Syst. Biol. 56(3):523–531, 2007

ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150701395340

WASABI: An Automated Sequence Processing System for Multigene Phylogenies

FRANK KAUFF,[1,2] CYMON J. COX,[1,3] AND FRANÇOIS LUTZONI[1]

[1]*Department of Biology, Duke University, Box 90338, Durham, NC 27708, USA; E-mail: fkauff@biologie.uni-kl.de (F.K.)*
[2]*Current Address: FB Biologie, Abt. Molekulare Phylogenetik, Technische Universität Kaiserslautern, Postfach 3049, 67653 Kaiserslautern, Germany*
[3]*Current Address: Department of Zoology, Natural History Museum, Cromwell Road, SW7 5BD, London, United Kingdom*

WASABI (Web Accessible Sequence Analysis for Biological Inference) is a software that provides the computational infrastructure for multiuser multigene phylogenetic projects. In addition to serving as a web-accessible sequence and voucher database, WASABI provides the functionality for automated sequence analysis, verification, and data set assembly. WASABI is unique as it supports the user by carrying out all steps from processing raw sequence data to a multigene sequence analysis within a strong phylogenetic context, and thus differs from tools for databasing and analyzing finalized gene sequences (e.g., ARB [Ludwig et al., 2004], TaxMan [www.nematodes.org/bioinformatics/Taxman/index.shtml]) or for the processing of ESTs (e.g., StackPack [www.egenetics.com/stackpack.html], PartiGene [Parkinson et al., 2004]). Although originally developed as a computational framework for the Assembling the Fungal Tree of Life (AFTOL) project (www.aftol.org, www.lutzonilab.net/aftol), part of the NSF-funded Assembling the Tree of Life initiative (ATOL, atol.sdsc.edu), WASABI can easily be adapted to the specific needs of most multiuser multigene sequencing projects. Currently, components of WASABI are adapted for three other Tree of Life projects: Assembling the Beetle Tree of Life (BToL, insects.oeb.harvard.edu/ATOL/people.htm), Cnidaria AToL (CnidTol, cnidarian.info), and Assembling the Tree of Eukaryotic Diversity (Eu-Tree, www.biology.uiowa.edu/eu_tree). WASABI should not be confused with the identically named Web Application for the Semantic Architecture of Biodiversity Informatics (Perry and Vieglais, 2006).

The AFTOL project, which was initiated in January 2003, is a collaboration centered upon five laboratories in four U.S. universities, with the international participation of more than 100 scientists. The aim of AFTOL is to improve our knowledge of fungal phylogeny by collecting sequence data from eight genetic loci for 1500 fungal taxa across all major fungal lineages, as well as developmental and ultrastructural data from a reduced set of selected taxa (Spatafora, 2005). AFTOL participants can contribute to this project by providing fresh specimens, culture strains, frozen material, DNA samples, primary sequence data for specific genes, or phenotypic data for targeted taxa. All information, including voucher information and the availability and location of material/samples, is attached to a DNA sample with a specific AFTOL number. These data are stored in a central database, managed by and accessed through WASABI (www.lutzonilab.net/aftol).

It is generally agreed that nucleotide sequence data derived from multiple unlinked loci are essential to resolve, with high confidence, relationships among a large number of species and ultimately assemble the tree of life (Lutzoni et al., 2004; James et al., 2006). Two main operations are required for large-scale phylogenetic studies—data acquisition and phylogenetic analysis. We define here the data acquisition process as all steps from the collection of samples in the field to the generation of concatenated data matrices of multiple data partitions (phenotypic and genomic) for which each data partition was tested for congruence until no significant conflicts were detected among data partitions; i.e., until the data are ready for final phylogenetic analyses. Most research efforts have been concentrated on improving phylogenetic analyses, with the resulting situation that data acquisition remains a time-consuming, manual, and error-prone process for large-scale multipartition phylogenetic studies. The limitations of this practice were revealed in a recent survey of 595 fungal phylogenetic trees found in 560 articles published from 1990 to 2003 (Lutzoni et al., 2004). Despite the enormous progress in the fields of genomics and information technology, as well as new theoretical and technical innovations derived from multidisciplinary research, 82% of these trees were based on single-locus data sets. Although the number of species included in published trees has generally increased over time, most studies have included fewer than 100 species, with an overall mean of 34.2 (±2.3) species per study. Few studies have focused on resolving relationships among orders of Fungi, with 354 of 595 (60%) trees conveying relationships within single orders.

One of the main reasons for this lack of performance in generating multilocus phylogenies for a large and broad set of taxa is the uncoordinated nature of data sequencing. For example, before AFTOL started depositing sequences in GenBank (2003), a search of GenBank revealed a total of 13,467 sequences for the Fungi. The maximum number of unique taxa with two loci sequenced at that time was 1010. The two loci were the nuclear small and large subunit ribosomal RNA genes (nucSSU and nucLSU rDNA). When sequences <600 bp

were excluded, all the definition line errors in GenBank were dealt with, and all nucSSU and nucLSU rDNA sequences generated by AFTOL by the end of 2003 were added to this compilation, this two-locus data set included 573 taxa. However, most of these taxa are members of the Ascomycota and Basidiomycota, with very few representatives from the other three phyla of the Fungi (Chytridiomycota, Glomeromycota, and Zygomycota). This number dropped to 253 taxa when adding the next most sequenced locus (mitSSU), and only the Ascomycota and Basidiomycota were represented in this data set. The latter was true for all data sets including more than two loci. The maximum number of taxa for which nucSSU, nucLSU, and a protein-coding gene (*RPB*2) were available was only 161 across the Fungi. The largest four-locus data set (nucSSU, nucLSU, mitSSU, and *RPB*2) available at that time included 103 species (Lutzoni et al., 2004). Other challenges faced when generating large multilocus phylogenetic trees across a broad diversity of taxa have been the lack of standardization and automation in the various data acquisition processes.

Assembling molecular data sets for phylogenetic studies once the raw data are generated by automated sequencers usually involves nucleotide base calling; assembly of overlapping reads into a contiguous sequence (contig); BLAST searches to confirm identity of sequences; alignments; delimitation of ambiguously aligned regions (Lutzoni et al., 2000), introns, and other indels; exclusion or recoding of these delimited regions; detection of significant conflicts among data partitions; concatenation of sequences from different loci for taxa shared by these single-locus alignments for multiple combination of loci (e.g., all two-, three-, and four-locus combination of loci); and adjustment of boundaries of ambiguously aligned regions for these different set of taxa associated with various combination of loci. Many of these steps are highly repetitive and their automation is relatively straightforward. However, in order to verify results and correct errors and inaccuracies, manual interaction is necessary.

Several software solutions exist for databasing and analyzing finalized gene sequences (e.g., ARB [Ludwig et al., 2004], TaxMan [www.nematodes.org/bioinformatics/Taxman/index.shtml]), for fully or partially automating the various steps of raw sequence analysis and sequence assembly (e.g., MAGIC-SPP [Liang et al., 2006], LUCY [Chou and Holmes, 2001]), or for processing ESTs (e.g., StackPack [www.egenetics.com/stackpack.html], PartiGene [Parkinson et al., 2004]). The focus of many of these software packages is on the rapid processing of large amounts of raw data, usually obtained through high-throughput DNA sequencing (EST sequencing, shotgun sequencing), rather than on the accuracy of each individual base of an individually sequenced gene. For a phylogenetic analysis, however, sequence accuracy is a major concern. Even for more complex phylogenetic projects that include multiple gene sequences for a large number of taxa (e.g., AFTOL), the actual sequence processing speed of base-calling or contig-assembly software is of less importance than the time the individual researcher spends on correcting sequences and performing other repetitive routine tasks related to gene sequence assembly. This time factor has been steadily increasing during recent years as the number of sequences generated per individual and per unit of time is growing. As a consequence, a system that automates the steps related to sequence assembly and sequence verification must give the user the possibility to efficiently check and correct intermediate and final results in order to ensure the highest possible degree of sequence accuracy. In addition, as sequencing often involves the use of highly "personalized" protocols and strategies, an automated raw sequence analysis as part of the computational infrastructure should not be mandatory but optional for the user, with the possibility to intervene. Consequently, any subsequent steps in sequence processing (alignment, etc.) have to have the capacity to handle both sequences resulting from an automated pipeline as well as manually entered sequences without discrimination.

The challenge of storing phylogenetic data increases drastically when research projects involve the collaboration of several laboratories and multiple sequence authors. Taxa initially intended for sequencing are likely to change during the lifetime of the project, and so are the targeted genes. Furthermore, if users can continuously provide different kinds of data (vouchers, raw sequence reads, finalized gene sequences), a central database that keeps track of all participants' entries is essential. In this context, storing molecular sequence data is only the most basic component of a database. It also provides a communication framework for the coordination and monitoring of ongoing research activities.

Access to such a system needs to be flexible, and preferably independent from any particular software and/or operating system, as the various computational environments on the users' sides are both already established and heterogeneous, with different users using different software. Depending on the given project's publishing policy, some of the data can be regarded as public (e.g., finalized gene sequences), whereas other information, such as intermediate results, will more likely be restricted to the participating researchers. Different portals (websites) to different aspects of the data and/or different levels of authorization must be established to allow data access and manipulation at the appropriate levels. Although the database contains data at different levels of completeness, users can then quickly select finalized data of a given subset of taxa and/or genes to perform a snapshot analysis of the available data at any time.

WASABI was designed to overcome limitations associated with the lack of automation, communication, coordination, and standardization intrinsic to the data acquisition process, by providing an online service to a worldwide community of systematists working on a large-scale phylogenetic problem. It consists of series of Python scripts (www.python.org) uniting publicly available programs to form an interactive bioinformatic pipeline for data storage and for high throughput of

high-quality large concatenated data sets. It provides a computational framework to facilitate the tasks related to the generation, storage, and analysis of molecular sequence data in a multiuser context, without restricting the user's ability to intervene with the automated process at any level (Fig. 1). WASABI's relational database serves as a central data storage for the AFTOL community and helps to coordinate the data generation for the project. The database includes a Web interface that allows easy access for the participants and the public and facilitates the immediate release of finished sequences to the public. In addition, WASABI features an automated pipeline for the processing of single-read sequence chromatograms into contig sequences, including a BLAST search and multiple quality checks across the entire WASABI pipeline. Finalized gene sequences for each genetic locus are continuously added to their respective alignments and subjected to a screening designed to detect significant topological conflicts among loci, which facilitates analyses from a "snapshot" of the project's data and allows the data sampling strategies to evolve. WASABI also provides an interface to perform a local BLAST search of single or multiple sequences against the AFTOL sequence database and facilitates the download of sequences from GenBank as well as the upload of newly generated sequences to GenBank (Fig. 1).

WASABI Components

The WASABI infrastructure consists of three main components corresponding to the three main operations

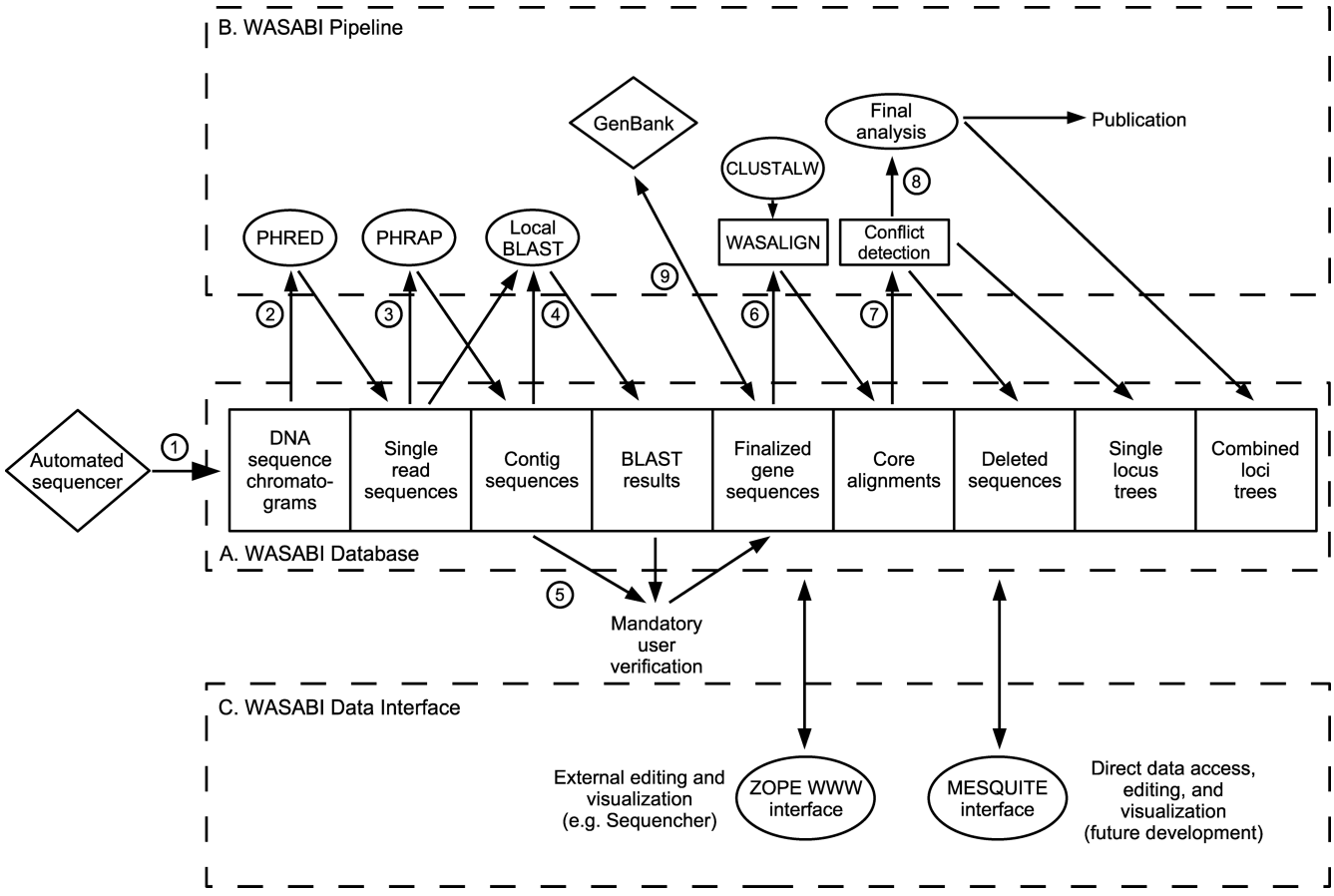

FIGURE 1. WASABI components and data flow (simplified). Arrows show data flow between WASABI components (rectangles), external software (ovals), and external services (diamonds). A, The WASABI database stores all data generated by WASABI, together with all intermediate and final results of the WASABI automated pipeline. B, WASABI pipeline. (1) DNA sequence chromatograms are uploaded from the sequencing facility into the WASABI database. (2) WASABI regularly queries the database for new chromatograms and performs base-calling with PHRED. (3) New single-read sequences are assembled into a contig sequence with PHRAP. (4) Single reads and contigs are subjected to a local BLAST; the results are stored in the WASABI database. (5) Users need to verify sequence accuracy manually before the sequence is considered final. (6) Finalized sequences are automatically aligned to their respective core alignments. (7) Single-gene alignments are tested for congruence. Erroneous sequences that were detected by the topology-based screening criterion and were confirmed by the user are removed from the core alignments and stored in the Deleted Sequences table of the WASABI database, together with an explanation for their removal. (8) Conflict-free single-locus data sets are combined and prepared for phylogenetic analysis. (9) Users of WASABI can select and download sequences from GenBank to be incorporated into the automated pipeline. Finalized sequences generated by WASABI can be uploaded into GenBank upon publication. C. WASABI data interface. Current access to the WASABI database is provided via the World Wide Web using a Zope application server as interface. Future development will allow direct data access, editing, and visualization through MESQUITE.

intrinsic to the data acquisition process for phylogenetic studies—data storing (WASABI database), automated data processing (WASABI pipeline), and data access, visualization, and editing interface (WASABI data interface) (Fig. 1).

Data Storing and Data Interface

The WASABI relational database stores all molecular data generated by AFTOL at all stages of data processing along the WASABI pipeline from sequence data capturing to the generation of multilocus data sets for large numbers of taxa. Specimen voucher information, which participants associated with each sequence and sample raw sequences data (ABI chromatograms), completed gene sequences, local BLAST output, sequencing primer data, results from the screening to detect topological conflicts, alignments, and erroneous sequences. Database access is provided by a password-protected Web interface implemented using a Zope application server (www.zope.org). Registered users are able to view all project data, but individual users are restricted to editing only their own data.

Voucher information is mandatory for all specimens sequenced by AFTOL, and each specimen is identified by a unique identification number (ID). This ID connects all data generated for a given specimen and links it to the specimen's voucher data, including names of participants that provided the samples and did the laboratory work to generate the sequences. DNA sequences can enter the database through the automated data capturing system (e.g., directly from the automated sequencers at Duke University into the WASABI database; Fig. 1), through a manual entry by a participant (e.g., unpublished DNA sequence generated from a different institution), or can be uploaded directly from GenBank by specifying the GenBank Accession Number or Identifier.

In addition to the regular voucher specimen entry procedure, users have the possibility to enter "GenBank placeholders" without submitting the otherwise mandatory voucher information. GenBank placeholders are for specimens that are not sequenced by AFTOL but that are of interest for the project and for which sequence data is available from GenBank; e.g., species sequenced in whole genome projects. Both procedures create a new entry in the Web page showing all the species for which the work is in progress or completed. Once this new entry has been established, the database is now ready to receive raw sequences for this taxon.

Automated Data Processing

WASABI provides an automated workflow to assemble completed gene sequences from raw sequence electrophoreograms ("ABI traces") for the targeted loci and to verify these sequences using a local BLAST. In order to use the automated contig assembly of WASABI, users must first provide the system with the necessary information to correctly identify and process the sequencing reaction data. For each set of sequencing reactions (typically a 96-well plate), a sample sheet is submitted via the Web interface of WASABI, which contains the unique AFTOL-ID of the specimen, the primer-ID for the targeted gene, and an optional name for each of the wells containing a sequencing reaction. One advantage and requirement of this procedure is the use of unique ID numbers for all primers used for a specific large-scale multiuser collaborative project without the need for the standardization of the primer names. Each researcher can enter a new primer in the database by providing a selected standardized name of the targeted gene or loci, their chosen name of this new primer, the primer's nucleotide sequence, its direction (forward versus reverse), and the name of the author entering this new primer. The database automatically gives a unique number to this new primer entry that all users can now use when filling the online form for submitting a 96-well plate.

The physical sequencing plate is transferred (mailed for laboratories outside of Duke University) to the Duke Center for Evolutionary Genomics. Once sequencing reactions have been processed by the automated sequencers, the resulting ABI traces are automatically uploaded into the WASABI database. WASABI queries the database in 10-min intervals for new data and starts the automated processing of new traces. Each sequencing reaction can be identified by the information from the previously submitted sampling sheets. Only the user who submitted the sequencing reactions is allowed to access the intermediate results.

Base-calling with PHRED.—WASABI uses PHRED (www.phrap.org) to perform the base-calling on the sequence chromatograms. PHRED assigns a quality score to each base of the resulting sequences, based on various read characteristics of the trace data (Ewing et al., 1998; Ewing and Green, 1998). For a sequence read to be considered successful, each base of the sequence must meet a minimum quality score, and the total sequence length must be at least 20 bases. Should these conditions not be met, a sequence read is considered as failed (about 16% of all processed reads for the AFTOL project did not pass this initial quality check) and will not be included in subsequent steps of the analysis but instead transferred to a designated table for low-quality sequence reads. Successful reads of sufficient quality are available for the assembly of contig sequences with PHRAP.

Contig assembly with PHRAP.—Sequence reads that meet all quality standards are assembled into contig sequences using PHRAP (www.phrap.org). The AFTOL-ID (for each DNA sample) and the primer-IDs are used to identify all available sequences for a specific combination of taxon and genetic locus. This includes sequence reads for the respective taxon/gene combination that are already present in the database from previous sequencing runs, which ensures that all available information is used to create a contig sequence.

PHRAP also assigns quality scores to each base of the contig sequence during the assembly process. The quality scores for the contig sequence for each position are examined, and, if below a given threshold score, the contig sequence is rejected. The sequence is also automatically rejected if it is not of sufficient length or if multiple

contig sequences have been assembled for a given taxon-gene combination. Due to the need for absolute sequence correctness in a phylogenetic context, the conditions for the quality check for the AFTOL project were set to be very rigorous, and consequently most contig sequences need to be manually inspected by their authors. Should a sequence author at any time decide to delete a single-read sequence used during a contig assembly, the contig sequence becomes invalid, and a new round of contig analysis, now omitting the deleted read, is automatically started.

BLAST of single-read and contig sequences.—To verify sequence identity, both single-read sequences and contig sequences are subjected to a local BLAST search. Potential cloning vector sequences are masked out before the BLAST search. If the phylum of the top five hits returned from the BLAST does not match the phylum given in the voucher information of the targeted specimen, the BLAST verification is considered as failed. This is one of many checkpoints of WASABI. This first checkpoint is meant to detect obvious lab errors (e.g., mislabeling, contaminated cultures, or DNA samples).

For the local BLAST, WASABI uses its own custom-built database, which is composed of three parts: (1) a random 20% subsample of all nonfungal sequences available from GenBank; (2) all fungal sequences with taxonomic names available from GenBank, which is semiautomatically updated on a monthly basis; and (3) all finalized sequences from the AFTOL database, including sequences not yet deposited in GenBank. This database is considerably smaller than the full nucleotide database from NCBI, yet it allows for 8 to 9 times faster screening with the local BLAST. In our experience, the 20% sample of nonfungal sequences is sufficient to estimate from the BLAST results whether a given sequence is of fungal origin or not. Nonfungal sequences will match within the 20% random sample of nonfungal sequences. In the current implementation of WASABI for AFTOL, the exact identity of nonfungal organisms is not of major importance, and the match within nonfungal organisms clearly identifies them as contaminations. Because this local BLAST database includes all the sequences generated by AFTOL that are not yet available in GenBank, it finds the closest matches within the fungi in order to identify possible fungal contaminations and identifies more accurately unknown fungi from environmental PCR studies. Other sequencing projects using WASABI may implement different strategies for creating customized blast databases or, provided that sufficient computational power is available, use the full nucleotide database from GenBank.

E-mail notification and access to results.—Upon completion of the base-calling, contig assembly, and BLAST, the sequence authors receive an e-mail that summarizes the results and provides the URLs to directly access them. The output of the local BLAST can be reviewed online; all other results can be downloaded and examined using Sequencher (Gene Codes Corporation, Ann Arbor, Michigan) or Consed (Gordon, 2004). Authors can manually edit their contig sequences and override any decisions made by PHRED or PHRAP and resubmit the edited DNA sequence to the database. Original results from the WASABI pipeline remain accessible to their authors even when a manually edited version of this sequence is inserted in the database. Contig sequences that passed all quality checks after the initial automated assembly enter the gene tables automatically but need to be verified by their authors before they can be used for subsequent analyses. Once a sequence has been verified by the author or entered manually, it will not be altered by the automated pipeline of WASABI. In addition, because primer information is part of the contig assembly, WASABI automatically stores all successful primers for each locus/taxon combinations and makes this information available to users.

Alignment.—The algorithm we developed for WASABI combines the profile alignment procedure with the principle of a new block-wise alignment method. The overarching principle of this algorithm follows the profile alignment procedure, where recently generated sequences are added to a previously established alignment. For each locus, WASABI stores a core alignment, which includes all aligned sequences included in the AFTOL database. New gene sequences, either previously assembled by WASABI or manually entered by AFTOL participants, are aligned to these continuously growing core-alignments. The core alignments are stored in NEXUS format (Maddison et al., 1997) in the database and include character sets with information about non-alignable regions and indels (e.g., introns, hypervariable regions with multiple short indels). This information is used to align new sequences to the core alignments more effectively using an algorithm that we developed for WASABI, which avoids any attempts to align regions that cannot be successfully aligned.

The initial core alignment for a locus can be aligned manually or automatically (e.g., ClustalW; Thompson et al., 1994), followed by a thorough manual inspection of the alignment to improve the alignment; i.e., by delimiting regions that cannot be aligned unequivocally (following the protocol by Lutzoni et al., 2000), as well as verifying the delimitations of various types of introns and indels in general. For the ribosomal RNA genes, the alignment is also improved by using the secondary structure and patterns of compensatory changes (Kjer, 1995) as an additional guide for assessing positional homology.

To align a new sequence to this core alignment with its various delimited regions, the WASABI alignment algorithm (WASALIGN) starts with the longest alignable block and aligns it to the corresponding part of the new sequence (Fig. 2). This can be done using an external alignment routine. The current implementation of the WASABI alignment procedure supports both ClustalW and the Biopython (www.biopython.org) built-in alignment module but can be adapted to support other available alignment programs. After the first and longest block has been aligned, both the sequence and the core-alignment are divided into two partitions, simplifying the subsequent alignment tasks of this procedure. The

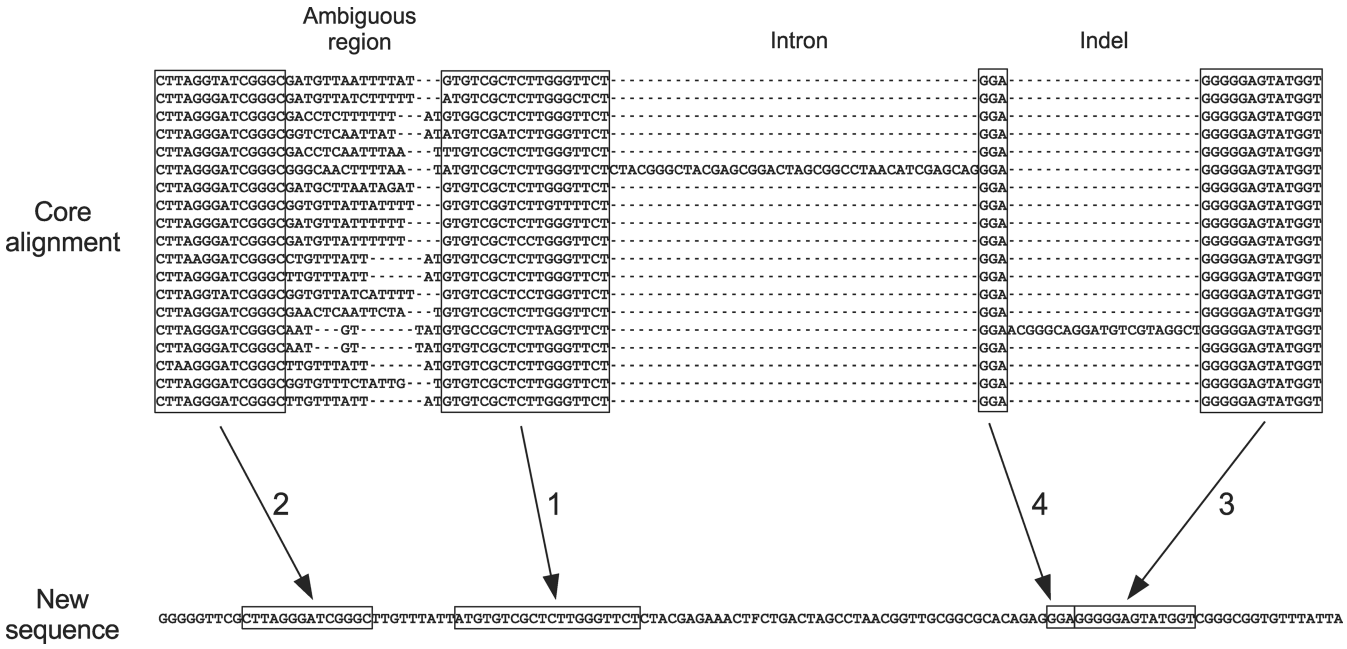

FIGURE 2. Alignment algorithm of WASALIGN. The information about nonalignable regions (ambiguous regions, introns) specified in the core alignment is used to delimit alignable blocks. (1) The largest alignable region is aligned to the new sequence. (2) As the second largest alignable region is located upstream of the first aligned region, only the upstream part of the new sequence is considered for alignment. (3) Accordingly, the third largest block is aligned downstream of the first aligned region. (4) The smallest region in this example is situated between regions 1 and 3 and is aligned only to the segment between the previously aligned regions 1 and 3 of the new sequence.

alignment of the next longest block from the core alignment is now restricted to the upstream or downstream partition of the new sequence. When the homologous region of the new sequence is aligned to the second longest block, the core alignment and new sequence have now three partitions, further simplifying the overall alignment problem. As the blocks of the core alignment become smaller, so are the remaining parts of the sequence to which they will have to be aligned, resulting in improved alignment efficiency and accuracy (Kauff, unpublished results). During the alignment, gaps are inserted into the alignable blocks of the core alignment as necessary, and the delimitations of the exclusion sets are adjusted. Although this alignment strategy is much slower compared to other available alignment software (mainly due to the current implementation in the Python scripting language, it can take several minutes to align a single sequence to a core alignment, depending on alignment length and number of ambiguous regions, introns, and indels present in the alignment), in WASABI the speed of the automated alignment process is of lesser importance than the time needed by the users to make necessary manual corrections.

Once new sequences have been aligned to the existing core matrices, the resulting alignments will become the new core alignments for future sequence additions. The core alignments are available for download and can be manually adjusted (if needed) with any NEXUS-compatible alignment editor (e.g., MESQUITE [Maddison and Maddison, 2005], SeaView [Galtier et al., 1996]), and reloaded into WASABI. As the quality of the alignments and the delimitation of ambiguous/unambiguous regions is critical for the success of the automated aligning procedure, the ability to modify the core alignments should only be available to a selected group of users and will eventually require the establishment of curators for alignments of each targeted locus. This procedure has the advantage of working like a ratchet where a given alignment is constantly improving and every change recorded. Previous core alignments remain accessible in WASABI.

Multilocus data set assembly and detection of topological conflicts among loci.—Every week, WASABI automatically assembles data sets for the maximum number of taxa shared among a set of loci, for all possible combinations of loci alignments, and screens for significant topological conflicts among genes using the reciprocal 70% bootstrap criterion (Mason-Gamer and Kellogg, 1996). This conflict detection procedure is currently implemented by conducting neighbor-joining bootstrap analyses using maximum likelihood distances with PAUP* (Swofford, 2002) on each single-gene matrix part of a combination set. Other methods could be used to generate the support values used to detect conflicts among loci. Bayesian methods are difficult to implement in an automated framework (determination of the number of generations necessary to reach stationarity, delimitation of burn-in), and current implementations of Bayesian Markov chain Monte Carlo methods can lead to high posterior probabilities for wrong relationships when internodes are very short (Alfaro et al., 2003; Lewis et al., 2005). In the

initial stages of the development of WASABI, tree search methods using maximum likelihood as the optimization criterion required too much computing time for large data sets. NJ bootstrap (using ML distances) is a conservative approach that seems to generate fewer false positives (Reeb et al., 2004) and has been successfully implemented in previous studies (Lutzoni et al., 2004; Reeb et al., 2004; Miadlikowska et al., 2006). WASABI can easily be adapted to implement other phylogenetic methods (e.g., using RAxML; Stamatakis, 2006; or Garli Zwickl, 2006) for the detection of topological conflicts.

The current implementation of WASABI assumes that there is a significant topological conflict when a group of taxa is supported (≥70% bootstrap) as monophyletic in one tree but supported as being nonmonophyletic in another single-locus tree. Rather than estimating incongruence as an overall quality of the compared trees, this method allows for the possible identification of incongruent taxa. For all conflicting groups, taxa likely to cause the conflict are identified and added as taxon sets to the NEXUS alignments to facilitate their exclusion in subsequent analyses. Users of WASABI need to go through this list of potential conflicting taxa and determine the cause (artifactual versus biological). The final decision whether a taxon is excluded or not from subsequent analyses and alignments is left to the user. If a DNA sequence is wrong (e.g., contaminant sequence, laboratory mishap, etc.), the sequence needs to be excluded from its alleged core alignment and from all future BLAST searches and phylogenetic analyses. In WASABI these erroneous sequences are stored separately (Fig. 1), with a note explaining their exclusion, for future references. Storing these erroneous sequences is important to prevent their reintroduction into the active WASABI database, which could happen if the sequence was acquired from GenBank in the first place, and to distinguish correct and incorrect single read and contig sequences stored in WASABI.

This checkpoint is much more sensitive than the initial BLAST procedure. As part of the AFTOL project, problematic sequences detected this way were usually the result of technical errors in the laboratory or contamination problems when PCR was performed directly on field specimens. Because the assembly of the fungal tree of life requires a broad sampling across all Fungi, conflicts due to recombination, lineage sorting, and horizontal transfer, for example, are assumed to be rare in these data sets.

Data sets without conflicts are then concatenated and ready for final analyses. These concatenated multilocus data sets are generated on a weekly basis in WASABI and available to the AFTOL participants for phylogenetic analyses. These are seen as the main products of WASABI. Preliminary phylogenetic analyses of these multilocus data sets (e.g., using RAxML and Garli) will be conducted automatically on a weekly basis on a cluster of computers and the resulting tree stored in the WASABI database. The latest preliminary trees with bootstrap values would be posted online for AFTOL participants.

Public Access

The public AFTOL website (www.aftol.org) offers public access for unregistered users to the final sequence data and taxon voucher information. Primer sequences and lists of primers that have been successfully used for the sequencing of AFTOL specimens are also publicly available from the public website.

Other Services Provided

BLAST.—Both registered and unregistered users can blast arbitrary sequence data against the sequences in the AFTOL database. The BLAST interface of WASABI allows the user to manually enter sequences or to upload a FASTA file containing multiple sequences. The results are returned as HTML files and can be viewed directly on the website or downloaded and opened locally in a Web browser. Similar to the standard output of an NCBI Web BLAST, all taxon references are HTML links to the WASABI database, which allows easy access to further detailed information about the particular sequence. This service is accessible at aftol.biology.duke.edu/pub/blast/blastUpload.

GenBank submissions.—To facilitate the submission of newly generated sequences to GenBank, WASABI can generate files in the format used by the program Sequin, the NCBI software required for Genbank submissions (www.ncbi.nlm.nih.gov/Sequin). Users can choose to generate Sequin-formatted files of all non-submitted sequences for which they are responsible or manually select a set of sequences. Using Sequin, new sequence data can be annotated and/or completed with information not available in the WASABI database prior to submission to GenBank.

Implementation

All scripts are written in Python (www.python.org). WASABI uses various modules from Biopython (www.biopython.org) to facilitate processing of sequence data and interaction with GenBank, and several modules of WASABI have been contributed to Biopython. Web services are managed by a Zope application server (www.zope.org) with Apache (www.apache.org) as HTTP server. The SQL database is implemented using PostgreSQL (www.postgresql.org) with Psycopg (initd.org/projects/psycopg1) as database adapter for Python. Base-calling and contig assembly are done with PHRED and PHRAP (www.phrap.org). The local BLAST was implemented using the NCBI toolkit software packages provided from NCBI (www.ncbi.nlm.nih.gov/BLAST).

The current implementation of WASABI runs under Linux on one dual-CPU PC used as Web server and database server and one single-CPU server executing the automated analyses and BLAST searches.

Future Development

WASABI is an ongoing project and under continuous development. The current architecture of WASABI was developed in close collaboration with AFTOL and

tailored to the specific needs of this project. However, the open architecture of WASABI makes adaptations for future extensions or for the specific needs of a given phylogenetic project particularly easy, and upcoming versions of WASABI will be extended to flexibly interact with preexisting sequence databases.

Establishing the MESQUITE-WASABI system.—Similar to WASABI, MESQUITE (Maddison and Maddison, 2005; www.mesquiteproject.org) has incorporated a sequence contig assembly function using PHRED and PHRAP in its new Chromaseq module. The developers of MESQUITE, a software package for evolutionary biology, and WASABI are cooperating to integrate the functionality of both systems. The chromatogram editor of MESQUITE (under development) can then be used to display the sequence reads, contigs, and associated chromatograms housed in WASABI's database. Users of WASABI can then comfortably access intermediate results of the automated data management pipeline using the graphical user interface of MESQUITE. After necessary adjustments have been made, the data can be resubmitted to WASABI, where any changes trigger new rounds of processing of the altered data. The connectivity between WASABI with MESQUITE will be further extended to other paths of the WASABI workflow for visualization and editing of alignments, visualization of conflict detections among data partitions, and interactions with phylogenetic searches. One of the ultimate goals of the MESQUITE-WASABI system will be to prepare high quality, large-scale, multilocus datasets for the latest analytical programs developed by CIPRES (www.phylo.org/architecture.html) and to further develop a closer interaction between WASABI, MESQUITE, and CIPRES.

The recent development in phylogenetics and phylogenetic software shows a clear trend away from interactive, menu-driven software and manual data preparation. Technologies associated with DNA sequencing are constantly improving (e.g., nanopore technologies and pyrosequencing), with the result that the lack of bioinformatic tools for sequence data management has become a major bottleneck in large-scale phylogenetic studies. Ways of analyzing molecular data change quickly, and the ever increasing speed of computers allows for a more flexible approach to data analysis, where an array of different programs and methodological approaches can be explored in a comparatively short amount of time. It is therefore essential that the data are processed in a way that enables an efficient communication between the numerous software applications that are involved in the many steps from the first base-calling to the final analysis. WASABI is a step in this direction, serving as a flexible and extensible interface between data generation, storage, and analysis.

ACKNOWLEDGMENTS

We thank all AFTOL participants and collaborators for comments and suggestions during the development of WASABI, David Maddison for guidance in future development of WASABI in concert with MESQUITE and CIPRES, Lisa Bukovnik from the Duke sequencing facility, Bill Rankin, John B. Pormann, and Sean Dilda from the Duke Shared Cluster Resource, all lab members, especially Molly McMullen, for their comments and their help with the manuscript. WASABI was developed through an NSF-ATOL grant (DEB-0228668) to F. Lutzoni and Rytas Vilgalys.

REFERENCES

Alfaro, M. E., S. Zoller, and F. Lutzoni. 2003. Bayes or bootstrap? A simulation study comparing the performance of Bayesian Markov chain Monte Carlo sampling and bootstrapping in assessing phylogenetic confidence. Mol. Biol. Evol. 20:255–266.

Chou, H.-H., and M. H. Holmes. 2001. DNA sequence quality trimming and vector removal. Bioinformatics 17:1093–1104.

Ewing, B., and P. Green. 1998. Base-calling of automated sequencer traces using *Phred*. II. Error probabilities. Genome Res. 8:186–194.

Ewing, B., L. Hillier, M. Wendl, and P. Green. 1998. Base-calling of automated sequencer traces using *Phred*. I. Accuracy assessment. Genome Res. 8:175–185.

Galtier, N., M. Guy, and C. Gautier. 1996. SeaView and Phylo-Win, two graphic tools for sequence alignment and molecular phylogeny. Comput. Appl. Biosci. 9:49–57.

Gordon, D. 2004. Viewing and editing assembled sequences using Consed. Section 11.2.1–11.2.43 *in* Current protocols in bioinformatics (A. D. Baxevanis and D. B. Davison, eds.). John Wiley & Co., New York.

James T. Y., F. Kauff, C. Schoch, P. B. Matheny, V. Hofstetter, C. J. Cox, G. Celio, C. Gueidan, E. Fraker, J. Miadlikowska, T. Lumbsch, A. Rauhut, V. Reeb, A. E. Arnold, A. Amtoft, J. E. Stajich, K. Hosaka, G.-H. Sung, D. Johnson, B. O'Rourke, M. Binder, J. M. Curtis, J. C. Slot, Z. Wang, A. W. Wilson, A. Schüßler, J. E. Longcore, K. O'Donnell, K. Mozley-Standridge, D. Porter, P. M. Letcher, M. J. Powell, J. W. Taylor, M. M. White, G. W. Griffith, D. R. Davies, J. Sugiyama, A. Y. Rossman, J. D. Rogers, D. H. Pfister, D. Hewitt, K. Hansen, S. Hambleton, R. A. Shoemaker, J. Kohlmeyer, B. Volkmann-Kohlmeyer, R. A. Spotts, M. Serdani, P. W. Crous, K. W. Hughes, K. Matsuura, E. Langer, G. Langer, W. A. Untereiner, R. Lücking, B. Büdel, D. M. Geiser, A. Aptroot, W. R. Buck, M. S. Cole, P. Diederich, C. Printzen, I. Schmitt, M. Schultz, R. Yahr, A. Zavarzin, D. S. Hibbett, F. Lutzoni, D. J. McLaughlin, J. W. Spatafora, R. Vilgalys. 2006. Reconstructing the early evolution of the fungi using a six gene phylogeny. Nature 443:818–822.

Kjer, K. M. 1995. Use of rRNA secondary structure in phylogenetic studies to identify homologous positions: An example of alignment and data presentation from the frogs. Mol. Phylogenet. Evol. 4:314–330.

Lewis, P. O., M. T. Holder, and K. E. Holsinger. 2005. Polytomies and Bayesian phylogenetic inference. Syst. Biol. 54:241–253.

Liang, C., F. Sun, H. Wang, J. Qu, R. M. Freeman Jr., L. H. Pratt, and M.-M. Cordonnier-Pratt. 2006. MAGIC-SPP, a database-driven DNA sequence processing package with associated management tools. BMC Bioinformatics 7:115.

Ludwig, W., O. Strunk, R. Westram, L. Richter, H. Meier, Yadhukumar, A. Buchner, T. Lai, S. Steppi, G. Jobb, W. Förster, I. Brettske, S. Gerber, A. W. Ginhart, O. Gross, S. Grumann, S. Hermann, R. Jost, A. König, T. Liss, R. Lüßmann, M. May, B. Nonhoff, B. Reichel, R. Strehlow, A. Stamatakis, N. Stuckmann, A. Vilbig, M. Lenke, T. Ludwig, A. Bode, and K.-H. Schleifer. 2004. ARB: A software environment for sequence data. Nucleic Acids. Res. 32:1363–1371.

Lutzoni, F., F. Kauff, C. J. Cox, D. McLaughlin, G. Celio, B. Dentinger, M. Padamese, D. Hibbett, T. Y. James, E. Baloch, M. Grube, V. Reeb, V. Hofstetter, C. Schoch, A. E. Arnold, J. Miadlikowska, J. Spatafora, D. Johnson, S. Hambleton, M. Crockett, R. Schoemaker, G.-H. Sung, R. Lücking, T. Lumbsch, K. O'Donnell, M. Binder, P. Diederich, D. Ertz, C. Gueidan, K. Hansen, R. C. Harris, K. Hosaka, Y.-W. Lim, B. Matheny, H. Nishida, D. Pfister, J. Rogers, A. Rossman, I. Schmitt, H. Sipman, J. Stone, J. Sugiyama, R. Yahr, and R. Vilgalys. 2004. Assembling the Fungal Tree of Life: Progress, classification, and evolution of subcellular traits. Am. J. Bot. 91:1446–1480.

Lutzoni, F., P. Wagner, V. Reeb, and S. Zoller. 2000. Integrating ambiguously aligned regions of DNA sequences in phylogenetic analyses without violating positional homology. Syst. Biol. 49:628–651.

Maddison, D. R., D. L. Swofford, and W. P. Maddison. 1997. NEXUS: An extensible file format for systematic information. Syst. Biol. 46:590–621.

Maddison, W. P., and D. R. Maddison. 2005. Mesquite: A modular system for evolutionary analysis. Version 1.06. http://mesquiteproject.org.

Mason-Gamer, R., and E. Kellogg. 1996. Testing for phylogenetic conflict among molecular data sets in the tribe Triticeae (Gramineae). Syst. Biol. 45:524–545.

Miadlikowska, J., F. Kauff, V. Hofstetter, E. Fraker, M. Grube, J. Hafellner, V. Reeb, B. P. Hodkinson, M. Kukwa, R. Lücking, G. Hestmark, M. A. G. Otalora, A. Rauhut, B. Büdel, C. Scheidegger, E. Timdal, S. Stenroos, I. Brodo, G. Perlmutter, D. Ertz, P. Diederich, J. C. Lendemer, E. Tripp, R. Yahr, P. May, C. Gueidan, A. E. Arnold, C. Robertson, and F. Lutzoni. 2006. New insights into classification and evolution of the Lecanoromycetes (Pezizomycotina, Ascomycota) from phylogenetic analyses of three ribosomal RNA- and two protein-coding genes. Mycologia 98(6).

Parkinson J, A. Anthony, J. Wasmuth, R. Schmid, A. Hedley, and M. Blaxter. 2004. PartiGene—constructing partial genomes. Bioinformatics 20:1398–1404.

Perry, S., and D. Vieglais. 2006. WASABI: Web Application for the Semantic Architecture of Biodiversity Informatics. Proceedings of the TDWG 2006. http://www.tdwg.org/proceedings/index.php/proceedings/article/view/56.

Reeb, V., F. Lutzoni, and C. Roux. 2004. Contribution of RPB2 to multilocus phylogenetic studies of the euascomycetes (Pezizomycotina, Fungi) with special emphasis on the lichen-forming Acarosporaceae and evolution of polyspory. Mol. Phylogenet. Evol. 32: 1036–1060.

Spatafora, J. W. 2005. Assembling the fungal tree of life (AFTOL). Mycol. Res. 109: 755–756.

Stamatakis, A. 2006. RAxML-VI-HPC: Maximum likelihood-based analyses with thousands of taxa and mixed models. Bioinformatics 22: 2688–2690.

Swofford, D. L. 2002. PAUP*: Phylogenetic analysis using parsimony (*and other methods). Version 4. Sinauer Associates, Sunderland, Massachusetts.

Thompson, J. D., D. G. Higgins, and T. J. Gibson. 1994. ClustalW: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22: 4673–4680.

Zwickl, D. J. 2006. Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion. PhD dissertation, The University of Texas at Austin. http://www.bio.utexas.edu/faculty/antisense/garli/Garli.html.

First submitted 8 August 2006; reviews returned 18 October 2006; final acceptance 28 November 2006
Associate Editor: Rod Page

Syst. Biol. 56(3):531–539, 2007

ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150701424546

Taxonomy in a Changing World: Seeking Solutions for a Science in Crisis

INGI AGNARSSON[1,3,4] AND MATJAŽ KUNTNER[2,3]

[1]*Departments of Zoology/Botany, University of British Columbia, 2370-6270 University Boulevard, Vancouver, BC, V6T 1Z4, Canada; E-mail: iagnarsson@gmail.com*
[2]*Institute of Biology, Scientific Research Centre of the Slovenian Academy of Sciences and Arts, Novi trg 2, P. O. Box 306, SI-1001 Ljubljana, Slovenia; E-mail: kuntner@gmail.com*
[3]*Department of Entomology, National Museum of Natural History, NHB-105, Smithsonian Institution, P.O. Box 37012, Washington DC, 20013-7012, USA*
[4]*Current Address: Department of Biology, University of Akron, Akron, OH 44325–3908, USA*

One of the fundamental quests of biology is learning what organisms inhabit the earth. To date approximately 2 million species have been described, with realistic estimates of actual diversity ranging from 4 to 12 million (Stork, 1997; Reaka-Kudla et al., 1997). But while species are disappearing at an ever increasing rate (Pimm and Raven, 2000; Thomas et al., 2004), species discovery and description—taxonomy—is facing a crisis (Wilson, 2004; Wheeler, 2004). Overcoming this "taxonomic impediment" (Rodman and Cody, 2003) is the primary goal of the ambitious and ongoing NSF PEET (Partnerships for Enhancing Expertise in Taxonomy) initiative (NSF, 1994), which has enjoyed much success in training a new generation of taxonomists (Rodman and Cody, 2003). To help estimate the impact of the NSF-PEET initiative and the status of taxonomy, we surveyed the trainees from the 1995 and 1997 NSF-PEET cohorts. PEET meetings have optimistically labeled the program as the *renaissance* of taxonomy (see also Wheeler, 2004). But as many PEET alumni (peetsters) are experiencing, taxonomic expertise is rarely required, or even relevant, when it comes to securing a job, especially in academia. Furthermore, most top-ranking evolutionary journals do not consider taxonomic revisions, and only allow species descriptions in exceptional cases of certain high-profile fossils and mammals (e.g., Jones et al., 2005; Gess et al., 2006). Further, some lower ranking journals reject taxonomic descriptions unless in a paper on a broader subject (e.g., the *Journal of Zoological Systematics and Evolutionary Research*; see author guidelines at http://www.blackwellpublishing.com/journals/jzs). Journals focusing on taxonomy typically have low measured impact, even the new and vibrant, rapid and interactive *Zootaxa*, which is enjoying an extraordinary and unprecedented growth among scientific journals and can be characterized as a "mega-journal" (Zhang, 2006; unofficial IF 2005 = 0.45). Taxonomic descriptions are—not necessarily by fact (see below), but by convention—low-impact scientific publications, barring those of newly discovered bird species, large mammals, or certain fossils.

Here we argue that an easily corrected mismeasure of the scientific impact of taxonomy—a convention not