BMC Bioinformatics

Software

Open Access

AIR: A batch-oriented web program package for construction of supermatrices ready for phylogenomic analyses

Surendra Kumar[1], Åsmund Skjæveland[1], Russell JS Orr[1], Pål Enger[1,2], Torgeir Ruden[2], Bjørn-Helge Mevik[2], Fabien Burki[3], Andreas Botnen[2] and Kamran Shalchian-Tabrizi*[1]

Address: [1]Microbial Evolution Research Group (MERG), Department of Biology, University of Oslo, Norway, [2]Centre of Information Technology, University of Oslo, Norway and [3]Department of Botany, University of British Columbia, Vancouver, British Columbia, Canada

Email: Surendra Kumar - surendra.kumar@bio.uio.no; Åsmund Skjæveland - asmund.skjaveland@bio.uio.no; Russell JS Orr - russell.orr@bio.uio.no; Pål Enger - pal.enger@usit.uio.no; Torgeir Ruden - t.a.ruden@usit.uio.no; Bjørn-Helge Mevik - b.h.mevik@usit.uio.no; Fabien Burki - burkif@interchange.ubc.ca; Andreas Botnen - andreas.botnen@gmail.com; Kamran Shalchian-Tabrizi* - Kamran@bio.uio.no

* Corresponding author

Published: 28 October 2009

BMC Bioinformatics 2009, **10**:357 doi:10.1186/1471-2105-10-357

Received: 21 April 2009
Accepted: 28 October 2009

This article is available from: http://www.biomedcentral.com/1471-2105/10/357

Abstract

Background: Large multigene sequence alignments have over recent years been increasingly employed for phylogenomic reconstruction of the eukaryote tree of life. Such supermatrices of sequence data are preferred over single gene alignments as they contain vastly more information about ancient sequence characteristics, and are thus more suitable for resolving deeply diverging relationships. However, as alignments are expanded, increasingly numbers of sites with misleading phylogenetic information are also added. Therefore, a major goal in phylogenomic analyses is to maximize the ratio of information to noise; this can be achieved by the reduction of fast evolving sites.

Results: Here we present a batch-oriented web-based program package, named AIR that allows 1) transformation of several single genes to one multigene alignment, 2) identification of evolutionary rates in multigene alignments and 3) removal of fast evolving sites. These three processes can be done with the programs AIR-**A**ppender, AIR-**I**dentifier, and AIR-**R**emover (AIR), which can be used independently or in a semi-automated pipeline. AIR produces user-friendly output files with filtered and non-filtered alignments where residues are colored according to their evolutionary rates. Other bioinformatics applications linked to the AIR package are available at the Bioportal http://www.bioportal.uio.no, University of Oslo; together these greatly improve the flexibility, efficiency and quality of phylogenomic analyses.

Conclusion: The AIR program package allows for efficient creation of multigene alignments and better assessment of evolutionary rates in sequence alignments. Removing fast evolving sites with the AIR programs has been employed in several recent phylogenomic analyses resulting in improved phylogenetic resolution and increased statistical support for branching patterns among the early diverging eukaryotes.

Background

A well-resolved phylogenetic tree demonstrating the relationships between species is one of the most important goals in evolutionary biology, and the fundament for comparative studies in many fields in life science. Multiple gene sequence data is increasingly being used to resolve phylogenetic relationships, and frequently more than 50 genes are being inferred to address key questions about the early evolution of eukaryotes [1-8]. Recent studies have for instance shown support for the grouping of known eukaryotes into a handful of supergroups [2,5,9-15]. The main reason for constructing multigene data instead of using single gene data in phylogenetic reconstruction is to collect enough information to improve the phylogenetic signal [9,16]. Accordingly, as the number of genes increases, the tendency is that phylogenetic relationships are better resolved and receive higher statistical support [2,5,16-18]. However, simply adding genes to an alignment to increase statistical support does not necessarily lead to more accurate results; inconsistencies in datasets may adversely lead to higher support for an incorrect topology. Reducing such stochastic errors is an important step in improving the phylogenetic resolution of the sequence data [16,19-21]. Consistency in the data may be improved by the removal of the fastest evolving sites; as such sites may have over-representation of substitution saturation causing homoplasies [22,23]. However, so far only a few bioinformatics program has been reported that allows for the concatenation of multiple single gene alignment files, identification of fast evolving sites and removal of fast evolving sites in accordance with the users needs.

Here we present a bioinformatics package, named AIR that combines all these possibilities. AIR is divided into three applications: AIR-**A**ppender, AIR-**I**dentifier and AIR-**R**emover (Figure 1). AIR-Appender performs separate processing of data by appending single gene alignment files to a multi-gene alignment. AIR-Identifier identifies fast evolving sites by calculating site-rates, and AIR-Remover removes fast evolving sites from an alignment. The AIR programs are interlinked with other applications useful in the field of phylogenomics (i.e., multi-gene BLAST, contig assembly of Sanger and 454 sequences, alignment and phylogeny) through the Bioportal at the University of Oslo.

Implementation

The AIR package is implemented on the Bioportal at the University of Oslo. The Bioportal is a web-based bioinformatics service freely available to academic users at the following URL: http://www.bioportal.uio.no/. The Bioportal uses SQL for maintaining information about users, files, databases, and jobs. The Bioportal resources are deployed on Linux with Apache HTTP server 2.2. The critical scripts to maintain the Bioportal, e.g. cron jobs scripts and post-processing scripts, are written in Perl v5.8, and python 2.3. The web-interface for all available applications on Bioportal is written in PHP 4.3.

Each user of the Bioportal has access to several file directories and file administration functions. All files used as input for analyses are stored in project folders defined by the users. Once the user has created a project folder they can upload data-files into its respective project folders. The user can then use the web interface created for each application on Bioportal to select their files, applications (here for example AIR-Appender, AIR-Identifier, or AIR-Remover) and parameter settings. For each analysis a working folder is created in the working directory 'job admin'. A 'copy home' function in the 'job admin' can be used to transfer files from working directories to project folders; hence result files from one process can be used as input files in subsequent analyses, and to link different applications in a semi-automated pipeline. For instance, alignments made by MAFFT [24] can be used for phylogenetic analyses by one of the available phylogenetic programs e.g. RAxML, Treefinder or MrBayes [25-27]. The Bioportal tutorial is available at the Bioportal website.

All successfully submitted Bioportal jobs are run in the background, the execution time of each process varies dependent on the file size and the nature of the selected applications. To keep track of the status of submitted jobs a manager module has been developed on the Bioportal; this updates the users about the current status of all jobs. Upon completion the results are returned to the respective working directory where files can then be downloaded in a compressed 'zip' format.

Currently the Bioportal is the largest high performance-computing environment in Norway. The available computer resources are 320 dedicated cores on the TITAN cluster at the University of Oslo. In addition, the Bioportal has access to all free or idle TITAN cores if needed (4000 at present). The TITAN cluster has LINUX nodes with 16 gigabytes of memory and 2× quadcore CPUs or 2× dual-core CPUs.

Results

Appending single gene alignments

AIR-Appender merges multiple single gene alignment files into one major multigene alignment; the program looks for species with identical names and subsequently merges these. If any of the single gene alignments are lacking taxa in relation to one another, the program will automatically replace the missing data with question marks '?'. The junction between genes will be marked with double hyphen for easy identification of the sequence borders. The resulting output of AIR-Appender is a single FASTA and PAML

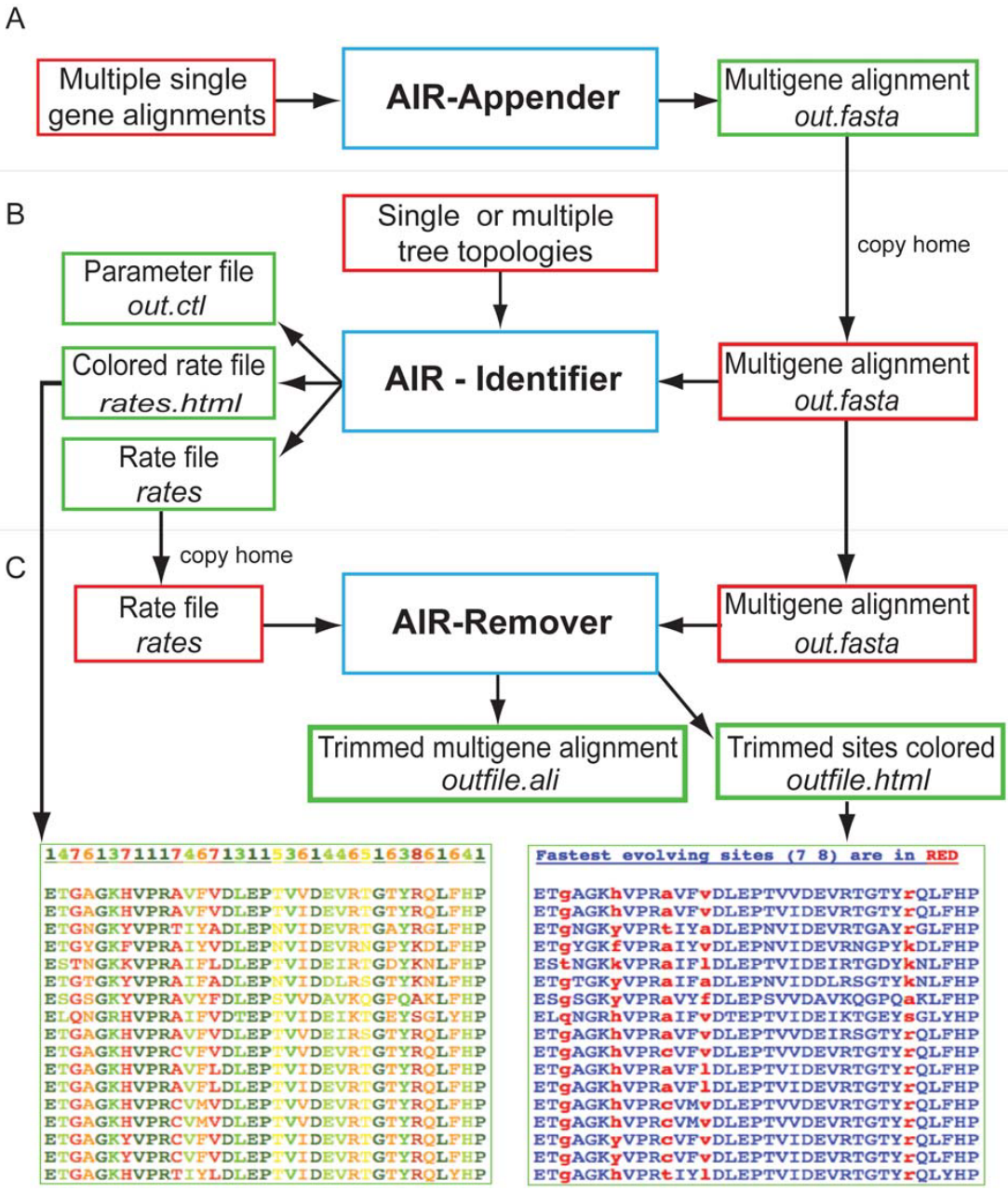

Figure 1

Overview of AIR-package. Overview of the functionalities and programs in the AIR-package installed on the Bioportal: The colored boxes depict input files (red), output files (green), and the AIR programs (Blue). Texts in *Italics* depict the filename and respective extension of output files of AIR programs. A) AIR-Appender uses several single gene alignments for construction of a multigene alignment. B) AIR-Identifier uses the output file from AIR-Appender and file containing one or more phylogenetic trees for calculating site rates and rate categories. C. AIR-Remover deletes fast evolving sites according to settings defined by the user. The output files from each of the AIR programs can be used in subsequent analysis by copying the files from the work directory to project folder on the Bioportal using the *copy home* function. Five main output files are produced by AIR. In which two are graphical html files with information about site rates and fast evolving sites (*rates.html*), and sites removed from the alignment (*outfile.html*). File 'rates.html' shows the rate categories as different colors (up to 8 categories), while '*outfile.html*' shows the removed sites in red color (e.g. category 7 and 8 removed are shown in red), and rest sites in blue. Files namely '*rates*' and '*out.ctl*' are produced by PAML programs, which are implemented in AIR-Identifier. While '*outfile.ali*' is the multigene alignment with fast evolving sites removed.

formatted file containing the multiple gene alignment (*out.fasta* in Figure 1); this can be used for downstream processing with AIR-Identifier (or other programs available on the Bioportal) or downloaded to a local computer as a compressed zip file.

Identifying site rate

After the user has made the multi-gene sequence file, site-rates (i.e. posterior mean values) can then be identified for nucleotides, codons and amino acids sequences with the program AIR-Identifier. AIR-Identifier applies the PAML programs codeml (for codon and amino acid sequences) and baseml (for nucleotide sequences) [28,29]. The control file (*out.ctl* in Figure 1) is critical as it is here that the user defines a set of parameters to be used for estimation of site rates by codeml or baseml. These programs are usually only available via the command line, and thus setting parameters for a successful run can be a cumbersome task. We have therefore developed AIR-Identifier as a user-friendly web interface for the PAML programs; here the users can define the parameters and their respective values (Figure 2). For instance, the evolutionary model for calculation of site-rates, and the number of rate categories (normally 8 categories) for the analysis can be defined. Users still have an option to use their own control file that can be uploaded to the Bioportal.

Two types of files are used to calculate the site rates: 1) a multigene alignment in FASTA format with file extension '.fasta' or PAML format, and 2) a corresponding file containing a phylogenetic tree. The tree file should be generated with a suitable phylogenetic programs; the codeml and baseml programs are not recommended to reconstruct trees (see the PAML manual [30]). The tree topologies accepted are typically specified using the parenthesis

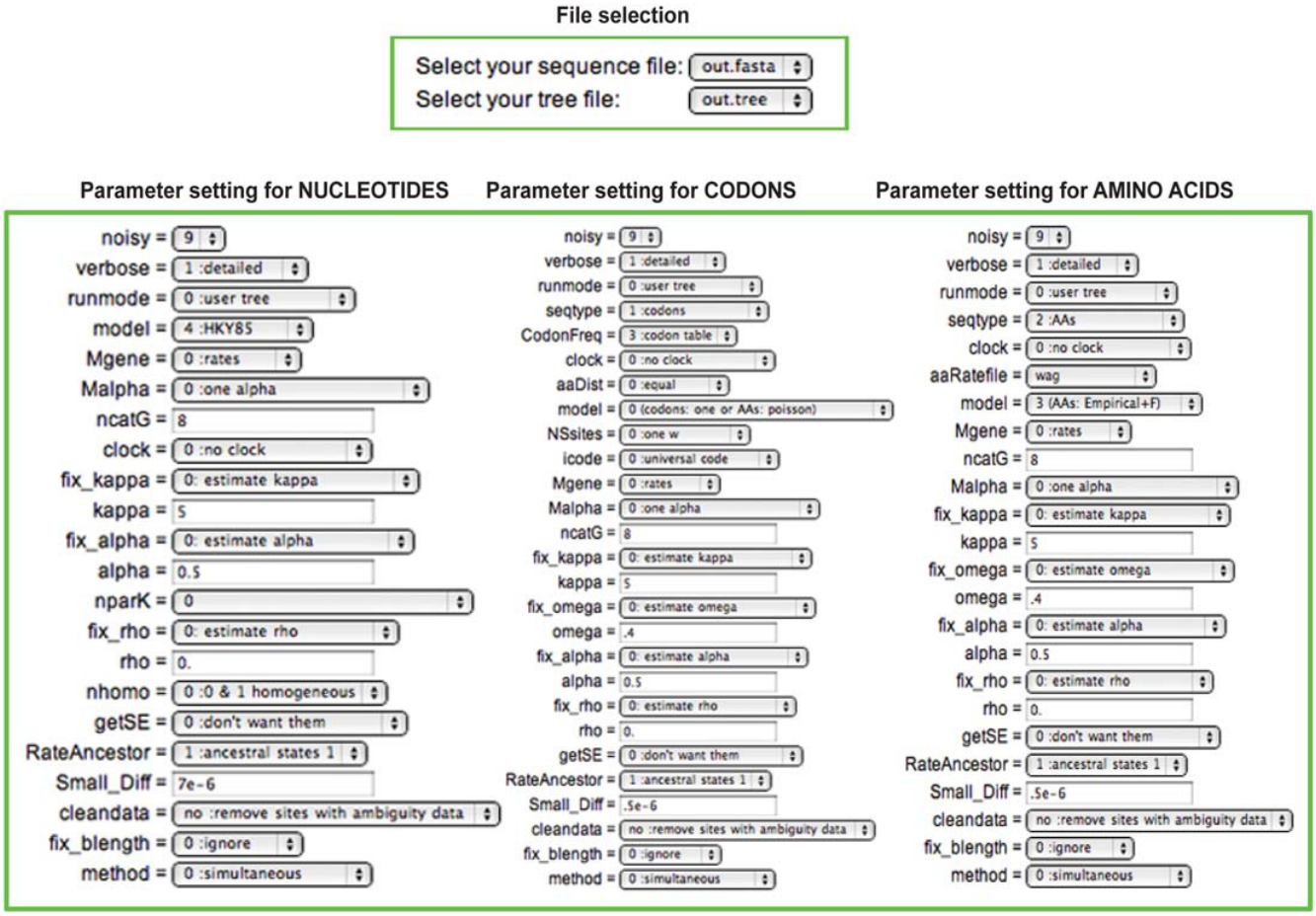

Figure 2

AIR-Identifier Web-Interface. AIR-Identifier web-interface on the Bioportal, where the user can select input files (i.e. sequence alignments and tree file containing phylogenetic trees) and parameters for three types of data; i.e. nucleotides, codons, and amino acids. The sequence files can be in FASTA or PAML format, while single or multiple trees in the tree file must be in Newick format and supplied in a single file.

notation such as the Newick tree format [31]. It should be noted that some widely used programs such as PAUP or MacClade [32,33] can produce tree files with limited compatibility, whereas other programs such as PHYLOBAYES v. 2.3 [34] or RAxML-VI-HPC [27] generate output files that are ready to use. Trees with or without branch length are accepted by AIR-Identifier.

It can often be difficult to decide which phylogeny should be used for estimating rates, especially when a dataset gives differing trees from different evolutionary models, parameters and tree searching algorithms. It has also been proposed that the selection of phylogeny can have a major impact on rate estimation [21]. For this reason we have constructed the AIR-Identifier to calculate site rates and rate categories from multiple phylogenetic trees.

The AIR-Identifier program produces two output files: 1) A rate file, which contains information about the evolutionary rate (rate category) for each site in the alignment (*rates* in Figure 1); 2) A html file (i.e. *rates.html* in Figure 1) that visually presents information about the rate pattern in the alignment and which allow the users to easily evaluate the importance of the various rate categories and the dispersal of the site rates along the alignment before sites are removed; the file also includes an graphical overview of the alignment where different rate categories have been color-coded.

Removing fast evolving sites

AIR-Remover is developed for the removal of fast evolving sites. The sites can be removed based on either site-rate or rate-category. The AIR-remover uses the alignment file and respective *rates* file obtained as output from AIR-Identifier. The users can then decide which of the rates and categories of fastest evolving sites should be removed. Multiple categories can be removed by using comma-separated numbers. The users can also remove sites that correspond to a fraction of the fastest evolving sites by defining a percentage of the total rate distribution; it is possible to remove e.g. the 5% fastest evolving sites (Figure 3). The AIR-Remover output files produces a main result file containing the ready to use alignment file (*outfile.ali* in Figure 1) and an html file (*outfile.html* in Figure 1) that enables the users to visualize the removed sites colored in red within their alignment.

Discussion and conclusion

The AIR package has been extensively used in recently published phylogenomic studies of deeply diverging eukaryote lineages [2,18]. In the study of Burki et al., 2008, a global eukaryote phylogeny was reconstructed from a dataset of 135 genes and 65 taxa, resulting in 73% bootstrap support for a single "megagroup" comprising nearly all photosynthetic lineages (including the supergroups Plantae, chromalveoalates and Rhizaria). When the fast evolving sites were identified and removed from the alignment with AIR, the same topology was recovered but with a substantially increased bootstrap support (97%) for the observed relationship. In the study of Minge et al. 2008, the evolutionary position of an enigmatic lineage named *Breviata* was in question using 78 genes and 38 taxa. The lineage was placed with strong bootstrap support as sister to the supergroup Amoebozoa, however statistical testing i.e. AU-test [35] of alternative placements in the eukaryote tree could not reject a sister relationship to another supergroup, the Excavata. Once fast evolving sites were removed using AIR the AU test could reject an affinity to the Excavata and additionally placed *Breviata* with the Amoebozoa with higher bootstrap support. Interestingly, the removal of additional fast evolving sites (altogether the 3 fastest rate categories) reduced the bootstrap support for the monophyly of *Breviata* and Amoebozoa, thus suggesting that the removal of too many categories or sites can reduce relevant phylogenetic information in the data. It demonstrates the need for detailed information about rates in the alignment provided by AIR.

The great need for efficient bioinformatic tools in reconstructing multi-gene alignments for phylogenomic inferences has over the last years been met by several new applications, such as Concatenator, IDEA, SCaFoS, IDEA and ASAP [36-40]. Several of these have overlapping functionalities with the AIR package, but the AIR is unique in combining key steps for constructing multi-gene alignments and evolutionary rate estimations. Most importantly AIR allows trimming of alignments according to the evolutionary rates and the users' preferences. Site rates estimation can be based on multiple phylogenies that account for uncertainties in the phylogeny. Several different criterions can be used for removing sites, either based on rate categories or site rates, which reduces the possibility of removing too many or few sites from the alignment. Monitoring of the site removal process is easy by using the colored alignment output files from the AIR.

In contrast to the vast majority of other programs, the AIR package is easily accessible on the web and does not require cumbersome installation on local computers. AIR is implemented on the Bioportal where users have their own file directories and can access several widely used programs in molecular evolution and ecology. The result files from the AIR programs can also be easily downloaded and applied in downstream analyses at other web-based bioinformatics services (such as http://www.phylo.org and http://bioweb2.pasteur.fr/). This makes the AIR package user-friendly and efficient. As AIR will process files on a large computer cluster, with the prospect of being linked to a larger grid infrastructure in

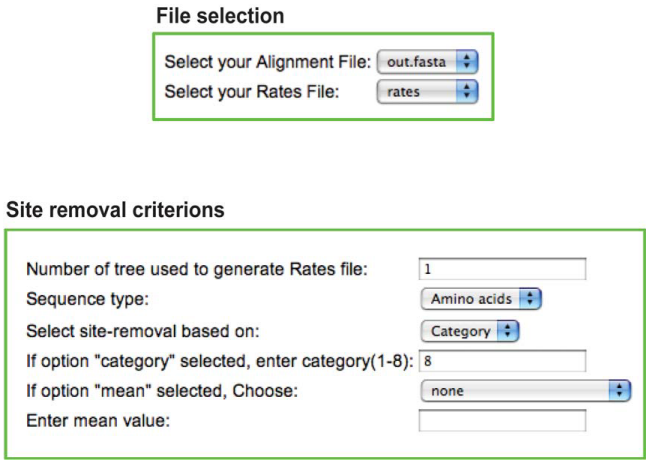

Figure 3
AIR-Remover Web-Interface. AIR-Identifier uses *rates* generated with AIR-Identifier (Figure 1) and the corresponding multigene alignment in PAML format. Sites can be removed on the basis of site rates or rate categories.

future, there is currently no restriction on the size of the input sequences.

Availability and requirements

Project name: AIR version 1.1

Project home page: http://www.bioportal.uio.no

Operating system(s): Platform independent

Programming language: SQL, Perl, Python and PHP

Other requirements: Apache webserver

License: GNU - GPL

Any restrictions to use by non-academics: AIR-Identifier uses PAML with license for academic use. Non-academic users still can use AIR-Appender and AIR-Remover at http://app3.titan.uio.no/biotools/. Test dataset for all programs of AIR is available at http://www.bioportal.uio.no/onlinemat/online_material.php.

Authors' contributions

SK conducted the programming of AIR-Appender, AIR-Identifier and AIR-remover, wrote the paper and implemented the applications on the Bioportal. ÅS contributed with programming of AIR-Appender. RO and FB tested the AIR programs and contributed with writing of the manuscript. PE contributed with programming and implementation of the AIR on the Bioportal. ÅS, PE, TR, BHM and AB programmed the Bioportal. KST funded and designed the project, supervised the process, wrote the first draft of the AIR paper. KST and AB initiated the Bioportal service, and KST is leading the development of the service. All authors read and approved the final manuscript.

Acknowledgements

We would like to thank Marianne Minge and Jon Bråte for valuable suggestions and testing of the AIR package. The Bioportal has been developed as collaboration between bioinformatics groups at USIT headed by Jostein Sundet and Hans Eide and a bioinformatics group in the KST lab. We thank Center of Technology at University of Oslo for maintenance of the TITAN clusters and Research Council of Norway for financing computers through AVIT and FUGE grants to a consortium headed by Kjetill S. Jakobsen at University of Oslo. This work is supported by University of Oslo start grant to KST and PhD for Surendra Kumar. The Bioportal service is financially supported by EMBIO, MLS and FUGE initiatives at University of Oslo.

References

1. Burki F, Pawlowski J: **Monophyly of Rhizaria and multigene phylogeny of unicellular bikonts.** *Mol Biol Evol* 2006, **23(10):**1922-1930.
2. Burki F, Shalchian-Tabrizi K, Pawlowski J: **Phylogenomics reveals a new 'megagroup' including most photosynthetic eukaryotes.** *Biol Lett* 2008, **4(4):**366-369.
3. Gadagkar SR, Rosenberg MS, Kumar S: **Inferring species phylogenies from multiple genes: concatenated sequence tree versus consensus gene tree.** *J Exp Zoolog B Mol Dev Evol* 2005, **304(1):**64-74.
4. Philippe H, Lartillot N, Brinkmann H: **Multigene analyses of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia.** *Mol Biol Evol* 2005, **22(5):**1246-1253.
5. Rodriguez-Ezpeleta N, Brinkmann H, Burger G, Roger AJ, Gray MW, Philippe H, Lang BF: **Toward resolving the eukaryotic tree: the phylogenetic positions of jakobids and cercozoans.** *Curr Biol* 2007, **17(16):**1420-1425.
6. Ruiz-Trillo I, Roger AJ, Burger G, Gray MW, Lang BF: **A phylogenomic investigation into the origin of metazoa.** *Mol Biol Evol* 2008, **25(4):**664-672.
7. Shalchian-Tabrizi K, Brate J, Logares R, Klaveness D, Berney C, Jakobsen KS: **Diversification of unicellular eukaryotes: cryptomonad colonizations of marine and fresh waters inferred from revised 18S rRNA phylogeny.** *Environ Microbiol* 2008, **10(10):**2635-2644.
8. Shalchian-Tabrizi K, Minge MA, Espelund M, Orr R, Ruden T, Jakobsen KS, Cavalier-Smith T: **Multigene phylogeny of choanozoa and the origin of animals.** *PLoS ONE* 2008, **3(5):**e2098.
9. Delsuc F, Brinkmann H, Philippe H: **Phylogenomics and the reconstruction of the tree of life.** *Nat Rev Genet* 2005, **6(5):**361-375.
10. Nikolaev SI, Berney C, Fahrni JF, Bolivar I, Polet S, Mylnikov AP, Aleshin VV, Petrov NB, Pawlowski J: **The twilight of Heliozoa and rise of Rhizaria, an emerging supergroup of amoeboid eukaryotes.** *Proc Natl Acad Sci USA* 2004, **101(21):**8066-8071.
11. Philippe H, Lopez P, Brinkmann H, Budin K, Germot A, Laurent J, Moreira D, Muller M, Le Guyader H: **Early-branching or fast-evolving eukaryotes? An answer based on slowly evolving positions.** *Proc Biol Sci* 2000, **267(1449):**1213-1221.
12. Burki F, Shalchian-Tabrizi K, Minge M, Skjaeveland A, Nikolaev SI, Jakobsen KS, Pawlowski J: **Phylogenomics reshuffles the eukaryotic supergroups.** *PLoS ONE* 2007, **2(8):**e790.
13. Shalchian-Tabrizi K, Kauserud H, Massana R, Klaveness D, Jakobsen KS: **Analysis of environmental 18S ribosomal RNA sequences reveals unknown diversity of the cosmopolitan phylum Telonemia.** *Protist* 2007, **158(2):**173-180.
14. Rodríguez-Ezpeleta N, Brinkmann H, Burey SC, Roure B, Burger G, Löffelhardt W, Bohnert HJ, Philippe H, Lang BF: **Monophyly of primary photosynthetic eukaryotes: green plants, red algae, and glaucophytes.** *Current Biology* 2005, **15(14):**1325-1330.
15. Keeling PJ: **Diversity and evolutionary history of plastids and their hosts.** *American Journal of Botany* 2004, **91:**1481-1493.

16. Dutilh BE, Huynen MA, Bruno WJ, Snel B: **The consistent phylogenetic signal in genome trees revealed by reducing the impact of noise.** *J Mol Evol* 2004, **58(5):**527-539.
17. Bapteste E, Brinkmann H, Lee JA, Moore DV, Sensen CW, Gordon P, Durufle L, Gaasterland T, Lopez P, Muller M, *et al.*: **The analysis of 100 genes supports the grouping of three highly divergent amoebae: Dictyostelium, Entamoeba, and Mastigamoeba.** *Proc Natl Acad Sci USA* 2002, **99(3):**1414-1419.
18. Minge AM, Silberman JD, Orr RJ, Cavalier-Smith T, Shalchian-Tabrizi K, Burki F, Skjaeveland A, Jakobsen KS: **Evolutionary position of breviate amoebae and the primary eukaryote divergence.** *Proc Biol Sci* 2009, **276(1657):**597-594.
19. Brinkmann H, Giezen M van der, Zhou Y, Poncelin de Raucourt G, Philippe H: **An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics.** *Syst Biol* 2005, **54(5):**743-757.
20. Pisani D: **Identifying and removing fast-evolving sites using compatibility analysis: an example from the Arthropoda.** *Syst Biol* 2004, **53(6):**978-989.
21. Rodriguez-Ezpeleta N, Brinkmann H, Roure B, Lartillot N, Lang BF, Philippe H: **Detecting and overcoming systematic errors in genome-scale phylogenies.** *Syst Biol* 2007, **56(3):**389-399.
22. Brinkmann H, Philippe H: **Archaea sister group of Bacteria? Indications from tree reconstruction artifacts in ancient phylogenies.** *Mol Biol Evol* 1999, **16(6):**817-825.
23. Burleigh JG, Mathews S: **Phylogenetic signal in nucleotide data from seed plants: implications for resolving the seed plant tree of life.** *American Journal of Botany* 2004, **91(10):**1599-1613.
24. Katoh K, Kuma K, Toh H, Miyata T: **MAFFT version 5: improvement in accuracy of multiple sequence alignment.** *Nucleic Acids Res* 2005, **33(2):**511-518.
25. Ronquist F, Huelsenbeck JP: **MrBayes 3: Bayesian phylogenetic inference under mixed models.** *Bioinformatics* 2003, **19(12):**1572-1574.
26. Jobb G, von Haeseler A, Strimmer K: **TREEFINDER: a powerful graphical analysis environment for molecular phylogenetics.** *BMC Evol Biol* 2004, **4:**18.
27. Stamatakis A: **RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models.** *Bioinformatics* 2006, **22(21):**2688-2690.
28. Yang Z: **PAML: a program package for phylogenetic analysis by maximum likelihood.** *Comput Appl Biosci* 1997, **13(5):**555-556.
29. Yang Z: **PAML 4: phylogenetic analysis by maximum likelihood.** *Mol Biol Evol* 2007, **24(8):**1586-1591.
30. Yang Z: 2007 [http://abacus.gene.ucl.ac.uk/software/pamlDOC.pdf].
31. **The Newick tree format** [http://evolution.genetics.washington.edu/phylip/newicktree.html]
32. Maddison WP, Maddison DR: **MacClade 4: Analysis of Phylogeny and Character Evolution.** Sinauer Associates, Sunderland, MA; 2000.
33. Swofford DL: **PAUP*: Phylogenetic Analysis Using Parsimony. (* and other methods).** In *ver. 4.0b10 edn* Sinauer Associates, Inc. Publishers, Sunderland, MA; 2003.
34. Lartillot N, Philippe H: **Computing Bayes factors using thermodynamic integration.** *Syst Biol* 2006, **55(2):**195-207.
35. Shimodaira H: **An approximately unbiased test of phylogenetic tree selection.** *Syst Biol* 2002, **51(3):**492-508.
36. Pina-Martins F, Paulo OS: **Cancatenator: Sequence Data Matrices Handling Made easy.** *Molecular Ecology Resource* 2008, **8(6):**1254-1255.
37. Egan A, Mahurkar A, Crabtree J, Badger JH, Carlton JM, Silva JC: **IDEA: Interactive Display for Evolutionary Analyses.** *BMC Bioinformatics* 2008, **9(1):**524.
38. Roure B, Rodriguez-Ezpeleta N, Philippe H: **SCaFoS: a tool for selection, concatenation and fusion of sequences for phylogenomics.** *BMC Evol Biol* 2007, **7(Suppl 1):**S2.
39. Felsenstein J: **PHYLIP (Phylogeny Inference Package) version 3.6.** Distributed by he author. Department of Genome Sciences, University of Washington, Seattle; 2005.
40. Sarkar IN, Egan MG, Coruzzi G, Lee EK, DeSalle R: **Automated simultaneous analysis phylogenetics (ASAP): an enabling tool for phlyogenomics.** *BMC Bioinformatics* 2008, **9:**103.