Deep arthropod phylogeny from 100 targeted nuclear coding-region sequences

(DEB-0120635; Genome-Enabled Environmental Sciences and Engineering / Biocomplexity in the Environment (BE): Integrated Research and Education in Environmental Systems)

Clifford W. Cunningham (Duke University), Joel W. Martin (Natural History Museum of Los Angeles County), Jerome C. Regier (University of Maryland Biotechnology Institute), Jeffrey W. Shultz (University of Maryland), & Jeffrey L. Thorne (North Carolina State University)

PROJECT SUMMARY UPDATE (November 15, 2004).

There are four major goals in this project. One goal (actually, goal2) is to generate a robust, higher-level phylogeny of Arthropoda to be inferred from primary sequence information present in multiple protein-coding regions of the nuclear genome. First, however, it is necessary to develop suitable genomic tools (goal 1). These tools consist of a bank of PCR primers that can consistently amplify orthologous sequences across Arthropoda and their near outgroups (i.e., Tardigrada and Onychophora in our project). Goal 1 is now largely completed, and we are on-target to complete goal 2 by the end of the granting period. The third major goal is to develop and to apply improved Bayesian methods to estimate absolute times of arthropod clade divergence from our multigenic sequence data and fossil evidence. Progress in software development is ongoing and current versions are already widely utilized. The newest software will be applied to the full arthropod data matrix to be completed in year 5 of the project. The fourth major goal was to develop and implement an after-school program on arthropod diversity and evolution for high school students in Los Angeles. This program was offered in year 1. Ideas, materials, and a dedicated Web site that resulted from program development continued to foster education about arthropods well after completion of the five-week program.

GOAL #1 -- PHYLOGENOMIC TOOLS.

We have sought to expand genomic information for arthropod phylogeny by developing PCR primers for a large number of protein-coding nuclear genes that can be consistently amplified. This study builds on previous explorations of arthropod phylogeny using elongation factor-1alpha, RNA polymerase II, and elongation factor-2 (4.3 kb total; see MPE 20: 136-148, 2001).

STEP 1 (completed). We have aligned coding regions from whole genomic sequences of human, fly, and roundworm. Putative orthologous genes were identified. In most cases, we confined analysis to orthologs present in all three taxa and for which any paralogs were highly divergent. In total, 595 genes were selected for further analysis.

STEP 2 (completed). Sequence alignments of these 595 genes were visually scanned for potential PCR primer sites. Typically, potential primer sites encoded six or more completely conserved amino acids that did not include leucine, arginine, or serine. Primers were made completely degenerate with respect to potential synonymous change, and overall levels of nucleotide degeneracy were typically at or below 128-fold. We tried to find at least three primer sites on individual genes to enable a hemi-nested amplification strategy, although this was not always possible. Amplicon sizes varied from ~300 bp to ~1200 bp. Primer pairs were identified for 159 of the 595 genes (27%). A total of 572 distinct primers were synthesized.

STEP 3 (completed). Within appropriate gene segments, all pairwise combinations of primers were tested for their ability to amplify five diverse panarthropods (i.e., one species each of Chelicerata, Myriapoda, Crustacea, Hexapoda, and Tardigrada). We chose an RT-PCR strategy rather than direct gene amplification in order to avoid introns, which complicate amplicon identification on gels, PCR amplification, and sequence assembly. The downside of this strategy is that some mRNA sequences may be temporally or spatially restricted, or ubiquitous but present in very low concentration, making amplification more difficult. Typically, the RT-PCR band (either visible or invisible) was gel isolated and reamplified from a nested primer site. The reamplified band was then gel isolated and sequenced directly. This direct approach allowed us to assess and to incorporate any polymorphism into the sequence. Our rule-of-thumb has been that while invisible RT-PCR products are tested for successful reamplification to yield visible bands, we do not reamplify invisible "bands" more than once. The number of false positives has been very low (~0.4%). Of the 159 genes for which primers were defined and tested, successful amplification was obtained for at least three of the five test taxa in 61 genes (38%), representing 68 nonoverlapping PCR segments. Successful amplifications used 188 different primers. Over the years, we have made improvements in PCR protocols that should be of interest to other molecular systematists (e.g., see Figure 1).

GOAL #1 -- PHYLOGENOMIC TOOLS (continued).

STEP 4 (completed). The number of species amplified for each of the 68 segments was increased from five to 13 -- two hexapods, five crustaceans (five different classes), two chelicerates, two myriapods, and one tardigrade. Overall, the data matrix is ~90% complete. The sequences have been preliminarily aligned, and although some indels are present, they are neither widespread nor do they (in most cases) present challenging alignment problems. The total number of nucleotide characters in 60 genes (and 66 separate fragments) is ~37,500. With EF-1alpha, Pol II, and EF-2 added (63 genes total), the total number of nucleotide characters is ~42,800. Sequences from two other genes were generally too polymorphic to assemble. Time permitting, we may clone amplicons of highly heterogenerous sequences to determine whether the heterogeneity is due mostly to synonymous change or whether we have amplified multiple, divergent paralogs.

STEP 5 (current and near-future). We are re-sequencing a small sampling of taxa to confirm the high quality of the data matrix. Shortly, we will perform a final alignment and begin analysis to characterize the evolutionary rates, among-site-rate-heterogeneity, and general phylogenetic informativeness of the individual genes. Early in 2005, we will make publically available the list of genes and primers, along with detailed protocols for amplification and recommendations as to the taxonomic levels for which particular genes seem most suited. The Cunningham lab has already used three of the genes (and 1460 bp) to robustly infer numerous relationships within a genus of dung beetles. A parallel study of two ribosomal + two mitochondrial genes (and 1844 bp) for the same taxa yielded much lower node support values. The Regier lab has sequenced all 60 of the gene regions from three diverse Lepidoptera (one non-ditrysian, one non-obtectomeran apodytrisian, and one apodytrisian) in order to assess their utility within that insect order and to identify genes of particular utility for resolving Mesozoic-age divergences, wherein lies much interest across Hexapoda. Manuscripts will be forthcoming.

GOAL #2 -- ARTHROPOD PHYLOGENETICS (next two years).

We will expand the sequencing of taxa to ~85 species for all 60 genes (+ EF-1alpha, Pol II, and EF-2). Phylogenetic analysis of the expanded data matrix will particularly aim to resolve the following controversial aspects of arthropod phylogeny:

a. the basal trichotomy (Chelicerata, Myriapoda, Pancrustacea),
b. pancrustacean higher-level relationships, including relationships among the traditional crustacean classes and the position of hexapods,
c. the monophyly (or, more likely, lack thereof) of Maxillopoda (Crustacea),
d. myriapod class relationships, and
e. arachnid ordinal relationships.

While the extensive effort expended in primer development (goal #1) has precluded sampling more than 85 taxa, we are hopeful that even this number will prove informative. Furthermore, we suspect that the genes developed in this study will prove useful to others who explore their utility in more focused and highly sampled investigations.

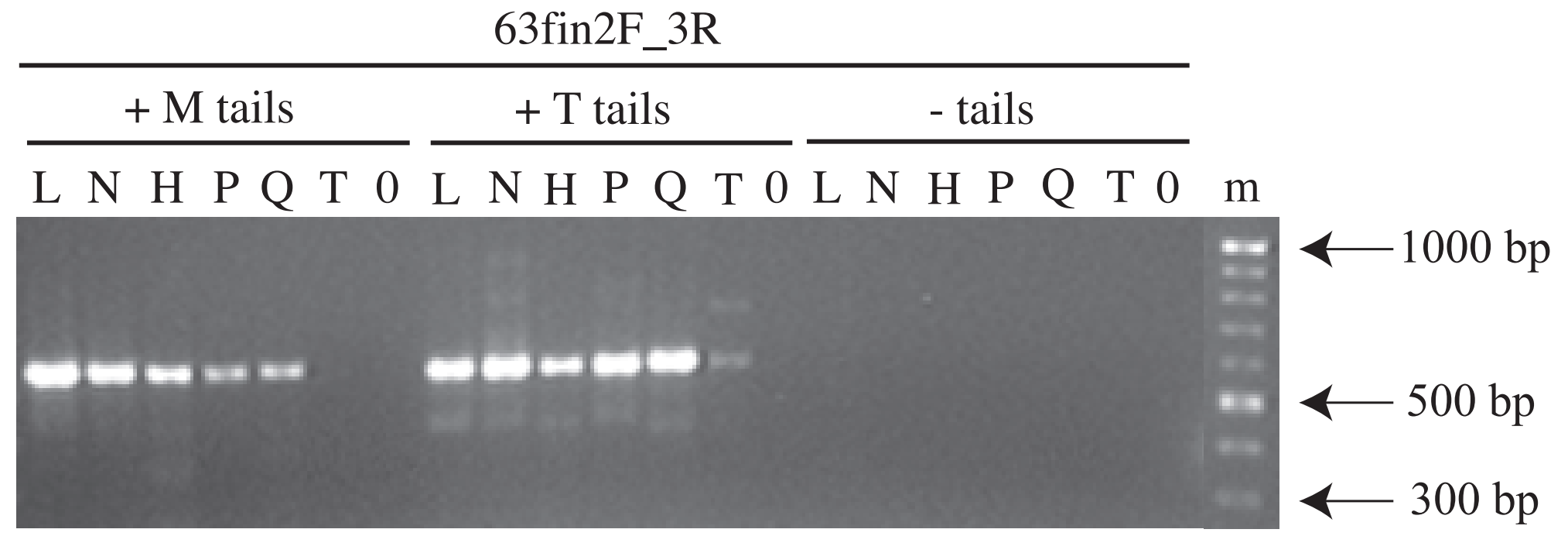

Figure 1. Increased yield of PCR product from degenerate primers (in this case, the primer pair "63fin2F_3R") with nondegenerate, nonhomologous 5' tails. Six diverse panarthropod taxa (L-T) plus a negative control (0) were amplified by RT-PCR using primers with and without nonhomologous 5' tails. The effect is specific to degenerate primers, is not dependent on the sequence of the 5' tail, and is very widespread (In Press).

GOAL #3 -- CLADE DIVERGENCE-TIME ESTIMATES (ongoing).

Co-PI Thorne and colleagues have continued to develop statistical methods and to modify their software for estimating divergence times from multigenic sequence data and fossil evidence (see Syst. Biol. 51: 689-72, 2002). Over 100 published studies have used the programs available via a web page (http://statgen.ncsu.edu/thorne/multidivtime.html). More recently, Thorne and colleagues have developed methods for separately estimating rates of synonymous and nonsynonymous change (rather than simply their ratio) as a means to better characterize the evolutionary process (MBE 21: 1201-1213, 2004). When our 63-gene sequence data matrix is completed for ~85 taxa in ~two years and phylogenetic analysis has yielded a robust phylogeny of Arthropoda (see goal #2), we will utilize the latest software iteration from co-PI Thorne to estimate key arthropod divergence times.

GOAL #4 -- EDUCATION IN ARTHROPOD DIVERSITY AND EVOLUTION (completed).

An after-school program entitled "Arthropod Diversity and Evolution" was designed and implemented for local high school students in the Los Angeles area (see course outline below). The class was offered in February and March of 2002 (= year 1 of current award). Co-PI Martin and colleagues worked extensively with education staff at the Natural History Museum of Los Angeles County. The course included tours of the museum displays, behind-the-scenes tours of arthropod collections, interactions with museum curators and collection managers, active participation in fieldwork, and lecture and lab sessions. The outline for the course and details of the activities that were planned for each session can be found on the project's web site at http://arthropods.nhm.org/ education.html.

COURSE OUTLINE

WEEK	TOPICS	WORKSHPS	LAB & FIELD TRIPS
1	1. Intro to the arthro. project Who are the arthropods? Arthropod diversity Live arthropods	Timeline	NHMLAC insect zoo NHMLAC invert. paleontology collections
	2. Evolution by nat. select.	Nat. select. Matrix building	NHMLAC Ent.
	Arthropods in review & matrix development		NHMLAC setup "Arthropod hatchery"
2	3. Arthropod morphology Insects		Grasshopper dissect. USC Seaver Library
	4. Terrestrial arthropods	ID keys to insect orders	Terrestrial field trip
3	5. Crustaceans	Curation	Crayfish dissection NHMLAC Crustacea collections
	6. Marine arthropods		
4	7. Chelicerates, millipedes centipedes, tardigrades, pycnogonids & Onychophora	Extract & amplify arthropod DNA.	Examine specimens NHMLAC Molec Lab
	8. Building phylogenies	MacClade demonstration	
5	9. Review & cleanup	Work on Web pg. Prepare presentations	
	10. Final class	Student oral presentations	Arthropod feast & arthropod Jeopardy

---

The ideas, materials, and the Web site were subsequently used for teacher training (e.g., presentations at conferences for high school science teachers) and for other arthropod-related activities (e.g., an Insect Fair).