspired by a hackathon demonstration project – integration of subtree extraction web-services into Mesquite [58], an extensible workbench for evolutionary analysis – and by the recent emergence of high-value data sets covering many species for many traits, e.g., the leaf economics spectrum data from the Global Plant Trait Network [59], or the mammal "phenomics" matrix of [60] (see also [61]). We will embed one or more such high-value data sets in an available analysis environment (e.g., Mesquite, R, iPlant Discovery Environment) modified to allow automated discovery of species trees. Using published studies as a guide, we will create recipes for various types of analyses, and offer screencasts that demonstrate the use of tools.

B.3.3. Autogenerate trees for web pages. As noted above (A.1), phylogenies would be a useful addition to many thousands of taxon pages, and sometimes are added using ad hoc methods. Dr. C. Parr of EOL [43] will join the design team as we work on 2 approaches to auto-generating trees for embedding in web pages. The first approach is to auto-extract names from existing page content (e.g., the Wikipedia "ant" page has 277 discoverable names, and we can recover a useful phylogeny with 88 matching species). The second approach is to generate a phylogeny for a higher taxon (e.g., "Carnivora"), including the case in which the taxon is too large to present, and a useful phylogeny requires a sampling approach, e.g., terminate the tree with families or genera, or choose 40 species based on a relevance metric such as the count of records in PubMed, EOL or **GBIF** (the Global Biodiversity Information Facility). We will build a client (executed when page content is updated) that discovers a phylogeny and returns a clickable HTML5 view with embedded links to species pages.

B.3.4. Autogenerate trees for gene-species tree reconciliation. Hackathon participants in [12] prototyped an application for gene-species tree reconciliation that allows the user to choose a starting gene tree, discover the corresponding list of species (using existing web services), obtain a species tree (using our web services), and carry out gene-species tree reconciliation (using a local reconciliation engine). Using an environment such as R, Python or BioPerl, we will implement a token annotation workflow that processes each gene in a genome by conducting a standard BLAST search to find homologs, obtains a species tree for reconciliation, and identifies orthologs.

B.3.5. PhyloGeoTastic and other use-cases and demonstrations. We will remain open to other use-cases and applications, including many possibilities suggested by hackathon discussions (e.g., a taxonomic name-completion widget, a mobile app for museums or zoos). One example that is part of our plan for engaging educators is based on a tool (PhyloGeoTastic) that uses subtree extraction to get a phylogeny based on species occurrence data. The user clicks on a map to select a custom geographic region, and the application responds by retrieving species occurrence records for that region (from iNaturalist or GBIF), then finding a tree for the implicated species.

B.4. Developing component services

Individual teams have the responsibility to develop core capacities (e.g., subtree extraction, spell-checking, scaling trees), work with the design team to develop APIs for the integrated system, and *implement those APIs so that the core capacities become available, via the API, as* ***component services*** *in the integrated system*.

B.4.1 TreeStores. The production TreeStore for this project will be provided by OpenTree, a multi-investigator project with ongoing support from NSF. It is already within OpenTree's mandate to ensure (1) strategies and standards for obtaining, licensing, versioning and updating trees; (2) ensuring that technologies for storage, referencing and retrieval are scalable for accessing multiple trees with ~10^6 nodes. OpenTree's database, implemented in neo4j [62], currently has about 4000 source trees, and a synthetic tree with 2.5 million species. These resources are available via a recently announced web services interface, and via flat files in github. The main challenge for OpenTree is to support the types of queries that emerge from our requirements gathering, and then

to implement the API that we develop together. To test our integrated system, it is important to have other TreeStores: for this purpose, we will also implement the supported parts of our API as a thin layer on CDAOStore [63] populated with a small set of high-value (i.e., very large) trees (e.g., [40, 41, 51, 64-71]), and on the existing web-services API of TreeBase [50].

B.4.2 Taxonomic Name Resolution Service. Among multiple web services offering validation or matching of names, the state of the art is a multi-featured spell-checking TNRSs stocked with information from multiple naming authorities [49]. The production TNRS for this project will be the Global Names Resolver (http://resolver.globalnames.org/), a part of the Global Names Architecture [48] that is already in operation, providing name-resolution services to projects such as **BHL** (Biodiversity Heritage Library). This work will be contracted to MBL, which houses the GNR team headed by Dmitry Mozzherin (letter), one of the authors of [49]. GNR will work with our design team to develop an API allowing name resolution to be an efficient part of our workflows. Necessary components of name resolution include name parsing, fuzzy matching, and normalization of synonyms where possible. Batch processing provides a means for context-based resolution of homonyms: a species name shared between taxa A and B is resolved (provisionally) in favor of B if the rest of the user's query refers to members of B.

B.4.3 Scaling of unscaled phylogenies. Converting topologies (tree structures without branch lengths) and phylograms (tree structures with branch lengths based on character change) to chronograms (trees with branches in time units) is a necessary step for many use cases for trees. There have been several methods developed to do this using fossil calibrations [72-74] and existing chronograms [75]. Currently, DateLife uses two methods. If presented with a list of taxa, it looks over a stored database of chronograms to return the age of the most recent common ancestor of all the taxa, complete with uncertainty. If presented with a topology, it uses the method of Eastman et al. [75] to stretch the input tree to match each stored chronogram. DateLife has numerous output options (tree structures, web pages, and raw numeric values) accessed via a RESTful [76] interface.

DateLife currently suffers from two issues that this project will address. The first is to implement the API to be designed in this project. The second is the scarcity of stored trees. Although DateLife has thousands of stored chronograms, they come from just 18 studies and have major gaps in taxonomic coverage (the comparable TimeTree project has over 1,000 studies, but does not release its data store for public use). OpenTree's treestore covers >4000 studies, but fewer than 1% have metadata indicating that they are chronograms. The DateLife team will find and add more chronograms to DateLife and collaborate with OpenTree to push data into its treestore.

B.4.4. Subtree extraction. OpenTree has implemented efficient subtree extraction and will offer it as a service using the API we develop. We also will create a stand-alone subtree extraction service based on available methods [12], with a system of caching so that trees can be passed by reference.

B.5. Developing an integrated system

The realization of the architecture illustrated in Figure 1 requires addressing a number of challenges. These challenges span both the domain of phyloinformatics and more traditional computer science.

B.5.1. A registry of services. Each client application identifies a controller, which describes the desired workflow to address the user needs. The workflow execution is realized through composition of web services. The first premise of this architecture is the presence of a **service registry**. The registry provides a centralized resource to publish the available services, advertising in a computable form functionalities and requirements of each service.

Several domain-independent platforms for service registry implementation have been proposed, mostly relying on UDDI [77]. In this project, we will deploy a service registry based on a public domain UDDI server (e.g, [78] or [79]). The service registry will be deployed on a dedicated server, to be acquired as part of this project and installed at NMSU. The domain-independent registry will rely on standard formats to advertise services (e.g., WSDL [80]); on top of this, we will build a layer to provide a domain-specific semantic description and classification of services, following established guidelines and languages (e.g., OWL-S [81]), and expressed in terms of a formal

ontology. The ontology should not only capture the operational behavior of the service (e.g., this service is a subtree extraction), but relevant properties, such as quality of services, input/output formats, scope limitations (e.g., operates only on fully resolved trees), and optional parameters (e.g., ability to select a specific database for name resolution). Pontelli and Stoltzfus have been involved in developing a domain ontology for evolutionary analysis [5, 6, 8], and in efforts to formalize a vocabulary and a skeleton ontology for a minimal-information standard, MIAPA [8, 82]. Nevertheless, the previous efforts are not complete and do not provide the specific focus and level of detail required for the configuration problem at hand. We propose to build on such efforts (in particular, [6]) to create an organized vocabulary, and then a formal (OWL-based) ontology for the annotation of services required for the proposed architecture.

B.5.2. Communicating commands, data, and metadata. The development of a web-service API to execute services (e.g., for RESTful services [76]), a primary focus of activity during the first 15 months, will provide the means of communication among parts of the system, which requires conveying data (e.g., a species name, the structure of a tree, a branch length), commands (including parameters, e.g., use only trees with publication year > 2007), and metadata (e.g., provenance of an extracted tree). The information conveyed among components must be machine-processable, richly structured, and extensible. The design will expand on the preliminary design in [12] and a provisional PhyloWS standard [83] (developed in an earlier hackathon and implemented in TreeBASE [50]). The prototype system in [12] passes all commands and parameters directly via URL (i.e., using GET), allows multiple output formats, and does not address metadata. In devising a more robust and complete API, we will follow a standard CRUD model, relying on HTTP messages to capture commands and parameters, and on XML for the representation of concrete data and metadata. In particular, we will rely on NeXML for the representation of data and specialized execution metadata (formalized using the previously mentioned ontology for methods). The design of NeXML [7] allows for arbitrarily complex semantic annotation of any data element using external vocabularies, such as those developed or extended for this project.

B.5.3. Workflow composition. The workflow associated to a controller, in general, identifies the core operations required to complete the desired goal (e.g., the steps described in Section B.1 or their expansion with additional analysis or presentations steps, such as a mashup with other metadata); furthermore, these are expressed within the context of a specific controller framework (e.g., Galaxy, Mesquite), that might expose to the designer operations at different levels of granularity. The process to generate an execution requires addressing the three challenges of (1) *configuring* the workflow, (2) *composing* the services, and (3) *enacting* the service composition.

Configuration: the problem of configuring a workflow is aimed at identifying all the specific services that need to be used to realize the workflow into a concrete executable entity. The core challenge is to transform each workflow operation in a query to the service registry, to identify all possible services that can accomplish the given operation. While several properties required to compose the query to the service registry will be obvious from the workflow description (e.g., the class to which the desired operation belongs to), others will need to be derived from the context in which the operation appears. For example, the nature of the query used to retrieve relevant phylogenies might provide information about the complexity of the trees to be processed, which might impact the choice of extraction operation (e.g., an optimized one to guarantee a certain response time). This type of "contextual" analysis is common in various areas of artificial intelligence (e.g., natural language processing [84-86]) but it has not been used to infer query structure for workflow configuration. We will address this problem using an action language description of the workflow (see the *composition* step below) and diagnosis algorithms (e.g., [87, 88]) to infer parameters of interest from the overall description of the complete workflow.

Composition: Once a pool of services has been identified, the composition step [52, 53] links services to produce an executable artifact. This requires addressing the following challenges: **(1)** The query to the service registry might potentially return multiple services for each operation; a challenge will be to identify the most appropriate service (e.g., in terms of reliability, response time); **(2)** Services need to be properly linked, using the APIs to connect inputs and outputs according to

the data flow prescribed by the workflow; **(3)** In order to establish these links, it might be necessary to introduce intermediate auxiliary operations (e.g., *shim* services for format translation); these might not be explicitly evident from the initial workflow description due to the higher level of abstraction and granularity used at the controller level. We propose to address these issues using technology drawn from the field of knowledge-based planning [89]. We adopt the view (pioneered in [90-92] and supported by several researchers [52, 93-95]) that each component of a workflow can be described as a *world-altering action*. Each action is associated with a precondition (e.g., describing the inputs) and a set of effects (e.g., outputs). Given a description of workflow components as actions, *planning algorithms* can be applied to automatically compose services that can properly communicate and represent an executable artifact. The initial workflow provides constraints on the plan being generated – it identifies the classes the actions should belong to (e.g., TNRS) and constraints on their order of execution. We will use state-of-the-art in planning technology (e.g., [96-100]) to address this problem. Because of the restricted nature of the problem to be addressed (e.g., the high-level structure of the workflow is known), the problem is efficiently solvable.

Enactment: we propose to develop libraries within existing codebases (e.g., R-phylo, BioPython) to support interaction with the registry and implement the APIs we develop. These libraries will be adequate representations of the executable workflow. Nevertheless, some challenges to enactment of the workflow remain. First of all, transactional behavior should be guaranteed, in light of possible failures of services, guided by monitoring of execution. Service failures should lead to a possible reconfiguration of the workflow (i.e., to replace a failed service with an alternative one, after warning the user), repeating the configuration and composition from the point of failure.

B.6. Developing Clients

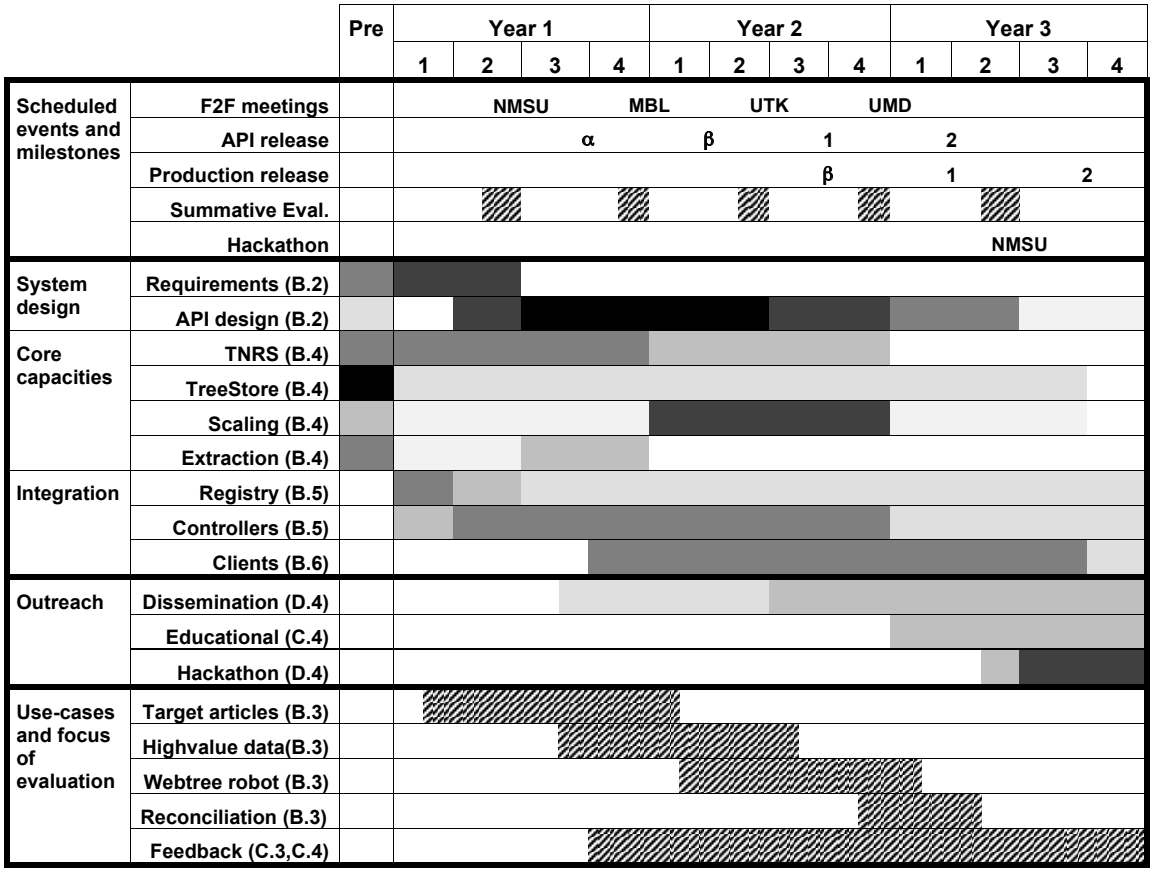
For this project we will develop, not only generalized controllers (above), but specialized clients, including a client with an interactive web interface (for the scientific papers use-case described in B.3.1, for classroom applications, etc.), an extension of an existing evolutionary analysis environment (for B.3.2), an embedded module invoked via a web content management system (for B.3.3), a script integrated in an R or BioPerl workflow for reconciliation (for B.3.4), and modifications to an existing PhyloGeoTastic client (for classroom applications; see B.3.5). Each client invokes a controller module (above). Here we describe only the generalized web client.

The most obvious public-facing outcome of the project will be a web program to request a tree using only a web browser to enter taxonomic inputs and constraints. We can imagine choosing *inputs* as follows: the user may enter a higher taxon, or a list of species names (using auto-completion), or process names from a remote source (identified by a URL) or an uploaded file. We can imagine designating *constraints* as follows: the user may wish to use only published source trees, or to use only source trees published in some recent interval; the user may wish to allow taxon substitution (at a desired level of species, genus, family), and so on. The choices of *outputs* might be a downloadable phylogeny (in various formats), a processing report, or an interactive tree rendered by an embedded viewer (e.g., Archaeopteryx [101]).

B.7 Milestones and Expected Deliverables

The table below conveys a timeline of scheduled events and milestones (above), followed by a chart showing how effort is distributed to different aspects of the projects, and showing the attention to use-cases used in formative and summative evaluation (not accounted are TreeStore efforts of our separately funded partner, OpenTree). Note that some aspects of the system have received considerable prior attention [11, 12, 49]. In year 1, as teams develop core capacities and build a framework for workflow composition, design efforts begin with requirements-gathering, then proceed to API design. In year 2, these activities continue, and the focus of development shifts toward integration and client applications, culminating in two more releases of the API, and the beta release of the production system in month 21, followed by a full release in month 27. The evaluation team performs a summative evaluation prior to each major design meeting. Formative evaluation takes place throughout the project, based on (1) use-cases analyzed by the evaluation team, and (2) community feedback (comments, bugs, feature requests) on designs and implementations, which

begins with the initial release of the API design. Outreach activities take place throughout the project (e.g., via social media), and become an important focus in year 3, with a hackathon in month 30 with follow-ups to leverage hackathon products, and classroom evaluation and improvement of the generalized web client and the PhyloGeoTastic tool.



C. Evaluation and Testing

Formal evaluation strategies will be applied to assess progress and evaluate the proposed architecture. *Formative* evaluation methods will be applied throughout the design and implementation of the different components. Note that core capacities of individual components can be tested separately from the overall architecture, via benchmarks, unit tests, and experimental comparisons. *Summative* evaluation of the integrated system will be achieved through: (1) applications of clients to use-cases described above (B.3); (2) distribution of all implemented services, APIs and libraries to collaborators and to the broad community. Both evaluation phases will include feedback mechanisms to communicate changes to the development team. Each PI will coordinate the evaluation of components within his area of expertise; Dr. Stoltzfus will coordinate the evaluation of the overall architecture, and consult with project staff on evaluations of individual components based on use-cases.

C.1 Evaluation Metrics

The evaluation process will rely on a number of metrics that will be used to assess progress towards the stated goals and effectiveness of the proposed technologies. Throughout the project we will consider both correctness metrics as well as quality metrics. Correctness metrics will be used to assess the capabilities of the individual components and their integration to correctly meet their stated purpose within set error tolerance levels (e.g., levels of precision in the scaling process). Quality metrics will be employed to assess quality and performance of the proposed technologies; examples of quality metrics include (1) traditional performance metrics (e.g., speed, memory usage,

bandwidth requirements); (2) coverage (e.g., range of trees that can be effectively extracted, data formats that can be processed); (3) accuracy (e.g., scaling precision, level of TNRS correction); (4) scope (e.g., range of clients and controllers that can be effectively developed). Rubrics, automated evaluation tools and other instruments will be developed to assess metrics values for each component of the project, and used to inform design and implementation changes.

C.2. Role of use cases

The evaluation team will take advantage of the expertise of project staff and outside consultants to develop tests based on the use-cases described above (expert consultants are listed for each use-case). Concrete tests for speed, accuracy, coverage, and capacity will be disseminated by the evaluation team to other project members, and used in formative and summative evaluations. Formative evaluations take place on a continual basis. The two main contexts for summative evaluations are (1) the face-to-face design meetings (every 6 months), which will begin with a presentation by the evaluation team of the current state of the project, based on available tests, and (2) the release of the beta and full production versions.

Some examples may serve to illustrate possible tests. The scope and ease of reproducing trees from past studies is an indicator of the capacity for supporting future studies: this is the premise of the “Generate Trees for Scientific Publications” use-case (B.3.1). Each study implicates a list of species, provides a published tree (useful as a standard of comparison), and may implicate criteria for tree selection (e.g., if the study infers a tree only from morphological data). Likewise, hundreds of studies re-use the Leaf Economics Spectrum data [11] invoked in the “Flexible analysis of high-value data sets” use-case (B.3.2): a systematic way to evaluate power is to assess how thoroughly such published studies can be replicated with our client application. The reconciliation use-case (B.3.5) will generate a large volume of distinct requests, which can be to assess the frequency of server errors, or the maximum load of queries that we can process within a given response time.

C.3. Engagement with phyloinformatics community

Our preliminary work is an example of successful engagement with a community interested in the informatics of phylogenies, comparative data, taxonomies, and biodiversity data. The proof-of-concept “Phylotastic” system was built by scientists who responded to an open call, leaving their “day jobs” to work on this project. The resulting publication [12] has been accessed >6,000 times.

For the proposed project, engagement with the evolutionary informatics community will take place through (1) regular communication with stakeholders and followers using an email list (currently with >80 hackathon participants and followers) that has been active for >30 months in project-related communications; (2) invited participation of use-case consultants in design meetings; (3) broadly distributed RFCs on API specifications, (4) feature requests and bug reports logged in our public code repositories, (5) the hackathon that takes place in year 3, (6) demonstrations and workshops at scientific meetings, (7) proposals for student engagement via Google Summer of Code.

The hackathon in year 3 provides a dual opportunity. As in past events, we will issue an open call and screen applicants based on diversity, qualifications, and potential. We will welcome resource-providers who wish to join the system, and those wishing to develop clients. The hackathon will result in 4 to 6 projects that reflect community interests. A shortcoming of past hackathons has been the lack of dedicated staff to follow up on promising ideas: for this project, we will dedicate significant resources to leverage those hackathon products with potential to inspire broader adoption of the system as a sustainable community infrastructure.

C.4. Integration of Research and Education

The project will provide several dimensions to the integration of research and education. The investigators will rely on both graduate and undergraduate students to sustain several development activities. Mentoring of students will follow the general guidelines of the post-doc management plan, with an appropriately greater emphasis on formal training, and a lesser emphasis on professional development. Participating students will be exposed to an inter-disciplinary working environment and gain knowledge of software engineering, evolutionary analysis, algorithm design, etc.

Two client applications, the generalized web client and PhyloGeoTastic (B.3, B.6) will be introduced into the classroom by educators, and their feedback will become part of our evaluation process for the effectiveness of client interfaces. The highly visual nature of PhyloGeoTastic (B.3.5) makes it suitable for educators to teach post-primary students about the evolution of communities, the geographic distribution of phylogenetic and taxonomic diversity, endemism, adaptive radiation, and other topics. Dr. Brian Sidlauskas (letter), who led the team that prototyped this idea, will join the design team while we improve the application. We will work with educators committed to using the software in biology courses at NMSU, UMD, and U. Mass College of Liberal Arts (letters: Nishiguchi, Haag, Himes). An additional dimension to the integration of research and education will derive from our existing work with K-12 schools in the DISSECT project (PI: Pontelli) and in the **YO-GUTC** (Young Women Growing Up Thinking Computationally) project (letter: Galves of NMSU, program director of YO-GUTC). This will allow us to bring tools developed in this proposed project to high school courses (e.g., Biology, Forensics) and summer camps (primarily focused on high school students from traditionally underrepresented groups); we are already working with Biology teachers at several High schools in Las Cruces through the two above mentioned projects. The collaboration with DISSECT and YO-GUTC will avail this project of the expertise of the external evaluators of DISSECT in assessing the impact of educational components.

D. Project Management

The organizational units of the project are the 3 individual centers (each headed by an investigator) supplemented by our OpenTree and Global Names collaborators; the *leadership team*, consisting of the 3 investigators plus Cranston (OpenTree) and Mozzherin (Global Names) and headed by Stoltzfus; the *design team*, led by Pontelli, consisting of the investigators plus key staff, collaborators, and use-case consultants; and a community of stakeholders and followers currently represented by an email list of ~80 people. The timeline for the project is in B.7 (above).

D.1 Personnel and Expertise

The project will develop as an interdisciplinary collaboration among computer scientists, evolutionary biologists, software developers, and stakeholders including resource-providers and consumers. The project will be led by Dr. Stoltzfus, who has extensive experience in evolution and bioinformatics, including theory [102-107], methods and application of phylogenetic analysis [13, 15, 19, 27, 108], software development [3, 7, 109, 110] and interoperability, including organizational leadership in multiple NESCent- and TDWG-sponsored activities focused on interoperability [4, 7, 12, 82, 111]. He will oversee and coordinate all aspects of project, and direct the research components focused on use cases, including the development of tests, summative evaluations, and client applications.

Dr. Pontelli provides expertise in software development (including large-scale projects, and an award-winning automated planner), planning, knowledge representation and reasoning, and high-performance computing [e.g., 55, 56, 91, 92, 96-100, 112, 113]. He has been active in the evolutionary informatics community [5, 8, 12, 63, 82]. In addition to leading the design team, Pontelli will lead the development of controllers for workflow composition, and serve as an expert on evaluating the performance and scalability of the software produced by the project.

Dr. O'Meara (DateLife) is an expert R programmer with extensive experience in development and application of phylogenetic methods, including fossil calibration [12, 14, 16-18, 20-26, 28-33, 111, 114].

The expertise of the PIs is supplemented by key collaborators on the design team (Cranston, lead PI of OpenTree; Mozzherin, lead developer of GNR).

D.2 Management Structure

The key responsibility of the leadership team is to ensure continued coordination of the collaborators, so as to ensure that the project remains focused on its major goals. The leadership team will meet by videoconference quarterly, and will meet in person every 6 months (for 3 to 4 days) as part of the face-to-face design meeting. However, coordination of development activities occurs mainly via the design team, led by Pontelli, which will have bi-weekly videoconferences to

discuss progress, set priorities, and consider all issues relevant to design and system integration. Each collaborating institution is responsible for providing core capacities, and has the authority to make implementation choices, but APIs are the responsibility of the design team.

We are experienced with, and will use, a variety of technologies for electronic communication and remote collaboration. The leadership team will have a private wiki and email list. The design team will have an email list and wiki that are public, and will be open (by invitation) to any non-staff who are committed to being part of the design process.

Coordination with related projects depends on personnel already in those projects: Mozzherin in Global Names Architecture; Cranston in OpenTree; Walls in iPlant; Parr in EOL; Pontelli and Stoltzfus in MIAPA efforts [82, 111] and CDAO [5]; Stoltzfus in NeXML [7].

D.3 Dissemination

All software, protocols, and architectures developed within this project will be disseminated as collaborative open-source projects. We will promote early dissemination of all software components being developed, under the terms of an OSI license [115]. Projects will be hosted on GitHub and, according to standard practices in the open-source community, will be open to non-project developers who show potential and commitment. This project also will contribute code to existing open-source projects (R, BioPerl, BioPython, NeXML, CDAO, MIAPA), using whatever repositories and licenses apply. We will communicate project announcements (updates, RFCs) with a larger community of stakeholders through a public wiki, a Google+ circle, a public email list (for anyone interested in following the project), and other means as appropriate. Traditional dissemination mechanisms also will be used, including journal publications, conference presentations and workshops (e.g., Evolution, TDWG, BOSC, SMBE), and visits to different institutions.

D.4 Sustainability Plan

Our plan for sustainability is to achieve, in 3 years, the critical mass that will lead in 2 years to widespread adoption in the realm of computable access to the ToL, as key players realize the potential of the system and commit resources to develop clients and implement APIs. The premise of this vision is that long-term sustainability depends on an effective architectural design, a committed and organized stakeholder community, and an ongoing registry. Achieving critical mass in our 3-year project depends on outcomes addressed above: (1) a robust production service; (2) effective demonstrations of scope and utility; (3) a core set of components and controllers that are open-source packages available for others to install, use, and adapt; (4) well documented reference implementations to guide other developers; (5) training, outreach, and dissemination activities to involve potential stakeholders and cultivate a community of practice to sustain the system.

Beyond the scope of this proposal, while the system is being adopted, it is important for us to maintain, update and improve the service registry in response to bug reports and feature requests. The NMSU team is committed to maintaining and updating the service registry (which has a small data footprint) indefinitely. The UTK team and GNR (MBL) will maintain production services for 3 years, and all teams commit to maintaining source code for production systems for 3 years.

To accelerate adoption, the investigators will pursue additional funding for new developments in areas such as (1) handling unrooted trees and anastomosing trees; (2) phylogenetic grafting; (3) accommodating the unique challenges of prokaryotic phylogenetics; and (4) managing trees as posterior distributions. Additionally, funding opportunities will be explored (e.g., SBIR/STTR) for applications with commercial potential (e.g., mobile apps for museums, services to validate names in published manuscripts, biomedical applications requiring tree-reconciliations).

References

1. Mora, C., D.P. Tittensor, S. Adl, A.G. Simpson, and B. Worm, *How many species are there on Earth and in the ocean?* PLoS biology, 2011. **9**(8): p. e1001127.
2. Cracraft, J., M. Donoghue, J. Dragoo, D. Hillis, and T. Yates. *Assembling the tree of life: harnessing life's history to benefit science and society*. 2002; Available from: http://ucjeps.berkeley.edu/tol.pdf.
3. Hladish, T., V. Gopalan, C. Liang, W. Qiu, P. Yang, and A. Stoltzfus, *Bio::NEXUS: a Perl API for the NEXUS format for comparative biological data.* BMC Bioinformatics, 2007. **8**: p. 191-201.
4. Lapp, H., S. Bala, J.P. Balhoff, A. Bouck, N. Goto, M. Holder, R. Hollan, A. Holloway, T. Katayama, P.O. Lewis, A. Mackey, B.I. Osborne, W.H. Piel, S.L. Kosakovsky Pond, A. Poon, W.G. Qiu, J.E. Stajich, A. Stoltzfus, T. Thierer, A.J. Vilella, R. Vos, C.M. Zmasek, D. Zwickl, and T.J. Vision, *The 2006 NESCent Phyloinformatics Hackathon: A field report.* Evolutionary Bioinformatics, 2007. **3**: p. 357-366.
5. Prosdocimi, F., B. Chisham, E. Pontelli, J.D. Thompson, and A. Stoltzfus, *Initial Implementation of a Comparative Data Analysis Ontology.* Evolutionary Bioinformatics, 2009. **5**: p. 47-66.
6. Panahiazar, M., A. Ranabahu, V. Taslimi, H. Yalamanchili, A. Stoltzfus, J. Leebens-Mack, and A.P. Sheth. *PhylOnt: A domain-specific ontology for phylogeny analysis*. in *Bioinformatics and Biomedicine (BIBM), 2012 IEEE International Conference on*. 2012.
7. Vos, R.A., J.P. Balhoff, J.A. Caravas, M.T. Holder, H. Lapp, W.P. Maddison, P.E. Midford, A. Priyam, J. Sukumaran, X. Xia, and A. Stoltzfus, *NeXML: rich, extensible, and verifiable representation of comparative data and metadata.* Systematic Biology, 2012. **61**(4): p. 675-89.
8. Lapp, H., E. Pontelli, A. Stoltzfus, and R. Walls, *An annotation ontology for validating minimum metadata reporting for phylogenetic analyses*, in *Annual International Conference on Intelligent Systems for Molecular Biology*. 2013.
9. Drew, B.T., R. Gazis, P. Cabezas, K.S. Swithers, J. Deng, R. Rodriguez, L.A. Katz, K.A. Crandall, D.S. Hibbett, and D.E. Soltis, *Lost branches on the tree of life.* PLoS Biol, 2013. **11**(9): p. e1001636.
10. Magee, A.F., M.R. May, and B.R. Moore, *The Dawn of Open Access to Phylogenetic Data.* PLOS ONE (accepted). 2014.
11. Stoltzfus, A., B. O'Meara, J. Whitacre, R. Mounce, E.L. Gillespie, S. Kumar, D.F. Rosauer, and R.A. Vos., *Sharing and Re-use of Phylogenetic Trees (and associated data) to Facilitate Synthesis.* BMC Research Notes, 2012. **5**: p. 574.
12. Stoltzfus, A., H. Lapp, N. Matasci, H. Deus, B. Sidlauskas, C.M. Zmasek, G. Vaidya, E. Pontelli, K. Cranston, R. Vos, C.O. Webb, L.J. Harmon, M. Pirrung, B. O'Meara, M.W. Pennell, S. Mirarab, M.S. Rosenberg, J.P. Balhoff, H.M. Bik, T.A. Heath, P.E. Midford, J.W. Brown, E.J. McTavish, J. Sukumaran, M. Westneat, M.E. Alfaro, A. Steele, and G. Jordan, *Phylotastic! Making tree-of-life knowledge accessible, reusable and convenient.* BMC bioinformatics, 2013. **14**: p. 158.
13. Stoltzfus, A., J.M. Logsdon, Jr., J.D. Palmer, and W.F. Doolittle, *Intron "sliding" and the diversity of intron positions.* Proc Natl Acad Sci U S A, 1997. **94**(20): p. 10739-44.
14. Farrell, B.D., A.S. Sequeira, B.C. O'Meara, B.B. Normark, J.H. Chung, and B.H. Jordal, *The evolution of agriculture in beetles (Curculionidae: Scolytinae and Platypodinae).* Evolution, 2001. **55**(10): p. 2011-27.
15. Qiu, W.G., N. Schisler, and A. Stoltzfus, *The evolutionary gain of spliceosomal introns: sequence and phase preferences.* Mol Biol Evol, 2004. **21**(7): p. 1252-63.

16. Driskell, A.C., C. Ane, J.G. Burleigh, M.M. McMahon, C. O'Meara B, and M.J. Sanderson, *Prospects for building the tree of life from large sequence databases.* Science, 2004. **306**(5699): p. 1172-4.
17. O'Meara, B.C., C. Ane, M.J. Sanderson, and P.C. Wainwright, *Testing for different rates of continuous trait evolution using likelihood.* Evolution, 2006. **60**(5): p. 922-33.
18. McBride, C.S., J.R. Arguello, and B.C. O'Meara, *Five Drosophila genomes reveal nonneutral evolution and the signature of host specialization in the chemoreceptor superfamily.* Genetics, 2007. **177**(3): p. 1395-416.
19. De Kee, D.W., V. Gopalan, and A. Stoltzfus, *A Sequence-based Model Accounts Largely for the Relationship of Intron Positions to Protein Structural Features.* Mol Biol Evol, 2007. **24**(10): p. 2158-68.
20. Collar, D.C., B.C. O'Meara, P.C. Wainwright, and T.J. Near, *Piscivory limits diversification of feeding morphology in centrarchid fishes.* Evolution, 2009. **63**(6): p. 1557-73.
21. Smith, S.A. and B.C. O'Meara, *Morphogenera, monophyly, and macroevolution.* Proc Natl Acad Sci U S A, 2009. **106**(36): p. E97-8; author reply E99-100.
22. Collar, D.C., J.A. Schulte, 2nd, B.C. O'Meara, and J.B. Losos, *Habitat use affects morphological diversification in dragon lizards.* J Evol Biol, 2010. **23**(5): p. 1033-49.
23. O'Meara, B.C., *New heuristic methods for joint species delimitation and species tree inference.* Syst Biol, 2010. **59**(1): p. 59-73.
24. Abercrombie, J.M., B.C. O'Meara, A.R. Moffatt, and J.H. Williams, *Developmental evolution of flowering plant pollen tube cell walls: callose synthase (CalS) gene expression patterns.* Evodevo, 2011. **2**(1): p. 14.
25. Beaulieu, J.M., D.C. Jhwueng, C. Boettiger, and B.C. O'Meara, *Modeling stabilizing selection: expanding the Ornstein-Uhlenbeck model of adaptive evolution.* Evolution, 2012. **66**(8): p. 2369-83.
26. Smith, S.A. and B.C. O'Meara, *treePL: divergence time estimation using penalized likelihood for large phylogenies.* Bioinformatics, 2012. **28**(20): p. 2689-90.
27. Yu, G. and A. Stoltzfus, *Population diversity of ORFan genes in Escherichia coli.* Genome Biology and Evolution, 2012. **4**(11): p. 1176-87.
28. Beaulieu, J.M., B.C. O'Meara, and M.J. Donoghue, *Identifying hidden rate changes in the evolution of a binary morphological character: the evolution of plant habit in campanulid angiosperms.* Syst Biol, 2013. **62**(5): p. 725-37.
29. Darrin Hulsey, C., B.P. Keck, H. Alamillo, and B.C. O'Meara, *Mitochondrial genome primers for Lake Malawi cichlids.* Mol Ecol Resour, 2013. **13**(3): p. 347-53.
30. Soltis, D.E., M.E. Mort, M. Latvis, E.V. Mavrodiev, B.C. O'Meara, P.S. Soltis, J.G. Burleigh, and R. Rubio de Casas, *Phylogenetic relationships and character evolution analysis of Saxifragales using a supermatrix approach.* Am J Bot, 2013. **100**(5): p. 916-29.
31. Jhwueng, D.C., S. Huzurbazar, B.C. O'Meara, and L. Liu, *Investigating the performance of AIC in selecting phylogenetic models.* Stat Appl Genet Mol Biol, 2014.
32. Williams, J.H., M.L. Taylor, and B.C. O'Meara, *Repeated evolution of tricellular (and bicellular) pollen.* Am J Bot, 2014. **101**(4): p. 559-71.
33. Zanne, A.E., D.C. Tank, W.K. Cornwell, J.M. Eastman, S.A. Smith, R.G. FitzJohn, D.J. McGlinn, B.C. O'Meara, A.T. Moles, P.B. Reich, D.L. Royer, D.E. Soltis, P.F. Stevens, M. Westoby, I.J. Wright, L. Aarssen, R.I. Bertin, A. Calaminus, R. Govaerts, F. Hemmings, M.R. Leishman, J. Oleksyn, P.S. Soltis, N.G. Swenson, L. Warman, and J.M. Beaulieu, *Three keys to the radiation of angiosperms into freezing environments.* Nature, 2014. **506**(7486): p. 89-92.
34. Walls, R.L., *Angiosperm leaf vein patterns are linked to leaf functions in a global-scale data set.* Am J Bot, 2011. **98**(2): p. 244-253.

35. Riek, A., *Allometry of milk intake at peak lactation.* Mammalian Biology Zeitschrift fur Saugetierkunde, 2011. **76**(1): p. 3-11.
36. Duarte, L.d.S., *Phylogenetic habitat filtering influences forest nucleation in grasslands.* Oikos, 2011. **120**(2): p. 208-215.
37. Burns, J.H. and S.Y. Strauss, *More closely related species are more ecologically similar in an experimental test.* Proceedings of the National Academy of Sciences of the United States of America, 2011. **108**(13): p. 5302-7.
38. Zhang, S.-B., J.W. Ferry Slik, J.-L. Zhang, and K.-F. Cao, *Spatial patterns of wood traits in China are controlled by phylogeny and the environment.* Global Ecology and Biogeography, 2011. **20**(2): p. 241-250.
39. Morlon, H., D.W. Schwilk, J.A. Bryant, P.A. Marquet, A.G. Rebelo, C. Tauss, B.J. Bohannan, and J.L. Green, *Spatial patterns of phylogenetic diversity.* Ecol Lett, 2011. **14**(2): p. 141-9.
40. Bininda-Emonds, O.R., M. Cardillo, K.E. Jones, R.D. MacPhee, R.M. Beck, R. Grenyer, S.A. Price, R.A. Vos, J.L. Gittleman, and A. Purvis, *The delayed rise of present-day mammals.* Nature, 2007. **446**(7135): p. 507-12.
41. Smith, S.A., J.M. Beaulieu, A. Stamatakis, and M.J. Donoghue, *Understanding angiosperm diversification using small and large phylogenetic trees.* Am J Bot, 2011. **98**(3): p. 404-14.
42. The Angiosperm Phylogeny, G., *An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG III.* Botanical Journal of the Linnean Society, 2009. **161**(2): p. 105-121.
43. Parr, C.S., N. Wilson, P. Leary, K.S. Schulz, K. Lans, L. Walley, J.A. Hammock, A. Goddard, J. Rice, M. Studer, J.T. Holmes, and R.J. Corrigan, Jr., *The Encyclopedia of Life v2: Providing Global Access to Knowledge About Life on Earth.* Biodivers Data J, 2014(2): p. e1079.
44. Doyon, J.P., V. Ranwez, V. Daubin, and V. Berry, *Models, algorithms and programs for phylogeny reconciliation.* Brief Bioinform, 2011. **12**(5): p. 392-400.
45. Leebens-Mack, J., T. Vision, E. Brenner, J. Bowers, S. Cannon, M. Clement, C. Cunningham, C. Depamphilis, R. DeSalle, J. Doyle, J. Eisen, X. Gu, J. Harshman, R. Jansen, E. Kellogg, E. Koonin, B. Mishler, H. Philippe, J. Pires, Y. Qiu, S. Rhee, K. Sjolander, D. Soltis, P. Soltis, D. Stevenson, K. Wall, T. Warnow, and C. Zmasek, *Taking the first steps towards a standard for reporting on phylogenies: Minimum information about a phylogenetic analysis (MIAPA).* Omics-a Journal of Integrative Biology, 2006. **10**(2): p. 231-237.
46. Webb, C.O. and M.J. Donoghue, *Phylomatic: tree assembly for applied phylogenetics.* Molecular Ecology Notes, 2005. **5**: p. 181–183
47. Martinson, H., K. Schneider, J. Gilbert, J. Hines, P. Hambäck, and W. Fagan, *Detritivory: stoichiometry of a neglected trophic level.* Ecological Research, 2008. **23**(3): p. 487-491.
48. Patterson, D.J., J. Cooper, P.M. Kirk, R.L. Pyle, and D.P. Remsen, *Names are key to the big new biology.* Trends Ecol Evol, 2010. **25**(12): p. 686-91.
49. Boyle, B., N. Hopkins, Z. Lu, J.A.R. Garay, D. Mozzherin, T. Rees, N. Matasci, M.L. Narro, W.H. Piel, S.J. Mckay, S. Lowry, C. Freeland, R.K. Peet, and B.J. Enquist, *The taxonomic name resolution service: an online tool for automated standardization of plant names.* BMC Bioinformatics 2013. **14** (16).
50. Piel, W., L. Chan, M. Dominus, J. Ruan, R. Vos, and V. Tannen. *TreeBASE v. 2: A Database of Phylogenetic Knowledge*. in *e-BioSphere 2009*. 2009. London.
51. Maddison, D., K.-S. Schulz, and W. Maddison, *The Tree of Life Web Project.* Zootaxa, 2007(1668): p. 19-40.
52. Rao, J. and X. Su, *A Survey of Automated Web Service Composition Methods*, in *Semantic Web Services and Web Process Composition*. 2004, Springer Verlag.

53. Pejman, E., Y. Rastegari, P. Esfahani, and A. Salajegheh, *Web Service Composition Methods: A Survey*, in *International Multiconference of Engineers and Computer Scientists*. 2012.
54. HIP working group. *Group Photo of Hackathon Participants (Tucson, 2013)*. 2013; Available from: http://www.evoio.org/wiki/File:Phylotastic_tucson_all_jazz.JPG.
55. Hug, S., J. Sandry, R. Vordermann, E. Pontelli, and B. Wright, *DISSECT: integrating computational thinking in the traditional K-12 curricula through collaborative teaching (abstract only)*, in *Proceeding of the 44th ACM technical symposium on Computer science education*. 2013, ACM: Denver, Colorado, USA. p. 742-742.
56. Sharifi, H., G. Rahnavard, and E. Pontelli, *An Ontology-Based Computational Thinking Framework*. Society for Information Technology & Teacher Education International Conference 2012, ed. P. Resta. 2012, Austin, Texas, USA: AACE. 39-46.
57. Stewart, C.A., G.T. Almes, and B.C. Wheeler, *Cyberinfrastructure Software Sustainability and Reusability: Report from an NSF-funded workshop*. 2010, Indiana University: Bloomington, IN.
58. Maddison, W. and D.R. Maddison. *Mesquite: a modular system for evolutionary analysis. Version 2.73.* 2010; Available from: http://mesquiteproject.org.
59. Wright, I.J., P.B. Reich, M. Westoby, D.D. Ackerly, Z. Baruch, F. Bongers, J. Cavender-Bares, T. Chapin, J.H. Cornelissen, M. Diemer, J. Flexas, E. Garnier, P.K. Groom, J. Gulias, K. Hikosaka, B.B. Lamont, T. Lee, W. Lee, C. Lusk, J.J. Midgley, M.L. Navas, U. Niinemets, J. Oleksyn, N. Osada, H. Poorter, P. Poot, L. Prior, V.I. Pyankov, C. Roumet, S.C. Thomas, M.G. Tjoelker, E.J. Veneklaas, and R. Villar, *The worldwide leaf economics spectrum.* Nature, 2004. **428**(6985): p. 821-7.
60. O'Leary, M.A., J.I. Bloch, J.J. Flynn, T.J. Gaudin, A. Giallombardo, N.P. Giannini, S.L. Goldberg, B.P. Kraatz, Z.X. Luo, J. Meng, X. Ni, M.J. Novacek, F.A. Perini, Z.S. Randall, G.W. Rougier, E.J. Sargis, M.T. Silcox, N.B. Simmons, M. Spaulding, P.M. Velazco, M. Weksler, J.R. Wible, and A.L. Cirranello, *The placental mammal ancestor and the post-K-Pg radiation of placentals.* Science, 2013. **339**(6120): p. 662-7.
61. Burleigh, J.G., K. Alphonse, A.J. Alverson, H.M. Bik, C. Blank, A.L. Cirranello, H. Cui, M. Daly, T.G. Dietterich, G. Gasparich, J. Irvine, M. Julius, S. Kaufman, E. Law, J. Liu, L. Moore, M.A. O'Leary, M. Passarotti, S. Ranade, N.B. Simmons, D.W. Stevenson, R.W. Thacker, E.C. Theriot, S. Todorovic, P.M. Velazco, R.L. Walls, J.M. Wolfe, and M. Yu, *Next-generation phenomics for the Tree of Life.* PLoS currents, 2013. **5**.
62. Robinson, I. and J. Webber, *Graph Databases*. 2013: O'Reilly Media.
63. Chisham, B., B. Wright, T. Le, T.C. Son, and E. Pontelli, *CDAO-Store: Ontology-driven Data Integration for Phylogenetic Analysis.* BMC Bioinformatics, 2011. **12**: p. 98.
64. Davies, T.J., T.G. Barraclough, M.W. Chase, P.S. Soltis, D.E. Soltis, and V. Savolainen, *Darwin's abominable mystery: Insights from a supertree of the angiosperms.* Proceedings of the National Academy of Sciences of the United States of America, 2004. **101**(7): p. 1904-9.
65. Federhen, S., *The NCBI Taxonomy database.* Nucleic acids research, 2012. **40**(Database issue): p. D136-43.
66. Goloboff, P.A., S.A. Catalano, J. Marcos Mirande, C.A. Szumik, J. Salvador Arias, M. Källersjö, and J.S. Farris, *Phylogenetic analysis of 73 060 taxa corroborates major eukaryotic groups.* Cladistics, 2009. **25**(3): p. 211-230.
67. Jetz, W., G.H. Thomas, J.B. Joy, K. Hartmann, and A.O. Mooers, *The global diversity of birds in space and time.* Nature, 2012. **491**(7424): p. 444-8.
68. Peters, R.S., B. Meyer, L. Krogmann, J. Borner, K. Meusemann, K. Schutte, O. Niehuis, and B. Misof, *The taming of an impossible child: a standardized all-in approach to the phylogeny of Hymenoptera using public database sequences.* BMC biology, 2011. **9**: p. 55.

69. Stevens, P.F. *Angiosperm Phylogeny Website*. 2008 [cited 2011 August 29]; Version 9, June 2008:[Available from: http://www.mobot.org/MOBOT/research/APweb/.
70. Wu, D., P. Hugenholtz, K. Mavromatis, R. Pukall, E. Dalin, N.N. Ivanova, V. Kunin, L. Goodwin, M. Wu, B.J. Tindall, S.D. Hooper, A. Pati, A. Lykidis, S. Spring, I.J. Anderson, P. D'Haeseleer, A. Zemla, M. Singer, A. Lapidus, M. Nolan, A. Copeland, C. Han, F. Chen, J.F. Cheng, S. Lucas, C. Kerfeld, E. Lang, S. Gronow, P. Chain, D. Bruce, E.M. Rubin, N.C. Kyrpides, H.P. Klenk, and J.A. Eisen, *A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea.* Nature, 2009. **462**(7276): p. 1056-60.
71. Yarza, P., M. Richter, J. Peplies, J. Euzeby, R. Amann, K.H. Schleifer, W. Ludwig, F.O. Glockner, and R. Rossello-Mora, *The All-Species Living Tree project: a 16S rRNA-based phylogenetic tree of all sequenced type strains.* Systematic and applied microbiology, 2008. **31**(4): p. 241-50.
72. Sanderson, M.J., *A nonparametric approach to estimating divergence times in the absence of rate constancy.* Mol Biol Evol, 1997. **14**: p. 1218–1231.
73. Sanderson, M.J., *Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach.* Mol Biol Evol, 2002. **19**(1): p. 101-9.
74. Drummond, A.J. and A. Rambaut, *BEAST: Bayesian evolutionary analysis by sampling trees.* BMC evolutionary biology, 2007. **7**: p. 214.
75. Eastman, J.M., L.J. Harmon, and D.C. Tank, *Congruification: support for time scaling large phylogenetic trees.* Methods in Ecology and Evolution, 2013. **4**(7): p. 688-691.
76. Rodriguez, A., *RESTful Web Services: The Basics*. 2008, IBM.
77. UDDI. *Online Community for the Universal Description, Discovery, and Integration OASIS Standard*. 2013 August, 2013]; Available from: http://uddi.xml.org.
78. Novell. *Novell Nsure UDDI Server*. 2013; Available from: www.novell.com/developer/ndk/uddiserver.html.
79. Apache.org. *JUDDI - An Open Source Implementation of the Universal Description, Discovery, and Integration Specification*. 2013 August, 2013]; Available from: http://juddi.apache.org.
80. World Wide Web Consortium. *Web Services Description Language (WSDL) Version 2.0*. 2007 August, 2013]; Available from: http://www.w3.org/TR/wsdl20-primer.
81. Martin, D., M. Burstein, J. Hobbs, O. Lassila, D. McDermott, S. McIlraith, S. Narayan, M. Paolucci, B. Parsia, T. Payne, E. Sirin, N. Srinivasan, and K. Sycara. *OWL-S: Semantic Markup for Web Services*. 2004 August, 2013]; Available from: http://www.w3.org/Submission/OWL-S/.
82. Cellinese, N., K. Cranston, H. Lapp, S. MacKay, E. Pontelli, and A. Stoltzfus. *First Phyloinformatics VoCamp*. 2009; Available from: http://www.evoio.org/wiki/VoCamp1.
83. Lapp, H. and R. Vos, *PhyloWS: Phyloinformatics Web Service API*. 2009, National Evolutionary Synthesis Center.
84. Gonzalez, G., J.C. Uribe, L. Tari, C. Baral, and C. Brophy, *Mining Gene-Disease Relationships from Biomedical Literature: Weighting Protein-Protein Interactions and Connectivity Measures*, in *Pacific Symposium in Bioinformatics*. 2007.
85. Baral, C., J. Dzifcak, and T.C. Son, *Using Answer Set Programming and Lambda Calculus to Characterize Natural Language Sentences with Normatives and Exceptions*, in *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence*. 2008, AAAI Press. p. 818-823.
86. Shinyama, Y. and S. Sekine, *Preemptive information extraction using unrestricted relation discovery*, in *Proceedings of HLT/NAACL*. 2006.
87. Balduccini, M. and M. Gelfond, *Diagnostic Reasoning with A-Prolog.* Theory and Practice of Logic Programming, 2003. **3**(4,5): p. 425–461.
88. Baral, C., S. McIlraith, and T.C. Son, *Formulating diagnostic problem solving using an action language with narratives and sensing*, in *Proceedings of the Seventh International*

Conference on Principles of Knowledge and Representation and Reasoning (KR'2000). 2000. p. 311-322.
89. Reiter, R., *KNOWLEDGE IN ACTION: Logical Foundations for Describing and Implementing Dynamical Systems*. 2001: MIT Press.
90. McIlraith, S. and T.C. Son, *Adapting Golog for Composition of Semantic Web Services*, in *Proceedings of the Eighth International Conference on Principles of Knowledge Representation and Reasoning (KR'2002)*. 2002, Morgan Kaufmann Publisher. p. 482–493.
91. Pontelli, E. and T.C. Son, *Developing Agents for Bioinformatics Applications: A Preliminary Design*, in *International Conference on Distributed Processing Techniques and Applications*. 2003.
92. Pan, Y., P.H. Tu, E. Pontelli, and T.C. Son, *Construction of an Agent-Based Framework for Evolutionary Biology: A Progress Report*, in *Declarative Agent Languages and Technologies II, Second International Workshop, DALT 2004, New York, NY, USA, July 19, 2004, Revised Selected Papers*, J.A. Leite, A. Omicini, P. Torroni, and P. Yolum, Editors. 2005, Springer. p. 92-111.
93. Gil, Y., E. Deelman, J. Blythe, C. Kesselman, and H. Tangmunanrunkit, *Artificial intelligence and grids: Workflow planning and beyond.* IEEE Intelligent Systems, 2004. **19**.
94. Koehler, J. and B. Srivastava, *Web Service Composition: Current Solutions and Open Problems*, in *ICAPS Workshop on Planning for Web Services*. 2003.
95. Zhao, W., K. Bhattacharya, B. Byant, F. Cao, and R. Hauser, *Transforming Business Process Models: Enabling Programming at a Higher Level*, in *IEEE International Conference on Services Computing*. 2005.
96. Pan, Y., E. Pontelli, and T.C. Son, *Bsis: An experiment in automating bioinformatics tasks through intelligent workflow construction.*, in *In Semantic e-Science*. 2010, Springer Verlag. p. 189-238.
97. Son, T. and E. Pontelli, *Planning with preferences using logic programming.* Theory and Practice of Logic Programming, 2006. **6**: p. 559-607.
98. Nguyen, H.-K., D.-V. Tran, T.C. Son, and E. Pontelli, *On Improving Conformant Planners by Analyzing Domain-Structures*, in *AAAI*. 2011.
99. To, S.T., T.C. Son, and E. Pontelli, *On the Effectiveness of Belief State Representation in Contingent Planning*, in *AAAI*. 2011.
100. Tran, V., K. Nguyen, T.C. Son, and E. Pontelli, *A conformant planner based on approximation: CpA(H).* ACM TIST, 2013. **4**(2): p. 36.
101. Zmasek, C. *Archaeopteryx: Visualization, Analysis, and Editing of Phylogenetic Trees*. 2012 [cited 2012 11/20/2012]; Available from: http://www.phylosoft.org/archaeopteryx/.
102. McCandlish, D.M. and A. Stoltzfus, *Modeling Evolution using the Probability of Fixation: History and Implications.* Quarterly Review of Biology, 2014. **in press**.
103. Milkman, R. and A. Stoltzfus, *Molecular evolution of the Escherichia coli chromosome. II. Clonal segments.* Genetics, 1988. **120**(2): p. 359-66.
104. Stoltzfus, A., *On the possibility of constructive neutral evolution.* J Mol Evol, 1999. **49**(2): p. 169-81.
105. Stoltzfus, A., *Mutation-Biased Adaptation in a Protein NK Model.* Mol Biol Evol, 2006. **23**(10): p. 1852-1862.
106. Stoltzfus, A., *Constructive neutral evolution: exploring evolutionary theory's curious disconnect.* Biol Direct, 2012. **7**(1): p. 35.
107. Yampolsky, L.Y. and A. Stoltzfus, *Bias in the introduction of variation as an orienting factor in evolution.* Evol Dev, 2001. **3**(2): p. 73-83.

108. Stoltzfus, A., J.F. Leslie, and R. Milkman, *Molecular evolution of the Escherichia coli chromosome. I. Analysis of structure and natural variation in a previously uncharacterized region between trp and tonB.* Genetics, 1988. **120**(2): p. 345-58.
109. Gopalan, V., W.G. Qiu, M.Z. Chen, and A. Stoltzfus, *Nexplorer: phylogeny-based exploration of sequence family data.* Bioinformatics, 2006. **22**(1): p. 120-121.
110. Stoltzfus, A., D. Spencer, and W.F. Doolittle, *Methods for Evaluating Exon-Protein Correspondences.* CABIOS, 1995. **11**(5): p. 509-515.
111. Stoltzfus, A., B. O'Meara, J. Whitacre, R. Mounce, E.L. Gillespie, S. Kumar, D.F. Rosauer, and R.A. Vos, *Sharing and re-use of phylogenetic trees (and associated data) to facilitate synthesis.* BMC Research Notes, 2012. **5**: p. 574.
112. Palu, A.D., E. Pontelli, and D. Ranjan, *Sequential and Parallel Algorithms for the NCA Problem on Pure Pointer Machines.* Theoretical Computer Science, 2006. **352**: p. 108-135.
113. Dovier, A., A. Formisano, and E. Pontelli, *Autonomous Agent Coordination: Action Languages meet CLP(FD) and Linda.* Theory and Practice of Logic Programming, 2013. **13**(2): p. 149-173.
114. O'Meara, B.C., *Evolutionary Inferences from Phylogenies: A Review of Methods.* Annual Review of Ecology, Evolution, and Systematics, 2012. **43**(1): p. 267-285.
115. Open Source Initiative. *The BSD 3-Clause License*. 2013 August 2013]; Available from: http://opensource.org/licenses/BSD-3-Clause.

Arlin Stoltzfus, Ph.D.

A. 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

B. Appointments

2006-present Associate Professor, Institute for Bioscience and Biotechnology Research (formerly CARB), University of Maryland and the National Institute of Standards and Technology, Rockville, MD

1999-2006 Assistant Professor, CARB, University of Maryland and the National Institute for Standards and Technology, Rockville, MD

C. Selected products

Five publications closely related to the proposed project

1. Gopalan, V., Qiu, W.G., Chen, M.Z., and **Stoltzfus, A**. 2006. Nexplorer: Phylogeny-based exploration of sequence family data. *Bioinformatics, 22:120-121*.
2. 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.
3. 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.
4. **Stoltzfus A**, O'Meara B, Whitacre J, Mounce R, Gillespie EL and others. 2012. Sharing and Re-use of Phylogenetic Trees (and associated data) to Facilitate Synthesis. *BMC Research Notes* 5: 574.
5. **Stoltzfus A**, Lapp H, Matasci N, Deus H, Sidlauskas B and others. 2013. Phylotastic! Making tree-of-life knowledge accessible, reusable and convenient. *BMC bioinformatics* **14**:158.

Five other publications

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. 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.
4. **Stoltzfus, A** and Yampolsky, Y. 2009. Climbing Mount Probable: Mutation as a cause of non-randomness in evolution. *J. Heredity* **100** (5): 637-47.
5. McCandlish, D. and **Stoltzfus, A**. 2014. Modeling evolution using the probability of fixation: history and implications. *Quart. Rev. Biol.* , in press.

Synergistic Activities

Co-leader of the NESCent HIP (Hackathons, Phylogeny, Informatics) working group, 2011 to present, focusing on the Phylotastic project to deliver tree-of-life knowledge

Co-leader of the NESCent Evolutionary Informatics working group, 2006 to 2009, focusing on improving interoperability through standards and technology (evoinfo.nescent.org).

Co-organizer of 6 NESCent hackathons from 2006 to 2014, empowering early-career

researchers with the skills and connections to improve interoperability.
Developer and project leader, Bio::NEXUS, an open-source Perl API for the NEXUS file format.
Developer and project leader, 2007 to 2012, Comparative Data Analysis Ontology (CDAO, www.evolutionaryontology.org)

D. Collaborations and Other Affiliations

Collaborators: Michael E. Alfaro (UCLA), James P. Balhoff (National Evolutionary Synthesis Center), Holly M. Bik (UC Davis), Brian O'Meara (Department of Ecology & Evolutionary Biology, Joseph W. Brown (Institute for Bioinformatics and Evolutionary Studies (IBEST), University of Idaho), Jason A Caravas (Wayne State University), Brandon Chisham (New Mexico State University), Karen Cranston (National Evolutionary Synthesis Center), Helena Deus (National University of Ireland), Emily L. Gillespie (Marshall University), Luke J. Harmon (University of Idaho), Tracy A. Heath (UC Berkeley), Mark T Holder (University of Kansas), Greg Jordan (Paperpile), Sudhir Kumar (Arizona State University), Hilmar Lapp (National Evolutionary Synthesis Center), Jim Leebens-Mack (University of Georgia), Naim Matasci (University of Arizona), David McCandlish (University of Pennsylvania), Emily Jane McTavish (University of Texas at Austin), Peter E. Midford (National Evolutionary Synthesis Center), Siavash Mirarab (University of Texas at Austin), John Moult (University of Maryland), Ross Mounce (University of Bath, UK), Ryan W. Norris (Lock Haven Univ.), Maryam Panahiazar (University of Georgia), Matthew W. Pennell (University of Idaho), Megan Pirrung (University of Colorado Denver), Enrico Pontelli (New Mexico State University), Anurag Priyam (Indian Institute of Technology Kharagpur, India), Ajith Ranabahu (Wright State University), Dan F. Rosauer (Yale), Michael S. Rosenberg (Arizona State University), Amit P. Sheth (Wright State University), Brian Sidlauskas (Oregon State University), Aaron Steele (U.C. Berkeley), Cory L. Strope (North Carolina State University), Jeet Sukumaran (University of Kansas), Gaurav Vaidya (University of Colorado Boulder), Rutger Vos (NCB Naturalis), Campbell O. Webb (Harvard), Mark Westneat (Field Museum of Natural History), Jamie Whitacre (NMNH), Xuhua Xia (University of Ottawa, Canada), Guoqin Yu (National Cancer Institute), Christian M. Zmasek (Sanford-Burnham Medical Research Institute).

Graduate and Post-Doctoral Advisors: W. Ford Doolittle (Dalhousie University, semi-retired), Roger Milkman (deceased).

Thesis Advisor and Post-graduate-Scholar Sponsor: Weigang Qiu (Assoc Prof, Hunter College, CUNY), Chengzhi Liang (CSHL), Danny DeKee (no scientific affiliation), Vivek Gopalan (Lockheed Martin NIAID Bioinformatics Support), Guoqin Yu (NCI), Ryan Norris (Asst Prof, Lock Haven).

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

- Interim Associate Dean, College of Arts & Sciences, NMSU, 2014-present.
- Department Head, Computer Science Dept., NMSU, 2/2009-6/2014.
- Director, NMSU Center of Research Excellence in Computational Biology, 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 Products

1. C. Sakama, T. Son, **E. Pontelli**. "A Logical Formulation for Negotiation among Dishonest Agents." International Joint Conference on Artificial Intelligence (IJCAI), pp. 1069-1074, 2011.
2. Y. Pan, **E. Pontelli**, S. Tran. "BSIS: An Experiment in Automating Bioinformatics Tasks Through Intelligent Workflow Construction", In Semantic e-Science, Springer Verlag, pp. 189-238, 2010.
3. **E. Pontelli,** D. Bevan, M. Chapman, J. He, J. MacCuish, N. MacCuish, D. Moreland, J. Pinto, X. Qin. "BIOPS Interactive, An E-Learning Platform Focused on Protein Structure and DNA", BioScene, 35(2), 2009.
4. K. Nguyen, V. Tran, T. Son, **E. Pontelli**. "A Conformant Planner Based On Approximation: CpA(H)", ACM Transactions on Intelligent Systems and Technology, 4(2), 2013.
5. B. Chisham, B. Wright, T. Le, T. Son, **E. Pontelli**. "CDAO-Store: Ontology-driven Data Integration for Phylogenetic Analysis," BMC Bioinformatics, 12:98, 2011.
6. C. Baral, G. Gelfond, **E. Pontelli**, T. Son. "Logic programming for finding models in the logics of knowledge and its applications", *Theory and Practice of Logic Programming*, 10(4-6), 2010.
7. B. Chisham, F. Prosdocimi, **E. Pontelli**, A. Stoltzfus, J. Thompson. "Initial Implementation of a Comparative Data Analysis Ontology", Evolutionary Bioinformatics, 5:47-66, 2009.
8. P. Tu, **E. Pontelli,** T. Son, S. To. "Applications of Parallel Processing Technologies in Heuristic Search Planning", *Concurrency and Computation*, 21(15):1928-1960, 2009.
9. F. Fioretto and **E. Pontelli**. "Constraint Programming in Community-based Gene Regulatory Network Inference", International Conference on Computational Methods in Systems Biology, Springer Verlag, 2013.
10. F. Campeotto, A. Dal Palù, A. Dovier, F. Fioretto, **E. Pontelli**. "A Constraint Solver for Flexible Protein Model." J. Artif. Intell. Res. (JAIR), 48: 953-1000, 2013.

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 serving as the director of the DISSECT (DIScovering SciencE through Computational Thinking) project, funded by the NSF GK-12 program, aimed at exploring infusion of computational thinking in the science middle school curricula, through graduate students presence in science classes and teachers training. I serve as member of the steering committee of the Computing Alliance of Hispanic Serving Institutions (CAHSI).
- I am a member of the of the Leadership team for the Engaging Hispanic/Latino Youth Collaboration, part of the Computer Science Collaborative Project (CSCP). I am the co-leader of

the New Mexico chapter of the National Girls Collaborative Project (NGCP). I have co-organized several New Mexico NGCP events. I have also co-chaired the 2012 New Mexico Celebration of Women in Computing Conference.

- I have published over 250 peer-reviewed publications in the areas of logic and constraint programming, computer science education, knowledge representation and reasoning, assistive technologies, parallel processing, and bioinformatics. I have served as Program Chair of the 2008 International Conference on Logic Programming and I have served as the PC 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). I am a member of the editorial board of the Artificial Intelligence Research (AIR) 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 been deeply involved with research and outreach activities in the field of assistive technologies and people with disabilities. I have served as the Vice-Chair of the ACM Special Interest Group on Accessible Computing (SIGACCESS), dealing with all aspects of computing for individuals with disabilities. I have chaired/co-chaired several Doctoral Consortia associated to the ASSETS conference (Computers and Accessibility) and the ICCHP conference (Computers and Handicaps). I am member of the editorial board of the ACM Transactions in Accessible Computing journal.
- I have created and directed for the first three years the Doctoral Consortium program associated to the International Conference on Logic Programming. I have organized and directed the 3rd International Summer School on Computational Logic, funded by CRAW/CDC, which took place in Las Cruces in July 2008.

Awards and Honors

- 2014, Regents Professor, New Mexico State University
- 2012, Manasse Scholar, New Mexico State University
- 2010, Best Paper Award, International Conference on Logic Programming
- 2008, First Prize, Non-deterministic Track, International Planning Competition
- 2006 and 2011 College of 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, 2002–2008.
- Department of Education, NIDDR program, 2000-2005.
- Department of Education, Graduate Assistants in Areas of National Need, 2003–2006.
- NSF, Research to Aid Persons with Disabilities, 2008–2011.
- NSF, Career Award, 1999–2004.

Recent Collaborators

Faculty: T. Son (NMSU), A. Stoltzfus (NIST), J. Thompson (U. Strasbourg), A. Dovier (U. of Udine), A. Dal Palu (U. Parma), C. Baral (Arizona State U.), M. Truszczynski (U. Kentucky), C. Sakama (Wakayama U.), A. Formisano (U. Perugia), I. Lee (Santa Fe Institute).

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. Wright, N. Alajarmeh (2014), F. Campeotto, F. Fioretto.

MS Students: 25 MS students graduated since 2006

Graduate Advisors: G. Gupta (U.T. Dallas), L. Slothouber (U. Houston)

Biographical Sketch
Brian C. O'Meara

Department of Ecology and Evolution
University of Tennessee
Knoxville, Tennessee

Phone: 865-408-8733
Fax: 865-974-3067
email: bomeara@utk.edu

Professional Preparation

Harvard University	Biology	B.A., 2001
University of California, Davis	Population Biology	Ph.D., 2008

Appointments

August 2009	Assistant Professor, Department of Ecology and Evolution University of Tennessee, Knoxville TN
Nov. 2007-2009	Postdoctoral Fellow, National Evolutionary Synthesis Center

Five Most Relevant Products

Zanne, A. E., D. C. Tank, W. K. Cornwell, J. M. Eastman, S. A. Smith, R. G. FitzJohn, D. J. McGlinn, **B. C. O'Meara**, A. T. Moles, P. B. Reich, D. L. Royer, D. E. Soltis, P. F. Stevens, M. Westoby, I. J. Wright, L. Aarssen, R. I. Bertin, A. Calaminus, R. Govaerts, F. Hemmings, M. R. Leishman, J. Oleksyn, P. S. Soltis, N. G. Swenson, L. Warman, and J. M. Beaulieu. 2014. Three keys to the radiation of angiosperms into freezing environments. Nature. 506(7486): 89-92

O'Meara, B.C. 2012. Evolutionary inferences from phylogenies: A review of methods. Annual Review of Ecology, Evolution, and Systematics. 43(1)

Beaulieu, J. M., D.-C. Jhwueng, C. Boettiger, and **B. C. O'Meara**. 2012. Modeling Stabilizing Selection: Expanding The Ornstein-Uhlenbeck Model Of Adaptive Evolution. Evolution 66: 2369-2383.

Smith, S.A., and **B.C. O'Meara**. 2009. Morphogenera, monophyly, and macroevolution. PNAS: 106:E97-E98

O'Meara, B.C., C. Ané, M.J. Sanderson, and P.C. Wainwright. 2006. Testing for different rates of continuous trait evolution using likelihood. Evolution 60: 922-933.

Up to Five Other Products

Stoltzfus, A., H. Lapp, N. Matasci, H. Deus, B. Sidlauskas, C. Zmasek, G. Vaidya, E. Pontelli, K. Cranston, R. Vos, C. Webb, L. Harmon, M. Pirrung, **B. O'Meara**, M. Pennell, S. Mirarab, M. Rosenberg, J. Balhoff, H. Bik, T. Heath, P. Midford, J. Brown, E. McTavish, J. Sukumaran, M. Westneat, M. Alfaro, A. Steele, and G. Jordan. 2013. Phylotastic! Making tree-of-life knowledge accessible, reusable and convenient. BMC Bioinformatics 14:1-17

Stotzfus, A, **B.C. O'Meara,** J. Whitacre, R. Mounce, E. Gillespie, S. Kumar, D. Rosauer, R. Vos. 2012. Sharing and re-use of phylogenetic trees (and associated data) to facilitate synthesis. BMC Research Notes 5(1): 574

Smith, S.A. and **B.C. O'Meara**. 2012. "treePL: Divergence time estimation using penalized likelihood for large phylogenies. Bioinformatics.

O'Meara, B. C. 2010. New Heuristic Methods for Joint Species Delimitation and Species Tree Inference. Systematic Biology 59:59-73.

Driskell, A. C., C. Ané, J. G. Burleigh, M. M. McMahon, **B. C. O'Meara**, and M. J. Sanderson. 2004. Phylogenetic utility of large sequence databases for building the tree of life. Science 306: 1172-1174

Synergistic Activities

1. Darwin Day Tennessee advisor
2. Curator of Phylogenetics task view for R
3. Organizer in Women in Science seminar series at UT Knoxville
4. Organizer of multiple hackathons
5. Developer of multiple software packages in R

Collaborators: Lonnie Aarssen (Queen's University (Ontario, Canada)), Robert Bertin (College of the Holy Cross), Carl Boettiger (UC Davis), Gordon Burghardt (U Tennessee, Knoxville), Gordon Burleigh (U Florida), Andre Calaminus (U Florida), Bryan Carstens (Ohio State U), David Collar (UC Santa Cruz), William Cornwell (U New South Wales), Karen Cranston (NESCent), Rafael de Casas (Duke), Helena Deus (DERI), Pam Diggle (U Colorado), Michael Donoghue (Yale), Jonathan Eastman (U Washington), Charlie Fenster (U Maryland), Richard FitzJohn (Macquarie), Damian Gessler (iPlant), Michael Gilchrist (U Tennessee, Knoxville), Emily Gillespie (Marshall U), Stephen Goff (iPlant), Rafael Govaerts (Royal Botanic Gardens, Kew), Matthew Hanlon (TACC), Luke Harmon (U Washington), Frank Hemmings (U New South Wales), Darrin Hulsey (U Tennessee, Knoxville), Benjamin Keck (U Tennessee, Knoxville), Sudhir Kumar (Arizona State), Hilmar Lapp (NESCent), Maribeth Latvis (U Florida), Michelle Leishman (Macquarie U), Andrew Lenards (iPlant), Jonathan Losos (Harvard), Eric Lyons (iPlant), Naim Matasci (iPlant), Evgeny Mavrodiev (U Florida), Daniel McGlinn (Utah State), Sheldon McKay (iPlant), Andrew Moffatt (U Tennessee, Knoxville), Angela Moles (U New South Wales), Mark Mort (U Kansas), Ross Mounce (U of Bath), Thomas Near (Yale), Jacek Oleksyn (U of Minnesota), Enrico Pontelli (New Mexico State), Peter Reich (U of Minnesota), Dan Rosauer (Australian National U), Dana Royer (Wesleyan University), James Schulte (Clarkson), Brian Sidlauskas (Oregon State), Stephen Smith (U Michigan), Stacey Smith (U Colorado), Douglas Soltis (U Florida), Pamela Soltis (U Florida), Conrad Stack (State College PA), Ann Stapleton (iPlant), Peter Stevens (U of MIssouri), Arlin Stoltzfus (NIST), Nathan Swenson (Michigan State U), David Tank (U Washington), Gaurav Vaidya (U Colorado Boulder), Matthew Vaughn (iPlant), Rutger Vos (NCB Naturalis), Peter Wainwright (UC Davis), Liya Wang (iPlant), Laura Warman (USDA Forest Service), Mark Westoby (Macquarie U), Jamie Whitacre (Smithsonian), Joseph Williams (U Tennessee, Knoxville), Ian Wright (Macquarie U), Amy Zanne (George Washington), Christian Zmasek (Sanford-Burnham Medical Research Institute)

Graduate Advisors and Postdoctoral Sponsors: Michael Sanderson (U Arizona), Michael Turelli (UC Davis), Phil Ward (UC Davis), Todd Vision (U North Carolina)

Thesis advisor (2) and Postgraduate-Scholar Sponsor (9)
Hugo Alamillo (North Seattle Community College), Barb Banbury (U Washington), Jeremy Beaulieu (NIMBioS), Jenn Bosco (U Tennessee, Knoxville), JJ Chai (Oak Ridge National Lab), Nathan Jackson (U Tennessee, Knoxville), Tony Jhwueng (Feng-Chia University), Michelle Lawing (Texas A&M), Ryan Martin (NIMBioS), Katie Massana (U Tennessee, Knoxville), Nick Matzke (NIMBioS)

Facilities, Equipment, and Other Resources (UMD)

Computer: Desktop computers and printers are available. UMD maintains a computer network available to all faculty and staff. Faculty and staff have access to a remote application server to run powerful programs like CLC Genomics Workbench, Matlab, and Topspin. This server doubles as a head node for a cluster of 204 processors, all running CentOS.

Office: There is an office for the PI, and cubicles for students and post-docs.

Other Resources: Dr. Arlin Stoltzfus, PI, will oversee all scientific and technical aspects of the proposed project. In addition, Dr. Stoltzfus will provide guidance to personnel participating in the project, and will participate in data analysis and software development. *All travel required for the PI will be covered outside of this grant.*

Facilities, Equipment, and Other Resources
New Mexico State University

Laboratory

The activities will be developed within two NMSU laboratories. The CREST laboratory is an ample space (that includes one student lab that can host 9 students and 2 offices reserved for doctoral students and visiting researchers) located on the second floor of Science Hall. The space has been committed by NMSU to the needs of CS research in computational biology and bioinformatics.

The second laboratory available to this project is the Knowledge representation, Logic, and Advanced Programming (KLAP) laboratory. Created in 2004, the laboratory supports both research projects in the areas of AI and bioinformatics as well as research activities in the areas of underrepresentation in computing and CS education.

Computers

The two laboratories are fully equipped with workstations (Linux, Windows, and MacOS) for all students and staff members associated to the project. KLAP provides access to a high-performance server with large storage capabilities to serve as archive and dissemination portal. KLAP also maintains two Beowulf clusters and a new hybrid platform for GPU computing. These computing facilities will be fully available to the needs of this project.

Offices

The NMSU department provides ample office space for the students and researchers involved in this project. Undergraduate and Graduate assistants will have seats in offices with dedicated workstations. The investigator has adequate office space in his home departments.

Other Resources

The NMSU CS department will provide the technical support for the computing infrastructure used by the project (e.g., backups, security, maintenance). The NMSU-CS department will provide the administrative support to manage the operations of the grant.

Facilities at University of Tennessee, Knoxville

The O'Meara lab maintains two servers used to host websites and databases, including the current implementation of DateLife. We also have automated backup to offsite storage. We have funds to use cloud hosting for DateLife as needed for scaling. There is office space needed to house the postdoc, including a Mac Pro computer with multiple monitors.

Data Management and Sharing Plan

Types of Data. This data management plan will be applied to all products of this this project – thus, the term “data” will be used to denote any of the outcomes of the proposed activities. We classify the products of the project in two categories: research and outreach products.

The **research products** can be summarized in the following major classes: (1) *Documents:* scientific papers and reports will document the research accomplishments, for distribution to the scientific community. Scientific papers will be prepared for presentation in national and international conferences and scientific journals. Papers and reports will document both the proposed infrastructure as well as the new scientific results enabled by it (e.g., outcomes of client applications). The reports will also include technical manuals for the tools developed by the research. (2) *Software:* this class of products includes all the software components developed; software products will materialize in the form of independent software tools (downloadable and installable, all with OSI-approved licenses), network-accessible web services, and formalized executable workflows. (3) *Scientific Data:* this class includes all the field and scientific data, collected using methodologies (e.g., formal data files), produced by the proposed research activities. Scientific data will also include metadata collected or generated by the project. (4) *Repositories:* while the creation of data repositories is not a goal of this project (e.g., we will use OpenTree and other tree-stores) the use of caching techniques to speed-up subtree extraction and other tree transformations will lead to a growing repository of phylogenies, that will be made available for other applications and uses.

The **outreach products** will include training and educational materials associated to the project. This may include: (1) *Online Materials:* including web-based tutorials and manuals, workshops, and online course materials (e.g., course modules) that use the proposed architecture; (2) *Documents:* tutorials, course materials and training manuals. Additionally, the investigators will work with high school and undergraduate instructors to introduce evolutionary-inspired analyses in their courses – by developing novel clients and associated course modules. Additional outreach products will include Web sites providing information about the research project, blogs, and social networks to enable interaction among students, researchers, and practitioners.

Standards for Data and Metadata. The activities are aimed at supporting the dissemination of data in the most inter-operable way. The team has an extensive expertise in the areas of ontologies, data management (e.g., retrieval, indexing), and data inter-operation for biological datasets. While the actual data produced will be encoded using the most popular formats (e.g., NEXUS or NeXML for phylogenies), the expertise in biological information management will facilitate the automated inter-conversion with the majority of commonly used formats for each type of data and the inter-operability among tools used in this effort. Such inter-conversion infrastructure will be made available as a web service, accessible using a programmatic interface as well as a Web portal. The investigators are leading ongoing efforts to standardize formats, ontologies, and APIs related to phyloinformatics (e.g., NeXML, CDAO, MIAPA, PhyloWS); we will use such emerging standards in this project. Standard formats will be employed for the description of educational materials, e.g., using the Shareable Content Object Reference Model to package course materials and enable its interaction with learning management systems (e.g., Blackboard).

Emphasis will be placed in providing metadata describing all products of the project - in the form of instances of formal ontologies. This component is essential to enable the use of cyber-infrastructures as platforms to allow access and dissemination of the products and for the effective execution of workflows and controllers. The use of metadata encoded using ontologies enables unambiguous interpretation of data, querying and search, and the use of automated reasoning techniques to support advanced uses of data and tools (e.g., composition of web services, inter-operation among data sets, composition of educational materials to meet specific pedagogical goals). Several ontologies have been developed for different areas addressed in this project (e.g.,

the Comparative Data Analysis Ontology, several ontologies from the BioPortal, ISO 19115 and FGDC metadata for geospatial data), and they will be used to provide formal descriptions of the products of the applications of the proposed architecture. Existing (e.g., the outcomes of the Minimal Information for Biological and Biomedical Investigations initiative) and novel ontologies will be used to describe experimental protocols and analysis pipelines. The formal description will also apply to the training products; we will use the IEEE Learning Object Metadata and practices from the IMS Learning Resource Metadata initiative.

Policies for Access and Sharing. All the tools, methodologies, and results produced will be made available to the broad community of scientists, students, and practitioners, throughout the project and beyond the ending of the funding period. All the tools and methodologies developed will be packaged for distribution as modules and services, representing complete pieces of knowledge; each module will be motivated by a specific research or educational goal and include all data, tools, and documents. GitHub will be used for the distribution of source code. Specific servers will be set up to serve as service registries, facilitating the community retrieval and use of all web services developed in this project. Data produced by the research and educational activities will be registered/deposited in publicly available repositories (e.g., OpenTree, TreeBase, ESA Data Registry, Dryad). All repositories will be accessible through web portals; the materials in the repositories will be annotated using ontologies, enabling sophisticated querying, retrieval, aggregation and integration across diverse data sets. The web presence (e.g., Google+ community) will also include directories of expertise of the investigators. The portals will be designed using techniques that promote access that is suitable to the most diverse audience—ranging from students seeking basic information to advanced scientists interested in cutting-edge research tools. This will include search forms (built around the terms of the ontologies) and graphical workflow systems. All of the resources will be provided through servers, hosting repositories of documents, data, software tools and web services. At the hardware level, the grant will acquire a dedicated servers, and the participating institutions are committed to the long-term maintenance and operation of such servers. The servers will have strong computing capabilities, large memory, and large data storage capabilities. The access to the infrastructure will be enabled by the large-bandwidth of the existing institutional networks.

Policies for Reuse, Redistribution, and Derivative Products. All the products developed by the proposed activities will be disseminated through open repositories (e.g., GitHub) and made available for reuse, reproduction, and modification with the least amount of restrictions. All the software, methodologies, and data will be distributed using licensing practices that reflect the principles of open source. In particular, we will adopt licensing models based on modified BSD 3-clause license, as reported by the Open Source Initiative organization. Similar licensing approaches will be applied to the educational materials, which will be freely distributed and made available under similar licensing schemes, encouraging modifications and redistribution.

Plans for Archiving and Preservation. NMSU Computer Science (NMSU-CS) will assume the responsibility of long-term archival and preservation of all data, tools and documents created. NMSU-CS will provide the infrastructure to archive all the data maintained in the project servers, and implement security policies to prevent data damage and inappropriate usage of resources. NMSU-CS runs a state-of-the-art automated backup infrastructure, powered by EMC NetWorker; this backup infrastructure has already been expanded, through support provided by existing projects to meet the needs of this new project. NMSU is committed to preserve the data and resources beyond the end of the funding period, ensuring continued accessibility to all the research products. NMSU will support the costs of maintaining the network, the servers, and the backup infrastructure beyond the funding period. Plans will be put in place to address any type of contingency, including the departure of team members and the transfer of preservation and archival duties to a different participating institution. We will also work closely with emerging foundations (e.g., the Phyloinformatics Research Foundation (PRF)).

Postdoctoral Researchers Mentoring Plan

The proposed project will include support for 2 post-doctoral researchers. One of them will be associated with the evaluation & client development team at UMD, and supervised by Dr. Stoltzfus, while one will be associated with the scaling team and supervised by Dr. O'Meara. The researchers will be recruited during the first semester of operation (UMD) or at least one semester before the researcher is supposed to start (UTK).

The project will implement a comprehensive plan to mentor and train the researchers. The training will be implemented partly through direct supervision and partly through formal training components (implemented both face-to-face as well as online). The overarching objective of the mentoring plan is to prepare the researchers to successfully enter the world of academia and research, and develop skills necessary to lead cutting-edge research and education activities in the field of computational biology and phyloinformatics. Mentoring will begin immediately upon hiring with a discussion of this document, and a development plan that will be reviewed and updated quarterly.

The training will be articulated in the following phases:

- *Phase I – Initial Skills Development:* During the initial training period, the researchers will participate in formal training programs that cover the following areas:
    - **Responsible conduct of research:** this will be provided by participating institutions as part of its comprehensive training program for graduate and post-doctoral researchers (provided in part through the Collaborative Institutional Training Initiative - CITI);
    - **Proposal writing:** this will be provided by workshops offered by the institutional research offices or via online courses;
    - **Writing in academia:** this will be provided by the investigators, supplemented by workshops available at UMD and UTK;
    - **Understanding collaborative science:** this training will be offered as a two-day workshop to all graduate students and post-doctoral researchers associated to the project; the workshop will be designed by Drs. Pontelli and Stoltzfus.

- *Phase II – Continual Skills Development:* during their tenure in the program, the researchers will be personally mentored by the relevant faculty member (Dr. Stoltzfus or Dr. O'Meara). Mentoring will take place in weekly meetings and as needed. The mentors will provide advice on aspects like proposal writing (and this will include involving the researchers in actual proposal writing tasks), academic careers, and communication and presentation skills. The researchers will make presentations during group meetings, project-wide research meetings, and at other venues such as professional meetings. The researchers will be encouraged to participate in additional training workshops. The researchers will also receive career advice by the project personnel, help in applying for faculty and research positions, and workshops on how to perform during job interviews.

- *Phase III – Practical Activities:* to allow the mentoring and training to be effective, the post-doctoral researchers will be required to conduct practical activities to refine their training. These will include mentored development of technical papers (including a formal peer-reviewing process by project personnel), participation in the development of grant proposals, and development of a poster or talk. Feedback will be provided by the mentors and by other members of the project.