Title

Evolutionary Informatics: Supporting Interoperability in Evolutionary Analysis

Short Title

Evolutionary Informatics

Project Leaders

Arlin Stoltzfus (arlin.stoltzfus@nist.gov), Research Biologist, NIST; and Fellow, Center for Advanced Research in Biotechnology, 9600 Gudelsky Drive, Rockville, MD 20850

Rutger Vos (rvosa@sfu.ca), Department of Biological Sciences, Simon Fraser University

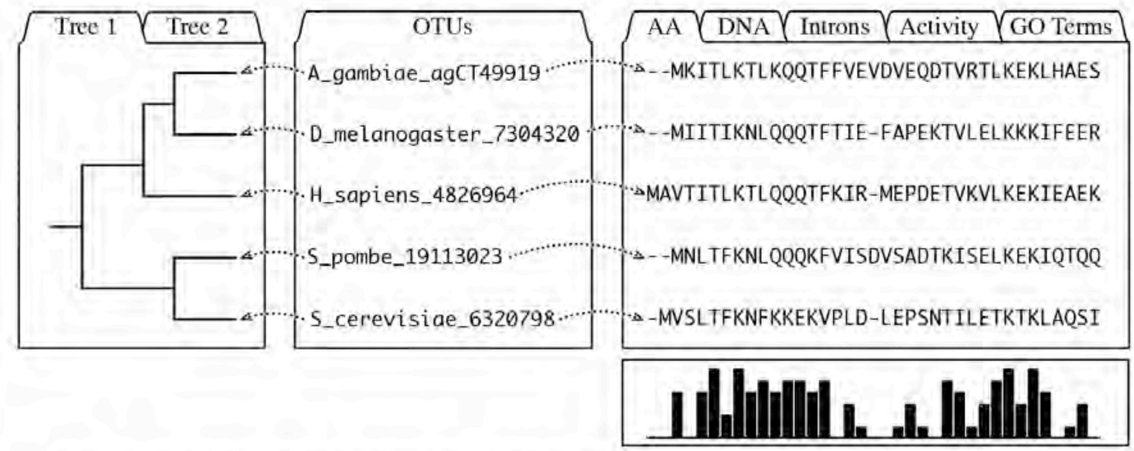
Project Summary

The continuing growth of bioinformatics and genomics presents an opportunity for expanding the application of evolutionary methods, with respect to both the amount and the variety of systematic comparative data. Though evolutionary biologists have developed powerful tools for inferring phylogenies, detecting selection, and so on, integrating evolutionary methodology into workflows in bioinformatics does not depend so much on the power of analysis tools as it does on a well developed informatics infrastructure: software and standards for data exchange, visualization, input-and-output, editing, control, and storage-and-retrieval. We propose a working group to facilitate (directly and indirectly) the development of this infrastructure. Through a series of four meetings, each with presentations, discussion, and actual software development, the working group will build on the foundation provided by current analysis tools and available standards. The working group will gather data sets that represent canonical use-cases; remedy inconsistencies in current NEXUS implementations, and resolve ambiguities; develop translators between NEXUS and other data formats; develop a back-end storage schema; and translate actual content from existing databases of sequence family data into a standard form. The working group will develop and implement revisions to the NEXUS standard (perhaps including an XML replacement), and will lay the ground-work for a new generation of standards by identifying different data models used in different types of evolutionary analyses, and by documenting use-cases that fall beyond the reach of current standards.

Introduction and Goals

Contemporary genomic analysis is an enormous arena of scientific activity that depends heavily on comparative biology— making inferences (e.g., "functional inferences") from similarities and differences in sequences, structures, regulatory networks, expression patterns, and so on. As such, it represents an opportunity to expand the scope of evolutionary analysis into areas of great practical importance. Evolutionary analysis, the natural framework for comparative analysis, differs from heuristic approaches based on machine-learning methods imported from computer science in that the items to be compared are not treated as independent entities, but as related entities that have evolved along paths of common descent (a tree) according to dynamics that reflect evolutionary genetics.

Ideally this framework converts questions about interpreting similarities and differences into theoretically well posed questions about evolutionary transitions in the states of characters, according to a character-state data model, or (here) CDAT for "character data and trees" (Figure). In the CDAT model the observable features of the "**OTU**s" (Operational Taxonomic Units) are the "states" of a set of underlying "characters", where the states may be discrete or continuous. NEXUS (Maddison, et al., 1997), which incorporates the "New Hampshire" tree standard, is the de facto standard file format for the CDAT model, though various formats exist for character data alone (e.g., alignment formats).


Tree 1
Tree 2
OTUs
AA
DNA
Introns
Activity
GO Terms
A_gambiae_agCT49919
D_melanogaster_7304320
H_sapiens_4826964
S_pombe_19113023
S_cerevisiae_6320798
--MKITLKTLKQQTFFVEVDVEQDTVRTLKEKLHAES
--MIITIKNLQQQTFTIE-FAPEKTVLELKKKIFEER
MAVTITLKTLQQQTFKIR-MEPDETVKVLKEKIEAEK
--MNLTFKNLQQQKFVISDVSADTKISELKEKIQTQQ
-MVSLTFKNFKKEKVPLD-LEPSNTILETKTKLAQSI

This powerful framework is used relatively rarely in large-scale or integrative analyses. Even top evolutionary researchers doing such analyses may avoid NEXUS in favor of ad hoc formats, and avoid a full phylogenetic model-based treatment in favor of simplified pairwise comparisons. Partly this reflects the fact that genomics is a new and data-rich discipline in which picking low-hanging fruit using heuristic methods is expedient, with no need to squeeze the most out of every observation (as in the traditionally data-poor field of evolutionary biology). However the main barrier to the wider use of evolutionary approaches in large-scale or integrative analyses is the lack of informatics support: evolutionary analysis is characterized by powerful applications software but weak informatics.

Effectively supporting the kind of automation and interoperability needed for large-scale and integrative analysis is not merely a matter of having a data exchange format: it also requires standards and software for visualization, input-and-output, editing (modifying a data set), program control, and storage-and-retrieval. The NEXUS format is sufficiently expressive for representing a wide variety of analyses, yet is barely known outside of molecular evolution and phylogenetics. Users complain that NEXUS is complex, but this is because, due to the lack of informatics support, every individual user is forced to learn the file format in detail in order to use it effectively. End-users and casual developers should never have to look at the file directly with a text viewer to know what data are in the file, because they should have visualization tools to display it; they should never have to enter or edit values by hand without an error-checking interface; no user should begin an analysis by composing a file on a keyboard, instead software should manage input from (and output to) other formats and data streams; and so on.

Mesquite is a notable step forward in filling out the concept of a generalized interface to evolutionary analysis, e.g., providing support for visualization and editing of data in the CDAT model, an interface to analysis tools, and input in a modified NEXUS format. However, a much broader community-based efforts is needed to propagate this kind of support, and enchance it with a schema for storage, more interfaces to external analysis tools, further data format standardization, rules for editing and other manipulations of data, and so on. Accordingly, this proposal focuses on the following goals:

1. Developing community cohesion and an appreciation of what informatics support entails.
2. Documenting use-cases illustrating data sets, analysis procedures, and interoperability.
3. Enhancing support for current conventions and standards that promote automation and interoperability.
4. Proposing and supporting revisions to current conventions and standards for the near-term.
5. Preparing the groundwork for a more sweeping revision of standards for the long-term.

Proposed Activities

The Evolutionary Informatics working group will carry out six types of activities: presentations, discussions, software development, writing, data-gathering, and dissemination. The major activities of presentations, discussions, and programming will take place during a series of four meetings held at half-year intervals at NESCent. On a voluntary basis, participants will carry out writing, data-gathering and dissemination activities between meetings. Below the relationship of these activities to the 5 goals are described.

Goal #1 is to develop community cohesion and a common appreciation of the nature and scope of informatics support for evolutionary analysis, in terms of software and standards for data exchange, visualization, input-and-output, editing, control, and storage-and-retrieval. All of the group's activities contribute to accomplishing this goal. Demonstration projects will show the benefits of interoperability and automation; the strict focus on informatics support (not scientific results, not algorithms) will make software developers aware of their roles in promoting interoperability and automation.

Goal #2 is to document use-cases of comparative analysis, including examples of interoperability, and of effective methods for visualization, editing, and so on. To develop standards requires a common understanding of the problem based on examples called "use cases". Typical use-cases are inference of a phylogenetic tree from character data, detecting positive selection, and reconstructing ancestral states. Display and editing of data and results are also use-cases, for which MacClade and Mesquite (for example) provide effective examples. HyPhy provides an example of a clean interface to specifying and evaluating models. Phylip programs have a clean command interface and stable input and output formats. Rutger Vos's Bio::Phylo provides an example of integration with BioPerl, providing programmatic access to BioPerl objects and methods.

Goal #3 is to provide enhanced support for the prevailing standards and conventions that provide effective support for use-cases and thus represent the current interface to evolutionary analysis data and software.

Data exchange format. Iglesias, et al. (see www.cs.nmsu.edu/~gupta/biology/nexus.ps) have written a complete Backus-Naur formalization of the NEXUS language, thus providing a basis for validating files and identifying inconsistencies in implementation of output formats. Examples include such things as the use of the Data Block (deprecated) format as an output format by ClustalW and BioPerl; unorthodox branch support values in MrBayes; and so on. With respect to inputs, one of us (AS) has developed a set of NEXUS files that test for specific flaws (e.g., the inability to process quoted strings), and a round-trip tester that determines whether two NEXUS files have the same content. The remedy for inconsistencies is to write patches to existing code where available. Writing such a patch is an example of a demonstration project enhancing support for a standard data exchange format.

Program control. Analyses carried out by programs with clean command interfaces can be automated, and if the input and output formats are standardized, operations can be chained together in complex pipelines or work-flows. A specific project in supporting program control would be to develop Perl-based CDAT interfaces to Phylip programs, or to software that can be controlled with a NEXUS private-block interface (e.g., MrBayes, PAUP*, or HyPhy). A Perl script that automates an analysis using one of these tools is a simple demonstration project. Java examples could be developed using Mesquite.

Input and Output. Ideally we can agree on a standard that all evolutionary analysis programs can use, but for the near-term, different formats are in use, and it is important to provide format translation. BioPerl provides a means for this (its NEXUS implementation is incomplete, but will be fixed prior to this project as part of a separate Bio::CDAT project by the AS group). Simple target formats for data conversion include common alignment formats (clustalw, Phylip, etc), PhyloXML, and the Pandit format of Whelan, et al. A specific example of a demonstration project is automated conversion of thousands of Pandit data sets to NEXUS format, and analysis of an evolutionary model with HyPhy.

Storage. Another specific project is to develop software libraries to link character-data-and-trees objects to data streams and to database schemas, such as that used in TreeBase. CIPRES is developing such a method using CORBA (see example by RV in Bio::Phylo). Prior BioPerl art in this arena includes the Bio::DB::GFF (Gene-Feature format) object, which provides an abstract interface to GFF sequence feature data in flat-files, various relational schemas, and DAS web-services.

Visualization. Visualization is an area in which we should make recommendations rather than hard standards. However, there is an opportunity to make valuable suggestions about visual conventions for representing ancestral state distributions (e.g., pie charts, sequence logos), branch support values, duplications vs. speciation events, and so on.

Goal #4 is to propose the updates, additions and extensions to prevailing standards that will facilitate automated and integrative analysis for the near-term. With respect to file formats, this could involve replacing NEXUS with an XML alternative based on the CDAT model. Achieving this goal entails developing an API library in some language (Perl, C++ and Java are standards), sample implementations of the revised format, and publishing a formal description. The revised format should incorporate various extensions and alternatives that have been implemented to circumvent short-comings of NEXUS and New Hampshire. For instance, it is important to allow arbitrary attributes to be assigned to sub-objects and sub-sub-objects defined within the CDAT object, such as data sets, models, individual characters, trees, nodes, etc. This requires implementing references within the file (rather than simply using literal names), easily possible in XML. For instance, nodes or branches of a tree may have attributes representing different types of branch support values, designation of duplication vs. speciation, dating estimates, etc (see PhyloXML and the NHX format used by ATV). Mesquite has a "project" concept that includes multiple blocks of character data and references implemented by novel "title" and "link" commands.

The kinds of demonstration projects for goal #4 are the same as those for goal #3, except now the possibilities are expanded. For instance, with the ability to store ancestral reconstructions (AS group has developed a NEXUS "History" block for this), it would be possible to use NEXUS as both input and output to software for ancestral state reconstruction.

Goal #5 is to lay the groundwork for a more sweeping revision of standards, as necessary. In the current proposal, we are able to specify projects in detail because our focus is mainly to consolidate existing standards and conventions, and to make incremental revisions that incorporate recent developments. Developing a new standard with community support is a much more difficult task. This task will require completing the first four goals of this project: community cohesion, prior standards, and documented use-cases. For instance, the first step in developing a new standard is to identify use-cases that fall outside the realm of operations currently supported. We will have to ask difficult question about the differing needs of different communities (Tree-of-Life project, model comparison, population-based analyses, integrative analyses using new kinds of data), support for special cases (non-bifurcating networks, high-dimensional data), and the feasibility of a standard that incorporates other standards (e.g., the sequence ontology, SO; Gene-Feature format, GFF).

Names of Proposed Participants

Each of the following individuals has agreed to be named as a possible participant.

- Dr. Jonathan Eisen, UC Davis Genome Center, UC Davis, CA
- Dr. Joe Felsenstein, Department of Genome Sciences and Department of Biology, University of Washington, Seattle, WA
- Dr. Mark Holder, School of Computational Science, Florida State University, Tallahassee, FL
- Dr. Sergei L. Kosakovsky Pond, Antiviral Research Center, University of California, San Diego, San Diego, CA
- Dr. Sudhir Kumar, Center for Evolutionary Functional Genomics, The Biodesign Institute, Arizona State University, Tempe, AZ
- Dr. Aaron Mackey, GlaxoSmithKline, King of Prussia, PA
- Dr. David Maddison, Department of Entomology, University of Arizona, Tucson, AZ
- Dr. Wayne Maddison, Departments of Zoology and Botany, University of British Columbia, Vancouver, BC (Canada)
- Dr. Weigang Qiu, Department of Biological Sciences, Hunter College of CUNY, New York, NY
- Dr. Andrew Rambaut, Zoology Department, University of Oxford, South Parks Road, Oxford, UK
- Dr. David L. Swofford, School of Computational Science, Florida State University, Tallahassee, FL
- Dr. Xuhua Xia, Biology Department, University of Ottawa, Ottawa, ON (Canada)
- Dr. Christian Zmasek, Burnham Institute for Medical Research, La Jolla, CA

Rationale for NESCent support

The rationale for NESCent support is that this is a community activity, specific to the field of evolutionary biology. We know that the evolutionary approach to comparative biology is not just different, but better. This proposal takes advantage of a nascent opportunity to expand the scope and practical importance of evolutionary analysis by attacking a weak point, informatics. Other projects do not have this kind of integrative focus.

Anticipated IT Needs

This project will require a versioning server, e.g., to support CVS, either from NESCent or from an open-source host such as SourceForge. During meetings of the working groups, participants will need networking, as well as internet conferencing to work with any remote participants. The dissemination activities require a web site and some support for maintaining the web site. The web site will provide documentation of the activities of the working group (e.g., electronic presentations), and will provide access to its results in terms of code revisions, demonstration projects,

Proposed Timetable

The Evolutionary Informatics Group will meet twice a year for two years, with the exact schedule to be determined later. Each meeting will last for 3-5 days and begin with presentations, proceed with a hack-a-thon, and end with progress reports. The proportion of the meeting devoted to software development and testing will increase as the project continues. The main focus of software development and testing is on goals 3 and 4. By the end of the second meeting, the group should be ready to draft a report, suitable for publication, proposing informatics standards for use in evolutionary analysis.

Anticipated Results

The tangible results of the activities of the Evolutionary Informatics Working Group include computer code, collated data sets, presentations, and a manuscript describing proposed standards and conventions, all of which will be made available on a group web site. The programming projects will include revisions to existing applications software, additions to software libraries such as BioPerl and Mesquite, and demonstration projects.

Biographical Sketch: Arlin Stoltzfus, Ph.D.

Professional Preparation

Grinnell College, Iowa, USA	English	B.A., *cum laude*, 1985
University of Iowa, Iowa, USA	Biology	Ph.D., 1991
Dalhousie Univ., Halifax, Canada	Molecular Evolution	1991-1999

Appointments

1999-present	Adjunct Assistant Professor, U. Maryland Biotechnology Institute
1999-present	Research Biologist, National Institute of Standards and Technology
1997	Teaching assistant, Genetics, Dalhousie University (Halifax, Nova Scotia)
1991-1999	Post-doctoral Fellow, Dalhousie University (Halifax, Nova Scotia)
1985-1991	Research assistant, Biology, University of Iowa (Iowa City, Iowa)
1985-1989	Teaching assistant, Biology, University of Iowa (Iowa City, Iowa)
1984	Teaching assistant, Biology, Grinnell College (Grinnell, Iowa)

Publications

Stoltzfus, A. Mutation-biased adaptation in a protein NK model. Mol Biol Evol (accepted).

Stoltzfus, A. Mutationism and the Dual Causation of Evolutionary Change. Evol Dev 8, 304-317 (2006).

Gopalan, V., Qiu, W. G., Chen, M. Z. & Stoltzfus, A. Nexplorer: phylogeny-based exploration of sequence family data. *Bioinformatics* 22, 120-121 (2006).

L.Y. Yampolsky and A. Stoltzfus. 2005. Untangling the Effects of Codon Mutation and Amino Acid Exchangeability. *Pacific Symposium on Biocomputing***,** 433-444

L.Y. Yampolsky and A. Stoltzfus. 2005. The Exchangeability of Amino Acids in Proteins. *Genetics* **170**, 1459-1472.

A. Stoltzfus. 2004. Molecular Evolution: Introns Fall into Place. *Curr. Biol.* 14:R351-2.

W.-G. Qiu, N.J. Schisler and A. Stoltzfus. 2004. Spliceosomal intron gain: sequence and phase preferences. *Molecular Biology and Evolution* 21:1252-63.

A. Stoltzfus. 2001. “Introns and Exons”, subject entry in The Encyclopedia of Genetics (Academic Press, San Diego).

L. Y. Yampolsky and A. Stoltzfus. 2001. Bias in the introduction of variation as an orienting factor in evolution, *Evol Dev,* 3, 73-83.

A. Stoltzfus. 1999. On the possibility of constructive neutral evolution. *J. Mol. Evol*. 49(2):169-181.

J.M. Logsdon, Jr., A. Stoltzfus, and W.F. Doolittle. 1998. Recent cases of spliceosomal intron gain? *Curr. Biol.* 8:R560-R563.

S. de Souza, W. Fischer, J.M. Logsdon, Jr., M. Long, W. Martin, and A. Stoltzfus. 1997. The great debate on the origin of introns. An electronic discussion in the inaugural issue of the electronic journal *HMS Beagle* (http://biomednet.com/hmsbeagle/1997/01/cutedge/overview.htm).

A. Stoltzfus, J. M. Logsdon, Jr., J. D. Palmer, and W. F. Doolittle. 1997. Intron "sliding" and the diversity of intron positions. *Proc. Natl. Acad. Sci. U.S.A*. 94: 10739-10744.

A. Stoltzfus, D.F. Spencer, M. Zuker, J.M. Logsdon, Jr.,and W.F. Doolittle. 1995. (Technical correspondence: Introns and the origin of protein-coding genes) *Science* 268:1367-1369.

A. Stoltzfus, D.F. Spencer, and W.F. Doolittle. 1995. Methods for evaluating exon-protein correspondences. *Computer Applications in the Biosciences* 11:509-15.

A. Stoltzfus, D.F. Spencer, M. Zuker, J.M. Logsdon, Jr.,and W.F. Doolittle. 1994. Testing the exon theory of genes: the evidence from protein structure. *Science* 265:202-7.
A. Stoltzfus. 1994. Origin of introns--early or late (Scientific Correspondence). *Nature* 369:526-7.
A. Stoltzfus and W.F. Doolittle. 1993. Slippery introns and globin gene evolution. *Current Biology* 3:215-217.
W.F. Doolittle and A. Stoltzfus. 1993. Molecular evolution: Genes-in-pieces revisited (News & Views). *Nature* 361:403.
A. Stoltzfus. 1991. *A survey of natural variation in the trp-tonB region of the E. coli chromosome.* Ph.D. Thesis, University of Iowa Press (Iowa City, Iowa).
A. Stoltzfus, J.F. Leslie, and R. Milkman. 1988. Molecular evolution of the Escherichia coli chromosome. I. Analysis of structure and natural variation in a previously uncharacterized region between trp and tonB. *Genetics* 120:345-58.
R. Milkman and A. Stoltzfus. 1988. Molecular evolution of the *Escherichia coli* chromosome. II. Clonal segments. *Genetics* 120:359-66

Synergistic Activities in the past 5 years

Delivered nine invited lectures at academic departments, conferences, and research institutes.
Mentored three interns from Montgomery Blair High School (magnet school for math, science and computing)
Mentored four undergraduate interns (Duke, Rice, UMCP) and one Master's student (UMBC)
Mentored three post-doctoral fellows who have gone on to positions in academia.
Organized local mini-symposium on gene duplication with speakers from NIH, NCBI, TIGR and CARB
Organized "chalk talks" on molecular evolution that brought in two dozen DC-area molecular evolutionists and computational biologists to speak with the Stoltzfus group at CARB
Developed a course module in bioinformatics for a course in Protein Structure and Function at UMCP (Molecular and Cellular Biology Program)
Participated in journal clubs in Systems Biology and Evo-Devo

Current external funding

"Intron Evolution: System for Phyloinformatic Analysis", NIH R01 LM007218, from the Computational Biology Program of the National Library of Medicine, to principal investigator Arlin Stoltzfus; $704K total direct costs; Oct. 2002-2006.

Collaborations and Other Affiliations

Collaborators: John M. Logsdon, Jr. (University of Iowa), Wei-Gang Qiu (Hunter College, CUNY), Nick Schisler (Furman University), Lev Yampolsky (East Tennessee State University), James R. Brown (GlaxoSmithKline), Sergei Kossakovsky Pond (

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

Thesis Advisor and Post-graduate-Scholar Sponsor: Wei-Gang Qiu (Hunter College, CUNY), Lev Yampolsky (East Tennessee State University),

Curriculum Vitae

Name: Rutger Aldo Vos
Date of birth: October 5th 1975, Amsterdam
Nationality: Dutch
Current address: 5879 Booth Avenue
Burnaby, British Columbia, V5H 3A9
Canada
Phone: +1 604 780 0190 (cell); +1 604 291 5625 (lab)
Fax: +1 604 291 3496
E-mail: rvosa@sfu.ca
Education: MSc. University of Amsterdam, 2000, PhD. Simon Fraser University, 2006.

Awards and fellowships

SFU Graduate Fellowship, 2002, 2003
SFSS Travel Award, 2003
SSB Systematic Biology Graduate Research Award, 2003
President's Research Stipend, 2005

Peer-reviewed publications

Vos, R.A., 2000. The generalist-to-specialist hypothesis in primate evolution. MSc Thesis, Institute for Biodiversity and Ecosystem Dynamics, University of Amsterdam, Amsterdam.

Rundle, H. D., F. Breden, C. Griswold, A. Ø. Mooers, **R. A. Vos**, and J. Whitton. 2001. Hybridization without Guilt: Gene Flow and the Biological Species Concept. *Journal of Evolutionary Biology* 14: 868-869.

Vos, R. A. 2003. Accelerated likelihood surface exploration: The likelihood ratchet. *Systematic Biology* 52: 368-373.

Vos, R. A. and A. Ø. Mooers, 2004. Reconstructing divergence times for supertrees: a molecular approach. *In:* O. R. P. Bininda-Emonds (Ed.), *Phylogenetic supertrees: Combining information to reveal the Tree of Life.* pp 281-299. Kluwer Academic Publishers, Dordrecht.

E. J. de Vries, **R. A. Vos**, G. Jacobs, and J. A. J. Breeuwer, in review. Western flower thrips preference for thrips-damaged leaves over fresh leaves enables uptake of symbiotic gut bacteria. *Entomologia Experimentalis et Applicata*

Vos, R. A. and A. Ø. Mooers, in revision. A dated MRP supertree for the order Primates. *Systematic Biology*

Vos, R. A., accepted. Accelerated Metropolis-Hastings coupled Markov chain Monte Carlo burn-in by iterative jackknifing. *Systematic Biology*

Non-refereed reports

Vos, R. A., 2003. Common procedures in phylogenetics. CIPRes internal report

Vos, R. A., 2004. An XML for phylogenetics. CIPRes internal report

Vos, R. A., 2005. Design patterns in phylogenetics: practical tree data structures and objects for serialization. CIPRes internal report.

EMPLOYMENT HISTORY

Position	Duration	Employer	Description
Research assistant	9/'98 – 6/'99	Institute for Systematics and Population Biology	Lab experiments (RAPD-PCR) tracking population divergence in laboratory strains of the Western Flower Thrips *Frankliniella occidentalis*.
Educational software developer	6/'99 – 9/'99	AMSTEL Institute	Development of educational software on mitosis and meiosis visualised with 3D microscope photographs.
Educational software developer	6/'00 – 12/'00	Biomedia	Website and software design for life sciences education.
Research assistant	1/'01 – 8/'06	Simon Fraser University	Assistant on a number of research projects involving phylogenetics.
Research assistant, postdoctoral research fellow	1/'04 –	University of British Columbia	International collaborator on CIPRES, contribution to architecture and database design.

TEACHING EXPERIENCE

Position	Duration	Employer	Description
Teaching Assistant	9/'99 – 6/'00	Zoological Museum of Amsterdam	For an international master course on phylogenetics I developed and taught a web-based lab on the systematics of lemurs.
Teaching Assistant	9/'01 – 12/'01	Simon Fraser University	BISC475: Biodiversity, lectures and seminars.
Teaching Assistant	9/'01 – 12/'01	Simon Fraser University	BISC100: Introduction to biology, labs and seminars.
Teaching Assistant	1/'02 – 4/'02	Simon Fraser University	BISC400: Evolution, seminars.
Teaching Assistant	9/'04 – 12/'04	Simon Fraser University	BISC475 : Biodiversity, lectures and seminars.

SERVICES TO THE SCIENTIFIC COMMUNITY

Co-chair (with Patrik Nosil), organizing committee for the 24th Annual Pacific Ecology and Evolution Conference (2003).

Webmaster for http://www.scientists-4-species.org which successfully lobbied for a strengthened Species At Risk Act (SARA; passed by Canadian Senate December 16th, 2002).

Reviewer for *Phylogenetic supertrees: Combining information to reveal the Tree of Life*, for *The American Naturalist*, for *Systematic Biology* and for *Bioinformatics*.

Guest speaker for Shad Valley on Human Evolution and on Phylogenetics.

Member of the Doctoral Students Advisory Group for Burnaby Mountain College.

Co-organizer departmental seminar series, Simon Fraser University.

Organizer of seminar series on Perl programming for biologists.

Further collaborations include CIPRES, IRMACS, Mammal Superteam, FAB* lab and VEG

Biographical Sketch: Arlin Stoltzfus, Ph.D.

Professional Preparation

Grinnell College, Iowa, USA	English	B.A., *cum laude*, 1985
University of Iowa, Iowa, USA	Biology	Ph.D., 1991
Dalhousie Univ., Halifax, Canada	Molecular Evolution	1991-1999

Appointments

1999-present	Adjunct Assistant Professor, U. Maryland Biotechnology Institute
1999-present	Research Biologist, National Institute of Standards and Technology
1997	Teaching assistant, Genetics, Dalhousie University (Halifax, Nova Scotia)
1991-1999	Post-doctoral Fellow, Dalhousie University (Halifax, Nova Scotia)
1985-1991	Research assistant, Biology, University of Iowa (Iowa City, Iowa)
1985-1989	Teaching assistant, Biology, University of Iowa (Iowa City, Iowa)
1984	Teaching assistant, Biology, Grinnell College (Grinnell, Iowa)

Publications

Stoltzfus, A. Mutation-biased adaptation in a protein NK model. Mol Biol Evol (accepted).

Stoltzfus, A. Mutationism and the Dual Causation of Evolutionary Change. Evol Dev 8, 304-317 (2006).

Gopalan, V., Qiu, W. G., Chen, M. Z. & Stoltzfus, A. Nexplorer: phylogeny-based exploration of sequence family data. *Bioinformatics* 22, 120-121 (2006).

L.Y. Yampolsky and A. Stoltzfus. 2005. Untangling the Effects of Codon Mutation and Amino Acid Exchangeability. *Pacific Symposium on Biocomputing***,** 433-444

L.Y. Yampolsky and A. Stoltzfus. 2005. The Exchangeability of Amino Acids in Proteins. *Genetics* **170**, 1459-1472.

A. Stoltzfus. 2004. Molecular Evolution: Introns Fall into Place. *Curr. Biol.* 14:R351-2.

W.-G. Qiu, N.J. Schisler and A. Stoltzfus. 2004. Spliceosomal intron gain: sequence and phase preferences. *Molecular Biology and Evolution* 21:1252-63.

A. Stoltzfus. 2001. "Introns and Exons", subject entry in The Encyclopedia of Genetics (Academic Press, San Diego).

L. Y. Yampolsky and A. Stoltzfus. 2001. Bias in the introduction of variation as an orienting factor in evolution, *Evol Dev,* 3, 73-83.

A. Stoltzfus. 1999. On the possibility of constructive neutral evolution. *J. Mol. Evol.* 49(2):169-181.

J.M. Logsdon, Jr., A. Stoltzfus, and W.F. Doolittle. 1998. Recent cases of spliceosomal intron gain? *Curr. Biol.* 8:R560-R563.

S. de Souza, W. Fischer, J.M. Logsdon, Jr., M. Long, W. Martin, and A. Stoltzfus. 1997. The great debate on the origin of introns. An electronic discussion in the inaugural issue of the electronic journal *HMS Beagle* (http://biomednet.com/hmsbeagle/1997/01/cutedge/overview.htm).

A. Stoltzfus, J. M. Logsdon, Jr., J. D. Palmer, and W. F. Doolittle. 1997. Intron "sliding" and the diversity of intron positions. *Proc. Natl. Acad. Sci. U.S.A*. 94: 10739-10744.

A. Stoltzfus, D.F. Spencer, M. Zuker, J.M. Logsdon, Jr.,and W.F. Doolittle. 1995. (Technical correspondence: Introns and the origin of protein-coding genes) *Science* 268:1367-1369.

A. Stoltzfus, D.F. Spencer, and W.F. Doolittle. 1995. Methods for evaluating exon-protein correspondences. *Computer Applications in the Biosciences* 11:509-15.

A. Stoltzfus, D.F. Spencer, M. Zuker, J.M. Logsdon, Jr.,and W.F. Doolittle. 1994. Testing the exon theory of genes: the evidence from protein structure. *Science* 265:202-7.

A. Stoltzfus. 1994. Origin of introns--early or late (Scientific Correspondence). *Nature* 369:526-7.

A. Stoltzfus and W.F. Doolittle. 1993. Slippery introns and globin gene evolution. *Current Biology* 3:215-217.

W.F. Doolittle and A. Stoltzfus. 1993. Molecular evolution: Genes-in-pieces revisited (News & Views). *Nature* 361:403.

A. Stoltzfus. 1991. *A survey of natural variation in the trp-tonB region of the E. coli chromosome.* Ph.D. Thesis, University of Iowa Press (Iowa City, Iowa).

A. Stoltzfus, J.F. Leslie, and R. Milkman. 1988. Molecular evolution of the Escherichia coli chromosome. I. Analysis of structure and natural variation in a previously uncharacterized region between trp and tonB. *Genetics* 120:345-58.

R. Milkman and A. Stoltzfus. 1988. Molecular evolution of the *Escherichia coli* chromosome. II. Clonal segments. *Genetics* 120:359-66

Synergistic Activities in the past 5 years

Delivered nine invited lectures at academic departments, conferences, and research institutes.
Mentored three interns from Montgomery Blair High School (magnet school for math, science and computing)
Mentored four undergraduate interns (Duke, Rice, UMCP) and one Master's student (UMBC)
Mentored three post-doctoral fellows who have gone on to positions in academia.
Organized local mini-symposium on gene duplication with speakers from NIH, NCBI, TIGR and CARB
Organized "chalk talks" on molecular evolution that brought in two dozen DC-area molecular evolutionists and computational biologists to speak with the Stoltzfus group at CARB
Developed a course module in bioinformatics for a course in Protein Structure and Function at UMCP (Molecular and Cellular Biology Program)
Participated in journal clubs in Systems Biology and Evo-Devo

Current external funding

"Intron Evolution: System for Phyloinformatic Analysis", NIH R01 LM007218, from the Computational Biology Program of the National Library of Medicine, to principal investigator Arlin Stoltzfus; $704K total direct costs; Oct. 2002-2006.

Collaborations and Other Affiliations

Collaborators: John M. Logsdon, Jr. (University of Iowa), Wei-Gang Qiu (Hunter College, CUNY), Nick Schisler (Furman University), Lev Yampolsky (East Tennessee State University), James R. Brown (GlaxoSmithKline), Sergei Kossakovsky Pond (

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

Thesis Advisor and Post-graduate-Scholar Sponsor: Wei-Gang Qiu (Hunter College, CUNY), Lev Yampolsky (East Tennessee State University),

Curriculum Vitae

Name:	Rutger Aldo Vos
Date of birth:	October 5th 1975, Amsterdam
Nationality:	Dutch
Current address:	5879 Booth Avenue Burnaby, British Columbia, V5H 3A9 Canada
Phone:	+1 604 780 0190 (cell); +1 604 291 5625 (lab)
Fax:	+1 604 291 3496
E-mail:	rvosa@sfu.ca
Education:	MSc. University of Amsterdam, 2000, PhD. Simon Fraser University, 2006.

Awards and fellowships

SFU Graduate Fellowship, 2002, 2003
SFSS Travel Award, 2003
SSB Systematic Biology Graduate Research Award, 2003
President's Research Stipend, 2005

Peer-reviewed publications

Vos, R.A., 2000. The generalist-to-specialist hypothesis in primate evolution. MSc Thesis, Institute for Biodiversity and Ecosystem Dynamics, University of Amsterdam, Amsterdam.

Rundle, H. D., F. Breden, C. Griswold, A. Ø. Mooers, **R. A. Vos**, and J. Whitton. 2001. Hybridization without Guilt: Gene Flow and the Biological Species Concept. *Journal of Evolutionary Biology* 14: 868-869.

Vos, R. A. 2003. Accelerated likelihood surface exploration: The likelihood ratchet. *Systematic Biology* 52: 368-373.

Vos, R. A. and A. Ø. Mooers, 2004. Reconstructing divergence times for supertrees: a molecular approach. *In:* O. R. P. Bininda-Emonds (Ed.), *Phylogenetic supertrees: Combining information to reveal the Tree of Life.* pp 281-299. Kluwer Academic Publishers, Dordrecht.

E. J. de Vries, **R. A. Vos**, G. Jacobs, and J. A. J. Breeuwer, in review. Western flower thrips preference for thrips-damaged leaves over fresh leaves enables uptake of symbiotic gut bacteria. *Entomologia Experimentalis et Applicata*

Vos, R. A. and A. Ø. Mooers, in revision. A dated MRP supertree for the order Primates. *Systematic Biology*

Vos, R. A., accepted. Accelerated Metropolis-Hastings coupled Markov chain Monte Carlo burn-in by iterative jackknifing. *Systematic Biology*

Non-refereed reports

Vos, R. A., 2003. Common procedures in phylogenetics. CIPRes internal report

Vos, R. A., 2004. An XML for phylogenetics. CIPRes internal report

Vos, R. A., 2005. Design patterns in phylogenetics: practical tree data structures and objects for serialization. CIPRes internal report.

EMPLOYMENT HISTORY

Position	Duration	Employer	Description
Research assistant	9/’98 – 6/’99	Institute for Systematics and Population Biology	Lab experiments (RAPD-PCR) tracking population divergence in laboratory strains of the Western Flower Thrips *Frankliniella occidentalis*.
Educational software developer	6/’99 – 9/’99	AMSTEL Institute	Development of educational software on mitosis and meiosis visualised with 3D microscope photographs.
Educational software developer	6/’00 – 12/’00	Biomedia	Website and software design for life sciences education.
Research assistant	1/’01 – 8/’06	Simon Fraser University	Assistant on a number of research projects involving phylogenetics.
Research assistant, postdoctoral research fellow	1/’04 –	University of British Columbia	International collaborator on CIPRES, contribution to architecture and database design.

TEACHING EXPERIENCE

Position	Duration	Employer	Description
Teaching Assistant	9/’99 – 6/’00	Zoological Museum of Amsterdam	For an international master course on phylogenetics I developed and taught a web-based lab on the systematics of lemurs.
Teaching Assistant	9/’01 – 12/’01	Simon Fraser University	BISC475: Biodiversity, lectures and seminars.
Teaching Assistant	9/’01 – 12/’01	Simon Fraser University	BISC100: Introduction to biology, labs and seminars.
Teaching Assistant	1/’02 – 4/’02	Simon Fraser University	BISC400: Evolution, seminars.
Teaching Assistant	9/’04 – 12/’04	Simon Fraser University	BISC475 : Biodiversity, lectures and seminars.

SERVICES TO THE SCIENTIFIC COMMUNITY

Co-chair (with Patrik Nosil), organizing committee for the 24th Annual Pacific Ecology and Evolution Conference (2003).

Webmaster for http://www.scientists-4-species.org which successfully lobbied for a strengthened Species At Risk Act (SARA; passed by Canadian Senate December 16th, 2002).

Reviewer for *Phylogenetic supertrees: Combining information to reveal the Tree of Life*, for *The American Naturalist*, for *Systematic Biology* and for *Bioinformatics*.

Guest speaker for Shad Valley on Human Evolution and on Phylogenetics.

Member of the Doctoral Students Advisory Group for Burnaby Mountain College.

Co-organizer departmental seminar series, Simon Fraser University.

Organizer of seminar series on Perl programming for biologists.

Further collaborations include CIPRES, IRMACS, Mammal Superteam, FAB* lab and VEG